

# 13. Properties of Least Squares

Spring 2023

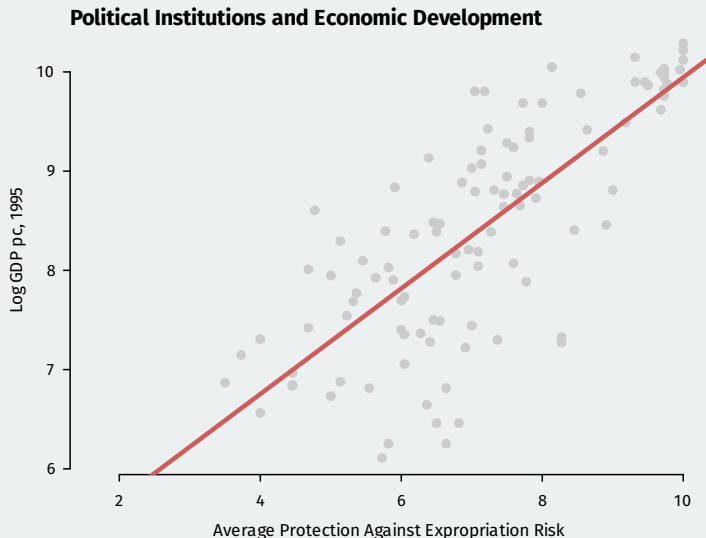
Matthew Blackwell

Gov 2002 (Harvard)

# Where are we? Where are we going?

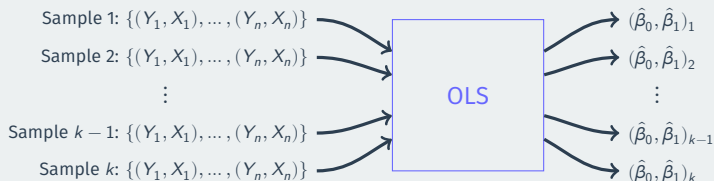
- Before: learned about CEFs and linear projections in the population.
- Last time: OLS estimator, its algebraic properties.
- Now: its statistical properties, both finite-sample and asymptotic.

# Acemoglu, Johnson, and Robinson (2001)



# Sampling distribution of the OLS estimator

- OLS is an estimator—we plug data into and we get out estimates.

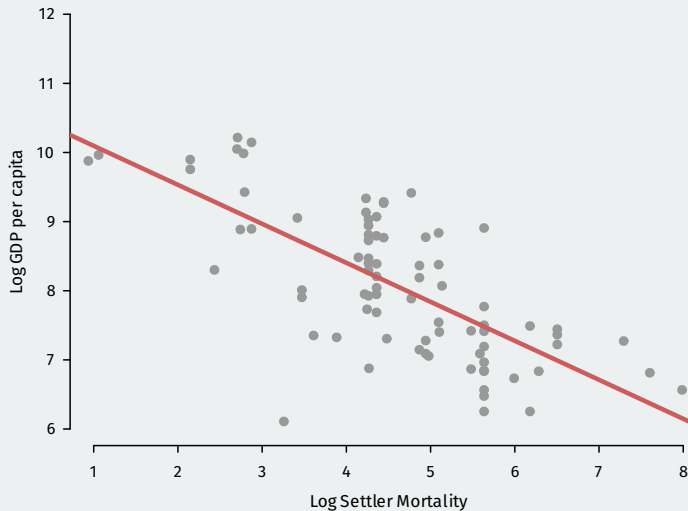


- Just like the sample mean or sample difference in means
- Has a sampling distribution, with a sampling variance/standard error.

