Stat 231 Chapter 6: Gaussian Response Models

Statistics

David Duan, 2019 Fall

Contents

1 Simple Linear Regression

1.1 Least Squares Estimate

1.2 Simple Linear Regression

1.3 Likelihood Function for α and β

1.4 Exploring the Slope

1.4.1 Estimator: Slope

1.4.2 Pivotal Quantity: Slope

1.4.3 Confidence Interval: Slope

1.4.4 Testing of Hypothesis: Slope

1.5 Exploring the Mean Response

1.5.1 Estimator: Mean Response

1.5.2 Pivotal Quantity: Mean Response

1.5.3 Confidence Interval: Mean Response

1.5.4 Confidence Interval: Intercept

1.6 Confidence Interval for an Individual Response Y at x

2 Gaussian Response Models and Model Checking

2.1 Gaussian Response Models

2.2 Model Checking

2.2.1 Method I: Scatter Plot with Fitted Line

2.2.2 Method II(a): Residual Plots

2.2.3 Method II(b): Standardized Residual Plots

2.2.4 Method III: Residual QQ Plot

2.2.5 Summary

3 Comparing Means of Two Populations

3.1 Two Sample Gaussian Problem (Equal Variance)

3.1.1 Estimator: Difference of Sample Means

3.1.2 Pivotal Quantity: Difference of Sample Means

3.1.3 Confidence Interval: Difference of Sample Means

3.1.4 Hypothesis Testing: Difference of Sample Means

3.1.5 Pivotal Quantity with Unequal SD: Difference of Sample Means

3.2 Paired Data

3.3 Summary

1 Simple Linear Regression

1.1 Least Squares Estimate

It is conventional to find the fitted line $y = \alpha + \beta x$ which minimizes the sum of squares of the distances between the observed points and the fitted line.

Def. 1.1.1 The method of *least squares* minimizes the sum of the squares of the *residuals*, the difference between an observed value and the fitted value provided by a model. Estimates of α and β, denoted $\hat{\alpha}$ and $\hat{\beta}$, are called the *least squares estimate (LSE)*.

Prop. 1.1.2 The least squares estimate for α and β are

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

Proof. We can find $\hat{\alpha}$ an $\hat{\beta}$ which minimizes the squared residuals $g(\alpha, \beta) = \sum_{i=1}^{n}(y_i - (\alpha + \beta x_i))^2$ by solving the simultaneous equations

$$\frac{\partial g}{\partial \alpha} = \frac{\partial}{\partial \alpha}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 = \sum_{i=1}^{n} 2(y_i - \alpha - \beta x_i)(-1) = 0 \implies \bar{y} - \alpha - \beta\bar{x} = 0$$

$$\frac{\partial g}{\partial \beta} = \frac{\partial}{\partial \beta}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 = \sum_{i=1}^{n} 2(y_i - \alpha - \beta x_i)(-x_i) = 0 \implies \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)(x_i) = 0$$

Substitute $\alpha = \bar{y} - \beta\bar{x}$ from (1) into (2), we get

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)x_i = 0 \implies \sum_{i=1}^{n}[y_i - (\bar{y} - \beta\bar{x}) - \beta x_i]x_i = \sum_{i=1}^{n}[y_i - \bar{y} - \beta(x_i - \bar{x})]x_i = 0.$$

Rearrange the equality, we get (1: Cancel x_i then multiply by $(x_i - \bar{x})$):

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} \overset{1}{=} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

and use $\alpha = \bar{y} - \beta\bar{x}$ to find $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$. □

Cor. 1.1.3 Let r be the sample correlation between x and y. Then

$$r = \hat{\beta}\sqrt{\frac{S_{xx}}{S_{yy}}}, \quad \hat{\beta} = r\sqrt{\frac{S_{yy}}{S_{xx}}}.$$

Proof. This follows directly from the definition of sample correlation and **Prop. 1.1.2**:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}, \quad -1 \le r \le 1$$

$$\text{where} \quad S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}_i). \quad \square$$

Ex. 1.1.4 Say we want to explore the relationship between Stat 230 and Stat 231 final grades. Suppose we determined $\hat{\beta} = 0.9944$ and $\alpha = -4.0667$ using calculations in **Note 1.1.2**. The fitted is therefore $y = -4.0667 + 0.9944x$. For example, if your final grade in Stat 230 was 75, then the least squares estimate of your final grade in Stat 231 is
$y = -4.0667 + 0.9944 \times 75 = 70.51$.

We will use this example throughout this section.

1.2 Simple Linear Regression

Ex. 1.2.1 Since not everyone who obtains a score of 75 in Stat 230 obtains a score of 70.51 in Stat 231, we want a statistical model that captures the uncertainty. More generally, it should model the variability in final grades for each Stat 230 final grade x.

