Linear Models Lecture 3: Estimation of the Linear Projection Model

Robert Gulotty

University of Chicago

February 20, 2026

Deriving the OLS estimator

- Recall, geometry can work in populations or samples.
- In the following slides we will derive
  - the OLS slope estimator $\hat{\boldsymbol{\beta}}$,
  - the OLS estimator for error variance,
  - Decomposition of OLS (Analysis of Variance), with R^2
- Next time, the FWL Theorem.

Review: Modeling the Stochastic Element in the Population

- The CEF is the regression of Y on X, but requires knowing the distribution of (Y, X), a fact about Populations.
- The linear projection model predicts Y as a linear function of $X = (X_1,\ X_2, \ldots, 1)$, approximating the CEF, but still a fact about Populations.

$$\beta = [\mathbb{E}(XX')]^{-1}\mathbb{E}[XY]$$

- Estimation involves using a model of a sample to make inferences about the population.
- A statistic is random variable constructed from the sample, can be used for description or inference.
- An estimator is a statistic used to estimate population parameters.

Samples

- We estimate the projection coefficient β with measurements of (Y, X), a *sample*.
- $\{(Y_i, X_i) : \ i \in 1, \ldots, \ n\}$, is the sample, governed by the distribution of (Y, X).
- From the next slide forward, $\boldsymbol{X}$ is going to be a matrix, and we will use lower case for vectors e.g. $\boldsymbol{e}$.

Identically Distributed.

- Main approach is to assume homogeneity: $\{(Y_i, \boldsymbol{x}_i) : \ i \in 1, \ldots, \ n\}$ are identically distributed.

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \beta_1 \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \beta_2 \begin{bmatrix} X_{21} \\ \vdots \\ X_{2n} \end{bmatrix} + \beta_3 \begin{bmatrix} X_{31} \\ \vdots \\ X_{3n} \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

$$= \begin{bmatrix} 1 & X_{21} & X_{31} \\ \vdots & \vdots & \vdots \\ 1 & X_{2n} & X_{3n} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

- Note, $E(Y_1) \neq E(Y_2)$.

Least Squares Estimator

- The linear projection coefficient β is the minimizer of the expected squared error

$$S(\beta) = \mathbb{E}[(Y - X'\beta)^2]$$

- The *moment estimator* of $S(\beta)$ is the sample average:

$$\hat{S}(\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i'\beta)^2$$

- $\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i'\beta)^2$ is called SSE(β), the **sum of squared errors**.
- The estimator $\hat{\beta}$ is the minimizer of $\hat{S}(\beta)$, as well as the minimizer of the SSE.
- $\hat{\beta}$ is called the least squares estimator, sometimes written $\hat{\beta}_n$.

Solving the minimization problem

- The form of $\hat{\beta}$ is the sample analogue to the population linear projection coefficient β.
- The solution will require some additional concepts associated with matrix/multivariate calculus.
  - Quadratic form of a matrix, connecting matrices with polynomial equations.
  - Definiteness of matrices, connecting matrices to convex functions.
  - Generalizing derivatives to cover matrices.

Finite Sample Assumptions of OLS estimator

- Assumption 1: The random variables $\{(Y_1,\ X_1),\ (Y_2,\ X_2), \ldots, (Y_n,\ X_n)$ are independent and identically distributed.
- Assumption 2: $Y = \boldsymbol{X}'\beta + e$, where $\mathbb{E}(e|\boldsymbol{X}) = 0$.
- Assumption 3: $\mathbb{E}(\boldsymbol{X}\boldsymbol{X}') > 0$ is invertible (with probability 1).
- Assumption 4*: $\mathbb{E}[e^2|\boldsymbol{X}] = \sigma^2(\boldsymbol{X}) = \sigma^2$.

Derivation of Least Squares Estimator

- The functional form of the least squares estimator can be derived via multivariate calculus.
- We look for a minimum of SSE(β).
- Take first derivatives to calculate first-order conditions and locate the critical values.
- Then we check we have the global minimum, by using the fact that positive definite matrices represent convex functions.