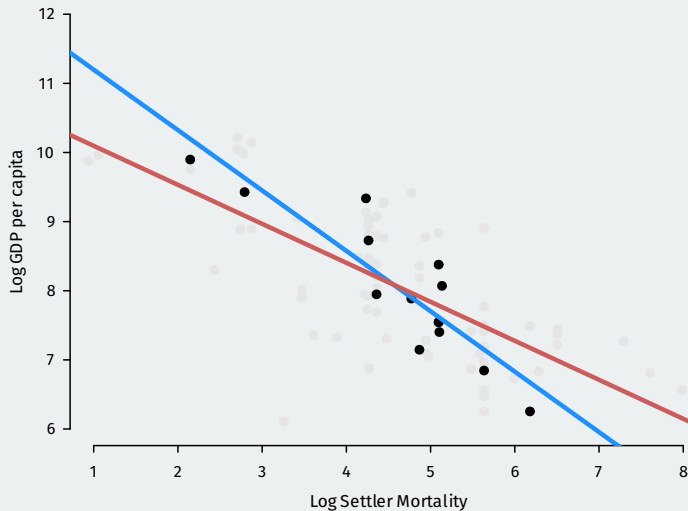
# Simulation procedure

- Let's take a simulation approach to demonstrate:
    - Pretend that the AJR data represents the population of interest
    - See how the line varies from sample to sample
1. Draw a random sample of size  $n = 30$  with replacement using `sample()`
  2. Use `lm()` to calculate the OLS estimates of the slope and intercept
  3. Plot the estimated regression line

# Population Regression



# Randomly sample from AJR



# Big picture

- We want finite-sample guarantees about our estimates.
  - Unbiasedness, exact sampling distribution, etc.
- But finite-sample results come at a price in terms of assumptions.
  - Unbiasedness: CEF is linear.
  - Exact sampling distribution: normal errors.
- Asymptotic results hold under much weaker assumptions, but require more data.
  - OLS consistent for the linear projection even with nonlinear CEF.
  - Asymptotic normality for sampling distribution under mild assumptions.
- Focus on two models:
  - **Linear projection model** for asymptotic results.
  - **Linear regression/CEF model** for finite samples.



# 1/ Linear projection model and Large-sample Properties

# Linear projection model

- We'll start at the most broad, fewest assumptions

## Linear projection model

1. For the variables  $(Y, \mathbf{X})$ , we assume the linear projection of  $Y$  on  $\mathbf{X}$  is defined as:

$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$

$$\mathbb{E}[\mathbf{X}e] = 0.$$

2. The design matrix is invertible, so  $\mathbb{E}[\mathbf{X}_i\mathbf{X}_i'] > 0$  (positive definite).

- Linear projection model holds under **very** mild assumptions.
  - Remember: not even assuming linear CEF!
  - Implies coefficients are  $\boldsymbol{\beta} = (\mathbb{E}[\mathbf{X}\mathbf{X}'])^{-1}\mathbb{E}[\mathbf{X}Y]$
- What properties can we derive under such weak assumptions?

# A very useful decomposition

$$\hat{\boldsymbol{\beta}} = \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i Y_i \right) = \boldsymbol{\beta} + \underbrace{\left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i e_i \right)}_{\text{estimation error}}$$

- OLS estimates are the truth plus some estimation error.
- Most of what we derive about OLS comes from this view.
- Sample means in the estimation error follow the law of large numbers:

$$\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \xrightarrow{p} \mathbb{E}[\mathbf{x}_i \mathbf{x}_i'] \equiv \mathbf{Q}_{\mathbf{xx}} \quad \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i e_i \xrightarrow{p} \mathbb{E}[\mathbf{x}_i e_i] = \mathbf{0}$$

- $\mathbf{Q}_{\mathbf{xx}}$  is invertible by assumption, so by the continuous mapping theorem:

$$\left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \xrightarrow{p} \mathbf{Q}_{\mathbf{xx}}^{-1} \implies \hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta} + \mathbf{Q}_{\mathbf{xx}}^{-1} \cdot \mathbf{0} = \boldsymbol{\beta},$$

# Consistency of OLS

## Theorem (Consistency of OLS)

Under the linear projection model and i.i.d. data,  $\hat{\beta}$  is consistent for  $\beta$ .

