Stats 110 Homework 1

Owen Lin

1. a. False
   b. False
   c. True
   d. False
   e. True
   f. False
2. a. Same, R is the correlation coefficient of X and Y, it measures how strong is the linear association between the two variables.
   b. Same, because R is the same, so R^2 is the same
   c. β_1 (slope) would be different (unless originally slope = 1)
   d. β_0 (y-intercept) would be different too, since X are Y switch position and the new y-intercept is the x-intercept before.
   e. Same, because R is the same and $T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$
   f. The residual for each each observation would be different because both variables changes as residual = predicted value - observed value
3. a. See Code Outputs

```
MidWestSales <- read.table("D:\\Coding\\Data\\Stats110\\MidwestSales.txt", fill = TRUE, header =
FALSE) # nolint
names(MidWestSales) <- c("id","price","sqft","bed","bath","ac","garage","pool","year","quality",
"style","lot","hwy") # nolint
model <- lm(price ~ sqft, data = MidWestSales)
summary(model)$coef
print(paste("y = ", model$coefficient[1], "+", model$coefficient[2], "x"))
```

```
##                Estimate   Std. Error   t value      Pr(>|t|)
## (Intercept) -81432.9464 11551.845727 -7.049345  5.744858e-12
## sqft           158.9502     4.874987 32.605261 8.284610e-128
## [1] "y =  -81432.9463955519 + 158.950230811551 x"
```

   b. For every one square footage increase of the house, we expect the price of the house to increase by 158.95 dollar
   c. H0: square footage has no linear relationship with price ($\beta_1 = 0$)
   Ha: square footage has a linear relationship with price ($\beta_1 \neq 0$)
   T-statistic: 32.605, which follows t-distribution with a degree of freedom of 520
   p-value = 8.284610e-128 < 0.05
   Thus, we reject the null and conclude that square footage has a significant linear relationship with price
   d. H0: square footage has no **positive** linear relationship with price ($\beta_1 \leq 0$)
   Ha: square footage has a significant **positive** linear relationship with price ($\beta_1 > 0$)
   T-statistic: 32.605, which follows t-distribution with a degree of freedom of 520
   p-value = 8.284610e-128/2 = 4.142305e-128 < 0.05

Thus, we reject the null and conclude that square footage has a significant **positive** linear relationship with price.

```
CI <- predict(model, list(sqft = 2000), interval = "c", level = 0.95)
PI <- predict(model, list(sqft = 2000), interval = "p", level = 0.95)
print(paste("The 95% confidence interval for the mean price when sqft=2000 is [", CI[2], ",", CI
[3], "]")) #nolint
print(paste("The 95% prediction interval for the price when sqft=2000 is [", PI[2], ",", PI[3],
"]")) # nolint
summary(MidWestSales$sqft)
```

```
## [1] "The 95% confidence interval for the mean price when sqft=2000 is [ 229220.673982746 , 24
3714.356472354 ]"
## [1] "The 95% prediction interval for the price when sqft=2000 is [ 80858.8455775154 , 392076.
184877585 ]"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     980    1701    2061    2261    2636    5032
```

e. We are 95% confident that the interval (229220.67, 243714.36) contains the **average** price when sqft = 2000

f. We are 95% confident that the interval (80858.85, 392076.18) contains the price when sqft = 2000.

g. The interval in part f would be narrower if we decreased the confidence level to 90%, because lower confidence means we try to predict more precisely.

h. No, it doesn't makes sense to predict the sale price of a house that is 8500 square feet, because the domain of our dataset is from 980 to 5032. There is no reason to believe that the linear trend continues to 8500 sqft houses.

i. The interval would cover all real number if the confidence level is increased to 100%, because there can always be outliers. It is unlikely to happen doesn't mean it will not happen. To be 100% confident, the interval needs to cover the entire domain.

j. The estimate of σ_ϵ is 79122.9. This is the square root of the variance in the observed price that is not explained by the variation in square footage of the house.

```
summary(model)$sigma
```

```
## [1] 79122.9
```

k. No, because the bigger house can varied more in price. It can be a big but plain house with low price or a big and fully decorated house with high price. On the other hand, the smaller house has less space for decoration => the price can't varies as much.

