Hypothesis Testing

Is the Relationship Real, or Just Noise?

Jake Anderson

March 21, 2026

Outline

1. The Question
2. Anatomy of a Hypothesis Test
3. Two-Sided Tests: Full Walkthrough
4. The p-Value
5. One-Sided Tests
6. Type I and Type II Errors
7. “Do Not Reject” $\neq$ “Accept”
8. Summary

Outline

Where We Left Off

From the food expenditure regression ($N = 40$, $df = 38$):

$$\widehat{\text{FOOD\_EXP}} = \underset{(43.41)}{83.42} + \underset{(2.09)}{10.21} \cdot \text{INCOME}$$

Last time we built a **95% confidence interval** for β_2:

$$10.21 \pm 2.024 \times 2.09 = [5.97,\ 14.45]$$

We are 95% confident that each additional \$100 of income raises food spending by between \$5.97 and \$14.45.

But what if someone asks a sharper question:

"Does income *really* affect food spending, or could $\beta_2 = 0$?"

Can You Tell Which Relationship Is Real?

Two regressions estimated from different datasets:

	Regression A	**Regression B**
Coefficient (b_2)	12.50	2.10
Std. Error ($\text{se}(b_2)$)	8.40	0.35

Regression A has the bigger coefficient. Does that make it the stronger evidence?

No. The coefficient alone tells you nothing about reliability. Regression A's estimate is noisy (b_2 could easily be zero), while Regression B's is precise.

$\Longrightarrow$ We need a formal measure of **evidence relative to noise**. That is what a hypothesis test provides.

From Estimation to Testing

A confidence interval tells us *where* β_2 plausibly lives.

A **hypothesis test** asks a yes/no question:

> *Is the data compatible with a specific claim about β_2, or should we reject that claim?*

Two questions we will answer today:

1. Does income affect food spending at all? (Is $\beta_2 = 0$?)
2. Does income raise food spending by more than $5.50 per $100? (Is $\beta_2 > 5.5$?)

$\Longrightarrow$ Same data, same estimates, but different questions yield different answers.

Outline

Five Ingredients

Every hypothesis test has five components:

1. **Null hypothesis** (H_0): the claim we put on trial
2. **Alternative hypothesis** (H_1): what we accept if we reject H_0
3. **Test statistic**: a number summarizing the evidence against H_0
4. **Rejection region** (or p-value): a decision rule
5. **Conclusion**: reject or do not reject H_0

Think of it as a trial: H_0 is "innocent until proven guilty." We need strong evidence to convict.

The Null and Alternative Hypotheses

Null hypothesis: $H_0 : \beta_k = c$, where c is a specific number.

We write β_k for any coefficient in the model; in our example, $k = 2$.

The null always contains an equality. It is our "default" belief until evidence says otherwise.

Alternative hypothesis: H_1 depends on the question:

Alternative	When to use
$H_1 : \beta_k > c$	Theory predicts a positive effect
$H_1 : \beta_k < c$	Theory predicts a negative effect
$H_1 : \beta_k \neq c$	We just want to know if β_k differs from c

Example: "Does income affect food spending?"

- $H_0 : \beta_2 = 0$ (income has no effect)
- $H_1 : \beta_2 \neq 0$ (income has *some* effect) $\leftarrow$ two-sided

The Test Statistic

If $H_0: \beta_k = c$ is true, substitute c for β_k in the t-formula:

$$t = \frac{b_k - c}{\text{se}(b_k)} \sim t_{(N-2)} \quad \text{if } H_0 \text{ is true}$$

(for simple regression with $K = 2$ parameters)

Interpretation:

- The numerator $b_k - c$ measures **how far** our estimate is from the null value
- The denominator $\text{se}(b_k)$ **scales** this distance by the estimation precision
- If H_0 is true, t should be close to zero (within the usual t-distribution spread)
- If H_0 is false, b_k will tend to be far from c, making $|t|$ large

$\Longrightarrow$ Large $|t|$ = evidence against H_0. But how large is “large enough”?

Outline

Testing $H_0: \beta_2 = 7.5$ (Consultant's Claim)

A consultant claims households spend $7.50 per additional $100 of income on food.