- Simple proof, but powerful result.
- OLS consistently estimates the linear projection coefficients,  $\beta$ .
  - No guarantees about what the  $\beta_j$  represent!
  - Best linear approximation to  $\mathbb{E}[Y \mid \mathbf{X}]$ .
  - If we have a linear CEF, then it's consistent for the CEF coefficients.
- Valid with no restrictions on  $Y$ : could be binary, discrete, etc.
- Not guaranteed to be unbiased (unless CEF is linear, as we'll see...)

# Central limit theorem, reminders

- We'll want to approximate the sampling distribution of  $\hat{\boldsymbol{\beta}}$ . CLT!
- Consider some sample mean of i.i.d. data:  $n^{-1} \sum_{i=1}^n g(\mathbf{X}_i)$ . We have:

$$\mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n g(\mathbf{X}_i) \right] = \mathbb{E}[g(\mathbf{X}_i)] \quad \text{var} \left[ \frac{1}{n} \sum_{i=1}^n g(\mathbf{X}_i) \right] = \frac{\text{var}[g(\mathbf{X}_i)]}{n}$$

- CLT implies:

$$\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n g(\mathbf{X}_i) - \mathbb{E}[g(\mathbf{X}_i)] \right) \xrightarrow{d} \mathcal{N}(0, \text{var}[g(\mathbf{X}_i)])$$

- If  $\mathbb{E}[g(\mathbf{X}_i)] = 0$ , then we have

$$\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n g(\mathbf{X}_i) \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n g(\mathbf{X}_i) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[g(\mathbf{X}_i)g(\mathbf{X}_i)'])$$

# Standardized estimator

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i e_i \right)$$

- Remember that  $(n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i')^{-1} \xrightarrow{p} \mathbf{Q}_{\mathbf{X}\mathbf{X}}^{-1}$  so we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx \mathbf{Q}_{\mathbf{X}\mathbf{X}}^{-1} \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i e_i \right)$$

- What about  $n^{-1/2} \sum_{i=1}^n \mathbf{x}_i e_i$ ? Notice that:
  - $n^{-1} \sum_{i=1}^n \mathbf{x}_i e_i$  is a sample average with  $\mathbb{E}[\mathbf{x}_i e_i] = 0$ .
  - Rewrite as  $\sqrt{n}$  times an average of i.i.d. mean-zero random vectors.
- Let  $\boldsymbol{\Omega} = \mathbb{E}[e_i^2 \mathbf{x}_i \mathbf{x}_i']$  and apply the CLT:

$$\left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i e_i \right) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Omega})$$

# Asymptotic normality

## Theorem (Asymptotic Normality of OLS)

Under the linear projection model,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_{\boldsymbol{\beta}}),$$

where,

$$\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{Q}_{\mathbf{XX}}^{-1} \boldsymbol{\Omega} \mathbf{Q}_{\mathbf{XX}}^{-1} = (\mathbb{E}[\mathbf{X}_i \mathbf{X}_i'])^{-1} \mathbb{E}[e_i^2 \mathbf{X}_i \mathbf{X}_i'] (\mathbb{E}[\mathbf{X}_i \mathbf{X}_i'])^{-1}$$

- $\hat{\boldsymbol{\beta}}$  is approximately normal with mean  $\boldsymbol{\beta}$  and variance  $\mathbf{Q}_{\mathbf{XX}}^{-1} \boldsymbol{\Omega} \mathbf{Q}_{\mathbf{XX}}^{-1} / n$
- $\mathbf{V}_{\hat{\boldsymbol{\beta}}} = \mathbf{V}_{\boldsymbol{\beta}} / n$  is the **asymptotic covariance matrix** of  $\hat{\boldsymbol{\beta}}$ 
  - Square root of the diagonal of  $\mathbf{V}_{\hat{\boldsymbol{\beta}}}$  = standard errors for  $\hat{\beta}_j$
- Allows us to formulate (approximate) confidence intervals, tests.

