Problem Set 2, Solutions

Heteroskedasticity, Consistency, and Time Series

EC 421: Introduction to Econometrics

Due *before* noon (11:59am) on Saturday, 09 February 2019

DUE Your solutions to this problem set are due *before* noon on Saturday, 09 February 2019. Your files must be uploaded to Canvas—including (1) your responses/answers to the question and (2) the R script you used to generate your answers. Each student must turn in her/his own answers.

OBJECTIVE This problem set has three purposes: (1) reinforce the econometrics topics we reviewed in class; (2) build your R toolset; (3) start building your intuition about causality and time series within econometrics.

Heteroskedasticity

1. We are interested in using OLS to estimate the model

$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{1}$$

where x_i is a categorical variable that takes the values 1, 2, or 3.

Suppose that we know $\text{Var}(u_i|x_i = 1) = 3$ and $\text{Var}(u_i|x_i = 2) = 3$. We do not know $\text{Var}(u_i|x_i = 3)$, *i.e.*, $\text{Var}(u_i|x_i = 3) = \sigma_3^2$ for some unknown parameter σ_3^2.

1a. What value must σ_3^2 take for our model to be homoskedastic?

ANSWER

If $\sigma_3^2 = 3$, then we have homoskedastic disturbances.

1b. If $\sigma_3^2 \neq 3$, is OLS still unbiased? Is it still the most efficient linear estimator? Explain your answer.

ANSWER

If $\sigma_3^2 \neq 3$, then our disturbances are heteroskedastic. In the presence of heteroskedasticity, OLS is still unbiased (we do not heteroskedasticity in our proof that OLS is unbiased). However, in the presence of heteroskedasticity, OLS is no longer the most efficient linear unbiased estimator. WLS is more efficient.

1c. *Goldfeld-Quandt* In order to test whether the data we will use to estimate (1) are homoskedastic/heteroskedastic, we will run a Goldfeld-Quandt test.

We estimate (1) for the upper one third of the dataset (sorted on x) and find SSE$_3$=1,000. We estimate (1) on the middle third and find SSE$_2$=800. Finally, we estimate (1) on the lower third and find SSE$_1$=600. Each of these three groups has 100 observations.

Conduct a Goldfeld-Quandt test. State your hypotheses, calculate the G-Q test statistic, determine the *p*-value, state your conclusion.

Hint: You can use the function `pf(q, df1, df2, lower.tail = F)` to calculate the probability of observing a value of `q` or greater in an F distribution with `df1` numerator degrees of freedom and `df2` denominator degrees of freedom.

ANSWER

Hypotheses: We are testing the null hypothesis H$_0$: $\sigma_1^2 = \sigma_3^2$ against the alternative hypothesis H$_a$: $\sigma_1^2 \neq \sigma_3^2$.

Test statistic: Our Goldfeld-Quandt test statistic is $F_{\text{GQ}} = \dfrac{\text{SSE}_3}{\text{SSE}_1} = \dfrac{1,000}{600} \approx 1.67$.

We test this test statistic against an $F_{98,98}$ distribution, which produces a p-value of

```
pf(q = 1000/600, df1 = 100-2, df2 = 100-2, lower.tail = F)
```

```
## [1] 0.006061599
```

Conclusion: At the 0.05 level, we reject H$_0$ and conclude that there is statistically significant evidence that σ_1^2 and σ_3^2 differ. Therefore we have statistically significant evidence of heteroskedasticity.

2. The dataset in this questions comes from "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States" by Chetty, Hendren, Kline, and Saez—published in *The Quarterly Journal of Economics* (QJE) in 2014. Our outcome variable will be *the probability that an individual born to parents in the bottom fifth of the income distribution makes it into the top fifth of the income distribution*. This measure differs from the main outcome in the paper, but it is also very interesting—and it helps simplify our problem set. An individual observation in this dataset represents a commuting zone in the United States.

