Week 7: Multiple Regression

Brandon Stewart[1]

Princeton

October 22, 24, 2018

[1] These slides are heavily influenced by Matt Blackwell, Adam Glynn, Justin Grimmer, Jens Hainmueller and Erin Hartman.

Where We've Been and Where We're Going...

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - Monday:
    - matrix form of linear regression
    - t-tests, F-tests and general linear hypothesis tests
  - Wednesday:
    - problems with *p*-values
    - agnostic regression
    - the bootstrap
- Next Week
  - break!
  - then . . . diagnostics
- Long Run
  - probability → inference → regression → causal inference

Questions?

1. Matrix Form of Regression
2. OLS inference in matrix form
3. Standard Hypothesis Tests
4. Testing Joint Significance
5. Testing Linear Hypotheses: The General Case
6. Fun With(out) Weights
7. Appendix: Derivations and Consistency
8. The Problems with *p*-values
9. Agnostic Regression
10. Inference via the Bootstrap
11. Fun With Weights
12. Appendix: Tricky *p*-value Example

The Linear Model with New Notation

- Remember that we wrote the linear model as the following for all $i \in [1, \ldots, n]$:

$$y_i = \beta_0 + x_i\beta_1 + z_i\beta_2 + u_i$$

- Imagine we had an n of 4. We could write out each formula:

$$\begin{aligned} y_1 &= \beta_0 + x_1\beta_1 + z_1\beta_2 + u_1 \quad \text{(unit 1)} \\ y_2 &= \beta_0 + x_2\beta_1 + z_2\beta_2 + u_2 \quad \text{(unit 2)} \\ y_3 &= \beta_0 + x_3\beta_1 + z_3\beta_2 + u_3 \quad \text{(unit 3)} \\ y_4 &= \beta_0 + x_4\beta_1 + z_4\beta_2 + u_4 \quad \text{(unit 4)} \end{aligned}$$

The Linear Model with New Notation

$$\begin{aligned}
y_1 &= \beta_0 + x_1\beta_1 + z_1\beta_2 + u_1 \quad \text{(unit 1)}\\
y_2 &= \beta_0 + x_2\beta_1 + z_2\beta_2 + u_2 \quad \text{(unit 2)}\\
y_3 &= \beta_0 + x_3\beta_1 + z_3\beta_2 + u_3 \quad \text{(unit 3)}\\
y_4 &= \beta_0 + x_4\beta_1 + z_4\beta_2 + u_4 \quad \text{(unit 4)}
\end{aligned}$$

- We can write this as:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \beta_1 + \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} \beta_2 + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix}$$

- Outcome is a linear combination of the the **x**, **z**, and **u** vectors

Grouping Things into Matrices

- Can we write this in a more compact form?
  Yes! Let **X** and β be the following:

$$\underset{(4\times 3)}{\mathbf{X}} = \begin{bmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ 1 & x_3 & z_3 \\ 1 & x_4 & z_4 \end{bmatrix} \quad \underset{(3\times 1)}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

Back to Regression

- $\mathbf{X}$ is the $n \times (k+1)$ design matrix of independent variables
- $\boldsymbol{\beta}$ be the $(k+1) \times 1$ column vector of coefficients.
- $\mathbf{X}\boldsymbol{\beta}$ will be $n \times 1$:

$$\mathbf{X}\boldsymbol{\beta} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \cdots + \beta_k \mathbf{x}_k$$

- We can compactly write the linear model as the following:

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times 1)}{\mathbf{X}\boldsymbol{\beta}} + \underset{(n\times 1)}{\mathbf{u}}$$

- We can also write this at the individual level, where $\mathbf{x}_i'$ is the ith row of $\mathbf{X}$:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

Multiple Linear Regression in Matrix Form

- Let $\widehat{\boldsymbol{\beta}}$ be the matrix of estimated regression coefficients and $\widehat{\mathbf{y}}$ be the vector of fitted values:

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{bmatrix} \qquad \widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

- It might be helpful to see this again more written out:

$$\widehat{\mathbf{y}} = \begin{bmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_n \end{bmatrix} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 1\widehat{\beta}_0 + x_{11}\widehat{\beta}_1 + x_{12}\widehat{\beta}_2 + \cdots + x_{1K}\widehat{\beta}_k \\ 1\widehat{\beta}_0 + x_{21}\widehat{\beta}_1 + x_{22}\widehat{\beta}_2 + \cdots + x_{2K}\widehat{\beta}_k \\ \vdots \\ 1\widehat{\beta}_0 + x_{n1}\widehat{\beta}_1 + x_{n2}\widehat{\beta}_2 + \cdots + x_{nK}\widehat{\beta}_k \end{bmatrix}$$

Residuals

- We can easily write the residuals in matrix form:

$$\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\beta}$$

- Our goal as usual is to minimize the sum of the squared residuals, which we saw earlier we can write:

$$\widehat{\mathbf{u}}'\widehat{\mathbf{u}} = (\mathbf{y} - \mathbf{X}\widehat{\beta})'(\mathbf{y} - \mathbf{X}\widehat{\beta})$$

OLS Estimator in Matrix Form

- Goal: minimize the sum of the squared residuals
- Take (matrix) derivatives, set equal to 0 (see Appendix)
- Resulting first order conditions:

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = 0$$

- Rearranging:

$$\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

- In order to isolate $\widehat{\boldsymbol{\beta}}$, we need to move the $\mathbf{X}'\mathbf{X}$ term to the other side of the equals sign.
- We've learned about matrix multiplication, but what about matrix "division"?

Back to OLS

- Let's assume, for now, that the inverse of $\mathbf{X}'\mathbf{X}$ exists
- Then we can write the OLS estimator as the following:

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- "ex prime ex inverse ex prime y" sear it into your soul.

Intuition for the OLS in Matrix Form

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- What's the intuition here?
- "Numerator" $\mathbf{X}'\mathbf{y}$: is approximately composed of the covariances between the columns of $\mathbf{X}$ and $\mathbf{y}$
- "Denominator" $\mathbf{X}'\mathbf{X}$ is approximately composed of the sample variances and covariances of variables within $\mathbf{X}$
- Thus, we have something like:

$$\widehat{\boldsymbol{\beta}} \approx (\text{variance of } \mathbf{X})^{-1}(\text{covariance of } \mathbf{X} \text{ \& } \mathbf{y})$$

i.e. analogous to the simple linear regression case!

Disclaimer: the final equation is exactly true for all non-intercept coefficients if you remove the intercept from $\mathbf{X}$ such that $\hat{\boldsymbol{\beta}}_{-0} = \text{Var}(\mathbf{X}_{-0})^{-1}\text{Cov}(\mathbf{X}_{-0}, \mathbf{y})$. The numerator and denominator are the variances and covariances if $\mathbf{X}$ and $\mathbf{y}$ are demeaned and normalized by the sample size minus 1.

1 Matrix Form of Regression

2 OLS inference in matrix form

3 Standard Hypothesis Tests

4 Testing Joint Significance

5 Testing Linear Hypotheses: The General Case

6 Fun With(out) Weights

7 Appendix: Derivations and Consistency

8 The Problems with *p*-values

9 Agnostic Regression

10 Inference via the Bootstrap

11 Fun With Weights

12 Appendix: Tricky *p*-value Example

OLS Assumptions in Matrix Form

1. Linearity: $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$
2. Random/iid sample: $(y_i, \mathbf{x}_i')$ are a iid sample from the population.
3. No perfect collinearity: $\mathbf{X}$ is an $n \times (k+1)$ matrix with rank $k+1$
4. Zero conditional mean: $\mathbb{E}[\mathbf{u}|\mathbf{X}] = \mathbf{0}$
5. Homoskedasticity: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
6. Normality: $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_n)$

Assumption 3: No Perfect Collinearity

Definition (Rank)

The rank of a matrix is the maximum number of linearly independent columns.

- In matrix form: $\mathbf{X}$ is an $n \times (k+1)$ matrix with rank $k+1$
- If $\mathbf{X}$ has rank $k+1$, then all of its columns are linearly independent
- …and none of its columns are linearly dependent implies no perfect collinearity
- $\mathbf{X}$ has rank $k+1$ and thus $(\mathbf{X}'\mathbf{X})$ is invertible
- Just like variation in X led us to be able to divide by the variance in simple OLS

Assumption 5: Homoskedasticity

- The stated homoskedasticity assumption is: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
- To really understand this we need to know what $\text{var}(\mathbf{u}|\mathbf{X})$ is in full generality.
- The variance of a vector is actually a matrix:

$$\text{var}[\mathbf{u}] = \Sigma_u = \begin{bmatrix} \text{var}(u_1) & \text{cov}(u_1, u_2) & \dots & \text{cov}(u_1, u_n) \\ \text{cov}(u_2, u_1) & \text{var}(u_2) & \dots & \text{cov}(u_2, u_n) \\ \vdots & & \ddots & \\ \text{cov}(u_n, u_1) & \text{cov}(u_n, u_2) & \dots & \text{var}(u_n) \end{bmatrix}$$

- This matrix is always symmetric since $\text{cov}(u_i, u_j) = \text{cov}(u_j, u_i)$ by definition.

Assumption 5: The Meaning of Homoskedasticity

- What does $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$ mean?
- $\mathbf{I}_n$ is the $n \times n$ identity matrix, σ_u^2 is a scalar.
- Visually:

$$\text{var}[\mathbf{u}] = \sigma_u^2 \mathbf{I}_n = \begin{bmatrix} \sigma_u^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_u^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_u^2 \end{bmatrix}$$

- In less matrix notation:
  - $\text{var}(u_i) = \sigma_u^2$ for all i (constant variance)
  - $\text{cov}(u_i, u_j) = 0$ for all $i \neq j$ (implied by iid)

Unbiasedness of $\hat{\beta}$

Is $\hat{\beta}$ still unbiased under assumptions 1-4? Does $E[\hat{\beta}] = \beta$?

$$\begin{aligned}
\hat{\beta} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \text{ (linearity and no collinearity)} \\
\hat{\beta} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) \\
\hat{\beta} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u} \\
\hat{\beta} &= \mathbf{I}\beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u} \\
\hat{\beta} &= \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u} \\
E[\hat{\beta}|\mathbf{X}] &= E[\beta|\mathbf{X}] + E[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
E[\hat{\beta}|\mathbf{X}] &= \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E[\mathbf{u}|\mathbf{X}] \\
E[\hat{\beta}|\mathbf{X}] &= \beta \text{ (zero conditional mean)}
\end{aligned}$$

So, yes!

A Much Shorter Proof of Unbiasedness of $\hat{\boldsymbol{\beta}}$

A shorter (but less helpful later) proof of unbiasedness,

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= E[\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}] \text{ (definition of the estimator)} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \text{ (expectation of y)} \\ &= \boldsymbol{\beta} \end{aligned}$$

Now we know the sampling distribution is centered on β we want to derive the variance of the sampling distribution conditional on X.

Rule: Variance of Linear Function of Random Vector

Recall that for a linear transformation of a random variable X we have $V[aX + b] = a^2 V[X]$ with constants a and b.

We will need an analogous rule for linear functions of random vectors.

Definition (Variance of Linear Transformation of Random Vector)

Let $f(\mathbf{u}) = \mathbf{A}\mathbf{u} + \mathbf{B}$ be a linear transformation of a random vector $\mathbf{u}$ with non-random vectors or matrices $\mathbf{A}$ and $\mathbf{B}$. Then the variance of the transformation is given by:

$$V[f(\mathbf{u})] = V[\mathbf{A}\mathbf{u} + \mathbf{B}] = \mathbf{A}V[\mathbf{u}]\mathbf{A}' = \mathbf{A}\Sigma_{\mathbf{u}}\mathbf{A}'$$

Conditional Variance of $\hat{\beta}$

$\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ and $E[\hat{\beta}|\mathbf{X}] = \beta + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = \beta$ so the OLS estimator is a linear function of the errors. Thus:

$$\begin{aligned}
V[\hat{\beta}|\mathbf{X}] &= V[\beta|\mathbf{X}] + V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V[\mathbf{u}|\mathbf{X}]((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \quad (\mathbf{X} \text{ is nonrandom given } \mathbf{X}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V[\mathbf{u}|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad \text{(by homoskedasticity)} \\
&= \sigma^2\mathbf{I}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

This gives the $(k+1) \times (k+1)$ variance-covariance matrix of $\hat{\beta}$.

