

Lecture 34: Chapter 13, Section 1 Two Quantitative Variables Inference for Regression

- Regression for Sample vs. Population
- Population Model; Parameters and Estimates
- Regression Hypotheses
- Test about Slope; Interpreting Output
- Confidence Interval for Slope

Looking Back: Review

- **4 Stages of Statistics**
 - Data Production (discussed in Lectures 1-4)
 - Displaying and Summarizing (Lectures 5-12)
 - Probability (discussed in Lectures 13-20)
 - Statistical Inference
 - 1 categorical (discussed in Lectures 21-23)
 - 1 quantitative (discussed in Lectures 24-27)
 - cat and quan: paired, 2-sample, several-sample (Lectures 28-31)
 - 2 categorical (discussed in Lectures 32-33)
 - **2 quantitative**

Regression Line and Residuals (Review)

Summarize linear relationship between explanatory (x) and response (y) values with line $\hat{y} = b_0 + b_1x$ minimizing sum of squared prediction errors $y_i - \hat{y}_i$ (called *residuals*). Typical residual size is

$$s = \sqrt{\frac{(y_1 - \hat{y}_1)^2 + \dots + (y_n - \hat{y}_n)^2}{n-2}}$$

- **Slope:** predicted change in response y for every unit increase in explanatory value x
- **Intercept:** predicted response for $x=0$

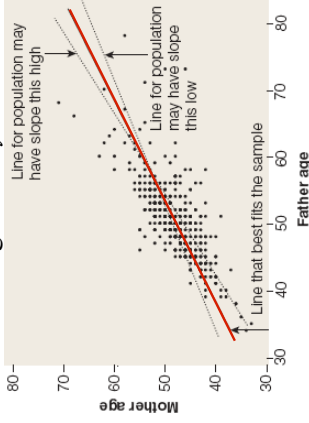
Note: this is the line that best fits the **sampled points**.

Regression for Sample vs. Population

- Can find line that best fits the *sample*.
- What does it tell about line that best fits *population*?

Example: Slope for Sample, Population

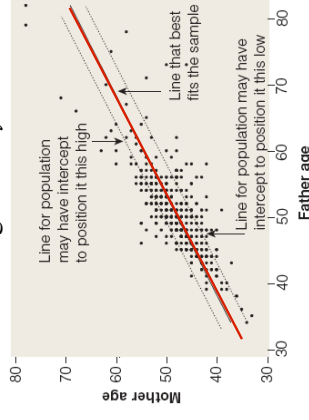
□ **Background:** Parent ages have $\hat{y} = 14.54 + 0.666x$, $s = 3.3$.



- **Question:** Is 0.666 the **slope** of the line that best fits relationship for *all* students' parents ages?
- **Response:** Slope β_1 of best line for *all* parents is

Example: Intercept for Sample, Population

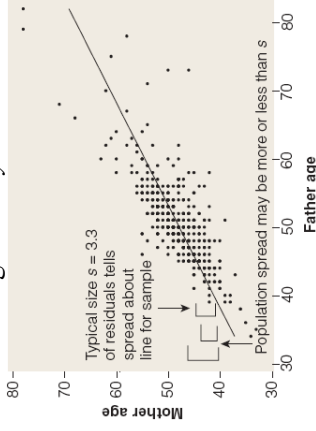
□ **Background:** Parent ages have $\hat{y} = 14.54 + 0.666x$, $s = 3.3$.



- **Question:** Is 14.54 the **intercept** of the line that best fits relationship for *all* students' parents ages?
- **Response:** Intercept β_0 of best line for *all* parents is

Example: Prediction Error for Sample, Pop.

□ **Background:** Parent ages have $\hat{y} = 14.54 + 0.666x$, $s = 3.3$.



- **Question:** Is 3.29 the typical **prediction error** size for the line that relates ages of *all* students' parents?
- **Response:** Typical residual size for best line for *all* parents is

Notation; Population Model; Estimates

σ : **typical residual size** for line best fitting linear relationship in population.