2a. Open up Rstudio, an `R` script, load whichever packages you want, and load the dataset contained in `dataPS02.csv`.

```
# Load 'pacman'
library(pacman)
# Load additional packages
p_load(tidyverse, broom, magrittr, ggplot2, ggthemes)
# Load the dataset
mobility_df <- read_csv("dataPS02.csv")
# Check the dataset
head(mobility_df)
```

```
## # A tibble: 6 x 6
##   prob_q5_q1 i_urban share_black share_middlecla… share_divorced
##        <dbl>   <int>       <dbl>            <dbl>          <dbl>
## 1     0.0621       1      0.0208            0.548         0.110
## 2     0.0537       1      0.0198            0.538         0.116
## 3     0.0731       0      0.0146            0.467         0.113
## 4     0.0563       1      0.0564            0.504         0.114
## 5     0.0446       1      0.174             0.500         0.0924
## 6     0.0519       0      0.224             0.538         0.0956
## # … with 1 more variable: share_married <dbl>
```

2b Describe the distribution of our main variable of interest (`prob_q5_q1`). You can provide statistical or graphical descriptions of this variable—try `summary(dataset$variable)` and `hist(dataset$variable)`, among others.

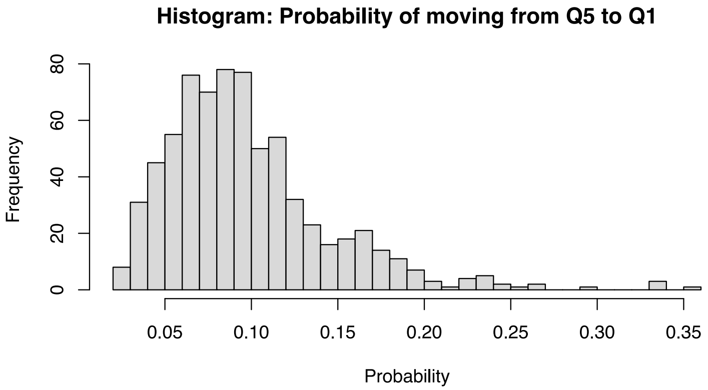
ANSWER

There is substantial variation in the probability an individual born to parents in the bottom fifth of the income distribution moves up to the top fifth. Some commuting zones nearly have zero probability, while others (the upper extremes) are approximately 30 percent probable. The median is approximately 0.089.

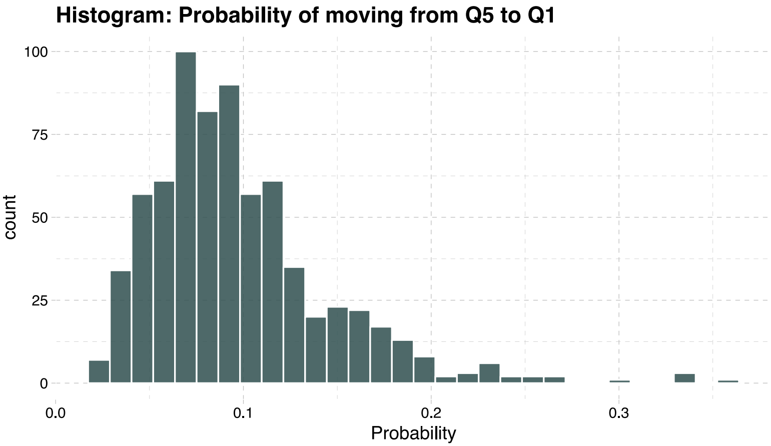
```r
# Summarize variable
summary(mobility_df$prob_q5_q1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02210 0.06588 0.08889 0.09761 0.11715 0.35714
```

```r
# A histogram using 'hist'
hist(
  mobility_df$prob_q5_q1,
  breaks = 25,
  col = "grey85",
  xlab = "Probability",
  main = "Histogram: Probability of moving from Q5 to Q1"
)
```



```r
# A histogram using 'ggplot'
ggplot(data = mobility_df, aes(x = prob_q5_q1)) +
  geom_histogram(fill = "darkslategrey", color = "white", alpha = 0.85) +
  xlab("Probability") +
  ggtitle("Histogram: Probability of moving from Q5 to Q1") +
  theme_pander()
```



2c. Regress the probability an individual moves from the bottom fifth of income to the top fifth of income (`prob_q5_q1`) on an intercept and the share of the commuting zone that is *middle class* (`share_middleclass`). Report your findings—the coefficients, brief interpretations of the coefficients, and whether the coefficients are statistically significant.

