Chapter 9: Regression

What do you get when you cross a statistician with a chiropractor?
You get an adjusted R squared from a BACKward regression problem!
~Gary Ramseyer, First Internet Gallery of Statistics Jokes

Learning Objectives

Upon completion of this chapter, students should know

- How to compute and make predictions using linear regression.
- How to compute and interpret regression coefficients.
- How to compute and interpret the standard error of the estimate.

Key Terms

Regression is a term used by statisticians to indicate a backward shift toward the mean when they are predicting an unknown value from a known value when the two values are correlated.

A **regression equation** is the equation for a straight line that is used to make predictions for variables with a linear relationship.

$\hat{Y} = bX + a$

Regression coefficients are computed values of *a* and *b* in a regression equation.

$b = \frac{\text{cov}_{XY}}{S_X^2} = r_{XY} \cdot \frac{S_Y}{S_X}$ or $b = \frac{\text{cov}_{XY}}{S_Y^2} = r_{XY} \cdot \frac{S_X}{S_Y}$ $a = \bar{Y} - (b \cdot \bar{X})$ or $a = \bar{X} - (b \cdot \bar{Y})$

Regression line is a diagonal line plotted from a regression equation. The regression line can be used to make predictions. Since there are two regression equations for predicting each pair of correlated variables, there are also two regression lines.

Standard error of the estimate is the standard deviation of the actual values of a variable from the predicted values.

$S_{XY} = S_X \cdot \sqrt{1 - r^2}$ or $S_{YX} = S_Y \cdot \sqrt{1 - r^2}$

Homoscedasticity is the assumption that the standard deviation of the Y values is the same for every value of X.

Lecture and Demonstration Aids

The topic of regression sometimes tends to slip into causality assumptions. Part of the confusion may occur because students are misinterpreting the coefficient of determination as meaning some portion of the variance was caused by the effect of the related variable. You may want to clarify the coefficient of determination indicates the amount of variance explained by the linear relationship only. Thus, if the coefficient of determination is high, a fairly useful prediction of the "related" variable can be made using regression procedures. Emphasis that the prediction is neither proof nor a guarantee.

Lecture and Demonstration Ideas

Quiz Score and Television Time. Refer to the example used in the last chapter to discuss regression terminology (see Transparency 9-1). To demonstrate the computation, ask students to predict the quiz score for a student that watches television 30 hours per week. The full solution is shown on Transparency 9-4 and the computed regression line is drawn on Transparency 9-5.

Slope of the Line. To help students visualize the formula for the slope of the regression line, ask students if they would be more confident in predictions made when a correlation coefficient was .90, .75, or .25? Their answers will reflect the extent they understand the origin of the variability. Since the variability in prediction is determined by $r_{XY} \cdot \frac{S_Y}{S_X}$, the error in prediction should be much less if the coefficient is high compared to low.

