STAT 614 - HW 7

Due: Monday, November 23, 2020 in Blackboard by 11:59pm.
Instructions: Please type your solutions in a separate document and upload the document in Blackboard as a pdf. I will not be collecting syntax for this assignment. You will need concepts from Chapters 7 & 8 on the simple linear regression model (in addition to past models!). HW 8 will address multiple regression.

Forced expiratory volume (FEV) is an index of pulmonary function that measures the volume of air expelled after 1 second of constant effort. The data set FEV.csv in Blackboard contains determinations of FEV for 654 children ages 3 through 19 who were seen in the Childhood Respiratory Disease (CRD) Study in East Boston, Massachusetts. These data are part of a longitudinal study to follow the change in pulmonary function over time in children. Variables in the data set are the participant ID number, Age (in years), FEV (in liters), Height (in inches), a binary Sex indicator (0 = female/1 = male), and Smoking status (0 = non-smoker/1 = current smoker).

1. Characterize the association between pulmonary function (FEV) and smoking status. To do this answer the following questions:

    a. Use the natural log transformation of FEV and examine an independent two-sample procedure to test for differences in the population mean LN(FEV) between the two smoking groups. Provide a brief summary of the model results. Is there evidence of an association between (transformed) pulmonary function and smoking status? If so, estimate the extent of the association (that is, give estimate and confidence interval for the parameter of interest).

    Assumption of two-sample t-procedure for difference of means:

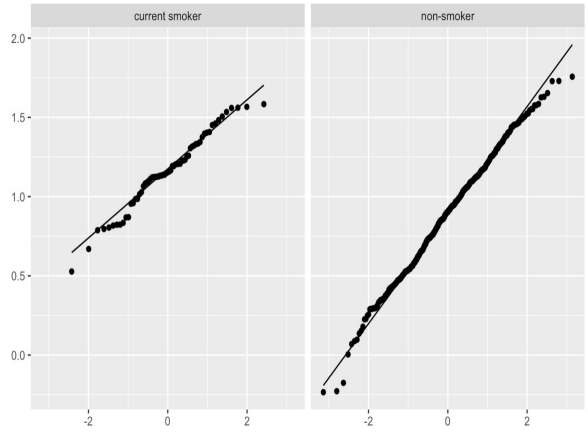
    - We have two samples of independent observations from two distinct populations: met
    - The samples are independent: met
    - Both populations are normally distributed with unknown mean and sd: the large p-value and qqplot are indicate that these two groups are normality.
    - Equal standard deviation: 0.23 vs. 0.33 - almost equal (variance is, too).

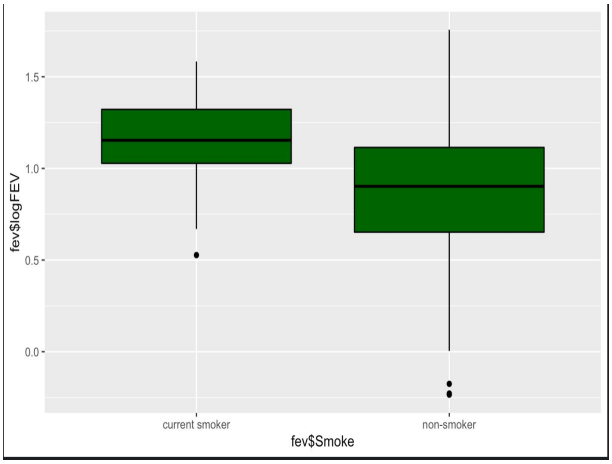
```
        Shapiro-Wilk normality test

data:  fev$logFEV[fev$Smoke == "non-smoker"]
W = 0.99599, p-value = 0.1389


        Shapiro-Wilk normality test

data:  fev$logFEV[fev$Smoke == "current smoker"]
W = 0.97567, p-value = 0.2283
```

```{r}
favstats(logFEV ~ Smoke, data = fev)
```

Smoke	min	Q1	median	Q3	max	mean	sd
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
current smoker	0.5270926	1.0278321	1.1534161	1.322022	1.583505	1.1604760	0.2342048
non-smoker	-0.2344573	0.6523252	0.9021918	1.114486	1.756650	0.8883953	0.3316671

2 rows | 1-8 of 10 columns





The study is met with a parametric two-sample t-test, so we set the mu1 = the population mean of the non-smoker group, and mu2 is the population mean of the current smoker group.

