**Problem Set #1,**

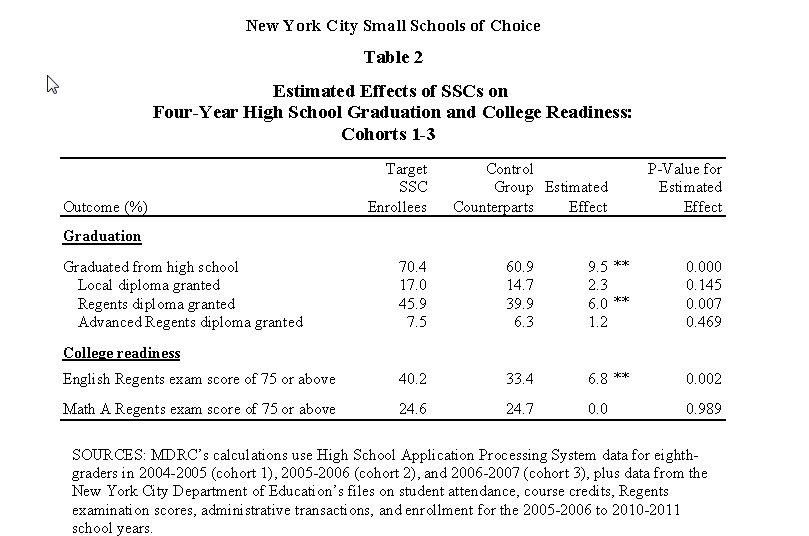
**Good luck on this problem set. I hope that you learn a lot while doing it. Just so we are clear: You need to work on this independently, consistent with the New School’s policy on academic honesty (responses may be text mined and assessed for similarity). Your responses should be in your words and based on your work. Be sure to answer all questions (and subquestions). Read carefully and start early. Make extra room as needed. Late submissions will lose points.**

**Enter Name Here: Douglas Locke**

1. **Understanding statistical significance and the general concepts around statistical inference. First, make sure to review the web video I prepared for you on statistical significance (in class 1 materials). Then, review the case below and answer the questions.**

**Researchers are currently analyzing the effects of an educational reform program in New York. The following table contains a series of real “T tests” computed to determine whether the program had an effect on educational outcomes. How is statistical significance used in this context? What is the benefit of this criterion? What does the P value tell you (use the first two rows of the table as examples—Graduated from High School and Local Diploma Granted)? Does a statistically significant effect indicate that the results are necessarily “substantial” or “policy relevant?” Does it mean that effects are “proven.”? The researchers concluded based on this pattern of results that the program was largely effective. Why did they come to that conclusion? Which outcomes did not have statistically significant differences? Does this mean that the program did not affect those measures? Below is some background on the program that was tested and below that is the table. You don’t need to interpret every outcome, just answer the questions above:**

In 2002, New York City embarked on an ambitious and wide-ranging series of education reforms. At the heart of its high school reforms were three interrelated changes: the institution of a districtwide high school choice process for all rising ninth-graders, the closure of 31 large, failing high schools with an average graduation rate of 40 percent, and the opening of more than 200 new small high schools. Over half of the new small schools created between the fall of 2002 and the fall of 2008 were intended to serve students in some of the district’s most disadvantaged communities and are located mainly in neighborhoods where large, failing high schools had been closed. The table below compares outcomes for students who attended these “small schools of choice,” or SSCs (so called because they are small, are academically nonselective, and were created to provide a realistic choice for students with widely varying academic backgrounds), to those who attended normal or “control” schools. Random assignment was used to determine who went to which schools. If you want more information see [here](http://www.mdrc.org/news/press-release/new-findings-show-new-york-city-s-small-high-schools-boost-college-enrollment), though this background is not necessary to answer the question.



<<<<<<Type your answers after here. Make as much room as you need>>>>>

How is statistical significance used in this context?

Statistical significance is used to determine if there is a non-random difference between the target “Small Schools” group and a control group.

What is the benefit of this criterion?

New York City would like to understand if the differences between these groups are large enough beyond what we might expect from a random sampling error.

What does the P value tell you (use the first two rows of the table as examples—Graduated from High School and Local Diploma Granted)?

