Probability and Statistics Notes series - Note Set 2

Regression Models

By: @0.5mins

Last update: September 30, 2025

Abstract

This note on regression model is developed based on Dr KY Liu's treatment in the course STAT3008 Applied Regression Analysis at CUHK, incorporating the textbook *Introduction to Linear Regression Analysis* by Montgomery, Peck and Vining, and other relevant materials. In this note, we will have a complete treatment of the implementation of regression models.

Contents

1 Simple linear regression

1.1 Interpolation

Suppose we have a data set $\{(x_i, y_i) : i = 1, 2, \cdots, n\} \subset \mathbb{R}^2$, where x-part is the independent variable and y-part is the dependent variable (so naturally x_i's are distinct). We wish to construct a model to predict the y values using the x values. A naive idea would be to find a polynomial $y = f(x)$ that passes through all data points, i.e. satisfies $y_i = f(x_i)$. To do so, we can use **Lagrange interpolation formula**. Define the **basis polynomials** by

$$f_i(x) = \prod_{\substack{k=1 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k}.$$

Note that they satisfy

$$f_i(x_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}.$$

Thus we have

$$y_i f_i(x_j) = \begin{cases} y_i & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}.$$

Summing up, we have the polynomial

$$f(x) = \sum_{i=1}^{n} y_i f_i(x)$$

which is a polynomial of degree $n - 1$ satisfying $y_i = f(x_i)$.

This is not a good model actually, because we usually obtain y_i by measurement, which means there is error. We need to control the error of the measurement from our model, but the Lagrange interpolation formula would take the error of the measurement into account as well. This is why we are going to study regression models in this note, because regression models also help us to handle the error using statistical methods.

1.2 Simple linear regression model

Let's build up the simple linear regression model. Consider a data set $\{(x_i, y_i) : i = 1, 2, \cdots, n\} \subset \mathbb{R}^2$, where x_i's are distinct. As lazy guys, we wish the data set to satisfy

$$y = \beta_0 + \beta_1 x$$

for some fixed constants β_i. Of course this is not always possible, so there should exist an error term ε such that

$$y = \beta_0 + \beta_1 x + \varepsilon. \tag{1}$$

This is called the **simple linear regression model**. In this equation,

- x is the independent variable, which is customarily called **regressor** variable in regression analysis, and
- y is the dependent variable, which is customarily called **response** variable in regression analysis.
- The **intercept** β_0 and the **slope** β_1 are unknown constants which we collectively refer them to as the **regression coefficients**.
- ε is the error term, which at this stage we assume $\varepsilon \sim N(0, \sigma^2)$ where σ is fixed.

The parameters β_i are unknown, and we need to estimate them by using sample data. This is where our data set comes into play. By our assumption, we may write

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

for which (x_i, y_i) are the sample data, and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. This is called the **sample regression model**.

In practice, the regressor x is controlled and measured with *negligible error*, while y is a random variable that we measure. Yes, y is a random variable, because $\varepsilon \sim N(0, \sigma^2)$, and β_i and x are fixed, so eventually $y \sim N(\beta_0 + \beta_1 x, \sigma^2)$:

$$\mathbb{E}[y|x] = \mathbb{E}[\beta_0 + \beta_1 x + \varepsilon] = \mathbb{E}[\beta_0] + \mathbb{E}[\beta_1 x] + \mathbb{E}[\varepsilon] = \beta_0 + \beta_1 x$$

$$\mathbb{V}(y|x) = \mathbb{V}(\beta_0 + \beta_1 x + \varepsilon) = \mathbb{V}(\varepsilon) = \sigma^2$$

Now we can see the interpretations of β_i:

- $\beta_0 = \mathbb{E}[y|x = 0]$ which is the mean of the distribution of the response y when $x = 0$, and
- $\beta_1 = [\beta_0 + \beta_1(x + 1)] - [\beta_0 + \beta_1 x] = \mathbb{E}[y|x + 1] - \mathbb{E}[y|x]$ which is the change in the mean of the distribution of y produced by a unit change in x.

1.2.1 Least-squares estimation

The **least-squares estimators** $\hat{\beta}_i$ of β_i are defined to be the minimizer of the sum of squares

$$S = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

What are the summands $(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$? Recall that y_i are the observed values of the response variable in our data set, and $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is called the **fitted simple linear regression model**. Using the fitted simple linear regression model, we can obtain a predicted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ for each x_i. The gap between y_i and $\hat{y}_i$ is called **residual**, which is denoted by

$$e_i = y_i - \hat{y}_i.$$

This e_i can be viewed as the *signed distance* from y_i to the corresponding point on the fitted simple linear regression model with the same x_i value. To minimize the *unsigned distance*, we consider the **sum of squares**. We prefer squares over absolute values because squares are smooth, but absolute values are not smooth and nonlinear.

Okay, after explaining the choice of estimator, we finally start deriving it. Here we provide an analytical method of derivation, and in a later section we will derive it by linear algebraic methods. The least-squares estimators $\hat{\beta}_i$ satisfy

$$\frac{\partial S}{\partial \hat{\beta}_0} = \frac{\partial S}{\partial \hat{\beta}_1} = 0 \tag{2}$$

as given by first derivative test. We compute the derivatives as follows:

$$\frac{\partial S}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = -2 \left\{ \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i \right\}$$

$$\frac{\partial S}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = -2 \left\{ \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \right\}$$

Thus we have

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \quad \text{and} \quad \sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2.$$

These equations are called the **least-squares normal equations**. Looking at the first normal equation, if we consider the sample means $\overline{x} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i$ and $\overline{y} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} y_i$, then it becomes

$$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x} \quad \Longrightarrow \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

Putting it into the second normal equation, we have

$$\begin{aligned} \sum_{i=1}^{n} x_i y_i &= (\overline{y} - \hat{\beta}_1 \overline{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\ &= n\overline{x}(\overline{y} - \hat{\beta}_1 \overline{x}) + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\ &= n\overline{x}\ \overline{y} + \hat{\beta}_1 \left\{ \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \right\} \end{aligned}$$

and hence

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\ \overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}.$$

Thus we have already obtained the least-squares estimators as

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\ \overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}. \tag{3}$$

Example 1.1. A study investigated whether the average number of tweets (or messages) per hour prior to the movie's release on Twitter.com could be used to forecast the opening weekend box office revenues of movies. The two variables of a sample of 23 movies were measured. Consider the simple linear regression model:

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

where y_i are the weekend box office revenues and x_i are the average number of tweets per hour. It is given that

$$\sum_{i=1}^{n} x_i^2 = 4933199 \quad \sum_{i=1}^{n} y_i^2 = 35626.09 \quad \sum_{i=1}^{n} x_i y_i = 396603.2$$
$$\sum_{i=1}^{n} x_i = 6980.65 \quad \sum_{i=1}^{n} y_i = 576.3$$

Compute the least-squares estimators of β_0 and β_1. Are the signs of the estimators sensible, based on the practical meanings of the regression coefficients?

Solution.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\ \overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{(396603.2) - (23)\left(\frac{6980.65}{23}\right)\left(\frac{576.3}{23}\right)}{(4933199) - (23)\left(\frac{6980.65}{23}\right)^2} \approx 0.07876722$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \approx \left(\frac{576.3}{23}\right) - (0.07876722)\left(\frac{6980.65}{23}\right) \approx 1.150158$$

- β_0 is the average weekend box office revenues when the tweet rate is 0. So the sign of $\hat{\beta}_0$ makes sense because, even no one talks about a movie, there should be someone who watched the movie.
- β_1 is the change in average weekend box office revenues when the tweet rate is increased by 1 unit. So the sign of $\hat{\beta}_1$ makes sense because more people talks about a movie implies the movie is more well-known, and so it makes sense that more people would watch the movie. ■

The formula we have for $\hat{\beta}_1$ looks rather messy. To write down a cleaner formula, let

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 \quad \text{and} \quad S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}). \tag{4}$$

Then

$$\begin{aligned} S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 &= \sum_{i=1}^{n} (x_i^2 - 2\overline{x}x_i + \overline{x}^2) \\ &= \sum_{i=1}^{n} x_i^2 - 2\overline{x}\sum_{i=1}^{n} x_i + n\overline{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2\overline{x}(n\overline{x}) + n\overline{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \end{aligned}$$

which is the denominator of $\hat{\beta}_1$. Similarly,

$$\begin{aligned} S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) &= \sum_{i=1}^{n} (x_i y_i - \overline{x}y_i - x_i\overline{y} + \overline{x}\ \overline{y}) \\ &= \sum_{i=1}^{n} x_i y_i - \overline{x}\sum_{i=1}^{n} y_i - \overline{y}\sum_{i=1}^{n} x_i + n\overline{x}\ \overline{y} \\ &= \sum_{i=1}^{n} x_i y_i - \overline{x}(n\overline{y}) - \overline{y}(n\overline{x}) + n\overline{x}\ \overline{y} \\ &= \sum_{i=1}^{n} x_i y_i - n\overline{x}\ \overline{y} \end{aligned}$$

which is indeed the numerator of $\hat{\beta}_1$. Therefore we have

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}. \tag{5}$$

1.2.2 Properties of the least-squares estimators

As said earlier, we choose to consider the sum of squares because it satisfies a range of clean properties, which we showcase as follows.