# Estimating OLS variance

$$\mathbf{V}_\beta = \mathbf{Q}_{\mathbf{X}\mathbf{X}}^{-1} \mathbf{\Omega} \mathbf{Q}_{\mathbf{X}\mathbf{X}}^{-1}$$

- Estimation of  $\mathbf{V}_\beta$  uses plug-in estimators.
  - Replace  $\mathbf{Q}_{\mathbf{X}\mathbf{X}} = \mathbb{E}[\mathbf{X}_i \mathbf{X}_i']$  with  $n^{-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i' = \mathbb{X}'\mathbb{X}/n$ .
  - Replace  $\mathbf{\Omega} = \mathbb{E}[e_i^2 \mathbf{X}_i \mathbf{X}_i']$  with  $n^{-1} \sum_{i=1}^n \hat{e}_i^2 \mathbf{X}_i \mathbf{X}_i'$
- Putting these together:

$$\begin{aligned}\widehat{\mathbf{V}}_\beta &= \left( \frac{1}{n} \mathbb{X}'\mathbb{X} \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \hat{e}_i^2 \mathbf{X}_i \mathbf{X}_i' \right) \left( \frac{1}{n} \mathbb{X}'\mathbb{X} \right)^{-1} \\ &= (\mathbb{X}'\mathbb{X})^{-1} \left( \frac{1}{n} \sum_{i=1}^n \hat{e}_i^2 \mathbf{X}_i \mathbf{X}_i' \right) (\mathbb{X}'\mathbb{X})^{-1}\end{aligned}$$

- Straightforward to show this is consistent:  $\widehat{\mathbf{V}}_\beta \xrightarrow{p} \mathbf{V}_\beta$ .
- Square root of the diagonal of  $\widehat{\mathbf{V}}_{\hat{\beta}} = n^{-1} \widehat{\mathbf{V}}_\beta$ :  
**heteroskedasticity-consistent (HC) SEs**

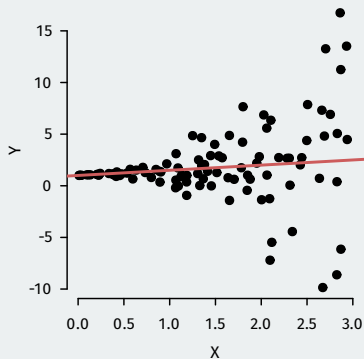


# Homoskedasticity

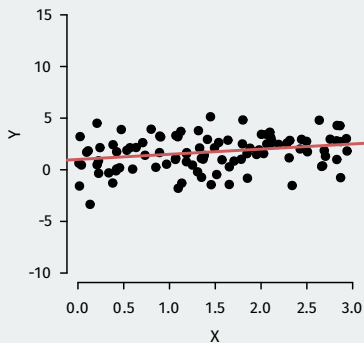
Assumption: Homoskedasticity

The variance of the error terms is constant in  $\mathbf{X}$ ,  $\mathbb{E}[e^2 | \mathbf{X}] = \sigma^2(\mathbf{X}) = \sigma^2$ .

**Heteroskedastic**



**Homoskedastic**



# Consequences of homoskedasticity

- Homoskedasticity implies  $\mathbb{E}[e_i^2 \mathbf{X}_i \mathbf{X}_i'] = \mathbb{E}[e_i^2] \mathbb{E}[\mathbf{X}_i \mathbf{X}_i'] = \sigma^2 \mathbf{Q}_{\mathbf{X}\mathbf{X}}$
- Simplifies the expression for the variance of  $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ :

$$\mathbf{V}_{\boldsymbol{\beta}}^{\text{lm}} = \mathbf{Q}_{\mathbf{X}\mathbf{X}}^{-1} \mathbb{E}[e_i^2] \mathbf{Q}_{\mathbf{X}\mathbf{X}} \mathbf{Q}_{\mathbf{X}\mathbf{X}}^{-1} = \sigma^2 \mathbf{Q}_{\mathbf{X}\mathbf{X}}^{-1}$$