The P value with “Graduated from High School” is telling us there is (almost) 0% chance the difference between the groups is due to randomness, and therefore there is statistical significance. The two stars (\*\*) indicates there is a < 5% chance the difference is due to randomness. It is not clear why with a p-value of .000 the author did not use (\*\*\*) three stars.

The P value associated with “Local Diploma Granted” is .145. This is telling us there is not statistical significance. There is a < 14.5% chance the difference between these groups is due to randomness.

Does a statistically significant effect indicate that the results are necessarily “substantial” or “policy relevant?”

No, not necessarily. Statistically significant results need to be interpreted in the context of the subject area or domain under study. Some results that appear to be statistically significant, that is, resulting in p-values < 5% themselves are due to randomness.

Does it mean that effects are “proven.”?

No. Statistically significant results merely reject the null hypothesis , which states “there is no difference between two groups”. It is not an affirmation of a causal hypothesis, because that is not what it is being tested.

The researchers concluded based on this pattern of results that the program was largely effective. Why did they come to that conclusion?

Judging by the limited set of variables in the problem, it is likely they came to the conclusion because they saw 3 of 6 variables showing that there is statistical significance in the results, with results favoring the small schools.

Which outcomes did not have statistically significant differences? Does this mean that the program did not affect those measures?

“Local Diploma Granted”, “Advanced Regent Diploma Granted”, “Math A Regent Scores of 75 or above” were not statistically significant. It does not mean the program did not affect those measures. We can only say that the effects were more likely to be due to chance.

**Suppose you work for a nonprofit trying to offer low interest credit cards to individuals who are cutoff from prime credit markets. You are trying to understand more about your risk exposure and how the characteristics of applicants can predict how much debt they will accumulate on their credit card. You collect data about the profile of borrowers when you opened their accounts. You then regress those variables (measured at the time of application) on their current credit card debt levels. This will help you understand, going forward, which borrowers are likely to accumulate the most credit card debt. Below are the descriptive statistics for the variables you can include in your model. Again, all variables are measured at the time of application with the exception of the dependent variable, which is measured at the current time.**

****

**#2. If you included all of these variables in a regression, what would the sample size be? \**

700. Valid N (listwise) is the number of cases that don’t have missing values. Regression requires variables without missing values.

We possibly could increase the regression to 850 but only if we did imputation on the missing values.

**#3 Based purely on the correlation matrix below, what appear to be the best candidates for independent variables to help predict credit card debt? List the variables in order of likelihood that they will be good predictors. What are the limitations of this approach to model building?**

**These 3 are the best candidates because they have coefficients > .3.**

1. Household Income in Thousands (.544)
2. Years with Current Employer (.428)
3. Age in Years (.326)

**These 2 have coefficients < .3 indicating they are very weak, and thus are not the best candidates:**

1. Previously Defaulted (.258)
2. Years at Current Address (.181)

Possibly, we may remove Previously Defaulted because the N is only 269 while the others are 329. So with these missing values we may either remove the variable or impute the missing values. Since it may be difficult to accurately impute this value, we may remove the variable from the regression.

What are the limitations of this approach to model building?

Correlation analysis assumes linear relationship between variables, normal distributions, and uniform variance. Correlation coefficients don’t tell us about relationships across more than 2 variables, so they are not helpful when analyzing 3 variables, hence the need to standardize.

****

**4 Among this potential set of predictors, do you have any concerns (based on this correlation matrix) about potential multicollinearity?**

Yes. The first three variables “Age in Years”, “Years with Current Employer” and “Years at Current Address” all have high potential for multicollinearity. People tend to live near where they work, so “Years at Current Address” and “Years with Current Exmployer” are likely to have multi-collinearity. These two variables not exactly the same, but some “overlapping” information. The same may be said of “Age in Years” with both “Years with current employer” and “Years at Current Address.” Some of these strong coefficients (.641 between “Years at Current Address” and “Age in Years” for example raise suspicion of multi-collinearity.

**5. You run three models as shown below in the “model summary”**

**a) The first model tried to predict credit card debt on the basis of household income. Interpret the R square in this case.**

32.5% of the variation in the variable “credit card debt” is explained by “household income”.