Proposition 1.1. $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combination of y_i's.

Proof. Recall that

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{S_{xx}}.$$

We manipulate the numerator S_{xy} as follows:

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

Therefore

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{S_{xx}} \\ &= \frac{\sum_{i=1}^n \{y_i(x_i - \overline{x})\}}{S_{xx}} \\ &= \sum_{i=1}^n \left\{ \frac{x_i - \overline{x}}{S_{xx}} \right\} y_i \\ &= \sum_{i=1}^n c_i y_i \qquad \text{where } c_i = \frac{x_i - \overline{x}}{S_{xx}} \end{aligned}$$

and hence

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = \frac{1}{n} \sum_{i=1}^n y_i - \overline{x} \sum_{i=1}^n c_i y_i = \sum_{i=1}^n \left\{ \frac{1}{n} - \overline{x} c_i \right\} y_i.$$

■

It is noteworthy that we have also deduced $\sum_{i=1}^n (x_i - \overline{x}) = 0$ in the middle of the proof, and similarly, we can deduce that $\sum_{i=1}^n (y_i - \overline{y}) = 0$. These facts will often be used in our proofs.

Using the linear property of $\hat{\beta}_i$ as well as the aforementioned fact, we can show that

Theorem 1.2. $\hat{\beta}_0$ and $\hat{\beta}_1$ are the unbiased estimators of the parameters β_0 and β_1 respectively.

Proof. First, by linearity, we can observe that

$$\begin{aligned} \mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\sum_{i=1}^n c_i y_i \right] &= \sum_{i=1}^n c_i \mathbb{E}[y_i] \\ &= \sum_{i=1}^n c_i \mathbb{E}[\beta_0 + \beta_1 x_i + \varepsilon_i] \\ &= \sum_{i=1}^n c_i \left\{ \mathbb{E}[\beta_0 + \beta_1 x_i] + \mathbb{E}[\varepsilon_i] \right\} \\ &= \sum_{i=1}^n c_i (\beta_0 + \beta_1 x_i) \\ &= \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i x_i \end{aligned}$$

Our goal is to show $\mathbb{E}[\hat{\beta}_1] = \beta_1$, which is equivalent to showing that $\sum_{i=1}^{n} c_i = 0$ and $\sum_{i=1}^{n} c_i x_i = 1$.

$$\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} \frac{x_i - \overline{x}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{S_{xx}} = 0$$

where here we used the fact that $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$. On the other hand,

$$\begin{aligned}
\sum_{i=1}^{n} c_i x_i = \sum_{i=1}^{n} c_i x_i - \overline{x}\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} c_i(x_i - \overline{x}) \\
&= \sum_{i=1}^{n} \left\{ \frac{x_i - \overline{x}}{S_{xx}} \right\} (x_i - \overline{x}) \\
&= \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{S_{xx}} \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x})^2 \\
&= \frac{S_{xx}}{S_{xx}} = 1
\end{aligned}$$

Thus we have shown that $\mathbb{E}[\hat{\beta}_1] = \beta_1$, which means $\hat{\beta}_1$ is an unbiased estimator of β_1. We can deduce $\mathbb{E}[\hat{\beta}_0] = \beta_0$ in a similar way as follows. Recall

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left\{ \frac{1}{n} - \overline{x} c_i \right\} y_i := \sum_{i=1}^{n} d_i y_i \qquad \text{where } d_n := \frac{1}{n} - \overline{x} c_i.$$

Then we can derive that

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_0] = \mathbb{E}\left[\sum_{i=1}^{n} d_i y_i \right] = \sum_{i=1}^{n} d_i \mathbb{E}[y_i] \\
&= \sum_{i=1}^{n} d_i \mathbb{E}[\beta_0 + \beta_1 x_i + \varepsilon_i] \\
&= \sum_{i=1}^{n} d_i \left\{ \mathbb{E}[\beta_0 + \beta_1 x_i] + \mathbb{E}[\varepsilon_i] \right\} \\
&= \sum_{i=1}^{n} d_i (\beta_0 + \beta_1 x_i) \\
&= \beta_0 \sum_{i=1}^{n} d_i + \beta_1 \sum_{i=1}^{n} d_i x_i
\end{aligned}$$

and so here we need to show $\sum_{i=1}^{n} d_i = 1$ and $\sum_{i=1}^{n} d_i x_i = 0$.

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} \left\{ \frac{1}{n} - \overline{x} c_i \right\} = 1 - \overline{x} \sum_{i=1}^{n} c_i = 1$$

$$\sum_{i=1}^{n} d_i x_i = \sum_{i=1}^{n} \left\{ \frac{1}{n} - \overline{x} c_i \right\} x_i = \frac{1}{n} \sum_{i=1}^{n} x_i - \overline{x} \sum_{i=1}^{n} c_i x_i = \overline{x} - \overline{x}(1) = 0$$

and we have already completed the proof that $\mathbb{E}[\hat{\beta}_0] = \beta_0$. ■

We can of course compute $\mathbb{V}(\hat{\beta}_i)$ in a similar fashion. Recall the following formulas: if $T = \sum_{i=1}^{n} a_i X_i$ and $W = \sum_{j=1}^{m} b_j Y_j$, then

$$\operatorname{Cov}(T, W) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \operatorname{Cov}(X_i, Y_j),$$

moreover,

$$\mathbb{V}(T) = \operatorname{Cov}(T, T) = \sum_{i=1}^{n} a_i^2 \mathbb{V}(X_i) + 2 \sum_{i<j} a_i a_j \operatorname{Cov}(X_i, X_j).$$

Proposition 1.3. $\mathbb{V}(\hat{\beta}_1) = \dfrac{\sigma^2}{S_{xx}}$ and $\mathbb{V}(\hat{\beta}_0) = \sigma^2 \left\{ \dfrac{1}{n} + \dfrac{\overline{x}^2}{S_{xx}} \right\}$

Proof. Recall that

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$$

Furthermore, as $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, we know $\operatorname{Cov}(\varepsilon_i, \varepsilon_j) = 0 \ \forall i \neq j$. Thus for $i \neq j$, we have

$$\begin{aligned} \operatorname{Cov}(y_i, y_j) &= \operatorname{Cov}(\beta_0 + \beta_1 x_i + \varepsilon_i, \beta_0 + \beta_1 x_j + \varepsilon_j) \\ &= \operatorname{Cov}(\varepsilon_i, \varepsilon_j) = 0 \end{aligned}$$

because $\beta_0 + \beta_1 x_i$ and $\beta_0 + \beta_1 x_j$ are constants. Hence we can compute that

$$\begin{aligned} \mathbb{V}(\hat{\beta}_1) = \sum_{i=1}^{n} c_i^2 \mathbb{V}(y_i) &= \sigma^2 \sum_{i=1}^{n} c_i^2 \\ &= \sigma^2 \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{S_{xx}^2} \\ &= \frac{\sigma^2}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \overline{x})^2 \\ &= \frac{\sigma^2}{S_{xx}^2} S_{xx} \\ &= \frac{\sigma^2}{S_{xx}} \end{aligned}$$

Now recall that $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$, and so

$$\mathbb{V}(\hat{\beta}_0) = \mathbb{V}(\overline{y} - \hat{\beta}_1 \overline{x}) = \mathbb{V}(\overline{y}) + \overline{x}^2 \mathbb{V}(\hat{\beta}_1) + 2\overline{x} \operatorname{Cov}(\overline{y}, \hat{\beta}_1)$$

Beware that $\overline{x}$ is a fixed constant. Next, note that $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, so

$$\mathbb{V}(\overline{y}) = \mathbb{V}\left(\frac{1}{n} \sum_{i=1}^{n} y_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{V}(y_i) = \frac{1}{n^2} (n\sigma^2) + 0 = \frac{\sigma^2}{n}.$$

On the other hand, we can compute that

$$\begin{aligned} \operatorname{Cov}(\overline{y}, \hat{\beta}_1) = \operatorname{Cov}\left(\frac{1}{n} \sum_{i=1}^{n} y_i, \sum_{j=1}^{n} c_j y_j \right) &= \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{c_j}{n} \operatorname{Cov}(y_i, y_j) \\ &= \sum_{i=1}^{n} \frac{c_i}{n} \mathbb{V}(y_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} c_i \sigma^2 \\ &= \frac{\sigma^2}{n} \sum_{i=1}^{n} c_i = 0 \end{aligned}$$

Putting these ingredients back, we have

$$\mathbb{V}(\hat{\beta}_0) = \mathbb{V}(\overline{y}) + \overline{x}^2 \mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{n} + \overline{x}^2 \frac{\sigma^2}{S_{xx}} = \sigma^2 \left\{ \frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right\}.$$

■

Remark. Note that y_i's are not iid, so the proof are not as clean as we expect!