Let Y be the Stat 231 final grade of a student drawn at random and Y_{75} denote the Stat 231 grade of a randomly chosen student who got a 75 in Stat 230. Assume $Y \sim G(\mu_{75}, \sigma_{75})$ where μ_{75} represents the mean Stat 231 final grade for students in the study population who obtained a final grade of 75 in Stat 230. We could then use this model along with the observed data for students who got a 75 in Stat 230 to obtain point and interval estimates for the mean μ_{75}.

Observe there exists a (roughly) linear relationship between Stat 230 and Stat 231 marks, i.e., the relationship between x and y doesn't seem to vary much as x changes. Thus, instead of having a different model for each "population" (one Y_m for each m), we can take advantage of the apparent relationship between x and y and propose a *single* model for ALL Stat 230 grades.

Let $\mu(x)$ represent the me mean Stat 231 final grade for students in the study population who obtained a final grade of x in Stat 230. We assume this takes a liner form in x: $\mu(x) = \alpha + \beta x$. For data (x_i, y_i), $i = 1, 2, \ldots, n$, we assume the model $Y_i \sim G(\alpha + \beta x_i, \sigma)$ for $i = 1, 2, \ldots, n$ independently and where the x_i $(i = 1, 2, \ldots, n)$ are assumed to be known constants.

This model is usually referred to as a *simple linear regression model*. Note we assumed that the standard deviation σ does not depend on x_i.

Note. 1.2.1 (Interpretation of Parameters) There are three unknown parameters in a simple linear regression model: α, β, σ.

- The parameter α represents the mean value of the response variate in the study population of individuals for whom their explanatory variate takes the value zero.
- The parameter β represents the increase in the mean value of the response variate in the study population for one unit increase in the value of the explanatory variate.
- Recall we assumed $Y_i \sim G(\alpha + \beta x_i, \sigma)$ for $i = 1, 2, \ldots, n$, so σ represents the variability in the response variate y in the study population for each value of the explanatory variate x.

Remark. In our example, β represents the *increase* in the mean Stat 231 final grade in the study population for one mark *increase* in Stat 230 final grade. *Be careful with the language here and try to avoid words like "change".*

1.3 Likelihood Function for α and β

In **Section 1.1**, we derived estimates for α and β using differentiation to minimize the squared residuals. We could also use maximum likelihood estimates.

Recall if we observed data from a distribution with p.d.f. $f(\mathbf{y};\theta)$ for $\theta \in \Sigma$, then the likelihood function θ based on the observed data $\mathbf{y} = (y_1, \ldots, y_n)$ is

$$L(\theta) = L(\theta; \mathbf{y}) = \prod_{i=1}^{n} f(y_i; \theta) \quad \theta \in \Omega.$$

Since our model is $Y_i \sim G(\alpha + \beta x_i, \sigma)$ for $i = 1, \ldots, n$ independently where the x_i $(i = 1, 2, \ldots, n$) are known constants, the likelihood function is

$$L(\alpha, \beta) = \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right) = \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right).$$

Assume (for now) that σ is known and ignoring constants with respect to α and β. To maximize $L(\alpha, \beta)$, we would minimize $\sum_{i=1}(y_i - \alpha - \beta x_i)^2$, but this is exactly the least squares problem! Thus, we get the following theorem, which states that *the maximum likelihood estimates are equivalent to the least squares estimates for simple linear regression models.*

Thm. 2.3.1 For the model $Y_i \sim G(\alpha + \beta x_i, \sigma)$ for $i = 1, \ldots, n$ independently where the x_i's are known constants, the MLE of α and β (often called the *regression parameters*) are given by

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

which are also the least squares estimates.

We have see that for a specific x, $y = \hat{\alpha} + \hat{\beta}x$ gives a point estimate of the mean response for an individual with observed explanatory variate x. To express uncertainty in this estimate, we want to consider both estimators for α and β.

Remark. We will focus on the following four steps for each unknown parameter β, α, μ:

1. Derive the maximum likelihood point estimator.
2. Find an appropriate pivotal pivotal quantity related to the estimator.
3. Construct confidence intervals using the pivotal quantity from step 2.
4. Test hypotheses related to the parameter of interest using the pivotal quantity from step 2.