ANSWER

```r
# Estimate the model
reg_2c ← lm(prob_q5_q1 ~ share_middleclass, data = mobility_df)
# Report the results
reg_2c %>% tidy()
```

```
## # A tibble: 2 x 5
##   term              estimate std.error statistic  p.value
##   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)        -0.0971    0.0104     -9.35 1.14e-19
## 2 share_middleclass   0.354     0.0187     18.9   5.55e-65
```

We estimate that the coefficient on the share of the middle class is approximately 0.354. This coefficient says that if the share middle class in a commuting zone increased by 1 percentage point (*e.g.*, from 23% to 24%), then we would expect the probability of moving from the bottom fifth to the top fifth of income to increase by 0.35%. Our estimate is statistically significant (different from zero) at the 5% level.

2d. Does it make sense to interpret the intercept in this case? Explain.

ANSWER

It does not make sense to interpert the intercept in this setting. The interpretation would be "the average mobility probability for a commuting zone with no middle-class population." In our data, the share of middleclass population ranges from 28% to 73%—zero percent is not reasonable (also evidenced by the fact that the intercept would suggest a negative probability).

2e. Plot the residuals from your regression in (2c) on the y axis and `share_middleclass` on the x axis. Do you see evidence of heteroskedasticity? Explain.

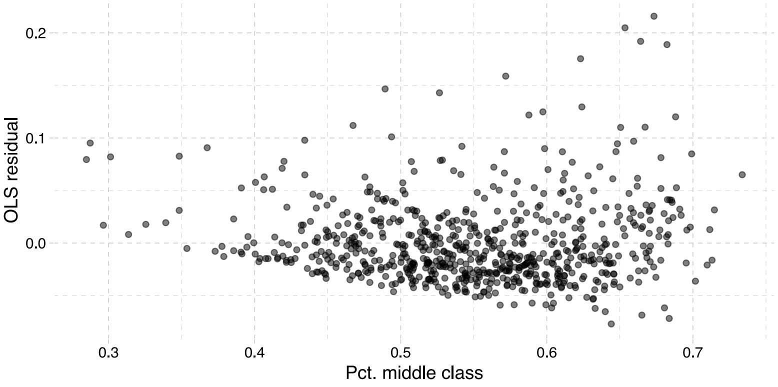
Hint$_1$: You can grab the residuals from a saved `lm` object by (1) using the `residuals()` function or (2) adding the suffix `$residuals` to the end of the `lm` object, *e.g.*, `my_reg$residuals` grabs the residuals from the `lm` object `my_reg`.

Hint$_2$: `plot(x = dataset$variable1, y = dataset$variable2)` makes quick and simple plots. You can also try `qplot()` from the package `ggplot2`, *i.e.*, `qplot(x = variable1, y = variable2, data = dataset)`.

ANSWER

Based upon the funnel-like figure below, heteroskedasticity seems likely.

```r
# Add residuals to the dataset
mobility_df %<>% mutate(e_2c = residuals(reg_2c))
# Plot with ggplot
ggplot(data = mobility_df, aes(x = share_middleclass, y = e_2c)) +
  geom_point(alpha = 0.5) +
  labs(
    x = "Pct. middle class", y = "OLS residual",
    main = "Visual inspection for heteroskedasticity in 2c."
  ) +
  theme_pander()
```



2f. Conduct a Breusch-Pagan test for heteroskedasticity in the regression model in (2c). Describe your hypotheses, the test statistic, the *p*-value, and your conclusion.

ANSWER

```r
# B-P regression
reg_2f <- lm(e_2c^2 ~ share_middleclass, data = mobility_df)
# B-P test statistic
lm_2f <- summary(reg_2f)$r.squared * 709
# B-P p-value
pchisq(q = lm_2f, df = 1, lower.tail = F)
```

```
## [1] 0.0005441366
```

Hypotheses Our Breusch-Pagan test here tests the hypotheses H$_0$ $\alpha_1 = 0$ vs. H$_a$ $\alpha_1 \neq 0$ for $e_i^2 = \alpha_0 + \alpha_1 x_i + v_i$ (where we are using e_i^2 to estimate u_i^2, which gives us an estimate for σ_i^2.) If we reject H$_0$, then we have evidence of heteroskedasticity.