In the followng we will state a theorem regarding the quality of the least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$:

Theorem 1.4 (Gauss-Markov theorem)**.** The least-squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and have minimum variance among all unbiased linear estimators.

We will prove it in a later section. A fun fact is that we often call these least-squares estimators **BLUE** - best unbiased linear estimator. By 'best', we mean they have the minimum variance. We will revisit this point in a greater generality later.

There are many other useful properties about the fitted simple linear regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ that are noteworthy:

(1) The sum of residuals $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$ vanishes, i.e.

$$\sum_{i=1}^{n} e_i = 0.$$

Proof. Recall $S = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$, so

$$\frac{\partial S}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = -2 \sum_{i=1}^{n} e_i.$$

Recall the least-squares condition $\dfrac{\partial S}{\partial \hat{\beta}_0} = 0$, so

$$\sum_{i=1}^{n} e_i = \frac{1}{-2} \frac{\partial S}{\partial \hat{\beta}_0} = 0$$

■

(2) The sum of observed values y_i and the sum of the fitted values $\hat{y}_i$ are equal, i.e.

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i.$$

This is immediate from property (1).

(3) The fitted model always pass through the **centroid** $(\overline{x}, \overline{y})$, i.e.

$$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}.$$

This is the geometric meaning of the first normal equation.

(4) $\sum_{i=1}^{n} x_i e_i = 0$

Proof. Recall $S = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$, so

$$\frac{\partial S}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = -2 \sum_{i=1}^{n} x_i e_i.$$

Recall the least-squares condition $\dfrac{\partial S}{\partial \hat{\beta}_1} = 0$, so

$$\sum_{i=1}^{n} x_i e_i = \frac{1}{-2} \frac{\partial S}{\partial \hat{\beta}_1} = 0$$

■

(5) $\sum_{i=1}^{n} \hat{y}_i e_i = 0$

Proof. $\sum_{i=1}^{n} \hat{y}_i e_i = \sum (\hat{\beta}_0 + \hat{\beta}_1 x_i) e_i = \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i e_i = 0$ ■

Problem 1.2 (STAT3008 (2023/24 S2) Midterm Q4)**.** Consider the simple linear regression model

$$y_i = \beta(x_i - \overline{x}) + \varepsilon_i$$

where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, 2, \cdots, 100$. A student proposed the following estimator of β:

$$\hat{\beta}^* = \frac{y_{100} - y_1}{x_{100} - x_1}.$$

(a) Show that $\hat{\beta}^*$ is an unbiased estimator of β.

(b) Show that $\hat{\beta}^*$ is a linear estimator of β.

(c) Find $\mathbb{V}(\hat{\beta}^*)$.

(d) Derive the least-squares estimator $\hat{\beta}_{LS}$ of β.

(e) Is there any advantage of using $\hat{\beta}_{LS}$ instead of $\hat{\beta}^*$?

Solution.

(a) Note that $y_i \sim N(\beta(x_1 - \overline{x}), \sigma^2)$, thus

$$\begin{aligned}
\mathbb{E}[\hat{\beta}^*] &= \mathbb{E}\left[\frac{y_{100} - y_1}{x_{100} - x_1}\right] \\
&= \frac{\mathbb{E}[y_{100}] - \mathbb{E}[y_1]}{x_{100} - x_1} \\
&= \frac{[\beta(x_{100} - \overline{x})] - [\beta(x_1 - \overline{x})]}{x_{100} - x_1} \\
&= \frac{\beta(x_{100} - x_1)}{x_{100} - x_1} = \beta
\end{aligned}$$

so $\hat{\beta}^*$ is an unbiased estimator of β.

(b) Clearly $\hat{\beta}^* = \left(\frac{-1}{x_{100} - x_1}\right) y_1 + \left(\frac{1}{x_{100} - x_1}\right) y_{100}$, so $\hat{\beta}^*$ is a linear estimator of β.

(c) Compute that

$$\mathbb{V}(\hat{\beta}^*) = \mathbb{V}\left(\frac{y_{100} - y_1}{x_{100} - x_1}\right) = \frac{\mathbb{V}(y_{100}) + V(y_1) - 2\operatorname{Cov}(y_1, y_{100})}{(x_{100} - x_1)^2} = \frac{2\sigma^2 - 2\operatorname{Cov}(y_1, y_{100})}{(x_{100} - x_1)^2}.$$

Now note that $\varepsilon_1, \varepsilon_{100}$ are independent, so

$$\operatorname{Cov}(y_1, y_{100}) = \operatorname{Cov}(\beta(x_1 - \overline{x}) + \varepsilon_1, \beta(x_{100} - \overline{x}) + \varepsilon_{100}) = \operatorname{Cov}(\varepsilon_1, \varepsilon_{100}) = 0.$$

Therefore

$$\mathbb{V}(\hat{\beta}^*) = \frac{2\sigma^2}{(x_{100} - x_1)^2}.$$

(d) First form the sum of squares

$$S = \sum_{i=1}^{100} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{100} (y_i - \hat{\beta}_{LS}(x_i - \overline{x}))^2.$$

The least-sqaures estimator $\hat{\beta}_{LS}$ is the minimizer of S. To find $\hat{\beta}_{LS}$, we solve

$$\frac{dS}{d\hat{\beta}_{LS}} = 0.$$

Compute that $\frac{dS}{d\hat{\beta}_{LS}} = -2\sum_{i=1}^{100} (x_i - \overline{x})(y_i - \hat{\beta}_{LS}(x_i - \overline{x}))$. Thus

$$\begin{aligned}
\sum_{i=1}^{100} (x_i - \overline{x})(y_i - \hat{\beta}_{LS}(x_i - \overline{x})) = 0 \quad &\Longrightarrow \quad \sum_{i=1}^{100} y_i(x_i - \overline{x}) - \hat{\beta}_{LS} \sum_{i=1}^{100} (x_i - \overline{x})^2 = 0 \\
&\Longrightarrow \quad \sum_{i=1}^{100} (y_i - \overline{y})(x_i - \overline{x}) - \hat{\beta}_{LS} \sum_{i=1}^{100} (x_i - \overline{x})^2 = 0 \\
&\Longrightarrow \quad S_{xy} - \hat{\beta}_{LS} S_{xx} = 0 \\
&\Longrightarrow \quad \hat{\beta}_{LS} = \frac{S_{xy}}{S_{xx}}
\end{aligned}$$

(e) By Gauss-Markov theorem, $\hat{\beta}_{LS}$ is the best linear unbiased estimator of β. By 'best' we mean $\hat{\beta}_{LS}$ is the linear unbiased estimator of β with minimal variance. In other words, $\mathbb{V}(\hat{\beta}_{LS}) < \mathbb{V}(\hat{\beta}^*)$. ■

Problem 1.3 (Modified from STAT3008 (2025/26 S1) Tut 2 Q1)**.** Consider the simple linear regression model

$$y_i = \beta x_i + \varepsilon_i$$

where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ for $i = 1, 2, \cdots, n$.

(a) Derive the least-squares estimator $\hat{\beta}$ of β.

(b) Show that $\hat{\beta}$ is an unbiased estimator of β.

(c) Show that the fitted regression line passes through the point

$$(\overline{x^2}, \overline{xy}) = \left(\frac{1}{n}\sum_{i=1}^n x_i^2, \frac{1}{n}\sum_{i=1}^n x_i y_i \right),$$

but the line does not pass through the point $(\overline{x}, \overline{y})$.

Solution.

(a) First form the sum of squares

$$S = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta} x_i)^2.$$

The least-sqaures estimator $\hat{\beta}$ is the minimizer of S. To find $\hat{\beta}$, we solve

$$\frac{dS}{d\hat{\beta}} = 0.$$

Compute that $\dfrac{dS}{d\hat{\beta}} = -2\sum_{i=1}^n (y_i - \hat{\beta} x_i) x_i = 2\hat{\beta}\sum_{i=1}^n x_i^2 - 2\sum_{i=1}^n x_i y_i$. Thus

$$2\hat{\beta}\sum_{i=1}^n x_i^2 - 2\sum_{i=1}^n x_i y_i = 0 \quad \Longrightarrow \quad \hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

(b) Note that $\mathbb{E}[y_i] = \mathbb{E}[\beta x_i + \varepsilon_i] = \beta x_i + \mathbb{E}[\varepsilon_i] = \beta x_i$. Hence

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}\left[\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \right] = \sum_{i=1}^n \left(\frac{x_i}{\sum_{\nu=1}^n x_\nu^2} \right) \mathbb{E}[y_i] = \sum_{i=1}^n \left(\frac{x_i}{\sum_{\nu=1}^n x_\nu^2} \right) (\beta x_i) = \beta \frac{\sum_{i=1}^n x_i^2}{\sum_{\nu=1}^n x_\nu^2} = \beta.$$

Therefore $\hat{\beta}$ is an unbiased estimator.

