# Introduction to Statistics - Homework #11

- Use Appendix C: Distribution tables (e.g. Z-table) from page 408 to 417 in the textbook if necessary.

# Exercise 8.33* - Husbands and wives, Part II

Data on **12 married couples** in Britain consist of the husband's height (`husband`, in cm) and the wife's height (`wife`, in cm). The wife's height is treated as the response variable, $y_i$, and the husband's height as the predictor variable, $x_i$. We consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2).$$

Answer the following questions based on the following R output.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.5200    16.8278   4.785 0.000740 ***
husband       0.4805     0.0941   5.109 0.000458 ***
---
Signif. codes: 0  '***'  0.001  '**'  0.01  '*'  0.05  '.'  0.1  ' '  1
Residual standard error: 2.324 on 10 degrees of freedom
Multiple R-squared: 0.7230, Adjusted R-squared: 0.6953
F-statistic: 26.10 on 1 and 10 DF, p-value: 0.0004583
```

(1) Fill in the values shown in the table below.

| $\hat{\beta}_0$ | $\hat{\beta}_1$ | **Estimated Regression Equation** | $MSE = \hat{\sigma}^2$ |
|---|---|---|---|
| | | | |

Solution: $\hat{\beta}_0 = 80.5200$ and $\hat{\beta}_1 = 0.4805$, and the fitted regression equation is $\hat{y} = 80.5200 + 0.4805x$. The residual standard error is $2.324$, so the mean squared error is $MSE = (2.324)^2 = 5.4010$.

(2) A married couple is selected in which the husband is $x_i = 185$ cm tall, and the wife's actual height is $y_i = 173$ cm. Use the model to predict the mean wife's height $\hat{y}_i$ at $x_i = 185$ and compute the residual $e_i$.

Solution: $\hat{y}_i = 80.5200 + 0.4805(185) = 169.4125 \text{ cm}$

$$e_i = y_i - \hat{y}_i = 173 - 169.4125 = 3.5875 \text{ cm}$$

(3) Using the R output from part (1), test whether the husband's height (`husband`, $x_i$) is a statistically significant predictor variable for the wife's height (`wife`, $y_i$) at the significance level $\alpha = 0.01$. Answer the following.

(a) State the null and alternative hypotheses.

Solution:

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

(b) Report the p-value, and complete the hypothesis test. State the conclusion in the context of the data.

Solution: From the R output, $t = 5.109, \quad p\text{-value} = 0.000458.$

Since $0.000458 < \alpha = 0.01$, we reject the null hypothesis.

**Conclusion:** At the 1% significance level, the husband's height is a statistically significant predictor variable for the wife's height.

# Exercise 8.23 - The Coast Starlight, Part II

The Coast Starlight Amtrak train runs from Seattle to Los Angeles. Using a linear model, we try to predict the amount of time it takes to travel from one stop to another ($y$, in minutes) using the distance between each stop ( $x$, in miles). The mean travel time from one stop to the next on the Coast Starlight is $\bar{y} = 129$ minutes, with a standard deviation of $s_y = 113$ minutes. The mean distance traveled from one stop to the next is $\bar{x} = 108$ miles with a standard deviation of $s_x = 99$ miles. The correlation between travel time and distance is $r = 0.636$.

(1) Compute $\hat{\beta}_0$ and $\hat{\beta}_1$. Write the regression equation using $\hat{\beta}_0$ and $\hat{\beta}_1$.

Solution: To compute the regression coefficients:

$$\hat{\beta}_1 = r \cdot \frac{s_y}{s_x} = 0.636 \cdot \frac{113}{99} \approx 0.7259$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 129 - 0.7259 \cdot 108 = 50.6028$$

Therefore, the regression equation is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 50.6028 + 0.7259x$$

(2) Calculate the coefficient of determination $R^2$. Interpret $R^2$ in the context of the application.

Solution:

$$R^2 = (0.636)^2 = 0.4045$$

About 40.45% of the variability in travel time is explained by the linear relationship with distance.

(3) The distance between Santa Barbara and Los Angeles is $x_i = 103$ miles. Use the model to estimate the average time $\hat{y}_i$ it takes to travel between these two cities.

Solution:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 50.6028 + 0.7259 \cdot 103 = 125.3705 \text{ minutes}$$

(4) It actually takes the Coast Starlight about $y_i = 168$ minutes to travel from Santa Barbara to Los Angeles. Calculate the residual $e_i$.

Solution:

$$e_i = y_i - \hat{y}_i = 168 - 125.3705 = 42.6295 \text{ minutes}$$

The actual travel time is 42.6295 minutes longer than the predicted travel time.

(5) Suppose Amtrak is considering adding a stop to the Coast Starlight $x = 500$ miles away from Los Angeles. Would it be appropriate to use this linear model to predict the travel time from Los Angeles to this point?

Solution:

**No.** A distance of $x = 500$ miles is far outside the range of the data used to fit the model. Using the model in this case would be an **extrapolation**.