1.4 Exploring the Slope

1.4.1 Estimator: Slope

From **Section 1.3**, the point estimate of β is

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) y_i.$$

The corresponding estimator is

$$\tilde{\beta} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) Y_i.$$

1.4.2 Pivotal Quantity: Slope

Since this is a linear combination of Gaussian random variables Y_i, by CLT, if $Y_i \sim G(\alpha + \beta x_i, \sigma)$ for $i = 1, \ldots, n$ independently where the x_i's are known constants, then the least squares estimator of β has distribution

$$\tilde{\beta} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) Y_i \sim G\left(\beta, \frac{\sigma}{\sqrt{S_{xx}}}\right).$$

Since σ^2 is usually unknown, we estimate it using the *mean squared error*

$$s_e^2 := \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 = \frac{1}{n-2} (S_{yy} - \hat{\beta} S_{xy}).$$

Note that s_e^2 is not the MLE of σ^2, but we use it to estimate σ^2 because $E[S_e^2] = \sigma^2$ where

$$S_e^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \tilde{\alpha} - \tilde{\beta} x_i)^2, \quad \tilde{\beta} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) Y_i, \quad \tilde{\alpha} = \overline{Y} - \tilde{\beta} - \bar{x}.$$

It can also be shown that $\dfrac{(n-2)S_e^2}{\sigma^2} \sim \chi^2_{n-2}$. Recall if $Z \sim G(0,1)$ and $U \sim \chi^2_k$ independently, then $T = \dfrac{Z}{\sqrt{U/k}} \sim t_k$. We can thus derive the following pivotal quantity to construct confidence intervals and test hypothesis for β:

$$\frac{\tilde{\beta} - \beta}{S_e / \sqrt{S_{xx}}} \sim t_{n-2}.$$

Warning. Note we have $n-2$ degrees of freedom here because the two unknowns x_i and y_i determine the MLE of α and β. Do not confuse this with the earlier chi-squared examples with $n-1$ degrees of freedom!

1.4.3 Confidence Interval: Slope

Using the pivotal quantity above, we can find a such that $P(-a \leq T \leq a) = p$ for $T \sim t_{n-2}$, so

$$p = P(-a \leq T \leq a) = P\left(-a \leq \frac{\tilde{\beta} - \beta}{S_e / \sqrt{S_{xx}}} \leq a\right).$$

We can rearrange this to help us find a $100p\%$ confidence interval for β:

$$P\left(\tilde{\beta} - a\frac{S_e}{\sqrt{S_{xx}}} \leq \beta \leq \tilde{\beta} + a\frac{S_e}{\sqrt{S_{xx}}}\right).$$

Hence, a $100p\%$ confidence interval for β is given by

$$\hat{\beta} \pm a\frac{s_e}{\sqrt{S_{xx}}}$$

where $P(T \leq a) = (1+p)/2$ and $T \sim t_{n-2}$.

1.4.4 Testing of Hypothesis: Slope

Define the test statistic for $H_0 : \beta = \beta_0$ using the pivotal quantity above:

$$D = \frac{|\tilde{\beta} - \beta_0|}{S_e/\sqrt{S_{xx}}}.$$

We know if $H_0 : \beta = \beta_0$ is true, then $D \sim T_{n-2}$. To test $H_0 : \beta = \beta_0$, the p-value is

$$\begin{aligned} p\text{-value} &= P(D \geq d; H_0) \\ &= P\left(\frac{|\tilde{\beta} - \beta_0|}{S_e/\sqrt{S_{xx}}} \geq \frac{|\hat{\beta} - \beta_0|}{s_e/\sqrt{S_{xx}}}\right) \\ &= P\left(|T| \geq \frac{|\hat{\beta} - \beta_0|}{s_e/\sqrt{S_{xx}}}\right) && T \sim t_{n-2} \\ &= 2\left[1 - P\left(T \leq \frac{|\hat{\beta} - \beta_0|}{s_e/\sqrt{S_{xx}}}\right)\right] && T \sim t_{n-2} \end{aligned}$$

Remark. There are a few interesting hypotheses.

1. $H_0 : \beta = 0$: "hypothesis of no linear relationship between the variates Y and x".
2. $H_1 : \beta = 1$: "hypothesis of perfect linear relationship between the variates Y and x".

1.5 Exploring the Mean Response

1.5.1 Estimator: Mean Response

Recall the point estimate for $\mu(x)$ is

$$\hat{\mu}(x) = \hat{\alpha} + \hat{\beta}x = (\overline{y} - \hat{\beta}\overline{x}) + \hat{\beta}x = \overline{y} + \hat{\beta}(x - \overline{x}).$$

The corresponding estimator is

$$\tilde{\mu}(x) = \overline{Y} + \tilde{\beta}(x - \overline{x}).$$

1.5.2 Pivotal Quantity: Mean Response

Let $Y_i \sim G(\alpha + \beta x_i, \sigma)$ for $i = 1, 2, \ldots, n$. It can be shown that:

$$\tilde{\mu}(x) = \sum_{i=1}^{n} \left(\frac{1}{n} + (x - \bar{x}) \frac{(x_i - \bar{x})}{S_{xx}} \right) Y_i \sim G\left(\mu(x), \sigma \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} \right)$$

where $\tilde{\mu}(x) = \tilde{\alpha} + \tilde{\beta}x$ and $\mu(x) = \alpha + \beta x$. Equivalently,

$$\frac{\tilde{\mu}(x) - \mu(x)}{\sigma \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim G(0,1)$$

Since we don't know σ, we use

$$\frac{\tilde{\mu}(x) - \mu(x)}{S_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim t_{n-2}$$

to construct confidence intervals for $\alpha + \beta x$.