(c) The fitted regression line is $\hat{y} = \hat{\beta} x$. Compute that

$$\hat{\beta}\overline{x^2} = \left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \right) \left(\frac{1}{n}\sum_{i=1}^n x_i^2 \right) = \frac{1}{n}\sum_{i=1}^n x_i y_i = \overline{xy},$$

so the fitted regression line passes through $(\overline{x^2}, \overline{xy})$. On the other hand,

$$\begin{aligned}
\overline{y} - \hat{\beta}\overline{x} &= \frac{1}{n}\sum_{i=1}^n y_i - \left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \right) \left(\frac{1}{n}\sum_{i=1}^n x_i \right) \\
&= \frac{1}{n}\sum_{i=1}^n y_i - \frac{1}{n}\sum_{i=1}^n \left(\frac{x_i \sum_{\nu=1}^n x_\nu}{\sum_{\nu=1}^n x_\nu^2} \right) y_i \\
&= \frac{1}{n}\sum_{i=1}^n \left(\frac{\sum_{\nu=1}^n x_\nu^2 - x_i \sum_{\nu=1}^n x_\nu}{\sum_{\nu=1}^n x_\nu^2} \right) y_i \\
&= \frac{1}{n}\sum_{i=1}^n \left(\frac{\sum_{\nu \neq i} x_\nu (x_\nu - x_i)}{\sum_{\nu=1}^n x_\nu^2} \right) y_i \neq 0
\end{aligned}$$

Therefore the fitted regression line does not pass through $(\overline{x}, \overline{y})$.

■

1.3 Point estimation about σ^2

Recall the model assumes $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, but this σ^2 is often unknown, so we would need to find some point estimators of σ^2.

If we are *able to observe* $\varepsilon_i = y_i - (\beta_0 + \beta_1 x) = y_i - \mathbb{E}[y_i]$, then an unbiased estimator of σ^2 is $\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2$:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\varepsilon_i^2] = \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{V}(\varepsilon_i) - (\mathbb{E}[\varepsilon_i])^2\right\} = \frac{1}{n}(n\sigma^2) = \sigma^2.$$

However, this assumption is too good to be true. In fact, we can *almost never observe* ε_i. In this case, we consider the **residual sum of squares**

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

We first simplify the expression of SS_{Res} as follows:

$$\begin{aligned}
SS_{Res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
&= \sum_{i=1}^{n}(y_i - (\overline{y} - \hat{\beta}_1\overline{x}) - \hat{\beta}_1 x_i)^2 \\
&= \sum_{i=1}^{n}((y_i - \overline{y}) - \hat{\beta}_1(x_i - \overline{x}))^2 \\
&= \sum_{i=1}^{n}(y_i - \overline{y})^2 - 2\hat{\beta}_1\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x}) + \hat{\beta}_1^2\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&= \sum_{i=1}^{n}(y_i - \overline{y})^2 - 2\hat{\beta}_1 S_{xy} + \hat{\beta}_1^2 S_{xx} \\
&= \sum_{i=1}^{n}(y_i - \overline{y})^2 - 2\hat{\beta}_1 S_{xy} + \hat{\beta}_1\left(\frac{S_{xy}}{S_{xx}}\right)S_{xx} \\
&= \sum_{i=1}^{n}(y_i - \overline{y})^2 - 2\hat{\beta}_1 S_{xy} + \hat{\beta}_1 S_{xy} \\
&= \sum_{i=1}^{n}(y_i - \overline{y})^2 - \hat{\beta}_1 S_{xy} \\
&= \sum_{i=1}^{n}(y_i - \overline{y})^2 - \hat{\beta}_1(\hat{\beta}_1 S_{xx}) \\
&= \sum_{i=1}^{n}(y_i - \overline{y})^2 - \hat{\beta}_1^2 S_{xx}
\end{aligned}$$

We call $\sum_{i=1}^{n}(y_i - \overline{y})^2$ the **total sum of squares**, and write

$$SS_T = \sum_{i=1}^{n}(y_i - \overline{y})^2,$$

that is

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SS_T - \hat{\beta}_1 S_{xy} = SS_T - \hat{\beta}_1^2 S_{xx} \tag{6}$$

Just like how we show $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$, we can show that

$$SS_T = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2.$$

It is useful to remember that

Theorem 1.5. $\dfrac{SS_{Res}}{\sigma^2} \sim \chi^2(n-2)$.

The proof will be supplied when we introduce the matrix notations. Now recall that if $X \sim \chi^2(r)$, then $\mathbb{E}[X] = r$ and $\mathbb{V}(X) = 2r$. That means

$$\mathbb{E}\left[\frac{SS_{Res}}{\sigma^2}\right] = n-2 \quad \Longrightarrow \quad \mathbb{E}[SS_{Res}] = (n-2)\sigma^2.$$

In other words, SS_{Res} is a biased estimator of σ^2. On the other hand, if we define the **residuals mean squares** as

$$MS_{Res} = \frac{SS_{Res}}{n-2},$$

then we can deduce that

$$\mathbb{E}[MS_{Res}] = \sigma^2$$

which gives an unbiased estimator of σ^2.

Example 1.4. Let's continue our previous example about the relationship between tweet rate and box office revenue, where we used the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \cdots, 23$$

where y_i are the weekend box office revenues and x_i are the average number of tweets per hour. It is given that

$$\sum_{i=1}^{n} x_i^2 = 4933199 \quad \sum_{i=1}^{n} y_i^2 = 35626.09 \quad \sum_{i=1}^{n} x_i y_i = 396603.2$$
$$\sum_{i=1}^{n} x_i = 6980.65 \quad \sum_{i=1}^{n} y_i = 576.3$$

We have already computed that $\hat{\beta}_0 \approx 1.150158$ and $\hat{\beta}_1 \approx 0.07876722$. Now estimate σ^2 of the error term ε_i.

Solution. An estimate is given by

$$\begin{aligned}
& MS_{Res} \\
&= \frac{SS_{Res}}{n-2} \\
&= \frac{1}{n-2}\left\{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2 - \hat{\beta}_1 S_{xy}\right\} \\
&= \frac{1}{n-2}\left\{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2 - \hat{\beta}_1\left(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\right)\right\} \\
&\approx \frac{1}{23-2}\left\{(35626.09) - (23)\left(\frac{576.3}{23}\right)^2 - (0.07876722)^2\left((396603.2) - (23)\left(\frac{6980.65}{23}\right)\left(\frac{576.3}{23}\right)\right)\right\} \\
&\approx 177.3297
\end{aligned}$$

■

1.4 Hypothesis testing on β_0 and β_1

1.4.1 Testing on β_1 - the t-test approach

Suppose we want to test the hypothesis that β_1 equals a given constant c. Then the hypotheses are

$$H_0 : \beta_1 = c, \qquad H_1 : \beta_1 \neq c.$$

A particularly important case is when $c = 0$, which we call **test of the significance of regression**. If we fail to reject H_0, that means there is no linear relationship between x and y.

Let's derive some test statistics. Recall $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, which implies $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Next, since $\hat{\beta}_1$ is a linear combination of y_i, we know $\hat{\beta}_1$ is also normally distributed. We have computed that

$$\mathbb{E}[\hat{\beta}_1] = \beta_1 \quad \text{and} \quad \mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}},$$

thus $\hat{\beta}_1 \sim N\left(\beta_1, \dfrac{\sigma^2}{S_{xx}}\right)$. Thus if the null hypothesis $\beta_1 = c$ is true, then

$$z^* = \frac{\hat{\beta}_1 - c}{\sqrt{\dfrac{\sigma^2}{S_{xx}}}} \sim N(0, 1)$$

given that we know σ^2. As we have said, this is almost impossible. Thus, we would consider using the unbiased estimator MS_{Res} instead. Recall $\dfrac{(n-2)MS_{Res}}{\sigma^2} = \dfrac{SS_{Res}}{\sigma^2} \sim \chi^2(n-2)$ and $z^* \sim N(0,1)$. Also, in a later section we will show that $\dfrac{(n-2)MS_{Res}}{\sigma^2}$ and z^* are independent. Thus the test statistic is

$$t^* = \frac{z^*}{\sqrt{\dfrac{(n-2)MS_{Res}}{\sigma^2} \times \dfrac{1}{n-2}}} = \frac{\hat{\beta}_1 - c}{\sqrt{\dfrac{MS_{Res}}{S_{xx}}}} \sim t(n-2).$$

We can reject H_0 if $|t^*| > t_{\alpha/2}(n-2)$, where $t_{\alpha/2}(n-2)$ is the critical value.

