ECOM20001
Econometrics 1

Lecture Note 6
Multiple Linear Regression - Testing

A/Prof David Byrne
Department of Economics
University of Melbourne

Stock and Watson: Chapter 7

Summary of Key Concepts

- Standard errors, hypothesis testing, confidence intervals with multiple linear regression
- Testing joint hypotheses
- F-statistics
- Overall regression F-statistic
- Homoskedasticity-only F-statistic
- Testing single restrictions involving multiple coefficients
- Confidence sets for multiple coefficients
- Model specification, omitted variable bias, coefficients of interest, control variables
- Conditional mean independence
- Choosing coefficients of interest and control variables
- Applications: policy, testing theory, exploration

Standard Errors with Multiple Linear Regression

- Recall from single linear regression, we had the following estimate of the sample variance of the OLS estimator $\hat{\beta}_1$:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$$

- Further recall that if n is large, we can approximate the population variance of β_1, $\sigma^2_{\beta_1}$ using the sample variance, $\hat{\sigma}^2_{\hat{\beta}_1}$
- We then computed the standard error of $\hat{\beta}_1$ from a single linear regression model as:

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}$$

Standard Errors with Multiple Linear Regression

- With multiple regression, all of these ideas carry over when we compute $SE(\hat{\beta}_j)$ for regressors $j = 0, 1, 2, \ldots, k$ no matter how many regressors are in the model (recall that k is the number of regressors)
- R produces the $SE(\hat{\beta}_j)$'s for you; the underlying mathematics, which requires the use of matrices, are investigated in ECOM 30002: Econometrics 2 and beyond
- Due to the LLN and the CLT, the marginal distribution of $\hat{\beta}_j$ is $N(\beta_j, \sigma^2_{\hat{\beta}_j})$, so we can compute the t-statistic for a hypothesised null value of $\beta_{j,0}$ with our familiar formula:

$$t^{act} = \frac{\hat{\beta}_j - \beta_{j,0}}{SE(\hat{\beta}_j)}$$

for regression coefficients $j = 0, 1, 2, \ldots, k$

Hypothesis Tests with a Single Coefficient

- Hypothesis testing for regression coefficient j follows a similar set of steps
- First, we state the null and alternative hypothesis, where we again primarily focus on 2-sided alternatives:

$$H_0 : \beta_j = \beta_{j,0} \quad \text{vs.} \quad H_1 : \beta_j \neq \beta_{j,0}$$

- By far, the most common test in practice has a hypothesised value of $\beta_{j,0} = 0$.
  - This is the null that no relationship exists between X_{ji} and Y_i versus the alternative that a relationship exists, holding other regressors constant

3 Steps for Testing Hypotheses About β_j

1. Compute the OLS estimate $\hat{\beta}_j$ and its standard error, $SE(\hat{\beta}_j)$ (using a statistics program like R)
2. Compute the t-statistic:

$$t^{act} = \frac{\hat{\beta}_j - \beta_{j,0}}{SE(\hat{\beta}_j)}$$

3a. Compute the *p*-value

$$p\text{-value} = 2\Phi(-|t^{act}|)$$

where recall Φ is the cumulative density function of the normal distribution.

3 Steps for Testing Hypotheses About β_j

(3a continued) Letting α be the level of significance of the test, we reject the null $H_0 : \beta_j = \beta_{j,0}$ if

$$p\text{-value} < \alpha$$

where typical values of α are 0.10, 0.5, 0.01

3b. Rather than using p-values to conduct the hypothesis test, we can equivalently using our t-statistic and critical values from the normal distribution to conduct the hypothesis test.

We reject the null $H_0 : \beta_1 = \beta_{j,0}$ in favour of $H_0 : \beta_j \neq \beta_{1,0}$ if

$$|t^{act}| > t^{\alpha}_{crit}$$

where $t^{\alpha}_{crit} = 1.65$ if $\alpha = 0.10$, $t^{\alpha}_{crit} = 1.96$ if $\alpha = 0.05$, $t^{\alpha}_{crit} = 2.58$ if $\alpha = 0.01$

Confidence Intervals

- Confidence intervals (with $1-\alpha$ level of confidence) with multiple linear regression follow exactly as before from inverting t-statistics and asking what null values $\beta_{j,0}$ would not be rejected at the α level of significance
- 90% confidence interval for β_j:

$$[\hat{\beta}_j - 1.65SE(\hat{\beta}_j), \hat{\beta}_j + 1.65SE(\hat{\beta}_j)]$$

- 95% confidence interval for β_j:

$$[\hat{\beta}_j - 1.96SE(\hat{\beta}_j), \hat{\beta}_j + 1.96SE(\hat{\beta}_j)]$$

- 99% confidence interval for β_j:

$$[\hat{\beta}_j - 2.58SE(\hat{\beta}_j), \hat{\beta}_j + 2.58SE(\hat{\beta}_j)]$$

Application to Test Score Data

- Using the test scores data, we run a multiple linear regression of test scores on class size and income:

$$\widehat{TestScore}_i = \underset{(9.20)}{60.41} - \underset{(0.31)}{5.47}\, ClassSize_i + \underset{(0.23)}{1.24}\, Income_i,\ SER = 9.20,\ \bar{R}^2 = 0.88$$

- We could further enrich the model and include, for example, a dummy variable $Urban_i = 1$ is a school is in an urban market with more than 100,000 people, and is equal to 0 otherwise

$$\widehat{TestScore}_i = \underset{(8.92)}{56.06} - \underset{(1.74)}{1.22}\, ClassSize_i + \underset{(0.22)}{1.30}\, Income_i + \underset{(2.08)}{5.21}\, Urban_i,$$
$$SER = 8.91,\ \bar{R}^2 = 0.89$$

Note: using slightly different dataset than in previous lectures, $n = 89$ schools

Application to Test Score Data

- 95% CI for the $ClassSize_i$ coefficient in the first regression with $Income_i$ as an additional regressor is:

$$[-5.47 - 1.96 \times 0.31, -5.47 + 1.96 \times 0.31] = [-6.08, -4.86]$$

- 95% CI for the $ClassSize_i$ coefficient in the second regression with $Income_i$ and $Urban_i$ as additional regressors is:

$$[-1.22 - 1.96 \times 1.74, -1.22 + 1.96 \times 1.74] = [-4.49, 2.05]$$

- Notice the substantial impact that including $Urban_i$ in the regression has on the coefficient estimate and confidence interval for the $ClassSize_i$ coefficient
  - Severe omitted variable bias on the $ClassSize_i$ coefficient from not controlling for $Urban_i$
  - Indicates that the $ClassSize_i$ coefficient in the first regression is due to urban students doing better on tests and having smaller class sizes

Tests of Joint Hypotheses

- Angry politician says:

  Fake news! Neither class size nor income affects student performance!

- This is an example of a joint hypothesis test involving two or more coefficients from our regression

$$TestScore_i = \beta_0 + \beta_1 ClassSize_i + \beta_2 Income_i + u_i, \quad i = 1, \ldots, n$$

- We can formally write down the joint test as:

$$H_0 : \beta_1 = 0 \text{ and } \beta_2 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0$$

- The test imposes two restrictions on the multiple regression model under the null: $\beta_1 = 0$ and $\beta_2 = 0$

Tests of Joint Hypotheses

- Generally, joint hypotheses can involve two or more restrictions on regression coefficients
- The null and alternative can be generally written as follows:
  - $H_0 : \beta_j = \beta_{j,0},\ \beta_m = \beta_{m,0}, \ldots$ for a total of q restrictions
  - H_1 : one or more of the q restrictions under H_0 does not hold

  where $\beta_j, \beta_m, \ldots$ refer to different regression coefficients, and $\beta_{j,0}, \beta_{m,0}, \ldots$ refer to the values of these coefficients under H_0
- In words, the alternative is that at least one of the equalities under H_0 does not hold
- Each respective null value $\beta_{j,0}, \beta_{m,0}, \ldots$ can be 0, a positive number, or a negative number, and we can test many such restrictions jointly (e.g., large q)

Tests of Joint Hypotheses

- Why can't we just test one hypothesis at a time?
- For instance, from the test score example

$$TestScore_i = \beta_0 + \beta_1 ClassSize_i + \beta_2 Income_i + u_i, \quad i = 1, \ldots, n$$

why not just test $\beta_1 \neq 0$ and then test $\beta_2 \neq 0$, where we:

1. reject the "first" null if $|t_1^{act}| > 1.96$, and then
2. reject the "second" null if $|t_2^{act}| > 1.96$

- This (incorrect) approach to joint testing relies on the fact that the marginal distributions of $\hat{\beta}_1$ and $\hat{\beta}_2$ are $N(\beta_1, \sigma^2_{\hat{\beta}_1})$ and $N(\beta_2, \sigma^2_{\hat{\beta}_2})$, respectively
- However, the problem with joint testing in this way is it fails to recognize that $\hat{\beta}_1$ and $\hat{\beta}_2$ are jointly distributed random variables, where in general $cov(\hat{\beta}_1, \hat{\beta}_2) \neq 0$
- Since the H_0 is based on joint values of β_1 and β_2, we need to construct test statistics based on their joint distribution, not their marginal distributions

F-Statistic

- The F-Statistic is used to test a joint hypothesis about regression coefficients
- With $q = 2$ restrictions of $\beta_1 = 0$ and $\beta_2 = 0$, the (complicated) F-Statistic formula is given by:

$$F = \frac{1}{2}\left(\frac{(t_1^{act})^2 + (t_2^{act})^2 - 2\hat{\rho}_{t_1^{act},t_2^{act}} t_1^{act} t_2^{act}}{1 - \hat{\rho}_{t_1^{act},t_2^{act}}}\right)$$

where $\hat{\rho}_{t_1^{act},t_2^{act}}$ is the correlation between t_1^{act} and t_2^{act}

  - If $\hat{\rho}_{t_1^{act},t_2^{act}} = 0 \rightarrow F$ is simply the average of the 2 squared t-statistics. Larger values of either t^{act} means we're more likely to reject the null.
  - If $\hat{\rho}_{t_1^{act},t_2^{act}} \neq 0 \rightarrow F$ formula adjusts to account for correlation in the t-statistics, but the same intuition holds

F-Statistic Distribution

- The general formula for the heteroskedasticity-robust F-Statistic testing q restrictions of the joint null hypothesis:
  - $H_0 : \beta_j = \beta_{j,0},\ \beta_m = \beta_{m,0}, \ldots$ for a total of q restrictions
  - H_1 : one or more of the q restrictions under H_0 does not hold

  is computed directly using statistical programs like R
- In large samples, under H_0, the following key result holds:

$$\text{the } F\text{-statistic is distributed } F_{q,\infty}$$

  This is notation the textbook uses for the distribution of the F-statistic in theory
- In practice, with $n < \infty$, the following key result holds:

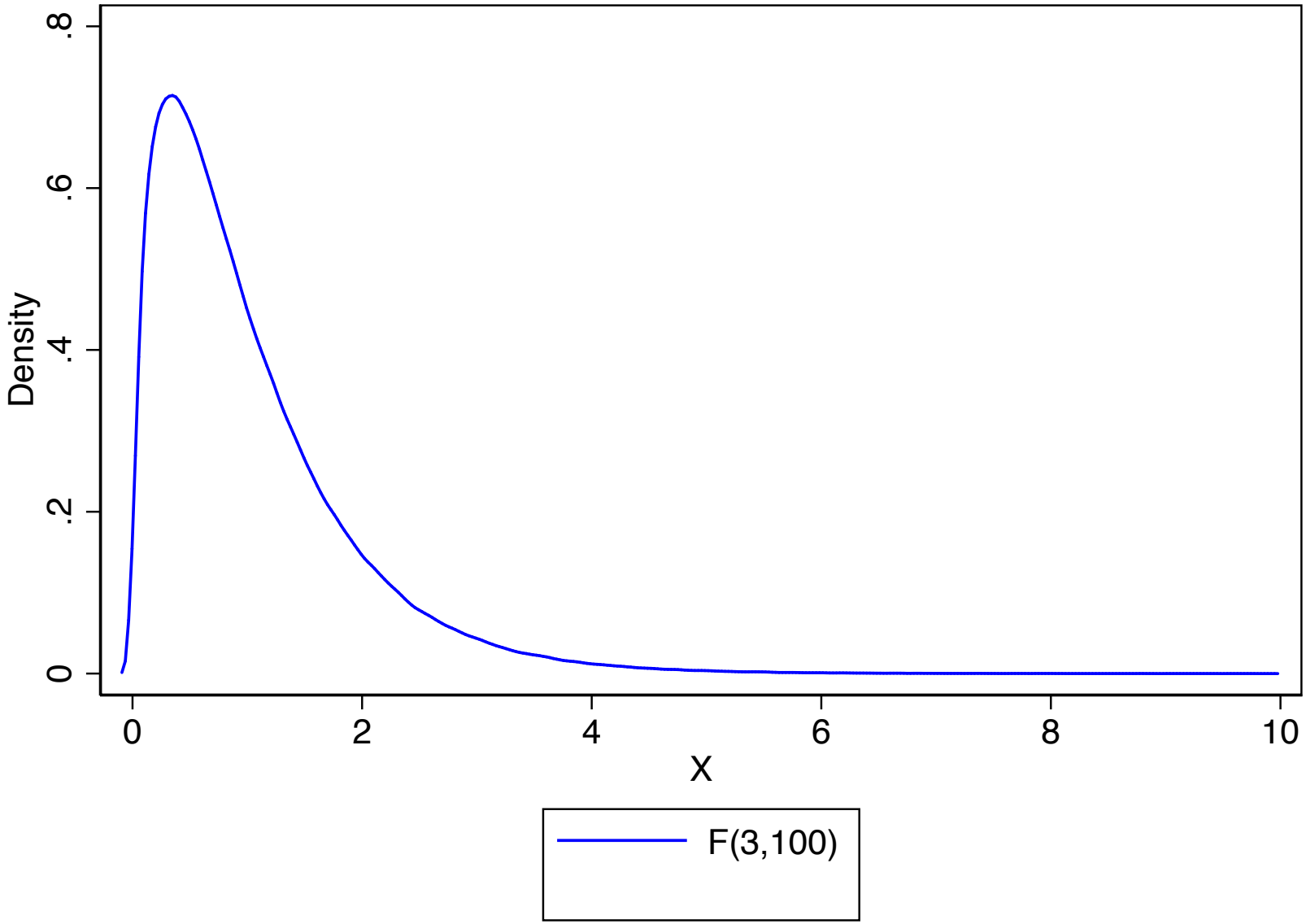
$$\text{the } F\text{-statistic is distributed } F_{q,n-k-1}$$

  where n is sample size and k is the number of regressors (not including the constant)

