Chapter 3 Assessment

Hannah Xiao Si Laws

Mar 18, 2016

Directions: Strike-through false statements using `~~strikethrough~~`. Bold all true statements and answers. By entering your name on the document you turn in, you are acknowledging that the work in the document is entirely your own unless specified otherwise in the document. Compile your document using `Knit PDF` and turn in a stapled hardcopy no later than 10am, Friday March 18, 2016. Create a directory named `ChapterThreeAssessment` inside your private class repository. Store this file inside the `ChapterThreeAssessment` directory. Use inline `R` expressions rather than hardcoding your numeric answers. Hand write the eight digit SHA for the document you turn in next to your name.

1. Why is linear regression important to understand? Select all that apply:

- ~~The linear model is often correct.~~
- ~~Linear regression is very extensible and can be used to capture nonlinear effects.~~
- **Simple methods can outperform more complex ones if the data are noisy.**
- **Understanding simpler methods sheds light on more complex ones.**

2. You may want to reread the paragraph on confidence intervals on page 66 of the textbook before trying this question (the distinctions are subtle). Which of the following are true statements? Select all that apply:

- ~~A 95% confidence interval formula is a random interval that is expected to contain the true parameter 95% of the time.~~
- ~~The true parameter is a random value that has 95% chance of falling in the 95% confidence interval.~~
- **I perform a linear regression and get a 95% confidence interval from 0.4 to 0.5. There is a 95% probability that the true parameter is between 0.4 and 0.5.**
- **The true parameter (unknown to me) is 0.5. If I repeatedly sample data and construct 95% confidence intervals, the intervals will contain 0.5 approximately 95% of the time.**

3. We run a linear regression and the slope estimate is 0.5 with estimated standard error of 0.2. What is the largest value of b for which we would NOT reject the null hypothesis that $\beta_1 = b$?

a. Assume a normal approximation to the t distribution, and that we are using the 5% significance level for a two-sided test; use two significant digits of accuracy.

```
B1 = 0.5
SE = 0.2

B1-(2*SE)
```

```
[1] 0.1
```

```
B1+(2*SE)
```

```
[1] 0.9
```

The answer is 0.9

b. Use a t distribution with 10 degrees of freedom, and assume that we are using the 5% significance level for a two-sided test; use two significant digits of accuracy.

```
be1= 2
se = .2
tvalue = 2.2282 #look this up with t distribution tables
round(be1 + (tvalue*se),1)
```

```
[1] 2.4
```

The answer is 2.4

4. Which of the following indicates a fairly strong relationship between X and Y?

- $R^2 = 0.9$
- ~~The p-value for the null hypothesis $\beta_1 = 0$ is 0.0001.~~
- ~~The t-statistic for the null hypothesis $\beta_1 = 0$ is 30.~~

5. Given the following:

```r
site <- "http://www-bcf.usc.edu/~gareth/ISL/Credit.csv"
Credit <- read.csv(file = site)
str(Credit)
```

```
'data.frame':   400 obs. of  12 variables:
 $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Income   : num  14.9 106 104.6 148.9 55.9 ...
 $ Limit    : int  3606 6645 7075 9504 4897 8047 3388 7114 3300 6819 ...
 $ Rating   : int  283 483 514 681 357 569 259 512 266 491 ...
 $ Cards    : int  2 3 4 3 2 4 2 2 5 3 ...
 $ Age      : int  34 82 71 36 68 77 37 87 66 41 ...
 $ Education: int  11 15 11 11 16 10 12 9 13 19 ...
 $ Gender   : Factor w/ 2 levels "Female"," Male": 2 1 2 1 2 2 1 2 1 1 ...
 $ Student  : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 2 ...
 $ Married  : Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 1 1 1 2 ...
 $ Ethnicity: Factor w/ 3 levels "African American",..: 3 2 2 2 3 3 1 2 3 1 ...
 $ Balance  : int  333 903 580 964 331 1151 203 872 279 1350 ...
```

```r
ModEthnic <- lm(Balance ~ Ethnicity, data = Credit)
summary(ModEthnic)
```

```

Call:
lm(formula = Balance ~ Ethnicity, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max 
-531.00 -457.08  -63.25  339.25 1480.50 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)          531.00      46.32  11.464   <2e-16 ***
EthnicityAsian       -18.69      65.02  -0.287    0.774    
EthnicityCaucasian   -12.50      56.68  -0.221    0.826    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 460.9 on 397 degrees of freedom
Multiple R-squared:  0.0002188, Adjusted R-squared:  -0.004818 
F-statistic: 0.04344 on 2 and 397 DF,  p-value: 0.9575
```

a. According to the balance vs ethnicity model (`ModEthnic`), what is the predicted balance for an Asian in the data set? (within 0.01 accuracy)

