# Notes on STAT-331: Applied Linear Models

*Unversity of Waterloo*

David Duan

Last Updated: May 2, 2021 (V1.0)

# Contents

# Chapter 1. Simple Linear Regression

## Section 1. Overview

**1.1.** Suppose we are given a set of data points $\{(x_1, y_1), \ldots, (x_n, y_n)\}$.

- How do we characterize the relationship between $x$ and $y$?
- How do we predict $y$ given $x$?
- How does the mean of $y$ change when $x$ increases by $a$?

We can answer questions like these with **simple linear regression** (SLR):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Intuitively, we are assuming that there exists some underlying linear relationship between the **covariate** $x_i$ and the **outcome** $y_i$, where the **regression coefficients** $\beta_0$ and $\beta_1$ are unknown. The **error term** $\varepsilon_i$ captures the difference between the actual value of $y_i$ and our prediction $\beta_0 + \beta_1 x_i$.

**1.2.** The model above is "simple" because there is only one explanatory variable $x$. Suppose now each sample $x_i$ has three covariates $x_{i1}, x_{i2}$, and $x_{i3}$. We generalize SLR to **multiple linear regression** (MLR), where each covariate $x_{ij}$ has a corresponding $\beta_j$ parameter:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i,$$

The meaning of $y_i$ and $\varepsilon_i$ remain the same; we just have more covariates to work with.

**1.3.** This course will focus on developing regression models in the following aspects:

- theoretically/mathematically: derive estimators;
- practically: how to fit these models in R;
- how to choose and compare a model, i.e., which covariates to include;
- how to evaluate the appropriateness of the model and assumptions.

## Section 2. Simple Linear Regression

**1.4. Remark:** We make the following assumptions (acronym: LINE):

- **L**inearity: there exists a linear relationship between $x$ and $y$.
- **I**ndependence: the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are independent.
- **N**ormality: the error terms have mean 0.
- **E**qual variance (aka **homoskedasticity**): all error terms share the same variance $\sigma^2$.

**1.5. Definition:** The general form of simple linear regression is given by

$$\boxed{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2).}$$

- $\beta_0, \beta_1, \sigma^2$: fixed, *unknown* parameters.
- $\varepsilon_i$: *unobserved* random error term.
- $y_i, x_i$ are observed data (we treat $x_i$ as fixed in this course).

Equivalently, we can write

$$\boxed{y_i \overset{\text{indep}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2).}$$

Note here $y_i$'s are independent but no longer have the same distribution because they have different means (depending on $x_i$).

**1.6. Example:** How to interpret $\beta_0$ and $\beta_1$? We make the following observations:

1. $\mathbb{E}[y_i \mid x_i] = \beta_0 + \beta_1 x_i$.
2. $\mathbb{E}[y_i \mid x_i = 0] = \beta_0$.
3. $\mathbb{E}[y_i \mid x_i = x^*] = \beta_0 + \beta_1 x^*$.
4. $\mathbb{E}[y_i \mid x_i = x^* + 1] = \beta_0 + \beta_1(x^* + 1) = \beta_0 + \beta_1 x^* + \beta_1$.
5. $\mathbb{E}[y_i \mid x_i = x^* + 1] - \mathbb{E}[y_i \mid x_i = x^*] = \beta_1$.

Therefore,

- By observation 2, $\beta_0$ is the average outcome when $x_0 = 0$.
- By observation 5, $\beta_1$ is the expected/average change in $y$ when $x$ moves by 1 unit.

## Section 3. SLR: Estimation

**1.7. Theorem:** *The LS estimators for $\beta_0$ and $\beta_1$ are given by*

$$\boxed{\begin{aligned}\hat{\beta}_0^{LS} &= \bar{y} - \hat{\beta}_1^{LS}\bar{x} \\ \hat{\beta}_1^{LS} &= \frac{(\sum_i x_i y_i) - n\bar{x}\bar{y}}{(\sum_i x_i^2) - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}}\end{aligned}}$$

*Proof.* The goal is to choose $\beta_0$ and $\beta_1$ that minimizes the sum of squared errors given by

$$S(\beta_0, \beta_1) := \sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Differentiate, set the partial derivatives to 0, and solve for $\beta_0$ and $\beta_1$:

$$\begin{aligned}\frac{\partial S(\delta_0, \delta_1)}{\partial \beta_0} &= \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1) \\ \frac{\partial S(\delta_0, \delta_1)}{\partial \beta_1} &= \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i)\end{aligned}$$

$$\begin{aligned}\text{(Set) } 0 &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \\ &= \left(\sum_{i=1}^{n} y_i\right) - n\beta_0 - \left(\beta_1 \sum_{i=1}^{n} x_i\right) \\ \implies \beta_0 &= \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) - \beta_1 \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \bar{y} - \beta_1 \bar{x}\end{aligned}$$

$$\begin{aligned}\text{(Set) } 0 &= \sum_{i=1}^{n} (y_i x_i - \beta_0 x_i - \beta_1 x_i^2) \\ &= \left(\sum_{i=1}^{n} y_i x_i\right) - \left(\beta_0 \sum_{i=1}^{n} x_i\right) - \left(\beta_1 \sum_{i=1}^{n} x_i^2\right) \\ &= \left(\sum_{i=1}^{n} y_i x_i\right) - (\bar{y} - \beta_1 \bar{x}) n\bar{x} - \left(\beta_1 \sum_{i=1}^{n} x_i^2\right) \qquad \text{plug in previous result} \\ &= \left(\sum_{i=1}^{n} y_i x_i\right) - n\bar{y}\bar{x} + \beta_1 n\bar{x}^2 - \left(\beta_1 \sum_{i=1}^{n} x_i^2\right) \\ &= \left(\sum_{i=1}^{n} y_i x_i\right) - n\bar{y}\bar{x} + \beta_1 \left(n\bar{x}^2 - \sum_{i=1}^{n} x_i^2\right) \\ \implies \beta_1 &= \frac{(\sum_{i=1}^{n} y_i x_i) - n\bar{y}\bar{x}}{(\sum_{i=1}^{n} x_i^2) - n\bar{x}^2} = \frac{S_{xy}}{S_{xx}} \qquad \text{See Proposition 1.24}\end{aligned}$$

□

**1.8. Theorem:** *The ML estimators for $\beta_0$ and $\beta_1$ coincide with the LS estimators.*

*Proof.* The **joint likelihood function** of $Y_1, \ldots, Y_n$ with $Y_i \overset{\text{indep}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$ is

$$\begin{aligned} L(\beta_0, \beta_1, \sigma) &= \prod_{i=1}^{n} f(y_i; \beta_0 + \beta_1 x_i, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right). \end{aligned}$$

The **log-likelihood function** is given by

$$\ell(\beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Maximizing the log-likelihood is equivalent to solving the following system of equations:

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_0} &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\right) = 0 \\ \frac{\partial \ell}{\partial \beta_1} &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\right) x_i = 0 \\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)\right)^2 = 0 \end{aligned}$$

Observe solving the first two equations is equivalent to minimizing the sum of squares! In other words, the ML estimators $\hat{\beta}_0^{\text{ML}}, \hat{\beta}_1^{\text{ML}}$ coincide with the LS estimators $\hat{\beta}_0^{\text{LS}}, \hat{\beta}_1^{\text{LS}}$. Therefore, we will remove the superscripts and simply call them $\hat{\beta}_0$ and $\hat{\beta}_1$. □

**1.9. Definition:** The **fitted values** $\hat{y}_i$ and the **residuals** $e_i$ are given by

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
- $e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$.

Note the residuals $e_i$ and the errors $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$ are not the same thing.

**1.10. Remark:** Solving the third equation, we obtain the ML estimator for $\sigma^2$:

$$\hat{\sigma}^2_{\text{ML}} = \frac{\sum_{i=1}^{n} e_i^2}{n}.$$

This is slightly different from the unbiased estimator for $\sigma^2$ (notice the $n-2$ in the denominator):

$$\boxed{\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}.}$$

This difference often doesn't matter when $n \geq 50$.

## Section 4. SLR: Inference

**1.11. Theorem:** *The estimator $\hat{\beta}_1$ follows the Normal distribution with parameters*

$$\boxed{\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right).}$$

*Proof.* Recall $y_i \overset{\text{indep}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$. Let us rewrite $\hat{\beta}_1$ as

$$\begin{aligned}\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} &= \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \\ &= \frac{\sum_i y_i (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} =: \sum_{i=1}^n w_i y_i, \qquad\qquad w_i := \frac{(x_i - \bar{x})}{\sum_i (x_i - \bar{x})}.\end{aligned}$$

Since we assumed that $x_i$'s are fixed, the variables $w_i$'s are fixed wrt the $y_i$'s. Thus, $\hat{\beta}_1$ is a linear combination independent Normal random variables $y_1, \ldots, y_n$. Moreover, all $y_i$'s share the same variance (homoskedasticity). By the Fact above, $\beta_i$ follows the normal distribution with parameters

$$\hat{\beta}_1 \sim N\left(\sum_{i=1}^n w_i(\beta_0 + \beta_1 x_i), \sigma^2 \sum_{i=1}^n w_i^2\right).$$

It remains to simplify the parameters.

$$\begin{aligned}\mathbb{E}[\hat{\beta}_1] &= \sum_{i=1}^n w_i \left(\beta_0 + \beta_1 x_i\right) \\ &= \sum_{i=1}^n \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \left(\beta_0 + \beta_1 x_i\right) \\ &= \beta_0 \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} + \beta_1 \frac{\sum_{i=1}^n x_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \\ &= 0 + \beta_1 \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \beta_1 \qquad\qquad \sum_{i=1}^n (x_i - \bar{x}) = 0\end{aligned}$$

$$\begin{aligned}\text{Var}[\hat{\beta}_1] &= \sigma^2 \sum_{i=1}^n w_i^2 \\ &= \sigma^2 \sum_{i=1}^n \left[\frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_j - \bar{x})^2}\right]^2 \\ &= \sigma^2 \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\left[\sum_{j=1}^n (x_j - \bar{x})^2\right]^2} \\ &= \sigma^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left[\sum_{j=1}^n (x_j - \bar{x})^2\right]^2} = \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}\end{aligned}$$

□

**1.12. Theorem:** *The estimator $\hat{\beta}_0$ follows the Normal distribution with parameters*

$$\boxed{\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\right).}$$

*Proof.*

$$\begin{aligned}
\mathbb{E}\left[\hat{\beta}_0\right] &= \mathbb{E}\left[\bar{y} - \hat{\beta}_1\bar{x}\right] = \mathbb{E}[\bar{y}] - \mathbb{E}[\hat{\beta}_1\bar{x}] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_i^n y_i\right] - \bar{x}\mathbb{E}\left[\hat{\beta}_1\right] \\
&= \frac{1}{n}\left(\sum_{i=1}^n \mathbb{E}[y_i]\right) - \bar{x}\beta_1 && \mathbb{E}[\hat{\beta}_1] = \beta_1 \\
&= \frac{1}{n}\left(\sum_{i=1}^n (\beta_0 + \beta_1 x_i)\right) - \bar{x}\beta_1 && \mathbb{E}[y_1] = \beta_0 + \beta_1 x_i \\
&= \beta_0 + \beta_1\bar{x} - \bar{x}\beta_1 = \beta_0
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}\hat{\beta}_0 &= \mathrm{Var}\left(\bar{y} - \hat{\beta}_1\bar{x}\right) \\
&= \mathrm{Var}(\bar{y}) - 2\mathrm{Cov}\left(\bar{y}, \hat{\beta}_1\bar{x}\right) + \mathrm{Var}\left(\hat{\beta}_1\bar{x}\right) \\
&= \frac{\sigma^2}{n} - 2\bar{x}\mathrm{Cov}\left(\bar{y}, \hat{\beta}_1\right) + \bar{x}^2\mathrm{Var}\hat{\beta}_1 && \text{See (1.25)} \\
&= \frac{\sigma^2}{n} - 2\bar{x}\mathrm{Cov}\left(\bar{y}, \hat{\beta}_1\right) + \bar{x}^2\frac{\sigma^2}{S_{xx}} \\
&= \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] - 2\bar{x}\mathrm{Cov}\left(\bar{y}, \hat{\beta}_1\right).
\end{aligned}$$

It remains to show that $\mathrm{Cov}(\bar{y}, \hat{\beta}_1) = 0$.

$$\begin{aligned}
\mathrm{Cov}\left(\bar{y}, \hat{\beta}_1\right) &= \mathrm{Cov}\left(\frac{1}{n}\sum_{i=1}^n y_i, \frac{\sum_i (x_i - \bar{x})\, y_i}{\sum_i (x_i - \bar{x})^2}\right) \\
&= \frac{1}{n\sum_i (x_i - \bar{x})^2}\mathrm{Cov}\left(\sum_i Y_i, \sum_i (x_i - \bar{x})\, Y_i\right) \\
&= \frac{1}{n\sum_i (x_i - \bar{x})^2}\sum_{i,j}\mathrm{Cov}\left(y_i, (x_i - \bar{x})\, y_j\right) && \mathrm{Cov}\left(y_i, (x_i - \bar{x})\, y_j\right) \propto \delta_{i,j} \\
&= \frac{1}{n\sum_i (x_i - \bar{x})^2}\sum_i (x_i - \bar{x})\,\mathrm{Var}(y_i) && \mathrm{Cov}(y_i, y_i) = \mathrm{Var}(y_i) \\
&= \frac{\sigma^2}{n\sum_i (x_i - \bar{x})^2}\sum_i (x_i - \bar{x}) && \mathrm{Var}(y_i) = \sigma^2 \\
&= 0 && \sum_i (x_i - \bar{x}) = 0
\end{aligned}$$

□

## Section 5. SLR: Confidence Interval

**1.13.** Let us derive a 95% confidence interval for $\beta_1$. Recall that

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \implies Z := \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1). \tag{1.1}$$

Suppose $\sigma$ is known. Then

$$\begin{aligned}
0.95 &= P(-1.96 \le Z \le 1.96) \\
&= P\left(-1.96 \le \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \le 1.96\right) \\
&= P\left(-1.96\frac{\sigma}{\sqrt{S_{xx}}} \le \hat{\beta}_1 - \beta_1 \le 1.96\frac{\sigma}{\sqrt{S_{xx}}}\right) \\
&= P\left(-1.96\frac{\sigma}{\sqrt{S_{xx}}} \le \beta_1 - \hat{\beta}_1 \le 1.96\frac{\sigma}{\sqrt{S_{xx}}}\right) \\
&= P\left(\hat{\beta}_1 - 1.96\frac{\sigma}{\sqrt{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + 1.96\frac{\sigma}{\sqrt{S_{xx}}}\right)
\end{aligned}$$

Thus, a 95% CI for $\beta_1$ is

$$\hat{\beta}_1 \pm 1.96\frac{\sigma}{\sqrt{S_{xx}}}.$$

In practice, $\sigma^2$ is often unknown. We can estimate it using the unbiased estimator $\hat{\sigma}^2$.

**1.14. Definition:** The **standard error** $\mathrm{SE}(\hat{\beta}_1)$ is an estimator of $\hat{\beta}_1$'s standard deviation:

$$\boxed{\mathrm{SE}(\hat{\beta}_1) := \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}.}$$

**1.15. Theorem:** *The confidence interval of $\hat{\beta}_1$ is given by*

$$\boxed{\hat{\beta}_1 \pm t_{1-\alpha/2,n-2}\mathrm{SE}(\hat{\beta}_1).}$$

*Proof.* Replacing $\sigma^2$ by $\hat{\sigma}^2$ in (1.1) gives the $t$-distributed pivotal quantity

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{\mathrm{SE}(\hat{\beta}_1)} \sim t_{(n-2)}.$$

A $100(1-\alpha)\%$ confidence interval is given by

$$1 - \alpha = \Pr\left(-q \le \frac{\hat{\beta}_1 - \beta_1}{\mathrm{SE}(\hat{\beta}_1)} \le q\right) = \Pr\left(\hat{\beta}_1 - q\frac{\hat{\sigma}}{\mathrm{SE}(\hat{\beta}_1)} \le \beta_1 \le \hat{\beta}_1 + q\frac{\hat{\sigma}}{\mathrm{SE}(\hat{\beta}_1)}\right)$$

Thus, a 95% CI for $\beta_1$ is

$$\hat{\beta}_1 \pm t_{1-\alpha/2,n-2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

where $t_{1-\alpha/2,n-2}$ is can be found with `qt(p = alpha/2, df = n-2)` in R. □

## Section 6. SLR: Hypotheses Testing

**1.16.** Suppose we want to test a null hypothesis $H_0 : \beta_1 = \theta_0$ against some alternative hypothesis $H_1 : \beta_1 \neq \theta_0$. For SLR, we often set

- $H_0 : \beta_1 = 0$: no linear relationship;
- $H_1 : \beta_1 \neq 0$: two-sided alternative.

