Chapter 3 Assessment

Huy Tu

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.

```
StdE <- 0.2
b_hat <- 0.5
z <- qnorm(.975) #get z-score
b_lower <- b_hat - z*StdE
b_upper <- b_hat + z*StdE
b_upper <- round(b_upper,digits = 2) #round to 2 digits
```

The largest value of b for which we would not reject the null hypothesis that $\beta_1 = b$ is b = 0.89

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.

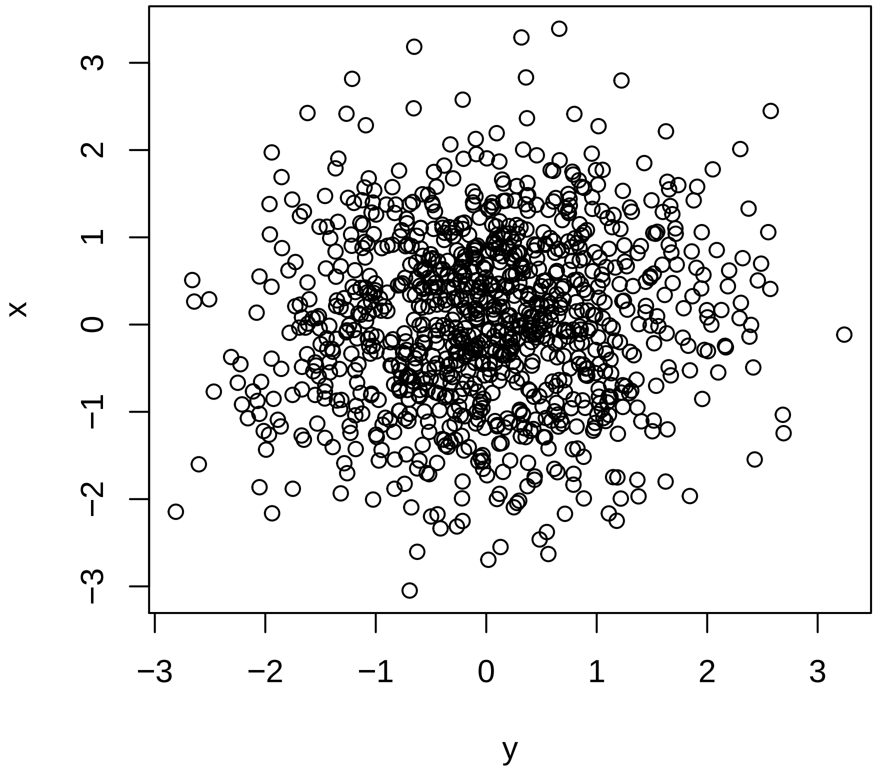
```
dfr <- 10
t <- qt(0.95,df=dfr ) #get t-score
b_upper <- b_hat + t*StdE
b_upper <- round(b_upper, digits = 2) #round to 2 digits
```

The largest value of b for which we would not reject the null hypothesis that $\beta_1 = b$ is b = 0.86

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.~~

```
set.seed(123)
y <- rnorm(1000)
x <- rnorm(1000)
plot(x~y)
```



```
mod <- lm(y~x)
summary(mod)
```

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7168 -0.6290 -0.0060  0.6451  3.2383 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  0.01252    0.03129   0.400  0.68909   
x            0.08494    0.03097   2.742  0.00621 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9885 on 998 degrees of freedom
Multiple R-squared:  0.007479,  Adjusted R-squared:  0.006484 
F-statistic:  7.52 on 1 and 998 DF,  p-value: 0.006211
```

5. Given the following:

```
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 ...
```

```
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
```

```
b0 <- coef(summary(ModEthnic))[1, 1]
b1 <- coef(summary(ModEthnic))[2, 1]
b2 <- coef(summary(ModEthnic))[3, 1]
c(b0, b1, b2)
```

```
[1] 531.00000 -18.68627 -12.50251
```

```r
AsianPredB <- b0 + b1 #the predicted balance for an Asian in the data set
AsianPredB <- round(AsianPredB,digits = 2)
AfAmPredB <- round(b0,digits = 2) #the predicted balance for an African American in the data set
```

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)

According to the balance vs ethnicity model (`ModEthnic`), the predicted balance for an Asian in the data set is 512.31.

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

According to the balance vs ethnicity model (`ModEthnic`), the predicted balance for an African American in the data set is 531.

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

```r
AsC <- predict(ModEthnic, newdata = data.frame(Ethnicity = "Asian"),
    interval = "confidence", level = 0.9)
```

A 90% confidence interval for Asians credit card balance is 437.0794784 to 587.5479725.

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

```r
AfAmC <- predict(ModEthnic, newdata = data.frame(Ethnicity = "African American"),
    interval = "prediction", level = 0.92)
```

A 92% prediction interval for Joe's (who is African American) credit card balance is -281.975745 to 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
#necessary variables's values
cN <- 4
cLimit <- 6000
edu <- 16
#necessary coefficients and intercept for the variables of the predicted model
C <- coef(summary(mod))[1, 1]
lC1 <- coef(summary(mod))[2, 1]
lC2 <- coef(summary(mod))[3, 1]
CC1 <- coef(summary(mod))[4, 1]
CC2 <- coef(summary(mod))[5, 1]
MYC <- coef(summary(mod))[6, 1]
SYC <- coef(summary(mod))[7, 1]
EC <- coef(summary(mod))[8, 1]
#predict the rating by using the model "mod"
ratingR <- C + cLimit*lC1 + (cLimit^2)*lC2 + cN*CC1 + (cN^2)*CC2 + (1)*MYC + (0)*SYC
ratingR
```

```
[1] 448.6916
```

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 is 448.691642.

