# Gov 2002: Final Practice

Spring 2021

## Contents

**Exercises** **1**
Problem 1: OLS properties and assumptions . . . . . 1
Problem 2: Interpreting regression results . . . . . 2
Problem 3: Grading an estimator . . . . . 4
Problem 4: Estimation with count processes . . . . . 5

**Solutions** **5**
Answer 1 . . . . . 5
Answer 2 . . . . . 6
Answer 3 . . . . . 8
Answer 4 . . . . . 10

## Exercises

### Problem 1: OLS properties and assumptions

For each of the following questions, state whether the statement is true or false. If the statement is false, give a reason why it is false or provide a counterexample to the statement.

1. A 95% confidence interval for a regression coefficient estimate will cover the true regression coefficient 95% of the time.
2. Heteroskedasticity in the regression errors leads to bias in the estimates of the regression coefficients and the standard errors for those estimates.
3. An unbiased estimator is always consistent.
4. Under the Gauss-Markov assumptions, OLS has the lowest variance of any estimator for the regression coefficients.
5. In a simple linear regression with a binary independent variable, linearity is not an assumption and is satisfied by definition.

6. In large samples, there is no downside to using heteroskedasticity-consistent standard errors.

7. Under assumptions 1-5 of OLS (including homoskedasticity), the variance-covariance matrix of the regression estimator, $\widehat{\boldsymbol{\beta}}$, is $\sigma_u^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix.

## Problem 2: Interpreting regression results

Around 1980, China adopted the so—called "one—child policy," in which parents were strongly encouraged to have only one child (except under exceptional circumstances, for example if their first child was disabled). A series of financial rewards for complying with the policy, and penalties for exceeding one child, were put in place. An important question is, just how responsive are individuals to such financial incentives when making their fertility decisions? Theory and intuition may give us some ideas, but ultimately, this is an empirical question. The following represents an attempt to answer this question with the tools learned in this course.

In 1983, rural localities in China were given freedom in choosing the size of the penalties for households exceeding the one child threshold. As a result, the penalties varied substantially across localities, from as low as 30 yuan, to as high as 4,000 yuan ($1 U.S.=5.5 yuan in 1992). We make use of micro-level data from a survey in which women were asked about the number of children they have given birth to. We can regress this variable on characteristics of the woman which determine fertility (like education, age, etc.) and the financial penalty in her locality for exceeding one child. The results from this regression are presented below. The two columns represent results from separate regressions. The numbers in the tables are the coefficients on the variables, with the t—statistics presented below the coefficients in parentheses. The reference category for educational attainment dummies are those with no elementary education.

### (a)

Focus first on column 1 and assume the model is "correct'' in the sense that you can adopt causal language. In detail, discuss the effects of the penalty on the number of children. That is, discuss the: exact interpretation, statistical significance, and substantive significance of the effect. (For comparison, the average wealth per person for households in the sample was 4,000 yuan.)

### (b)

Column 1 again: If a local administrator believes this model, and wants to lower average fertility in a village by 1 child, by how much should he/she increase the penalty for couples exceeding one child? (give approximate value, or show how you would compute this quantity).

| Model | (1) | (2) |
|---|---|---|
| Age (in years) | .021 | .023 |
| | (.38) | (.40) |
| Dummy for completed elementary School | −.011 | .040 |
| | (−0.193) | (.712) |
| Dummy for completed middle school or more | −.052 | .043 |
| | (−.855) | (.697) |
| Penalty (in 1,000 yuan) | −0.468 | −0.338 |
| | (−3.63) | (−2.00) |
| Household wealth (in 1,000 yuan) | .255 | .513 |
| | (4.81) | (7.28) |
| Penalty * Household wealth | | .005 |
| | | (2.23) |
| Constant | 3.3 | 5.07 |
| | (30.4) | (24.2) |
| $R^2$ | 0.058 | 0.111 |
| Number of Observations | 5532 | 5532 |

Table 1: Results of regression of number of children on covariates

**(c)**

In column 2, we interact household wealth with the size of the penalty. Why would one want to do that, and what is the interpretation of the results?

**(d)**

For column 2, what is the standard error of the coefficient on the penalty? Provide an interpretation for what this standard error means.

**(e)**

Suppose we want to test the joint hypothesis that all of the slope coefficients in the second regression are zero. Can you set up the relevant test statistic from the information given in the table? If yes, set up the statistic and calculate a p-value. If not, explain what additional information you would need to perform the test.

**(f)**

Do you think a causal interpretation of the regression coefficient on penalty is appropriate here? Discuss why or why not.

**(g)**

If you were given a modest amount of funding (enough to hire two graduate RAs for a year) to improve the data for this study, where would you focus your efforts? What would be the benefits of this?

