Stephanie Mecham

BIOSTATS 522.

1/22/18

BIOSTAT 522 HW #2 (Due Tues, Jan 23, 2018, in class)

For this assignment, use a SAS dataset called “surgery”. This is a real dataset with messy data issues like most data you will see in real world. Submit any SAS code used to do the assignment.

The data are from 261 patients with a major surgery, followed for five post-operative days. For this assignment, we will only use two variables: **pain** and **anxiety**. Both variables are self-reported, and were assessed on post-op day 5 using a visual analog scale that can give values ranging from 0 to 10 where 0 corresponds to min possible level and 10 corresponds to max possible level. The goal for now is to assess the relationship between **pain** and **anxiety**.

To start, if you downloaded the SAS dataset (named surgery.sas7bdat) from canvas to, for example, ‘M:\class\bios522-w18\datasets’ folder in M drive, then one way to start this assignment is by submitting the following SAS code to access the dataset.

**/\* Below saves a new temporary dataset named “tmp” using surgery.sas7bdat. \*/**

libname homework "M:\class\bios522-w18\datasets\";

**data** tmp;

set homework.surgery;

**run**;

You likely need to use PROC MEANS, PROC CORR and PROC REG, and some plotting commands for this assignment. You might feel that some questions are repetitive, and that is because many are indeed related questions asked in different ways.

1. Find appropriate descriptive statistics for **pain** and **anxiety** and describe for each what you find. Obtaining **N, mean and standard deviation** may be sufficient, but feel free to take a look at the distribution of the variables and other descriptive statistics as you want to practice SAS. This dataset will be used for future assignments.

**Pain:**

**N: 206**

**Mean: 3.746**

**St. deviation: 2.51**

**Anxiety:**

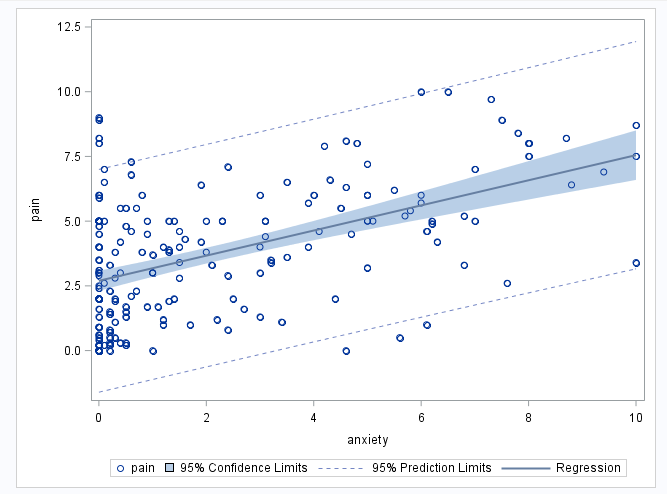
**N: 202**

**Mean: 2.102**

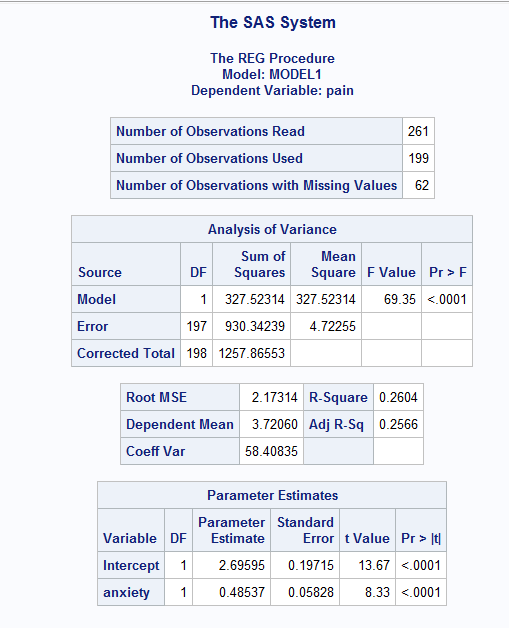
**St. deviation: 2.64**

1. Assess graphically the relationship between **pain** and **anxiety** by plotting **pain** on the y-axis and **anxiety** on the x-axis. Describe in one sentence what you assessed from the graph and submit the graph.