**b) The second model adds a dummy variable which indicates whether or not you have a high school diploma. Based on the change in R square and F, did adding this variable improve the model?**

Based only the change in R Square, the second model does not explain the variation of “credit card debt” much better (the difference between R Square is in Model 2 (.327) and Model 1 (.325) is only .002 which is extremely negligible. The F statistic tells us if the model as a whole has statistical significance. Because the second model Significant F. Change is +.139 however (and above our rule of thumb significance level of .05), we can say that adding this variable did not significantly improve the prediction.

**c) The third model added several other predictors. Taken as a whole, is this model better than model 2?**

Yes. Looking at model 3 with its several new predictors, as a whole, the significant F change is measured at .000, which tell us these predictors, taken as a group, have significantly improve the model. We can see the R Square is now .618, a +.290 a very strong improvement over models both 1 and 2.

**Model Summary(d)**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | | | | |
|  |  |  |  |  | R Square Change | F Change | df1 | df2 | Sig. F Change |
| 1 | .570(a) | .325 | .324 | 1.74054 | .325 | 336.268 | 1 | 698 | .000 |
| 2 | .572(b) | .327 | .325 | 1.73905 | .002 | 2.198 | 1 | 697 | .139 |
| 3 | .786(c) | .618 | .613 | 1.31679 | .290 | 87.447 | 6 | 691 | .000 |

a Predictors: (Constant), Household income in thousands

b Predictors: (Constant), Household income in thousands, No High School Diploma

c Predictors: (Constant), Household income in thousands, No High School Diploma, Debt to income ratio (x100), Years at current address, Previously defaulted, Age in years, Years with current employer, Other debt in thousands

d Dependent Variable: Credit card debt in thousands

**6 What does the output below tell you?**

An ANOVA analysis was run.

3 models were run to predict a dependent variable of Credit Card Debt (in thousands).

We know in model 1, there were was only a single variable (Household income).

In model 2, we had 2 variables, Household Income & High School Diploma.

In model 3, we had several predictor variables.

**The “Sig” value on the F-Statistic for all three models tell us that all 3 models are statistically significant, as all three Sig values are < .05.**

Breaking this down further, we can see each model contains “Regression” “Residual” and Total.

The first line “Regression” tell us about the variance that can be explained by the Model’s independent variables.

The second line “Residual” tells us about the variance that can NOT be explained by the model’s independent variables. In other words, the “Residual” tell us about the model’s error.

The third line “Total” explains the “Regression” (model) + “Residual” (error).

The regression degrees of freedom line tell us about the number of co-efficient in the model minus 1.

The Residual degrees of freedom is the total minus the regression degrees of freedom.

The Mean Squares are the sum of Squares divided by the degrees of freedom.

****

**7 These questions applies to the coefficients table which is shown on the next page. All questions apply to model #3, except where otherwise indicated.**

1. **Using an alpha value of .05, list which variables are statistically significant?**

1. Household income in thousands
2. Years with Current Employer
3. Debt to Income Ratio (x100)
4. Previously Defaulted

1. **Interpret the meaning of the unstandardized beta values for the variables which are statistically significant.**

1. Household income in thousands

.026 means that for every $1000 increase in household income, there is a $26 increase in Credit Card debt is predicted, while all other variables stay constant.

1. Years with Current Employer

.047 means that for every year increase with a current employer, there is a $47 increase in Credit Card debt predicted, while all other variables stay constant.

1. Debt to Income Ratio (x100)

.113 means that for every unit increase in the Debt to Income Ratio (x100), there is a $113 increase in Credit Card debt predicted, while all other variables stay constant.

1. Previously Defaulted

.941 means that for every unit increase in Previously Defaulted (or an easier way to say this is that when Previously Defaulted = 1), there is a $914 increase in Credit Card debt predicted, while all other variables stay constant.

