ECON 453/STAT 433: ECONOMETRICS
Fall 2023
Project Overview

Goal of the Project

Your goal for this project is to produce a document that indicates your level of competence in the field of econometrics. The end result of your statistical analysis will be a series of multiple linear regression models that explore the relationships between several variables. The project will also involve the use of procedures you have learned throughout your training in statistics.

Your final product will be a 10-15 page paper that demonstrates your abilities in collecting data, running statistical procedures, interpreting output, and presenting your findings in a clear, insightful manner.

You should be treating this as if it were a report you were handing in to a boss or to an instructor as a large part of your grade in a course. This means that you should proofread and present figures and numbers that can be easily interpreted and read by a wide audience. Even though I will be familiar with your project by the time you hand in your final draft, you should write the final product as if you are handing it to someone who has never before discussed the project with you.

Throughout the time you are working on this project you are highly encouraged to discuss your progress with me and ask any questions that you have as they come up.

Points of Emphasis

There are three main factors I want to emphasize when it comes to how your projects will be evaluated:

(1) *Statistical competence* – are you properly setting up models, testing for problems, and interpreting results correctly?
(2) *Presentation* – are your tables and figures neat and easy to read? Do you discuss background information/context that is important to understanding your project? Do you make the paper engaging and highlight important and/or interesting information for the reader?
(3) *Reflection* – Do you have a reasonable justification for the way you have set up your model? Do you have a realistic overall sense of how the work you have completed answers the question at hand? Do you have appropriate suggestions for how someone could build from this work going forward?

Deadlines for the Project

Friday October 20th – Project Proposal: you need to submit a 1 to 2-page typed proposal that indicates the subject of your project. This should describe the goal of the project. It should also provide me with some specifics about the data you will use for your project. What is(are) the source(s) of your dataset? What is your response variable(s)? What are your explanatory variables? How many observations will you expect to include? Finally, you should give a brief discussion about the types of regression models you are thinking of running, and your hypotheses regarding your findings. Why do you think these variables are important and what do you expect to find in the results? It is not a bad idea to discuss your topic with me before you submit your proposal. The project proposal will be submitted through Canvas. This is worth 5% of your final project grade. ***Note that if you do not meet this deadline, you will still need to provide a proposal for your project before I will grade your final project.***

Friday November 17th – Submit your dataset: you need to submit a copy of your dataset (preferably in Excel or gretl format) to me (through Canvas). The purpose of this deadline is to ensure that you have collected a dataset that will allow you to finish the project on time. It will also allow me to check the accuracy and validity of the data you have collected. This is worth 5% of your final project grade. ***Note***

that if you do not meet this deadline, you will still need to provide a copy of your dataset before I will grade your final project.

Thursday December 14th – Final Draft of Project: you will submit a document that meets the requirements listed below. This final draft should address any suggestions and comments provided from the proposal and dataset stages. All tables and figures presented in your project should be labeled and neat. There should be no obvious grammatical or spelling errors remaining. This final draft is worth 90% of your final project grade.

Requirements

Your project should include each of the following:

- Summary statistics for the variables included in your dataset, and a discussion of the summary statistics that adds value for the reader.
- At least one scatterplot or other visualization of the data
- Estimating at least three multiple linear regression equations with at least three explanatory variables (not all models need to include at least three explanatory variables). The relationships between the models are up to you, but ideally, they will be related. Some suggestions for creating multiple models:
    - Try different forms of a key explanatory variable (linear, quadratic, log, categorical)
    - Try different combinations of explanatory variables.
    - Try different subsamples of the data (based on characteristics, time periods, levels of a key variable, etc.)
    - Try different modeling options (e.g., pooled cross-section vs. fixed effects)
    - Test different theories within the same broad topic.
    - These are not the only options, just some suggestions.
- A discussion of the overall validity/explanatory power of your regression equation(s)
- A discussion of the significance of the explanatory variables included in your models.
- A discussion of the problems (such as multicollinearity, heteroskedasticity, or serial correlation) that may be present in your analysis. You should test for these problems if possible (in at least one of the models) and discuss the results.
- Your dataset should include at least 25 observations.
    - Note: this is a very small number, ideally you will have many more than this.

