STA 674 - Assignment 1

Matt Massey

1/24/2022

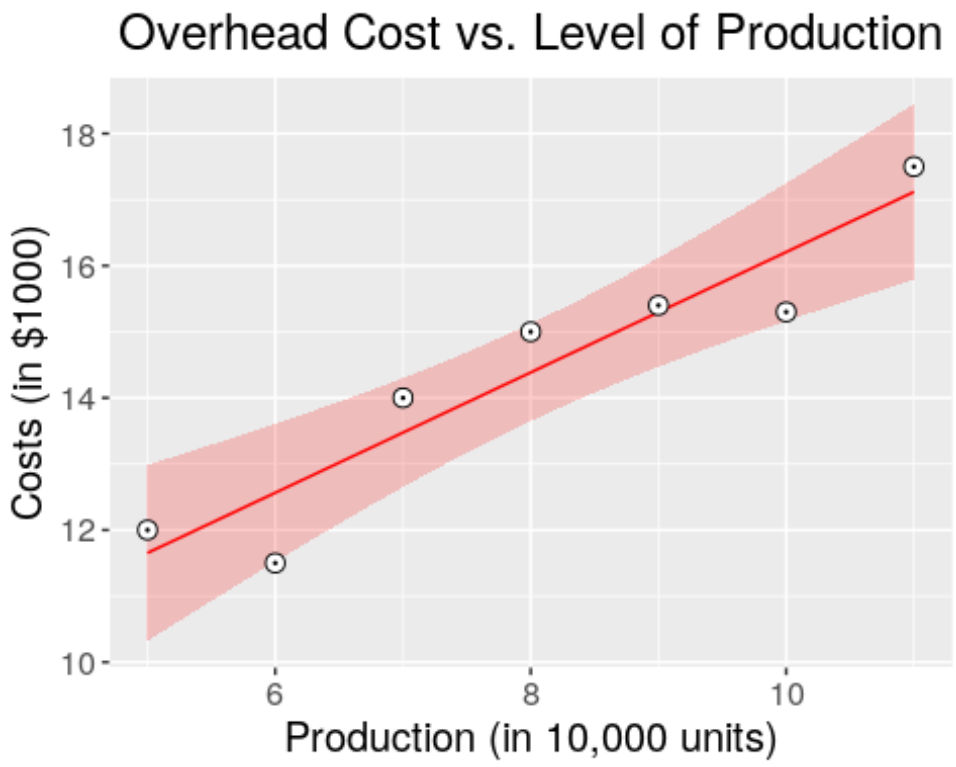
Question 1

1A.

```
# import ggplot2 library
library(ggplot2)

# add data and create dataframe for plotting
y = c(12, 11.5, 14, 15, 15.4, 15.3, 17.5)
x = c(5:11)
plant = data.frame(x, y)

# create scatterplot and linear regression model from dataframe
ggplot(plant, aes(x, y)) + geom_smooth(method="lm", color='red', size=0.5,
fill='red', alpha=0.2) +  geom_point(aes(x, y), shape=21, size=3, fill='white') +
geom_point(aes(x,y), size=0.2) + labs(x="Production (in 10,000 units)", y = "Costs
(in $1000)") + theme(text=element_text(size=15), plot.title=element_text(hjust=0.5))
+ ggtitle("Overhead Cost vs. Level of Production")
## `geom_smooth()` using formula 'y ~ x'
```




```
plant.lm = lm(y~x, data = plant)

# calculate model statistics
summary(plant.lm)
```

```
##
## Call:
## lm(formula = y ~ x, data = plant)
##
## Residuals:
##       1       2       3       4       5       6       7
##  0.3464 -1.0643  0.5250  0.6143  0.1036 -0.9071  0.3821
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.1000     1.1838   5.997  0.00185 **
## x             0.9107     0.1436   6.344  0.00144 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7597 on 5 degrees of freedom
## Multiple R-squared:  0.8895, Adjusted R-squared:  0.8674
## F-statistic: 40.24 on 1 and 5 DF,  p-value: 0.001437
```

1B. The least-squares regression model is shown by the red line in the figure above and defined by the equation:

Costs = 0.9107(Production) + 7.1000

where Costs is in units of $1000 and Production is in units of 10,000. The model suggests a positive correlation between Costs versus Production, where Costs increase $910.7 for every 10,000 units produced.

1C. The standard errors of b_0 (intercept) and b_1 (slope) are 1.1838 and 0.1436, respectively.

