DSC 423
Assignment 1

Based on Prerequisites and Modules 1 and 2

Ronaldlee Ejalu

Student ID: 2020637

I have completed this work independently and the solutions given are entirely my own work.
You submission must be submitted as a PDF.

1. Short Essay (10 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. (5 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?

      Yes, it a good regression model since 69% of the variability in the dependent variable is explained by our model.

      R-squared will always take on values between 0 and 1.

      R-squared less than 0.5 means moving towards underfitting of the model.

      R-squared greater than 0.9 means moving towards Overfitting of the model so a range of R-squared between 0.6 to 0.9 means that our model can explain a good amount of variance in the dependent variable which is predicted from any of the available independent variable in our dataset.

      If you cannot tell, what additional information do you need? Explain.

      If I can not tell, look at the two different models and compare their R-squared to determine which one is giving the best information.
   b. (5 pts.) Research and then explain the "regression fallacy". Provide at least one example.

      Regression fallacy occurs when one mistakes regression to the mean, which is a statistical phenomenon, for a causal relationship. The frequency of accidents on I-94 West fell after speed cameras were installed. Therefore, the speed cameras have improved road safety. Speed Cameras are installed after a road incurs an exceptionally high number of accidents, and this value usually falls (regression to mean) immediately afterwards. Many speed camera proponents attribute this fall in accidents to the speed camera, without observing the overall trend.

2. Short Essay (5 pts.) Consider the following two scenarios. A) take a simple random sample of 100 graduate students at DePaul university and b) take a simple random sample of 100 graduate students

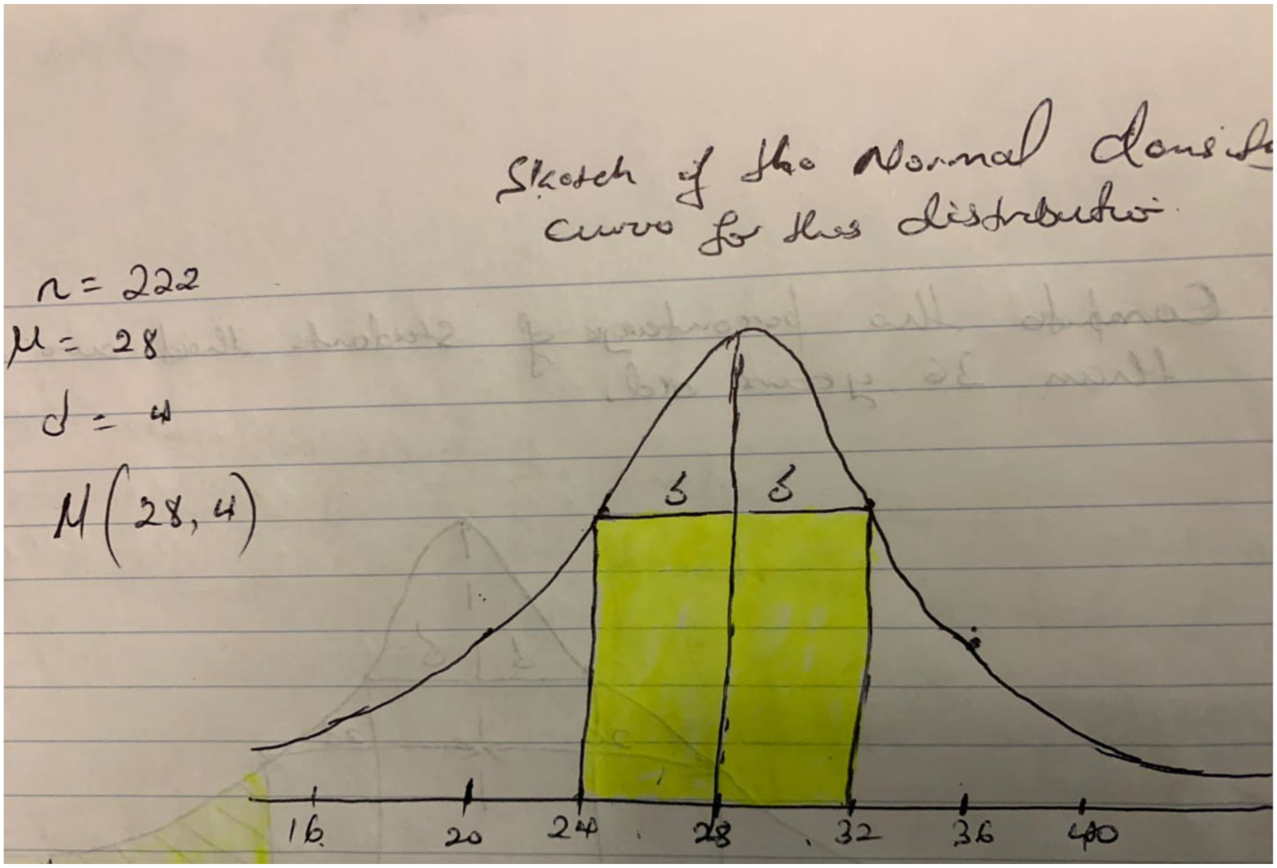
studying Data Science. For each sample you record the amount spent on textbooks used for classes. Which sample do you expect to have the smaller standard deviation? Explain your answer.