4. a. Yes, latitude has a linear association with the mortality because the p-value for β_1 is 3.31e-23 < 0.05

```
skincancer <- read.table("D:\\Coding\\Data\\Stats110\\skincancer.txt", fill = TRUE, header = TRU
E) #nolint
head(skincancer)
model2 <- lm(Mort ~ Lat, data = skincancer)
summary(model2)
```

```
##         State  Lat Mort Ocean  Long
## 1     Alabama 33.0  219     1  87.0
## 2     Arizona 34.5  160     0 112.0
## 3    Arkansas 35.0  170     0  92.5
## 4  California 37.5  182     1 119.5
## 5    Colorado 39.0  149     0 105.5
## 6 Connecticut 41.8  159     1  72.8
##
## Call:
## lm(formula = Mort ~ Lat, data = skincancer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.972 -13.185   0.972  12.006  43.938
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 389.1894    23.8123   16.34  < 2e-16 ***
## Lat          -5.9776     0.5984   -9.99 3.31e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.12 on 47 degrees of freedom
## Multiple R-squared:  0.6798, Adjusted R-squared:  0.673
## F-statistic:  99.8 on 1 and 47 DF,  p-value: 3.309e-13
```

```
CI <- predict(model2, list(Lat = 40), interval = "c", level = 0.99)
PI <- predict(model2, list(Lat = 40), interval = "p", level = 0.99)
# print(confint(model2, level = 0.99)) CI for estimators
print(paste("The 99% confidence interval for the mean mortality rate when Lat=40 is [", CI[2],
",", CI[3], "]")) #nolint
print(paste("The 99% prediction interval for the mortality rate when Lat=40 is [", PI[2], ",", P
I[3], "]")) # nolint
```

```
## [1] "The 99% confidence interval for the mean mortality rate when Lat=40 is [ 142.71481544150
9 , 157.453027066775 ]"
## [1] "The 99% prediction interval for the mortality rate when Lat=40 is [ 98.2421431690755 , 2
01.925699339209 ]"
```

b. We are 99% confident that the interval (142.71, 157.45) contains the **average** mortality rate when latitude is 40.
c. We are 99% confident that the interval (98.24, 201.93) contains the mortality rate when latitude is 40.
d. The center of the confidence interval and prediction interval is the same, because the only difference between the confidence and the prediction interval lies in standard error.
e. The width of the confidence interval is narrower compare to the prediction interval, because the standard error for the prediction interval for the actual mortality rate has an extra MSE term (which is strictly positive)

5. a. Sample size is increased => prediction interval will be narrower
   b. If Xp gets closer to $\bar{X}$ => prediction interval will be narrower
   c. If the variability of the response variable decreases => prediction interval will be narrower

d. The average of the response is increased => prediction interval stay the same (but shift upward)

6. $$R^2 = \frac{SSR}{SSTO} = \frac{SSR}{(SSR + SSE)} = \frac{110}{110 + 40} = 0.73333$$

73% of the variation in Y is explained by X

7. a. H0: Resting pulse rate has no linear relationship with smoking status ($\beta_1 = 0$)
Ha: Resting pulse rate has a linear relationship with smoking status ($\beta_1 \neq 0$)
T-statistic: 2.429, which follows t-distribution with a degree of freedom of 230
p-value = 0.0159 < 0.05
Thus, we reject the null and conclude that resting pulse rate has a linear relationship with smoking status

```
pulse <- read.table("D:\\Coding\\Data\\Stats110\\Pulse.txt", fill = TRUE, header = TRUE) #nolint
pulse$Smoker <- ifelse(pulse$Smoke == 1, "Yes", "No")
# head(pulse)
# ls(pulse)
model3 <- lm(Rest ~ Smoke, data = pulse)
summary(model3)
```

```
##
## Call:
## lm(formula = Rest ~ Smoke, data = pulse)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.791  -6.041  -0.791   6.209  38.209
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   67.791      0.686  98.826   <2e-16 ***
## Smoke          4.978      2.049   2.429   0.0159 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.845 on 230 degrees of freedom
## Multiple R-squared:  0.02502,    Adjusted R-squared:  0.02078
## F-statistic: 5.902 on 1 and 230 DF,  p-value: 0.01589
```

b. The t-statistic is the same as the test from part a (t = 2.429), and so is the p-value and the conclusion.

```
t.test(pulse$Rest[pulse$Smoker=="Yes"], pulse$Rest[pulse$Smoker=="No"], var.equal=TRUE) #nolint
```

```
##
##  Two Sample t-test
##
## data:  pulse$Rest[pulse$Smoker == "Yes"] and pulse$Rest[pulse$Smoker == "No"]
## t = 2.4294, df = 230, p-value = 0.01589
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.9405909 9.0153463
## sample estimates:
## mean of x mean of y
##  72.76923  67.79126
```

c. Population model: $y = \beta_0 + \beta_1 * X1 + \beta_2 * X2 + \epsilon$, Y: resting pulse rate, x1: weight, x2: smoking status estimated regression equation: $y = 78.247 - 0.067X1 + 6.043X2$

```
model4 <- lm(Rest ~ Wgt + Smoke, data = pulse)
round(coef(summary(model4)), 3)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   78.247      3.221  24.296    0.000
## Wgt           -0.067      0.020  -3.319    0.001
## Smoke          6.043      2.031   2.975    0.003
```

d. Without an interaction term, we are assuming that the slope on weight with respect to resting pulse rate is the same for smoker and non-smoker. In other words, the smoking status is only an additive parallel shift.

