ANA 510 Spring Term II 2024

Final Examination

This exam has XXX parts. There are 1,000 points available for your raw score that when you final score is computed will be scaled to 100 possible.

# Part I – Sum of Squares / Least Squares Regression and Goodness of Fit

I am using notation as follows:

* SST or or is the total sum of squares
* SSE or is the explained sum of squares
* SSR or is the residual sum of squares

Using the milner dataset provided, answer the following questions. Develop a “restricted” simple linear regression model using Sales (Sales) as a function of Advertising Expense (AdvExp); i.e. Sales is the dependent variable, Advertising Expense the independent variable or regressor. You can use any program to complete this. I set this up in an Excel spreadsheet with the following columns in order to more quickly manipulate the data.



1. What is SST?
2. What is SSE?
3. What is SSR?
4. What is R2?
5. What is the adjusted R2?
6. What is the MSE (mean square error)?
7. What is the standard deviation of the independent variable?
8. Establish a confidence interval for the slope. The lower boundary is \_\_\_\_\_\_ and the upper boundary is \_\_\_\_\_\_ .

Now, develop an “unrestricted” multivariable linear model using Sales as a function of Accounts, AdvExp, Poten, and Share. A description of these variables is in the milner-var file. Remember that an F-test can be conducted using SSR-unrestricted and SSR-restricted in the appropriate formula.

1. Conduct a goodness of fit test to determine the veracity of the variables in the unrestricted model. What is the value of the F-statistic?
2. What is the value of F-critical?
3. Based on your goodness of fit analysis, do all the variables in the unrestricted model need to be in the final model? Yes/No

# Part II – EDA and Feature Selection Using the Pima Indians Diabetes Dataset

## About the Pima Indians Diabetes Dataset

### Context

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

### Content

The datasets consists of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

### Acknowledgements

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). [Using the ADAP learning algorithm to forecast the onset of diabetes mellitus](http://rexa.info/paper/04587c10a7c92baa01948f71f2513d5928fe8e81). *In Proceedings of the Symposium on Computer Applications and Medical Care* (pp. 261--265). IEEE Computer Society Press.

(Retrieved from Kaggle at <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database> December 6, 2022.)

The Pima Indians diabetes dataset is a well-known dataset and is used in many different types of courses in data science, particularly in machine learning. For example, rather than conducting a logistic regression you will find this dataset used in classification studies to classify whether or not someone in the dataset has diabetes. These methods of conducting an analysis are obviously closely related. Artificial Neural Networks have been developed to analyze this dataset too. Words of warning, the specific question on this exam were developed by me for this exam. Don’t spend too much time looking for information on the Internet. You have everything you need from prior assignments and the midterm exam to complete this part of the Final.

Using the Pima Indians diabetes dataset, consider “feature selection” as one way of reducing the dimensionality of a dataset and determining which independent variables should be included in a model as well as which should be removed. A nice discussion of feature selections methods is included in the paper “A survey on feature selection methods,” by Chandrashekar (Chandrashekar, 2014). One of the most common feature selection involves the use of the Pearson correlation coefficient. In effect, independent variables that are highly correlated with the dependent variable are considered important. Independent variables that are highly correlated with another independent variable are considered for elimination. The trick is to establish exactly what is meant by “highly correlated”. We’ll look at this in the questions that follow.

Since considering correlation between variables is usually done when you first look at your dataset, or when you are conducting exploratory data analysis (EdA), do that to answer the following questions.