Question 2

```
# Note: file hw1_div3.csv was uploaded to RStudio, then accessed

#Calculate least-squares regression model for div3
div3 = read.csv("hw1_div3.csv", header = T)
div3.lm = lm(DIVYIELD~EPS, data = div3)
result.div3 = summary(div3.lm)
result.div3
##
## Call:
## lm(formula = DIVYIELD ~ EPS, data = div3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4894 -1.6827 -0.0619  1.2196  4.1746
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0336     0.5405   3.762  0.00054 ***
## EPS           0.3740     0.2395   1.562  0.12624
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.85 on 40 degrees of freedom
## Multiple R-squared:  0.05747,    Adjusted R-squared:  0.0339
## F-statistic: 2.439 on 1 and 40 DF,  p-value: 0.1262

# test statistic for sample with 42 points
test_statistic = qt(p=.025, df=40, lower.tail=FALSE)
cat("Test statistic at alpha=0.05: ", test_statistic)
## Test statistic at alpha=0.05:  2.021075
```

2A. The least-squares regression model is defined by the equation:

Dividend yield = 0.3740(EPS) + 2.0336

2B. The plot shown on the handout suggests a positive linear correlation. To test this hypothesis, we assume a null hypothesis (H_0) of 0 and an alternative hypothesis (H_A) not equal to 0. A significance level (α) of 0.05 is used to construct a 95% confidence interval for this two-tailed hypothesis test. If the t value of the slope of the regression falls within the range of test statistic values, then we will fail to reject the null hypothesis; otherwise we reject the null hypothesis. The test statistic (n=42, df=2) is +/-2.021075. Therefore, the t value of the regression (1.562) is within the range of critical values and we fail to reject the null hypothesis that there is no linear correlation between Dividend Yield and EPS. This is echoed by the P value of 0.12624 being greater than our significance level of 0.05, and a confidence interval ranging from negative to positive values and includes the null hypothesis value of 0.

2C. The 95% confidence interval of B1 (slope) is -0.11000 to 0.85791, and for B0 (intercept) it is 0.94121 to 3.12607

Question 3

```
# Note: file hw1_SALESAD3.csv was uploaded to RStudio, then accessed

#Calculate least-squares regression model for SALESAD3
salesad3 = read.csv("hw1_SALESAD3.csv", header = T)
salesad3.lm = lm(SALES~ADV, data = salesad3)
result.salesad3 = summary(salesad3.lm)
result.salesad3
##
## Call:
## lm(formula = SALES ~ ADV, data = salesad3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1543.52  -150.34    22.81   192.88  1542.63
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -57.281    509.750  -0.112    0.912
## ADV           17.570      1.642  10.702 3.11e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 594.8 on 18 degrees of freedom
## Multiple R-squared:  0.8642, Adjusted R-squared:  0.8566
## F-statistic: 114.5 on 1 and 18 DF,  p-value: 3.111e-09

# test statistic for sample with 20 points
test_statistic = qt(p=.025, df=18, lower.tail=FALSE)
cat("Test statistic at alpha=0.05: ", test_statistic)
## Test statistic at alpha=0.05:  2.100922
```

3A. The plot shown on the handout suggests a positive correlation. To test this hypothesis, we assume a null hypothesis (H_0) of 0 and an alternative hypothesis (H_A) not equal to 0. A significance level (α) of 0.05 is used to construct a 95% confidence interval for this two-tail test. If the t value of the slope of the regression falls within the range of test statistic values, then we will fail to reject the null hypothesis; otherwise we reject the null hypothesis. The test statistic (n=20, df=2) is +/- 2.100922. Therefore, the t value of the regression (10.702) is well beyond the bounds of the critical values and we reject the null hypothesis that there is no correlation between Sales and Advertising This is echoed by the small P value of 3.111e-09 below our significance level of 0.05, and a confidence interval with only positive values.

3B. The regression model is given by the equation:

Sales = 17.570(Advertising) - 57.281

where Sales and Advertising are both expressed in units of $100. The regression model suggests a positive correlation between money spent on Advertising and the Sales of the firm's major product, where Sales increases by $1757 for every $100 increase in Advertising.

3C. If Advertising were increased by $1000 (10 units), the predicted Sales would increase by $17,570 (175.7 units).

Question 4
In Question 1B it was determined that the regression model equation is defined by:

Costs = 0.9107(Production) + 7.1000

where Costs is in thousands of dollars and Production is in units of 10,000.

