Spring 2024 Term II – ANA 510 Statistical Methods for Data Science and Analytics

You have been given a data file and related publication about the data. A complete description of the data and values for selected fields are included in the related publication[[1]](#footnote-1). These data have been and can be used for many different types of analyses. For example, these data have been used for classification analyses. From the previous classification analyses you know that the dependent variable is "target," where target can be one of three categories; graduate, dropout, or enrolled. You will study a variety of methods for classification and the relationship between classification and (logistic) regression in another course.

Because you have the related publication you know that various methods have been used for classification because of the authors' perception that the data are skewed toward the category of "target" that comprises "graduate". However because of your Data Analytics program of study, you also know that these data appear to be skewed because of some apparent distribution to the data. You have already handled similar datasets and know (generally) what to do to build a model that accurately handles and describes this distribution.

# Part I – General ANA 500 Review Questions

Questions 1 through and including 40 are from the ANA 500 Foundations of Analytics review. These are easy and not intended to take a lot of time. However, there can be some confusion about terminology and acronyms. Therefore, I’ve put the following text to help you with this.

Ordinary least squares (regression) is, at its heart, concerned with how well we can build a linear model to fit the data. The acronyms SST, SSE, etc. are discussed further below. A simple linear (regression) model is given by the equation . It is based on minimizing the sum of the squared residuals, i.e. the square of the distance from the regression line to the respective observation. Much of how well we can build that model has to do with dispersion, or the variance inherent in the data. As you know, the variance of a sample is usually designated as “s”. Remember that Greek letters are typically used for populations while

Roman or Latin letters are used for samples. We compute the variance as

This shows that the variance is actually the average squared deviations from the mean. Note that in this equation the mean (or y-bar) is used. Compare this to the mean square error or MSE.

The mean square error or MSE is computed as

where the MSE is the average squared deviations from the “estimated” mean and there are degrees of freedom since we are estimating both .

The standard deviation is the square root of the variance. The standard deviation is a measure of the spread of a group of numbers from the mean. The standard deviation will have the same units of measure as the original data, e.g. sqft.

One of the problems with the ordinary least squares method, also known as the sum of squares method, is the confusion with the acronyms used. For example, the sum of squares total is usually known as SST (or TSS for total sum of squares). However, the acronym SSE can be used for either the sum of squares error or the sum of squares explained. Unfortunately, the sum of squares explained can also be known as SSR standing for sum of squares regression. But, SSR can also be used to designate the residual sum of squares which is actually the sum of squares error. Because of this confusion I’ll try to use the entire term and not an acronym for the values you’ll need to find. In general the total sum of squares is equal to the sum of squares explained plus the sum of squares error. To be thorough I’ll provide the following equations for you to use to build your ordinary least squares model.

Using SSE for sum of squares (residual) error or sum of squares residual.

Using SSR for sum of squares regression or sum of squares explained.

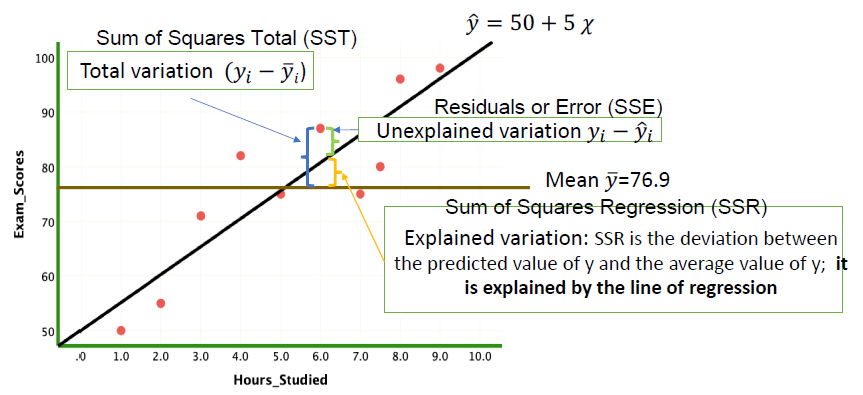
Using SST for total sum of squares.

Observed value

Value estimated, e.g. on the regression line

Mean value of a sample

This is shown on the figure from your VoiceThread lectures, Lecture 3, shown below.



Furthermore, we can find R-squared or R2 using the sums of squares, i.e. (NOTE! SST can also be written as .)

You can find the Pearson Correlation Coefficient by the following equations.

Questions 41 through and including 50 have to do with model building and “exclusion restrictions”. Specifically, these questions walk you through building a model for the mlb1 data set using a variety of parameters (independent variables) available. We’ve done this a lot already so it really should not be a problem. This time, however, we’ll use the output from the models built to look at a different way to determine which variables should remain in the model.

Remember that we have tested models in a number of different ways. In fact, gretl (if you use it) outputs the F statistic and its related P-value as part of its routine model output that we’ve used as an evaluation of how good a model is overall. The tests we’ve done include evaluating variables to determine which should remain in the model and which do not affect the dependent variable so can be removed from the model.