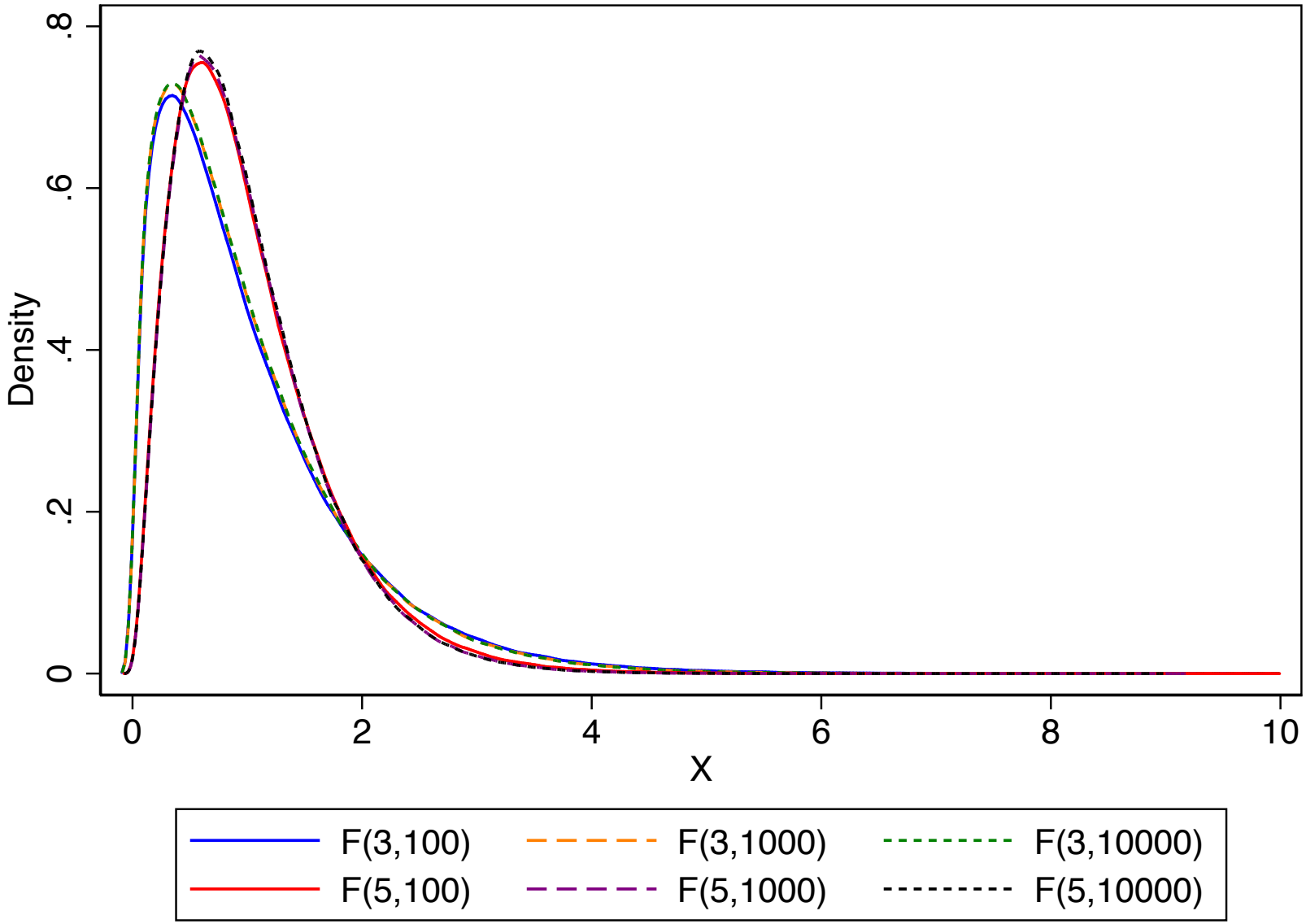
F-Statistic Distribution

- Recall way back from Lecture 2 on probability that F distributions are denoted $F_{m,n}$ where m and n are the degrees of freedom of the distribution (numerator and denominator, respectively)
  - that is, just like a Normal Distribution has two parameters for its mean μ and variance σ^2, the F distribution has two parameters m and n which define its shape
- R reports F-statistics with q and $n - k - 1$ degrees of freedom (so distributed $F_{q,n-k-1}$) whenever a joint hypothesis test is performed using a regression model with k regressors, estimated using a sample with n observations, and where q restrictions are imposed under H_0
- The mathematics in computing F-statistics are explored in graduate school econometrics subjects

F-Statistic distribution graphically



F-Statistic distribution graphically



F-Statistic p-value

- The p-value of the F-statistic can be computed using its large-sample approximation of $F_{q,n-k-1}$
- If F^{act} is the value of F-statistic we computed for a joint hypothesis test, the p-value under the null hypothesis is:

$$p\text{-value} = \Pr[F_{q,n-k-1} > F^{act}]$$

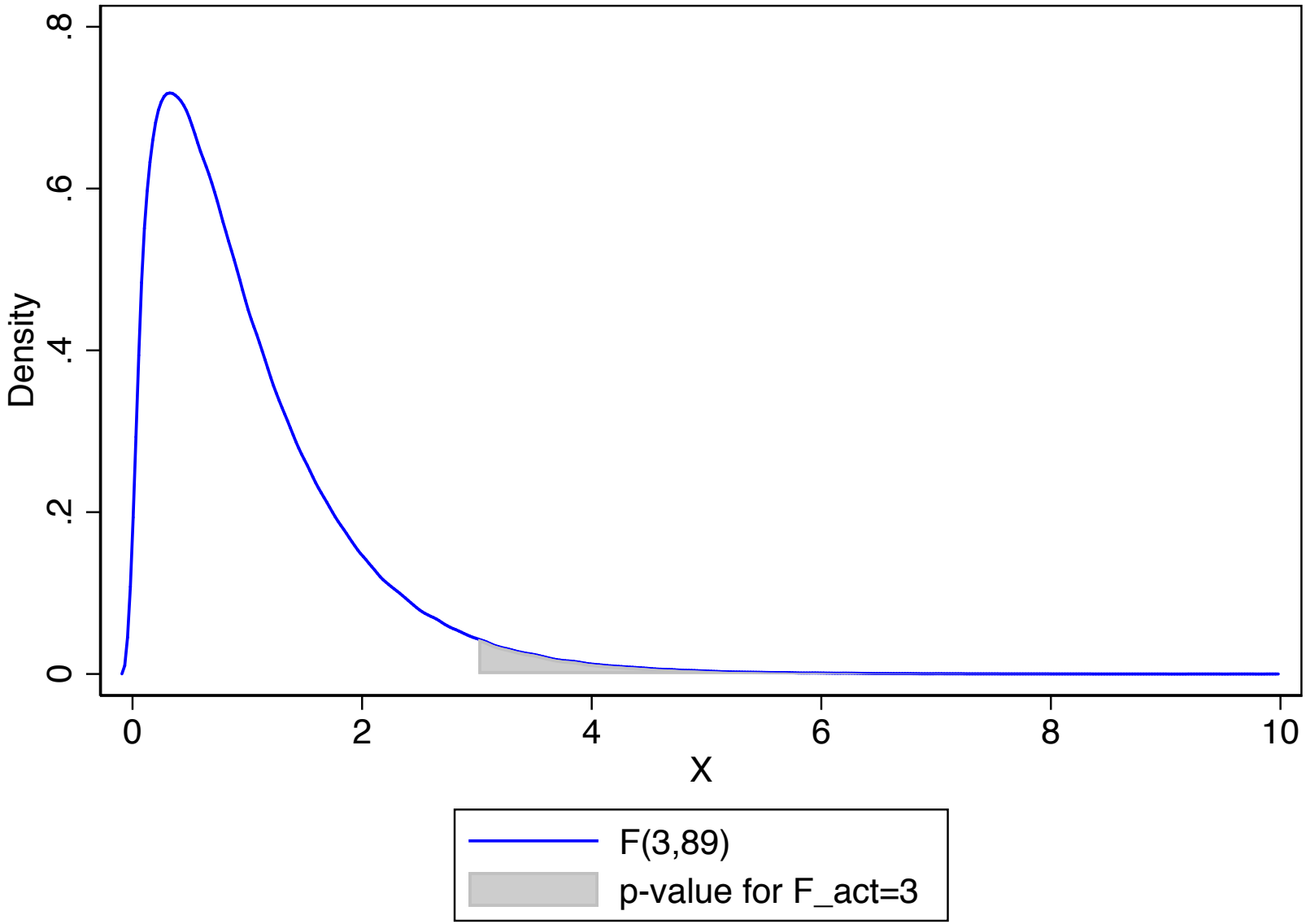
- Let $G(x; q, n-k-1)$ be the cumulative density function of the $F_{q,n-k-1}$ distribution evaluated at x. The p-value for the F-statistic for a null that imposes q restrictions is computed:

$$p\text{-value} = 1 - G(F^{act}; q, n-k-1)$$

- In words, the p-value is the area under the density of the $F_{q,n-k-1}$ distribution to the right of the F^{act} test statistic that we computed

