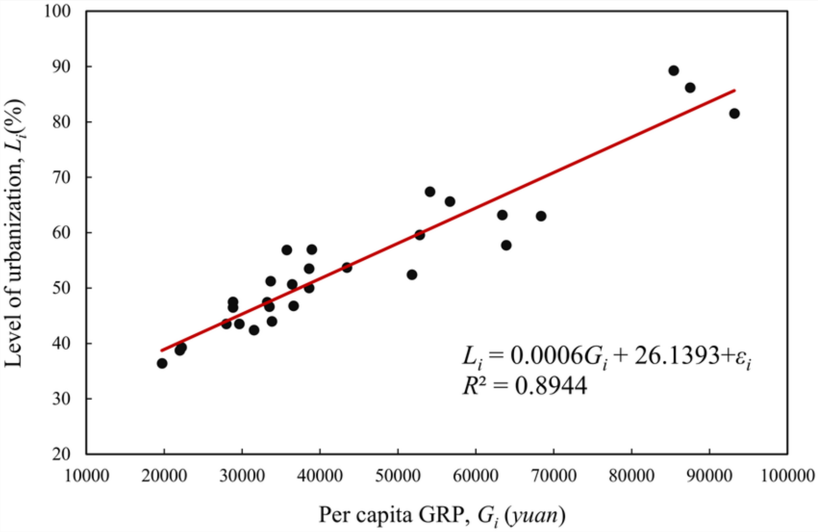
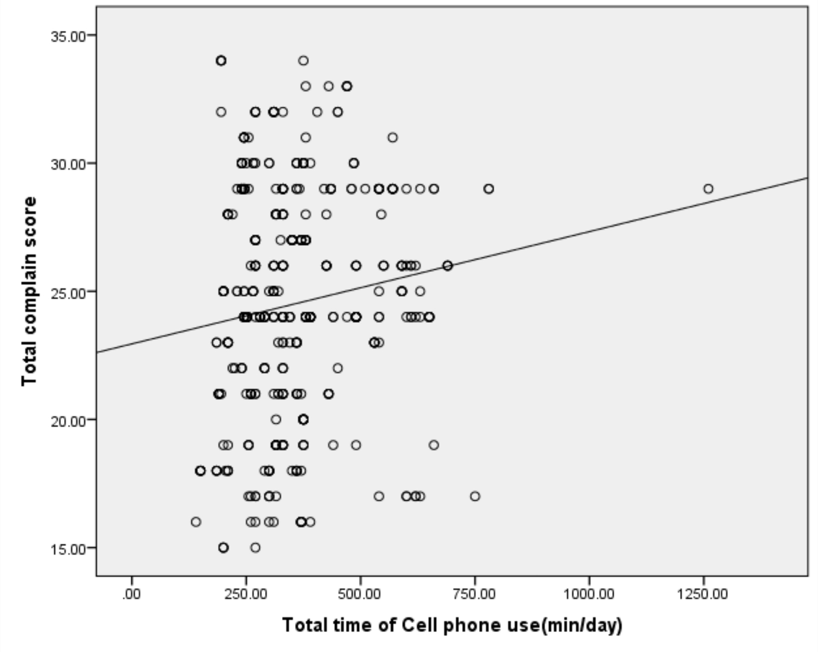
Introduction to Regression


Level of urbanization, L_i(%)
Per capita GRP, G_i (*yuan*)
$L_i = 0.0006G_i + 26.1393+\varepsilon_i$
$R^2 = 0.8944$
100
90
80
70
60
50
40
30
20
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000

coef = .58144999, se = .09536607, t = 6.1


35.00
30.00
25.00
20.00
15.00
Total complain score
.00
250.00
500.00
750.00
1000.00
1250.00
Total time of Cell phone use(min/day)

Idea: Draw a line given a sample of data

What does this line *mean*?

Under what assumptions?

The Mechanics of OLS Regression

- The OLS algorithm
- Statistical assumptions
- Statistical guarantees

Unit Plan

Goals of This Week

At the end of this week, you will:

1. Understand that regression is the plug-in estimator of the best linear predictor (BLP).
2. Understand overall model fit and use an F-test to assess whether a candidate model is performing better than a baseline model.
3. Understand how to appropriately interpret regression coefficients, including measures of certainty and uncertainty, and use Wald tests to assess whether coefficients are different from zero.
4. Lay a foundation for careful interpretation.