Calculus Derivation of Least Squares Estimator

$$
\begin{aligned}
SSE(\beta) &= (\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta) && \text{(Definition of } \boldsymbol{e}\text{)} \\
&= (\boldsymbol{y}' - \beta'\boldsymbol{X}')(\boldsymbol{y} - \boldsymbol{X}\beta) && \text{(Transpose rules)} \\
&= \boldsymbol{y}'\boldsymbol{y} - \beta'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\beta + \beta'\boldsymbol{X}'\boldsymbol{X}\beta && \text{(Distributive property)} \\
\min_{\beta} SSE(\beta) &= \min_{\beta} \underbrace{\boldsymbol{y}'\boldsymbol{y}}_{\text{constant}} - \underbrace{2\boldsymbol{y}'\boldsymbol{X}\beta}_{\text{linear in } \beta} + \underbrace{\beta'\boldsymbol{X}'\boldsymbol{X}\beta}_{\text{quadratic form}}
\end{aligned}
$$

Calculus Derivation of Least Squares Estimator

$$\begin{aligned}
\frac{\partial \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} &= \mathbf{0} - \boldsymbol{X}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{X} + (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}} \\
&= -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} && \text{(Symmetric)} \\
-2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} &= \mathbf{0} && \text{(find critical point)} \\
\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} &= \boldsymbol{X}'\boldsymbol{y}. && \text{(normal equations)} \\
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} && \text{(If } \boldsymbol{X}'\boldsymbol{X} \text{ can be inverted (A3).)}
\end{aligned}$$

Second order condition

Take second derivative, and check that it is positive

$$\frac{\partial - 2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}}{\partial \hat{\boldsymbol{\beta}}'} = 2(\boldsymbol{X}'\boldsymbol{X})'$$

Which is positive so long as $\boldsymbol{X}'\boldsymbol{X}$ is positive definite (same assumption as A1).
$\hat{\boldsymbol{\beta}}$ minimizes the sum of squared errors, assuming $\boldsymbol{X}$ is of full rank.

OLS in R: The Formula in Action

```r
library(carData)
X <- cbind(1, Prestige$education, Prestige$income)
y <- Prestige$prestige

# The matrix formula from the slides:
solve(t(X) %*% X) %*% t(X) %*% y

# What lm() computes:
coef(lm(prestige ~ education + income, data=Prestige))
```

The two are identical: `lm()` implements $\hat{\beta} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$.

Projection Matrix

- There is a geometric interpretation for the minimization problem that uses the idea of "projection".
- The following matrix is called a "projection matrix"

$$\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

- Properties of $\boldsymbol{P}$
  - $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{X}$
  - If $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$, then $\boldsymbol{P}\boldsymbol{X}_1 = \boldsymbol{X}_1$.
  - $\boldsymbol{P} = \boldsymbol{P}'$, that is, $\boldsymbol{P}$ is symmetric.
  - $\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$, that is, $\boldsymbol{P}$ is idempotent.
  - $\boldsymbol{P}\boldsymbol{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{y}}$, called the fitted value. This is why $\boldsymbol{P}$ is called the "hat matrix".

Projection onto column of 1

- If $\boldsymbol{X} = \mathbf{1}$, then $\boldsymbol{P}_1 = \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' = \frac{1}{n}\mathbf{1}\mathbf{1}'$
- So if we project $\boldsymbol{y}$ onto $\boldsymbol{P}_1$, we get a vector repeating the sample mean:

$$\boldsymbol{P}_1\boldsymbol{y} = \frac{1}{n}\mathbf{1}\mathbf{1}'\boldsymbol{y} = \mathbf{1} \cdot \frac{1}{n}\sum_{i=1}^{n} Y_i = \mathbf{1}\bar{Y}$$

- **Takeaway:** The simplest regression (intercept only) projects $\boldsymbol{y}$ onto the sample mean. Every additional regressor refines this projection.
- The annihilator $\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{y} - \mathbf{1}\bar{Y}$ *demeans* $\boldsymbol{y}$—it removes the part explained by the intercept.

Orthogonal Projection

- The following matrix is called an "orthogonal projection matrix", or the Annihilator Matrix.

$$\begin{aligned} \boldsymbol{M} &= \boldsymbol{I}_n - \boldsymbol{P} \\ &= \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' \end{aligned}$$

- $\boldsymbol{M}$ and $\boldsymbol{X}$ are orthogonal:

$$\begin{aligned} \boldsymbol{M}\boldsymbol{X} &= (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{X} \\ &= \boldsymbol{X} - \boldsymbol{P}\boldsymbol{X} \\ &= \boldsymbol{X} - \boldsymbol{X} \\ &= 0 \end{aligned}$$

- We can define the residuals from projecting onto a column of 1s: $\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{y} - \mathbf{1}\bar{Y}$, this demeans $\boldsymbol{y}$.

