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.

R-squared us the coefficient of determination which is the proportion of the variance independent variable that is predicted from the independent variable.
R-squared = 0.69 means about 70% of variability is explained by the model.
To define whether the regression model with R-squared = 0.69 is a good model, we need to set up the main objective for the regression model first.
If the main objective for the regression model is to explain the relationship between response variable and independent variable, then the r-square would be irrelevant to define a good regression model.
However, if the main objective for the regression model is to predict the response variable by using independent variables, then the higher r-square means that the model could be more accurate.

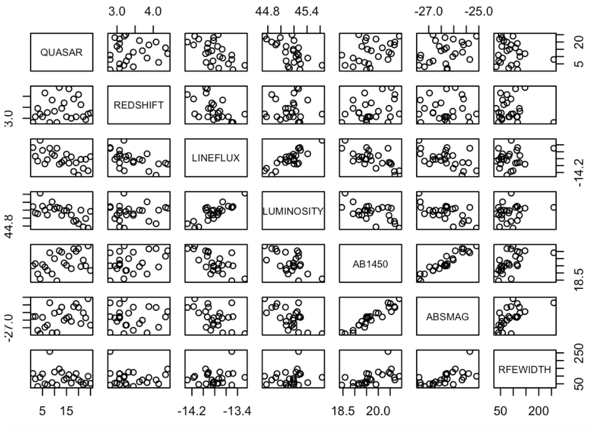
b. (10 pts.) Research and then explain the “regression fallacy”. Provide at least one example.
The regression (or regressive) fallacy is an informal fallacy. It assumes that something has returned to normal because of corrective actions taken while it was abnormal. This fails to account for natural fluctuations. It is frequently a special kind of the post hoc fallacy.(from Wiki)
If Peter had a fever the day before yesterday, then Peter took the medicine today and got better, so the reason that Peter got better is because he took the medicine today.

In reality, even if Peter did not take the medine, he could be getting better naturally, which means that we could know if Peter getting better is because of the effects of the medicine.

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.

plot(QUASAR)




```
> model_1 <- lm(RFEWIDTH~REDSHIFT, data = QUASAR)
> summary(model_1)

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
```

```
> model_2 <- lm(RFEWIDTH~LINEFLUX, data = QUASAR)
> summary(model_2)

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


> model_3 <- lm(RFEWIDTH~LUMINOSITY, data = QUASAR)
> summary(model_3)

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


> model_4 <- lm(RFEWIDTH~AB1450, data = QUASAR)
> summary(model_4)

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


> model_5 <- lm(RFEWIDTH~ABSMAG, data = QUASAR)
> summary(model_5)

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
```

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.

Model_1:
The best estimator for β1 slope is -7.013
The best estimator for β0 y-intercept is 112.115
T-value is -0.342
P-value is 0.735

R-squared: 0.005073

Form:linear
Strength : mild
Interaction: negative

R-squared 0.5% is quite low , only 0.5% in the variability is explained by the regression model.

Model_2:
The best estimator for β1 slope is   41.83
The best estimator for β0 y-intercept is 665.77
T-value is 1.025
P-value is 0.3162
R-squared: 0.04365

Form:linear
Strength : mild
Interaction: positive

R-squared 4% is quite low , only 4% in the variability is explained by regression model.

Model_3:

The best estimator for β1 slope is   45.78
The best estimator for β0 y-intercept is -1978.21
T-value is 0.928
P-value is 0.3629
R-squared: 0.03611

Form:linear
Strength : mild
Interaction: positive

R-squared 3.6% is quite low , only 3.6% in the variability is explained by regression model.

Model_4:
The best estimator for β1 slope is   38.31
The best estimator for β0 y-intercept is -667.31
T-value is 3.158
P-value is 0.004399

R-squared: 0.3024

Form:linear
Strength : mild
Interaction: positive

R-squared 30% is doable , 30% in the variability is explained by regression model.

Model_5:
The best estimator for β1 slope is   44.63
The best estimator for β0 y-intercept is 1263.64
T-value is 3.695
P-value is 0.001197

R-squared: 0.3724

Form:linear
Strength : mild
Interaction: positive

R-squared 37% is acceptable , 37% in the variability is explained by the regression model.

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

I would evaluate these five models by R - square since R-square represents the percent of variability in the dependent variable explained by the dependent variable.
The highest R-squared among 5 models is the fifth model whose R-squared is 37% which may be sufficient if there is extreme variability in the dataset.