1. The number of observations in the dataset is \_\_\_\_\_\_ and the number of independent variables is \_\_\_\_\_\_ .
2. \_\_\_\_\_ is the dependent variable in the diabetes dataset.
3. Is this variable binary? Yes/No
4. What is the proportion of Pima Indians in the diabetes dataset that have diabetes (as a percentage, rounded)?
5. What is the average measured level of Glucose of all those included in the diabetes dataset?
6. Based on the data type of the dependent variable, which type of regression will you conduct?
   1. Linear
   2. Log-linear
   3. Log-quadratic
   4. Logistic / Logit
   5. Multivariable Linear
7. The most highly correlated independent variables are \_\_\_\_\_ and \_\_\_\_\_ with a Pearson correlation coefficient of \_\_\_\_\_\_.
8. The next most highly correlated independent variables are \_\_\_\_\_ and \_\_\_\_\_\_ with a Pearson correlation coefficient of \_\_\_\_\_\_ .
9. \_\_\_\_\_\_ is the independent variable most highly correlated with the dependent variable with a correlation coefficient of \_\_\_\_ .
10. Of the independent variables age and pregnancies, since both have been associated with the onset of diabetes is it a good idea to eliminate either even though they are the most highly correlated independent variables? Yes/No
11. Which, if any, of the following variables may not meet one or more of the typical assumptions or requirements of regression, e.g. normality?
    1. Pregnancies
    2. Glucose
    3. BloodPressure
    4. SkinThickness
    5. Insulin
    6. BMI
    7. DiabetesPedigreeFunction
    8. Age
    9. Outcome

# Part III – Logistic Regression Using the Pima Indians Diabetes Dataset

The questions below continue from the prior section on EDA and Feature Selection.

1. Build the appropriate model using the diabetes dataset with Outcome as the dependent variable and Glucose as the independent variable. Based on this model, what number of cases was correctly predicted?
2. Which of the following variables are statistically significant at the 0.05 level?
   1. Pregnancies
   2. Glucose
   3. BloodPressure
   4. SkinThickness
   5. Insulin
   6. BMI
   7. DiabetesPedigreeFunction
   8. Age
   9. Outcome
3. What is the average marginal effect for change in Glucose? (Since we are going to be talking about very small numbers round your answer to four decimal places here.)
4. What is the standard deviation for this average marginal effect? (Since we are going to be talking about very small numbers round your answer to four decimal places here.)
5. Now build the appropriate model using the diabetes dataset with all variables. Based on this model what number of cases was correctly predicted?
6. What is the average marginal effect for change in Glucose now? (Since we are going to be talking about very small numbers round your answer to four decimal places here.)
7. What is the standard deviation for this average marginal effect? (Since we are going to be talking about very small numbers round your answer to four decimal places here.)

# Part IV – Forward-Selection and Backward-Elimination Procedures Using the Pima Indians Diabetes Dataset

Using the Pima Indian diabetes dataset again, conduct a forward-selection procedure to determine a “best” regression model for this dataset using this method. Last, generate possibly another “best” regression model for this dataset using the backward-elimination procedure. You can use any program to complete this part of the Final examination. Use your results from feature selection, forward-selection, and backward-elimination to answer the following questions.

The following are all “fill-in multiple blanks” questions beginning with feature selection.

## Forward-Selection Procedure

1. Using the forward-selection procedure, the independent variable \_\_\_\_\_\_ can be used to build the “best” single variable model.
2. Feature-selection using correlation coefficients and the forward-selection procedure both produce the same single variable model. True/False
3. There are \_\_\_\_\_ independent variables that should be included in the final forward-selection regression model.
4. The first choice of independent variables is \_\_\_\_\_ with a t-ratio of \_\_\_\_\_.
5. The second choice of independent variables is \_\_\_\_\_ with a t-ratio of \_\_\_\_\_.
6. The third choice of independent variables is \_\_\_\_\_ with a t-ratio of \_\_\_\_\_.
7. The fourth choice of independent variables is \_\_\_\_\_ with a t-ratio of \_\_\_\_\_.
8. The fifth choice of independent variables is \_\_\_\_\_ with a t-ratio of \_\_\_\_\_.
9. From the final forward-selection regression model, the coefficient of the first choice is \_\_\_\_\_.
10. From the final forward-selection regression model, the coefficient of the second choice is \_\_\_\_\_.
11. From the final forward-selection regression model, the coefficient of the third choice is \_\_\_\_\_.
12. From the final forward-selection regression model, the coefficient of the fourth choice is \_\_\_\_\_.
13. From the final forward-selection regression model, the coefficient of the fifth choice is (rounded to three decimal places) \_\_\_\_\_.
14. From the final forward-selection regression model, the coefficient of the intercept is \_\_\_\_\_.

