Lecture 35: Chapter 13, Section 2
Two Quantitative Variables
Interval Estimates

- PI for Individual Response, CI for Mean Response
- Explanatory Value Close to or Far from Mean
- Approximating Intervals by Hand
- Width of PI vs. CI
- Guidelines for Regression Inference

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

Correlation and Regression *(Review)*

- Relationship between 2 quantitative variables
  - Display with scatterplot
  - Summarize:
    - Form: linear or curved
    - Direction: positive or negative
    - Strength: strong, moderate, weak

  If form is linear, correlation *r* tells direction and strength.

  Also, equation of least squares regression line lets us predict a response $\hat{y}$ for any explanatory value *x*.

Population Model; Parameters and Estimates

Summarize linear relationship between sampled *x* and *y* values with line $\hat{y} = b_0 + b_1 x$ minimizing sum of squared residuals $y_i - \hat{y}_i$. Typical residual size is

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

Model for population relationship is $\mu_y = \beta_o + \beta_1 x$ and responses vary normally with standard deviation σ

- Use b_o to estimate β_o
- Use b_1 to estimate β_1
- Use s to estimate σ

Looking Back: *Our hypothesis test focused on slope.*

Regression Null Hypothesis *(Review)*

- $H_o : \beta_1 = 0$

  →no population relationship between *x* and *y*

Test statistic $t = \frac{b_1 - 0}{SE_{b_1}}$

P-value is probability of *t* this extreme, if H_0 true (where *t* has *n*-2 df)

Confidence Interval for Slope *(Review)*

Confidence interval for β_1 is

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

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

If *n* is large, 95% confidence interval is

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

If CI does not contain 0, reject H_0, conclude *x* and *y are* related.

Interval Estimates in Regression

Seek **P**rediction and **C**onfidence **I**ntervals for

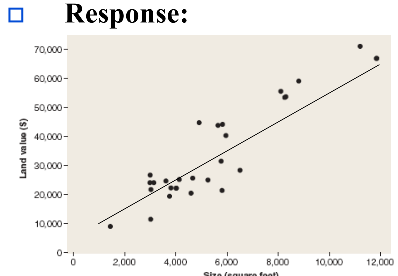
- Individual response to given *x* value (PI)
  - For large *n*, approx. 95% PI: $\hat{y} \pm 2s$
- Mean response to subpopulation with given *x* value (CI)
  - For large *n*, approx. 95% CI: $\hat{y} \pm 2\frac{s}{\sqrt{n}}$

Both intervals centered at predicted *y*-value $\hat{y}$.

These approximations may be poor if *n* is small or if given *x* value is far from average *x* value.

Example: *Reviewing Data in Scatterplot*

- **Background**: Property owner feels reassessed value $40,000 of his 4,000 sq.ft. lot is too high. For random sample of 29 local lots, means are 5,619 sq.ft. for size, $34,624 for value. Regression equation $\hat{y}$ =1,551+5.885*x*, *r* =+0.927, *s*=$6,682.
- **Question:** Where would his property appear on scatterplot?
- **Response:**



***A Closer Look:** His lot is smaller than average but valued higher than average; some cause for concern because the relationship is strong and positive. But it's not perfect, so we seek statistical evidence of an unusually high value for the lot's size.*

Example: *An Interval Estimate*

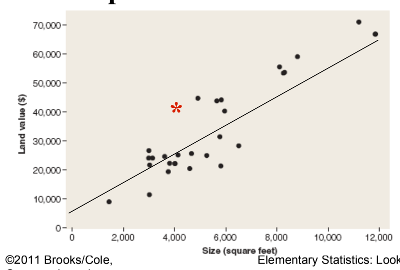
- **Background**: Property owner feels reassessed value $40,000 of his 4,000 sq.ft. lot is too high. For random sample of 29 local lots, means are 5,619 sq.ft. for size, $34,624 for value. Regression equation $\hat{y}$ =1,551+5.885x, r =+0.927, s=$6,682.
- **Questions:** What range of values are within two standard errors of the predicted value for 4,000 sq.ft.? Does $40,000 seem too high?
- **Responses:** Predict $\hat{y}$ = ______________________