1. **List the variables in order of their importance in predicting current levels of credit card debt.**

1. Household income in thousands (0.471 Standardized Coefficient)
2. Debt to Income Ratio (x100) (.366)
3. Previously Defaulted (.181)
4. Years with Current Employer. (.150)

1. **Does the effect of household income seem to “hold up” as other variables are added to the model? (for this you’ll have to look at all three models)**

It hold up well enough. In Model 2, the “No High School Diploma” variable is not significant, and therefore we can reject the addition of this new variable. In Model 3, with multiple variables added, the standardized coefficient moves from .563 (original from Model 1) to .471. (new in Model 3). The .092 difference is small, and in model 3, the “Household income” variable is still the strongest variable, with 2nd strongest being Debt to Income Ratio (.366).

****

**#8. Do the patterns of correlations and the VIF statistics that follow provide any basis for concern about multicollinearity?**

None of the VIFs are > 10, which would indicate a high amount of multi-collinearity. We may do nothing else to these models and safely ignore multi-collinearity.

However, some of the VIFs are between 1 – 10 which would indicate there is at least some multicollinearity occurring in the model. Based on these results, I would consider dropping “Other Debt in thousands” as it has the highest VIF, the lowest Tolerance, and see what happens to the model.

****

**#9. Do the casewise diagnostics (below) provide any basis for concern? (in general and/or specific cases). Explain**

The mean of the residuals is 0**,** which is a sign the residuals are normally distributed, a sign of model goodness of fit.

There are approximately 32 of these residuals listed which is about 4.6% of the N in the model (~700). So this seems a reasonable number of outliers and about what we expect (5%). This is because the std. residual is a z-score, or a measure of standard deviation. We expect 95% of the std. residuals to be within 2 z-scores of the mean.

There are 15 of these residuals more with std. residual > 2 (or < -2) again, this is ~ 2% of the data set and about what we might expect.

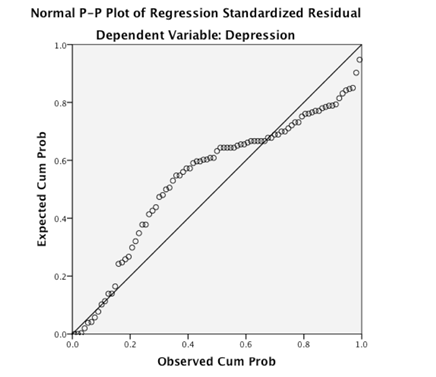
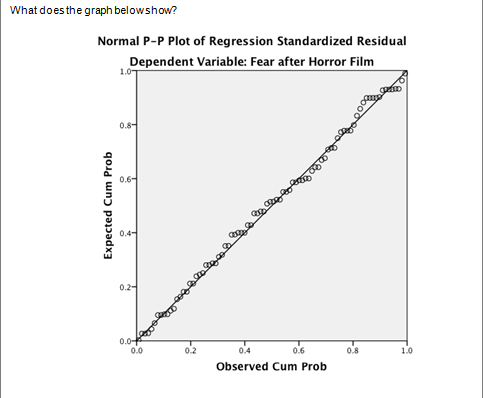
Of the 4 Debt to Income Ratio <= 5 cases, 3 have Std Residuals ~ 2 indicating the model handles them reasonably well.

Cases above 3 error terms should be investigated. I would start investigating those >5, then the > 4, and then the > 3. Cases 344 & 467 (the above 5’s ) should be investigated first.

**\**

****

1. **Compare these two plots. What are they intended to display? Which looks more problematic? What might this suggest?**



<<<<<<Type your answers after here. Make as much room as you need>>>>>

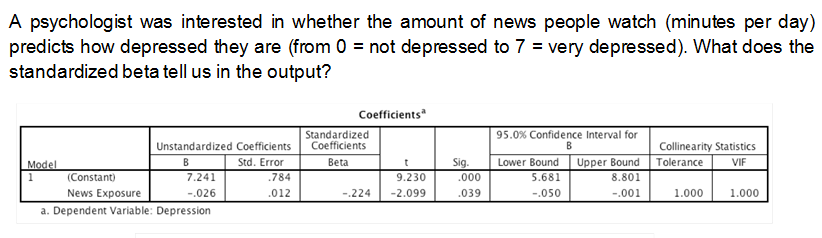
This P-P plot helps us to observe if the distribution of residuals is normal.