Test statistic We calculate a B-P test statistic of approximately 11.96.

***p*-value** Under the distribution of a χ_1^2, the implied *p*-value for our LM statistic (the probability of seeing this test statistic or greater) is approximately 0.00054.

Conclusions Because our *p*-value is less than our standard significance of 0.05, we reject the null hypothesis $(\alpha_1 = 0)$—there is statistically significant evidence at the 5% level that $\alpha_1 \neq 0$, meaning there is statistically significant evidence of a relationship between e_i^2 and x_i^2 (the commuting zone's share of middle class residents). Therefore, we have statistically significant evidence of heteroskedasticity.

2g. Conduct a White test for heteroskedasticity in the regression model in (2c). Describe your hypotheses, the test statistic, the *p*-value, and your conclusion.

Hint: To square the variable `x` in `lm()`, we write `lm(y ~ x + I(x^2), data = dataset)`.

ANSWER

Hypotheses Our White test in this question tests the hypotheses H$_0$ $\alpha_1 = \alpha_2 = 0$ vs. H$_a$ $\alpha_1 \neq 0$ *or* $\alpha_2 \neq 0$, where $e_i^2 = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + v_i$ (where, again, we are using e_i^2 to estimate u_i^2, which gives us an estimate for σ_i^2.) If we reject H$_0$, then we have evidence of heteroskedasticity.

Test statistic We calculate a White test statistic of approximately 11.96.

***p*-value** Under the distribution of a χ_2^2, the implied *p*-value for our LM statistic (the probability of seeing this test statistic or greater) is approximately 0.00000094.

Conclusions Because our *p*-value is less than our standard significance of 0.05, we reject the null hypothesis $(\alpha_j = 0)$—there is statistically significant evidence at the 5% level that either $\alpha_1 \neq 0$ or $\alpha_2 \neq 0$. Therefore we find statistically significant evidence of a relationship between e_i^2 and x_i^2 (the commuting zone's share of middle class residents). We have statistically significant evidence of heteroskedasticity.

2h. Let's imagine that we think heteroskedasticity is present. Estimate heteroskedasticity-robust standard errors. Do your standard errors change? What about the coefficients? Why is this the case?

Hint: To do this, use the `felm()` function in the `lfe` package. `felm()` takes a regression formula just like `lm()`. Then use `summary(., robust = T)` to show the heteroskedasticity-robust standard errors.

Example:

```
# The regression
some_reg ← felm(y ~ x, data = fake_data)
# Print the coefficients w/ het-robust standard errors
summary(some_reg, robust = T)
```

ANSWER

```
# Load the 'lfe' package
p_load(lfe)
# Same regression as in (2c)—but with 'felm'
reg_2h ← felm(prob_q5_q1 ~ share_middleclass, data = mobility_df)
# Print the coefficients w/ and w/out het-robust standard errors
reg_2h %>% summary(robust = T)
reg_2h %>% summary(robust = F)
```

```
## Coefficients:
##                    Estimate Robust s.e t value Pr(>|t|)
##  (Intercept)       -0.09714    0.01191  -8.159 1.55e-15 ***
##  share_middleclass  0.35412    0.02226  15.912  < 2e-16 ***

## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
##  (Intercept)       -0.09714    0.01039  -9.349   <2e-16 ***
##  share_middleclass  0.35412    0.01870  18.934   <2e-16 ***
```

The estimated coefficients are the same across the two sets of estimates (with and without heteroskedasticity-robust standard errors), because they both use OLS to estimate the coefficients. The standard errors change because they use different estimators for the standard errors—a heteroskedasticity-robust estimator and an estimator that assumes homoskedasticity. The heteroskedasticity-robust standard errors are slightly larger.