  Approximate range of plausible values for individual 4,000 sq.ft. lot is ______________________________

  ____

Example: *Interval Estimate on Scatterplot*

- **Background**: A homeowner's 4,000 sq.ft. lot is assessed at $40,000. Predicted value is $25,091 and predicted range of values is ($11,727, $38,455).
- **Question:** Where do the prediction and range of values appear on the scatterplot?
- **Response:**



Prediction Interval vs. Confidence Interval

- Prediction interval corresponds to 68-95-99.7 Rule for ***data***: where an individual is likely to be.
  - PI is wider: individuals vary a great deal
- Confidence interval is ***inference*** about mean: range of plausible values for mean of sub-population.
  - CI is narrower: can estimate mean with more precision
- Both PI and CI in regression utilize info about x to be more precise about y (PI) or mean y (CI).

Example: *Prediction or Confidence Interval*

- **Background**: Property owner feels reassessed value $40,000 of his 4,000 sq.ft. lot is too high. Based on a random sample of 29 local lots, software was used to produce interval estimates when size equals 4,000 sq.ft.

```
Predicted Values for New Observations
New Obs     Fit     SE Fit          95.0% CI              95.0% PI
1          25094      1446    (   22127,    28060)  (   11066,    39121)
Values of Predictors for New Observations
New Obs       Size
1             4000
```

- **Questions:** What is the "Fit" value reporting? Which interval is relevant for the property owner's purposes: CI or PI?
- **Responses:** *Fit* is ______________________________

  The____is relevant: he wants to show that his individual lot is over-assessed.

Examples: *Series of Estimation Problems*

- Based on sample of male weights, estimate
  - weight of individual male
  - mean weight of all males

  No regression needed.
- Based on sample of male hts and weights, estimate
  - weight of individual male, 71 inches tall
  - mean weight of all 71-inch-tall males
  - weight of individual male, 76 inches tall
  - mean weight of all 76-inch-tall males

  Examples use data from sample of college males.

Example: *Estimate Individual Wt, No Ht Info*

- **Background**: A sample of male weights have mean 170.8, standard deviation 33.1. Shape of distribution is close to normal.
- **Question:** What interval should contain the weight of an individual male?
- **Response:** Need to know distribution of weights is approximately normal to apply 68-95-99.7 Rule: Approx. 95% of individual male weights in interval

  ____

Practice: 13.20e p.659

Example: *Estimate Mean Wt, No Ht Info*

- **Background**: A sample of 162 male weights have mean 170.8, standard deviation 33.1.
- **Questions:**
  - What interval should contain the mean weight of all males?
  - How does it compare to this interval for an individual male's weight? $170.8 \pm 2(33.1) = (104.6, 237.0)$
- **Responses:**
  - Need to know _____________ to construct approximate 95% confidence interval for mean:
  - Interval for **mean** involves division by square root of ***n*** → ____________ than interval for individual

Practice: 13.20f p.659

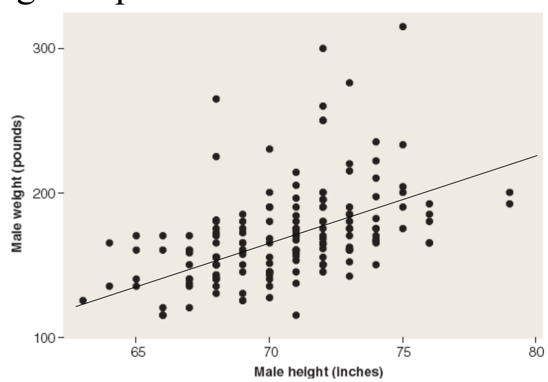
Examples: *Series of Estimation Problems*

- Based on sample of male weights, estimate
  - weight of individual male
  - mean weight of all males
- Based on sample of male heights and weights, est
  - weight of individual male, 71 inches tall
  - mean weight of all 71-inch-tall males
  - weight of individual male, 76 inches tall
  - mean weight of all 76-inch-tall males

  Need regression

Examples: *Series of Estimation Problems*

The next 4 examples make use of regression on height to produce interval estimates for weight.



