ECON 310: Course Notes

Scott W. Hegerty
Department of Economics
Northeastern Illinois University
Chicago, IL 60625
S-Hegerty@neiu.edu

April 19, 2022

1 Introduction

This document summarizes some of the concepts and material for Economics 310, *Business and Economic Statistics II*, which is offered for the second eight weeks of the Spring 2022 semester at NEIU. As a writing-intensive course, there are three main goals.

First, a set of new statistical techniques are introduced. These are commonly used in a number of disciplines–and by me, so I use some examples from my own work. Second, students will practice writing for different audiences, as well as giving and incorporating feedback to improve their (and others') writing. Third, students will have the opportunity to add to their professional toolkit, including learning new software, writing professional documents, and improving their professional profile. They will have the option to use the *R* software, apply LaTeXor *RMarkdown*, or set up a GitHub page, but these last ones are not required.

These notes will be updated throughout the semester–their final form will only be known after the material is actually covered. In particular, more might be added regarding the statistical tools.

1.1 Prerequisites and Expectations

Students in this course are expected to have completed an introductory statistics course at the level of NEIU's Economics 220, *Business and Economic Statistics I*, or the equivalent. They should be familiar with summary statistics such as mean and standard deviation, measures of association such as correlation, basic distributions such as the Normal and t-distributions, and hypothesis testing. We will cover a little bit more about regression analysis in this class, since that does not always get covered in the intro courses.

Students should also be willing to spend time working with data and writing polished reports. The primary aim of this course is to incorporate solid data analysis and interpretation with well-written, professional English. Either or both of these components might take you longer than you think. But the vast majority of the work in this class is in the preparation and editing of three projects–there aren't really any lecture-heavy tests. So plan accordingly!

1.2 Statistical Software

The primary software package that I will use is *R*. This has a huge variety of applications, and is widely used in the field. Students with good *R* skills will have an advantage on the job market; since it is free to use, it has surpassed many of the paid packages. But you have to learn the coding and practice a lot to get good at it.

As another (free) alternative, I also allow (and may briefly show how to use) the software *gretl*. This has drop-down menus and might look familiar to students without a programming background. I have "solved" all the assignments using both packages.

Students who know Python, Eviews or Stata (all of which are also widely used) might be able to do their assignments using their preferred software, but I might need more supporting material if you don't get the same answers I do.

1.3 Readings

One decent open-source introductory text is *OpenStax Introductory Statistics* by Barbara Illowsky and Susan Dean. It is freely available online at: openstax.org/details/books/introductory-statistics.

If you go through the NEIU Library, Springer ebooks are free. I don't really lecture out of these, but they are good for the underlying math:

Thomas W. MacFarland and Jan M. Yates. 2018. *Introduction to Nonparametric Statistics for the Biological Sciences Using R* (1st. ed.). Springer Publishing Company, Incorporated.

Everitt, Brian, and Torsten Hothorn. 2011. *An Introduction to Applied Multivariate Analysis with R*. New York: Springer.

I may base my own lectures out of these that I have on my shelf:

Myles Hollander and Douglas A. Wolfe. 1999. *Nonparametric Statistical Methods*. Chichester: Wiley.

Alvin C. Rencher. 2002. *Methods of Multivariate Analysis*. New York: Wiley.

1.4 Supplementary Materials

In addition to these readings, I have material both on the GitHub site for this course (https://github.com/hegerty/ECON343 and two others:

Economics 343, *Macroeconomic Data Analysis*
https://github.com/hegerty/ECON343

Economics 346, *Applied Economic Statistics Using R*
https://github.com/hegerty/ECON346

ECON 343 has materials on writing reports, producing documents, and laying out graphics and text, and ECON 346 has material on basic statistics and using R–including an introduction to regression analysis.