****Note: These are the basic requirements for the project. You should not assume that meeting these requirements will result in a 100% grade for the project. The project is graded based on the effort and skill you have demonstrated in analyzing the topic you have chosen.**

Below is an example of how I typically think of organizing a paper such as the one you will produce. There is some flexibility here, and please ask if you have any questions about how to organize your study.

Project Outline

(1) Introduction

Provide a basic description of the relationships and issues that you are analyzing. In the introduction you should give any background knowledge necessary to set up your project. You should also provide some intuition as to why the topic you are covering is important, practical, or relevant. Finally, you should highlight the main findings of your analysis.

(2) Data and Descriptive Statistics

In this section you should be describing the source(s) you used to collect the data for your sample, and describing the main variables that you are measuring. Did you restrict the sample? Did you clean the data, omit outliers? These are the kinds of things that should be addressed in

this section. You should include some discussion of how reliable your dataset is, and whether or not the method in which data was collected will influence the results of your statistical analysis. Next, you should be providing summary statistics (mean, standard deviation, minimum and maximum, and any others you find relevant (median, for example)) for each of the variables included in your regression. These statistics should be provided in a neat table that is properly labeled. You should provide a brief discussion of these statistics. For example, are there any potential problems with skew, outliers, etc. that might be present based on these statistics? If possible, provide any other descriptive information that is helpful as part of your overall presentation.

This section will also include a descriptive analysis of the relationship between variables. This will include an examination of correlation coefficients and at least one scatterplot or other data visualization. You should be interpreting these things for the reader as you provide them.

(3) Empirical Strategy

Here you will discuss the main focus of your project, the multiple linear regression models. You should provide at least one equation like the following:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \mu_i$$

And then indicate what each of these variables represents. Here you will discuss how you specified the models. For example, how did you scale the variables? Did you log any of the variables? You should be outlining the series of models you are estimating, and clarifying what you hope to investigate with these models. Your empirical strategy section should also discuss your hypotheses regarding the explanatory variables, or the overall relationships being examined. What effect do you think each explanatory variable will have on the response variable, and why?

(4) Results

You will next present the results from your regression models that you explained in the previous section. These results should be presented in one or more tables that include at least the estimated coefficients, the standard errors of the coefficients, and the value of adjusted-R^2. You will then discuss the results. You should be discussing each of the coefficients in terms of sign, significance, and magnitude, though you do not need to give a detailed discussion of every variable (can make statements like: “all of the year dummy variables are positive and statistically significant”, for example). The key to this discussion is to ensure that you are giving the reader an intuitive understanding of the findings of your model. Together with discussing statistical significance, you should be discussing whether or not your results indicate practical significance.

As part of the results section, you should also be discussing the validity of at least one of your regressions. This can include – based on the nature of your dataset – considering the likelihood of problems with multicollinearity, heteroskedasticity, and/or serial correlation. How big of a problem might these be, and is there anything that can/should be done about it?

(5) Conclusion

Here you should summarize what your project has shown. You should focus on summarizing the results from your regression Models. If possible, you should discuss how these results might be useful for policies, etc. You should also include an honest discussion of the limitations of your analysis, as well as ways in which your project could be revised or extended in the future.

Example

We are interested in finding out how people perform when they are under pressure. Specifically, when there is a lot of money on the line, do people tend to excel, or do they tend to choke? In order to test this information, we look at how professional golfers putt, and see if how well they do differs based on how much money is on the line. To do this, we collect data on each shot taken in PGA tour events from 2004 to 2012 from the ShotLink data system. We focus on the last hole of each tournament.