Reading: The Golem of Prague

Reading: The Golem of Prague

This is a placeholder for a reading call. We're just placing it here for organization.

- Read Sections 1.0 and 1.1 of *Statistical Rethinking*, which we have provided a copy of in PDF form from the publisher.
- *Statistical Rethinking* is a great book and reference that you should consider later in your data science and statistics path.

Regression, a Statistical Golem

Regression, a Statistical Golem

- Regression—like all models—is a tool.
- We put tools to use toward a data scientific purpose; however, tools are only tools.
- Use of a straight-edge, scale, and T-square doesn't make one an architect any more than use of `{insert language}` or `{insert technique}` makes one a data scientist.

The Machinery of OLS Regression

- OLS regression is a plug-in estimator for the best linear predictor (BLP).
- The BLP is the lowest mean squared error (MSE) estimator, out of all linear functions.

Applications of OLS Regression

Regression is fantastically versatile

- Under some circumstances, regression has explainable internal weights (coefficients) that are of interest.
- Under other circumstances, regression identifies causal effects.
- Under many circumstances, regression is the *de facto* baseline estimator.

Elements of a Linear Model

Elements in a Linear Model

Linear model A.K.A. linear predictor

A **linear model** is a representation of a random variable (Y) as a linear function of other random variables $\boldsymbol{X} = (X_1, X_2, ..., X_k)$.

Y	$X_1, X_2, ..., X_k$
Target	Features
Outcome	Predictors
Dependent variable	Independent variables
Output	Inputs
Response	Controls
Left-hand side (LHS)	Right-hand side (RHS)
⋮	⋮

The Linear Model Formula

The linear model formula

$$\begin{aligned}\hat{Y} &= g(X_1, X_2, ..., X_k) \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k\end{aligned}$$

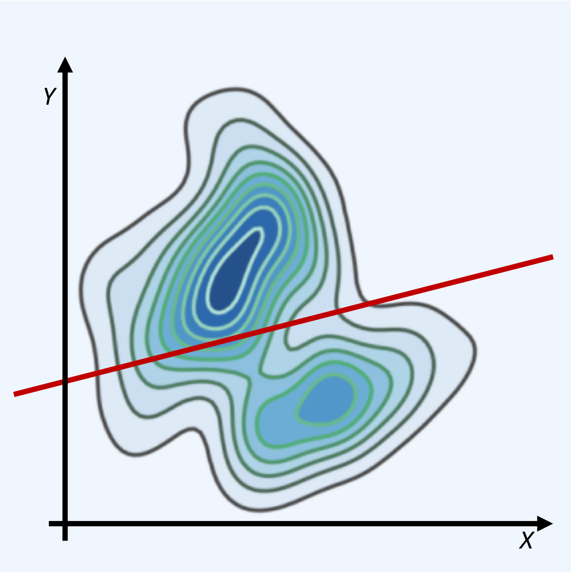
A Linear Model Example

Brunch in Berkeley

$$\widehat{\text{Avocados}} = 2 + 1 \cdot \text{Lemons} + 2 \cdot \text{Loaves\_Bread}$$

Review: Outcome, Prediction, and Error

Review: Outcome, Prediction, and Error



Concept Check: Making Predictions with a Linear Model

Concept Check: Making Predictions with a Linear Model

- Students will be given a model and (x,y) and compute prediction and error.

Metric Inputs

Interpreting Model Weights (Coefficients)

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Interpretation of coefficients

- If X_i changes by ΔX_i units, the predicted value of the target, $\hat{Y}$ changes by $\beta_i \cdot \Delta X_i$ units.
- If X_i and X_j change by ΔX_i and ΔX_j respectively, then the predicted value of the target, $\hat{Y}$ changes by $(\beta_i \Delta X_i) + (\beta_j \Delta X_j)$.

Ceteris paribus: all else equal

Interpreting Model Coefficients: Example

Does this model say peacocks with longer tails fly slower?

$$\text{air\_speed} = 4.3 - 1.2 \cdot \text{tail\_length} + 0.8 \cdot \text{muscle\_mass}$$

Photo by Thimindu Goonatillake CC BY-SA 2.0

Categorical Inputs

Categorial Inputs, Part I

What if, rather than *numeric* inputs, we had *categorical* inputs?