e.

```
# ls(summary(model4))
summary(model4)
rsq <- summary(model4)$r.squared
sigma <- summary(model4)$sigma
print(paste("The coefficient of determination (multiple R^2) is", round(rsq, 3))) #nolint
print(paste("The estimate of stardard error on the error term is", round(sigma, 3))) #nolint
```

```
##
## Call:
## lm(formula = Rest ~ Wgt + Smoke, data = pulse)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.872  -6.207  -0.719   5.794  37.128
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 78.24692    3.22061  24.296  < 2e-16 ***
## Wgt         -0.06697    0.02017  -3.319  0.00105 **
## Smoke        6.04288    2.03136   2.975  0.00325 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.638 on 229 degrees of freedom
## Multiple R-squared:  0.06978,    Adjusted R-squared:  0.06165
## F-statistic: 8.589 on 2 and 229 DF,  p-value: 0.0002531
##
## [1] "The coefficient of determination (multiple R^2) is 0.07"
## [1] "The estimate of stardard error on the error term is 9.638"
```

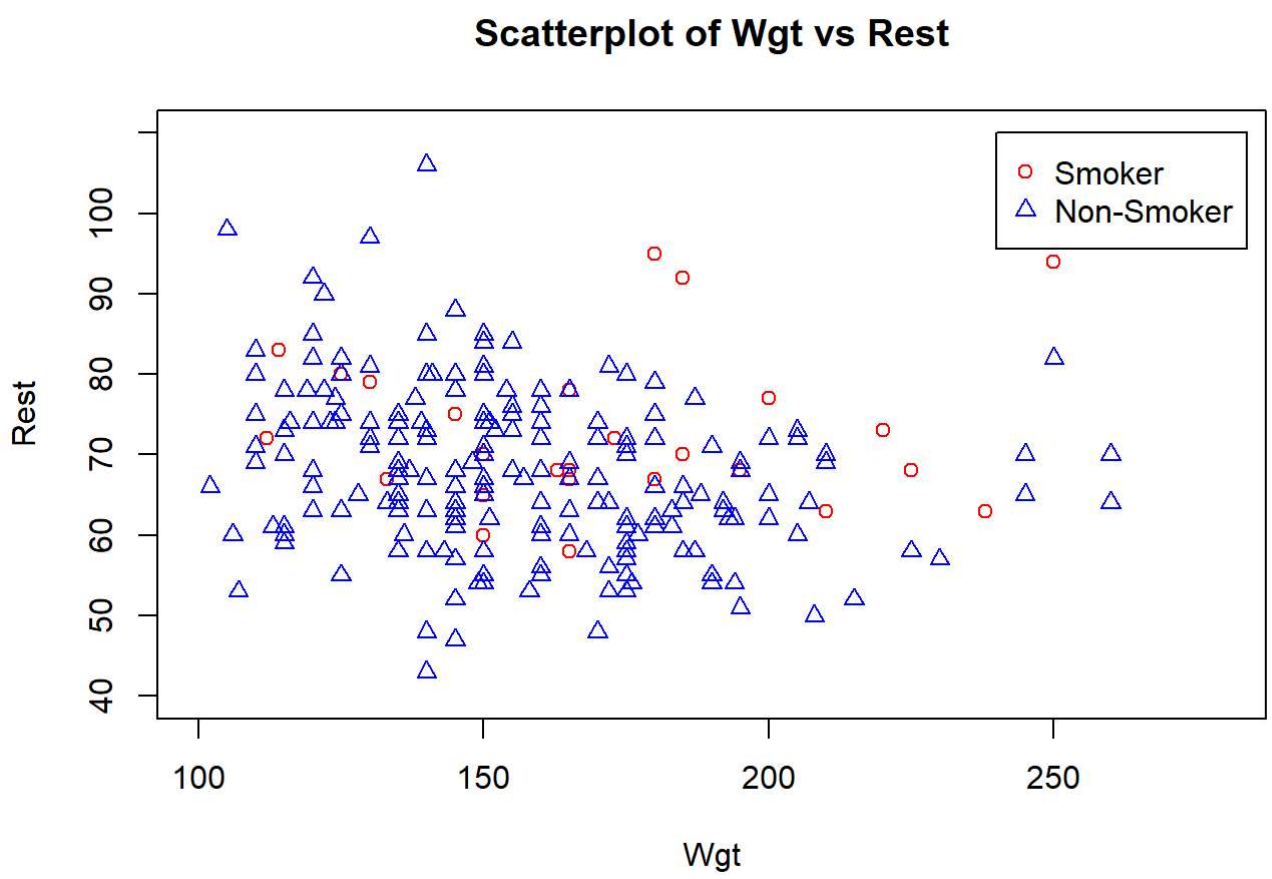
f. There are 229 + 3 = 232 observations

g. H0: Smoking status does not affect resting pulse rate ($\beta_2 = 0$)
   Ha: Smoking status does affect resting pulse rate ($\beta_2 \neq 0$)
   T-statistic: 2.975, which follows t-distribution with a degree of freedom of 229
   p-value = 0.00325 < 0.05
   Thus, we reject the null and conclude that Smoking status does affect resting pulse rate

h. We can't conclude that smoking causes lower resting pulse rate because the data did not come from not an controlled experiment.

i. For Non-smoker, there seems to be some negative association between weight and resting pulse rate. For smokers, there are more noise in the data and we can't really tell if there's an association.

```
plot(pulse$Wgt[pulse$Smoker=="Yes"] , pulse$Rest[pulse$Smoker=="Yes"], xlab="Wgt", #nolint
       ylab ="Rest", main="Scatterplot of Wgt vs Rest", col="red", xlim=c(100,280), ylim=c(40,11
0)) #nolint
points(pulse$Wgt[pulse$Smoker=="No"] , pulse$Rest[pulse$Smoker=="No"], col = "blue", pch = 24) #
nolint
legend(240,110,legend=c("Smoker", "Non-Smoker"),pch=c(1,24), col=c("red","blue")) #nolint
```



j. Population Model: $y = \beta_0 + \beta_1 * X1 + \beta_2 * X2 + \beta_3 * X1 * X2 + \epsilon$, Y: resting pulse rate, x1: weight, x2: smoking status, x3: interaction term

```
model5 <- lm(Rest ~ Wgt + Smoke + Wgt*Smoke, data = pulse)
round(coef(summary(model5)), 3)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   80.430      3.464  23.222    0.000