- Estimated variance of  $\hat{\boldsymbol{\beta}}$  under homoskedasticity

$$s^2 = \frac{1}{n-k} \sum_{i=1}^n \hat{e}_i^2 \quad \widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}^{\text{lm}} = \frac{1}{n} s^2 \left( \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i' \right)^{-1} = s^2 (\mathbb{X}' \mathbb{X})^{-1}$$

- LLN implies  $s^2 \xrightarrow{P} \sigma^2$  and so  $n \widehat{\mathbf{V}}_{\hat{\boldsymbol{\beta}}}^{\text{lm}}$  is consistent for  $\mathbf{V}_{\boldsymbol{\beta}}^{\text{lm}}$

# Notes on skedasticity

- Homoskedasticity: strong assumption that isn't needed for consistency.
- Software: almost always reports  $\widehat{\mathbf{V}}_{\hat{\beta}}^{\text{lm}}$  by default.
  - e.g. `lm( )` in R or `reg` in Stata.
- Separate commands for HC SEs  $\widehat{\mathbf{V}}_{\hat{\beta}}$ 
  - Use `{sandwich}` package in R or `, robust` in Stata.
- If  $\widehat{\mathbf{V}}_{\hat{\beta}}^{\text{lm}}$  and  $\widehat{\mathbf{V}}_{\hat{\beta}}$  differ a lot, maybe check modeling assumptions (King and Roberts, PA 2015)
- Lots of “flavors” of HC variance estimators (HC0, HC1, HC2, etc).
  - Mostly small, ad hoc changes to improve finite-sample performance.

# Inference with OLS

- Inference is basically the same as any asymptotically normal estimator.
- Let  $\widehat{V}_{\beta_j}$  be the estimated SE for  $\hat{\beta}_j$ .
  - Square root of  $j$ th diagonal entry:  $\sqrt{[\widehat{\mathbf{V}}_{\hat{\beta}}]_{jj}}$
- Hypothesis test of  $\beta_j = \beta_0$ :

$$\text{general t-statistic} = \frac{\hat{\beta}_j - \beta_0}{\sqrt{\widehat{V}_{\beta_j}}} \quad \text{“usual” t-statistic} = \frac{\hat{\beta}_j}{\sqrt{\widehat{V}_{\beta_j}}}$$

- Use same critical values from the normal as usual  $z_{\alpha/2} = 1.96$ .
- 95% (asymptotic) confidence interval for  $\hat{\beta}_j$ :

$$[\hat{\beta}_j - 1.96 \widehat{V}_{\beta_j}, \hat{\beta}_j + 1.96 \widehat{V}_{\beta_j}]$$

- Software often uses  $t$  critical values instead of normal (we'll see why).

## **2/** Inference for Multiple Parameters

# Inference for interactions

$$m(x, z) = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3$$

- **Partial** or **marginal** effect of  $X$  at  $Z$ :  $\frac{\partial m(x, z)}{\partial x} = \beta_1 + z\beta_3$
- Estimate it by plugging in the estimated coefficients:  $\frac{\partial \widehat{m}(x, z)}{\partial x} = \widehat{\beta}_1 + z\widehat{\beta}_3$
- What if we want the variance of this effect for any value of  $Z$ ?

$$\mathbb{V} \left( \frac{\partial \widehat{m}(x, z)}{\partial x} \right) = \mathbb{V} [\widehat{\beta}_1 + z\widehat{\beta}_3] = \mathbb{V}[\widehat{\beta}_1] + z^2 \mathbb{V}[\widehat{\beta}_3] + 2z \text{cov}[\widehat{\beta}_1, \widehat{\beta}_3]$$

- Use the estimated covariance matrix:

$$\widehat{\mathbb{V}} \left( \frac{\partial \widehat{m}(x, z)}{\partial x} \right) = \widehat{V}_{\beta_1} + z\widehat{V}_{\beta_3} + 2z\widehat{V}_{\beta_1\beta_3}$$

# Tests of multiple coefficients

$$m(X, Z) = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3$$

- What about a test of no effect of  $X$  ever? Involves 2 coefficients:

$$H_0 : \beta_1 = \beta_3 = 0$$

- Alternative:  $H_1 : \beta_1 \neq 0$  or  $\beta_3 \neq 0$
- We would like a test statistic that is large when the null is implausible.
  - What about  $\hat{\beta}_1^2 + \hat{\beta}_3^2$ ?
  - Distribution depends on the variance/covariance of the coefficients.
  - Need to normalize like the t-statistic.