- Information that says it belongs to one category or another, but doesn't provide a value to that category?
- **Reminder**: There are four "levels" of information – (1) Categorical; (2) Ordinal; (3) Interval; (4) Ratio

Categorical Inputs, Part II

How is this represented in a model?

- The model aims only to distinguish one category from another, and so switches from stacked *labels* to *one-hot encoded*, or *dummy* variables.
- Practically, in this model, one of the labels is identified as the *baseline*, *default*, or *omitted* category
- Indicators for *alternate* levels mark changes from the baseline to the alternate level.

Categorical Inputs, Part II

Learnosity: Interpreting Model Weights

This is a placeholder for a Learnosity activity.

- have to change this, so it's about interpretation, not prediction
- In this activity, students will be given a fitted model that conforms with the data that they have for the peacock
- They will make predictions *first* from the data, and then
- Second, from newly created data to see how the predictions change

Part 2: Selecting a Linear Model with OLS

OLS is Regression for Estimating the BLP

Fitting Linear Models

Linear regression: an algorithm for fitting a linear model given a sample of data

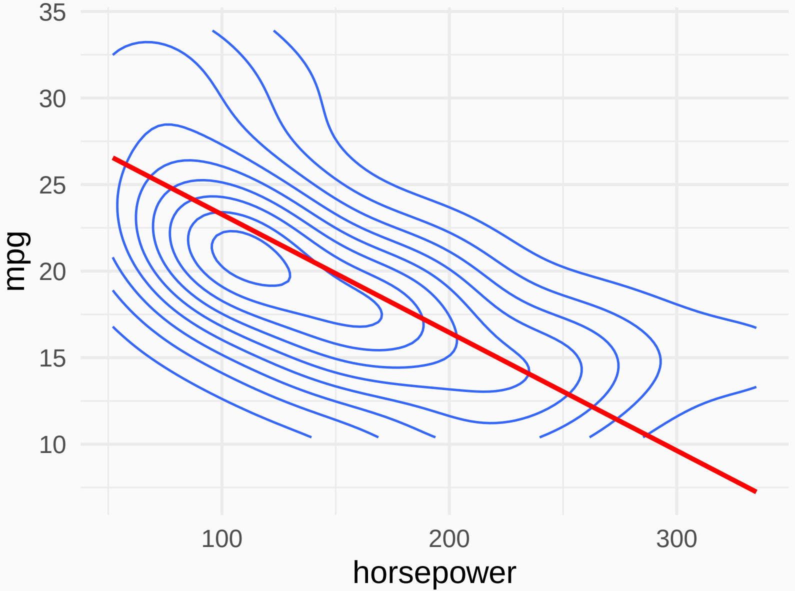
- Ordinary least squares (OLS) regression
- Quantile regression
- Regularized regression
    - Lasso
    - Ridge regression

Understanding OLS

Ordinary least squares (OLS) regression

- The most well-known type of linear regression
- A foundation for many other types of regression
- Key goal: estimating the best linear predictor (BLP)

The BLP Minimizes Expected Squared Error



The Best Linear Predictor

The BLP (population regression function)

The best linear predictor is defined by the function

$$g(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)$ are chosen to minimize the expected squared error.

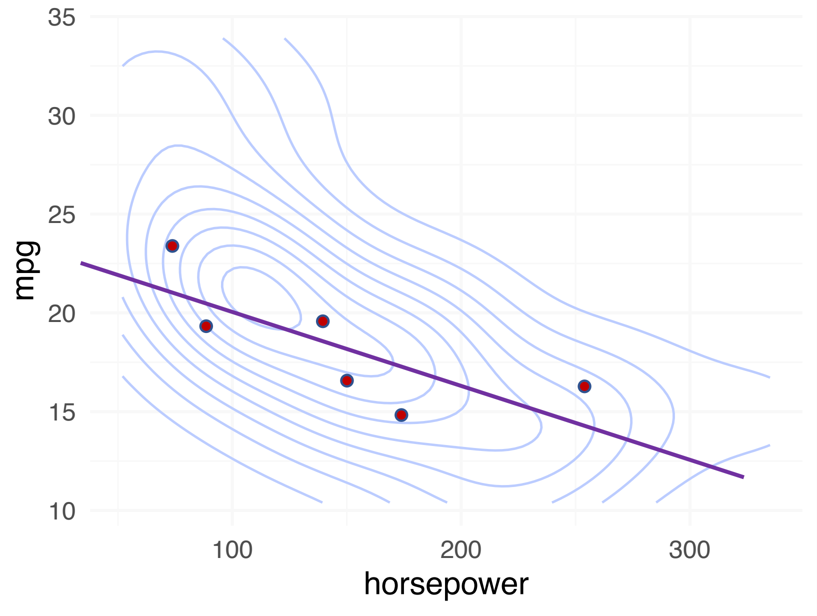
$$\min_{(b_0,...,b_k)} \mathrm{E}\left[(Y - (b_0 + b_1 X_1 + ... + b_k X_k))^2\right]$$

