OLS Asymptotics

Paul Schrimpf

UBC
Economics 326

February 7, 2017

References

- Wooldridge (2013) chapter 5
- Stock and Watson (2009) chapter 18
- Angrist and Pischke (2014) appendix of chapter 2
- Review of asymptotics
    - Wooldridge (2013) appendix C
    - Menzel (2009) especially VI-IX

Section 1

Motivation

Motivation

- Our six regression assumptions,

MLR.1 (linear model)
MLR.2 (independence) $\{(x_{1,i}, x_{2,i}, y_i)\}_{i=1}^n$ is an independent random sample
MLR.3 (rank condition) no multicollinearity: no $x_{j,i}$ is constant and there is no exact linear relationship among the $x_{j,i}$
MLR.4 (exogeneity) $\mathrm{E}[\epsilon_i | x_{1,i}, \ldots, x_{k,i}] = 0$
MLR.5 (homoskedasticity) $\mathrm{Var}(\epsilon_i | X) = \sigma_\epsilon^2$
MLR.6 $\epsilon_i | X \sim N(0, \sigma_\epsilon^2)$

especially MLR.6 (and to a lesser extent MLR.1 and MLR.4) are often implausible

- Requiring OLS to only be consistent instead of unbiased will let us relax MLR.1 and MLR.4
- We will use the Central limit theorem to relax assumption MLR.6 and still perform inference (t-tests and F-tests)

Review of asymptotic inference

- Idea: use limit of distribution of estimator as $N \to \infty$ to approximate finite sample distribution of estimator
- Notation:
  - Sequence of samples of increasing size n, $S_n = \{(y_1, x_1), ..., (y_n, x_n)\}$
  - Estimator for each sample $\hat{\theta}$ (implicitly depends on n)

Section 2

Consistency

Review of convergence in probability

- $\hat{\theta}$ converges in probability to θ if for every $\epsilon > 0$,
$$\lim_{n\to\infty} \mathrm{P}\left(\left|\hat{\theta} - \theta\right| > \epsilon\right) = 0$$
denote by $\operatorname{plim} \hat{\theta} = \theta$ or $\hat{\theta} \xrightarrow{p} \theta$
- Show using a law of large numbers: if $y_1, ..., y_n$ are i.i.d. with mean μ; or if $y_1, ..., y_n$ have finite expectations and $\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^n \mathrm{Var}(y_i) = 0$ is finite, then $\bar{y} \xrightarrow{p} \mathrm{E}[Y]$
- Properties:
  - $\operatorname{plim} g(\hat{\theta}) = g(\operatorname{plim} \hat{\theta})$ if g is continuous (continuous mapping theorem (CMT))
  - If $\hat{\theta} \xrightarrow{p} \theta$ and $\hat{\zeta} \xrightarrow{p} \zeta$, then (Slutsky's lemma)
    - $\hat{\theta} + \hat{\zeta} \xrightarrow{p} \theta + \zeta$
    - $\hat{\theta}\hat{\zeta} \xrightarrow{p} \theta\zeta$
    - $\frac{\hat{\theta}}{\hat{\zeta}} \xrightarrow{p} \frac{\theta}{\zeta}$
- $\hat{\theta}$ is a consistent estimate of θ if $\hat{\theta} \xrightarrow{p} \theta$

Consistency of OLS

- Bivariate regression of y on x
- Slope:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}) y_i}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}$$

- Working with the numerator:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}) y_i = & \left(\frac{1}{n}\sum_{i=1}^n x_i y_i\right) - \bar{x}\frac{1}{n}\sum_{i=1}^n y_i \\ = & \left(\frac{1}{n}\sum_{i=1}^n x_i y_i\right) - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)\left(\frac{1}{n}\sum_{i=1}^n y_i\right) \\ & \text{using LLN} \\ \xrightarrow{p} & \mathrm{E}[xy] - \mathrm{E}[x]\mathrm{E}[y] = \mathrm{Cov}(x, y) \end{aligned}$$