# Alternative test for one coefficient

- Usually t-test of  $H_0 : \beta_j = b_0$  based on the t-statistic:

$$t = \frac{\hat{\beta}_j - b_0}{\widehat{\text{se}}(\hat{\beta}_j)},$$

- Reject when  $|t| > c$  for some critical value  $c$  from the standard normal.
- Equivalent test based rejects when  $t^2 > c^2$

$$t^2 = \frac{(\hat{\beta}_j - b_0)^2}{\mathbb{V}[\hat{\beta}_j]} = \frac{n(\hat{\beta}_j - b_0)^2}{[\mathbf{V}_{\beta}]_{jj}}$$

- Because  $t \xrightarrow{d} \mathcal{N}(0, 1)$ , we'll have  $t^2$  converging to a  $\chi_1^2$  distribution
  - Reminder:  $\chi_k^2$  is the sum of  $k$  squared standard normals.
  - Could get the critical value for  $t^2$  directly from  $\chi_1^2$ .



# Rewriting hypotheses with matrices

- We can rewrite the null hypothesis as  $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{c}$  where,

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \mathbf{c} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

- $\mathbf{L}$  has  $q$  rows or restriction and  $k + 1$  columns (one for each coefficient)
- Estimated version of the constraint:  $\mathbf{L}\hat{\boldsymbol{\beta}}$
- By the Delta method, under the null hypothesis we have

$$\sqrt{n}(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{L}\boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, \mathbf{L}'\mathbf{V}_{\boldsymbol{\beta}}\mathbf{L}).$$

- In this case:

$$\sqrt{n} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_3 \end{pmatrix} \xrightarrow{d} \mathcal{N} \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} [\mathbf{V}_{\boldsymbol{\beta}}]_{[11]} & [\mathbf{V}_{\boldsymbol{\beta}}]_{[13]} \\ [\mathbf{V}_{\boldsymbol{\beta}}]_{[31]} & [\mathbf{V}_{\boldsymbol{\beta}}]_{[33]} \end{bmatrix} \right)$$

- If this covariance matrix were identity, then these would be standard normal and  $\hat{\beta}_1^2 + \hat{\beta}_3^2$  would be  $\chi^2_2$  under the null

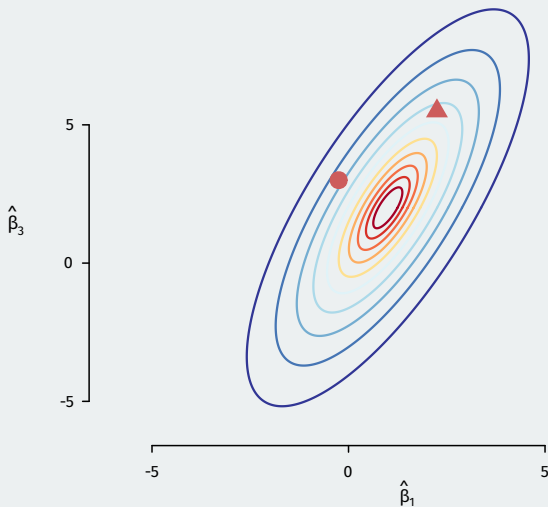
# Wald statistic

- Under the null,  $\sqrt{n}(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c}) \xrightarrow{d} \mathcal{N}(0, \mathbf{L}'\mathbf{V}_{\boldsymbol{\beta}}\mathbf{L})$
- $(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c})'(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c})$  is the squared deviations from the null.
  - Problem: doesn't account for variance/covariance of the estimated coefficients.
- **Wald statistic** normalize by the covariance matrix:

$$W = n(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c})'(\mathbf{L}'\widehat{\mathbf{V}}_{\boldsymbol{\beta}}\mathbf{L})^{-1}(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c})$$