In a normal distribution curve, it is known that the probability distribution function is inversely proportional to the standard deviation. This means that the more the standard deviation, the less will be the probability implying the normal distribution curve will be flatter or the less the standard deviation, the more the probability will be.

In scenario b, there is more probability of using books thus making the standard deviation small; hence scenario b will have small standard deviation.

3. Empirical rule (10 pts.) The 222 students enrolled in online-learning courses offered by a college ranged from 18 to 64 years of age. The mean age was 28 with standard deviation equal to 4. Use the 68-95-99.7 rule to answer the following questions:
   a. (5 pts.) Compute the percentage of students that are between 24 and 32 years old. Show your work.

      N = 222
      μ=28
      Standard deviation = 4
      N(28, 4)



Sketch of the Normal densit
curve for this distribution

$n = 222$

$\mu = 28$

$\sigma = 4$

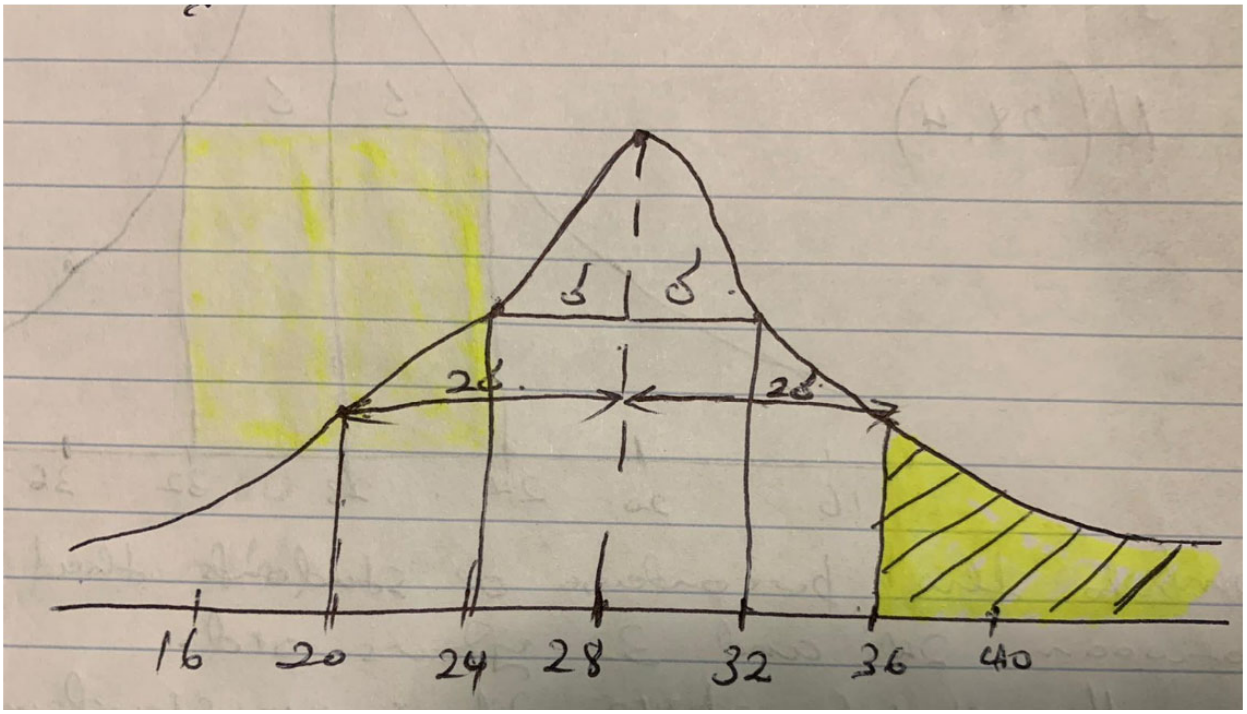
$N(28, 4)$

We also know from the empirical rule that 68 percent of observations are within one standard deviation, so the green area is 68 percent.

Therefore, 68 percent of the Students are between 24 and 32 years old.

b. (5 pts.) Compute the percentage of students that are older than 36 years. Show yourwork.

We know from the empirical rule that 95 percent of the observations are within two standard deviations below and two standard deviations above. So this area under the curve accounts for 95 percent of the data as shown below:



Remember, the area under the curve is always one and the area of the entire curve is 100 percent, that leaves 5 percent left over (100 - 95) = 5

We also know that the 5 percent must be in tails.

We know that the distribution is symmetric so the five percent must be evenly distributed between two tails.

$$5\,/\,2 = 2 \cdot 5\%$$

So, 2.5 percent of the students are older than 36 years old.

4. Z-scores (5 pts.) Monthly sale figures for a particular e-retailer tend to be normally distributed with mean equal to 150 thousand dollars and a standard deviation of 35 thousand dollars. Use the normal distribution to determine the top 1% monthly sale figure (a.k.a. 99th percentile)? Show your work.

$$\rho = 0.99$$

We find the z-scores through the z- score table

Z=2.33

We know that $z = \frac{(x-\mu)}{\delta}$

We interested in finding x, which is the top 1 percent month sales

$$z\sigma = x - \mu$$

$$x = \mu + z\delta$$

$$= 150 + 2.33 * 35 = 231550$$

The top 1 percent monthly sales is $231,550

5. Hypothesis Testing (10 pts.) A network provider investigated the number of blocked intrusions to its network, and found that there were, on average, 45 blocked intrusions per day. After a change in firewall settings, the mean number of intrusions during the next 35 days was 42 with a standard deviation equal to 15.5. Perform a hypothesis test to determine if the change in firewall settings reduced the number of intrusions. Show your work.

