Math 229 **First Exam** **Due 2:00 PM on Tuesday, 2/26/2019**

This is a take-home exam. ***Please insert your answers into this document.*** *You may use a calculator, your notes, my notes, and R. You may not discuss any aspect of this exam with anybody but me until after 3 pm on 2/26. Neatness and clarity will be appreciated. Do show the major steps in your work. Correct answers may not receive full credit without supporting work. Please insert only as much computer output as is necessary to support your answer. Please write and sign the Honor Pledge [I have neither given nor received help on this exam] below. Please do not answer questions on this page.*

Honor Pledge: **I have neither given nor received help on this exam.**

Name: **Ellen Stanton**

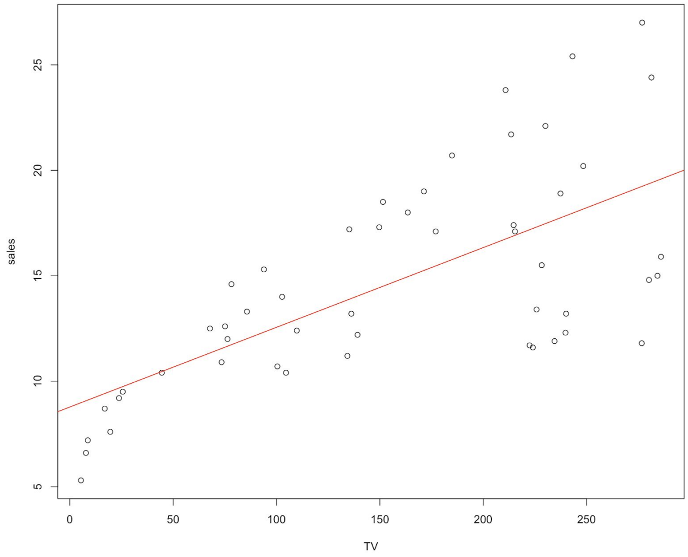
The data set *adsales.csv* is based on sales by a large manufacturer of small appliances. The data consists of sales (in thousands of appliances), TV advertising budget, radio advertising budget, and newspaper advertising budget (all in thousands of dollars) for a random sample of 50 world-wide markets.

(a) Verify (graphically and numerically) that TV is the best single predictor of sales.

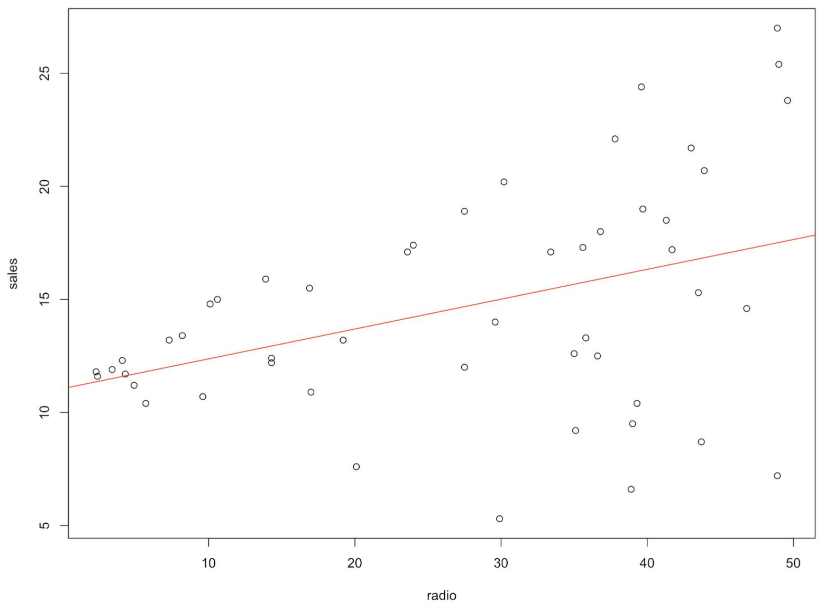
```
cor(adsales)
                 TV      radio   newspaper       sales
TV        1.0000000 -0.3228970 -0.18491639  0.66885009
radio    -0.3228970  1.0000000  0.30889948  0.40455360
newspaper -0.1849164  0.3088995  1.00000000 -0.02896018
sales     0.6688501  0.4045536 -0.02896018  1.00000000
```

Numerically, we can see that TV is the greatest predictor of sales, for TV and sales have a correlation coefficient of .669. Sales and radio have a correlation coefficient of .404, and sales and newspaper have a correlation coefficient of -.029, both of which are weaker (since they're closer to zero) than the correlation of sales with TV. Because TV has the highest correlation coefficient with sales, it is the best single predictor of sales.