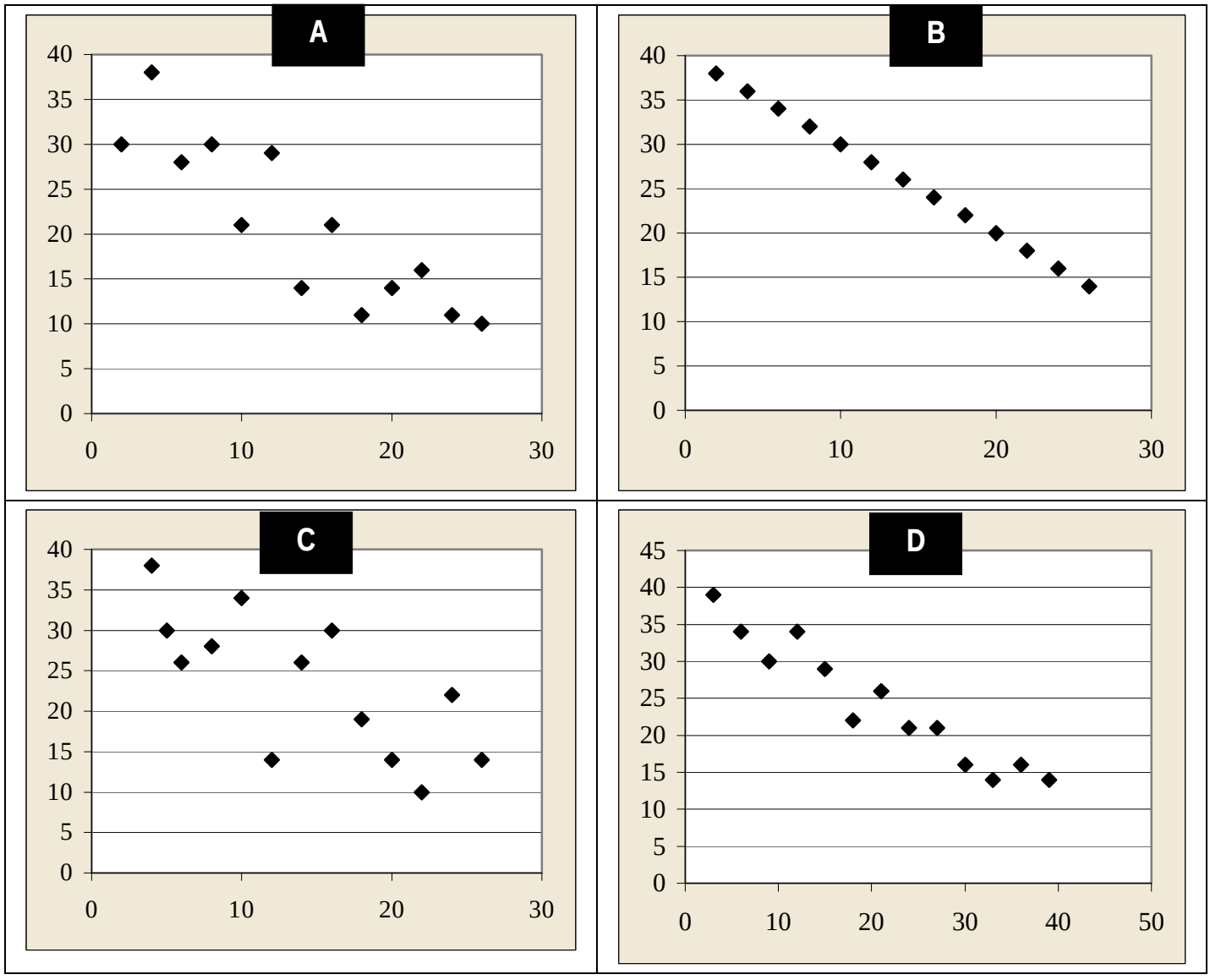
Instructional Video. *Against All Odds: Inside Statistics***.** Program Eight, "Describing Relationships" discusses linear relationships and regression. The three segments are most appropriate for regression and discuss egression lines. These videos are produced by the Consortium for Mathematics and Its Applications and Chedd-Angier (1989) and available through Annenberg/CPB.

Active-Learning Activities

Recognizing Variance. Distribute Handout 9-A. Ask students to examine and rank order the scatterplots that would have the least to the most variability in predictions. Next, ask students to estimate a regression line and indicate the variability. This will help students recognize data patterns and more fully understand the relationship of correlation to regression.

Handout 9-A.

Variability in Predictions



Transparency 9-1.

Regression Equation (to predict Y)

$$\hat{Y} = bX + a$$

$\hat{Y}$ = the predicted value of *Y* (Y Hat)

b = the slope of the regression line (the amount of change in *Y* associated with a 1-unit change in *X*)

