

Statistical View of Least Squares

Joyee Ghosh

May 23, 2006

An Introduction to Regression

Purpose of Regression

Some Examples

Least Squares

Simple Linear Regression

Purpose of Regression

- ▶ Suppose we have two variables x and y

Purpose of Regression

- ▶ Suppose we have two variables x and y
- ▶ To explore the relationship between two variables

Purpose of Regression

- ▶ Suppose we have two variables x and y
- ▶ To explore the relationship between two variables
- ▶ To predict the value of one variable as the other changes

Purpose of Regression

- ▶ Suppose we have two variables x and y
- ▶ To explore the relationship between two variables
- ▶ To predict the value of one variable as the other changes
- ▶ Identify unusual observations

Some Examples

- ▶ example 1. Income and Education

Some Examples

- ▶ example 1. Income and Education
- ▶ example 2. Icecream consumption and temperature

Some Examples

- ▶ example 1. Income and Education
- ▶ example 2. Icecream consumption and temperature
- ▶ example 3. ozone levels and air pollution in New York City

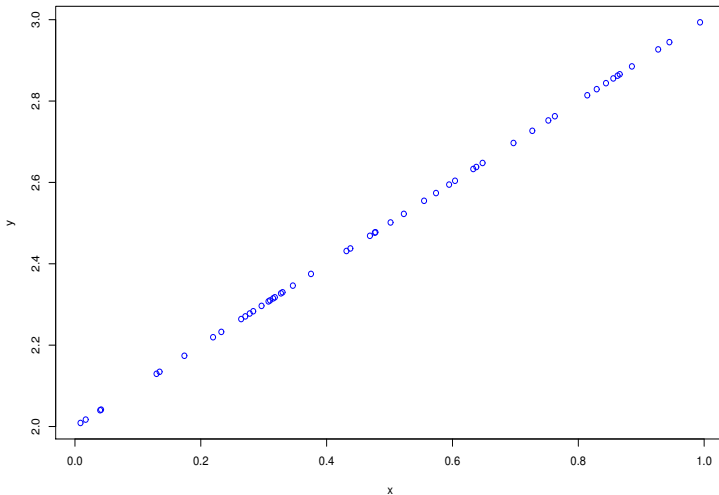
Some Examples

- ▶ example 1. Income and Education
- ▶ example 2. Icecream consumption and temperature
- ▶ example 3. ozone levels and air pollution in New York City
- ▶ Usually we denote the independent or explanatory variable by 'x' and the dependent or response variable by 'y'

Some Examples

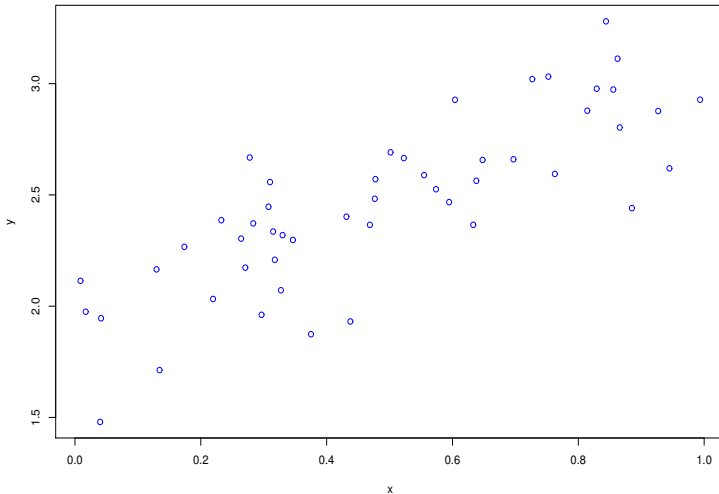
- ▶ example 1. Income and Education
- ▶ example 2. Icecream consumption and temperature
- ▶ example 3. ozone levels and air pollution in New York City
- ▶ Usually we denote the independent or explanatory variable by 'x' and the dependent or response variable by 'y'
- ▶ Can you identify the 'x' and the 'y' here?