1. $H_0: \beta_2 = 7.5, \quad H_1: \beta_2 \neq 7.5$
2. $t = \dfrac{b_2 - 7.5}{\text{se}(b_2)} = \dfrac{10.21 - 7.5}{2.09} = 1.29$
3. $\alpha = 0.05$. Critical values: ± 2.024

Rejection rule (two-tail): Reject H_0 if $|t| \geq t_c$

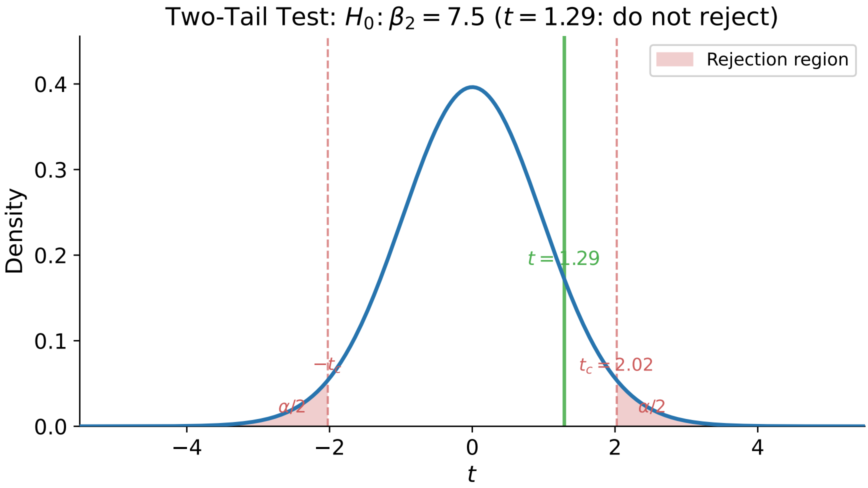
$|1.29| < 2.024 \implies$ **Do not reject H_0.**

$p = 2 \cdot P(t_{(38)} \geq 1.29) = 0.2033 > 0.05$. Same conclusion.

$7.5 \in [5.97,\ 14.45]$? Yes. Same conclusion again.

$\implies$ The data are consistent with $\beta_2 = 7.5$. But they are also consistent with $\beta_2 = 8.5$ ($t = 0.82$), or $\beta_2 = 6.5$ ($t = 1.77$). Not rejecting does **not** prove the null is true.

Visualizing a Non-Rejection



$t = 1.29$ falls in the non-rejection region. The estimate $b_2 = 10.21$ is close enough to 7.5, relative to

Testing $H_0 : \beta_2 = 0$ (The Significance Test)

Question: Does income affect food spending at all?

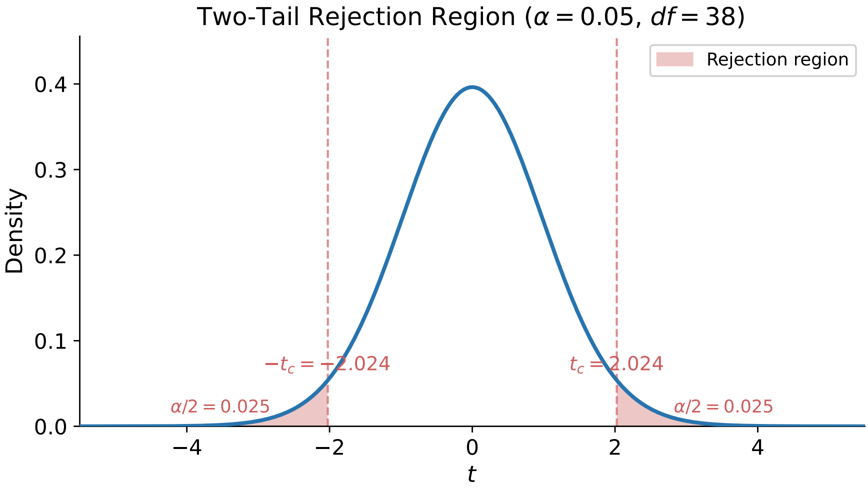
1. $H_0 : \beta_2 = 0, \quad H_1 : \beta_2 \neq 0$
2. Test statistic: $t = \dfrac{b_2 - 0}{\text{se}(b_2)} = \dfrac{10.21}{2.09} = 4.88$
3. $\alpha = 0.05$. Critical value: $t_c = t_{(0.975,\,38)} = 2.024$

Rejection rule (two-tail): Reject H_0 if $|t| \geq t_c$

$|4.88| = 4.88 \geq 2.024?$

Yes. Reject H_0. There is a statistically significant relationship between income and food expenditure.

Visualizing the Rejection Region



With $t = 4.88$, we land deep in the rejection region. The data are not compatible with $\beta_2 = 0$.