Now we’ll use this same output in a new way of computing the F statistic in order to consider whether a group of variables remain in a model or not. This method considers the original “complete” model to be an “unrestricted” model. That is, the model is complete because it includes all the variables we think should be in the model. Then, we remove a group of variables, perhaps these variables are all related in some way as are the “performance statistics” in the mlb1 dataset, and build a second model as a “restricted” model. This model is “restricted” in the sense that we do not include, or in other words we restrict, the group of variables in the model that we want to evaluate in terms of whether or not they affect the dependent variable.

Keep in mind that if a variable or a group of variables does/do not affect the dependent variable it/they are not doing us any good to keep in the model. I have included some background that you may want with respect to unrestricted versus restricted models in the script provided. In addition, you may want to go online to find more about this. I’ve done that myself and find that there is considerable difference in what different researchers are defining as the “degrees of freedom” associated with the unrestricted versus restricted models. See these two websites to learn more about this:

<https://www.econometrics.blog/post/understanding-the-f-statistic/#:~:text=The%20difference%20between%20them%20is,regression%20is%20free%20to%20set%20>.

<https://online.stat.psu.edu/stat462/node/135/>

Use the “degrees of freedom” as defined in the script for the Midterm Exam. Then, you can use whatever definition you want for degrees of freedom. I suggest that you be sure to include the way you are defining degrees of freedom in your report of an analysis to avoid confusion and maintain reproducibility.

# Part II Evaluating Regression Models

To complete this part of the Midterm Exam you will use the White Wines dataset in the whiteWines.gdt file. As always first, take a look at the data. Are there any missing values? Anything interesting or sort-of off when you consider the descriptive and/or summary statistics? Follow the directions in each of the questions to complete this part of the exam.

1. (Parts a-c) What is the dependent variable in this dataset? (Hint, if you are having trouble with this question think about what variable could be a function of all other variables.) What is the data type of this variable? Is the distribution for this variable normal?
2. (Parts a-b) Does multicollinearity exist? If so, which variables are highly correlated. Be sure to check all the variables that are highly correlated. (Remember that I use 0.6 and above as a rule of thumb for high correlation. And, don’t worry about pairing up these variables right now. Just check the variable if it is highly correlated with another variable.)
3. Generate an OLS model (Model 1) for the dataset using all independent variables as parameters in the model. This will be your “unrestricted” model so save the variables required for additional computations! Are the intercept and all coefficients statistically significant?
4. Based on the R-squared value for this model, does the model explain most of the variation in the data?
5. Check the variable or variables you found that are highly correlated and are NOT statistically significant.
6. Build a new OLS model for this dataset and include Volatile Acidity, Residual Sugar, Sulphates, pH, Free Sulfur Dioxide, Density, and Alcohol as independent variables. Now are the intercept and all coefficients statistically significant?
7. Did the value of R-squared improve?
8. Using some information published on the Internet to conduct feature selection (to choose which variables to use as independent variables), build an OLS regression model (Model 2) using Residual Sugar, Chlorides, Total Sulfur Dioxide, pH, and alcohol as the independent variables. Based on the value of R-squared, did the model improve in terms of explaining the variation in the data?
9. ~~It appears that some of the variables still included in the model do not help explain the variation in the data. Next, try this, now create a subset of the dataset in which you can consider the observations for Quality = 3 and/or 4.~~
10. It appears that some of the variables still included in the model do not help explain the variation in the data.  Or, as I like to say, "Something hinky is going on..."  Sample the dataset to find the observations that make the most difference to the quality of the wine.  That is, first create a subset of the dataset for observations where quality is greater than 7; and, second create another subset where quality is less than 5.  This should give you some idea of the variables differences between the worst and best wines.  Which variables appear to make the most difference to the quality of white wines?  Using Volatile Acidity, Citric Acid, Residual Sugar, and Free Sulfur Dioxide as independent variables build an OLS model.  Based on the value of R-squared, does the model explain more of the variation in the data? Or in other words, did the R-squared value show this model improved?
11. Using Volatile Acidity, Citric Acid, Residual Sugar, and Free Sulfur Dioxide as independent variables and the formula you used in Part I of this exam to compute the F statistic, evaluate whether or not all the variables in the unrestricted model should be included. Do all variables need to remain in this model?

Notes: This is a somewhat frustrating answer. It doesn’t appear that any set of variables performs better than just slightly being able to explain the variation in the data. However, part of this may be because we are using what is an ordinal categorical variable as the dependent variable for multivariable regression, not the best choices! You see, you can almost always get some output even if it isn’t very good.

# Part III Logistic Regression

The global financial crisis of 2007-2008 has highlighted the importance of transparency and rigor in banking practices. As the availability of credit has been limited, banks are increasingly tightening their lending systems and turning to machine learning to more accurately identify risky loans. In this case, machine learning (methods) include decision trees, neural networks, and more.