To estimate $V[\hat{\beta}|\mathbf{X}]$, we replace σ^2 with its unbiased estimator $\hat{\sigma}^2$, which is now written using matrix notation as:

$$\hat{\sigma}^2 = \frac{\sum_i \hat{u}_i^2}{n-(k+1)} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-(k+1)}$$

Sampling Variance for $\hat{\beta}$

Under assumptions 1-5, the variance-covariance matrix of the OLS estimators is given by:

$$V[\hat{\beta}|\mathbf{X}] = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} =$$

$$\begin{array}{c|cccccc}
 & \widehat{\beta}_0 & \widehat{\beta}_1 & \widehat{\beta}_2 & \cdots & \widehat{\beta}_k \\
\hline
\widehat{\beta}_0 & V[\widehat{\beta}_0] & \text{Cov}[\widehat{\beta}_0, \widehat{\beta}_1] & \text{Cov}[\widehat{\beta}_0, \widehat{\beta}_2] & \cdots & \text{Cov}[\widehat{\beta}_0, \widehat{\beta}_k] \\
\widehat{\beta}_1 & \text{Cov}[\widehat{\beta}_0, \widehat{\beta}_1] & V[\widehat{\beta}_1] & \text{Cov}[\widehat{\beta}_1, \widehat{\beta}_2] & \cdots & \text{Cov}[\widehat{\beta}_1, \widehat{\beta}_k] \\
\widehat{\beta}_2 & \text{Cov}[\widehat{\beta}_0, \widehat{\beta}_2] & \text{Cov}[\widehat{\beta}_1, \widehat{\beta}_2] & V[\widehat{\beta}_2] & \cdots & \text{Cov}[\widehat{\beta}_2, \widehat{\beta}_k] \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
\widehat{\beta}_k & \text{Cov}[\widehat{\beta}_0, \widehat{\beta}_k] & \text{Cov}[\widehat{\beta}_k, \widehat{\beta}_1] & \text{Cov}[\widehat{\beta}_k, \widehat{\beta}_2] & \cdots & V[\widehat{\beta}_k]
\end{array}$$

Recall that standard errors are the square root of the diagonals of this matrix.

Overview of Inference in the General Setting

- Under assumption 1-5 in large samples:

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{SE}[\widehat{\beta}_j]} \sim N(0,1)$$

- In small samples, under assumptions 1-6,

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{SE}[\widehat{\beta}_j]} \sim t_{n-(k+1)}$$

- Estimated standard errors are:

$$\widehat{SE}[\hat{\beta}_j] = \sqrt{\widehat{\text{var}}[\widehat{\boldsymbol{\beta}}]_{jj}}$$

$$\widehat{\text{var}}[\widehat{\boldsymbol{\beta}}] = \widehat{\sigma}^2_u(\mathbf{X}'\mathbf{X})^{-1}$$

$$\widehat{\sigma}^2_u = \frac{\widehat{\mathbf{u}}'\widehat{\mathbf{u}}}{n-(k+1)}$$

- Thus, confidence intervals and hypothesis tests proceed in essentially the same way.

Properties of the OLS Estimator: Summary

Theorem

Under Assumptions 1–6, the $(k+1) \times 1$ vector of OLS estimators $\hat{\boldsymbol{\beta}}$, conditional on $\mathbf{X}$, follows a multivariate normal distribution with mean $\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$:

$$\hat{\boldsymbol{\beta}}|\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$$

- *Each element of $\hat{\boldsymbol{\beta}}$ (i.e. $\hat{\beta}_0, ..., \hat{\beta}_{k+1}$) is normally distributed, and $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ as $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$*
- *Variances and covariances are given by $V[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$*
- *An unbiased estimator for the error variance σ^2 is given by*

$$\widehat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n - (k+1)}$$

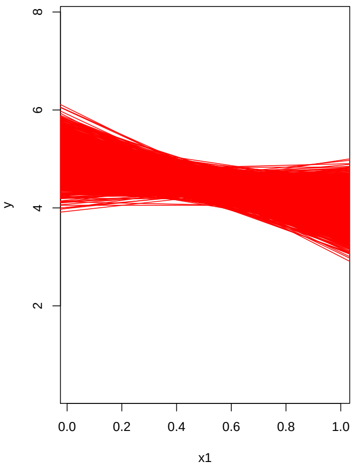
- *With a large sample, $\hat{\boldsymbol{\beta}}$ approximately follows the same distribution under Assumptions 1–5 only, i.e., without assuming the normality of $\mathbf{u}$.*

Implications of the Variance-Covariance Matrix

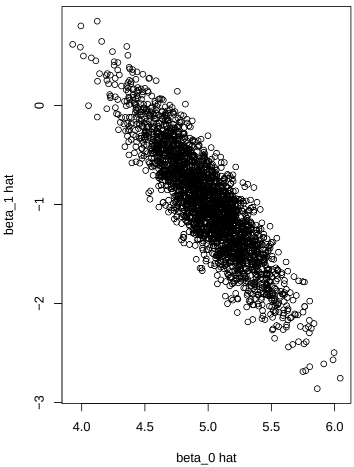
- Note that the sampling distribution is a joint distribution because it involves multiple random variables.
- This is because the sampling distribution of the terms in $\hat{\beta}$ are correlated.
- In a practical sense, this means that our uncertainty about coefficients is correlated across variables.

Multivariate Normal: Simulation

$Y = \beta_0 + \beta_1 X_1 + u$ with $u \sim N(0, \sigma_u^2 = 4)$ and $\beta_0 = 5$, $\beta_1 = -1$, and $n = 100$:



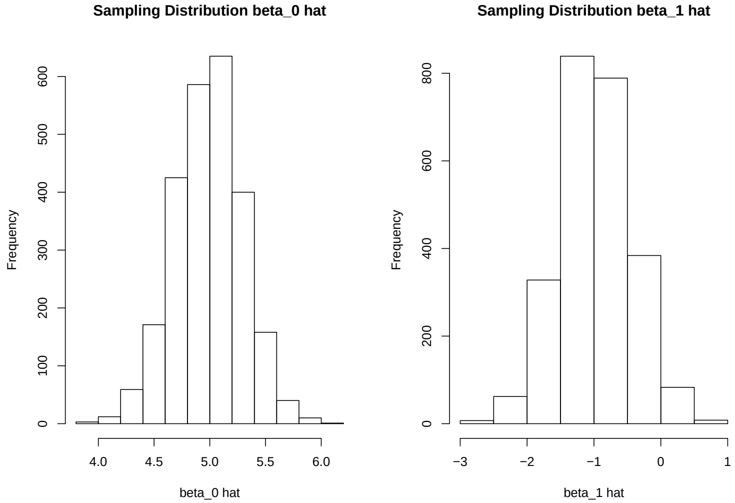
Sampling distribution of Regression Lines



Joint sampling distribution

Marginals of Multivariate Normal RVs are Normal

$Y = \beta_0 + \beta_1 X_1 + u$ with $u \sim N(0, \sigma_u^2 = 4)$ and $\beta_0 = 5$, $\beta_1 = -1$, and $n = 100$:



1. Matrix Form of Regression
2. OLS inference in matrix form
3. Standard Hypothesis Tests
4. Testing Joint Significance
5. Testing Linear Hypotheses: The General Case
6. Fun With(out) Weights
7. Appendix: Derivations and Consistency
8. The Problems with *p*-values
9. Agnostic Regression
10. Inference via the Bootstrap
11. Fun With Weights
12. Appendix: Tricky *p*-value Example

Running Example: Chilean Referendum on Pinochet

- The 1988 Chilean national plebiscite was a national referendum held to determine whether or not dictator Augusto Pinochet would extend his rule for another eight-year term in office.
- Data: national survey conducted in April and May of 1988 by FLACSO in Chile.
- Outcome: 1 if respondent intends to vote for Pinochet, 0 otherwise. We can interpret the β slopes as marginal "effects" on the probability that respondent votes for Pinochet.
- Plebiscite was held on October 5, 1988. The No side won with 56% of the vote, with 44% voting Yes.
- We model the intended Pinochet vote as a linear function of gender, education, and age of respondents.

Hypothesis Testing in R

Model the intended Pinochet vote as a linear function of gender, education, and age of respondents.

R Code

```
> fit <- lm(vote1 ~ fem + educ + age, data = d)
> summary(fit)
~~~~~

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4042284  0.0514034   7.864 6.57e-15 ***
fem          0.1360034  0.0237132   5.735 1.15e-08 ***
educ        -0.0607604  0.0138649  -4.382 1.25e-05 ***
age          0.0037786  0.0008315   4.544 5.90e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4875 on 1699 degrees of freedom
Multiple R-squared: 0.05112,       Adjusted R-squared: 0.04945
F-statistic: 30.51 on 3 and 1699 DF,  p-value: < 2.2e-16
```

The t-Value for Multiple Linear Regression

- Consider testing a hypothesis about a single regression coefficient β_j:

$$H_0 : \beta_j = c$$

- In the simple linear regression we used the t-value to test this kind of hypothesis.
- We can consider the same t-value about β_j for the multiple regression:

$$T = \frac{\hat{\beta}_j - c}{\hat{SE}(\hat{\beta}_j)}$$

- How do we compute $\hat{SE}(\hat{\beta}_j)$?

$$\hat{SE}(\hat{\beta}_j) = \sqrt{\widehat{V}(\hat{\beta}_j)} = \sqrt{\widehat{V}(\hat{\boldsymbol{\beta}})_{(j,j)}} = \sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{(j,j)}}$$

where $\mathbf{A}_{(j,j)}$ is the (j,j) element of matrix $\mathbf{A}$.
That is, take the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ and square root the diagonal element corresponding to j.

Getting the Standard Errors

R Code

```
> fit <- lm(vote1 ~ fem + educ + age, data = d)
> summary(fit)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4042284  0.0514034   7.864 6.57e-15 ***
fem          0.1360034  0.0237132   5.735 1.15e-08 ***
educ        -0.0607604  0.0138649  -4.382 1.25e-05 ***
age          0.0037786  0.0008315   4.544 5.90e-06 ***
---
```

We can pull out the variance-covariance matrix $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ in `R` from the `lm()` object:

R Code

```
> V <- vcov(fit)
> V
              (Intercept)           fem          educ           age
(Intercept)  2.642311e-03 -3.455498e-04 -5.270913e-04 -3.357119e-05
fem         -3.455498e-04  5.623170e-04  2.249973e-05  8.285291e-07
educ        -5.270913e-04  2.249973e-05  1.922354e-04  3.411049e-06
age         -3.357119e-05  8.285291e-07  3.411049e-06  6.914098e-07

> sqrt(diag(V))
 (Intercept)          fem         educ          age
0.0514034097 0.0237132251 0.0138648980 0.0008315105
```

Using the t-Value as a Test Statistic

The procedure for testing this null hypothesis ($\beta_j = c$) is identical to the simple regression case, except that our reference distribution is t_{n-k-1} instead of t_{n-2}.

1. Compute the t-value as $T = (\hat{\beta}_j - c)/\hat{SE}[\hat{\beta}_j]$
2. Compare the value to the critical value $t_{\alpha/2}$ for the α level test, which under the null hypothesis satisfies

$$P\left(-t_{\alpha/2} \leq T \leq t_{\alpha/2}\right) = 1 - \alpha$$

3. Decide whether the realized value of T in our data is unusual given the known distribution of the test statistic.
4. Finally, either declare that we reject H_0 or not, or report the p-value.

Confidence Intervals

To construct confidence intervals, there is again no difference compared to the case of $k = 1$, except that we need to use t_{n-k-1} instead of t_{n-2}

Since we know the sampling distribution for our t-value:

$$T = \frac{\hat{\beta}_j - c}{\hat{SE}[\hat{\beta}_j]} \sim t_{n-k-1}$$

So we also know the probability that the value of our test statistics falls into a given interval:

$$P\left(-t_{\alpha/2} \leq \frac{\widehat{\beta}_j - \beta_j}{\hat{SE}[\hat{\beta}_j]} \leq t_{\alpha/2}\right) = 1 - \alpha$$

We rearrange:

$$\left[\widehat{\beta}_j - t_{\alpha/2}\hat{SE}[\hat{\beta}_j]\, ,\, \widehat{\beta}_j + t_{\alpha/2}\hat{SE}[\hat{\beta}_j]\right]$$

and thus can construct the confidence intervals as usual using:

$$\hat{\beta}_j \pm t_{\alpha/2} \cdot \hat{SE}[\hat{\beta}_j]$$

Confidence Intervals in R

R Code

```
> fit <- lm(vote1 ~ fem + educ + age, data = d)
> summary(fit)
~~~~~

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4042284  0.0514034   7.864 6.57e-15 ***
fem          0.1360034  0.0237132   5.735 1.15e-08 ***
educ        -0.0607604  0.0138649  -4.382 1.25e-05 ***
age          0.0037786  0.0008315   4.544 5.90e-06 ***
---
```

R Code

```
> confint(fit)
                   2.5 %      97.5 %
(Intercept)  0.303407780  0.50504909
fem          0.089493169  0.18251357
educ        -0.087954435 -0.03356629
age          0.002147755  0.00540954
```

Testing Hypothesis About a Linear Combination of β_j

R Code

```
> fit <- lm(REALGDPCAP ~ Region, data = D)
> summary(fit)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        4452.7      783.4   5.684 2.07e-07 ***
RegionAfrica      -2552.8     1204.5  -2.119   0.0372 *
RegionAsia          148.9     1149.8   0.129   0.8973
RegionLatAmerica   -271.3     1007.0  -0.269   0.7883
RegionOecd         9671.3     1007.0   9.604 5.74e-15 ***
```

- $\hat{\beta}_{Asia}$ and $\hat{\beta}_{LAm}$ are close. So we may want to test the null hypothesis:

$$H_0 : \ \beta_{LAm} = \beta_{Asia} \ \Leftrightarrow \ \beta_{LAm} - \beta_{Asia} = 0$$

against the alternative of

$$H_1 : \ \beta_{LAm} \neq \beta_{Asia} \ \Leftrightarrow \ \beta_{LAm} - \beta_{Asia} \neq 0$$

- What would be an appropriate test statistic for this hypothesis?

