Unit 9: Inference for Regression (Slopes)

Merrick Fanning

August 8, 2025

Unit 9 Overview

- Review: least-squares regression from Unit 2 (slope b, intercept a, r, s, r^2)
- Conditions for regression inference (LINER)
- Sampling distribution of b; standard error SE_b
- t-interval for slope β
- t-test for slope: $H_0: \beta = 0$ vs. H_a (directional or two-sided)
- Reading and interpreting computer output
- Worked examples with the **same datasets/figures from Unit 2**

From Description (Unit 2) to Inference (Unit 9)

- Model: $Y = \alpha + \beta x + \varepsilon$, with $\varepsilon \sim \text{Normal}(0, \sigma)$.
- Fit to sample: $\hat{y} = a + bx$ where $b = r\dfrac{s_y}{s_x}$ and $a = \overline{y} - b\overline{x}$.
- s (residual SD): $s = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$.
- In Unit 9 we ask: what does our sample slope b tell us about the *population* slope β?

We will reuse your Unit 2 figures (scatterplots, residuals) and now add intervals/tests for β.

Conditions for Regression Inference: LINER

L — Linearity: The mean relationship is linear. *Check:* scatterplot and residual plot (no curve).
I — Independence: Observations are independent (by design); for random sampling/assignment.
N — Normality of Residuals: Residuals are approximately normal. *Check:* histogram or NPP of residuals.
E — Equal Variance: Constant spread of residuals across x (homoscedastic).
R — Randomness: Data arise from a random process (random sample/assignment).

If these are reasonably met, we may proceed with t-procedures for β*.*

Sampling Distribution of the Sample Slope b

Under the model assumptions, the sampling distribution of b is approximately:

$$t\text{-distributed with } df = n - 2, \quad E[b] = \beta, \quad SE_b = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}.$$

- s is the residual standard deviation; $s = \sqrt{\dfrac{\sum(y_i - \hat{y}_i)^2}{n-2}}$.
- The spread of b shrinks when n is larger and when x has more spread ($\sum(x - \bar{x})^2$ big).

t-Interval and t-Test for the Slope β

Confidence Interval for β (level C):

$$b \pm t^{\star}_{df=n-2}\, SE_b$$

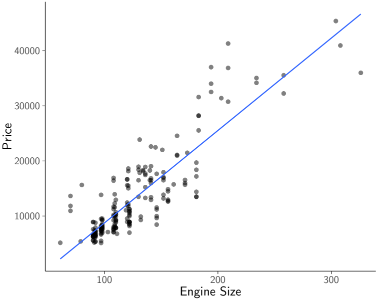
Hypothesis Test for β (two-sided):

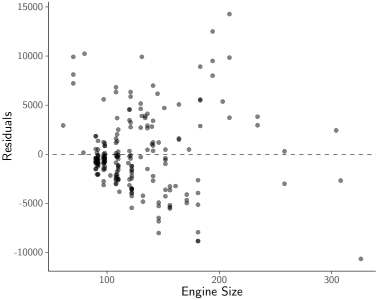
$$H_0 : \beta = 0 \quad \text{vs} \quad H_a : \beta \neq 0, \qquad t = \frac{b-0}{SE_b},\ df = n-2.$$

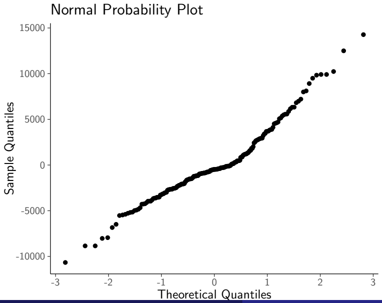
p-**value**: area in t_{n-2} beyond $|t|$ (double tail for two-sided).

Interpretation rule-of-thumb: If the CI for β excludes 0, the test at the matching α rejects H_0.