1.5.3 Confidence Interval: Mean Response

A $100p\%$-confidence interval for $\mu(x) = \alpha + \beta x$, i.e., the mean response at x is

$$\hat{\mu}(x) \pm a s_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}} = \hat{\alpha} + \hat{\beta}x \pm a s_e \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

where $P(T \leq a) = (1+p)/2$ and $T \sim t_{n-2}$.

1.5.4 Confidence Interval: Intercept

Since $\mu(0) = \alpha + \beta(0) = \alpha$, a $100p\%$ confidence interval for α is given by

$$\hat{\alpha} \pm a s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}.$$

Note if $\bar{x}$ is large in magnitude (i.e., the x_i are typically large), the confidence interval for α will be very wide. This will be concerning if the value $x = 0$ is a value of interest, but often it is not.

1.6 Confidence Interval for an Individual Response Y at x

The confidence interval for $\mu(x)$ we constructed in **Section 1.5** answered the following question: *Based on my population of n observations, what is a plausible range of values for the average Stat 231 grade of all students who scored x in Stat 231?* The question we will now consider is: *Based on my population of n observations what is a plausible range of values for the Stat 231 grade of a new student who scored x in Stat 230?*

Let Y represent a potential observation for given value of x. We then have $Y = \mu(x) + R$ where $R \sim G(0, \sigma)$ independent of $Y_1, \ldots, Y_n$. We have established that $Y \sim G(\alpha + \beta x, \sigma)$ and

$$\tilde{\mu}(x) \sim G\left(\alpha + \beta x, \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right).$$

We now want the distribution of $Y - \tilde{\mu}(x)$, the error in the point estimator of Y.

Rewrite it as $Y - \tilde{\mu}(x) = Y - \mu(x) + \mu(x) - \tilde{\mu}(x) = R + [\mu(x) - \tilde{\mu}(x)]$, since R is independent of $\mu(x)$ (because it is not connected to the existing sample), the equation above is therefore the sum of independent, normally distributed random variables, so it is also normally distributed.

Let's calculate its mean and variance:

$$E[Y - \tilde{\mu}(x)] = E(R + [\mu(x) - \tilde{\mu}(x)]) = E(R) + E[\mu(x)] - E[\tilde{\mu}(x)] = 0 + \mu(x) - \mu(x) = 0$$
$$Var(Y - \tilde{\mu}(x)) = Var(Y) + Var[\tilde{\mu}(x)] = \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right] = \sigma^2\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]$$

Therefore, we get

$$Y - \tilde{\mu}(x) \sim G\left(0, \sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}\right).$$

Using a similar argument,

$$\frac{Y - \tilde{\mu}(x)}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim G(0,1).$$

Since we don't know σ, we use

$$\frac{Y - \tilde{\mu}(x)}{S_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}} \sim t_{n-2}.$$

The corresponding interval is

$$\hat{\alpha} + \hat{\beta}x \pm as_e\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

where $P(T \leq a) = (1+p)/2$ and $T \sim t_{n-2}$. We call this a $100p\%$ *prediction interval* instead of a confidence interval, because here Y is not a parameter by a "future" observation.

Remark. Compare the $100p\%$ prediction interval for a future observation Y and the $100p\%$ confidence interval for $\mu(x)$: the only different is the extra "1" inside the square root.

2 Gaussian Response Models and Model Checking

2.1 Gaussian Response Models

The simple linear models mentioned above belongs to a larger family of models.

Def. 2.1.1 A *Gaussian response model* is of the form $Y_i \sim G(\mu(\mathbf{x}_i), \sigma)$ for $i = 1, 2, \ldots, n$ independently where the $\mathbf{x}_i$, $i = 1, 2, \ldots, n$ are assumed to be known constants (possibly vectors).

Observe we assumed that the mean $E[Y_i] = \mu(\mathbf{x}_i)$ depends on the explanatory variate $\mathbf{x}_i$ but the standard deviation $sd(Y_i) = \sigma$ does not.

We can also write this as $Y_i = \mu(\mathbf{x}_i) + R_i$ where $R_i \sim G(0, \sigma)$, $i = 1, 2, \ldots, n$ independently. Framed this way, Y_i is the sum of two components:

1. $\mu(\mathbf{x}_i)$ is a deterministic component (i.e., not a random variable).
2. R_i is a random component (i.e., is a random variable).

In many examples, the deterministic component takes the form

$$E[Y_i] = \mu(\mathbf{x}_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

so that $E[Y_i]$ is a linear function of a vector of explanatory variates $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ for unit i and the unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$. We call these models *linear regression models*, in which the β_j are the *regression coefficients* and the x_i are called *covariates*.

