Prediction Intervals and Goodness of Fit

How Much Can Our Predictions Be Off, and How Well Does the Model Fit?

Jake Anderson

March 21, 2026

Where We Left Off

In Topic 8, we predicted food expenditure for a household earning $2,000/week:

$$\hat{y}_0 = 83.42 + 10.21 \times 20 = \$287.62$$

But how much could that prediction be off?

We built confidence intervals for **parameters** (β_1, β_2) in Topics 9–10. Now we need intervals for **predictions**.

And a broader question: does this model even **fit well**? Income alone is supposed to explain food spending, but how much variation does it actually capture?

$\Longrightarrow$ Today: prediction intervals, R^2, and what they tell us (and what they don't).

Outline

Outline

Two Questions About Prediction

When we predict $\hat{y}_0 = 287.62$ at $x_0 = 20$, we might want to know:

Question 1: What is the **average** food expenditure for all households earning $2,000/week?

- This asks about $E(y|x_0) = \beta_1 + \beta_2 x_0$
- Uncertainty comes only from estimating the line (Topic 9)

Question 2: What will **one particular** household earning $2,000/week actually spend?

- This asks about $y_0 = \beta_1 + \beta_2 x_0 + e_0$
- Uncertainty comes from estimating the line **and** from the individual error e_0

$\Longrightarrow$ Question 2 is harder. The answer is always **less precise**.

The Forecast Error

Define the **forecast error** as the difference between the actual outcome and our prediction:

$$f = y_0 - \hat{y}_0$$

Expanding (where b_1, b_2 are the OLS estimates):

$$f = \underbrace{(\beta_1 + \beta_2 x_0 + e_0)}_{\text{actual}} - \underbrace{(b_1 + b_2 x_0)}_{\text{predicted}}$$

$\mathrm{E}(f) = 0$, so $\hat{y}_0$ is an **unbiased** predictor of y_0.

Under our standard regression assumptions, $\hat{y}_0$ is the **Best Linear Unbiased Predictor** (BLUP): no other linear unbiased predictor has a smaller forecast error variance.

Variance of the Forecast Error

Predicting for **one person** is harder than predicting the **average**, because individuals are noisy. Even if we knew the regression line perfectly, any single household's spending still scatters around that line by e_0.

The variance of the forecast error captures both sources of uncertainty:

$$\text{Var}(f) = \sigma^2 \left[\underbrace{1}_{\substack{\text{individual} \\ \text{error } e_0}} + \underbrace{\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}_{\text{estimation uncertainty}} \right]$$

Compare this to the variance for $\text{E}(y|x_0)$ (confidence interval, Topic 9):

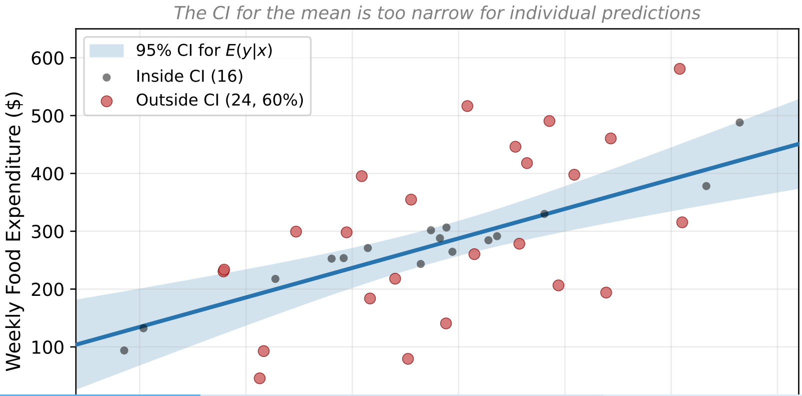
$$\text{Var}(\hat{y}_0) = \sigma^2 \left[\underbrace{\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}_{\text{estimation uncertainty only}} \right]$$

$\Longrightarrow$ The prediction variance has an **extra** σ^2 from the individual error e_0. This term does not shrink

Why the CI for the Mean Fails as a Prediction Interval

Suppose you told a client: "Based on our model, a household earning \$2,000/week spends between \$260 and \$315 on food."

That range is the **confidence interval for the average**. It is precise, but it captures where the *population mean* sits, not where *individual households* land.



The Prediction Interval

Replace σ^2 with $\hat{\sigma}^2$ and take the square root:

$$\text{se}(f) = \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]}$$

A $100(1-\alpha)\%$ **prediction interval** for y_0 is:

$$\boxed{\hat{y}_0 \pm t_c \cdot \text{se}(f)}$$

where t_c is the critical value from $t_{(N-2)}$.