Three Equivalent Decision Methods

There are three ways to reach the same conclusion. All three always agree.

1. **Rejection region:** Reject if $|t| \geq t_c$

   $|4.88| \geq 2.024$? Yes $\implies$ Reject

2. ***p*-value:** Reject if $p \leq \alpha$

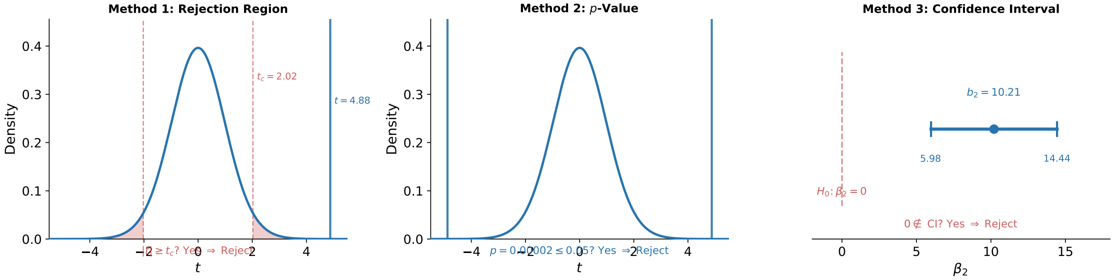
   $p = P(|t_{(38)}| \geq 4.88) = 0.00002 \leq 0.05$? Yes $\implies$ Reject

3. **Confidence interval:** Reject if the null value c falls outside the CI

   $0 \in [5.97,\ 14.45]$? No $\implies$ Reject

$\implies$ These are not three different tests. They are three windows onto the same **same test**.

Three Methods: Visual Comparison



All three panels represent the same test ($H_0 : \beta_2 = 0$, $\alpha = 0.05$). Each reaches "Reject" through a different lens: comparing t to t_c, comparing p to α, or checking whether 0 falls inside the CI.

Why Do the Three Methods Always Agree?

The logic is circular by construction:

- The CI is $b_k \pm t_c \cdot \text{se}(b_k)$
- The rejection region rejects when $|t| = |b_k - c| / \text{se}(b_k) \geq t_c$
- Rearranging: $|b_k - c| \geq t_c \cdot \text{se}(b_k)$

This is exactly the condition that c lies **outside** the interval $b_k \pm t_c \cdot \text{se}(b_k)$.

And the p-value $\leq \alpha$ precisely when $|t| \geq t_c$.

$\implies$ For a two-sided test at level α: reject via $|t|$, reject via p-value, reject via CI are all the same statement, rearranged.

Comparing the Two Tests

	Test 1: Consultant	Test 2: Significance
H_0	$\beta_2 = 7.5$	$\beta_2 = 0$
H_1	$\beta_2 \neq 7.5$	$\beta_2 \neq 0$
t	$\frac{10.21 - 7.5}{2.09} = 1.29$	$\frac{10.21 - 0}{2.09} = 4.88$
p-value	0.2033	0.00002
Decision	Do not reject	Reject

$\Longrightarrow$ The same b_2 and $\text{se}(b_2)$ produce opposite conclusions because the **distance to the null** differs. The t-statistic measures how many standard errors b_2 is from c.

Outline

What Does a p-Value Measure?

You have already used the p-value to make a decision. Now let's understand exactly what it measures.

Definition

The p-value is the probability of observing a test statistic **at least as extreme** as the one we calculated, **assuming H_0 is true**.

Small p = the observed t would be very unlikely under H_0
$\implies$ strong evidence against H_0

Large p = the observed t is not unusual under H_0
$\implies$ no reason to doubt H_0

Decision rule:

$$p \leq \alpha \implies \text{Reject } H_0 \qquad p > \alpha \implies \text{Do not reject } H_0$$

