Final Exam, Math 463, Spring 2017

Instructions. READ CAREFULLY.:

(i) The work you turn in must be your own. You may not discuss the final with anyone, either in the class or outside the class. (You may of course consult with me for any clarification.)

(ii) The exam is designed to be completed requiring only your course notes. You may consult written material on R or background material, but you must list any written (or online) source consulted not solely related to R, including specific pages.

Failure to follow the above will be considered an instance of academic dishonesty.

(iii) Your final must be clearly written and legible. **I will not grade problems which are sloppily presented and such problems will receive a grade of** 0. If you are unable to write legibly and clearly, use a word processor. Use of a word processor is highly recommended. Budget time for writing up your solutions. I will not accept any files created with a camera, e.g. mobile device.

(iv) Think about your exposition. Your answer is only correct if I can understand what you have done. Pay attention to style, grammar, and spelling.

(v) Include all computer code you have used in a single separate file. **Do not include the code or unformatted computer output in the main body of your final**. **If I have to consult your computer code to determine what you have done, you have not provided sufficient explanation in the text.**

(vi) **Late finals will not be accepted**.

(vii) **Include the following statement at top of your exam if it is truthful: "By submitting this final, I certify that I have followed exactly the rules outlined on the front of the exam."** Exams not containing this statement will be given a grade of 0.

Problem	Points
1	10
2	10
3	10
4	10
5	10
6	10
TOTAL	60

Problem 1. Consider the data available at

```
> ozone =
+   read.table("http://pages.uoregon.edu/dlevin/DATA/ozo.txt",
+                      header=T)
```

The variables `temp` and `humidity` give daily temperature and humidity readings. The variable `HO` indicates if ozone levels are high. Use the data to fit a probit model: Here let $Y_i = 1$ if and only if the ozone level is high, and write $\boldsymbol{x}^{(i)} = (1, \texttt{temp}_i, \texttt{humidity}_i)$.

$$\mathbb{P}(Y_i = 1 \mid \boldsymbol{x}^{(i)}) = \Phi(\boldsymbol{x}^{(i)}\boldsymbol{\beta}),$$

where Φ is the normal cdf. Provide the fitted coefficients and their standard errors. (The R function `glm` can be used to fit a probit. Be sure to specify `family=binomial(link="probit")`.)

(a) What is the estimated probability of a high ozone day if the temperature is 95 degrees and the humidity is 80%?

(b) Find a 95% confidence interval for the linear predictor $\beta_0 + 95\beta_1 + 80\beta_2$ at these values.

(c) Find a 95% confidence interval for the probability of high ozone at these values.

Note that if `f` is the fitted probit model (using `glm`), then `summary(f)$cov.unscaled` gives the approximate covariance matrix $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$. Alternatively, `predict` can give fitted values and their standard errors, for given covariates.

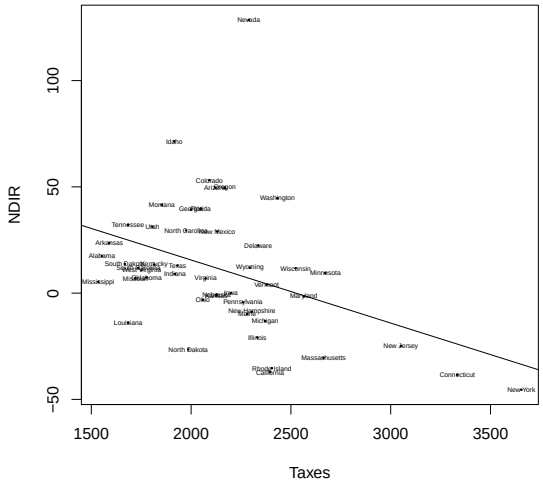

Figure 1: Immigration vs. taxes.

Problem 2. This problem concerns the data available at the location specified in the R code below:

```
> di = read.table("http://pages.uoregon.edu/dlevin/DATA/DI.txt",
+                 header=T,row.names=1,sep="\t")
> plot(NDIR~Taxes, data=di, pch=19, cex=0.2)
> text(di$Taxes,di$NDIR,row.names(di), cex=0.4)
> g = lm(NDIR~Taxes, data=di)
> abline(g)
```

Figure 1 is a scatterplot of net immigration to states against income tax. (The data is aggregrated over a few years in the early 90's.)

Do people move because of tax rates? Use the data in the file (see above) to discuss this question.

Problem 3. Recall that Instrumental Variables Least Squares (IVLS) requires variables $\boldsymbol{Z}$ which are independent of the error terms $\boldsymbol{\varepsilon}$. (Such variables are called *exogenous*.) Succesful application hinges on this assumption. Can this be verified from the data? This problem explores this question.

Suppose that

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,. \tag{1}$$

Assume $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$. Let $\boldsymbol{Z}$ be a random n-vector, and suppose that $\mathrm{Cov}(Z_i, \varepsilon_i) = \rho$. The triples $(Z_i, X_i, \varepsilon_i)$ are i.i.d. as triples for $i = 1, 2, \ldots, n$.

(a) Show that

$$n^{-1}\sum_{i=1}^{n} Z_i\varepsilon_i \to \rho\,. \tag{2}$$

(b) Show that if n is large enough, *and you can observe* $\boldsymbol{\varepsilon}$, you can test $H_0 : \rho = 0$ with power 0.99 against the alternative $H_1 : |\rho| > 0.001$.

