9 Lecture 10: Feb 5

Last time

- Sum of squares

Today

- R-square
- Statistical model of SLR

Sample correlation coefficient

Definition: The sample correlation coefficient r_{xy} of the paired data (x_1, y_1), (x_2, y_2), ..., (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:

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

- 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:

$$\begin{aligned} \text{TSS} &= 4753.8 \\ \text{RSS} &= 418.87 \\ \text{RegSS} &= 4334.9 \end{aligned}$$

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 $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$. Equivalently, the conditional distribution of the response variable is normal: $y_i \overset{iid}{\sim} N(\beta_0 + \beta_1 x_i, \sigma_\epsilon^2)$.
- *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 9.1 shows the assumptions of linearity, constant variance, and normality in SLR model.

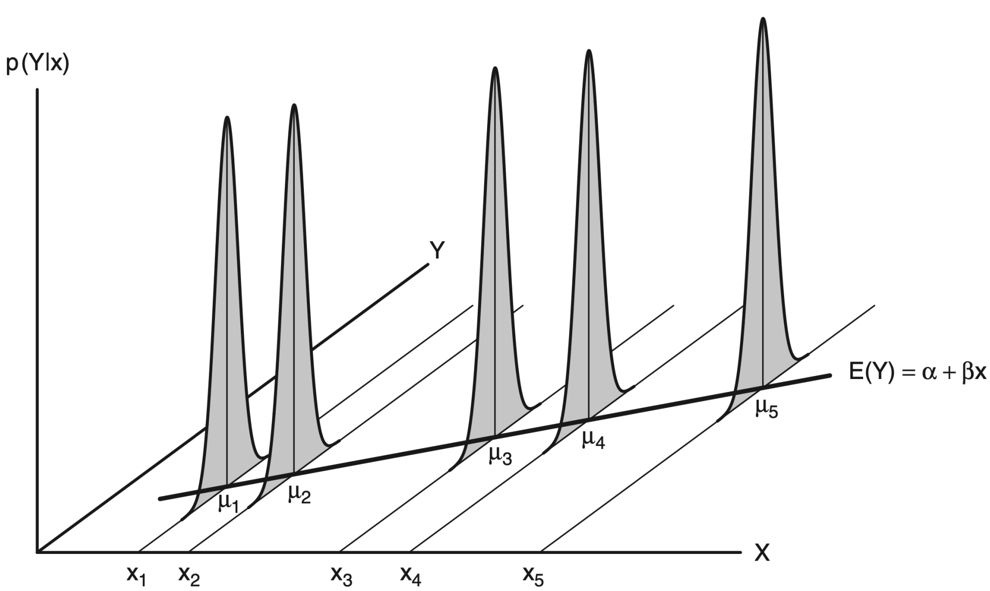

Figure 9.1: The assumptions of linearity, constant variance, and normality in simple regression. The graph shows the conditional population distributions $\Pr(Y|x)$ of Y for several values of the explanatory variable X, labeled as $x_1, x_2, \ldots, x_5$. The conditional means of Y given x are denoted $\mu_1, \ldots, \mu_5$.

Properties of the Least-Squares estimator

Under the strong assumptions of the simple regression model, the sample least squares coefficients $\hat{\beta}_{ls}$ have several desirable properties as estimators of the population regression coefficients β_0 and β_1:

- The least-squares intercept and slope are *linear estimators*, in the sense that they are linear functions of the observations y_i.
  Proof:

- The sample least-squares coefficients are *unbiased estimators* of the population regression coefficients:

  $$\mathbf{E}\left(\hat{\beta}_0\right) = \beta_0$$
  $$\mathbf{E}\left(\hat{\beta}_1\right) = \beta_1$$

  Proof:

- Both $\hat{\beta}_0$ and $\hat{\beta}_1$ have simple sampling variances:

  $$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}$$
  $$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2}$$

  Proof:

- Rewrite the formula for $\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{(\mathrm{n}-1)\mathrm{S}_\mathrm{X}^2}$, we see that the sampling variance of the slope estimate will be small when
  - The error variance σ_ϵ^2 is small
  - The sample size n is large
  - The explanatory-variable values are spread out (i.e. have a large variance, S_X^2)
- (Gauss-Markov theorem) Under the assumptions of linearity, constant variance, and independence, the least-squares estimators are BLUE (Best Linear Unbiased Estimator), that is they have the smallest sampling variance and are unbiased. (show this)
  Proof:

- Under the full suite of assumptions, the least-squares coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are the maximum-likelihood estimators of β_0 and β_1. (show this)
  Proof:

- Under the assumption of normality, the least-squares coefficients are themselves normally distributed. Summing up,

$$\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma_\epsilon^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2})$$
$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma_\epsilon^2}{\sum (x_i - \bar{x})^2})$$