Testing Hypothesis About a Linear Combination of β_j

R Code

```
> fit <- lm(REALGDPCAP ~ Region, data = D)
> summary(fit)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        4452.7      783.4   5.684 2.07e-07 ***
RegionAfrica      -2552.8     1204.5  -2.119   0.0372 *
RegionAsia          148.9     1149.8   0.129   0.8973
RegionLatAmerica   -271.3     1007.0  -0.269   0.7883
RegionOecd         9671.3     1007.0   9.604 5.74e-15 ***
```

- Let's consider a t-value:

$$T = \frac{\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}}{\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia})}$$

  We will reject H_0 if T is sufficiently different from zero.

- Note that unlike the test of a single hypothesis, both $\hat{\beta}_{LAm}$ and $\hat{\beta}_{Asia}$ are random variables, hence the denominator.

Testing Hypothesis About a Linear Combination of β_j

- Our test statistic:

$$T = \frac{\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}}{\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia})} \sim t_{n-k-1}$$

- How do you find $\hat{SE}(\hat{\beta}_{LAm} - \hat{\beta}_{Asia})$?
- Is it $\hat{SE}(\hat{\beta}_{LAm}) - \hat{SE}(\hat{\beta}_{Asia})$? No!
- Is it $\hat{SE}(\hat{\beta}_{LAm}) + \hat{SE}(\hat{\beta}_{Asia})$? No!
- Recall the following property of the variance:

$$V(X \pm Y) = V(X) + V(Y) \pm 2\text{Cov}(X, Y)$$

  Therefore, the standard error for a linear combination of coefficients is:

$$\hat{SE}(\widehat{\beta}_1 \pm \widehat{\beta}_2) = \sqrt{\widehat{V}(\widehat{\beta}_1) + \widehat{V}(\widehat{\beta}_2) \pm 2\widehat{\text{Cov}}[\widehat{\beta}_1, \widehat{\beta}_2]}$$

  which we can calculate from the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$.

- Since the estimates of the coefficients are correlated, we need the covariance term.

Example: GDP per capita on Regions

R Code

```
> fit <- lm(REALGDPCAP ~ Region, data = D)
> V <- vcov(fit)
> V
                 (Intercept) RegionAfrica RegionAsia RegionLatAmerica
(Intercept)         613769.9    -613769.9  -613769.9        -613769.9
RegionAfrica       -613769.9    1450728.8   613769.9         613769.9
RegionAsia         -613769.9     613769.9  1321965.9         613769.9
RegionLatAmerica   -613769.9     613769.9   613769.9        1014054.6
RegionOecd         -613769.9     613769.9   613769.9         613769.9
                 RegionOecd
(Intercept)       -613769.9
RegionAfrica       613769.9
RegionAsia         613769.9
RegionLatAmerica   613769.9
RegionOecd        1014054.6
```

Example: GDP per capita on Regions

We can then compute the test statistic for the hypothesis of interest:

R Code

```
> se <- sqrt(V[4,4] + V[3,3] - 2*V[3,4])
> se
[1] 1052.844
>
> tstat <- (coef(fit)[4] - coef(fit)[3])/se
> tstat
RegionLatAmerica
      -0.3990977
```

$$t = \frac{\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}}{\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia})} \quad \text{where}$$

$$\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}) = \sqrt{\widehat{V}(\widehat{\beta}_{LAm}) + \widehat{V}(\widehat{\beta}_{Asia}) - 2\widehat{\text{Cov}}[\widehat{\beta}_{LAm}, \widehat{\beta}_{Asia}]}$$

Plugging in we get $t \approx -0.40$. So what do we conclude?

We cannot reject the null that the difference in average GDP resulted from chance.

Aside: Adjusted R^2

R Code

```
> fit <- lm(vote1 ~ fem + educ + age, data = d)
> summary(fit)
~~~~~

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4042284  0.0514034   7.864 6.57e-15 ***
fem          0.1360034  0.0237132   5.735 1.15e-08 ***
educ        -0.0607604  0.0138649  -4.382 1.25e-05 ***
age          0.0037786  0.0008315   4.544 5.90e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4875 on 1699 degrees of freedom
Multiple R-squared: 0.05112,        Adjusted R-squared: 0.04945
F-statistic: 30.51 on 3 and 1699 DF,  p-value: < 2.2e-16
```

Aside: Adjusted R^2

- R^2 often used to assess in-sample model fit. Recall

  $$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}$$

  where SS_{res} are the sum of squared residuals and the SS_{tot} are the sum of the squared deviations from the mean.
- Perhaps problematically, it can be shown that R^2 always stays constant or increases with more explanatory variables
- So, how do we penalize more complex models? Adjusted R^2
- This makes R^2 more 'comparable' across models with different numbers of variables, but the next section will show you an even better way to approach that problem in a testing framework.
- Still since people report it, let's quickly derive adjusted R^2,

Aside: Adjusted R^2

- Key idea: rewrite R^2 in terms of variances

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}/n}{\text{SS}_{\text{tot}}/n}$$
$$= 1 - \frac{\tilde{\mathbb{V}}(\text{SS}_{\text{res}})}{\tilde{\mathbb{V}}(\text{SS}_{\text{tot}})}$$

  where $\tilde{\mathbb{V}}$ is a biased estimator of the population variance.

- What if we replace the biased estimator with the unbiased estimators

$$\hat{\mathbb{V}}(\text{SS}_{\text{res}}) = \text{SS}_{\text{res}}/(n-k-1)$$
$$\hat{\mathbb{V}}(\text{SS}_{\text{tot}}) = \text{SS}_{\text{tot}}/(n-1)$$

- Some algebra gets us to

$$R^2_{adj} = R^2 - \underbrace{(1-R^2)\frac{k-1}{n-k}}_{\text{model complexity penalty}}$$

- Adjusted R^2 will always be smaller than R^2 and can sometimes be negative!

1 Matrix Form of Regression

2 OLS inference in matrix form

3 Standard Hypothesis Tests

4 **Testing Joint Significance**

5 Testing Linear Hypotheses: The General Case

6 Fun With(out) Weights

7 Appendix: Derivations and Consistency

8 The Problems with *p*-values

9 Agnostic Regression

10 Inference via the Bootstrap

11 Fun With Weights

12 Appendix: Tricky *p*-value Example

F Test for Joint Significance of Coefficients

- In research we often want to test a joint hypothesis which involves multiple linear restrictions (e.g. $\beta_1 = \beta_2 = \beta_3 = 0$)
- Suppose our regression model is:

$$Voted = \beta_0 + \gamma_1 FEMALE + \beta_1 EDUCATION +$$

$$\gamma_2 (FEMALE \cdot EDUCATION) + \beta_2 AGE + \gamma_3 (FEMALE \cdot AGE) + u$$

  and we want to test

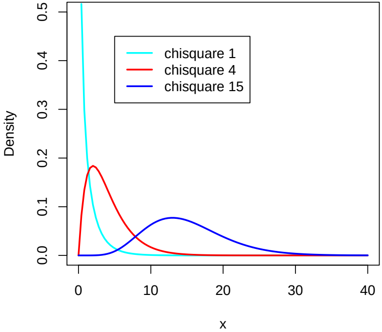
$$H_0 : \; \gamma_1 = \gamma_2 = \gamma_3 = 0.$$

- Substantively, what question are we asking?

  → Do females and males vote systematically differently from each other? (Under the null, there is no difference in either the intercept or slopes between females and males).
- This is an example of a joint hypothesis test involving three restrictions: $\gamma_1 = 0$, $\gamma_2 = 0$, and $\gamma_3 = 0$.
- If all the interaction terms and the group lower order term are close to zero, then we fail to reject the null hypothesis of no gender difference.
- F tests allows us to to test joint hypothesis

The χ^2 Distribution

- To test more than one hypothesis jointly we need to introduce some new probability distributions.
- Suppose $Z_1, ..., Z_n$ are n i.i.d. random variables following $\mathcal{N}(0,1)$.
- Then, the sum of their squares, $X = \sum_{i=1}^n Z_i^2$, is distributed according to the χ^2 distribution with n degrees of freedom, $X \sim \chi^2_n$.



Properties: $X > 0$, $E[X] = n$ and $V[X] = 2n$. In R: `dchisq()`, `pchisq()`, `rchisq()`

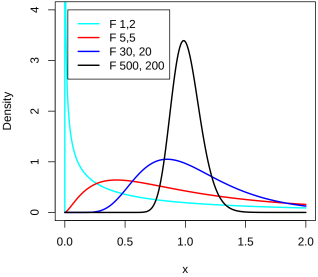
The F distribution

The F distribution arises as a ratio of two independent chi-squared distributed random variables:

$$F = \frac{X_1/df_1}{X_2/df_2} \sim \mathcal{F}_{df_1,df_2}$$

where $X_1 \sim \chi^2_{df_1}$, $X_2 \sim \chi^2_{df_2}$, and $X_1 \perp\!\!\!\perp X_2$.

df_1 and df_2 are called the numerator degrees of freedom and the denominator degrees of freedom.



In R: `df()`, `pf()`, `rf()`

F Test against $H_0: \ \gamma_1 = \gamma_2 = \gamma_3 = 0.$

The F statistic can be calculated by the following procedure:

1. Fit the Unrestricted Model (UR) which *does not* impose H_0:

   $$Vote = \beta_0 + \gamma_1 FEM + \beta_1 EDUC + \gamma_2(FEM * EDUC) + \beta_2 AGE + \gamma_3(FEM * AGE) + u$$

2. Fit the Restricted Model (R) which *does* impose H_0:

   $$Vote = \beta_0 + \beta_1 EDUC + \beta_2 AGE + u$$

3. From the two results, compute the F Statistic:

   $$F_0 = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

   where SSR=sum of squared residuals, q=number of restrictions, k=number of predictors in the unrestricted model, and n= # of observations.

   Intuition:

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

The F statistics have the following sampling distributions:

- Under Assumptions 1–6, $F_0 \sim \mathcal{F}_{q,n-k-1}$ regardless of the sample size.
- Under Assumptions 1–5, $qF_0 \overset{a.}{\sim} \chi^2_q$ as $n \to \infty$ (see next section).

Unrestricted Model (UR)

R Code

```
> fit.UR <- lm(vote1 ~ fem + educ + age + fem:age + fem:educ, data = Chile)
> summary(fit.UR)
~~~~~

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.293130   0.069242   4.233 2.42e-05 ***
fem          0.368975   0.098883   3.731 0.000197 ***
educ        -0.038571   0.019578  -1.970 0.048988 *
age          0.005482   0.001114   4.921 9.44e-07 ***
fem:age     -0.003779   0.001673  -2.259 0.024010 *
fem:educ    -0.044484   0.027697  -1.606 0.108431
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.487 on 1697 degrees of freedom
Multiple R-squared: 0.05451,        Adjusted R-squared: 0.05172
F-statistic: 19.57 on 5 and 1697 DF,  p-value: < 2.2e-16
```

Restricted Model (R)

R Code

```
> fit.R <- lm(vote1 ~ educ + age, data = Chile)
> summary(fit.R)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4878039  0.0497550   9.804  < 2e-16 ***
educ        -0.0662022  0.0139615  -4.742 2.30e-06 ***
age          0.0035783  0.0008385   4.267 2.09e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4921 on 1700 degrees of freedom
Multiple R-squared:  0.03275,       Adjusted R-squared:  0.03161
F-statistic: 28.78 on 2 and 1700 DF,  p-value: 5.097e-13
```

F Test in R

R Code

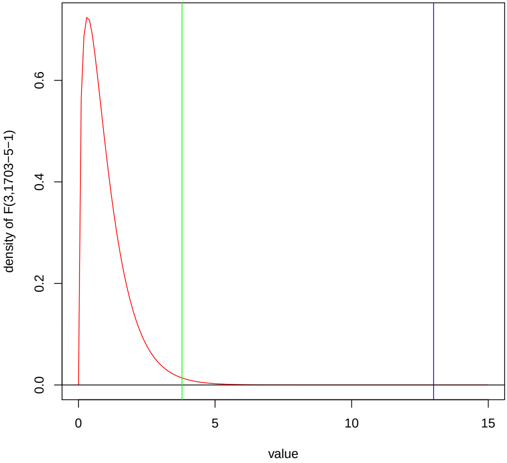
```
> SSR.UR <- sum(resid(fit.UR)^2)  # = 402
> SSR.R <- sum(resid(fit.R)^2)    # = 411

> DFdenom <- df.residual(fit.UR)  # = 1703
> DFnum <- 3

> F <- ((SSR.R - SSR.UR)/DFnum) / (SSR.UR/DFdenom)
> F
[1] 13.01581

> qf(0.99, DFnum, DFdenom)
[1] 3.793171
```

Given above, what do we conclude?