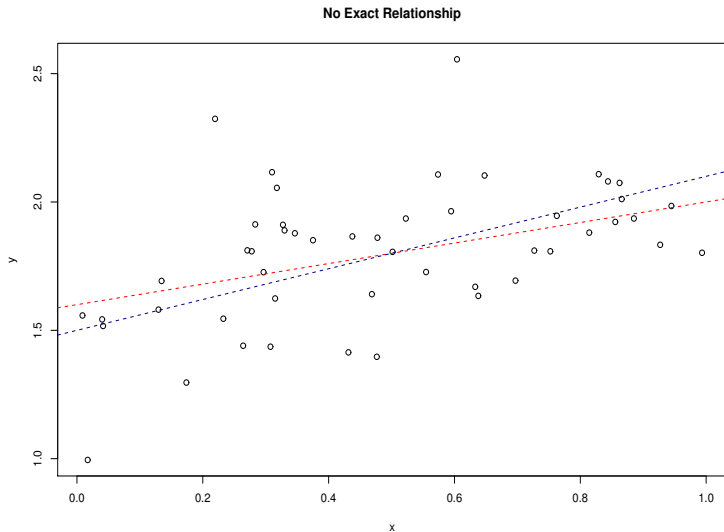
Exact Relationship: $y = 2+x$



- ▶ Usually for real data there is no exact relationship
- ▶ Hence no single line will pass through all the points in the scatterplot

No Exact Relationship





Least Squares

- ▶ Different people might draw different lines using eye estimation
- ▶ We need to draw a line that doesn't depend on our guess
- ▶ error = observed y - predicted y

$$\hat{e}_i = y_i - \hat{y}_i \quad (1)$$

Least Squares

- ▶ Different people might draw different lines using eye estimation
- ▶ We need to draw a line that doesn't depend on our guess
- ▶ error = observed y - predicted y

$$\hat{e}_i = y_i - \hat{y}_i \quad (1)$$

- ▶ We want to minimize the aggregate error

Least Squares

- ▶ Different people might draw different lines using eye estimation
- ▶ We need to draw a line that doesn't depend on our guess
- ▶ error = observed y - predicted y

$$\hat{e}_i = y_i - \hat{y}_i \quad (1)$$

- ▶ We want to minimize the aggregate error
- ▶ One common way: **Least Squares** to minimize the sum of squared errors

Mathematical Representation

Suppose we have n observations

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (2)$$

- ▶ y_i : value of the response for the i^{th} observation
- ▶ x_i : value of the predictor variable for the i^{th} observation
- ▶ β_0 : intercept
- ▶ β_1 : slope
- ▶ ϵ_i : random error term corresponding to i^{th} observation

Estimation Using Least Squares

We want to minimize the sum of squared errors

$$E = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (3)$$

- ▶ We want to find that pair of (β_0, β_1) for which E is smallest

Estimation Using Least Squares

We want to minimize the sum of squared errors

$$E = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (3)$$

- ▶ We want to find that pair of (β_0, β_1) for which E is smallest
- ▶ we'll call them $\hat{\beta}_0$ $\hat{\beta}_1$

Estimation Using Least Squares

We want to minimize the sum of squared errors

$$E = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (3)$$

- ▶ We want to find that pair of (β_0, β_1) for which E is smallest
- ▶ we'll call them $\hat{\beta}_0$ $\hat{\beta}_1$
- ▶ Let's find them analytically

Normal Equations

To find $\hat{\beta}_0$ $\hat{\beta}_1$ we take the partial derivatives of E with respect to β_0 and β_1 and get the following equations called normal equations:

$$\frac{\delta E}{\delta \beta_0} = 0 \quad (4)$$

$$\frac{\delta E}{\delta \beta_1} = 0 \quad (5)$$

- ▶ problem 1. Solve these two equations simultaneously to obtain $\hat{\beta}_0$ $\hat{\beta}_1$

Assumptions made for doing Statistical Inference

- ▶ Now once we have got the estimates we may want to see how good they are that is how close are our estimates and the true values