**The scatter plot shows a slight positive linear relationship between anxiety and pain.**

****

1. Fit a regression model with Y = **pain** and X = **anxiety**. Paste the SAS output, write out the model you found, and interpret estimates you found for the intercept (β0) and slope (β1).



**Model:**

**y = 2.70+0.49x where y = reported pain score and x= reported anxiety score**

**Interpretations:**

**For patients that reported ‘0’ on the anxiety scale, we can expect a pain score of approx.. 2.70. The 0.49 coefficient represents the estimated change in pain score per anxiety score. The slightly positive slope indicates a weak positive relationship between the patients’ anxiety self-report and pain self-report.**

1. Provide **N and %** of the study cohort for the following questions. Based on your analysis so far, how many non-missing **pain** data do you have in this dataset? How about those with non-missing **anxiety** data? And how many patients provided data for the scatter plot in problem #2? Lastly, how many observations are used in the regression model?

**Non-missing pain data = 206 (206/261 \*100 =78.93%)**

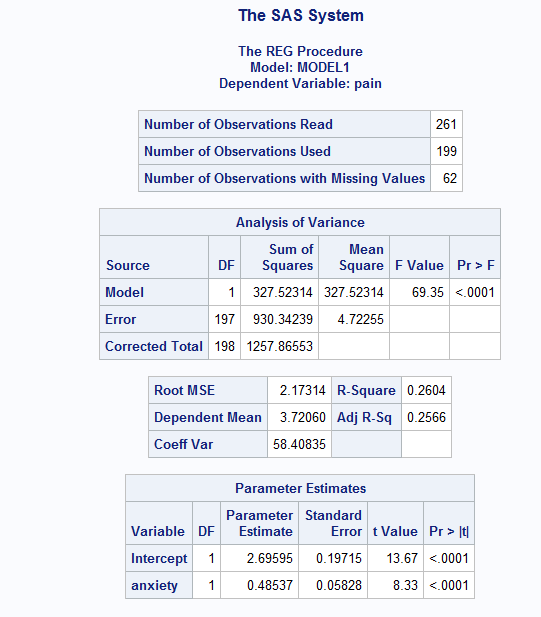
**Non-missing anxiety data= 202 (202/261 \* 100= 77.39%)**

**Scatter plot included data from 199 patients (199/261 \*100 =76.25%)**

**Regression model also included data from 199 patients (76.25%)**

5) Under the regression model, the errors, ei, are assumed to follow a Normal distribution with mean 0 and standard deviation, σ. Find the estimate of σ for the model you fit.

**Estimate: 2.173**



6) In one or two sentences, compare the estimates for σfrom the regression model with the sample standard deviation of the variable **pain** and describe in one sentence why and how they are different.

**Regression model σ estimate: 2.173**

**Pain standard deviation: 2.513**

**The sample standard deviation of pain is 0.34 larger than the estimate found in the regression model. This has to do with the fact that the statistical formulas for obtaining these two metrics are a little different, so they are close but not entirely identical (regression estimates will be different than sample population calculations).**

Whenever asked to do a test, state H0 and H1, write the test statistic (formula), state the test statistic’s reference distribution, compute the test statistic and p-value, and draw a conclusion.

7) Test to assess if **pain** is associated with **anxiety**.