## Backward-Elimination Procedure

1. There are \_\_\_\_\_ independent variables that should be removed for the final backward-elimination regression model.
2. For the first round of backward-elimination, the independent variable to remove is \_\_\_\_\_ with a t-ratio of \_\_\_\_\_.
3. For the second round of backward-elimination, the independent variable to remove is \_\_\_\_\_ with a t-ratio of \_\_\_\_\_.
4. For the third round of backward-elimination, the independent variable to remove is \_\_\_\_\_ with a t-ratio of \_\_\_\_\_.
5. Which of the following (independent) variables are statistically significant in both the forward-selection and backward-elimination final models?
   1. Pregnancies
   2. Glucose
   3. BloodPressure
   4. SkinThickness
   5. Insulin
   6. BMI
   7. DiabetesPedigreeFunction
   8. Age
   9. Outcome

# Part V Model Building and Pitfalls

We’ve looked at model building from several perspectives, many included pitfalls or challenges to model building. Use the EmployeeSalaries in .csv format to answer the following questions. Note, the EmployeeSalaries dataset was obtained from Kaggle at <https://www.kaggle.com/datasets/anninasimon/employee-salary-dataset> on December 6, 2022. You can visit that webpage to get full details of the dataset and citation information. This dataset was selected because data about certain things, e.g. wages, prices, etc., typically have a “certain” distribution. We’ve looked at this in your assignments in ANA 510. Just apply what you’ve learned from those assignments to these questions!

1. Is the EmployeeSalaries dataset sufficiently large to conduct an analysis? Yes/No
2. What is the dependent variable in the EmployeeSalaries dataset?
3. What are some of the problems with meeting the assumptions of regression you can determine by conducting some exploratory data analysis (EDA) such as considering correlations, creating scatter plots, etc.? (Note that you will also want to also do some preliminary ordinary least squares regressions to consider all choices.) Select all the choices below that reflect your observations.
   1. Multicollinearity
   2. Heteroscedasticity or Heteroskedasticity
   3. Linearity
   4. Normality
   5. Independence
   6. Lack of accuracy in the data
   7. Lack of precision in the data
   8. Sampling errors in the data
4. What would be your first approach to correct some of the problems with the EmployeeSalaries dataset? Select all the “best” choice below that reflect your understanding of these approaches and are based on your previous exploratory data analysis, etc.
   1. Quit, there is no way to correct data this bad.
   2. Refer the data back to its author with a request for better sampling.
   3. Take the log of the independent variable or variables.
   4. Take the log of the dependent variable.
   5. Square the independent variable or variables.
   6. Square the dependent variable.
5. First, build an OLS model using the dependent variable and the independent variables Exper and Age. In addition, try a typical approach to improving OLS models, i.e. build an OLS model of the EmployeeSalaries data using the log of the dependent variable and compare that to an OLS model. How much has the value of R-squared improved? (Simply take the difference between the two values of R-squared and round to two decimal places.)
6. But has the model really been improved? Consider how the q-q plot has changed from the original OLS model to the model using the log of the dependent variable. Consider the center of the q-q plot. Does it show improvement? Yes/No
7. Using the same output as Q52, Consider the tails of the q-q plots. Does using the log of the dependent variable improve the behavior of the tails? Yes/No
8. Using a forward-selection procedure, which independent variable(s) \_\_\_\_\_\_ should be included in your final model for the EmployeeSalaries dataset?
   1. Age
   2. Experience\_Years
   3. Both
   4. Neither
9. In the forward-selection process, the t-ratio value for Age is \_\_\_\_\_\_ . 8.47
10. In the forward-selection process, the t-ratio value for Experience\_Years is \_\_\_\_\_\_. 8.06

## Two-Stage Least Squares Regression Models

1. Generate a TSLS model for the EmployeeSalaries dataset using the log of the dependent variable, the log of the independent variable you found using the forward-selection procedure, and use the remaining independent variable Age as an instrument . Is the independent variable statistically significant? Yes/No
2. Using the output from the model described in Q59, does the q-q plot appear to be improved? Yes/No
3. Select the TSLS model below that you have found “best” fits the EmployeeSalaries dataset.
   1. OLS model without modifying any variables
   2. OLS model using the log of the dependent variable
   3. TSLS model using the log of the dependent variable, the log of the independent variable selected from the forward-selection procedure, and the remaining independent variable as an instrument
   4. TSLS model using the log of the dependent variable, the log of the independent variable Exper, and the independent variable Age as an instrument.