Consistency of OLS

- Similarly

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \xrightarrow{p} \text{Var}(x)$$

- Then by Slutsky's lemma, $\hat{\beta}_1 \xrightarrow{p} \frac{\text{Cov}(x,y)}{\text{Var}(x)}$
- Recall that $\frac{\text{Cov}(x,y)}{\text{Var}(x)}$ is equal to the population regression coefficient
    - The bivariate population regression of Y on X is

$$(\beta_0, \beta_1) = \arg\min_{b_0, b_1} \text{E}[(Y - b_0 - b_1 X)^2]$$

i.e. $\beta_1 = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$ and $\beta_0 = \text{E}[y] - \beta_1 \text{E}[x]$

- Thus, OLS consistently estimates the population regression under very weak assumptions

Consistency of OLS

- We only need to assume $\frac{1}{n}\sum_{i=1}^n x_i \xrightarrow{p} \mathrm{E}[x]$, $\frac{1}{n}\sum_{i=1}^n y_i \xrightarrow{p} \mathrm{E}[y]$, $\frac{1}{n}\sum_{i=1}^n x_i^2 \xrightarrow{p} \mathrm{E}[x^2]$, and $\frac{1}{n}\sum_{i=1}^n x_i y_i \xrightarrow{p} \mathrm{E}[xy]$. There are multiple versions of the law of large numbers that would make this true. The details of LLNs are not important for this course, so we will be slightly imprecise and say that this is true assuming x_i and y_i have finite second moments and are not too dependent

Theorem

Assume $y_i, x_{i1}, \ldots, x_{ik}$ have finite second moments and observations are not too dependent then OLS consistently estimates the population regression of y on x_1, ..., x_k

Consistency of OLS

- Recall that the population regression is the minimal mean square error linear approximation to the conditional expectation function, i.e.

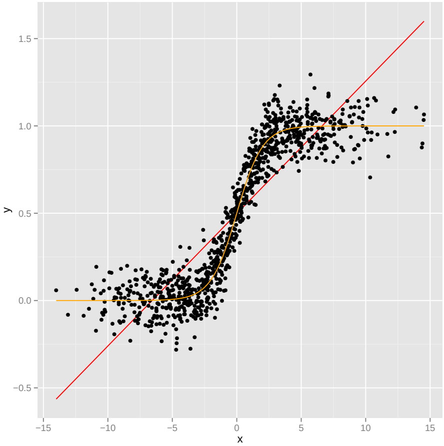
$$\underbrace{\arg\min_{b_0,b_1} \mathrm{E}\left[(Y-(b_0+b_1X))^2\right]}_{\text{population regression}} = \arg\min_{b_0,b_1} \underbrace{\mathrm{E}_X\left[(\mathrm{E}[Y|X]-(b_0+b_1X))^2\right]}_{\text{MSE of linear approximation to } \mathrm{E}[Y|X]}$$

- Population regression (and the conditional expectation function) might not be (and often is not) the model you want to estimate
- Population regression (and the conditional expectation function) are not causal
- If we have a true linear model,

$$y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i} + \epsilon_i$$

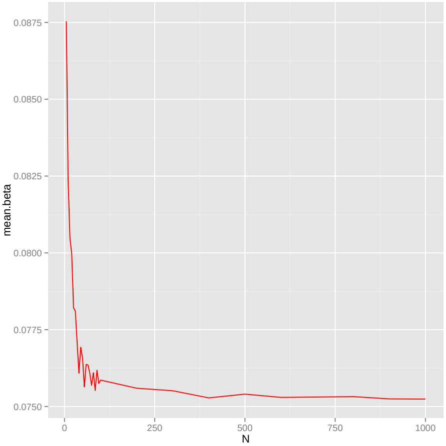
then OLS is consistent for β_j if $\mathrm{E}[\epsilon_i x_i] = 0$
  - $\mathrm{E}[\epsilon_i x_i] = 0$ is a weaker assumption than $\mathrm{E}[\epsilon_i | x_i] = 0$.