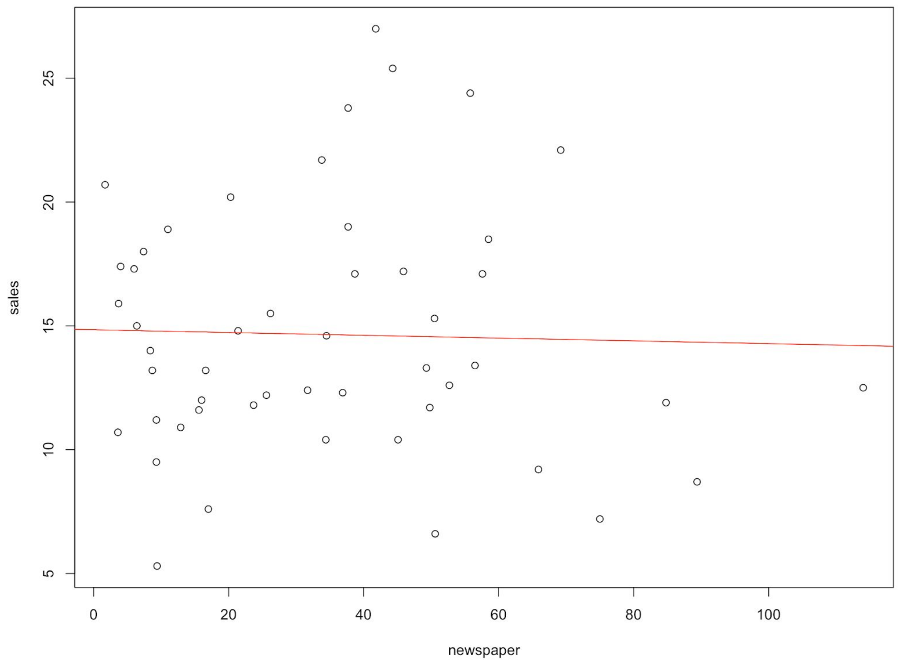
```
plot(sales~TV,data=adsales)
abline(lm(sales~TV,data=adsales),col="red")
```



```
plot(sales~radio,data=adsales)
abline(lm(sales~radio,data=adsales),col="red")
```



```
plot(sales~newspaper,data=adsales)
abline(lm(sales~newspaper,data=adsales),col="red")
```



The graphs shown visualize and confirm what we found numerically. TV has the strongest correlation with sales, since the points on the scatterplot fall (vertically) closest to the regression line of sales on TV, meaning that the residuals are smallest, meaning that the

model fits best. Radio is the next best single predictor of sales, since its points have smaller residuals than newspaper’s, though greater than TV’s. Note- discussion of points and their vertical distance from the regression line is of course about general trends overall, not specific points- also note that the graphs have the same scale on the y-axis for sales so these distances can be seen by the eye.

(b) Obtain the equation of the regression line predicting sales from TV spending. Interpret the slope of the line.

```
salesTV <- lm(sales~TV,data=adsales)
salesTV

Call:
lm(formula = sales ~ TV, data = adsales)

Coefficients:
(Intercept)           TV  
    8.77713      0.03778  
```

The equation of the regression line predicting sales from TV spending is:
Predicted Sales^ (in thousands of appliances) = 8.77713 + .03778(TV spending in thousands of dollars).
The slope of the line tells us that for each additional thousand dollars of TV advertising budget, the predicted sales increases by .03778 thousand appliances (or 37.78 appliances).

(d) How much of the variability in sales can be associated with TV spending?

```
summary(salesTV)

Call:
lm(formula = sales ~ TV, data = adsales)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.4320 -2.4362 -0.1515  2.8678  7.7604 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 8.777128   1.081498   8.116 1.47e-10 ***
TV          0.037784   0.006062   6.233 1.10e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.747 on 48 degrees of freedom
```

```
Multiple R-squared:  0.4474,  Adjusted R-squared:  0.4358
F-statistic: 38.86 on 1 and 48 DF,  p-value: 1.102e-07

.4474*100
[1] 44.74
```

44.74% of the variability in sales can be associated with TV spending.

