9. Linear Regression and Correlation

Data: *y* – *a* quantitative response variable
x – a quantitative explanatory variable
(Chap. 8: Recall that both variables were *categorical*)

For example,
y = annual income, *x* = number of years of education
y = college GPA, *x* = high school GPA (or perhaps SAT)

We consider:

- *Is there* an association? (test of *independence*)
- *How strong* is the association? (uses *correlation*)
- How can we describe the *nature* of the relationship, e.g., by using *x* to predict *y*? (*regression equation, residuals)*

Linear Relationships

Linear Function (Straight-Line Relation):

$$y = \alpha + \beta x$$

expresses *y* as *linear function* of *x* with *slope* β and *y-intercept* α.

For each 1-unit increase in *x,* y increases β units

$\beta > 0 \Rightarrow$ Line slopes upward (*positive* relationship)

$\beta = 0 \Rightarrow$ Horizontal line (*y* does not depend on *x*)

$\beta < 0 \Rightarrow$ Line slopes downward (*negative* relation)

Example: Economic Level and CO2 Emissions

OECD (Organization for Economic Development, www.oecd.org): Advanced industrialized nations "committed to democracy and the market economy."

oecd-data file (from 2004) on p. 62 of text and at text website www.stat.ufl.edu/~aa/social/

- Let *y* = carbon dioxide emissions (per capita, in metric tons)

  Ranges from 5.6 in Portugal to 22.0 in Luxembourg

  mean = 10.4, standard deviation = 4.6

- *x* = gross domestic product (GDP, in thousands of dollars per capita)

  Ranges from 19.6 in Portugal to 70.0 in Luxembourg

  mean = 32.1, standard deviation = 9.6

The relationship between *x* and *y* can be approximated by

$$y = 0.42 + 0.31x.$$

- At $x = 0$, predicted CO2 level $y = 0.42 + 0.31x = 0.42 + 0.31(0) = 0.42$ (irrelevant, because no GDP values near 0)

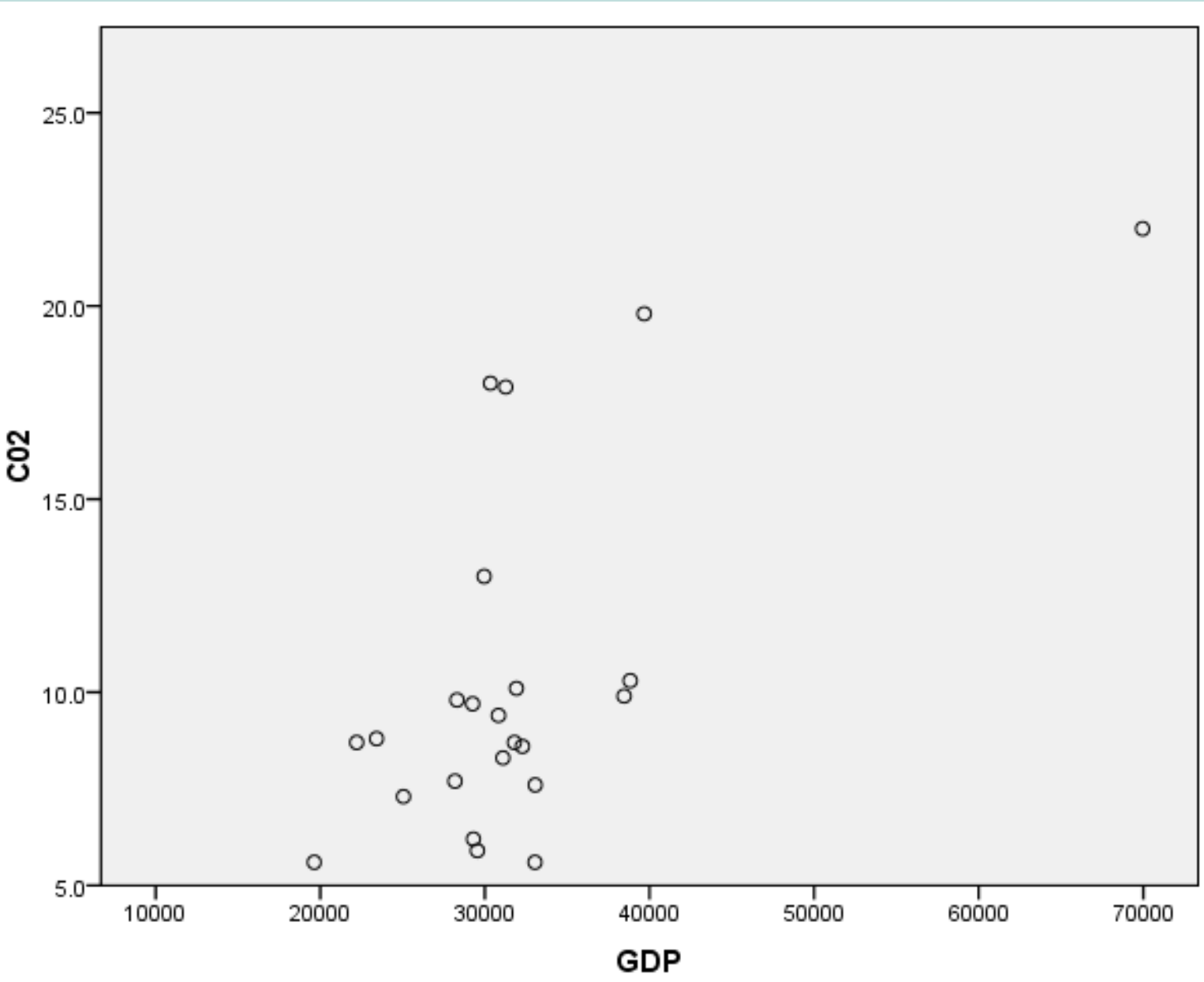
- At $x = 39.7$ (value for U.S.), predicted CO2 level $y = 0.42 + 0.31(39.7) = 12.7$ (actual = 19.8 for U.S.)