4A.

```
# calculate predicted mean at a given value of x
n = 7 ###(number of observation)
xm = 8 ###(8 in 10000 units)
xbar = mean(x) ###(defined in Question 1)
# determine point estimate
ym = 7.1 + 0.9107 * 8 ###regression equation from Question 1 with b1=0.9107 and
b0=7.1
cat("Point estimate: ", ym, "\n")
## Point estimate:  14.3856

# calculate confidence interval of the mean at given value of x
# mean square error taken from linear model output
MSE.plant.lm = anova(plant.lm)[2,3]
#creating a function to compute confidence interval for a value of y, on average
est.interval = function(x, yhat, xm, alpha = 0.05, MSE)
{
  n = length(x)
  xbar = mean(x)
  interval = yhat + c(-1, 1) * qt(alpha/2, n-2, lower.tail = F) * sqrt( MSE*  (1/n +
(xm-xbar)^2/((n-1)*sd(x)^2 )) )
  return(interval)
}
# run new CI function
cat("Confidence interval: ", est.interval(x, ym, xm,alpha=0.05, MSE.plant.lm))
## Confidence interval:  13.64753 15.12367
```

If Production reaches 80,000 (8 units on regression), the mean overhead cost is estimated to be $14,385.6. The 95% confidence interval for overhead is between $13,647.53 to $15,123.67, given that production is at 80,000.

4B.

```
# calculate predicted observation at a given value of x...same value as predicted
average in 4A
cat("Point estimate: ", ym, "\n")
## Point estimate:  14.3856
# calculate prediction interval for predicted observation at a given value of x
# create function to compute prediction interval
pre.interval = function(x, yhat, xm, alpha = 0.05, MSE)
{
  n = length(x)
  xbar = mean(x)
  interval = yhat + c(-1, 1) * qt(alpha/2, n-2, lower.tail = F) * sqrt(MSE *(1 + 1/n
+ (xm-xbar)^2/((n-1)*sd(x)^2 )))
  return(interval)
}
```

```r
#run new PI function
cat("Prediction interval: ", pre.interval(x, ym, xm, alpha = 0.05, MSE.plant.lm))
## Prediction interval:  12.29802 16.47318
```

Overhead costs for a single production run of 80,000 would be predicted to be $14,385.6. The 95% prediction interval for a single run is between $12,298.02 to $16,473.18.

4C.
The predicted mean and single observation are calculated with the same value of x (Production) and are both best estimated by the regression model with the equation found in Question 1. Therefore, we see the same answer of 14.3856 for both point estimates. There is, however, a difference in the confidence interval (CI) of the mean and prediction interval (PI) of the single observation. Specifically, the CI is smaller than the PI, which will always be true due to the equations used to calculate their standard errors. The CI represents the error associated with estimating a parameter (the mean), which is represented by the regression line. The PI represents the variation we expect when predicting a single new observation at a given value.

Question 5
In question 2A it was determined that the regression model equation is defined by:

Dividend = 0.3740(EPS) + 2.0336

5A.

```
# calculate predicted observation at $2 earnings per share and 95% prediction
interval using regression from Question 2
Xm1 = 2 ### EPS=2
EPS = div3$EPS

# point estimate at EPS=2
Ym1 = 2.0336 + 0.374 * Xm1
cat("Point estimate: ", Ym1, "\n")
## Point estimate:  2.7816
# 95% PI using function from Question 4B
# mean square error
MSE.div3.lm = anova(div3.lm)[2, 3]
# PI
div3.2.pre = pre.interval(EPS, Ym1, Xm1, alpha = 0.05, MSE.div3.lm )
cat("Prediction interval: ", div3.2.pre)
## Prediction interval:  -1.00135 6.56455
```

Using the regression equation above, a firm with $2 earnings per share is predicted to have a dividend yield of 2.7816. The 95% prediction interval would range from $-1.00135 to $6.56455.

5B.

```
# calculate predicted observation at $5 earnings per share and 95% prediction
interval using regression from Question 2
Xm2 = 5 ### EPS=5

# point estimate at EPS=5
Ym2 = 2.0336 + 0.374 * Xm2
cat("Point estimate: ", Ym2, "\n")
## Point estimate:  3.9036
# 95% PI
div3.5.pre = pre.interval(EPS, Ym2, Xm2, alpha = 0.05, MSE.div3.lm)
cat("Prediction interval: ", div3.5.pre)
## Prediction interval:  -0.162776 7.969976
```

A firm with $5 earnings per share is predicted to have a dividend yield of 3.9036. The 95% prediction interval would range from $-0.162776 to $7.969976.

5C.
No values in questions 5A and 5B remained the same because each question used a different value of x (EPS) for their predictions, and with a positive slope of the regression model we cannot have the same predicted value of y. As such, predicted values and their prediction intervals are all different. The prediction intervals do overlap, which suggests that it would be reasonable to expect the same y values (Dividend) for either $2 or $5 earnings per share. This overlap is expected based on the summary statistics calculated in question 2, which conclude that there is no significant linear relationship between the dividend yields and earnings.