PSTAT 126 HW 2

Tamjid Islam

4/15/2020

1. This problem uses the wblake data set in the alr4 package. This data set includes samples of small mouth bass collected in West Bearskin Lake, Minnesota, in 1991. Interest is in predicting length with age. Finish this problem without using lm().

```
# install.packages('alr4')
library(alr4)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

(a) Compute the regression of length on age, and report the estimates, their standard errors, the value of the coefficient of determination, and the estimate of variance. Write a sentence or two that summarizes the results of theses computations.

```
library(alr4)
data(wblake)
avg1 <- mean(wblake$Age)
avg1
```

```
## [1] 4.202733
```

```
avg2 <- mean(wblake$Length)
avg2
```

```
## [1] 192.9704
```

```
Sxy <- sum((wblake$Age-avg1)*(wblake$Length-avg2))
Sxy
```

```
## [1] 52610.64
```

```
Sxx <-sum((wblake$Age-avg1)^2)
Sxx
```

```
## [1] 1734.957
```

```r
slope <- Sxy / Sxx
slope
```

```
## [1] 30.32389
```

```r
intercept <- avg2 - (avg1*slope)
intercept
```

```
## [1] 65.52716
```

```r
y_hat <- intercept + slope*wblake$Age
n <- length(wblake$Length)
SSE <- sum((wblake$Length-y_hat)^2)
var <- SSE/(n-2)
var
```

```
## [1] 820.5847
```

```r
se_age <- sqrt(var) / (sqrt(Sxx))
se_age
```

```
## [1] 0.6877291
```

```r
se_length <- sqrt(var) * (sqrt(1/n + (avg1)^2/Sxx))
se_length
```

```
## [1] 3.197388
```

```r
SSTO <- sum((wblake$Length-avg2)^2)
R_2 <- 1- (SSE / SSTO)
R_2
```

```
## [1] 0.816477
```

The regression of length on age is $\hat{Y} = 65.52716 + 30.32389x$. The estimates of this regression is the slope and intercept. The estimate on the slope is ($b_1 = 30.32389$) and the estimate on the intercept is ($b_0 = 65.52716$). The standard error for age is 0.6877291 and standard error for length is 3.197388. The coefficient of determination is 0.816477 implying the proportion of the variation to length. The estimate of the varience is 820.5847.

(b) Obtain a 99% confidence interval for β_1 from the data. Interpret this interval in the context of the data.

```r
t_pct <- qt(p = .995, df = n - 2)
ci_b1_99 <- slope + c(-1, 1)*t_pct*se_age
ci_b1_99
```

```
## [1] 28.54465 32.10313
```

We are 99% confident that the small mouth bass collected grows or has a length between 28.54465 and 32.10313 for every unit increase in age.

(c) Obtain a prediction and a 99% prediction interval for a small mouth bass at age 1. Interpret this interval in the context of the data.

```r
t_pct <- qt(p = .995, df = n - 2)
t_pct
```

```
## [1] 2.587126
```

```r
x = 1
n <- length(wblake$Length)
y_hat1 <- intercept + slope * x
y_hat1
```

```
## [1] 95.85105
```

```r
new_se <- sqrt(var) * (sqrt(1 + 1/n + ((x-avg1)^2)/Sxx))
new_se
```

```
## [1] 28.76292
```

```r
ci_pred_99 <- y_hat1 + c(-1, 1)*t_pct*new_se
ci_pred_99
```

```
## [1]  21.43775 170.26436
```

We are 99% confident that the length of the small mouth bass at age 1 will be between 21.43775 and 170.26436.

2. This problem uses the data set Heights data set in the alr4 package. Interest is in predicting dheight by mheight.

(a) Compute the regression of dheight on mheight, and report the estimates, their standard errors, the value of the coefficient of determination, and the estimate of the variance.