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
#necessary variables's values
cN <- 2
cLimit <- 12000
edu <- 8
#necessary coefficients and intercepts for the variables of the predicted model
C <- coef(summary(mod))[1, 1]
lC1 <- coef(summary(mod))[2, 1]
lC2 <- coef(summary(mod))[3, 1]
CC1 <- coef(summary(mod))[4, 1]
CC2 <- coef(summary(mod))[5, 1]
MYC <- coef(summary(mod))[6, 1]
SYC <- coef(summary(mod))[7, 1]
EC <- coef(summary(mod))[8, 1]
#predict the rating by using the model "mod"
ratingR <- C + cLimit*lC1 + (cLimit^2)*lC2 + cN*CC1 + (cN^2)*CC2 + (1)*MYC + (0)*SYC
ratingR
```

```
[1] 844.2286
```

The Rating for an individual that has a credit card limit of $12,000, has 2 credit cards, is married, and is not a student, and has an eighth grade education is 844.2286339.

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

```r
#Stephanie helped me with the code of this one
CreditMY <- subset(Credit, Credit$Married == "Yes")
tval <- t.test(CreditMY$Rating, conf.level = .90, alternative = "two.sided")
lowerMY <- tval$conf.int[1]
upperMY <- tval$conf.int[2]
```

We are 90% confident that the true parameter of Rating for a married person would fall between 342.801888 and 376.1123977.

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 
```

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
TVN <- 25
RadioN <- 1
RadioNC <- coef(modSales)[3]  #coefficient of Radio variable for the model for sales vs TV interacted w
TVvsRC <- coef(modSales)[4]  #coefficient of TV interacted with Radio variable for the model for sales 

Effect <- RadioN * RadioNC + TVN * RadioN * TVvsRC
Effect <- round(Effect, digits = 4)  #round to 4 digits
```

According to the model for sales vs TV interacted with radio (`modSales`), the effect of an additional 1 unit of radio advertising if TV = 25 is 0.056.

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

```
TVN <- 300
RadioN <- 1
RadioNC <- coef(modSales)[3]  #coefficient of Radio variable for the model for sales vs TV interacted w
TVvsRC <- coef(modSales)[4]  #coefficient of TV interacted with Radio variable for the model for sales 

Effect <- RadioN * RadioNC + TVN * RadioN * TVvsRC
Effect <- round(Effect, digits = 4)  #round to 4 digits
```

According to the model for sales vs TV interacted with radio (`modSales`), the effect of an additional 1 unit of radio advertising if TV = 300 is 0.3548.

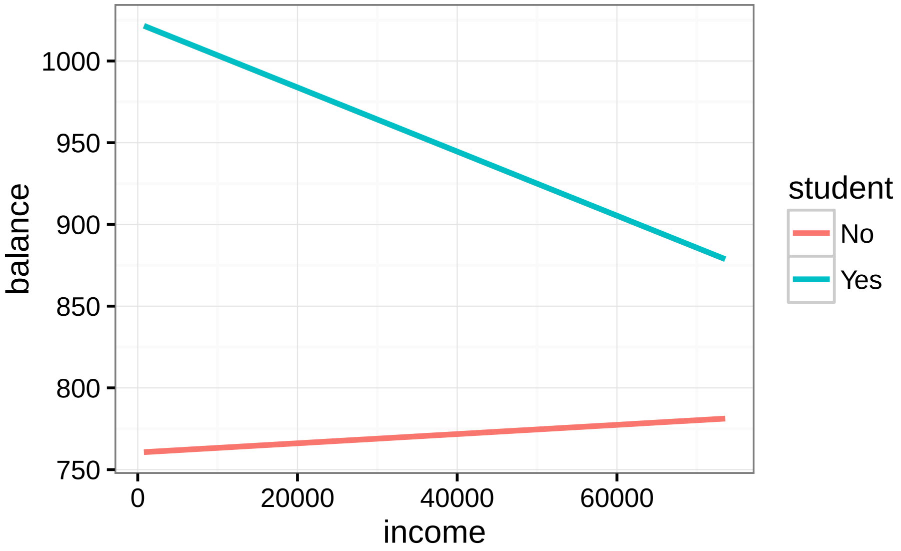
8. What is the difference between `lm(y ~ x*z)` and `lm(y ~ I(x*z))`, when `x` and `z` are both numeric variables?

```
#Kevin helps me understand the solutions for this problem
```

- **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()
```



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.~~