Moving from bottom left corner to the upper right, we look to see if the % of the errors is consistent with what we expect, that is a normal distribution.

The graph is on the left looks good, but the graph on the right is problematic.

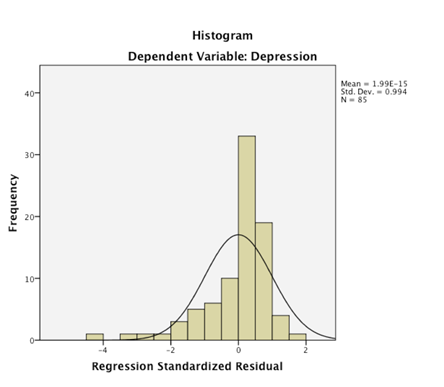
It suggests the model is mostly unreliable. There errors follow a pattern that looks more like a polynomial. If linear regression was used, perhaps the best model may be non-linear.

11..



Based on the output? is news exposure a significant predictor of depression?

But do you have any concerns about the graph below?



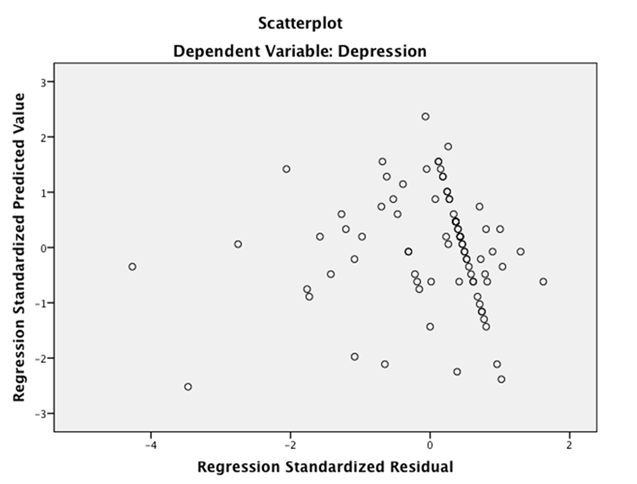
<<<<<<Type your answers after here. Make as much room as you need>>>>>

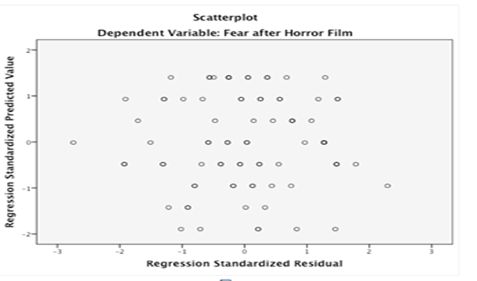
The standardized beta -.224 gives us a measure of the overall strength of relationship between News Exposure and Depression. It is actually a Z-Score. News Exposure goes up by 1 standard deviation and Depression will go down .224 standard deviations. With just one variable in the model, the -.224 should also be the correlation coefficient between News Exposure and Depression.

Given that the significance value of this measure is < .05 we can yes, news exposure appears to be a significant predictor of depression.

Considering that the graph shows the errors are not normally distributed, we should seriously doubt the prediction reliability of this model. We should expect the errors to be clustered around zero, and instead there are too many low values. We can say the model seems to be influenced by a small number of cases.

1. Compare these two plots. What are they trying to show? Which looks more problematic? Why do we worry about this?





<<<<<<Type your answers after here. Make as much room as you need>>>>>

These two charts are showing the scattering of errors in 2 different models.

We expect the errors to be clustered evenly around zero for models that predict well across the entire range of dependent values.

In the top chart, the values are skewed negative, with residual values ranging 0 to -4, and positive values of >0 to 2. It looks slightly heteroskedastic, that at the high range of the dependent variable looks to have a higher range of errors.

In the second chart, we see the values are more even clustered around 0.

We can say this model seems to predict better across the range of predicted values.

However, the ordinal nature of the predictions does mean we see even amounts of spacing in the plot.

Compounding things, it seems some circles are darker than others, suggesting over-plotting and perhaps hiding some observations in the plot.

Perhaps the observations need to have jittering added (a random number assigned to its X and Y position) to more fully observe all observations uniquely.