Descriptive Statistics:

TABLE 1

Descriptive Statistics

	Mean	Standard Deviation	Minimum	Maximum
Putt Made	60.71%	0.49	0	1
Value of Putt (tens of thousands of dollars)	1.89	4.38	0	68.40
Distance to Pin (feet)	11.78	14.69	0	108.83
Player Age	35.48	6.63	18.03	64.38
Money Earned in Previous Year (millions of dollars)	1.22	1.24	0	10.91
Money Earned in Career (millions of dollars)	7.83	9.57	0	94.82
Total Putts Gained	0.92	3.35	-14.45	13.84
Uphill	51.00%	0.50	0	1
Number of Putts Observed	23,596			
Number of Players	568			
Number of Tournaments	210			
Number of Courses	50			
Years	2004-2012			

You want to include a straightforward table such as this (does not have to be formatted like this) that summarizes the key variables in the analysis. This is important to help people understand the methods and results. For example, the key question in this project is about monetary pressure of a shot. The key variable, then, is the "Value of Putt" which represents the difference in tournament rankings (and monetary prize) between making and missing a putt. The average value of a putt in the sample is $18,900. There is one shot in the sample, however, where the difference in prize money between a make and a miss is $684,000!

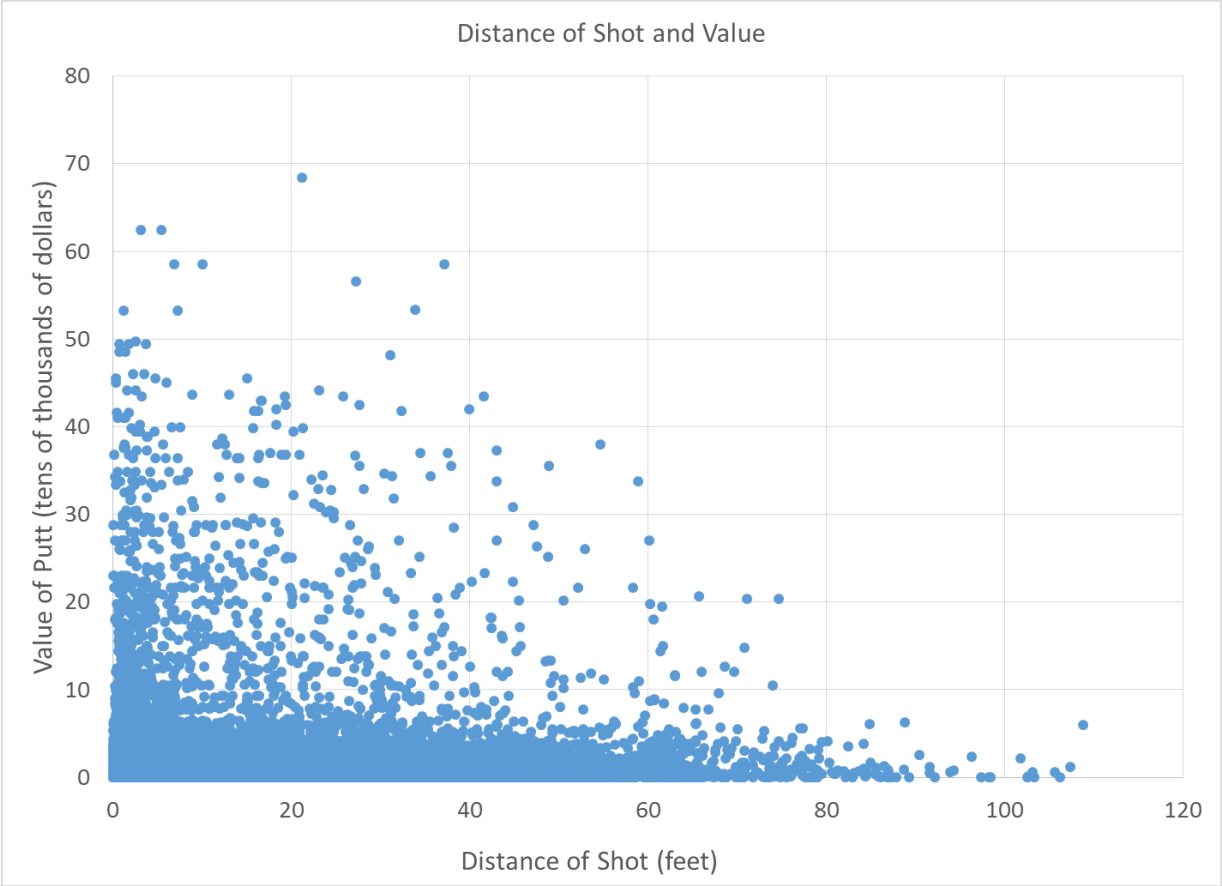
You also want to consider ways to add value for the reader. For example, the table below provides some additional descriptive information that provides a first glance at the main question (before we get to the regression analysis). This provides some initial evidence that players may "choke" under pressure.