Example: nonlinear CEF: $\hat{\beta}$ biased but consistent estimator of population regression



Code

Example: nonlinear CEF: $\hat{\beta}$ biased but consistent estimator of population regression



Code

When is OLS not consistent?

- OLS is always a consistent estimator of the population regression
- OLS might not be consistent if the model we want to estimate is not the population regression
- Examples:
  - Omitted variables: want to estimate

    $$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i$$

    but $x_{2,i}$ not observed, so estimate

    $$y_i = \beta_0^s + \beta_1^s x_{1,i} + u_i$$

    instead
  - Causal effect: want slope to be the causal effect of x on y
  - Economic model: e.g. production function

Section 3

Asymptotic normality

Review of central limit theorem

- Let F_n be the CDF of $\hat{\theta}$ and W be a random variable with CDF F
- $\hat{\theta}$ converges in distribution to W, written $\hat{\theta} \xrightarrow{d} W$, if $\lim_{n\to\infty} F_n(x) = F(x)$ for all x where F is continuous
- Central limit theorem: Let $\{y_1, ..., y_n\}$ be i.i.d. with mean μ and variance σ^2 then $Z_n = \sqrt{n}\left(\bar{y}_n - \mu\right)$ converges in distribution to a $N(0, \sigma^2)$ random variable
  - As with the LLN, the i.i.d. condition can be relaxed if additional moment conditions are added; we will not worry too much about the exact assumptions needed
  - For non-i.i.d. data, if $\mathrm{E}[y_i] = \mu$ for all i and $v = \lim_{n\to\infty} \mathrm{E}\left[\left(\frac{1}{n}\sum_{i=1}^n y_i - \mu\right)^2\right]$ exists (and some technical conditions are met) then

$$\sqrt{n}\left(\bar{y}_n - \mu\right) \xrightarrow{d} N(0, v)$$

Review of central limit theorem

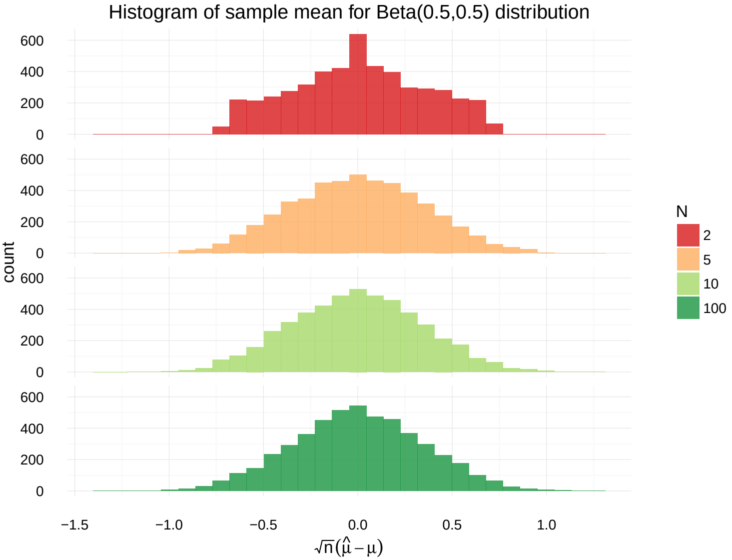
- Properties:
  - If $\hat{\theta} \xrightarrow{d} W$, then $g(\hat{\theta}) \xrightarrow{d} g(W)$ for continuous g (continuous mapping theorem (CMT))
  - Slutsky's theorem: If $\hat{\theta} \xrightarrow{d} W$ and $\hat{\zeta} \xrightarrow{p} c$, then (i) $\hat{\theta} + \hat{\zeta} \xrightarrow{d} W + c$, (ii) $\hat{\theta}\hat{\zeta} \xrightarrow{d} cW$, and (iii) $\hat{\theta}/\hat{\zeta} \xrightarrow{d} W/c$