**H0 : β1 = 0 (self-reported pain scores have no association with self-reported anxiety scores)**

**H1 : β1 ≠ 0 (self-reported pain scores are associated with self-reported anxiety scores)**

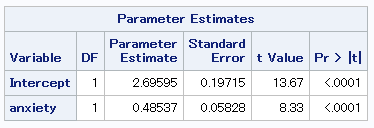
**Test statistic formula: Use a t-test. T = β1 hat / SE (β1 hat)**

**Test statistic’s reference distribution: 197 degrees of freedom (n-2 or 199-2=197).**

**Test statistic: 8.33**

**P-value: <0.0001**

**Conclusion: Because our p-value is lower than the standard alpha significance level of 0.05, we can reject the null hypothesis and conclude that self-reported pain scores are indeed associated with self-reported anxiety scores.**



8) What are the predicted pain and residual of a patient who had anxiety score of 2 and pain score of 8 based on the fitted model?

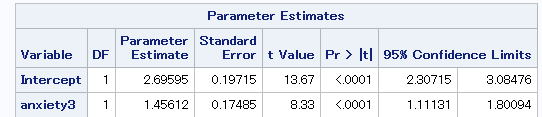
**Predicted pain: 3.68 (y=2.7+(4.9\*2) = 3.68))**

**Predicted residual: 4.32 (8-3.68 = 4.32)**

9) Compute a 95% confidence interval for the mean change in pain score per three unit increase in anxiety based on the model.

**Estimate: 1.456**

**95% CI: (1.111, 1.801)**



10) Compute **by hand** a 95% confidence interval for mean pain when anxiety = 4 and a 95% prediction interval (PI) for pain when anxiety = 4. Briefly contrast the CI vs. PI.

**Estimate: 2.7 + 0.49(4) = 4.66**

**CI: μ(x0) = μhat(x0) ± tn-2, 0.975 σhat\*√ (1/n + (x0 - x̅)2/SSx)**

**= 4.66 ± 1.9721\*2.65\*√(1/199 + (4– 2.11)2/1390)**

**CI: 4.66 (4.20, 5.12)**

**PI: Y(x0) = Yhat(x0) ± tn-2, 0.975 σhat\*√ (1 + 1/n + (x0 - x̅)2/SSx)**

**= 4.66 ± 1.9721\*2.65\*√ (1 + 1/199 + (4– 2.11)2/1390)**

**PI: 4.66 (-0.586, 9.91)**

**We can see from above that the confidence interval is much narrower than the prediction interval, which makes sense because prediction intervals take into account uncertainty plus the scatter/distribution of data, whereas confidence intervals just tell us about uncertainty in capturing the true mean.**

11) What fraction of the variation in **pain** is explained by **anxiety**?

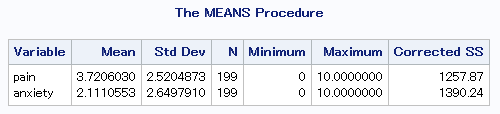
**26.04% of the variation in pain can be explained by anxiety, because the regression analysis resulted in an R2 of 0.2604, or 26.04%.**

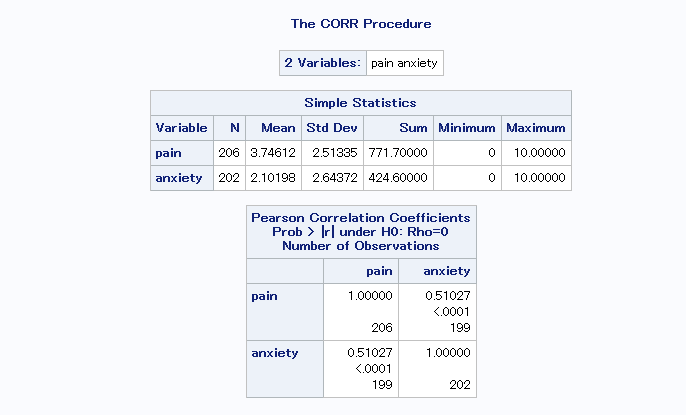


12) What fraction of the variation in **anxiety** is explained by the variation in **pain**? Compare your answer to this problem with what you found in the problem above.