- For each increase of 1 thousand dollars in per capita GDP, CO2 use predicted to increase by 0.31 metric tons per capita

- But, this linear equation is just an approximation. The correlation between *x* and *y* for these nations was 0.64, not 1.0

(It is even less, 0.41, if we take out the outlier observation for Luxembourg.)

Likewise, we would not expect to be able to predict annual income perfectly using years of education or to predict college GPA perfectly using high school GPA.


C02
25.0
20.0
15.0
10.0
5.0
10000
20000
30000
40000
50000
60000
70000
GDP

Effect of variable coding?

Slope and intercept depend on units of measurement.

- If x = GDP measured in dollars (instead of thousands of dollars), then

$$y = 0.42 + 0.00031x$$

because a change of $1 has only 1/1000 the impact of a change of $1000 (so, the slope is multiplied by 0.001).

- If y = CO2 output in kilograms instead of metric tons

(1 metric ton = 1000 kilograms), with x in dollars, then

$$y = 1000(0.42 + 0.00031x) = 420 + 0.31x$$

Suppose x changes from U.S. dollars to British pounds and 1 pound = 2 dollars. What happens?

Probabilistic Models

- In practice, the relationship between *y* and *x* is not "perfect" because *y* is not completely determined by *x*. Other sources of variation exist.
  - We let $\alpha + \beta x$ represent the *mean* of *y*-values, as a function of *x.*
  - We replace equation

    $y = \alpha + \beta x$ by $E(y) = \alpha + \beta x$ (for population)

  (Recall E(*y*) is the "expected value of *y",* which is the mean of its probability distribution.)

  e.g., if *y* = income, *x* = no. years of education, we regard $E(y) = \alpha + \beta(12)$ as the mean income for everyone in population having 12 years education.

- A *regression function* is a mathematical function that describes how the mean of the response variable *y* changes according to the value of an explanatory variable *x.*

- A *linear* regression function is part of a *model* (a simple representation of reality) for summarizing a relationship. A linear model is OK if the true relationship is approximately linear, not OK if highly nonlinear.