   Because the population standard deviation is unknown , use of the z-test and z-critical values based on the Central limit theorem is not possible so we use test static t and the t- critical values.

   Step number 1:

   $H_{0\,:}$ μ = 45

   H_a: $_:$ μ = 45

   Test statistic:

$$t = \frac{(\bar{y}-\mu_0)}{\left({}^{s}/_{\sqrt{n}}\right)}$$

$$t = \frac{(42-45)}{\left(\frac{15\cdot 5}{\sqrt{35}}\right)} = -1.145$$

$$t = -1 \cdot 145$$

The critical value as the significant level is unknown so we take the significant level as

Alpha (α) = 0.05

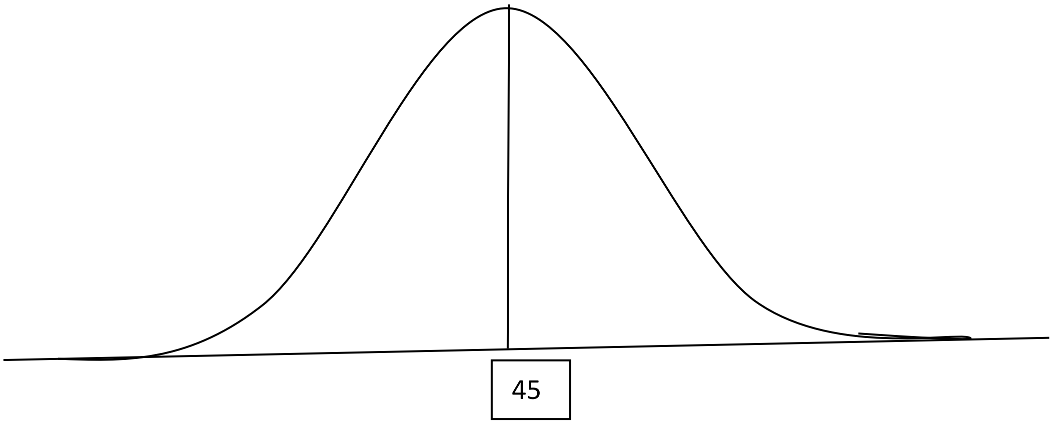
Df(degrees of freedom) = n – 1

$df = 35 - 1$ = 34

Since it is left tail test:

Critical value of t using t table we check value at df = 34

and $\alpha = 0.05$. $t_{0.05,34} = -1.69$



So the critical value of t = -1.69

Our test criteria is:

Reject the null hypothesis if t < -t(0.05, 34)

The test statistic t = -1.145

Using the test criteria above:

-1.145 > 1.69 therefore we cannot reject the null hypothesis.

Since we cannot reject the null hypothesis, there is not sufficient evidence to conclude that change in firewall settings reduced the number of intrusions.

