LINEAR REGRESSION MODELS: Homework #2

Due on September 25, 2017

Professor Jingchen Liu

Fan Yang
UNI: fy2232

Problem 1

a.

My answer is yes. There is a linear association between Y and X. Because in the diagram above, we can find 95% confidence interval. And the estimated slope value is 0.755048, which lies in the 95% confidence interval [0.452886, 1.05721]. What's more, it is now obvious that the implied level of significance is 95%.

b.

The plausible reason for this question is 'dollar sales cannot be negative even if the population in a district is zero.' But actually population in a district can not be 0. What's more, x=0 is not included in our sample space. So there is no sense considering the value of sales when x=0.

Problem 2

a.
The summary of model in Problem 1.19 lists as below:

```
> summary(lm.pr19)

Call:
lm(formula = da$V1 ~ da$V2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.74004 -0.33827  0.04062  0.44064  1.22737

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
da$V2        0.03883    0.01277   3.040  0.00292 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

As we already know the confidence interval of β_1 : $[\widehat{\beta}_1 - \xi_{0.005} SD(\beta_1), \widehat{\beta}_1 + \xi_{0.005} SD(\beta_1)]$
Since $\xi_{0.005} = 2.58$, and we know $SD(\beta_1) = 0.01277$, therefore, the CI is:
$[0.03883 - 2.58 \times 0.01277, 0.03883 + 2.58 \times 0.01277]$, which is
$[0.0058834, 0.0717766]$

this interval does not include 0.
if the interval includes 0, which means GPA(Y) can not be predicted by ACT test score(X) since they shows little asso ciation. And this conclusion could reach a confidence level of 99%.

b.
two alternatives are:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

test statistic is

$$t^* = \frac{\widehat{\beta}_1}{sd(\beta)}$$

when $t^* > t(1-\alpha/2, n-2)$, we reject H_0; otherwise, conclude H_0.

Calculate t^* first: $t^* = \frac{\widehat{\beta}_1}{sd(\beta)} = \frac{0.03883}{0.01277} = 3.04072$

And we can get from table that $t(1-\alpha/2, n-2) = t(1-0.005, 118) \approx 2.58 < t^* = 3.04072$.

Therefore, we can conclude that $\beta_1 \neq 0$.

Which means linear association exists between student's ACT score (X) and GPA.

c.

The two-sided P-value for the sample outcome is obtained by first finding the onesided P-value, $Pr(|t(118)| > t* = 3.04072)$.

We use R to compute that this probability is about 0.003.

Since the P-value is less 0.01, which is the level of significance.Therefore we can conclude H_1 directly.

Problem 3

a. we use R to get the summary of the model:

```
Call:
lm(formula = d22$V1 ~ d22$V2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1500 -2.2188  0.1625  2.6875  5.5750

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 168.60000    2.65702   63.45  < 2e-16 ***
d22$V2        2.03438    0.09039   22.51 2.16e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The regression function is $Y_i = 168.6 + 2.03438X_i$

We already know that $\frac{\widehat{\beta}_1 - \beta_1}{sd(\widehat{\beta}_1)}$ follows $t(n-2)$

Thus, $[\widehat{\beta}_1 - t(0.005, 14)sd(\beta_1), \widehat{\beta}_1 + t(0.005, 14)sd(\beta_1)]$ is the confidence interval.

$\widehat{\beta}_1 = 2.03438$ and $t(0.005, 14) = 2.98$ and $sd(\beta_1) = 0.09039$

So the confidence interval is$[2.03438 - 0.09039 \times 2.98, 2.03438 + 0.09039 \times 2.98]$

which is $[1.765018, 2.303742]$

This is also interval estimate for change in the mean hardness.

b. two alternatives are:

$$H_0 : \beta_1 = 2$$
$$H_1 : \beta_1 \neq 2$$

test statistic is

$$t^* = \frac{\widehat{\beta}_1 - \beta_1}{sd(\widehat{\beta}_1)}$$

when $t^* > t(1 - \alpha/2, n - 2)$, we reject H_0; otherwise, conclude H_0.

Calculate t^* first: $t^* = \dfrac{\widehat{\beta}_1 - \beta_1}{sd(\widehat{\beta}_1)} = \dfrac{2.03438 - 2}{0.09039} = 0.3803518$

And we can get from table that $t(1 - \alpha/2, n - 2) = t(1 - 0.005, 14) \approx 2.98 > t^* = 0.3803518$.

Therefore, we can not reject H_0. So we conclude that $\beta_1 = 2$.