Computing the p-Value

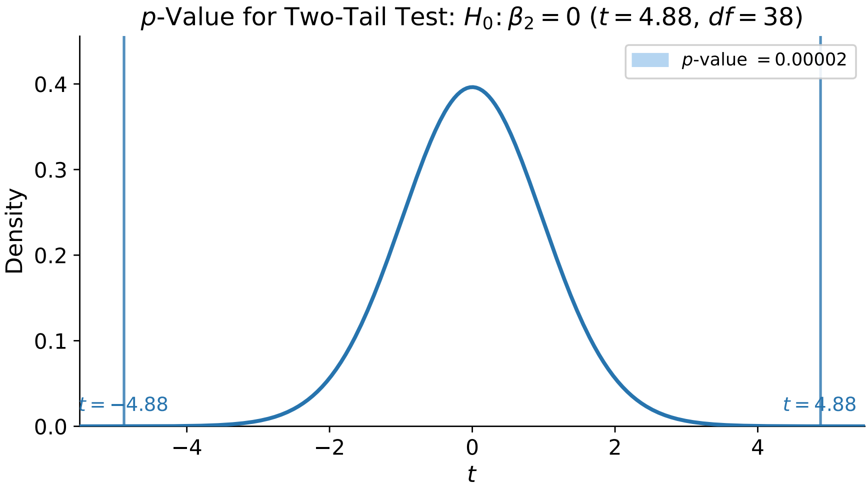
The direction of H_1 determines which tail(s) to measure:

Alternative	**p-value formula**	**Tail(s)**
$H_1 : \beta_k > c$	$p = P(t_{(N-2)} \geq t)$	Right
$H_1 : \beta_k < c$	$p = P(t_{(N-2)} \leq t)$	Left
$H_1 : \beta_k \neq c$	$p = 2 \cdot P(t_{(N-2)} \geq \lvert t \rvert)$	Both

Memory rule: The alternative points to the tail. Two-sided $\implies$ both tails $\implies$ double the one-tail probability.

Software regression output reports the two-tail p-value for $H_0 : \beta_k = 0$ by default. If you need a one-tail p-value: divide by 2, **but only when the sign of t agrees with your alternative** (e.g., $t > 0$ for $H_1 : \beta_k > 0$). If t points in the opposite direction, the one-tail $p > 0.50$ and you cannot reject.

p-Value for $H_0: \beta_2 = 0$ (Two-Tail)



$p = 0.00002$. The shaded area is so small it is barely visible. At any conventional α (0.01, 0.05, 0.10),

Outline

When Theory Points in One Direction

Sometimes economic theory tells us the sign of the effect:

- Income should *increase* food spending $\implies H_1 : \beta_2 > 0$
- A subsidy should *decrease* price $\implies H_1 : \beta_k < 0$

One-sided tests concentrate all α in a **single tail**, making it easier to reject in the predicted direction.

Right-tail test ($H_1 : \beta_k > c$):

$$\text{Reject } H_0 \text{ if } t \geq t_{(1-\alpha,\, N-2)}$$

Left-tail test ($H_1 : \beta_k < c$):

$$\text{Reject } H_0 \text{ if } t \leq -t_{(1-\alpha,\, N-2)}$$

Memory rule: The rejection region is in the direction of the inequality in H_1.

Right-Tail Example: $H_0 : \beta_2 \leq 5.5$

A supermarket chain will build a new store only if there is strong evidence that households spend more than $5.50 per $100 of additional income on food.

1. $H_0 : \beta_2 \leq 5.5, \quad H_1 : \beta_2 > 5.5$

   We write $\leq$ to indicate the region where we would fail to reject, but the test is conducted at the boundary $\beta_2 = 5.5$.

2. $t = \dfrac{10.21 - 5.5}{2.09} = 2.25$

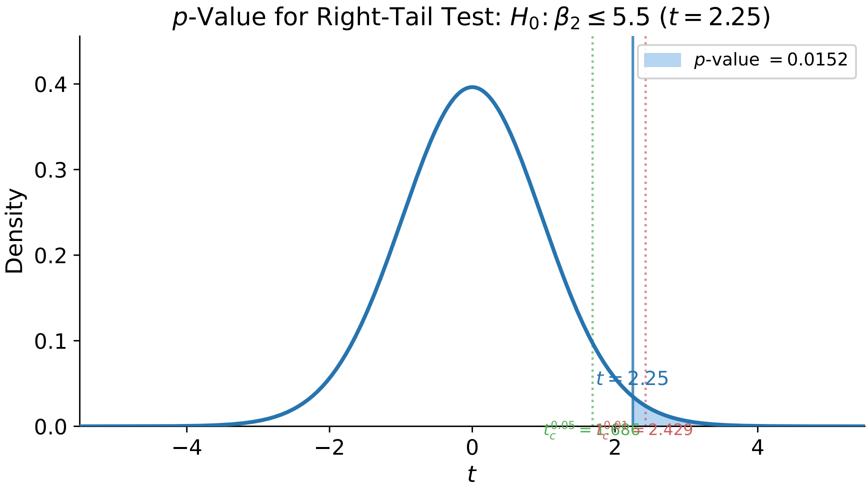
3. $\alpha = 0.01$. Critical value: $t_{(0.99,\,38)} = 2.429$. Reject if $t \geq 2.429$.

