Simple Linear Regression

- The response variable Y is linearly related to one explanatory variable X. That is,

$$Y_i = (a + bX_i) + \epsilon_i. \qquad i = 1, 2, \ldots, n.$$

Assumptions:

1. The mean of ϵ_i is 0 and the variance of ϵ_i is σ^2.

2. The random errors ϵ_i are uncorrelated.

3. a and b are parameters.

4. X_i is a known constant.

- **Equation of the Least-Squares Regression Line .** Suppose we have data on an explanatory variable x and a response variable y for n individuals. The means and standard deviations of the sample data are $\bar{x}$ and s_x for x and $\bar{y}$ and s_y for y, and the correlation between x and y is r. The equation of the least-squares regression line of y on x is

$$\hat{y} = \hat{a} + \hat{b}x$$

with *slope*

$$\hat{b} = \frac{SS_{xy}}{SS_{xx}} = \frac{(\Sigma xy) - \frac{1}{n}(\Sigma x)(\Sigma y)}{(\Sigma x^2) - \frac{1}{n}(\Sigma x)^2} = r\frac{s_y}{s_x} \tag{1}$$

and *intercept*

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \tag{2}$$

- The **fitted** (or **predicted**) **values** $\hat{y}_i$'s are obtained by successively substituting the x_i's into the estimated regression line: $\hat{y} = \hat{a} + \hat{b}x_i$. The **residuals** are the vertical deviations, $e_i = y_i - \hat{y}_i$, from the estimated line.

- The **error sum of squares**, (equivalently, **residual sum of squares**) denoted by SSE, is

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{a} + \hat{b}x_i)]^2 \tag{3}$$

$$= \sum y_i^2 - \hat{a}\sum y_i - \hat{b}\sum x_i y_i \tag{4}$$

and the estimate of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \frac{(n-1)s_y^2(1-r^2)}{n-2}. \tag{5}$$

- The **coefficient of determination**, denoted by r^2, is the amount of the variation in y that is explained by the regression line.

$$r^2 = 1 - \frac{SSE}{SST}, \qquad \text{where, } SST = SS_{yy} = \sum (y_i - \bar{y}) \tag{6}$$

$$= \frac{SST - SSE}{SST} = \frac{\text{explained variation}}{\text{total variation}} \tag{7}$$

- **Inference for b.**

1. Test statistic:

$$\frac{\hat{b} - b}{SE_{\hat{b}}} \sim t(n-2) \qquad SE_{\hat{b}} = \frac{s}{s_x\sqrt{n-1}} = \frac{\hat{b}\sqrt{1-r^2}}{r\sqrt{n-2}}$$

2. Confidence Interval: $\hat{b} \pm t_{\alpha/2} SE_{\hat{b}}$

- **Mean Response of Y at a specified value x^*, ($\mu_{Y|x^*}$).**

  1. **Point Estimate.** For a specific value $x*$, the estimate of the **mean** value of Y is given by

  $$\hat{\mu}_{Y|x^*} = \hat{a} + \hat{b}x^*$$

  2. **Confidence Interval.** For a specific value $x*$, the $(1-\alpha)100\%$ confidence interval for $\mu_{Y|x^*}$ is given by

  $$\hat{\mu}_{Y|x^*} \pm t_{\alpha/2;(n-2)} SE_{\hat{\mu}}$$

  where, $SE_{\hat{\mu}} = s\sqrt{\dfrac{1}{n} + \dfrac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$, and $s = s_y\sqrt{\dfrac{(n-1)(1-r^2)}{n-2}}$

- **Prediction of Y at a specified value x^*.**

  1. **Point Estimate.** For a specific value $x*$, the predicted value of Y is given by

  $$\hat{y} = \hat{a} + \hat{b}x^*$$

  2. **Prediction Interval.** For a specific value $x*$, the $(1-\alpha)100\%$ prediction interval is given by

  $$\hat{y} \pm t_{\alpha/2;(n-2)} SE_{\hat{y}}$$

  where, $SE_{\hat{y}} = s\sqrt{1 + \dfrac{1}{n} + \dfrac{(x^* - \bar{x})^2}{(n-1)s_x^2}}$, and $s = s_y\sqrt{\dfrac{(n-1)(1-r^2)}{n-2}}$

- **Blood Pressure Measurements.** To see if there is a linear relationship between the Systolic and Diastolic blood pressure of a person, the following measurements from 14 randomly selected individuals were recorded.

Person	1	2	3	4	5	6	7
Systolic (x)	138	130	135	140	120	125	120
Diastolic (y)	82	91	100	100	80	90	80
Person	8	9	10	11	12	13	14
Systolic (x)	130	130	144	143	140	130	150
Diastolic (y)	80	80	98	105	85	70	100

  1. Determine the correlation coefficient r.

  2. What can you say about the linear relationship of x and y? Is it a strong linear relationship.

  3. Determine the coefficient of determination r^2. Explain the meaning of this quantity in this context.

  4. Determine the regression line.

5. Test $H_0 : b = 0$ vs. $H_1 : b \neq 0$

6. Construct a 95% confidence interval for b.

7. Find an estimate to the expected diastolic blood pressure (μ_y) for people with a systolic reading of 122.

8. Construct a 95% confidence interval for μ_y for people with a systolic reading of 122.

9. Find the best predicted diastolic blood pressure for a person with a systolic reading of 122.

10. Construct a 95% confidence interval for $\hat{y}$ for a person with a systolic reading of 122.

- **Homework Problem :** Consider the following data of 10 production runs of a certain manufacturing company.

Production run	1	2	3	4	5	6	7	8	9	10
Lot size (x)	30	20	60	80	40	50	60	30	70	60
Man-Hours (y)	73	50	128	170	87	108	135	69	148	132

1. Determine the correlation coefficient r.

2. What can you say about the linear relationship of x and y? Is it a strong linear relationship.

3. Determine the coefficient of determination r^2. Explain the meaning of this quantity in this context.

4. Determine the regression line.

5. Using $\alpha = 0.01$, test $H_0 : b = 0$ vs. $H_1 : b \neq 0$

6. Construct and interpret a 99% confidence interval for b.

7. Find an estimate for the **mean** number of man-hours ($\hat{\mu}_{Y|x^*}$) required to produce a lot size 100.

8. Construct a 95% confidence interval for the **mean** number of man-hours ($\mu_{Y|x^*}$) required to produce a lot size 100.

9. Predict the number of man-hours ($\hat{y}$) required to produce a lot size 100.

10. Construct a 95% prediction interval for the number of man-hours ($\hat{y}$) required to produce a lot size 100.

- **Homework problems:**
  Section 11.3/11.4: (pp. 615-616) # 33, 34.
  Section 11.5: (pp. 621-624) # 47, 48, 49.
  Section 11.8: (pp. 641-642) # 86, 87.