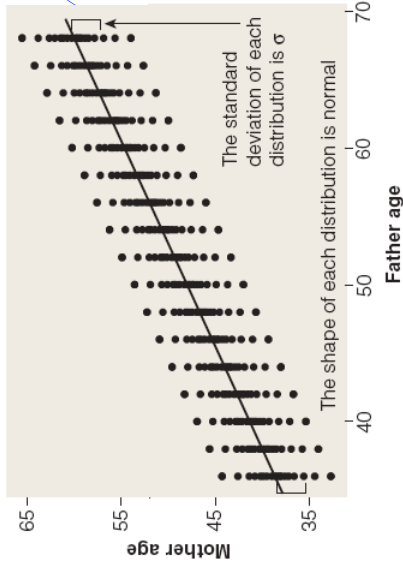
$$\mu_y = \beta_0 + \beta_1 x: \text{population mean response}$$

to any x . Responses vary normally about μ_y with standard deviation σ

Parameter	Estimate
β_0	b_0
β_1	b_1
σ	s

Population Model

Each distribution of mother ages is centered at the mean response to all such father ages (on the population regression line)



Estimates

Parameter	Estimate
β_0	b_0
β_1	b_1
σ	s

- Intercept and spread: point estimates suffice.
- Slope is focus of regression inference (hypothesis test, sometimes confidence interval).

Regression Hypotheses

- $H_0 : \beta_1 = 0 \rightarrow \mu_y = \beta_0 + \beta_1 x$
 \rightarrow no population relationship between x and y
- $H_a : \beta_1 \begin{cases} > \\ < \\ \neq \end{cases} 0$
 $\rightarrow x$ and y are related for population (and relationship is positive if $>$, negative if $<$)

Example: Point Estimates and Test about Slope

- Background:** Consider parent age regression:

The regression equation is

$$\text{MotherAge} = 14.5 + 0.666 \text{ FatherAge}$$

431 cases used 15 cases contain missing values

Predictor	Coef	SE Coef	T	P
Constant	14.542	1.317	11.05	0.000
FatherAge	0.66576	0.02571	25.89	0.000
S = 3.288		R-Sq = 61.0%	R-Sq(adj) = 60.9%	

- Questions:** What are parameters of interest and accompanying estimates? What hypotheses will we test?
- Responses:** For $\mu_y = \beta_0 + \beta_1 x$, estimate

- Parameter ___ with ___
- Parameter ___ with ___
- Parameter ___ with ___
- Test H_0 : ___ vs. H_a : ___

Suspect relationship.
Practice: 13.9, 6.48

Key to Solving Inference Problems (Review)

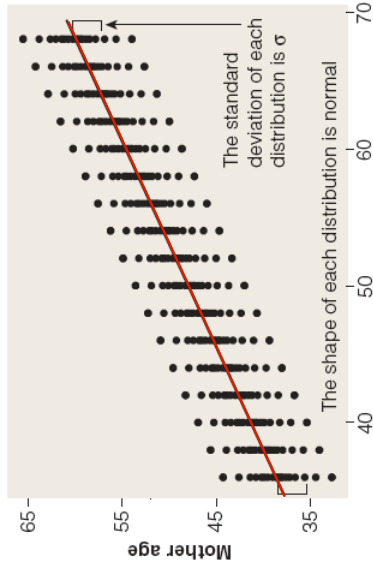
(1 quantitative variable) For a given population mean μ , standard deviation σ , and sample size n , needed to find probability of sample mean \bar{X} in a certain range:

Needed to know **sampling distribution** of \bar{X} in order to perform inference about μ .

Now, to perform inference about β_1 , need to know **sampling distribution** of b_1 .