Example 1.5. Let's continue our previous example about the relationship between tweet rate and box office revenue, where we used the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \cdots, 23$$

where y_i are the weekend box office revenues and x_i are the average number of tweets per hour. It is given that

$$\sum_{i=1}^{n} x_i^2 = 4933199 \quad \sum_{i=1}^{n} y_i^2 = 35626.09 \quad \sum_{i=1}^{n} x_i y_i = 396603.2$$
$$\sum_{i=1}^{n} x_i = 6980.65 \quad \sum_{i=1}^{n} y_i = 576.3$$

We have already computed that $\hat{\beta}_0 \approx 1.150158$, $\hat{\beta}_1 \approx 0.07876722$ and $MS_{Res} \approx 177.3297$. Test for the significance of the regression of the model. Use 5% level of significance.

Solution. The hypotheses are

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0.$$

The test statistic is

$$t^* = \frac{\hat{\beta}_1 - 0}{\sqrt{\dfrac{MS_{Res}}{S_{xx}}}} \approx \frac{0.07876722}{\sqrt{\dfrac{177.3297}{(4933199) - (23)\left(\dfrac{6980.65}{23}\right)^2}}} \approx 9.92333$$

Note that the critical value is $t_{0.025}(23-2) = 2.080$. Since $t^* > t_{0.025}(21)$, H_0 is rejected, and we conclude that there is a linear relationship between the revenue and tweet rate. ■

1.4.2 Testing on β_0

Suppose we want to test the hypothesis that β_0 equals a given constant d. Then the hypotheses are

$$H_0 : \beta_0 = d, \qquad H_1 : \beta_0 \neq d.$$

Similar to the argument above, by noting that

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

we have the test statistic

$$t^* = \frac{\hat{\beta}_0 - d}{\sqrt{MS_{Res}\left\{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right\}}} \sim t(n-2).$$

1.4.3 ANOVA

The genius statistician Fisher invented the technique of **analysis of variance** (abbrev. ANOVA), which can be used for testing the significance of regression. The basic idea is to split SS_T as follows:

$$\begin{aligned} SS_T = \sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) + \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \end{aligned}$$

Note that the cross term vanishes by the previous properties of the residual:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \sum_{i=1}^{n} e_i(\hat{y}_i - \overline{y}) = \sum_{i=1}^{n} e_i\hat{y}_i - \overline{y}\sum_{i=1}^{n} e_i = 0.$$

Also note that the first term is simply SS_{Res}. We will customarily call the last term **regression sum of squares**, denote it by

$$SS_{Reg} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$$

because it measures the spread of $\hat{y}$ about the sample mean $\overline{y}$. So now we can write the identity as

$$SS_T = SS_{Res} + SS_{Reg}.$$

Let's analyze the meaning of the identity. SS_T measures the variation of the data about their mean - this is trivial because $SS_T = \sum_{i=1}^{n}(y_i - \overline{y})^2$ just differs from the sample variance by a constant multiple. This variation can be split into two parts:

- $SS_{Reg} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$ which measures the variation of the predicted value $\hat{y}_i$ about the sample mean $\overline{y}$. If the regression model is doing a good job of predicting the observed values, then the predicted values $\hat{y}_i$ should be sufficiently close to the observed values y_i, and hence they will also vary around the sample mean $\overline{y}$. Therefore we say it is 'the variation explained by the model'.
- $SS_{Res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ which measures the variation of the residual, i.e. the gap between the measured y_i and corresponding predicted value $\hat{y}_i$, so we say it is 'the variation that cannot be explained by the model', as oppose to SS_{Reg}.

So now we have the identity $SS_T = SS_{Res} + SS_{Reg}$. But then in the previous subsection we has deduced that

$$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy} = SS_T - \hat{\beta}_1^2 S_{xx}.$$

By comparing the identities, we have

$$SS_{Reg} = \hat{\beta}_1 S_{xy} = \hat{\beta}_1^2 S_{xx}. \tag{7}$$

Note that $\mathbb{E}[SS_{Reg}] = \mathbb{E}[\hat{\beta}_1^2 S_{xx}] = S_{xx}\mathbb{E}[\hat{\beta}_1^2] = S_{xx}\{\mathbb{V}(\hat{\beta}_1) + (\mathbb{E}[\hat{\beta}_1])^2\} = S_{xx}\left\{\frac{\sigma^2}{S_{xx}} + \beta_1^2\right\} = \sigma^2 + \beta_1^2 S_{xx}$.

We can now proceed to the derivation of the F-test. Since we wish to test the significance of the regression, our hypotheses are

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0.$$

Recall $\dfrac{SS_{Res}}{\sigma^2} \sim \chi^2(n-2)$. Also, in a later section we will show that

- If H_0 is true, i.e. $\hat{\beta}_1 = 0$, then $\dfrac{SS_{Reg}}{\sigma^2} \sim \chi^2(1)$, and
- SS_{Res} and SS_{Reg} are independent.

Thus the test statistic

$$F^* = \frac{\dfrac{SS_{Reg}}{1}}{\dfrac{SS_{Res}}{n-2}} = \frac{MS_{Reg}}{MS_{Res}} \sim F(1, n-2)$$

where $MS_{Reg} = \dfrac{SS_{Reg}}{1} = SS_{Reg}$ is the **regression mean squares**. We can reject H_0 if $F^* > F_\alpha(1, n-2)$, where $F_\alpha(1, n-2)$ is the critical value.

In practice, we often summarize the test in an ANOVA table:

Source of variation	Sum of squares	Degree of freedom	Mean square	F^*
Regression	SS_{Reg}	1	MS_{Reg}	$\dfrac{MS_{Reg}}{MS_{Res}}$
Residual	SS_{Res}	$n-2$	MS_{Res}	
Total	SS_T	$n-1$		

Example 1.6. Let's continue our previous example about the relationship between tweet rate and box office revenue, where we used the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \cdots, 23$$

where y_i are the weekend box office revenues and x_i are the average number of tweets per hour. It is given that

$$\sum_{i=1}^{n} x_i^2 = 4933199 \quad \sum_{i=1}^{n} y_i^2 = 35626.09 \quad \sum_{i=1}^{n} x_i y_i = 396603.2$$
$$\sum_{i=1}^{n} x_i = 6980.65 \quad \sum_{i=1}^{n} y_i = 576.3$$

We have already computed that $\hat{\beta}_0 \approx 1.150158$ and $\hat{\beta}_1 \approx 0.07876722$. Test for the significance of the regression of the model by the F-test. Use 5% level of significance.

Solution. The hypotheses are

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0.$$

Next, the sum of squares are

$$SS_T = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 = (35626.09) - (23)\left(\frac{576.3}{23}\right)^2 \approx 21186.01652$$

$$SS_{Reg} = \hat{\beta}_1 S_{xy} \approx (0.07876722)\left(396603.2 - (23)\left(\frac{6980.65}{23}\right)\left(\frac{576.3}{23}\right)\right) = 17462.09338$$

Source of variation	Sum of squares	Degree of freedom	Mean square	F^*
Regression	17462.09338	1	17462.09338	98.47249
Residual	3723.92314	$23-2$	177.32967	
Total	21186.01652	$23-1$		

The test statistic is

$$F^* = \frac{MS_{Reg}}{MS_{Res}} = \frac{17462.09338}{177.32967} = 98.47249$$

Note that the critical value is $F_{0.05}(1, 23-2) \approx F_{0.05}(1, 20) = 4.35$. Since $F^* > F_{0.05}(1, 21)$, H_0 is rejected, and we conclude that there is a linear relationship between the revenue and tweet rate. ■

1.4.4 Cofficient of determination

As we have said, SS_{Reg} measures the variation explained by the regression model. Therefore it makes sense to define the **coefficient of determination**

$$R^2 = \frac{SS_{Reg}}{SS_T}$$

which measures the proportion of variation explained by the regression model.

Example 1.7. Let's continue our previous example about the relationship between tweet rate and box office revenue, where we used the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \cdots, 23$$

where y_i are the weekend box office revenues and x_i are the average number of tweets per hour. It is given that

$$\sum_{i=1}^{n} x_i^2 = 4933199 \quad \sum_{i=1}^{n} y_i^2 = 35626.09 \quad \sum_{i=1}^{n} x_i y_i = 396603.2$$

$$\sum_{i=1}^{n} x_i = 6980.65 \quad \sum_{i=1}^{n} y_i = 576.3$$

We have already obtained the ANOVA table

Source of variation	Sum of squares	Degree of freedom	Mean square	F^*
Regression	17462.09338	1	17462.09338	98.47249
Residual	3723.92314	$23-2$	177.32967	
Total	21186.01652	$23-1$		

Compute the coefficient of determination.