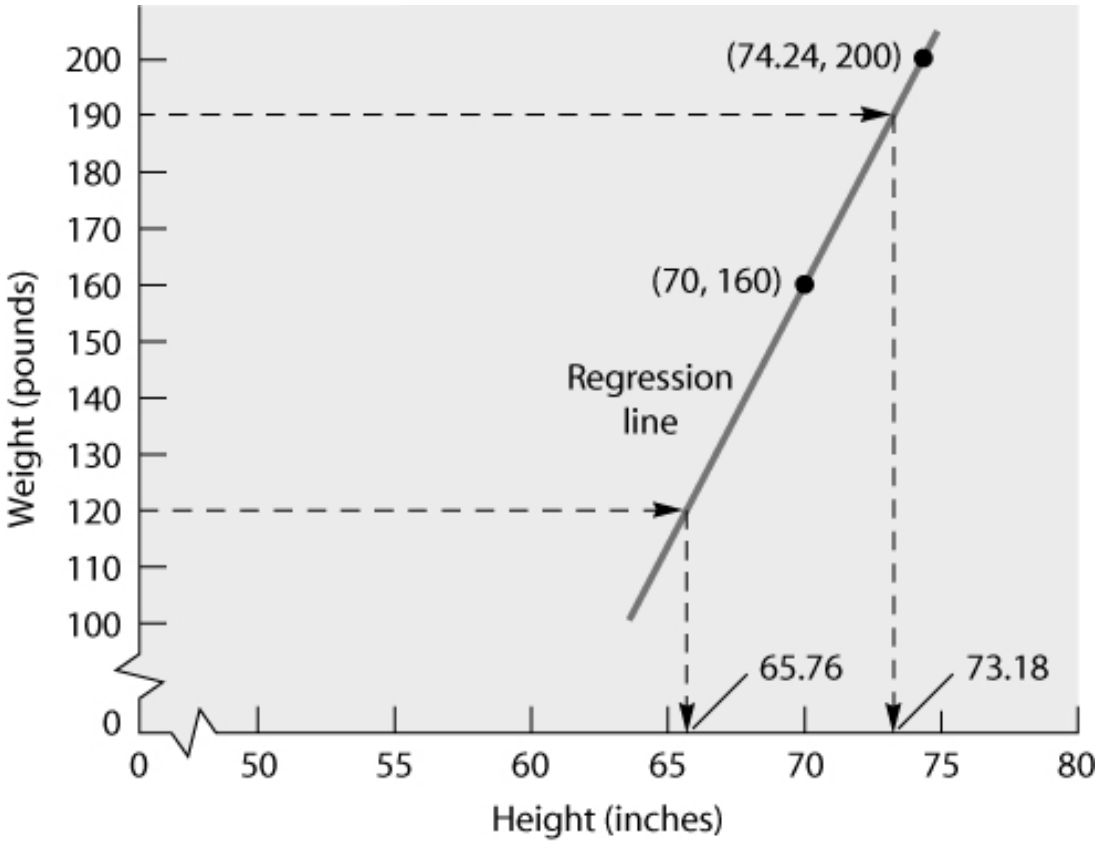
a = the *Y* intercept (the predicted value of *Y* when *X* = 0)

X = the value of *X* used to predict *Y*

Note: *b* and *a* are called "regression coefficients"

