STAT 505/506 MS Comprehensive Exam–In Class
5 or 9 jan 2017

Instructions: This exam is closed notes and closed book though you may use a calculator. Please answer these questions on separate sheets of paper **making sure to include your name at the top of each page, number the pages, and only write on ONE SIDE of the page**. Due to the length of the exam complete sentences are not necessary. Please look over the entire exam before beginning. Good luck!

1. **Mortar**. Mortar is used to "glue" bricks together and is made by mixing water and plaster and then letting the mixture cure. A strong mortar is important to hold together the bricks; strength depends on the composition of plaster, type of water, and length of time the mortar is allowed to cure. In one study the effects of the type of water (soft or hard) and length of curing time (3 days, 7 days, and 28 days) were evaluated. Nine samples of a small amount of plaster were randomly assigned to each of the six combinations of water type and curing time leading to a total of 54 observations. The tensile strength for each sample was then measured.

   (a) Write out the hypothesized effects model for this example indicating which effects are fixed and which are random.

   (b) Provide an example of an estimable function of the parameters AND explain why it is estimable.

(c) What is the OLS estimator of $\mu_{hard,27days}$?

(d) Why is the OLS estimator often used? That is, explain what properties the OLS estimator possesses.

(e) The investigator in charge of this study decides to provide you with some additional information about this study. The samples of plaster used were taken from nine available bags of plaster which the investigator believes are representative of all bags of plaster. A small amount from each of the nine bags was randomly assigned to each of the six combinations of water and curing time. With this information write out the hypothesized model of interest clearly identifying which factors are fixed and which are random.

(f) What is the (population) correlation between two samples of plaster from the same bag?

2. **PSID 1982**. The Panel Study on Income Dynamics (PSID) is a longitudinal study that began in 1968 with a nationally representative sample of over 18,000 individuals from 5,000 families in the US. Information from these families and their descendants have been collected over tim. In this particular analysis we will consider the cross-section of data from 1982. The purpose of this analysis is to build a model with log(wage) as the response and 11 potential explanatory variables: `experience` (years of full time work experience), `weeks` (weeks worked), `occupation` (two levels: white collar and blue collar), `industry` (does or does not work in manufacturing), `south` (does the individual reside in the south), `smsa` (does the individual reside in a standard metropolitan statistical area), `married` (two levels: married or not married), `gender` (two levels: male, female), `union` (is the individual's wage set by a union contract), `education` (years of education), and `ethnicity` (two levels: African American, not African American). Attached is some useful and not so useful output. Consider the model:

$$\begin{aligned} log(wage)_i = \quad & \beta_0 \quad + \beta_1 Education_i + \beta_2 Experience_i + \beta_3 Weeks_i + \beta_4 Married_i \\ & + \quad \beta_5 Gender + \beta_6 Ethnicity + \beta_7 Union_i + \epsilon_i. \end{aligned}$$

Use the available output to answer the following.

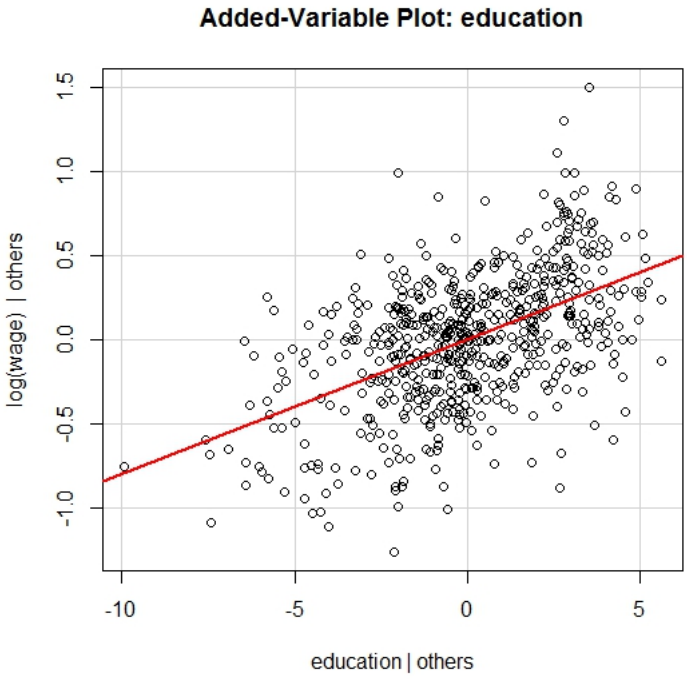
(a) An undergraduate researcher calculated the slope coefficient for education using a simple linear regression as 0.0717. When the undergrad sees your results, she is concerned that she made a calculation error. Was an error made? If so explain what the error is. If not, explain why the results differ.

