Stat 6021: Homework Set 1

Diana McSpadden

9/11/2020

Name: H. Diana McSpadden

UID: hdm5s

Assignment: Homework Set 2

Studied with Chelsea Alvarado, David Fuentes, Caprill Wright, and Abby Bernhardt

Question 1

(R required) For this question, you will use the dataset "Copier.txt" for this question. This is the same data set that you used in the last homework. The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, Serviced is the number of copiers serviced and Minutes is the total number of minutes spent by the service person.

It is hypothesized that the total time spent by the service person can be predicted using the number of copiers serviced. Fit an appropriate linear regression and answer the following questions:

Knowns n = 45 Serviced == number of copy machines serviced on a call Minutes == number minutes spent by service person on the call

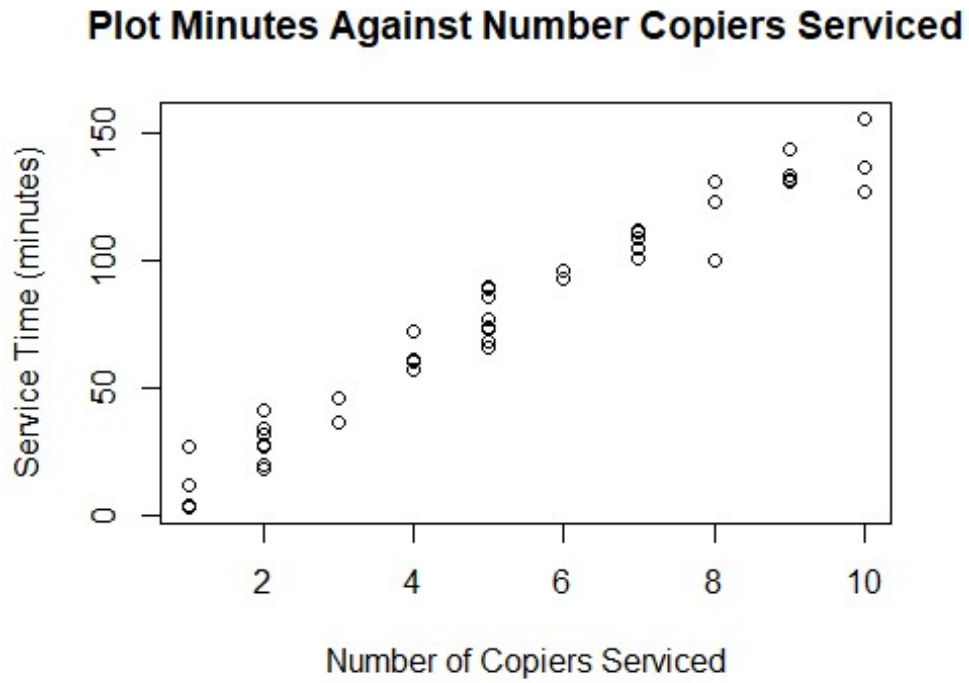
1a. Produce an appropriate scatterplot and comment on the relationship between the total time spent by the service person and the number of copiers serviced.

```
#getwd()
copierData <- read.table('copier.txt', sep='\t', header = TRUE)
head(copierData,5)
```

```
##   Minutes Serviced
## 1      20        2
## 2      60        4
## 3      46        3
## 4      41        2
## 5      12        1
```

```
attach(copierData)

#use plot(...) to create a scatterplot
plot(x=Serviced, y=Minutes, main='Plot Minutes Against Number Copiers Service
d', xlab = 'Number of Copiers Serviced', ylab = 'Service Time (minutes)')
```



1b. What is the correlation between the total time spent by the service person and the number of copiers serviced? Interpret this correlation contextually.

```
corvalue <- cor(x=Serviced,y=Minutes)
print(paste("Correlation between Service Time and Number of Copiers: ", corva
lue))
```

```
## [1] "Correlation between Service Time and Number of Copiers:  0.9785169817
01943"
```

Because the correlation value is so close to 1, i.e. 0.9785, and the scatter plot supports a positive linear relationship between service time and copiers serviced, there is evidence of

a linear relationship between the predictor variable, number of copiers serviced, and the response variable, service time.

1c. Can the correlation found in part 1b be interpreted reliably? Briey explain.

Yes, it can be reliably interpreted after also analyzing the scatter plot to confirm a linear relationship. Both confirm a linear relationship. The correlation is not a reliable predictor unless both the scatter plot and correlation value are reviewed.