TABLE 2

Percentage of Putts Made by Value of Putt

Value of Putt	Observations	Made
0 to $999	6,941	65.06%
$1,000 to $9,999	7,181	61.98%
$10,000 to $24,999	4,895	57.36%
$25,000 to $49,999	2,993	56.87%
$50,000 to $99,999	792	53.41%
$100,000 or more	794	53.53%
All Values	**23,596**	**60.71%**

Next, we want to look at some of the data to help provide more context to the reader. In my case, the dependent variable is a binary variable (make the putt or not), so it doesn't make much sense to do a scatterplot with that. Instead, I will look at how the value of the putt variable relates to the main factor that affects the success rate of a putt – the distance. Instead of a scatterplot, I could have chosen to do a histogram of the value of putt variable (to give more detail) or another data visualization.



Below is my table of correlations for some of the key variables in the analysis. Here we see whether or not the variables are directly related. This can help me check for issues like collinearity.

Table 3: Correlations

	Made	Value of Putt	Distance	Money Earned Previous Year
Made	1			
Value of Putt	-0.044	1		
Distance	-0.649	0.042	1	
Money Earned Previous Year	0.006	0.152	0.003	1

Empirical Strategy

To estimate the impact of financial pressure, we focus on the last hole of the tournament, where it is easier to identify the direct value of a shot. The value is determined by looking at the difference in prize money between the position the player finishes in if he makes the shot, and the position he finishes in if he misses the shot.

The equation I am trying to find is:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \mu_i$$

Where the variables stand for:

$$y = \text{Putt made (0 or 1)}$$
$$x_1 = \text{Value of putt (tens of thousands of dollars)}$$
$$x_2 = \text{Distance of putt (feet)}$$
$$x_3 = \text{Money earned by player in previous year (in millions of dollars)}$$

Hypotheses:

$$H_0 : \beta_1 = 0 \qquad H_0 : \beta_2 = 0 \qquad H_0 : \beta_3 = 0$$
$$H_A : \beta_1 < 0 \qquad H_A : \beta_2 < 0 \qquad H_A : \beta_3 > 0$$

I expect that the more money is involved with a particular shot, the less likely an individual is to make that shot (I expect people to choke under pressure). Obviously, the greater the distance of a particular shot, the less likely it is to be made. Finally, I expect that the more successful a player is in his career, the more likely that indvidual is to make a given shot.

Table 3: Regression Results

Variable	Coefficient	Standard Error	t-statistic	p-value	Lower 95% Interval	Upper 95% Interval
Intercept	0.8595	.0039	217.88	0.0000	.8518	.8673
Value of Putt	-0.0020	.0006	-3.61	0.0003	-.0031	-.0009
Distance	-0.0215	.0001	-130.79	0.0000	-.0219	-.0212
Money Earned	0.0043	.0020	2.19	0.0289	.0004	.0082
R^2 = 0.4215						
Obs = 23,596						

The numbers indicate that, for example, if the monetary value of a putt increases by $10,000, the probability that shot is made decreases by 0.2 percentage points. Another way to say this is that if a player is taking a shot worth $0, and a shot worth $50,000 from the exact same spot, we expect the likelihood he makes the second shot to be 1 percentage point lower, all else equal. We also see that longer shots are made less often, and that players that earned more money in the past are more likely to make the current shot. I would make a note in my project that I tested for heteroskedasticity and found a problem (p-value = 0.000000), so the standard errors reported have been corrected for heteroskedasticity.