**Like above, 26.04% of the variation in anxiety can be explained by the variation in pain. This is a feature of linear regression, because R2 represents the proportion of variability in Y explained by the linear relationship between X and Y. That is, fraction of variation in Y due to X will be the same as X due to Y in simple linear regression.**

13) [SLR3 note] Now you will **hand calculate** the **intercept** and **slope estimates** for the model where Y = **pain** and X = **anxiety** using only what you obtain from PROC MEANS and PROC CORR. Read below hints first to avoid making mistakes. Please paste the output and show your work for the hand calculations.





**β1hat = r \* SD(Y)/SD(X)**

**= 0.51027 \* (2.5204873/2.6497910)**

**= 0.4854**

**β0 = Ybar - β1hat(Xbar)**

**= 3.74612 – (0.4854\*2.10198)**

**= 2.726**

**Model: y = 2.726 + 0.4854x where y is reported pain and x is reported anxiety.**

Hint: You will need the numerical values for quantities like r, n, SD(X) (or SSX which is not SD(X), but is related), etc. Getting SSX is not a requirement for this problem, but to clearly see the difference between SD(X) and SSX, it might be worth writing out how you can calculate SSX if you know N and SD of X.

**What you need to be careful here and with any dataset:** If you use PROC MEANS to get summary statistics (e.g., means) of any variables (such as X and Y), it will give you the **summary statistics of X using all non-missing X**, and **the summary statistics of Y using all non-missing Y**. And the number of observations with non-missing X and the number of observations with non-missing Y can be very different.

The intercept and slope from a regression model will be based on persons for whom pairs of X and Y are non-missing. Hence, the SD(X) or SSX from PROC MEANS for the variable X can be different from the SD(X) or SSX to be used for the regression model, *if some of the observations with X variable have missing Y variable, or vice versa*.

Below are two blocks of SAS codes that you can use to obtain the mean and SD of X or even SSX (under Corrected SS column). The second block of PROC MEANS restricts the calculation to a subset that contains only the observations where values of both **pain** and **anxiety** are non-missing. Make sure to use the correct values.

/\* Gives you summary statistics for pain and anxiety \*/

**proc** **means** data=tmp mean std n min max css;

var pain anxiety;

**run**;

/\*Gives summary statistics for pain and anxiety, but only for the data used in the regression model, i.e., where both pain and anxiety are non-missing. \*/

**proc** **means** data=tmp mean std n min max css;

where pain NE **.** and anxiety NE **.** ;

/\* The above subsets the dataset to those with non-missing pain and anxiety. \*/

var pain anxiety;

**run**;

**CODE:**

\*Biostats 522 Homework #2 | Stephanie Mecham |;

libname Hw2 "C:\Users\smecham\Desktop\Hw2";

**run**;

**data** Hw2.homework;

set Hw2.surgery;

keep idn anxiety pain;

**run**;

\*Question 1 | Descriptive Statistics;

**proc** **univariate** data= Hw2.homework;

**run**;

\*Question 2 | Scatterplot;

**proc** **sgplot** data= Hw2.homework;

scatter x=anxiety y=pain;

reg x=anxiety y=pain / cli clm;

**run**;

\*Question 3 | Regression;

**proc** **reg** data= Hw2.homework;

model pain = anxiety;

**run**;

\*Question 9 | Creating a new factor of anxiety;

**data** Hw2.homework;

set Hw2.homework;

anxiety3= anxiety/**3**;

**run**;

**proc** **contents** data=Hw2.homework;

**run**;

**proc** **reg** data=hw2.homework;

model pain=anxiety3 / clb;

output out=regout\_anxiety3 p=pred l95m=lcl u95m=ucl l95=lpl u95=upl;

**run**;

\*Question 13;

**proc** **means** data=Hw2.homework mean std n min max css;

where pain NE **.** and anxiety NE **.** ;

**run**;