6. QUASAR (10 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

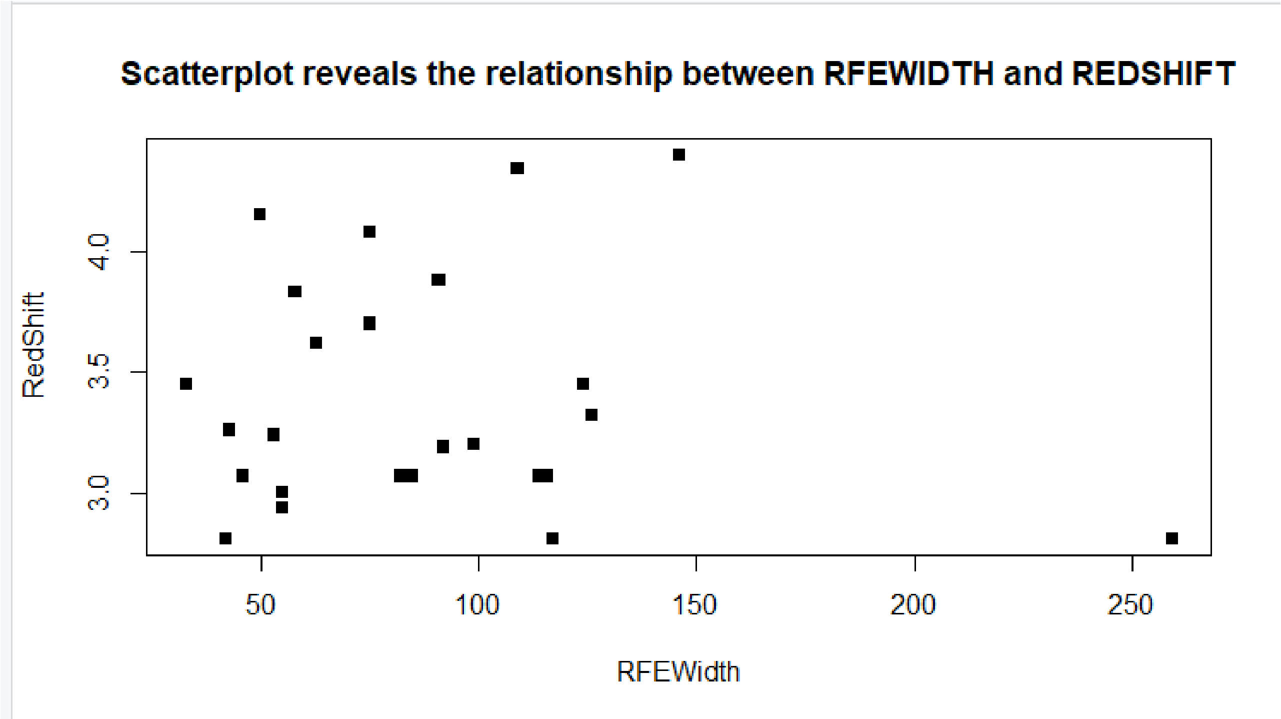
X4 - AB1450 Magnitude

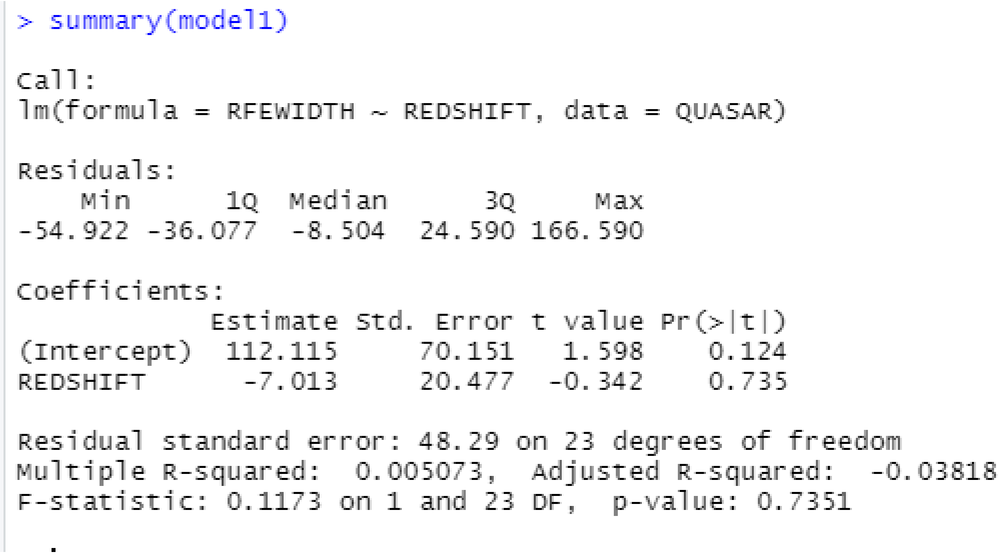
X5 - Absolute Magnitude

Y1 - Rest frame Equivalent Width

a. (5 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.
b. (5 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. What is the best model? Assume your audience is a fellow DSC423 student

