DSC 423: Data Analysis and Regression

Assignment 2:

Name: Adarsh Shankar

Student ID: 2117611

Honor Statement: "I have completed this work independently. The solutions given are entirely my own work."

1. Short Essay (20 pts.) For each of these questions, your audience are persons that are not experts in statistics. Write with complete sentences and paragraphs. Cite any references that you use.

a. (10 pts.) Imagine you fit a regression model to a dataset and find that R-squared = 0.69. Is this a good regression model or not? If you cannot tell, what additional information do you need? Explain.

Ans:

R-squared score of 69%, or 0.69, is considered to be moderate. Any number above 0.7 is regarded favourably, the best models have values over 0.9, and models with R-squared values below 0.5 are despised. R-squared ranges from 0% to 100% always: 0% means that no variability in the response data around the mean is explained by the model. 100% means that all of the variability in the response data around the mean is explained by the model. The better the model matches your data, in general, the greater the R-squared. Values above 0.5, on the other hand, are considered to indicate a significant or strong connection. To determine what R-squared values are typically used in your field of study, you'll need to do some research.

b. (10 pts.) Research and then explain the "regression fallacy". Provide at least one example.

Ans:

The regression (or regressive) fallacy is a logical fallacy in which regression towards the mean is perceived as being caused by a particular cause rather than as a normal fluctuation. It is often a particular variation of the post hoc fallacy. {Academickids.com is the source.} The regression fallacy, to put it simply, is when someone assumes they have fixed or improved something that had a flaw in it without offering a convincing justification. Citation: (https://academickids.com/encyclopedia/index.php/Regressive_fallacy) For instance, the nation's GDP increased after the new President was elected. The public praised this audacious move. Despite the fact that there is no logical connection between individuals who believe it to be true, this is a fallacy.

2. QUASAR (30 pts.) -- A quasar is a distant celestial object (at least four billion light-years away) that provides a powerful source of radio energy. The Astronomical Journal (July 1995) reported on a study of 90 quasars detected by a deep space survey. The survey enabled astronomers to measure several different quantitative characteristics of each quasar, including:

X1 - Redshift

X2 - Line Flux

X3 - Line Luminosity

X4 - AB1450 Magnitude

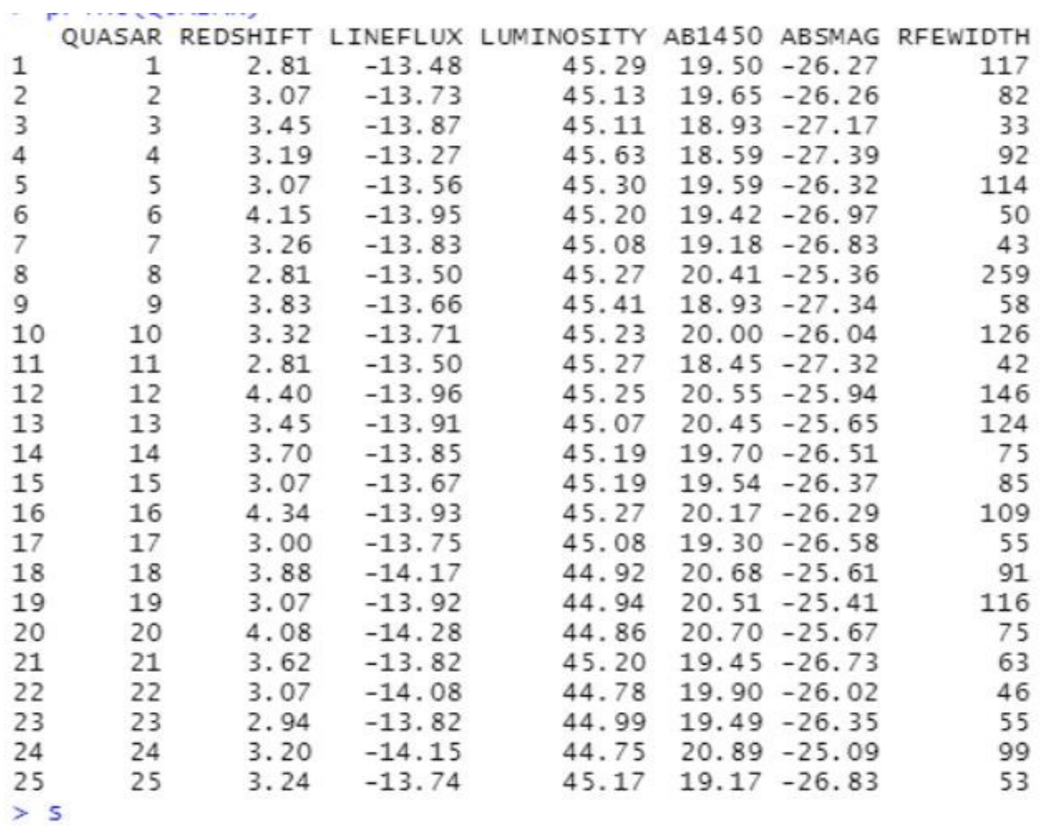
X5 - Absolute Magnitude

Y1 - Rest frame Equivalent Width

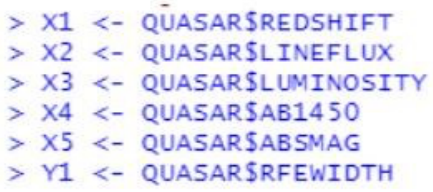
a). (10 pts.) Use R to perform a regression analysis on the QUASAR dataset (found on the D2L). For each of the explanatory variables create a regression model and copy/paste it into your submission.

Ans:

QUASAR = read. delim("C: /users/Adarsh/Downloads/QUASAR. txt")
> print (QUASAR)



```
> print(QUASAR)
   QUASAR REDSHIFT LINEFLUX LUMINOSITY AB1450 ABSMAG RFEWIDTH
1       1     2.81   -13.48      45.29  19.50 -26.27      117
2       2     3.07   -13.73      45.13  19.65 -26.26       82
3       3     3.45   -13.87      45.11  18.93 -27.17       33
4       4     3.19   -13.27      45.63  18.59 -27.39       92
5       5     3.07   -13.56      45.30  19.59 -26.32      114
6       6     4.15   -13.95      45.20  19.42 -26.97       50
7       7     3.26   -13.83      45.08  19.18 -26.83       43
8       8     2.81   -13.50      45.27  20.41 -25.36      259
9       9     3.83   -13.66      45.41  18.93 -27.34       58
10     10     3.32   -13.71      45.23  20.00 -26.04      126
11     11     2.81   -13.50      45.27  18.45 -27.32       42
12     12     4.40   -13.96      45.25  20.55 -25.94      146
13     13     3.45   -13.91      45.07  20.45 -25.65      124
14     14     3.70   -13.85      45.19  19.70 -26.51       75
15     15     3.07   -13.67      45.19  19.54 -26.37       85
16     16     4.34   -13.93      45.27  20.17 -26.29      109
17     17     3.00   -13.75      45.08  19.30 -26.58       55
18     18     3.88   -14.17      44.92  20.68 -25.61       91
19     19     3.07   -13.92      44.94  20.51 -25.41      116
20     20     4.08   -14.28      44.86  20.70 -25.67       75
21     21     3.62   -13.82      45.20  19.45 -26.73       63
22     22     3.07   -14.08      44.78  19.90 -26.02       46
23     23     2.94   -13.82      44.99  19.49 -26.35       55
24     24     3.20   -14.15      44.75  20.89 -25.09       99
25     25     3.24   -13.74      45.17  19.17 -26.83       53
> s
```



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

Rest frame Equivalent Width and REDSHIFT

model1 <- lm(Y1 ~ X1)

summary(model1)

Rest frame Equivalent Width and Line Flux

model2 <- lm(Y1 ~ X2)

summary(model2)

Rest frame Equivalent Width and Line Luminosity

model3 <- lm(Y1 ~ X3)

summary(model3)

