Patrick Keener
ID#: 1385832

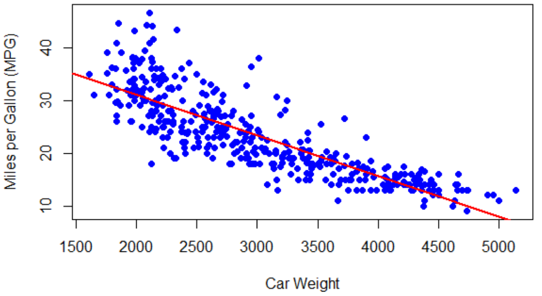
Assignment 1

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

1a.

The data was modeled using linear regression, which is a technique used to estimate relationships between two variables. Using this model, a predictor variable can be used as input and the model will output an estimate of the variable we are trying to predict. The model itself is designed to return a straight line that passes as close to every outcome as possible, as shown in figure 1[1], where the red line is the regression line (model output).

Figure 1. Relationship between MPG and Car Weight



The fact that the line runs as close through every outcome as possible is not by itself enough evidence to suggest a model should be used, since "as close as possible" may still not be very close. Additional insight comes from investigating two other factors: a measure of how much of the data is explained by the model, which is represented by R-squared, and a measure of the confidence that the variables are predictive, which is measured with the t-statistic.

The model has an R-squared of .69, which suggests that almost 70% of the variation seen in the data can be explained by the model. Unfortunately, no t-statistic was provided, which is necessary to determine whether the variable is reliable. Additionally, determination of whether a model is "good" is dependent on the context it is used in. Slightly more than 30% of the variation in the data is not accounted for in the model. In many circumstances that may be acceptable, however if the model was being used to help decide what kind of medical treatment to use it would not be.

1b.

The regression fallacy is a logical fallacy with statistical roots. If some events are random and normally distributed, then after a time at either extreme they should naturally revert towards the mean[2]. For example, in order for an athlete to be selected for the Sports Illustrated cover they have likely had an exceptionally good year- an outlier. It is only natural that the following year or years they should not do as well, however instead of this being seen as a normal progression of events it is instead attributed to the "Sports Illustrated cover jinx".

One notable case occurred in 1957 when the Oklahoma football team was featured. The next game they lost to Notre Dame; however, this was after an impressive 47-game winning streak[3]. That winning streak was almost certainly an outlier since, to this day, it is still the longest winning streak in NCAA D1 college football history.

Figure 2 Sports Illustrated, Jan 21, 2002

[1] Graph generated using the "Cars" dataset provided for the class.

[2] "Beware the Regression Fallacy", Psychology Today website, published Jan 04, 2014, written by Scott Lilienfeld https://www.psychologytoday.com/us/blog/the-skeptical-psychologist/201401/beware-the-regression-fallacy

[3] From the Wikipedia article on the Sports Illustrated Cover Jinx. https://en.wikipedia.org/wiki/Sports_Illustrated_cover_jinx

2.

There are a few things to consider: First, the Data Science ("DS") students are a sample of the graduate population. Second, we can not consider the DS sample representative, or their mean representative, of the graduate population.

In credit modeling we segment by a number of things: geographic location, risk of individual client, type of business, etc., and the difference in risk between segments can be dramatic (which is why we do the segmentation to begin with). A similar phenomenon occurs in graduate school, where the segmentation is along concentrations/majors.

Some concentrations, like Computational Finance, have notoriously expensive books. Other concentrations may have costs allocated elsewhere- for example, a fine arts student may have their costs weighted towards painting supplies or musical instruments. Additionally, the graduate population as a whole includes doctoral and potentially post-doc student who have either very expensive books or no books at all.

While the graduate student population as a whole is characterized by heterogeneity, the DS segment is almost by definition characterized by homogeneity. Therefore, the variance of textbook costs among the DS population should be lower than the variance of textbook costs among the graduate population as a whole.

3.

There are 222 students enrolled in online-learning, ages 18-64. The mean age 28 and the standard deviation is 4. Use 68-95-99.7 rule to answer the questions.

a) Find the percentage of students between 24 and 32 years old:
   1. The first step is to analyze the problem. We are looking at two values, 24 and 32, and are attempting to find the values in between. Additionally, the mean lies precisely in the middle of them, which means we are looking for a two-sided distribution.
   2. The next step is to determine how many standard deviations from the mean each side is so we can apply the rule. The formula to compute standard deviations (also known as the z-score) is: z-score $= \frac{x-\bar{x}}{\sigma}$. Using 24 and 32 to for x, 28 as x-bar (mean), and 4 for standard deviation we get: $\frac{24-28}{4}\ and \frac{32-28}{4}$, which return -1 and 1, respectively.
   3. Next, notice that 1 standard deviation of the mean corresponds to 68%.
   4. **Therefore, 68% of the students should be between the ages of 24 and 32.**

b) Compute the percentage of students older than 36:
   1. Similar to part a, begin by analyzing the problem. The "greater than" piece tells us that the solutions will be the distance between the value at 100%. Additionally, 36 is to the right of the mean, which means each probability will need to be divided by 2.
   2. Next, we determine the number of standard deviations from the mean. $\frac{36-28}{4} = 2$ standard deviations, which corresponds to 95%.
   3. Since we are looking for only those students older than 36, we are looking for the area to the right of 95%, which is easily computed as 100% - 95% = 5%.
   4. Finally, since we are only looking at students on one side of the distribution, we divide 5% by 2.
   5. **Therefore, approximately 2.5% of the students are older than 36.**

4.