Relationship between residuals and disturbances

$$
\begin{aligned}
\hat{\boldsymbol{e}} &= \boldsymbol{y} - \hat{\boldsymbol{y}} \\
&= \boldsymbol{y} - \boldsymbol{X}\hat{\beta} \\
&= \boldsymbol{y} - \boldsymbol{P}\boldsymbol{y} \\
&= \boldsymbol{M}\boldsymbol{y} \\
&= \boldsymbol{M}(\boldsymbol{X}\beta + \boldsymbol{e}) \\
&= \boldsymbol{M}\boldsymbol{X}\beta + \boldsymbol{M}\boldsymbol{e} \\
&= \boldsymbol{M}\boldsymbol{e}
\end{aligned}
$$

Projection and Residuals in R

```r
mod <- lm(prestige ~ education + income, data=Prestige)
X <- cbind(1, Prestige$education, Prestige$income)
P <- X %*% solve(t(X) %*% X) %*% t(X)
M <- diag(nrow(X)) - P

# P*y = fitted values
all.equal(as.vector(P %*% y), fitted(mod))  # TRUE

# M*y = residuals
all.equal(as.vector(M %*% y), resid(mod))   # TRUE
```

$\boldsymbol{P}$ and $\boldsymbol{M}$ are not abstract notation—they are the matrices R uses to compute `fitted()` and `resid()`.

Why Does Projection Help?

- We want to estimate the systematic part $\boldsymbol{X}\beta$. Two candidates:
  1. Use $\boldsymbol{y}$ itself (just report the raw data).
  2. Use $\boldsymbol{P}\boldsymbol{y} = \hat{\boldsymbol{y}}$ (the projection onto the column space of $\boldsymbol{X}$).
- Compare their variances as estimators of $\boldsymbol{X}\beta$:

$$\text{Var}(\boldsymbol{y}|\boldsymbol{X}) = \sigma^2\boldsymbol{I}$$
$$\text{Var}(\boldsymbol{P}\boldsymbol{y}|\boldsymbol{X}) = \sigma^2\boldsymbol{P}$$

- The difference is $\sigma^2(\boldsymbol{I} - \boldsymbol{P}) = \sigma^2\boldsymbol{M}$, which is positive semi-definite.
- So $\text{Var}(\boldsymbol{y}) \geq \text{Var}(\boldsymbol{P}\boldsymbol{y})$: projection *removes noise* without distorting the signal, because $\boldsymbol{P}\boldsymbol{X}\beta = \boldsymbol{X}\beta$.

Variance of the regression errors

- We want to measure the precision of our regression estimates.
- The error (or disturbance) of an observation is the deviation of a value from its theoretical mean, cannot be observed.
- The coefficients inherit that uncertainty from the theoretical model.
- Our estimates have additional uncertainty, related to the fact we only have a sample.
- Residuals are the difference between observations and estimates, can be observed.

The bane of statistics

- The mean squared error, or MSE, is calculated on the computed residuals, not the unobservable errors.
- Residuals, although observed, have a distribution that is not identical to the population disturbances.
- We will try to stick to disturbances and residuals to make distinctions.

Estimation of Error Variance

- Natural estimator: plug residuals $\hat{e}_i$ into the sample variance formula:

$$\hat{\sigma}^2 = n^{-1}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} = n^{-1}(\boldsymbol{My})'(\boldsymbol{My}) = n^{-1}\boldsymbol{y}'\boldsymbol{My} = n^{-1}\boldsymbol{e}'\boldsymbol{Me}$$

  where the last step uses $\boldsymbol{M} = \boldsymbol{M}'\boldsymbol{M}$ and $\boldsymbol{MX} = 0$.

- But $\hat{\sigma}^2$ underestimates the true variance $\tilde{\sigma}^2 = n^{-1}\boldsymbol{e}'\boldsymbol{e}$:

$$\tilde{\sigma}^2 - \hat{\sigma}^2 = n^{-1}\boldsymbol{e}'\boldsymbol{e} - n^{-1}\boldsymbol{e}'\boldsymbol{Me} = n^{-1}\boldsymbol{e}'\boldsymbol{Pe} \geq 0$$

  because $\boldsymbol{P}$ is positive semi-definite.