Rest frame Equivalent Width and AB1450 Magnitude

model4 <- lm(Y1 ~ X4)

summary(model4)

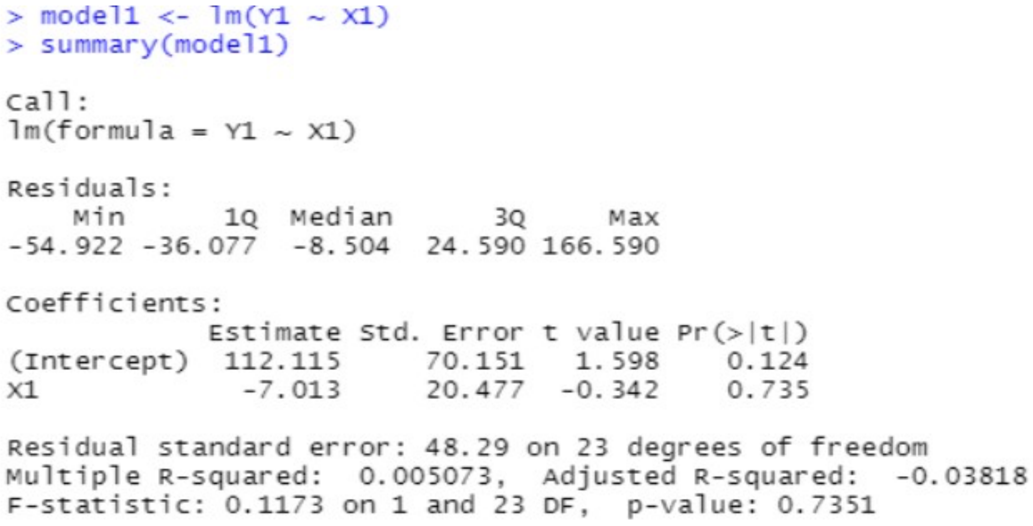
Rest frame Equivalent Width and Absolute Magnitude

model5 <- lm(Y1 ~ X5)

summary(model5)

b). (10 pts.) Evaluate your models. For each discuss how well they predict the dependent variable. Your description should begin by reporting basic facts about your model; but should also include an analysis of the findings.