- In practice, we use data to check whether a particular model is plausible (e.g., by looking at a scatterplot) and to estimate model parameters.

Estimating the linear equation

- A *scatterplot* is a plot of the *n* values of (*x, y*) for the *n* subjects in the sample
- Looking at the scatterplot is first step of analysis, to check whether linear model seems plausible

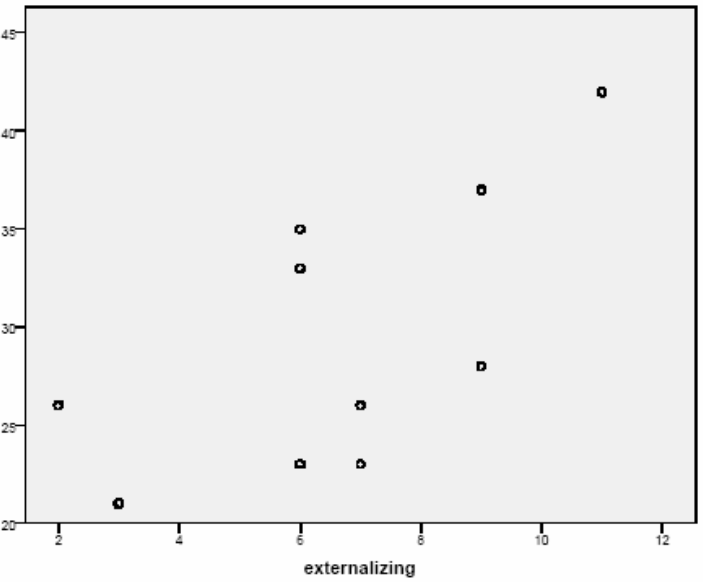
Example: Are externalizing behaviors in adolescents (e.g., acting out in negative ways, such as causing fights) associated with feelings of anxiety?

(Nolan et al., *J. Personality and Social Psych.,* 2003)

Data (some)

Subject	Externalizing (*x*)	Anxiety *(y)*
1	9	37
2	7	23
3	7	26
4	3	21
5	11	42
6	6	33
7	2	26
8	6	35
9	6	23
10	9	28

As exercise, conduct analyses with *x, y* reversed



Variables

	Anxiety (*y*)	Externalizing (*x*)
mean	29.4	6.6
std. dev.	7.0	2.7

- How to choose the line that “best fits” the data?

  – Criterion: Choose line that minimizes sum of squared vertical distances from observed data points to line. This is called the **least squares prediction equation**.

Solution (using calculus):

Denote estimate of α by a, estimate of β by b, estimate of $E(y)$ and the prediction for y by $\hat{y}$. Then,

$$\hat{y} = a + bx$$

with

$$b = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} \quad \text{and} \quad a = \overline{y} - b\overline{x}$$

Example: What causes $b > 0$ or $b < 0$?

Subject	Externalizing (*x*)	Anxiety *(y)*
1	9	37
2	7	23

Numerator of *b* is $\Sigma(x_i - \bar{x})(y_i - \bar{y})$

The contribution of subjects 1 and 2 to *b* is

(9 - 6.6)(37 - 29.4) + (7 - 6.6)(23 - 29.4)

positive negative

Motivation for formulas:

- If observation has both *x* and *y* values above means, or both values below means, then

  $(x - \bar{x})(y - \bar{y})$ is positive. Slope estimate $b > 0$ when most observations like this.

- $a = \bar{y} - b\bar{x}$ means that

  $\bar{y} = a + b\bar{x}$

- i.e., predicted value of *y* at mean of *x* is mean of *y.* The prediction equation passes through the point with coordinates ($\bar{x}$, $\bar{y}$).

Results for anxiety/externalizing data set

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	18.407	4.901		3.756	.006
	externalizing	1.666	.692	.648	2.408	.043

a. Dependent Variable: anxiety

Least squares estimates are

a = 18.407 and b = 1.666.