The goal is to characterize how much evidence we have against $H_0$, or how "extreme" our data are relative to $H_0$. We can test the null hypothesis with the $t$-statistic

$$T := \frac{\hat{\beta}_1 - \theta_0}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{(n-2)}.$$

Assuming $H_0$ is true, what's the probability to have some as extreme or more than what we observe?

$$\Pr(|T| \geq |t_{\text{obs}}|) = 2\Pr(T \geq |t_{\text{obs}}|) = 2[1 - \Pr(T \leq -|t_{\text{obs}}|)].$$

We typically reject the null hypothesis at the 5% level, i.e., reject $H_0$ if $p < 0.05$. Would we accept $H_0$ if $p > 0.05$? **No, we simply would not have enough evidence to reject.**

**1.17. Remark:** Does this mean $\Pr(\beta_1 = 0) = p$? No. Instead, it means under the null hypothesis, i.e., assuming $\beta_1 = 0$, the probability of a test statistic as extreme as the one observed is equal to $p$. That's why a small $p$-value is evidence against the null, since it would be particularly "rare" under the null.

**1.18. Remark:** Note that a $100(1-\alpha)\%$ CI (e.g., 0.95) corresponds with a hypothesis test with a $100\alpha\%$ significance level (e.g., 0.05), i.e., we will derive a similar conclusion. In particular, if we reject $H_0$ at the 0.05-level (i.e., when the $p$-value is less than 0.05), then the 95% CI will not contain the value of 0.

## Section 7. SLR: Estimation of Mean Response

**1.19. Theorem:** *Given new $x_0$, the estimated mean response is given by*

$$\hat{\mu}_0 = \beta_0 + \beta_1 x_0 \sim N\left(\mu_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right).$$

*Proof.* The mean response for an arbitrary $x_0$ is given by

$$\hat{\mu}_0 = \mathbb{E}[y \mid x_0] = \hat{\beta}_0 + \hat{\beta}_1 x_0 = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_0 = \bar{y} + \hat{\beta}_1 (x_0 - \bar{x}).$$

The estimate of mean response is unbiased:

$$\mathbb{E}[\hat{\mu}_0] = \mathbb{E}[\hat{\beta}_0 + \hat{\beta}_1 x_0] = \mathbb{E}[\hat{\beta}_0] + \mathbb{E}[\hat{\beta}_1] x_0 = \beta_0 + \beta_1 x_0 =: \mu_0.$$

The variance is given by

$$\begin{aligned}
\operatorname{Var}[\hat{\mu}_0] &= \operatorname{Var}\left[\hat{\beta}_0 + \hat{\beta}_1 x_0\right] \\
&= \operatorname{Var}\left[\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) + \hat{\beta}_1 x_0\right] \\
&= \operatorname{Var}\left[\bar{y} + \hat{\beta}_1 (x_0 - \bar{x})\right] \\
&= \operatorname{Var}\left[\left(\sum_{i=1}^n \frac{1}{n} y_i\right) + \left(\sum_{i=1}^n \frac{(x_i - \bar{x})}{S_{xx}} y_i\right)(x_0 - \bar{x})\right] \\
&= \operatorname{Var}\left[\sum_{i=1}^n \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}}\right) y_i\right] \qquad \star \\
&= \sum_{i=1}^n \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}}\right)^2 \sigma^2 \\
&= \sigma^2 \sum_{i=1}^n \left(\frac{1}{n^2} + \frac{(x_i - \bar{x})^2 (x_0 - \bar{x})^2}{S_{xx}^2} + 2\frac{1}{n}\frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}}\right) \\
&= \sigma^2 \left(\sum_{i=1}^n \frac{1}{n^2} + \sum_{i=1}^n \frac{(x_i - \bar{x})^2 (x_0 - \bar{x})^2}{S_{xx}^2} + 2\sum_{i=1}^n \frac{1}{n}\frac{(x_i - \bar{x})(x_0 - \bar{x})}{S_{xx}}\right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}^2} S_{xx} + 2\frac{1}{n}\frac{(x_0 - \bar{x})}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x})\right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)
\end{aligned}$$

Note in the 5th line of the derivation of variance (labeled $\star$), we see that $\hat{\mu}_0$ is a linear combination of Normal random variables $y_i$, so $\mu_0$ is also Normal:

$$\hat{\mu}_0 \sim N\left(\mu_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right).$$

□

**1.20. Note:** From above, we know that

$$\frac{\hat{\mu}_0 - \mu_0}{\sigma\sqrt{\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)}} \sim N(0,1) \quad \text{and} \quad \frac{\hat{\mu}_0 - \mu_0}{\hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)}} \sim t_{n-2}.$$

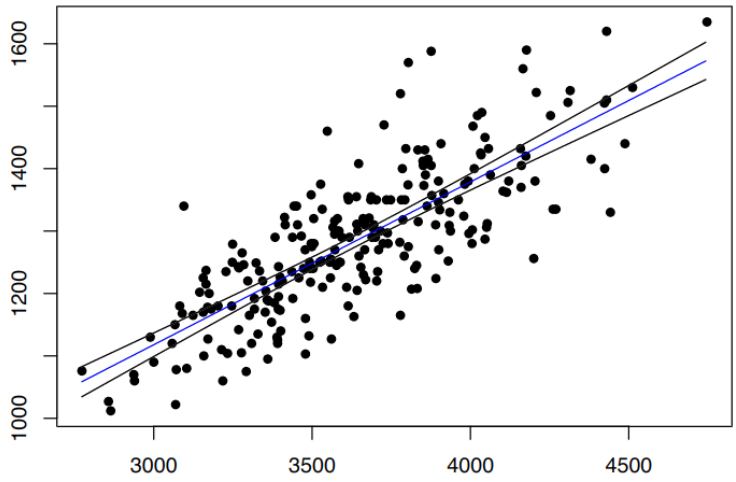
Thus, a 95% CI is given by

$$0.95 = P\left(-t_{n-2,1-\frac{\alpha}{2}} \leq \frac{\hat{\mu}_0 - \mu_0}{\hat{\sigma}\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)^{1/2}} \leq t_{n-2,1-\frac{\alpha}{2}}\right)$$

In general, a $100(1-\alpha)\%$ CI is given by

$$\boxed{\hat{\mu}_0 \pm t_{n-2,1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)}.}$$

Note that the CIs get bigger as $x \to \infty$ and $x \to -\infty$ as we have fewer data points there.




Note that many points fall outside of the CI. What if we don't just care about the mean, but also the predictions? That is, even if we got the mean absolutely perfect, the new points wouldn't fall directly on the line!

## Section 8. SLR: Prediction of a Single Response

**1.21. Note:** Suppose we want to predict the response for a new covariate value:

$$y_{\text{new}} = \beta_0 + \beta_1 x_{\text{new}} + \varepsilon_{\text{new}}.$$

Define the predicted value $\hat{y}_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}$ and prediction error $\hat{y}_{\text{new}} - y_{\text{new}}$. Let's quantify the prediction error.

$$\begin{aligned} E\left[\hat{y}_{\text{new}} - y_{\text{new}}\right] &= E\left[\left(\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}\right) - \left(\beta_0 + \beta_1 x_{\text{new}} + \varepsilon_{\text{new}}\right)\right] \\ &= \beta_0 + \beta_1 x_{\text{new}} - \left(\beta_0 + \beta_1 x_{\text{new}}\right) = 0 \end{aligned}$$

Note that $\hat{y}_{\text{new}}$ and $y_{\text{new}}$ are independent, because the former is a linear combination of the known $y_i$'s while the latter has nothing to do with those. Moreover, $\hat{y}_{\text{new}}$ is Normal as $y_i$'s are Normal.

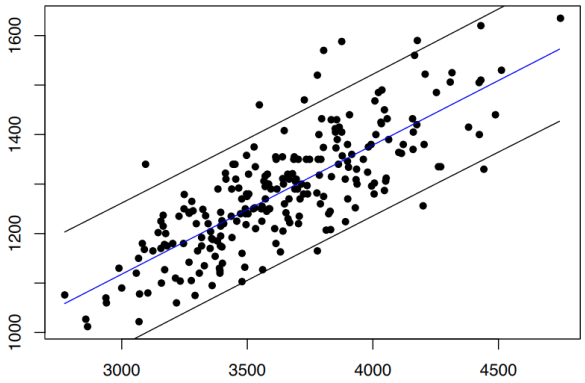
$$\begin{aligned} \text{Var}\left[\hat{y}_{\text{new}} - y_{\text{new}}\right] &= \text{Var}\left[\left(\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}\right) - y_{\text{new}}\right] \\ &= \text{Var}\left[\left(\hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}\right)\right] + \text{Var}\left[y_{\text{new}}\right] \\ &= \left[\sigma^2\left(\frac{1}{n} + \frac{\left(x_{\text{new}} - \bar{x}\right)^2}{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}\right)\right] + \left[\sigma^2\right] \\ &= \sigma^2\left(1 + \frac{1}{n} + \frac{\left(x_{\text{new}} - \bar{x}\right)^2}{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}\right) \end{aligned}$$

Using the same approach as above, we have

$$\frac{\hat{y}_{\text{new}} - y_{\text{new}}}{\sigma\sqrt{\left(1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)}} \sim N(0,1) \quad \text{and} \quad \frac{\hat{y}_{\text{new}} - y_{\text{new}}}{\hat{\sigma}\sqrt{\left(1 + \frac{1}{n} + \frac{(x_{\text{new}} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)}} \sim t_{n-2}.$$

Thus, a $100(1-\alpha)\%$ **prediction interval** is given by

$$\boxed{\hat{y}_{\text{new}} \pm t_{n-2,\left(1-\frac{\alpha}{2}\right)}\hat{\sigma}\sqrt{\left(1 + \frac{1}{n} + \frac{\left(x_{\text{new}} - \bar{x}\right)^2}{\sum_{i=1}^n \left(x_i - \bar{x}\right)^2}\right)}.}$$




Note the margin of error of PI is much wider compared to the previous CI.

## Section 9. Appendix

**1.22. Definition:** Let $\bar{x}, \bar{y}$ denote the mean of $x$'s and $y$'s. Define

$$\boxed{\begin{aligned} S_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 \\ S_{xy} &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \\ S_{yy} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \end{aligned}}$$

**1.23. Lemma:** *Let $\bar{x}$ be the mean of $\{x_1, \ldots, x_n\}$. Then*

$$\boxed{\sum_{i=1}^{n}(x_i - \bar{x}) = 0.}$$

*Proof.* Observe that

$$\begin{aligned} \sum_{i=1}^{n}(x_i - \bar{x}) &= \left[\sum_{i=1}^{n} x_i\right] - n\bar{x} \\ &= \left[\sum_{i=1}^{n} x_i\right] - n\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \left[\sum_{i=1}^{n} x_i\right] - \left[\sum_{i=1}^{n} x_i\right] = 0. \end{aligned}$$

□

**1.24. Proposition:** *We have the following equalities for $S_{xx}$ and $S_{xy}$:*

$$\boxed{\begin{aligned} S_{xx} &= \left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2 \\ S_{xy} &= \left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}. \end{aligned}}$$

*Proof.* Observe that

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) \\ &= \left[\sum_{i=1}^{n} x_i(x_i - \bar{x})\right] - \left[\bar{x}\sum_{i=1}^{n}(x_i - \bar{x})\right] && \bar{x} \text{ does not depend on } i \\ &= \sum_{i=1}^{n} x_i(x_i - \bar{x}) && \sum_{i=1}^{n}(x_i - \bar{x}) = 0 \end{aligned}$$

$$= \left[\sum_{i=1}^{n} x_i^2\right] - \left[\bar{x}\sum_{i=1}^{n} x_i\right] \qquad \bar{x} \text{ does not depend on } i$$

$$= \left[\sum_{i=1}^{n} x_i^2\right] - \bar{x}(n\bar{x}) \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \implies \sum_{i=1}^{n} x_i = n\bar{x}$$

$$= \left[\sum_{i=1}^{n} x_i^2\right] - n\bar{x}^2$$

The second property can be derived using a similar approach (Exercise). □

**1.25. Lemma:**

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}\left(X_i\right) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

# Chapter 2. Multiple Linear Regression

## Chapter Highlight

Let $\mathbf{y} = (y_1, \ldots, y_n)$ be a random vector. Then

$$\boxed{\begin{aligned} \mathbb{E}[\mathbf{y}] &= [\mathbb{E}[y_i]]_{1\le i\le n} \in \mathbb{R}^{n\times 1}, \\ \mathrm{Var}[\mathbf{y}] &= [\mathrm{Cov}(y_i, y_j)]_{1\le i,j\le n} \in \mathbb{R}^{n\times n}. \end{aligned}}$$

In particular, $V$ is *symmetric* and *positive semidefinite*.

Properties of $\mathbb{E}$, Var, and Cov:

$$\boxed{\begin{aligned} \mathbb{E}[\mathbf{a}^T\mathbf{y} + c] &= a^T\boldsymbol{\mu} + c \in \mathbb{R} \\ \mathrm{Cov}(\mathbf{a}^T\mathbf{y} + c, \mathbf{b}^T\mathbf{y} + d) &= \mathbf{a}^T\mathbf{V}\mathbf{b} \in \mathbb{R} \\ \mathbb{E}[\mathbf{A}\mathbf{y} + \mathbf{b}] &= \mathbf{A}\mathbb{E}[\mathbf{y}] + \mathbf{b} = \mathbf{A}\boldsymbol{\mu} \in \mathbb{R}^k \\ \mathrm{Var}(\mathbf{A}\mathbf{y} + \mathbf{b}) &= \mathbf{A}\mathrm{Var}(\mathbf{y})\mathbf{A}^T = \mathbf{A}\mathbf{V}\mathbf{A}^T \in \mathbb{R}^{k\times k}. \end{aligned}}$$

## Section 1. Review: Linear Algebra and Calculus

**2.1. Remark:** It's often a lot easier to understand formulas intuitively in higher-dimensional spaces once you know their sizes/dimensions (sanity check!). I will try to label the dimensions of vectors and spaces as much as possible. **Warning**: There will be abuse of notations for random variables, e.g., I will label a random vector $\mathbf{x}$ with three elements as $\mathbf{x} \in \mathbb{R}^3$.

**2.2. Note:** We briefly review some facts about matrices. Let $\mathbf{A}, \mathbf{B}, \mathbf{C}$ be matrices.

- $[\mathbf{C}^T]_{ij} = [\mathbf{C}]_{ij}$.
- $\mathbf{C}$ is **symmetric** if $\mathbf{C}^T = \mathbf{C}$.
- $(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T$.
- If a square matrix $\mathbf{B}$ is non-singular, then $\mathbf{B}\mathbf{B}^{-1} = \mathbf{B}^{-1}\mathbf{B} = \mathbf{I}$.
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ is both are non-singular square matrices.
- $(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$.
- $\mathrm{tr}(\mathbf{A}) = \sum_j^n a_{jj}$ for square matrix $\mathbf{A}$.
- $\mathrm{tr}(c\mathbf{A} + \mathbf{B}) = c \cdot \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$.
- $\mathrm{tr}(\mathbf{A}^T) = \mathrm{tr}(\mathbf{A})$.
- $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$.

**2.3. Note:** We briefly review some matrix calculus.

- Let $\mathbf{y} = (y_1, \ldots, y_k) \in \mathbb{R}^k$ and $f : \mathbb{R}^k \to \mathbb{R}$ be a function of $\mathbf{y}$. Then
$$\frac{\partial f}{\partial \mathbf{y}} = \begin{bmatrix} \dfrac{\partial f}{\partial y_1} \\ \vdots \\ \dfrac{\partial f}{\partial y_k} \end{bmatrix} \in \mathbb{R}^{k\times 1}$$
- If $z = \mathbf{a}^T\mathbf{y} \in \mathbb{R}$ where $\mathbf{a} = (a_1, \ldots, a_k) \in \mathbb{R}^k$ is a column vector, then
$$\frac{\partial z}{\partial \mathbf{y}} = \mathbf{a} \in \mathbb{R}^{k\times 1}.$$
- If $z = \mathbf{y}^T A\mathbf{y} \in \mathbb{R}$ where $A \in \mathbb{R}^{k\times k}$, then
$$\frac{\partial z}{\partial \mathbf{y}} = (A + A^T)\mathbf{y} \in \mathbb{R}^{k\times 1}.$$
In particular, if $A$ is symmetric, then
$$\frac{\partial z}{\partial \mathbf{y}} = 2A\mathbf{y} \in \mathbb{R}^{k\times 1}.$$

## Section 2. Random Vectors

**2.4. Definition:** A **random vector** is a vector of random variables. Let $\mathbf{y} = (y_1, \ldots, y_n)$ be a random vector. The **mean** of $\mathbf{y}$ is

$$\boxed{\mathbb{E}[\mathbf{y}] = \begin{bmatrix} \mathbb{E}[y_1] \\ \vdots \\ \mathbb{E}[y_n] \end{bmatrix} \in \mathbb{R}^{n \times 1}.}$$

The **variance** of $\mathbf{y}$ is given by the **covariance matrix**:

$$\begin{aligned} \mathrm{Var}(\mathbf{y}) = \mathbf{V} &= \mathbb{E}[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T] \\ &= \begin{bmatrix} \mathrm{Var}(y_1) & \mathrm{Cov}(y_1, y_2) & \cdots & \mathrm{Cov}(y_1, y_n) \\ \mathrm{Cov}(y_2, y_1) & \mathrm{Var}(y_2) & \cdots & \mathrm{Cov}(y_1, y_n) \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(y_n, y_1) & \mathrm{Cov}(y_n, y_2) & \cdots & \mathrm{Var}(y_n) \end{bmatrix} \in \mathbb{R}^{n \times n} \end{aligned}$$

In particular,

$$\boxed{\mathbf{V}_{ij} = \mathrm{Cov}(y_i, y_j).}$$

**2.5. Proposition:** *Let $V = \mathrm{Var}(\mathbf{y})$ be the covariance matrix of $\mathbf{y}$.*

- $\mathbf{V}$ *is* ***symmetric****, i.e.,* $\mathbf{V}_{ij} = \mathbf{V}_{ji}$.
- $\mathbf{V}$ *is* ***positive semidefinite****, i.e.,* $\forall \mathbf{a} \in \mathbb{R}^n$, $\mathbf{a}^T \mathbf{V} \mathbf{a} \geq 0$.