Example: *Predict Individual Wt, Given Av. Ht*

- **Background**: Male hts: mean about 71 in. Wts: s.d. 33.1 lbs. Regression of wt on ht has r =+0.45, p=0.000. Regression line is $\hat{y} = -188 + 5.08x$ and s = 29.6 lbs.
- **Questions:** How much heavier is a sampled male, for each additional inch in height? Why is $s < s_y$? What interval should contain the weight of an individual 71-inch-tall male? (Got interval estimates for x=71.)

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	172.83	2.35	(168.20, 177.47)	(114.20, 231.47)

- **Responses:**
  - For each additional inch, sampled male weighs__lbs more.
  - $s < s_y$ because wts vary ____about line than about mean.
  - Look at____for x = 71: ________________

Example: *Approx. Individual Wt, Given Av. Ht*

- **Background**: Male hts: mean about 71 in. Wts: s.d. 33.1 lbs. Regression of wt on ht has r =+0.45, p=0.000. Regression line is $\hat{y} = -188 + 5.08x$ and s = 29.6 lbs. Got interval estimates for wt when ht=71:

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	172.83	2.35	(168.20, 177.47)	(114.20, 231.47)

- **Questions:**
  - How do we *approximate* interval estimate for wt. of an individual 71-inch-tall male *by hand*?
  - Is our approximate close to the true interval?
- **Responses:**
  - Predict y for x=71:______________________________

    Approx. PI=_____________________________________
  - Close?____

Example: *Est Mean Wt, Given Average Ht*

- **Background**: Male hts: mean about 71 in. Wts: s.d. 33.1 lbs. Regression of wt on ht has r =+0.45, p=0.000. Regression line is $\hat{y} = -188 + 5.08x$ and s = 29.6 lbs.
- **Questions:**
  - What interval should contain mean weight of all 71-inch-tall males?

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	172.83	2.35	(168.20, 177.47)	(114.20, 231.47)

  - How do we *approximate* the interval *by hand*? Is it close?
- **Response:**
  - Software→____ for x=71
  - Predict y for x=71: $\hat{y} = -188 + 5.08(71) = 172.7$

    Approx.

    Close? ____

Example: *Estimate Wt, Given Tall vs. Av. Ht*

- **Background**: Regression of male wt on ht produced equation $\hat{y} = -188 + 5.08x$
  For height 71 inches, estimated weight is
  $\hat{y} = -188 + 5.08(71) = 172.7$
- **Question:** How much heavier will our estimate be for height 76 inches?
- **Response:** Since ____________, predict___more lbs for each additional inch; ____more lbs for 76, which is 5 additional inches:
  Instead of weight about 173, estimate weight about __________

Example: *Est Individual Wt, Given Tall Ht*

- **Background**: Regression of male weight on height has r =+0.45, p=0.000→strong evidence of moderate positive relationship. Reg. line $\hat{y} = -188 + 5.08x$ and s=29.6 lbs. Got interval estimates for x=76.

```
New Obs     Fit     SE Fit        95.0% CI              95.0% PI
1         198.21      4.88    ( 188.58,  207.84)  ( 138.97,  257.45)
```

- **Questions:** What interval should contain the weight of an individual male, 76 inches tall? How does the interval compare to the one for ht=71?

```
New Obs     Fit     SE Fit        95.0% CI              95.0% PI
1         172.83      2.35    ( 168.20,  177.47)  ( 114.20,  231.47)
```

- **Responses:**
  - ___for x=76
  - Predicted wt (fit) about ___lbs more for x=76 than for 71: ______________ (5 more lbs per additional inch).