## Wgt           -0.081      0.022  -3.719    0.000
## Smoke        -10.031      9.824  -1.021    0.308
## Wgt:Smoke      0.095      0.057   1.672    0.096
```

k. Estimated regression equation: $y = 80.430 - 0.081X1 + 10.031X2 + 0.095X2 * X3$

l. R square increase and standard deviation of the error term decreased as we add the interaction term.

```
rsq <- summary(model5)$r.squared
sigma <- summary(model5)$sigma
summary(model5)
ls(model5)
print(paste("The coefficient of determination (multiple R^2) is", round(rsq, 3))) #nolint
print(paste("The estimate of stardard error on the error term is", round(sigma, 3))) #nolint
```

```
##
## Call:
## lm(formula = Rest ~ Wgt + Smoke + Wgt * Smoke, data = pulse)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.097  -6.182  -0.752   5.832  36.903
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  80.43031    3.46360  23.222  < 2e-16 ***
## Wgt          -0.08095    0.02177  -3.719 0.000252 ***
## Smoke       -10.03147    9.82355  -1.021 0.308258
## Wgt:Smoke     0.09473    0.05665   1.672 0.095863 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.6 on 228 degrees of freedom
## Multiple R-squared:  0.08105,    Adjusted R-squared:  0.06896
## F-statistic: 6.703 on 3 and 228 DF,  p-value: 0.0002356
##
##  [1] "assign"        "call"          "coefficients"  "df.residual"
##  [5] "effects"       "fitted.values" "model"         "qr"
##  [9] "rank"          "residuals"     "terms"         "xlevels"
## [1] "The coefficient of determination (multiple R^2) is 0.081"
## [1] "The estimate of stardard error on the error term is 9.6"
```

m. H0: The effect of weight on resting pulse rate is the same for smokers and non-smokers ($\beta_3 = 0$)
Ha: The effect of weight on resting pulse rate differs for smokers and non-smokers ($\beta_3 \neq 0$)
T-statistic: 1.672, which follows t-distribution with a degree of freedom of 228
p-value = 0.0959 > 0.05
At 95 confident level, we fail to reject the null. Whether the effect of weight on resting pulse rate differs for smokers and non-smokers is still inconclusive.

n. No, the value of R square is 0.07, which means only 7 percents of variation in resting pulse rate is explained by the weight.

o. We can infer that people who weight more exercise less, which can be a factor that decreases resting pulse rate.

p. H0: Adding weight as a variable doesn't improve R square
Ha: Adding weight as a variable does improve R square significantly