- Projection "absorbs" some of the error into fitted values, so residuals understate the true noise.

Estimation of σ_e^2

- Our estimator will be $s_{\hat{e}}^2 = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}/(N-K)$.
- Proof of unbiasedness of $s_{\hat{e}}^2$:

$$\begin{aligned}
\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} &= \boldsymbol{e}'\boldsymbol{M}'\boldsymbol{M}\boldsymbol{e} = \boldsymbol{e}'\boldsymbol{M}\boldsymbol{e} && (\boldsymbol{M} \text{ symmetric, idempotent}) \\
E[\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}|\boldsymbol{X}] &= E[\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e}|\boldsymbol{X}] \\
&= E[\text{tr}(\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e})|\boldsymbol{X}] && (\text{scalar} = \text{its own trace}) \\
&= E[\text{tr}(\boldsymbol{M}\boldsymbol{e}\boldsymbol{e}')|\boldsymbol{X}] && (\text{tr}(\boldsymbol{ABC})=\text{tr}(\boldsymbol{CAB})) \\
&= \text{tr}(\boldsymbol{M}\, E[\boldsymbol{e}\boldsymbol{e}'|\boldsymbol{X}]) && (\boldsymbol{M} \text{ fixed given } \boldsymbol{X}) \\
&= \text{tr}(\boldsymbol{M}\sigma^2\boldsymbol{I}) = \sigma^2\text{tr}(\boldsymbol{M}) = \sigma_e^2(N-K) && (\text{A4, tr}(\boldsymbol{M})=N-K)
\end{aligned}$$

Standard Deviation and Standard Error

- Random variables have a mean and a standard deviation, the latter quantifies variability.
- Statistics (like the sample mean) are calculated from a random sample to make inferences.
- Statistics have a sampling distribution with a mean and standard deviation.
- The standard deviation of an estimated statistic is called the "standard error".
- $\sqrt{s_e^2}$ is called the (computed) standard error of the regression
- $\sqrt{s_e^2}$ is a biased estimator for σ, but it is **consistent**.
- Similarly, the standard deviation of $\hat{\beta}$ is estimated using the standard errors.

  That is, we have a tool for estimating parameters which itself has parameters to estimate!

Interpretation of Residual Standard Error

- summary(lm()) reports s_e as the Residual standard error or "sigma"
- STATA reports this quantity as "Root Mean Squared Error".
- s_e is on the same scale as the outcome.
- If $y > 0$, it can make sense to calculate an *average prediction error rate*: $\frac{s_e}{\bar{y}}$.
- Example: if we are predicting houses that are on average 1 million dollars, being generally off by \$50k is less bad than if we are predicting houses that are \$100k.

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

$$
\begin{aligned}
V(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= V((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}|\boldsymbol{X}) && \text{(Normal Equation)} \\
&= V(\boldsymbol{A}\boldsymbol{y}|\boldsymbol{X}) && \text{(Pick } \boldsymbol{A} \text{ as placeholder)} \\
&= \boldsymbol{A}V(\boldsymbol{y}|\boldsymbol{X})\boldsymbol{A}' && (\text{Var}(cX) = c^2\text{Var}(X)) \\
&= \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}' \\
&= \boldsymbol{A}(\sigma^2\boldsymbol{I})\boldsymbol{A}' && \text{(Assumption A4)} \\
&= \sigma^2\boldsymbol{A}\boldsymbol{A}' \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
\hat{V}(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= s_e^2(\boldsymbol{X}'\boldsymbol{X})^{-1} && \text{(Standard Error of } \hat{\boldsymbol{\beta}}\text{)}
\end{aligned}
$$

Standard Errors in R: Connecting to the Formula

```r
mod <- lm(prestige ~ education + income, data=Prestige)

# The formula from the slides: s^2 (X'X)^{-1}
sigma(mod)^2 * solve(t(X) %*% X)

# What R reports:
vcov(mod)

# Standard errors = sqrt of the diagonal
sqrt(diag(vcov(mod)))
coef(summary(mod))[, "Std. Error"]  # identical
```

Every standard error in `summary(lm())` comes from $\sqrt{\text{diag}(s_e^2(\boldsymbol{X'X})^{-1})}$.