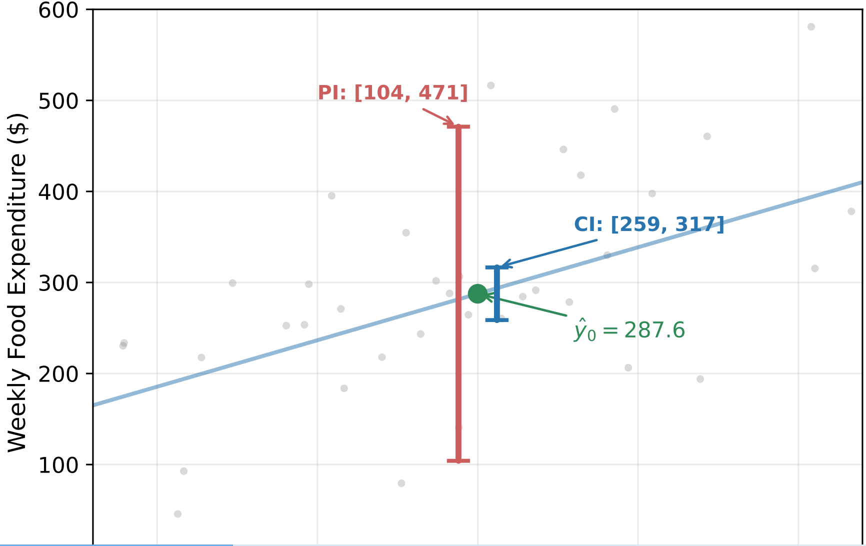
Food expenditure example ($x_0 = 20$, $N = 40$, $t_{0.975,38} = 2.024$):

$$\text{se}(f) = 90.63$$

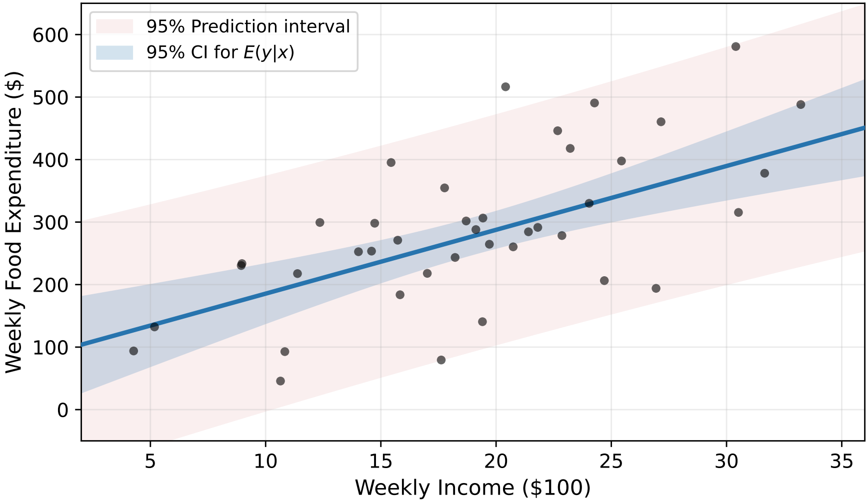
$$287.62 \pm 2.024 \times 90.63 = [\$104,\ \$471]$$

$\Longrightarrow$ We predict between $104 and $471. That range is enormous.

Why Prediction Intervals Are Wider Than CIs for the Mean

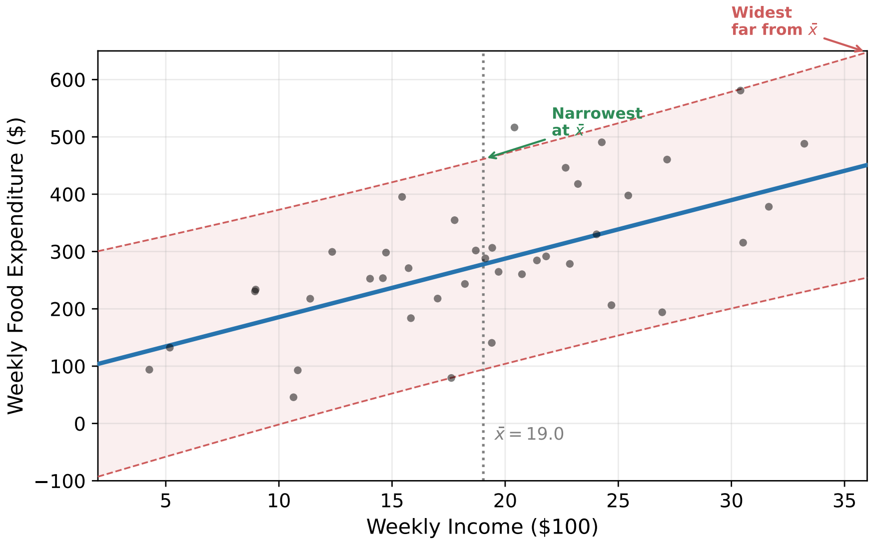


Both Intervals on the Same Plot



The blue band (CI for the mean) shrinks toward the line as $N \to \infty$. The red band (prediction