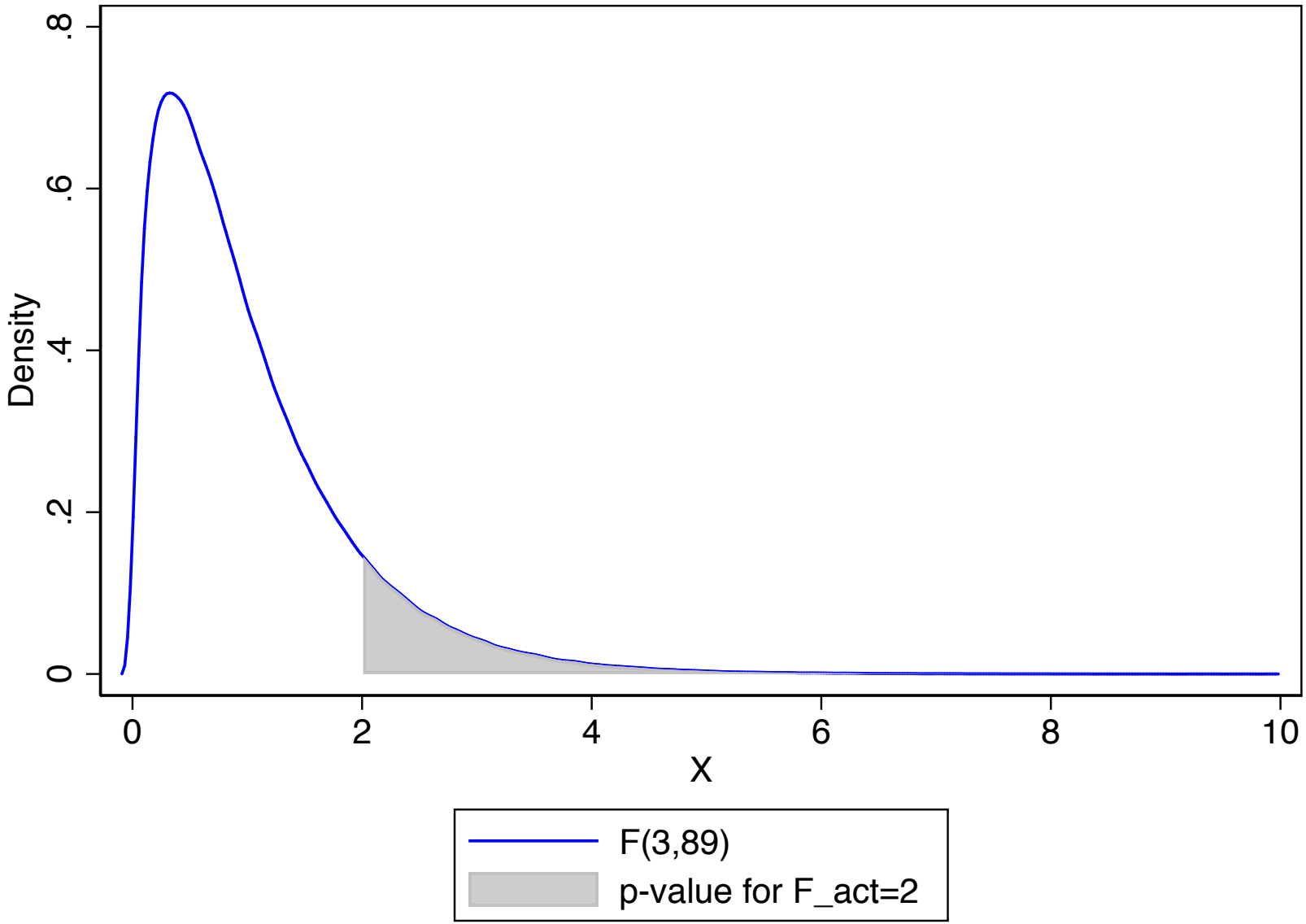
F-Statistic p-value graphically

$q = 3, k = 10, n = 100, n - k - 1 = 89, F^{act} = 3$



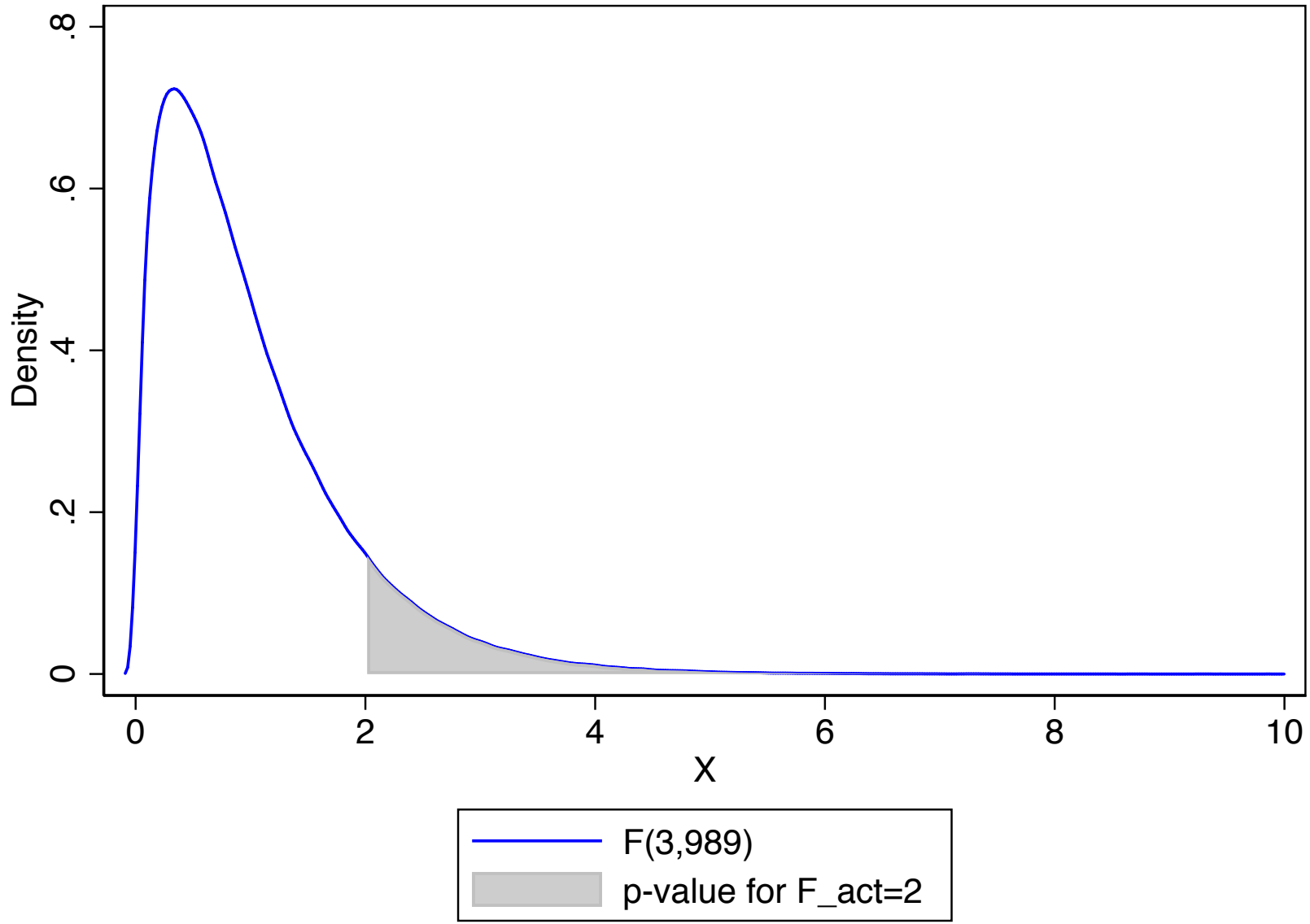
F-Statistic p-value graphically

$q = 3, k = 10, n = 100, n - k - 1 = 89, F^{act} = 2$



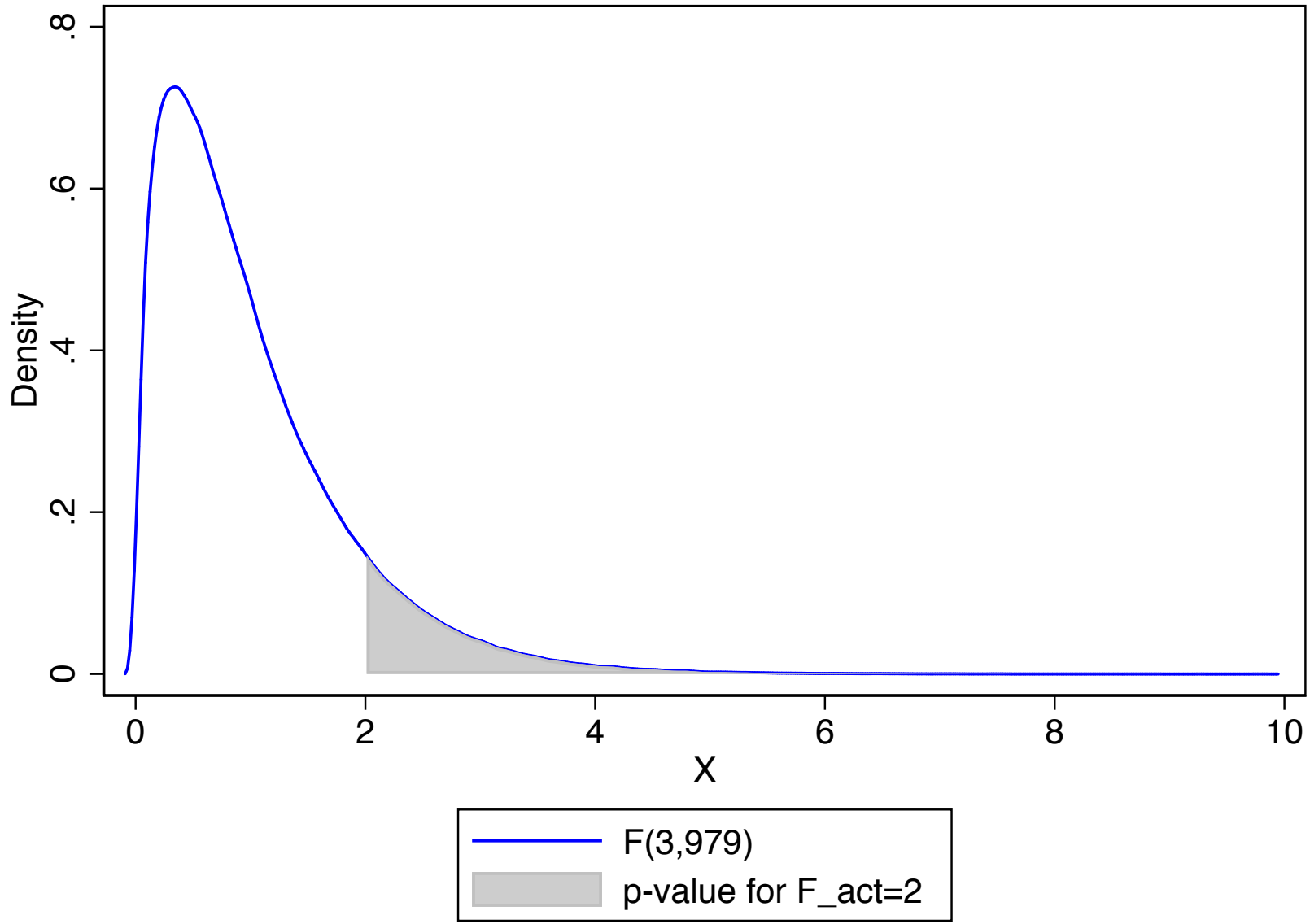
F-Statistic p-value graphically

$q = 3, k = 10, n = 1000, n - k - 1 = 989, F^{act} = 2$



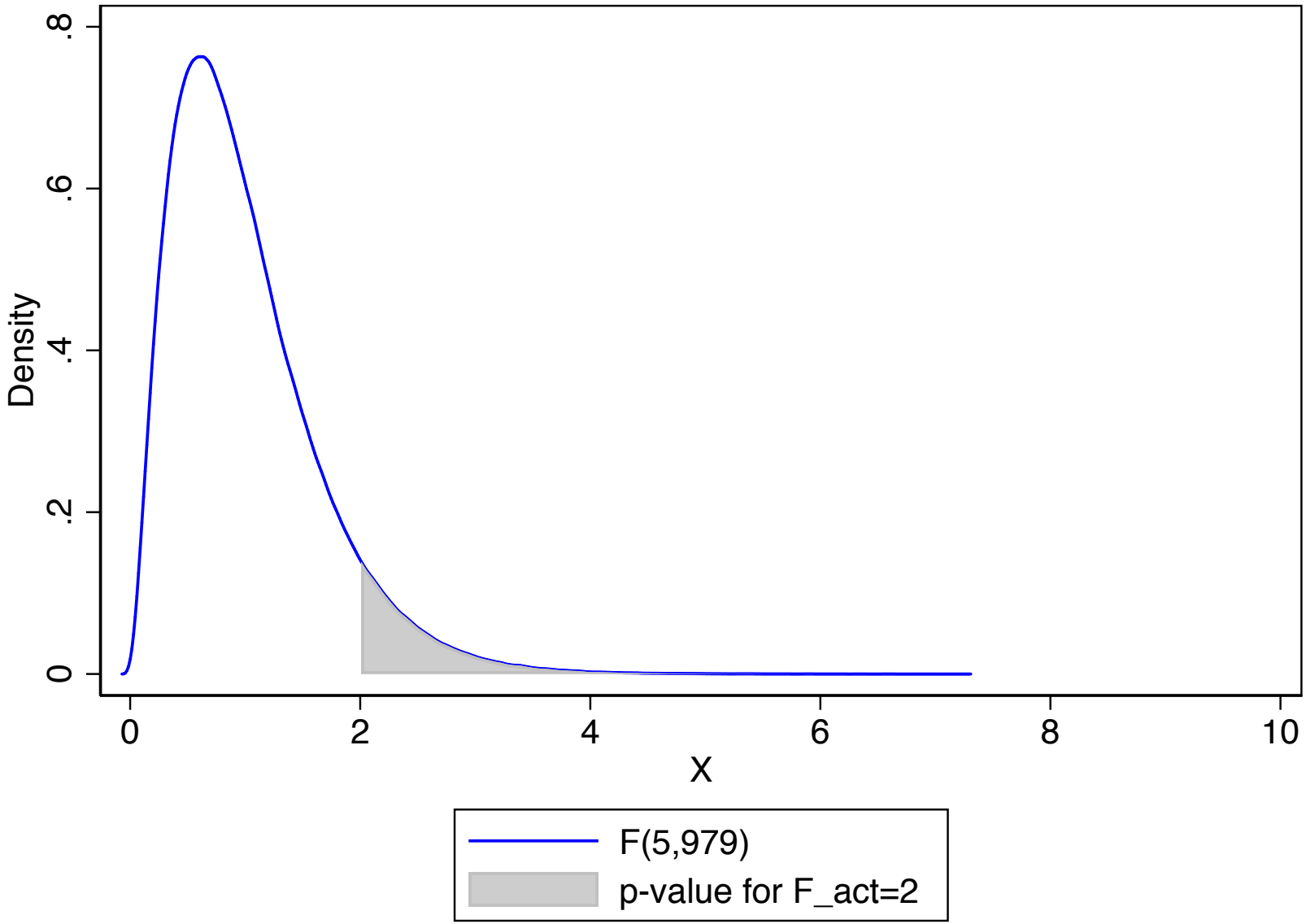
F-Statistic p-value graphically

$q = 3, k = 20, n = 1000, n - k - 1 = 979, F^{act} = 2$



F-Statistic p-value graphically

$q = 5, k = 20, n = 1000, n - k - 1 = 979, F^{act} = 2$



Hypotheses testing with F-Statistics

- As with individual coefficient hypothesis testing, we reject the joint null hypothesis if p-value $< \alpha$ where α is the level of significance.
- Equivalently, we can reject the null if F^{act} is bigger than the critical value F^{α} from the $F_{q,n-k-1}$ distribution, where F^{α} is defined as:

$$\Pr[F_{q,n-k-1} > F^{\alpha}] = 1 - G(F^{\alpha}; q, n-k-1) = \alpha$$

where $G(x; q, n-k-1)$ is the cumulative density function of a F-distribution with q and $n-k-1$ degrees of freedom evaluated at x