Monthly sales figures have a mean of \$150k and standard deviation of \$35k; it is normally distributed. What is the 99th percentile monthly sales figure?

1. First, analyze the problem. The goal is to find the value corresponding to 99%, which means the solution will be the distance between 0% and 99%. Additionally, 99% is to the right of the mean (50%) so we can use a positive z-score chart to easily find the solution. It should be noted, though, that we are looking at a single-sided distribution. If looking at a negative z-score chart we would find the value corresponding to 1%, not .5% as we would for a two-sided distribution.
2. Next, find out how many standard deviations correspond to 99%. The z-score table I found shows that 99% of values are found between 0 and 2.33 standard deviations.
3. **3.** Multiple the standard deviation by 2.33 and add this to the mean, which results in a **99th percentile monthly sales figure of \$231,550**

5.

Average intrusions of 45 per day (blocked). Changed setting, new mean over 35 days was 42 blocks per day with a standard deviation of 15.5. Perform hypothesis test.

1. First, analyze the problem. The problem is hypothesis testing. The sample size for the new mean is 35.
2. Design the null hypothesis: Test the likelihood that 42 is a new distribution, therefore we will make the hypothesis that it is not.
    a. Null Hypothesis: H_0: $\mu = \mu_0$
3. Next, calculate the sample standard deviation, which is $\frac{15.5}{\sqrt{35}}$ = 2.62
4. Using this new standard deviation, we calculate the z-score: $\frac{45-42}{2.62} = 1.145$
5. Assuming a required α of 5%, we look for a z-score corresponding to 2.5% which, given the empirical rule, is 2 standard deviations.
6. **6. Since 1.145 standard deviations is less than 2 standard deviations, we fail to reject the null hypothesis.**

6.

a. Regression analysis- regression models

Red Shift

```
Call:
lm(formula = q$RFEWIDTH ~ q$REDSHIFT)

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
q$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
```

Line Flux

```
Call:
lm(formula = q$RFEWIDTH ~ q$LINEFLUX)

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
q$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
```

Luminosity

```
Call:

lm(formula = q$RFEWIDTH ~ q$LUMINOSITY)

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
q$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
```

AB1450 Magnitude

```
Call:
lm(formula = q$RFEWIDTH ~ q$AB1450)

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 * 
q$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
```

Absolute Magnitude

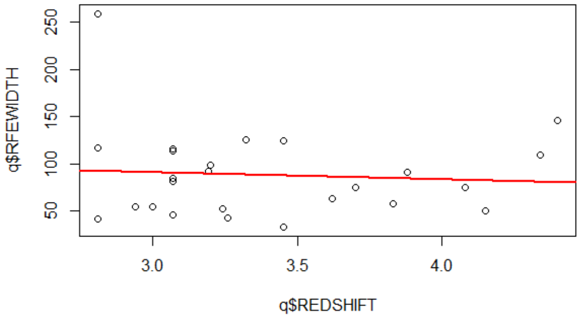
```
Call:
lm(formula = q$RFEWIDTH ~ q$ABSMAG)

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 ***
q$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
```

b. Evaluate the models

Red Shift – Red shift is a poor explanatory variable. Neither the intercept nor slope pass the t-test; the intercept with a .124 and the slope with a whopping .735. Additionally, the R-squared is .005073, so there is virtually zero explanatory power. Finally, the p-value is .7351 vs an allowed value of <.05, which is far outside of tolerance.

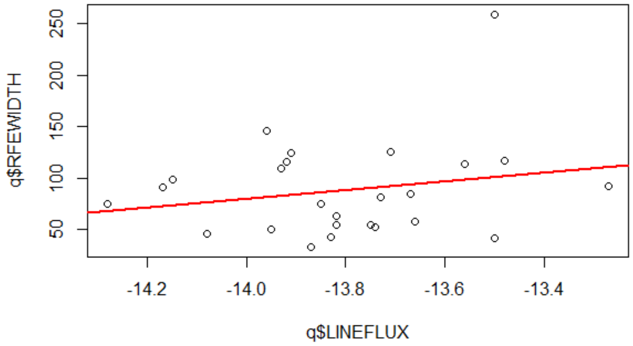
	Estimate	Pr(>|t|)
Intercept (ß$_0$)	ß$_0$ = 112.115	.124
RedShift (ß$_1$)	ß$_1$ = -7.013	.735

R^2 = .005073 p-value: .7351



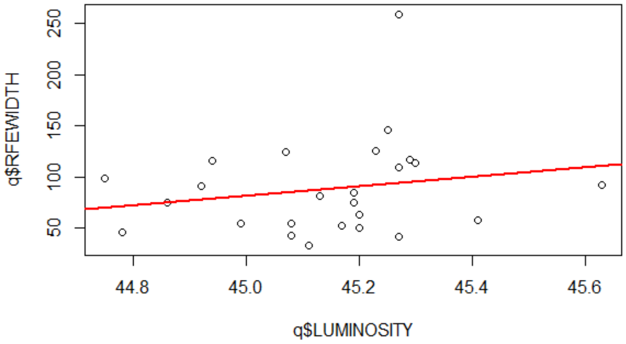
Line Flux – Line Flux is not an acceptable explanatory variable since the neither the slope or intercept pass the t-test (probabilities of .315 and .25, respectively). The R^2 implies that only 4.365% of the variation is explained by the variable, and last the p-value is .3162, above the threshold of .05.