Solution.

$$R^2 = \frac{17462.09338}{21186.01652} \approx 82.4227\%$$

■

1.5 Interval estimations

1.5.1 Confidence interval for β_0 and β_1

Recall that

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \sim t(n-2).$$

Thus we can form the following probabilistic statement:

$$\begin{aligned}
1-\alpha &= \mathbb{P}\left\{ -t_{\alpha/2}(n-2) \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \leq t_{\alpha/2}(n-2) \right\} \\
&= \mathbb{P}\left\{ -t_{\alpha/2}(n-2)\sqrt{\frac{MS_{Res}}{S_{xx}}} \leq \hat{\beta}_1 - \beta_1 \leq t_{\alpha/2}(n-2)\sqrt{\frac{MS_{Res}}{S_{xx}}} \right\} \\
&= \mathbb{P}\left\{ \hat{\beta}_1 - t_{\alpha/2}(n-2)\sqrt{\frac{MS_{Res}}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2}(n-2)\sqrt{\frac{MS_{Res}}{S_{xx}}} \right\}
\end{aligned}$$

Therefore a $100(1-\alpha)\%$ confidence interval for β_1 is given by

$$\hat{\beta}_1 - t_{\alpha/2}(n-2)\sqrt{\frac{MS_{Res}}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2}(n-2)\sqrt{\frac{MS_{Res}}{S_{xx}}}$$

Similarly, since

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{MS_{Res}\left\{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right\}}} \sim t(n-2),$$

a $100(1-\alpha)\%$ confidence interval for β_0 is given by

$$\hat{\beta}_1 - t_{\alpha/2}(n-2)\sqrt{MS_{Res}\left\{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right\}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2}(n-2)\sqrt{MS_{Res}\left\{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right\}}.$$

1.5.2 Confidence interval for σ^2

Recall that $\frac{SS_{Res}}{\sigma^2} \sim \chi^2(n-2)$. Thus we can form the following probabilistic statement:

$$\begin{aligned}
1-\alpha &= \mathbb{P}\left\{ \chi^2_{1-\alpha/2}(n-2) \leq \frac{SS_{Res}}{\sigma^2} \leq \chi^2_{\alpha/2}(n-2) \right\} \\
&= \mathbb{P}\left\{ \frac{SS_{Res}}{\chi^2_{\alpha/2}(n-2)} \leq \sigma^2 \leq \frac{SS_{Res}}{\chi^2_{1-\alpha/2}(n-2)} \right\}
\end{aligned}$$

and so a $100(1-\alpha)\%$ confidence interval for σ^2 is given by

$$\frac{SS_{Res}}{\chi^2_{\alpha/2}(n-2)} \leq \sigma^2 \leq \frac{SS_{Res}}{\chi^2_{1-\alpha/2}(n-2)}$$

1.5.3 Confidence interval for $\mu_{y|x_0}$

An important usage of regression models is to estimate $\mu_{y|x_0} = \mathbb{E}[y|x_0]$, the **mean response** $\mathbb{E}[y]$ for a fixed value x_0 of the regressor x. Recall in the beginning, we have computed that

$$\mu_{y|x_0} = \beta_0 + \beta_1 x_0,$$

so an unbiased estimator of $\mu_{y|x_0}$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 = \hat{y}_0.$$

To obtain a $100(1-\alpha)\%$ confidence interval, we first note that $\hat{y}_0$ is normally distributed, because $\hat{\beta}_i$ are linear combinations of normally distributed y_i, so is $\hat{y}_0$. Now we compute its variance:

$$\begin{aligned}
\mathbb{V}(\hat{y}_0) &= \mathbb{V}(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\
&= \mathbb{V}(\overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x_0) \\
&= \mathbb{V}(\overline{y} + \hat{\beta}_1 (x_0 - \overline{x})) \\
&= \mathbb{V}(\overline{y}) + (x_0 - \overline{x})^2 \mathbb{V}(\hat{\beta}_1) + 2(x_0 - \overline{x}) \operatorname{Cov}(\overline{y}, \hat{\beta}_1) \\
&= \mathbb{V}\left(\frac{1}{n} \sum_{i=1}^{n} y_i \right) + (x_0 - \overline{x})^2 \mathbb{V}(\hat{\beta}_1) && \text{Recall } \operatorname{Cov}(\overline{y}, \hat{\beta}_1) = 0 \\
&= \mathbb{V}\left(\frac{1}{n} \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i + \varepsilon_i) \right) + (x_0 - \overline{x})^2 \mathbb{V}(\hat{\beta}_1) \\
&= \mathbb{V}\left(\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \right) + (x_0 - \overline{x})^2 \mathbb{V}(\hat{\beta}_1) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{V}(\varepsilon_i) + (x_0 - \overline{x})^2 \mathbb{V}(\hat{\beta}_1) && \text{Recall } \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2) \\
&= \frac{1}{n^2} (n\sigma^2) + (x_0 - \overline{x})^2 \left(\frac{\sigma^2}{S_{xx}} \right) \\
&= \sigma^2 \left\{ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}
\end{aligned}$$

Now recall $\dfrac{(n-2)MS_{Res}}{\sigma^2} \sim \chi^2(n-2)$ and note that $\dfrac{\hat{y}_0 - \mu_{y|x_0}}{\sqrt{\mathbb{V}(\hat{y})}} \sim N(0,1)$. We can also show their independence, and therefore we have

$$\frac{\hat{y}_0 - \mu_{y|x_0}}{\sqrt{MS_{Res} \left\{ \dfrac{1}{n} + \dfrac{(x_0 - \overline{x})^2}{S_{xx}} \right\}}} \sim t(n-2).$$

Thus we can form the following probabilistic statement:

$$\begin{aligned}
1 - \alpha &= \mathbb{P}\left\{ -t_{\alpha/2}(n-2) \leq \frac{\hat{y}_0 - \mu_{y|x_0}}{\sqrt{MS_{Res} \left\{ \dfrac{1}{n} + \dfrac{(x_0 - \overline{x})^2}{S_{xx}} \right\}}} \leq t_{\alpha/2}(n-2) \right\} \\
&= \mathbb{P}\left\{ -t_{\alpha/2}(n-2) \sqrt{MS_{Res} \left\{ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}} \leq \hat{y}_0 - \mu_{y|x_0} \leq t_{\alpha/2}(n-2) \sqrt{MS_{Res} \left\{ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}} \right\} \\
&= \mathbb{P}\left\{ \hat{y}_0 - t_{\alpha/2}(n-2) \sqrt{MS_{Res} \left\{ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}} \leq \mu_{y|x_0} \leq \hat{y}_0 + t_{\alpha/2}(n-2) \sqrt{MS_{Res} \left\{ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}} \right\}
\end{aligned}$$

Therefore a $100(1-\alpha)\%$ confidence interval for $\mu_{y|x_0}$ is given by

$$\hat{y}_0 - t_{\alpha/2}(n-2) \sqrt{MS_{Res} \left\{ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}} \leq \mu_{y|x_0} \leq \hat{y}_0 + t_{\alpha/2}(n-2) \sqrt{MS_{Res} \left\{ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}}.$$

Carefully examining the expressions, we will realize that the width of the confidence interval is a function of x_0. The width is minimized when $x_0 = \overline{x}$ and widens as $|x_0 - \overline{x}|$ increases. Hence if x_0 is far away from the original data set, the confidence interval would become wider, so our prediction would have a larger variation.

1.5.4 Prediction interval

When we predict a future response y_0 of a regressor that is outside the original data set, our confidence interval should be wider than the one in the previous one because we also need to take the error of observation into account. Thus we will derive another confidence interval as follows:

Fix the value x_0 of the regressor variable of interest. Then

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

is a point estimate of the response y_0. Note very carefully that here y_0 is a random variable where $y_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$. To obtain a confidence interval for the future observation y_0, or what we customarily call a **prediction interval** for the future observation y_0, we consider the random variable

$$\psi = y_0 - \hat{y}_0.$$

ψ is a sum of two normally distributed random variables, so ψ is also normal. We can compute that

$$\begin{aligned}
\mathbb{E}[\psi] = \mathbb{E}[y_0 - \hat{y}_0] &= \mathbb{E}[y_0] - \mathbb{E}[\hat{y}_0] \\
&= (\beta_0 + \beta_1 x_0) - \mathbb{E}[\hat{\beta}_0 + \hat{\beta}_1 x_0] \\
&= (\beta_0 + \beta_1 x_0) - (\mathbb{E}[\hat{\beta}_0] + \mathbb{E}[\hat{\beta}_1] x_0) \\
&= (\beta_0 + \beta_1 x_0) - (\beta_0 + \beta_1 x_0) \\
&= 0
\end{aligned}$$

Also, logically y_0 and $\hat{y}_0$ are independent because a future observation y_0 should not rely on its estimator. Under this assumption, we can compute that

$$\begin{aligned}
\mathbb{V}(\psi) = \mathbb{V}(y_0 - \hat{y}_0) &= \mathbb{V}(y_0) + \mathbb{V}(\hat{y}_0) - 2\,\mathrm{Cov}(y_0, \hat{y}_0) \\
&= \mathbb{V}(y_0) + \mathbb{V}(\hat{y}_0) \\
&= \sigma^2 + \sigma^2 \left\{ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\} \\
&= \sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}
\end{aligned}$$

