Lecture 08: Statistical Inference for Regression

CS109A: Introduction to Data Science

Harvard University

■ **Course:** CS109A: Introduction to Data Science

■ **Lecture:** Lecture 08

■ **Instructor:** Pavlos Protopapas, Kevin Rader, Chris Gumb

■ **Objective:** Understanding uncertainty in regression coefficients through bootstrapping, building confidence intervals, evaluating predictor significance using t-tests and p-values, and distinguishing confidence intervals from prediction intervals

Contents

1 Introduction: Why Do We Need Inference?

Lecture Overview

So far, we've learned how to *fit* models and *predict* outcomes. But fitting a model gives us just one estimate—one set of coefficients based on one sample of data. How certain can we be about these numbers?

This lecture answers a critical question: **"How much should we trust our model?"**

Key Topics:

- **Accuracy of Estimates**: How precise are our $\hat{\beta}$ values?
- **Bootstrapping**: Simulating "parallel universes" to measure uncertainty
- **Confidence Intervals**: A range where the true β likely falls
- **Feature Importance**: Which predictors actually matter?
- **Statistical Significance**: Is the effect real or just random noise?
- **Prediction Intervals**: How uncertain are our predictions $\hat{y}$?

1.1 The Consultant Scenario

Professor Protopapas sets up a vivid scenario:

Example: The $10

Imagine you're a consultant who built a model for advertising:

$$\hat{y} = 1.01x + 0.05$$

Where x = TV advertising budget (in thousands), y = sales (in thousands).

Your interpretation: "For every $1,000 spent on TV ads, sales increase by $1,010. Net profit: $10 per $1,000 invested."

The question: If you go to your boss and ask for $10,000 for your analysis, will they pay?

The doubt: That coefficient 1.01 came from *one* sample of data. What if you had collected data on a different day? You might have gotten 1.03, or 0.98, or something completely different!

The core issue: How do we *quantify* and *communicate* our uncertainty?

1.2 Sources of Uncertainty

Before measuring uncertainty, let's understand where it comes from:

Definition: Two Types of Error

1. Irreducible Error (Aleatoric Error, ϵ)

- Inherent randomness in the system
- Even with perfect model, there's noise we can't eliminate
- Example: Same ad budget on different days yields different sales due to weather, competitor

actions, random human behavior

2. Reducible Error

- **Model misspecification**: We assumed linear but reality is curved
- **Limited samples**: We only observed one "realization" of reality
- Can be reduced with better models or more data

For this lecture: We bundle everything into ϵ (epsilon)—the error term that makes our $\hat{\beta}$ uncertain.

2 The Thought Experiment: Parallel Universes

2.1 What If We Could Repeat the Experiment?

To understand how much $\hat{\beta}$ varies, imagine this scenario:

The Parallel Universe Thought Experiment

Scenario: We know the true relationship $y = f(x) + \epsilon$. But due to error ϵ, each measurement is slightly different from the true value.

Process:

1. **Universe 1**: Collect data (with random error). Fit model. Get $\hat{\beta}^{(1)}$.
2. **Universe 2**: Collect new data (different random error). Fit model. Get $\hat{\beta}^{(2)}$.
3. **Universe 3**: Collect new data. Get $\hat{\beta}^{(3)}$.
4. ... repeat 100 times ...

Result: We get 100 different $\hat{\beta}$ values. We can plot their histogram!

Insight: If the histogram is **narrow**, our estimate is precise. If **wide**, it's uncertain.

Example: Visualizing the Spaghetti Plot

If we plot all 100 fitted regression lines from our parallel universes, we get a "spaghetti plot"—many lines clustered together.

- **Tight spaghetti**: All lines are similar → Low variance → Confident estimate
- **Wild spaghetti**: Lines spread everywhere → High variance → Uncertain estimate

(This is exactly what we saw in Lecture 07 when discussing bias-variance!)