*Proof.* The first claim follows from the fact that the Cov operator is symmetric. For the second claim, observe that

$$\mathbf{a}^T \mathbf{V} \mathbf{a} = \mathbf{a}^T \mathbb{E}[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T] \mathbf{a} = \mathbb{E}[\mathbf{a}^T (\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})^T \mathbf{a}] \overset{\star}{=} \mathbb{E}[((\mathbf{y} - \boldsymbol{\mu})^T \mathbf{a})^2] \geq 0$$

where $\star$ follows from the fact that $\mathbf{a}^T(\mathbf{y} - \boldsymbol{\mu})$ and $(\mathbf{y} - \boldsymbol{\mu})^T \mathbf{a}$ are scalars. □

**2.6. Note:** Recall the following facts. Let $a_i, b_i, c, d \in \mathbb{R}$ be constants, $y_i$ be random variables, and $z = \sum_{i=1}^n a_i y_i + c$, $u = \sum_{i=1}^n b_i y_i + d$ be linear combinations of $y_i$'s; $z, u \in \mathbb{R}$. Then

$$\boxed{\begin{aligned} \mathbb{E}[z] &= \sum_{i=1}^n a_i \mathbb{E}[y_i] + c \in \mathbb{R} \\ \mathrm{Cov}(z, u) &= \sum_{i=1}^n \sum_{j=1}^n a_i b_j \mathrm{Cov}(y_i, y_j) \in \mathbb{R}. \end{aligned}}$$

Equivalently in matrix notation, $z = \mathbf{a}^T \mathbf{y} + c \in \mathbb{R}$, $u = \mathbf{b}^T \mathbf{y} + d \in \mathbb{R}$, then

$$\boxed{\begin{aligned} \mathbb{E}[\mathbf{a}^T \mathbf{y} + c] &= a^T \boldsymbol{\mu} + c \in \mathbb{R} \\ \mathrm{Cov}(\mathbf{a}^T \mathbf{y} + c, \mathbf{b}^T \mathbf{y} + d) &= \mathbf{a}^T \mathbf{V} \mathbf{b} \in \mathbb{R} \end{aligned}}$$

where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{y}]$ and $\mathbf{V} = \mathrm{Var}(\mathbf{y})$. We now consider their multivariate counterparts.

**2.7. Note:** Consider a random vector $\mathbf{z} = (z_1, \ldots, z_k)^T$ of $k$ linear combinations of random $\mathbf{y}$:

$$\begin{aligned} z_1 &= a_{11}y_1 + a_{12}y_2 + \cdots + a_{1n}y_n \\ z_2 &= a_{21}y_1 + a_{22}y_2 + \cdots + a_{2n}y_n \\ &\vdots \\ z_k &= a_{k1}y_1 + a_{k2}y_2 + \cdots + a_{kn}y_n \end{aligned}$$

We can equivalently write $\mathbf{z} = \mathbf{A}\mathbf{y} \in \mathbb{R}^k$ for $\mathbf{A} \in \mathbb{R}^{k\times n}, [\mathbf{A}]_{ij} = a_{ij}$. Then

$$\begin{aligned} \mathbb{E}[\mathbf{A}\mathbf{y}] &= \mathbf{A}\mathbb{E}[\mathbf{y}] = \mathbf{A}\boldsymbol{\mu} \in \mathbb{R}^k \\ \mathrm{Var}(\mathbf{A}\mathbf{y}) &= \mathbb{E}[(\mathbf{A}\mathbf{y} - \mathbb{E}[\mathbf{A}\mathbf{y}])(\mathbf{A}\mathbf{y} - \mathbb{E}[\mathbf{A}\mathbf{y}])^T] \\ &= \mathbb{E}[\mathbf{A}(\mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{A}(\mathbf{y} - \mathbb{E}[\mathbf{y}]))^T] \\ &= \mathbb{E}[\mathbf{A}(\mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^T A^T] \\ &= \mathbf{A}\mathbb{E}[(\mathbf{y} - \mathbb{E}[\mathbf{y}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^T]A^T \\ &= \mathbf{A}\mathrm{Var}(\mathbf{y})\mathbf{A}^T \\ &= \mathbf{A}\mathbf{V}\mathbf{A}^T \in \mathbb{R}^{k\times k} \end{aligned}$$

where $\boldsymbol{\mu} = \mathbb{E}[\mathbf{y}]$ and $\mathbf{V} = \mathrm{Var}(\mathbf{y})$. In other words, you can pull out a matrix of constants from the expectation and the variance operator much like what you do with vectors. We summarize this result into the following proposition (with an extra bias term $\mathbf{b}$).

**2.8. Theorem:** *Let $y \in \mathbb{R}^n$ and $\mathbf{A} \in \mathbb{R}^{k\times n}$. Then*

$$\boxed{\begin{aligned} \mathbb{E}[\mathbf{A}\mathbf{y} + \mathbf{b}] &= \mathbf{A}\mathbb{E}[\mathbf{y}] + \mathbf{b} = \mathbf{A}\boldsymbol{\mu} \in \mathbb{R}^k \\ \mathrm{Var}(\mathbf{A}\mathbf{y} + \mathbf{b}) &= \mathbf{A}\mathrm{Var}(\mathbf{y})\mathbf{A}^T = \mathbf{A}\mathbf{V}\mathbf{A}^T \in \mathbb{R}^{k\times k}. \end{aligned}}$$

## Section 3. Multivariate Normal Distribution

**2.9. Definition:** A vector $\mathbf{y}$ has a **multivariate normal distribution** $\text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if its density function has the form

$$\boxed{f(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\}}$$

where $\mathbb{E}[\mathbf{y}] = \boldsymbol{\mu}$ and $\text{Var}(\mathbf{y}) = \boldsymbol{\Sigma}$.

**2.10. Example:** Let $\mathbf{z} = (z_1, \ldots, z_n) \in \mathbb{R}^n$ be a random vector of iid standard normal random variables, i.e., $z_i \overset{\text{iid}}{\sim} N(0,1)$ for all $i$'s. Then for any $\mathbf{A} \in \mathbb{R}^{k\times n}$,

$$\mathbf{y} = \mathbf{A}\mathbf{z} + \boldsymbol{\mu} \in \mathbb{R}^k \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\mathbb{E}[\mathbf{y}] = \boldsymbol{\mu}$ and $\text{Var}(\mathbf{y}) = \boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T$.

**2.11. Proposition:** *Some nice properties of MVN:*

- *Linearity: If* $\mathbf{u} = \mathbf{C}\mathbf{y} + \mathbf{d}$*, then*

$$\mathbf{u} \sim MVN(\mathbf{C}\boldsymbol{\mu} + \mathbf{d}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^T).$$

- *Marginal distribution: If* $\tilde{\mathbf{y}} = (y_1, \ldots, y_m)^T \subseteq \mathbf{y}$ *is a vector subset of* $\mathbf{y}$*, then* $\tilde{\mathbf{y}}$ *is MVN-distributed. In particular, every* $y_j \in \mathbf{y} \sim N(\mu_j, \Sigma_{jj})$ *is normally distributed.*
- *Conditional distribution: If* $\mathbf{u} = (\mathbf{y}_1^T, \mathbf{y}_2^T)^T \sim MVN$ *(i.e., breaking a column vector* $\mathbf{u}$ *into two pieces), then* $\mathbf{y}_1^T \mid \mathbf{y}_2^T$ *is MVN-distributed.*
- *Independence: If* $\Sigma_{ij} = 0$*, then* $y_i$ *and* $y_j$ *are independent.*
  - *Note this only holds for Normal variables: independence* $\implies$ Cov $= 0$ *always holds, but the other direction is generally false (but true for MVN).*

# Section 4. Multiple Linear Regression

**2.12. Definition:** The **multiple linear regression** (MLR) model is given by

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_P x_{iP} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right) \\ &\iff \\ y_i \mid x_i &\overset{\text{indep}}{\sim} N\left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \sigma^2\right) \end{aligned}$$

- $(x_i, y_i)$: the $i$th observation, but now we have $P$ covariates instead of just 1.
- The meaning of other symbols remain the same.
- Assume $p < n$, or we have more variates than observations.

**2.13. (Cont'd):** Equivalently, we can write

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1P} \\ 1 & x_{21} & x_{22} & \dots & x_{2P} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nP} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_P \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

or more compactly,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \text{MVN}(\mathbf{0}, \sigma^2\mathbf{I}) \iff \mathbf{y} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I),$$

where

- $\mathbf{X}$ is the **design matrix**,
- $\boldsymbol{\beta}$ is the **parameter vector**,
- $\boldsymbol{\varepsilon}$ is the **error vector**, and
- $\mathbf{y}$ is the **response vector**.

Note $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$, where each row represents a sample and each column correspond to a covariate.

**2.14. Note:** How to interpret the regression coefficients:

- $\beta_0$ is the mean outcome when all variates are set to 0.
- $\beta_j$ represents the difference in mean outcome for a 1-unit change in the $j$th variate $x_j$, *holding other covariates* fixed.

## Section 5. MLR: Least Squares Estimation

**2.15. Theorem:** *The LS estimators for $\boldsymbol{\beta}$ is given by*

$$\boxed{\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}.}$$

*Proof.* We wish to minimize the sum of squares:

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \qquad\qquad \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}, \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} \in \mathbb{R} \end{aligned}$$

Taking its derivative with respect to the vector $\boldsymbol{\beta}$, we get

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{y} + (\mathbf{X}^T\mathbf{X} + \mathbf{X}^T\mathbf{X})\boldsymbol{\beta}$$

Note the last term comes from the derivative of the quadratic form

$$\frac{\partial}{\partial \mathbf{y}}(\mathbf{y}^T\mathbf{A}\mathbf{y}) = (\mathbf{A} + \mathbf{A}^T)\mathbf{y}.$$

Now set the derivative to 0,

$$\begin{aligned} -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} &= 0 \\ \left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta} &= \mathbf{X}^T\mathbf{y} \\ \implies \hat{\boldsymbol{\beta}} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}. \end{aligned}$$

Note the inverse exists iff $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}$ has full column rank (i.e., the columns of $\mathbf{X}$ are linearly independent). Thus, we require $n \geq p+1$. □

**2.16. Remark:** Maximum likelihood gives the same estimators. We omit the derivation.

**2.17. Theorem:** *The LS estimator $\hat{\boldsymbol{\beta}}$ has the following properties:*

$$\boxed{\begin{aligned} \hat{\boldsymbol{\beta}} &\sim \mathrm{MVN}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \\ \hat{\beta}_j &\sim N(\beta_j, \sigma^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}) \end{aligned}}$$

*In particular,* $\mathbf{V} = (\mathbf{X}^T\mathbf{X})^{-1}$.

*Proof.*

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbb{E}[\mathbf{y}] && \text{Linearity of } \mathbb{E} \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}) \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

$$\begin{aligned}
\text{Var}[\hat{\boldsymbol{\beta}}] &= \text{Var}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right] \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\,\text{Var}[\mathbf{y}]\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\,\text{Var}[\mathbf{y}]\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\
&= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\sigma^2\mathbf{I}\right)\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\
&= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\
&= \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}
\end{aligned}$$

Finally, since $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is a linear combination of $\mathbf{y} \sim$ MVN, $\hat{\boldsymbol{\beta}}$ is also MVN. The second statement follows from the *marginal distribution* property of MVN. □

**2.18. Theorem:** *The unbiased estimator of $\sigma^2$ is given by*

$$\boxed{\hat{\sigma}^2 = \frac{1}{n-(p+1)}\mathbf{e}^T\mathbf{e}.}$$

*Proof.* Omitted. □

**2.19. Lemma:** $(\mathbf{X}^{\mathbf{X}})^{-1}$ *is symmetric.*

*Proof.* $[(\mathbf{X}^T\mathbf{X})^{-1}]^T = [(\mathbf{X}^T\mathbf{X})^T]^{-1} = (\mathbf{X}^T\mathbf{X})^{-1}$. □

## Section 6. MLR: Fitted Values and Residuals

**2.20. Definition:** Let $\hat{\boldsymbol{\beta}}$ be the LS estimator of $\boldsymbol{\beta}$. The **fitted values** is defined as

$$\begin{aligned}\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right] \\ &= \left[\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right]\mathbf{y} =: \mathbf{H}\mathbf{y}.\end{aligned}$$

The matrix $\boxed{\mathbf{H} := \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T}$ is called the **Hat matrix**, as applying $\mathbf{H}$ to $\mathbf{y}$ yields $\hat{\mathbf{y}}$ ("adding a hat to $\mathbf{y}$"). You should be familiar with the property $\boxed{\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}.}$

**2.21. Proposition:** *The Hat matrix $\mathbf{H}$ is symmetric and idempotent (i.e., a projection matrix).*

*Proof.* $\mathbf{HH} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{H}.$ □

**2.22. Corollary:** $\mathbf{I} - \mathbf{H}$ *is symmetric and idempotent (i.e., a projection matrix).*

*Proof.* $(\mathbf{I}-\mathbf{H}) = \mathbf{I}^T - \mathbf{H}^T = (\mathbf{I}-\mathbf{H})$. Also, $(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) = \mathbf{II} - 2\mathbf{H} + \mathbf{HH} = \mathbf{I} - \mathbf{H}$. □

**2.23. Proposition:** $\boxed{\mathbb{E}[\hat{\mathbf{y}}] = \mathbf{X}\boldsymbol{\beta}, \mathrm{Var}[\hat{\mathbf{y}}] = \sigma^2\mathbf{H}.}$

*Proof.*

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{y}}] &= \mathbb{E}[\mathbf{H}\mathbf{y}] \\ &= \mathbf{H}\mathbb{E}[\mathbf{y}] \\ &= \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{X}\boldsymbol{\beta}\end{aligned}$$

$$\begin{aligned}\mathrm{Var}[\hat{\mathbf{y}}] &= \mathrm{Var}[\mathbf{H}\mathbf{y}] \\ &= \mathbf{H}\,\mathrm{Var}[\mathbf{y}]\mathbf{H}^T \\ &= \mathbf{H}\sigma^2\mathbf{I}\mathbf{H} \\ &= \sigma^2\mathbf{H}\end{aligned}$$

□

**2.24. Definition:** Define **residuals** as $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{H}\mathbf{y} = \boxed{(\mathbf{I}-\mathbf{H})\mathbf{y}}$.

**2.25. Remark:** Note that the sum of residuals is zero:

$$\begin{bmatrix}\sum_{i=1}^n e_i \cdot 1 \\ \sum_{i=1}^n x_{i1}e_i \\ \vdots \\ \sum_{i=1}^n x_{ip}e_i\end{bmatrix} = \mathbf{X}^T\mathbf{e} = \mathbf{X}^T(\mathbf{y}-\mathbf{H}\mathbf{y}) = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{y} = \mathbf{0}.$$

**2.26. Proposition:** $\boxed{\mathbb{E}[\mathbf{e}] = \mathbf{0}, \mathrm{Var}[\mathbf{e}] = \sigma^2(\mathbf{I}-\mathbf{H}).}$

*Proof.*

$$\begin{aligned}\mathbb{E}[\mathbf{e}] &= \mathbb{E}[(\mathbf{I}-\mathbf{H})\mathbf{y}] \\ &= (\mathbf{I}-\mathbf{H})\mathbb{E}[\mathbf{y}] \\ &= \left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)(\mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}\end{aligned}$$

$$\begin{aligned}\mathrm{Var}[\mathbf{e}] &= \mathrm{Var}[(\mathbf{I}-\mathbf{H})\mathbf{y}] \\ &= (\mathbf{I}-\mathbf{H})\,\mathrm{Var}[\mathbf{y}](\mathbf{I}-\mathbf{H})^T \\ &= \sigma^2(\mathbf{I}-\mathbf{H})\end{aligned}$$

□

**2.27. Note:** Recall $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ are both linear combinations of $\mathbf{y}$. Since $\mathbf{y}$ is MVN-distributed, the vector obtained by stacking rows of $\hat{\boldsymbol{\beta}}$ on top of the rows of $\mathbf{e}$,

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \\ (\mathbf{I} - \mathbf{H})\mathbf{y} \end{pmatrix} = \begin{pmatrix} (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ \mathbf{I} - \mathbf{H} \end{pmatrix} \mathbf{y}$$

is also MVN-distributed. We now explore the relationship between $\hat{\boldsymbol{\beta}}$ and $\mathbf{e}$.

