Chapter 8
Assessing Relationships with Correlation and Regression

Section 8.2
The Regression Model

School of Business Administration
Portland State University

- Regression Analysis
  - Inferential Analysis

8.2c
Inferential Analysis

Inferential Considerations

What happens in the population?

- There is a question *always* to be asked of any descriptive statistic, such as b_0 or b_1
- How close is the calculated descriptive statistic likely to be to the unknown, but desired, population value?
- **Key Concept**: If a new sample of X and y values were taken, a *different* value of the slope coefficient b_1 would be obtained
- Like any descriptive statistic, the slope coefficient b_1 has a standard error, the standard deviation of the statistic across repeated samples, here s_{b_1}
- The basis of statistical inference is to understand the sampling variability of the descriptive statistic, that is, its standard error

Population Regression Coefficients

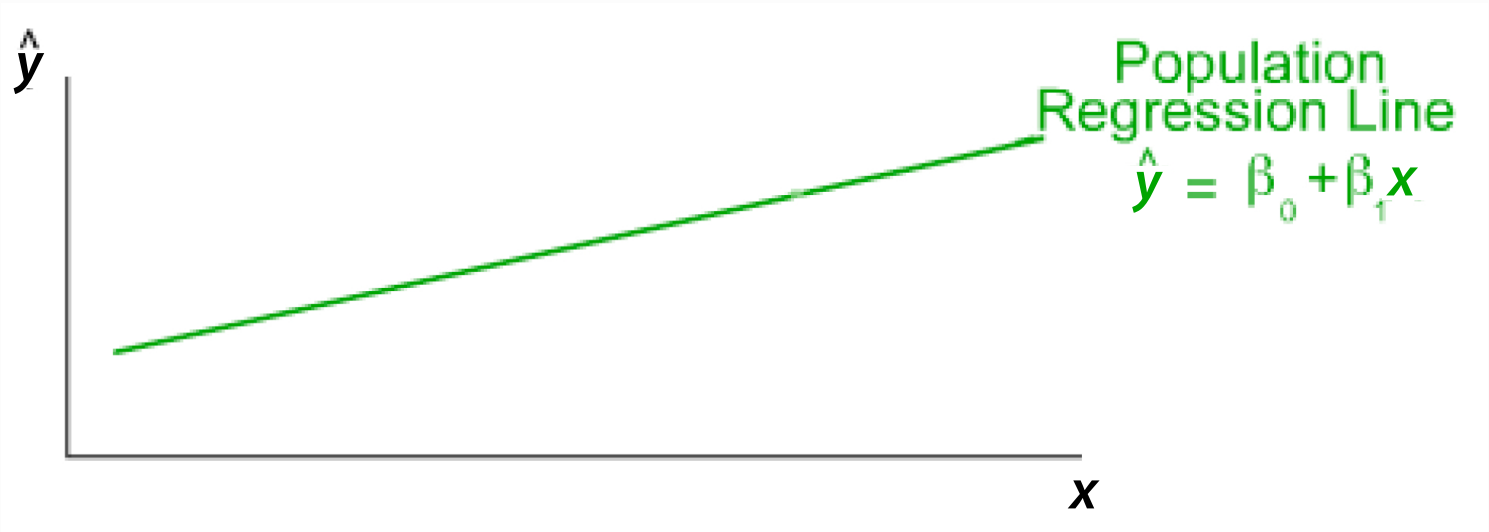
What happens in the population?

- Each of the sample regression coefficients, the y-intercept, b_0, and slope coefficient, b_1, has a corresponding population value
- **Notation**: The coefficients of the population regression line are β_0 and β_1, which define