H0: mu1 = mu2

Ha: mu1 ≠ mu2

σ = 0.05

```
        Welch Two Sample t-test

data:  logFEV by Smoke
t = 8.4751, df = 94.957, p-value = 2.983e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2083467 0.3358146
sample estimates:
mean in group current smoker     mean in group non-smoker
                   1.1604760                    0.8883953
```

With a small p-value we have evidence to reject the null hypothesis in favor of the alternative hypothesis that the population mean of the non-smoker group and the current smoker group are different with 95 % confidence intervals. Furthermore, the current

smoker group is greater than non-smoker group 0.208 to 0.335 liters of forced expiratory volume, and the difference in sample mean is 0.272 with 95 % CI.

b. Are you surprised by the results in (a)? (Note that this is the unadjusted association.)
I am not surprised. With the large sample sizes for each group we could see there are different transformed FEV liters between two groups through exploratory data analysis.

2. The smoking status variable is an indicator variable in that it takes the value 0 for non-smokers and 1 for current smokers. Non-smokers are considered the reference group. Even though this is a categorical (i.e. qualitative or grouping) variable, we can use indicator variables to designate groups in the regression procedure. (This is solely due to the 0/1 status of the variable!) Fit the simple linear regression model of FEV (use the natural log transformed FEV) with smoking status as the explanatory variable.

Simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

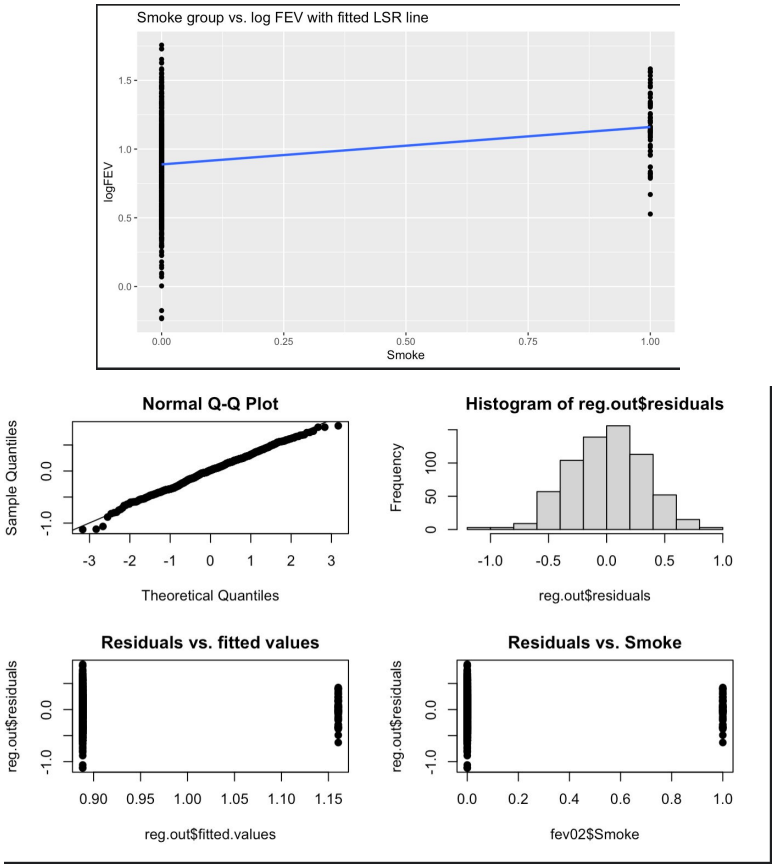
$$\varepsilon_i \sim N(0, \sigma^2)$$

a. Test the null hypothesis of no association between smoking status and FEV in the simple linear regression model.

Assumptions of simple linear regression model:

1. independent observation
2. Normally distribution
3. Equal variances
4. No influential outliers
5. Linear association between (mean) y and x. That is, residual : ri = yi - yhat i.

As the data met the t-test assumptions, we just need to check the residual with the same data. And we can know residuals are consistent with normality, and no influential outliers. Thus, the transformed data (logFEV) met the assumptions.



