FINAL EXAM
MATH 463 - SPRING 2017

ALEX THIES

1. Assignment

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

Solution.

(a) We compute the following,
(b) We compute the following,
(c) We compute the following,

□

1.2. **Problem 2.** This problem concerns the data available at the location specified in the R code below: Figure !!! 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 FIGURE , we observe that there appears to be negative correlation. We WORDS and find that in the large model none of the variables are highly significant, so we have to consider WORDS. 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. If we were to find negative correlation among regions, and

additionally have high confidence that taxes is a significant factor, then it would be reasonable to assert that people move because of taxes. □

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 $i = 1, 2, \ldots, n$.

(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?

Solution.

(a) From the definition
(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 $\hat{\boldsymbol{\beta}}$:

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

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.

(a)
(b)
(c)
(d)

□

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 $\mathrm{Var}(\hat{\beta}|x)$,

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

Thus the standard deviation in terms of σ and x is $\mathrm{SD}(\hat{\beta}|x) = \sigma X\sqrt{(X'X)^{-1}}$, which is not $\sigma^2\boldsymbol{I}_{n\times n}$.

(c) Same as 4, read slide on GLS
(d) Read the books
(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. □

E-mail address: athies@uoregon.edu