## Problem 3: Grading an estimator

Suppose you observe $n$ independent observations $X_1, \ldots, X_n$. Each $X_i$ has a normal distribution with mean $\mu$ and variance $\frac{\sigma^2}{c_i}$ where $\sigma^2$ is a known constant and $c_1, \ldots, c_n$ are known positive constants specific to each observation $i$. That is,

$$X_i \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{c_i}\right)$$

You are interested in estimating the unknown constant $\mu$.

(a) A common estimator of $\mu$ is the sample mean $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$. However, a friend suggests an alternative estimator $\bar{X}_w$.

$$\bar{X}_w = \frac{\sum_{i=1}^{n} \sqrt{c_i} X_i}{\sum_{i=1}^{n} \sqrt{c_i}}$$

Find the expectation of $\bar{X}_w$.

(b) Is $\bar{X}_w$ biased?

(c) Derive the variance of $\bar{X}_w$.

(d) What is the sampling distribution of $\bar{X}_w$ when $n$ is large?

(e) What assumption must we make about $c_i$ to ensure that $\bar{X}_w$ is a *consistent* estimator for $\mu$?

(f) This set-up frequently appears in "meta-analyses" of previously run studies, where $c_i$ represents the sample size for each study. Suppose you have the following six estimates of $\mu$ obtained from sample means from six different and independent studies. Also suppose that it is known that the true population variance $\sigma^2 = 100$.

For these six studies generate a combined point estimate for $\mu$ using your friend's sample mean estimator $\bar{X}_w$. Construct a 95% confidence interval for the mean using this estimator. Interpret your confidence interval.

| Study Number | Point Estimate | Sample Size |
|---|---|---|
| 1 | 5.165 | 40 |
| 2 | 1.128 | 120 |
| 3 | 2.324 | 60 |
| 4 | 6.159 | 800 |
| 5 | 9.591 | 20 |
| 6 | 8.932 | 300 |

Table 2: Sample mean estimates and sample sizes for six independent studies

### Problem 4: Estimation with count processes

Let $Y$ be a random variable denoting the number of steps a person takes in a minute. Let $p$ be the probability that a person is sitting in a given minute, and when the person is not sitting we can model the number of steps they take using a Poisson model.

(a) Write down the PDF of $Y$?

(b) What is $E[Y]$ as a function of $p$?

(c) Is $Y$ an unbiased estimator of $\lambda$?

(d) What is $Var(Y)$? Use Eve's Law.

## Solutions

### Answer 1

1. A 95% confidence interval for a regression coefficient estimate will cover the true regression coefficient 95% of the time. (**TRUE**)
2. Heteroskedasticity in the regression errors leads to bias in the estimates of the regression coefficients and the standard errors for those estimates. (**FALSE - recall from section code that estimates of coefficients themselves don't change, biased for the standard errors**)
3. An unbiased estimator is always consistent. (**FALSE - using the first observation for the mean is unbiased, but inconsistent; more conceptually, the former is a small-sample property, the latter is a large-sample property.**)
4. Under the Gauss-Markov assumptions, OLS has the lowest variance of any estimator for the regression coefficients. (**FALSE - it has the lowest variance of linear unbiased estimators**)
5. In a simple linear regression with a binary independent variable, linearity is not an assumption and is satisfied by definition. (**TRUE**)

6. In large samples, there is no downside to using heteroskedasticity-consistent standard errors. (**FALSE - as we saw in section, HC SEs are generally larger than normal SEs; more conceptually, the HC SEs are inefficient relative to OLS if homoskedasticity is actually true; as a result, you might nullify a hypothis that is actually not null.**)

7. Under assumptions 1-5 of OLS (including homoskedasticity), the variance-covariance matrix of the regression estimator, $\widehat{\boldsymbol{\beta}}$, is $\sigma_u^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. (**FALSE - this is tricky: this is true of the variance-covariance matrix of the *error*, but as for the variance-covariance matrix of the *coefficients*, there will be some covariance between difference regression coefficient estimates**)

## Answer 2

### (a)

In expectation, an increase in the penalty by 1000 yuan decreases the number of children that a couple has by 0.468. In other words, the expected difference in the number of children had by two women living in regions with penalties that differ by 1000 yuan is 0.468.

This estimate is statistically significant at the .05 level (the absolute t-statistic is greater than 1.96). Substantively though, a very large penalty is needed to reduce expected fertility by 1 child - an increase in the penalty equal to half of the average wealth of households (2,000 yuan), in expectation, reduces the number of children by slightly less than 1 (0.936).