(e) Obtain the appropriate ANOVA table. How does your answer in part (d) relate to the sums of squares in the ANOVA table?

```
anova(salesTV)
Analysis of Variance Table

Response: sales
          Df Sum Sq Mean Sq F value    Pr(>F)
TV         1 545.49  545.49  38.856 1.102e-07 ***
Residuals 48 673.86   14.04
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(545.49)/(545.49+673.86)
[1] 0.4473613
```

The answer in part (d) of 44.74% shows that the coefficient of determination is .4474, as was found using `summary(salesTV)`. The coefficient of determination can also be calculated from the sums of squares in the ANOVA table, as the sum of squares due to the regression divided by the total sum of squares (SSREG/SSTOT). My code above confirms this result.

(f) Provide a rough interpretation of the residual standard error in this case.

```
sqrt(14.04)
[1] 3.746999
```

The residual standard error is the square root of the mean square of residuals and is also found in the output of `summary(salesTV)` pasted in part (d). In this case, it is 3.747. This is roughly how far, on average, the points are vertically from the regression line. Units in this case are number of appliances in thousands. So, the 50 sample markets are on average 3.747 thousand appliances, or 3747 appliances, off from the regression line between TV spending and sales. This measures the lack of fit of the model.

(g) The TV expenditure in Boston is $5,000 higher than the TV expenditure in Richmond, VA. How much higher would you predict the sales in Boston to be compared to Richmond.?

```
.03778*5000
[1] 188.9
```

I would predict that the sales in Boston is 188.9 appliances higher than the sales in Richmond. If TV expenditure in Richmond is x thousand dollars, it's 5+x thousand dollars in Boston. Plugging 5+x in for TV spending in the regression line and then subtracting what you get when you plug in x for TV spending will find the difference in thousands of appliances. In this case, it's .03778*5, which is .1889 thousand or 188.9 appliances. I found this by multiplying: .03778 appliances/dollar times 5000 dollars = 188.9 appliances.

(h) Provide the three components/conditions of the population model in this case. There is no need to check the validity of these conditions.

Linearity: There is a linear relationship between TV spending (x) and mean sales of appliances (μ_y).
Constant Standard Deviation: The standard deviation of sales among markets with the same TV spending is designated by a σ which does not vary with TV spending.
Normality: The distribution of sales among markets with the same TV spending is normal with a mean $\mu_{s|TV}=\beta_0+\beta_1$(TV spending) and standard deviation σ.

Note: In our data, TV spending is in thousands of dollars and sales is in thousands of appliances.

(i) Obtain a 95% confidence interval for the population slope (β_1). Give an interpretation of your interval.

The results of the following command were displayed in part (d) but have been reproduced here to show where the numbers used in my calculations were found.

```
summary(salesTV)

Call:
lm(formula = sales ~ TV, data = adsales)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.4320 -2.4362 -0.1515  2.8678  7.7604 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 8.777128   1.081498   8.116 1.47e-10 ***
TV          0.037784   0.006062   6.233 1.10e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.747 on 48 degrees of freedom
Multiple R-squared:  0.4474,  Adjusted R-squared:  0.4358 
F-statistic: 38.86 on 1 and 48 DF,  p-value: 1.102e-07
```

```
df <- 48
b1 <- .037784
se <- .006062
t <- qt(.975,df)
me <- t*se
lb <- b1-me
ub <- b1+me
lb;ub
[1] 0.02559553
[1] 0.04997247
```

```
confint(salesTV)
                 2.5 %      97.5 %
(Intercept) 6.60262968 10.95162611
TV          0.02559666  0.04997165
```

```
1000*lb;1000*ub
[1] 25.59553
[1] 49.97247
```

β_1 is the population slope, which means it’s the mean increase in appliance sales per additional dollar of TV spending. A 95% confidence interval for the population slope β_1 is .02560 appliances per dollar of TV spending to .04997 appliances per dollar of TV spending. This translates to a mean increase of 25.60 to 49.97 appliances per additional thousand dollars of TV spending. We’re 95% confident that β_1, the mean increase in appliance sales for every $1 increase in TV spending lies between .02560 and .04997 appliances. Consequently, we’re 95% confident that the mean increase in appliance sales for every $1000 increase in TV spending lies between 25.60 and 49.97 appliances.

(j) Based upon your interval in part (i), is it plausible that β_1 could be 0? Explain.