- Similar to dividing by the SE for the t-test
- Squared distance of observed values from the null, weighted by the distribution of the parameters under the null

# Weighting by the distribution



# Wald test

$$W = n (\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c})' (\mathbf{L}'\widehat{\mathbf{V}}_{\boldsymbol{\beta}}\mathbf{L})^{-1} (\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{c})$$

- Asymptotically under the null  $W \xrightarrow{d} \chi_q^2$  where  $q$  is rows of  $\mathbf{L}$ 
  - $q$  is the number of linear restrictions in the null
- **Wald test:** reject when  $W > w_\alpha$ , where  $\mathbb{P}(W > w_\alpha) = \alpha$  under the null.
  - Use  $\chi_q^2$  distribution for critical values, p-values
- Typical software output: **F-statistic**  $F = W/q$ 
  - p-values and critical values come from  $F$  distribution with  $q$  and  $n - k - 1$  dfs.
  - As  $n \rightarrow \infty$ ,  $F_{q, n-k-1} \xrightarrow{d} \chi_q^2$  so asymptotically similar to Wald under homoskedasticity (slightly more conservative).
  - No justification for  $F$  test under heteroskedasticity.
  - “Usual” F-test reports test of all coef = 0 except intercept (pointless?)

# Wald test steps

1. Choose a Type I error rate,  $\alpha$ .
  - Same interpretation: rate of false positives you are willing to accept
2. Calculate the rejection region for the test (one-sided)
  - Rejection region is the region  $W > w_\alpha$  such that  $\mathbb{P}(W > w_\alpha) = \alpha$
  - We can get this from R using the `qchisq()` function
3. Reject if observed statistic is bigger than critical value
  - Use `pchisq()` to get p-values if needed.
  - When applied to a single coefficient, equivalent to a t-test.
  - Use packages like `{aod}` or `{clubSandwich}` in R.

# Multiple testing

- Separate t-tests for each  $\beta_j$ :  $\alpha$  of them will be significant by chance.
- Illustration:
  - Randomly draw 21 variables independently.
  - Run a regression of the first variable on the rest.
- By design, no effect of any variable on any other.

# Multiple test example

```
noise <- data.frame(matrix(rnorm(2100), nrow = 100, ncol = 21))
summary(lm(noise))
```

```
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.028039   0.113820  -0.25   0.8061
## X2          -0.150390   0.112181  -1.34   0.1839
## X3           0.079158   0.095028   0.83   0.4074
## X4          -0.071742   0.104579  -0.69   0.4947
## X5           0.172078   0.114002   1.51   0.1352
## X6           0.080852   0.108341   0.75   0.4577
## X7           0.102913   0.114156   0.90   0.3701
## X8          -0.321053   0.120673  -2.66   0.0094 **
## X9          -0.053122   0.107983  -0.49   0.6241
## X10          0.180105   0.126443   1.42   0.1583
## X11          0.166386   0.110947   1.50   0.1377
## X12          0.008011   0.103766   0.08   0.9387
## X13          0.000212   0.103785   0.00   0.9984
## X14          -0.065969   0.112214  -0.59   0.5583
## X15          -0.129654   0.111575  -1.16   0.2487
## X16          -0.054446   0.125140  -0.44   0.6647
## X17          0.004335   0.112012   0.04   0.9692
## X18          -0.080796   0.109853  -0.74   0.4642
## X19          -0.085806   0.118553  -0.72   0.4713
## X20          -0.186006   0.104560  -1.78   0.0791 .
## X21          0.002111   0.108118   0.02   0.9845
## ---
## Signif. codes:
##  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.999 on 79 degrees of freedom
## Multiple R-squared:  0.201, Adjusted R-squared: -0.00142
## F-statistic: 0.993 on 20 and 79 DF,  p-value: 0.48
```

# Multiple testing gives false positives

- 1 out of 20 variables significant at  $\alpha = 0.05$
- 2 out of 20 variables significant at  $\alpha = 0.1$
- Exactly the number of false positives we would expect.
- But notice the F-statistic: the variables are not **jointly** significant
- **Bonferroni correction**: use p-value cutoff  $\alpha/m$  where  $m$  is the number of hypotheses.
  - Example:  $0.05/20 = 0.0025$
  - Ensures that the family-wise error rate (probability of making at least 1 Type I error) is less than  $\alpha$ .



