15. Properties of Least Squares

Spring 2021

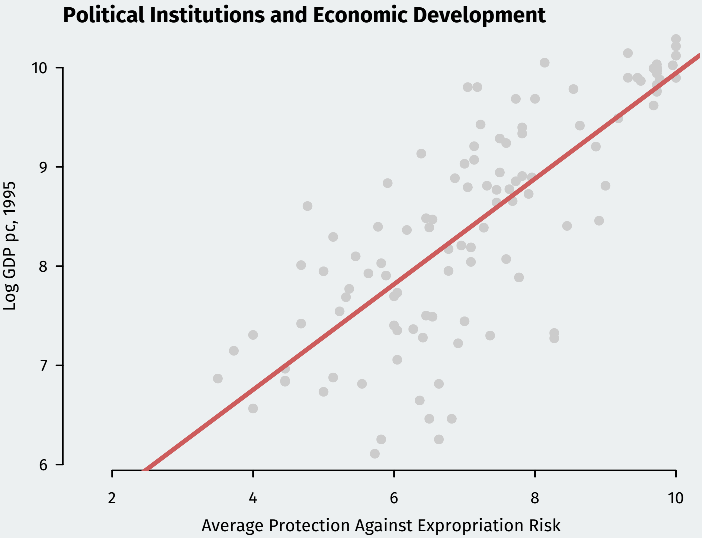
Matthew Blackwell

Gov 2002 (Harvard)

Where are we? Where are we going?

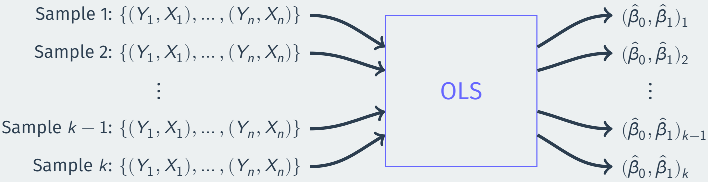
- Before: learned about CEFs and linear projections in the population.
- Last time: OLS estimator, its algebraic properties.
- Now: its statistical properties, both finite-sample and asymptotic.

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.

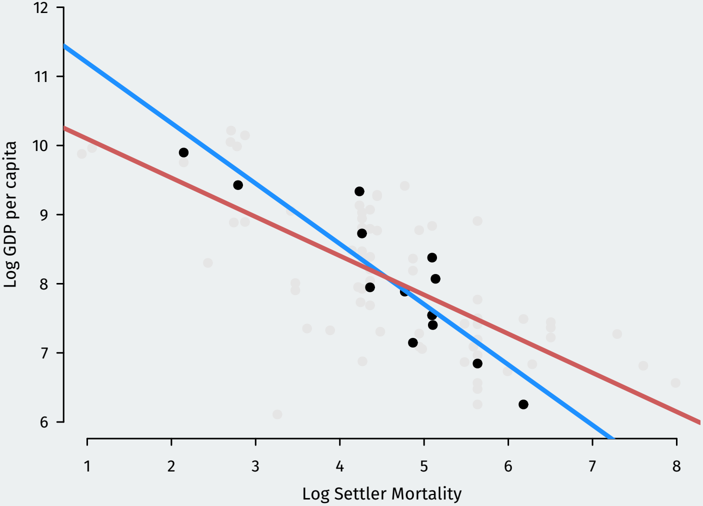
Simulation procedure

- 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

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

- 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:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i' \xrightarrow{p} \mathbb{E}[\mathbf{X}_i\mathbf{X}_i'] \equiv \mathbf{Q}_{\mathbf{XX}} \qquad \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_ie_i \xrightarrow{p} \mathbb{E}[\mathbf{X}e] = \mathbf{0}$$

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

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

Consistency of OLS

Theorem (Consistency of OLS)

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

- Simple proof, but powerful result.
- OLS consistently estimates the linear projection coefficients, $\boldsymbol{\beta}$.
  - No guarantees about what the β_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:

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

- 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

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

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

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

- 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,

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

where,

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

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

Estimating OLS variance

$$\mathbb{V}[\hat{\boldsymbol{\beta}}] = \frac{1}{n}\mathbf{V}_{\hat{\boldsymbol{\beta}}} = \mathbf{Q}_{\mathbf{XX}}^{-1}\boldsymbol{\Omega}\mathbf{Q}_{\mathbf{XX}}^{-1}$$