That is,

$$\hat{y} = 18.41 + 1.67x$$

Interpretations

- 1-unit increase in *x* corresponds to predicted increase of 1.67 in anxiety score.
- *y*-intercept of 18.4 is predicted anxiety score for subject having $x = 0$.
- The value $b = 1.67 > 0$ corresponds to a positive *sample* association between the variables.
- … but, sample size is small, with lots of variability, and it is not clear there would be a positive association for a corresponding *population*.

Residuals (prediction errors)

- For an observation, difference between observed value of *y* and predicted value $\hat{y}$ of *y*,

$$y - \hat{y}$$

  is called a **residual** (vertical distance on scatterplot)

Example: Subject 1 has *x* = 9, *y* = 37.

Predicted anxiety value is 18.41 + 1.67(9) = 33.4.

Residual = $y - \hat{y}$ = 37 – 33.4 = 3.6

Residual *positive* when *y* > predicted value

Residual *negative* when *y* < predicted value

The sum (and mean) of the residuals = 0.

Prediction equation has "least squares" property

- Residual sum of squares (i.e., sum of squared errors):

$$SSE = \sum (y_i - \hat{y}_i)^2 = \Sigma[y_i - (a + bx_i)]^2$$

- The "least squares" estimates *a* and *b* provide the prediction equation with minimum value of SSE

- For $\hat{y} = 18.41 + 1.67x$ software tells us SSE = 254.18.

  Any other equation, such as $\hat{y} = 19 + 1.7x$ has a larger value for SSE.

The Linear Regression Model

- Recall the linear regression model is *E(y)* = $\alpha + \beta x$ *(probabilistic* rather than *deterministic).* This says that the mean of the *conditional distribution* of *y* at each fixed value of *x* follows a straight line.

- The model has another parameter σ that describes the variability of the conditional distributions; that is, the variability of y values for all subjects having the same *x*-value.

- The estimate of the conditional standard deviation of y is

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

Example: We have SSE = 254.2 based on $n = 10$.

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{254.18}{8}} = \sqrt{31.77} = 5.6$$

At any fixed level of *x* (externalizing behaviors), the estimated standard deviation of anxiety values is 5.6

(Called “Std. Error of the Estimate” in SPSS printout, a very poor label)

- *df* = *n* – *2* is *degrees of freedom* for the estimate *s* of σ.

(n – *2* because we estimated 2 parameters to get predictions)

- The ratio SSE*/(n-2*) is called the **mean square error** and often denoted by MSE.

- The **total sum of squares** about the sample mean of *y* decomposes into the sum of the **residual (error) sum of squares** and the **regression sum of squares**

$$\Sigma(y_i - \bar{y})^2 = \Sigma(y_i - \hat{y}_i)^2 + \Sigma(\hat{y}_i - \bar{y})^2$$

TSS = SSE + Regression SS

We'll see that regression is more effective in predicting *y* using *x* when SSE is relatively small, regression SS is relatively large.

Software shows sums of squares in an "ANOVA" (analysis of variance) table

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	184.222	1	184.222	5.798	.043[a]
	Residual	254.178	8	31.772		
	Total	438.400	9			

a. Predictors: (Constant), externalizing

b. Dependent Variable: anxiety

Example: (text, p. 267, study in undergraduate research journal by student at Indiana Univ. of South Bend)

- Sample of 50 college students in an introductory psychology course reported *y* = high school GPA and *x* = weekly number of hours watching TV

- The study reported $\hat{y} = 3.44 - 0.03x$

- Software reports:

	Sum of Squares	df	Mean Square
Regression	3.63	1	3.63
Residual	11.66	48	.24
Total	15.29	49	

- The estimate of the conditional std dev is

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}} = \sqrt{\frac{11.66}{48}} = 0.49$$

i.e., predict GPA's vary around 3.44 – 0.03*x* with a standard deviation of 0.49

e.g., at *x* = 10 hours of TV watching, conditional dist of GPA is estimated to have mean of

3.44 – 0.03(10) = 3.14

and a standard deviation of 0.49.