**2.28. Theorem:**

$$\boxed{\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \mathbf{e} \end{bmatrix} \sim \text{MVN}\left( \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \sigma^2 \begin{bmatrix} (\mathbf{X}^T\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{I} - \mathbf{H}) \end{bmatrix} \right).}$$

*Moreover,*

1. $\hat{\boldsymbol{\beta}} \sim \text{MVN}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)$
2. $\mathbf{e} \sim \text{MVN}\left(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H})\right)$, *and*
3. $\hat{\boldsymbol{\beta}}$ *and* $\mathbf{e}$ *are independent.*

*Proof.* We already proved Claim 1. For Claim 2 and 3, it suffices to prove that the vector has the claim distribution, as $\boldsymbol{\Sigma}_{22} = \text{Var}[\mathbf{e}]$ and $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21} = 0$ indicates variables $\hat{\boldsymbol{\beta}}$ and $\mathbf{e}$ are independent.

$$\begin{aligned}
\mathbb{E}[\mathbf{e}] &= \mathbb{E}[(\mathbf{I} - \mathbf{H})\mathbf{y}] \\
&= (\mathbf{I} - \mathbf{H})E[\mathbf{y}] \\
&= (\mathbf{I} - \mathbf{H})\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{H}\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{0}
\end{aligned}$$

$$\begin{aligned}
\text{Var}\begin{bmatrix} \hat{\beta} \\ \mathbf{e} \end{bmatrix} &= \text{Var}\left[\begin{pmatrix} \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \\ (\mathbf{I} - \mathbf{H}) \end{pmatrix} \mathbf{y}\right] \\
&= \begin{pmatrix} \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \\ (\mathbf{I} - \mathbf{H}) \end{pmatrix} \text{Var}[\mathbf{y}] \begin{pmatrix} \left(\mathbf{X}^T\mathbf{X}\right)^{-1}X^T \\ (\mathbf{I} - \mathbf{H}) \end{pmatrix}^T \\
&= \sigma^2 \begin{pmatrix} \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \\ (\mathbf{I} - \mathbf{H}) \end{pmatrix} \begin{pmatrix} \left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T & (\mathbf{I} - \mathbf{H})^T \end{pmatrix} \qquad \text{Var}[\mathbf{y}] = \sigma^2\mathbf{I} \\
&= \sigma^2 \begin{pmatrix} \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \\ (\mathbf{I} - \mathbf{H}) \end{pmatrix} \begin{pmatrix} \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} & (\mathbf{I} - \mathbf{H}) \end{pmatrix} \\
&= \sigma^2 \begin{pmatrix} \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} & \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{H}) \\ (\mathbf{I} - \mathbf{H})\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} & (\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) \end{pmatrix} \\
&= \sigma^2 \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}
\end{aligned}$$

Now

$$\mathbf{A} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

$$\begin{aligned}\mathbf{B} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{I}-\mathbf{H}) \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right) \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{0} = \mathbf{C}^T\end{aligned}$$

$$\begin{aligned}\mathbf{D} &= (\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})^T \\ &= \left(\mathbf{I}\mathbf{I}^T - \mathbf{I}\mathbf{H}^T - \mathbf{H}\mathbf{I}^T + \mathbf{H}\mathbf{H}^T\right) \\ &= (\mathbf{I}-2\mathbf{H}-\mathbf{H}) = (\mathbf{I}-\mathbf{H})\end{aligned}$$

□

## Section 7. MLR: Deriving $t$-Statistic*

**2.29. Remark** (Review on eigen-decomposition)**:** Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ with $n$ linearly independent eigenvectors $q_i$, $1 \le i \le n$. Then $\mathbf{A}$ can be factorized as $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$ where $\mathbf{Q} \in \mathbb{R}^{n\times n}$, whose $i$th column is the eigenvector $q_i$ of $\mathbf{A}$, and $\mathbf{\Lambda}$ is the diagonal matrix whose diagonal elements are the corresponding eigenvalues, $\Lambda_{ii} = \lambda_i$. Only diagonalizable matrices can be factorized in this way.

**2.30. Note:** So far, we have proved that

$$\hat{\beta} \sim N\left(\beta, \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right) \Longrightarrow \hat{\beta}_j \sim N\left(\beta_j, \sigma^2 V_{jj}\right).$$

If we can show that

1. $\frac{1}{\sigma^2}\mathbf{e}^T\mathbf{e} \sim \chi^2_{n-(p+1)}$, and
2. it is independent of $\hat{\boldsymbol{\beta}}$,

then we obtain the following $t$-statistic, which can be used for constructing confidence intervals and hypothesis testing. Note we did something similar for SLR but we didn't give a mathematical proof back then.

$$\frac{\frac{\hat{\beta}_j-\beta_j}{\sqrt{\sigma^2 V_{jj}}}}{\sqrt{\left(\frac{1}{\sigma^2}\mathbf{e}^T\mathbf{e}\right)/(n-(p+1))}} = \frac{\hat{\beta}_j-\beta_j}{\sqrt{\hat{\sigma}^2 V_{jj}}} \sim t_{n-(p+1)}.$$

Intuitively, we have $n-(p+1)$ degrees of freedom because we have $n$ data points and we are trying to estimate $p+1$ regression parameters. We now show the math behind this.

**2.31. (Cont'd):** The second condition is easy. Since $\mathbf{e}$ is independent of $\hat{\boldsymbol{\beta}}$, $\frac{1}{\sigma^2}\mathbf{e}^T\mathbf{e}$ as a function of $\mathbf{e}$ is also independent of $\hat{\boldsymbol{\beta}}$. Now for the first condition, recall that $\mathbf{e} = (\mathbf{I}-\mathbf{H})\mathbf{y}$. Consider the eigen-decomposition $\mathbf{I}-\mathbf{H} = \mathbf{\Gamma}^T\mathbf{D}\mathbf{\Gamma}$ where $\mathbf{\Gamma}^{-1} = \mathbf{\Gamma}^T$ and

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_n \end{bmatrix}$$

is the diagonal matrix whose diagonal contains the eigenvalues of $(\mathbf{I}-\mathbf{H})$. Define $\tilde{\mathbf{e}} = \mathbf{\Gamma}\mathbf{e}$. Then

$$\begin{aligned} \mathbb{E}[\tilde{\mathbf{e}}] &= \mathbb{E}[\mathbf{\Gamma}\mathbf{e}] = \mathbf{\Gamma}\mathbb{E}[\mathbf{e}] = \mathbf{0} \\ \mathrm{Var}[\tilde{\mathbf{e}}] &= \mathrm{Var}[\mathbf{\Gamma}\mathbf{e}] \\ &= \mathbf{\Gamma}\,\mathrm{Var}[\mathbf{e}]\mathbf{\Gamma}^T \qquad\qquad \mathrm{Var}[\mathbf{A}\mathbf{e}] = \mathbf{A}\mathrm{Var}[\mathbf{e}]\mathbf{A}^T \\ &= \sigma^2\mathbf{\Gamma}(\mathbf{I}-\mathbf{H})\mathbf{\Gamma}^T \qquad\qquad \mathrm{Var}[\mathbf{e}] = \mathbf{I}-\mathbf{H} \\ &= \sigma^2\mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{D}\mathbf{\Gamma}\right)\mathbf{\Gamma}^T \\ &= \sigma^2\mathbf{D} \end{aligned}$$

Thus, $\tilde{\mathbf{e}} \sim \mathrm{MVN}(\mathbf{0}, \sigma^2\mathbf{D})$ as $\mathbf{e} \sim \mathrm{MVN}$, which implies

$$\tilde{e}_i \overset{\text{indep}}{\sim} N(0, \sigma^2[\mathbf{D}]_{ii}) = N(0, \sigma^2\lambda_i^2).$$

**2.32. Remark** (Review on $\chi^2$ Distributions)**:** For standard normal rvs $Z_i \overset{\text{iid}}{\sim} N(0,1)$,

$$X = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n.$$

**2.33. (Cont'd):** Next,

$$\tilde{\mathbf{e}}^T\tilde{\mathbf{e}} = (\mathbf{\Gamma e})^T(\mathbf{\Gamma e}) = \mathbf{e}^T\mathbf{\Gamma}^T\mathbf{\Gamma e} = \mathbf{e}^T\mathbf{e},$$

so we can write

$$\frac{1}{\sigma^2}\mathbf{e}^T\mathbf{e} = \frac{1}{\sigma^2}\tilde{\mathbf{e}}^T\tilde{\mathbf{e}} = \sum_{i=1}^{n}\left(\frac{\tilde{e}_i}{\sigma}\right)^2 = \sum_{i=1}^{n} Z_i^2, \qquad Z_i \overset{\text{indep}}{\sim} N(0,\lambda_i^2).$$

Thus, $\frac{1}{\sigma^2}\mathbf{e}^T\mathbf{e}$ is a sum of squared independent normally distributed rvs. To show

$$\frac{1}{\sigma^2}\mathbf{e}^T\mathbf{e} \sim \chi^2_{(n-(p+1))},$$

we need to show that $n-(p+1)$ of the eigenvalues $\lambda_j$'s are equal to 1, and all others are equal to 0. (Indeed, if $\lambda_j = 0$, then $Z_j \sim N(0,0)$ becomes a constant.) We know that $(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) = \mathbf{I}-\mathbf{H}$. This gives

$$\begin{aligned}(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) &= \mathbf{I}-\mathbf{H}\\ (\mathbf{\Gamma}^T\mathbf{D\Gamma})(\mathbf{\Gamma}^T\mathbf{D\Gamma}) &= (\mathbf{\Gamma}^T\mathbf{D\Gamma})\\ \mathbf{\Gamma}^T\mathbf{DD\Gamma} &= \mathbf{\Gamma}^T\mathbf{D\Gamma},\end{aligned}$$

i.e., $\mathbf{DD} = \mathbf{D}$ and thus $\lambda_j^2 = \lambda_j$. Thus all $\lambda_j$ are either 0 or 1. Next,

$$\begin{aligned}\sum_j \lambda_j = \operatorname{tr}(\mathbf{D}) &= \operatorname{tr}\left(\mathbf{D\Gamma\Gamma}^T\right) && \text{trace is similarity-invariant}\\ &= \operatorname{tr}\left(\mathbf{\Gamma}^T\mathbf{D\Gamma}\right) && \text{invariant under cyclic permutation}\\ &= \operatorname{tr}(\mathbf{I}-\mathbf{H})\\ &= \operatorname{tr}(\mathbf{I}) - \operatorname{tr}(\mathbf{H}) && \text{trace is linear}\\ &= n - \operatorname{tr}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\\ &= n - \operatorname{tr}(\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}) && \text{invariant under cyclic permutation}\\ &= n - \operatorname{tr}(\mathbf{I}_{p+1}) && \mathbf{X}\in\mathbb{R}^{n\times(p+1)} \implies \mathbf{X}^T\mathbf{X}\in\mathbb{R}^{(p+1)(p+1)}\\ &= n-(p+1)\end{aligned}$$

This concludes our proof.

**2.34. Note:** This entire section is optional. The only thing you need to remember is that

$$\boxed{\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 V_{jj}}} \sim t_{n-(p+1)}.}$$

Moreover, the standard error of $\hat{\beta}_j$ is given by

$$\boxed{\operatorname{SE}(\hat{\beta}_j) = \hat{\sigma}\sqrt{V_{jj}}.}$$

## Section 8. MLR: Hypothesis Testing

**2.35. Note:** Suppose we want to test a null hypothesis $H_0 : \beta_j = \theta_0$ against some alternative hypothesis $H_1 : \beta_j \neq \theta_0$. Our goal is to characterize how much evidence we have against $H_0$, or more intuitively, how *extreme* are our data relative to $H_0$. Under $H_0$ (i.e., if $H_0$ holds), then

$$T := \frac{\hat{\beta}_j - \theta_0}{\hat{\sigma}\sqrt{V_{jj}}} \sim t_{n-p-1}.$$

Below we discuss two approaches for hypothesis testing.

**2.36. (Cont'd):** First, we can compute the $p$-value and compare it against $\alpha$.

1. Given observed value
$$T_{\text{obs}} := \frac{\hat{\beta}_j - \theta_0}{\hat{\sigma}\sqrt{V_{jj}}} \sim t_{n-p-1},$$
2. Compute the $p$-value $p = \Pr(|T| \geq |T_{\text{obs}}|) = 2\Pr(T \geq T_{\text{obs}})$ given by
`p <- 2 * pt(T_obs, df = n-p-1, lower.tail=FALSE).`
Note the `pt` call gives you the $p$-value against a one-sided alternative.
3. If $p < \alpha$, reject $H_0$ (at $\alpha$).

**2.37. (Cont'd):** Alternatively, we can compute the quantile, known as the **critical value**, of the test statistic $T$ that gives a $p$-value of $\alpha$, then compare our observed value with this threshold.

1. Given observed value
$$T_{\text{obs}} := \frac{\hat{\beta}_j - \theta_0}{\hat{\sigma}\sqrt{V_{jj}}} \sim t_{n-p-1},$$
2. Compute the threshold by
`q <- qt(p = alpha/2, df=n-p-1).`
3. If $|T_{\text{obs}}| < t_{n-p-1,1-\alpha/2} =$ `q`, reject $H_0$ (at $\alpha$).

**2.38. Theorem:** *A* $(100-\alpha)\%$ *CI for* $\beta_j$ *is*

$$\boxed{\hat{\beta}_j \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{V_{jj}}.}$$

*Proof.* Omitted. □

**2.39. Note:** We can never guarantee that any single CI contains the true value. However, as we repeatedly construct CIs, about $(100-\alpha)\%$ of them will contain the true value.

## Section 9. MLR: Estimating Mean Response

**2.40.** For an arbitrary vector of covariates $\mathbf{x}_0 = [1, x_{01}, x_{02}, \ldots, x_{0p}]$, the mean response is

$$\mu_0 = \mathbb{E}[\mathbf{y}_0 \mid \mathbf{x}_0] = \mathbf{x}_0\boldsymbol{\beta}.$$

We can estimate this as $\hat{\mu}_0 = \mathbf{x}_0\hat{\boldsymbol{\beta}}$. We now look at the properties of this estimator.

**2.41. Proposition:**

$$\boxed{\begin{aligned}\mathbb{E}[\hat{\mu}_0] &= \mathbf{x}_0\boldsymbol{\beta} \\ \operatorname{Var}[\hat{\mu}_0] &= \sigma^2\mathbf{x}_0\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_0^T\end{aligned}}$$

*Proof.*

$$\begin{aligned}\mathbb{E}[\hat{\mu}_0] &= \mathbb{E}[\mathbf{x}_0\hat{\boldsymbol{\beta}}] \\ &= \mathbf{x}_0\mathbb{E}[\hat{\boldsymbol{\beta}}] \\ &= \mathbf{x}_0\boldsymbol{\beta}\end{aligned}$$

$$\begin{aligned}\operatorname{Var}[\hat{\mu}_0] &= \operatorname{Var}\left(\mathbf{x}_0\hat{\boldsymbol{\beta}}\right) \\ &= \mathbf{x}_0\operatorname{Var}(\hat{\boldsymbol{\beta}})\mathbf{x}_0^T \\ &= \mathbf{x}_0\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_0^T \\ &= \sigma^2\mathbf{x}_0\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_0^T\end{aligned}$$

□

**2.42. Note:** By the same logic as before,

$$\frac{\hat{\mu}_0 - \mu_0}{\sigma\sqrt{\mathbf{x}_0\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_0^T}} \sim N(0,1),$$

$$\frac{\hat{\mu}_0 - \mu_0}{\hat{\sigma}\sqrt{\mathbf{x}_0\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_0^T}} \sim t_{n-p-1},$$

and a $100(1-\alpha)\%$ CI is given by

$$\boxed{\hat{\mu}_0 \pm t_{n-p-1,1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{\mathbf{x}_0\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_0^T}.}$$

## Section 10. MLR: Prediction

**2.43. Note:** For a new response

$$y_{\text{new}} = \mathbf{x}_{\text{new}}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}_{\text{new}}\,,$$

our prediction is

$$\hat{y}_{\text{new}} = \mathbf{x}_{\text{new}}\,\hat{\boldsymbol{\beta}}.$$

**2.44. Proposition:**

$$\boxed{\begin{aligned}\mathbb{E}[\hat{y}_{\text{new}}] &= \mathbf{x}_{\text{new}}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{new}\\ \text{Var}[\hat{y}_{\text{new}}\,] &= \sigma^2\mathbf{x}_{\text{new}}\,\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_{\text{new}}^T\end{aligned}}$$

*Proof.*

$$\begin{aligned}\mathbb{E}\left[\hat{y}_{\text{new}}\,\right] &= \mathbb{E}\left[\mathbf{x}_{\text{new}}\,\hat{\boldsymbol{\beta}}\right]\\ &= \mathbf{x}_{\text{new}}\,\mathbb{E}[\hat{\boldsymbol{\beta}}]\\ &= \mathbf{x}_{\text{new}}\,\boldsymbol{\beta}\end{aligned}$$

$$\begin{aligned}\text{Var}\left[\hat{y}_{\text{new}}\,\right] &= \text{Var}\left(\mathbf{x}_{\text{new}}\,\hat{\boldsymbol{\beta}}\right)\\ &= \mathbf{x}_{\text{new}}\,\text{Var}(\hat{\boldsymbol{\beta}})\mathbf{x}_{\text{new}}^T\\ &= \mathbf{x}_{\text{new}}\,\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_{\text{new}}^T\\ &= \sigma^2\mathbf{x}_{\text{new}}\,\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_{\text{new}}^T\end{aligned}$$

□

**2.45. Note:** Since $y_{\text{new}}$ and $\hat{y}_{\text{new}}$ are independent and normally-distributed, we have

$$\frac{y_{\text{new}} - \hat{y}_{\text{new}}}{\sigma\sqrt{1 + \mathbf{x}_{\text{new}}\,\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_{\text{new}}^T}} \sim N(0,1),$$

$$\frac{y_{\text{new}} - \hat{y}_{\text{new}}}{\hat{\sigma}\sqrt{1 + \mathbf{x}_{\text{new}}\,\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_{\text{new}}^T}} \sim t_{n-p-1}$$

Thus, a $100(1-\alpha)\%$ prediction interval for $y_{\text{new}}$ is

$$\boxed{\hat{y}_{\text{new}} \pm t_{n-p-1,1-\alpha/2}\hat{\sigma}\sqrt{1 + \mathbf{x}_{\text{new}}\,\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{x}_{\text{new}}^T}\,.}$$

## Section 11. MLR: Categorical Covariates

**2.46.** Let `weight` be continuous and `fishpart` be categorical with four possible values:

- none (`N`)
- muscle tissue only (`M`)
- muscle tissue and sometimes whole fish (`MW`)
- whole fish (`W`)

Let `MeHg`(the concentration of methyl mercury extracted from hair sample) be the (continuous) response variable. For simplicity, let us ignore `weight` for now and only model the relationship between `fishpart` and `MeHg`. How should we encode `fishpart`? We will see that the way we encode the categorical covariates imposes assumptions on our model. In particular, it affects how we interpret the model parameters.

