Lecture 3

Stat102, Spring 2007

- Chapter 3.1 – 3.2:
  - Introduction to regression analysis
  - Linear regression as a descriptive technique
  - The least-squares equations
- Chapter 3.3
  - Sampling distribution of b_0, b_1.
 - Continued in next lecture

1

Regression Analysis

Galton's classic data on heights of parents and their child (952 pairs)

- Describes the relationship between child's height (y) and the parents' height (x).
- Predict the child's height given parents height.

Parent ht	Child ht
73.60	72.22
72.69	67.72
72.85	70.46
71.68	65.13
70.62	61.20
70.23	63.10
70.74	64.96
70.73	66.43
69.47	63.10
68.26	62.00
65.88	61.31
64.90	61.36
64.80	61.95
64.21	64.96

And more

2

1
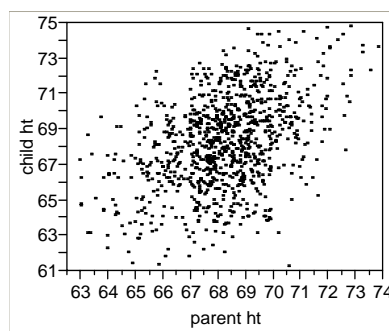
Uses of Regression Analysis

- Description: Describe the relationship between a dependent variable y (child's height) and explanatory variables x (parents' height).
- Prediction: Predict dependent variable y based on explanatory variables x.

3

Model for Simple Regression Model

- Consider a population of units on which the variables (y,x) are recorded.
- Let $\mu_{y|x}$ denote the conditional mean of y given x.
- The goal of regression analysis is to estimate $\mu_{y|x}$.
- Simple linear regression model:

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

4

Simple Linear Regression Model

- Model (more details later)

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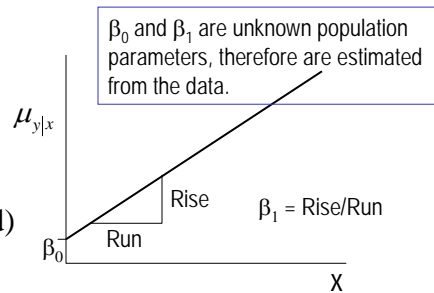
y = dependent variable

x = independent variable

β_0 = y-intercept

β_1 = slope of the line

e = error (normally distributed)

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

β_0 and β_1 are unknown population parameters, therefore are estimated from the data.

$\mu_{y|x}$

Rise

β_1 = Rise/Run
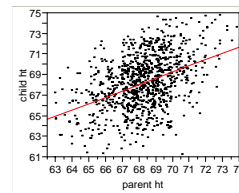
Run

β_0

X

5

Interpreting the Coefficients

- The slope β_1 is the change in the mean of y that is associated with a one unit change in x

   *e.g.,for each extra inch for parents, the average heights of the child increases by 0.6 inch.*

- The intercept is the estimated mean of y for x=0.

    However, this interpretation should only be used when the data contains observations with x near 0. Otherwise it is an extrapolation of the model which can be unreliable (Section 3.7.2).

child ht = 26.46 + 0.6 parent ht

6

Estimating the Coefficients

- The estimates are determined from
  - observations: $(x_1, y_1), \ldots, (x_n, y_n)$.
  - by calculating sample statistics.
  - Correspond to a straight line that cuts into the data.
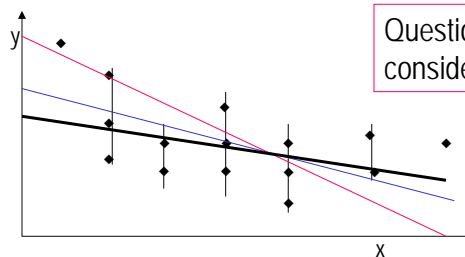
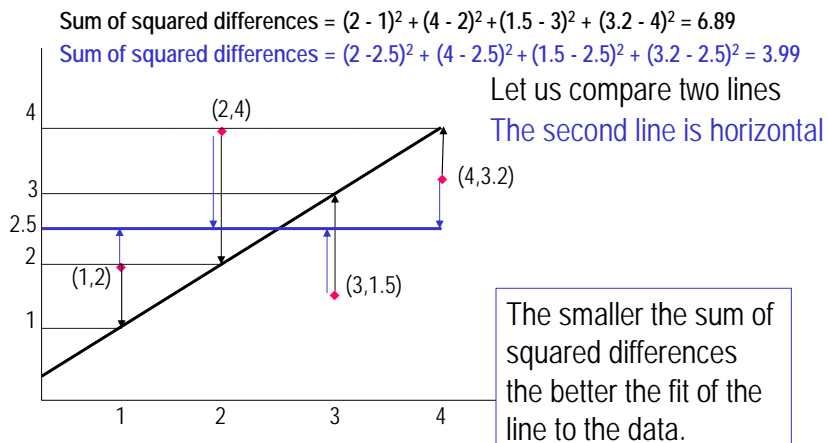Question: What should be considered a good line?