2.2 The Problem: We Can't Actually Visit Parallel Universes

This thought experiment is beautiful, but impossible in practice. We only have **one** dataset. We can't go back in time and recollect data under different random conditions.

Solution: **Bootstrapping**—a clever trick to simulate parallel universes using only the data we have!

3 Bootstrapping: Creating "Parallel Universes"

3.1 The Core Idea

Definition: Bootstrapping

Bootstrapping is a resampling technique that creates many simulated datasets by randomly sampling *with replacement* from our original data.

Key insight: Our original dataset represents the "population" (as best we know it). By resampling from it, we create variations that mimic what we'd see in parallel universes.

3.2 The Ball-in-Bucket Analogy

Example: Understanding Sampling with Replacement

Setup: You have a bucket with 5 numbered balls: {1, 3, 5, 8, 9}

Goal: Create a new "parallel universe" dataset of size 5

Process (Sampling WITH Replacement):

1. Reach in, randomly grab ball #8. Record it. **Put it back.**
2. Reach in again. Grab ball #8 again (possible because we replaced it!). Record.
3. Grab ball #3. Record. Put back.
4. Grab ball #5. Record. Put back.
5. Grab ball #1. Record.

Result:

- Original: {1, 3, 5, 8, 9}
- Bootstrap sample: {8, 8, 3, 5, 1}

Notice:

- Ball #8 appears *twice*
- Ball #9 doesn't appear at all
- This is exactly what we want—random variation!

Why "With Replacement"?

If we sampled *without* replacement, we'd just get the original dataset back (in a different order). That wouldn't create any variation!

Sampling *with* replacement means each draw is independent, and we naturally get variations where some points appear multiple times and others don't appear at all.

3.3 The Full Bootstrap Procedure

Bootstrap Algorithm for Confidence Intervals

Input: Original dataset of size n

Procedure:

1. Set number of bootstrap samples: S (typically 1000-10000)
2. For $s = 1$ to S:
   (a) Create bootstrap sample: randomly select n points from original data *with replacement*
   (b) Fit regression model to this bootstrap sample
   (c) Record coefficients: $\hat{\beta}_0^{(s)}, \hat{\beta}_1^{(s)}, \ldots$
3. Now you have S values of each coefficient
4. Analyze the distribution of these values

Output: Distribution of $\hat{\beta}$ values from which we can compute confidence intervals

3.4 Building Confidence Intervals from Bootstrap

Once we have S bootstrap estimates of $\hat{\beta}_1$, we can build a confidence interval:

Definition: Percentile Method for Confidence Intervals

95% Confidence Interval:

1. Sort all S bootstrap $\hat{\beta}$ values from smallest to largest
2. Find the 2.5th percentile (lower bound)
3. Find the 97.5th percentile (upper bound)
4. The interval between these is your 95% CI

In Python:

```
lower = np.percentile(bootstrap_betas, 2.5)
upper = np.percentile(bootstrap_betas, 97.5)
confidence_interval = [lower, upper]
```

Example: Computing Bootstrap CI

You ran 1000 bootstrap samples and got 1000 values of $\hat{\beta}_1$.

After sorting: [11.50, 12.26, 12.81, ..., 15.21]

- 2.5th percentile (25th value): 12.80
- 97.5th percentile (975th value): 13.71

95% CI: [12.80, 13.71]

Interpretation: We are 95% confident that the true β_1 lies between 12.80 and 13.71.

3.5 Standard Error

Definition: Standard Error

The **standard error** (SE) of $\hat{\beta}$ is simply the **standard deviation** of the bootstrap distribution.

$$SE_{\hat{\beta}} = \text{std}(\hat{\beta}^{(1)}, \hat{\beta}^{(2)}, \ldots, \hat{\beta}^{(S)})$$

If we assume the distribution is approximately normal, we can approximate:

$$95\%\ \text{CI} \approx [\bar{\beta} - 2 \cdot SE, \quad \bar{\beta} + 2 \cdot SE]$$

where $\bar{\beta}$ is the mean of bootstrap estimates.