2 Tools for the Course

2.1 Software

I mention a few software packages above, but a good professional skillset in Economics requires software competency. One way to know which software to learn is to look at job postings for positions you'd like. But in general, *R* and Python are worth learning–some people try to good at both, but I'd rather put time into getting *really* good at one and maybe know a little bit about the other. Of the two, *R* is more suited for statistics and econometrics, while Python is a good "all around" software package and is used a lot in machine learning.

SAS is still used a lot in business, but it is expensive and not used in academia. As a result, it is hard to learn in a school setting. I personally think the first step–learning to code at all–makes picking up a second programming language a lot easier.

In general, "drag-and-drop" software packages are not as useful on your résumé as programming-based ones. If you put down "gretl," employers might not really care. The same goes for better-known ones such as SPSS. Even if you use Eviews, people might want you to script commands. Remember that your résumé is part reflecting actual knowledge, and part *signaling*, so that you show you can learn and that your are worth hiring.

Practicing data analysis takes time. You might spend an hour on something easy (as I did the other day, trying to make a .bbl file for a LaTeX reference page!). Anything that seems like a waste of time now will help you get better at the same thing next time. So don't cut corners, especially if you are just learning to code.

2.2 Writing

Writing a professional report involves taking the statistical analysis you have worked on and summarizing it in a way that meets your and the reader's goals.

You need to know your audience, know which results to include and how to explain them, organize everything effectively, and write well.

Your audience will differ in terms of statistical training, background, and even political leanings. That will help drive your writing. People who really know stats will want to see if your methods are good. Others might just want to see what you, the economist, used your skills to come up with. It is tricky to figure you the right spot between too basic and too analytical. Some reports have side boxes or appendices where they put formal models.

You *never* need to include every single statistics result you came up with. Even within a single table–I always say in other classes never to include raw regression output. You need to know which results matter. One way to learn is to look at other professional papers and reports.

You also might run way more analyses than you need. Some are just to look at the data and to get an idea of what you are doing. You might try different versions of a model. Or, you might completely change direction and not use something just because. Ignore the "sunk cost fallacy" and don't feel like you need to use it anyway. One thing I do is to wait a day and look at my Results section–just the graphs and tables. If I can't remember why they are important, they probably don't need to be there.

One big role for an economist is explaining results in a way that your audience understands. You have to take what is there, package it, and come up with some sort of actionable conclusion. People fail to do these in different ways.

Sometimes people see what they want to see. Is a statistic significant? Sometimes it's not (or it is, but at like 10 percent), but that would make the whole argument fall apart. Or it is statistically significant, but the actual number is so small it has no practical meaning. Or, if something is significant, but the "opposite" of what you expected, you can't just pretend it didn't happen.

Once you have a set of results, you have to present them in a way that helps your audience. What is the best way to do this? It could be more visual, or you might want to report tables. I talk about this, as well as how to make good figures and tables, in ECON 343. You also have to verbally "tie together" everything into a cohesive narrative.

One big mistake is when people just list out what they found. You have to explain what everything means, and what to do with it. If you get different results for a bunch of different countries, you should be able to offer some sort of explaination why. Or at least acknowledge the differences. Sometimes a writer comes off like they couldn't even find them on a map. You might have to do some background research to do so convincingly.

Your writing should be your best, even on a "rough draft." Editors like to focus on bigger things and not have to correct misspellings and things that should have been taken care of beforehand. In business, there might be a professional editor who makes corrections or even formats your document. In that

case you have to be extra-careful that it ends up looking the way you want it to. A draft might also circulate to different departments (including Legal!) to make sure everything is approved. In an academic setting, it might just be you doing the editing. You may need to consult a grammar reference, have a friend/relative look it over, and (if it's super-important, like a dissertation) even hire a professional editor. One easy trick is to set aside your draft for a day and then look it over one last time.

Don't be afraid to rewrite a section, even entirely, even though it might hurt to throw that work away. Sometimes sections can be combined to be more concise; you might even see parts get repeated. Be sure to watch for abbreviatuions and "jargon" and make sure the words make sense. Also make sure you attribute your material appropriately.