Predictions Are Best Near $\bar{x}$



What Makes the Forecast More Precise?

$$\text{Var}(f) = \sigma^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

Factor	**Change**	**Effect on** Var(f)
σ^2	Less noise in y	Smaller
N	More observations	Smaller
$\sum (x_i - \bar{x})^2$	More x-variation	Smaller
$(x_0 - \bar{x})^2$	Predict near $\bar{x}$	Smaller

Even with infinite data, the "1" in the brackets stays. That reflects the inherent variability of individual outcomes around the population regression line.

$\Longrightarrow$ Prediction intervals never collapse to a point. They converge to $\hat{y}_0 \pm t_c \cdot \sigma$ as $N \to \infty$.

Outline

How Well Does the Model Fit?

Our food model predicts between \$104 and \$471 for a \$2,000/week household. That seems imprecise.

A natural follow-up: **how much of the variation in food expenditure does income actually explain?**

To answer this, we decompose the total variation in y into two pieces:

- The part explained by the regression
- The part left unexplained (residuals)

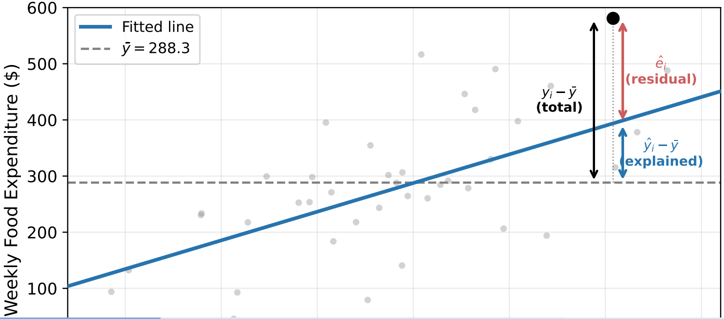
Decomposing Each Observation

Every observed value can be written as:

$$y_i = \hat{y}_i + \hat{e}_i$$

Subtract $\bar{y}$ from both sides:

$$\underbrace{y_i - \bar{y}}_{\text{total deviation}} = \underbrace{(\hat{y}_i - \bar{y})}_{\text{explained by } x} + \underbrace{\hat{e}_i}_{\text{unexplained}}$$



The Sum of Squares Identity

Square and sum over all observations:

$$\underbrace{\sum_{i=1}^{N}(y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^{N}\hat{e}_i^2}_{SSE}$$

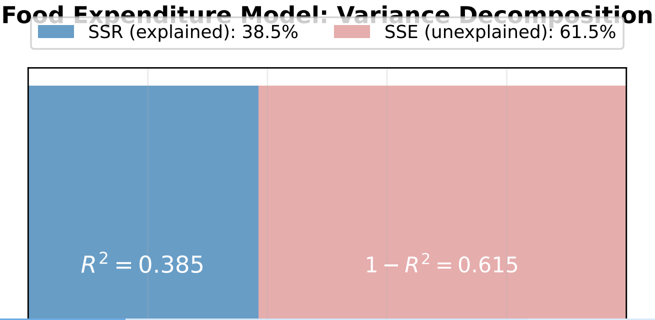
Term	Name	Measures
SST	Total Sum of Squares	Total variation in y about $\bar{y}$
SSR	Regression Sum of Squares	Variation explained by the model
SSE	Error Sum of Squares	Variation left unexplained

This identity requires an **intercept** in the model. The cross-product $\sum(\hat{y}_i - \bar{y})\hat{e}_i$ vanishes because the OLS normal equations guarantee $\sum \hat{e}_i = 0$ and $\sum x_i\hat{e}_i = 0$.

Food Expenditure: The Numbers

From the food expenditure regression:

$$SST = 495{,}132 \quad \text{(total variation)}$$
$$SSE = 304{,}505 \quad \text{(unexplained)}$$
$$SSR = SST - SSE = 190{,}627 \quad \text{(explained)}$$



Outline

Defining R^2

The **coefficient of determination** is the fraction of total variation explained:

$$\boxed{R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}}$$

Properties:

- $0 \leq R^2 \leq 1$ (always between 0 and 1)
- $R^2 = 1$: perfect fit (all points on the line, $SSE = 0$)
- $R^2 = 0$: x explains nothing ($SSR = 0$, fitted line is flat at $\bar{y}$)

For the food expenditure model:

$$R^2 = 1 - \frac{304{,}505}{495{,}132} = 0.385$$

R^2 and the Sample Correlation

In simple linear regression, R^2 has a direct link to the sample correlation coefficient r_{xy}:

$$R^2 = r_{xy}^2$$

where $r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$ measures the linear association between x and y.