Example: *Approx. Individual Wt for Tall Ht*

- **Background**: Regression of male weight on height has r =+0.45, p=0.000→strong evidence of moderate positive relationship. Reg. line $\hat{y} = -188 + 5.08x$ and s=29.6 lbs. Got interval estimates for x=76.

```
New Obs     Fit     SE Fit        95.0% CI              95.0% PI
1         198.21      4.88    ( 188.58,  207.84)  ( 138.97,  257.45)
```

- **Questions:**
  - How do we *approximate* the prediction interval by hand?
  - Is it close to the true interval?
- **Responses:**
  - Predict y for x=76:____________________
  Approx. PI=____________________
  - Close? ____

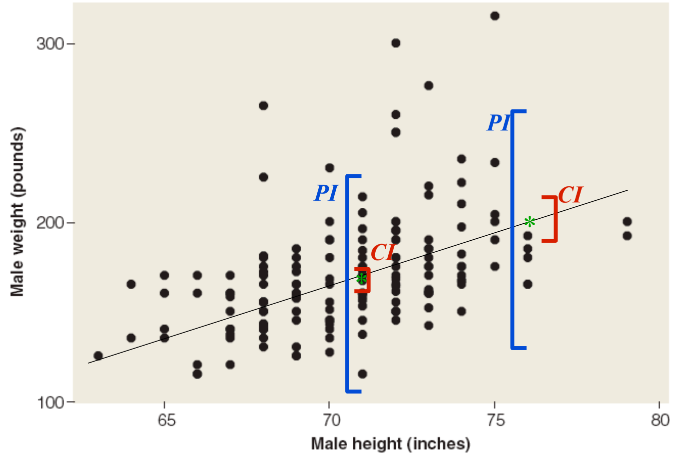
Example: *Est Mean Wt, Given Tall Ht*

- **Background**: Regression of 162 male wts on hts has r =+0.45, p=0.000→strong evidence of moderate positive relationship. Reg. line $\hat{y} = -188 + 5.08x$ and s=29.6 lbs. Got interval estimates for x=76.

```
New Obs     Fit     SE Fit        95.0% CI              95.0% PI
1         198.21      4.88    ( 188.58,  207.84)  ( 138.97,  257.45)
```

- **Questions:**
  - What interval should contain mean wt of all 76-in males?
  - How do we *approximate* the interval by hand? Is it close?
- **Responses:**
  - Refer to ____
  - Predict y for x=76: $\hat{y} = -188 + 5.08(76) = 198.1$
    ____
    Close? ________