Hypotheses testing with F-Statistics

- It's easier to use p-values and not F^{α} critical values for evaluating joint hypothesis tests using the F-statistic.
  - the critical value $F^{\alpha}_{q,n-k-1}$ depends on the sample size (n), number of regressors (k), and number of restrictions (q) and thus is not the same for any given sample, regression and joint hypothesis test
- For these reasons, statistics programs like R report p-values with joint hypothesis tests based on the F-statistic

Overall Regression F-Statistic with $q = k$ Restrictions

- The overall regression F-statistic tests the joint hypothesis that all of the slope coefficients in the regression are 0
- Formally, the null and alternative hypotheses are:
  - $H_0 : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_k = 0$
  - $H_1 : \beta_j \neq 0$ for at least one j, $j = 1, \ldots, k$
- In words: under the null, none of the regressors explains any of the variation in Y_i except for β_0 (which is the mean of Y_i under the null)
- If we fail to reject H_0, it is typically interpreted as rejecting the entire regression model. In other words, it is statistically useless for explaining Y_i
- The overall regression F-statistic has a $F_{k,n-k-1}$ distribution under the null, and we reject the null if its corresponding p-value is less than the chosen level of significance

F-statistic with $q = 1$ Restriction

- When there is only $q = 1$ restriction, the F-statistic tests a single restriction
- The joint null hypothesis is simply the null hypothesis on a single regression coefficient
- In this case, the F-statistic is equal to the square of the t-statistic for the corresponding single restriction hypothesis test, and it is distributed $F_{1,n-k-1}$ under the null

Application of F-statistics to Test Score Data

- Regressions using test score data, class size, and income

$$\widehat{TestScore}_i = \underset{(9.20)}{60.41} - \underset{(0.31)}{5.47}\, ClassSize_i + \underset{(0.23)}{1.24}\, Income_i,\ SER = 9.20,\ \bar{R}^2 = 0.88$$

- Overall regression F-statistic, that tests
  - $H_0 : \beta_1 = 0, \beta_2 = 0$; $H_1 : \beta_1 \neq 0$ and/or $\beta_2 \neq 0$
  - In words: The null hypothesis is that "the entire model is statistically useful"
  - $n - k - 1 = 90 - 2 - 1 = 87$; $q = 2$
  - $F^{act}_{2,87} = 333.18$ with p-value < 0.00001
  - Strongly reject the null, which implies the regression is useful, statistically, for explaining variation in $TestScore_i$

Application of F-statistics to Test Score Data

- Regressions using test score data, class size, and income

$$\widehat{TestScore}_i = \underset{(9.20)}{60.41} - \underset{(0.31)}{5.47}\, ClassSize_i + \underset{(0.23)}{1.24}\, Income_i,\ SER = 9.20,\ \bar{R}^2 = 0.88$$

- Another example: joint test that $\beta_1 = -1$ and $\beta_2 = 1$
  - $H_0 : \beta_1 = -1, \beta_2 = 1$; $H_1 : \beta_1 \neq -1$ and/or $\beta_2 \neq 1$
  - In words: The null hypothesis is that "a 1-person increase in class size leads to a 1-unit decrease in test scores, and a \$1,000-increase in income leads to a 1-unit increase in test scores"
  - $n - k - 1 = 90 - 2 - 1 = 87$; $q = 2$
  - $F^{act}_{2,87} = 166.63$ with p-value < 0.00001
  - Strongly reject the null, meaning that either a 1-person increase in class size does not lead to a 1-unit decrease in test score, or a \$1,000-increase in income does not lead to a 1-unit increase in test scores (or both!)

Homoskedasticity-Only F-Statistic

- We can also interpret hypothesis joint tests that use F-Statistic as addressing the following question:

  Does relaxing the q restrictions that constitute the null hypothesis improve the fit of the regression by enough that this improvement is unlikely to be the result merely of random sampling variation if the null hypothesis is true?

- More simply: if I relax my restrictions, does the model fit improve more than by chance?
  - Intuitive links: large F-Statistic $\rightarrow$ large increases in the R^2 from relaxing the q restrictions

Homoskedasticity-Only F-Statistic

- If u_i is homoskedastic, there is in fact a formula directly connecting the F-Statistic to the R^2 from regressions that impose and do not impose the null hypothesis
- Studying this formula is instructive for understanding what joint hypothesis tests based on the F-Statistic do, even though we do not typically use the formula in practice because we work under the assumption of heteroskedasticity

Homoskedasticity-Only F-Statistic

- We compute the homoskedasticity-only F-Statistic using two regressions

1. Restricted regression: run the regression imposing: $H_0 : \beta_j = \beta_{j,0},\ \beta_m = \beta_{m,0}, \ldots$ for a total of q restrictions
   - compute the SSR from this regression, $SSR_{restricted}$
2. Unrestricted regression: run the regression not imposing any of the constraints in H_0
   - compute the SSR from this regression, $SSR_{unrestricted}$

- If the SSR is sufficiently smaller under the unrestricted regression than the restricted regression, the joint test of the null H_0 is rejected

Homoskedasticity-Only F-Statistic

- The homoskedasticity-only F-statistic of the unrestricted regression is:

$$F^{act} = \frac{(SSR_{restricted} - SSR_{unrestricted})/q}{SSR_{unrestricted}/(n-k-1)}$$

  which is distributed $F_{q,n-k-1}$ where k is the number of regressors (excluding the constant regressor) in the unrestricted regression

- An alternative, equivalent, formula is given by:

$$F^{act} = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n-k-1)}$$

Homoskedasticity-Only F-Statistic

$$F^{act} = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k - 1)}$$

- Intuition: if there is a big drop in the R^2 when the q restrictions are jointly imposed, then it is unlikely that we would obtain as large a drop in the model fit as we found from relaxing the restrictions if the null were in fact true
- Similar intuition holds with the heteroskedasticity-robust F-statistic for testing joint restrictions, but it does not have as simple a formula

Application of Homoskedasticity-Only F-Statistic

- Regression model including full set of regressors (unrestricted):

$$\widehat{TestScore}_i = \underset{(8.92)}{56.06} - \underset{(1.74)}{1.22}\, ClassSize_i + \underset{(0.22)}{1.30}\, Income_i + \underset{(2.08)}{5.21}\, Urban_i,$$

$$SER = 8.91,\ R^2 = 0.89,\ \bar{R}^2 = 0.89$$

- Example: restricted model imposing $\beta_2 = 0$, $\beta_3 = 0$:

$$\widehat{TestScore}_i = \underset{(3.18)}{108.41} - \underset{(0.29)}{6.45}\, ClassSize_i,$$

$$SER = 10.62,\ R^2 = 0.85,\ \bar{R}^2 = 0.84$$

- Homoskedasticity-Only F-Statistic is:

$$F^{act} = \frac{(0.89 - 0.85)/2}{(1 - 0.89)/(90 - 3 - 1)} = 15.64$$

which has a p-value of $\Pr[F_{2,87} > 15.64] < 0.001$

Testing Single Restrictions Involving Multiple Coefficients

- We can also conduct a hypothesis test involving linear relations between multiple coefficients
- A common example is tests of the form $\beta_1 = \beta_2$, which is equivalent to a test that imposes $\beta_1 - \beta_2 = 0$
- Formally stating the null and alternative hypotheses:

$$H_0 : \beta_1 = \beta_2 \quad \text{vs.} \quad H_1 : \beta_1 \neq \beta_2$$

Testing Single Restrictions Involving Multiple Coefficients

$$H_0 : \beta_1 = \beta_2 \quad \text{vs.} \quad H_1 : \beta_1 \neq \beta_2$$

- Here, there is a single restriction so $q = 1$, but the restriction involves multiple coefficients (e.g., β_1 and β_2)
- Notice this is different that the tests we have been considering so far involving $q > 1$ restrictions, where we restrict regression coefficients to be equal to particular numbers
  - Those tests did not impose restrictions that coefficients are a linear function of each other, for example $\beta_1 = \beta_2$

Testing Single Restrictions Involving Multiple Coefficients

- Statistical packages like R typically have commands designed to test restrictions involving multiple coefficients
- The result is an F-statistic because $q = 1$ with a $F_{1,n-k-1}$ distribution under the null
- Useful result: the square of a $N(0,1)$ random variable is distributed $F_{1,n-k-1}$
- So the 95% percentile of the $F_{1,n-k-1}$ distribution is $1.96^2 = 3.84$, which approximately represents the 5% critical value for conducting tests involving a single restriction with multiple coefficients if n is large

Testing Single Restrictions Involving Multiple Coefficients

- You can also transform the regression of interest if you cannot test a restriction directly using a statistics package
- Suppose you want to conduct the following test:

$$H_0 : \beta_1 = \beta_2 \quad \text{vs.} \quad H_1 : \beta_1 \neq \beta_2$$

with the regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- Transform the regression by substracting and adding $\beta_2 X_{1i}$:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_2 X_{1i} - \beta_2 X_{1i} + u_i$$

which equals

$$Y_i = \beta_0 + \underbrace{(\beta_1 - \beta_2)}_{\gamma_1} X_{1i} + \beta_2 \underbrace{(X_{1i} + X_{2i})}_{W_i} + u_i$$

Testing Single Restrictions Involving Multiple Coefficients

- Now redefine the regression to be

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i$$

  where $W_i = X_{1i} + X_{2i}$
- We can then estimate this transformed regression, and test the null that $\gamma_1 = 0$ against the alternative $\gamma_1 \neq 0$ using a t-statistic for the single-coefficient test
- Notice how by transforming the regression, we have transformed the problem from testing a restriction involving two coefficients into a restriction on a single coefficient
- The 95% CI for the difference in coefficients $\gamma_1 = \beta_1 - \beta_2$ is:

$$[\hat{\gamma} - 1.96SE(\hat{\gamma}_1), \hat{\gamma} + 1.96SE(\hat{\gamma}_1)]$$

Application of Tests Involving Multiple Coefficients

- Recall from the start of the lecture note that we obtain the following regression results:

$$\widehat{TestScore}_i = \underset{(8.92)}{56.06} - \underset{(1.74)}{1.22}\, ClassSize_i + \underset{(0.22)}{1.30}\, Income_i + \underset{(2.08)}{5.21}\, Urban_i$$

  where recall we are estimating with a sample of $n = 89$ observations.
- Suppose we wanted to test the null that $\beta_1 = \beta_2$
- Using a joint hypothesis test, we obtain $F^{act} = 2.38$ which is distributed $F_{1,85}$ and has a p-value=0.14
- So we fail to reject the null that $\beta_1 = \beta_2$ at the $\alpha = 0.05$ level

Application of Tests Involving Multiple Coefficients

- Consider the following transformed regression model:

$$\begin{aligned} TestScore_i &= \beta_0 + \beta_1 ClassSize_i + \beta_2 Income_i + \beta_3 Urban_i + u_i \\ &= \beta_0 + \beta_1 ClassSize_i + \beta_2 Income_i + \beta_3 Urban_i + u_i + \beta_2 ClassSize_i - \beta_2 ClassSize_i \\ &= \beta_0 + (\beta_1 - \beta_2) ClassSize_i + \beta_2 (Income_i + ClassSize_i) + \beta_3 Urban_i + u_i \\ &= \beta_0 + \gamma ClassSize_i + \beta_2 W_i + \beta_3 Urban_i + u_i \end{aligned}$$

where $\gamma = (\beta_1 - \beta_2)$ and $W_i = Income_i + ClassSize_i$

- Running the transformed regression we obtain:

$$\widehat{TestScore}_i = \underset{(8.92)}{56.06} - \underset{(1.69)}{2.52}\, ClassSize_i + \underset{(0.22)}{1.30}\, W_i + \underset{(2.08)}{5.21}\, Urban_i$$