No, it’s not plausible that β_1 could be zero, given that the interval from part (i) does not contain zero. The interval in part (i) shows that we are 95% confident that β_1 is between

.02560 and .04997 appliances sold per additional dollar of TV spending. If given a null hypothesis that states that there's no linear relationship between TV spending and sales (which implies β_1=0), we would reject it at the 95% confidence level.

(j) Obtain 95% confidence intervals for the mean sales for markets where the TV expenditure was \$10,000, \$15,000, \$20,000, and \$25,000, respectively. Why do the width of your intervals vary?

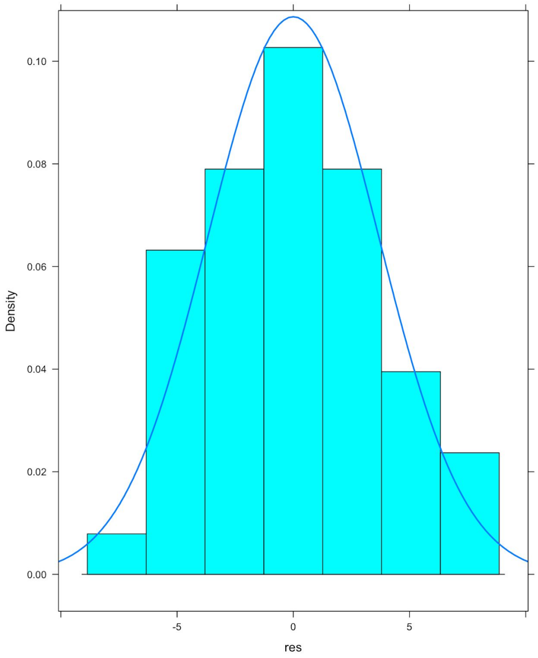
```
k <- data.frame(TV=c(10,15,20,25))
p <- predict(salesTV,newdata=k,interval="confidence")
p <- cbind(TV=c(10,15,20,25),p)
width <- p[,4]-p[,3]
p <- cbind(p,width)
p
  TV      fit      lwr      upr    width
1 10 9.154969 7.085854 11.22408 4.138231
2 15 9.343890 7.326769 11.36101 4.034242
3 20 9.532811 7.567170 11.49845 3.931281
4 25 9.721732 7.807017 11.63645 3.829431

mean(adsales$TV)
[1] 155.538
```

**A 95% confidence interval for TV spending of \$10,000 is 7.09 to 11.22 thousand appliance.
A 95% confidence interval for TV spending of \$15,000 is 7.33 to 11.36 thousand appliance.
A 95% confidence interval for TV spending of \$20,000 is 7.57 to 11.50 thousand appliances.
A 95% confidence interval for TV spending of \$25,000 is 7.81 to 11.64 thousand appliances.**

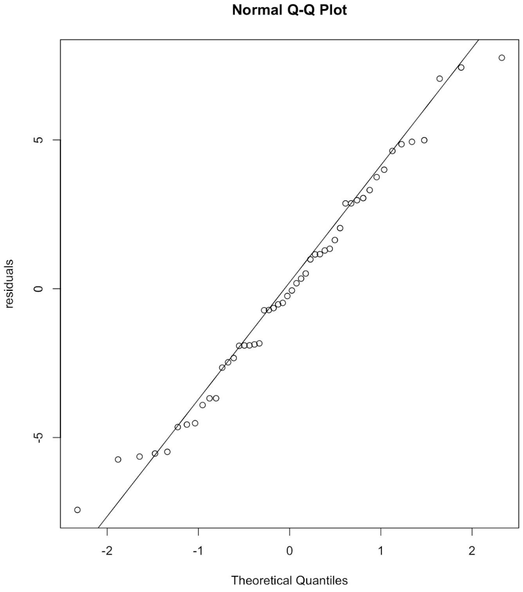
The widths of the intervals vary because the width of a confidence intervals for $\mu_{y|x}$ is always narrower for an x value closer to the sample mean of x (x-bar) and wider for an x value farther from x-bar. Essentially, the further x is from the mean value of x in the sample, the wider the confidence interval will be. In this case, our x values are 10, 15, 20, and 25, and x-bar is 155.538. So, the interval width should be narrowest for 25, wider for 20, even wider for 15, and widest for 10, because 25 is closest to 155.538 and 10 is furthest. This is indeed the case, with the 95% confidence interval for markets with \$25,000 of TV spending having a width of 3.83 thousand appliances, the 95% confidence interval for markets with \$20,000 of TV spending having a width of 3.93 thousand appliances, the 95% confidence interval for markets with \$15,000 of TV spending having a width of 34.03 thousand appliances, and the 95% confidence interval for markets with \$10,000 of TV spending having a width of 4.14 thousand appliances.

