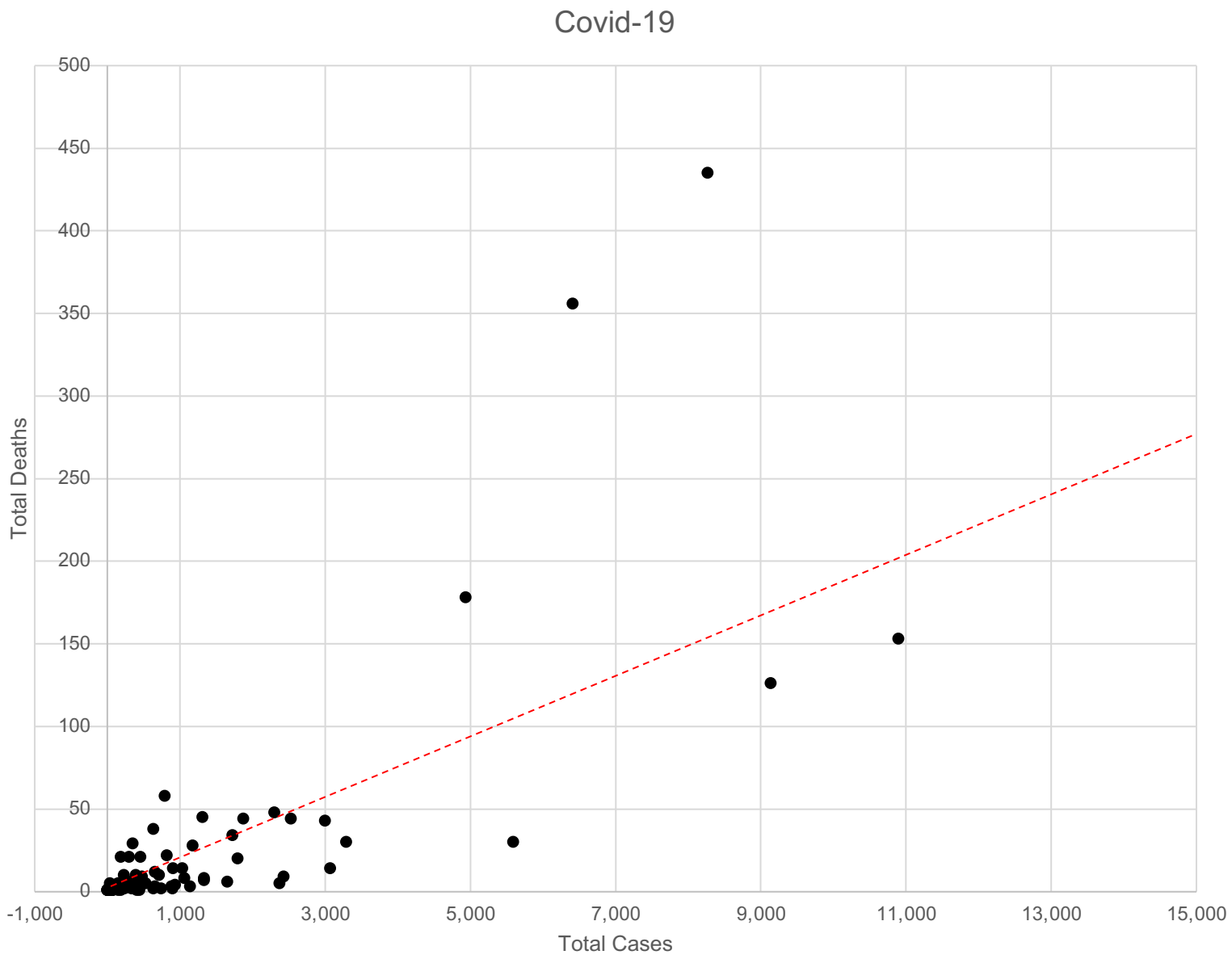
Nile University

Spring 2024

CIT651 – Introduction to Machine Learning and Statistical Data Analysis

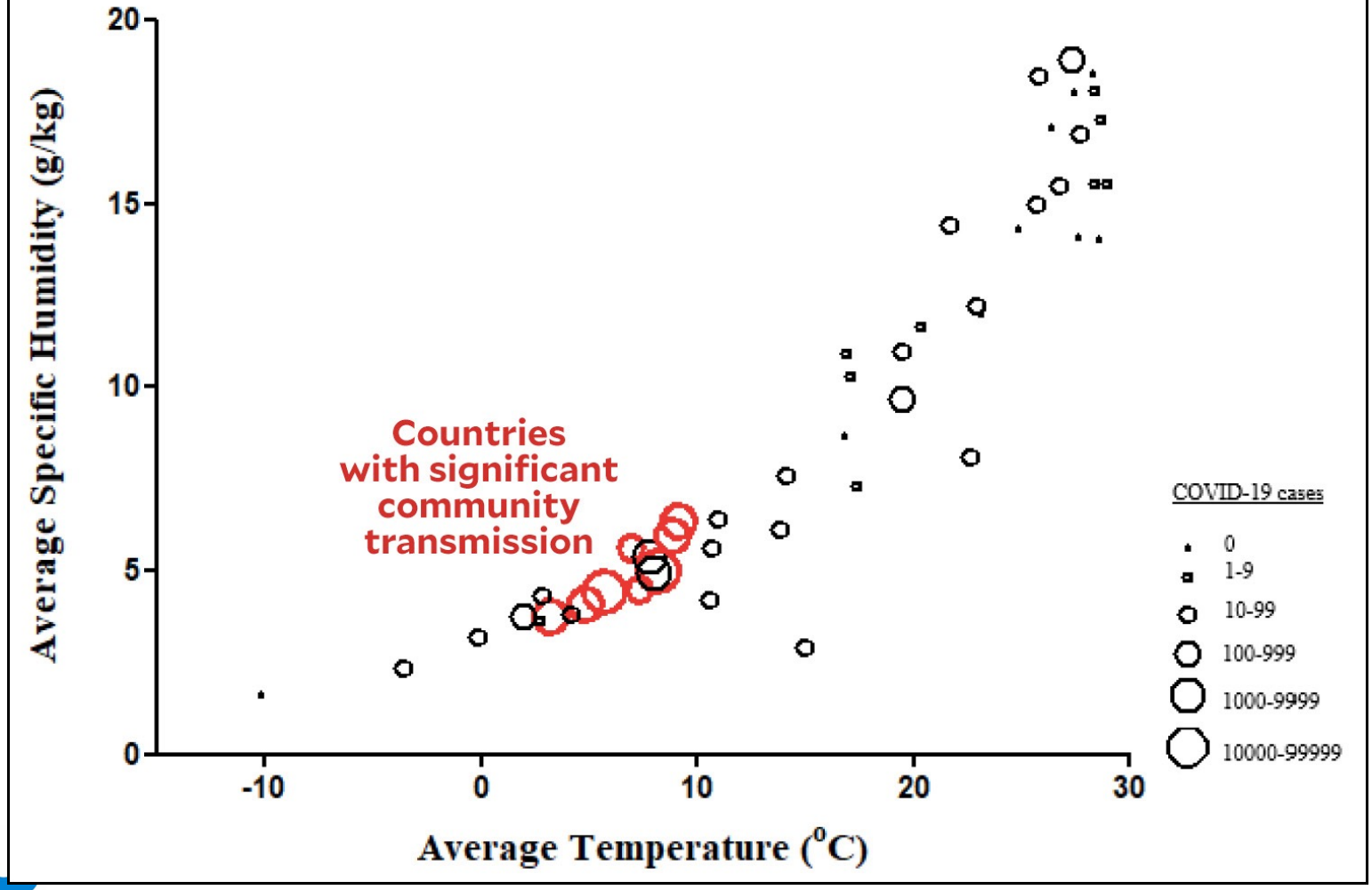
Lec 4 - Statistics Basics – Regression

Mustafa Elattar


Covid-19
Total Deaths
500
450
400
350
300
250
200
150
100
50
0
-1,000
1,000
3,000
5,000
7,000
9,000
11,000
13,000
15,000
Total Cases

