STAT 308 – Take Home Exam Solutions

This take home exam is worth a total of 50 points, and counts for half of your Exam 1 grade, along with the in class portion of the exam. For the problems in which calculations are needed, please include your `R` code with your answers. Turn in this exam to Sakai by Thursday, October 6 at 2:30 pm.

1. For the following question, we will define

- β_0 and β_1 as the population intercept and slope for the linear model with Y as the response and X as the explanatory variable.

  For a random sample of $n = 40$, we obtain the following sample statistics:

$$\bar{x} = 2, \bar{y} = -3, s_x = 2.5, s_y = 4.5, r = -0.6.$$

a. [2 pts] Calculate estimates of the least squares regression line, $\hat{\beta}_0$ and $\hat{\beta}_1$.

```
xbar <- 2
ybar <- -3
s_x <- 2.5
s_y <- 4.5
r <- -0.6
beta1 <- r*s_y/s_x
beta0 <- ybar - beta1*xbar
beta1
```

```
## [1] -1.08
```

```
beta0
```

```
## [1] -0.84
```

$\hat{y} = -0.84 - 1.08x$

b. [2 pts] Calculate the sum of squared errors, SSE, and the estimate of the regression variance, $s^2_{Y|X}$. (Hint: recall that SSE is the percent of the total sums of squares not explained by the linear model).

```
n <- 40
SST <- (n-1)*s_y^2
SSE <- SST * (1-r^2)
s_yx2 <- SSE/(n-2)
SSE
```

```
## [1] 505.44
```

```
s_yx2
```

```
## [1] 13.30105
```

SSE $= 505.44$ $s^2_{y|x} = 13.30$

c. [2 pts] Calculate the standard errors of the regression estimates, $s_{\hat{\beta}_0}$ and $s_{\hat{\beta}_1}$.

```
s_beta1 <- sqrt(s_yx2/((n-1)*s_x^2))
s_beta1
```

```
## [1] 0.2335988
```

```
s_beta0 <- sqrt(s_yx2)*sqrt((1/n) + xbar^2/((n-1)*s_x^2))
s_beta0
```

```
## [1] 0.742159
```

$s_{\hat{\beta}_0} = 0.742$, $s_{\hat{\beta}_1} = 0.234$

d. [2 pts] Calculate a test statistic and p-value for testing $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 > 0$.

```
t <- beta1/s_beta1
t
```

```
## [1] -4.623311
```

```
p <- pt(t,df=n-2,lower.tail=FALSE)
p
```

```
## [1] 0.9999786
```

$t = -4.62$, $p - value = 0.99998$

For the following questions, we will use the dataset `nfl` contains information about the regular season games from the 2021 regular season, including

- `Week`: The Week the game took place
- `Day`: The Day of the week the game took place
- `Date`: The Date of the game
- `Time`: The time the game started (in Eastern Time)
- `AwayTeam`: The game's away team
- `HomeTeam`: The game's home team
- `YdsDiff`: The difference in the offensive yards between the home team and away team

- `TODiff`: The difference in turnovers between the home team and away team
- `PtsDiff`: The difference in points scored between the home team and away team

2. Suppose we think that the point difference is not dependent on any other explanatory variables.

a. [2 pts] What do we expect the point differential between a randomly selected home and away team to be?

```
nfl <- read.csv("../data/nfl.csv")
mean(nfl$PtsDiff)
```

```
## [1] 1.713235
```

We expect a randomly selected NFL game to have a difference of 1.71 points between the home and away teams.

b. [4 pts] Calculate a 99% confidence interval for the expected point differential. Interpret this interval in the context of the problem.

```
est <- mean(nfl$PtsDiff)
n <- length(nfl$PtsDiff)
se <- sd(nfl$PtsDiff)/sqrt(n)
alpha = 0.01
crit <- qt(1-alpha/2,df=n-1)
est + c(-1,1)*crit*se
```

```
## [1] -0.7141366  4.1406072
```

We are 99% confident that the true average difference in home team and away teams scores is between -0.714 and 4.14 points.

c. [4 pts] Bookmakers often maneuver their point spreads to account for the belief that, on average, home teams win by three points. Perform a hypothesis test on the counterargument that home teams do not, on average, win by three points. Be sure to formally state all the necessary information to perform a formal hypothesis test. Assume $\alpha = 0.05$.

$$H_0 : \mu = 3$$
$$H_a : \mu \neq 3$$

```
t.test(nfl$PtsDiff,conf.level=0.95,mu=3)
```