Demonstration of CLT

```r
N <- c(1,2,5,10,20,50,100)
simulations <- 5000
means <- matrix(0,nrow=simulations,ncol=length(N))
for(i in 1:length(N)) {
  n <- N[i]
  dat <- matrix(runif(n*simulations),
                nrow=simulations, ncol=n)
  means[,i] <- (apply(dat, 1, mean) - 0.5)*sqrt(n)
}

# Plotting
df <- data.frame(means)
df$n <- N
df <- melt(df)
cltPlot <- ggplot(data=df,aes(x=value, fill=variable))
  geom_histogram(alpha=0.2, position="identity") +
  scale_x_continuous(name=expression(sqrt(n)(bar(x)-mu
  scale_fill_brewer(type="div",palette="RdYlGn",
                    name="N",label=N)
cltPlot
```

Demonstration of CLT



Asymptotic normality of OLS

- Bivariate regression of y on x
- Slope:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}) y_i}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}$$

- Consider $\sqrt{n}(\hat{\beta}_1 - \beta_1)$, where β_1 is the population regression coefficient
- Can always write y in terms of the population regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where by construction $\mathrm{E}[\epsilon_i x_i] = 0$

Asymptotic normality of OLS

- Then,

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_1 - \beta_1) =& \sqrt{n}\left(\frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})y_i}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} - \beta_1 \right) \\
=& \sqrt{n}\left(\frac{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \epsilon_i)}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} - \beta_1 \right) \\
=& \frac{\sqrt{n}\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}
\end{aligned}$$

- Already showed that $\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 \xrightarrow{p} \mathrm{Var}(x)$
- Need to apply CLT to $\sqrt{n}\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})\epsilon_i$
  - $\mathrm{E}[(x_i - \bar{x})\epsilon_i] = 0$

Asymptotic normality of OLS

- With homoskedasiticity,

$$\begin{aligned} \mathrm{Var}\left((x_i - \bar{x})\epsilon_i\right) = & \mathrm{E}\left[\mathrm{Var}\left((x_i - \bar{x})\epsilon_i | x\right)\right] + \mathrm{Var}\left(\underbrace{\mathrm{E}[(x_i - \bar{x})\epsilon_i | x]}_{=0}\right) \\ = & \mathrm{E}\left[(x_i - \bar{x})^2 \sigma_\epsilon^2\right] \\ \approx & \mathrm{Var}(x)\sigma_\epsilon^2 \end{aligned}$$

- Can conclude that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (x_i - \bar{x})\epsilon_i \xrightarrow{d} N(0, \mathrm{Var}(x)\sigma_\epsilon^2)$$

Asymptotic normality of OLS

- By Slutsky's theorem,

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) = \frac{\sqrt{n}\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\xrightarrow{d} N\left(0, \frac{\sigma_\epsilon^2}{\mathrm{Var}(x)}\right)$$

or equivalently,

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma_\epsilon^2}{n\mathrm{Var}(x)}}} \xrightarrow{d} N(0,1)$$

Asymptotic normality of OLS

- Again by slutsky's lemma can replace σ_ϵ^2 and $\mathrm{Var}(x)$ by consistent estimators, and

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}_\epsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \xrightarrow{d} N(0,1)$$