2.3 Other Professional Tools

If you are on the job market, you will need to "package" yourself as professionally as possible, so that you stand out and get noticed. Some economists think their skills will sell themselves, but usually no one is that special and everyone still needs to interview.

A good résumé is important. Make sure that it lists real skills–including specific statistics courses–in addition to your degree. You can highlight some non-Economics jobs, but don't overdo it. If you had a cashiering job, point out the responsibilities you had, such as being entrusted with money. But you don't need to say "Duties: Worked the cash register." Also make sure any degree you list is actually correct: If you get a Bachelor of Arts in Economics, it isn't a "Bachelor's of Arts in International Economics."

You may wish to have website–but if you do, make sure it looks at least remotely professional. It might be worth purchasing a domain name. You can link that to a Google Sites page; a lot of statistics professionals put their site on a GitHub page, along with their code and data.

You might also want to write reports in LaTeX. I have written elsewhere that it mostly serves as a signal, since it is far less efficient than simply using Word. It is good, however, if you have a lot of equations. But in many mathematical disciplines, you will not be taken seriously if you use Word. My advice is if you are serious about learning to code, it's not that much extra work to learn this software.

3 Statistical Concepts

This is a very general overview; You should read the recommended texts and also use the .R files and watch the videos.

3.1 Nonparametric Methods

Traditional statistics are *parametric* they have parameters that serve as restrictions. One big restriction is that data are supposed to be normally distributed. If this isn't the case, any inferences might be incorrect. Nonparametric measures have fewer assumptions. In that way, they are simpler. But they often require a lot of computation to calculate. This used to be more of a challenge, but your laptop is easily able to do these. A lot of nonparametric measures use associations between observations, and look at associations such as differences and ranks.

For example, the nonparametric version of the (Pearson) correlation is the Spearman ρ, which takes the rank of X and the rank of the corresponding Y. If the #1 ranked X is paired with the #1 ranked Y (and so forth), the differences between ranks will all be zero, so they sum to zero. 1 minus this sum would give a perfect correlation. If correlation is not perfect, the ranks do not need to be exactly equal, but ρ will still be high.

$$1 - \frac{\sum_{i=1}^{n} 6D_i^2}{n(n^2-1)} \tag{1}$$

Here, D_i is a difference in ranks between R_1, which is X's position from low to high, and S_i, which is Y's position. If ranks #4 and #5 for X or Y have the same value, these "ties" are both assigned a rank of 4.5. If the #1-ranked X and the #4-ranked Y go together, then D equals -3. This is done for all pairs. Hollander and Wolfe (1999, pp. 394-397) provide more detail.

Kendall's τ captures a similar idea in a different way. Here, every possible pair of (X, Y) combinations in compared. If both X and Y increase or decrease, the value of Q will be positive. If they go in opposite directions, Q will be negative. Lots of matches will lead to a large sum.

$$Q((X_i, Y_i), (X_j, Y_j)) = \begin{cases} 1, & if Q((X_i, Y_i), (X_j, Y_j)) > 0 \\ 0, & if Q((X_i, Y_i), (X_j, Y_j)) < 0 \end{cases} \tag{2}$$

This gives the statistic

$$K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q((X_i, Y_i), (X_j, Y_j)) \tag{3}$$

Similar concepts covered in this class include nonparametric ANOVA (the Kruskall-Wallis test for differences in group means) Here, all N observations from all k samples are combined, with each element X_i given its particular ranking $r_i j$ amoing the joint sample.

$$R_j = \sum_{i=1}^{n_j} r_{ij} \tag{4}$$

These are summed for each of the k groups. If H is significantly large, you reject the null hypothesis of no differences among groups.

$$H = (\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j}) - 3(N-1) \tag{5}$$

This follows a χ^2 distribution with $k-1$ degrees of freedom.