Note: *Conditional* std. dev. *s* differs from *marginal* std. dev. of *y,* which we studied in Chap. 3 and is based on variability about $\bar{y}$ and ignores *x* in describing variability of *y*

Example: y = GPA, x = TV watching

We found $s = 0.49$ for estimated conditional standard deviation of GPA

Estimated *marginal* standard deviation of GPA is

$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{15.29}{49}} = 0.56$$

Normally cond. std. dev. s < marginal std. dev. s_y

How can they be dramatically different?

(picture)

Measuring association: The correlation

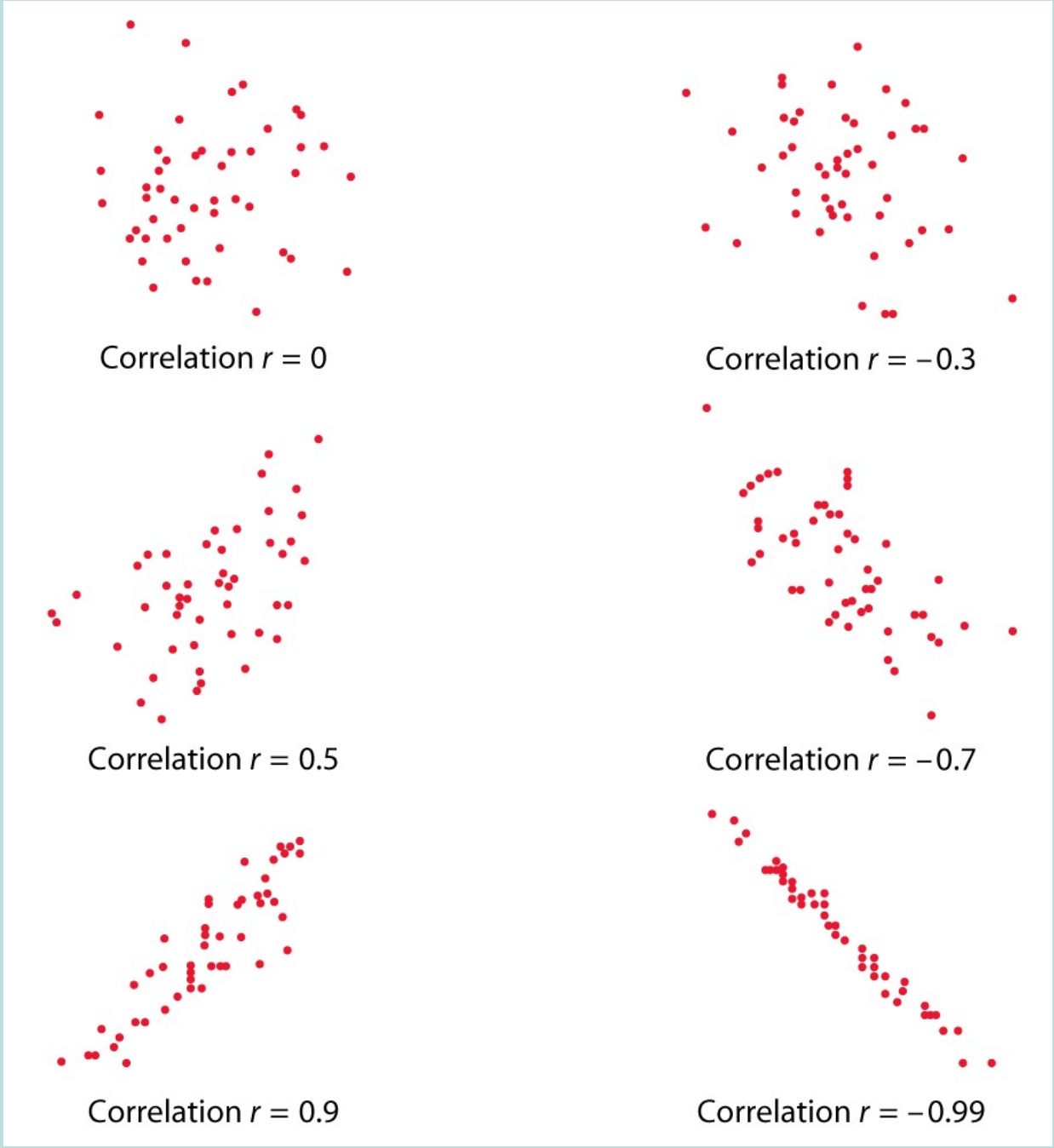
- Slope of regression equation describes the direction of association between *x* and *y,* but…
  - The magnitude of the slope depends on the units of the variables
  - The correlation is a *standardized* slope that does not depend on units
  - *Correlation r* relates to slope *b* of prediction equation by

