

## Chapter 3

### Linear Regression

**The Problem** To “fit” a linear equation  $y = b_0 + b_1x$  to a set of data points.

- Plot the following points:

$x$	1	1	2	4
$y$	1	3	2	6

- On the graph, draw a line which fits the data pretty well.
- We want to use this line to predict the values of the data points.

$y$  = the observed value of  $y$  corresponding to  $x$ .

$\hat{y}$  = the  $y$ -value predicted to correspond to  $x$  by the equation.

- For example, consider the line  $y = 1.33x + 0.33$ . Fill in the following table.

$x$	$y$	$\hat{y}$
1		
1		
2		
4		

- How do we decide if one line fits better than the other?

- The **Residual**, or error, of the *observed* value  $y$  is the difference

$$e = y - \hat{y},$$

the difference between the observed and predicted value of  $y$ .

- Consider the two lines:  $y = 1.33x + 0.33$  and  $y = 1.25x + 0.5$ . Fill in the following table:

$$y = 1.33x + 0.33$$

$x$	$y$	$\hat{y}$	$e$	$e^2$
1				
1				
2				
4				

$$y = 1.25x + 0.5$$

$x$	$y$	$\hat{y}$	$e$	$e^2$
1				
1				
2				
4				

- **Least Squares Criterion** The straight line that best fits a set of data points is the one having the smallest possible sum of squared errors.
- **Regression Line** The straight line that best fits a set of data points according to the least squares criterion.
- **Regression Equation** The equation of the regression line.
- **Notation:**

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{1}{n} \left( \sum x_i \right) \left( \sum y_i \right)$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \left( \sum y_i \right)^2 / n$$

- **The Regression Equation:** For a set of  $n$  data points, the regression equation is

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

where

$$b_1 = \frac{S_{xy}}{S_{xx}}$$

and

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

**Example** Find the regression equation for the data points:

$x$	1	1	2	4
$y$	1	3	2	6

Use this to estimate the value of  $\hat{y}$  when  $x = 3$ .

**Example** Consider the data on Shoe Sizes:

Shoe Size	Height
4	62
7	64
7.5	66
9	68
9.5	68
10.5	69
9	69
10	70
11	71
11.5	72

Find the least-squares regression line, with shoe size as the predictor variable and height as the response variable.

**Example** The number of points and rebounds made by Wilt Chamberlain in the 13 years that he played (we exclude the 1969-1970 season in which he was injured) are shown in the table:

Season	Team	Points	Rebounds
1959-1960	Philadelphia	2707	1941
1960-1961	Philadelphia	3033	2149
1961-1962	Philadelphia	4029	2052
1962-1963	San Francisco	3585	1946
1963-1964	San Francisco	2948	1787
1964-1965	San Francisco	2534	1673
1965-1966	Philadelphia	2649	1943
1966-1967	Philadelphia	1956	1957
1967-1968	Philadelphia	1992	1952
1968-1969	Los Angeles	1664	1712
1969-1970	Los Angeles	injured	
1970-1971	Los Angeles	1696	1493
1971-1972	Los Angeles	1213	1572
1972-1973	Los Angeles	1084	1526

**Scatter Diagram** A graph that shows the relationship between two variables measured on the same individual. The predictor variable is plotted on the horizontal axis and the response variable on the vertical axis.

**Example** Use SPSS to make a scatter diagram of Wilt Chamberlain's points vs rebounds. Use points as the predictor variable and rebounds as the response variable.

- The data is located on the U: drive in the file

**MT Student File Area/dgarth/stat190/  
wiltchamberlain.sav**

- In SPSS, run the command **Graphs/Interactive/Scatterplot**
- Choose a simple scatter plot.
- Put the variable **points** on the X axis, and the variable **rebounds** on the Y axis.
- To put a title on the plot, label axes, etc, right click on the graph and select *chart object*
- To plot the regression line,

- Select **Fit**
- Select Regression
- If you forget the regression line, you can add it by performing the following.
  - Right click on the graph
  - Select **Interactive Graph Object**
  - Select **Insert/Fit Line/Regression**

**Example** Use SPSS to find the least squares regression line for the Wilt Chamberlain data. Use Points as the predictor variable and Rebounds as the response variable.

- Select

### Analyze/Regression/Linear

- Put **points** in the independent variable box, and **rebounds** in the dependent variable box.

### Inferences for the Regression Line

#### Population Regression Line

$$\mu_{y|x} = \beta_0 + \beta_1 x$$

- $\mu_{y|x}$  is the mean of  $y$  for the given value of  $x$ .

#### Sample Regression Line

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

**Question:** What kind of inferences can we make about the population regression line based on the sample regression line? ie, how well does the sample regression line approximate the population regression line?