I also mention nonparametric bivariate regression, which has some good (albeit limited) uses. True nonparametric regression is much more sophisticated.

The Thiel-Sen estimator takes the $N = n(n-1)/2$ possible slope values for the equation $Y = \alpha + \beta X + \epsilon$:

$$S_{ij} = (Y_j - Y_i)/(X_j - X_i) \tag{6}$$

and then takes the median value for its estimate:

$$\hat{\beta} = \ median \ \{S_i j, i \leq i < j \leq n\} \tag{7}$$

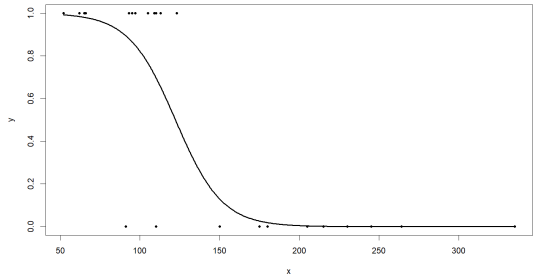
To get the intercept α, you use the median values for the previous equation as follows:

$$\hat{\alpha} = median \ \{ \ Y_i - \hat{\beta} X_i \} \tag{8}$$

3.2 Logistic Regression

Logistic regression is different from OLS because the dependent variable Y is only either 0 or 1. The goal is to see what variable explain which group an observation can be classified into. Similar methods are commonly used, but there is a slightly different interpretation than when Y is "continuous."

Figure 1: Logistic Regression Curve.



Generated using provided R dataset *mtcars*

The basic equation is written as:

$$Pr(Y = 1|X_1, X_2, ..., X_k) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)}} \quad (9)$$

The "logistic curve" goes from "no" (0) values to "yes" (1) values smoothly.

Besides interpreting the basic output, economists often look at "marginal effects." But in this class, where many students need to understand basic OLS first, we will focus on sign and significance.

Remember that regression output gives the estimated coefficient as well as three related values. The main one is the standard error; a coefficient needs to be at least two standard errors away from zero if it is to be considered significant. The t-statistic basically shows how many s.e.'s the coefficient is from zero, calculated as $t = \hat{\beta}/se$. The *p-value* represents the fraction of the t-distribution that lies beyond the t-statistic. Since there won't be much left past a large t-value, the corresponding t-statistic will be very low.

For example, in an estimation of the equation $Y = \alpha + \beta X + \gamma Z + \epsilon$ (from ECON 346) shows that the intercept and X have significantly positive effects on Y, while Z has no significant effect.

Figure 2: OLS Regression Output.

```
Call:
lm(formula = y ~ x + z)

Residuals:
    Min      1Q  Median      3Q     Max
-423.43  -87.22   -2.74   98.02  474.81

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9.447     11.299  -0.836  0.40328
x             12.693      4.141   3.065  0.00224 **
z              1.459      2.078   0.702  0.48272
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 135.3 on 997 degrees of freedom
Multiple R-squared:  0.01005,   Adjusted R-squared:  0.008067
F-statistic: 5.062 on 2 and 997 DF,  p-value: 0.006496
```

Generated using the provided R file

The intercept and X have relatively small standard errors, so the t-statistics are large, and the corresponding p-values are practically zero. But since the standard error is (coincidentally) the exact same size as the coefficient estimate,

the t-statistic equals 1. Because the 95% confidence interval (±2 standard errors) crosses zero, we cannot say that the coefficient is positive. As a result, the p-value is far too high to claim significance.

One other note: When you make your own regression table, the primary numbers to take from this table are the coefficient; only one of the s.e., t-statistic, or p-value (and clearly state which one), R-squared; and maybe N. You should format your own tables, and never just copy/paste regression output.

