FINAL EXAM
MATH 463 - SPRING 2017

ALEX THIES

Introduction. By submitting this final, I certify that I have followed exactly the rules outlined on the front of the exam.

1. Assignment

1.1. **Problem 1.** Consider the data available at

```
rm(list=ls())
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 $\text{Cov}(\hat{\boldsymbol{\beta}})$. Alternatively, `predict` can give fitted values and their standard errors, for given covariates.

Solution.

(a) We compute the following in R,

$$\begin{aligned}\mathbb{P}(X\beta|\beta_{\text{Temp}} = 95, \beta_{\text{Humid}} = 80) &= \Phi(x_0 + \beta_{\text{Temp}}x_1 + \beta_{\text{Humid}}x_2),\\ &= 0.98.\end{aligned}$$

(b) We compute a 95% following confidence interval for $X\beta$ given the specified values for temperature and humidity in R,

$$(1.52, 2.74).$$

(c) Utilizing the endpoints of the previous confidence interval, we compute the following 95% confidence interval for high ozone given the values of temperature and humidity,

$$(\Phi(1.52), \Phi(2.74)) = (0.94, 1)\,.$$

□

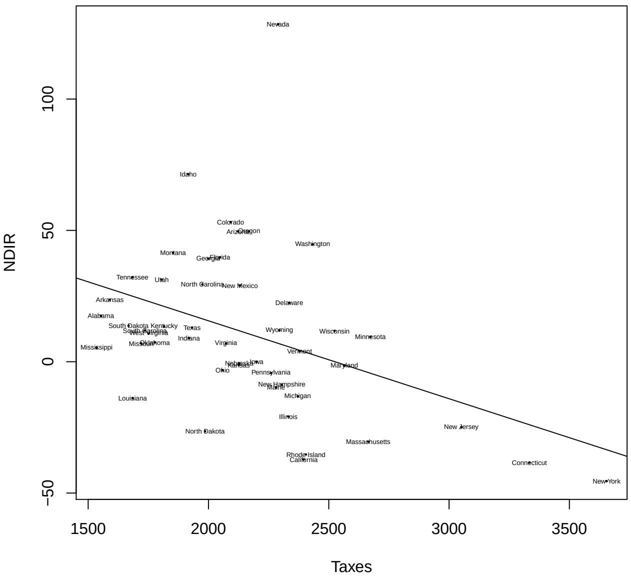

Figure 1. Immigration vs. taxes

1.2. **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.

Solution. Upon inspection of the regression of Taxes onto NDIR illustrated in 1, we observe that there appears to be negative correlation. Looking at the summary of this model (Table 1) we observe that none of the variables are highly significant, including taxes.

Imagine you are an overly taxed New Yorker looking to move, are you more likely to traverse the Continental United States for a lessened tax burden, or over the border say Pennsylvania, or Vermont? Suppose we treat region as a categorical variable and fit different slopes and intercepts according to region. Figure 2 seems to indicate that one's tax burden will spur them to move, depending on where they live. Figure 3 illustrates that the regions which are experiencing the most out-migration are the two regions which have the highest tax burden in favor of regions with a lower tax burden. We can observe the ANOVA table in Table 2 to see that

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	82.1620	152.5896	0.54	0.5938
Unemp	-1.4120	5.3201	-0.27	0.7923
Wage	-0.3928	4.1049	-0.10	0.9243
Crime	0.0196	0.0249	0.79	0.4373
Income	-0.0018	0.0020	-0.88	0.3832
Metrop	-0.0058	0.3610	-0.02	0.9872
Poor	-2.5828	2.2324	-1.16	0.2554
Taxes	-0.0131	0.0180	-0.73	0.4723
Educ	0.6199	1.7484	0.35	0.7251
BusFail	14.9455	48.3529	0.31	0.7591
Temp	-0.2168	1.2619	-0.17	0.8646
RegionNortheast	-16.8286	17.7235	-0.95	0.3491
RegionSouth	16.5672	16.7952	0.99	0.3309
RegionWest	36.8358	13.8137	2.67	0.0116

Table 1. Summary of Big Model

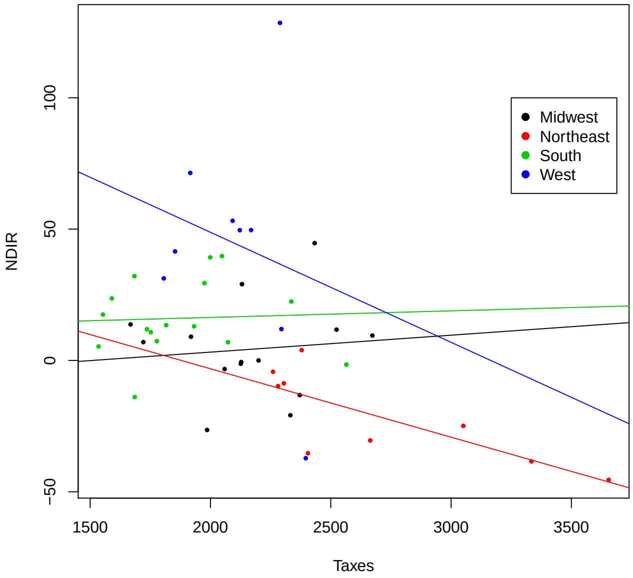