Slopes b_1 from Random Samples Vary

Each distribution of mother ages is centered at the mean response to all such father ages (on the population regression line)

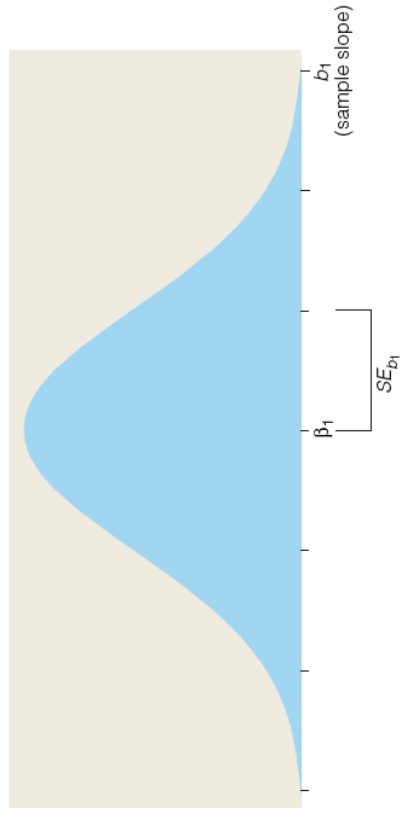


Distribution of Sample Slope

As a random variable, sample slope b_1 has

- Mean β_1
- s.d. $\approx SE_{b_1} = \frac{s}{\sqrt{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}}$
 - Residuals large \rightarrow slope hard to pinpoint
 - Residuals small \rightarrow slope easy to pinpoint
- Shape approximately normal if responses vary normally about line, or n large

Distribution of Sample Slope



Distribution of Standardized Sample Slope

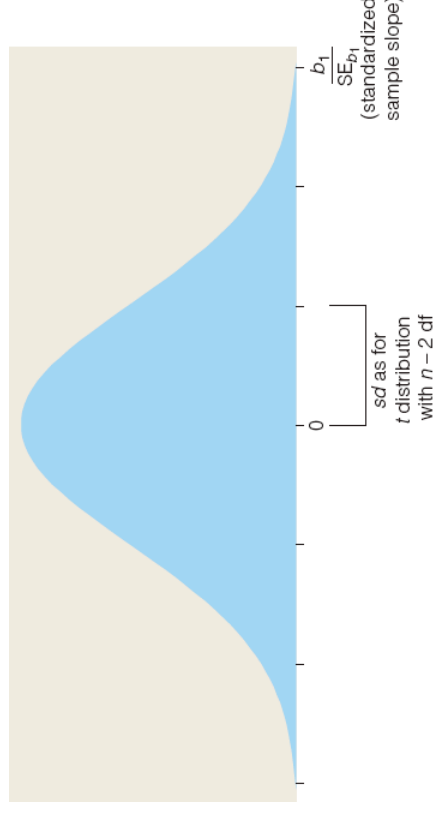
$$\begin{aligned} \text{Standardize } b_1 \text{ to } t &= \frac{b_1 - \beta_1}{SE_{b_1}} \\ &= \frac{b_1 - 0}{SE_{b_1}} \text{ if } H_0 \text{ is true.} \end{aligned}$$

For large enough n , t follows t distribution with $n-2$ degrees of freedom.

- b_1 close to 0 $\rightarrow t$ not large $\rightarrow P$ -value not small
- b_1 far from 0 $\rightarrow t$ large $\rightarrow P$ -value small

Sample slope far from 0 gives evidence to reject H_0 , conclude population slope not 0.

Distribution of Standardized Sample Slope



Example: Regression Output (Review)

- Background:** Regression of mom and dad ages:

The regression equation is

$$\text{MotherAge} = 14.5 + 0.666 \text{ FatherAge}$$

431 cases used 15 cases contain missing values

Predictor	Coef	SE Coef	T	P
Constant	14.542	1.317	11.05	0.000
FatherAge	0.66576	0.02571	25.89	0.000
S = 3.288		R-Sq = 61.0%		R-Sq(adj) = 60.9%

- Question:** What does the output tell about the relationship between mother' and fathers' ages in the **sample**?

- Response:**
 - Line _____ best fits sample (slope pos).
 - Sample relationship _____ : $r =$ _____
 - Typical size of prediction errors for sample is _____

Example: Regression Inference Output

- Background:** Regression of 431 parent ages:

Predictor	Coef	SE Coef	T	P
Constant	14.542	1.317	11.05	0.000
FatherAge	0.66576	0.02571	25.89	0.000
S = 3.288		R-Sq = 61.0%		R-Sq(adj) = 60.9%

- Question:** What does the output tell about the relationship between mother' and fathers' ages in the **population**?