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

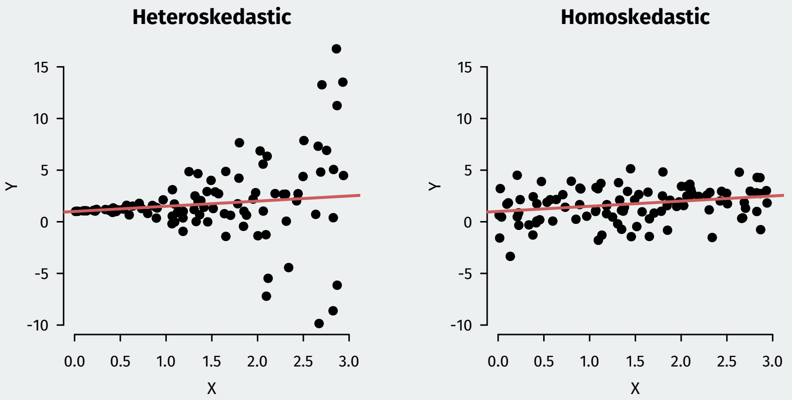
$$\begin{aligned}\widehat{\mathbb{V}}[\hat{\boldsymbol{\beta}}] &= \frac{1}{n}\left(\frac{1}{n}\mathbb{X}'\mathbb{X}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i^2\mathbf{X}_i\mathbf{X}_i'\right)\left(\frac{1}{n}\mathbb{X}'\mathbb{X}\right)^{-1} \\ &= \left(\mathbb{X}'\mathbb{X}\right)^{-1}\left(\sum_{i=1}^{n}\hat{e}_i^2\mathbf{X}_i\mathbf{X}_i'\right)\le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

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

Homoskedasticity

Assumption: Homoskedasticity

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



Consequences of homoskedasticity

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

$$\mathbf{V}_{\hat{\boldsymbol{\beta}}}^0 = \mathbf{Q}_{\mathbf{XX}}^{-1}\mathbb{E}[e_i^2]\mathbf{Q}_{\mathbf{XX}}\mathbf{Q}_{\mathbf{XX}}^{-1} = \sigma^2\mathbf{Q}_{\mathbf{XX}}^{-1}$$

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

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n}\hat{e}_i^2 \qquad \widehat{\mathbb{V}}^0[\hat{\boldsymbol{\beta}}] = \frac{1}{n}s^2\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1} = s^2\left(\mathbb{X}'\mathbb{X}\right)^{-1}$$

- LLN implies $s^2 \xrightarrow{p} \sigma^2$ and so $n\widehat{\mathbb{V}}^0[\hat{\boldsymbol{\beta}}]$ is consistent for $\mathbf{V}_{\hat{\boldsymbol{\beta}}}^0$

Notes on skedasticity

- Homoskedasticity: strong assumption that isn't needed for consistency.
- Software: almost always reports $\widehat{\mathbb{V}}^0[\hat{\boldsymbol{\beta}}]$ by default.
    - e.g. `lm()` in R or `reg` in Stata.
- Separate commands for HC SEs $\widehat{\mathbb{V}}[\hat{\boldsymbol{\beta}}]$
    - Use **`{sandwich}`** package in R or **`,robust`** in Stata.
- If $\widehat{\mathbb{V}}^0[\hat{\boldsymbol{\beta}}]$ and $\widehat{\mathbb{V}}[\hat{\boldsymbol{\beta}}]$ 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 jth diagonal entry: $\sqrt{[\widehat{\mathbb{V}}[\hat{\boldsymbol{\beta}}]]_{jj}}$
- Hypothesis test of $\beta_j = \beta_0$:

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

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

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

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

2/ Inference for Multiple Parameters

Inference for interactions

$$\mathbb{L}[Y|\mathbf{X}] = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3$$

- **Partial** or **marginal** effect of X at Z: $\frac{\partial \mathbb{L}[Y|\mathbf{X}]}{\partial X} = \beta_1 + Z\beta_3$
- Estimate it by plugging in the estimated coefficients: $\frac{\partial \widehat{\mathbb{L}[Y|\mathbf{X}]}}{\partial X} = \hat{\beta}_1 + Z\hat{\beta}_3$
- What if we want the variance of this effect for any value of Z?