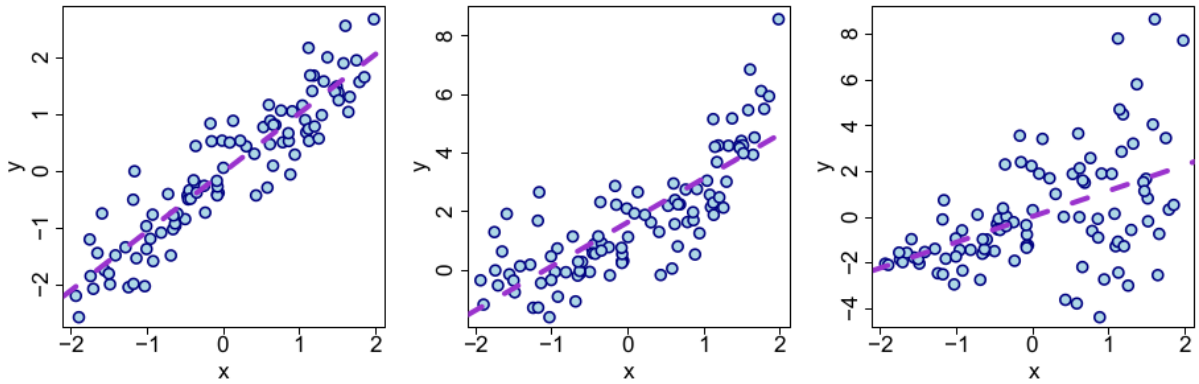
2.2 Model Checking

There are two main assumptions we make for Gaussian linear response models:

1. Y_i (given covariates x_i) has a Gaussian distribution with standard deviation σ which does not depend on the covariates.
2. $E[Y_i] = \mu(\mathbf{x}_i)$ is a linear combination of known covariates $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ and the unknown regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$.

We can check the assumptions in the following ways.

2.2.1 Method I: Scatter Plot with Fitted Line



A scatter plot of the data with the fitted line superimposed indicates how well the model fits the data. We can ask the following two questions to test our hypotheses:

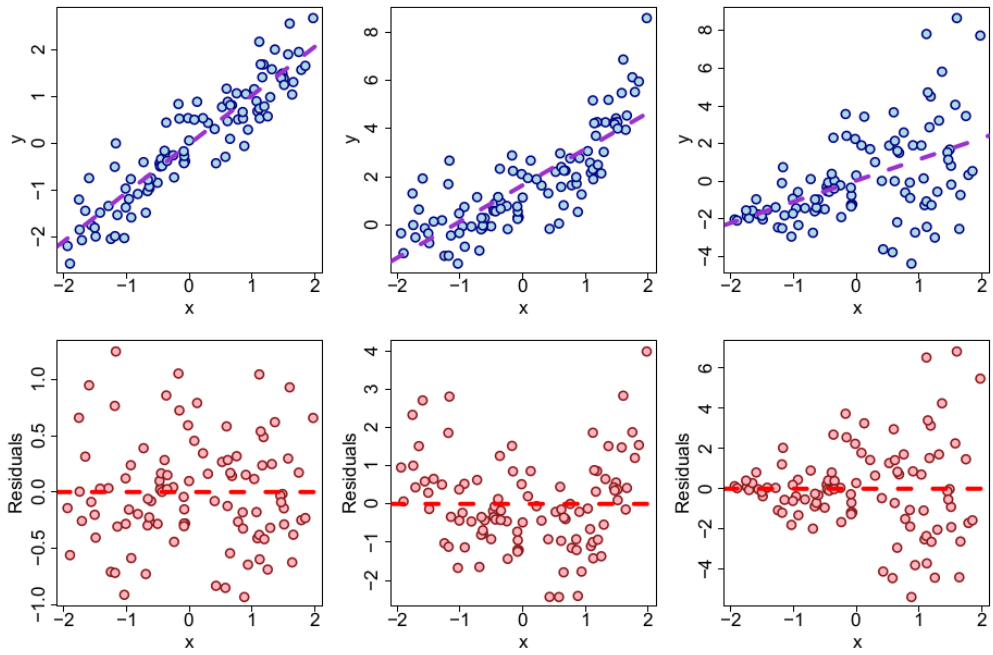
1. Do the points seem to fit reasonably along a straight line? (i.e., is $E[Y_i] = \mu(x)$ a linear function of x?)
2. Are the points generally "spread out" to the same extent regardless of x? (i.e., does σ depend on x?)

2.2.2 Method II(a): Residual Plots

If we have more than one covariate, scatter plots aren't super useful. Instead, we look at *residual plots*. For simple linear regression, define

- the *fitted response* of x_i as $\hat{\mu}_i = \hat{\alpha} + \hat{\beta} x_i$, and
- the *residual* after the model has been fitted to the data as $\hat{r}_i = y_i - \hat{\mu}_i$.