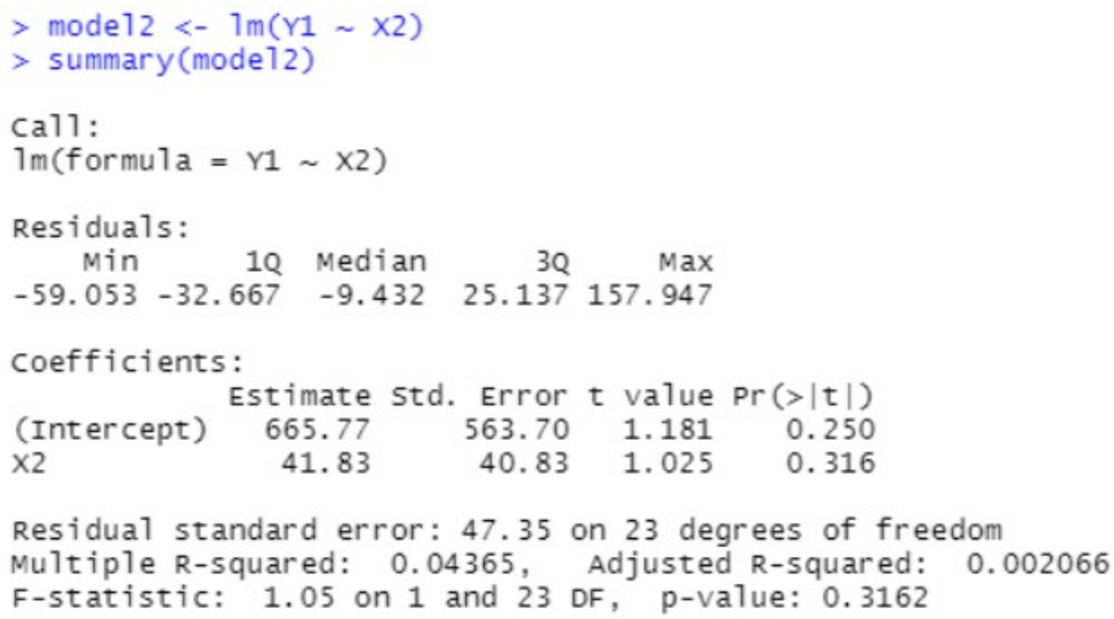
Rest frame Equivalent Width and REDSHIFT



```
> 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's intercept is 112.115, and its R-squared value is 0.5073%, meaning that only that much of the weight variability is explained. This model's R square value is low, making it ineffective.

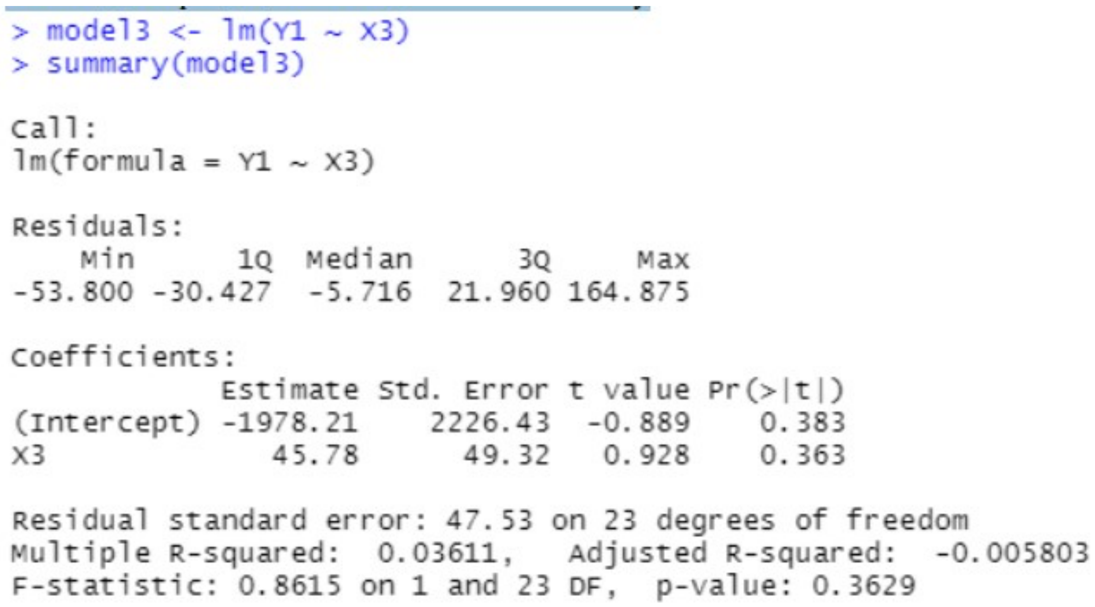
Rest frame Equivalent Width and Line Flux



