PROJECT (100 POINTS)

Due Thursday, Dec. 7, 2023

The project consists of a multiple linear regression analysis on a data set of your choosing. You will be expected to perform the analysis and write the report in groups of one to three people.

FINDING THE DATA

You are encouraged to find a data set that represents a problem of interest to you. You may select a data set from a textbook (except ones we have or will analyze in class), or you may use data from the internet, from another class, or that you have collected on your own. The table below gives some websites with interesting data. Your data set should have **at least four potential explanatory variables and at least twenty observations**.

Real data sets are more difficult to analyze than the data found in a textbook. Additionally, the more predictor variables involved, the more complex the analysis will be (and possibly more problematic). For these reasons, your grade on the project will depend on how well your analysis meets the challenges of your data set. Thus, **I expect you to do more analysis with a simple data set**. For any data set, you will have to address the standard issues in multiple linear regression: the overall fit of the model, variable selection, residual diagnostics, and prediction.

Data Source <Address>
EPA (Water Quality) https://www.waterqualitydata.us/
EPA (eGRID) https://www.epa.gov/egrid/data-explorer
United Nations (Human Development) https://hdr.undp.org/data-center
CDC - National Center for Health Statistics http://www.cdc.gov/nchs/fastats/default.htm
USDA https://www.ers.usda.gov/data-products/
US Dept of Education http://nces.ed.gov/programs/digest/
National Center for Education Statistics (Colleges) http://nces.ed.gov/ipeds/datacenter/
US Census Bureau https://www.census.gov/data.html
Stock Market http://finance.yahoo.com/ or http://www.nasdaq.com/ (choose your stock symbol or index then click historical data)
Bureau of Justice Statistics http://www.bjs.gov/index.cfm?ty=dca
NOAA Climate Data https://www.ncei.noaa.gov/cdo-web/ you have to click to “buy” the data, but there is no cost
Sports: https://espn.com/ (choose your sport then click Stats), or try MLB.com or NFL.com

Project Proposal

Your group should provide me with a brief project description including the question(s) of interest, a brief description of the data, and the data set[†] (with references, if necessary). In addition, you should include the results of the overall F-test, R^2, and the results of the individual t-tests for the marginal effects. The model you analyze should include all your potential explanatory variables (although you may leave out any indicator or interaction variables) and will generally be your initial model in the final report. The writing of your project proposal may be informal.

The proposal is **due Sept. 28** and is worth 10 points of your final project grade.

Analyzing the Data

You will be exposed to many of the procedures and analyses in multiple linear regression. You are not expected to apply every technique we discuss in class to your data set. However, you are required to address the following components. Note: if you consider more than one model in your analysis, the required components need only be done for one model (not necessarily the same model for all components).

(Minimum) Required Components:

- What are your variables: data description, definition of response and predictor variables, definition of indicator or interaction variables (if present), source of the data, and the value of $\mathbf{X}^T\mathbf{X}$ for your data
- Is your model useful in predicting the response: overall F-test and R^2 (or R^2_{adj})
- Which individual variables are important to predicting the response: individual t-tests for marginal effects
- How can your model be used for predicting future responses: (point) prediction of response
- Does your model satisfy the assumptions of linear regression: residual plots for linearity and variation, at least one check of normality (histogram, Q-Q plot, statistical tests for normality)

Again, the complexity of your data set may determine which techniques can be used, and simpler data sets are expected to be accompanied by more complete analyses. A project that involved the simplest of data sets and only the required components listed above would likely earn a low B or C. To earn a higher grade, you must include some of the following components. (There is not a specific formula here, but you should plan on doing *at least 4 or 5* additional components to earn an A).

Additional Components:

- more complicated data: either data that is harder to collect, involves more variables, or includes interactions and/or indicator variables
- more descriptive statistics: scatter plots, individual descriptive statistics for each variable
- more about individual variables or subsets of variables and model comparison: interpretation of estimated coefficients, confidence intervals for individual coefficients or simultaneous confidence interval, checks for multicollinearity,

[†] You may submit your project description and/or data electronically.

partial *F*-tests, conditional sums of squares tests, AIC/BIC comparisons, model selection: all subsets, forward, backwards, or step-wise techniques

- calculations done using matrix form rather than (or in addition to) R output
- more about prediction: confidence interval for mean response, prediction interval for new response
- more checks for assumptions: tests for independence (if appropriate), checks for outliers and influential observations
- corrections for failed assumptions: transformations, WLS, modeling autocorrelation

Written Report

Your final report should include a description of your question of interest, the data considered, the analyses performed, and the results of your analyses. **Note: It is *not* sufficient to give output without discussing the output in your written text.** (For example, if you include the Coefficients Table, but do not talk about the the individual *t*-tests in your written report, I will not count that as having included the individual *t*-tests.) At the same time, for any elements you discuss in your written report, you should include output that supports that discussion.