# 3/ Linear Regression Model and Finite-sample Properties

# Standard linear regression model

- Standard textbook model: **correctly specified linear CEF**
  - Designed for finite-sample results.

## Assumption: Linear Regression Model

1. The variables  $(Y, \mathbf{X})$  satisfy the the linear CEF assumption.

$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$

$$\mathbb{E}[e \mid \mathbf{X}] = 0.$$

2. The design matrix is invertible  $\mathbb{E}[\mathbf{X}\mathbf{X}'] > 0$  (positive definite).

- Basically this assumes the CEF of  $Y$  given  $\mathbf{X}$  is linear.
- We continue to maintain  $\{(Y_i, \mathbf{X}_i)\}$  are i.i.d.

# Properties of OLS under linear CEF

- Linear CEFs imply stronger finite-sample guarantees:

1. **Unbiasedness:**  $\mathbb{E}[\hat{\beta} \mid \mathbb{X}] = \hat{\beta}$

2. **Conditional sampling variance:** let  $\sigma_i^2 = \mathbb{E}[e_i^2 \mid \mathbf{X}_i]$

$$\mathbb{V}[\hat{\beta} \mid \mathbb{X}] = (\mathbb{X}'\mathbb{X})^{-1} \left( \sum_{i=1}^n \sigma_i^2 \mathbf{x}_i \mathbf{x}_i' \right) (\mathbb{X}'\mathbb{X})^{-1}$$

- Useful when linearity holds by default (discrete  $X$  in experiments, etc)

# Linear CEF under homoskedasticity

- Under homoskedasticity, we have a few other finite-sample results:
3. **Conditional sampling variance:**  $\mathbb{V}[\hat{\beta} \mid \mathbb{X}] = \sigma^2 (\mathbb{X}'\mathbb{X})^{-1}$
  4. **Unbiased variance estimator:**  $\mathbb{E} [\hat{\mathbb{V}}^0[\hat{\beta}] \mid \mathbf{X}] = \sigma^2 (\mathbb{X}'\mathbb{X})^{-1}$
  5. **Gauss-Markov:** OLS is the best linear unbiased estimator of  $\beta$  (BLUE). If  $\tilde{\beta}$  is a linear estimator,

$$\mathbb{V}[\tilde{\beta} \mid \mathbb{X}] \geq \mathbb{V}[\hat{\beta} \mid \mathbb{X}] = \sigma^2 (\mathbb{X}'\mathbb{X})^{-1}$$

- For matrices,  $\mathbf{A} \geq \mathbf{B}$  means that  $\mathbf{A} - \mathbf{B}$  is positive semidefinite.
- A matrix  $\mathbf{C}$  is p.s.d. if  $\mathbf{x}'\mathbf{C}\mathbf{x} \geq 0$ .
- Upshot: OLS will have the smaller SEs than any other linear estimator.

# Normal regression model

- Most parametric:  $Y \sim \mathcal{N}(\mathbf{X}'\boldsymbol{\beta}, \sigma^2)$ .
  - Normal error model since  $e = Y - \mathbf{X}'\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma^2)$ .
- Rarely believed, but allows for exact inference for all  $n$ .
  - $(\hat{\beta}_j - \beta_j)/\widehat{\text{se}}(\hat{\beta}_j)$  follows a  $t$  distribution with  $n - k$  degrees of freedom.
  - $F$  statistics follows  $F$  distribution exactly rather than approximately.
- Software often implicitly assumes this for p-values.
- With reasonable  $n$ , asymptotic normality has the same effect.