  i.e. usual t-statistic is asymptotically normal

- Similar reasoning applies to multivariate regression

Theorem

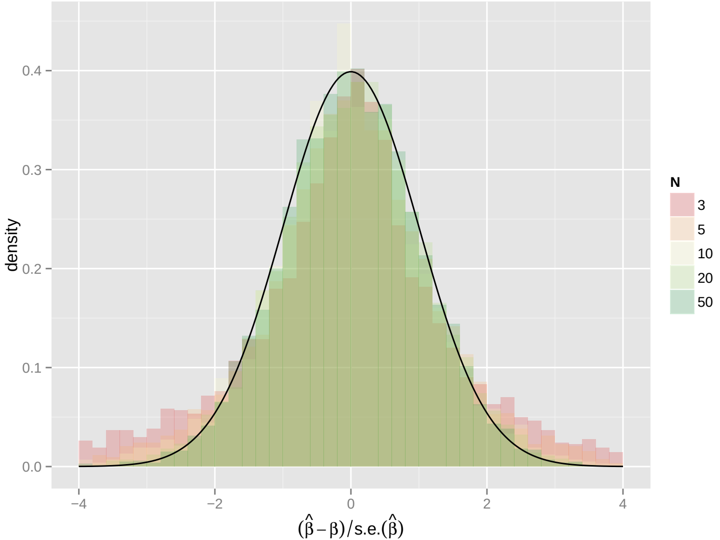
Assume MLR.1-3, MLR.5, and MLR.4': $\mathrm{E}[\epsilon_i x_{i,j}] = 0 \forall j$, then OLS is asymptotically normal with

$$\sqrt{n}\begin{pmatrix}\hat{\beta}_0 - \beta_0 \\ \vdots \\ \hat{\beta}_k - \beta_k\end{pmatrix} \xrightarrow{d} N\left(0, \Sigma\right)$$

and in particular

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\frac{\hat{\sigma}_\epsilon^2}{\sum_{i=1}^n \tilde{x}_{ji}^2}}} \xrightarrow{d} N(0, 1)$$

Demonstration



Code

Large sample inference

- OLS asymptotically normal $\Rightarrow$ in large sample we can use the usual t and F statistics for inference without assuming $\epsilon_i | X \sim N$
- E.g. test $H_0 : \beta_j = \beta_j^*$ against $H_a : \beta_j \neq \beta_j^*$ at significance level α in

$$y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i} + \epsilon_i$$

assuming MLR.1-3, MLR.4', and MLR.5

  - t-statistic:

$$\hat{t} = \frac{\hat{\beta}_j - \beta_j^*}{\sqrt{\frac{\hat{\sigma}_\epsilon^2}{\sum_{i=1}^n \tilde{x}_{ji}^2}}} \xrightarrow{d} N(0,1)$$

  - p-value:

$$p = \mathrm{P}(|t| \geq |\hat{t}|) = 2\Phi(-|\hat{t}|)$$

Large sample inference

- Since $\lim_{n\to\infty} F_{t,n-k-1}(x) = \Phi(x)$ it is also valid to use t-distribution CDF instead of normal distribution CDF
- Reject H_0 if $p < \alpha$
    - Because p-value is based on asymptotic distribution instead of exact finite sample distribution, the test will not exactly have the correct size

$$\mathrm{P}(\text{reject } H_0 \text{ if it is true}) \neq \alpha$$

however it will have the correct size for large samples

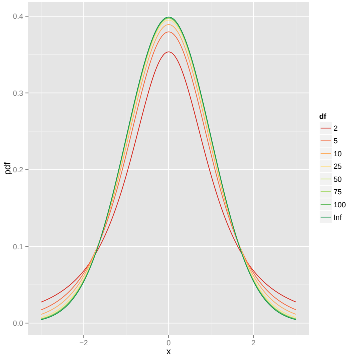
$$\lim_{n\to\infty} \mathrm{P}(\text{reject } H_0 \text{ if it is true}) = \alpha$$