Great Things About the BLP

The best linear predictor...

- minimizes MSE out of all linear models.
- captures an infinitely complex distribution in a few parameters.
- can be estimated with much less data compared to a probability density.
- is easy to reason about.
- is easy to communicate to others, helping knowledge advance.
- has a closed form solution that is relatively easy to work with.

Applying the Plug-In Principle



OLS Regression Is the BLP Plug-In Estimator

Plug-in strategy

The OLS regression line is given by

$$\hat{g}(\mathbf{X}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$ are given by

$$\min_{(b_0,\dots,b_k)} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - (b_0 + b_1 X_{[1]i} + \dots + b_k X_{[k]i})\right)^2.$$

An Analogy with the Mean

Given only Y	Given Y and X
E[Y] minimizes MSE out of all numbers.	the BLP minimizes MSE out of all linear models.
We can't compute E[Y] without knowing the distribution.	We can't compute the BLP without knowing the distribution.
$\bar{X}$ is the plug-in estimator for E[Y].	OLS is the plug-in estimator for the BLP.

Coming Up Soon...

We still have to:

- solve the minimization problem
- show that OLS is *consistent* for the BLP

Learnosity: You Minimize It!

Learnosity: You Minimize It!

Note: This is a Learnosity Activity. We're just placing it here for organization.

This is the activity that is currently coded

`regression_fit_2d_exercise/`. **Note that we would like to expand this to ask students to work through several examples in a row. Presently we have a single example made; expanding this to a broader set is relatively easy. In the expanded set, we should ensure that we have some complexity in the data—e.g., a sine curve.**

Reading: OLS Regression Estimates the BLP

Reading: Linear Regression is a Plug-in Estimator for the BLP

Note: this is a reading call, we're just placing it here for organization.

Read pages 143–147 of *Foundations of Agnostic Statistics*.

Choosing Assumptions for OLS Regression

When Does OLS "Work"?

The OLS regression line is

$$\hat{g}(\mathbf{X}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$$

where $\hat{\beta}$ are chosen to minimize squared residuals.

When Does OLS "Work"?

The OLS regression line is

$$\hat{g}(\mathbf{X}) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$$

where $\hat{\beta}$ are chosen to minimize squared residuals.

Different assumptions

$$\Downarrow$$

Different statistical guarantees

More data $\Longrightarrow$ less restrictive assumptions

More data $\Longrightarrow$ easier to assess assumptions

The large-sample model (not an official name)

Just two assumptions:

- I.I.D.
- Unique BLP exists

Asymptotic behavior as $n \to \infty$ provides considerable guarantees.

OLS in a Small Sample

The classical linear model

- A parametric model—fully specifies $f_{Y|X}$
- Traditional starting point for regression
- Even with extensive transformations, may be hard to justify assumptions

Guarantees come from strict assumptions.

OLS in Very Small Samples

Special difficulties when $n <\sim 15$

- No help from asymptotics
- Not enough data to assess CLM

Randomization inference

- A framework for testing (restrictive) null hypotheses

Rules of Thumb for OLS Assumptions

Rules of thumb

In general, you might reason about data and regression models in the following way.