Thus it follows that

$$\frac{y_0 - \hat{y}_0}{\sqrt{\sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}}} \sim N(0, 1).$$

Now recall $\dfrac{(n-2)MS_{Res}}{\sigma^2} \sim \chi^2(n-2)$. With their independence, we have

$$\frac{y_0 - \hat{y}_0}{\sqrt{MS_{Res} \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}}} \sim t(n-2)$$

and hence

$$\begin{aligned}
1 - \alpha &= \mathbb{P}\left\{ -t_{\alpha/2}(n-2) \le \frac{y_0 - \hat{y}_0}{\sqrt{MS_{Res} \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}}} \le t_{\alpha/2}(n-2) \right\} \\
&= \mathbb{P}\left\{ -t_{\alpha/2}(n-2) \sqrt{MS_{Res} \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}} \le y_0 - \hat{y}_0 \le t_{\alpha/2}(n-2) \sqrt{MS_{Res} \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}} \right\} \\
&= \mathbb{P}\left\{ \hat{y}_0 - t_{\alpha/2}(n-2) \sqrt{MS_{Res} \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}} \le y_0 \le \hat{y}_0 + t_{\alpha/2}(n-2) \sqrt{MS_{Res} \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}} \right\}
\end{aligned}$$

Therefore a $100(1-\alpha)\%$ confidence interval for y_0 is given by

$$\hat{y}_0 - t_{\alpha/2}(n-2) \sqrt{MS_{Res} \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}} \le y_0 \le \hat{y}_0 + t_{\alpha/2}(n-2) \sqrt{MS_{Res} \left\{ 1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right\}}.$$

Example 1.8. Let's continue our previous example about the relationship between tweet rate and box office revenue, where we used the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \cdots, 23$$

where y_i are the weekend box office revenues and x_i are the average number of tweets per hour. It is given that

$$\sum_{i=1}^{n} x_i^2 = 4933199 \quad \sum_{i=1}^{n} y_i^2 = 35626.09 \quad \sum_{i=1}^{n} x_i y_i = 396603.2$$
$$\sum_{i=1}^{n} x_i = 6980.65 \quad \sum_{i=1}^{n} y_i = 576.3$$

We have already computed that $\hat{\beta}_0 \approx 1.150158$ and $\hat{\beta}_1 \approx 0.07876722$, and have already obtained the ANOVA table

Source of variation	Sum of squares	Degree of freedom	Mean square	F^*
Regression	17462.09338	1	17462.09338	98.47249
Residual	3723.92314	$23-2$	177.32967	
Total	21186.01652	$23-1$		

Construct 95% confidence interval for $\mu_{y|x_0}$ and a 95% prediction interval on a future value of a movie with Tweet rate 100.

Solution.

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \approx (1.150158) + (0.07876722)(100) = 9.02688$$

$$MS_{Res}\left\{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\} \approx (177.32967)\left\{\frac{1}{23} + \frac{\left(100 - \frac{6980.65}{23}\right)^2}{4933199 - (23)\left(\frac{6980.65}{23}\right)^2}\right\} \approx 10.31933828$$

$$MS_{Res}\left\{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\} \approx (177.32967)\left\{1 + \frac{1}{23} + \frac{\left(100 - \frac{6980.65}{23}\right)^2}{4933199 - (23)\left(\frac{6980.65}{23}\right)^2}\right\} \approx 187.6490083$$

$$t_{0.05/2}(23-2) = 2.080$$

- The 95% confidence interval for $\mu_{y|x_0}$ is

$$\left[\hat{y}_0 - t_{\alpha/2}(n-2)\sqrt{MS_{Res}\left\{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\}}, \hat{y}_0 + t_{\alpha/2}(n-2)\sqrt{MS_{Res}\left\{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\}}\right]$$
$$\approx [2.345144811, 15.70861519]$$

- The 95% prediction interval is

$$\left[\hat{y}_0 - t_{\alpha/2}(n-2)\sqrt{MS_{Res}\left\{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\}}, \hat{y}_0 + t_{\alpha/2}(n-2)\sqrt{MS_{Res}\left\{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\}}\right]$$
$$\approx [-19.46600805, 37.51976805]$$

■

1.6 Doing regression analysis in R

We can certainly perform regression analysis in R. Suppose we want to study the relationship between the final exam score and the total score for a course. Suppose the regressor variable is the final exam score, denoted by `Final`, and the response variable is the total score, denoted by `Total`. Suppose we already have a dataset in the file `score.csv`. [1] We first import the dataset into R by

```
> score.data <- read.csv(file.choose())
```

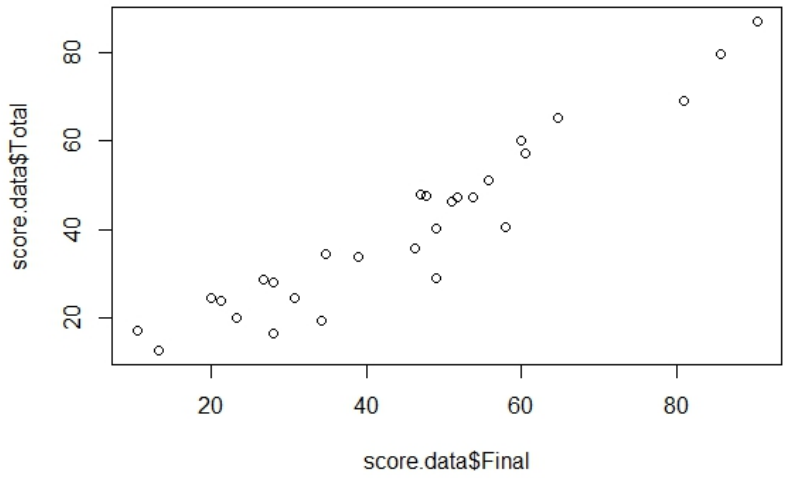
and select the file `score.csv`. Then we would have created a data fram with two columns: `Final` and `Total`. When we type

```
> score.data
```

we can see the data frame printed out in a nice column format. Then we can type

```
> plot(score.data$Final, score.data$Total)
```

to create a plot of `Total` against `Final`:



At first glance we can already predict that a linear relationship exists, but we need to confirm this relationship as we have done before. We set up the linear regression by typing

```
> score.regression <- lm(Total ~ Final, data=score.data)
```

Here the function `lm()` stands for linear model, and the argument `Total ~ Final` specifies the regressor and response by the format

`response ~ regressor`,

while the argument `data=score.data` specifies the data frame.

[1] The dataset is available at `https://github.com/half-a-minute/prob_n_stat_notes_proj/blob/main/reg/score.csv`.

The real meat of doing regression analysis in R is in the summary function:

```
> summary(score.regression)

Call:
lm(formula = Total ~ Final, data = score.data)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9795  -3.4401   0.8372   4.5110   7.5003

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.34884    2.84851   0.122    0.903
Final        0.89382    0.05766  15.503 1.19e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.188 on 26 degrees of freedom
Multiple R-squared:  0.9024,  Adjusted R-squared:  0.8986
F-statistic: 240.3 on 1 and 26 DF,  p-value: 1.189e-14
```

It is important to understand how to interpret this large chunk of information. Let's study them one by one.

The first chunk

```
Call:
lm(formula = Total ~ Final, data = score.data)
```

just prints out the original call to the `lm()` function, while the second chunk

```
Residuals:
     Min       1Q   Median       3Q      Max
-14.9795  -3.4401   0.8372   4.5110   7.5003
```

gives the five-number summary of residuals $e_i = y_i - \hat{y}_i$.