**2.47.** Naively, we could use numbers $0, 1, 2, 3$ to encode $\texttt{N}, \texttt{M}, \texttt{MW}, \texttt{W}$ (so that $\texttt{fishpart}_i \in \{0, 1, 2, 3\}$ for each $i$) and use

$$\texttt{MeHg}_i = \beta_0 + \beta_1 \texttt{fishpart}_i + \varepsilon, \qquad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

This model implicitly assumes that the difference between each consecutive factor of `fishpart` is the same. Some assumptions we made include:

- the mean difference of `MeHg` between people of group $i$ and people of group $i+1$ is always $\beta_1$;
- the mean difference of `MeHg` between people of group $i$ and people of group $i+2$ is $2\beta_1$;
- the mean difference of `MeHg` between people of group 3 and people of group 0 is $3\beta_1$, etc.

It is easy to see that if we had used other numbers (instead of 0 to 3) to encode the groups, then the model assumptions will be different.

**2.48.** We often don't want to make assumption about the relative differences between categories. A more flexible alternative is to use **indicator functions** and write

$$\begin{aligned}\texttt{MeHg}_i \sim \gamma_{\texttt{N}} \cdot \mathbf{1}\left[\texttt{fishpart}_i = \texttt{N}\right] &+ \gamma_{\texttt{M}} \cdot \mathbf{1}\left[\texttt{fishpart}_i = \texttt{M}\right] \\ &+ \gamma_{\texttt{MW}} \cdot \mathbf{1}\left[\texttt{fishpart}_i = \texttt{MW}\right] \\ &+ \gamma_{\texttt{W}} \cdot \mathbf{1}\left[\texttt{fishpart}_i = \texttt{W}\right] + \epsilon_{\text{i}}, \quad \epsilon \overset{\text{iid}}{\sim} N(0, \sigma^2).\end{aligned}$$

We essentially fitted four models based on `fishpart` with same variance but different mean; no assumption about relative differences between categories are made here:

$$\begin{aligned}\texttt{MeHg} \mid \{\texttt{fishpart=N}\} &\sim N\left(\gamma_{\texttt{N}}, \sigma^2\right) \\ \texttt{MeHg} \mid \{\texttt{fishpart=M}\} &\sim N\left(\gamma_{\texttt{M}}, \sigma^2\right) \\ \texttt{MeHg} \mid \{\texttt{fishpart=MW}\} &\sim N\left(\gamma_{\texttt{MW}}, \sigma^2\right) \\ \texttt{MeHg} \mid \{\texttt{fishpart=W}\} &\sim N\left(\gamma_{\texttt{W}}, \sigma^2\right)\end{aligned}$$

Another way to interpret this is that we are fitting four models with different intercepts and 0 slope (as `fishpart` is the only covariate here), i.e., they are horizontal lines at $y = \gamma_X$ with $X \in \{\texttt{N}, \texttt{M}, \texttt{MW}, \texttt{W}\}$.

**2.49.** We can replace the first term $\gamma_{\texttt{N}} \cdot \mathbf{1}\left[\texttt{fishpart}_i = \texttt{N}\right]$ with a $\beta_0$ and replace all $\gamma$'s with $\beta$'s. The resulting model will look more familiar to us:

$$\begin{aligned}\texttt{MeHg}_i \sim \beta_0 &+ \beta_{\texttt{M}} \cdot \mathbf{1}\left[\texttt{fishpart}_i = \texttt{M}\right] \\ &+ \beta_{\texttt{MW}} \cdot \mathbf{1}\left[\texttt{fishpart}_i = \texttt{MW}\right] \\ &+ \beta_{\texttt{W}} \cdot \mathbf{1}\left[\texttt{fishpart}_i = \texttt{W}\right] + \epsilon_i, \quad \epsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)\end{aligned}$$

The relationship between $\gamma$'s and $\beta$'s are given below:

- $\gamma_{\texttt{N}} = \beta_0$;
- $\gamma_{\texttt{M}} = \beta_0 + \beta_{\texttt{M}}$;
- $\gamma_{\texttt{MW}} = \beta_0 + \beta_{\texttt{MW}}$;
- $\gamma_{\texttt{W}} = \beta_0 + \gamma_{\texttt{W}}$.

and

- $\beta_{\texttt{0}} = \gamma_{\texttt{N}}$;
- $\beta_{\texttt{N}} = \gamma_{\texttt{M}} - \gamma_{\texttt{N}}$;
- $\beta_{\texttt{MW}} = \gamma_{\texttt{MW}} - \gamma_{\texttt{N}}$;
- $\beta_{\texttt{W}} = \gamma_{\texttt{W}} - \gamma_{\texttt{N}}$;

Interpretation of $\gamma$'s and $\beta$'s:

- $\gamma_X$ represents the mean `MeHg` for people of group $X$;
- $\beta_0$ represents the mean `MeHg` for people of group $N$, known as the **reference group**;
- $\beta_X$ represents the difference of the mean `MeHg` between group $X$ and the reference group.

**2.50.** Let us add the continuous covariate `weight`back. We can encode a regression model where expected `MeHg` is *linear* in `weight` for each level of `fishpart`, with common slope but different intercepts as follows:

$$\begin{aligned}\texttt{MeHg}_i \sim \gamma_1\texttt{weight} &+ \gamma_{\texttt{N}} \cdot \mathbf{1}\left[\texttt{fishpart}_i = \texttt{N}\right] \\ &+ \gamma_{\texttt{M}} \cdot \mathbf{1}\left[\texttt{fishpart}_i = \texttt{M}\right] \\ &+ \gamma_{\texttt{MW}} \cdot \mathbf{1}\left[\texttt{fishpart}_i = \texttt{MW}\right] \\ &+ \gamma_{\texttt{W}} \cdot \mathbf{1}\left[\texttt{fishpart}_i = \texttt{W}\right] + \epsilon_{\text{i}}, \quad \epsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)\end{aligned}$$

Interpretation of the parameters:

- $\gamma_1$ is the mean difference of `MeHg` for one unit of change in `weight`, holding `fishpart` constant.

$$\gamma_1 = \mathbb{E}[y \mid \texttt{weight} = w^*, \texttt{fishpart} = X] - \mathbb{E}[y \mid \texttt{weight} = w^* - 1, \texttt{fishpart} = X]$$

- $\gamma_X$ is the mean `MeHg` of people of group $X$ ($\texttt{fishpart}_i = X$), holding `weight` at 0.

$$\gamma_X = \mathbb{E}[y \mid \texttt{weight} = 0, \texttt{fishpart} = X].$$

This model consists of four submodels with different intercepts ($\gamma_X$'s) but a common slope ($\gamma_1$).

**2.51.** The corresponding $\beta$ model is given below:

$$\begin{aligned}\texttt{MeHg}_i \sim \beta_0 &+ \beta_1 \texttt{weight} \\ &+ \beta_{\texttt{M}} \cdot \mathbf{1}\left[\texttt{fishpart}_i = \texttt{M}\right] \\ &+ \beta_{\texttt{MW}} \cdot \mathbf{1}\left[\texttt{fishpart}_i = \texttt{MW}\right] \\ &+ \beta_{\texttt{W}} \cdot \mathbf{1}\left[\texttt{fishpart}_i = \texttt{W}\right] + \epsilon_i, \quad \epsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)\end{aligned}$$

Interpretation of the parameters:

- $\beta_0$ is the mean outcome of the reference group, holding `weight` at 0:

$$\beta_0 = \mathbb{E}[y \mid \texttt{weight} = 0, \texttt{fishpart} = \texttt{N}].$$

- $\beta_1$ is the mean difference of `MeHg` for one unit change in `weight`, holding `fishpart` constant:

$$\beta_1 = \mathbb{E}[y \mid \texttt{weight} = w^*, \texttt{fishpart} = X] - \mathbb{E}[y \mid \texttt{weight} = w^* - 1, \texttt{fishpart} = X].$$

- $\beta_X$ is the mean difference of `MeHg` between group $X$ and the reference group, holding `weight` constant:

$$\beta_X = \mathbb{E}[y \mid \texttt{weight} = w^*, \texttt{fishpart} = X] - \mathbb{E}[y \mid \texttt{weight} = w^*, \texttt{fishpart} = N].$$

This model consists of four submodels with different intercepts ($\beta_0$ or $\beta_0 + \beta_X$) and a common slope ($\beta_1$). The actual graph will be the same as the $\gamma$-model.

## Section 12. MLR: Hypotheses Testing (Categorical Covariates)

**2.52.** Suppose we want to test whether the average `MeHg` varies by `fishpart` adjusted for `weight`. There are two equivalent null hypotheses:

1. $\gamma_N = \gamma_M = \gamma_{MW} = \gamma_W$.
2. $\beta_M = \beta_{MW} = \beta_W = 0$.

The second is simpler for testing, so we'll proceed with the $\beta$-model from here on out.

**2.53.** To compare one group to the reference group:

$$\frac{\hat{\beta}_M - 0}{\text{SE}(\hat{\beta}_M)} \sim N(0,1)$$

To compare two non-reference groups:

$$\frac{\hat{\beta}_M - \hat{\beta}_{NW}}{\text{SE}(\hat{\beta}_M - \hat{\beta}_{MW})} \sim N(0,1)$$

where

$$\begin{aligned}\text{Var}(\hat{\beta}_\text{M} - \hat{\beta}_\text{MW}) &= \text{Var}(\hat{\beta}_\text{M}) + \text{Var}(\hat{\beta}_\text{MW}) - 2\,\text{Cov}(\hat{\beta}_\text{M}, \hat{\beta}_\text{MW}) \\ &= \sigma^2\,(V_{3,3} + V_{4,4} - 2V_{3,4})\end{aligned}$$

Don't forget to estimate $\sigma^2$ by $\hat{\sigma}^2$ and plug in $V = (X^TX)^{-1}$ as the covariance matrix

**2.54.** Suppose now we want to compare more than two groups. For example, does mean `MeHg` vary by `fishpart`, adjusted for `weight`? The null is given by

$$H_0 : \beta_* = (\beta_M, \beta_{MW}, \beta_W)^T = \mathbf{0}.$$

Recall that

$$\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1}) \implies \hat{\beta}_* \sim N(\beta, \sigma^* V_*)$$

where $V_*$ is the corresponding sub-matrix.

**2.55. Theorem** (Cholesky Decomposition)**:** *Any covariance matrix $V$ can be uniquely decomposed as $V = LL^T$ where $L$ is a lower triangular matrix with non-negative entries $L_{ii} \geq 0$ on the diagonal. When $V$ is positive-definite, then $L_{ii} > 0$.*

**2.56.** Let $L$ be a lower-triangular matrix such that $\sigma^2 V_* = LL^T$ and define $Z = L^{-1}(\hat{\beta}_* - \beta_*)$. Note that $Z \sim N(0, I)$:

$$\begin{aligned}\mathbb{E}[Z] &= L^{-1}\mathbb{E}[\hat{\beta}_*] - L^{-1}\mathbb{E}[\beta_*] = L^{-1}\beta_* = L^{-1}\beta_* = 0 \\ \text{Var}[Z] &= \text{Var}[L^{-1}(\hat{\beta}_* - \beta_*)] \\ &= \text{Var}[L^{-1}\hat{\beta}_*] \\ &= L^{-1}\text{Var}(\hat{\beta}_*)(L^{-1})^T = L^{-1}\sigma^2 V_*(L^{-1})^T = L^{-1}LL^T(L^{-1})^T = I\end{aligned}$$

Let $q$ be the dimension of $\beta_*$. Consider the sum of $q$ squared standard normals:

$$\begin{aligned}\sum_{j=1}^{q} Z_j^2 &= Z^T Z \\ &= (\hat{\beta}_* - \beta_*)^T (L^{-1})^T L^{-1} (\hat{\beta}_* - \beta_*) \\ &= (\hat{\beta}_* - \beta_*)^T (LL^T)^{-1} (\hat{\beta}_* - \beta_*) \\ &= \frac{1}{\sigma^2} (\hat{\beta}_* - \beta_*)^T (V_*)^{-1} (\hat{\beta}_* - \beta_*). \quad (LL^T)^{-1} = (\sigma^2 V_*)^{-1}\end{aligned}$$

Thus, under $H_0$, we have that

$$\frac{1}{\sigma^2} (\hat{\beta}_*)^T (V_*)^{-1} (\hat{\beta}_*) = \sum_{j}^{q} Z_j^2 \sim \chi_q^2$$

and this is independent of (shown previously)

$$\frac{n-(p+1)}{\sigma^2} \hat{\sigma}^2 \sim \chi^2_{n-(p+1)},$$

Define an F-statistic:

$$F = \frac{\frac{1}{\sigma^2} (\hat{\beta}_*)^T (V_*)^{-1} (\hat{\beta}_*)/q}{\frac{n-(p+1)}{\sigma^2} \hat{\sigma}^2/(n-(p+1))} = \frac{(\hat{\beta}_*)^T (V_*)^{-1} (\hat{\beta}_*)}{q\hat{\sigma}^2}$$

obtained by dividing each random variable by its degree of freedom. Under $H_0$,

$$F = \frac{(\hat{\beta}_*)^T (V_*)^{-1} (\hat{\beta}_*)}{q\hat{\sigma}^2} \sim F(q, n-(p+1))$$

**2.57. Definition** (F-distribution)**:** Let $X_1 \sim \chi^2_{\nu_1}$ and $X_2 \sim \chi^2_{\nu_2}$ be independent. Then

$$W = \frac{x_1/\nu_1}{x_2/\nu_2}$$

has an F-distribution:

$$\begin{aligned}W &\sim F(\nu_1, \nu_2) \\ f(w) &= \frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)} \left( \nu_1^{\nu_1} \nu_2^{\nu_2} \frac{w^{\nu_1-2}}{(\nu_2+\nu_1 w)^{(\nu_1+\nu_2)}} \right)^{1/2}\end{aligned}$$

**2.58.** We can test $H_0$ by comparing $F$ to the corresponding $F$ distribution.

```
pf(F_obs, df1=3, df2=n-p-1, lower.tail=FALSE)
```

## Section 13. MLR: Intersections and Non-Linearities

**2.59.** Consider the model from last section:

$$\texttt{MeHg}_i \sim \beta_0 + \beta_1 \texttt{weight}_i + \beta_{\texttt{M}} \texttt{M}_i + \beta_{\texttt{MW}} \texttt{MW}_i + \beta_{\texttt{W}} \texttt{W}_i + \epsilon_i, \quad \epsilon \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)$$

where $X_i = \mathbf{1}[\texttt{fishpart}_i = X]$ denotes the corresponding indicator function. Recall this implies common slope for weight for any value of `fishpart` (i.e., parallel lines with different intercepts). What if we want different intercepts and different slopes?

**2.60.** Consider the following model.

$$\begin{aligned}\texttt{MeHg}_i \sim \beta_0 &+ \beta_1 \texttt{weight}_i + \beta_{\texttt{M}} \texttt{M}_i + \beta_{\texttt{MW}} \texttt{MW}_i + \beta_{\texttt{W}} \texttt{W}_i \\ &+ \beta_{1\texttt{M}} \texttt{weight}_i \texttt{M}_i + \beta_{1\texttt{MW}} \texttt{weight}_i \texttt{MW}_i + \beta_{1\texttt{W}} \texttt{weight}_i \texttt{W}_i + \epsilon_i, \quad \epsilon \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)\end{aligned}$$

We added three columns to our design matrix. More specifically, the design matrix looks as follows:

- The first column consists of just 1s.
- The second column contains weights, real numbers.
- The next three columns correspond to $\texttt{M}_i, \texttt{MW}_i, \texttt{W}_i$, so either 0 or 1.
- The last three columns are the products of $X_i \cdot \texttt{weight}_i$. If $X_i = 0$ then the entry is 0; otherwise the entry is $\texttt{weight}_i$.