$F_0 = 13$ is greater than the critical value for a .01 level test. So we *reject* the null hypothesis.

Null Distribution, Critical Value, and Test Statistic

Note that the F statistic is always positive, so we only look at the right tail of the reference F (or χ^2 in a large sample) distribution.



F Test Examples I

The F test can be used to test various joint hypotheses which involve multiple linear restrictions. Consider the regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$

We may want to test:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

- What question are we asking?

  → Does any of the X variables help to predict Y?

- This is called the omnibus test and is routinely reported by statistical software.

Omnibus Test in R

R Code

```
> summary(fit.UR)
~~~~~

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.293130   0.069242   4.233 2.42e-05 ***
fem          0.368975   0.098883   3.731 0.000197 ***
educ        -0.038571   0.019578  -1.970 0.048988 *
age          0.005482   0.001114   4.921 9.44e-07 ***
fem:age     -0.003779   0.001673  -2.259 0.024010 *
fem:educ    -0.044484   0.027697  -1.606 0.108431
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.487 on 1697 degrees of freedom
Multiple R-squared: 0.05451,        Adjusted R-squared: 0.05172
F-statistic: 19.57 on 5 and 1697 DF,  p-value: < 2.2e-16
```

Omnibus Test in R with Random Noise

R Code

```
> set.seed(08540)
> p <- 10; x <- matrix(rnorm(p*1000), nrow=1000)
> y <- rnorm(1000); summary(lm(y~x))
~~~~~

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0115475  0.0320874  -0.360   0.7190
x1          -0.0019803  0.0333524  -0.059   0.9527
x2           0.0666275  0.0314087   2.121   0.0341 *
x3          -0.0008594  0.0321270  -0.027   0.9787
x4           0.0051185  0.0333678   0.153   0.8781
x5           0.0136656  0.0322592   0.424   0.6719
x6           0.0102115  0.0332045   0.308   0.7585
x7          -0.0103903  0.0307639  -0.338   0.7356
x8          -0.0401722  0.0318317  -1.262   0.2072
x9           0.0553019  0.0315548   1.753   0.0800 .
x10          0.0410906  0.0319742   1.285   0.1991
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.011 on 989 degrees of freedom
Multiple R-squared:  0.01129,       Adjusted R-squared:  0.001294
F-statistic: 1.129 on 10 and 989 DF,  p-value: 0.3364
```

F Test Examples II

The F test can be used to test various joint hypotheses which involve multiple linear restrictions. Consider the regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$

Next, let's consider:

$$H_0 : \beta_1 = \beta_2 = \beta_3$$

- What question are we asking?
  $\rightarrow$ Are the coefficients X_1, X_2 and X_3 different from each other?

- How many restrictions?
  $\rightarrow$ Two ($\beta_1 - \beta_2 = 0$ and $\beta_2 - \beta_3 = 0$)

- How do we fit the restricted model?
  $\rightarrow$ The null hypothesis implies that the model can be written as:

  $$Y = \beta_0 + \beta_1 (X_1 + X_2 + X_3) + ... + \beta_k X_k + u$$

  So we create a new variable $X^* = X_1 + X_2 + X_3$ and fit:

  $$Y = \beta_0 + \beta_1 X^* + ... + \beta_k X_k + u$$

Testing Equality of Coefficients in R

R Code

```
> fit.UR2 <- lm(REALGDPCAP ~ Asia + LatAmerica + Transit + Oecd, data = D)
> summary(fit.UR2)
~~~~~

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1899.9      914.9   2.077   0.0410 *
Asia          2701.7     1243.0   2.173   0.0327 *
LatAmerica    2281.5     1112.3   2.051   0.0435 *
Transit       2552.8     1204.5   2.119   0.0372 *
Oecd         12224.2     1112.3  10.990   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3034 on 80 degrees of freedom
Multiple R-squared: 0.7096,       Adjusted R-squared: 0.6951
F-statistic: 48.88 on 4 and 80 DF,  p-value: < 2.2e-16
```

Are the coefficients on *Asia*, *LatAmerica* and *Transit* statistically significantly different?

Testing Equality of Coefficients in R

R Code

```
> D$Xstar <- D$Asia + D$LatAmerica + D$Transit
> fit.R2 <- lm(REALGDPCAP ~ Xstar + Oecd, data = D)

> SSR.UR2 <- sum(resid(fit.UR2)^2)
> SSR.R2 <- sum(resid(fit.R2)^2)

> DFdenom <- df.residual(fit.UR2)

> F <- ((SSR.R2 - SSR.UR2)/2) / (SSR.UR2/DFdenom)
> F
[1] 0.08786129

> pf(F, 2, DFdenom, lower.tail = F)
[1] 0.9159762
```

So, what do we conclude?
The three coefficients are statistically indistinguishable from each other, with the p-value of 0.916.

t Test vs. F Test

Consider the hypothesis test of

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

What ways have we learned to conduct this test?

- Option 1: Compute $T = (\hat{\beta}_1 - \hat{\beta}_2)/\hat{SE}(\hat{\beta}_1 - \hat{\beta}_2)$ and do the t test.
- Option 2: Create $X^* = X_1 + X_2$, fit the restricted model, compute $F = (SSR_R - SSR_{UR})/(SSR_R/(n-k-1))$ and do the F test.

It turns out these two tests give identical results. This is because

$$X \ \sim \ t_{n-k-1} \quad \Longleftrightarrow \quad X^2 \ \sim \ \mathcal{F}_{1,n-k-1}$$

- So, for testing a single hypothesis it does not matter whether one does a t test or an F test.
- Usually, the t test is used for single hypotheses and the F test is used for joint hypotheses.

Some More Notes on F Tests

- The F-value can also be calculated from R^2:

$$F = \frac{(R^2_{UR} - R^2_R)/q}{(1 - R^2_{UR})/(n-k-1)}$$

- F tests only work for testing nested models, i.e. the restricted model must be a special case of the unrestricted model.

  For example F tests cannot be used to test

$$Y = \beta_0 + \beta_1 X_1 \qquad\qquad + \beta_3 X_3 + u$$

against

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \qquad\quad + u$$

Some More Notes on F Tests

Joint significance does not necessarily imply the significance of individual coefficients, or vice versa:

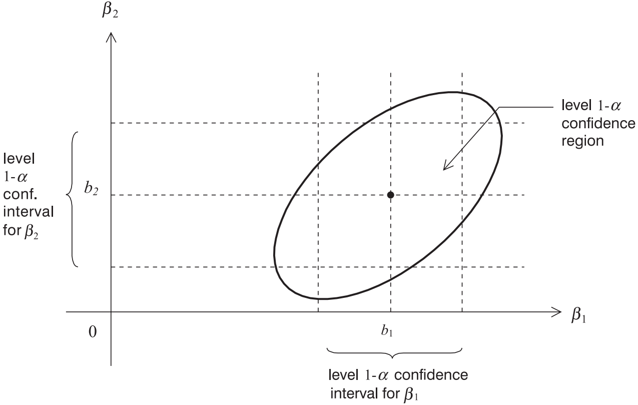

β_2
level 1-α confidence region
level 1-α conf. interval for β_2
b_2
β_1
0
b_1
level 1-α confidence interval for β_1

Figure 1.5: t- versus F-Tests

Image Credit: Hayashi (2011) *Econometrics*

1. Matrix Form of Regression
2. OLS inference in matrix form
3. Standard Hypothesis Tests
4. Testing Joint Significance
5. **Testing Linear Hypotheses: The General Case**
6. Fun With(out) Weights
7. Appendix: Derivations and Consistency
8. The Problems with *p*-values
9. Agnostic Regression
10. Inference via the Bootstrap
11. Fun With Weights
12. Appendix: Tricky *p*-value Example

Limitation of the F Formula

Consider the following null hypothesis:

$$H_0: \ \beta_1 = \beta_2 = \beta_3 = 3$$

or

$$H_0: \ \beta_1 = 2\beta_2 = 0.5\beta_3 + 1$$

Can we test them using the F test?
To compute the F value, we need to fit the restricted model. How?

- Some restrictions are difficult to impose when fitting the model.
- Even when we can, the procedure will be ad hoc and require some creativity.
- Is there a general solution?

General Procedure for Testing Linear Hypotheses

- Notice that any set of q linear hypotheses can be written as

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

  where
  - $\mathbf{R}$ is a $q \times (k+1)$ matrix of prespecified coefficients on $\boldsymbol{\beta}$ (hypothesis matrix)
  - $\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \cdots \ \beta_k]'$
  - $\mathbf{r}$ is a $q \times 1$ vector of prespecified constants

- Examples:

$$\beta_1 = \beta_2 = \beta_3 = 3 \Leftrightarrow \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} \Leftrightarrow \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}$$

$$\beta_1 = 2\beta_2 = 0.5\beta_3 + 1 \Leftrightarrow \begin{bmatrix} \beta_1 - 2\beta_2 \\ \beta_1 - 0.5\beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \Leftrightarrow \begin{bmatrix} 0 & 1 & -2 & 0 \\ 0 & 1 & 0 & -0.5 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Wald Statistic

- Let's consider testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, a set of q linear restrictions.
- If H_0 is true, $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ should be zero except for sampling variability.
- To formally evaluate the statistical significance of the deviation from zero, we must transform $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ to a statistic that can be compared to a reference distribution.
- It turns out that the following Wald statistic can be used:

$$W = \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)' \cdot \left[\hat{\sigma}^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} \cdot \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)$$

- Looks complicated? Let's figure out why this makes sense:
  - The first and last components give the sum of squares of the components of $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$. This summarizes its deviation from zero.
  - The middle component is the variance of $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$. This standardizes the sum of squares to have variance one.
- We know $\hat{\boldsymbol{\beta}}$ is approximately normal $\Rightarrow$ $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ should also be normal $\Longrightarrow$ W should therefore be ... χ^2 distributed!

Sampling Distribution of the Wald Statistic

Theorem (Large-Sample Distribution of the Wald Statistic)

Under Assumptions 1–5, as $n \to \infty$ the distribution of the Wald statistic approaches the chi square distribution with q degrees of freedom:

$$W \overset{d}{\to} \chi^2_q \quad \textit{as} \quad n \to \infty$$

Theorem (Small-Sample Distribution of the Wald Statistic)

Under Assumptions 1–6, for any sample size n the Wald statistic divided by q has the F distribution with $(q, n-k-1)$ degrees of freedom:

$$W/q \sim \mathcal{F}_{q,n-k-1}$$

- $q\mathcal{F}_{q,n-k-1} \overset{d}{\to} \chi^2_q$ as $n \to \infty$, so the difference disappears when n large.

