Solution to HW 2

Guanhua FANG

October 4, 2017

2.1

a. Confidence interval of slope excludes zero, it shows there is a positive linear relationship between Y and X. Though CI of intercept includes zero, it is often not a problem when you center the predictor X.

b. One reason is that because population(x) being zero is not within the slope of the model, it's meaningless though sales(y) interval contains negative numbers. The second thing is that negative lower confidence limit shows that it is not suitable to use linear model for modeling the relationship between sales product and populations. It is sign to let us to consider other models for modeling the relationship population and sales.

2.4

a. $\hat{\beta}_1 = 0.0388$, the 99 percent confidence interval will be $(0.0388 + t_{0.005}(118)\text{sd}(b_1), 0.0388 + t_{0.995}(118)\text{sd}(b_1))$, that is, $(0.005, 0.072)$. Zero is not included in confidence interval. This is important for director of admission since they can select the better candidate based on ACT score if there exists a positive relation between ACT score and GPA.

b. Null is $\beta_1 = 0$, the alternative is $\beta_1 \neq 0$. Then, the t statistic value will be $t^* = 0.0388/0.0128 = 3.04$. The $t_{0.995}(118) = 2.62$. Since $t^* > t_{0.995}(118)$. Hence, the null is rejected.

c. The P-value of corresponding test is 0.003 which is smaller than 0.01. Hence, the test is significant.

2.7

a. The estimated regression line is $y = 2.03x + 168.60$. Hence, the change of mean hardness is 2.03. Recall the formula that $\frac{b_1 - \beta_{10}}{\text{sd}(b1)} \sim t(n-2)$. Hence, the 99% confidence interval is $(b_1 + t_{0.005}(n-2)\text{sd}(b_1), b_1 + t_{0.995}(n-2)\text{sd}(b_1))$, that is, $(1.77, 2.30)$. The probability of true slope β_1 between 1.77 and 2.30 is 99%.

b. The null hypothesis is $\beta_1 = 2$, the alternative is $\beta_1 \neq 2$. The statistics we construct is

$$\frac{b_1 - \beta_{10}}{\text{sd}(b_1)},$$

where $b_1 = 2.30$, $\text{sd}(b_1) = 0.0904$ and $\beta_{10} = 2$. Then, the test statistic value is 0.38. We know that the critical cut-off is 2.977. Hence, the null is failed to reject. The P-value of the test is 0.71.

c. We calculate the distribution under the alternative $\beta_1 = 2.3$.

$$\begin{aligned} & \frac{b_1 - \beta_{10}}{\text{sd}(b_1)} \\ = & \frac{(b_1 - \beta_{10})/\sigma(b_1)}{\text{sd}(b_1)/\sigma(b_1)} \\ = & \frac{(b_1 - \beta_{1a} + \beta_{1a} - \beta_{10})/\sigma(b_1)}{\text{sd}(b_1)/\sigma(b_1)} \end{aligned}$$

We know that $\beta_{1a} = 0.23$, $\beta_{10} = 0.2$ and $\sigma(b_1) = 0.1$. Then, the distribution under alternative is $t(3, 14)$ with non central parameter equal to 3. Hence, the power is $P(t(3, 14) > 2.14 \text{ or } t(3, 14) < -2.14)$ which is equal to 0.53.

2.12

Recall that $\sigma^2\{\text{pred}\} = \sigma^2 + \sigma^2\{\hat{Y}_h\}$, hence $\sigma^2\{\text{pred}\}$ is always larger than σ^2 and cannot go to zero as long as there exists noise.

For $\sigma^2\{\hat{Y}_h\}$, we recall that

$$\sigma^2\{\hat{Y}_h\} = \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_i (X_i - \bar{X})^2} \right]. \tag{1}$$

Therefore, the first term goes to zero as n increases. For the second term, it could be away from zero when some of X_i's are really large compare with others. However, if X_i comes from certain distribution, $g(X)$. Then, the second term goes to zero by the law of large number.

The difference lies in that the test error cannot be avoided, but the training error can be shrunk under certain conditions.