$$\hat{y} = \beta_0 + \beta_1 x$$

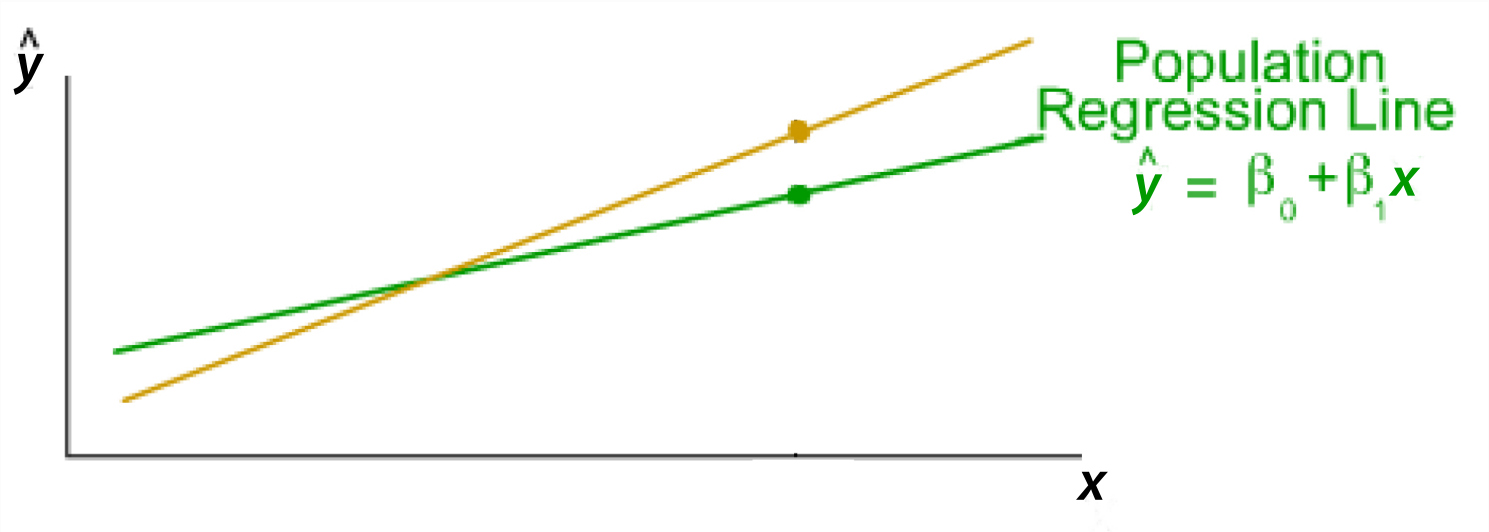
- For OLS regression estimates, unknown, but desired, population coefficients, β_0 and β_1, minimize $\sum e_i^2$ for the *entire* population

Population Regression Line



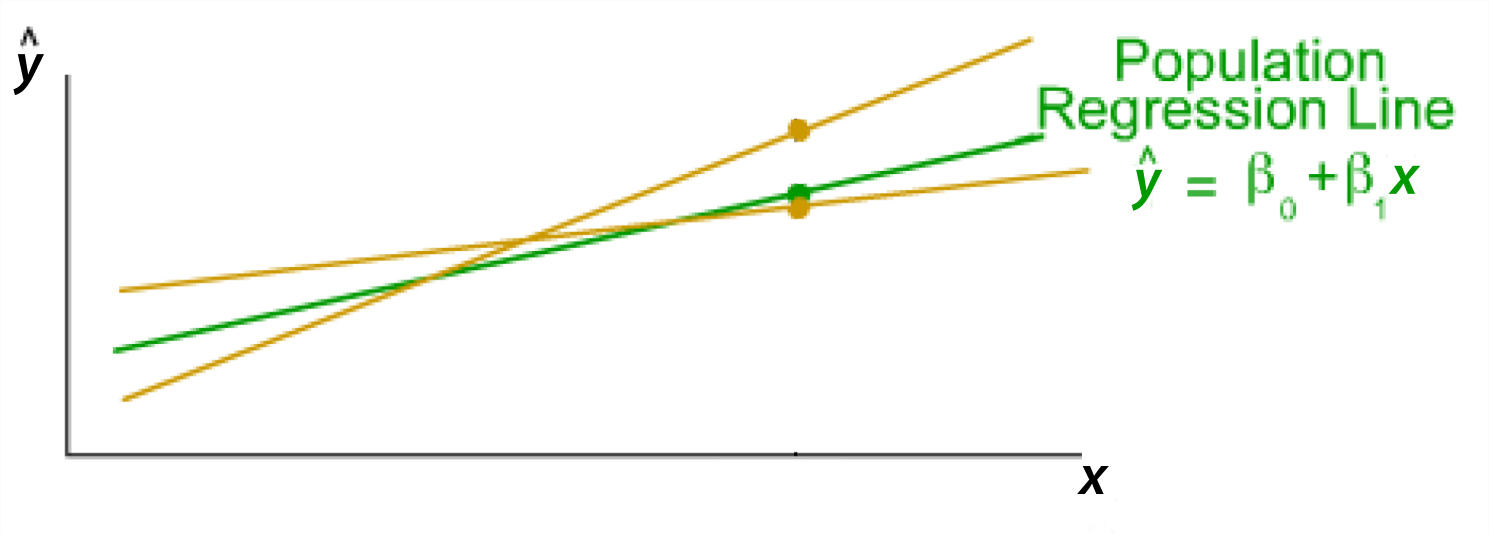
- The population coefficients, β_0 and β_1 describe the population regression line, which unfortunately cannot be directly observed