There are many ways we could expand this project into multiple regression models. One thing to do is to incorporate different levels of "fixed effects" to account for potential omitted variable bias. This is something we will talk about during the second half of the course. When I do this, I might try to format using the idea of a "model table", to be more efficient in my presentation:

TABLE 3

Regression Results

	(1)	(2)	(3)	(4)
Value of Putt (tens of thousands of dollars)	-0.0019***	-0.0018***	-0.0019***	-0.0017***
	[0.0004]	[0.0004]	[0.0004]	[0.0004]
Player Age	-0.0004	0.0605**	-0.0090***	
	[0.0003]	[0.0253]	[0.0017]	
Money Earned in Previous Year (millions of dollars)	0.0026	-0.0003	-0.0002	
	[0.0018]	[0.0024]	[0.0024]	
Total Putts Gained	0.0089***	0.0087***	0.0088***	0.0089***
	[0.0006]	[0.0007]	[0.0007]	[0.0007]
Uphill	0.0156***	0.0152***	0.0164***	0.0150***
	[0.0040]	[0.0039]	[0.0041]	[0.0040]
Constant	1.0708***	-1.3064	1.5042***	1.0523***
	[0.0149]	[1.0160]	[0.0734]	[0.0082]
Course Fixed Effects	X	X		X
Year Fixed Effects	X	X		
Player Fixed Effects		X	X	
Course-Year Fixed Effects			X	
Player-Year Fixed Effects				X
Observations	23,596	23,596	23,596	23,596
Adjusted R-squared	0.613	0.615	0.615	0.613

Notes: Results from estimating linear probability model where dependent variable is equal to 1 if putt is made and 0 if not. Standard errors, clustered at tournament level, displayed in brackets. Each regression includes a seventh-order polynomial for distance, the estimated coefficients of which are not presented in the table. *** indicates significance at 1% level, ** indicates significance at 5% level, * indicates significance at 10% level.

This project has shown that individuals tend to underperform when they are facing large amounts of monetary pressure. This is true when considering the distance of the shot as well as how well the player has performed recently. One limitation of this work is that it is not clear how well this indicates how individuals will perform under pressure in other settings. Can we extend examples from the sports world to predict how workers will perform under pressure in the business world? Going forward, it is important to control for other factors, such as the difficulty of a particular course, in order to be sure the results are valid. An interesting extension would be to see if those that have faced many pressure situations throughout their career are less likely to underperform as a result.

Some other examples of project ideas:

- Explaining GDP in the U.S. over time by looking at other macroeconomic variables (unemployment, inflation, etc.)

- Predicting economic growth across countries using variables about health (life expectancy, infant mortality rate), education (literacy, average years of education), % of economy by industry, corruption index, etc.

- Predicting attendance at sporting events based on winning %, opponents winning %, time of day, population of city, etc.

- Surveying students and asking about GPA and time use, and using time use to predict GPA. How many hours of sleep do you get on average? How much time do you spend on social media per day? How much time do you spend studying per day? What gender are you?

- Using historical election data to predict what percentage of the vote the Republican presidential candidate will receive. Some explanatory variables you might want to include: demographics of population (% of minorities, age distribution), economic indicators (unemployment, GDP), whether or not the current president at the time of the election is a Republican.

- Predicting differences in crime rates across Idaho counties by looking at differences in demographic variables, differences in economic performance, differences in police spending (per capita), etc.

Some things to keep in mind when choosing topics

- Try to pick something you are interested in

- Try to pick something for which there will be enough data readily available
    - Be flexible/creative in variables used
        - i.e., do not make up your mind on a specific topic/set of variables until you know what data is available.
        - Example: what has caused income inequality in the US to increase?
            - Macro time-series?
            - Regional variation?
            - Individual level?
            - How did the financial crisis/recession impact income inequality?
    - How far back is the data available?
    - Can you download the data directly?