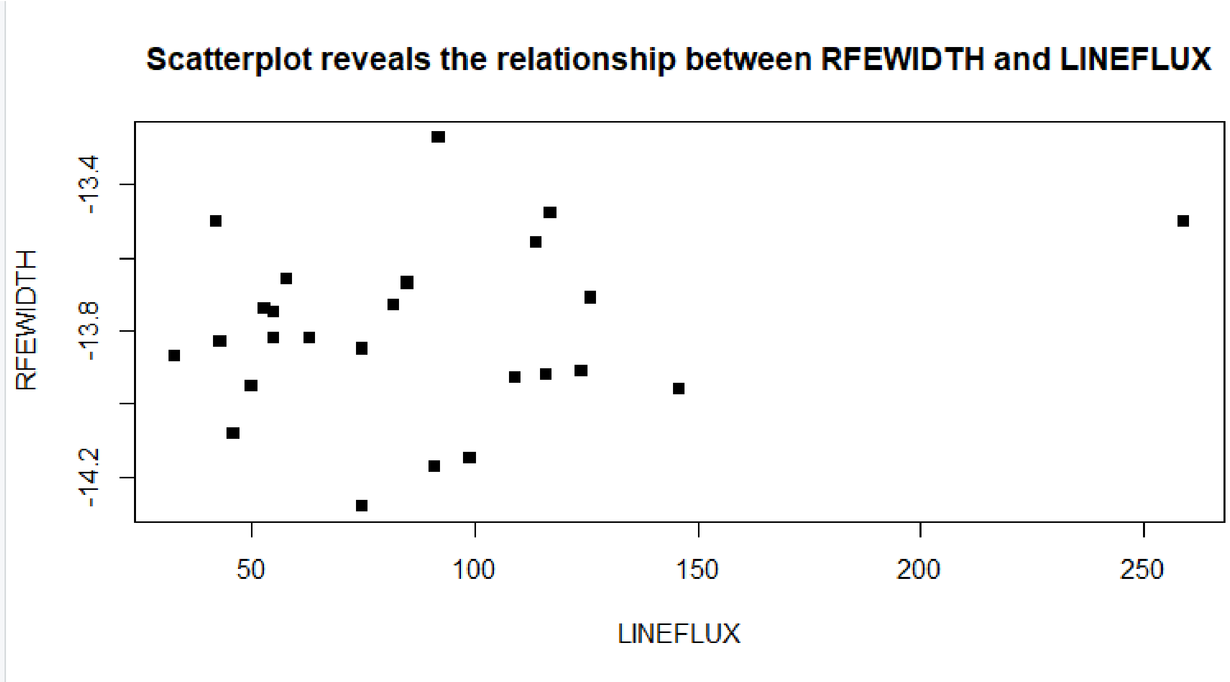
For X1-RedShift and Y1

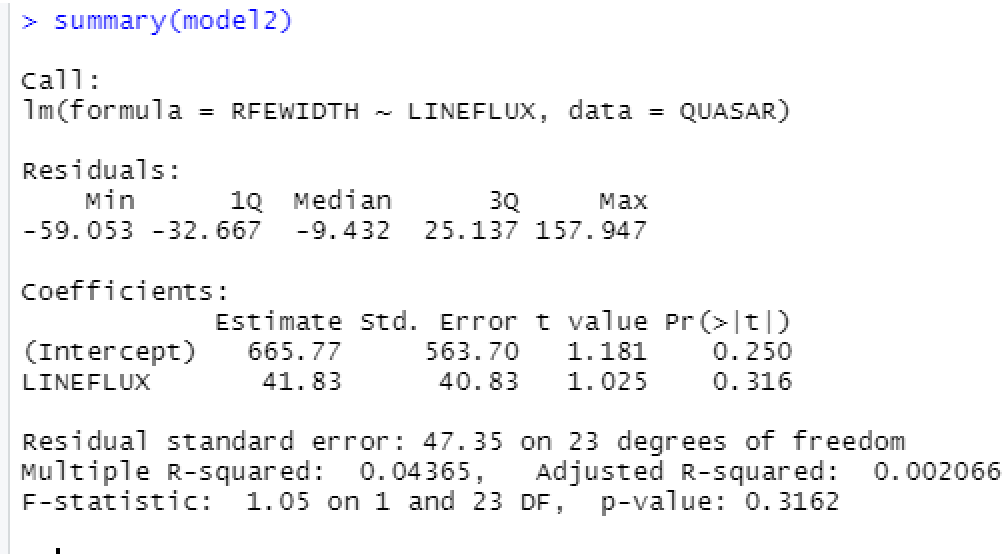




```
> summary(model1)

Call:
lm(formula = RFEWIDTH ~ REDSHIFT, data = QUASAR)

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
REDSHIFT      -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

  .
```

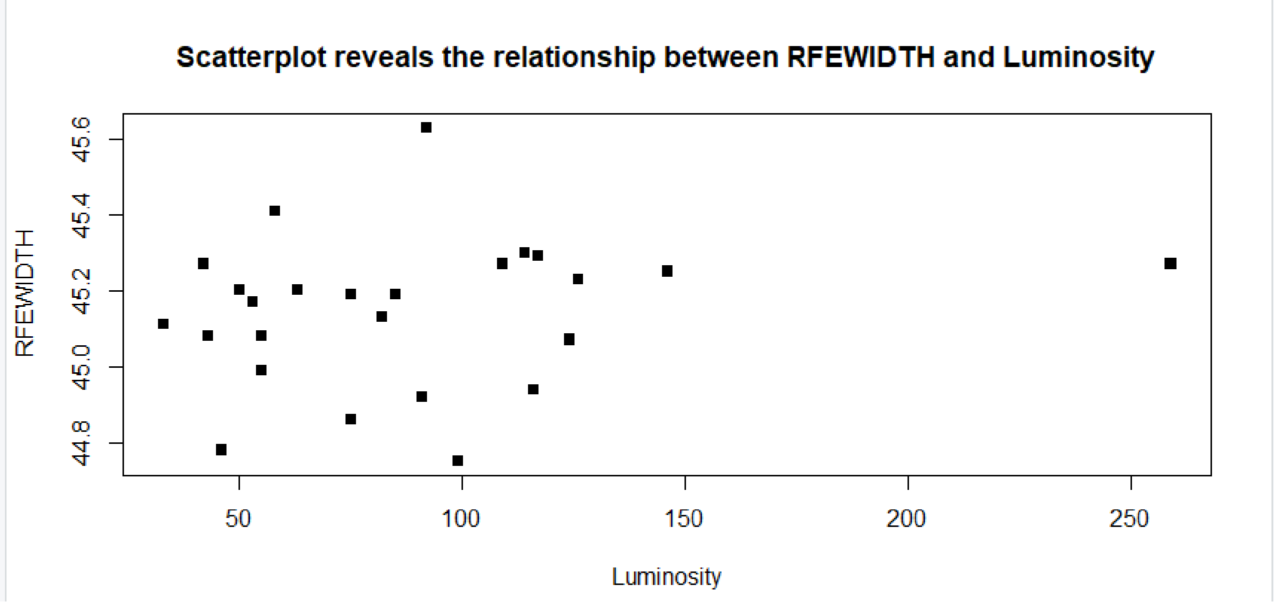
In this model, the p-values are out of range since they are bigger than 0.05. Therefore, we fail to reject the null hypothesis, that beta one is equal to zero. Adjusted R-Square is pretty low with -3.8 percent, so this is not a good model to be used to explain the variability of the dependent variable. The t-values are also greater than 0.05 therefore we are unable to reject the null hypothesis, that the beta associated with the independent variable, RedShift, equals zero. We are in effect saying, we don't know if the independent variable, REDSHIFT, should be used to predict the dependent variable, RFEWIDTH.