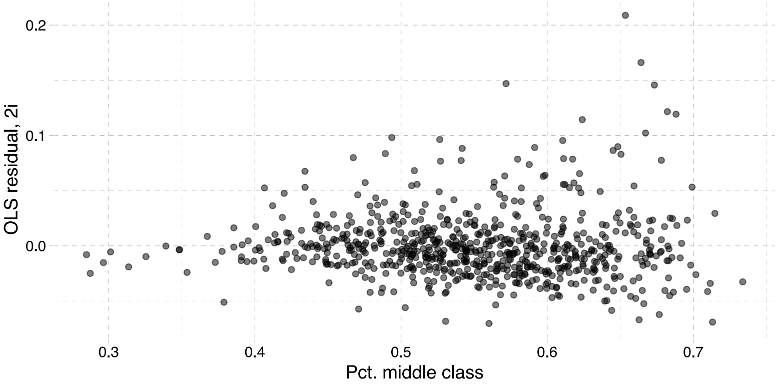
2i. As we discussed in class, we can introduce heteroskedasticity by mis-specifying our regression model. Try adding the additional variables from this dataset into the regression (possibly also adding interactions or squared explanatory variables). Then plot the new residuals against share middleclass (`share_middleclass`). *Briefly* describe which regressions you ran and whether it affected the appearance of heteroskedasticity.

Note: You do not need to formally test for heteroskedasticity.

ANSWER

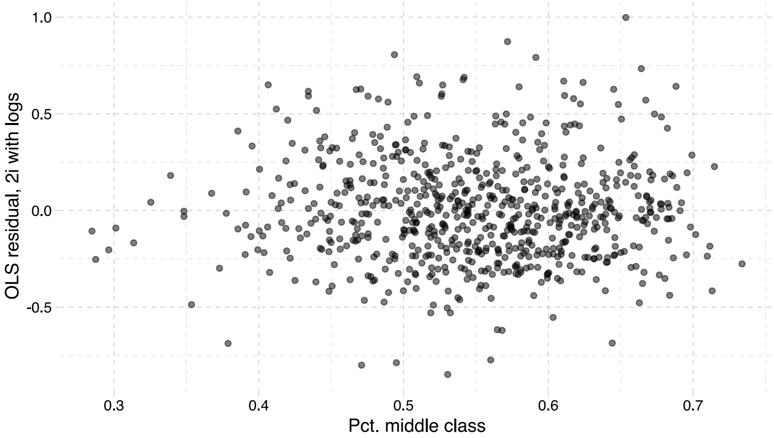
If we stick with the outcome variable as a level, then heteroskedasticity appears likely, even if we include all of the variables in the dataset, their squares, and the two-way interactions.

```
# Regression with all variables, quadratics, and interactions
reg_2i ← lm(
  prob_q5_q1 ~
  i_urban +
  share_black + I(share_black^2) +
  share_middleclass + I(share_middleclass^2) +
  share_divorced + I(share_divorced^2) +
  share_married + I(share_married^2) +
  share_black:share_middleclass + share_black:share_divorced + share_black:share_married +
  share_middleclass:share_divorced + share_middleclass:share_married +
  share_divorced:share_married,
  data = mobility_df
)
# Add residuals to dataset
mobility_df$e_2i ← residuals(reg_2i)
# Plot residuals against share_middleclass
# Plot with ggplot
ggplot(data = mobility_df, aes(x = share_middleclass, y = e_2i)) +
  geom_point(alpha = 0.5) +
  labs(
    x = "Pct. middle class", y = "OLS residual, 2i",
    main = "Visual inspection for heteroskedasticity in 2i."
  ) +
  theme_pander()
```



However, if we take the log of our previous outcome variable, things start to look much more homoskedastic.

```r
# Regression with all variables, quadratics, and interactions
reg_2i_log ← lm(
  log(prob_q5_q1) ~
  i_urban +
  share_black + I(share_black^2) +
  share_middleclass + I(share_middleclass^2) +
  share_divorced + I(share_divorced^2) +
  share_married + I(share_married^2) +
  share_black:share_middleclass + share_black:share_divorced + share_black:share_married +
  share_middleclass:share_divorced + share_middleclass:share_married +
  share_divorced:share_married,
  data = mobility_df
)
# Add residuals to dataset
mobility_df$e_2i_log ← residuals(reg_2i_log)
# Plot residuals against share_middleclass
# Plot with ggplot
ggplot(data = mobility_df, aes(x = share_middleclass, y = e_2i_log)) +
  geom_point(alpha = 0.5) +
  labs(
    x = "Pct. middle class", y = "OLS residual, 2i with logs",
    main = "Visual inspection for heteroskedasticity in 2i."
  ) +
  theme_pander()
```



2j. Should we take the regression in (2c) be *causal*? Explain your answer. If we cannot interpret the regression as causal, can we still learn something interesting here? Explain.