H0: $\beta 1 = 0$ vs. Ha: $\beta 1 \neq 0$

The null hypothesis states that the slope is equal to zero, and the alternative hypothesis states that the slope is not equal to zero. From the model summary, the model p-value and predictor's p-value are less than the significance level, so we know we have a statistically significant model. Also, the R-squared is 0.05975 so we have 5.97 % of the Variation in logFEV can be explained by the linear relationship with Smoke, and the F-test is statistically significant.

```r
reg.out <- lm(logFEV ~ Smoke, data = fev02)
summary(reg.out)
```

```
Call:
lm(formula = logFEV ~ Smoke, data = fev02)

Residuals:
     Min       1Q   Median       3Q      Max
-1.12285 -0.22803  0.01238  0.21777  0.86825

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.88840    0.01333  66.668  < 2e-16 ***
Smoke        0.27208    0.04227   6.437 2.36e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3234 on 652 degrees of freedom
Multiple R-squared:  0.05975,   Adjusted R-squared:  0.05831
F-statistic: 41.43 on 1 and 652 DF,  p-value: 2.364e-10
```

b. Interpret the slope coefficient.

Coefficients: For each variable and the intercept, a weight is produced and that weight has other attributes like the standard error, a t-test value and significance.
In this SLM , we have found a 95 % confidence interval for the slope of Smoke.
Par: $\beta 1$ = slope of Smoke variable
$\hat{\beta}0 = 0.88840$ $SE(\hat{\beta}0) = 0.01333$
$\hat{\beta}1 = 0.27208$ $SE(\hat{\beta}1) = 0.04227$
MSE = 0.1046
Test statistics $T = 0.27208/0.04227 = 6.437 \sim t(652)$

1. Estimate: This is the weight given to the variable. In the simple regression case (one variable plus the intercept), for every one level increase in Smoke (0 for non-smokers and 1 for current smokers), the model predicts an increase of 0.27208 FEV / liters.
2. Std. Error: Tells us precisely how the estimate was measured. It’s really only useful for calculating the t-value.
3. t-value and Pr(>[t]): The t-value is calculated by taking the coefficient /Std. Error. t-value is then used to test whether or not the coefficient is significantly different from zero. If it isn’t significant, then the coefficient really isn’t adding anything to the model and could be dropped or investigated further. For this model, Pr(>|t|) is the significance level.

c. Compare your results to those in part 1a. You should draw identical conclusions. Do you?(Same estimated difference in mean LN(FEV) between non-smokers and smokers, same confidence interval, same p-value.)
The main difference is that t-tests involve the use of categorical predictors, while linear regression involves the use of continuous predictors, so we have changed Smoke variables to numeric type before we start to do the linear regression model.

	Two-sample t-procedure	linear regression model
estimated difference in mean	0.272	0.272 (slope coefficient)
95 % confidence interval	0.208 ~ 0.335	0.189~ 0.354
p-value	2.983e-13	2.364e-10

The 0.0025 cut-off for t-distribution with 652 degrees of freedom is qt(0.025, df = 652) = -1.963609. Hence a 95 % CI for β1 had endpoints: 0.272est - 1.964(0.042se) = 0.189 & 0.272est + 1.964(0.042se) = 0.354. That is, we are 95 % confident the population slope β1 is between 0.189~ 0.354. Moreover, both p-values of two models are provided H0 ≠ Ha, and have an equal difference in mean, which suggest evidence the result of two models are the same.

d. Explain how to use a regression model with indicator variables to include a three (or more) category explanatory variable. For example, if smoking status was 0 for never smoked, 1 for past smokers, and 2 for current smokers, how would you incorporate the three smoking levels into a regression model? (Note: This gives us a way to incorporate both quantitative and qualitative variables into a model!)

For a variable with n categories there are always n - 1 dummy (indicator) variables. In this question, we have three categories, so we have 3-1 = 2 indicator variables.

$y = \beta 0 + \beta 1x1 + \beta 2x2$

smoking status	x1	x2
never smoked	0	1
past smokers	1	0
current smokers	0	0

In this question, we don't have to create an indicator variable to represent the "independent" category of smoking status. If x1 = 0 and x2 = 0, we can directly know the smoking status is neither never smoked nor past smoker. Therefore, current smokers must be independent and this model can still work.
Reference: https://stattrek.com/multiple-regression/dummy-variables.aspx