Standard Deviation vs. Standard Error

Standard Deviation (SD): Measures spread of *data points* around their mean

- How spread out are the y values in our dataset?

Standard Error (SE): Measures spread of an *estimate* across hypothetical samples

- How much would $\hat{\beta}$ vary if we re-did the study many times?

In bootstrapping, SE is calculated as the standard deviation of the bootstrap $\hat{\beta}$ values.

4 Evaluating Predictor Significance

Now that we can quantify uncertainty in $\hat{\beta}$, we can ask more sophisticated questions about which predictors matter.

4.1 The Naive Approach: Largest Coefficient

Example: Advertising Data - Three Predictors

Suppose we have three predictors: TV, Radio, Newspaper.
After bootstrapping, we get:

Predictor	Mean $\hat{\beta}$	Std Dev (SE)
Newspaper	0.10	0.10
TV	0.05	0.005
Radio	-0.05	0.10

Naive question: Which predictor is most important?
Naive answer: Newspaper! It has the largest $|\hat{\beta}|$.
But wait... Look at the uncertainty! Newspaper's true β could be anywhere from -0.10 to +0.30. TV's is tightly concentrated around 0.05.

4.2 The t-test: Signal-to-Noise Ratio

We need a metric that considers **both** the coefficient value (signal) and its uncertainty (noise):

Definition: The t-test Statistic ($\hat{t}$)

The **t-test statistic** measures how many standard errors the coefficient is from zero:

$$\hat{t} = \frac{\text{Mean}(\hat{\beta})}{\text{SE}(\hat{\beta})} = \frac{\bar{\beta}}{\sigma_{\hat{\beta}}}$$

Interpretation:

- Large $|\hat{t}|$: Coefficient is far from zero relative to uncertainty → Strong signal
- Small $|\hat{t}|$: Coefficient is close to zero relative to uncertainty → Weak/uncertain signal

Note: Professor Protopapas uses $\hat{t}$ ("t-hat") because the classical t-statistic includes $\sqrt{n}$, which we omit for simplicity since n is fixed.

Example: Revisiting Advertising Data with t-test

Using the data from before:

Predictor	**Mean** $\hat{\beta}$	**SE**	$\hat{t} = \frac{\bar{\beta}}{SE}$
Newspaper	0.10	0.10	**1.0**
TV	0.05	0.005	**10.0**
Radio	-0.05	0.10	**-0.5**

New ranking by $|\hat{t}|$:

1. TV: $|\hat{t}| = 10$ (most important!)
2. Newspaper: $|\hat{t}| = 1$
3. Radio: $|\hat{t}| = 0.5$

Even though TV has the *smallest* coefficient, it's the *most reliably* non-zero!

Feature Importance Rankings Change!

The California Housing Price dataset example:

By coefficient magnitude $|\bar{\beta}|$:

1. Average Bedrooms
2. Median Income
3. Average Rooms

By t-statistic $|\hat{t}|$:

1. **Median Income**
2. **House Age**
3. Latitude

The coefficient-based ranking can be misleading! Average Bedrooms has a large coefficient but high uncertainty, so it drops in the t-statistic ranking.

5 Statistical Significance: The p-value

5.1 The Key Question

We've found that Median Income has the highest $\hat{t}$ score. But there's still a nagging question:

"What if ALL my predictors are junk, and Median Income is just the 'least bad'?"

We need to test whether the observed effect is **real** or just **random chance**.

5.2 Hypothesis Testing Framework

Definition: Hypothesis Testing

Null Hypothesis (H_0): The predictor has *no* effect on the outcome.

- Mathematically: $\beta = 0$
- Any non-zero $\hat{\beta}$ we observed was pure luck/noise

Alternative Hypothesis (H_1): The predictor *does* have an effect.

- Mathematically: $\beta \neq 0$

Strategy: Assume H_0 is true. Calculate how "surprising" our observed $\hat{t}$ would be under this assumption.

