Homework 2

Gather data where there is a numerical response variable, and one numerical potential predicting variable. Your sample should consist of at least 30 observations to start with (your final chosen model might be based on fewer observations, of course, if you have decided to omit unusual observations). Analyze the data in a full and appropriate fashion using any methods and techniques that are appropriate, and discuss what you find. Examples of this type of data analysis would include: examining predicting baseball team winning percentages for last season from team earned run average for all of the teams in Major League Baseball; examining predicting stock returns from net profits as a percentage of revenues for a sample of publicly-traded companies; or examining predicting revenue for each of 50 different stores of a retail chain from the average income in the neighborhood of that store.

Homework due: September 29 by 10:00 PM. Upload it as either a pdf file or an MS-Word document through the “Assignments” tab on NYU Classes. Please include your name in the name of the document file.

Important Notes

When doing your analysis of these data, or any other analyses you do later in the semester, remember what is required of a good analysis and report. I do not want a quick and routine analysis; a good job will show understanding of the problem and possible solutions and techniques to consider. The technical results should be stated clearly, with supporting documentation, and conclusions should be presented in a concise, informative, clearly written, non-technical summary. The class handouts, including the Purchasing Power Parity handout we’ve looked at in class, are examples of what I’m talking about. In particular, a good report should state the problem at hand (and why it is interesting), include description of the data (including its source), the questions of interest, the implications of any graphical or statistical methodology used, discussion of the results obtained, including any adjustments to the data or corrective actions that might have been taken (with explanations of the observed patterns if possible), implications of your results for the real-world problem, and a summary of what you’ve learned from the analysis. You should report and summarize the implications of the “best” model you come up with even if you ultimately decide that it is deficient for some reason (you should also discuss why you think it’s deficient, and what you might be able to do to fix it, even if you don’t have the tools yet to do so); that is, for any of the homeworks you should do the best you can do, and say why it’s not good enough if necessary, but don’t say that since you can’t get it completely right you can’t do anything at all. Note that this includes if you believe that
there is no relationship between the response and the predictor; you should still describe the implications of your model. There is no minimum or maximum number of pages for such a report; it should be as long as it takes to tell the story you need to tell.

Remember that the goal of this, and any statistical analysis you do for this class, is to discover what is actually going on in the data, not to get a statistically significant regression. Your criterion in choosing a data set should be that it is reasonable to expect that the target variable might be related to the predicting variable(s). If it turns out that the target is not related to the predictor(s) (after careful analysis), that’s fine — you’ve learned something. There is no such thing as a “smallest allowable $R^2$” in this, or any other, assignment. You should summarize what the various aspects of the fitted regression imply (coefficient estimates, $R^2$, etc.), but if the correct conclusion is that there doesn’t seem to be evidence of a relationship, then that is what you should report. Similarly, the fact that there might be violations of assumptions does not mean that there is a problem with using the data set; as long as you address those model violations as best you can using the methods we have discussed in class up to that time, that is all that is necessary.

I recognize that this is a homework assignment, and you should too. You shouldn’t be spending many hours trying to find the perfect data set, one that is interesting, exciting, fundamentally important, and so on. As long as there is a reasonably interesting question that can be answered using statistical (regression) methods, the data set is adequate for our purposes. Believe me, I do understand the difficulties in getting your own data, compared with being given “canned” data; but please also believe me, you will find that the rewards really do outweigh the costs. You should also know that the next homework is going to be a multiple regression assignment requiring at least four predictors in the initial model, and you are welcome to build on your analysis here for that assignment. If you plan on doing that, it would be a good idea for you to gather all of the variables now, rather than have to go back later to get them.

You will find the analysis is much more interesting, rewarding, and educational when based on data that you care about, so I strongly encourage you to use data that are of genuine interest to you, whether that means finance data, sports data, medical data, or any other kind of data (job-related data are obvious possibilities, and you can request any level of confidentiality with respect to those data that you wish). There are a vast number of data sources on the World Wide Web. I have created a web site with links to many data sources on the web, with descriptions of the types and formats of data available. The location is http://people.stern.nyu.edu/jsimonof/classes/datalink.html. You may not, however, in this homework or ANY future homework, take your data from a textbook, a journal article that includes a regression analysis of the data, or an online digest of data sets that have been put together for educational or research purposes or use as regression analysis examples, including (but not

© 2020, Jeffrey S. Simonoff

2
limited to) the Data and Story Library (DASL), the UCLA Statistics Online Computational Resource (SOCR), StatCrunch, StatLib, RPubs, Kaggle, crowdflower, data.world, or the UC Irvine Machine Learning Repository. (Note that the DASL files can be found all over the internet, so it is up to you to make sure that you don’t use one inadvertently.) You may not use data sets that have been created as part of data analysis competitions such as the KDD Cup or the Yelp Academic data sets, and you should not use data sets that are included with implementations of software, such as being included as part of an R package. This also includes digests of data sets from textbooks and articles specifically gathered together to be examples of regression data, including those put together by faculty members at colleges or universities, even if no analyses are provided along with the data (that is, don’t try to find data by doing a Google search of “regression data” or use data from a web page with a title like “Regression data sets”). Simply put, the idea of these homework analyses is that you think of a problem that interests you, and you gather data that you can analyze to examine that problem in an appropriate way.