Decomposition

- $\boldsymbol{M}$ makes the least squares residuals:

$$\boldsymbol{My} = \boldsymbol{y} - \boldsymbol{Py} = \boldsymbol{y} - \boldsymbol{X}\hat{\beta} = \hat{\boldsymbol{e}}$$

- We can rewrite $\boldsymbol{y}$ in terms of

$$\boldsymbol{y} = \boldsymbol{Py} + \boldsymbol{My} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{e}}$$

- This decomposition is orthogonal,

$$\hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} = (\boldsymbol{PY})'(\boldsymbol{MY}) = \boldsymbol{y}'\boldsymbol{PMY} = 0$$

$$\begin{aligned}\boldsymbol{Y'Y} &= (\hat{\boldsymbol{y}} + \hat{\boldsymbol{e}})'(\hat{\boldsymbol{y}} + \hat{\boldsymbol{e}}) = (\hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}} + \hat{\boldsymbol{Y}}'\hat{\boldsymbol{e}} + \hat{\boldsymbol{e}}'\hat{\boldsymbol{Y}} + \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}) \\ &= \hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}} + \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} \\ \sum_{i=1}^{n} Y_i^2 &= \sum_{i=1}^{n} \hat{Y}_i^2 + \sum_{i=1}^{n} \hat{e}_i^2\end{aligned}$$

Analysis of Variance

- Subtract $\bar{Y}$ from both sides of the decomposition:

$$\boldsymbol{y} - \mathbf{1}\bar{Y} = (\hat{\boldsymbol{y}} - \mathbf{1}\bar{Y}) + \hat{\boldsymbol{e}}$$

- Take inner products. The cross terms vanish when $\boldsymbol{X}$ contains a constant:

$$(\hat{\boldsymbol{y}} - \mathbf{1}\bar{Y})'\hat{\boldsymbol{e}} = \hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} - \bar{Y}\mathbf{1}'\hat{\boldsymbol{e}} = 0 - 0 = 0$$

- So we get the **analysis of variance** (ANOVA) decomposition:

$$\|\boldsymbol{y} - \mathbf{1}\bar{Y}\|^2 = \|\hat{\boldsymbol{y}} - \mathbf{1}\bar{Y}\|^2 + \|\hat{\boldsymbol{e}}\|^2$$

$$\underbrace{\sum(Y_i - \bar{Y})^2}_{\text{SST}} = \underbrace{\sum(\hat{Y}_i - \bar{Y})^2}_{\text{SSR (regression)}} + \underbrace{\sum \hat{e}_i^2}_{\text{SSE (error)}}$$

Goodness-of-Fit: R^2

- Dividing through by SST:

$$1 = \frac{\text{SSR}}{\text{SST}} + \frac{\text{SSE}}{\text{SST}}$$
$$R^2 \equiv \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

- R^2 is the proportion of the variance in Y that is linearly explained by $\boldsymbol{X}$.
- $0 \leq R^2 \leq 1$: equals 0 when $\boldsymbol{X}$ explains nothing; equals 1 when $\hat{e}_i = 0$ for all i.
- We can always make $R^2 = 1$ by adding enough linearly independent columns to $\boldsymbol{X}$ (one per observation). This does not mean the model is good.

R^2 in Practice

- R^2 measures *descriptive fit*, not causal validity. A high R^2 does not mean the coefficients are unbiased; a low R^2 does not mean the model is useless.
- Typical values vary by field:
  - Cross-sectional micro data (e.g. earnings regressions): $R^2 \approx 0.2$–0.4
  - Aggregate time-series (e.g. GDP growth models): $R^2 \approx 0.7$–0.9
  - Experimental data with noisy outcomes: R^2 can be very low and the treatment effect still precisely estimated.
- In R: `summary(lm(y ~ x))$r.squared` reports R^2.
- The **adjusted** $R^2 = 1 - \frac{n-1}{n-k}(1 - R^2)$ penalizes adding regressors and can decrease when a variable adds no explanatory power.

Naming Conventions: A Warning

- Different textbooks use SSE and SSR with *opposite* meanings:

	Hansen / this course	**Some other texts**
$\sum(\hat{Y}_i - \bar{Y})^2$	SSR (regression)	ESS or SSR
$\sum \hat{e}_i^2$	SSE (error)	RSS or SSE

- The underlying math is always the same: SST = (explained) + (unexplained).
- When reading papers or other textbooks, check which convention is in use.