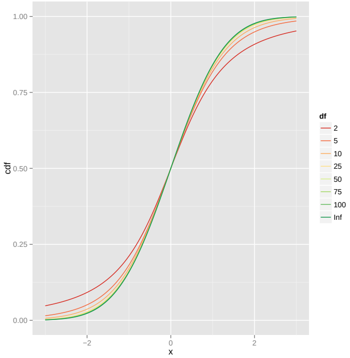
Large sample inference

t distribution as degrees of freedom increases

PDF

CDF





Large sample inference

- E.g. 95% confidence interval for β_j
  - We know

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\frac{\hat{\sigma}_\epsilon^2}{\sum_{i=1}^n \tilde{x}_{ji}^2}}} \xrightarrow{d} N(0,1)$$

so

$$\mathrm{P}\left(\Phi^{-1}(0.025) \leq \frac{\hat{\beta}_j - \beta_j}{\sqrt{\frac{\hat{\sigma}_\epsilon^2}{\sum_{i=1}^n \tilde{x}_{ji}^2}}} \leq \Phi^{-1}(0.975)\right) \to 0.95$$

$$\mathrm{P}\left(\begin{array}{c}\hat{\beta}_j + \sqrt{\frac{\hat{\sigma}_\epsilon^2}{\sum_{i=1}^n \tilde{x}_{ji}^2}}\Phi^{-1}(0.025) \leq \beta_j \leq \\ \leq \hat{\beta}_j + \sqrt{\frac{\hat{\sigma}_\epsilon^2}{\sum_{i=1}^n \tilde{x}_{ji}^2}}\Phi^{-1}(0.975)\end{array}\right) \to 0.95$$

and we can use the same confidence interval as before

$$\hat{\beta}_j \pm s.e.(\hat{\beta}_j)\Phi^{-1}(0.025)$$

Large sample inference

- As above, using $F^{-1}_{t,n-k-1}$ instead of Φ^{-1} is valid
- As above, the confidence interval is only guaranteed to have correct coverage probability in large samples

- E.g. testing $H_0 : \beta_2 = 0$ and $\beta_3 = 0$ against $H_a : \beta_2 \neq 0$ or $\beta_3 \neq 0$ in

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \epsilon_i$$

  - F-statistic (LR version):

$$\hat{F} = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

  where

  - SSR_r = sum of squared residuals from restricted model, i.e. regressing y_i on just $x_{1,i}$
  - SSR_{ur} = sum of squared residuals from unrestricted model, i.e. regressing y_i on $x_{1,i}$, $x_{2,i}$, and $x_{3,i}$

Large sample inference

- $q = 2 =$ number of restrictions

- Asymptotic normality of $\hat{\beta}$ implies

$$qF \xrightarrow{d} \chi^2(q)$$

- Asymptotic p-value:

$$p = \mathrm{P}(F \geq \hat{F}) = 1 - F_{\chi^2(q)}(q\hat{F})$$

where $F_{\chi^2(q)}$ is CDF of $\chi^2(q)$ distribution
    - Since $\lim_{n\to\infty} F_{F(q,n-k-1)}(x) = F_{\chi^2(q)}(qx)$, can use F distribution instead of χ^2

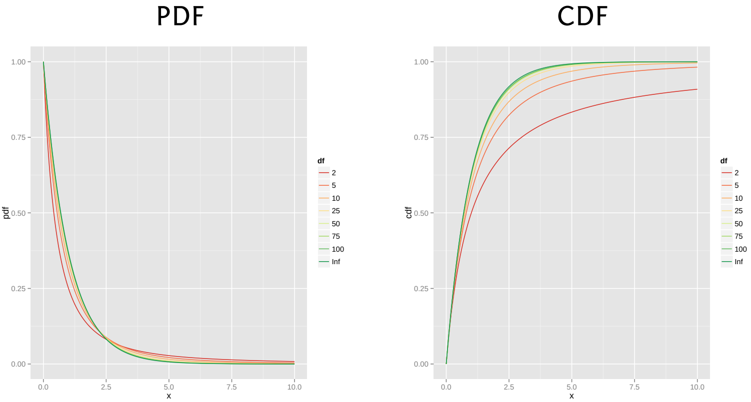
- Same is true for Wald version of F-statistic