Figure 2. Immigration vs. taxes, by region

indeed, when split by the category of region, people may in fact move because of taxes.

Some concerns that I have with the data in question are that factors such as income and wage may be correlated, if not linearly then in some fashion. I have the same concern with unemployment, business failure rate, and education. It also seems likely that the errors here are not necessarily independent of the variables.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	47	46469.15				
2	46	38769.06	1	7700.09	12.98	0.0009
3	43	25089.12	3	13679.94	7.69	0.0004
4	40	23733.04	3	1356.08	0.76	0.5222

Table 2. ANOVA

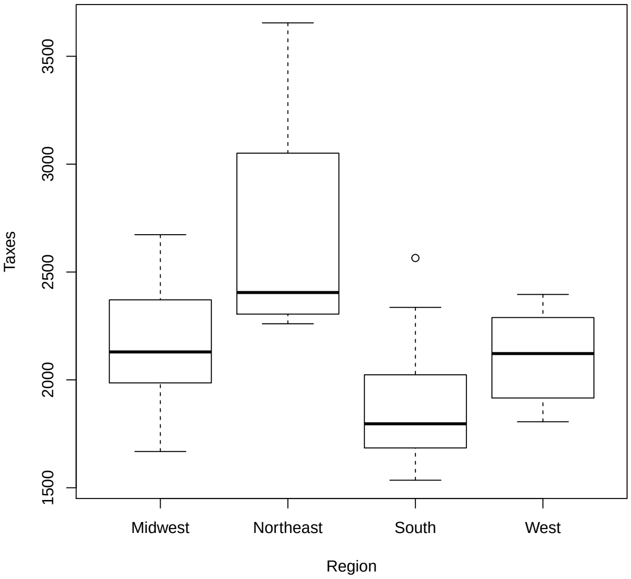

Figure 3. Region as Taxes

Many things have an effect on factors such as business failure rate, and education that are not taken into consideration by the model. Moreover, factors which might serve to answer the question posed in this problem could be things like the rate of change of the GDP of a state, the robust-ness of the welfare state, and the type of taxes which are being summed into the taxes variable. It seems unlikely that renters would be effected by increases in property tax, whereas they would be effected by increases in income, or sales tax. This question may be more easily answered given a newer set of observational data, especially given that Kansas has recently undergone massive tax cuts, effectively running an experiment for us. □

1.3. **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 $i = 1, 2, \dots, n$.

(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?

Solution.

(a)

(b)

(c)

(d)

(e)

□

1.4. **Problem 4.** Suppose that

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

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

(a) If $n = 10, \boldsymbol{\beta} = (0.2, 0.5)$ and $\sigma = 1$, use simulation to estimate the mean-square error of both $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$:

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

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 $n = 100000$?

Solution. We perform the requested simulations and produce the following results,

n	ρ	τ	MSE($\hat{\boldsymbol{\beta}}$)	MSE($\tilde{\boldsymbol{\beta}}$)	SE($\hat{\boldsymbol{\beta}}$)	SE($\tilde{\boldsymbol{\beta}}$)	Bias $\hat{\boldsymbol{\beta}}$	Bias $\tilde{\boldsymbol{\beta}}$
10	0.8	1	0.13	0.14	0.1	0.1	-0.09	-0.09
10	0.3	1	0.16	0.15	0.12	0.13	0.1	0.07
10	0	1	0.23	0.26	0.22	0.26	-0.08	0.01
10000	0.8	1	0	0	0	0	-3.59×10^{-4}	0
10000	0.3	1	0	0.01	0	0	0	0
10000	0	1	0.01	0.01	0	0	0.01	0.01
10	0.8	50	0.2	0.19	0.17	0.19	0.1	0.06
10	0.3	50	0.15	0.18	0.15	0.17	-0.03	-0.05
10	0	50	0.36	0.41	0.19	0.22	-0.3	-0.35
10000	0.8	50	0.01	0.01	0	0	-0.01	-0.01