Sample size	Required assumptions
$100 \leq n$	Large-sample linear model
$15 \leq n < 100$	Classical linear model
$5 \leq n < 15$	Randomization inference

The Bivariate OLS Solution

Existing Content (in distinct format)

9.5 Deriving the Bivariate OLS Estimators

Consistency of Bivariate OLS Under the Large-Sample Model

Continuous mapping theorem: Let $(S^{(1)}, S^{(2)}, S^{(3)}, \ldots)$ and $(T^{(1)}, T^{(2)}, T^{(3)}, \ldots)$ be sequences of random variables. Let $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ be a function that is continuous at $(a, b) \in \mathbb{R}^2$. If $S^{(n)} \xrightarrow{p} a$ and $T^{(n)} \xrightarrow{p} b$, then $g(S^{(n)}, T^{(n)}) \xrightarrow{p} g(a, b)$

Assumptions: 1) I.I.D. 2) Unique BLP exists ($\mathrm{V}(X) > 0$)

Continuous mapping theorem: Let $(S^{(1)}, S^{(2)}, S^{(3)}, \ldots)$ and $(T^{(1)}, T^{(2)}, T^{(3)}, \ldots)$ be sequences of random variables. Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a function that is continuous at $(a, b) \in \mathbb{R}^2$. If $S^{(n)} \xrightarrow{p} a$ and $T^{(n)} \xrightarrow{p} b$, then $g(S^{(n)}, T^{(n)}) \xrightarrow{p} g(a, b)$

Assumptions: 1) I.I.D. 2) Unique BLP exists ($\mathrm{V}(X) > 0$)

$$
\begin{aligned}
x^{(n)} &= (x_1, x_2, \ldots, x_n) \quad y^{(n)} = (y_1, y_2, \ldots, y_n) \\
S^{(n)} &= \widehat{\mathrm{cov}}(x^{(n)}, y^{(n)}) \quad T^{(n)} = \widehat{\mathrm{V}}(x^{(n)}) \\
\hat{\beta}_1^{(n)} &= S^{(n)}/T^{(n)} \\
S^{(n)} &\xrightarrow{p} \mathrm{cov}[X, Y], \qquad T^{(n)} \xrightarrow{p} \mathrm{V}[X] \\
g(c, d) &= c/d \text{ is continuous where } d \neq 0 \\
\hat{\beta}_1^{(n)} &= g(S^{(n)}, T^{(n)}) \xrightarrow{p} g(\mathrm{cov}[X, Y], \mathrm{V}[X, Y]) = \beta_1
\end{aligned}
$$

The Matrix Formulation of a Linear Model

The Matrix Formulation of a Linear Model

- Insert content from previous version of course: 10.7 Matrix Form of the Linear Model
- This content leads into the next lightboard of the derivation of the OLS normal equations

Reading: The Matrix Solution For OLS Regression

Reading: The Matrix Solution For OLS Regression

Read section 4.1.3, which is on pages 147 - 151.

The Multiple OLS Solution

The Multiple OLS Solution

- Pull in the lightboard called *Matrix Derivation of the OLS Estimator*.
- In the next iteration of the course, pull in a geometric derivation of the OLS coefficients.

Sample Moment Conditions

Review: Population Moment Conditions

Population: Let ϵ represent error from the BLP.

Version 1: $\mathrm{E}[\epsilon] = 0, \mathrm{E}[X_j \epsilon] = 0$ for all j.

Version 2: $\mathrm{E}[\epsilon] = 0, \mathrm{cov}[X_j, \epsilon] = 0$ for all j.

Sample Moment Conditions

Sample: Let $\boldsymbol{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$ represent OLS residuals.

Sample Moment Conditions

Sample: Let $\boldsymbol{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$ represent OLS residuals.

$\mathbb{X}^T \boldsymbol{Y} = \mathbb{X}^T \mathbb{X} \beta, \qquad 0 = \mathbb{X}^T (\boldsymbol{Y} - \mathbb{X}\beta) = \mathbb{X}^T \boldsymbol{e}$

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ X_{[1]1} & X_{[1]2} & \dots & X_{[1]n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{[k]1} & X_{[k]2} & \dots & X_{[k]n} \end{bmatrix} \cdot \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

$\sum e_i = 0$, $\sum X_{[j]i} e_i = 0$. or $\widehat{\text{cov}}(\boldsymbol{X}_{[j]}, \boldsymbol{e}) = 0$