We have not really proven too much in this section. It has provided you with some exercises to help build your intuition about how to work with variables to build better models. One of the issues with this particular dataset is that age and experience are highly correlated. As an employee ages he/she acquires more experience. In any event, it isn’t sufficient to just consider part of the results of an analysis such as the value of R-squared. You must also consider how much the overall model is improved using a variety of tests.

This part of the exam uses a completely different dataset, i.e. the SchoolingReturns dataset. If is attached as an Excel file for you to download and use. The variables are identified in the table below as:

|  |  |
| --- | --- |
| 1 | wage {Raw wages in 1976 (in cents per hour).} |
| 2 | education {Education in 1976 (in years).} |
| 3 | experience {Years of labor market experience, computed as \code{age - education - 6}.} |
| 4 | ethnicity {Factor indicating ethnicity. Is the individual African-American (\code{"afam"}) or not (\code{"other"})?} |
| 5 | smsa {Factor. Does the individual reside in a SMSA (standard metropolitan statistical area) in 1976?} |
| 6 | south {Factor. Does the individual reside in the South in 1976?} |
| 7 | age {Age in 1976 (in years).} |
| 8 | nearcollege {Factor. Did the individual grow up near a 4-year college?} |
| 9 | nearcollege2 {Factor. Did the individual grow up near a 2-year college?} |
| 10 | nearcollege4 {Factor. Did the individual grow up near a 4-year public or private college?} |
| 11 | enrolled {Factor. Is the individual enrolled in college in 1976?} |
| 12 | married {factor. Is the individual married in 1976?} |
| 13 | education66 {Education in 1966 (in years).} |
| 14 | smsa66 {Factor. Does the individual reside in a SMSA in 1966?} |
| 15 | south66 {Factor. Does the individual reside in the South in 1966?} |
| 16 | feducation {Father's educational attainment (in years). Imputed with average if missing.} |
| 17 | meducation {Mother's educational attainment (in years). Imputed with average if missing.} |
| 18 | fameducation {Ordered factor coding family education class (from 1 to 9).} |
| 19 | kww {Knowledge world of work (KWW) score.} |
| 20 | iq {Normed intelligence quotient (IQ) score} |
| 21 | parents14 {Factor coding living with parents at age 14: both parents, single mother, step parent, other} |
| 22 | library14 {Factor. Was there a library card in home at age 14?} |

1. How many observations are there in the SchoolingReturns dataset?
2. How many total variables are there in the SchoolingReturns dataset?

To keep things simple we will just use the first eight variables in the SchoolingReturns dataset.

1. What is the dependent variable in the SchoolingReturns dataset?

Let’s do some EDA on this dataset to see what we have.

1. Is the dependent variable, wage, right skewed as expected? Yes/No
2. What is the mean age?
3. Based on the summary statistics for “nearcollege” would you say:
   1. More live near a college.
   2. Fewer live near a college.
   3. Cannot tell anything about “nearcollege”
   4. I’m not sure how to tell anything about “nearcollege” from these statistics
4. Based on the output from the OLS and TSLS models, the percent returning to school as given by the variable “education” is:
   1. Realistic for the TSLS Regression model
   2. Realistic for the OLS Regression model
   3. Realistic for both models
   4. Not realistic for either model
5. The percent returning to school given by the variable education in the TSLS Regression model is \_\_\_\_\_\_. Multiply by 100 and round to two decimal places to report as a percentage.
6. Can you detect a **significant improvement** in the q-q plots from the OLS model to the TSLS Regression model? Yes/No

We know that we have corrected the skew in the dependent variable. We know that we have properly accounted for at least some of the endogeneity in the data. Given time we could do a better job of model building, don’t you think? There is actually a lot more to learn about Two-stage Least Squares Regression. And, you can play around with these datasets or use them as you want.

I hope that this has given you a sense of the different methods for building models using statistics and causal inference. I also hope that you’ll save these gretl scripts. Remember that gretl is both open source and based on the C programming language. You may find the syntax useful in the future! You’ll find many more ways to build models when you take ANA 540 Advanced Data Analytics which covers machine learning.

Best,

Marvine