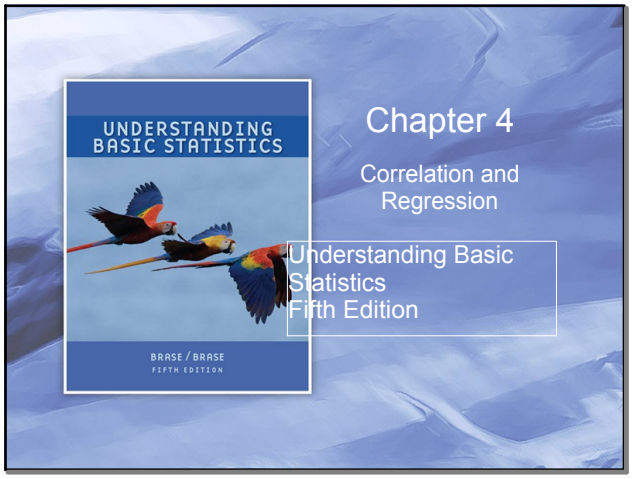
Login your clickers & yes calculators

Have out your 4.2 vocabulary

and pages 130 - 133 to correct

Oct 9-10:39 AM



Oct 8-11:45 AM

Linear Regression

- Linear Regression - a mathematical technique for creating a linear model for paired data.
- Based on the "least-squares" criterion of best fit.

 4 | 20

Oct 8-11:45 AM

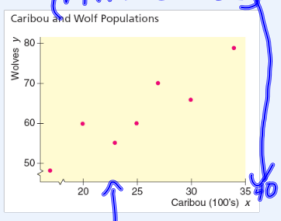
Caribou and wolf populations in Denali National Park

Questions

- Do the data points have a linear relationship?
- How do we find an equation for the best fitting line?
- Can we predict the value of the response variable for a new value of the predictor variable?
- What fractional part of the variability in *y* is associated with the variability in *x*?

Extrapolation (After - outside)

Interpolation (Inside - within)



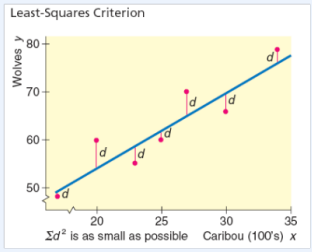
 4 | 21

Oct 8-11:45 AM

Least-Squares Criterion

Least-squares criterion

The sum of the squares of the vertical distances from the data points (x, y) to the line is made as small as possible.



Oct 8-11:45 AM

How to find the equation for the least-squares line $\hat{y} = a + bx$

Obtain a random sample of n data pairs (x, y), where x is the *explanatory variable* and y is the *response variable*.

1. Using the data pairs, compute Σx, Σy, Σx^2, Σy^2, and Σxy. Then compute the sample means $\bar{x}$ and $\bar{y}$.
2. With n = sample size, Σx, Σy, Σx^2, Σy^2, Σxy, $\bar{x}$, and $\bar{y}$, you are ready to compute the slope b and intercept a using the computation formulas

Slope: $$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} \quad (3)$$

Intercept: $$a = \bar{y} - b\bar{x} \quad (4)$$

Be careful! The notation Σx^2 means first square x and then calculate the sum, whereas $(\Sigma x)^2$ means first sum the x values, then square the result.

Oct 8-11:45 AM

3. The equation of the least-squares line computed from your sample data is

$$\hat{y} = a + bx \quad (5)$$

Note: Inferences for the population slope (Section 11.4) require the data pairs to have a *bivariate normal distribution*. That is, for a fixed value of x, the y values should have a normal distribution (or at least a mound-shaped and symmetric distribution), and for a fixed value of y, the x values should have their own (approximately) normal distribution. Chapter 6 discusses normal distributions.

Oct 8-11:45 AM

Properties of the Regression Equation

- The point $(\bar{x}, \bar{y})$ is always on the least-squares line.
- The slope tells us the amount that *y* changes when *x* increases by one unit.

Oct 8-11:45 AM

Illustration

Caribou (x, in hundreds) and wolf (y) populations

x	30	34	27	25	17	23	20
y	66	79	70	60	48	55	60

x	y	x^2	y^2	xy
30	66	900	4356	1980
34	79	1156	6241	2686
27	70	729	4900	1890
25	60	625	3600	1500
17	48	289	2304	816
23	55	529	3025	1265
20	60	400	3600	1200
$\Sigma x = 176$	$\Sigma y = 438$	$\Sigma x^2 = 4628$	$\Sigma y^2 = 28{,}026$	$\Sigma xy = 11{,}337$

Oct 8-11:45 AM

Illustration

x	y	x^2	y^2	xy
30	66	900	4356	1980
34	79	1156	6241	2686
27	70	729	4900	1890
25	60	625	3600	1500
17	48	289	2304	816
23	55	529	3025	1265
20	60	400	3600	1200
$\Sigma x = 176$	$\Sigma y = 438$	$\Sigma x^2 = 4628$	$\Sigma y^2 = 28{,}026$	$\Sigma xy = 11{,}337$

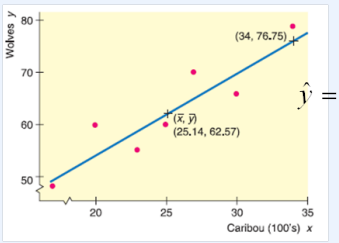
$$b = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = \frac{7(11{,}337) - (176)(438)}{7(4628) - (176)^2} = \frac{2271}{1420} \approx 1.60$$

$$a = \bar{y} - b\bar{x} \approx 62.57 - 1.60(25.14) \approx 22.35$$

Oct 8-11:45 AM

Illustration

Least-squares linear relationship between caribou and wolf populations:



$$\hat{y} = 22.35 + 1.60x$$

$y = 1.60x + 22.35$

Oct 8-11:45 AM

Critical Thinking: Making Predictions

- We can simply plug in x values into the regression equation to calculate y values.

Predicting $\hat{y}$ values for x values that are **between** observed x values in the data set is called **interpolation.**

Predicting $\hat{y}$ values for x values that are **beyond** observed x values in the data set is called **extrapolation.**

- Extrapolation may produce unrealistic forecasts.

Oct 8-11:45 AM

Coefficient of Determination

- Another way to gauge the fit of the regression equation is to calculate the coefficient of determinatio r^2.

1). Compute r. Simply square this value to get r^2.
2). r^2 is the fractional amount of total variation in y that can be explained using the linear model.
3). $1 - r^2$ is the fractional amount of total variation in that is due to random chance (or possibly due to lurking variables).

Oct 8-11:45 AM

Coefficient of Determination

1 Answer?

The linear correlation coefficient for a set of paired data is $r = 0.86$.

What fractional amount of the total variation in y is due to random chance and/or to lurking variables?

a). 0.86 b). 0.14 c). 0.74 d). 0.26

Oct 8-11:45 AM

1. Article 4?
2. Ch 4 Review
   15-20 minutes
3. Ch. 3 Test Correction
   Due Friday
4. Statistical Study Part 2
   Due Friday

Apr 24-8:35 AM