Assumptions made for doing Statistical Inference

- ▶ Now once we have got the estimates we may want to see how good they are that is how close are our estimates and the true values
- ▶ In order to do that we have to make some more assumptions

Assumptions made for doing Statistical Inference

- ▶ Now once we have got the estimates we may want to see how good they are that is how close are our estimates and the true values
- ▶ In order to do that we have to make some more assumptions
- ▶ Remember our model was

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (6)$$

Assumptions made for doing Statistical Inference

- ▶ Now once we have got the estimates we may want to see how good they are that is how close are our estimates and the true values
- ▶ In order to do that we have to make some more assumptions
- ▶ Remember our model was

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (6)$$

- ▶ We now assume $\epsilon_i \sim N(0, \sigma^2)$

Assumptions made for doing Statistical Inference

- ▶ Now once we have got the estimates we may want to see how good they are that is how close are our estimates and the true values
- ▶ In order to do that we have to make some more assumptions
- ▶ Remember our model was

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (6)$$

- ▶ We now assume $\epsilon_i \sim N(0, \sigma^2)$
- ▶ we also assume ϵ_i and ϵ_j are independent for $i \neq j$

Checking Model Assumptions

- ▶ Always do some exploratory data analysis before making inference

Checking Model Assumptions

- ▶ Always do some exploratory data analysis before making inference
- ▶ The results will not make sense if the assumptions are terribly violated

Checking Model Assumptions

- ▶ Always do some exploratory data analysis before making inference
- ▶ The results will not make sense if the assumptions are terribly violated
- ▶

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (7)$$

\hat{y}_i 's are called fitted values

Checking Model Assumptions

- ▶ Always do some exploratory data analysis before making inference
- ▶ The results will not make sense if the assumptions are terribly violated



$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (7)$$

\hat{y}_i 's are called fitted values



$$\hat{e}_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (8)$$

\hat{e}_i 's are called residuals

Checking Model Assumptions

- ▶ Always do some exploratory data analysis before making inference
- ▶ The results will not make sense if the assumptions are terribly violated



$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (7)$$

\hat{y}_i 's are called fitted values



$$\hat{e}_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (8)$$

\hat{e}_i 's are called residuals

- ▶ The residuals are used to check if any of the model assumptions have been violated

Checking Model Assumptions

- ▶ We will look at the plot of the residuals vs. the fitted values

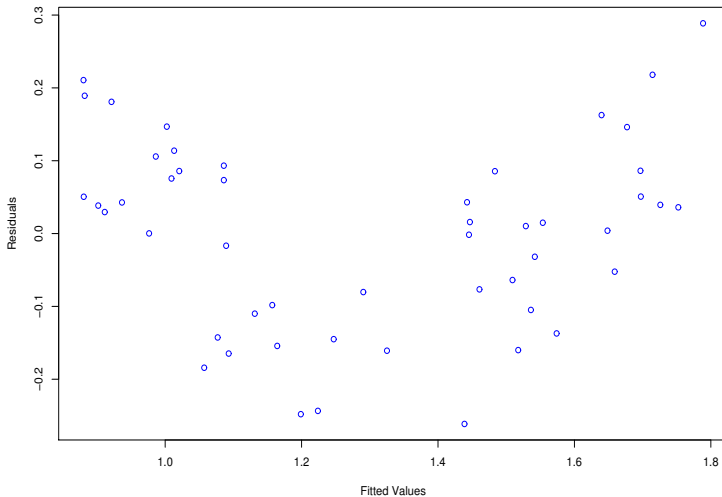
Checking Model Assumptions

- ▶ We will look at the plot of the residuals vs. the fitted values
- ▶ If the assumption of independence really holds we would expect the residuals to fluctuate in a random pattern around 0