For the food data: $r_{xy} = 0.62$, so $R^2 = 0.62^2 = 0.385$.

R^2 can also be expressed as:

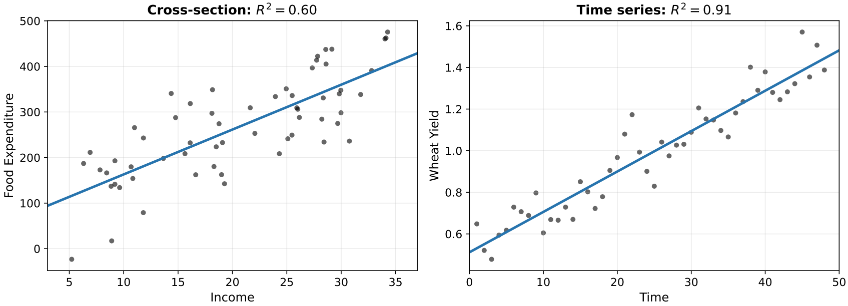
$$R^2 = \text{Corr}(y_i, \hat{y}_i)^2$$

the squared correlation between observed and fitted values. This version generalizes to multiple regression.

Is $R^2 = 0.385$ Good or Bad?

It depends on the context, not on some universal threshold.

Data type	Typical R^2 range
Cross-sectional (micro): wages, spending	$0.10 - 0.40$
Time series: GDP, yields	$0.60 - 0.95$
Experimental data	$0.70 - 0.99$



Limitations of R^2

R^2 measures "fraction of variation explained." It does **not** tell you whether:

1. **The model is correctly specified.** A high R^2 with a wrong functional form still gives misleading predictions.
2. **The coefficients are meaningful.** An insignificant slope ($t < 2$) in a model with $R^2 = 0.90$ from a strong time trend means x is not doing the explaining.
3. **Important variables are included.** $R^2 = 0.385$ tells you income matters, but it does not tell you what the missing 61.5% is.
4. **The model predicts well out of sample.** R^2 is in-sample. A model can overfit the observed data and fail on new data.

$\Longrightarrow$ Evaluate a model by its coefficients (signs, magnitudes, significance), not by R^2 alone.

Outline

Why Units Can Be Misleading

Two researchers study the same food expenditure data but measure income differently:

	Researcher A	Researcher B
Income units	$100/week	$/week
Reported slope	10.21	0.1021

At first glance, these look like contradictory results. One slope is 100 times larger.

But plug in the same household (earning $2,000/week):

$$\text{A:} \quad 83.42 + 10.21 \times 20 = 287.62$$

$$\text{B:} \quad 83.42 + 0.1021 \times 2{,}000 = 287.62$$

$\Longrightarrow$ Same prediction, same model. The coefficients just reflect different measurement units.

Scaling Rules

Rescale x by multiplying by c (i.e., $x^* = cx$):

Quantity	**Changes?**	**How?**
Slope b_2	Yes	Divided by c
$\text{se}(b_2)$	Yes	Divided by c
Intercept b_1	No	
t-statistics	No	
R^2	No	

In the food example: $x^* = 100 \cdot x$ (from \$100s to dollars), so the slope is divided by 100: $10.21/100 = 0.1021$.

Rescale y by multiplying by c (i.e., $y^* = cy$):

- All coefficients multiplied by c, all standard errors multiplied by c
- t-statistics and R^2: unchanged

$\Longrightarrow$ Scaling is a change in **units**, not a change in the relationship. Statistical conclusions (t-stats, p-values, R^2) are invariant to units.

Outline

What to Check When Someone Hands You Regression Output

1. How wide is the prediction interval?

- A point prediction alone is incomplete. The interval tells you how useful the prediction actually is.

2. What does R^2 mean in context?

- $R^2 = 0.385$ for cross-sectional micro data is typical. The same R^2 for experimental data would be a red flag.

3. Are the coefficients significant?

- A high R^2 with insignificant slopes means something else (a trend, a confound) is driving the fit.

4. Is the functional form appropriate?

- R^2 can be high even when the model is wrong. Residual plots catch what R^2 misses.

$\Longrightarrow$ Prediction intervals, R^2, coefficient significance, and residual diagnostics are complementary. No single number summarizes model quality.

Thank you!

jakeanderson@g.ucla.edu