Consistency of Multiple OLS

Consistency of Multiple OLS

Assumptions: 1) I.I.D. 2) Unique BLP exists

In population: $\beta = \mathrm{E}[X^T X]^{-1} \mathrm{E}[X^T Y]$

Consistency of Multiple OLS

Assumptions: 1) I.I.D. 2) Unique BLP exists

In population: $\beta = \mathrm{E}[X^T X]^{-1} \mathrm{E}[X^T Y]$

$$\hat{\beta}^{(n)} = \left(\frac{1}{n}\mathbb{X}^{(n)T}\mathbb{X}^{(n)}\right)^{-1}\frac{1}{n}\mathbb{X}^{(n)T}\boldsymbol{Y}^{(\boldsymbol{n})}$$

$$\frac{1}{n}\mathbb{X}^{(n)T}\boldsymbol{Y}^{(\boldsymbol{n})} = \frac{1}{n}\begin{bmatrix} | & | & & | \\ x_1 & x_2 & \dots & x_n \\ | & | & & | \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} x_i^T y_i$$

$$\frac{1}{n}\mathbb{X}^T\mathbb{X} = \frac{1}{n}\begin{bmatrix} | & | & & | \\ x_1 & x_2 & \dots & x_n \\ | & | & & | \end{bmatrix}\begin{bmatrix} — & x_1 & — \\ — & x_2 & — \\ & \vdots & \\ — & x_n & — \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} x_i^T x$$

Consistency of OLS

$$\text{WLLN} \implies \frac{1}{n}\sum_{i=1}^{n} x_i^T x_i \xrightarrow{p} \mathrm{E}[X^T X] \qquad \frac{1}{n}\sum x_i^T y_i \xrightarrow{p} \mathrm{E}[X^T Y]$$

$$\text{CMT} \implies \hat{\beta} = \beta + (\frac{1}{n}\sum_{i=1}^{n} x_i^T x_i)^{-1}(\frac{1}{n}\mathbb{X}^T\epsilon) \xrightarrow{p} \beta + 0 = \beta$$

Unique Variation and Regression Anatomy

How Can We Understand a Specific $\hat{\beta}_i$?

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{Y}$$

Partialling Out

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_k X_k$$

Step 1: Regress X_1 on other Xs

$$\hat{X}_1 = \hat{\delta}_0 + \hat{\delta}_2 X_2 + ... + \hat{\delta}_k X_k + r_1$$

Step 2: Regress Y on the residuals from Step 1

$$\hat{Y} = \hat{\gamma}_0 + \hat{\beta}_1 r_1$$

Regression anatomy: $\hat{\beta}_1 = \frac{\widehat{\text{cov}}(\boldsymbol{Y}, \boldsymbol{r_1})}{\widehat{\text{V}}(\boldsymbol{r_1})}$

Deriving the Regression Anatomy Formula

Deriving the Regression Anatomy Formula

Use content from old course:

10.5 Regression Anatomy

Segment for Consideration: Applying the Regression Anatomy Formula

Applying the Regression Anatomy Formula

Consider using the old concept check: 10.6 Applying the Regression Anatomy Formula

Interpreting Model Coefficients: Warnings

$$\widehat{Wage} = \beta_0 + \beta_1 Age + \beta_2 Birth\_Year$$

What does it mean to hold *Age* constant while increasing *Birth_Year*?

Evaluating the Large-Sample Linear Model

The Large-Sample Linear Model

- I.I.D. data
- A unique BLP exists

What Does I.I.D. Mean?

Imagine selecting each new datapoint...

- from the same distribution
- with no memory of any past datapoints

Common Violations of Independence

- Clustering
  - Geographic areas
  - School cohorts
  - Families
- Strategic Interaction
  - Competition among sellers
  - Imitation of species
- Autocorrelation
  - One time period may affect the next

How can observing one unit provide information about some other unit?

A Unique BLP Exists

A BLP exists:

- $\text{cov}[X_i, X_j]$ and $\text{cov}[X_i, Y]$ are finite (no heavy tails)

The BLP is unique:

- No perfect collinearity
- $E[X^T X]$ is invertible

$\implies$ No X_i can be written as a linear combination of the other X's.

Perfect Collinearity Example 1

$$\widehat{Price} = .5\ Donuts + 0.0\ Dozens$$

or

$$\widehat{Price} = 0.0\ Donuts + 6.0\ Dozens$$

Perfect Collinearity Example 2

$$\widehat{Voters} = 200\, Positive\_Ads + 100\, Negative\_Ads + 0\, Total\_Ads$$

or

$$\widehat{Voters} = 100\, Positive\_Ads + 0\, Negative\_Ads + 100\, Total\_Ads$$

Goodness of Fit

Goodness of fit: How well does a model fit the data?