If you are unsure about whether a data set is appropriate, ask me to make sure. A list of types and sources of data that have been used by students in the class in previous years can be found at people.stern.nyu.edu/jsimonof/classes/2301/sources.pdf. You must include discussion of the source of your data as part of your report; if the data are from the World Wide Web, give the appropriate URL(s), and if they are from a book, include photocopies of the page(s) from which the data come. Do not merely say “The data come from the Census Bureau”; give the precise source at the Census web site or in a Census publication.

You can analyze a data set that is a time series (data where the observations are ordered in time, such as, for example, modeling the change in a stock index from inflation rate for 40 consecutive quarters) if you wish, but I don’t recommend it; if you do so you should recognize that such data have the possibility of the occurrence of additional problems related to autocorrelation. While I would not expect you to be able to address such problems (yet), their presence could make your analysis more challenging, and you would need to at least recognize the implications of those issues. Note that if you do use time series data, you must order the observations forward in time, not reverse; that is, for example, the row for January followed by the row for February followed by the row for March, and so on. Still, I suggest that you stick to cross-sectional data (data where the observations can be viewed as a sample from some population) for the next couple of assignments, but that is up to you (Homework 4 will be an explicitly time series assignment). If you do decide to do an analysis on time series data, you should not use time itself as your only predictor. A good way to think about what your data should look
like is that the rows of your data matrix should be able to be viewed as a sample from some population — a sample of cities where each row is a different city, a sample of countries where each row is a different country, a sample of companies where each row is a different company, a sample of states where each row is a different state, a sample of people where each row is a different person, a sample of teams where each row is a different team, and so on. If you cannot look at each of your observations and think of it as having been drawn from some well-defined population, you do not have appropriate regression data.

If your observations constitute the entire set of members of some group, such as all teams in a sports league or all states in the United States, you still need to treat this as a sample from a population (if you actually have all of the members of a population in your data there are no inferential questions to ask, as you know all of the properties of and parameters associated with that population). This can usually be reasonably done by treating your data as a snapshot in time of a stable process; that is, as the set of values at one time point, where your interest is in understanding the underlying process and making predictions about future time points.

I am not expecting you (in this or future homeworks) to use methodologies that we have not discussed in class, unless I explicitly say otherwise. However, if you choose to do so (because you have been exposed to them in another class, for example, or did a search on the internet and found out about them), you must apply them and interpret the results fully and correctly. If you do not do so you could be penalized, so you should consider very carefully whether you want to pursue such analyses, as opposed to just saying that a correct analysis is beyond the scope of what we have discussed in class up to that time.

I have one more piece of advice about how you input percentage data into Minitab or R. It is possible to enter in data values as percents in Minitab, including the percent (%) sign. If you do this, Minitab does the reasonably sensible thing of internally converting the variable to a proportion, as in 15% to 0.15 (Excel does this also, by the way). This is fine, and all analyses using the variable will be correct, but it can lead to confusion if you try to interpret output in terms of percentages, rather than proportions. For example, if you have a predictor variable in a regression coded this way, the slope coefficient does not correspond to the estimated expected change in the response associated with a one percentage point change in the predictor, but rather a change of 1, which corresponds to a 100 percentage point change in the predictor. For this reason, I urge you to stay away from coding the variable this way. Since for data of this type it is likely that it is percentage point changes that are of interest, all that means is that you should enter a value like 15% as 15, leaving out the % signs for all of the values. It’s even easier with respect to R — if you include a % sign in the data R will treat the variable as a text variable and it will be impossible to do any numerical analysis using it, so you should not do this under any circumstances.
If you are constructing your data set as a text (.csv) file, particularly for input to R, don’t include commas in the numbers, as that will confuse the input function (i.e., use 1234567, not 1,234,567).

I encourage you to ask me questions about this analysis, and further analyses throughout the semester, either via Zoom or by e-mail. I’ll be happy to address any specific questions you might have (although it’s possible the answer will be a variation on “You have to figure that out for yourself.”). Please don’t, however, ask me to review and critique an entire report that you’ve written, as I am not willing to do that.