The third chunk is very important as it contains many useful information:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.34884    2.84851   0.122    0.903
Final        0.89382    0.05766  15.503 1.19e-14 ***
```

In the column `Estimate`, we can find the least-squares estimates of β_i. In this example,

$$\hat{\beta}_0 = 0.34884, \qquad \hat{\beta}_1 = 0.89382$$

and so the fitted regression line is

$$\hat{y} = 0.34884 + 0.89382x.$$

The column `Std. Error` gives the standard error in their t-tests, i.e. the denominators of the test statistics. The column `t value` gives the test statistics, while `Pr(>|t|)` gives the p-value of the tests. These values are calculated to test the hypotheses

$$H_0 : \beta_i = 0, \qquad H_1 : \beta_i \neq 0.$$

In general, our tests are carried at a 5% level of significance. So when the p-value is < 0.05 indicating statistical significance, there will be some asterisks `*` next to the p-value, where more asterisks indicate a lower p-value as explained by the significance codes

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this example, the p-value is $1.19 \times 10^{-14} \ll 0.05$, so there is a linear relationship between the total score and the final score.

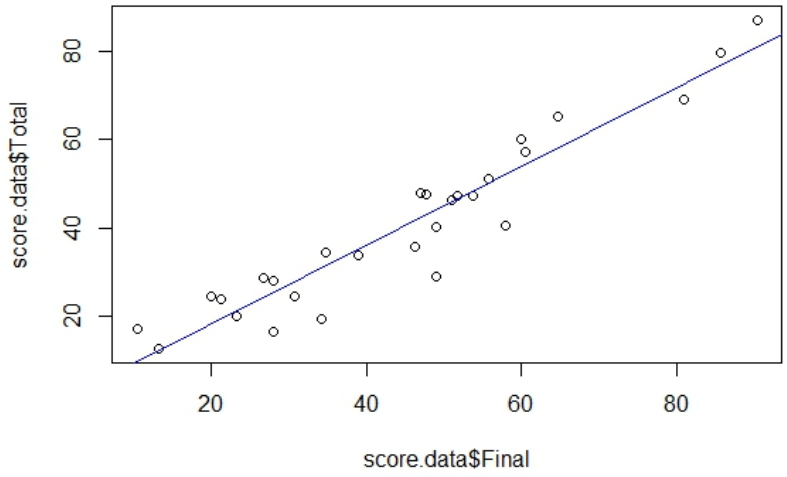
The last chunk is rather intuitive, yet it still contains many useful information:

```
Residual standard error: 6.188 on 26 degrees of freedom
Multiple R-squared:  0.9024,  Adjusted R-squared:  0.8986
F-statistic: 240.3 on 1 and 26 DF,  p-value: 1.189e-14
```

`Residual standard error` is $\sqrt{MS_{Res}}$, while `Multiple R-squared` is R^2, the coefficient of determination. `F-statistic` is the test statistic when we use the ANOVA approach.

Finally we can add the fitted regression line into the plot:

```
> abline(score.regression, col="blue")
```



Example 1.9. An ad-testing company, Video Board Tests, Inc., interviewed 4,000 adults to survey television advertisements. The people interviewed were regular product users asked to name a commercial they had seen for a product category in the past week. The number of retained impressions per week (in millions) is the response variable y_i, and the amount of money spent on advertising (in millions) is the regressor x_i. A random sample of 20 firms provided the data for this study. Below are some summary statistics of the data:

$$\sum_{i=1}^{n} x_i^2 = 111175.8 \quad \sum_{i=1}^{n} y_i^2 = 51176.09 \quad \sum_{i=1}^{n} x_i y_i = 63076.53$$

$$\sum_{i=1}^{n} x_i = 1031.5 \quad \sum_{i=1}^{n} y_i = 812.1$$

Below is the output given by R (note: some original outputs are removed and replaced by `XXX`):

```
Call:
lm(formula = y ~ x, data = q1)

Residuals:
    Min      1Q  Median      3Q     Max
-42.405 -12.632  -7.888   9.515  50.761

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)      XXX     7.4623     XXX      XXX
x                XXX     0.1001     XXX      XXX
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.1 on XXX degrees of freedom
Multiple R-squared:  XXX, Adjusted R-squared:  XXX
F-statistic: XXX on XXX and XXX DF,  p-value: XXX
```

(a) Find the fitted simple linear regression model.

(b) Using a t-test with a 5% level of significance, is there a significant relationship between the amount a company spends on advertising and retained impressions?

(c) Using a F-test with a 5% level of significance, is there a significant relationship between the amount a company spends on advertising and retained impressions?

(d) Calculate the proportion of variation in the retained impressions that is explained by the amount a company spends on advertising.

(e) What are the 95% confidence and prediction intervals for the number of retained impressions for MCI, which spent \$26.9 million on advertising?

Solution.

(a) Compute that

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n x_i y_i - n\overline{x}\,\overline{y}}{\sum_{i=1}^n x_i^2 - n\overline{x}^2} = \frac{63076.53 - (20)\left(\frac{1031.5}{20}\right)\left(\frac{812.1}{20}\right)}{111175.8 - (20)\left(\frac{1031.5}{20}\right)^2} \approx 0.365537532$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = \left(\frac{812.1}{20}\right) - (0.365537532)\left(\frac{1031.5}{20}\right) \approx 21.75240179$$

and so the fitted linear regression model is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 0.365537532x + 21.75240179$$

(b) The hypotheses are

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0.$$

The test statistic is

$$t^* = \frac{0.365537532 - 0}{0.1001} \approx 3.651723596$$

The critical value is $t_{0.05/2}(20-2) = 2.101$. Since $t^* > t_{0.025}(18)$, H_0 is rejected, and we can conclude that there is a linear relationship.

(c) The hypotheses are still

$$H_0 : \beta_1 = 0, \qquad H_1 : \beta_1 \neq 0.$$

The sums of squares are

$$SS_T = \sum_{i=1}^n (y_i - \overline{y})^2 = \sum_{i=1}^n y_i^2 - n\overline{y}^2 = (51176.09) - (20)\left(\frac{812.1}{20}\right)^2 \approx 18200.7695$$

$$SS_{Reg} = \hat{\beta}_1 S_{xy} = (0.365537532)\left(63076.53 - (20)\left(\frac{1031.5}{20}\right)\left(\frac{812.1}{20}\right)\right) \approx 7746.644095$$

Source of variation	Sum of squares	Degree of freedom	Mean square	F^*
Regression	7746.644095	1	7746.644095	13.33823618
Residual	10454.12541	$20-2$	580.7847447	
Total	18200.7695	$20-1$		

The test statistic is

$$F^* = \frac{MS_{Reg}}{MS_{Res}} = \frac{7746.644095}{580.7847447} = 13.33823618$$

Note that the critical value is $F_{0.05}(1, 20-2) \approx F_{0.05}(1, 20) = 4.35$. Since $F^* > F_{0.05}(1, 21)$, H_0 is rejected, and we conclude that there is a linear relationship.

Remark. A quick check:
$\sqrt{MS_{Res}} \approx \sqrt{580.7847447} \approx 24.09947603$, which agrees with the given `residual standard error`.

(d)

$$R^2 = \frac{SS_{Reg}}{SS_T} = \frac{7746.644095}{18200.7695} = 0.42562179$$

(e) Compute that

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \approx 21.75240179 + 0.365537532 \times 26.9 \approx 31.5853614$$

$$MS_{Res}\left\{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\} \approx (580.7847447)\left\{\frac{1}{20} + \frac{\left(26.9 - \frac{1031.5}{20}\right)^2}{111175.8 - (20)\left(\frac{1031.5}{20}\right)^2}\right\} \approx 35.13853548$$

$$MS_{Res}\left\{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\} \approx (580.7847447)\left\{1 + \frac{1}{20} + \frac{\left(26.9 - \frac{1031.5}{20}\right)^2}{111175.8 - (20)\left(\frac{1031.5}{20}\right)^2}\right\} \approx 615.9232802$$

$$t_{0.05/2}(20 - 2) = 2.101$$

- The confidence interval is

$$\left[\hat{y}_0 - t_{\alpha/2}(n-2)\sqrt{MS_{Res}\left\{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\}}, \hat{y}_0 + t_{\alpha/2}(n-2)\sqrt{MS_{Res}\left\{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\}}\right]$$

$$\approx [19.13110275, 44.03962005]$$

- The prediction interval is

$$\left[\hat{y}_0 - t_{\alpha/2}(n-2)\sqrt{MS_{Res}\left\{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\}}, \hat{y}_0 + t_{\alpha/2}(n-2)\sqrt{MS_{Res}\left\{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right\}}\right]$$

$$\approx [-20.55683993, 83.72756273]$$

■

2 Multiple linear regression

2.1 Multiple linear regression model

In practice, there are multiple regressors that contribute to the response. For example, the yield of ammonia under the Haber process can be affected by gas pressure and the surface area of the catalyst. A multiple regression model that might describe this relationship is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon,$$

where y is the yield, x_1 is the gas pressure and x_2 is the surface area of the catalyst. This is what we call a **multiple linear regression model** with two regressors.

In general, the response y may be related to k regressors. In this case, the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

is a **multiple linear regression model** with k regressors. The parameters β_i are called **regression coefficients**. The interpretations of β_i are still the same:

- β_0 is the expected value of the response y when $x_1 = x_2 = \cdots = x_n = 0$.
- β_i $(i \neq 0)$ is the expected change in the response y per unit change in x_i when all of the other regressors x_j $(j \neq i)$ are held constants. Thus the parameters β_i $(i \neq 0)$ are called **partial regression coefficients**.

Moreover, the term **linear** is used because the model is a linear function of the regression coefficients.

Example 2.1.

Linear models	• $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$ • $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$ • $\log(y) = \beta_0 + \beta_1 \sin(x) + \varepsilon$
Nonlinear models	• $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_1 \beta_2 x_1 x_2 + \varepsilon$ • $\log(y) = \sin(\beta_0 + \beta_1 x) + \varepsilon$