Examples: *PI and CI for Wt; Ht=71 or 76*

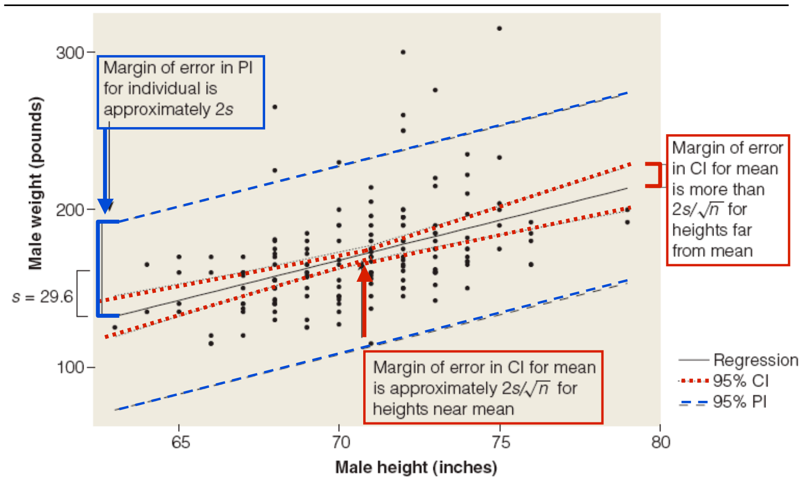


Interval Estimates in Regression *(Review)*

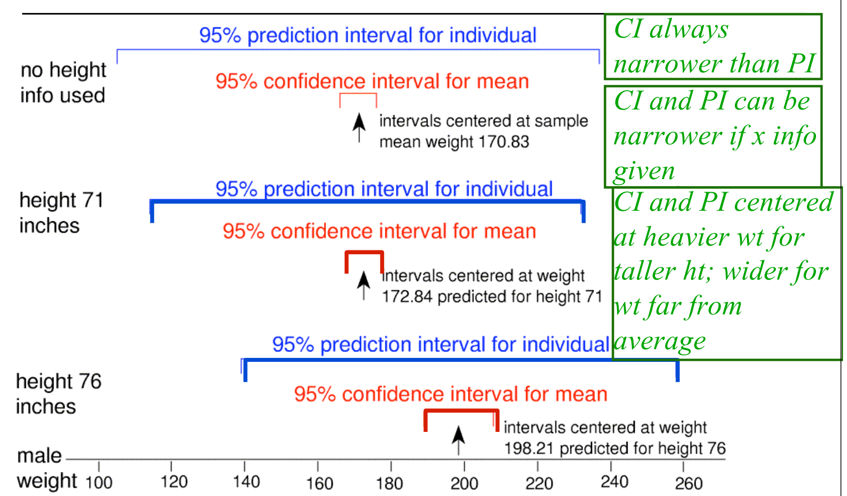
Seek interval estimates for

- Individual response to given x value (PI)
  - For large n, approx. 95% PI: $\hat{y} \pm 2s$
- Mean response to subpopulation with given x value (CI)
  - For large n, approx. 95% CI: $\hat{y} \pm 2\frac{s}{\sqrt{n}}$
- Intervals approximately correct *only for x values close to mean*; otherwise wider
  - Especially CI much wider for x far from mean

PI and CI for *x* Close to or Far From Mean



Summary of Example Intervals

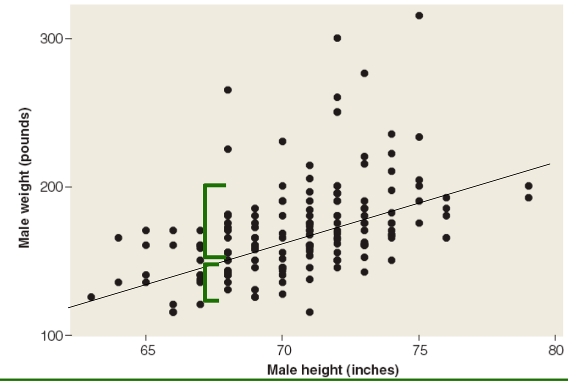


Example: *A Prediction Interval Application*

- **Background**: A news report stated that Michael Jackson was a fairly healthy 50-year-old before he died of an overdose. "His 136 pounds were in the acceptable range for a 5-foot-9 man…"
- **Question:** Based on the regression equation $\hat{y} = -188 + 5.08x$ and s=29.6 lbs, would we agree that 136 lbs. is not an unusually low weight?
- **Response:** For x = 69, predict y = ______________
  Our PI is ______________________; his weight 136 ______________________

***A Closer Look:** Our PI is a bit misleading because the distribution of weights is actually somewhat right-skewed, not normal. More of the spread reported in s=29.6 comes about from unusually heavy men, and less from unusually light men.*

Example: *A Prediction Interval Application*



***A Closer Look:** Our PI is a bit misleading because the distribution of weights is actually somewhat right-skewed, not normal. More of the spread reported in s=29.6 comes about from unusually heavy men, and less from unusually light men.*

Guidelines for Regression Inference

- Relationship must be linear
- Need random sample of independent observations
- Sample size must be large enough to offset non-normality
- Need population at least 10 times sample size
- Constant spread about regression line
- Outliers/influential observations may impact results
- Confounding variables should be separated out

Lecture Summary
(Inference for Quan→Quan; PI and CI)

- Interval estimates in regression: PI or CI
  - Non-regression PI (individual) and CI (mean)
  - Regression PI and CI for x value near mean or far
  - Approximating intervals by hand
  - Width of PI vs. CI
  - Guidelines for regression inference