Grammar, spelling, and general readability do count. It's a good idea to have someone outside of your group read your final report (you could trade with another group in class or take it to the Writing Center). You can also show me a sample.

At the **minimum**, your project should include

- *relevant* computer output (graphs, coefficients tables, ANOVA tables, *etc.*)

You must include enough output to show from where your results come. For example, if you give the regression equation in your text—which you should, then you should include the Coefficients Table in your output (but you could delete the columns you don't use). You do not need to include the original data nor output for models that you tried but did not discuss in the report. You also do not need to include the R commands that generated your output, but you may include them if you wish. As you are likely to have considered multiple models, be sure to label your output (this will also help you be able to refer to your output from your text).

- final results of any hand calculations (matrix calculation, partial *F*-tests, confidence intervals, predictions)

You should show enough detail that I can reconstruct and verify your calculations (*i.e.* show all the values that were used in your calculations), but you do not need to show all the steps of your work.

- interpretation of results (purpose of analysis, interpretation of calculated values, conclusions regarding hypothesis tests)

In many situations you have a choice of procedure; you should give some explanation how and/or why you made the choice you did. (Note: simplicity is a perfectly valid reason; I just want you to acknowledge any trade-offs.) Even if a procedure was a required component, you should talk about what the analysis is used for (why would someone want to do this analysis, in general), how the procedure works, and any special considerations or interpretations relevant to your data. For example, instead of saying "we performed the global F test and found the model to be significant ($p<0.05$)", you should say something more like "The global F-test is used to evaluate whether the regression model is useful in explaining or predicting the response. Specifically, the null hypothesis states…. In the ANOVA table, we find F = 123.4, p=0.000 (see Table 4). Thus, our result is significant ($p<0.05$), and we conclude that our regression model is indeed useful for explaining our response."

- an overall project summary

This should include a discussion of what you learned about your data and the relationship with the response, but also about the process of data collection and analysis. You may also include a discussion of any difficulties you had, things you might have done differently, or what you liked or didn't like about the project.

Your report should be **typed**. There is no specific length requirement, but I expect most of you will have 7 – 15 pages (depending on graphs, output, *etc.*). Graphs and tables may be included throughout the text or may be collected as an appendix, as long as they are **labeled and referred to** in the text of the document. Graphs may be done by computer or drawn *neatly* by hand.

Grading

Grading of the project report will address appropriateness, accuracy, and depth of analyses as well as writing (including style, spelling, and grammar). The project (and proposal) is worth a total of 100 points, assigned as follows.

Component	Points
(1) project proposal, **due Sept. 28**	...10
(2) are the required components present	...25
(a) definition of response, predictors, and indicator and interaction variables (if any)	(6)
(b) value of $\mathbf{X}^T\mathbf{X}$	(8)
(b) overall (ANOVA) *F*-test	(6)
(c) R^2 or R^2_{adj}	(4)
(d) individual *t*-tests for marginal effects	(6)
(e) (point) prediction of response	(6)
(f) residual plots for linearity and variation	(8)
(g) check of normality (histogram, Q-Q plot, or a statistical test for normality)	(6)
(3) are calculations and interpretations correct, is there full explanation of purpose and meaning (and not just reporting of values)	...25
(4) additional components or analysis	...15
(5) spelling, typos, and grammar	...10
(6) writing style, organization, and general readability: is there an introduction and conclusion, are there transitions between sections and	...15

paragraphs, are sentences confusing, brief, logical, or insightful ***NOTE: If you are missing any of the required analyses, your project will automatically be deducted 4-8 points for that element (the combined points for having the element and the calculation/interpretation of the element; see values in the parentheses, above). Remember, you must actually discuss an element in the written text for me to consider it as being present.***	