$$\mathbb{V}\left(\frac{\partial \widehat{\mathbb{L}[Y|\mathbf{X}]}}{\partial X}\right) = \mathbb{V}\left[\hat{\beta}_1 + Z\hat{\beta}_3\right] = \mathbb{V}[\hat{\beta}_1] + Z^2\mathbb{V}[\hat{\beta}_3] + 2Z\text{cov}[\hat{\beta}_1, \hat{\beta}_3]$$

- Use the estimated covariance matrix:

$$\hat{\mathbb{V}}\left(\frac{\partial \widehat{\mathbb{L}[Y|\mathbf{X}]}}{\partial X}\right) = \widehat{V}_{\beta_1} + Z\widehat{V}_{\beta_3} + 2Z\widehat{V}_{\beta_1\beta_3}$$

Tests of multiple coefficients

$$\mathbb{L}[Y|\mathbf{X}] = \beta_0 + X\beta_1 + Z\beta_2 + XZ\beta_3$$

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

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

- 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.
    - Compare model fit when we do and do not impose the null hypothesis.

Restricted and unrestricted models

- Estimates and SSR from **unrestricted model** (alternative is true):

$$\widehat{Y}_i = \hat{\beta}_0 + X_i\hat{\beta}_1 + Z_i\hat{\beta}_2 + X_iZ_i\hat{\beta}_3 \qquad SSR_u = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

- Estimates and SSR from **restricted model** (null is true):

$$\widetilde{Y}_i = \tilde{\beta}_0 + Z_i\tilde{\beta}_2 \qquad SSR_r = \sum_{i=1}^{n}(Y_i - \widetilde{Y}_i)^2$$

- If the null is true, we'd expect SSR_u to not be much bigger than SSR_r.
    - SSR mechanically increases even if you add noise, but not that much.

F statistic

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

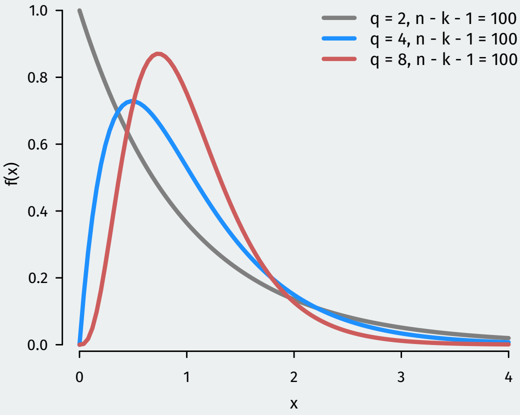
- Components:
  - $(SSR_r - SSR_u)$: increase in residual variation dropping β_1, β_3
  - $q =$ number of restrictions (numerator degrees of freedom)
  - $n-k-1$: denominator/unrestricted degrees of freedom
- Intuition:

$$\frac{\text{increase in prediction error}}{\text{original prediction error}}$$

- Under the null, F is just sampling noise!
- Asymptotic distribution of F:

$$\frac{(SSR_r - SSR_u)/q}{SSR_u/(n-k-1)} \xrightarrow{d} F_{q,n-(k+1)}$$

F distribution



- Ratio of two χ^2 (Chi-squared) distributions

F-test steps

1. Choose a Type I error rate, α.
    - 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 $F > c$ such that $\mathbb{P}(F > c) = \alpha$
    - We can get this from R using the `qf()` function
3. Reject if observed statistic is bigger than critical value
    - Use `pf()` to get p-values if needed.

- When applied to a single coefficient, equivalent to a t-test.

F statistic for all variables

- **"The" F-test**: null of all coefficients except the intercept being 0.
- In that case, the restricted model is just:

$$Y = \beta_0 + e$$

- And the estimate here would just be sample mean ($\hat{\beta}_0 = \overline{Y}$)
- The SSR_r then would just be the total sum of squares in Y:

$$SSR_r = SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

- Often reported with regression output.

Multiple testing

- Separate t-tests for each β_j: α 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

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

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

- 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**: $\mathbb{V}[\hat{\boldsymbol{\beta}} \mid \mathbb{X}] = \sigma^2 \left(\mathbb{X}'\mathbb{X}\right)^{-1}$
4. **Unbiased variance estimator**: $\mathbb{E}\left[\widehat{\mathbb{V}}^0[\hat{\boldsymbol{\beta}}] \mid \mathbf{X}\right] = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}$
5. **Gauss-Markov**: OLS is the best linear unbiased estimator of $\boldsymbol{\beta}$ (BLUE). If $\tilde{\boldsymbol{\beta}}$ is a linear estimator,

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

  - 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.