```
> pf(3.1, 2, 500,lower.tail=F) [1] 0.04591619

> pchisq(2*3.1, 2,lower.tail=F) [1] 0.0450492

> pf(3.1, 2, 50000,lower.tail=F) [1] 0.04505786
```

Testing General Linear Hypotheses in R

In R, the `linearHypothesis()` function in the car package does the Wald test for general linear hypotheses.

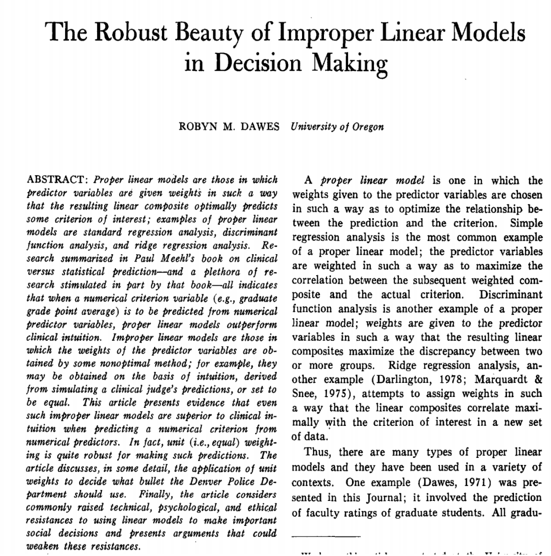
R Code

```
> fit.UR2 <- lm(REALGDPCAP ~ Asia + LatAmerica + Transit + Oecd, data = D)
> R <- matrix(c(0,1,-1,0,0, 0,1,0,-1,0), nrow = 2, byrow = T)
> r <- c(0,0)
> linearHypothesis(fit.UR2, R, r)
Linear hypothesis test

Hypothesis:
Asia - LatAmerica = 0
Asia - Transit = 0

Model 1: restricted model
Model 2: REALGDPCAP ~ Asia + LatAmerica + Transit + Oecd