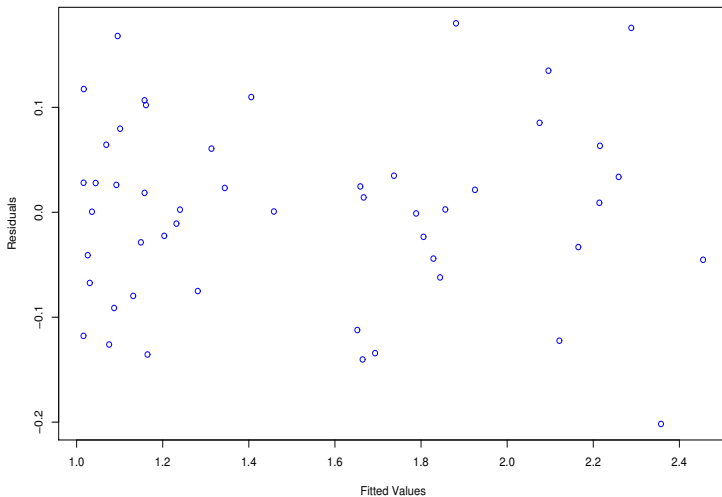
Checking Model Assumptions

- ▶ We will look at the plot of the residuals vs. the fitted values
- ▶ If the assumption of independence really holds we would expect the residuals to fluctuate in a random pattern around 0
- ▶ A curved pattern may indicate that the relationship between y and x is not linear

Residual Plot Where y is a Quadratic Function of x



Residual Plot After Making the Explanatory Variable x^2



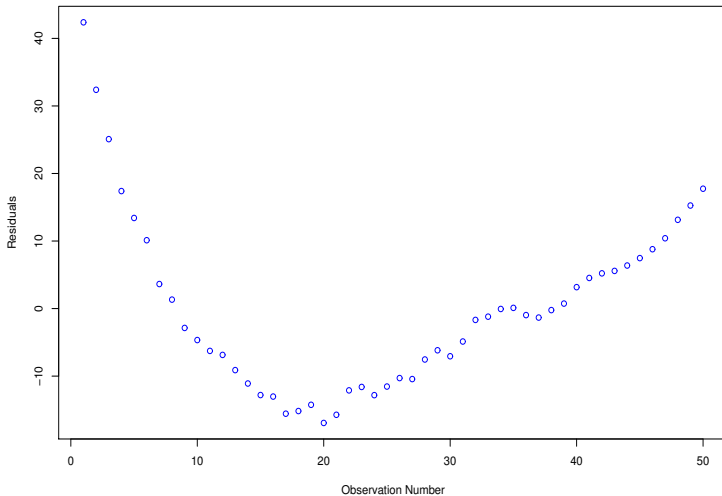
Checking Model Assumptions

- ▶ If there's some distinct pattern like a sinusoidal curve we'll start worrying that they are correlated

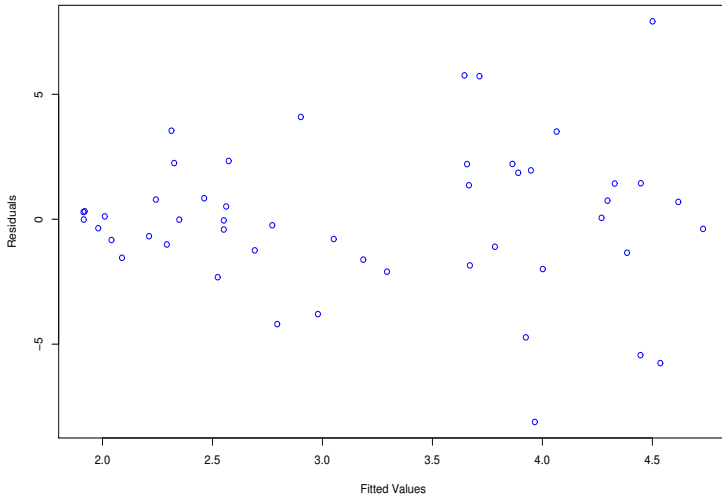
Checking Model Assumptions

- ▶ If there's some distinct pattern like a sinusoidal curve we'll start worrying that they are correlated
- ▶ To check whether the assumption of constant variance holds we also look at the residual plots