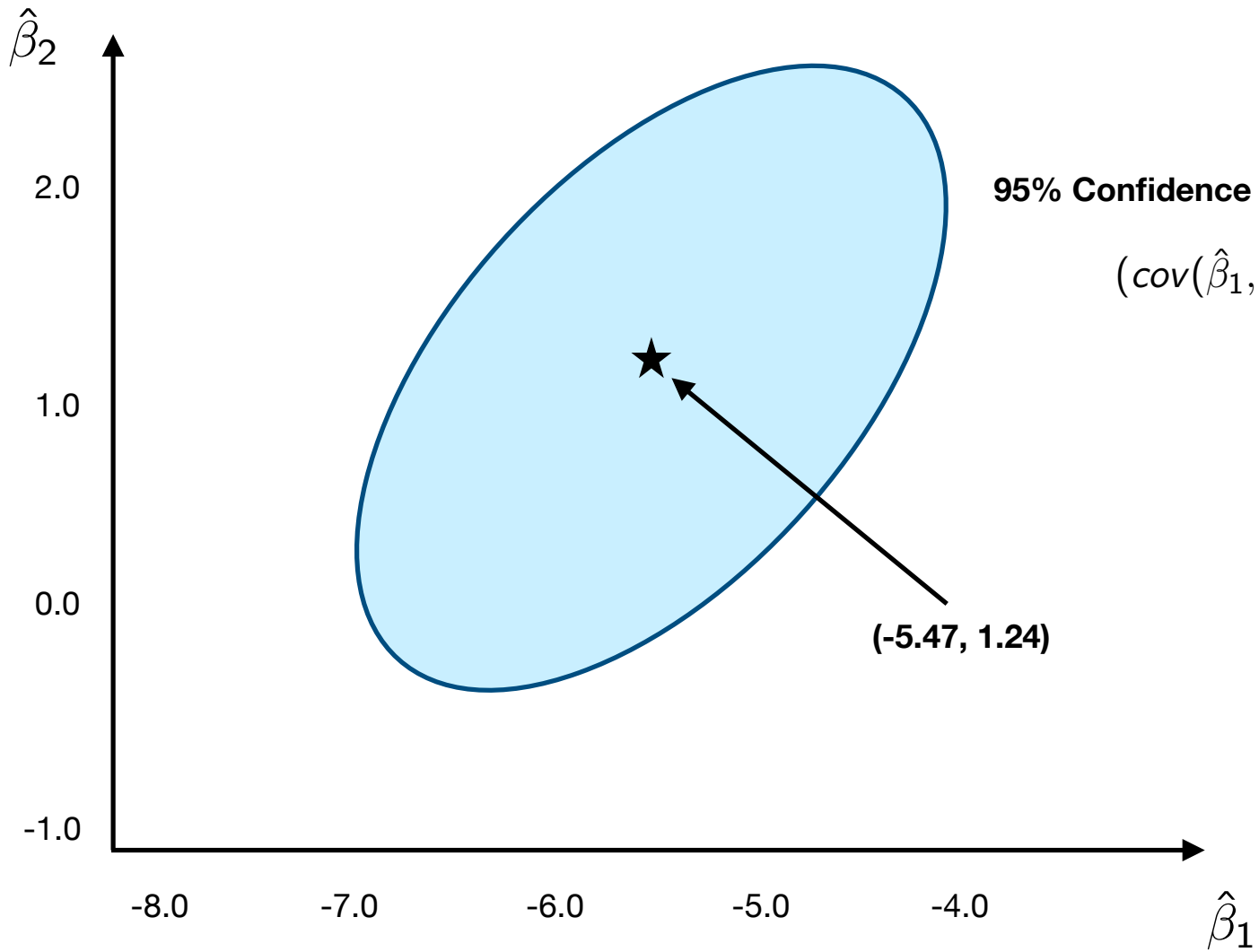
- The t-statistic on $ClassSize_i$ for the 2-sided test of the null $H_0 : \gamma = 0$ vs $H_1 : \gamma \neq 0$ is $t = -1.49$ with p-value=0.14
- The test result is exactly the same as the p-value from the joint test of $\beta_1 = \beta_2$ on the previous slide. This is because the single-coefficient test on this slide is identical to the joint test on the previous slide.

Confidence Sets for Multiple Coefficients

- Related to confidence intervals, you can construct confidence sets for multiple coefficients
- The 95% confidence set for two or more coefficients contains the set of hypothesized population values for the coefficients that cannot be jointly rejected at the 95% level of significance
- For example, suppose you want to construct a confidence set of two coefficients, β_1 and β_2
- You would construct a 95% confidence set based on the joint null hypothesis that $\beta_1 = \beta_{1,0}$ and $\beta_2 = \beta_{2,0}$ for all possible $(\beta_{1,0}, \beta_{2,0})$ pairs that you would not jointly reject at the 5% level using the F-statistic
- Similar to how 95% confidence intervals are based on the t-statistic, the 95% confidence set is based on the F-statistic
- When the set is constructed based on two coefficients, the confidence set is an ellipse

Confidence Sets for Multiple Coefficients Graphically

Regression: $\widehat{TestScore}_i = \underset{(9.20)}{60.41} - \underset{(0.31)}{5.47}\, ClassSize_i + \underset{(0.23)}{1.24}\, Income_i$

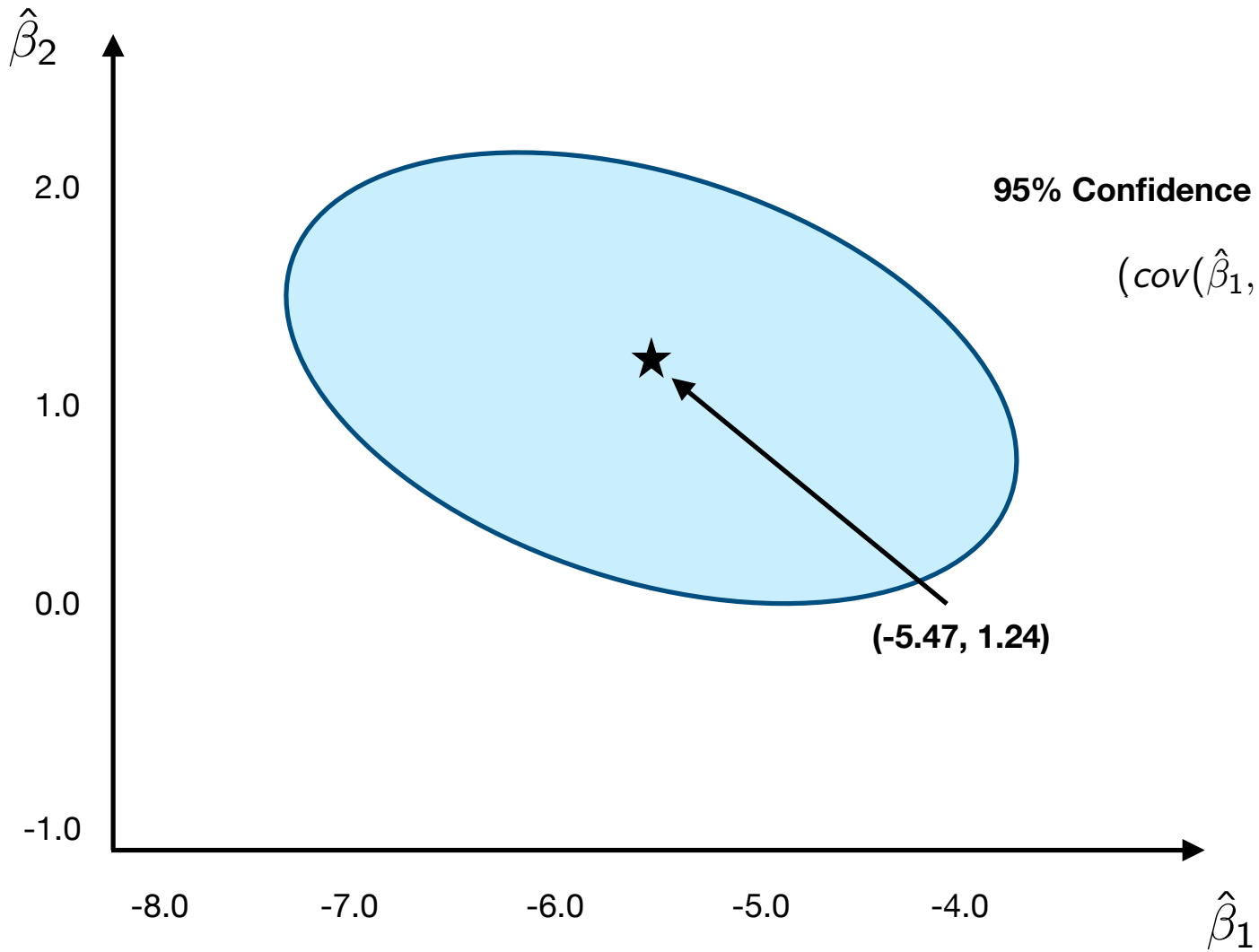


95% Confidence Set For: $(\hat{\beta}_1, \hat{\beta}_2)$

$(cov(\hat{\beta}_1, \hat{\beta}_2) > 0)$

Confidence Sets for Multiple Coefficients Graphically

Regression: $\widehat{TestScore}_i = \underset{(9.20)}{60.41} - \underset{(0.31)}{5.47}\, ClassSize_i + \underset{(0.23)}{1.24}\, Income_i$



95% Confidence Set For: $(\hat{\beta}_1, \hat{\beta}_2)$

$(cov(\hat{\beta}_1, \hat{\beta}_2) < 0)$

Model Specification for Multiple Regression

- With so many potential options, how do we know which regressors to include in a regression model?
- How do we know we have the "right" model estimated?
- The key guiding principle in developing econometric models to address an empirical question is to obtain an unbiased estimate of the causal effect of interest
- It is not soley about pure statistical measures of fit such as R^2 or $\bar{R}^2$ strategies)

Omitted Variable Bias and Control Variables

- In many econometric studies, multiple linear regressions roughly take on the following form:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i$$

  where the coefficient of interest is in red and the control variable coefficients are in blue
- The goal is to include a sufficient set of control variables in the regression to remove all sources of omitted variable bias in β_1
- In general, the coefficients on the control variables, $\beta_2, \ldots, \beta_k$, will be biased and not interperable as causal effects
- But this is not the primary aim of including control variables in a regression; they are there to hold all other factors fixed in obtaining an unbiased estimate of coefficient of interest, β_1
- This common approach to model building requires us to modify the first Least Squares assumption, which recall is $E[u_i|X_{1i}, X_{2i}, \ldots, X_{ki}] = 0$

Conditional Mean Independence

- When focusing on a regression with a coefficient of interest and control variables, we can replace the Independence assumption with a different assumption that ensures an unbiased estimate of the coefficient of interest
- For illustration, consider a regression where X_{1i} is the variable of interest and X_{2i} is the control variable
- The modified Least Squares assumption, Conditional Mean Independence says the conditional expectation of u_i given X_{1i} and X_{2i} is independent of X_{1i} (but it can potentially depend on X_{2i}).
- That is, conditional mean independence assumes:

$$E(u_i|X_{1i}, X_{2i}) = E(u_i|X_{2i})$$

Conditional Mean Independence

- If we only assume conditional mean independence:

$$E(u_i|X_{1i}, X_{2i}) = E(u_i|X_{2i})$$

OLS estimates based on the following regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

yield an unbiased β_1 estimate and a potentially biased β_2 estimate

Conditional Mean Independence

- Conditional mean independence says that once you control for X_{2i} in the regression, X_{1i} can be treated as if it were randomly assigned distributed observations $i = 1, \ldots, n$ and hence uncorrelated with u_i
- That is, after controlling for X_{2i}, X_{1i} has no predictive power for u_i in the regression as the conditional mean of u_i
- When a control variable like X_{2i} is included in the regression, it controls for both it own direct causal effect on Y_i as well as any indirect influence of omitted factors that would otherwise bias our OLS estimate of the coefficient of interest, β_1
- The same intuition and definitions hold for the inclusion of multiple regressors, $X_{2i}, \ldots, X_{ki}$, in the regression.
  - Instead of just having X_{2i} as controls on this and the previous slide, have the list $X_{2i}, \ldots, X_{ki}$
  - Collectively, the list of controls is used to ensure conditional mean independence with the variable of interest, X_{1i}

How to Decide on What is a Variable of Interest?

- There are typically three sources through which a dependent variable and coefficient of interest are defined

1. Policy: there is an outcome Y_i that is potentially impacted by a variable of interest X_{1i}, where X_{1i} can be influenced by policy decisions
   - Y_i =road congestion, X_{1i}=road toll price
   - Y_i =Pollution, X_{1i}=carbon tax
2. Testing Economic Theory: economic theory predicts a relationship between an outcome variable Y_1 and a variable of interest X_{1i}
   - Y_i =demand for a product, X_{1i}=price of a product
   - Y_i =employment, X_{1i}=minimum wage
3. Exploring New Phenomena: empirical investigations of a plausible link between an outcome variable Y_1 and a variable of interest X_{1i}, but that have not been closely considered previously by policy or economic theory
   - Y_i =voting in favour of women issues, X_{1i}=having a daughter
   - Y_i =billionaires in the country, X_{1i}=country openness to trade

How to Decide on What is a Variable of Interest?