512.31

b. What is the predicted balance for an African American? (within .01 accuracy)

531

c. Construct a 90% confidence interval for the average credit card balance for Asians.

```
pp<-confint(ModEthnic, level = .9);pp
```

```
                          5 %      95 %
(Intercept)         454.6343 607.36565
EthnicityAsian     -125.8866  88.51403
EthnicityCaucasian -105.9526  80.94756
```

The 90% confidence interval for the average credit card balance for Asians is -125.8865769, 88.5140279

d. Construct a 92% prediction interval for Joe's (who is African American) credit card balance.

```
dd<-predict(ModEthnic, newdata = data.frame(Ethnicity = "African American"), interval = "pred", level =
```

```
  fit       lwr      upr
1 531 -281.9757 1343.976
```

The 92% prediction interval for the average credit card balance for Joe is 531, -281.975745, 1343.975745

6. Given the following:

```r
mod <- lm(Rating ~ poly(Limit, 2, raw = TRUE) + poly(Cards, 2, raw = TRUE) +
            Married + Student + Education, data = Credit)
summary(mod)
```

```
Call:
lm(formula = Rating ~ poly(Limit, 2, raw = TRUE) + poly(Cards,
    2, raw = TRUE) + Married + Student + Education, data = Credit)

Residuals:
     Min       1Q   Median       3Q      Max
-27.8814  -6.8317  -0.3358   6.5136  25.9925

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.579e+01  3.816e+00   6.760 5.01e-11 ***
poly(Limit, 2, raw = TRUE)1  6.529e-02  7.506e-04  86.984  < 2e-16 ***
poly(Limit, 2, raw = TRUE)2  1.320e-07  6.297e-08   2.096   0.0368 *
poly(Cards, 2, raw = TRUE)1  7.615e+00  1.301e+00   5.855 1.01e-08 ***
poly(Cards, 2, raw = TRUE)2 -3.972e-01  1.783e-01  -2.228   0.0264 *
MarriedYes                   2.295e+00  1.043e+00   2.199   0.0285 *
StudentYes                   3.159e+00  1.693e+00   1.866   0.0628 .
Education                   -2.774e-01  1.627e-01  -1.705   0.0889 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.09 on 392 degrees of freedom
Multiple R-squared:  0.9958,    Adjusted R-squared:  0.9957
F-statistic: 1.334e+04 on 7 and 392 DF,  p-value: < 2.2e-16
```

a. Use `mod` to predict the `Rating` for an individual that has a credit card limit of \$6,000, has 4 credit cards, is married, and is not a student, and has an undergraduate degree (`Education` $= 16$).

```r
ff<-predict(mod, newdata = data.frame(Limit = 6000, Cards = 4, Married = "Yes", Student = "No", Educati
```

444.2525851

b. Use `mod` to predict the `Rating` for an individual that has a credit card limit of \$12,000, has 2 credit cards, is married, is not a student, and has an eighth grade education (`Education` $= 8$).

```r
ee<-predict(mod, newdata = data.frame(Limit = 12000, Cards = 2, Married = "Yes", Student = "No", Educati
```

842.0091055

c . Construct and interpret a 90% confidence interval for β_5 (a married person).

```r
cc<-confint(mod, level = .9)
```

The confidence interval for Married People is 5.4705001, 9.7591851

7. Given the following:

```r
site <- "http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv"
Advertising <- read.csv(file = site)
str(Advertising)
```

```
'data.frame':   200 obs. of  5 variables:
 $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ TV       : num  230.1 44.5 17.2 151.5 180.8 ...
 $ Radio    : num  37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2.6 ...
 $ Newspaper: num  69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...
 $ Sales    : num  22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...
```

```r
modSales <- lm(Sales ~ TV + Radio + TV:Radio, data = Advertising)
summary(modSales)
```

```

Call:
lm(formula = Sales ~ TV + Radio + TV:Radio, data = Advertising)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.3366 -0.4028  0.1831  0.5948  1.5246 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.750e+00  2.479e-01  27.233   <2e-16 ***
TV          1.910e-02  1.504e-03  12.699   <2e-16 ***
Radio       2.886e-02  8.905e-03   3.241   0.0014 ** 
TV:Radio    1.086e-03  5.242e-05  20.727   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9435 on 196 degrees of freedom
Multiple R-squared:  0.9678,    Adjusted R-squared:  0.9673 
F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