$$\hat{F} = \frac{1}{q} \begin{pmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix}^T \begin{pmatrix} \widehat{\mathrm{Var}}(\hat{\beta}_2) & \widehat{\mathrm{Cov}}(\hat{\beta}_2, \hat{\beta}_3) \\ \widehat{\mathrm{Cov}}(\hat{\beta}_2, \hat{\beta}_3) & \widehat{\mathrm{Var}}(\hat{\beta}_3) \end{pmatrix}^{-1} \begin{pmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix}$$

$$= \frac{1}{q} \frac{\hat{\beta}_2^2 \widehat{\mathrm{Var}}(\hat{\beta}_3) + \hat{\beta}_3 \widehat{\mathrm{Var}}(\hat{\beta}_2) - 2\widehat{\mathrm{Cov}}(\hat{\beta}_2, \hat{\beta}_3)\hat{\beta}_2\hat{\beta}_3}{\widehat{\mathrm{Var}}(\hat{\beta}_2)\widehat{\mathrm{Var}}(\hat{\beta}_3) - \widehat{\mathrm{Cov}}(\hat{\beta}_2, \hat{\beta}_3)^2}$$

Large sample inference

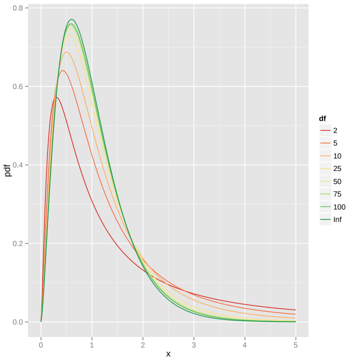
$F(2, df)$ distribution as degrees of freedom increases



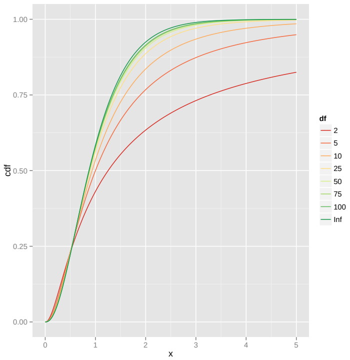
Large sample inference

$F(5, df)$ distribution as degrees of freedom increases

PDF



CDF



F-test in R

```
rm(list=ls())
library(lmtest) ## for lrtest() and waldtest()

k <- 3
n <- 1000
beta <- matrix(c(1,1,0,0), ncol=1)
x <- matrix(rnorm(n*k),nrow=n,ncol=k)
e <- runif(n)*2-1 ## U(-1,1)
y <- cbind(1,x) %*% beta + e

## LR form of F-test
df <- data.frame(y,x)
unrestricted <- lm(y ~ X1 + X2 + X3, data=df)
restricted <- lm(y ~ X1,data=df)
F <- (sum(restricted$residuals^2) -
      sum(unrestricted$residuals^2))/2 /
  (sum(unrestricted$residuals^2)/(n-k-1))
p <- 1-pf(F,2,n-k-1)
```

F-test in R

```r
## or use anova
anova(restricted,unrestricted)
## or lrtest (uses chi2 instead of F distributio
lrtest(unrestricted, restricted)

## Wald form
Fw <- 0.5*coef(unrestricted)[c("X2","X3")] %*%
  solve(vcov(unrestricted)[c("X2","X3"),
                           c("X2","X3")]) %*%
  coef(unrestricted)[c("X2","X3")]
pw <- 1-pf(Fw,2,n-k-1)
## Should have F == Fw and p==pw

## automated Wald test
waldtest(unrestricted, restricted, test="F")
```

Code

References

Angrist, Joshua D and Jörn-Steffen Pischke. 2014. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press.

Menzel, Konrad. 2009. "14.30 Introduction to Statistical Methods in Economics." Massachusetts Institute of Technology: MIT OpenCourseWare. URL `http://ocw.mit.edu/courses/economics/14-30-introduction-to-statistical-methods-in-econom` `lecture-notes/`.

Stock, J.H. and M.W. Watson. 2009. *Introduction to Econometrics, 2/E*. Addison-Wesley.

Wooldridge, J.M. 2013. *Introductory econometrics: A modern approach*. South-Western.