5.3 The p-value Concept

Definition: p-value

The **p-value** is the probability of observing a test statistic *as extreme or more extreme* than what we actually observed, *assuming H_0 is true.*

$$p\text{-value} = P(|t_{\text{random}}| \geq |t^*_{\text{observed}}| \mid H_0 \text{ is true})$$

In plain English: If there were truly no relationship, how often would random data produce a $\hat{t}$ value this large (or larger)?

5.4 How to Calculate p-value

Example: The p-value Calculation Process

Step 1: Generate random data (no relationship between x and y)
Step 2: Fit a model and calculate $\hat{t}$
Step 3: Repeat many times to build a distribution of "random $\hat{t}$ values"
Step 4: See where your actual $\hat{t}$ falls in this distribution
Shortcut: This distribution is the well-known **Student's t-distribution**. We don't need to simulate—we can compute directly!
Visualization: Plot the t-distribution. Your p-value is the area in the "tails" beyond your observed

$\hat{t}$ value (both positive and negative tails, since we consider absolute values).

5.5 Interpreting p-values

p-value Interpretation Guide

Large p-value (e.g., $p = 0.50$):

- "If there were no real effect, there's a 50% chance we'd see this result by luck."
- This is very plausible under H_0.
- **Conclusion**: Cannot reject H_0. No evidence the predictor matters.

Small p-value (e.g., $p = 0.01$):

- "If there were no real effect, there's only a 1% chance we'd see this result by luck."
- This is very unlikely under H_0.
- **Conclusion**: Reject H_0. The predictor is **statistically significant**.

Convention: We use $\alpha = 0.05$ as the threshold.

- $p < 0.05$: Reject H_0 (significant)
- $p \geq 0.05$: Cannot reject H_0 (not significant)

Common p-value Misconceptions

WRONG: "$p = 0.03$ means there's a 3% probability that H_0 is true."

CORRECT: "$p = 0.03$ means *if* H_0 were true, we'd see results this extreme only 3% of the time."

The p-value is about the *data*, not the hypothesis!

6 Prediction Intervals vs. Confidence Intervals

Now we shift from uncertainty in *coefficients* to uncertainty in *predictions*.

6.1 The Spaghetti Plot of Predictions

Example: Visualizing Prediction Uncertainty

From our S bootstrap samples, we have S different models:

$$\hat{f}^{(1)}(x), \hat{f}^{(2)}(x), \ldots, \hat{f}^{(S)}(x)$$

Plotting all S regression lines gives us a "spaghetti plot" for predictions.

For any specific x value (e.g., TV budget = \$200K):

- We get S different predicted values
- We can compute a histogram of these predictions
- We can compute the 2.5th and 97.5th percentiles $\rightarrow$ **Confidence Interval for** $f(x)$

6.2 The Key Distinction

Very Important: Confidence Interval vs. Prediction Interval

Confidence Interval (CI): Uncertainty in the *mean response* $f(x)$

- "Where does the *average* sales fall for TV budget = \$200K?"
- Only accounts for uncertainty in $\hat{\beta}$

Prediction Interval (PI): Uncertainty in a *new individual observation* y

- "What sales will a *specific new store* get with TV budget = \$200K?"
- Accounts for uncertainty in $\hat{\beta}$ **AND** the irreducible error ϵ

Key fact: **Prediction Interval is ALWAYS wider than Confidence Interval!**

Why? Because an individual y includes noise ϵ on top of the mean:

$$y = f(x) + \epsilon$$

Example: CI vs. PI Example

For TV Budget = \$200,000:

95% Confidence Interval for $f(x)$: [\$16.5M, \$17.5M]

- "The *average* sales for stores with this budget is 95% likely to be in this range."

95% Prediction Interval for y: [\$14.0M, \$20.0M]

- "A *specific new* store with this budget is 95% likely to have sales in this range."
- Much wider because we added uncertainty from ϵ!