(b) Provide an interpretation of the coefficient associated with `union` on the original scale of the data and in context of the problem.

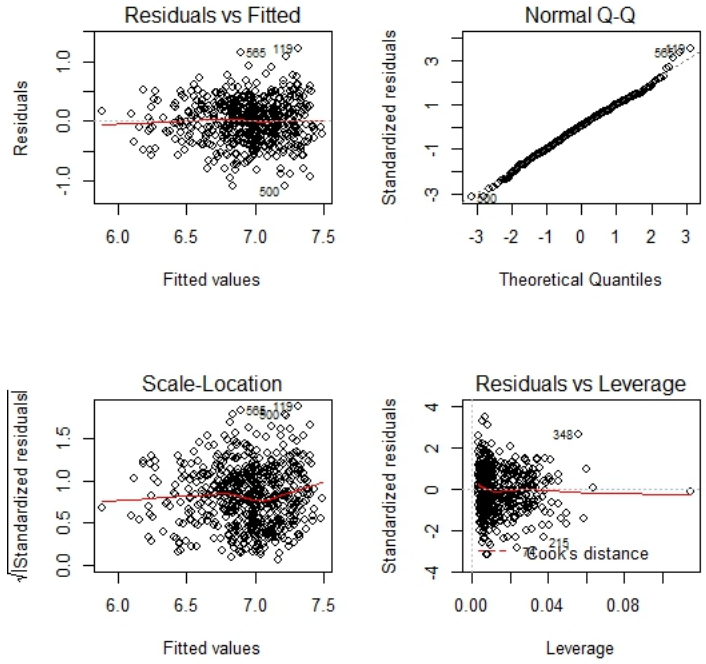
(c) If possible conduct the test of $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. If not possible, explain how you would conduct the test and explain what distribution your test statistic follows.

(d) One purpose of this study is to predict wages for individuals that were not sampled. In the output for this problem are several diagnostic plots and VIF values are also provided. Based on these, do you have any concerns about the fitted model? If so, discuss your concerns. If not, explain why not.

(e) One concern raised by a fellow researcher is that the relationship between log(wages) and education is often non-linear such that as education increases, log(wages) increase at a decreasing rate. Below is the added variable (or partial regression plot) for the variable education. Using this plot explain whether education should be included in the model and if the relationship should be modeled as being nonlinear.



```
> m1<-lm(log(wage)~education+experience+weeks+married+gender
+ +ethnicity+union,data=PSID1982)
> summary(m1)

Call:
lm(formula = log(wage) ~ education + experience + weeks + married +
    gender + ethnicity + union, data = PSID1982)

Residuals:
     Min       1Q   Median       3Q      Max
-1.09654 -0.23459  0.01555  0.23422  1.21750

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.513941   0.171859  32.084  < 2e-16 ***
education      0.080380   0.005496  14.624  < 2e-16 ***
experience     0.006795   0.001387   4.899 1.25e-06 ***
weeks          0.003870   0.002838   1.364 0.173208
marriedyes     0.096087   0.052055   1.846 0.065413 .
genderfemale  -0.309227   0.064442  -4.799 2.03e-06 ***
ethnicityafam -0.175167   0.057116  -3.067 0.002263 **
unionyes       0.107387   0.031660   3.392 0.000741 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3474 on 587 degrees of freedom
Multiple R-squared:  0.3795,    Adjusted R-squared:  0.3721
F-statistic: 51.29 on 7 and 587 DF,  p-value: < 2.2e-16

> vif(m1)
 education experience      weeks    married     gender  ethnicity      union
  1.157518   1.102745   1.065823   2.096836   2.045924   1.078382   1.147306
```