ANSWER

We probably should not apply a causal interpretation to our estimated coefficients in (2c). There are likely many omitted variables that are (1) correlated with *share middleclass* and (2) affect the probability an individual moves from the first fifth to the upper fifth of the income distribution. One example may be school quality within the commuting zone.

Another potential example is the share of the commuting zone that is married. For example, the correlation between share married and share middleclass is 0.53. If share married affects our outcome variable (the probability an individual growing up in the lowest fifth of the income distribution moves into the top fifth), then our estimate on share middleclass will suffer from omitted-variable bias. Specifically, if we think share married positively affects our outcome variable, then our coefficient should be an overestimate of the true effect of *share middleclass*. Let's try including *share married* to see what happens.

```
# The results with only share_middle
reg_2c %>% tidy()
```

```
## # A tibble: 2 x 5
##   term              estimate std.error statistic  p.value
##   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)        -0.0971    0.0104     -9.35 1.14e-19
## 2 share_middleclass   0.354     0.0187     18.9  5.55e-65
```

```
# The results from adding in share_married
lm(prob_q5_q1 ~ share_middleclass + share_married, data = mobility_df) %>% tidy()
```

```
## # A tibble: 3 x 5
##   term              estimate std.error statistic  p.value
##   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)         -0.208    0.0177    -11.7  3.93e-29
## 2 share_middleclass    0.270    0.0212     12.7  1.45e-33
## 3 share_married        0.274    0.0362      7.56 1.27e-13
```

Just as we predicted: By including *share married*, the estimated 'effect' of *share middleclass* decreases considerably.

We might guess that *share black* would also (1) correlate with *share middleclass* and (2) affect our outcome variable. Because the correlation between *share middleclass* and *share black* is negative (correlation of -0.64), and because *share black* may have a downward effect on the probability an individual moves from the lowest to the highest fifth of the income distribution, we would again expect the estimated effect of *share middleclass* to overstate the actual effect due to omitted variable bias. Let's see.

```
## # A tibble: 4 x 5
##   term              estimate std.error statistic  p.value
##   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)        -0.137     0.0219     -6.25 6.95e-10
## 2 share_middleclass   0.204     0.0242      8.45 1.72e-16
## 3 share_married       0.225     0.0368      6.11 1.61e- 9
## 4 share_black        -0.0796    0.0151     -5.28 1.74e- 7
```

Again, we see that the estimated coefficient on share middleclass drops.

While we probably do not want to take these regression coefficients as *causal*, they can certainly still be interesting. The regression in 2c tells us that the probability an individual moves from the lowest fifth to the highest fifth of the income distribution changes substantially around the country—specifically correlating strongly with the share of the commuting zone that is middleclass. Even when we control for other variables, this correlation remains (though it diminishes). This is interesting/important because we (hopefully) want to help children who grow up below the poverty line to move into higher income levels throughout their life.

Consistency

3. Compare/contrast the concepts **expected value** and **probability limit**.

ANSWER The **expected value** describes the mean of a distribution. When we're considering estimators, the **expected value** gives us the mean of the estimator's distribution for a fixed sample size. The **probability limit** also can describe the behavior of an estimator, but it shows what the estimator converges (collapses) to as the sample size gets very big (*i.e.*, infinity).

4. What does it mean if the estimator $\hat{\theta}$ is consistent for θ?

ANSWER If $\hat{\theta}$ is consistent for θ, then the probability limit of $\hat{\theta}$ is θ. In other words, as the sample size approaches infinity, the distribution of $th\hat{e}ta$ collapses to a spike at θ.

5. What is required for an omitted variable to make the OLS estimator $\hat{\beta}_j$ inconsistent for β_j?

ANSWER If we've omitted a variable that (1) correlates with x_j and (2) affects our outcome variable y, the our estimate for the effect of x_j on y (*i.e.*, $\hat{\beta}_j$) will be inconsistent due to omitted-variable bias.

6. Imagine that we are interested in the following model

$$\text{Health}_i = \beta_0 + \beta_1 \text{Money}_i + \beta_2 \text{Happiness}_i + u_i$$

but we are unable to measure an individual's *happiness*.