The $\hat{r}_i$ can be thought of as observed values of R_i in the model $Y_i = \mu_i + R_i$ where $R_i \sim G(0, \sigma)$, $i = 1, 2, \ldots, n$ independently. Then if the model is correct, residuals $\hat{r}_i$ should behave roughly like a random sample from a $G(0, \sigma)$ distribution. (It can be shown that the LSE of regression parameters implies $\sum_{i=1}^{n} \hat{r}_i = 0$, i.e., the average of our residuals is always zero.)



If the model assumptions hold, then a plot of the points $(x_i, \hat{r}_i)$ should like more or less within a horizontal band or belt around the line $\hat{r}_i = 0$ showing no obvious pattern.

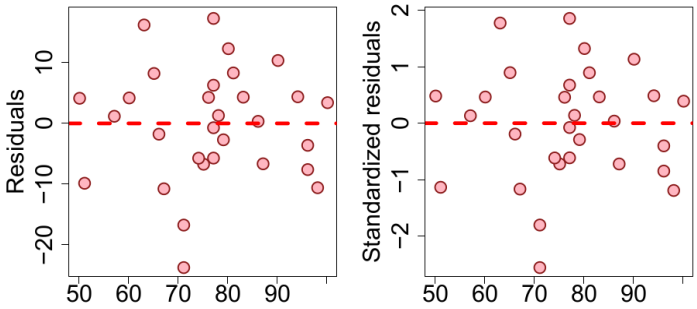
2.2.3 Method II(b): Standardized Residual Plots

Recall the $\hat{r}_i$ can be thought of as observed values of R_i in the model $Y_i = \mu_i + R_i$ where $R_i \sim G(0, \sigma)$, $i = 1, 2, \ldots, n$ independently. The variance in our residuals depends on σ, so different datasets will result in more/less variable residuals.

Define the *standardized residuals*

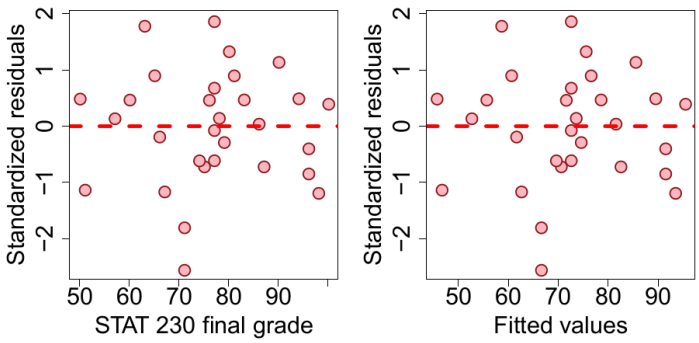
$$\hat{r}_i^* = \frac{\hat{r}_i}{s_e} = \frac{y_i - \hat{\mu}_i}{s_e} \quad i = 1, 2, \ldots, n.$$

If we plot $(x_i, \hat{r}_i^*)$ instead of $(x_i, \hat{r}_i)$, the plot will look the same, but be "scaled".



In fact, the $\hat{r}_i^*$ values should lie in the range $(-3, 3)$ as they'll be approximately $G(0, 1)$.

We could also plot $(\hat{\mu}_i, \hat{r}_i^*)$ where $\hat{\mu}_i = \hat{\beta}_0 + \sum_{j=1}^{n} \hat{\beta}_j x_{ij}$ are the fitted values. Such a plot can be used to check the assumption about the form of the mean $E[Y_i] = \mu(x_i)$.



If the assumption $E[Y_i] = \mu(x_i) = \alpha + \beta x_i$ is reasonable, then we should see approximately a horizontal band around the line $\hat{r}_i^* = 0$.

The key differences between residual (standard residual) plots and scatter plots are:

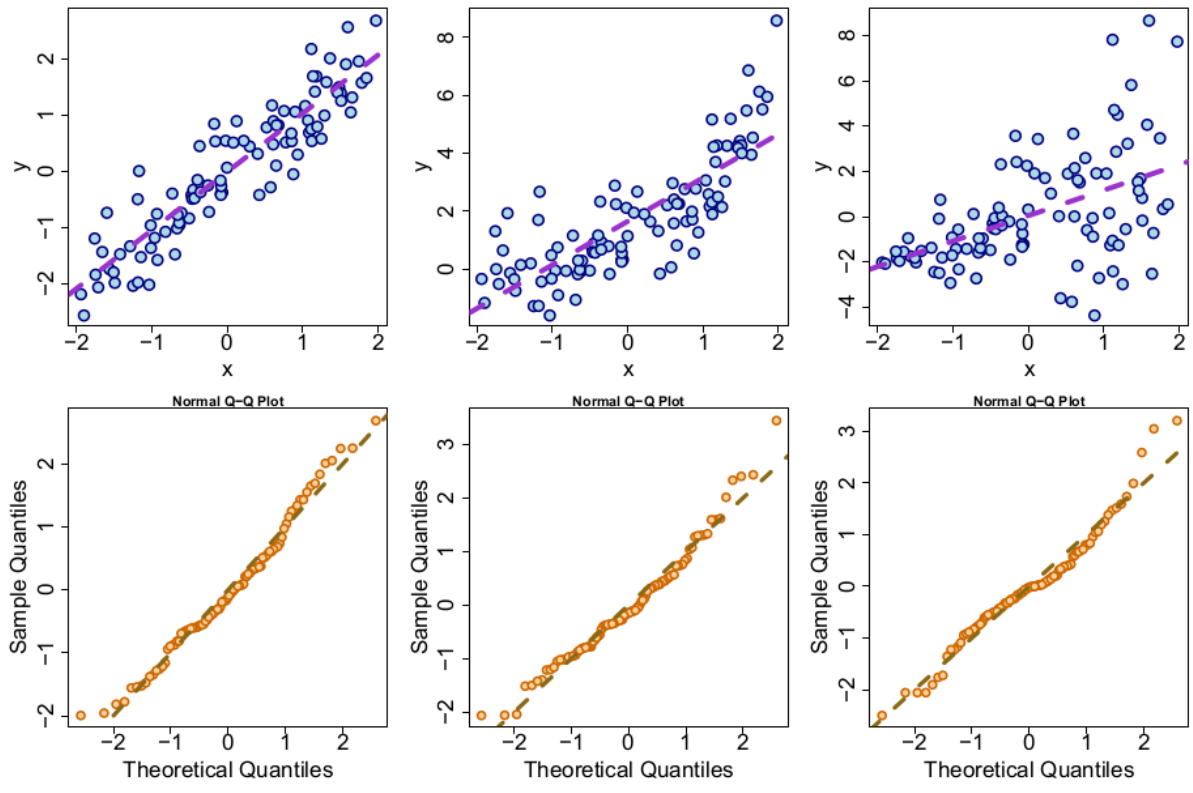
1. Residual plots are more general as they can be used when we have more than one covariate.
2. Residual plots can make visualization easier: assessing whether the points lie along a horizontal line rather than an angled line.

2.2.4 Method III: Residual QQ Plot

Since our assumed model (for the standardized residuals) is

$$\frac{R_i}{\sigma} = \frac{Y_i - \mu_i}{\sigma} \sim G(0,1),$$

the $\hat{r}_i^*$ should roughly represent a sample from the $G(0,1)$ distribution. Thus, we could also check the Gaussian assumption using a QQ plot.



If the model assumptions hold, we should see approximately a straight line.

2.2.5 Summary

Here is a useful summary for model checking:

1. In plots of the points $(\hat{\mu}_i, \hat{r}_i^*)$ or $(\hat{\mu}_i, \hat{r}_i)$:
    a. A distinctive pattern suggests that the assumed form for $E[Y_i] = \mu(x_i)$ may be inappropriate.
    b. If variability in $\hat{r}_i^*$ (or $\hat{r}_i$) is bigger for large values of $\hat{\mu}_i$ than for small values of $\hat{\mu}_i$ (or vice versa), then there is evidence to suggest that the assumption of constant variance $Var(V_i) = Var(R_i) = \sigma^2$ does not hold.

    Both of these can be assessed in a scatter plot as well in the case of simple linear regression, but it is usually more difficult to do so.
2. If the points in the Q-Q plot for residuals do not lie on a straight line, this suggests the Gaussian assumption may not hold.

3 Comparing Means of Two Populations

3.1 Two Sample Gaussian Problem (Equal Variance)

Suppose for sample sizes n_1 and n_2 we have

$$\begin{aligned} Y_{1i} &\sim G(\mu_1, \sigma) \quad \text{for } i = 1, 2, \ldots, n_1 \text{ independently} \\ Y_{2i} &\sim G(\mu_2, \sigma) \quad \text{for } i = 1, 2, \ldots, n_2 \text{ independently} \end{aligned}$$

It can be shown to be a special case of the Gaussian response model.

Suppose we want to test the hypothesis $H_0 : \mu_1 = \mu_2$. Rewrite this as $H_0 : \mu_1 - \mu_2 = 0$ so H_0 is in the "standard form" $H_0 : \theta = \theta_0$.

