



Chapter 12

Simple Linear Regression



Introduction

- Exam Score vs. Hours Studied Scenario
- Regression Analysis
 - used to **quantify** the relation between 2 (or more) variables so you can **predict** the value of one variable based on the value of another
 - develop an equation to predict the value of a **dependent** variable based on the value of one or more **independent** variables
- Correlation Analysis
 - measures the **strength of linear relation** between a pair of variables
 - if you plan to **predict** Y from X, they ought to be **related**!



Simple vs. Multiple Regression

- Simple Regression Analysis
 - use a **single** independent variable to predict the dependent variable
 - estimated Score = $40.0816 + 1.4966(\text{Hours})$
 - $r^2 = .7432$
- Multiple Regression Analysis
 - use **multiple** independent variables to predict the dependent variable
 - the set of independent variables should be **independent of one another** and each should be highly related to the dependent variable
 - estimated Score = $33.914 + 3.472(\text{GPA}) - 1.698(\text{Absences}) + 1.395(\text{Hours})$
 - $r^2 = .7654$

3



Characterizing Relationships

- Direct Relation
 - line of best fit has **positive** slope
- Inverse Relation
 - line of best fit has **negative** slope
- Deterministic (Functional) Relation
 - "100% pure" relation between the pair of variables
 - there is **no scatter** with respect to line of best fit, so the value of Y can be determined exactly (without error) based on value of X
- Stochastic (Statistical, Random) Relation
 - a "less than perfect" relation between the pair of variables
 - since variables other than X impact Y, there is **scatter** with respect to line of best fit and there will be error when use x to predict y
- How characterize the apparent relation between Exam Score and Hours Studied?

4



Simple Linear Regression Model

- Population Linear Regression Equation

$$y = \beta_0 + \beta_1 x + e$$

- e represents the combined effects of other variables and is assumed to have mean of 0 and variance of σ^2

- Sample Linear Regression Equation

$$\hat{y} = b_0 + b_1 x$$

5



Least Squares Method: Line Of Best Fit

- The sample regression line won't perfectly fit the sample points... there will be **errors in fit**. Why?

$$\text{error in fit} = \text{residual} = (y - \hat{y})$$

- Provides the best fitting line in the sense that it has the **minimum** amount of **squared deviation between each observed value and the corresponding point on the regression line**
- Minimizes the sum of **squared** residuals in order to:
 - prevent (+) and (-) errors from cancelling
 - draws added attention to any large errors
 - prefers to make several small errors in order to avoid large errors

6



Least Squares Method: Line Of Best Fit

- Properties of the Least Squares regression equation
 - 1) b_0 and b_1 are unbiased estimators of β_0 and β_1
 - 2) line passes through the point (\bar{x}, \bar{y})
 - 3) the sum of the residuals is zero $\sum (y - \hat{y}) = 0$
 - 4) the sum of the squared residuals is minimized $\sum (y - \hat{y})^2 = \text{minimum}$

$$\text{slope} = b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$\text{y intercept} = b_0 = \bar{y} - b_1 \bar{x}$$

- Exam Score vs. Hrs Studied
 - the sample regression equation is: _____
- compute the predicted values
- compute the residuals and squared residuals

7



Conditional Distribution Of y

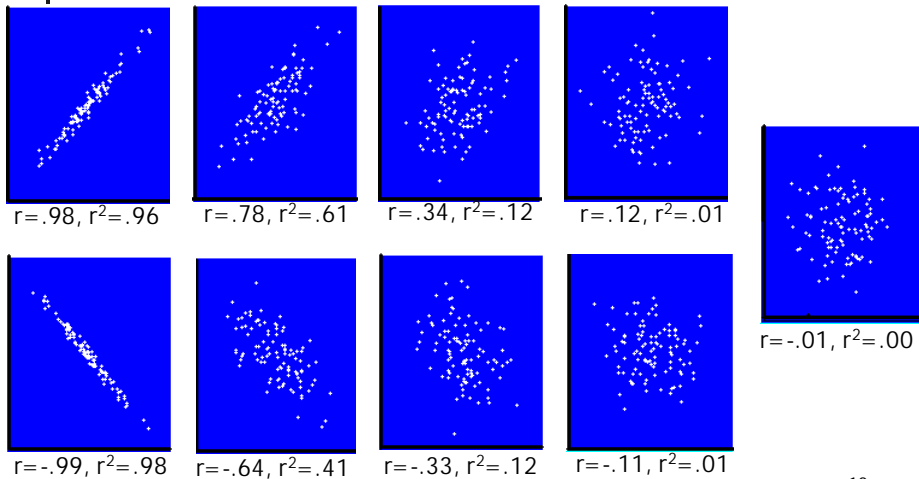
- Figure 12.8 on page 511
- Why is y variable at any given value x?
- Distribution of y is assumed Normal with mean = \hat{y}
- The regression equation is the line which connects the mean value of y at each value of x

