Ximan Liu
1935858
DSC 423
HW2

I have completed this work independently. The solutions given are entirely my own work.

【Question 1】

1a)

“R-squared value is 0.69” represents that 69% of the variability in weight is explained by the model. Meanwhile, 0.69 is lower than 0.7, which means that is a relatively lower R-squared value. The performance of the model may be poor, in that case we may need more judgment conditions.

There is no double that, for statistics, R-squared value is very important. It explains how well the explanation power of the regression model (with the scale of 0 - 100%); In other words, we can say, how well the regression model fits the dataset.

In that case, the greater the R-squared value, the better performance it may have. But this is not the only condition that determines the quality of the model. For example, we still need more technical data like residual plots to determine whether it is a biased model.

1b)

The regression fallacy is a fallacy that presupposed someone/something has done the corrections on someone/something when there is an error, and the result is back to normal. It lacks the explanation and consideration of natural variables.

For example, Joe had a fever one week after got the COVID vaccination. It could be the side effect of the vaccine, but it could also be catching cold or flu.

Another example, she was drowsy after taking melatonin and pain reliever. It is difficult to explain whether it was due to melatonin or painkillers that caused the drowsiness.

Citation:

Wikimedia Foundation. (2021, May 11). Regression fallacy. Wikipedia. https://en.wikipedia.org/wiki/Regression_fallacy.

【Question 2】

2a)

```
Y1 <- QUASAR$RFEWIDTH
X1 <- QUASAR$REDSHIFT
X2 <- QUASAR$LINEFLUX
X3 <- QUASAR$LUMINOSITY
X4 <- QUASAR$AB1450
X5 <- QUASAR$ABSMAG

model1 <- lm(Y1 ~ X1)
```

summary(model1)

model2 <- lm(Y1 ~ X2)
summary(model2)

model3 <- lm(Y1 ~ X3)
summary(model3)

model4 <- lm(Y1 ~ X4)
summary(model4)

model5 <- lm(Y1 ~ X5)
summary(model5)

2b)
<1>
model1 <- lm(Y1 ~ X1)
summary(model1)

```
Call:
lm(formula = Y1 ~ X1)

Residuals:
    Min      1Q  Median      3Q     Max 
-54.922 -36.077  -8.504  24.590 166.590 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  112.115     70.151   1.598    0.124
X1            -7.013     20.477  -0.342    0.735

Residual standard error: 48.29 on 23 degrees of freedom
Multiple R-squared:  0.005073,  Adjusted R-squared:  -0.03818 
F-statistic: 0.1173 on 1 and 23 DF,  p-value: 0.7351
```

This model represents a regression relationship between Y1(Rest frame Equivalent Width) and X1(REDSHIFT).
The intercept of this model is 112.115.
Based on R-squared, we can say, 0.5073% of the variability in weight is explained by the model. Since it represents coefficient of determination, and the R2 value is lower than 0.70, we do not consider it as a good model.

<2>
model2 <- lm(Y1 ~ X2)
summary(model2)

```
Call:
lm(formula = Y1 ~ X2)

Residuals:
    Min      1Q  Median      3Q     Max 
-59.053 -32.667  -9.432  25.137 157.947 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   665.77     563.70   1.181    0.250
X2             41.83      40.83   1.025    0.316

Residual standard error: 47.35 on 23 degrees of freedom
Multiple R-squared:  0.04365,	Adjusted R-squared:  0.002066 
F-statistic:  1.05 on 1 and 23 DF,  p-value: 0.3162
```

This model represents a regression relationship between Y1(Rest frame Equivalent Width) and X2(Line Flux).
The intercept of this model is 665.77.
Based on R-squared, we can say, 4.365% of the variability in weight is explained by the model. Since it represents coefficient of determination, and the R2 value is lower than 0.70, we do not consider it as a good model.

<3>
model3 <- lm(Y1 ~ X3)
summary(model3)

```
Call:
lm(formula = Y1 ~ X3)

Residuals:
    Min      1Q  Median      3Q     Max
-53.800 -30.427  -5.716  21.960 164.875

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1978.21    2226.43  -0.889    0.383
X3             45.78      49.32   0.928    0.363

Residual standard error: 47.53 on 23 degrees of freedom
Multiple R-squared:  0.03611,   Adjusted R-squared:  -0.005803
F-statistic: 0.8615 on 1 and 23 DF,  p-value: 0.3629
```

This model represents a regression relationship between Y1(Rest frame Equivalent Width) and X3(Line Luminosity).
The intercept of this model is -1978.21.
Based on R-squared, we can say, 3.611% of the variability in weight is explained by the model. Since it represents coefficient of determination, and the R2 value is lower than 0.70, we do not consider it as a good model.

<4>
model4 <- lm(Y1 ~ X4)
summary(model4)

```
Call:
lm(formula = Y1 ~ X4)

Residuals:
    Min      1Q  Median      3Q     Max 
-50.630 -24.405  -3.409   7.946 144.479 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -667.31     239.42  -2.787   0.0105 * 
X4             38.31      12.13   3.158   0.0044 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.44 on 23 degrees of freedom
Multiple R-squared:  0.3024,	Adjusted R-squared:  0.2721 
F-statistic: 9.972 on 1 and 23 DF,  p-value: 0.004399
```

This model represents a regression relationship between Y1(Rest frame Equivalent Width) and X4(AB1450 Magnitude).
The intercept of this model is -667.31.
Based on R-squared, we can say, 30.24% of the variability in weight is explained by the model. Since it represents coefficient of determination, and the R2 value is higher than 0.70, we consider it as a good model.

<5>
model5 <- lm(Y1 ~ X5)
summary(model5)

```
Call:
lm(formula = Y1 ~ X5)

Residuals:
    Min      1Q  Median      3Q     Max 
-56.281 -22.287  -7.592  18.770 127.261 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1263.64     318.22   3.971 0.000605 ***
X5             44.63      12.08   3.695 0.001197 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.36 on 23 degrees of freedom
Multiple R-squared:  0.3724,	Adjusted R-squared:  0.3451 
F-statistic: 13.65 on 1 and 23 DF,  p-value: 0.001197
```

This model represents a regression relationship between Y1(Rest frame Equivalent Width) and X5(Absolute Magnitude).
The intercept of this model is 1263.64.
Based on R-squared, we can say, 37.24% of the variability in weight is explained by the model. Since it represents coefficient of determination, and the R2 value is higher than 0.70, we consider it as a good model.

2c)
Model 5 (Y1: Rest frame Equivalent Width ~ X5: Absolute Magnitude) is the best models.
It has relatively the highest R-squared value among all models and has relatively lower p-value. Thus, we consider it as the best models among all the others.