6a. If we simply omit *happiness* and estimate the equation

$$\text{Health}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Money}_i + e_i$$

in which direction should we expect our estimate for β_1 to be biased? Explain your answer.

ANSWER If we think that $\beta_2 > 0$ (meaning, on average, money increases health) and $\text{Cov}(\text{Money}, \text{Happiness})$ (money and happiness are positively correlated), then our estimate $\hat{\beta}_1$ will overestimate the effect of money on health.

6b. Instead of omitting happiness, we decide to use a proxy for happiness—an individual's self-reported feeling of happiness (on a scale 1–10).

$$\text{Health}_i = \hat{\beta}_0 + \hat{\beta}_1 \text{Money}_i + \hat{\beta}_2 (\text{Survey Happiness})_i + e_i$$

Should we expect our estimate $\hat{\beta}_2$ to over- or under-estimate the true value of β_2. Explain.

ANSWER This *surveyed* happiness measure likely contains measurement error, relative to an individual's true happiness, which will lead to attenuation bias—meaning $\hat{\beta}_2$ will be biased toward zero.

Time series

7. Give an example in which a **static time-series model** might be appropriate. Briefly explain why it is appropriate.

ANSWER If time periods are *academic years*, and our model is

$$(\text{EC421 Grades})_t = \beta_0 + \beta_1(\text{Hours Studied for EC421})_t + u_t$$

In this model, the average grade each year in EC421 likely does not depend upon grades or hours studied in other time periods—only on the amount of studying for EC421 in the same period.[†]

8. Give an example in which a **static time-series model** would not be appropriate. Write down a **dynamic model** that would deal with the shortcomings of the failed static model.

ANSWER Similar to our discussion in class… Let the time period t represent a month.

$$\text{Births}_t = \beta_0 + \beta_1\text{Population}_t + \beta_2\text{Income}_t + u_t$$

This statitic model does not make very much sense—births today probably only weakly depend (or maybe do not depend at all) on population today or income today. Instead, population and income in several lags would be important—as would prior numbers of births—*e.g.*,

$$\begin{aligned}\text{Births}_t = &\beta_0 + \\ &\beta_1\text{Population}_{t-12} + \beta_2\text{Population}_{t-11} + \beta_3\text{Population}_{t-10} + \beta_4\text{Population}_{t-9} + \\ &\beta_5\text{Population}_{t-8} + \beta_6\text{Population}_{t-7} + \beta_7\text{Population}_{t-6} + \\ &\beta_8\text{Income}_{t-12} + \beta_9\text{Income}_{t-11} + \beta_{10}\text{Income}_{t-10} + \beta_{11}\text{Income}_{t-9} + \\ &\beta_{12}\text{Income}_{t-8} + \beta_{13}\text{Income}_{t-7} + \beta_{14}\text{Income}_{t-6} + \\ &\beta_{15}\text{Births}_{t-12} + \beta_{16}\text{Births}_{t-11} + \beta_{17}\text{Births}_{t-10} + \beta_{18}\text{Births}_{t-9} + \\ &\beta_{19}\text{Births}_{t-8} + \beta_{20}\text{Births}_{t-7} + \beta_{21}\text{Births}_{t-6} + u_t\end{aligned}$$

9. Why are dynamic models with lagged dependent variables biased with OLS? Which of our assumptions do they violate?

ANSWER Dynamic models with lagged dependent variables create a situation where our disturbances in one time period (*e.g.*, t) are correlated with an explanatory variable in another period (*e.g.*, $t+1$), which violates our exogeneity assumption. To see this point, write out a simple dynamic model for two consecutive time periods

$$\begin{aligned}y_t &= \beta_0 + \beta_1 y_{t-1} + u_t \\ y_{t+1} &= \beta_0 + \beta_1 y_t + u_{t+1}\end{aligned}$$

Now notice that u_t, a **disturbance**, correlates with y_t in the top equation, and y_t is an **explanatory variable** in the second equation. Thus, we have a correlation between our disurbance and an explanatory variable—violating our assumption of exogeneity.

[†] Admittedly, this static model may still be wrong if EC421 grades in t depend on the hours studied for EC320 in another year.