- R^2
- Adjusted R^2
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)

Breaking Down Variance

Total variance = explained variance + residual variance

Defining R^2

$$R^2 = 1 - \frac{\hat{\mathbb{V}}(\hat{\epsilon})}{\hat{\mathbb{V}}(\boldsymbol{Y})} = 1 - \frac{\text{residual variance}}{\text{total variance}}$$

$$\text{For OLS: } R^2 = \frac{\text{explained variance}}{\text{total variance}}$$

How much of the variation in the outcome does the model explain?

R is Correlation

$$R^2 = \frac{\hat{\mathbb{V}}(\hat{\boldsymbol{Y}})}{\hat{\mathbb{V}}(\boldsymbol{Y})}$$

Understanding Sums of Squares

Total sum of squares: $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

Explained sum of squares: $\text{ESS} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

Residual sum of squares: $\text{RSS} = \sum_{i=1}^{n}(\hat{\epsilon}_i)^2$

$$\text{For OLS: TSS} = \text{ESS} + \text{RSS}$$

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Things to Remember About R^2

- Adding variables always makes R^2 go up.
- With many variables, consider alternatives.
    - Adjusted R^2
- R^2 is not a measure of practical significance.
    - For example, regress hospital admissions on being shot
- A low R^2 is a negative, but assess it in context.

Measuring Uncertainty in Regression Estimates

Introduction to Measuring Uncertainty

- You've run your regression, and produced your BLP estimate.
- How much *could* the coefficients that you are reporting change if you had drawn a different sample?

Example: Becomming a Man

- **Becomming a Man** is a randomized controlled trial that recruited 7,500 male youth in Chicago. [link]
- The program, among many other efforts, taught to think about circumstances before engaging.
- Youth randomized to receive this treatment were:
    - 28-35% less likely to be arrested for any reason
    - 45-50% less likely to be arrested for violent crimes.
    - 12-19% more likely to graduate high-school

Example: Predicting School Achievement

- Gellman, Hill and Vehtari (2020) write about 343 Portuguese high-school students drawn from across the country.
- Predict final-year math grade (as a percentage) and use a large number of predictors:
    - **Demographic Features**: family size, health, parent's education
    - **Past Education Features**: past-class failures, educational support, paid classes
    - **Educational Cost Features**: home-to-school travel time, home internet access
    - **Social Features**: going out with friends, weekday and weekend alcohol consumption

Describing Uncertainty as "Resolution"

- Measured data is a point-in-time, limited view of a complex reality.
- With limited data, we can only make out "fuzzy" patterns
- With more data, we can see more detail
- With different data, we can see different details

Describing Uncertainty as "Resolution"

Describing Uncertainty as "Resolution"

Review: Review: Statistics, Distributions, and Uncertainty

Reading: Robust Standard Errors

Reading: Robust Standard Errors

Read section 4.2 Inference, stopping at the bottom of page 153.

Robust Standard Errors

Robust Standard Errors, Part I

- Regression coefficients are random variables
- As such, they have sampling variance, just as any other realized random variable
- Recall you've used the sampling variance of the sample mean, $\hat{V}[\overline{X}]$, and the related concept, the standard error of the sample mean, $\hat{\sigma}[\overline{X}] = \sqrt{\hat{V}[\overline{X}]}$.
- Here, we introduce the equivalent concept in OLS regression: $\hat{\sigma}[\hat{\beta}_k] = \sqrt{\hat{V}[\hat{\beta}_k]}$

Robust Standard Errors, Part II

- On the bottom of page 152, the authors seem to, from nowhere, divine the statement

$$\left(E[X^TX]\right)^{-1} E[\epsilon^2 X^TX] \left(E[X^TX]\right)^{-1}$$

- Let's slow that down just a little bit.

Robust Standard Errors, Part III

Robust Standard Errors, Part IV

- What information is contained in: $X^T X$?

Robust Standard Errors, Part V

- How do changes in data *shape* increase or decrease the standard errors of regression coefficients?