```r
library(alr4)
data(Heights)
fit <- lm(Heights$dheight ~ Heights$mheight)
fit
```

```
##
## Call:
## lm(formula = Heights$dheight ~ Heights$mheight)
##
## Coefficients:
##     (Intercept)  Heights$mheight
##         29.9174           0.5417
```

```r
summary(fit)
```

```
##
## Call:
## lm(formula = Heights$dheight ~ Heights$mheight)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.397 -1.529  0.036  1.492  9.053
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     29.91744    1.62247   18.44   <2e-16 ***
## Heights$mheight  0.54175    0.02596   20.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.266 on 1373 degrees of freedom
## Multiple R-squared:  0.2408, Adjusted R-squared:  0.2402
## F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16
```

```r
summary(fit)$sigma^2
```

```
## [1] 5.136167
```

The regression of length on age is $\hat{Y} = 29.9174 + 0.5417x$. The estimates of this regression is the slope and intercept. The estimate on the slope is ($b_1 = 0.5417$) and the estimate on the intercept is ($b_0 = 29.9174$). The standard error for mheight is 0.02596 and standard error for dheight is 1.62247. The coefficient of determination is 0.2408 implying the proportion of the variation to dheight. The estimate of the varience is 5.136167.

(b) For this problem, give an interpretation for β_0 and β_1.

β_0 and β_1 are involving the population intercept and population slope in a simple regression model respectively. In this case for β_0, the expected mean value of dheight is 29.9174 when mheight is zero and for β_1 dheight increases by 0.5417 for every one unit of mheight.

(c) Obtain a prediction and a 99% prediction interval for a daughter whose mother is 64 inches tall.

```r
fit <- lm(dheight ~ mheight, data = Heights)
fit
```

```
##
## Call:
## lm(formula = dheight ~ mheight, data = Heights)
##
## Coefficients:
## (Intercept)      mheight
##     29.9174       0.5417
```

```r
summary(fit)
```

```
##
## Call:
## lm(formula = dheight ~ mheight, data = Heights)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.397 -1.529  0.036  1.492  9.053
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.91744    1.62247   18.44   <2e-16 ***
## mheight      0.54175    0.02596   20.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.266 on 1373 degrees of freedom
## Multiple R-squared:  0.2408, Adjusted R-squared:  0.2402
## F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16
```

```r
new <- data.frame(mheight = 64)
ans <- predict(fit,new, se.fit = TRUE, interval = 'prediction', level = 0.99, type = 'response')
ans
```

```
## $fit
##        fit      lwr      upr
## 1 64.58925 58.74045 70.43805
```

```
##
## $se.fit
## [1] 0.07313503
##
## $df
## [1] 1373
##
## $residual.scale
## [1] 2.266311
```

We are 99% confident that the daughter's height will be between 58.74045 and 70.43805 inches when the mother's height is 64 inches.

3. The simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, i = 1, . . . , n can also be written as

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Using matrix notations, the model is

$$Y = X\beta + \epsilon. \tag{1}$$

In this problem, we will show that that the least squares estimate is given by:

$$b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} (X'X)^{-1} X'Y$$

(a) Using straightforward matrix multiplication, show that

$$(X'X) = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{pmatrix} = n \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & S_{xx}/n + \bar{x}^2 \end{pmatrix}$$

$$(X'Y) = \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^n x_i Y_i \end{pmatrix} = \begin{pmatrix} n\bar{Y} \\ S_{xy} + n\bar{x}\bar{Y} \end{pmatrix}$$

Solving for (X'X)

$$(X'X) = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{pmatrix} \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & n * \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2 + n\bar{x}^2}{n} \end{pmatrix}$$

$$= n \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & \frac{\sum_{i=1}^n (x_i - \bar{x}) + n\bar{x}^2}{n} \end{pmatrix} = n \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & S_{xx}/n + \bar{x}^2 \end{pmatrix}$$

Solving for (X'Y)

$$(X'Y) = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{pmatrix} \begin{pmatrix} Y_1 & Y_2 \cdots Y_n \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{pmatrix} = \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^{n} x_i Y_i \end{pmatrix} = \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^{n} x_i Y_i - n\bar{x}\bar{Y} + n\bar{x}\bar{Y} \end{pmatrix} = \begin{pmatrix} n\bar{Y} \\ \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) + n\bar{x}\bar{Y} \end{pmatrix} = \begin{pmatrix} n\bar{Y} \\ S_{xy} + n\bar{x}\bar{Y} \end{pmatrix}$$

(b) Using the identity

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

for a 2 × 2 matrix, show that

$$(X'X)^{-1} = \frac{1}{S_{xx}} \begin{pmatrix} S_{xx}/n + \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

Applying the identity theorem

$$(X'X)^{-1} = n \begin{pmatrix} 1 & \bar{x} \\ \bar{x} & S_{xx}/n + \bar{x}^2 \end{pmatrix}^{-1} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & S_{xx} + n\bar{x}^2 \end{pmatrix}^{-1} = \frac{1}{S_{xx}} \begin{pmatrix} S_{xx}/n + \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

(c) Combine your answers from (a) and (b) to show that

$$b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = (X'X)^{-1} X'Y \tag{2}$$

where $b_1 = S_{xy}/S_{xx}$ and $b_0 = \bar{Y} - b_1\bar{x}$ are the least squares estimates from simple linear regression

$$(X'X)^{-1} X'Y = \frac{1}{S_{xx}} \begin{pmatrix} S_{xx}/n + \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \begin{pmatrix} n\bar{Y} \\ S_{xy} + n\bar{x}\bar{Y} \end{pmatrix}$$

$$= \frac{1}{S_{xx}} \begin{pmatrix} S_{xx}\bar{Y} + n\bar{x}^2\bar{Y} - \bar{x}S_{xy} - n\bar{x}^2\bar{Y} \\ -n\bar{x}\bar{Y} + S_{xy} + n\bar{x}\bar{Y} \end{pmatrix} = \begin{pmatrix} \bar{Y} - \bar{x}S_{xy}/S_{xx} \\ S_{xy}/S_{xx} \end{pmatrix} = \begin{pmatrix} \bar{Y} - \bar{x}S_{xy}/S_{xx} \\ b_1 \end{pmatrix} = \begin{pmatrix} \bar{Y} - \bar{x}b_1 \\ b_1 \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = b$$

(d) Simulate a data set with n = 100 observation units such that $Y_i = 1 + 2x_i + \epsilon_i$, i = 1, . . . , n. ϵ_i follows the standard normal distribution, i.e., a normal distribution with zero mean and unit variance. Use the result in (c) to compute b_0 and b_1. Show that they are the same as the estimates by lm(). Start with generating x as

n = 100
x = seq(0, 1, length = n)

(Hint: check the help page of rnorm() about how to simulate normally distributed random variables. Use solve() to get an inverse matrix and use t() to get a transpose matrix.)