	Estimate	Pr(>|t|)
Intercept (ß$_0$)	ß$_0$ = 665.77	.25
RedShift (ß$_1$)	ß$_1$ = 41.83	.316
R^2 = .04365		p-value: .3162



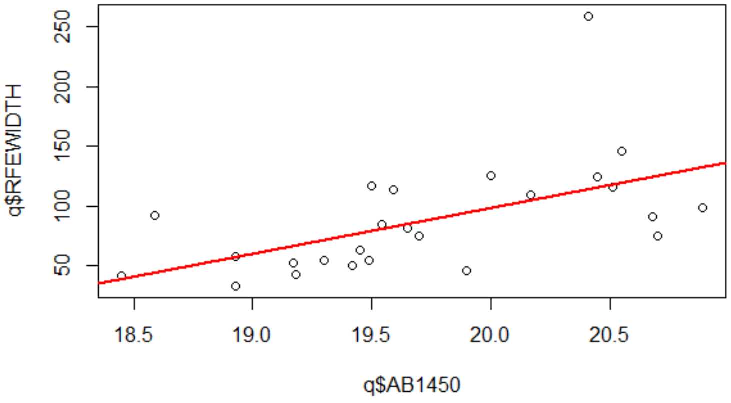
Luminosity – Luminosity does not meet the standards to be used as an explanatory variable. The t-test acceptance threshold is 5%; when used as a predictor variable, luminosity produces probabilities of .383 and .363 for the intercept and slope, respectively. Additionally, the R-squared is .03611, which implies virtually no predictive power, and the p-value is .3629 which is far above the threshold as well.

	Estimate	Pr(>|t|)
Intercept (ß$_0$)	ß$_0$ = -1978.21	.383
RedShift (ß$_1$)	ß$_1$ = 45.78	.363
R^2 = .03611		p-value: .3629



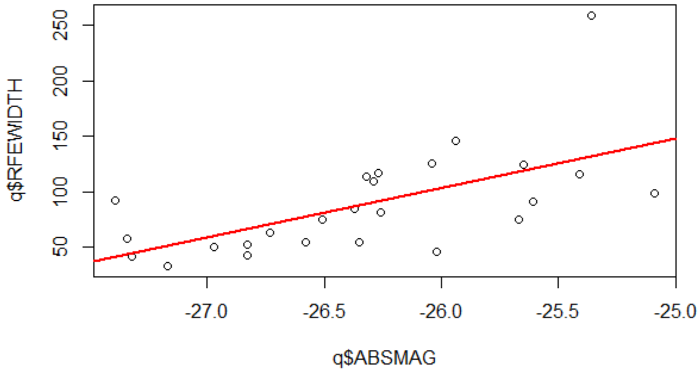
AB1450 Magnitude - AB1450 Magnitude has the potential to be used as an explanatory variable. It explains about 30% of the variation in the dataset (R^2 of .3024), and satisfies the t-test (.0105 and .0044 vs .05 for the intercept and slope, respectively). Finally, the p-value is .004399 which is within tolerance. The y-intercept is far below 0 at -667.31, and the slope itself is 38.31.

	Estimate	Pr(>|t|)
Intercept (ß$_0$)	ß$_0$ = -667.31	.0105
RedShift (ß$_1$)	ß$_1$ = 38.31	.0044
R^2 = .3024		p-value: .004399



Absolute Magnitude – Absolute magnitude has the potential to be used as an explanatory variable. The intercept and slope pass the t-test, at values of .000605 and .001197, respectively, vs a threshold of <.05. The variable explains about 37% of the variation in the data (R^2 of .3724), and the p-value is .001197, which is within tolerance.

	Estimate	Pr(>|t|)
Intercept (ß$_0$)	ß$_0$ = 1263.64	.000605
RedShift (ß$_1$)	ß$_1$ = 44.63	.001197
R^2 = .3724		p-value: .001197



Analysis

Of the five variables tested, only two passed the t-test: AB1450 magnitude and absolute magnitude. Of these, absolute magnitude had a higher R^2, .3724 (vs .3024). Looking into the variables further we find that the standard error for AB1450 is 12.13, and the standard error for absolute magnitude is 12.08. Finally, the residual standard error for AB1450 is 40.44, vs 38.36 for absolute magnitude. Overall, absolute magnitude is the better predictor variable.

Additionally, a model using both absolute magnitude and ab1450 was tested. All variables failed the t-test.