**2.61.** To see that this model gives different mean and different intercepts, observe that if $\texttt{fishpart}_i = \texttt{N}$, the mean outcome is given by

$$\mathbb{E}[\texttt{MeHg}_i \mid \texttt{weight}_i, \texttt{N}_i = 1] = \beta_0 + \beta_1 \texttt{weight}_i,$$

so the mean is linear in $\texttt{weight}_i$ with intercept $\beta_0$. For $\texttt{fishpart}_i = \texttt{MW}$, the mean outcome is

$$\begin{aligned}\mathbb{E}[\texttt{MeHg}_i \mid \texttt{weight}_i, \texttt{MW}_i = 1] &= \beta_0 + \beta_1 \texttt{weight}_i + \beta_{1\texttt{MW}} \texttt{weight}_i + \beta_{\texttt{MW}} \\ &= (\beta_0 + \beta_{\texttt{MW}}) + (\beta_1 + \beta_{1\texttt{MW}}) \texttt{weight}_i.\end{aligned}$$

The mean is still linear in $\texttt{weight}_i$, but with a different slope and a different intercept. Terms like $\texttt{weight}_i \texttt{MW}_i$, where different covariates are multiplied together, are called **interaction terms**. They are preferable here as they allow different slopes.

**2.62. Note:** Time for the interpretation of the parameters. The parameter $\beta_1$ is the mean difference in `MeHg` for one unit change of `weight`, provided that $\texttt{fishpart}_i = \texttt{N}$

$$\begin{aligned}\beta_1 &= \mathbb{E}[\texttt{MeHg}_i \mid \texttt{weight}_i = x^* + 1, \texttt{N}_i = 1] - \mathbb{E}[\texttt{MeHg}_i \mid \texttt{weight}_i = x^*, \texttt{N}_i = 1] \\ &= (\beta_0 + \beta_1(x^* + 1)) - (\beta_0 + \beta_1 x^*).\end{aligned}$$

$\beta_{1\texttt{MW}} + \beta_1$ is the mean difference in `MeHg` for one unit change of `weight`, given that $\texttt{fishpart}_i = \texttt{MW}$.

$$\begin{aligned}\beta_1 &= \mathbb{E}[\texttt{MeHg}_i \mid \texttt{weight}_i = x^* + 1, \texttt{MW}_i = 1] - \mathbb{E}[\texttt{MeHg}_i \mid \texttt{weight}_i = x^*, \texttt{MW}_i = 1] \\ &= (\beta_0 + \beta_{MW} + \beta_1(x^* + 1) + \beta_{1\texttt{MW}}(x^* + 1)) - (\beta_0 + \beta_1 x^* + \beta_1 x^* + \beta_{1\texttt{MW}} x^*).\end{aligned}$$

Thus, $\beta_{1\texttt{MW}}$ is the difference between "mean difference in `MeHg` for one unit change of `weight`, provided that $\texttt{fishpart}_i = \texttt{MW}$" and "mean difference in `MeHg` for one unit change of `weight`, provided that $\texttt{fishpart}_i = \texttt{N}$". It is the increase of the *slope* compared to the slope of the reference group.

**2.63. Note** (Interactions of continuous covariates)**:** Let $x_{i1}$ and $x_{i2}$ be continuous covariates and consider the following model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$$

We can bruteforce this by observing that

$$\beta_1 = \mathbb{E}[y_i \mid x_{i2} = 0, x_{i1} = x^* + 1] - \mathbb{E}[y_i \mid x_{i2} = 0, x_{i1} = x^*],$$

i.e., it is the average change of outcome for one unit change of $x_{i1}$, holding $x_{i2} = 0$. This is not very intuitive. A better way to interpret this is to observe that

- at every level of $x_2$, the conditional mean outcome is linear in $x_1$;
- at every level of $x_2$, the intercept and slope of $x_1$ are different.

Since the change in mean outcome due to one unit change in $x_1$ varies with $x_2$, it's better to fix a set of $x_{i2}$'s and then report the corresponding $\beta_3$ as the average change of outcome for one unit change of $x_{i1}$.

**2.64. Note** (More Flexible Models)**:** Sometimes a simple linear model does not fit the data well. One way to make the model more flexible is to include a quadratic term for $x$:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i.$$

Note here the change in mean outcome for a one unit change in $x_i$ varies with $x_i$. To test whether the quadratic model is more appropriate than the simple linear one, the null hypothesis is $H_0 : \beta_2 = 0$. Beyond polynomial terms, linear regression can be specified flexibly:

$$y_i = \sum_{j=1}^{p} \beta_j f_j(x_i) + \epsilon_i$$

where $f_j(\cdot)$ are arbitrary functions of $x_i$. However, there is a tradeoff between *fit* and *interpretability*.

**2.65. Note** (Hierarchical Principle)**:**

- If there is a higher order interaction term, include main effects (and lower order interaction terms), i.e.:
  - If including $x_1 \cdot x_2$, include also $x_1$ and $x_2$ (main effects).
  - If including $x_1 \cdot x_2 \cdot x_3$, include also $x_1 \cdot x_2$ and $x_2 \cdot x_3, x_1 \cdot x_3$, and main effects.
- If there is a higher order polynomial term, include main effects and lower order terms
  - If including $x^3$, include also $x^2$ and $x$.

Otherwise can have unexpected interpretations/implications.

**2.66. Example:** Consider the model $y_i = \beta_0 + \beta_2 x_i^2 + \varepsilon_i$. Now suppose we shift the exposure by some fixed amount $b$, e.g., center the $x_i$ to have mean 0 (so b = $\bar{x}$):

$$\begin{aligned} y_i &= \beta_0 + \beta_2 (x_i - b)^2 + \epsilon_i \\ &= \beta_0 + \beta_2 \left(x_i^2 - 2x_i b + b^2\right) + \epsilon_i \\ &= \left(\beta_0 + b^2 \beta_2\right) + (-2b\beta_2) x_i + \beta_2 x_i^2 + \epsilon_i \end{aligned}$$

Suddenly, there is now a linear term, simply because of a shift!

## Section 14. Analysis of Variance and $R^2$

**2.67. Motivation:** Recall the **sample variance** is given by

$$s^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2.$$

Suppose we wish to quantify how much of the variability in the outcome $\mathbf{y}$ is explained by our model. That is, we want to decompose the sum of squares $\sum (y_i - \bar{y})^2$ into two parts, one for the variance we can explain with our model and one for the variance we cannot explain.

**2.68. Definition** (ANOVA Decomposition)**:** Define the following terms:

- **Total sum of squares**, SSTotal, quantifies how much the data points $y_i$ vary around their mean $\bar{y}$.

$$\text{SSTotal} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = (\mathbf{y} - \bar{y}\mathbf{1})^T(\mathbf{y} - \bar{y}\mathbf{1}) = \mathbf{y}^T\mathbf{y} - n\bar{y}^2.$$

- **Regression sum of squares**, SSReg, quantifies how far the estimated regression model $\hat{y}_i$ is from the horizontal "no relationship line", the sample mean $\bar{y}$.

$$\text{SSReg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = (\mathbf{H}\mathbf{y} - \bar{y}\mathbf{1})^T(\mathbf{H}\mathbf{y} - \bar{y}\mathbf{1}) = \mathbf{y}^T\mathbf{H}\mathbf{y} - n\bar{y}^2.$$

- **Residual sum of squares**, SSRes, quantifies how much the data points $y_i$ vary around the regression estimates $\hat{y}_i$.

$$\text{SSRes} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = (\mathbf{y} - \mathbf{H}\mathbf{y})^T(\mathbf{y} - \mathbf{H}\mathbf{y}) = \mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y}.$$

**2.69. Theorem** (ANOVA Decomposition)**:** $\text{SSTotal} = \text{SSReg} + \text{SSRes}$.

*Proof.* Observe that

$$\begin{aligned}
\text{SSTotal} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} ([y_i - \hat{y}_i] + [\hat{y}_i - \bar{y}])^2 \\
&= \sum_{i=1}^{n} [y_i - \hat{y}_i]^2 + \sum_{i=1}^{n} [\hat{y}_i - \bar{y}]^2 + 2\sum_{i=1}^{n} [y_i - \hat{y}_i]\,[\hat{y}_i - \bar{y}] \\
&= \text{SSRes} + \text{SSReg} + 0
\end{aligned}$$

where we used the fact that

$$\sum_{i=1}^{n} [y_i - \hat{y}_i]\,[\hat{y}_i - \bar{y}] = \sum_{i=1}^{n} e_i\,[\hat{y}_i - \bar{y}] = \mathbf{e}^T[\hat{\mathbf{y}} - \bar{y}\mathbf{1}] = \mathbf{e}^T\mathbf{X}\hat{\boldsymbol{\beta}} - \bar{y}\mathbf{e}^T\mathbf{1} = 0 - 0 = 0.$$

as $\mathbf{e}^T\mathbf{X} = \mathbf{e}^T\mathbf{1} = \mathbf{0}$. □

**2.70. Definition:** The **coefficient of decomposition** defined as

$$R^2 = \frac{\text{SSReg}}{\text{SSTotal}} = 1 - \frac{\text{SSRes}}{\text{SSTotal}} \in [0, 1]$$

is the proportion of the variance in the outcome that is explained by our regression model.

**2.71. Note:**

- Since $R^2$ is a proportion, it is always a scalar between 0 and 1.
- If $R^2 = 1$, then $\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2$, so all data points fall on our regression line and the model perfectly captures all variability of the data.
- If $R^2 = 0$, then $\sum(\hat{y}_i - \bar{y})^2 = 0$, so $\hat{y}_i = \bar{y}$ for every $i$, i.e., our regression model is just the line of mean $\bar{y}$ and no variance is explained by our model.
- Thus, a higher $R^2$ indicates that more variability in the outcome is explained by our model.

**2.72. Remark:** In SLR, $R^2 = r^2$ where $r$ is the coefficient of correlation.

**2.73. Note** (F-Test for Model Significance)**:** Suppose we want to test the significance of the regression model, i.e., is there any relationship between the outcome and *at least one* covariate? Consider the following hypotheses:[1]

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$
$$H_1 : \exists i \in \{1, \ldots, p\} : \beta_j \neq 0.$$

We can conduct an F-test with the SS decomposition. Under the null,

$$\frac{\text{SSReg}}{\sigma^2} \sim \chi^2_p$$
$$\frac{\text{SSRes}}{\sigma^2} \sim \chi^2_{n-(p+1)}$$

and they're independent, so we can define the $F$-statistic

$$F = \frac{\text{SSReg}/p}{\text{SSRes}/n - (p+1)} \sim F_{p,n-(p+1)}.$$

Reject null if $p$-value $< \alpha$. If we reject $H_0$, we conclude that at least one of the regression coefficients is non-zero. Otherwise, we don't have enough evidence to conclude that none of $\beta_j$ is important.

**2.74. Note** (ANOVA Table)**:** We can summarize everything into a table:

| Source | SS | df | MS | F |
|---|---|---|---|---|
| Regression | $\text{SSReg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ | $p$ | $\text{MSReg} = \frac{\text{SSReg}}{p}$ | $\frac{\text{MSReg}}{\text{MSRes}} = \frac{\text{SSReg}}{p} / \frac{\text{SSRes}}{n-(p+1)}$ |
| Residuals | $\text{SSRes} = \sum_{i=1}^n (y_i - \hat{y})^2$ | $n-(p+1)$ | $\text{MSRes} = \frac{\text{SSRes}}{n-(p+1)}$ | |
| Total | $\text{SSTotal} = \sum_{i=1}^n (y_i - \bar{y})^2$ | $n-1$ | | |

[1] Note that we are not imposing any condition on the intercept $\beta_0$.

**2.75. Note** (F-Test for a Subset of Covariates)**:** Suppose we want to test $\beta_j = 0$ for $q$ of the $p$ covariates. Intuitively, we are testing whether these $q$ covariates are not very useful in our model:

$$H_0 : \beta_{k_1} = \cdots = \beta_{k_q} = 0$$
$$H_1 : \exists i \in \{k_1, \ldots, k_q\} : \beta_{k_i} \neq 0.$$

Suppose we fit the full model as before and additionally fit a *reduced model* under the null. The *additional variation* explained by the identified $q$ covariates

$$\text{SSReg(Full)} - \text{SSReg(Reduced)}$$

has $q$ degrees of freedom. Under $H_0$, we thus have

$$\frac{(\text{SSReg(Full)} - \text{SSReg(Reduced)})/q}{\text{SSRes}/n - (p+1)} \sim F_{q,n-(p+1)}$$

Note that F-Test for Model Significance can be viewed as a special case of this where we are testing the significance of all $p$ covariates.

**2.76. Note** (F-Test for General Linear Hypothesis)**:** We can use the same infrastructure to test a broader class of null hypothesis, called **general linear hypotheses**, all of the form

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$$
$$H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{0}$$

Here $\mathbf{C} \in \mathbb{R}^{\ell \times (p+1)}$ is a matrix of rank $r$ representing the hypotheses. In particular, $\ell$ denotes the number of linear constraints. For example, given a model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i,$$

the matrices corresponding to the null hypotheses are given by

$$H_0 : \beta_1 = \beta_2 = 0 \iff \mathbf{C} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$
$$H_0 : \beta_1 = \beta_2 \iff \mathbf{C} = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 \end{bmatrix}$$

Fit the full model and the reduced model (the model under $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$) and then construct $F$ statistic:

$$\frac{(\text{SSReg(Full)} - \text{SSReg(Reduced)})/r}{\text{SSRes}/n - (p+1)} \sim F_{r,n-(p+1)}$$

where $r = \text{rank}(C)$.

## Section 15. Multicollinearity and Variance Inflation Factor

**2.77. Motivation: Multicollinearity** is a phenomenon in which one covariate in a MLR model can be linearly predicted from the others with a substantial degree of accuracy. In this situation, the coefficient estimates of the MLR model may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set; it only affects calculations regarding individual predictors. That is, a multivariate regression model with collinear predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others.

**2.78. Definition: Collinearity** is a linear association between two covariates. Two variables are **perfectly collinear** if there is an exact linear relationship between them. **Multicollinearity** refers to a situation in which more than two covariates in a MLR model are highly linearly related.

**2.79. Note** (Perfect Multicollinearity in OLS)**:** Recall OLS requires *no multicollinearity*, i.e., there cannot exist an exact (non-stochastic) linear relation among the covariates, because in that case the design matrix $\mathbf{X}$ has less than full rank, and therefore the moment matrix $\mathbf{X}^T\mathbf{X}$ cannot be inverted. Under these circumstances, for a general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the OLS estimators $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ does not exist.

**2.80. Note** (VIF)**:** To detect pairwise collinearity, we may plot the correlation matrix and look for correlation values which are close to one. It's much hard to detect multicollinearity. One intuitive thing to do is to try to predict one covariate $x_j$ using the rest with a MLR model, i.e., let $x_j$ be the true outcome and consider the model

$$x_j = \mathbf{X}_{-j}\boldsymbol{\alpha} + \boldsymbol{\epsilon}^*$$

where $\mathbf{X}_{-j}$ denotes the matrix obtained by removing the column corresponding to $x_j$ from $\mathbf{X}$ and $\boldsymbol{\alpha}, \boldsymbol{\epsilon}^*$ play the role of $\boldsymbol{\beta}, \boldsymbol{\varepsilon}$, respectively. The fitted values are

$$\hat{x}_j = \mathbf{X}_{-j}\hat{\boldsymbol{\alpha}}$$

where $\hat{\boldsymbol{\alpha}}$ is estimator of $\boldsymbol{\alpha}$. Recall in SLR, $r^2_{yx} = R^2$. It turns out that in MLR, $r^2_{y,\hat{y}} = R^2$. Therefore, the coefficient of correlation between the true $y$ and the fitted $\hat{y}$ is exactly $R^2$. In particular, this is true in the regression of $x_j$ on $\mathbf{X}_{-j}$. Thus, we could examine $r^2_{x_j,\hat{x}_j}$ which is equal to the $R^2_j$ for the regression of $x_j$ on $X_{-j}$. This motivates the following definition.

**2.81. Definition:** Let $R_j$ the coefficient of correlation for the regression on $x_j$ using other covariates. The **variance inflation factor** (VIF) defined by

$$\text{VIF}_j := \frac{1}{1 - R^2_j}$$

quantifies the severity of multicollinearity in a regression analysis. It provides an index that measures how much the variance of an estimated regression coefficient is increased due to collinearity, i.e., how much it is increased relative to the ideal case in which all covariates are uncorrelated (i.e., the columns of $X$ are orthogonal).

**2.82. Note:** Consider the MLR $y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$. It turns out that the variance of the estimator of $\beta_j$ can be expressed as

$$\text{Var}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{\sum (x_{ij} - \bar{x}_j)^2} \times \frac{1}{1 - R_j^2}.$$

We make the following observations:

- $\sigma^2$: greater scatter in the data around the regression surface leads to proportionately more variance in the coefficient estimates.
- $\sum (x_{ij} - \bar{x}_j)^2$: greater variability in a particular covariate leads to proportionately less variance in the corresponding coefficient estimate.