$$r = b(s_x/s_y)$$

where s_x and s_y are sample standard deviations of *x* and *y.*

Properties of the correlation

- *r* is *standardized slope* in sense that *r* reflects what *b* equals if $s_x = s_y$
- *-1* ≤ *r* ≤ *+1,* with *r* having same sign as *b*
- *r* = 1 or -1 when all sample points fall exactly on prediction line, and *r* describes *strength of linear association*
- *r* = 0 when *b* = 0 (can happen when assoc., but not linear)
- The larger the absolute value, the stronger the assoc.


Correlation r = 0
Correlation r = −0.3
Correlation r = 0.5
Correlation r = −0.7
Correlation r = 0.9
Correlation r = −0.99

Examples

- For y = anxiety and x = externalizing behavior,
  $\hat{y}$ = 18.41 + 1.67x, and s_x = 2.7, s_y = 7.0.

  The correlation equals
  $r = b(s_x/s_y) = 1.67(2.7/7.0) = 0.648$
  (moderate positive association)

- For y = high school GPA and x = TV watching, we'll see that r = - 0.49 (moderate negative association)

- Beware: Prediction equation and r can be sensitive to outliers (Recall OECD data and effect of Luxembourg observation)

Correlation implies that predictions regress toward the mean

- When *x* goes up 1, predicted y changes by *b*
- When *x* goes up s_x, the predicted *y* changes by

$$s_x b = r s_y$$

A 1 standard deviation increase in *x* corresponds to predicted change of *r* standard deviations in *y.*

y is predicted to be "closer" to its mean than *x* is to its mean; i.e., there is *regression toward the mean* (Francis Galton 1885)

Example: *x* = parent height, *y* = child height

- Be careful not to be fooled by effects that merely represent regression toward the mean.

Examples:

Exercise 9.50 on effect of special tutoring for students who do poorly on midterm exam

Exercise 9.51 how the lightest readers tend to read more at a later time, the heaviest readers tend to read less at a later time.

r^2 = proportional reduction in error

- When we use *x* in the prediction equation to predict *y,* a summary measure of prediction error is

  sum of squared errors $SSE = \Sigma(y - \hat{y})^2$

- When we predict *y* without using *x,* best predictor is sample mean of *y,* and summary measure of prediction error is

  total sum of squares $TSS = \Sigma(y - \bar{y})^2$

Predictions using *x* get “better” as SSE decreases relative to TSS

- The *proportional reduction in error* in using *x* to predict *y* (via the prediction equation) instead of using sample mean of *y* to predict *y* is

$$r^2 = \frac{TSS - SSE}{TSS} = \frac{\Sigma(y - \bar{y})^2 - \Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$

- i.e., the proportional reduction in error is the square of the correlation!

- This measure is sometimes called the **coefficient of determination,** but more commonly just

"*r*-squared"

Example: high school GPA and TV watching

	Sum of Squares	df	Mean Square
Regression	3.63	1	3.63
Residual	11.66	48	.24
Total	15.29	49	

So, $r^2 = (15.29 - 11.66)/15.29 = 3.63/15.29 = 0.237$

There is a 23.7% reduction in error when we use x = TV watching to predict y = high school GPA.

"23.7% of the variation in high school GPA is explained by TV watching."

The correlation r is the negative square root of 0.237 (because $b < 0$), which is $r = -0.49$.