Population Regression Line



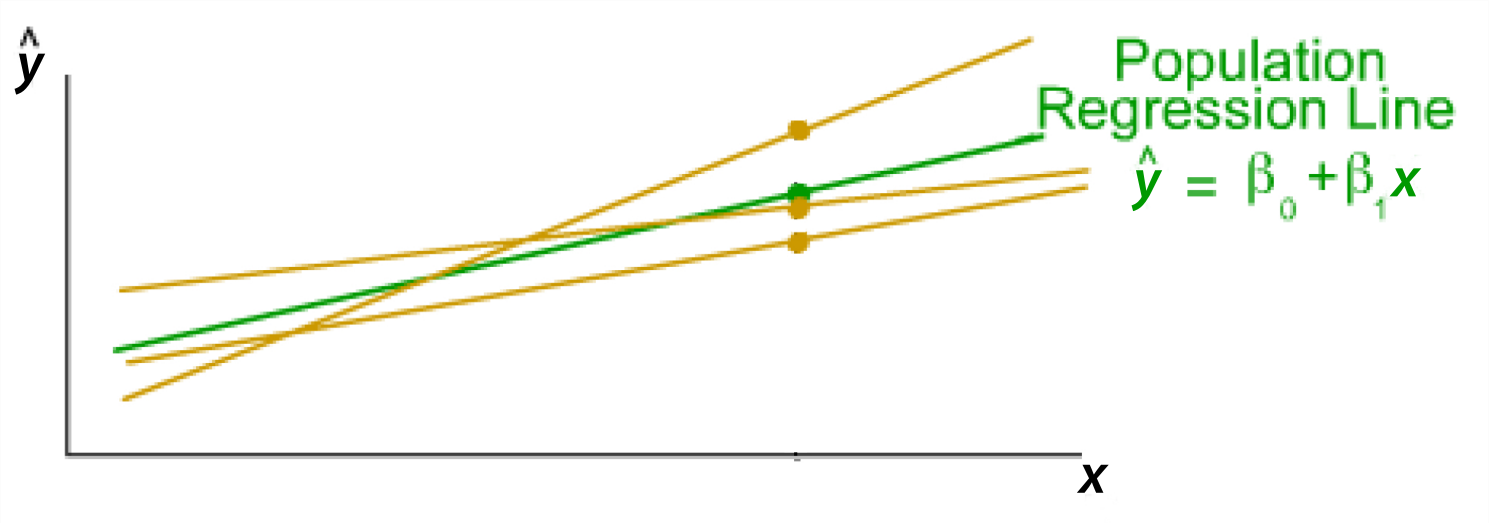
- Obtain a random sample and calculate the sample regression line, which differs from the population line because of sampling error

Population Regression Line



- Draw a second random sample and get another sample regression line because of sampling error, which leads to each sample regression line different from another

Population Regression Line



- Conceptually, the sample regression line randomly fluctuates from random sample to random sample even though only one sample is typically taken and only one regression line estimated

Inference of the Slope Coefficient: Focus on 0

What is the value of the population slope coefficient?

- The sample slope coefficients, b_1, the fitted values $\hat{y}_i$, and modeling errors or residuals, e_i, are all descriptive statistics
- Regarding the slope coefficient specifically: Is there a relationship *in the population* between the predictor variable, X, and the response variable y?
- For the *population* regression model,

$$\hat{y} = \beta_0 + \beta_1 x_1$$

- The focus of the inferential test for the population slope coefficient β_1 is on 0
  - If $\beta_1 < 0$, a negative relationship of X and y
  - If $\beta_1 = 0$, no relationship of X and y
  - If $\beta_1 > 0$, a positive relationship of X and y

Logic of Inference for the Slope Coefficient

What is the value of each population slope coefficient?