The remaining term, $1/(1-R_j^2)$, is the VIF. It reflects all other factors that influence the uncertainty in the coefficient estimates. The VIF equals 1 when the vector $X_j$ is orthogonal to every other column of the design matrix. By contrast, the VIF is greater than 1 when the vector $X_j$ is not orthogonal to all columns of the design matrix for the regression of $X_j$ on the other covariates. Finally, note that the VIF is invariant to the scaling of the variables.

# Chapter 3. Model Building

## Section 16. Model Fit

In this section, we discuss four model building principles:

- Interpretability.
- Parsimony.
- Goodness of fit.
- Predictive accuracy.

**3.1. Note** (Interpretability)**:** When the goal of regression analysis is to make inferences about the relationship between $y$ and one or more covariates, our model is useful to the extent that it can be interpreted. There is often a tradeoff between complexity and interpretability. We may have to do more work to make a complex model more interpretable, e.g., plotting fitted values and reporting mean differences for specific contrasts.

**3.2. Note** (Parsimony)**:** We prefer models with fewer parameters for the following reasons:

- Interpretability: adjusting for more covariates makes interpreting $\beta_1$ more difficult.
- Precision: as we include irrelevant predictors, $p$ increases, so the variance increases:

$$p \uparrow \implies \hat{\sigma}^2 = \frac{\sigma_i(y_i - \hat{y}_i)}{n - (p+1)} \uparrow .$$

- Prediction: as we include more predictors, the SE could also increase:

$$p \uparrow \implies \mathrm{SE}(\hat{y}_{new}) = \hat{\sigma}\sqrt{1 + x_{\text{new}}^T (X^T X)^{-1} x_{\text{new}}} \uparrow .$$

**3.3. Note** (Goodness of Fit)**:** We cover some criteria for measuring goodness of fit:

1. $R^2$
2. Adjusted $R^2$
3. Mean squared error
4. AIC and related criteria

**3.4. (Cont'd)** ($R^2$)**:** Recall that

$$R^2 = \frac{\text{SSReg}}{\text{SSTotal}} = 1 - \frac{\text{SSRes}}{\text{SSTotal}}$$

can be viewed as the proportion of variability explained by the model. It has a small problem: $R^2$ will never decrease when more variables are added. The intuition is as follows. Recall the OLS estimators $\hat{\beta}$ minimizes $\text{SSRes} = \sum_i (y_i - \hat{y}_i)^2$ where $\hat{y}_i$ is in the column space of $X$. Now increasing the column space of $X$ (resulted from adding more covariates) increases the space over which we are minimizing. Thus, in the larger space we could never do worse than in the reduced space. This makes comparing models of different size difficult, as it would always favor the larger model.

**3.5. (Cont'd)** (Adjusted $R^2$)**:** Instead, we can use the **adjusted** $R^2$ given by

$$R^2_{\text{adj}} = 1 - \frac{\text{SSRes}/(n-(p+1))}{\text{SSTotal}/n-1}$$

where $p$ denotes the number of covariates. Intuitively, SSRes is non-decreasing, but $p$ increases with the number of variables. Thus, if SSRes decreases only slightly or not at all, it could be outweighed by increasing the degrees of freedom used and decrease $R^2_{\text{adj}}$. If SSRes decreases a lot, it can outweigh the increase in degrees of freedom and increase $R^2_{\text{adj}}$. Thus, this measure is more useful for comparing models of different size: we prefer model with higher $R^2_{\text{adj}}$. We pay the cost of interpretability: this no longer presents the proportion of variance explained.

**3.6. (Cont'd)** ($R^2_{\text{adj}}$ vs MSE)**:** Observe that minimizing $\hat{\sigma}$ is equivalent to maximizing $R^2_{\text{adj}}$:

$$R^2_{\text{adj}} = 1 - \frac{\text{SSRes}/(n-p-1)}{\text{SSTotal}/n-1} = 1 - \frac{\hat{\sigma}^2}{\text{MSTotal}}.$$

Thus, we could equivalently choose the model with lowest $\hat{\sigma}$.

**3.7. (Cont'd)** ($R^2_{\text{adj}}$ vs $R^2$)**:** Observe that

$$\begin{aligned}
R^2_{adj} &= 1 - \frac{\text{SSRes}\,/(n-p-1)}{\text{SSTotal}\,/n-1} \\
&= 1 - \frac{n-1}{n-p-1} \cdot \frac{\text{SSRes}}{\text{SSTotal}} \\
&= 1 - \left(\frac{n-1}{n-p-1}\right)\left(1-R^2\right) \\
&= 1 - \left(1 + \frac{p}{n-p-1}\right)\left(1-R^2\right) = R^2 - \left(\frac{p}{n-p-1}\right)\left(1-R^2\right)
\end{aligned}$$

Since the second term is positive, $R^2_{\text{adj}}$ is a penalized version of $\mathbb{R}^2$. In particular,

$$n \to \infty \implies R^2_{\text{adj}} \to R^2.$$

**3.8. (Cont'd)** (AIC)**:** In the same sense, we could use other penalized criteria such as the **Akaike Information Criterion** (AIC). Recall that $\hat{\beta}$ is the MLE under the assumption of normality. Define

$$\text{AIC} = -2\log\mathcal{L}(\hat{\theta}) + 2k$$

where $\hat{\theta}$ denotes the MLE of $\theta$ parameters and $k$ is the number of all parameters to be estimated (including intercept and $\sigma^2$). We prefer models with lower AIC. Indeed, we want to maximize the log-likelihood, hence minimize the first term. Note this is subject to a penalty term for the number of parameters.

**3.9. (Cont'd)** (BIC)**:** The **Bayesian information criterion**

$$\text{BIC} = -2\log\mathcal{L}(\hat{\theta}) + k\log(n)$$

gives a larger penalty term and takes the sample size into consideration. Other information criteria exist and all aim to balance fit with a penalty for more variables.

## Section 17. Model Building: Automatic Selection

**3.10. Motivation:** Suppose we have $p$ covariates and we want to find which subset of variables to adjust for in our regression model. We could consider all possible subsets and choose the best. This requires fitting $2^p$ models and is often not doable.

**3.11. Note** (Forward Selection)**:** Consider the following greedy method.

1. Start with model $M_0$ containing no covariate:
$$M_0 : y = \beta_0 + \varepsilon.$$
2. Fit all $p$ models including exactly one covariate and pick the one that performs best (according to some criteria, e.g., $p$-value, BIC, etc.). The current model is then
$$M_1 : y = \beta_0 + \beta_1 x_* + \varepsilon.$$
3. Fit all $p-1$ models including exactly two covariates, one being the $x_*$ from step 1. Pick the one that performs best. The current model is then
$$M_2 : y = \beta_0 + \beta_1 x_* + \beta_2 x_{**} + \varepsilon.$$
4. Continue until including the next variable stops improving significantly.

Note that once a variable enters the parameter set, it is never removed. Since we consider $(p+i-1)$ at step $i$, we at most consider $p!$ models. However, there is no guarantee that this is the globally optimal model.

**3.12. Note** (Backward Elimination)**:** A related greedy approach is as follows.

1. Start with all covariates included.
2. Drop the least important covariate according to our criteria. Note if covariates are categorical, this would require an $F$-test (to determine if the categorical covariate as a whole is relevant).
3. Continue until dropping a variable doesn't significantly improve our measure.

Note that once a variable is removed, it cannot be added back into the model. This can sometimes perform better than forward selection. However, if $p$ is very large (e.g., $p > n$), we may not be able to fit the model. We need a compromise between these two approaches.

**3.13. Note** (Stepwise Selection)**:**

1. Start with a model $M_0 : y = \beta_0 + \varepsilon$ containing no covariate.
2. Add one covariate according to the same criteria as in forward selection.
3. Assess whether any of the covariates should be removed as in backward selection.
4. Repeat 2 and 3 until the most recently added covariate is removed.

**3.14. Remark:** These methods are fairly primitive. We will soon cover some more modern approaches, e.g., LASSO.

**3.15. Remark** (Problems with Variable Selection)**:** Inference is not valid in the final model. The intuition is that we are using the data to select the model, so the model will underestimate the uncertainty (i.e., SE too small). Classical inference requires the model/hypothesis to be fixed.

## Section 18. Overfitting and Cross Validation

**3.16. Note:** Criteria like AIC, BIC attempt to approximate *out-of-sample* prediction error based on *in-sample error*, SSRes, and explicitly penalize the number of parameters to avoid overfitting.

**3.17. Note** (Holdout)**:** Split $n$ observations into a training set $S_{\text{train}}$ of $n_{\text{train}}$ observations and a test set $S_{\text{test}}$ of size $n_{\text{test}} = n - n_{\text{train}}$.

- Fit our model to the training set and get estimates $\hat{\beta}_{\text{train}}$.
- Use these to predict $y^{\text{new}}$ with $\hat{y}^{\text{new}} = (x^{\text{new}})^T \hat{\beta}$.
- Measure prediction accuracy with MSE (or its square root, RMSE)

$$\frac{1}{n_{\text{test}}} = \sum_{i \in S_{\text{test}}} (y_i^{\text{new}} - \hat{y}_i^{\text{new}})^2.$$

**3.18. Note** ($k$-Fold Cross Validation)**:** The downside of splitting data is that we aren't using all the data to fit the model.

1. Randomly divide the data into $K$ parts.
2. Fir the model on $(K-1)$ of the $K$ folds (leaving the $k$th out).
3. Predict outcomes on the $k$th part (as a test set $S_{\text{test},k}$) and compute

$$\text{MSE}_k = \frac{1}{n_k} \sum_{i \in S_{\text{test},k}} (y_i^{\text{new}}, \hat{y}_i^{\text{new}})^2.$$

4. Repeat 2-3, leaving out each of the $K$ folds once.
5. Compute

$$\text{MSE}_{cv} = \frac{1}{K} \sum_{k=1}^{K} \text{MSE}_k.$$

**3.19. Note:** Setting $K = n$, this becomes **leave-one-out cross validation**. That is, we fit the model to the entire data except we leave one observation out, and compute $\hat{y}_{i,(-i)}$. The MSE is given by

$$(y_i - \hat{y}_{i,(-i)})^2.$$

In linear regression, it turns out this is equal to the square of the PRESS statistic:

$$\frac{e_i}{1 - h_i},$$

where $h_i$ is the $i$th diagonal of the hat matrix, $H = X(X^T X)^{-1} X^T$. Hence, we can compute LOO-CV MSE without refitting the model $K = n$ times (1 time is enough). (Note this is not true for more complicated methods.)

**3.20. Note:** When reporting prediction error for final model after selection, we typically think about 3 types of data: training, validation, and test.

- Training set: Used to fit the model.

- Validation set: Compare predictions to validation outcomes, choose model with lowest RMSE. Refit best model to training + validation set.
- Test set: Estimate prediction error based on final model fit.

If you are doing cross validation, you can think of this as splitting the data into a training and test set, and then repeatedly splitting the training set further into a training set and validation set.

## Section 19. Model Building: LASSO and Shrinkage Methods

**3.21. Motivation:** Recall AIC and BIC penalize models for having a large number of covariates. The intuition behind LASSO and other shrinkage methods is that these methods incorporate a penalty directly into the estimation procedure. Recall the objective of OLS was

$$\min \quad \sum_i (y_i - x_i^T \beta)^2.$$

We now add a penalty term, so our new objective function becomes

$$\min \quad \sum_i (y_i - x_i^T \beta)^2 + \text{Penalty}.$$

**3.22. Note** (LASSO)**:** LASSO estimator is defined as the estimate of $\beta$ that minimizes the following:

$$\min \quad \sum_i (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Note we do not penalize the intercept estimate $\beta_0$. In other words, we are adding an $L_1$ penalty term with a hyperparameter $\lambda$. This penalty term has the effect of shrinking parameter estimates toward zero. The choice of $L_1$ is particularly convenient since it shrinks certain values all the way to zero. This is automatic variable selection. In fact, the minimization problem above is equivalent to

$$\min \quad \sum_i (y_i - x_i^T \beta)^2 \quad s.t. \quad \sum_{j=1}^{p} |\beta_j| \leq t,$$

so the penalty term really corresponds to a constraint.

**3.23. Note** (Fitting LASSO)**:** Different values of $\lambda$ will then lead to different model fits. To tune the hyperparameter $\lambda$, we could fit various models using different $\lambda$ values then do cross validation to choose the best one. R will do this automatically.

**3.24. Note** (Ridge Regression)**:**

$$\min \quad \sum_i (y_i - x_i^T \beta)^2 + \lambda \sqrt{\sum_{j=1}^{p} \beta_j^2}.$$

Ridge regression also shrinks estimates, but unlike LASSO, it doesn't shrink them all the way to zero.

**3.25. Note** (Relaxed LASSO)**:** Here's some more recent development:

- Fit LASSO and obtain the optimal $\lambda$ via CV.

$$\min \quad \sum_i (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

- Fit LASSO to the subset of covariates whose coefficients were not set to zero:

$$\min \sum_{i}^{n} \left(y_i - x_i^{*T}\beta^*\right)^2 + \phi\lambda \sum_{l} |\beta_l^*|$$

  The hyperparameter $\phi$ allows us to tune the ultimate level of shrinkage:

  - $\phi = 1$ gives the LASSO estimator.
  - $\phi = 0$ correspond to OLS estimate on the subset of selected variables.
  - $0 < \phi < 1$ allow different level of shrinkage, independent of the selection.

**3.26. Note:** Other shrinkage estimators include *elastic net*, *fused LASSO*, *group LASSO*, etc.

# Chapter 4. Regression Diagnostics

## Section 20. Regression Diagnostics: Residuals

**4.1. Motivation:** Recall $\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ only relies on linearity; the other three assumptions were not necessary for unbiased estimates. What about SEs?

$$\begin{aligned}\text{Var}(\hat{\boldsymbol{\beta}}) &= \text{Var}\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right) \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\,\text{Var}(\mathbf{y})\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\end{aligned}$$

Under independence and homoskedasticity, we have $\text{Var}[\mathbf{y}] = \sigma^2\mathbf{I}$:

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}.$$

But if either assumption is not met, our variance estimates (and hence SEs, CIs, etc.) be incorrect.

**4.2. (Cont'd):** A quick note on Normality. Without this assumption, $\hat{\boldsymbol{\beta}}$ is no longer a linear transformation of a MVN vector, hence it is no longer normally distributed and our CIs and test are not necessarily valid. However, in large samples, $\hat{\boldsymbol{\beta}}$ is *approximately* normal due to CLT, so we can get away with valid inference spite non-normal errors in "large-enough" samples. We just replace $t_{n-p-1,\alpha/2}$ with $z_{\alpha/2}$.

**4.3. (Cont'd):** Prediction intervals are sensitive to all 4 assumption. In particular, it explicitly require Normality:

$$y_{\text{new}} \sim N(x_{\text{new}}^T\boldsymbol{\beta}, \sigma^2).$$

Without normality, our predictions are still unbiased, but the prediction intervals are invalid.

**4.4. (Cont'd):** Each nice feature of regression relies on one of our assumption (to varying degrees). Once we have fit a model, we need some tools to diagnose whether our assumptions are broken.

**4.5. Note** (Assessing Normality)**:** One of the best tools for diagnostics is to visualize residuals. Recall **ordinary residuals**: $e_i = y_i - \hat{y}_i$. Define **studentized residuals** as

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_i}}$$

where $h_i$ is the $i$th diagonal of $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. The intuition is as follows.

$$\mathbf{e} = (\mathbf{I}-\mathbf{H})\mathbf{y} \sim N(0, \sigma^2(\mathbf{I}-\mathbf{H})).$$

Thus, $e_i \sim N(0, \sigma^2(1-h_i))$. Since $e_i$ have different variances, it is difficult to learn anything about their distribution. By contrast, $e_i/\sqrt{1-h_i}$ has constant variance $\sigma^2$, so they should look normally distributed when plotted. (Note: in practice, we estimate $\hat{\sigma}$, so the studentized residuals are really $t$-distributed.)

**4.6. Note** (Assessing Heteroskedasticity)**:** We can plot residuals against fitted values. This

can detect some mean-variance relationships, e.g., if there is higher variance for larger fitted values.

**4.7. Note** (Assessing Independence)**:** It could be difficulty to visualize independence unless you have something like time-series data. Instead, we can consider how data were collected.

**4.8. Note** (Assessing Linearity in SLR)**:** Consider SLR $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. The linearity assumption states that $\mathbb{E}[y_i] = \beta_0 + \beta_1 x_i$. The residual is given by $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$. We could do two things:

1. Plot $y_i$ against $x_i$, which should look linear.
2. Plot $e_i$ against $x_i$. This should look fairly random and the existence of conspicuous pattern may indicate non-linearity. It is sometimes easier to identify linearity with this plot.

**4.9. Note** (Assessing Linearity in MLR)**:** For MLR, plotting $y_i$ against $x_i$ ignores the effect of all the other covariates. Instead, we could use partial regression plots. To assess linearity in $x^*$:

1. Regress $y$ on all other covariates $x_j = x^*$, get fitted values from this model fit, and then compute the residuals $e_y$.
2. Regression $x^*$ on all other covariates $x_j = x^*$, get fitted values from this model fit, and then compute the residuals $e_{x^*}$.
3. Plot $e_y$ against $e_{x^*}$.