$2.25 < 2.429 \implies$ **Do not reject H_0.**

Insufficient evidence at the 1% level. The supermarket is not built.

Why $\alpha = 0.01$? The cost of a Type I error (building an unprofitable store) is high. A conservative α raises the evidence bar.

p-Value for the Right-Tail Test



$p = 0.0152$. At $\alpha = 0.01$: $0.0152 > 0.01$, do not reject. At $\alpha = 0.05$: $0.0152 < 0.05$, reject. The

One-Tail vs. Two-Tail: Comparison

Testing $H_0 : \beta_2 = 0$ with $\alpha = 0.05$:

	Two-Tail	**Right-Tail**
H_1	$\beta_2 \neq 0$	$\beta_2 > 0$
Critical value	± 2.024	1.686
p-value	$2 \times P(t \geq 4.88) = 0.00002$	$P(t \geq 4.88) = 0.00001$

Differences:

- One-tail has a **lower critical value** (1.686 vs. 2.024) $\Longrightarrow$ easier to reject
- One-tail p-value is **half** the two-tail p-value, but only when t has the sign predicted by H_1. If t points away from H_1 (e.g., $t < 0$ when $H_1 : \beta_k > 0$), you cannot reject regardless of α

$\Longrightarrow$ Use one-sided only when theory gives a clear directional prediction *before* seeing the data. Otherwise, use two-sided.

Outline

Two Ways to Be Wrong

	H_0 is actually true	H_0 is actually false
Reject H_0	Type I error	Correct
Do not reject H_0	Correct	Type II error

Type I error (false positive): We "convict an innocent" H_0.

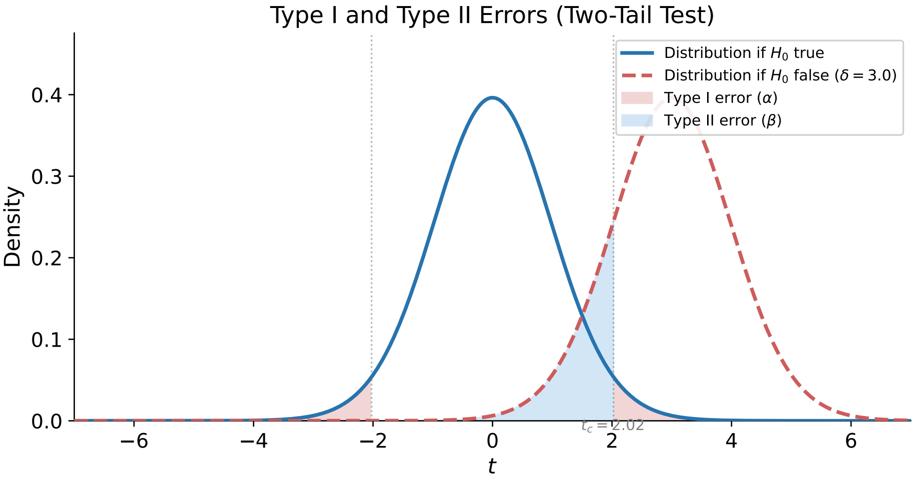
- $P(\text{Type I}) = \alpha$ (the significance level we choose)

Type II error (false negative): We "let a guilty H_0 go free."

- $P(\text{Type II}) = \beta$ (this β is unrelated to the regression coefficient β_k)
- Depends on the true parameter value; not directly controlled

Power $= 1 - \beta =$ probability of correctly rejecting a false H_0.

Type I and Type II: Visual



Lowering α (moving t_c outward) shrinks the Type I region but **enlarges** the Type II region. There is always a tradeoff.

Concrete Interpretation

In our food expenditure example, testing $H_0: \beta_2 = 0$ at $\alpha = 0.05$:

Type I error: We conclude income affects food spending when it actually does not.

- Probability: 5% (we chose this)

Type II error: We conclude the data are consistent with $\beta_2 = 0$ when income actually does affect spending.

- Probability: depends on the true β_2 and N
- If the true β_2 is very large, Type II error is unlikely (strong signal)
- If the true β_2 is close to zero, Type II error is likely (weak signal)

$\Longrightarrow$ "Do not reject" is weaker than "reject." It means the data cannot distinguish β_2 from zero, not that β_2 *is* zero.