Properties of r^2

- Since $-1 \le r \le +1$, $0 \le r^2 \le 1$
- Minimum possible SSE = 0, in which case $r^2 = 1$ and all sample points fall exactly on prediction line
- If $b = 0$, then

$$a = \overline{y} - b\overline{x} = \overline{y}$$

so

$$\hat{y} = a + bx = \overline{y}$$

and so TSS = SSE and $r^2 = 0$.

- r^2 does not depend on units, or distinction between x, y

Inference about slope (β) and correlation (ρ)

Assumptions:

- The study used randomization in gathering data
- The linear regression equation $E(y) = \alpha + \beta x$ holds
- The standard deviation σ of the conditional distribution is the same at each *x*-value.
- The conditional distribution of *y* is normal at each value of *x* (least important, especially for two-sided inference with relatively large *n*)

Test of independence of *x* and *y*

- Parameter: Population slope in regression model (β)
- Estimator: Least squares estimate *b*
- *Estimated* standard error: $se = \frac{s}{\sqrt{\sum (x - \bar{x})^2}} = \frac{s}{s_X \sqrt{n-1}}$

  decreases (as usual) as *n* increases
- H_0: independence is H_0: $\beta = 0$ (under assumptions above)
- H_a can be two-sided H_a: $\beta \neq 0$

  or one-sided, H_a: $\beta > 0$ or H_a: $\beta < 0$
- Test statistic $t = (b - 0)/se$, with $df = n - 2$

Example: Anxiety/externalizing behavior revisited

From SPSS output below, $t = 1.666/0.692 = 2.41$,

$df = n - 2 = 10 - 2 = 8$, two-sided P-value = 0.043. (Interpret)

Considerable evidence against H_0: $\beta = 0$. It appears there is positive association in population between externalizing behaviors and feelings of anxiety (a CI will show this).

For H_a: $\beta > 0$, *P*-value = right-tail probability above $t = 2.41$, which is 0.043/2 = 0.02.

(Note: "Standardized coeff." in a bivariate analysis is the

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	18.407	4.901		3.756	.006
	externalizing	1.666	.692	.648	2.408	.043

a. Dependent Variable: anxiety

Confidence interval for slope β

- A CI for β has form $b \pm t(se)$

where *t-score* has *df* = *n-2* and is from *t*-table with half the error probability in each tail. (Same *se* as in test)

Example: $b = 1.666$, $se = 0.692$

With $df = 8$, for 95% CI, *t*-score = $t_{0.025} = 2.306$

95% CI for β is $1.666 \pm 2.306(0.692)$, or (0.07, 3.26).

We conclude that association in population is positive, with slope in this range (wide CI because *n* so small)

(Recall *y* = anxiety has mean = 29, std. dev. = 7
x = externalizing behavior has mean = 6.6, std. dev. = 2.7)

- What is effect of 3-unit increase in x = externalizing behavior?
  (nearly a standard deviation increase in x)

  Estimate is now 3b, which has 3(se), and we have 3b ± 3t(se), which is 3(0.07, 3.26) = (0.2, 9.8).

- Conclusion of two-sided test about H_0: $\beta = 0$ is consistent with conclusion of corresponding CI, with error prob. α that is the significance level of test.

Example: Two-sided P-value = 0.04, so reject H_0: $\beta = 0$ at 0.05 level and conclude there is an association. Likewise, 95% CI for β does not contain 0 as a plausible value for β.

What if reverse roles of variables?

(Now, *y* = externalizing behavior, *x* = anxiety

Prediction equation changes (-0.82 + 0.25*x*)
Correlation stays same (*r* = 0.648)
Result of *t* test is same (*t* = 2.41, *P* = 0.043 two-sided)

(How about bivariate analyses of categorical var's?
What changes and what stays the same?)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-.817	3.157		-.259	.802
	anxiety	.252	.105	.648	2.408	.043

a. Dependent Variable: externalizing

Some comments

- Equivalent test of independence uses H_0: $\rho = 0$, where ρ is popul. correlation that sample correlation *r* estimates (Useful when no distinction between response, explanatory variables)
- Test statistic

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Example: $r = 0.648$, $n = 10$, so $t = 0.648/0.269 = 2.41$, $df = 8$.