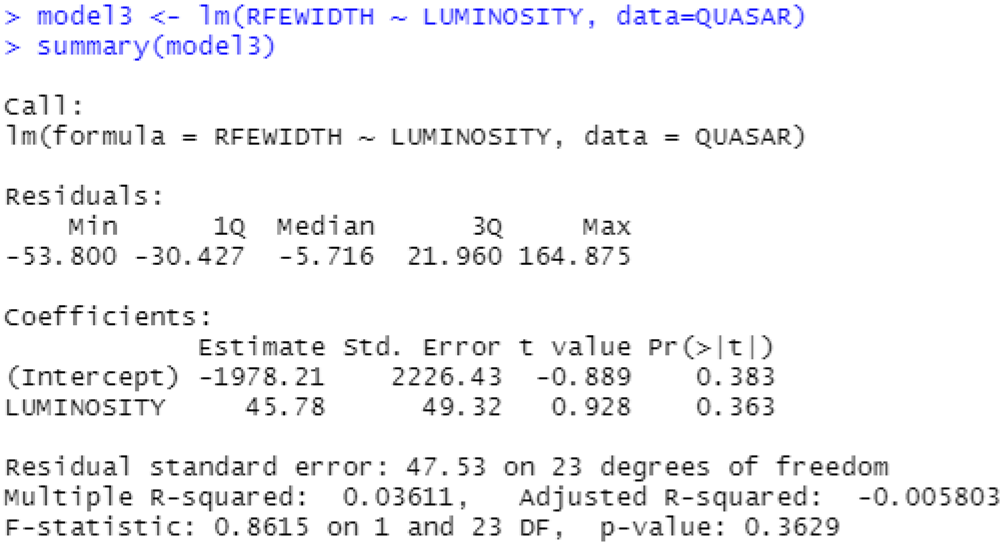




```
> summary(model2)

Call:
lm(formula = RFEWIDTH ~ LINEFLUX, data = QUASAR)

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
LINEFLUX       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
```

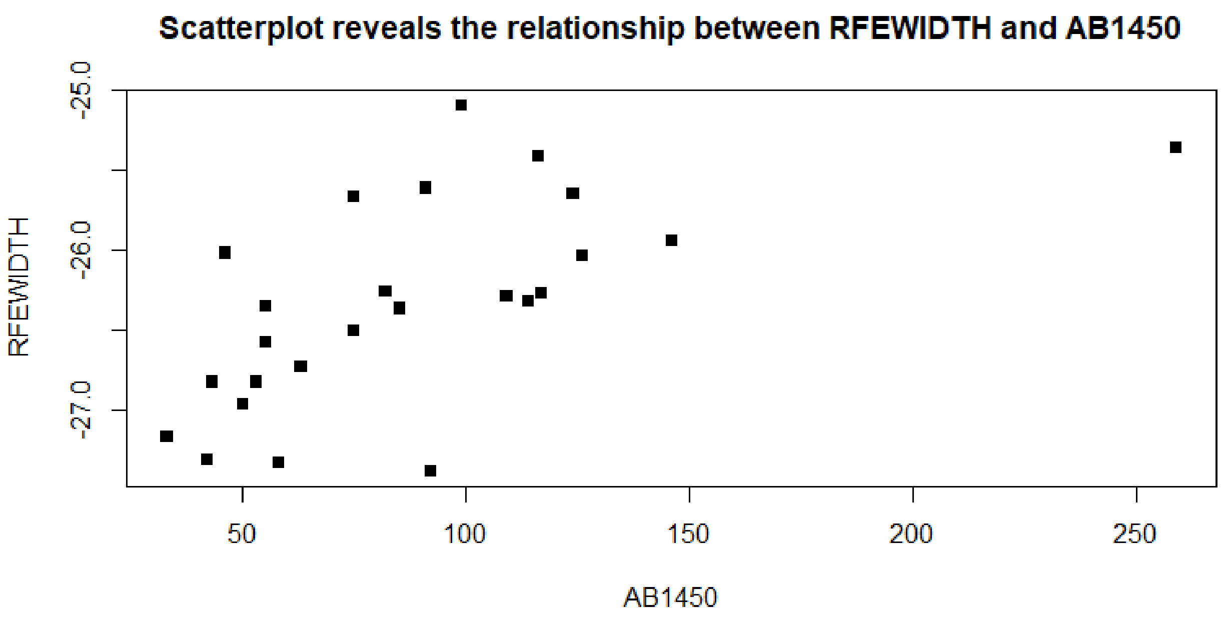
Still in model2 above, the p-values are out of range since they are greater than 0.05. Therefore, we fail to reject the null hypothesis, that beta one is equal to zero. Adjusted R-Square is pretty low with 0.002 percent, so this is not a good model to be used to explain the variability of the dependent variable. The t-values are also greater than 0.05 therefore we are unable to reject the null hypothesis, that the beta associated with the independent variable, LINEFLUX, equals zero. We are in effect saying, we don't know if the independent variable, LINEFLUX, should be used to predict the dependent variable, RFEWIDTH.