### (b)

In order to reduce expected fertility by 1 child, the penalty should be increased by approximately 2,136.8 Yuan.

$$\begin{aligned} -1 &= -0.468 \times \Delta\text{penalty} \\ \frac{-1}{-0.468} &= \Delta\text{penalty} \\ \Delta\text{penalty} &= 2.1368 \end{aligned}$$

### (c)

Because of the diminishing marginal utility of wealth, the penalty may have a larger effect on poorer families than on wealthier families. That is, wealthier families may be more willing to pay a large penalty in order to have an additional child compared to a poorer family that does not have the means. Including an interaction term allows the estimated effect of the penalty to vary with respect to household wealth.

We estimate that a 1,000 yuan increase in household wealth, reduces the expected negative effect of a 1, 000 yuan penalty on number of children by .005. This estimate is statistically significant at the .05 level (the t-statistic is greater than 1.96).

## (d)

The standard error is the coefficient divided by the t-statistic: $\frac{-0.338}{-2} = 0.169$

## (e)

When you hear joint hypothesis, you should think F-test. In order to construct the F-statistic for the F-test of all regression coefficients, we can use the $R^2$ value of the fitted model:

$$F = \frac{\frac{RSS_r - RSS_u}{q}}{\frac{RSS_u}{n-k-1}}$$

where $q$ is the number of restrictions and $k$ is the number of coefficients in the full model. In the F-test for all coefficients, the restricted $RSS$ is equal to the total sum of squares $TSS$ (i.e. $RSS_r = TSS$). Also, $q = k = 6$. Therefore for our particular test

$$F = \frac{\frac{(TSS - RSS)}{6}}{\frac{RSS}{5525}}$$

which has an F distribution with df 6 and 5525. We can re-write the statistic as:

$$F = \frac{(TSS - RSS)}{RSS} \times \frac{5525}{6}$$

We know by definition that $R^2 = 1 - \frac{RSS}{TSS} = \frac{(TSS - RSS)}{TSS}$. Therefore,

$$\frac{R^2}{1 - R^2} = \frac{(TSS - RSS)}{TSS} \times \frac{TSS}{RSS} = \frac{(TSS - RSS)}{RSS}$$

and,

$$F = \frac{R^2}{1 - R^2} \times \frac{5525}{6} = \frac{0.111}{1 - 0.111} \times \frac{5525}{6} \approx 114.97$$

The $p$-value is:

```
df(114.97, 6, 5525)
```

```
## [1] 8.815221e-137
```

(f)

This answer will vary - the general answer should be no (unless there's a very well justified yes). Problems include unoberved confounding - e.g. selection with respect to localities imposing the penalty (localities likely assign the penalty depending on what they know residents can tolerate), also selection on the part of households choosing where to live (and which penalties they will pay.

(g)

Again, there are many potential answers. If you mentioned potentially observable confounders in (f), you could decide to collect observations of those. Repeated observations on localities over time (looking at changes in the penalties) would also be a valid answer (to enable a diff-in-diff design).

## Answer 3

(a)

$$
\begin{aligned}
E\left[\bar{X}_w\right] &= E\left[\frac{\sum_{i=1}^n \sqrt{c_i} X_i}{\sum_{i=1}^n \sqrt{c_i}}\right] \\
&= \frac{1}{\sum_{i=1}^n \sqrt{c_i}} E\left[\sum_{i=1}^n \sqrt{c_i} X_i\right] \\
&= \frac{1}{\sum_{i=1}^n \sqrt{c_i}} \left[\sum_{i=1}^n E[\sqrt{c_i} X_i]\right] \\
&= \frac{1}{\sum_{i=1}^n \sqrt{c_i}} \left[\sum_{i=1}^n \sqrt{c_i} E[X_i]\right] \\
&= \frac{1}{\sum_{i=1}^n \sqrt{c_i}} \left[\sum_{i=1}^n \sqrt{c_i} \mu\right] \\
&= \mu \times \frac{1}{\sum_{i=1}^n \sqrt{c_i}} \left[\sum_{i=1}^n \sqrt{c_i}\right] \\
&= \mu \times \frac{\sum_{i=1}^n \sqrt{c_i}}{\sum_{i=1}^n \sqrt{c_i}} \\
&= \mu
\end{aligned}
$$

(b) No, in part (a) we found that $E[\bar{X}_w] = \mu$, and hence the bias is 0.