P-value = 0.043 for $H_a: \rho \neq 0$. (This formula useful for showing that $|t|$ increases, *P*-value decreases, as $|r|$ and/or n increase, and that t does not depend on response/explanatory distinction)

- CI for ρ more complex because *r* has a highly skewed sampling distribution when $|\rho|$ gets closer to 1.0 (text, exercise 9.64)

- Linear regression is a *model*: We don't truly expect *exactly* a linear relation with constant variability, but it is often a good and simple approximation in practice.

- Extrapolation beyond observed range of *x*-values dangerous. For *y* = high school GPA and *x* = weekly hours watching TV, $\hat{y} = 3.44 - 0.03x$. If observe *x* between 0 and 30, say, does not make sense to plug in *x* = 100 and get predicted GPA = 0.44.

- Observations are very influential if they take extreme values (small or large) of *x* and fall far from the linear trend the rest of the data follow. These can unduly affect least squares results.

Example of effect of outlier

- For data on *y* = anxiety and *x* = externalizing behavior, subject 5 had *x* = 11, *y* = 42. Suppose data for that subject had been incorrectly entered in data file as *x* = 110 and *y* = 420.

- Instead of $\hat{y} = 18.41 + 1.67x$, we get
  $\hat{y} = 5.14 + 3.76x$
- Instead of *r* = 0.64, get *r* = 0.998

- Suppose *x* entered OK but *y* entered as 420.
  Then $\hat{y} = -109.1 + 26.7x$, and *r* = 0.58.

- Correlation biased downward if only narrow range of *x*-values sampled. (see picture for why this happens)

Thus, *r* mainly useful when both *X* and *Y* randomly sampled.

Example (p. 286): How strong is association between *x* = *SAT* exam score and *y* = college GPA at end of second year of college? We'll find a very weak correlation if we sample only Harvard students, because of the very narrow range of *x*-values.

- An alternative way of expressing the regression model

$$E(y) = \alpha + \beta x$$

is

$$y = \alpha + \beta x + \varepsilon,$$

where ε is a population residual (error term) that varies around 0 (see p. 287 of text)

Software reports SS values, test results in an ANOVA (analysis of variance) table

The *F* statistic in the ANOVA table is the square of the *t* statistic for testing H_0: $\beta = 0$, and it has the same P-value as for the two-sided *t* test. This is a more general statistic needed when a hypothesis contains *more than one* regression parameter (Chapter 11).

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	184.222	1	184.222	5.798	.043[a]
	Residual	254.178	8	31.772		
	Total	438.400	9			

a. Predictors: (Constant), externalizing

b. Dependent Variable: anxiety

Some review questions for Chapter 9

1. What do we mean by a *regression model?*

2. What is a *residual*, and how do residuals relate to summarizing regression prediction errors, "least squares," *r*-squared, residuals for a contingency table?

3. What is the effect of units of measurement on least squares prediction equation, correlation, inferences such as *t* test about slope? What's effect of interchanging *x* and *y*?

4. In what sense is the correlation a *standardized slope?*

How can you interpret it in this way, and how can you interpret it in terms of the effect of a standard deviation change in *x*?

5. What is meant by “regression toward the mean” and what are some of its implications? (What would have to happen for there to be *no* regression toward the mean?)

6. When is the correlation misleading as a summary measure of association for quantitative variables? Why?

7. What is meant by a “conditional distribution” (quantitative, categorical cases)? What is a “conditional” standard deviation and when is it much less than a “marginal” standard deviation?

8. How do you do a test of independence for (a) two quantitative variables? (b) two categorical variables? (c) a quantitative and a categorical variable (that is binary)?

9. If for a given data set, predictor x_1 has a stronger correlation with y than does x_2 then how do their SSE values compare? How do their TSS values compare?