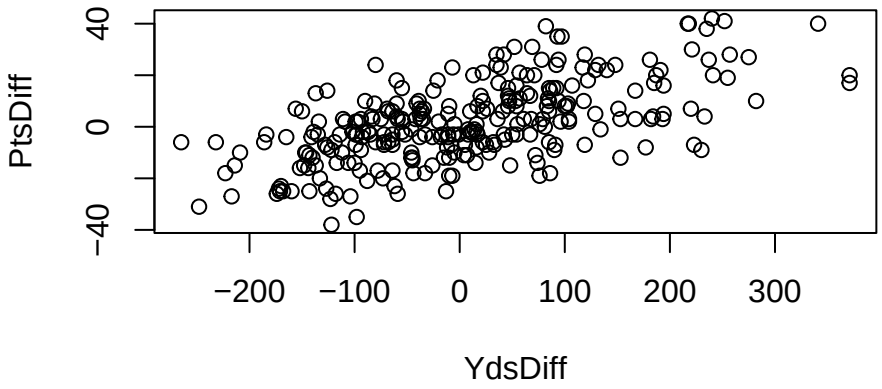
```
##
##  One Sample t-test
##
## data:  nfl$PtsDiff
## t = -1.3751, df = 271, p-value = 0.1702
## alternative hypothesis: true mean is not equal to 3
## 95 percent confidence interval:
##  -0.1289904  3.5554610
## sample estimates:
## mean of x
##  1.713235
```

$t = -1.375$, $p - value = 0.17$. Because $p - value > \alpha$, we fail to reject H_0 and we can conclude that there is not statistically significant evidence that the average point difference between home and away teams is not equal to 3 points.

3. Now, suppose we think that the observed point differentials can be better explained through a linear relationship with the difference in offensive yards gained.

a. [2 pts] Create a scatterplot of yard differential vs. point differential. Comment on the validity of a linear relationship.

```
plot(PtsDiff ~ YdsDiff,nfl)
```



The scatterplot shows a moderately strong linear relationship, so a linear model appears to be valid.

b. [2 pts] Calculate a least squares regression line using R. Formally state the regression line.

```
mod <- lm(PtsDiff ~ YdsDiff,nfl)
summary(mod)
```

```
##
## Call:
## lm(formula = PtsDiff ~ YdsDiff, data = nfl)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.4658  -9.5369  -0.1682   7.8160  31.2914
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.179596   0.740676   1.593    0.112
## YdsDiff     0.079621   0.006219  12.802   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.2 on 270 degrees of freedom
## Multiple R-squared:  0.3777, Adjusted R-squared:  0.3754
## F-statistic: 163.9 on 1 and 270 DF,  p-value: < 2.2e-16
```

$\hat{Y} = 1.18 + 0.08x$

c. [2 pts] Determine if there are any influential observations in the given dataset. If you determine that there are, state your new least squares regression line.

```
studres <- rstudent(mod)
n <- nrow(nfl)
crit <- qt(1-0.01/2,df=n-2-1)
max(abs(studres)) > crit
```

```
## [1] TRUE
```

```
id <- which.max(abs(studres))

mod2 <- lm(PtsDiff ~ YdsDiff,nfl[-id,])
studres <- rstudent(mod2)
crit <- qt(1-0.01/2,df=n-2-2)
max(abs(studres)) > crit
```

```
## [1] FALSE
```

Line 202 is an influential observation, so we will remove that value.

```
summary(mod2)
```

```
##
## Call:
## lm(formula = PtsDiff ~ YdsDiff, data = nfl[-id, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.4294  -9.4496  -0.1084   7.8167  29.2524
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.068086   0.734157   1.455    0.147
## YdsDiff     0.079006   0.006159  12.829   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.07 on 269 degrees of freedom
## Multiple R-squared:  0.3796, Adjusted R-squared:  0.3773
## F-statistic: 164.6 on 1 and 269 DF,  p-value: < 2.2e-16
```

$\hat{Y} = 1.068 + 0.079x$

For the remainder of this exam, use either the regression line in (b) if you determine there are no influential observations, or the line in (c) if you determine that there are influential observations.

d. [2 pts] What is the estimate of the regression variance?

```
s2 <- deviance(mod2)/df.residual(mod2)
s2
```

```
## [1] 145.6409
```

$s^2 = 145.64$

e. [2 pts] Interpret the estimate of the population slope in the context of the problem.

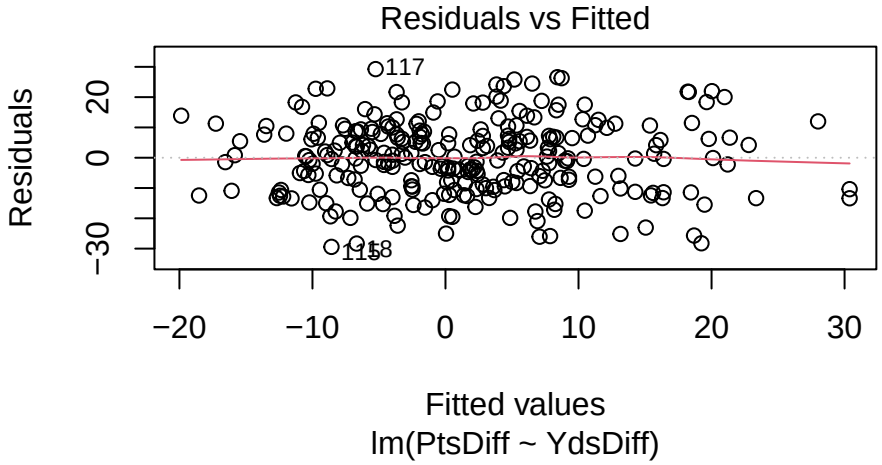
When the difference in yards between the home and away team increases by 1 yard, the expected point difference increases by 0.079 yards.

f. [2 pts] Interpret the estimate of the population intercept in the context of the problem. Does this interpretation make sense? Why or why not?

