Statistical Methods II
Midterm II
Take-Home

This examination covers the most important things we covered during the second third of the course. I will not tell you which test to use; use the correct one(s). I will not tell you to provide a graph; provide a well-labeled graph.

Make sure your graphs are well-labeled. The axes need to be labeled. The units need to be labeled. The graph needs to be titled (`main`). The axis values (`las`) need to be horizontal. Make sure the margins for your graphs look good and do not cut off axis labels.

Your examimation answers must be nicely typed. The answers should be as long as they need to be in order to fully answer the question. Grammar counts. Explain in detail what test you are using, what it tests, what the null hypothesis of the test is, what the conclusion of the test is (with test statistic, degrees of freedom, and p-value in parentheses). Explain a lot. Re-read your answers and make sure they logically answer the question posed. Have someone else read your answers.

In your answers, include statistics appropriately. Finally, make sure you provide the name for the test, not the `R` function. The only place I should see anything in `R` 'speak' is the appendix.

When you turn in this examination on Tuesday, attach your `R` script to the back of the pages as an appendix. The graphs need to be woven in your narrative; that is, meaningfully refer to them in the text, explain what the graph tells us, and number the graphs. You can still include them all at the end of the homework if you wish (before the `R` Appendix), or you can put them in the body of your assignment.

Start each answer on a new page. Only print on one side of the paper.

Finally, as usual, if you have any questions or issues, let me know as soon as possible. The worst I can do is not answer your question. You have access to all non-living sources.

Problem 1 ⟦5⟧

Let us start out with something rather easy. We spent most of this section learning about generalized linear models (GLMs). I did spend one day talking about ordinary least squares, however. During that lecture, I laid out several reasons GLMs were *superior* to (general) linear models (OLS). Name and explain any *two*.

Problem 2 [[5]]

One of the assumptions about GLMs is that the residuals are distributed Normally. Explain how you would test this assumption. Remember, use the name of the tests (and explain what you are looking for); do not give the `R` function name.

Problem 3 ⟦5⟧

For the binomial distribution, we always used the canonical link. It is not the only allowed link, however; there are many others that may be used. For instance, we can use the probit link, the cauchit link, the log-log link or the complementary log-log link. What are three requirements for a link for the binomial family to be allowable?

Problem 4 ⟦5⟧

On March 24, 2011, we created a ROC curve to test the goodness of my terrorism model. What does the ROC curve indicate about models (in general)? What type of models can undergo ROC analysis? What does the area under the ROC curve indicate?

Problem 5 [[5]]

On March 8, 2011, we used a non-canonical link for the Gaussian distribution (model `mod.gaus`). Why did we do that? Why was that link allowable when the canonical link for the Gaussian is the identity link?

Problem 6 ⟦20⟧

We have dealt quite a bit with the `gdpcap` data. For a refresher, check Assignment 3. My research hypothesis is that the GDP per capita is significantly affected by both the level of honesty in the government and the democracy level. In your model that uses just these two variables (in some combination and/or power), predict the GDP per capita of a State that has an honesty in government score of 5 and a democracy level of 5.

Note. *Your answer in the write-up will be one sentence. However, the points will be earned in the* R *appendix.*

Problem 7 [[20]]

The Federal Bureau of Investigation (FBI) contacts each of the 50 states (and the District of Columbia) for crime statistics. The Census Bureau (a part of the US Department of Commerce) carries out surveys designed to determine other aspects of the 50 states (plus the District of Columbia). I have combined various bits of data from these two sources in the `crime` data file. This file contains measures for 1990 (suffixed by '90') and for 2000 (suffixed by '00' or not suffixed).

Existing literature suggests that the current violent crime rate is determined in large part by the violent crime rate a decade ago, as well as the current urbanization level, current GSP per capita in the state, and the conservatism in the state.

⋆⋆⋆

Create an appropriate model using these four independent variables. How well did your model predict the violent crime rate in Oklahoma? According to the data, the pertinent values for Oklahoma are `vcrime00` = 497.8; `vcrime90` = 547.5; `urbanp2000` = 65.3; `gspcap00` = 26352.36; and `conserve` = 0.12.

Note. *Again, your answer in the write-up will be one sentence. However, the points will be earned in the* `R` *appendix.*

PROBLEM 8 ⟦20⟧

Finally, we have worked a lot with election datasets. First, there was extensive work with the 2010 Sri Lankan presidential election. Then we had assignments that had us attempt to detect vote fraud in the 2009 Afghan presidential election and the 2011 Egyptian parliamentary election. Let us build on these since you should have have a good feeling for election data. The new election data deals with the 2011 South Sudanese referendum on independence from the north.

⋆⋆⋆

A referendum for independence from Sudan was held from the 9th through the 15th of January 2011. *A priori* expectations were that the referendum would pass easily. The referendum was one of the consequences of the 2005 Naivasha Agreement between the Khartoum central government and the Sudan People's Liberation Army/Movement (SPLA/M). The referendum commission published the final results, with 98.83% voting in favor of independence. The reported turnout exceeded 100% in 10 of the 79 counties. As the referendum passed, the date for the creation of an independent state is 9 July 2011.

⋆⋆⋆

The fact that the reported turnout exceeded 100% in 10 of the 79 counties is interesting, but not *prima facie* evidence of vote fraud. In many countries, citizens are registered in their home district to vote, but may vote anywhere in the country, with their vote counting as a vote in the district in which they placed the vote, not in their home district. In the United States, we may physically vote in Stillwater, but we need to send our vote back home to be counted. Thus, turnout in excess of 100% is evidence of vote fraud in the United States, but not in most of the world.

Let us work the same magic as previous homework assignments. Finding a statistically significant relationship between the proportion of the votes invalidated and the proportion of the votes in favor of independence (Secession) will serve as evidence of electoral fraud in this election.

Make sure you explicitly state the null and research (alternative) hypotheses. Also make sure you test your model. Remember, election data will be overdispersed, so there is no need to alter the research model.

Your graph needs to be proportion of vote declared invalid (Y) against the proportion of the vote in favor of independence (X). Include a prediction curve and the 95% confidence band on that prediction.