Intuitively, we are isolating the $y \sim x^*$ relationship, after adjusting for the other covariates.

## Section 21. Fixing Problems and Weighted Least Squares

**4.10. Motivation:** We have seen how violating assumptions can make our results invalid and how to assess whether our assumptions are broken. How do we fix our models?

**4.11. Note** (Fixing Linearity)**:** Suppose linearity isn't met. We might consider transforming $x_j$, e.g., using $\log(x_j)$ or quadratic $x_j^2$. However, this can change the interpretation.

**4.12. Note** (Fixing Independence)**:** Violations of independence require more advanced regression methods. If estimates are still unbiased but standard errors are broken, we can replace SEs with more robust alternatives. Alternatively, we can explicitly model the dependence structure.

**4.13. Note** (Fixing Normality)**:** Violations of normality might not be a big deal, especially if we have a large sample size. However, normality is required for valid prediction intervals. Solutions include could consider transforming $Y$, e.g., use $\log(Y)$. This again changes interpretation (but might not be a problem if we only care about predictions). We could consider other regression approaches, e.g., GLMs, etc.

**4.14. Note** (Fixing Homoskedasticity)**:** If our error are heteroskedastic, we have a few options:

- Transform outcome.
- Weighted least squares.
- Bootstrap.

**4.15. Note** (Weighted Least Squares)**:** Suppose we have heteroskedasticity:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \text{ s.t. } \boldsymbol{\epsilon} \sim N(0, \boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}$$

Likelihood:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[\frac{-1}{2\sigma_i^2}\left(y_i - x_i^T\boldsymbol{\beta}\right)^2\right].$$

Maximizing the likelihood is equivalent to:

$$\min w_i \left(y_i - x_i^T\boldsymbol{\beta}\right)^2,$$

where $w_i = \frac{1}{\sigma_i^2}$ This is **weighted least squares**, as opposed to ordinary least squares.

**4.16. (Cont'd)** (WLS in Matrix Notation)**:** In matrix notation, we can write

$$\min(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{W}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad \text{where } \mathbf{W} = \text{diag}\,(w_1, \dots, w_n)$$

Taking $\mathbf{W}$ as fixed for the moment:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}} \left[ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right] \\
&= \frac{\partial}{\partial \boldsymbol{\beta}} \left[ \mathbf{y}^T \mathbf{W} \mathbf{y} - \mathbf{y}^T \mathbf{W} \mathbf{X} \boldsymbol{\beta} - \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{W} \mathbf{y} + \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{W} \mathbf{X} \boldsymbol{\beta} \right] \\
\mathbf{0} &= \left[ -2\mathbf{X}^T \mathbf{W} \mathbf{y} + 2 \left( \mathbf{X}^T \mathbf{W} \mathbf{X} \right) \boldsymbol{\beta} \right] \\
\mathbf{X}^T \mathbf{W} \mathbf{y} &= \left( \mathbf{X}^T \mathbf{W} \mathbf{X} \right) \boldsymbol{\beta} \\
\left( \mathbf{X}^T \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y} &= \hat{\boldsymbol{\beta}}_W
\end{aligned}$$

Here $\hat{\boldsymbol{\beta}}_W$ is our WLS estimator. Compare it to the OLS estimator $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

**4.17. Note** (Properties of WLS Estimator)**:**

$$\begin{aligned}
\mathbb{E}\left[ \hat{\boldsymbol{\beta}}_W \right] &= \mathbb{E}\left[ \left( \mathbf{X}^\top \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y} \right] \\
&= \left( \mathbf{X}^T \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W} \mathbb{E}[\mathbf{y}] \\
&= \left( \mathbf{X}^T \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X} \boldsymbol{\beta} = \boldsymbol{\beta}
\end{aligned}$$

$$\begin{aligned}
\operatorname{Var}\left[ \hat{\boldsymbol{\beta}}_W \right] &= \operatorname{Var}\left[ \left( \mathbf{X}^\top \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y} \right] \\
&= \left( \mathbf{X}^T \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{W} \operatorname{Var}[\mathbf{y}] \mathbf{W}^T X \left( \mathbf{X}^T \mathbf{W} \mathbf{X} \right)^{-1} \\
&= \left( \mathbf{X}^\top \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{W} \boldsymbol{\Sigma} \mathbf{W}^T \mathbf{X} \left( \mathbf{X}^T \mathbf{W} \mathbf{X} \right)^{-1} \\
&= \left( \mathbf{X}^\top \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W} \mathbf{X} \left( \mathbf{X}^\top \mathbf{W} \mathbf{X} \right)^{-1} \\
&= \left( \mathbf{X}^\top \mathbf{W} \mathbf{X} \right)^{-1}
\end{aligned}$$

**4.18. Note** (Alternative View of WLS)**:** Alternatively, let $\mathbf{W}^{1/2} = \operatorname{diag}(w_1^{1/2}, \ldots, w_n^{1/2})$ where $w_i = 1/\sigma_i^2$. We could pre-multiply our model by $\mathbf{W}^{1/2}$:

$$\begin{aligned}
\mathbf{W}^{1/2} \mathbf{y} &= \mathbf{W}^{1/2} \mathbf{X} \boldsymbol{\beta} + \mathbf{W}^{1/2} \epsilon \\
\mathbf{y}_w &:=: \mathbf{X}_w \boldsymbol{\beta} + \boldsymbol{\epsilon}_w
\end{aligned}$$

The main benefit is to observe that $\boldsymbol{\epsilon}_w \sim \mathcal{N}(0, I)$:

$$\begin{aligned}
\mathbb{E}\left[ \boldsymbol{\epsilon}_w \right] &= 0 \\
\operatorname{Var}\left[ \boldsymbol{\epsilon}_w \right] &= \operatorname{Var}\left( \mathbf{W}^{1/2} \boldsymbol{\epsilon} \right) \\
&= \mathbf{W}^{1/2} \operatorname{Var}(\boldsymbol{\epsilon}) \mathbf{W}^{1/2} \\
&= \mathbf{W}^{1/2} \boldsymbol{\Sigma} \mathbf{W}^{1/2} = \mathbf{I}
\end{aligned}$$

In other words, we could achieve $\hat{\boldsymbol{\beta}}_W$ be applying OLS of $\mathbf{y}_w$ on $\mathbf{X}_w$.

**4.19. Note** (Fitting WLS)**:** In practice, we often don't know $\mathbf{W}$. Instead, we estimate $\sigma_i^2$ via

$e_i^2$. We can do this in a few ways:

- Directly: set $\sigma_i^2 \leftarrow e_i^2$ (unstable).
- Binning: estimate a single $\sigma_i^2$ for a group of observations.
- Model $\sigma_i^2$:
  - E.g., $|e_i| = \alpha_0 + \alpha_1 \hat{y}_i + \epsilon'$ and then $\hat{\sigma}_i^2 = |\hat{e}_i|^2$
  - E.g., $e_i^2 = \alpha_0 + \alpha_1 \hat{y}_i + \epsilon'$ and then $\hat{\sigma}_i^2 = \widehat{e_i^2}$
  - Could also regress against covariates instead of fitted values.

But how do we get $e_i = y_i - \hat{y}_i$ without first estimating $\hat{y}_i$? This becomes a bit circular. We now discuss the iterative reweighted LS algorithm:

1. Fit OLS $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and get fitted values $\hat{y}_i$ residuals $e_i$.
2. Using fitted values and residuals to estimate $\sigma_i^2$ as described above, then set $w_i = 1/\sigma_i 62$.
3. Fit WLS: $\hat{\boldsymbol{\beta}}_W = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$. Update fitted values and residuals.
4. Repeat until $\hat{\boldsymbol{\beta}}_W$ converges.

## Section 22. Outliers

**4.20. Outliers** are unusual or extreme observations.

**4.21. Note** (Detecting $X$-Outliers)**:** Consider a single covariate. Intuitively, outliers are points far from the mean, i.e., $x_i$ with large $|x_i - \bar{x}|$. How do we generalize this to MLR?

**4.22. Note** (Leverage)**:** Recall that hat matrix $\mathbf{H}$, which adds a hat to $\mathbf{y}$:

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}.$$

The **leverage** for the $i$th observation, $h_i$, is defined as the $i$th diagonal of $H$. Here's some intuition:

$$\hat{y}_i = \mathbf{H}_i\mathbf{y} = [h_{i1}, \ldots, h_{in}] \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \sum_{j=1}^{n} h_{ij}y_j = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j.$$

In words, $\hat{y}_i$ is a weighted average of our outcomes and the leverage $h_{ii}$ determines how much $Y_i$ contributes to the $i$th fitted value.

**4.23. (Cont'd):** Now recall that $\text{Var}[e_i] = \sigma^2(1 - h_i)$. If $h_i$ is large (close to 1), then $\text{Var}[e_i]$ is small, $|e_i|$ is small, and $\hat{y}_i$ is close to $y_i$.

**4.24. (Cont'd)** (Leverage in SLR)**:** Some intuition from SLR:

$$\begin{aligned}
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\
&= \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) \\
&= \sum_j \frac{1}{n} y_j + (x_i - \bar{x}) \frac{\sum_j y_j (x_j - \bar{x})}{S_{xx}} \\
&= \left[ \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right] y_i + \sum_{j \neq i} \left[ \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{S_{xx}} \right] y_j
\end{aligned}$$

Thus,

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}$$

which is large when $(x_i - \bar{x})^2$ is large, i.e., when $x_i$ is far from $\bar{x}$.

**4.25. (Cont'd)** (Leverage in MLR)**:** The story is analogous in MLR. Here's a statistical rule of thumb: label a point "high leverage" (i.e., is an outlier) if $h_i > 2\bar{h}$, where

$$\bar{h} = \frac{1}{n} \sum_{i=1}^{n} h_i = \frac{1}{n} \operatorname{tr}(\mathbf{H}) = \frac{p+1}{n},$$

where in the last equality we used the fact that the trace of an idempotent matrix is equal to its rank.

**4.26.** Leverage tells us whether $\hat{y}_i$ is close to $y_i$. It might be a problem if $y_i$ is also an outlier.

**4.27. Note:** Recall ordinary residuals $e_i = y_i - \hat{y}_i$. Since $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y} \sim N(0, \sigma^2(\mathbf{I} - \mathbf{H}))$, $e_i \sim N(0, \sigma^2(1 - h_i))$, i.e., $e_i$ have different variances, so it is difficult to work with. Instead, we can use studentized residuals

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_i}}.$$

The intuition is that $r_i$ has constant variance, so they should look normally distributed when plotted. In practice, we estimate $\hat{\sigma}$, so the studentized residuals are really $t$-distributed. We can look closely at observations with large $|r_i|$.

**4.28. Note:** Note that studentized residuals contain $\hat{\sigma}$. The larger the $h_i$, the smaller the studentized residuals. Thus, high leverage points can have lower studentized residuals. Also, the larger the $\hat{\sigma}$, the smaller the studentized residuals. Thus, $y$-outliers with large residuals would themselves contribute to a large $\hat{\sigma}$. We may think that large $e_i$ would cause a large $r_i$, but large $e_i$ would make the denominator large as well! Outliers can hide from us!

**4.29. (Cont'd):** To combat this, we can use leave-one-out or Jacknife residuals

$$e_{i(-i)} = y_i - \hat{y}_{i(-i)}$$

where $\hat{y}_{i(-i)}$ is the fitted value for the $i$th observation based on fitting the model *without* $y_i$. Likewise, we can compute studentized Jacknife residuals

$$r_{i(-i)} = \frac{e_{i(-i)}}{s_i}$$

where $s_i$ is the appropriate standard deviation so that $r_{i(-i)}$ has constant variance. Intuitively, this approach removes the $i$th observation from affecting the fitted value, so we can recognize them more easily.

**4.30. (Cont'd):** To avoid fitting the model $n$ times, we present some computational simplification. It can be shown that

$$r_{i(-i)} = \frac{e_i}{\sqrt{\hat{\sigma}^2_{(-i)}\,(1 - h_i)}} = r_i \left[\frac{(n - p - 2)}{n - p - 1 - r_i^2}\right]^{1/2},$$

which we can extract from a single model fit.

**4.31. Note:** Should we remove outliers? We should if we believe the observation is in some sense incorrect, e.g., data entry error, or an observation from a different population. However, we should not by default remove outliers; it's more useful to examine its impact on our results.

## Section 23. Influence

**4.32.** Previously, we saw that *high leverage* observations are those that have potential to impact our regression. In this section, we look at *influential* observations, those that strongly impact our regression. How do we quantify the impact of an observation on our regression model? One possible approach is to compare the model fit to the full data, to a model fit to the whole data *except for the ith observation.* This has the same intuition as for the Jackknife residuals last time. We now look at some other approaches, including DFFITS, Cook's distance, and DFBetas.

**4.33. Note** (DFFITS)**:** Define

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(-i)}}{\sqrt{\hat{\sigma}^2_{(-i)} h_i}}$$

where

$$\begin{aligned}\hat{\mathbf{y}}_{(-i)} &= \mathbf{X}\hat{\boldsymbol{\beta}}_{(-i)} \\ &= \mathbf{X}((\mathbf{X}^T_{(-i)}\mathbf{X}_{(-i)})^{-1}\mathbf{X}^T_{(-i)}\mathbf{y}_{(-i)})\end{aligned}$$

and $\hat{\sigma}^2_{(-i)}$ is the MSE for model fit to all observations except $y_i$. Intuitively, we look at the scaled difference between the fitted value for $y_i$ and what we would have gotten if we hadn't observed $y_i$. A large value of DF-Fits suggests that the fitted value changes substantially.

**4.34. (Cont'd):** It can be shown that

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(-i)}}{\sqrt{\hat{\sigma}^2_{(-i)} h_i}} = r_{i(-i)}\sqrt{\frac{h_i}{1-h_i}},$$

which is a function of $r_{i(-i)}$ and $h_i$. Thus, it incorporates information about both $x$-outliers (see $h_i$) and $y$-outliers (see $r_{i(-i)}$) and does not require us to refit models.

**4.35. (Cont'd):** Statistical rule of thumb: We say an observation is influential if its DFFITS satisfies

$$|\text{DFFITS}_i| > 2\sqrt{\frac{p+1}{n}}.$$

**4.36. Note** (Cook's Distance)**:** We define

$$D_i = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)}))^T(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)}))}{\hat{\sigma}^2 \times (p+1)} = \frac{\sum_j^n (\hat{y}_j - \hat{y}_{j(-i)})^2}{\hat{\sigma}^2 \times (p+1)}$$

where

$$\begin{aligned}\hat{\mathbf{y}}_{(-i)} &= \mathbf{X}\hat{\boldsymbol{\beta}}_{(-i)} \\ &= \mathbf{X}((\mathbf{X}^T_{(-i)}\mathbf{X}_{(-i)})^{-1}\mathbf{X}^T_{(-i)}\mathbf{y}_{(-i)})\end{aligned}$$

Intuitively, this is a scaled measure of averaged squared distance between fitted values with and without $y_i$, i.e., how the $i$th observation affects the fitted values.

**4.37. (Cont'd):** It turns out that we can write

$$D_i = \frac{r_i^2}{p+1}\frac{h_i}{1+h_i}.$$

Thus, it also incorporates information about both $x$-outliers and $y$-outliers and we don't need to refit the model.

**4.38. (Cont'd):** Statistical rule of thumb: Compare $D_i$ with $F_{p+1,N-p-1}$. A large percentile (e.g., 50th or above) indicates large effect on the fit.

**4.39.** DFFITS measures the $i$th observation's impact on its fitted value; Cook's distance measure $i$th observation's impact on all fitted values. Sometimes what we really care about is estimating $\beta$. We now present DFBETAS, which measures the impact of the $i$th observation on coefficient estimates.

**4.40. Note** (DFBETAS)**:** The DFBETAS measure for the $i^{\text{th}}$ observation's influence on the $k^{\text{th}}$ coefficient $\beta_k$

$$\text{DFBETAS}_{k,i} = \frac{\hat{\beta}_k - \hat{\beta}_{k(-i)}}{\sqrt{\hat{\sigma}^2_{(-i)}V_{kk}}}$$

where $\hat{\beta}_{k(-i)}$ is the $k^{th}$ element of $\hat{\boldsymbol{\beta}}_{(-i)}$ :

$$\hat{\boldsymbol{\beta}}_{(-i)} = (\mathbf{X}^T_{(-i)}\mathbf{X}_{(-i)})^{-1}\mathbf{X}^T_{(-i)}\mathbf{y}_{(-i)}$$

Note that the denominator is computed under full data fit.

**4.41. (Cont'd):** Statistical rule of thumb: Large values indicate large impact on estimation of $\beta_k$. In particular, we say an observation is influential if

$$|\text{DFBETAS}_{k,i}| > \frac{2}{\sqrt{n}}.$$

**4.42.** What to do about highly influential points? If we have reason to suspect that they are in some sense incorrect, we could exclude them. More broadly, it is good practice to report them, e.g., how do the results look with and without these points, and are our conclusions substantially different?