The two-sided P-value for the sample outcome is obtained by first finding the onesided P-value, $Pr(|t(14)| > t* = 0.3803518)$.

We use R to compute that this probability is about 0.7093927. So P-value is 0.7093927.

c.

From textbook equation (2.27) we know that

$$\delta = \frac{|\beta_1 - \beta_{10}|}{\sigma(\widehat{\beta}_1)}$$
$$= \frac{0.3}{\sigma(0.1)} = \frac{0.3}{\sigma(0.1)} = 3$$

And $\alpha = 0.01$ $\delta = 3$, we can find the power using table: Power=0.53.

Problem 4

Let's derive $\sigma^2\{\widehat{Y}_h\}$ first:

$\sigma^2\{\widehat{Y}_h\} = \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right] = \sigma^2 \left[\frac{1}{n} + \frac{\frac{1}{n}(X_h - \overline{X})^2}{\frac{1}{n}\sum(X_i - \overline{X})^2}\right]$

as n becomes large, $\frac{1}{n} \to 0$

According to law of large number, when n is large enough

$\frac{1}{n}\sum(X_i - \overline{X})^2 \to E((X_i - \overline{X})^2)$; While $E(X_i - \overline{X}) = 0$

Thus $\frac{1}{n}\sum(X_i - \overline{X})^2 \to var(X)$

Therefore $\sigma^2\{\widehat{Y}_h\} \to \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{n \times var(X)}\right] \to \sigma^2 [0 + 0] = 0$

Now we will consider $\sigma^2\{pred\}$:

$\sigma^2\{pred\} = \sigma^2 + \sigma^2\{\widehat{Y}_h\}$

Although $\sigma^2\{\widehat{Y}_h\} \to 0$ while n becomes large,

σ^2 is variance of the distribution of Y at $X = X_h$ and will always lager than 0

So $\sigma^2\{pred\}$ can not be close to 0

We can conclude that with sample size becomes larger, we can get accurate mean of Y; but the prediction mean could not be accurately got no matter how large n is.

Problem 5

a.

The $1-\alpha$ confidence limits are:

$$\widehat{Y}_h \pm t(1-\alpha/2; n-2)s\{\widehat{Y}_h\}$$

where

$\widehat{Y}_h = 2.11405 + 0.03883 * 28 = 3.20129$

$$s^2\{\widehat{Y}_h\} = \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right]$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2}\right]$$

While $\widehat{\sigma}^2 = MSE = 0.3882848$

$$\frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2} = \frac{(28 - 24.725)^2}{\sum(X_i - 24.725)^2} = 0.004506707$$

So $s\{\widehat{Y}_h\} = \sqrt{0.3882848 * (\frac{1}{118} + 0.004506707)} = 0.07099602$

And we know $t(0.975, 118) = 1.980272$, therefore the confidence limits is

$3.20129 \pm 1.980272 * 0.07099602$ which is $[3.060699,\ 3.341881]$

b.

The $1-\alpha$ confidence interval is:

$$\widehat{Y}_h \pm t(1-\alpha/2; n-2)s\{pred\}$$

We already know that:

$\widehat{Y}_h = 3.20129 \quad MSE = 0.3882848 \quad s\{\widehat{Y}_h\} = 0.07099602$ and $t(0.975, 118) = 1.980272$

Then we need to get $s\{pred\}$

$$s^2\{pred\} = MSE + s^2\{\widehat{Y}_h\} = 0.3882848 + 0.07099602^2 = 0.3933252$$

so$s\{pred\} = 0.6271564$

Therefore the confidence limits is

$3.20129 \pm 1.980272 * 0.6271564$ which is $[1.95935,\ 4.44323]$

c.

The prediction interval in part (b) is wider than the confidence interval in part (a).

Because the prediction for new observation involves new error randomness. In order to get the same confidence level, it must be wider.

d.

The two boundary values at $1-\alpha$ confidence level is:

$\widehat{Y}_h \pm Ws\{\widehat{Y}_h\}$

where

$W^2 = 2F(1-\alpha; 2, n-2) = 2F(0.95; 2, 118) = 6.146181$

So $W = 2.479149$

We already know that:

$\widehat{Y}_h = 3.20129 \quad MSE = 0.3882848 \quad s\{\widehat{Y}_h\} = 0.07099602$

Therefore the confidence band are

$3.02528 \leq \beta_0 + \beta_1 X_h \leq 3.3773$

The confidence band is wider at this point than the confidence interval in part (a)
Because the bands must cover all possible values no matter what observation is, which means the same level of confidence interval estimate of the mean must be included in the bands.