Least Squares Regression Line

- What is a good estimate of the line?
- A good estimated line should predict y well based on x.
  - Least absolute value regression line: Line that minimizes the absolute values of the prediction errors in the sample. Good criterion but hard to compute.
  - Least squares regression line: Line that minimizes the squared prediction errors in the sample. Good criterion and easy to compute.

The Least Squares (Regression) Line

Sum of squared differences = $(2 - 1)^2 + (4 - 2)^2 + (1.5 - 3)^2 + (3.2 - 4)^2 = 6.89$

Sum of squared differences = $(2 - 2.5)^2 + (4 - 2.5)^2 + (1.5 - 2.5)^2 + (3.2 - 2.5)^2 = 3.99$

Let us compare two lines

The second line is horizontal

(2,4)

(4,3.2)

(1,2)

(3,1.5)

The smaller the sum of squared differences the better the fit of the line to the data.

9

The Estimated Coefficients

To calculate the **estimates** of the coefficients of the line that minimizes the sum of the squared differences between the data points and the line, use the formulas:

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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

The regression equation that **estimates** the equation of the simple linear regression **model** is:

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

10

5

Example Heights (cont.)

- For simple linear regression analysis in JMP:
  - Click "Analyze, Fit Y by X"; then put child ht in Y and parent ht in X and click "OK".
  - Then click red triangle next to "Bivariate Fit" and click "Fit Line".
  - Some commands we will use later can now be found in the red triangle next to "Linear Fit"

11

Example Heights (cont)

Based on our observations, find b_1 and b_0

- The summary statistics for parent hts and child hts:

Child hts   Parent hts

Mean	68.20	Mean	68.27
Std Dev	2.60	Std Dev	1.79
Std Err Mean	0.084	Std Err Mean	0.0580
upper 95% Mean	68.37	upper 95% Mean	68.38
lower 95% Mean	68.04	lower 95% Mean	68.15
N	952	N	952

- For the regression line –

  From JMPIN $b_1 = 0.61$
  $$b_0 = \overline{Y} - b_1\overline{X} = 68.20 - 0.61 \times 68.27 = 26.55$$
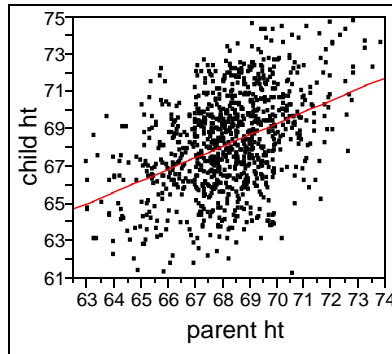
- The LS equation is
  $$\hat{y} = 26.55 + 0.61x$$

12

6

JMP Output

Bivariate Fit of child ht By parent ht

Linear Fit
child ht = 26.456 + 0.612 parent ht

13

JMP Output (cont)

Note the values of b_0, b_1 in the "parameter estimates" table
The other output entries will be explained later

Summary of Fit

RSquare	0.177
RSquare Adj	0.176
Root Mean Square Error	2.357
Mean of Response	68.202
Observations	952

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1136.50	1136.50	204.59
Error	950	5277.28	5.56	Prob > F
C. Total	951	6413.78		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	26.456	2.920	9.06	<.0001
parent ht	0.612	0.043	14.30	<.0001

