Keiland Pullen DSC 423 | Spring 2021
ID:1977120

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

1a.)

Before examining an R-squared value of 0.69, several data science terms must be discussed. First, regression analysis is defined as a statistical technique that is used to find relationships between the dependent variable and independent variables in a given dataset. In simpler terms, regression it is a way to identify patterns or trends in data. With respect to data, the Australian Bureau of Statistics states that "variables are characteristics, numbers or quantity that can be measured or counted" (https://www.abs.gov.au/websitedbs/D3310114.nsf/home/statistical+language+-+what+are+variables) . Examples of data variables can be age, height, weight, income, area code, etc. A model can be defined as a representation of data. Linear regression is a type of model where there exists a linear relationship between a dependent variable and independent variables. Basically, linear regression can be used to predict the value of one variable based on the value of other variables. What does R-squared mean? According to the Corporate Finance Institute, R-Squared is defined as a statistical measure that is "the proportion of variance in the dependent variable that can be explained by the independent variable." (https://corporatefinanceinstitute.com/resources/knowledge/other/r-squared/) In short, R-squared displays whether or not the data is a good fit in the regression model. If presented with an R-squared value of 0.69, would this value be considered a good fit for the model? The answer really depends on the type of data and the prediction that we're trying to make. If we are trying to predict the number of males over six-feet tall in DePaul University's incoming freshmen class, then an R-squared value of 69% would be good. However, if we're trying to predict how well a heart medication will perform for a specified age group, then an R-squared value of 69% may not be sufficient. In this case, more independent variables may need to be included in the model.

1b.)

The web-site Logically Fallacious describes a regression fallacy as "Ascribing a cause where none exists in situations where natural fluctuations exist while failing to account for these natural fluctuations" (https://www.logicallyfallacious.com/logicalfallacies/Regression-Fallacy). This definition can be restated to describe a regression fallacy as a flawed logic that makes an assumption that something has improved or returned to its original state based on some action that has no real impact. For example, a Depaul student doe not study for a mid-term and receives a grade of "C-". For the final exam, the student studies extra hours every day and eats one magic apple every morning. This results in the student passing the final exam with a grade of "A+". The student may attribute the passing grade to eating a magic apple every day and not consider the extra study time. In this example, the flawed logic is that the magic apple was the cause of the improved grade – when in fact the magic apple had no impact on the grade. "The Regression Fallacy" web-page (https://www.fallacyfiles.org/regressf.html), tells us that regression fallacy is the result of the misinterpretation of "regression to the mean". "Regression to the mean" is described as "the tendency of a variable characteristic in a population to move away from extreme values toward the average value". The impact of this misinterpretation is that it confuses true causes and effects and puts more weight on generalities.

2.)

Standard deviation is the amount of variance among the mean. In the case of the scenarios presented, I would expect the second scenario of 100 graduate students studying Data Science to have the smaller standard deviation. The average amount of textbooks for Data Science students would roughly be the same, compared to a larger spread for the averages of textbooks for all graduate programs. The first scenario contains students from all of Depaul's graduate programs. As such, there may be programs that require more textbooks such as Education, English or Mathematics compared with programs that may require fewer textbooks such as Physical Education. Also, as a current Data Science student, there is a degree of bias in this answer.

3)

The empirical rule or the 68-95-99.7 rule allows us to estimate the percentages of data up to 3 standard deviations away from the mean.

For this problem the range is 18 to 64, the mean is 28 and the SD is 4.

3a.)

The percentage of students between the ages of 24 and 32 is **68%**. This is calculated because we know the mean is 28 and the SD is 4:

28 + 4 (or 1 standard deviation) = 32
28 – 4 (or 1 standard deviation) = 24

According to the empirical rule, the 68% of the data falls within one standard deviation of the man.

3b.)

The percentage of students older than 36 years is **2.5%**.

For this problem, we know that 36 is two standard deviations away from the mean. The empirical rule states that two deviations away

28 + 4 + 4 (or 2 standard deviations) = 36

95% of the ages are between 20 and 36

This tells us that there are 5% of the ages lower than 16 and higher than 36. Thus, ½ of 5% is 2.5%

4.)

In this problem, we know the following:

mean = 150 (or $150,000)
sd = 35 (or $35,000)

We are looking for a value in the 99th percentile or top 1%.

Using a z-table (http://www.z-table.com/), a percentage of .9901 has a z-score of 2.33.

To determine the top 1%, we can then do the following: z-score * sd + mean

2.33 * 35 + 150 = 81.55 + 150 = 231.55

Based on the above calculations, the top 1% sales figure should be **$231,550**.

5.)

For this problem, a two-tailed hypothesis test is used:

N = 35 days
$\bar{x} = 42$
$\sigma = 15.5$
$\alpha = .05$ (95% confidence level)
z - score = 1.96

H_0: $\mu = 45$ – the null hypothesis is that the number of intrusion is 45
H_1: $\mu \neq 45$ – the alternative hypothesis is that the number of intrusions is not equal to 45.

$$z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{N}}}$$

$$z = \frac{(42 - 45)}{\frac{15.5}{\sqrt{35}}} = \frac{-3}{2.6199} = 1.145$$

$z = 1.145 < 1.96$

Based on the above calculations, we fail to reject the null hypothesis.

6.)

6a.)

```
> model1 <- lm(RFEWIDTH ~ REDSHIFT , data = QUASAR)
> 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

> >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
> model2 <- lm(RFEWIDTH ~ LINEFLUX, data = QUASAR)
> 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

> >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
> 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

> >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
> 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

> >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
> 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
```

6b.)

Five models were generated for the Quasar data set. The first model looks at the REDSHIFT variable. The adjusted R-squared for this model is -0.03818 which immediately tells us that this variable is insignificant and that an additional term should be added to the model or the sample size should be increased. The second model is for the LINEFLUX variable. The summary for this model has a high p-value of 0.3 and this immediately tells us that this model is a bad fit. This model's t-test values are also relatively high. The third model for LUMINOSITY also has an adjusted r-squared that is negative. The p-value and t-tests are both high. It is obvious that Model 3 is not a good fit. Again, a negative r-squared value tells us that this variable is insignificant in this model. The fourth model for AB1450 has a p-value that is less than 5%, its t-tests look good, however, its adjusted r-squared is low – which states that 27% of the variability of this variable is explained by this model. The fifth model, ABSMAG, has a good p-value of 0.0012 and a good t-test compared to each of the previous models. The adjusted r-squared value is 35%, which is better than the fourth model.

Based on the 5 models, the best model would be model #5, the ABSMAG model.