3. **Military Coups in Sub-Saharan Africa**. Following the decolonization of Africa, many African nations gained independence from their European colonizers in the late 1950s and early 1960s. Unfortunately, many of these nations faced great instability due to coups of the government. Bratton and van der Walle (1997) collected data on 47 sub-Saharan African nations on the characteristics of post-colonial political regimes. One aspect of this study was to study the factors affecting regime stability. The response of interest is `miltcoups` the number of successful military coups from independence to 1989. Potential factors include:

   `oligarchy`: number of years the country was ruled by military oligarchy from independence to 1989

   `pollib`: measure of political liberalization (0 = no civil rights, 1 = limited civil rights, 2 = full civil rights)

   `parties`: number of legal political parties in 1993

   `pctvote`: percent voting in last election

   `popn`: population in millions in 1989

   `size`: area in 1000 square km

   `numelec`: total number of legislative and presidential elections

   `numregim`: number of regime types

   (a) The researchers decide to use Poisson regression for their analysis. Explain why the researchers likely chose this option.

   (b) The researchers are not sure what variables to include in their model but are concerned about potential issues with multicollinearity. Concisely explain why the presence of multicollinearity is problematic in an analysis.

(c) For better or worse, the researchers decide to include the variables `oligarchy`, `pollib`, and `parties` in their model. Provide an interpretation of the impact of the variable `oligarchy` on the response *in context of the problem*.

(d) Upon further inspection, the researchers notice that the p-value associated with `pollib` for countries with limited civil rights is "large" and they debate whether or not it should be removed from the model. What do you recommend and why?

(e) During the time frame of this study, Ethiopia, a country on the Horn of Africa, was ruled for 10 years as an oligarchy with 17 political parties and its citizens had full civil rights. Based on this information, what is the predicted number of military coups?

(f) One common concern in Poisson regression models is the presence of overdispersion. Define what overdispersion is and explain whether it appears that it present in this analysis.

```
> cor(africa,use="complete")
              miltcoup   oligarchy     pollib     parties      pctvote
miltcoup   1.000000000  0.60723367 -0.3394218  0.31105966  0.009148621
oligarchy  0.607233668  1.00000000 -0.1290684  0.15888330  0.011192170
pollib    -0.339421835 -0.12906840  1.0000000  0.17200032  0.197183046
parties    0.311059659  0.15888330  0.1720003  1.00000000 -0.139611628
pctvote    0.009148621  0.01119217  0.1971830 -0.13961163  1.000000000
popn       0.359306969  0.34925939 -0.1542419 -0.07516457 -0.173196451
size       0.135278859  0.30189567 -0.1847529  0.06717680 -0.097070052
numelec    0.030250230 -0.19042153 -0.1659638  0.33215436  0.233321177
numregim   0.255534509  0.46160153 -0.1618926  0.19228370  0.152832037
                popn        size     numelec    numregim
miltcoup   0.35930697  0.13527886  0.03025023  0.25553451
oligarchy  0.34925939  0.30189567 -0.19042153  0.46160153
pollib    -0.15424186 -0.18475293 -0.16596383 -0.16189264
parties   -0.07516457  0.06717680  0.33215436  0.19228370
pctvote   -0.17319645 -0.09707005  0.23332118  0.15283204
popn       1.00000000  0.42715960  0.05890562 -0.14043668
size       0.42715960  1.00000000  0.20616525  0.02213111
numelec    0.05890562  0.20616525  1.00000000  0.22990580
numregim  -0.14043668  0.02213111  0.22990580  1.00000000

Call:
glm(formula = miltcoup ~ oligarchy + parties + pollib, family = poisson(link = "log"),
    data = africa)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3609  -1.0407  -0.3153   0.6145   1.7536

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.207981   0.445679   0.467   0.6407
oligarchy    0.091466   0.022563   4.054 5.04e-05 ***
parties      0.022358   0.009098   2.458   0.0140 *
pollib1     -0.495414   0.475645  -1.042   0.2976
pollib2     -1.112086   0.459492  -2.420   0.0155 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 65.945  on 35  degrees of freedom
Residual deviance: 32.822  on 31  degrees of freedom
AIC: 107.63

Number of Fisher Scoring iterations: 5
```