In this part of the Midterm Exam, you will explore whether or not logistic regression provides better results than machine learning. You can find more information about this dataset and other analyses conducted on it in the UC, Irvine Data Repository at [http://archive.ics.uci.edu/ml/datasets/South+German+Credit+%28UPDATE%29#](http://archive.ics.uci.edu/ml/datasets/South+German+Credit+%28UPDATE%29) and at RPubs at <https://rpubs.com/virag_l23/679613> and <https://rstudio-pubs-static.s3.amazonaws.com/270406_204e9868261b41edb7d8f78943276937.html>. Note that there are two data files available. The second is an update of the original file. I am uploading a publication about the dataset and the issues with it that you can refer to, “South German Credit Data: Correcting a Widely Used Data Set” by Grömping (2019). Using the creditData.gdt datafile, conduct logistic regression to identify risky bank loans.

As usual, look at your data. How many observations do you have (i.e. is it a small dataset)? How many variables do you have? What is the dependent variable? What data type is the dependent variable? To keep the number of questions low, I won’t ask about these things this time. You are responsible for carrying out this preliminary exploratory data analysis.

In fact, various researchers have used different sets of the parameters available. So, don’t look for others to have used these parameters to obtain their answers! In fact, I have spent several hours looking for an exact one-to-one solution for the specific variables I’ve chosen to confirm the model’s output. I have not been able to find such a reference (aargh!). So in the absence of better information, I believe that the dataset I have is miscoded and that in the dependent variable 0 = “good” or “no default” and 1=”bad” or “defaulted” rather than the other way around. That is, this appears to be the opposite of what some or even many researchers have used. But, if the coding in this dataset has been reversed then the results of the models match much better in terms of percentages predicted. Therefore, use 0 = “good” or “no default” and 1=”bad” or “defaulted” as the coding for the dependent variable!

Just to give you some insight without you having to read the papers/blogs on the websites mentioned before, the use of decision trees does not do a lot better than the multivariable regression for evaluating the quality of white wine did in Part II. Without boosting, only 8% of loans that defaulted were accurately predicted. The code used, the C5.0 algorithm did a much better job of predicting loans that would not default. When boosting was added, only 31 mistakes out of 900 observations were made. Let’s see how logistic regression compares to this.

1. Before anything else choose which parameters (independent variables) you will use to build your logistic regression model. To start, I will suggest (and the test is built around using) the following as parameters:

* AccountBalance
* DurationOfCredMonths
* PaymentPerCredit
* CreditAmount
* ValueSavingsStocks
* ConcurrentCredits
* NoOfDependents

What is the best type model for these data?

1. Build a logistic regression model, Model1, and use its output to answer the related questions. Are the intercept and all the coefficients statistically significant?
2. What percentage of defaults did the model accurately predict? (Note that this is only for defaults = 0. This can be compared to the results using the C5.0 algorithm without that resulted in 16% accuracy.)

The results for logistic regression are not quite as good as the machine learning/decision tree results were (16% by the author I am currently looking at). On the other hand, the decision tree model was built using all variables whereas the variables used in the logistic regression model were a best guess, so not really too bad overall for the first attempt.

1. Out of 1000 observations, what are the number of cases “correctly predicted” by Model1?
2. One of the problems with current credit rating systems is that they tend to not be very reliable in terms of predicting defaults, a Type II error. Output from Model2 indicates that \_\_\_\_\_\_ instances were predicted to be “good” or “no default” when in fact those instances had defaulted.
3. The problem with Type II errors can be mitigated by decreasing the number of observations.
4. Type II errors can be decreased by increasing the level of significance. Of course, then the number of Type I errors will increase.
5. Given the output from Model2, the coefficients for the variables AccountBalance, ValueSavingsStocks, and ConcurrentCredits indicate that as a loan applicant’s account status, the value of his/her savings, and the less he/she currently owes on other credit accounts indicates his/her credit worthiness will decrease.
6. Now look at the average marginal effects (AME) and some probabilities, i.e. the probability that a debtor will default given some very basic data to consider. (Note that the fact a value is output for the AME with respect to the intercept is meaningless.) Since we will reuse some of the functions from the script from PS3 and add a bunch more functions, I’ll try to keep this simple. The dependent variable Creditability will be a function of the intercept and the independent variables AccountBalance, CreditAmount, ValueSavingsStocks, and ConcurrentCredits. The logic is that the more ability a debtor has to repay the less trouble he/she would have repaying a loan. And, the more the debtor borrows the more trouble he/she would have repaying that amount, etc.

Using the script provided, build a simpler logistic regression model, Model2, using the variables listed above. What is the average marginal effect of a loan applicant’s current bank account balance, i.e. AccountBalance?

1. The average marginal effect for AccountBalance means that a unit increase in a loan applicants accounts status will result in a 0.11 increase in the applicant’s credit worthiness or Creditability. (This is where the miscoding gets really sticky. This result is consistent with my discussion about potential miscoding above and I think consistent with results by other researchers.)
2. Considering average marginal effects again, if a loan applicant requests a higher loan amount that will have a nearly negligible effect on an applicant’s credit worthiness or Creditability.
3. Now, considering the 95% confidence intervals, we can be 95% certain that within the true population a loan applicant will likely not currently have a checking or savings account.

1. Realinho, V., Machado, J., Baptista, L., and Martins, M.V. (2022) “Predicting Student Dropout and Academic Success”. ***Data 7***, 146. DOI: 10.3390/data7110146. [↑](#footnote-ref-1)