In Lesson 3.1 you must run a linear regression and an ANOVA in a software package other than Python.

**Any assignment submitted in Python will be a zero.**

**If you do not remember how to run a regression and/or ANOVA please return to the online notes for STAT 500 or the equivalent course you completed as a prerequisite for STAT 487**

Use the [health data](https://drive.google.com/file/d/1zLF7t6DiitfSUPQUd6gSZ1hw90je0Sa-/view?usp=sharing) for the Unit 3 Activities.

The health data is adapted from medicaid data from the U.S. (a copy of the data dictionary is contained on CANVAS in the ‘Course Data Set Dictionaries’ module.)   The data represent medicaid data averages by state.  The variables you will be working with have been adapted.  Specifically the amounts in the data have been divided by 10,000 to be more visually appealing.  Also, the total cost variable is the total actual cost divided by the number of beneficiaries for the state.

Please do not draw any actual conclusions from this data, it is for illustrative purposes only.

Task 1

Conduct an OLS Regression, as a point of emphasis, **in a software package of your choice other than Python**.  Specifically, your model will predict TotCostPer from TestUsers (the number of medicaid beneficiaries who file a claim for medical testing)  and Average Age (the average age of medicaid beneficiaries in the state).  To complete this assignment you must:

1. Conduct univariate descriptive statistics for all three variables and conduct univariate graphs for all three variables.

Testusers is heavily right skewed. While age is somewhat normally distributed.

Chart, histogram

Description automatically generated

1. Conduct bivariate descriptive statistics among all possible combinations of all three variables and conduct bivariate graphs for all three variables.

Text

Description automatically generated

Calendar

Description automatically generated

1. Conduct an OLS regression
   1. (if you cannot remember how to run a regression please see your class notes from STAT 500 or the equivalent course you completed)

Text

Description automatically generated with medium confidence

1. Check the linearity, independence, normality, and equal variance assumptions for the regression by using graphs
   1. (if you cannot remember how to use 4 in 1 plots to visually examine the regression assumptions please see your class notes from STAT 500 or the equivalent course you completed)

Linearity:

* Checking using the plot of residuals vs fitted. No trend is obvious here.

Independence:

* We can run the predicted vs. time to check autocorrelation but durbin should work better.

Normality:

* Normal QQ plot shows all points near the diagonal line => mostly normal.

Equal Variance:

* Scale-location plot shows not too horizontal. The assumption might be slightly violated.

Diagram, schematic

Description automatically generated

1. Check the linearity, independence, normality, and equal variance assumptions by using the appropriate tests for each assumption
   1. (if you cannot remember how to use tests to test the regression assumptions please see your class notes from STAT 500 or the equivalent course you completed)

Linearity:

* Checking using the plot of residuals vs fitted. No trend is obvious here.

Independence:

* Durbin-Watson p-value > 0.05 => errors are NOT autocorrelated

Normality:

* Shapiro-Wilk p-value > 0.05 => residuals are normally distributed

Equal Variance:

* NCV p-value > 0.05 => equal variance satisfied

Text

Description automatically generatedText

Description automatically generated

Text

Description automatically generated

1. If the assumptions are not met, run a robust regression or transform your data (if you transform your data you will have to check the assumptions a second time).  If no transformations are necessary then state ‘No transformations necessary’

Text, letter

Description automatically generated

1. Interpret your final model including:
   1. Coefficients/significance test for each predictor
   2. Overall model significance

From the Wald test below, we can see that Testusers and Age are both statistically significant different from 0. Thus, both are significant predictor of Total Cost.

Robust model



The MSE of the robust model is 1490.549. From the plot of the error, no strange behavior is detected – normal and no trend in residuals plot.

Graphical user interface, text, application, email

Description automatically generated

Chart, diagram

Description automatically generated

Task 2

Conduct a Two-Way Main Effect ANOVA, as a point of emphasis, **in a software package of your choice other than Python**.  Specifically you will determine if there are significant differences in TotCostPer among the two “Female” groups and/or significant differences among the 10 “FedRegion” groups.  (Note:  Female is 1 if the state has a majority of female medicaid beneficiaries; FedRegion are the codes for the standard federal region, a 10 category classification system the federal government uses to geographically classify the states.  To complete this assignment you must:

To complete this assignment you must:

1. Conduct univariate descriptive statistics for all three variables and conduct univariate graphs for all three variables.

For two categorical variables, the barplot provides the count. We see a lot of 1 in Female variable while FedRegion has almost even distribution across all categories.

Chart, bar chart, histogram

Description automatically generated

1. Conduct bivariate descriptive statistics and conduct bivariate graphs for all three variables.

Focusing on the bivariate between each categorical and TotCostPer. We see group 1 in Female has a much higher TotCostPer versus the 0 group. Similarly, TotCostPer also varies across the FedRegion groups

We also remove color for FedRegion because 10 colors are too much!

Chart, box and whisker chart

Description automatically generated

1. Conduct an two way main effects ANOVA

Anova result is given below. Note that we use type III to account for UNEQUAL sample size across factors.. Interpretation is given in 14.

Text

Description automatically generated

1. Check the independence, normality, and equal variance assumptions for the ANOVA model
   1. (if you cannot remember how to use plots to visually examine the ANOVA assumptions please see your class notes from STAT 500 or the equivalent course you completed)

Independence:

* Assuming data is collected through time. Need index of collection to really test here.

Normality:

* Most of the points stay close to the diagonal line. There are some deviations at the lower tail but not too much to worry about. Assumption should be met as well.

Equal Variance:

* Scale-location has a bit of a curve here but not too drastic. Also, the residual vs fit show good horizontal upper and lower bands. Assumption should be good here.

Chart

Description automatically generated

1. Check the independence, normality, and equal variance assumptions by using the appropriate tests for each assumption
   1. (if you cannot remember how to use tests to test the ANOVA  assumptions please see your class notes from STAT 500 or the equivalent course you completed)

Independence:

* Assume all data points are collected independently.

Normality:

* Shapiro-Wilk test, p-value > 0.05 => normality also met.

Equal variance:

* Using Levene test, p-value > 0.05 => no evidence different variance => assumption met.

Above assumptions can also be seen through plot.

Text

Description automatically generated

Text

Description automatically generated

1. If the assumptions are not met, run a robust model or transform your data (if you transform your data you will have to check the assumptions a second time)

Assumptions are met!

1. Interpret your final model including:
   1. Significance of each main effect ; post hoc test if appropriate
   2. Overall model significance

First if we assume the interaction model, we see that the interaction term is NOT significant and thus reduce to two variables (Additive model).

However, the additive model shows only Female is significant predictor as well! Thus, in fact we should be using only Female here

The model is significant with only Female variable but it didn’t explained much variance! One interesting to note is that the group of Female = 1 has a bit higher Total Cost compared to other group! It is indeed significant as well..

Text

Description automatically generated

Text

Description automatically generated

To submit this assignment, insert screenshots of the output from your statistical software into this word document and include your written conclusions

**Submissions without output and/or interpretation will be graded as zero**