```
n = 100
x <- seq(0, 1, length = n)
y <- 1 + 2*x + rnorm(n)
X <- cbind(1,x)
Y <- matrix(y)
solve <- solve(t(X)%*%X)%*%t(X)%*%Y
solve
```

```
##       [,1]
##   1.008850
## x 2.130289
```

```
fit <- lm(y ~ x)
fit
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##       1.009        2.130
```

The estimates are the same as the lm() function. They are always subject to change with this specific code due to the fact that there are 100 random values that will always be generated.

4. This problem uses the UBSprices data set in the alr4 package. The international bank UBS regularly produces a report (UBS, 2009) on prices and earnings in major cities throughout the world. Three of the measures they include are prices of basic commodities, namely 1 kg of rice, a 1 kg loaf of bread, and the price of a Big Mac hamburger at McDonalds. An interesting feature of the prices they report is that prices are measured in the minutes of labor required for a "typical" worker in that location to earn enough money to purchase the commodity. Using minutes of labor corrects at least in part for currency fluctuations, prevailing wage rates, and local prices. The data file includes measurements for rice, bread, and Big Mac prices from the 2003 and the 2009 reports. The year 2003 was before the major recession hit much of the world around 2006, and the year 2009 may reflect changes in prices due to the recession.

The first graph below is the plot of Y = rice2009 versus x = rice2003, the price of rice in 2009 and 2003, respectively, with the cities corresponding to a few of the points marked.

(a) The line with equation Y = x is shown on this plot as the solid line. What is the key difference between points above this line and points below the line?

The key difference is that countries/points above the line had an increase in rice price from 2003 compared to 2009. The countries/points under the line had a decrease in rice price from 2003 compared to 2009.

(b) Which city had the largest increase in rice price? Which had the largest decrease in rice price?

Vilnius had the largest increase in rice price. Mumbai had the largest decrease in rice price.

(c) Give at least one reason why fitting simple linear regression to the figure in this problem is not likely to be appropriate.

The reason why fitting a simple linear regression would not likely be appropriate is beacuse there isn't much scatter even tho there seems to be a upward trend.There is much more of a cluster rather than a scatter. Therefore, there is not a good linear relationship.

(d) The second graph represents Y and x using log scales. Explain why this graph and the previous graph suggests that using log scales is preferable if fitting simple linear regression is desired. The linear model is shown by the dashed line.

This graph is preferable with fitting a simple linear regression because it is more normally distributed and shows an upward trend linearly in relation to 2003 rice price and 2009 rice price.