Figure 3: Logistic Regression Output.

```
glm(formula = BANKCOUNTB ~ PERCVAC + PERC25K, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7735  -0.7499  -0.5454  -0.3826   2.3205

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.053575   0.089597 -11.759  < 2e-16 ***
PERCVAC     -0.014824   0.001922  -7.711 1.25e-14 ***
PERC25K     -0.002797   0.007041  -0.397    0.691
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2035.5  on 2140  degrees of freedom
Residual deviance: 1943.2  on 2138  degrees of freedom
  (5 observations deleted due to missingness)
AIC: 1949.2
```

Generated using the provided R file

The R output for a Logistic regression looks similar. Here, I use real data for Chicago block groups and bank locations, and assign the value 1 for block groups that contain at least one bank and 0 for the rest. I use census data for the vacancy rate and the percentage of households making less than $25,000 per year. The idea is that areas with higher vacancies or poorer residents, all else equal, will be less likely to have a bank. Only the vacancy rate plays a significant role. I used the standard *glm()* command for this estimation.

I found an alternative package, *rms*, which provides a number of diagnostic statistics. Note that the estimates are the same. One of the diagnostics is "pseudo-R^2." This is based on the ratio of two log likelihoods–one for the full mode, and one with the intercept:

$$R^2_{pseudo} = 1 - \frac{log(L_{full})}{log(L_{inpt})} \tag{10}$$

Generally, "model fit" involves minimizing statistics such as log likelihood or the Akaike Information Criterion. So a better model fit means a relatively low LL_{full} and a large value when subtracted from one. But this statistic is not very high is this model. In class, we estimate other specifications as well.

Figure 4: Logistic Regression Output Using the "rms" package.

```
> lrm(BANKCOUNTB~PERCVAC+PERC25K,data=data)
Frequencies of Missing Values Due to Each Variable
BANKCOUNTB    PERCVAC    PERC25K
         0          5          5

Logistic Regression Model

 lrm(formula = BANKCOUNTB ~ PERCVAC + PERC25K, data = data)


                       Model Likelihood    Discrimination    Rank Discrim.
                             Ratio Test           Indexes          Indexes
 Obs          2141    LR chi2      92.27    R2       0.069    C       0.626
  0           1750    d.f.             2    g        0.650    Dxy     0.252
  1            391    Pr(> chi2) <0.0001    gr       1.916    gamma   0.253
 max |deriv| 2e-13                          gp       0.082    tau-a   0.075
                                            Brier    0.144

           Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept -1.0536 0.0896 -11.76 <0.0001
 PERCVAC   -0.0148 0.0019  -7.71 <0.0001
 PERC25K   -0.0028 0.0070  -0.40 0.6912
```

Generated using the provided R file

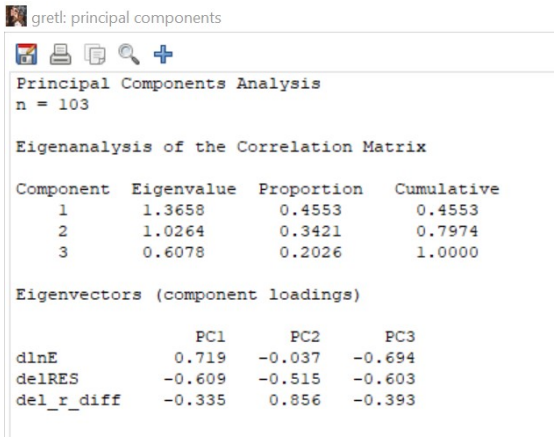
3.3 Principal Component Analysis

PCA is related to *Factor Analysis*, which classifies variables by statistical similarities. A company could give a 30-question customer-service survey, for example. Some of the questions might get at the same thing (such as "friendliness" or "responsiveness"), so the variables can be reduced into sets of common factors.

PCA can be used to convert multiple variables into a single index. "Components" are basically projections onto a different set of coordinates. Different "principal components" can be created, but they have zero correlation with one another. There can be as many principal components as there are variables in the dataset (in the case below, there can be as many as three). Each one explains as much of the variance of the original series as possible; the first explains the most. The way to tell how many have statistical meaning is to look at the *eigenvalues* and only use those with values above one. The *loadings* describe how much each original variable contributes to the new variable.