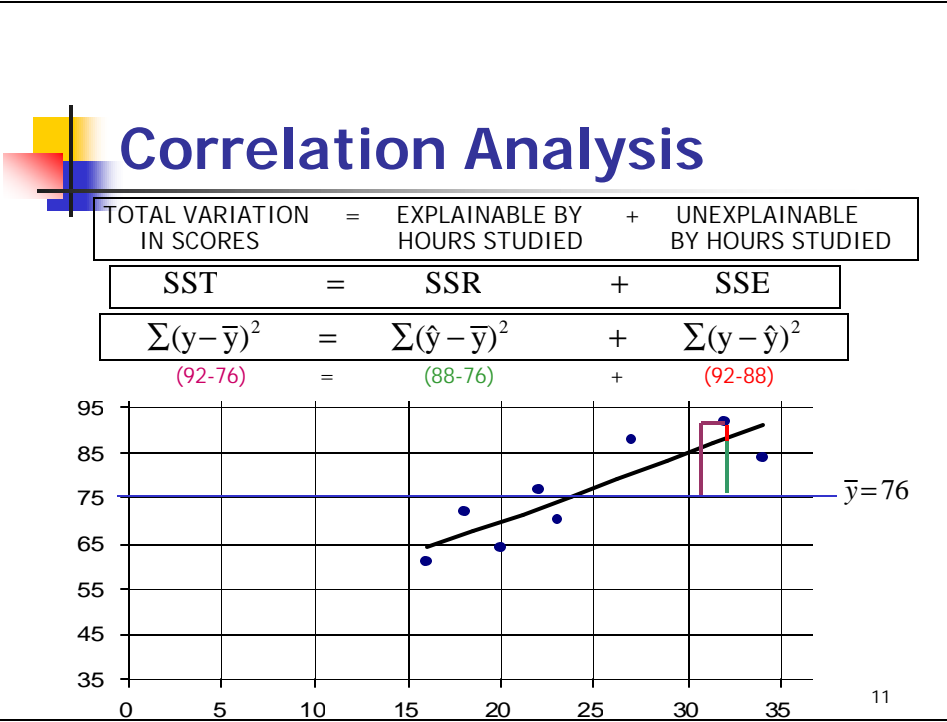
8

Correlation Analysis Concepts

- Measures the **strength of linear relation** between two variables
- If you intend to use X to predict Y, how strongly related are they?
- The slope of the sample regression equation was +1.4965 so these variables seem to “move together”
- The mean exam score was 76 and variation among student scores was $s=11.2504$
 - **some** of the variation in scores can be **explained** by taking into account hours studied

Strength of Relationship





Correlation Analysis

TOTAL VARIATION IN SCORES	=	EXPLAINABLE BY HOURS STUDIED	+	UNEXPLAINABLE BY HOURS STUDIED
SST	=	SSR	+	SSE
$\sum (y - \bar{y})^2$	=	$\sum (\hat{y} - \bar{y})^2$	+	$\sum (y - \hat{y})^2$

- Exam Score vs. Hours Studied
 - SST = _____ SSR = _____ SSE = _____

12



Coefficient Of Determination

- Measures the **proportion** of variation in variable y that is explained by variable x
- Indicates how well the sample regression line fits the sample data
- ρ^2 estimated by r^2
- $0 \leq r^2 \leq 1$

$$r^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{\text{SSR}}{\text{SST}} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$