2.13

a. $\hat{Y}_{28} = 3.20$, $\sigma^2(\hat{Y}_{28}) = 0.005$. Hence, the confidence interval for $\hat{Y}_{28}$ is $(3.06, 3.34)$.

b. $Y_{pred,28} = 3.20$, $\sigma^2_{pred} = 0.393$. Hence, the confidence interval for $Y_{pred,28}$ is $(1.96, 4.44)$.

c. The prediction interval is much wider than the confidence interval in part (a). This is because that there will be additional noise when a new observation comes in.

d. $\hat{Y}_{28}$ is still be 3.20. The band is $(3.2 - \sqrt{2F(0.975, 2, 118)}\sigma(\hat{Y}_{28}), 3.2 + \sqrt{2F(0.975, 2, 118)}\sigma(\hat{Y}_{28}))$, that is, $(3.00, 3.40)$. The confidence band is wider than the confidence interval in part (a). This is because that the band is for the regression line which consists two components β_0 and β_1. The interval is for the fixed fitted point which is linear combination of β_0 and β_1. Hence, we can use narrower interval to estimate the fix point, but need wider band to estimate the whole line.

2.51

Recall that $b_0 = \bar{Y} - \bar{X}b_1$, $\mathrm{E}Y_i = \beta_0 + \beta_1 X_i$, $\mathrm{E}\bar{Y} = \beta_0 + \beta_1 \bar{X}$ and

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}.$$

Then, we get that

$$\mathrm{E}b_1 = \frac{\sum(X_i - \bar{X})(\mathrm{E}Y_i - \mathrm{E}\bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{\sum(X_i - \bar{X})(X_i - \bar{X})\beta_1}{\sum(X_i - \bar{X})^2} = \beta_1.$$

Then, we have $\mathrm{E}b_0 = \mathrm{E}\bar{Y} - \bar{X}\mathrm{E}b_1 = \beta_0 + \beta_1\bar{X} - \beta_1\bar{X} = \beta_0$. Hence, b_0 is unbiased.

2.52

We want to compute $\mathrm{var}b_0$, that is,

$$\begin{aligned}
\mathrm{var}\{b_0\} &= \mathrm{var}\{\bar{Y} - \bar{X}b_1\} \\
&= \mathrm{var}\{\bar{Y} - \bar{X}\frac{\sum_i(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i(X_i - \bar{X})^2}\} \\
&= \mathrm{var}\{\sum_i \big[\frac{1}{n} - \frac{\bar{X}(X_i - \bar{X})}{\sum_i(X_i - \bar{X})^2}\big]Y_i\} \\
&= \sum_i \big\{\frac{1}{n^2} + \frac{\bar{X}^2(X_i - \bar{X})^2}{\big[\sum_i(X_i - \bar{X})^2\big]^2} - \frac{2}{n}\frac{\bar{X}(X_i - \bar{X})}{\sum_i(X_i - \bar{X})^2}\big\}\sigma^2 \\
&= \big[\frac{1}{n} + \frac{\bar{X}^2}{\sum_i(X_i - \bar{X})^2}\big]\sigma^2
\end{aligned}$$

This is what we want. Since, b_0 can be seen as $b_0 + b_1 \cdot 0$ when $X_h = 0$. Hence, if we plug zero into equation (2.29b), it becomes the above formula. Therefore, variance of b_0 is the special case of variance (2.29b).

2.63

$\hat{\beta}_1$ for region 1 is 0.52. The corresponding interval is $(0.46, 0.58)$. $\hat{\beta}_2$ for region 2 is 0.24. The corresponding interval is $(0.19, 0.28)$. $\hat{\beta}_3$ for region 3 is 0.33. The corresponding interval is $(0.29, 0.38)$. $\hat{\beta}_4$ for region 4 is 0.44. The corresponding interval is $(0.36, 0.51)$. From the above display, it seems that the slope of region 1 is little bit higher than region 2 - 4. If we want to draw a more formal conclusion, we need to do a two sample comparison test which has been covered yet.