14

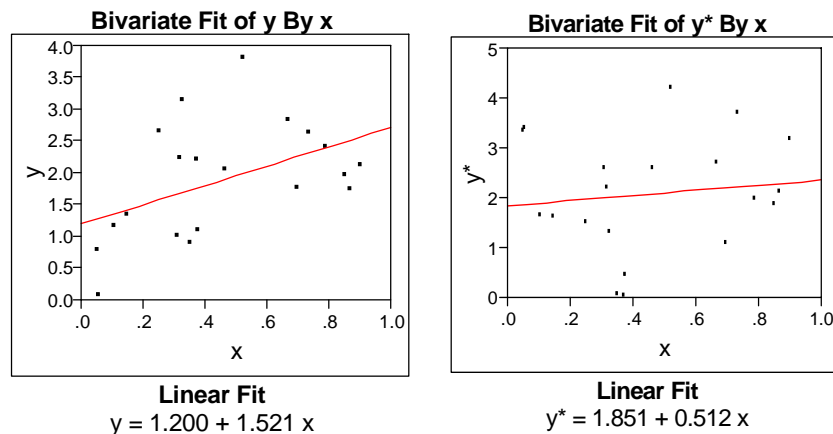## Ordinary Linear Model Assumptions

- Properties of errors under ideal model:
  - $\mu_{y|x} = \beta_0 + \beta_1 x$ for all x.
  - $y_i = \beta_0 + \beta_1 x_i + e_i$ for all x_i
  - The distribution of $e_i \mid x_i$ is normal.
  - $e_1, \ldots, e_n$ are independent.
  - $E(e_i \mid x_i) = 0$ and $Var(e_i \mid x_i) = \sigma_e^{\,2}$

- Equivalent definition: For each x_i, y_i has a normal distribution with mean $\beta_0 + \beta_1 x_i$ and variance $\sigma_e^{\,2}$. Also, $y_1, \cdots, y_n$ are independent.

15

Sampling Distribution of b_0, b_1

- The "sampling distribution" of b_0, b_1 is the probability distribution of the estimates over repeated samples $y_1, .., y_n$ from the ideal linear regression model with fixed values of β_0, β_1 and σ_e^2 and $x_1, .., x_n$.
- "Standardregression.jmp" contains a simulation of pairs $(x_1, y_1), .., (x_n, y_n)$ from a simple linear regression model with $\beta_0 = 1, \beta_1 = 2, \sigma_e^2 = 1.$ AND
- It contains another simulation labeled $(x_1, y_1^*), .., (x_n, y_n^*)$ from the same model.
- Notice the difference in the estimated coefficients calculated from the y's and from the y^*'s.

16

Bivariate Fit of y By x

Linear Fit
y = 1.200 + 1.521 x

Bivariate Fit of y* By x

Linear Fit
y* = 1.851 + 0.512 x

Two outcomes from "standardregression.jmp"
Each data set comes from the model with $\beta_0 = 1$, $\beta_1 = 2$, $\sigma_e^2 = 1$
The values of $x_1, ..., x_{20}$ are the same in both data sets

17

Sampling Distribution (Details)

- b_0 and b_1 have easily described normal distributions
- Sampling distribution of b_0 is normal with

$$E(b_0) = \beta_0 \quad \text{(Hence the estimate is "unbiased")}$$

$$Var(b_0) = \sigma_e^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} \right) \text{ where } s_x^2 \triangleq \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$$

- Sampling distribution of b_0 is normal with

$$E(b_1) = \beta_1 \quad \text{(Hence the estimate is "unbiased")}$$

$$Var(b_1) = \frac{\sigma_e^2}{(n-1)s_x^2}$$

18

Typical Regression Analysis

1. Observe pairs of data $(x_1,y_1),\ldots,(x_n,y_n)$ that are a sample from population of interest.
2. Plot the data.
3. Assume simple linear regression model assumptions hold.
4. Estimate the true regression line $\mu_{y|x} = \beta_0 + \beta_1 x$ by the least squares line $\hat{\mu}_{y|x} = b_0 + b_1 x$
5. Check whether the assumptions of the ideal model are reasonable (Chapter 6, and next lecture)
6. Make inferences concerning coefficients β_0, β_1 and make predictions ($\hat{y} = b_0 + b_1 x$)

19

Notes

Formulas for the least squares equations:

   1.   The equations for b_0 and b_1 are easy to derive. Here is a derivation that involves a little bit of calculus:

   It is desired to minimize the sum of squared errors. Symbolically, this is

$$SSE(b_0,b_1) = \sum_i \left(y_i - (b_0 + b_1 x_i) \right)^2.$$

The minimum occurs when $0 = \dfrac{\partial}{\partial b_1} SSE(b_0,b_1)$ and $0 = \dfrac{\partial}{\partial b_0} SSE(b_0,b_1)$.

