

7 Lecture 7: Feb 9

Last time

- SLR in Matrix Form

Today

- Simple correlation
- The statistical model of the SLR (JF chapter 6)

Some properties:

- (a) $\sum x_i \hat{\epsilon}_i = 0$.
- (b) $\sum \hat{y}_i \hat{\epsilon}_i = 0$ (HW1).

Proof:

Other quantities in Matrix Form

Fitted values

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 + \hat{\beta}_1 x_1 \\ \hat{\beta}_0 + \hat{\beta}_1 x_2 \\ \vdots \\ \hat{\beta}_0 + \hat{\beta}_1 x_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \mathbf{X} \hat{\boldsymbol{\beta}}$$

Hat matrix

$$\hat{\mathbf{Y}} = \mathbf{X} \hat{\boldsymbol{\beta}}$$

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

$$\hat{\mathbf{Y}} = \mathbf{H} \mathbf{Y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ is called “hat matrix” because it turns \mathbf{Y} into $\hat{\mathbf{Y}}$.

Davis’s data example

For Davis’s data, we have

$$n = 101$$

$$\bar{y} = \frac{5780}{101} = 57.228$$

$$\bar{x} = \frac{5731}{101} = 56.743$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 4435.9$$

$$\sum (x_i - \bar{x})^2 = 4539.3,$$

so that

$$\begin{aligned}\hat{\beta}_1 &= \frac{4435.9}{4539.3} = 0.97722 \\ \hat{\beta}_0 &= 57.228 - 0.97722 \times 56.743 = 1.7776\end{aligned}$$

Figure 7.1 shows Davis's data on the measured and reported weight in kilograms of 101 women who were engaged in regular exercise.

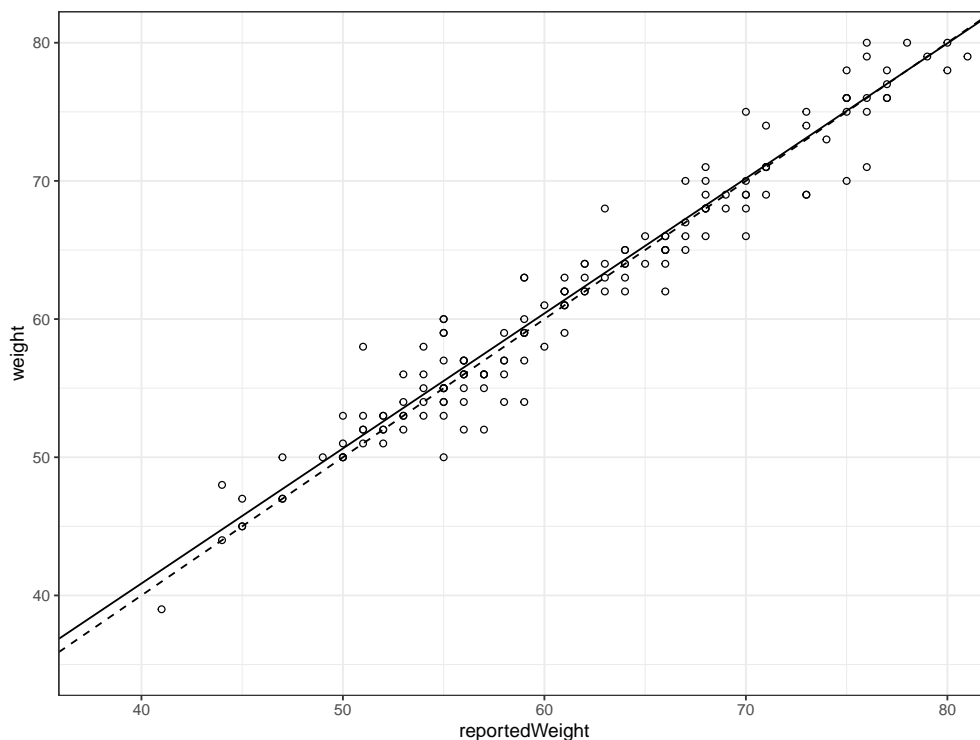


Figure 7.1: Scatterplot of Davis's data on the measured and reported weight of 101 women. The dashed line gives $y = x$. The solid line gives the least squares line $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