1d. Obtain the 95% confidence interval for the slope, B1.

```
# create a linear model
copierModel <- lm(Minutes~Serviced)
copierModel

##
## Call:
## lm(formula = Minutes ~ Serviced)
##
## Coefficients:
## (Intercept)     Serviced
##     -0.5802      15.0352

#summary(copierModel)

# get the 95% CI for the Serviced coef
CIB1 = confint(copierModel, "Serviced", level = 0.95)

print(paste("95% CI of slope: ", CIB1[1], " - ", CIB1[2]))

## [1] "95% CI of slope:  14.0610098266507  -  16.0094862569002"
```

1e. Suppose a service person is sent to service 5 copiers. Obtain an appropriate 95% interval that predicts the total service time spent by the service person.

```
## 95% CI for response when x=5
newdata<-data.frame(Serviced=5)
#print(newdata) # print newdata to understand what we get back

prediction = predict.lm(copierModel,newdata,level=0.95, interval="prediction"
)
#prediction

print(paste("The 95% CI for service minutes for a service call for 5 copiers
is: ", prediction[2], " - ", prediction[3], " minutes"))

## [1] "The 95% CI for service minutes for a service call for 5 copiers is:
56.42132504534  -  92.7708420564876  minutes"
```

1f. What is the value of the residual for the first observation? Interpret this value contextually.

```
# get all the residuals from the model
copierResiduals <- copierModel$residual
#copierResiduals

# get the first residual
print(paste("The value of the first residual is: ", copierResiduals[1]))
```

```
## [1] "The value of the first residual is:  -9.49033942558754"
```

1f Interpretation The residual is the difference between the observed first data point, and the model's prediction for the first data point. So, the difference between y0 for the first x value, and y-hat for the first x value:

Residual1 = (y1 - Y1-hat)

So, the first value in our data set is **9.49 fewer minutes** than E(Y|x) for a service call for 2 copiers (the first x data point).

1g. What is the average value of the all the residuals? Is this value surprising (or not)? Briefly explain.

```
meanCopierResidual = mean(copierModel$residual)

print(paste("The mean of the residuals is: ", meanCopierResidual, " minutes."
))
```

```
## [1] "The mean of the residuals is:  -2.61220443424174e-16  minutes."
```

1g Explanation

No, this is not surprising. The mean of the residuals is 0, and -0.00000000000000026122 is basically 0 with some expected rounding error. Last week we proved mathematically that the sum of the residuals is 0, so the mean (sum(residuals)/n) == (0/n) == 0. This is the result of the definition of BLUE.

Question 2 (No R required)

A substance used in biological and medical research is shipped by airfreight to users in cartons of 1000 ampules. The data consist of 10 shipments. The variables are number of times the carton was transferred from one aircraft to another during the shipment route (transfer), and the number of ampules found to be broken upon arrival (broken).

We want to fit a simple linear regression. A simple linear regression model is fitted using R. The corresponding output from R is shown next, with some values missing.

Call: lm(formula = broken ~ transfer)

Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 10.2000 0.6633 _____ _____ ***transfer 4.0000 0.4690 _____ _____***

Residual standard error: 1.483 on 8 degrees of freedom

Analysis of Variance Table Response: broken Df Sum Sq Mean Sq F value Pr(>F) transfer 1 160.0 160.0 _____ _____ *** Residuals 8 17.6 2.2

Additional Knowns

Predictor Variable, x: Number of transfers

Response variable, y: Number of broken ampules

n = 10

x-bar = 1

Sxx, sum(((x-i) - (x-bar)^2) == 10

2a. Carry out a hypothesis test to assess if there is a linear relationship between the variables of interest.

H0: B1 = 0

Ha: B1 != 0

Our estimated equation is: E(y|x) = 10.2 + 4(x)

The equation has B1-hat == 4.

F0 = 160 / 2.2 = 72.7272, this is a pretty large number, and implies a t-value = sqrt(72.7272) = 8.528

Using the t-value comparison, 8.528 is greater than 2.306 so we rejecting H0, and determine there is a linear relationship.

This is evidence for a linear relationship between the variables.

2b. Calculate a 95% confidence interval that estimates the unknown value of the population slope.