```
> 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
```

Only that much of the weight variability is explained by this model, which has an intercept of 665.77 and an R-squared value of 4.365%. This model's R square value is low, making it ineffective.

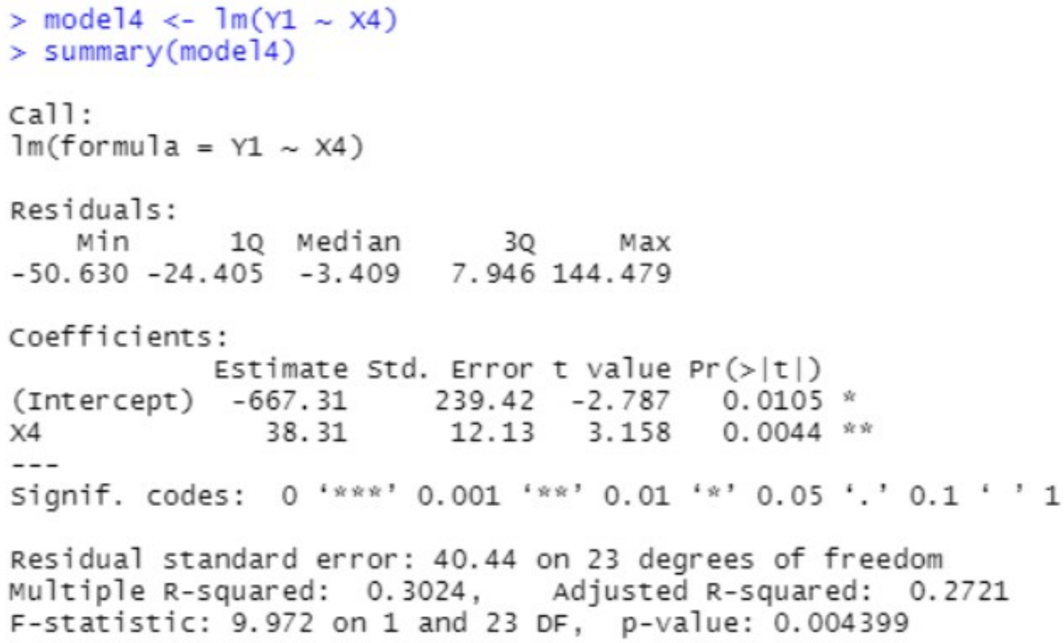
Rest frame Equivalent Width and Line Luminosity



```
> 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's intercept is -1978.21, and its R-squared value is 3.611%, meaning that only that much of the weight variability is explained. This model's R square value is low, making it ineffective.

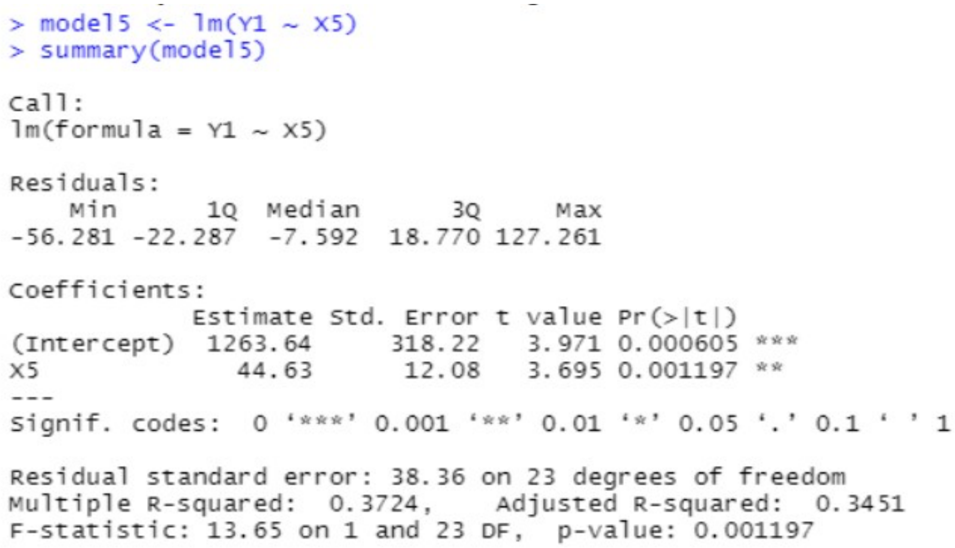
Rest frame Equivalent Width and AB1450 Magnitude



```
> 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
```

The intercept of this model is -667.31 and the R-squared value is 30.24%, only that much of variability of weight is explained. Since the R square value is 30.24% this model is effective.

Rest frame Equivalent Width and Absolute Magnitude



```
> 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
```

Only this much of the weight variability is explained by the R-squared value of 37.24% and the intercept of 1263.64. This model is regarded as the most effective because it has the highest R square value when compared to all the others.

c). (10 pts.) Of the models you built, what is the "best" model? Explain. Assume your audience is a fellow DSC423 student

pause frame In comparison to the previous models, Equivalent Width and Absolute Magnitude (Model 5) had the lowest p-value and highest R-squared value.

The biggest amount of weight variability can be explained by this model. As a result, Model 5 is regarded as the best.