3.1.1 Estimator: Difference of Sample Means

Recall the maximum likelihood estimator of μ_1 and μ_2:

$$\tilde{\mu}_1 = \overline{Y_1} = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{1i}, \quad \tilde{\mu}_2 = \overline{Y_2} = \frac{1}{n_1} \sum_{i=1}^{n_2} Y_{2i}$$

A point estimator of the difference $\mu_1 - \mu_2$ is thus

$$\tilde{\mu}_1 - \tilde{\mu}_2 = \overline{Y}_1 - \overline{Y}_2.$$

3.1.2 Pivotal Quantity: Difference of Sample Means

Define estimators

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{1i} - \overline{Y}_1)^2, \quad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_{2i} - \overline{Y}_2)^2.$$

Observe S_1^2 is the point estimator of σ^2 based on only the Y_{1i} and S_1^2 is the point estimator of σ^2 based on only the Y_{2i}.

We define the *pooled estimator of variance*, which is obtained by "pooling" from the two estimators of σ^2 from the two samples:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (Y_{1i} - \overline{Y}_1)^2 + \sum_{i=1}^{n_2} (Y_{2i} - \overline{Y}_2)^2 \right)$$

Note that S_p^2 is not the maximum likelihood estimator of σ^2; we use it because $E[S_p^2] = \sigma^2$.

Recall $Y \sim G(\mu, \sigma) \implies \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. In the two population case, we have

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} \sim \chi^2_{n_1 + n_2 - 2}.$$

Using a similar argument as the one population case, we get

$$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

3.1.3 Confidence Interval: Difference of Sample Means

A $100p\%$ confidence interval for $\mu_1 - \mu_2$ is therefore given by

$$\overline{y}_1 - \overline{y}_2 \pm a s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $P(T \leq a) = \frac{1+p}{2}$ and $T \sim t_{n_1+n_2-2}$.

3.1.4 Hypothesis Testing: Difference of Sample Means

Suppose μ_1 and μ_2 are the *true* values of the population means. We frame our hypothesis test as "*Suppose* $\mu_1 - \mu_2 = 0$*, are we surprised by what we observed in our sample?*" That is, if the null hypothesis is $H_0 : \mu_1 - \mu_2 = 0$, then

$$\frac{\overline{Y}_1 - \overline{Y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

and we can see how unusual it would be for our observed value of this test statistic to be from such a t distribution.

Define the test statistic and observed value (since t distribution is symmetric, we can use the absolute value trick again):

$$D = \frac{\overline{Y}_1 - \overline{Y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad d = \frac{\overline{y}_1 - \overline{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

The p-value is then given by

$$p\text{-value} = 2[1 - P(T \leq d)]$$

where $T \sim t_{n_1+n_2-2}$.

3.1.5 Pivotal Quantity with Unequal SD: Difference of Sample Means

Now suppose we want to examine $H_0 : \sigma_1 = \sigma_2$ using a likelihood ratio test. If n_1 and n_2 are large, e.g., ≥ 30, then we can use the approximate pivotal quantity

$$\frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim G(0,1) \quad \text{approximately}$$

to construct confidence intervals and test hypotheses for the mean difference $\mu_1 - \mu_2$. For example, an approximate 95% confidence interval for $\mu_1 - \mu_2$ would be

$$\overline{y}_1 - \overline{y}_2 \pm 1.96\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

3.2 Paired Data

Often experimental studies designed to compare means are conducted with *pairs of units*, where the responses within a pair are not independent. For a paired experiment, if $Var(Y_{1i}) = \sigma_1^2$ and $Var(Y_{2i}) = \sigma_2^2$, then

$$Var(\overline{Y}_1 - \overline{Y}_2) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n} - 2Cov(Y_{1i}, Y_{2i}).$$

If $Cov(Y_{1i}, Y_{2i}) > 0$, then $Var(\overline{Y}_1 - \overline{Y}_2)$ is smaller for than for an unpaired experiment.

To make inferences about $\mu_1 - \mu_2$, we analyze the within-pair differences $Y_i = Y_{1i} - Y_{2i}$ for $i = 1, 2, \ldots, n$ by assuming $Y_i \sim G(\mu_1 - \mu_2, \sigma)$, $i = 1, \ldots, n$ independently. We can then use the one sample analysis we used previously for analyzing a random sample from a $G(\mu, \sigma)$ distribution, just with $\mu = \mu_1 - \mu_2$.

3.3 Summary

For tests of $H_0 : \mu_1 = \mu_2$ (or equivalently, $H_0 : \mu_1 - \mu_2 = 0$) where $Y_{1i} \sim G(\mu_1, \sigma)$ and $Y_{2i} \sim G(\mu_2, \sigma)$,

- Unpaired data:

$$d = \frac{|\overline{y}_1 - \overline{y}_2 - 0|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad p = 2[1 - P(T \leq d)], T \sim t_{n_1+n_2-2}.$$

- Paired data: define $Y_i = Y_{1i} - Y_{2i} \sim G(\mu, \sigma)$ and test $H_0 : \mu = 0$:

$$d = \frac{\overline{y} - 0}{s/\sqrt{n}}, \quad p = 2[1 - P(T \leq d)], T \sim t_{n-1}.$$