10000	0.3	50	0.01	0.01	0	0	0	0.01
10000	0	50	0	0	0	0	-7.93×10^{-4}	0

□

1.5. **Problem 5.** Suppose that

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

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

(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 $\sigma = 10$.

Solution.

(a) No, in order for the OLS estimator for any β to be unbiased we must assume that $\boldsymbol{\varepsilon}_i$ are I.I.D., we do not make that assumption in this case.

(b) No, we proceed with the definition of $\mathrm{Var}(\hat{\beta}|x)$ and find that the standard errors are not $\mathrm{SE}(\hat{\beta}) = \sigma^2 \mathbb{I}_{n\times n}$. We compute the following, note that $\mathrm{Var}(Y|X) = \mathrm{Var}(\varepsilon\,|X)$.

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}|x) &= \mathrm{Var}\left[(X'X)^{-1} X'Y|X\right], \\
&= (X'X)^{-1} X' \,\mathrm{Var}\,[Y|X] \left[(X'X)^{-1} X'\right]', \\
&= (X'X)^{-1} X' \,\mathrm{Var}\,[\varepsilon\,|X] \left[(X'X)^{-1} X'\right]', \\
&= (X'X)^{-1} X'\sigma^2 X^2 \left[(X'X)^{-1} X'\right]', \\
&= \sigma^2 \left\{(X'X)^{-1} X'X \cdot X \left[(X'X)^{-1} X'\right]'\right\}.
\end{aligned}$$

It is at this point where my matrix algebra skills fail me as I am crunched for time. However, I am certain that the terms inside the brackets do not simplify to the identity matrix, thus the standard error is not $\sigma^2 \mathbb{I}_{n\times n}$.

(c)

(d)

(e)

□

1.6. **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.

Solution. Given that the constraint that the power is 0.95, we can compute σ by first determining the numerator of the non-centrality parameter δ and the appropriate critical value $F^\star$, and then using a custom R function to determine an appropriate σ by trial and error. In order to compute the numerator of δ, and critical value we do the following in R

```
V3perp <- residuals(lm(V3~.-V4, data = X))
V4perp <- residuals(lm(V4~.-V3, data = X))
d1.num <- ((0.1^2)*sum(V3perp^2) + (0.1^2)*sum(V4perp^2)
           + (2*0.1*0.1)*sum(V3perp*V4perp))

fstar <- qf(0.95, 2, 4)
```

Doing this yields that the numerator of δ for the first matrix is $\delta_{\text{Num}} = 0.16$ and for the second matrix, $\delta_{\text{Num}} = 0.05$. We then use `d1.num`, and $F^\star =$ `fstar` in a custom function to trial and error our way to an appropriate σ.

```
powah <- function(crit,del,sig){1-pf(crit,2,4,I(del/sig**2))}
```

After much trial and error we find that a reasonable standard deviation for the first model is $\sigma_1 \doteq 0.06$ and for the second model $\sigma_2 \doteq 0.04$. I suspect that these differ because the models in question appear to have little in common beyond their dimension. □

E-mail address: `athies@uoregon.edu`