**Disturbances** For a data point  $(x_i, y_i)$ , the disturbance is the quantity  $e_i$  satisfying

$$y_i = \beta_0 + \beta_1 x_i + e_i.$$

#### Assumptions About the Population Regression Line

- The mean of the disturbances is 0. Thus, the population regression equation is linear.
- The variance of each  $e_i$  has the same value, say  $\sigma_e^2$ .
- The  $e_i$  are normally distributed.
- The  $e_i$  are independent.

See **Figure 3.12** of the text

## Inferences about $\beta_0$ and $\beta_1$

- Notice that  $b_0$  and  $b_1$  are random variables, and therefore have sampling distributions
  - The sampling distributions for  $b_0$  and  $b_1$  are the probability distributions for the set of all possible coefficients that could be obtained in regression lines from all possible samples of points with fixed  $x$  coordinates  $x_1, \dots, x_n$ .

### Sampling Distribution for $b_1$

- $\mu_{b_1} = \beta_1$  ( $b_1$  is unbiased)
- $\sigma_{b_1}^2 = \frac{\sigma_e^2}{\sum (x_i - \bar{x})^2} = \frac{\sigma_e^2}{(n-1)s_x^2}$ ,

where  $s_x$  is the sample standard deviation of the  $x$  coordinates of the points in the sample

- The sampling distribution of  $b_1$  is normally distributed

### Sampling Distribution for $b_0$

- $\mu_{b_0} = \beta_0$  ( $b_0$  is unbiased)
- $\sigma_{b_0}^2 = \sigma_e^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right) = \sigma_e^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right)$ ,

where  $s_x$  is the sample standard deviation of the  $x$  coordinates of the points in the sample

- The sampling distribution of  $b_0$  is normally distributed

**Estimating  $\sigma_e^2$**  Usually we use the sample variance around the regression line.

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

### Confidence Intervals for $b_1$

$$b_1 \pm t_{\alpha/2} \cdot s_{b_1} = b_1 \pm t_{\alpha/2} \cdot \frac{s_e}{\sqrt{S_{xx}}}$$

where  $df = n - 2$ .

### Confidence Intervals for $b_0$

$$b_0 \pm t_{\alpha/2} \cdot s_{b_0} = b_0 \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

where  $df = n - 2$

**Example** Construct confidence intervals for  $\beta_1$  and  $\beta_0$  in the shoe sizes example.

### Hypothesis Testing for $\beta_1$ and $\beta_0$ .

- We want to test the hypotheses:

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

- If  $\beta_1 = 0$ , there is no linear relationship between  $x$  and  $y$ .
  - ie,  $x$  is not useful for predicting  $y$
- If  $\beta_1 \neq 0$ , then there is a linear relationship between  $x$  and  $y$ .
  - ie,  $x$  is useful for predicting  $y$ .

## Logic Behind the Hypothesis Test

- Assume that  $\beta = 0$ .
- Take sample data and find the least squares regression line for the sample data,

$$\hat{y} = b_0 + b_1x$$

- Find the probability that, if  $\beta_1 = 0$ , that we would get the value that we found for  $b_1$ .
- If this probability is small, reject our assumption that  $\beta_1 = 0$ . If the probability is large, do not reject it.

## T-test for the utility of a Regression

**Assumptions:** The four assumptions for regression inferences are met.

- **Hypotheses**

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

- **Test Statistic**

$$t = \frac{b_1}{s_e/\sqrt{S_{xx}}}$$

- For samples of size  $n$ , each with the same values of

$$x_1, x_2, \dots, x_n,$$

this variable has the  $t$ -distribution with  $n - 2$  degrees of freedom.

**Example** The National Center for Health Statistics publishes data on heights and weights in *Vital and Health Statistics*. A random sample of 11 males age 18-24 years gave the following data, where  $x$  denotes height, in inches, and  $y$  denotes weight, in pounds.

x	65	67	71	71	66	75	67	70	71	69	69
y	175	133	185	163	126	198	153	163	159	151	155



- Find a confidence interval estimate for the number of rebounds, assuming he scores no points.

### Coefficient of Determination

Define the three quantities:

- $SST$  = The total variation in the observed incomes

$$SST = \sum (y_i - \bar{y})^2$$

We call  $SST$  the **Total Sum of Squares**

- $SSR$  = The amount of variation in the observed incomes explained by the regression.

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

We call  $SSR$  the **Regression Sum of Squares**.

- $SSE$  = The amount of variation in the observed values of the response variable not explained by the regression:

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

We call  $SSE$  the **Error Sum of Squares**.