- Exam Score vs. Hrs Studied



Coefficient Of Correlation

- ρ estimated by r
- $-1 \leq r \leq +1$

$$r = (\text{sign of } b_1) \sqrt{r^2}$$

- Interpretation: There is a **(strength)** **(direct or inverse)** correlation between **(variable X)** and **(variable Y)**
- Exam Score vs. Hrs Studied

Value of r	Strength of correlation
.9 to 1	very high
.7 to .9	high
.5 to .7	moderate
.3 to .5	weak
.0 to .3	little if any



Coefficient Of Correlation

- When working with multiple variables, common to obtain the correlation between **each pair** of variables
 - a triangular **correlation matrix**
- Can investigate whether or not the potential independent variables are truly **independent** of one another

Student	Hours	Score	Gender	GPA	Absences
Adams	20	64	0	3.1	2
Baker	16	61	0	2.9	1
Clinton	34	84	1	3.3	1
Dole	23	70	0	3	2
Edwards	27	88	1	3.2	1
Fox	32	92	1	3	0
Gore	18	72	0	3.1	0
Hale	22	77	0	3.1	3

	Score	Hours	GPA
Hours	0.862		
GPA	0.489	0.566	
Absences	-0.343	-0.234	0.028

15



Limitations Of Regression Analysis

- Regression/Correlation cannot prove cause-and-effect relationships
 - Brightman article
- Don't use the regression model to predict beyond range of observed X-values

16

Mean Square Error & Standard Error of Estimate

- Measures amount of **scatter** around the **regression line**

- Serves as an estimate of σ^2

$$\text{M.S.E.} = \frac{\text{SSE}}{n-2} = \frac{\sum (y - \hat{y})^2}{n-2}$$

- Standard Error of Estimate

- Square root of MSE
- Serves as an estimate of σ

$$s_{\text{est}} = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$$

- Used for **inference** regarding the regression line
 - hypothesis tests
 - interval estimates
- Exam Score vs. Hrs Studied

17

t-Test for Significance of the Slope

- b_1 estimates β_1
- $H_0: \beta_1 = 0$ no relation between the two variables
- $H_A: \beta_1 \neq 0$ is a relation between the two variables
- test statistic = b_1 whose sampling distr follows t_{n-2}
- Standard Error of the Slope
 - measures ROSE when use b_1 to estimate β_1

$$s_{b_1} = \frac{\sqrt{\text{M.S.E.}}}{\sqrt{\sum (x - \bar{x})^2}} = \frac{s_{\text{est}}}{\sqrt{\sum (x - \bar{x})^2}}$$

- Exam Score vs. Hrs Studied

18



Interval Estimation In Regression Analysis

- What score would you predict for students who study 30 hours?
- We've estimated that the **mean** score of all students studying 30 hours is ~ 85 . This is a point estimate based on a sample of $n=8$.
- The estimate could be in error due to 2 sources:
 - 1) sampling error
 - since b_0 and b_1 are sample results, they may be biased
 - we're not certain where the true population regression equation is
 - 2) stochastic relation
 - wherever the true population regression equation actually is, there is **scatter around it** due to the combined effects of other variables

19



Confidence Interval Estimate of the Mean Value of y

- Estimate the **mean** value for y at a given value of x
- Standard Error of the Conditional Mean (nib)
 - accounts for sampling error in estimating b_0 and b_1 which would affect our predicted value

$$s_{\hat{y}} = s_{\text{est}} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

$$\text{CI for } y = \hat{y} \pm t_{n-2} s_{\hat{y}}$$

- Pg. 529: notice that the width of the **confidence band** increases as you predict further away from \bar{x} -bar
- Exam Score vs. Hours Studied
 - 95% CI for the **mean** score of students who study 30 hours

20



Prediction Interval Estimate of an Individual Value of y

- Estimate an **individual** value for y at a given x
- Standard Error of the Forecast (s_{ind})
 - accounts for sampling error and the fact that there is dispersion around the regression line

$$s_{\text{ind}} = s_{\text{est}} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

$$\text{PI for } y_{\text{ind}} = \hat{y} \pm t_{n-2} s_{\text{ind}}$$

- Pg. 531: notice that PI bands are wider than CI bands and that each is wider as you predict further away from x-bar
- Exam Score vs. Hours Studied
 - 95% PI for **individual** score of a particular student who studies 30 hrs