Residual Plot Where the Observations are Correlated



Residual Plot for Non-constant Variance



Checking Model Assumptions

- ▶ If an observation lies outside the general pattern of the other observations it is called an outlier

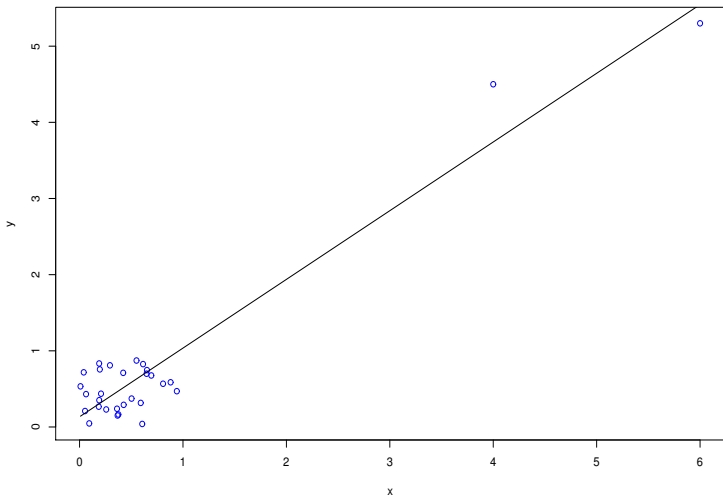
Checking Model Assumptions

- ▶ If an observation lies outside the general pattern of the other observations it is called an outlier
- ▶ An outlier has the potential to change the Least Squares Estimates dramatically

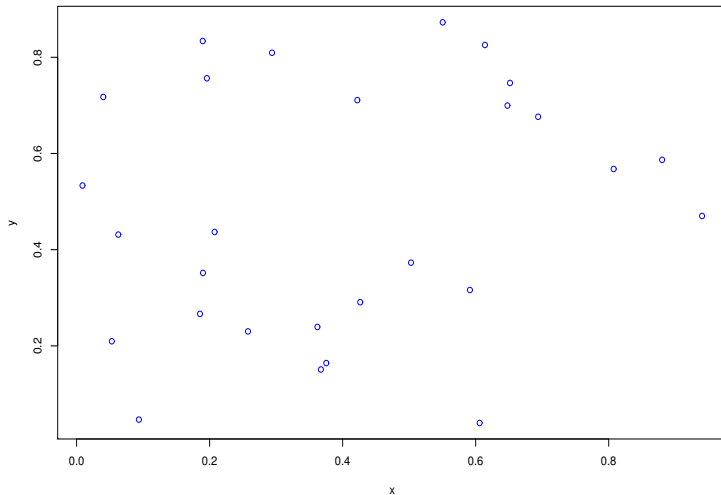
Checking Model Assumptions

- ▶ If an observation lies outside the general pattern of the other observations it is called an outlier
- ▶ An outlier has the potential to change the Least Squares Estimates dramatically
- ▶ Ideally we should make a separate analysis after excluding the outlier to see how much effect it has on the regression line and report both

Scatterplot and Fitted Regression Line



Scatterplot Without the Two Outliers



Matrix Representation

The regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

can also be written using matrix notation as:

$$y = X\beta + \epsilon \tag{9}$$

- ▶ How does X look like here?

Matrix Representation

The normal equations are

$$X'X\beta = X'y \quad (10)$$

estimate of β :

$$\hat{\beta} = (X'X)^{-1}X'y \quad (11)$$

residuals:

$$e = Y - \hat{y} = y - X\hat{\beta} \quad (12)$$

Problem 2.

Now we have all the tools to actually implement these using matlab

- ▶ Download the dataset “smoke.txt”
- ▶ Look at the scatterplot first to see if doing linear regression seems appropriate
- ▶ Find the least square estimates of the model parameters
- ▶ Look at residual plots
- ▶ Do the assumptions seem valid here?
- ▶ Look at the graph of fitted values superimposed on the observed values
- ▶ How does the fit look?