(c)

$$
\begin{aligned}
Var(\bar{X}_w) &= Var\left(\frac{\sum_{i=1}^{n}\sqrt{c_i}X_i}{\sum_{i=1}^{n}\sqrt{c_i}}\right) \\
&= \frac{1}{\left(\sum_{i=1}^{n}\sqrt{c_i}\right)^2} Var\left(\sum_{i=1}^{n}\sqrt{c_i}X_i\right) \\
&= \frac{1}{\left(\sum_{i=1}^{n}\sqrt{c_i}\right)^2} \sum_{i=1}^{n} Var(\sqrt{c_i}X_i) \\
&= \frac{1}{\left(\sum_{i=1}^{n}\sqrt{c_i}\right)^2} \sum_{i=1}^{n} c_i Var(X_i) \\
&= \frac{1}{\left(\sum_{i=1}^{n}\sqrt{c_i}\right)^2} \sum_{i=1}^{n} c_i \times \frac{\sigma^2}{c_i} \\
&= \frac{1}{\left(\sum_{i=1}^{n}\sqrt{c_i}\right)^2} \sum_{i=1}^{n} \sigma^2 \\
&= \frac{n\sigma^2}{\left(\sum_{i=1}^{n}\sqrt{c_i}\right)^2}
\end{aligned}
$$

(d) By the CLT, the distribution of $\bar{X}_w$ is:

$$\bar{X}_w \sim N\left(\mu, \frac{n\sigma^2}{(\sum_{i=1}^{n}\sqrt{c_i})^2}\right)$$

(e) For consistency to hold, $Var(\bar{X}_w) \xrightarrow{n} 0$. We're also given in the problem setup that each $c_i$ is some positive constant such that $c_i > 0$. Therefore to get the variance to go to 0, we need the denominator to go to infinity faster than the numerator. The additional assumption we need to make is that $\sum_{i=1}^{n}\sqrt{c_i}$ *diverges* as $n \to \infty$.

(f)

Our point estimate of $\mu$ using $\bar{X}_w$, $\bar{x}_w$, is:

$$\begin{aligned}\bar{x}_w &= \frac{\sqrt{40} \times 5.165 + \sqrt{120} \times 1.128 + \sqrt{60} \times 2.324 + \sqrt{800} \times 6.159 + \sqrt{20} \times 9.591 + \sqrt{300} \times 8.932}{\sqrt{40} + \sqrt{120} + \sqrt{60} + \sqrt{800} + \sqrt{20} + \sqrt{300}} \\ &\approx \frac{434.826}{75.102} \\ &\approx 5.790\end{aligned}$$

The standard error of our estimator is

$$\begin{aligned}\sqrt{Var(\bar{X}_w)} &= \sqrt{\frac{6 \times 100}{\left(\sqrt{40} + \sqrt{120} + \sqrt{60} + \sqrt{800} + \sqrt{20} + \sqrt{300}\right)^2}} \\ &= 0.326\end{aligned}$$

So our 95% confidence interval for this estimate is $[5.790 \pm 1.96 \times 0.326] = [5.151, 6.429]$. We are 95% confident that the true mean lies within the interval $[5.151, 6.429]$. That is, in repeated samples, we would draw a sample where the 95% confidence interval covered the true mean 95% of the time.

## Answer 4

(a) Let $Z$ be an indicator variable for whether the person is sitting. This is known as a Zero-inflated Poisson model. By LOTP:

$$\Pr(Y = 0) = \Pr(Y = 0|Z = 1)\Pr(Z = 1) + \Pr(Y = 0|Z = 0)\Pr(Z = 0) = p + (1 - p)e^{-\lambda}$$

And for $y \in 1, 2, ...,$

$$\Pr(Y = y) = (1 - p)\frac{\lambda^y e^{-\lambda}}{y!}$$

(b) By LIE:

$$E[Y] = E[Y|Z = 1] \cdot \Pr(Z = 1) + \underbrace{E[Y|Z = 0]}_{\text{Expectation of a Poisson rv}} \cdot \Pr(Z = 0)$$

$$= 0 + \lambda(1 - p) = \lambda(1 - p)$$

(c) In part (b) we found that $E[Y] = \lambda(1-p) \leq \lambda$. The estimator is unbiased only in the case where $p = 0$.

(d) By Eve's law:

$$Var(Y) = E[Var(Y|Z)] + Var(E[Y|Z])$$

Where,

$$E[Var(Y|Z)] = 0 + \lambda \cdot \Pr(Z = 0) = \lambda \cdot (1-p)$$

Next, it is useful to note that $Y = (1-Z) \cdot X$, where $X \sim Pois(\lambda)$. Thus:

$$Var(E[Y|Z]) = Var(E[(1-Z) \cdot X|Z])$$

$$Var((1-Z) \cdot E[X])$$

$$= \lambda^2 \cdot p \cdot (1-p)$$

Thus, putting this together we find that

$$Var(Y) = \lambda \cdot (1-p) + \lambda^2 \cdot p \cdot (1-p)$$

$$= \lambda(1-p)(1+\lambda p)$$