  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1     82 738141635
2     80 736523836  2   1617798 0.0879  0.916
```

Conclusion

- Multiple regression is much like the regression formulations we have already seen.
- We showed how to estimate the coefficients and get the variance covariance matrix.
- Much of the hypothesis test infrastructure ports over nicely, plus there are new joint tests we can use.
- Appendix contains material on:
  - Derivation for the estimator (+ some of the math for this)
  - Proof of consistency

Fun Without Weights



The Robust Beauty of Improper Linear Models in Decision Making

ROBYN M. DAWES *University of Oregon*

ABSTRACT: *Proper linear models are those in which predictor variables are given weights in such a way that the resulting linear composite optimally predicts some criterion of interest; examples of proper linear models are standard regression analysis, discriminant function analysis, and ridge regression analysis. Research summarized in Paul Meehl's book on clinical versus statistical prediction—and a plethora of research stimulated in part by that book—all indicates that when a numerical criterion variable (e.g., graduate grade point average) is to be predicted from numerical predictor variables, proper linear models outperform clinical intuition. Improper linear models are those in which the weights of the predictor variables are obtained by some nonoptimal method; for example, they may be obtained on the basis of intuition, derived from simulating a clinical judge's predictions, or set to be equal. This article presents evidence that even such improper linear models are superior to clinical intuition when predicting a numerical criterion from numerical predictors. In fact, unit (i.e., equal) weighting is quite robust for making such predictions. The article discusses, in some detail, the application of unit weights to decide what bullet the Denver Police Department should use. Finally, the article considers commonly raised technical, psychological, and ethical resistances to using linear models to make important social decisions and presents arguments that could weaken these resistances.*

A *proper linear model* is one in which the weights given to the predictor variables are chosen in such a way as to optimize the relationship between the prediction and the criterion. Simple regression analysis is the most common example of a proper linear model; the predictor variables are weighted in such a way as to maximize the correlation between the subsequent weighted composite and the actual criterion. Discriminant function analysis is another example of a proper linear model; weights are given to the predictor variables in such a way that the resulting linear composites maximize the discrepancy between two or more groups. Ridge regression analysis, another example (Darlington, 1978; Marquardt & Snee, 1975), attempts to assign weights in such a way that the linear composites correlate maximally with the criterion of interest in a new set of data.

Thus, there are many types of proper linear models and they have been used in a variety of contexts. One example (Dawes, 1971) was presented in this Journal; it involved the prediction of faculty ratings of graduate students. All gradu-

Improper Linear Models

- Proper linear model is one where predictor variables are given optimized weights in some way (for example through regression)
- Meehl (1954) *Clinical Versus Statistical Prediction: A Theoretical Analysis and Review of the Evidence* argued that proper linear models outperform clinical intuition in many areas.
- Dawes argues that even improper linear models (those where weights are set by hand or set to be equal), outperform clinical intuition.
- Equal weight models are argued to be quite robust for these predictions

Example: Graduate Admissions

- Faculty rated all students in the psych department at University of Oregon
- Ratings predicted from a proper linear model of student GRE scores, undergrad GPA and selectivity of student's undergraduate institution. Cross-validated correlation was .38
- Correlation of faculty ratings with average rating of admissions committee was .19
- Standardized and equally weighted improper linear model, correlated at .48

Other Examples

- Self-assessed measures of marital happiness: modeled with improper linear model of (`rate of lovemaking - rate of arguments`): correlation of .40
- Einhorn (1972) study of doctors coding biopsies of patients with Hodgkin's disease and then rated severity. Their rating of severity was essentially uncorrelated with survival times, but the variables they coded predicted outcomes using a regression model.

Other Examples

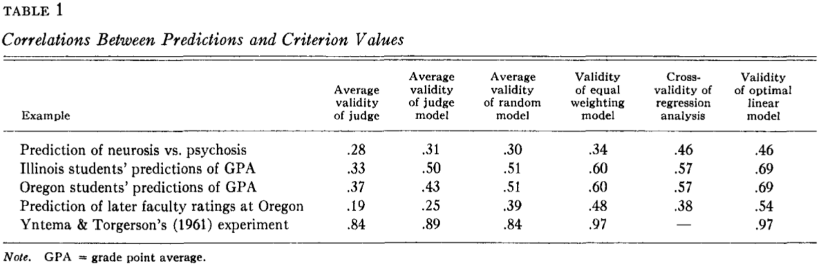


TABLE 1

Correlations Between Predictions and Criterion Values

Example	Average validity of judge	Average validity of judge model	Average validity of random model	Validity of equal weighting model	Cross-validity of regression analysis	Validity of optimal linear model
Prediction of neurosis vs. psychosis	.28	.31	.30	.34	.46	.46
Illinois students' predictions of GPA	.33	.50	.51	.60	.57	.69
Oregon students' predictions of GPA	.37	.43	.51	.60	.57	.69
Prediction of later faculty ratings at Oregon	.19	.25	.39	.48	.38	.54
Yntema & Torgerson's (1961) experiment	.84	.89	.84	.97	—	.97

Note. GPA = grade point average.

Column descriptions:

- C1) average of human judges
- C2) model based on human judges
- C3) randomly chosen weights preserving signs
- C4) equal weighting
- C5) cross-validated weights
- C6) unattainable optimal linear model

The Argument

- "People – especially the experts in a field – are much better at selecting and coding information than they are at integrating it." (573)
- The choice of variables is extremely important for prediction!
- This parallels a piece of folk wisdom in the machine learning literature that a better predictor will beat a better model every time.
- People are good at picking out relevant information, but terrible at integrating it.
- The difficulty arises in part because people in general lack a strong reference to the distribution of the predictors.
- Linear models are robust to deviations from the optimal weights (see also Waller 2008 on "Fungible Weights in Multiple Regression")

My Thoughts on the Argument

- Particularly in prediction, looking for the true or right model can be quixotic
- The broader research project suggests that a big part of what quantitative models are doing predictively, is focusing human talent in the right place.
- This all applies because predictors well chosen and the sample size is small (so the weight optimization isn't great)
- It is a fascinating paper!

1. Matrix Form of Regression
2. OLS inference in matrix form
3. Standard Hypothesis Tests
4. Testing Joint Significance
5. Testing Linear Hypotheses: The General Case
6. Fun With(out) Weights
7. Appendix: Derivations and Consistency
8. The Problems with *p*-values
9. Agnostic Regression
10. Inference via the Bootstrap
11. Fun With Weights
12. Appendix: Tricky *p*-value Example

Gradient

Let $v = v(\mathbf{u})$ be a scalar-valued function $\mathbb{R}_n \to \mathbb{R}_1$ where $\mathbf{u}$ is a $(n \times 1)$ column vector. For example: $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ where $\mathbf{c} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$

Definition (Gradient)

We can define the column vector of partial derivatives

$$\frac{\partial v(\mathbf{u})}{\partial \mathbf{u}} = \begin{bmatrix} \partial v / \partial u_1 \\ \partial v / \partial u_2 \\ \vdots \\ \partial v / \partial u_n \end{bmatrix}$$

This vector of partial derivatives is called the gradient.

Vector Derivative Rule I (linear functions)

Theorem (differentiation of linear functions)

Given a linear function $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ of an $(n \times 1)$ vector $\mathbf{u}$, the derivative of $v(\mathbf{u})$ w.r.t. $\mathbf{u}$ is given by

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{c}$$

This also works when $\mathbf{c}$ is a matrix and therefore v is a vector-valued function.

For example, let $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ where $\mathbf{c} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$, then

$$v = \mathbf{c}'\mathbf{u} = 0 \cdot u_1 + 1 \cdot u_2 + 3 \cdot u_3$$

and

$$\frac{\partial v}{\partial \mathbf{u}} = \begin{bmatrix} \partial v/\partial u_1 \\ \partial v/\partial u_2 \\ \partial v/\partial u_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix} = \mathbf{c}$$

Hence,

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{c}$$

Vector Derivative Rule II (quadratic form)

Theorem (quadratic form)

Given a $(n \times n)$ *symmetric matrix* $\mathbf{A}$ *and a scalar-valued function* $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ *of* $(n \times 1)$ *vector* $\mathbf{u}$*, we have*

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{A}'\mathbf{u} + \mathbf{A}\mathbf{u} = 2\mathbf{A}\mathbf{u}$$

For example, let $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$.Then $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ is equal to

$$\begin{aligned} v &= [3 \cdot u_1 + u_2, \; u_1 + 5 \cdot u_2] \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \\ &= 3u_1^2 + 2u_1u_2 + 5u_2^2 \end{aligned}$$

and

$$\frac{\partial v}{\partial \mathbf{u}} = \begin{bmatrix} \partial v / \partial u_1 \\ \partial v / \partial u_2 \end{bmatrix} = \begin{bmatrix} 6u_1 + 2u_2 \\ 2u_1 + 10u_2 \end{bmatrix} = 2 \cdot \begin{bmatrix} 3u_1 + 1u_2 \\ 1u_1 + 5u_2 \end{bmatrix} = 2\mathbf{A}\mathbf{u}$$

Hessian

Suppose v is a scalar-valued function $v = f(\mathbf{u})$ of a $(k+1) \times 1$ column vector $\mathbf{u} = \begin{bmatrix} u_1 & u_2 & \cdots & u_{k+1} \end{bmatrix}'$

Definition (Hessian)

The $(k+1) \times (k+1)$ matrix of second-order partial derivatives of $v = f(\mathbf{u})$ is called the Hessian matrix and denoted

$$\frac{\partial v^2}{\partial \mathbf{u} \partial \mathbf{u}'} = \begin{bmatrix} \frac{\partial v^2}{\partial u_1 \partial u_1} & \cdots & \frac{\partial v^2}{\partial u_1 \partial u_{k+1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial v^2}{\partial u_{k+1} \partial u_1} & \cdots & \frac{\partial v^2}{\partial u_{k+1} \partial u_{k+1}} \end{bmatrix}$$

Note: The Hessian is symmetric.

The above rules are used to derive the optimal estimators in the appendix slides.

Derivatives with respect to $\tilde{\boldsymbol{\beta}}$

$$\begin{aligned} S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y}) &= (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}} \end{aligned}$$

$$\frac{\partial S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}}$$

- The first term does not contain $\tilde{\boldsymbol{\beta}}$
- The second term is an example of rule I from the derivative section
- The third term is an example of rule II from the derivative section

And while we are at it the Hessian is:

$$\frac{\partial^2 S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\boldsymbol{\beta}}'} = 2\mathbf{X}'\mathbf{X}$$

Solving for $\hat{\boldsymbol{\beta}}$

$$\frac{\partial S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}}$$

Setting the vector of partial derivatives equal to zero and substituting $\hat{\boldsymbol{\beta}}$ for $\tilde{\boldsymbol{\beta}}$, we can solve for the OLS estimator.

$$\begin{aligned}
\mathbf{0} &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\
-2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= -2\mathbf{X}'\mathbf{y} \\
\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}'\mathbf{y} \\
\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\
\mathbf{I}\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\
\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}
\end{aligned}$$

Note that we implicitly assumed that $\mathbf{X}'\mathbf{X}$ is invertible.

Consistency of $\hat{\beta}$

To show consistency, we rewrite the OLS estimator in terms of sample means so that we can apply LLN.

First, note that a matrix cross product can be written as a sum of vector products:

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i \quad \text{and} \quad \mathbf{X}'\mathbf{y} = \sum_{i=1}^{n} \mathbf{x}_i' y_i$$

where $\mathbf{x}_i$ is the $1 \times (k+1)$ row vector of predictor values for unit i.

Now we can rewrite the OLS estimator as,

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\sum\nolimits_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i\right)^{-1} \left(\sum\nolimits_{i=1}^{n} \mathbf{x}_i' y_i\right) \\
&= \left(\sum\nolimits_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i\right)^{-1} \left(\sum\nolimits_{i=1}^{n} \mathbf{x}_i'(\mathbf{x}_i\boldsymbol{\beta} + u_i)\right) \\
&= \boldsymbol{\beta} + \left(\sum\nolimits_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i\right)^{-1} \left(\sum\nolimits_{i=1}^{n} \mathbf{x}_i' u_i\right) \\
&= \boldsymbol{\beta} + \left(\frac{1}{n}\sum\nolimits_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i\right)^{-1} \left(\frac{1}{n}\sum\nolimits_{i=1}^{n} \mathbf{x}_i' u_i\right)
\end{aligned}$$

Consistency of $\hat{\boldsymbol{\beta}}$

Now let's apply the LLN to the sample means:

$$\begin{aligned}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i\right) &\xrightarrow{p} E[\mathbf{x}_i'\mathbf{x}_i], \text{ a } (k+1)\times(k+1) \text{ nonsingular matrix.}\\ \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i'u_i\right) &\xrightarrow{p} E[\mathbf{x}_i'u_i] \;=\; 0, \text{ by the zero cond. mean assumption.}\end{aligned}$$

Therefore, we have

$$\begin{aligned}\text{plim}(\hat{\boldsymbol{\beta}}) &= \boldsymbol{\beta} + \left(E[\mathbf{x}_i'\mathbf{x}_i]\right)^{-1}\cdot 0\\ &= \boldsymbol{\beta}.\end{aligned}$$

We can also show the asymptotic normality of $\hat{\boldsymbol{\beta}}$ using a similar argument but with the CLT.

References

- Wooldridge, Jeffrey. *Introductory econometrics: A modern approach*. Cengage Learning, 2012.

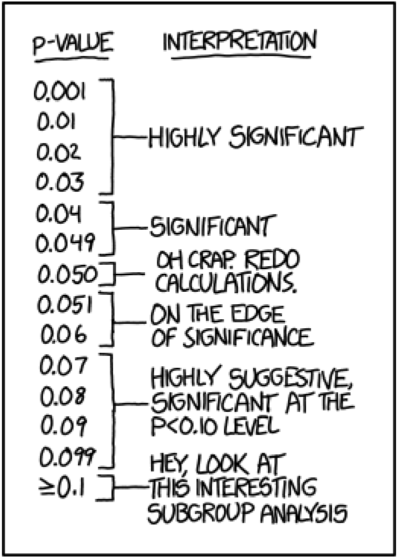
Where We've Been and Where We're Going...

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - Monday:
    - matrix form of linear regression
    - t-tests, F-tests and general linear hypothesis tests
  - Wednesday:
    - problems with *p*-values
    - agnostic regression
    - the bootstrap
- Next Week
  - break!
  - then . . . diagnostics
- Long Run
  - probability → inference → regression → causal inference

Questions?

1. Matrix Form of Regression
2. OLS inference in matrix form
3. Standard Hypothesis Tests
4. Testing Joint Significance
5. Testing Linear Hypotheses: The General Case
6. Fun With(out) Weights
7. Appendix: Derivations and Consistency
8. **The Problems with p-values**
9. Agnostic Regression
10. Inference via the Bootstrap
11. Fun With Weights
12. Appendix: Tricky p-value Example

p-values (courtesy of XKCD)


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

The value of the *p*-value

Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.

Ronald Fisher (1935)

> In social science (and I think in psychology as well), the null hypothesis is almost certainly false, false, false, and you don't need a p-value to tell you this. The p-value tells you the extent to which a certain aspect of your data are consistent with the null hypothesis. A lack of rejection doesn't tell you that the null hypothesis is likely true; rather, it tells you that you don't have enough data to reject the null hypothesis.
>
> *Andrew Gelman (2010)*

Problems with p-values

- p-values are extremely common in the social sciences and are often the standard by which the value of the finding is judged.
- p-values are not:
  - an indication of a large substantive effect
  - the probability that the null hypothesis is true
  - the probability that the alternative hypothesis is false
- a large p-value could mean either that we are in the null world OR that we had insufficient power

So what is the basic idea?

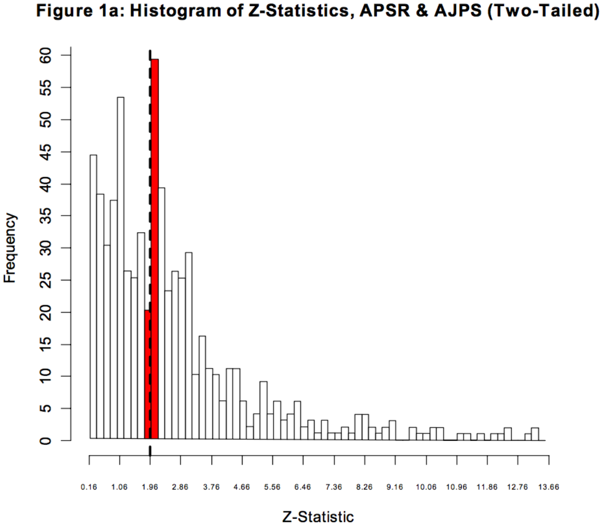
The idea was to run an experiment, then see if the results were consistent with what random chance might produce. Researchers would first set up a 'null hypothesis' that they wanted to disprove, such as there being no correlation or no difference between groups. Next, they would play the devil's advocate and, assuming that this null hypothesis was in fact true, calculate the chances of getting results at least as extreme as what was actually observed. This probability was the P value. The smaller it was, suggested Fisher, the greater the likelihood that the straw-man null hypothesis was false.
(Nunzo 2014, emphasis mine)

I've got 99 problems. . .

p-values are hard to interpret, but even in the best scenarios they have some key problems:

- they remove focus from data, measurement, theory and the substantive quantity of interest
- they are often applied outside the dichotomous/decision-making framework where they make some sense
- significance isn't even a good filter for predictive covariates (Ward et al 2010, Lo et al 2015)
- they lead to publication filtering on arbitrary cutoffs.

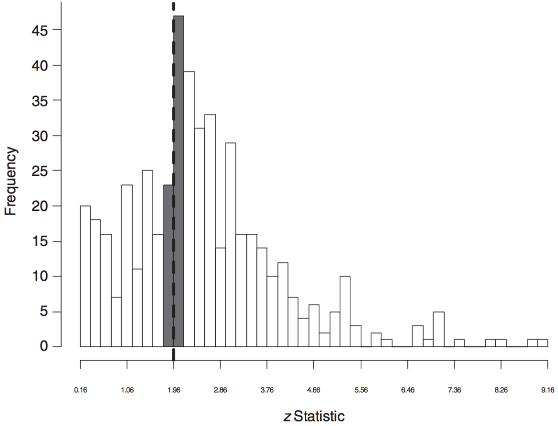
Arbitrary Cutoffs



Gerber and Malhotra (2006) Top Political Science Journals

Arbitrary Cutoffs

Figure 1
Histogram of *z* Statistics From the *American Sociological Review*, the *American Journal of Sociology*, and *The Sociological Quarterly* (Two-Tailed)



Gerber and Malhotra (2008) Top Sociology Journals

Arbitrary Cutoffs

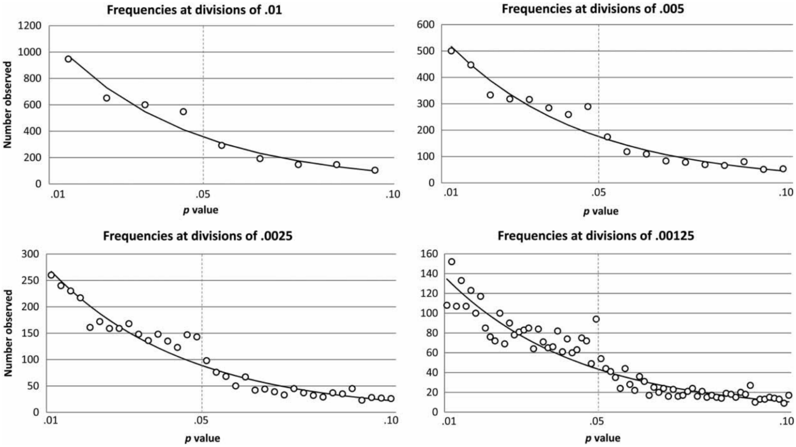
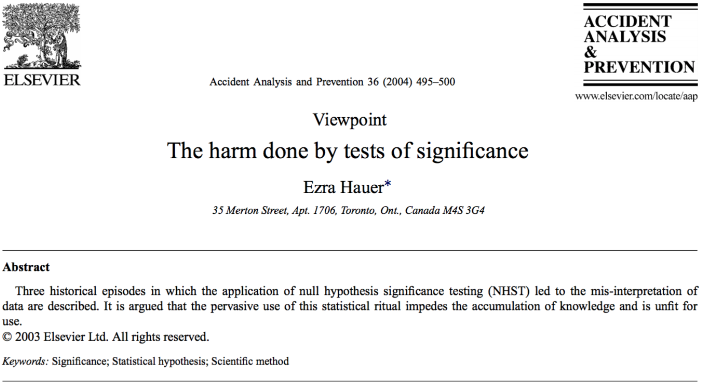


Figure 1.. *The graphs show the distribution of 3,627* p *values from three major psychology journals.*

Masicampo and Lalande (2012) Top Psychology Journals

Still Not Convinced?
The Real Harm of Misinterpreted *p*-values



ELSEVIER

Accident Analysis and Prevention 36 (2004) 495–500

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Viewpoint

The harm done by tests of significance

Ezra Hauer*

35 Merton Street, Apt. 1706, Toronto, Ont., Canada M4S 3G4

Abstract

Three historical episodes in which the application of null hypothesis significance testing (NHST) led to the mis-interpretation of data are described. It is argued that the pervasive use of this statistical ritual impedes the accumulation of knowledge and is unfit for use.

Keywords: Significance; Statistical hypothesis; Scientific method

Example from Hauer: Right-Turn-On-Red

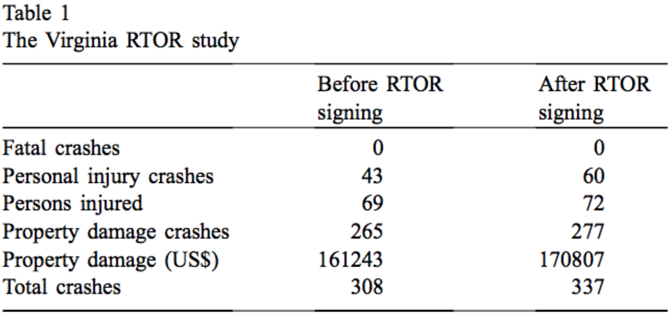


Table 1
The Virginia RTOR study

	Before RTOR signing	After RTOR signing
Fatal crashes	0	0
Personal injury crashes	43	60
Persons injured	69	72
Property damage crashes	265	277
Property damage (US$)	161243	170807
Total crashes	308	337

The Point in Hauer

- Two other interesting examples in Hauer (2004)
- Core issue is that lack of significance is not an indication of a zero effect, it could also be a lack of power (i.e. a small sample size relative to the difficulty of detecting the effect)
- On the opposite end, large tech companies rarely use significance testing because they have huge samples which essentially always find some non-zero effect. But that doesn't make the finding significant in a colloquial sense of important.

p-values and Confidence Intervals

p-values one of most used tests in the social sciences–and you're telling me not to rely on them?

Basically, yes.

What's the matter with you?

- Two reasons not to worship p-values [of many]
  1) Statistical: they represent a very specific quantity under a null distribution. If you don't really care about rejecting just that null, then you should focus on providing more information
  2) Substantive: p-values are divorced from your quantity of interest–which almost always should relate to how much an intervention changes a quantity of social scientific interest (newspaper rule)

p-values and Confidence Intervals

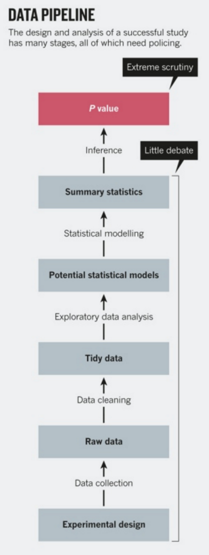
But I want to assess the probability that my hypothesis is true–why can't I use a p-value?

1) Me too, good luck.
2) That's not what p-values measure
3) No one study is going to eliminate an entire hypothesis; even if that study generates a really small p-value, you'd probably want an entirely different infrastructure

Instead, show quantities you care about with confidence intervals.

Don't misinterpret, or rely too heavily, on your p-values. They are evidence against your null, not evidence in favor of your alternative.

But Let's Not Obsess Too Much About *p*-values



From Leek and Peng (2015) "*P* values are just the tip of the iceberg" *Nature*.

1. Matrix Form of Regression
2. OLS inference in matrix form
3. Standard Hypothesis Tests
4. Testing Joint Significance
5. Testing Linear Hypotheses: The General Case
6. Fun With(out) Weights
7. Appendix: Derivations and Consistency
8. The Problems with p-values
9. **Agnostic Regression**
10. Inference via the Bootstrap
11. Fun With Weights
12. Appendix: Tricky p-value Example

Regression as parametric modeling

Let's summarize the parametric view we have taken thus far.

- Gauss-Markov assumptions:
  - (A1) linearity, (A2) i.i.d. sample, (A3) full rank $\mathbf{X}_i$, (A4) zero conditional mean error, (A5) homoskedasticity.
  - basically, assume the model is right
- $\rightsquigarrow$ OLS is BLUE, plus (A6) normality of the errors and we get small sample SEs and BUE.
- What is the basic approach here?
  - A1 defines a linear model for the conditional expectation:

$$E[Y_i|\mathbf{X}_i] = \mu_i = \mathbf{X}_i'\boldsymbol{\beta}$$

  - A4-6, define a probabilistic model for the conditional distribution of Y_i given X_i:

$$Y_i|X_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

  - A3 covers the edge-case that the βs are indistinguishable.
  - A2 ensures that we observe independent samples for estimation.

Agnostic views on regression

$$[Y_i|X_i] \sim N(X_i'\beta, \sigma^2)$$

- These assumptions assume we know a lot about how Y_i is 'generated'.
- Justifications for using OLS (like BLUE/BUE) often invoke these assumptions which are unlikely to hold exactly.
- Alternative: take an agnostic view on regression.
  - use OLS without believing these assumptions.
  - lean on two things: A2 i.i.d. sample, asymptotics (large-sample properties)
- Lose the distributional assumptions, focus on the conditional expectation function (CEF):

$$\mu(x) = \mathbb{E}[Y_i|X_i = x] = \sum_y y \cdot \mathbb{P}[Y_i = y|X_i = x]$$

- NB: this makes no statement about whether or not the CEF you are looking at is the 'right' one.

Justifying linear regression

- Define linear regression:

$$\beta = \arg\min_b \mathbb{E}[(Y_i - X_i' b)^2]$$

- The solution to this is the following:

$$\beta = \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[X_i Y_i]$$

- Note that the is the population coefficient vector, not the estimator yet.
- In other words, even a non-linear CEF has a “true” linear approximation, even though that approximation may not be great.

Regression anatomy

- Consider simple linear regression:

$$(\alpha, \beta) = \underset{a,b}{\arg\min}\, \mathbb{E}\left[(Y_i - a - bX_i)^2\right]$$

- In this case, we can write the population/true slope β as:

$$\beta = \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[X_i Y_i] = \frac{\text{Cov}(Y_i, X_i)}{\mathrm{Var}[X_i]}$$

- With more covariates, β is more complicated, but we can still write it like this.
- Let $\tilde{X}_{ki}$ be the residual from a regression of X_{ki} on all the other independent variables. Then, β_k, the coefficient for X_{ki} is:

$$\beta_k = \frac{\text{Cov}(Y_i, \tilde{X}_{ki})}{\mathrm{Var}(\tilde{X}_{ki})}$$

Justification 1: Linear CEFs

- Justification 1: if the CEF is linear, the population regression function is it. That is, if $E[Y_i|X_i] = X_i'b$, then $b = \beta$.
- When would we expect the CEF to be linear? Two cases.
  1. Outcome and covariates are multivariate normal.
  2. Linear regression model is saturated.
- A model is saturated if there are as many parameters as there are possible combination of the X_i variables.

Saturated model example

- Two binary variables, X_{1i} for marriage status and X_{2i} for having children.
- Four possible values of X_i, four possible values of $\mu(X_i)$:

$$
\begin{aligned}
E[Y_i|X_{1i}=0, X_{2i}=0] &= \alpha \\
E[Y_i|X_{1i}=1, X_{2i}=0] &= \alpha + \beta \\
E[Y_i|X_{1i}=0, X_{2i}=1] &= \alpha + \gamma \\
E[Y_i|X_{1i}=1, X_{2i}=1] &= \alpha + \beta + \gamma + \delta
\end{aligned}
$$

- We can write the CEF as follows:

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

Saturated model example

$$E[Y_i|X_{1i}, X_{2i}] = \alpha + \beta X_{1i} + \gamma X_{2i} + \delta(X_{1i}X_{2i})$$

- Basically, each value of $\mu(X_i)$ is being estimated separately.
  - $\rightsquigarrow$ within-strata estimation.
  - No borrowing of information from across values of X_i.
- Requires a set of dummies for each categorical variable plus all interactions.
- Or, a series of dummies for each unique combination of X_i.
- This makes linearity hold mechanically and so linearity is not an assumption.

Saturated model example

- Washington (AER) data on the effects of daughters.
- We'll look at the relationship between voting and number of kids (causal?).