Problem 6

$$
\begin{aligned}
b_0 &= \overline{Y} - b_1\overline{X} \\
E(b_0) &= E(\overline{Y} - b_1\overline{X}) \\
&= E(\beta_0 + \beta_1\overline{X} + \overline{\varepsilon} - b_1\overline{X}) \\
&= E(\beta_0 + \overline{X}(\beta_1 - b_1)) \\
&= \beta_0 + \overline{X}(\beta_1 - E(b_1)) \quad (b_1 \text{ is unbiased.}) \\
&= \beta_0
\end{aligned}
$$

Thus, b_0 is a unbiased estimator of β_0.

Problem 7

$$
\begin{aligned}
\sigma^2\{b_0\} &= var(\overline{y} - b_1\overline{X}) \\
&= var(\overline{y}) + var(b_1\overline{X}) - 2cov(\overline{y}, b_1\overline{X}) \\
&= var(\overline{y}) + \overline{X}^2 var(b_1) - 2\overline{X}^2 cov(\overline{y}, b_1) \quad (\text{using } (2.31), then :) \\
&= var(\overline{y}) + \overline{X}^2 var(b_1) \\
&= var(\frac{1}{n}\sum y_i) + \overline{X}^2 \frac{\sigma^2}{\sum(X_i - \overline{X})^2} \\
&= \frac{1}{n^2}\sum var(y_i) + \frac{\sigma^2\overline{X}^2}{\sum(X_i - \overline{X})^2} \\
&= \frac{1}{n^2}\sum \sigma^2 + \frac{\sigma^2\overline{X}^2}{\sum(X_i - \overline{X})^2} \\
&= \frac{1}{n}\sigma^2 + \frac{\sigma^2\overline{X}^2}{\sum(X_i - \overline{X})^2} \\
&= \sigma^2(\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2})
\end{aligned}
$$

$$\sigma^2\{\widehat{Y}_h\} = \sigma^2(\frac{1}{n} + \frac{(X_h - \overline{X})^2}{\sum(X_i - \overline{X})^2})$$

while X_h is 0 $\sigma^2\{\widehat{Y}_h\} = \sigma^2\{b_0 + b_1 * 0\} = \sigma^2\{b_0\} = \sigma^2(\frac{1}{n} + \frac{\overline{X}^2}{\sum(X_i - \overline{X})^2})$

That's to say, variance (2.22b) is a special case of variance (2.29b) when $X_h = 0$

Problem 8

the $1 - \alpha$ confidence limits for β_1 are:

$$b_1 \pm t(1 - \alpha/2; n - 2)sd\{b_1\}$$

For the **first** region

```
> b_1 <- lm.cdi1$coefficients[2]
> b_1
522.1588
> t <- qt(0.95,lm.cdi1$df.residual)
> sd2 <- sum((lm.cdi1$residuals)^2)/lm.cdi1$df.residual / sum((X-mean(X))^2)
> b_1-t*sqrt(sd2)
460.5177
> b_1+t*sqrt(sd2)
583.8
```

interval estimate of β_1 is [460.5177, 583.8]

For the **second** region

```
> b_1 <- lm.cdi2$coefficients[2]
> b_1
238.6694
> t <- qt(0.95,lm.cdi2$df.residual)
> sd2 <- sum((lm.cdi2$residuals)^2)/lm.cdi2$df.residual / sum((X-mean(X))^2)
> b_1-t*sqrt(sd2)
193.4858
> b_1+t*sqrt(sd2)
283.853
```

interval estimate of β_1 is [193.4858, 283.853]

For the **third** region

```
> b_1 <- lm.cdi3$coefficients[2]
> b_1
330.6117
> t <- qt(0.95,lm.cdi3$df.residual)
> sd2 <- sum((lm.cdi3$residuals)^2)/lm.cdi3$df.residual / sum((X-mean(X))^2)
> b_1-t*sqrt(sd2)
285.7076
> b_1+t*sqrt(sd2)
375.5158
```

interval estimate of β_1 is [285.7076, 375.5158]

For the **forth** region

```
> b_1 <- lm.cdi4$coefficients[2]
> b_1
440.3157
> t <- qt(0.95,lm.cdi4$df.residual)
> sd2 <- sum((lm.cdi4$residuals)^2)/lm.cdi4$df.residual / sum((X-mean(X))^2)
> b_1-t*sqrt(sd2)
364.7585
> b_1+t*sqrt(sd2)
515.8729
```

interval estimate of β_1 is [364.7585, 515.8729]

Different regression lines of the 4 regions have different slopes as shown above. Their values of bound of interval vary from each other very much.