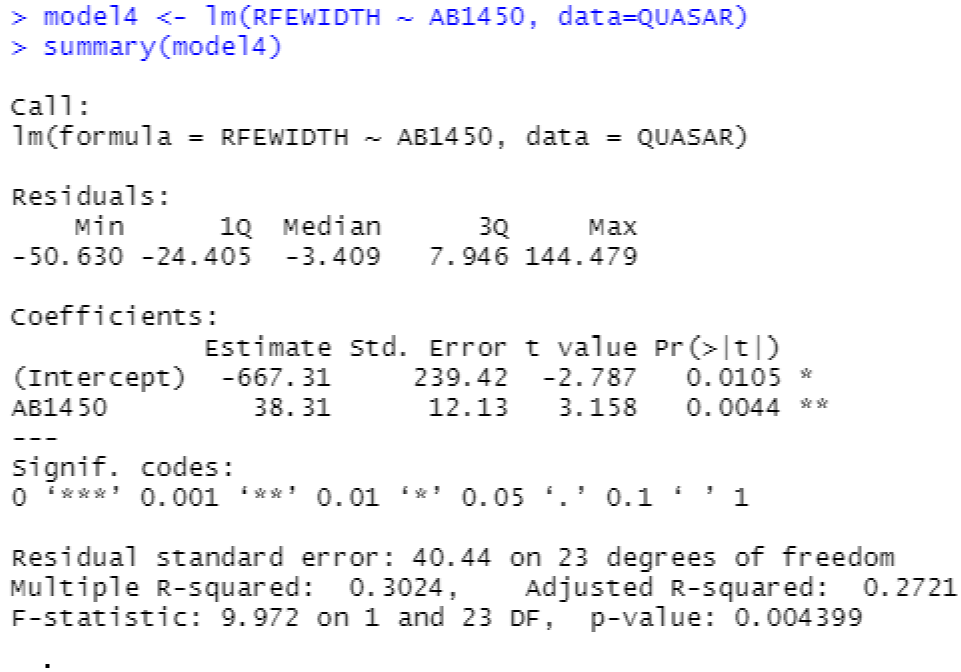




```
> model3 <- lm(RFEWIDTH ~ LUMINOSITY, data=QUASAR)
> summary(model3)

Call:
lm(formula = RFEWIDTH ~ LUMINOSITY, data = QUASAR)

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
LUMINOSITY     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
```

Still in model3 above, the p-values are out of range since they are greater than 0.05. Therefore, we fail to reject the null hypothesis, that beta one is equal to zero. Adjusted R-Square is pretty low with -0.58 percent, so this is not a good model to be used to explain the variability of the dependent variable. The t-values are also greater than 0.05 therefore we are unable to reject the null hypothesis, that the beta associated with the independent variable, LUMINOSITY, equals zero. We are in effect saying, we don't know if the independent variable, LUMINOSITY, should be used to predict the dependent variable, RFEWIDTH.





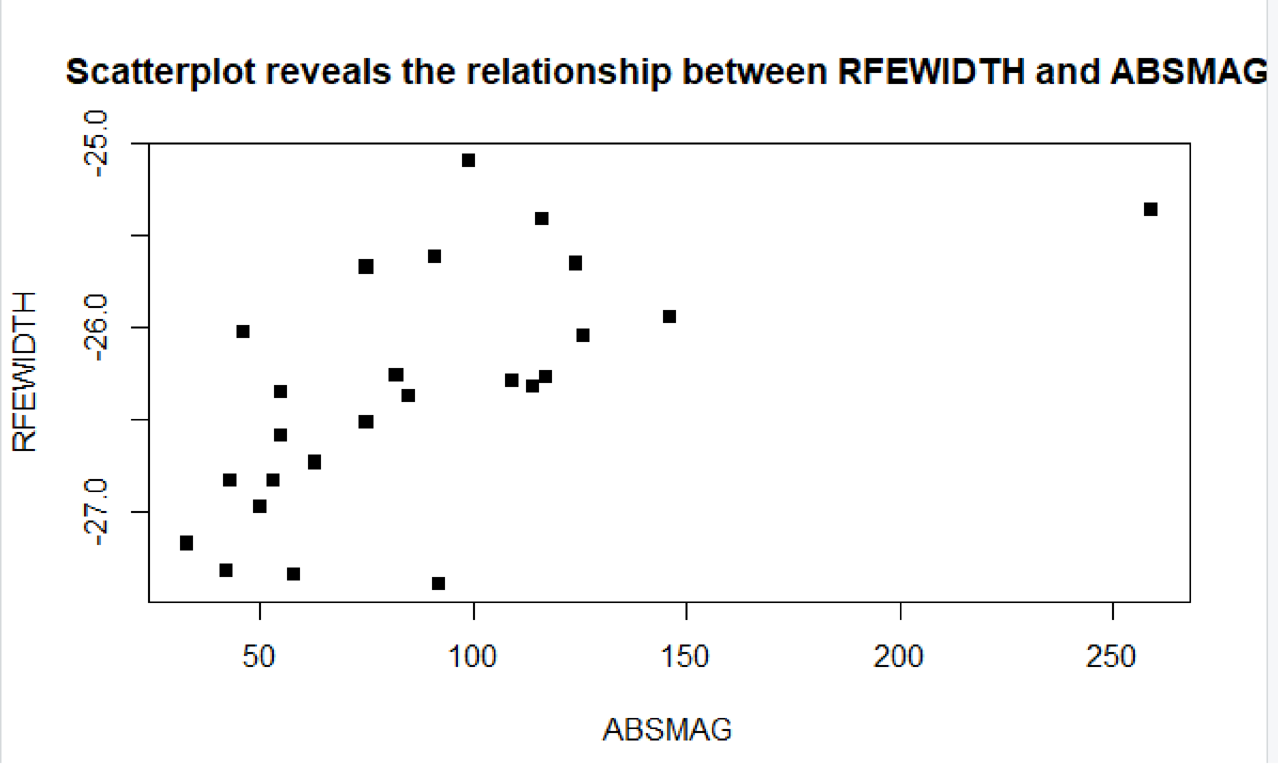
```
> model4 <- lm(RFEWIDTH ~ AB1450, data=QUASAR)
> summary(model4)

Call:
lm(formula = RFEWIDTH ~ AB1450, data = QUASAR)

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 * 
AB1450         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
```