```r
girls <- foreign::read.dta("girls.dta")
head(girls[, c("name", "totchi", "aauw")])
```

```
##                    name totchi aauw
## 1   ABERCROMBIE, NEIL      0  100
## 2   ACKERMAN, GARY L.      3   88
## 3 ADERHOLT, ROBERT B.      0    0
## 4    ALLEN, THOMAS H.      2  100
## 5  ANDREWS, ROBERT E.      2  100
## 6        ARCHER, W.R.      7    0
```

Linear model

```
summary(lm(aauw ~ totchi, data = girls))
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    61.31       1.81   33.81   <2e-16 ***
## totchi         -5.33       0.62   -8.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42 on 1733 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.0408, Adjusted R-squared:  0.0403
## F-statistic: 73.8 on 1 and 1733 DF,  p-value: <2e-16
```

Saturated model

```
summary(lm(aauw ~ as.factor(totchi), data = girls))
```

```
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)             56.41       2.76   20.42  < 2e-16 ***
## as.factor(totchi)1       5.45       4.11    1.33   0.1851
## as.factor(totchi)2      -3.80       3.27   -1.16   0.2454
## as.factor(totchi)3     -13.65       3.45   -3.95  8.1e-05 ***
## as.factor(totchi)4     -19.31       4.01   -4.82  1.6e-06 ***
## as.factor(totchi)5     -15.46       4.85   -3.19   0.0015 **
## as.factor(totchi)6     -33.59      10.42   -3.22   0.0013 **
## as.factor(totchi)7     -17.13      11.41   -1.50   0.1336
## as.factor(totchi)8     -55.33      12.28   -4.51  7.0e-06 ***
## as.factor(totchi)9     -50.41      24.08   -2.09   0.0364 *
## as.factor(totchi)10    -53.41      20.90   -2.56   0.0107 *
## as.factor(totchi)12    -56.41      41.53   -1.36   0.1745
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41 on 1723 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.0506, Adjusted R-squared:  0.0446
## F-statistic: 8.36 on 11 and 1723 DF,  p-value: 1.84e-14
```

Saturated model minus the constant

```
summary(lm(aauw ~ as.factor(totchi) - 1, data = girls))
```

```
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## as.factor(totchi)0     56.41       2.76   20.42   <2e-16 ***
## as.factor(totchi)1     61.86       3.05   20.31   <2e-16 ***
## as.factor(totchi)2     52.62       1.75   30.13   <2e-16 ***
## as.factor(totchi)3     42.76       2.07   20.62   <2e-16 ***
## as.factor(totchi)4     37.11       2.90   12.79   <2e-16 ***
## as.factor(totchi)5     40.95       3.99   10.27   <2e-16 ***
## as.factor(totchi)6     22.82      10.05    2.27   0.0233 *
## as.factor(totchi)7     39.29      11.07    3.55   0.0004 ***
## as.factor(totchi)8      1.08      11.96    0.09   0.9278
## as.factor(totchi)9      6.00      23.92    0.25   0.8020
## as.factor(totchi)10     3.00      20.72    0.14   0.8849
## as.factor(totchi)12     0.00      41.43    0.00   1.0000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41 on 1723 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.587,  Adjusted R-squared:  0.584
## F-statistic:  204 on 12 and 1723 DF,  p-value: <2e-16
```

Compare to within-strata means

- The saturated model makes no assumptions about the between-strata relationships.
- Just calculates within-strata means:

```
c1 <- coef(lm(aauw ~ as.factor(totchi) - 1, data = girls))
c2 <- with(girls, tapply(aauw, totchi, mean, na.rm = TRUE))
rbind(c1, c2)
```