```r
coef(modSales)
```

```
(Intercept)          TV       Radio    TV:Radio 
6.750220203 0.019101074 0.028860340 0.001086495 
```

```r
b0<-coef(summary(modSales))[1,1]
b1<-coef(summary(modSales))[2, 1]
b2<-coef(summary(modSales))[3,1]
b3<-coef(summary(modSales))[4,1]
```

a. According to the model for sales vs TV interacted with radio (`modSales`), what is the effect of an additional 1 unit of radio advertising if TV = 25? (with 4 decimal accuracy)

```r
rr<-round(b2+b3*25,4)
```

0.056

b. What if TV = 300? (with 4 decimal accuracy)

```r
rer<-round(b2+b3*300,4)
```

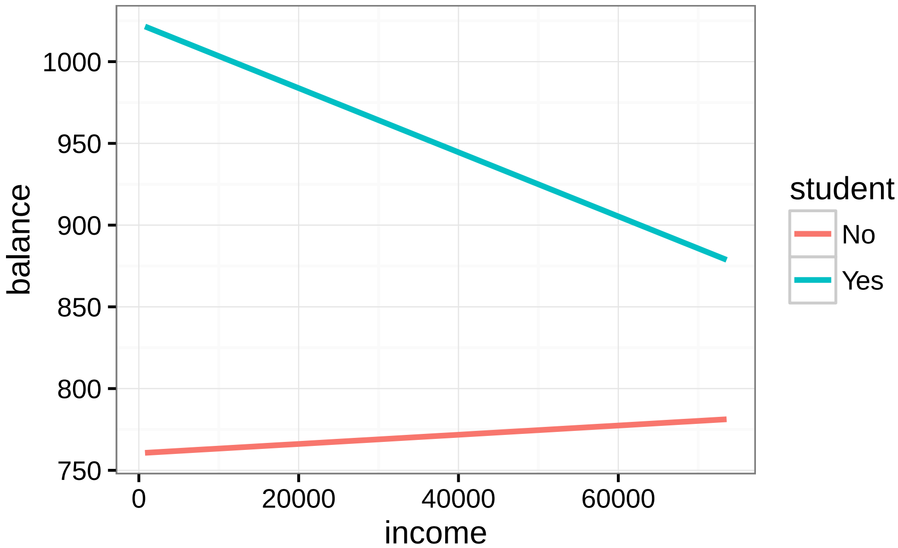
0.3548

8. What is the difference between `lm(y ~ x*z)` and `lm(y ~ I(x*z))`, when `x` and `z` are both numeric variables?

- **The first one includes an interaction term between `x` and `z`, whereas the second uses the product of `x` and `z` as a predictor in the model.**
- ~~The second one includes an interaction term between `x` and `z`, whereas the first uses the product of `x` and `z` as a predictor in the model.~~
- ~~The first includes only an interaction term for `x` and `z`, while the second includes both interaction effects and main effects.~~
- ~~The second includes only an interaction term for `x` and `z`, while the first includes both interaction effects and main effects.~~

9. Given the following model:

```r
modBalance <- lm(balance ~ student + income + student:income, data = Default)
library(ggplot2)
ggplot(data = Default, aes(x = income, y = balance, color = student)) +
  geom_smooth(method = "lm", se = FALSE, fullrange = TRUE) +
  theme_bw()
```




```r
summary(modBalance)
```

```

Call:
lm(formula = balance ~ student + income + student:income, data = Default)

Residuals:
     Min       1Q   Median       3Q      Max
-1004.74  -347.04   -10.84   320.42  1724.01

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.605e+02  2.323e+01  32.733  < 2e-16 ***
studentYes         2.625e+02  4.256e+01   6.169 7.15e-10 ***
income             2.819e-04  5.633e-04   0.501    0.617
studentYes:income -2.243e-03  2.007e-03  -1.118    0.264
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 473.6 on 9996 degrees of freedom
Multiple R-squared:  0.04157,	Adjusted R-squared:  0.04128
F-statistic: 144.5 on 3 and 9996 DF,  p-value: < 2.2e-16
```

Which of the following statements are true?

- **In the `modBalance` model, the estimate of β_3 is negative.**

- ~~One advantage of using linear models is that the true regression function is often linear.~~
- ~~If the F statistic is significant, all of the predictors have statistically significant effects.~~
- ~~In a linear regression with several variables, a variable has a positive regression coefficient if and only if its correlation with the response is positive.~~