   Hence we need

$$0 = \frac{\partial}{\partial b_1} SSE(b_0,b_1) = -2\sum x_i \left(y_i - (b_0 + b_1 x_i) \right) \text{ and}$$

$$0 = \frac{\partial}{\partial b_0} SSE(b_0,b_1) = -2\sum \left(y_i - (b_0 + b_1 x_i) \right).$$

   These are two linear equations in the two unknowns b_0 and b_1. Some algebraic manipulation shows that the solution can be written in the desired form –

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \text{ and } b_0 = \bar{y} - b_1 \bar{x}.$$

20

2. A NICE FACT that's sometimes useful:

      a. The least squares line passes through the point $(\bar{x}, \bar{y})$.

To see this note that if $x = \bar{x}$ then the corresponding point on the least squares line is $\hat{y} = b_0 + b_1 \bar{x}$. Substituting the definition of b_0 yields $\hat{y} = (\bar{y} - b_1 \bar{x}) + b_1 \bar{x} = \bar{y}$, as claimed.

      b. The equation for the least squares line can be re-written in the form

$$y - \bar{y} = b_1 (x - \bar{x}).$$

3. There are other useful ways to write the equations for b_0 and b_1. Recall that the sample covariance is defined as

$$Cov(\{x_i, y_i\}) = \frac{1}{n-1} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \triangleq S_{xy}, \text{ say.}$$

Similarly, the sample correlation coefficient is

$$\frac{S_{xy}}{\sqrt{S_x^2 S_y^2}} \triangleq R, \text{ say.}$$

[$S_x^2 = s_x^2$ is defined on overhead 18, and S_y^2 is defined similarly.]

    Thus,

$$b_1 = \frac{S_{xy}}{S_x^2} = \frac{S_y}{S_x} \frac{S_{xy}}{\sqrt{S_x^2 S_y^2}} = \frac{S_y}{S_x} R.$$

21

History of Galton's Data:

4. Francis Galton gathered data about heights of parents and their children, and published the analysis in 1886 in a paper entitled "Regression towards mediocrity *[sic]* in hereditary stature". In the process he coined the term "Regression" to describe the straight line that summarizes the type of relational data that may appear in a scatterplot.

    He did not use our current least-squares technique for finding this line; instead he used a clever analysis whose final step is to fit the line by eye. He estimated the slope of the regression line as $2/3$

    Further work in the next decades by Galton and by K. Pearson, Gossett (*writing as* "A. Student") and others connected Galton's analysis to the least squares technique earlier invented by Gauss (1809), and also derived the relevant sampling distributions needed to create a statistical regression analysis.

5. The data we use for our analysis is packaged with the JMP program disk. It is not exactly Galton's original data. We believe it is a version of the data set prepared by S. Stigler (1986) as a minor modification of Galton's data. In order for the data to plot nicely, Stigler "jittered" the data. He also included some data that Galton did not. The data listed as "**Parent height**" in this data set is actually the ***average of both parents' heights***, after adjusting the mothers' heights as discussed in the next note.

22

6. Galton did not know how to separately treat men's and women's heights in order to produce the kind of results he wanted to look at. SO (after looking at the structure of the data) *he multiplied all female heights by 1.08*. This puts all the heights on very nearly the same scale, and allowed him to treat mens' and womens' heights together, without regard to sex.

[Instead of doing this Galton could have divided the mens' heights by 1.08; or he could have achieved a similar effect by *dividing* the male heights by 1.04 **and** *multiplying* the female ones by 1.04. *Why didn't he use one of these other schemes?*]

7. Galton did not use modern random-sampling methods to obtain his data. Instead, he obtained his data "through the offer of prizes" for the "best extracts from their own family records" obtained from individual family correspondents. He summarized the data in a journal that is now in the Library of the University College of London. Here is what the first half of p. 4 looks like. (According to Galton's notations one should "add 60 inches to every entry in the Table".)

23

Half of p4 of Galton's Journal
(note the approximate heights for some records, and the entries "tall" and "deformed")

This photocopy, as well as much of the above discussion is taken from Hanley, J. A. (2004), "Transmuting women into men: Galton's family data on human stature", *The Amer. Statistician*, **58**, p237-243. Another excellent reference is Stigler, S. (1986) "The English breakthrough: Galton" in *The History of Statistics: The Measurement of Uncertainty before 1900*, Harvard Univ. Press.

24