## Computing Formulas for Sums of Squares

- $SST = S_{yy}$
- $SSR = \frac{S_{xy}^2}{S_{xx}}$
- $SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$

**Mean Squares** a sum of squares divided by the degrees of freedom

- **Mean Square Error**  $MSE = \frac{SSE}{n-2} = s_e^2$
- **Mean Square due to regression**  $MSR = \frac{SSR}{1}$

**Coefficient of Determination**  $r^2 = \frac{SSR}{SST} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$

## Interpretations of $r^2$

- $r^2$  is the proportion of the variation in the observed values of the response variable explained by the regression.
- If  $r^2$  is near 0, then the regression equation is not useful for making predictions.
- If  $r^2$  is near 1, then the regression equation is very useful for making predictions.

## Regression Identity

$$SST = SSR + SSE$$

**Example** Compute  $SSE$ ,  $SSR$ , and  $SST$ , and the coefficient of determination for the shoe size data.

## Example

- Compute the coefficient of determination for the Wilt Chamberlain data.

- What percentage of variation in the data is explained by the regression equation?

## Linear Correlation Coefficient $r = \pm\sqrt{r^2}$

Facts about  $r$ .

- $r$  is between  $-1$  and  $1$ .
- If  $r$  is positive, then  $x$  and  $y$  are **positively linearly correlated**, ie,  $y$  tends to increase as  $x$  increases.
- If  $r$  is negative, then  $x$  and  $y$  are **negatively linearly correlated**.
- If  $r$  is close to  $1$  or  $-1$ , then there is a strong linear relationship.
- If  $r$  is close to  $0$ , there is not a strong linear relationship.

## T-test for the utility of a Regression Using the $F$ Statistic

**Assumptions:** The four assumptions for regression inferences are met.

- **Hypotheses**

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

- **Test Statistic**

$$F = \frac{MSR}{MSE}$$

- For samples of size  $n$ , each with the same values of

$$x_1, x_2, \dots, x_n,$$

this variable has the  $F$ -distribution with 1 numerator and  $n - 2$  denominator degrees of freedom.

- Use a Tables B.3,B.4, and B.5 in the appendix to compute test statistics

**Example** An economist is interested in the relationship between the disposable income of a family and the amount of money spent annually on food. A sample of 8 middle income families revealed the following, where  $x$  denotes disposable income, in thousands of dollars, and  $y$  denotes food expenditure, in hundreds of dollars.

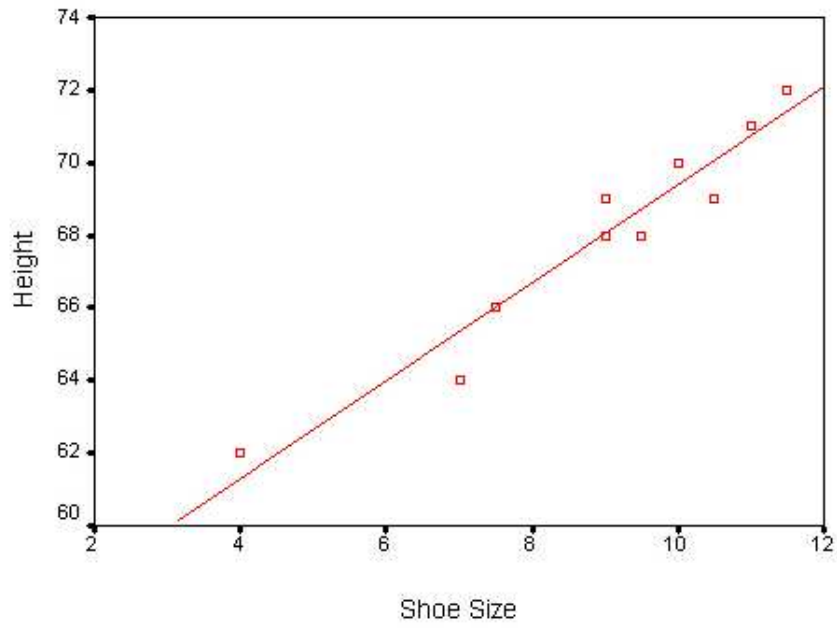
x	30	36	27	20	16	24	19	25
y	55	60	42	40	37	26	39	43

Do the data provide sufficient evidence to conclude that disposable income is useful as a predictor of annual food expenditure for middle income families?  $\alpha = 0.01$ .

### Comparison of $t$ test and $F$ test

- The  $F$  test only works for  $H_0 : \beta_1 = 0$  versus  $H_a : \beta_1 \neq 0$ . It does not work for a left or right tailed test.
- Either approach can be used when  $y$  is related to a single variable.
- $F = t^2$
- In multiple regression (chapter 4), the  $F$  statistic will sometimes be necessary, as the  $t$  test won't always work

Scatter Plot of Height by Shoe Size



# Scatter Plot of Wilt Chamberlain's Points Vs Rebounds