The Tradeoff

If we lower α...	Type I	Type II
$\alpha = 0.10$	10%	Lower
$\alpha = 0.05$	5%	Moderate
$\alpha = 0.01$	1%	Higher

Stricter α demands more evidence to reject $\implies$ fewer false positives but more false negatives.

Recall the supermarket example ($\alpha = 0.01$): we chose a strict threshold because the cost of a false positive (building an unprofitable store) was high.

$\implies$ The choice of α should reflect the relative costs of the two types of errors in your specific context.

Outline

A Common Mistake

Wrong: "We accept H_0."
Correct: "We do not reject H_0."

Why does this distinction exist?

When we tested $H_0 : \beta_2 = 7.5$, we did not reject ($t = 1.29$, $p = 0.20$).

But we also would not reject $H_0 : \beta_2 = 8.5$ ($t = 0.82$), or $H_0 : \beta_2 = 6.5$ ($t = 1.77$), or any value inside $[5.97, \ 14.45]$.

$\Longrightarrow$ Failing to reject does not single out $\beta_2 = 7.5$ as the truth. It just says 7.5 is **one of many values** the data cannot distinguish from $b_2 = 10.21$.

Statistical vs. Economic Significance

"Statistically significant" and "economically meaningful" are different things.

Large sample, tiny effect:

- Suppose $b_2 = 0.001$ with $\text{se}(b_2) = 0.0001$
- $t = 10.0 \implies$ highly significant
- But the economic effect is negligible: $0.001 per $100 of income

Small sample, large effect:

- Suppose $b_2 = 15$ with $\text{se}(b_2) = 10$
- $t = 1.5 \implies$ not significant at $\alpha = 0.05$
- But $15 per $100 of income would be economically important

$\implies$ Always report the **magnitude** of b_k alongside the t-statistic. Significance tells you whether the effect is distinguishable from zero; it does not tell you whether the effect is large enough to care about.

Outline

Decision Flowchart

When you encounter a hypothesis test, follow these steps:

1. **Directional question?**
   - Yes (theory predicts a sign) $\implies$ one-tail test ($H_1 : \beta_k > c$ or $< c$)
   - No (just testing whether β_k differs from c) $\implies$ two-tail test ($H_1 : \beta_k \neq c$)
2. **Null value?** Identify c and compute $t = (b_k - c)/\operatorname{se}(b_k)$
3. **Decide** via any one of three equivalent methods:
   - Rejection region: compare $|t|$ (or t) to t_c
   - p-value: compare p to α
   - Confidence interval: check if c is inside or outside the CI
4. **Interpret:** Reject $\implies$ evidence against H_0. Do not reject $\implies$ data are compatible with H_0 (but also with other values in the CI).

Hypothesis Testing: Summary

1. **Set up:** $H_0 : \beta_k = c$ vs. one of three alternatives ($>$, $<$, $\neq$)
2. **Compute:** $t = \dfrac{b_k - c}{\text{se}(b_k)}$, which follows $t_{(N-2)}$ under H_0
3. **Decide** using any one of three equivalent methods:
   - Rejection region: compare $|t|$ (or t) to t_c
   - p-value: compare p to α
   - Confidence interval: check if c is inside or outside the CI
4. **Interpret:**
   - Reject $\implies$ evidence against H_0 (strong conclusion)
   - Do not reject $\implies$ data are compatible with H_0 (weak conclusion)
   - "Do not reject" $\neq$ "accept"
5. **Errors:** Type I (α, controlled) vs. Type II (β, depends on true value)

Software Output: What to Look For

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	83.4160	43.4102	1.9216	0.0622
INCOME	10.2096	2.0933	4.8774	0.0000

- The "t-Statistic" column $=$ Coefficient / Std. Error (testing $H_0: \beta_k = 0$)
- The "Prob." column $=$ **two-tail** p-value for $H_0: \beta_k = 0$
- For a one-tail test in the observed direction: divide by 2, but only if t has the sign predicted by H_1. If t points opposite to H_1, you cannot reject

$\Longrightarrow$ INCOME: $p \approx 0.0000$ means the relationship is significant at any conventional level.
$\Longrightarrow$ Intercept: $p = 0.0622$ means not significant at 5%, but significant at 10%.

Thank you!

jakeanderson@g.ucla.edu