- Some outcomes and coefficients of interest touch on all three motivations (policy, theory, exploration)
- Exploration of new phenomena in many cases involves interdisciplinary research, and can provide new insights for new policies and economic theories
  - cynics often label exploratory empirical research "pop science"
- Regardless of motivation, the outcome and variable of interest anchors your empirical analysis with a focus on obtaining an unbiased estimate of the regression coefficient of interest

How to Choose What Control Variables to Include

- The challenge with finding control variables is to convincingly remove omitted variable bias in the estimate of the coefficient of interest, and determining which control variables are most important to include in the regression
- Assessing the role of control variables in a regression analysis typically involves two groups of regressors:
- Base specification regressors are those which are the key set of control variables in the regression. These are determined by expert knowledge of the problem at hand, economic theory, or policy discussion of variables that determine Y_i
- Alternative specification regressors are those which are less obvious as control variables, but that you nonetheless check to see what impact their inclusion as controls has on the coefficient of interest

Interpreting the R^2 and $\bar{R}^2$ in Practice

- R^2 and $\bar{R}^2$ are useful for diagnosing the fit of your regression model, and the importance of additional control variables in explaining the data as you build your econometric model
- 4 pitfalls to be mindful of when using R^2 and $\bar{R}^2$
  1. An increase in R^2 or $\bar{R}^2$ does not necessarily mean that an added variable is statistically significant. We use t-statistics to assess this.
  2. A high R^2 or $\bar{R}^2$ does not mean that the regressors are a true cause of the dependent variable. It just means they are predictive of the dependent variable.
  3. A high R^2 or $\bar{R}^2$ does not mean that there is no omitted variable bias in the coefficient of interest. You can have omitted variable bias if R^2 and $\bar{R}^2$ is high. You can have no omitted variable if R^2 and $\bar{R}^2$ is low.
  4. A high R^2 or $\bar{R}^2$ does not necessarily mean that you have the most appropriate set of regressors, nor does having a low R^2 or $\bar{R}^2$ necessarily mean that you have an inappropriate set of regressors

Policy: Application to Test Scores and Class Size

- In developing an econometric analysis, it is useful to progressively add control variables to the regression to see their influence on the coefficient of interest and model fit
- For our policy-based application that studies test scores and class size, we can run the following sets of regressions:

$$TestScore_i = \beta_0 + \beta_1 ClassSize_i + u_i$$
$$TestScore_i = \beta_0 + \beta_1 ClassSize_i + \beta_2 Income_i + u_i$$
$$TestScore_i = \beta_0 + \beta_1 ClassSize_i + \beta_2 Income_i + \beta_3 Urban_i + u_i$$

- We focus on the magnitude and statistical significance of the β_1 coefficient estimate in each regression, as well as the $\bar{R}^2$

Policy: Test Scores and Class Size

The Effect of Class Size on Test Scores

	(1)	(2)	(3)
$ClassSize_i$	-6.45**	-5.47**	-1.22
	(0.29)	(0.31)	(1.67)
$Income_i$		1.24**	1.30**
		(0.23)	(0.22)
$Urban_i$			5.21*
			(2.01)
Constant	108.41**	60.41**	56.06**
	(3.18)	(9.20)	(9.07)
Adj. R-Squared	0.844	0.883	0.890
Observations	90	90	90

Notes: Dependent variable is the average test score in a class on VCE economics out of 100. Heteroskedasticity robust standard errors are reported in parentheses.
** $p < 0.01$,* $p < 0.05$

- The ** and * indicate rejecting the null that a given coefficient is equal to 0 at the 1% and 5% level, respectively
- So the column (2) -5.47 coefficient for $ClassSize_i$ is statistically significantly different from 0 at the 1% level of significance
- But the column (3) -1.22 coefficient for $ClassSize_i$ is not statistically significantly different from 0 at the 1% or 5% level, which can be interpreted as $ClassSize_i$ having no impact of $TestScore_i$

Policy: Test Scores and Class Size

The Effect of Class Size on VCE Economics Test Scores

	(1)	(2)	(3)
Number of Students in the Class	-6.45**	-5.47**	-1.22
	(0.29)	(0.31)	(1.67)
Average Household Income (1000's)		1.24**	1.30**
		(0.23)	(0.22)
Classroom in Location with > 100,000 People			5.21*
			(2.01)
Constant	108.41**	60.41**	56.06**
	(3.18)	(9.20)	(9.07)
Adj. R-Squared	0.844	0.883	0.890
Observations	90	90	90

Notes: Dependent variable is the average test score in a class on VCE economics out of 100. Heteroskedasticity robust standard errors are reported in parentheses.
** $p < 0.01$,* $p < 0.05$

- This is an even better table than the table on the previous slide because the regressors are described more clearly, and not just in terms of their names in the regression which are less meaningful to a reader
- Ideally, tables and figures should be completely self explanatory based on their titles and all labels

Testing Economic Theory: Application to Petrol Prices and Consumer Search

- A key prediction from economic theory states that when there is more dispersion in prices, people will search more for deals
- Example: Coles and Woolworths
  - if Coles and Woolworths had the same prices for everything, then you would never check prices at both stores for deals
  - if Coles and Woolworths offer different prices, then you have an incentive to check prices at both stores to find a deal
- Economists' call the behaviour of "looking for deals" consumer search
- Prediction from theory to test: $\uparrow$ price dispersion $\rightarrow \uparrow$ consumer search

Testing Economic Theory: Petrol Prices and Search

- Search theory is important: it underpins tools for studying firms' market power in the field of industrial organisation, for studying individuals' education and job choices in the field of labour economics, and for understanding the impact of the internet on market outcomes (e.g., prices, quantities, product variety) and consumer welfare
- Yet, despite being a foundational theoretical concept in economics, there is actually very little direct empirical evidence on whether consumers actually behave according to the predictions of search theory
- The reason for this is simple: measuring both search behaviour and price dispersion is hard

Testing Economic Theory: Petrol Prices and Search

- Dependent variable Y_i: number of daily website hits on an online petrol price disclosure website in WA called Fuelwatch
- Variable of interest X_{i1}: daily standard deviation of petrol prices in Perth
- This research is an example of empirical industrial organization, which is my main area of research
- Reference: Byrne, D. P. and N. de Roos (2017): "Consumer Search in Retail Gasoline Markets," *Journal of Industrial Economics*, 65(1), 2017.

Testing Economic Theory: Petrol Prices and Search

Home | About FuelWatch | My FuelWatch | Price Search | FuelWatch News | Fuel Information | For Industry

FuelWatch Quick Search - Results

Product: ULP
Metro Region: Any Metro
Suburbs: Any Suburb (including surrounding suburbs)
Brands: Any Brand
Country Region: None
Date: Today and tomorrow

Refine Search | New Search

Best prices available from 6:00am for today and tomorrow

Today	Tomorrow	Product	Brand	Name Mouse over Name for details	Address	Suburb/Town	Map
142.8	157.9	ULP	Caltex Woolworths	Caltex Woolworths Greenwood	37 Canham Way (cnr Wanneroo Rd)	GREENWOOD	
142.8	157.9	ULP	Caltex Woolworths	Caltex Woolworths South Lake	752 North Lake Rd (Cnr Berrigan Drive)	SOUTH LAKE	
145.8	157.9	ULP	Caltex Woolworths	Caltex Woolworths Canning Vale	Cnr Amherst & Nicholson Rds	CANNING VALE	
145.8	156.9	ULP	Gull	Gull North Perth	311 Fitzgerald St	NORTH PERTH	
145.8	156.9	ULP	Gull	Gull Rockingham	40 Kent Street	ROCKINGHAM	
145.8	157.9	ULP	Caltex Woolworths	Caltex Woolworths Warnbro Fair	202 Warnbro Sound Avenue	WARNBRO	
146.1	146.1	ULP	Coles Express	Coles Express Casuarina (WA)	Cnr Thomas & Johnson Rds	BERTRAM	
146.1	146.1	ULP	Coles Express	Coles Express Clarkson	Cnr Marmion Ave & Pensacola Tce	CLARKSON	
146.1	146.1	ULP	Coles Express	Coles Express Fremantle	101 Hampton Rd	FREMANTLE	
146.7	155.7	ULP	United	United Leda	Cnr Feilman Drive & Gilmore Avenue	LEDA	
146.7	156.7	ULP	United	United Mindarie	Cnr Marmion & Hester Ave	MINDARIE	
146.7	155.7	ULP	United	United Mt Lawley	791 Beaufort Street	MT LAWLEY	
146.7	156.7	ULP	United	United Quinns Rocks	Cnr Tapping Way & Quinns Road	QUINNS ROCKS	
146.7	155.7	ULP	United	United South Lake	49 Berrigan Drive	SOUTH LAKE	
147.5	156.7	ULP	United	United Hepburn Heights	Cnr Walter Padbury Blvd & Hepburn Av	PADBURY	
147.7	156.5	ULP	United	United Kewdale	382-412 Orrong Rd	KEWDALE	
147.7	155.7	ULP	United	United Lexia	779 Gnangara Rd	LEXIA	
147.7	156.5	ULP	United	United Northbridge	31 Fitzgerald Street	NORTHBRIDGE	
147.7	155.7	ULP	United	United Roleystone	11 Wygonda Rd	ROLEYSTONE	
148.5	156.7	ULP	United	United Spearwood	Cnr Stock Rd & Barrington St	SPEARWOOD	

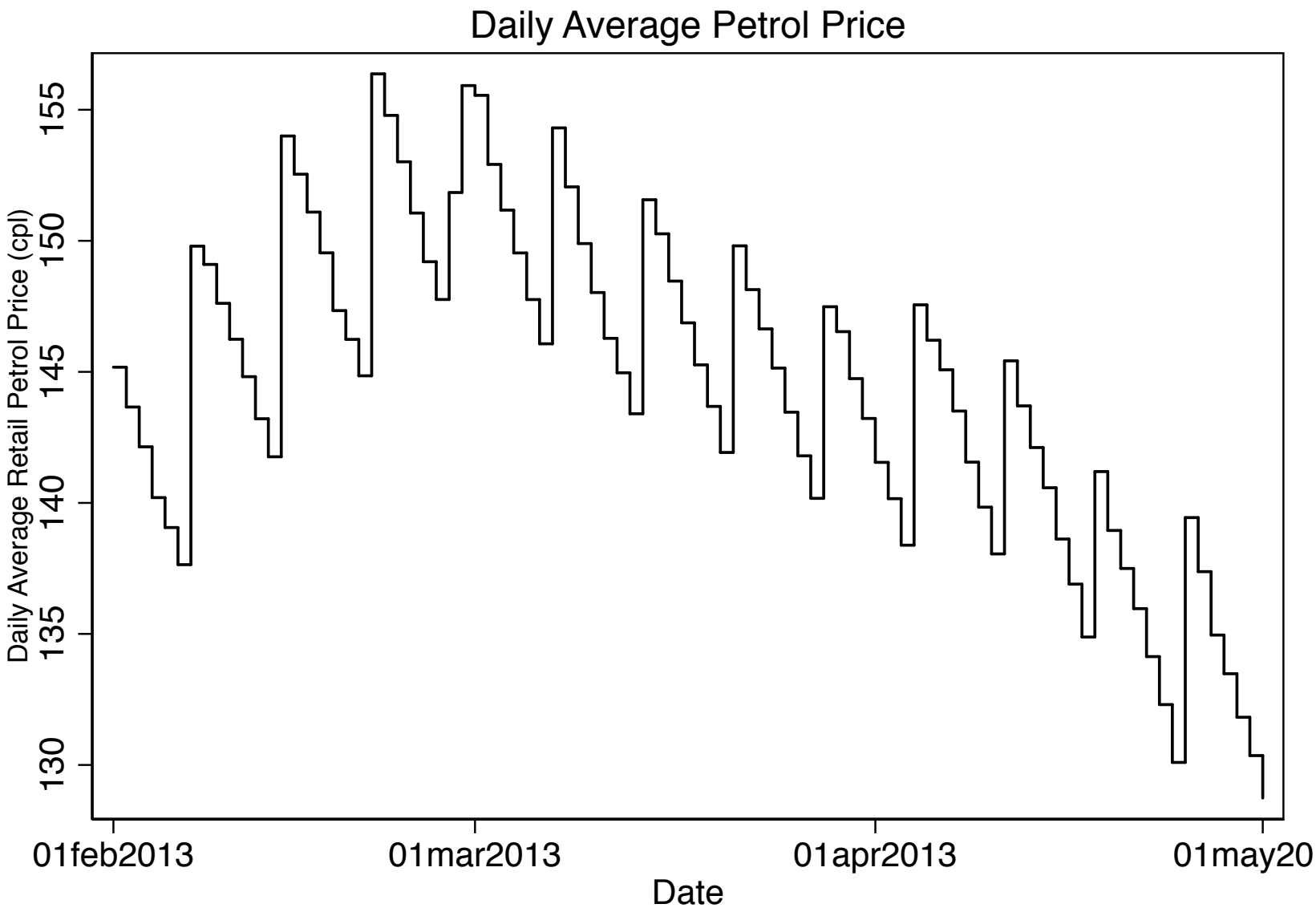
* Unmanned Site (Credit card charges may apply)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Items per page 20