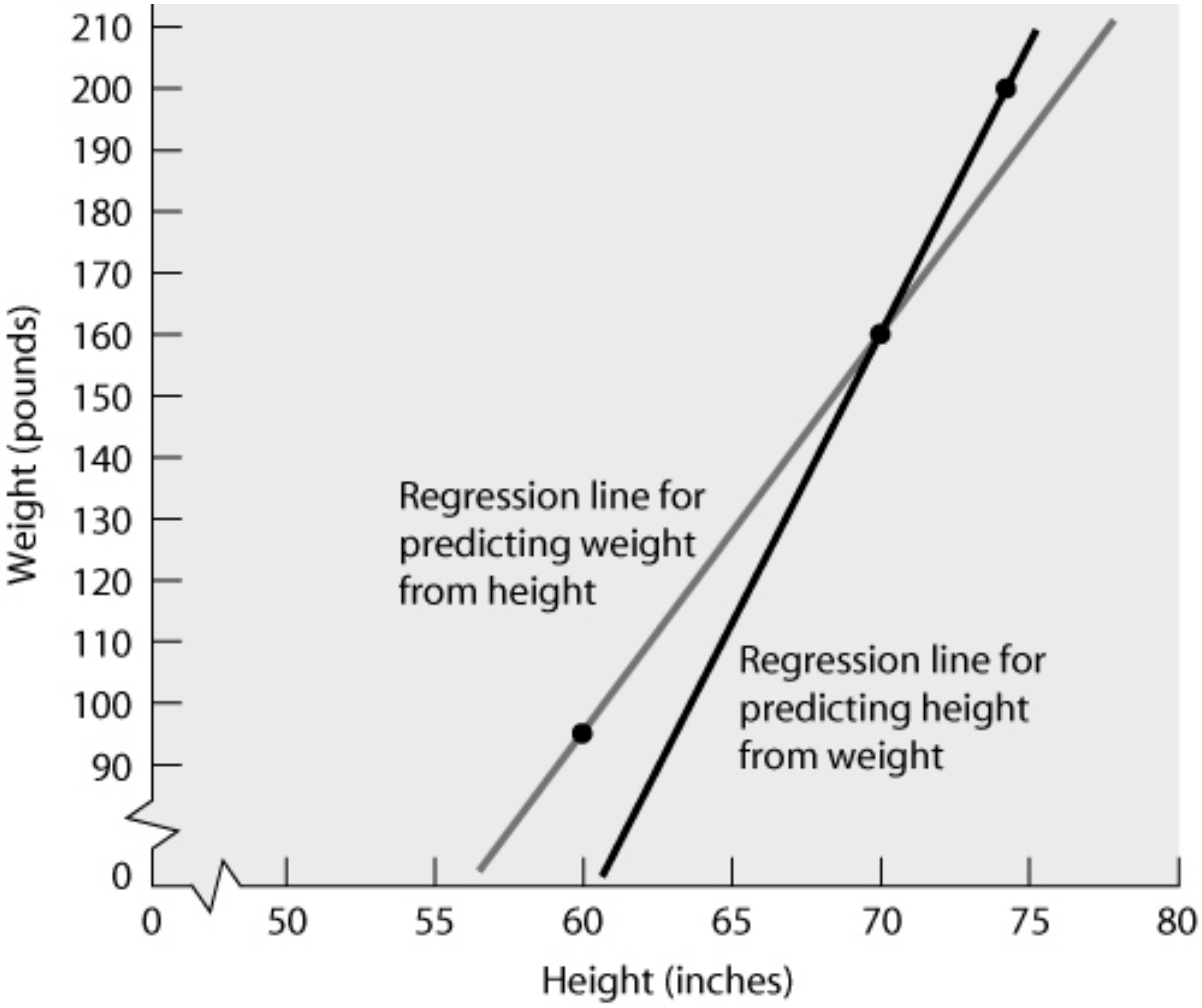
Transparency 9-2.

Predict Weight From Height Example (Text)



Transparency 9-3.

Two Different Regression Lines (Text Example)



Transparency 9-4.

Example: Predict the quiz score of a student who spends 30 hours a week watching television.

(X) TV Hours	(Y) Quiz Scores
$\bar{X} = 18.60$	$\bar{Y} = 73.50$
$S_X = 7.604$	$S_Y = 13.826$

r = -.817

Compute *b*: $b = r_{XY} \cdot \frac{S_Y}{S_X} = -.817 \cdot \frac{13.826}{7.604} = -1.486$

Compute *a*: $a = \bar{Y} - (b \cdot \bar{X}) = 73.50 - (-1.486 \cdot 18.60) = 101.14$

Regression Equation = $\hat{Y} = (-1.486)X + 101.14$

$$56.56 = (-1.486)30 + 101.14$$

Compute Standard Error of the Estimate

$$S_{YX} = S_Y \cdot \sqrt{1 - r^2} = 13.826 \cdot \sqrt{1 - (-.817)^2} = (13.826)(.577) = 7.978$$

The predicted quiz score is 56.56 points $\pm$ 7.978 points

Transparency 9-5.

Regression Line = $\hat{Y} = (-1.486)X + 101.14$