(k) As best you can, check the validity of the three conditions for the population model. One of the conditions appears to be violated. Which one and what is the evidence?

```
res <- resid(salesTV)
histogram(~res,fit="normal")
```

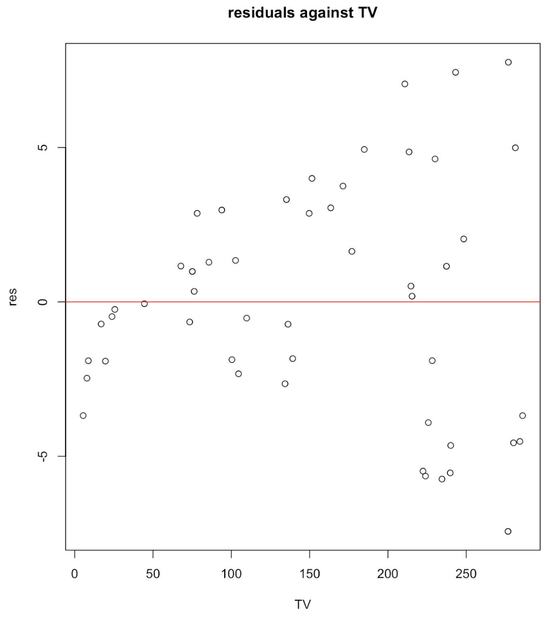


```
qqnorm(res,ylab="residuals")
qqline(res)
```



The normality condition is met. Both the histogram and the Normal Probability Plot show that the distribution of residuals is approximately normal. The histogram fits the superimposed normal curve well and the NPP is roughly linear along the superimposed line, though the residual values at either extreme of TV sales are a bit farther from the line than we'd like.

```
plot(res~TV,data=adsales,main="residuals against TV")
abline(h=0,col="red")
```

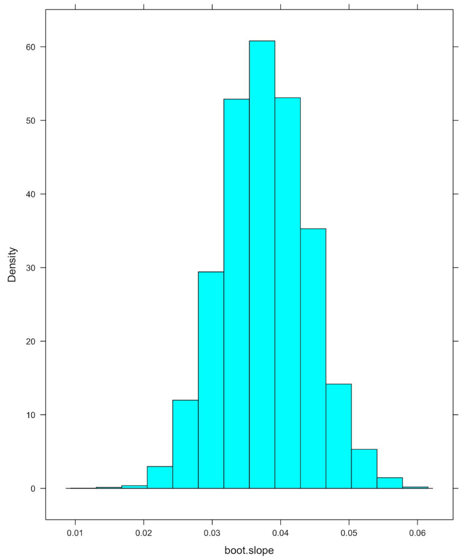


The linearity condition is met because for each approximate value of TV, the residuals follow a pattern of random noise around 0, meaning that they're distributed somewhat equally above and below the line residual=0.
The equal standard deviation condition is not met because for each approximate value of TV, the residuals have different spreads. It seems as if the spread of residuals increases as the value of TV spending increases.

(l) Repeat part (i) using a bootstrap approach.

```
library(mosaic)
boot.slope <- numeric(10000)
for (i in 1:10000)
{
  s <- adsales[sample(1:50, 50, replace=T),]
  l <- lm(sales~TV,data=s)
  c <- coef(l)
  boot.slope[i] <- c[2]
}
```

```
histogram(~boot.slope)
```



```
qdata(~boot.slope,p=c(.025,.975))
        quantile     p
2.5%  0.02589624 0.025
97.5% 0.05044540 0.975

lb <- 2*.037784-.05044540
ub <- 2*.037784-.02589624
lb;ub
[1] 0.0251226
[1] 0.04967176
```

A 95% bootstrap confidence interval for the population slope β_1 is .02512 appliances per additional dollar of TV spending to .04967 appliances per additional dollar of TV spending. This translates to 25.12 to 49.67 appliances per additional thousand dollars of TV spending. We're 95% confident that β_1, the mean increase in appliance sales for every $1 increase in TV spending lies between .02512 and .04967 appliances. Consequently, we're 95% confident that the mean increase in appliance sales for every $1000 increase in TV spending lies between 25.12 and 49.67 appliances.