204 Prices found

Testing Economic Theory: Petrol Prices and Search



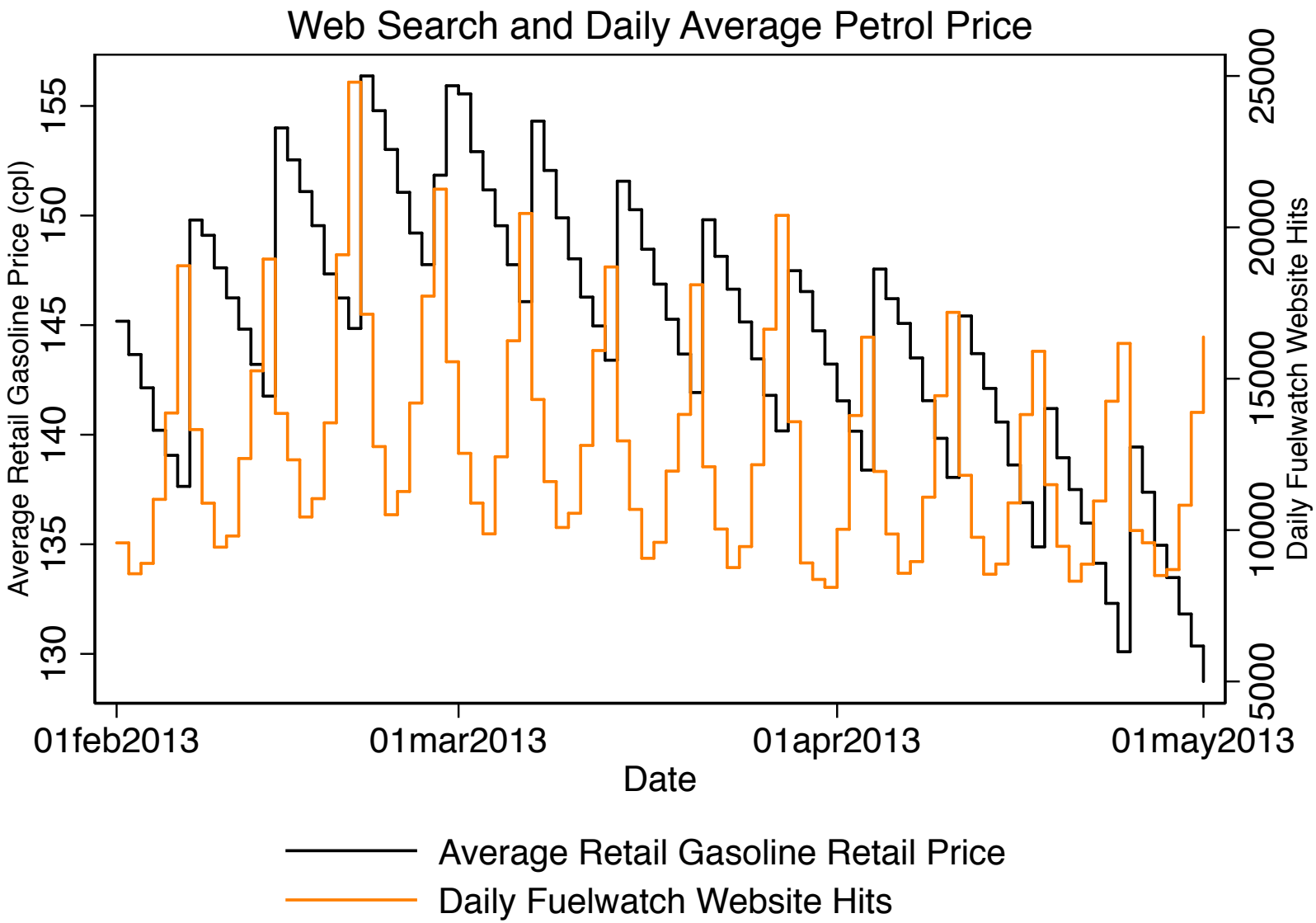
Testing Economic Theory: Petrol Prices and Search

Table I
DESCRIPTIVE STATISTICS

Day of the Cycle	Fuelwatch Website Visits		Price Changes		Price Dispersion	
(Mon) 3 days before jump	11718.80	(1053.70)	-1.76	(0.21)	2.56	(0.43)
(Tue) 2 days before jump	14778.25	(1864.02)	-1.62	(0.24)	2.61	(0.45)
(Wed) 1 day before jump	17991.29	(2907.60)	-1.64	(0.20)	2.96	(0.40)
(Thu) Price jump day	12965.93	(1687.83)	9.44	(2.03)	5.71	(0.81)
(Fri) 1 day after jump	10417.66	(1796.57)	-1.18	(0.66)	3.93	(0.89)
(Sat) 2 days after jump	9062.16	(968.03)	-1.76	(0.49)	3.17	(0.77)
(Sun) 3 days after jump	9462.62	(852.76)	-1.58	(0.28)	2.76	(0.57)

Notes to Table I: $N = 381$. Sample averages and standard deviations (in parentheses) by day of the cycle reported. Price jump days classified as those where the daily change in the average price across stations is positive.

Testing Economic Theory: Petrol Prices and Search



Testing Economic Theory: Petrol Prices and Search

- The multiple linear regression equation we use to study the impact of price dispersion on consumer search on date t is:

$$\text{Search}_t = \alpha_0 + \sum_{\tau=-3}^{3} \alpha_{1\tau} d_\tau + \alpha_2 \sigma_{p_t} + \sum_{k=t+1}^{t-1} \alpha_{3k} \Delta p_t + \mathbf{X}_t \beta + \epsilon_t$$

  - d_τ is a dummy variable that equals 1 if date t is τ days before/after a price jump
  - σ_{p_t} is the standard deviation of prices on date t
  - $\Delta p_t = p_t - p_{t-1}$ is the change in average daily petrol prices between dates t and $t+1$
  - $\mathbf{X}_t$ contains various control variables for weather (max daily temperature, dummy for rain), holidays and holiday weekend dummies, and week-of-year dummies to control for seasonal fluctuations in search
  - ϵ_t is the econometric error term
- The main coefficient of interest is α_2, the relationship between consumer search and price dispersion

Testing Economic Theory: Petrol Prices and Search

Fuelwatch Website Visits, Price Changes and Price Dispersion

	(1)	(2)	(3)		
2 days before price jump	2.834**			2.798**	2.785**
	(0.227)			(0.233)	(0.234)
1 day before price jump	6.197**			5.928**	3.671**
	(0.319)			(0.303)	(0.989)
Price Jump Day	1.219**			-0.987*	-2.740**
	(0.146)			(0.447)	(0.923)
1 day after price jump	-0.963**			-2.006**	-2.421**
	(0.159)			(0.224)	(0.594)
2 days after price jump	-2.559**			-3.002**	-2.926**
	(0.147)			(0.169)	(0.176)
3 days after price jump	-2.280**			-2.406**	-2.369**
	(0.089)			(0.104)	(0.097)
Price Standard Deviation			-0.772**	0.709**	0.518**
			(0.206)	(0.153)	(0.148)
Δp_{t+1}		0.588**	0.602**		0.207*
		(0.030)	(0.032)		(0.095)
Δp_t		0.134**	0.329**		0.197*
		(0.015)	(0.053)		(0.090)
Δp_{t-1}		-0.051**	0.022		0.050
		(0.012)	(0.025)		(0.048)
Observations	381	381	381	381	381
Adj. R-Squared	0.916	0.608	0.628	0.921	0.924

Notes: Dependent variable is the number of daily web searches on the Fuelwatch website in terms of 1000's of hits. Heteroskedasticity robust standard errors are reported in parentheses. ** $p < 0.01$,* $p < 0.05$

Exploring New Phenomena: Application to the Baby Bonus and Birth Timings

- We will consider a separate example here that studies the impact that the Australian baby bonus on birth timings
- Reference: Gans, Joshua S. and Andrew Leigh (2007): "Born on the First of July: An (Un)natural Experiment in Birth Timing," *Journal of Public Economics*, 93(2), 246-263.

Baby Bonus and Birth Timings

- Children born on or after July 1, 2004 would receive a $3000 "Baby Bonus" per baby
  - For the median household, this is about 2.8 weeks of post-tax income
- Policy was announced 7 weeks before July 1, 2004

Baby Bonus and Birth Timings

- An exchange between Hon. Tim Lester (Labor MP) and Hon. Kay Patterson (Liberal MP, Minister for Family and Community Services) in 2004:

TIM LESTER: Minister, with the benefit of hindsight, would it have been better to have announced and introduced this policy on the same day?

KAY PATTERSON: This policy is a bonus to families.

TIM LESTER: That doesn't answer my question, though, with respect, Minister. Would it have been better to have announced and introduced this policy at the same time?

KAY PATTERSON: I believe this is a fantastic policy for mothers, they're going to get $3000 –

TIM LESTER: Minister, that still doesn't answer my question with respect. Would it have been better to have announced and introduced this policy at the same time?

KAY PATTERSON: Well if I thought that mothers would put their babies at risk, but I don't believe mothers would put them at risk.[1]

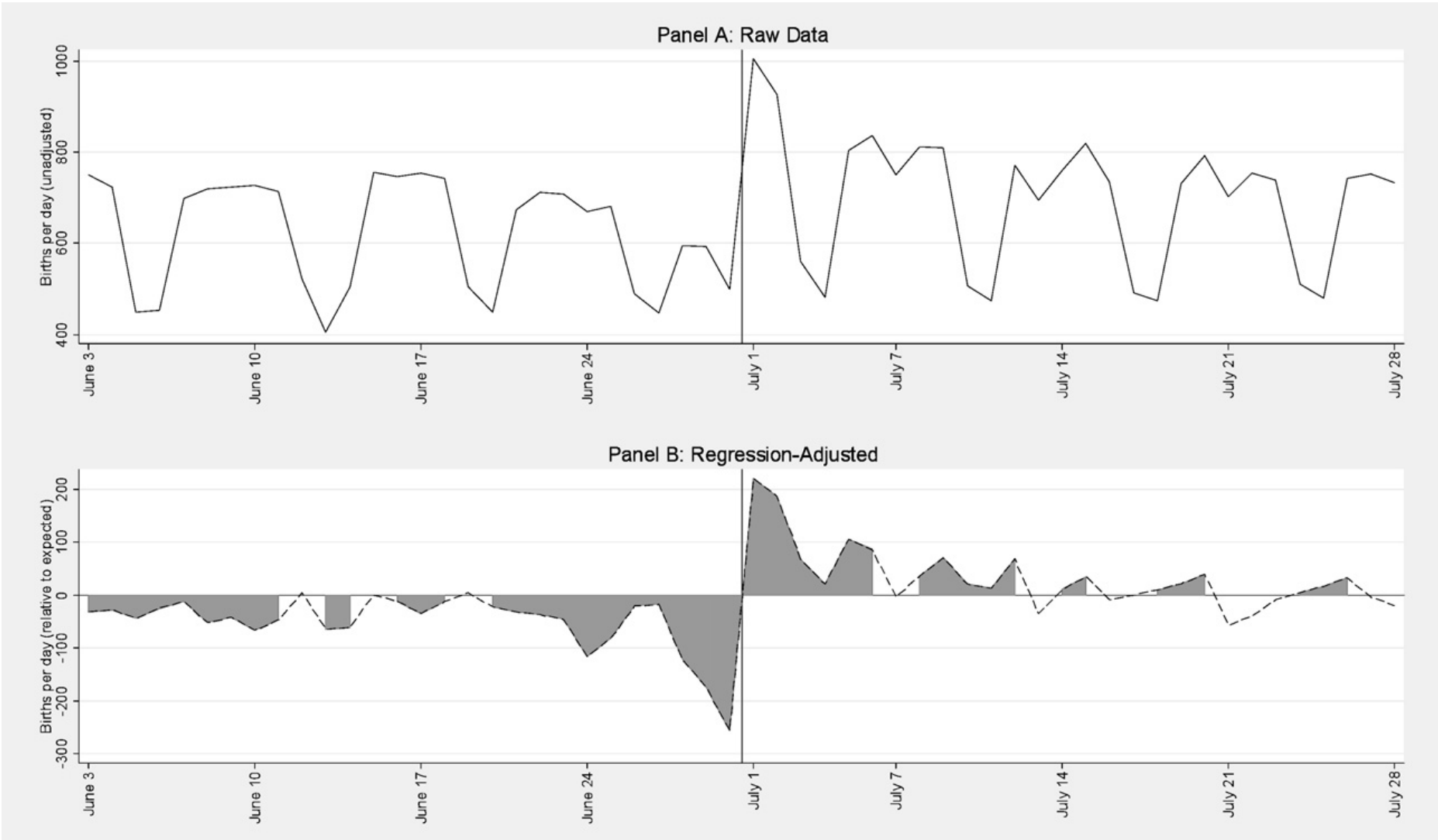
Baby Bonus and Birth Timings

- Using regressions and individual-level birth data, the authors estimate that about 1000 births were "moved" so parents were eligible for the Baby Bonus

Baby Bonus and Birth Timings

- Data set is from the Australiab Bureau of Statistics (ABS)
- Dependent variable: daily data on the number of Australian births between 1975 and 2004
  - 232,682 births in 1975, 245,143 births in 2004
- Comes from state and territory hospital birth registries
- Data does not include unregistered births
- Let's look at the raw births data between June 3, 2004 and July 28, 2004 in Panel A on the following slide