- Think about what type of data you want (or can get)
    - Cross-sectional: Individuals at one point in time
    - Time-series: One individual at several points in time
    - Panel data: Same individuals at several points in time (more difficult to work with)

Some places you might go to get data and/or ideas:

- Bureau of Labor Statistics (http://bls.gov/)
  - Historical data on the U.S. economy
    - Inflation, unemployment, labor productivity, etc.
    - Data at various geographic levels

- Bureau of Economic Analysis (http://bea.gov/)
  - Macroeconomic U.S. Data
    - Trade, GDP
    - Data at State/regional levels as well

- Center for Disease Control (https://data.cdc.gov/)
  - Data on injury, illness in the U.S.
  - Data on health behaviors (alcohol use at state level)

- Data.gov (http://www.data.gov/)
  - The US Government's open data
  - Many datasets at many levels (federal, state, local) on a variety of topics

- Digest of Education Statistics (https://nces.ed.gov/programs/digest/)
  - From US National Center of Education Statistics (NCES)
  - Information for all levels of educations, covers many years
  - For recent years, much of the information can be downloaded directly into Excel

- Integrated Postsecondary Education Data System (IPEDS - https://nces.ed.gov/ipeds/datacenter/)
  - Information specifically on postsecondary institutions
  - A wealth of information for virtually every school over many years
  - Examples of data available include: applications, enrollments, expenditures, graduation rates, retention rates, tuition, financial aid, and many other topics

- FBI – Uniform Crime Reporting (http://www.ucrdatatool.gov/)
  - Crime statistics
  - Data on crime rates by type of crime/geographic region

- Gallup Poll (http://www.gallup.com/home.aspx)
  - Data on opinions on a wide arrange of topics
    - Examples: presidential approval rating, economic confidence, concern about terrorism, etc.)
  - Related: Gallup-Healthways Well-being Index: http://www.healthways.com/solution/default.aspx?id=1125
    - Rankings of states, communities on well-being and other factors

- The General Social Survey (https://gss.norc.org/Get-The-Data)
  - Information on a number of social topics as well as demographics and behavioral information.
  - Examples of information include how religious the person is, whether they own a gun, how they feel about climate change, life satisfaction, and many other things.
  - Survey conducted many years, beginning in 1972.

- U.S Census (http://www.census.gov/)
  - U.S. Demographic Variables
    - Race, gender, age

        - Education, income
        - Mobility
        - Home ownership
    - State and county level variables for cross-sectional analysis
    - The factfinder can be a useful tool: http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml
- Integrated Public Use Microdata Samples (IPUMS - https://usa.ipums.org/usa/)
    - Individual level data from the Census samples
    - Information from the decennial census going back to 1850
    - Information from the American Community Survey (ACS) annually from 2000 to 2015
    - Information about the household, person – income, marital status, education, travel time to work, and so on.
    - Note, these are very large files and can be difficult to work with, please see me if you are thinking of using this kind of information.
        - As an example, the 2013 ACS contains observations for more than 3 million individuals.

- Inter-University Consortium for Political and Social Research (http://www.icpsr.umich.edu/icpsrweb/ICPSR/)
    - Vast amounts of data relating to social sciences
        - Aging
        - Education
        - Crime
        - Elections

- World Bank Development Indicators (https://data.worldbank.org/indicator)
    - Many variables measured across countries and over time
    - Information on health outcomes, economic outcomes, energy usage etc.
    - Data is easy to obtain and format

- Washington State Liquor and Cannabis Board (http://www.liq.wa.gov/)
    - Information on sales and tax collected by county

- Idaho Indicators (http://indicatorsidaho.org/)
    - Run by the UI Extension
    - Information at county level in Idaho
    - Includes state and county rankings for many variables (teen birth rates, housing affordability, labor force participation rate, etc.)

- The Library of Congress, US Election Statistics (https://www.loc.gov/rr/program/bib/elections/statistics.html)
    - Data on election results at federal and state levels
    - Links to many sources that provide information on voter turnout, and many other things

- College Scorecard (https://collegescorecard.ed.gov/)
    - Data collected from the federal government using financial aid records
    - Information on costs and outcomes (graduation rate, salary after graduating)
    - Also provides data on characteristics of the schools

- The Internet (http://www.google.com/)