Simple correlation

Having calculated the least squares line, it is of interest to determine how closely the line fits the scatter of points. There are many ways of answering it. The standard deviation of the residuals, S_E , often called the *standard error of the regression* or the *residue standard error*, provides one sort of answer. Because of estimation considerations, the variance of the residuals is defined using *degrees of freedom* $n - 2$:

$$S_\epsilon^2 = \frac{\sum \hat{\epsilon}_i^2}{n - 2}.$$

The residual standard error is,

$$S_\epsilon = \sqrt{\frac{\sum \hat{\epsilon}_i^2}{n - 2}}$$

For the Davis's data, the sum of squared residuals is $\sum \hat{\epsilon}_i^2 = 418.87$, and thus the standard error of the regression is

$$S_\epsilon = \sqrt{\frac{418.87}{101 - 2}} = 2.0569\text{kg}.$$

On average, using the least-squares regression line to predict measured weight from reported weight results in an error of about 2 kg.

Sum of squares:

- Total sum of squares (TSS) for Y: $\text{TSS} = \sum (y_i - \bar{y})^2$
- Residual sum of squares (RSS): $\text{RSS} = \sum (y_i - \hat{y}_i)^2$
- regression sum of squares (RegSS): $\text{RegSS} = \text{TSS} - \text{RSS} = \sum (\hat{y}_i - \bar{y})^2$
- $\text{RegSS} + \text{RSS} = \text{TSS}$

Sample correlation coefficient

Definition: The sample correlation coefficient r_{xy} of the paired data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ is defined by

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}) / (n - 1)}{\sqrt{\sum (x_i - \bar{x})^2 / (n - 1) \times \sum (y_i - \bar{y})^2 / (n - 1)}} = \frac{s_{xy}}{s_x s_y}$$

s_{xy} is called the sample covariance of x and y :

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$s_x = \sqrt{\sum (x_i - \bar{x})^2 / (n - 1)}$ and $s_y = \sqrt{\sum (y_i - \bar{y})^2 / (n - 1)}$ are, respectively, the sample standard deviations of X and Y .

Some properties of r_{xy} :

- r_{xy} is a measure of the linear association between x and y in a dataset.
- correlation coefficients are always between -1 and 1 :

$$-1 \leq r_{xy} \leq 1$$

- The closer r_{xy} is to 1 , the stronger the positive linear association between x and y
- The closer r_{xy} is to -1 , the stronger the negative linear association between x and y
- The bigger $|r_{xy}|$, the stronger the linear association
- If $|r_{xy}| = 1$, then x and y are said to be perfectly correlated.
- $\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = r_{xy} \frac{s_y}{s_x}$

R-square

The ratio of RegSS to TSS is called the *coefficient of determination*, or sometimes, simply “r-square”. It represents the proportion of variation observed in the response variable y which can be “explained” by its linear association with x .

- In simple linear regression, “r-square” is in fact equal to r_{xy}^2 . (But this isn’t the case in multiple regression.)
- It is also equal to the squared correlation between y_i and \hat{y}_i . (This is the case in multiple regression.)

For Davis’s regression of measured on reported weight:

$$\text{TSS} = 4753.8$$

$$\text{RSS} = 418.87$$

$$\text{RegSS} = 4334.9$$

Thus,

$$r^2 = \frac{4334.9}{4753.8} = 1 - \frac{418.87}{4753.8} = 0.9119$$

The statistical model of Simple Linear Regression

Standard statistical inference in simple regression is based on a *statistical model* that describes the population or process that is sampled:

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

where the coefficients β_0 and β_1 are the *population regression parameters*. The data are randomly sampled from some population of interest.

- y_i is the value of the response variable
- x_i is the explanatory variable
- ϵ_i represents the aggregated omitted causes of y (i.e., the causes of y beyond the explanatory variable), other explanatory variables that could have been included in the regression model, measurement error in y , and whatever component of y is inherently random.