Baby Bonus and Birth Timings

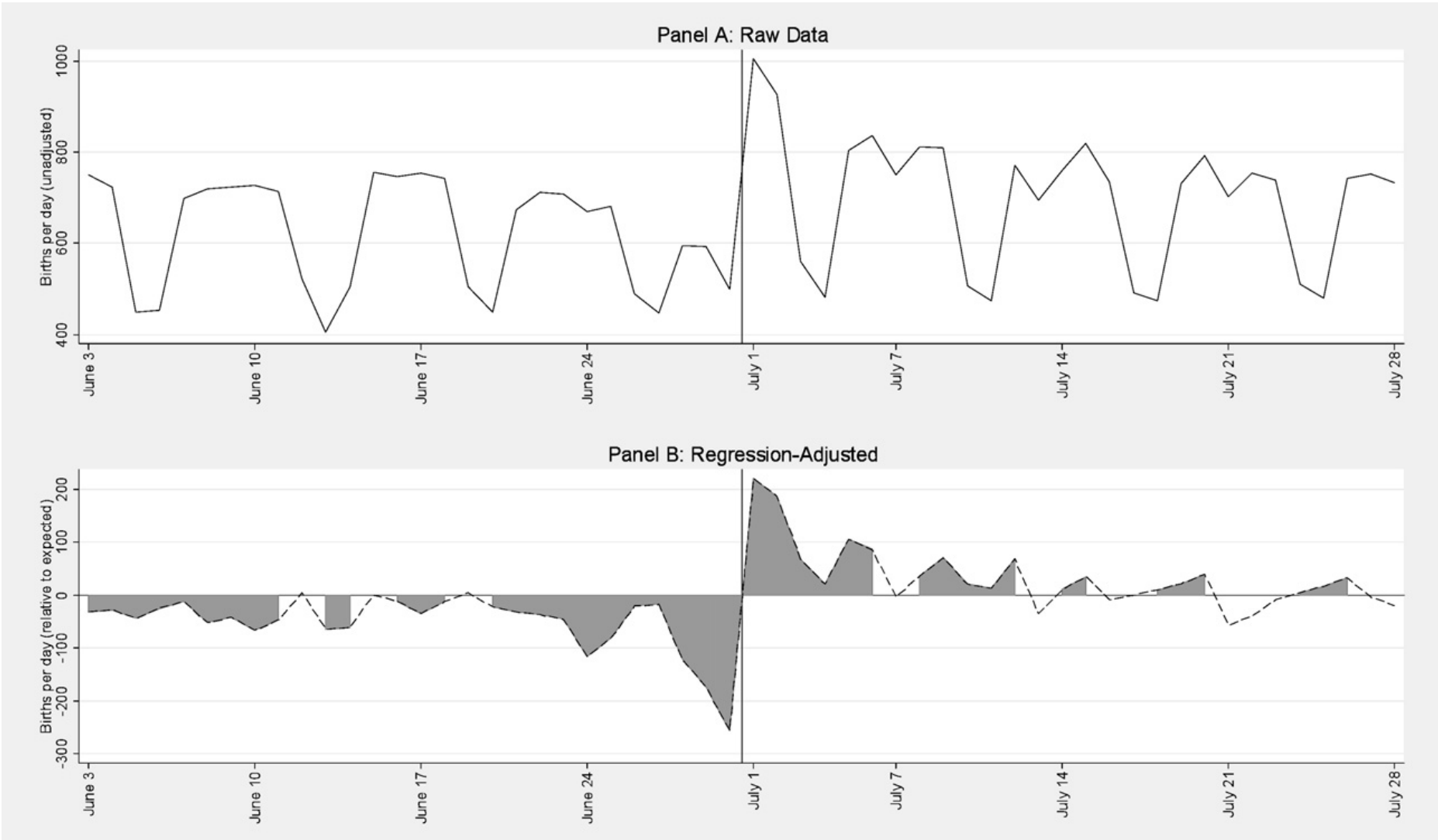


Baby Bonus and Birth Timings

$$\text{Births}_i = I_i^{\text{Year}} \times I_i^{\text{Day of Week}} + I_i^{\text{Day of Year}} + I_i^{\text{Public Holiday}} + \varepsilon_i$$

- Above is a regression model that predicts births for each date i using dummy variables only
- First (interaction) of Year and Day of Week dummy variables: controls for year-specific days of the week (e.g., Thursdays in 2004, Fridays in 2005, etc.)
- Second dummy variable: controls for day of the year (e.g. July 1, Sep 14, Dec 12 etc.)
- Third dummy variable: controls for public holidays that do not always fall on the same day of the week or day of the year
- The residual from this regression highlights unpredictable births which we can compare to the actual births in Panel B of the figure to visually highlight the Baby Bonus impact

Baby Bonus and Birth Timings



Baby Bonus and Birth Timings

$$\text{Births}_i = I_i^{\text{Baby Bonus}} + I_i^{\text{Year}} \times I_i^{\text{Day of Week}} + I_i^{\text{Day of Year}} + I_i^{\text{Public Holiday}} + \varepsilon_i$$

- Same regression as before except the first dummy variable for the Baby Bonus for dates after July 1, 2004
- Dummy variable for Baby Bonus equals one if the date in the dataset is after July 1, 2004
- They run the regressions using dates July 1 $\pm$ 7 days from each year between 1975 and 2004
  - so the regression has 30 years $\times$ 14 days/year $=$ 420 days
- Other sample windows: $\pm$ 14 days (840 obs), $\pm$ 21 days (1260 obs), $\pm$ 28 days (1680 obs)
- Regression coefficient on Baby Bonus estimates the impact of Baby Bonus on births within a window around July 1, 2004

Baby Bonus and Birth Timings

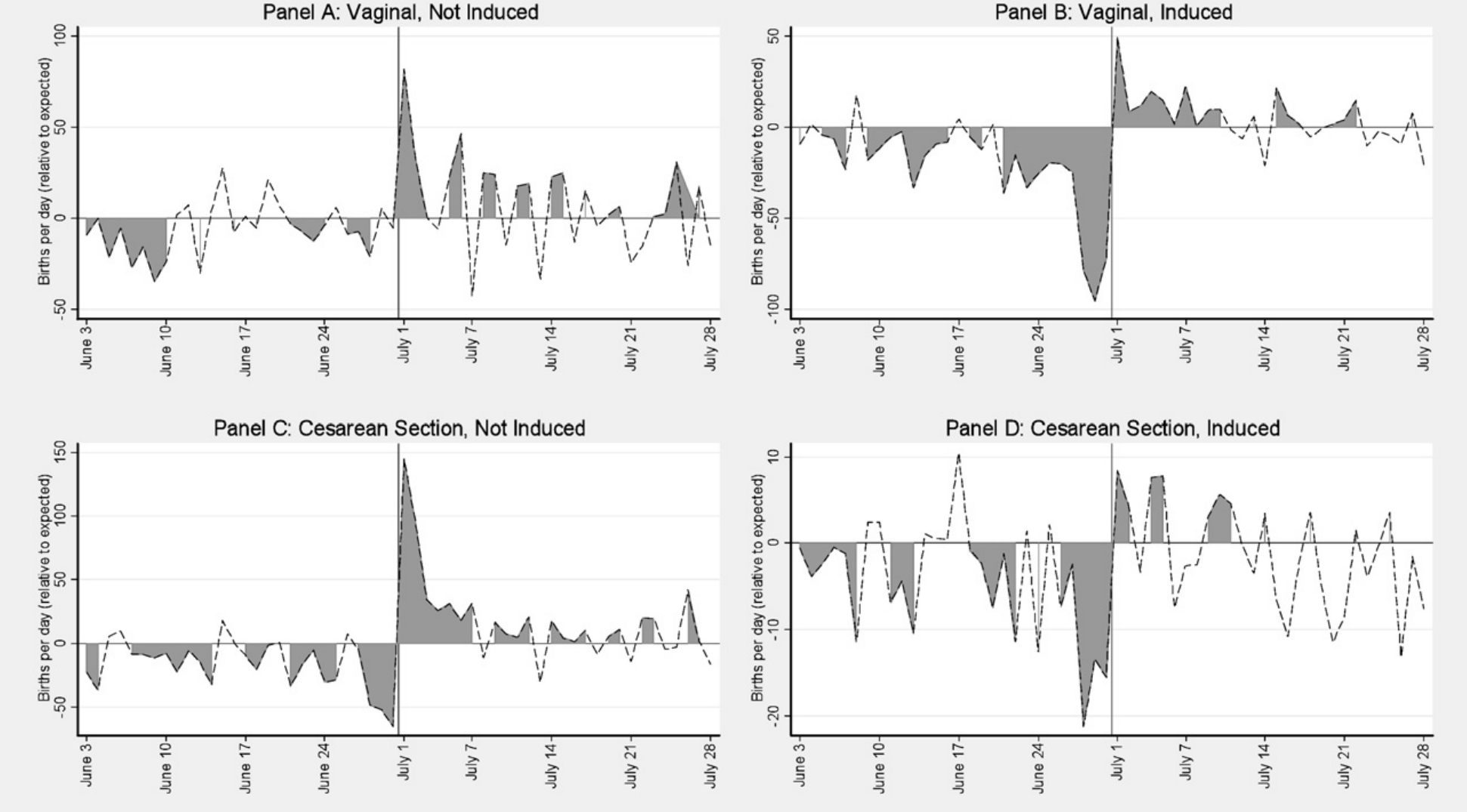
Window	(1)	(2)	(3)	(4)
	±7 days	±14 days	±21 days	±28 days
Panel A: Dependent variable is number of births				
Baby Bonus	210.507***	131.382***	101.624***	83.602***
	[15.911]	[11.626]	[9.579]	[8.386]
Observations	420	840	1260	1680
R-squared	0.97	0.94	0.94	0.93
Number of births moved	*737*	*920*	*1067*	*1170*
Panel B: Dependent variable is ln(number of births)				
Baby Bonus	0.300***	0.187***	0.147***	0.123***
	[0.023]	[0.017]	[0.014]	[0.013]
Observations	420	840	1260	1680
R-squared	0.97	0.95	0.94	0.94
Share of births moved	*16%*	*10%*	*8%*	*6%*

Notes: Standard errors in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%. Sample is daily births within the relevant window from 1975–2004. All specifications include day of year, public holiday, and year × day of week fixed effects. Window denotes the number of days before and after the start of July. For example, the ±7 day window covers the last seven days of June and the first seven days of July. Number of births moved is $W\beta/2$, where W is the number of days in the window. Share of births moved is $\exp(\beta/2)-1$.

Baby Bonus and Birth Timings

- How were the births moved?
- The authors conduct a similar analysis considering four types of pregnancies
  - vaginal, not induced (50% of pregnancies)
  - vaginal, induced (20% of pregnancies)
  - cesarean section, induced (5% of pregnancies)
  - cesarean section, not induced (5% of pregnancies)
- Data only available for 2004-2005
- The data shows two ways in births were delayed
  - substantial reduction in birth inductions just before July 1, 2004 (which can sometimes be delayed)
  - large jump in cesarean sections just after July 1, 2004 (which can be scheduled)

Baby Bonus and Birth Timings



Baby Bonus and Birth Timings

- Re-run the main regression for estimating the Baby Bonus effect using the following subsamples based on types of births
  - vaginal, not induced (50% of pregnancies)
  - vaginal, induced (20% of pregnancies)
  - cesarean section, induced (5% of pregnancies)
  - cesarean section, not induced (5% of pregnancies)
- Slightly different set of controls: regressors are day-of-week dummies and public holiday dummies
  - we cannot include year-specific day of the week dummies or day of the year dummies because only one-year of data are available on types of births
  - you would run into the dummy variable trap if you tried to estimate include these sets of dummy variables in the regression

Baby Bonus and Birth Timings

Window	(1)	(2)	(3)	(4)
	±7 days	±14 days	±21 days	±28 days
Panel A: Vaginal, not induced				
Baby Bonus	24.429	16.429*	12.647*	13.661**
	[14.651]	[9.088]	[6.963]	[5.679]
Observations	14	28	42	56
R-squared	0.67	0.38	0.32	0.28
Number of births moved	*86*	*115*	*133*	*191*
Panel B: Vaginal, induced				
Baby Bonus	66.143***	39.714***	33.186***	25.250***
	[11.587]	[8.553]	[6.363]	[5.311]
Observations	14	28	42	56
R-squared	0.94	0.86	0.87	0.87
Number of births moved	*232*	*278*	*348*	*354*
Panel C: Cesarean section, not induced				
Baby Bonus	86.429***	51.214***	37.412***	32.448***
	[19.798]	[13.489]	[10.089]	[7.856]
Observations	14	28	42	56
R-squared	0.9	0.8	0.82	0.83
Number of births moved	*303*	*358*	*393*	*454*
Panel D: Cesarean section, induced				
Baby Bonus	12.143**	7.643***	4.304*	3.089*
	[3.700]	[2.547]	[2.184]	[1.710]
Observations	14	28	42	56
R-squared	0.85	0.76	0.72	0.73
Number of births moved	*43*	*54*	*45*	*43*

Dependent Variable is the Number of Births by Various Procedures.

Notes: Standard errors in brackets. * significant at 10%; ** significant at 5%; *** significant at 1%. All specifications are based on data from 2004 only, and include day of week fixed effects and an indicator for public holidays. Window denotes the number of days before and after the start of July. For example, the ±7 day window covers the last seven days of June and the first seven days of July. Number of births moved is $W\beta/2$, where W is the number of days in the window.

What's remaining in the subject?

- We are now tooled up with multiple linear regression, the central tool for econometric analysis
- The remainder of the subject builds around this main tool
  - Nonlinear regression
  - Evaluating econometric analyses
  - Field and natural experiments
  - Time series analysis