HW5

Ting Huang

February 28, 2016

Question 2.1

(a)

Not quite. The the 95% confidence interval of the slope doesn't contain zero, and therefore there is indeed an association between X and Y. However, the available information is not enough to state that the association is linear in nature.

(b)

The value $X = 0$, which means there is no people in the marketing districts, is not in the scope of the model. So the incercept doesn't provide meaningful information in this case. Moreover, the 95% confidence limit contain both positive and negative values. Therefore we don't need to worry about the negative lower confidence limit for the intercept.

Question 2.2

We can't conclude that there is no linear association between X and Y since the H_0 hypothesis is $\beta_1 \leq 0$. It is possible that $\beta_1 < 0$, which indicates there is a negative association between X and Y. More generally, we never 'accept' H_0, but say that there is no evidence against H_0 (it is still possible that the null hypothesis is still wrong, but the dataset does not have enough statistical power to reject it).

Question 2.3

The statement is incorrect. The two-sided p-value for estimated slope is 0.91, which is much larger than the commom signigicant level 5%. So we fail to reject the null hypothesis $\beta_1 = 0$, and conclude that there is no evidence against no association between X and Y. A more accurate conclusion is that there is no evidence that the spending impacts the sales.

Question 2.7

(a)

```
X<-c(16.0,16.0,16.0,16.0,24.0,24.0,24.0,24.0,32.0,32.0,32.0,32.0,40.0,40.0,40.0,40.0)
Y<-c(199.0,205.0,196.0,200.0,218.0,220.0,215.0,223.0,237.0,234.0,235.0,230.0,250.0,248.0,253.0,246.0)
plastic<-data.frame("elapse_time"=X,"hardness"=Y)
plastic.lm <- lm(hardness ~ elapse_time, data=plastic)
residuals<-resid(plastic.lm)
MSE=sum(residuals^2)/(16-2)
s2_b1=MSE/sum((X-mean(X))^2)
s2_b1
```

```
## [1] 0.008171038
```

```
s_b1=sqrt(s2_b1)
s_b1 #Standard error of b1
```

```
## [1] 0.09039379
```

For a 99 percent confidence interval, we require t(0.995;14)=2.976843. The 99% confidence interval is

$$(2.034 - 2.976843 * 0.0904, 2.034 + 2.976843 * 0.0904) = (1.764893, 2.303107)$$

The confidence interval doesn't contain zero value, so there is an association between the elapsed time and the hardness of the plastic.

(b)

$H_0 : \beta = 2$ and $\beta \neq 2$.

The test statistic is:

$$t^* = \frac{b_1 - 2}{s\{b_1\}} = \frac{2.034 - 2}{0.0904} = 0.3761062$$

. The decision rule for significant level 1% is $|t^*| \leq t(0.995; 14)$. The p-value is $2 \cdot P(t > t^*) = 0.7024056$. There is no evidence against H_0.

(c)

$\delta = \frac{0.3}{\sigma\{b_1\}} = 3$ and $df = 14$. So the power is $P\{|t^*| > t(1 - \alpha/2; 14)\} = 0.5279026$.

Question 2.9

Because the value of $s^2\{\hat{Y}_h\}$ depends on the specific observation X_h. So for each X_i in the sample, there exists a corresponding $s^2\{\hat{Y}_h\}$. It is not necessary to print all the $s^2\{\hat{Y}_h\}$, especially when the sample size is large.

Question 2.10

(a)

A prediction interval for a new observation is appropriate since we predict the humidity level in this greenhouse tomorrow.

(b)

A mean response is appropriate since we estimate the average cost of the families.

(c)

A prediction interval for a new observation is appropriate since we preidct the consume of electricity next month.

Question 2.11

Yes. There is difference. In $E\{Y_h\}$, we estimate the mean of the distribution of Y. In $Y_{h(new)}$, we predict an individual instance drawn from the distribution of Y_h. In the mean of m values of $Y_{h(new)}$, we predict the outcome of m new instances drawn from the distribution of Y_h. In the first case, the uncertainty comes from the estimation of the regression line. In the last two cases, the uncertainty comes from both the estimation of the regression line, and of the variation of the individual instances around the line.

Question 2.16

(a)

$$\hat{Y_h} = 168.6 + 2.034 \cdot 30 = 229.62$$

$$s^2\{\hat{Y_h}\} = 10.45893[\frac{1}{16} + \frac{(30-28)^2}{1280}] = 0.6863673$$

$$s\{\hat{Y_h}\} = 0.8284729$$

Given $t(0.99; 14) = 2.624494$, 98% confidence interval is

$$229.62 - 2.624494 * 0.8284729 \leq E\{Y_h\} \leq 229.62 + 2.624494 * 0.8284729$$

$$227.4457 \leq E\{Y_h\} \leq 231.7943$$

(b)

$$\hat{Y_h} = 168.6 + 2.034 * 30 = 229.62$$

$$s^2\{pred\} = s^2\{\hat{Y_h}\} + MSE = 0.6863673 + 10.45893 = 11.1453$$

$$s\{pred\} = 3.338458$$

Given $t(0.99; 14) = 2.624494$, 98% confidence interval is

$$229.62 - 2.624494 * 3.338458 \leq Y_{h(new)} \leq 229.62 + 2.624494 * 3.338458$$

$$220.8582 \leq Y_{h(new)} \leq 238.3818$$

(c)

$$\hat{Y_h} = 168.6 + 2.034 * 30 = 229.62$$

$$s^2\{predmean\} = s^2\{\hat{Y_h}\} + \frac{MSE}{10} = 0.6863673 + 1.045893 = 1.73226$$

$$s\{predmean\} = 1.316153$$

Given $t(0.99; 14) = 2.624494$, 98% confidence interval is

$$229.62 - 2.624494 * 1.316153 \leq Y_{h(new)} \leq 229.62 + 2.624494 * 1.316153$$

$$226.1658 \leq Y_{h(new)} \leq 233.0742$$

(d)

The prediction interval in part (c) is narrower than that in part (b) because part (c) involves a prediction of the mean hardness for 10 newly molded test items. The variance in part (c) is smaller than that of part (b).

(e)

$$\hat{Y_h} = 168.6 + 2.034 * 30 = 229.62$$

$$s\{\hat{Y_h}\} = 0.8284729$$

$$W^2 = 2F(1-\alpha; 2, n-2) = 2F(0.99; 2, 14) = 2 * 5.24075 = 10.4815$$

$$W = 3.237514$$

98% confidence band is

$$229.62 - 3.237514 * 0.8284729 \leq Y_{h(new)} \leq 229.62 + 3.237514 * 0.8284729$$

$$226.9378 \leq Y_{h(new)} \leq 232.3022$$

The confidence interval in part (d) is wider than that in part (a). The W multiple is larger than the t multiple because the confidence band must encompass the entire regression line, whereas the confidence limits apply only at the single level X_h.

Question 2.17

In first case, α level should be greater than 0.33 in order to reject the null hypothesis.

When $\alpha = 0.01 <$ P-value, we fail to reject the null hypothesis.