- Basic logic of confidence intervals and hypothesis tests for means and mean difference remains unchanged
- Here the statistic of interest is the sample slope coefficient, b_1, with corresponding population value β_1
- Inference is based on the t-cutoff, $t_{.025}$, and corresponding degrees of freedom, df
  - first pass through data calculates two values: b_0 and b_1
  - second pass through the *same* data to calculate residuals with the *two* previously estimated values, for b_0 and b_1
  - effective sample size, or degrees of freedom, is $df = n - 2$ for single predictor regression model

Hypothesis Test for the Slope Coefficient

How many estimated standard errors from the null?

- **Hypothesis Test for** β_1: Test the null hypothesis that $\beta_1 = 0$ with the t-statistic, the estimated number of standard errors the obtained b_1 is from 0, and then compare the resulting p-value to a pre-specified value of α, such as $\alpha = 0.05$
  - Null hypothesis: $H_0 : \beta_1 = 0$
  - Alternative hypothesis: $H_1 : \beta_1 \neq 0$
- If $p > .05$, conclude no difference detected from 0, that is, a relationship was not detected
- If $p < .05$, conclude a difference detected from 0, and, more informally, state the direction of the relation, positive or negative

Regression Output: Hypothesis Tests

Selectively edited output for the OLS estimates of the model

- The inferential analysis of a statistic begins with its estimated standard error, which, for the slope coefficient, is $s_{b_1} = 1.799$

	Estimate	Std Err	t-value	p-value
(Intercept)	-193.452	126.340	-1.531	0.164
Ht	5.202	1.799	2.892	0.020

- The corresponding t-value is $t_{b_1} = \dfrac{b_1 - 0}{s_{b_1}} = \dfrac{5.202}{1.799} = 2.892$
- **Statistical Decision**: p-value $= 0.020 < \alpha = 0.05$, so reject the null hypothesis of no relation of Heigth and Weight and, because $b_1 > 0$, conclude that $\beta_1 > 0$
- **Interpretation**: As Height increases, on average, Weight increases

Confidence Interval for the Slope Coefficient

What is the estimated value of the slope coefficient?

- **Confidence Interval** for the slope coefficient: Range of plausible values of β_1
- To construct the confidence interval, move $t_{.025}$ standard errors on either side of the sample slope coefficient

  β_1 within $b_1 \pm (t_{.025})(s_{b_1})$ for $df = n - 2$
- Is 0 is in the confidence interval?
  - If 0 is in the interval, then no relation detected between the corresponding predictor variable X and response variable y
  - If 0 is *not* in the interval, then all plausible values of the population slope coefficient, β_1, are either negative or positive, so either a negative or a positive relation has been detected between X and y

R: `lessR reg` Output, Confidence Intervals

Selectively edited output for the OLS estimates of the model

```
             Lower 95%   Upper 95%
(Intercept)    -484.794      97.889
         Ht       1.054       9.349
```

- **Margin of Error**: $(t_{.025;n-2})(s_{b_1}) = (2.306)(1.799) = 4.15$
- The 95% confidence interval for β is

  from $5.202 - 4.15 = 1.05$ to $5.202 + 4.15 = 9.35$

- This confidence interval sets the likely bounds of the population slope coefficient, β, so apply the definition of β to each end of the interval
- **Interpretation**: At the 95% level of confidence, for each one inch increase in Height, on average, weight likely increases somewhere from slightly more than 1 pound to $9\frac{1}{3}$ pounds
- Margin of error is very large, but no surprise for just $n = 10$

Epilogue

Ultimate purpose is to interpret the statistical results

- The interpretation ...
  - generalizes the results from the sample to the population
  - without the use of jargon, explained in relaxed, conversational English
- The output by itself is worthless for business decision making
- The numbers on the output need to be translated into meaningful conclusions
- Use statistical jargon like "null hypothesis" and "p-value" for statistical reasoning, but no jargon for interpreting the results
- Statistical inference generalizes to the population
  - Hypothesis test tells us if a relationship exists, from which we can infer the direction, + or -
  - Confidence interval estimates the extent of the relationship, a range of plausible values of the population slope coefficient

- The End