When the difference between the home and away teams yards is 0, the expected point difference is 1.068 points. This does make sense because it is possible for teams to have the same number of yards in the game.

g. [2 pts] Determine if the assumption of heteroscedasticity is violated.

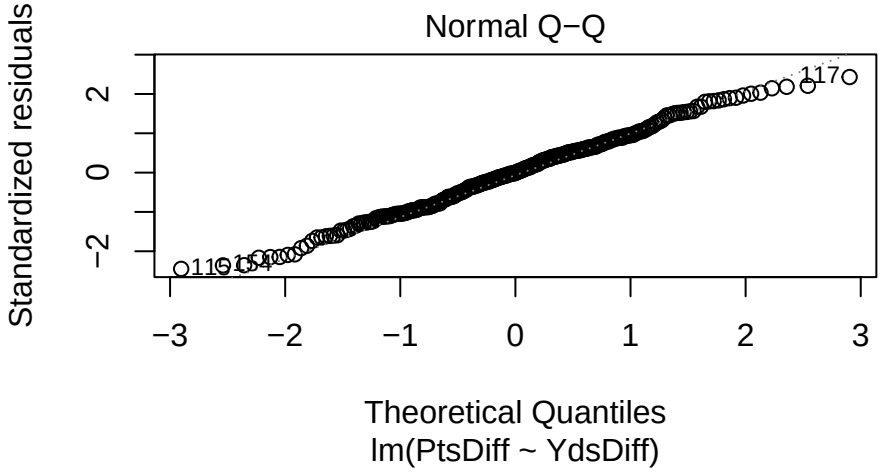
```
plot(mod2,1)
```



The residuals vs. fitted values plot appears to be randomly scattered around 0, so the assumption of homoscedasticity is not violated.

h. [2 pts] Determine if the assumption of normally distributed residuals is violated.

```
plot(mod2,2)
```



The points appear to fall close to the 45 degree line, so the assumption of normally distributed residuals does not appear to be violated.

i. [4 pts] Find the r^2 for this linear regression model. Interpret the r^2 in the context of the problem.

```
summary(mod2)$r.squared
```

```
## [1] 0.3795761
```

37.96% of the variation in point differential can be explained by its linear relationship with yard differential.

j. [4 pts] Calculate a 99% confidence interval for the population slope. Interpret this interval in the context of the problem.

```
confint(mod2,level=0.99)
```

```
##                   0.5 %     99.5 %
## (Intercept) -0.83648454 2.97265589
## YdsDiff      0.06302899 0.09498224
```

We are 99% confident that when the yard differential increases by 1 yard, the true average point differential increases by between 0.063 and 0.094 points.

k. [4 pts] Perform a hypothesis test for a positive linear relationship between yard differential and point differential. Be sure to formally state all the necessary information to perform a formal hypothesis test. Assume $\alpha = 0.01$.

$$H_0 : \beta_1 = 0$$
$$H_0 : \beta_1 > 0$$

```
summary(mod2)
```

```
##
## Call:
## lm(formula = PtsDiff ~ YdsDiff, data = nfl[-id, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.4294  -9.4496  -0.1084   7.8167  29.2524
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.068086   0.734157   1.455    0.147
## YdsDiff     0.079006   0.006159  12.829   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.07 on 269 degrees of freedom
## Multiple R-squared:  0.3796, Adjusted R-squared:  0.3773
## F-statistic: 164.6 on 1 and 269 DF,  p-value: < 2.2e-16
```

$t = 12.829$, p-value = <2e-16/2 = <1e-16. Because p-value $\leq \alpha$, we reject H_0 and say there is significant evidence of a positive linear relationship between yard differential and point differential.

l. [4 pts] Calculate a prediction as well as a 90% prediction interval for the point differential when the home team has 100 more yards than the away team. Interpret the prediction interval in the context of the given problem.

```
newdata=data.frame(YdsDiff=100)
predict(mod2,newdata,interval="prediction",level=0.9)
```

```
##        fit       lwr      upr
## 1 8.968648 -11.00971 28.94701
```

Prediction = 8.969. We are 90% confident that when a randomly selected game has a home team with 100 more yards than the road team, the point differential will be between -11.01 and 28.95 points.