In this example, there are two principal components with Eigenvalues greater than one. The first principal component ($PC1$) explains 45% of the variance in the 3-component series. The loadings differ and sign and size between $PC1$ and $PC2$, but neither match the theory. When I have used them to create an index

Figure 5: Principal Component Analysis Using gretl.



of "Exchange Market Pressure" that combines currency depreciations, reserve losses, and interest-rate hikes), they tend not to work any better than an index that divides each component by its own standard deviation and adding them all together. In fact, PCA performs worse, since only the second component should have a negative sign.

Each of these three techniques will be the basis of both a class example and a project. You can use the *R* code from class to help on your own assignment, but make sure you carefully modify it to match what you are doing. For example, the *PCA* example is a time-series, but the assignment is geography-based. It makes no sense to put quarterly values on block groups. In fact, if you had a shapefile of Chicago block groups and corresponding identifiers, you could map your new index using GIS software–or even *R*.

4 Working on a Project

Completing a class assignment is different than a project you think up yourself. If there are clear goals, make sure you know what the professor wants. Read–and re-read–any rubric, and make sure you cover every point that is asked for.

Also make sure that your data analysis "makes sense." You will not need to "make up" numbers to fill in any blanks. Your results should look at least remotely like what you covered in class. If you take logs of inflation rates, for examples, the negative numbers won't calculate–so if you get missing values, you probably did something wrong. As you go, you should check every step. Sometimes you might want to get started right away, but check any variables you calculate yourself (such as per capita GDP, using nominal and population) carefully. One minor error might make the entire project incorrect, and then you will have to start over!

Check and double-check your written explanation against your results. They should explain the exact same things. Don't expect people to get everything from your graphs and tables, but at the same time, some readers skip to the graphs and tables and only want those.

If you have supporting literature, make sure you cite it appropriately. I don't know if Economics has any specific style, so in my class, you can go with the one you know best. In the professional world, companies or journals can have their own "house" style, so you will have to format your documents appropriately.

Expect to write multiple drafts before you turn anything in. Sometimes you should worry less about grammar and spelling at first, and instead focus on getting your ideas out without stopping. But then you will have to read and re-read the document, making corrections not only for spelling and grammar, but also to make sure it reads well and that everything is in order. Sometimes an idea that made sense when you wrote it makes less sense once you read your document from start to finish. Or maybe you used a term that you explain halfway through, because you didn't write the paper in order.

If you use equations in Word, make sure they are correct. You at a minimum should use "Equation Editor" or something similar–not Symbols. Check other papers to see how they are done. I also write elsewhere about fonts and colors.

Once you turn in a rough draft, it should not be too rough. You should have done your best to catch any basic mistakes and should have your ideas as worked-out as possible. Editors do not like having to do clean-up. Their job is to make decent work even better, or to catch harder-to-find errors.

If you are peer-reviewing someone else's work, your job is to help make it better. You might bring a fresh perspective, so if something doesn't make sense that's a perfect time to point that out. Think of questions that haven't been answered, as well as pointing out any errors. It is a good skill to be able to do this in a professional manner, since a lot of writers aren't excited by criticism!

You can learn a lot be being a peer reviewer. You can see new ideas and ways of approaching a problem. If you and a classmate have the same assignment, you might see there are multiple "correct" ways of doing a project. You don't necessarily have to change your paper after the fact. But it will give you more things to think about for the next one.

5 Summing Up

Hopefully, you will be able to use the skills you learned in this class in future courses and projects. You will definitely need to keep going in studying statistics, but a lot of the concepts covered in ECON 310 are general. If you can interpret data, test hypotheses, and explain your results appropriately, you will have gotten a great start in the field.