Regression Example: Price vs. Engine Size







	Dependent variable:
	price
enginesize	167.698***
Constant	−8,005.446***
Observations	205
R^2	0.764
Adjusted R^2	0.763
Residual Std. Error	3,889.454 (df = 203)
F Statistic	657.640*** (df = 1; 203)

*p<0.1; **p<0.05; ***p<0.01

Regression Conditions (LINER) — Car Price vs Engine Size

- **L — Linearity:** Scatterplot shows a straight-line trend with no obvious curvature; residual plot shows no systematic pattern. ✓
- **I — Independence:** Random sampling/assignment assumed. If sampling *without* replacement, verify the 10% condition: $n \leq 0.1N$.
  - If the population is *all individual cars/listings* in the market, N is huge, so $n = 205 \ll 0.1N$ — mark ✓.
  - If the population is *model types in a year*, N may be only a few hundred, so $n = 205$ may violate $n \leq 0.1N$ — do *not* mark ✓; note the limitation.
- **N — Normality of residuals:** Residual histogram and normal probability plot are roughly symmetric/linear; no heavy tails or extreme outliers. ✓
- **E — Equal variance (Homoscedasticity):** Residuals have an approximately constant vertical spread across engine sizes; no "fan" shape. ✓
- **R — Randomness:** Data treated as a random sample of comparable car models; no evidence of selection or measurement bias. ✓

Conclusion: All LINER conditions appear reasonably met, so t-procedures for the slope β are appropriate.

95% Confidence Interval for Slope b

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-8005.4455	873.2207	-9.17	0.0000
enginesize	167.6984	6.5394	25.64	0.0000

From regression output:

$$b = 167.6984, \quad SE_b = 6.5394, \quad df = 203$$

Critical value for 95% CI:

$$t^*_{203,\,0.025} \approx 1.972$$

$$\text{CI: } b \pm t^* \cdot SE_b$$

$$167.6984 \pm 1.972(6.5394) = 167.6984 \pm 12.89$$

$$\boxed{(154.81,\ 180.59)}$$

Interpretation: We are 95% confident that each additional unit of engine size is associated with an increase of between about \$154.81 and \$180.59 in car price.

Hypothesis Test for Slope b

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-8005.4455	873.2207	-9.17	0.0000
enginesize	167.6984	6.5394	25.64	0.0000

Test:

$$H_0 : \beta = 0 \quad \text{vs} \quad H_a : \beta \neq 0$$

From regression output:

$$t = \frac{167.6984 - 0}{6.5394} \approx 25.64, \quad df = 203$$

Two–tailed p-value:

$$p = 2 \cdot P\left(t_{203} > 25.64\right) \approx 0.0000$$

Decision: Since $p \ll 0.05$, reject H_0.

Conclusion (in context): Assuming the true slope is 0 (meaning engine size has no association with price in the population), the probability of getting a sample slope of 167.6984 or more extreme purely by random chance is essentially 0. This provides very strong evidence that engine size and price are positively associated in the population.

Beyond AP Stats: Regression in the Future

Where this shows up later:

- **College Statistics** — deeper inference methods, more formal derivations of formulas.
- **Economics, Psychology, Biology, Engineering** — regression is a primary analysis tool for real-world data.
- **Data Science & Machine Learning** — regression ideas are the backbone of predictive modeling.

Extensions beyond simple linear regression:

- **Multiple Linear Regression (MLR)** — modeling a response using several explanatory variables at once.
- **Polynomial Regression** — modeling curved relationships by adding higher-order terms.
- **Logistic Regression** — modeling the probability of a binary outcome (yes/no, pass/fail).
- **Generalized Linear Models (GLMs)** — extending regression to many types of outcomes.
- **Machine Learning methods** — regularization (ridge, lasso), decision trees, random forests, and neural networks build on regression concepts.

Takeaway: What you've learned here — interpreting slopes, checking conditions, making inferences — is the foundation for far more powerful statistical tools.