```
##      0  1  2  3  4  5  6  7   8 9 10 12
## c1 56 62 53 43 37 41 23 39 1.1 6  3  0
## c2 56 62 53 43 37 41 23 39 1.1 6  3  0
```

Other justifications for OLS

- **Justification 2**: $X_i'\beta$ is the best linear predictor (in a mean-squared error sense) of Y_i.
  - Why? $\beta = \arg\min_b \mathbb{E}[(Y_i - X_i' b)^2]$
- **Justification 3**: $X_i'\beta$ provides the minimum mean squared error linear approximation to $E[Y_i|X_i]$.
- Even if the CEF is not linear, a linear regression provides the best linear approximation to that CEF.
- Don't need to believe the assumptions (linearity) in order to use regression as a good approximation to the CEF.
- Warning if the CEF is very nonlinear then this approximation could be terrible!!

The error terms

- Let's define the error term: $e_i \equiv Y_i - X_i'\beta$ so that:

$$Y_i = X_i'\beta + [Y_i - X_i'\beta] = X_i'\beta + e_i$$

- Note the residual e_i is uncorrelated with X_i:

$$\begin{aligned}
\mathbb{E}[X_i e_i] &= \mathbb{E}[X_i(Y_i - X_i'\beta)] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}[X_i X_i'\beta] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}\left[X_i X_i' \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i Y_i]\right] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}[X_i X_i'] \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i Y_i] \\
&= \mathbb{E}[X_i Y_i] - \mathbb{E}[X_i Y_i] = 0
\end{aligned}$$

- No assumptions on the linearity of $\mathbb{E}[Y_i|X_i]$.

OLS estimator

- We know the population value of β is:

$$\beta = \mathbb{E}[X_i X_i']^{-1}\mathbb{E}[X_i Y_i]$$

- How do we get an estimator of this?
- Plug-in principle $\rightsquigarrow$ replace population expectation with sample versions:

$$\hat{\beta} = \left[\frac{1}{N}\sum_i X_i X_i'\right]^{-1} \frac{1}{N}\sum_i X_i Y_i$$

- If you work through the matrix algebra, this turns out to be:

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

Asymptotic OLS inference

- With this representation in hand, we can write the OLS estimator as follows:

$$\hat{\beta} = \beta + \left[\sum_i X_i X_i'\right]^{-1} \sum_i X_i e_i$$

- Core idea: $\sum_i X_i e_i$ is the sum of r.v.s so the CLT applies.
- That, plus some simple asymptotic theory allows us to say:

$$\sqrt{N}(\hat{\beta} - \beta) \rightsquigarrow N(0, \Omega)$$

- Converges in distribution to a Normal distribution with mean vector 0 and covariance matrix, Ω:

$$\Omega = \mathbb{E}[X_i X_i']^{-1} \mathbb{E}[X_i X_i' e_i^2] \mathbb{E}[X_i X_i']^{-1}.$$

- No linearity assumption needed!

Estimating the variance

- In large samples then:

$$\sqrt{N}(\hat{\beta} - \beta) \sim N(0, \Omega)$$

- How to estimate Ω? Plug-in principle again!

$$\widehat{\Omega} = \left[\sum_i X_i X_i'\right]^{-1} \left[\sum_i X_i X_i' \hat{e}_i^2\right] \left[\sum_i X_i X_i'\right]^{-1}.$$

- Replace e_i with its emprical counterpart (residuals) $\hat{e}_i = Y_i - X_i'\hat{\beta}$.
- Replace the population moments of X_i with their sample counterparts.
- The square root of the diagonals of this covariance matrix are the "robust" or Huber-White standard errors (we will return to this in a few classes).

The Agnostic Statistics Perspective

- The key insight here is that we can derive estimators for properties of the conditional expectation function under somewhat weaker assumptions
- They still rely heavily on large samples (asymptotic results) and independent samples.
- See the Aronow and Miller textbook for a great explanation and defense of this worldview (also Lin's 2013 'Agnostic notes on regression adjustments to experimental data').
- We will come back to re-thinking the implications for finite samples during the diagnostics classes.

1. Matrix Form of Regression
2. OLS inference in matrix form
3. Standard Hypothesis Tests
4. Testing Joint Significance
5. Testing Linear Hypotheses: The General Case
6. Fun With(out) Weights
7. Appendix: Derivations and Consistency
8. The Problems with *p*-values
9. Agnostic Regression
10. Inference via the Bootstrap
11. Fun With Weights
12. Appendix: Tricky *p*-value Example

The Bootstrap

How do we do inference if we don't know how to construct the sampling distribution for our estimator?

Idea of the Bootstrap: Use the empirical CDF (eCDF) as a plug-in for the CDF, and resample from that.

We are pretending our sample eCDF looks sufficiently close to our true CDF, and so we're sampling from the eCDF as an approximation to repeated sampling from the true CDF. This is called a resampling method.

1. Take a with replacement sample of size n from our sample.
2. Calculate our would-be estimate using this bootstrap sample.
3. Repeat steps 1 and 2 many (B) times.
4. Using the resulting collection of bootstrap estimates to calculate estimates of the standard error or confidence intervals (more later).

Why The Bootstrap

If we have a very complicated estimator, such as

- $\overline{X}_1^4/(\overline{X}_2 - \overline{X}_3)^2$ where samples 2 and 3 are not drawn independently
- $(\hat{\beta}_0 + \hat{\beta}_1)^3/(2 + \hat{\beta}_2^2)$

Then the bootstrap is very useful because we don't have to derive the analytical expectation and variance, we can calculate them in the bootstrap.

Bootstrap is really useful when:

- you want to avoid making a normal approximation
- you want to avoid making certain assumptions (i.e. homoskedasticity)
- you are considering an estimator for which an analytical variance estimator is hard or impossible to calculate.

This is the closest thing to magic I will show you all semester.

Two ways to calculate intervals and *p*-values

- Using normal approximation intervals and *p*-values, use the estimates from step 4.

$$\left[\overline{X} - \Phi^{-1}(1-\alpha/2) * \hat{\sigma}_{boot}, \overline{X} + \Phi^{-1}(1-\alpha/2) * \hat{\sigma}_{boot}\right]$$

  - Intuition check: the standard error is just the standard deviation of the bootstrap replicates. There is not square root of n. Why?

- Percentile method for the CI: Sort B bootstrap estimates from smallest to largest. Grab the values at $\alpha/2 * B$ and $1 - \alpha/2 * B$ position.
  - Percentile method does not rely on normal approximation but requires very large B and thus more computational time.

Example: Linear Regression

We skimmed over the sampling distribution of the variance parameter in a linear regression earlier.

It turns out that $\widehat{\sigma}^2 \sim \chi^2_{n-(K+1)}$.

But instead we'll use Bootstrap:

1) Sample from data set, with replacement n times, $\tilde{\boldsymbol{X}}$
2) Calculate $f(\tilde{\boldsymbol{X}})$ (in this case a regression)
3) Repeat B times, form distribution of statistics
4) Calculate confidence interval by identifying $\alpha/2$ and $1-\alpha/2$ value of statistic. (percentile method)

The Bootstrap More Formally

- What we are discussing is the nonparametric bootstrap
- $y_1, \ldots, y_n$ are the outcomes of independent and identically distributed random variables $Y_1, \ldots, Y_n$ whose PDF and CDF are denoted by f and F.
- The sample is used to make inferences about an estimand, denoted by θ using a statistic T whose value in the sample is t.
- If we observed F, statistical inference would be very easy, but instead we observe $\hat{F}$, which is the empirical distribution that put equal probabilities n^{-1} at each sample value y_i.
  - Estimates are constructed by the plug-in principle, which says that the parameter $\theta = t(F)$ is estimated by $\hat{\theta} = t(\hat{F})$. (i.e. we plug in the ECDF for the CDF)
  - Why does this work? Sampling distribution entirely determined by the CDF and n, WLLN says the ECDF will look more and more like the CDF as n gets large.

When Does the Bootstrap Fail?

Bootstrap works in a wide variety of circumstances, but it does require some regularity conditions and it can fail with certain types of data and estimators:

- Bootstrap fails when the sampling distribution of the estimator is non-smooth. (e.g. max and min).
- Dependent data: nonparametric bootstrap assumes data so independent so will not work with time series data or other dependent structures.
  - For clustered data, standard bootstrap will not work, but the block bootstrap will work. In the block bootstrap, clusters are resampled (not necessarily units) with replacement.
  - More on this later.
- Many other variants that may be right for certain situations: studentized intervals, jackknife, parametric bootstrap, bag of little bootstraps, bootstrapping for complex survey designs, etc.

Fox Chapter 21 has a nice section on the bootstrap, Aronow and Miller (2016) covers the theory well.

Today in Summary

- The difficulty of interpreting p-values
- Agnostic Regression: how will regression behave in large samples if we don't really buy the assumptions
- Bootstrap: a close to assumption free way of constructing confidence intervals
- Appendix contains a fun example of the difficulty of thinking through p-values.

Next “Week” of Classes (Three Classes)

- What can go wrong and how to fix it → Diagnostics
- Day 1 (M): Unusual and Influential Data → Robust Estimation
- Day 2 (W): Nonlinearity → Generalized Additive Models
- Day 3 (M): Unusual Errors → Sandwich Standard Errors
- Reading:
    - Angrist and Pishke Chapter 8 (‘Nonstandard Standard Error Issues’)
    - Optional: Fox Chapters 11-13
    - Optional: King and Roberts “How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It.” *Political Analysis*, 2, 23: 159179.
    - Optional: Aronow and Miller Chapters 4.2-4.4 (Inference, Clustering, Nonlinearity)

1 Matrix Form of Regression

2 OLS inference in matrix form

3 Standard Hypothesis Tests

4 Testing Joint Significance

5 Testing Linear Hypotheses: The General Case

6 Fun With(out) Weights

7 Appendix: Derivations and Consistency

8 The Problems with *p*-values

9 Agnostic Regression

10 Inference via the Bootstrap

11 **Fun With Weights**

12 Appendix: Tricky *p*-value Example

Fun With Weights

Aronow, Peter M., and Cyrus Samii. "Does Regression Produce Representative Estimates of Causal Effects?." *American Journal of Political Science* (2015).[2]

- Imagine we care about the possibly heterogeneous causal effect of a treatment D and we control for some covariates X?
- We can express the regression as a weighting over individual observation treatment effects where the weight depends only on X.
- Useful technology for understanding what our models are identifying off of by showing us our effective sample.

[2] I'm grateful to Peter Aronow for sharing his slides, several of which are used here.

How this works

We start by asking what the estimate of the average causal effect of interest converges to in a large sample:

$$\hat{\beta} \xrightarrow{p} \frac{E[w_i \tau_i]}{E[w_i]} \text{ where } w_i = (D_i - E[D_i|X])^2 \,,$$

so that $\hat{\beta}$ converges to a reweighted causal effect. As $E[w_i|X_i] = \text{Var}[D_i|X_i]$, we obtain an average causal effect reweighted by conditional variance of the treatment.

Estimation

A simple, consistent plug-in estimator of w_i is available: $\hat{w}_i = \tilde{D}_i^2$ where $\tilde{D}_i$ is the residualized treatment. (the proof is connected to the partialing out strategy we showed last week)

Easily implemented in R:

```
wts <- (d - predict(lm(d~x)))^2
```

Implications

- Unpacking the black box of regression gives us substantive insight
- When some observations have no weight, this means that the covariates completely explain their treatment condition.
- This is a feature, not a bug, of regression: we can't learn anything from those cases anyway (i.e. it is automatically handling issues of common support).
- The downside is that we have to be aware of what happened!

Application

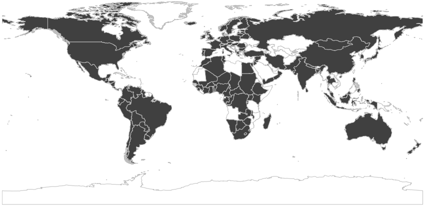
Jensen (2003), "Democratic Governance and Multinational Corporations: Political Regimes and Inflows of Foreign Direct Investment."

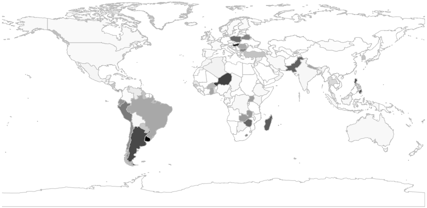
Jensen presents a large-N TSCS-analysis of the causal effects of governance (as measured by the Polity III score) on Foreign Direct Investment (FDI).

The nominal sample: 114 countries from 1970 to 1997.

Jensen estimates that a 1 unit increase in polity score corresponds to a 0.020 increase in net FDI inflows as a percentage of GDP ($p < 0.001$).

Nominal and Effective Samples





Over 50% of the weight goes to just 12 (out of 114) countries.

Broader Implications

When causal effects are heterogeneous, we can draw a distinction between "internally valid" and "externally valid" estimates of an Average Treatment Effect (ATE).

- "Internally valid": reliable estimates of ATEs, but perhaps not for the population you care about
  - randomized (lab, field, survey) experiments, instrumental variables, regression discontinuity designs, other natural experiments
- "Externally valid": perhaps unreliable estimates of ATEs, but for the population of interest
  - large-N analyses, representative surveys

Broader Implications

Aronow and Samii argue that analyses which use regression, even with a representative sample, have no greater claim to external validity than do [natural] experiments.

- When a treatment is "as-if" randomly assigned conditional on covariates, regression distorts the sample by implicitly applying weights.
- The effective sample (upon which causal effects are estimated) may have radically different properties than the nominal sample.
- When there is an underlying natural experiment in the data, a properly specified regression model may reproduce the internally valid estimate associated with the natural experiment.

1. Matrix Form of Regression
2. OLS inference in matrix form
3. Standard Hypothesis Tests
4. Testing Joint Significance
5. Testing Linear Hypotheses: The General Case
6. Fun With(out) Weights
7. Appendix: Derivations and Consistency
8. The Problems with *p*-values
9. Agnostic Regression
10. Inference via the Bootstrap
11. Fun With Weights
12. Appendix: Tricky *p*-value Example

Still Not Convinced? A Tricky Example: *p*-values

Morris 1987[3]

> *Mr. Allen the candidate for political Party A will run against Mr. Baker of Party B for office. Past races between these parties for this office were always closer, and it seems this one will be no exception- Party A candidates always have gotten between 40% and 60% of the vote and have won about half of the elections. Allen needs to know for $\theta =$the proportion of voters favoring him today, whether H_0: $\theta < .5$ or $H_1 : \theta > .5$ is true. A random sample of n voters is taken, with Y voters favoring Allen. The population is large and it is justifiable to assume that $Y \sim Bin(n, \theta)$, the binomial distribution. The estimate $\hat{\theta} = Y/n$ will be used.*

[3] From a Comment on Berger and Sellke "Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence

Example: *p*-values

Question: Which of three outcomes, all having the same p value, would be most encouraging to candidate Allen?

(a) $Y = 15, n = 20, \hat{\theta} = .75, (.560, .940)$

(b) $Y = 115, n = 200, \hat{\theta} = .575, (.506, .644)$

(c) $Y = 1046, n = 2000, \hat{\theta} = .523(.501, .545)$

Note: The p values are all about .021*.*

Example: *p*-values

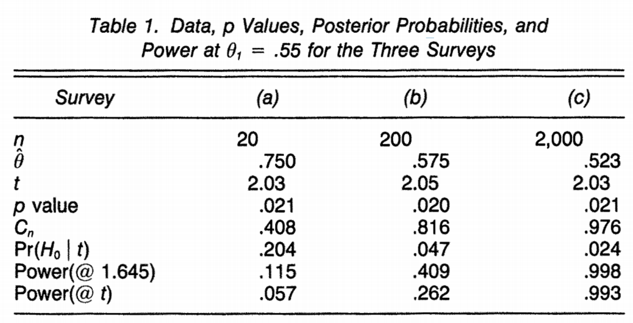


Table 1. Data, p Values, Posterior Probabilities, and Power at θ_1 = .55 for the Three Surveys

Survey	*(a)*	*(b)*	*(c)*
n	20	200	2,000
$\hat{\theta}$	.750	.575	.523
t	2.03	2.05	2.03
p value	.021	.020	.021
C_n	.408	.816	.976
$\Pr(H_0 \mid t)$	.204	.047	.024
Power(@ 1.645)	.115	.409	.998
Power(@ t)	.057	.262	.993

- the *p*-values corresponds to $Pr(H_0|t)$ only when good power obtains at typical H_1 parameter values.
- alternatively, simulate the thing you care about