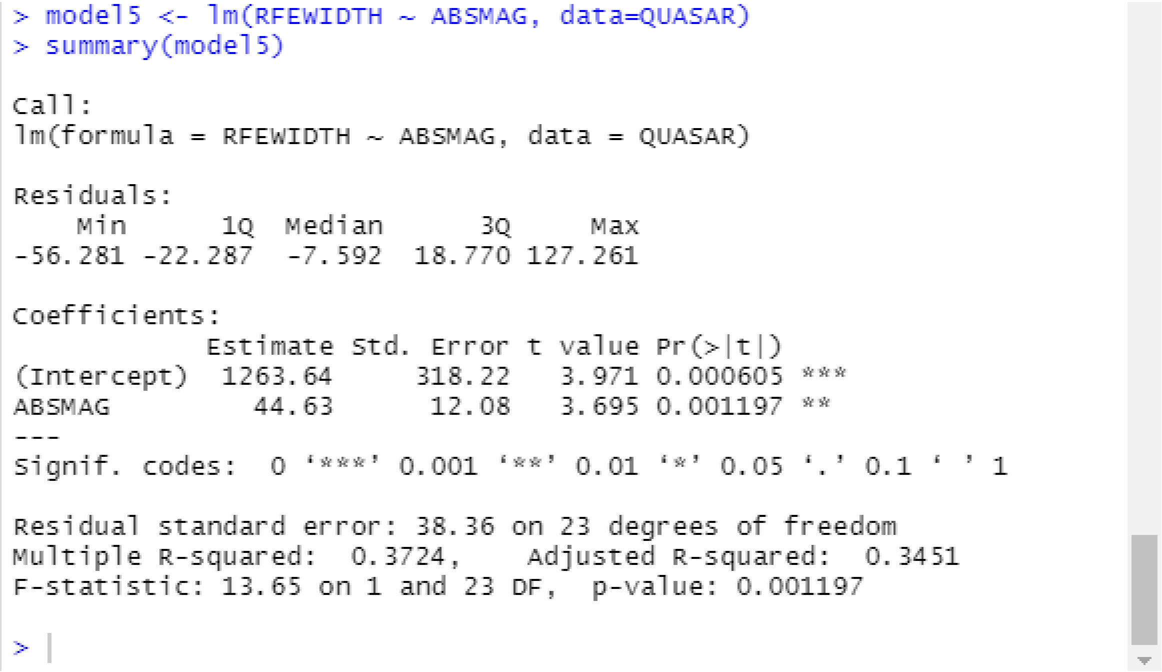
.

In model4 above, the p-values is 0.4 percent. That means, if our null hypothesis is true, beta one should be equal to zero, then there is only a 0.4 percent chance that we would observe this outcome. This seems quite unlikely, so we are going to reject our null hypothesis that beta one is equal to zero and accept the alternative the beta one doesn’t equal to zero.

Adjusted R-squared of 27 percent looks goods when compared to the previous models, 1, 2 and 3 though it is below 50 fifty percent.

The t-values look good so we reject the null hypothesis that the beta associated with the independent variable, AB1450, equals zero and accept the alternative that beta one is not equal to zero, so we go ahead use the estimation of 38.31.





```
> model5 <- lm(RFEWIDTH ~ ABSMAG, data=QUASAR)
> summary(model5)

Call:
lm(formula = RFEWIDTH ~ ABSMAG, data = QUASAR)

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 ***
ABSMAG         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

> 
```

In model5 above, the p-values is 0.11 percent. That means, if our null hypothesis is true, beta one should be equal to zero, then there is only a 0.11 percent chance that we would observe this outcome. This seems quite unlikely, so we are going to reject our null hypothesis that beta one is equal to zero and accept the alternative the beta one doesn't equal to zero.

Adjusted R-squared of 34.5 percent looks goods when compared to the previous models, 1, 2, 3 and 4 though it is still below 50 fifty percent. The adjusted R-Squared went up to 34.5 percent. 34 percent of the variability in RFEWIDTH is explained by the model.

The t-values look good so we reject the null hypothesis that the beta associated with the independent variable, ABSMAG, equals zero and accept the alternative that beta one is not equal to zero, so we go ahead use the estimation of 44.63.

Model 5 is the best model since it has the biggest Adjusted R-Squared of 34.5 percent 34.5 percent of the availability in RFEWIDTH is explained by the model.