B1-hat - t-alpha/2,n-2(se(B1-hat)) <= B1-hat <= B1-hat + t-alpha/2,n-2(se(B1-hat))

4 - t-(.975,8)(se(B1-hat)) <= B1-hat <= 4 + t(.975,8)(se(B1-hat))

4 - (2.306)(se(B1-hat)) <= B1-hat <= 4 + (2.306)(se(B1-hat)) # t-value from textbook for 0.025 and 8 df

se(B1-hat) = 0.4690

se(B1-hat) = 4 / 4.6904

So, 4 - (2.306)(0.4690) <= B1-hat <= 4 + (2.306)(0.4690)

4 - 1.0815 <= B1-hat <= 4 + 1.0815

95% CI for population slope is 2.918 - 5.0814

2c. A consultant believes the mean number of broken ampules when no transfers are made is different from 9.

Conduct an appropriate hypothesis test (state the hypotheses statements, calculate the test statistic, and write the corresponding conclusion in context, in response to his belief).

H0: B0 = 9

Ha: B0 != 9

t0 = (B0-hat - 9) / se(B0-hat)

se(B0-hat) == sqrt((MS-res)(1/n + ((x-bar)^2/Sxx))

t0 = (10.2 - 9) / sqrt((2.2) * (.1 + .1)) # MS-res of 2.2 comes from ANOVA table

t0 = (10.2 - 9) / sqrt(.44)

t0 = 1.2 / .6633

t0 = 1.8091 < 2.306004

Answer:

This test statistic, 1.8091, is smaller than t(0.975,8), a 95% confidence level which is 2.306. We do not reject the null hypothesis. There is not evidence to suggest the intercept does not equal 9.

2d. Calculate a 95% confidence interval for the mean number of broken ampules and a 95% prediction interval for the number of broken ampules when the number of transfers is 2.

MEAN CI: CI for mean number of broken ampules

E(y|2) +- t(.0975,8)(sqrt((MS-res)(1/n + (2 - 1)/Sxx))

(10.2 + (4*2)) +- (2.306)sqrt((2.2)(.1 + ((2 - 1)^2/10)) # t value for 0.025 with df = 8 is from textbook, and MS-res is 2.2

18.2 +- (2.306) * sqrt((2.2)(.1 + .1))

18.2 +- (2.306) * sqrt(2.2 * .2)

18.2 +- (2.306) * (0.6633)

18.2 +- 1.53

95% CI for mean ampules broken when there are 2 transfers is 16.67 - 19.73

PREDICTED VALUE CI: CI for predicted number of broken ampules when number of transfers is 2

10.2 + 4(2) +- (2.306004)sqrt((2.2)(1 + .1 + ((2 - 1)^2/10))

18.2 +- (2.306004)(sqrt((2.2 * 1.2)))

18.2 +- (2.306004)(1.6248)

18.2 +- 3.7468

95% CI for the predicted number of ampules broken when there are 2 transfers is 14.4532 - 21.9468

2e. What happens to the intervals from the previous part when the number of transfers is 1? (Describe what happens without calculating)

The range of the two confidence intervals will be smaller, and I can describe why both logically and mathematically:

Logically: 1 transfer is the mean number of transfers, so our predictions will be better, thus a smaller interval, because our predictive power is better at the mean of x.

Mathematically: The correction (+/-) to y0 will be smaller because the numerator to part of the correction is (x0 - (x-bar)), which will be 1 - 1 == 0.

2f. What is the value of the F statistic for the ANOVA table?

F0 = 160 / 2.2 = **72.7272**

2g. Calculate the value of R2, and interpret this value in context.

The R2, Coef of determination, formula is: SS-R / SS-T, i.e. the regression error divided by the total error and can be used as a measure of how well variations in x explain variations in y.

R2 = 160 / (160 + 17.6)

R2 = 160 / 177.6 == **.9009**

This value is close to 1, and implies that our model does demonstrate a linear relationship between number of transfers and ampules broken. R2 can be unreliable without seeing the data and a scatter plot of the data.

Question 3 (No R required)

Suppose that the population slope for a straight-line relationship between y and x is 0.

3a. Describe how the straight line would look in a plot of y versus x.

Answer

The line would be a straight, perfectly horizontal line at some unknown y-intercept.

3b. Explain why a slope that is equal to 0 would indicate that y and x are not linearly related, and why a slope that is not equal to 0 would indicate that y and x are linearly related.

Answer

A slope of zero describes a y value that does not change, no matter the value of x. x could be anything, and y remains stable. Y is independent of x. An example is the speed of light, no matter what the hair color of a person is, the speed of light remains the same. That relationship has a slope of zero.

If the slope is not equal to 0, then the value of y does change, to some degree, as the value of x changes. As the change in x increases/decreases from the mean of x, the change in y will also increase/decrease from y's mean. That change expresses some type of linear relationship with a non-0 slope.