- Response:** To test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 > 0$ focus on _____ line of numbers (about slope, not intercept)

- Estimate for slope of line best fitting population: _____
- Standard error of sample slope: _____
- Stan. sample slope: _____ = 0.000 where t has $df =$ _____
- P -value: _____
- Reject H_0 ? _____ Variables related in population? _____

Strength of Relationship or of Evidence

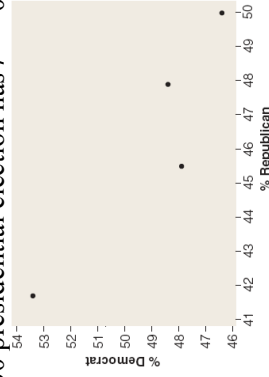
- Can have weak/strong evidence of weak/strong relationship.
- Correlation r tells strength of relationship (observed in **sample**)
 - $|r|$ close to 1 \rightarrow relationship is strong
- P -value tells strength of evidence that variables are related in **population**.
 - P -value close to 0 \rightarrow evidence is strong

Example: Strength of Relationship, Evidence

- **Background:** Regression of students' mothers' on fathers' ages had $r=+0.78$, $p=0.000$.
- **Question:** What do these tell us?
- **Response:**
 - r fairly close to 1 \rightarrow _____
 - P -value 0.000 \rightarrow _____
 - We have _____ evidence of a _____ relationship between students' mothers' and fathers' ages in general.

Example: Strength of Evidence; Small Sample

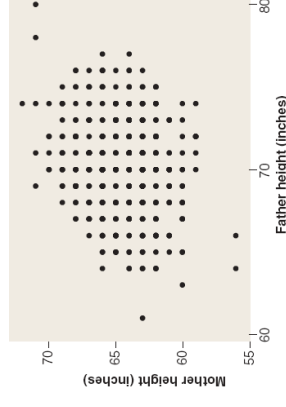
- **Background:** % voting Dem vs. % voting Rep for 4 states in 2000 presidential election has $r = -0.922$, P -value 0.078.



- **Question:** What do these tell us?
- **Response:** We have _____ evidence (due to _____) of a _____ relationship in the population of states.

Example: Strength of Evidence; Large Sample

- **Background:** Hts of moms vs. hts of dads have $r = +0.225$, P -value 0.000.



- **Question:** What do these tell us?
- **Response:** There is _____ evidence (due to _____) of a _____ relationship in the population.

Distribution of Sample Slope (Review)

As a random variable, sample slope b_1 has

- Mean β_1
- s.d. $\approx SE_{b_1} = \frac{s}{\sqrt{(x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}}$
- Shape approximately normal if responses vary normally about line, or n large

To construct confidence interval for unknown population slope β_1 use b_1 as estimate, SE_{b_1} as estimated s.d., and t multiplier with $n-2$ df.

Confidence Interval for Slope

Confidence interval for β_1 is

$$b_1 \pm \text{multiplier}(SE_{b_1})$$

where multiplier is from t dist. with $n-2$ df.

If n is large, 95% confidence interval is

$$b_1 \pm 2(SE_{b_1}).$$

Example: Confidence Interval for Slope

- **Background:** Regression of 431 parent ages:

Predictor	Coef	SE	Coef	T	P
Constant	14.542	1.317	11.05	0.000	
FatherAge	0.66576	0.02571	25.89	0.000	

$S = 3.288$ $R\text{-Sq} = 61.0\%$ $R\text{-Sq}(\text{adj}) = 60.9\%$

- **Question:** What is an approximate 95% confidence interval for the slope of the line relating mother's age and father's age for all students?

- **Response:** Use multiplier _____

We're 95% confident that for population of age pairs, if a father is 1 year older than another father, the mother is on average between _____ and _____ years older.

Note: Interval _____ \leftrightarrow Rejected H_0 .

Lecture Summary (Inference for $Quan \rightarrow Quan$: Regression)

- Regression for sample vs. population
 - Slope, intercept, sample size
- Regression hypotheses
- Test about slope
 - Distribution of sample slope
 - Distribution of standardized sample slope
- Regression inference output
 - Strength of relationship, strength of evidence
- Confidence interval for slope