Key assumptions of SLR

The key assumptions of the SLR model concern the behavior of the errors, equivalently, the distribution of y conditional on x :

- *Linearity*. The expectation of the error given the value of x is 0: $\mathbf{E}(\epsilon) \equiv \mathbf{E}(\epsilon|x_i) = 0$. And equivalently, the expected value of the response variable is a linear function of the explanatory variable: $\mu_i \equiv \mathbf{E}(y_i) \equiv \mathbf{E}(y_i|x_i) = \mathbf{E}(\beta_0 + \beta_1 x_i + \epsilon_i|x_i) = \beta_0 + \beta_1 x_i$.

- *Constant variance.* The variance of the errors is the same regardless of the value of x : $\mathbf{Var}(\epsilon|x_i) = \sigma_\epsilon^2$. The constant error variance implies constant conditional variance of y on given x : $\mathbf{Var}(y|x_i) = \mathbf{E}((y_i - \mu_i)^2) = \mathbf{E}((y_i - \beta_0 - \beta_1 x_i)^2) = \mathbf{E}(\epsilon_i^2) = \sigma_\epsilon^2$. (Question: why the last equal sign?)
- *Normality.* The errors are independent identically distributed with Normal distribution with mean 0 and variance σ_ϵ^2 . Write as $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma_\epsilon^2)$. Equivalently, the conditional distribution of the response variable is normal: $y_i \stackrel{iid}{\sim} N(\beta_0 + \beta_1 x_i, \sigma_\epsilon^2)$.
- *Independence.* The observations are sampled independently.
- *Fixed X , or X measured without error and independent of the error.*
 - For experimental research where X values are under direct control of the researcher (i.e. X 's are fixed). If the experiment were replicated, then the values of X would remain the same.
 - For research where X values are sampled, we assume the explanatory variable is measured without error and the explanatory variable and the error are independent in the population from which the sample is drawn.
- *X is not invariant.* X 's can not be all the same.

Figure 7.2 shows the assumptions of linearity, constant variance, and normality in SLR model.

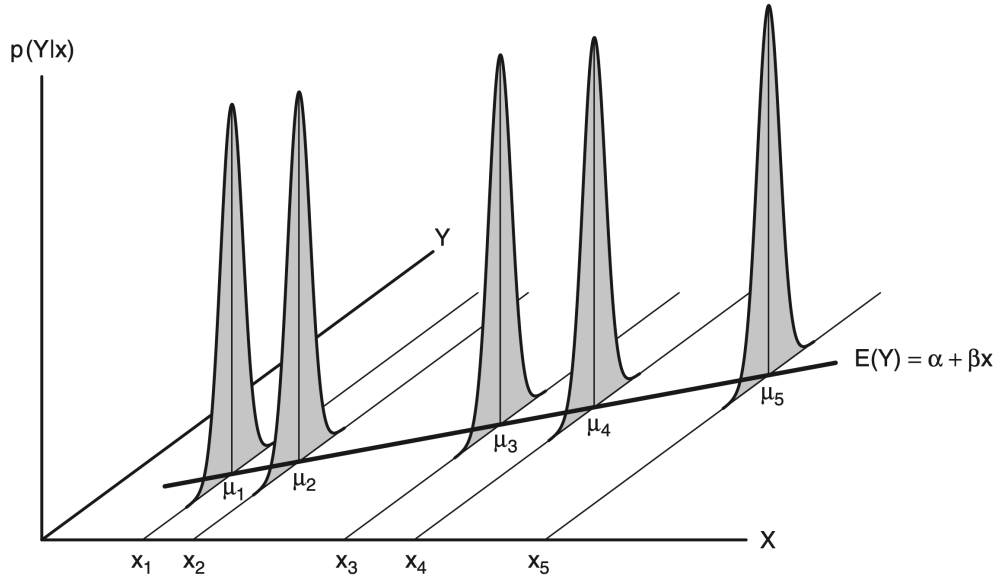


Figure 7.2: The assumptions of linearity, constant variance, and normality in simple regression. The graph shows the conditional population distributions $\Pr(Y|x)$ of Y for several values of the explanatory variable X , labeled as x_1, x_2, \dots, x_5 . The conditional means of Y given x are denoted μ_1, \dots, μ_5 .