Hypothesis Tests

Testing Improvement in a Model

Testing Improvements in a Model: The F-Test

Testing Individual Coefficients

Hypothesis Tests for OLS

Are the relationships we see in our regression true of the population, or just consequences of sampling variation?

Most common tests:

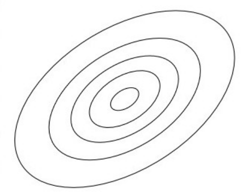
- Testing single coefficients
    - Usually $H_0 : \beta_i = 0$
- Testing multiple coefficients
    - Usually $H_0 : \beta_i = \beta_j = ... = 0$

Asymptotic Normality of OLS

Theorem

When the data points are I.I.D. from a distribution with unique BLP. $\sqrt{n}(\hat{\beta} - \beta)$ converges in distribution to a multivariate normal distribution.

- Each β_i is asymptotically normal.
- Linear combinations (e.g., $\beta_i - \beta_j$) are asymptotically normal.



Testing a Single OLS Coefficient

Large-sample testing procedure

Let s_i be the robust standard error estimate for β_i.
Then, under $H_0 : \beta_i = \mu_0$,

$$z = \frac{\hat{\beta}_i - \mu_0}{s_i}$$

is asymptotically distributed $N(0, 1)$.

- Computed automatically in statistical software

The t-Test for OLS Coefficients

In practice, we call the statistic t and test against a T-distribution.

- Asymptotically, the Z and T distributions are equal.
- Under the classical linear model assumptions, the T distribution is *exact* for small samples.

Practical Significance in OLS

Is the relationship between X_i and Y of a magnitude we should care about?

Two common effect size measures:

- $\hat{\beta}_i$
- Coefficient after standardizing X_i or Y

Practical Significance Example 1

$$\widehat{Price} = 82,213 + 25,134\ Bedrooms + 8\ Gnomes$$

"Garden gnomes have a statistically significant relationship with house price ($t = 2.1$, $p = .03$)."

Practical Significance Example 1

$$\widehat{Price} = 82,213 + 25,134\ Bedrooms + 8\ Gnomes$$

"Garden gnomes have a statistically significant relationship with house price ($t = 2.1$, $p = .03$)."

"The predicted price only changes by \$8 per garden gnome, a negligible amount compared to the total house price. For comparison, the model predicts that it would take over 3,000 gnomes to match the effect of a single bedroom."

Practical Significance Example 2

$$\widehat{Agility} = 132 + 3.4\,Sleep\_Hours$$

"We found evidence that more sleep is related to a higher agility score ($t = 3.8$, $p < .001$)."

"Each extra hour of sleep is associated with an extra 3.4 points."

Practical Significance Example 2 (cont.)

Let $S\_Agility = \frac{Agility - \overline{Agility}}{\sqrt{\widehat{var}(Agility)}}$

$$\widehat{S\_Agility} = -1.21 + 0.92\ Sleep\_Hours$$

"Each extra hour of sleep is predicted to increase agility by 0.92 standard deviations."

Testing Multiple Coefficients

Testing Multiple Coefficients

	Dependent Variable: Crimes per 1000
Density	8.414***
	(1.140)
Federal Wage	0.027
	(0.030)
Service Wage	−0.008
	(0.007)
Manufacturing wage	0.003
	(0.019)
Intercept	10.768
	(11.964)

$$\text{Full model: } \textit{Crime} = \widehat{\beta_0} + \hat{\beta}_1 \textit{Density} + \hat{\beta}_2 \textit{Fed\_Wage} + \hat{\beta}_3 \textit{Ser\_Wage} + \hat{\beta}_4 \textit{Man\_Wage} + \hat{\epsilon}_f$$

$$\text{Restricted model: } \textit{Crime} = \hat{\beta}_0 + \hat{\beta}_1 \textit{Density} + \hat{\epsilon}_r$$

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$$

Idea: if H_0 is true, the full model should not be much better at explaining the outcome

Total variance = explained variance + error variance

$$f = \frac{df_f}{df_r - df_f} \frac{\mathsf{V}[\epsilon_r] - \mathsf{V}[\epsilon_f]}{\mathsf{V}[\epsilon_f]}$$

F-Test Results

	RSS	Df	Sum of Sq	F	Pr(>F)
Full	14526				
Wages	14924	-3	-397.57	0.7846	0.5057