6.3 The Funnel Shape

Both CI and PI have a characteristic "funnel" or "hourglass" shape:

- **Narrowest at** $\bar{x}$: Near the center of our data, we have the most information
- **Widens at extremes**: Far from the center, small errors in slope get amplified

This is because the regression line "pivots" around the data center $(\bar{x}, \bar{y})$.

7 Model Comparison Summary: When to Use What

This lecture began with comparing our three main model types:

Aspect	Linear Reg.	Polynomial Reg.	kNN
Type	Parametric	Parametric	Non-parametric
Interpretability	High	Moderate	Low
Coefficient meaning	Clear	Complex	None
Computational cost	Low	Moderate	High

Table 1: *Model comparison summary*

Key points:

- **Linear regression**: Fast (closed-form solution), interpretable

- **Polynomial regression**: Flexible but design matrix grows quickly
- **kNN**: Computationally expensive (must compute distances to all training points for every prediction)

8 Quick Reference Summary

Lecture 08 Quick Reference Card

1. Bootstrapping

- Create S "parallel universe" datasets by sampling *with replacement*
- Fit model to each $\rightarrow$ get distribution of $\hat{\beta}$
- **Confidence Interval**: Use 2.5th and 97.5th percentiles
- **Standard Error**: Standard deviation of bootstrap $\hat{\beta}$ values

2. Feature Importance

- **Naive**: Rank by $|\hat{\beta}|$ (ignores uncertainty!)
- **Better**: Rank by $|\hat{t}| = |\bar{\beta}|/SE$ (signal-to-noise ratio)

3. Statistical Significance

- **Null hypothesis** H_0: $\beta = 0$ (no effect)
- **p-value**: Probability of seeing our result if H_0 is true
- $p < 0.05 \rightarrow$ Reject $H_0 \rightarrow$ Significant!

4. CI vs. PI

- **Confidence Interval (CI)**: Uncertainty in mean $f(x)$
- **Prediction Interval (PI)**: Uncertainty in individual $y = f(x) + \epsilon$
- **PI is always wider** (includes ϵ variance)

9 Common Questions and Answers

Q: How many bootstrap samples (S) should I use?

A: Typically 1,000-10,000. More is better but has diminishing returns. For rough estimates, 1,000 is fine. For precise confidence intervals, use 10,000+.

Q: How does bootstrapping relate to overfitting?

A: Great question! Bootstrapping helps us understand coefficient uncertainty, but doesn't prevent overfitting. You should still use regularization (Ridge/Lasso) and cross-validation. In fact, you can bootstrap a Ridge regression model—apply regularization within each bootstrap iteration.

Q: When should I use bootstrap CIs vs. analytical formulas?

A: Analytical formulas (which assume normality) are faster but make assumptions. Bootstrap is more general—it works even when data isn't normal. Professor Protopapas prefers bootstrap because it's "assumption-free." In practice, both give similar results for large samples.

Q: Why does the confidence/prediction interval have a funnel shape?

A: The regression line is best constrained at the data center $(\bar{x}, \bar{y})$. Away from the center, small errors in the slope get amplified. Think of it like a see-saw pivoting at the center—small tilts at the pivot become large movements at the ends.

Q: Is $p < 0.05$ a universal rule?

A: No! It's a convention that works in many contexts, but:

- Some fields (particle physics) use much stricter thresholds
- Multiple testing requires adjustments (Bonferroni correction)
- Effect size matters too—a tiny effect can be "significant" with enough data

10 Looking Ahead

This lecture introduced key concepts that will be developed further:

- Kevin Rader will cover the **probabilistic foundations** of these ideas
- Formal hypothesis testing with assumptions about the error distribution
- Connection to the t-distribution and degrees of freedom
- These concepts extend to classification (logistic regression) and beyond

The key takeaway: Always quantify and communicate uncertainty. A point estimate without a confidence interval is incomplete!