Hint: By the CLT, the test statistic

$$\sqrt{n}\left(n^{-1}\sum_{i=1}^{n} Z_i\varepsilon_i - \rho\right) \approx N(0, \kappa)$$

where $\kappa = \mathrm{Var}(Z_1\varepsilon_1)$.

Thus, with enough data, you can determine with high probability if the errors $\boldsymbol{\varepsilon}$ are correlated with $\boldsymbol{Z}$. (Provided you can observe $\boldsymbol{\varepsilon}$. In most applications, $\boldsymbol{\varepsilon}$ is unobservable, however.)

(c) Let $\boldsymbol{e}$ be the residuals from the OLS fit in (1). Find the limit

$$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} Z_i e_i = \lim_{n\to\infty} n^{-1}\langle \boldsymbol{Z}, \boldsymbol{e}\rangle\,.$$

Is it the same as the limit in (2)?

(d) Can you then use the residuals $\boldsymbol{e}$ to determine $\mathrm{Cov}(Z_1, \varepsilon_1)$?

(e) If "no" what does this say about the ability to verify exogeniety (independence from error term) of instrumental variables?

Problem 4. Suppose that $(Z_{i,1}, Z_{i,2}, \delta_i, \varepsilon_i)$ are IID (as 4-tuples) jointly normal with mean 0. Assume that $(Z_{i,1}, Z_{i,2})$ is independent of $(\delta_i, \varepsilon_i)$, the components of Z_i are independent with variance 1, and $\mathrm{Var}(\varepsilon_i) = \sigma^2$ and $\mathrm{Var}(\delta_i) = \tau^2$. Let $\rho = \mathrm{cor}(\delta_i, \varepsilon_i)$. Let

$$X_i = Z_{i,1} + 2Z_{i,2} + \delta_i\,.$$

Suppose that

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}\,.$$

The variables $\boldsymbol{Z}_1, \boldsymbol{Z}_2$ are instruments used to estimate β. Let $\tilde{\beta}$ denote the IVLS estimator of β. Let $\hat{\beta}$ denote the OLS estimator of β.

(a) If $n = 10, \beta = 0.5$ and $\sigma = 1 = \tau = 1$, use simulation to estimate the mean-square error of both $\tilde{\beta}$ and $\hat{\beta}$:

$$\sqrt{\mathbb{E}_\beta[\|\tilde{\beta} - \beta\|^2]}, \quad \sqrt{\mathbb{E}_\beta[\|\hat{\beta} - \beta\|^2]}$$

Do this for $\rho = 0.8, 0.3, 0$.

(b) Do the same for $n = 10000$. Repeat both for $\tau = 50$.

(c) For $n = 10$ and $n = 10000$: Estimate the standard errors for both estimators. Which one is larger? Estimate the bias for both estimators. Which one is larger?

(d) Which estimator is better when $n = 10$. When $n = 100$? When $n = 100000$?

Problem 5. Suppose that

$$\boldsymbol{Y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x} + \boldsymbol{\varepsilon},$$

where $\{\varepsilon_i\}$ are uncorrelated, and $\mathrm{Var}(\varepsilon_i) = \sigma^2 \times x_i^2$.

(a) Is the OLS estimator for β_1 unbiased?

(b) Are the standard errors reported for the OLS estimator correct? (That is, good estimates of the actual standard deviation of the OLS estimator when applied to data generated from this model.) Give an expression for the standard deviation of the OLS estimators for this model, in terms of σ and $\boldsymbol{x}$.

(c) Write down explicitly the GLS estimator of $\boldsymbol{\beta}$ in terms of $\boldsymbol{Y}$ and $\boldsymbol{x}$.

(d) Suppose that instead, $\mathrm{Var}(\varepsilon_i) = \sigma^2 a_i$, where a_i is a constant that takes on one of four variables depending on a categorical variable w_i. Describe the strategy of the feasible GLS estimator.

(e) Suppose that $\{X_i\}_{i=1}^n$ are i.i.d. $N(0,1)$. Suppose also that w_i is each equally likely to take on any of its four values. Assume that the truth is $(a_1, a_2, a_3, a_4) = (1,2,4,8)$. For $n = 25$ and $n = 1000$, use simulation to estimate the true standard error of the OLS estimator and the true standard error of the feasible GLS estimator (implement the strategy above.) Estimate the bias in the reported standard error when using OLS from the true standard error of the OLS estimate (is it zero?).

Assume that $Y_i = 3.2 + 2.4x_i + \varepsilon_i$ and $\sigma = 10$.

Problem 6. Suppose that

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix}$$

Let $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Assume the errors are i.i.d. $N(0, \sigma^2)$. Find σ so that the power of the F-test of

$$H_0 : \beta_3 = \beta_4 = 0$$

is 0.95 against the alternative $\beta_3 = \beta_4 = 0.1$.

Do the same with the matrix

$$\boldsymbol{X} = \begin{bmatrix} -1.70 & -1.45 & -0.55 & -0.85 \\ -0.09 & -0.01 & 0.02 & 1.32 \\ -1.03 & -1.27 & -1.47 & 0.24 \\ -0.49 & 0.39 & -1.26 & -0.57 \\ -0.42 & -2.25 & -0.93 & 0.02 \\ 0.45 & 0.66 & 0.15 & 1.41 \\ 0.33 & 0.33 & 0.92 & -0.36 \\ 0.31 & -0.78 & 0.72 & 0.34 \end{bmatrix}$$

If the answer differs, explain why.