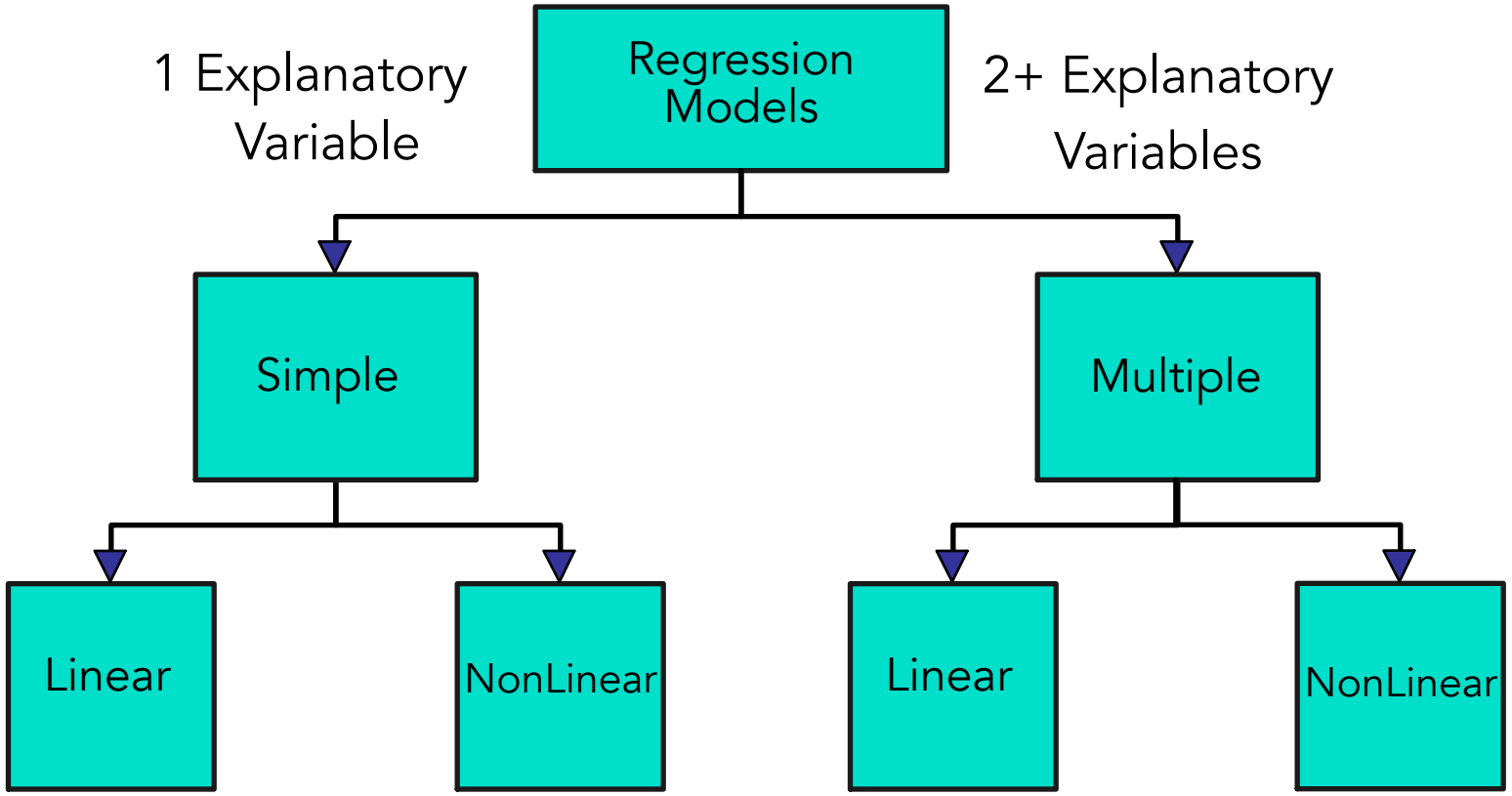
Spread of COVID-19

Mostly restricted to a narrow band of temperature and specific humidity

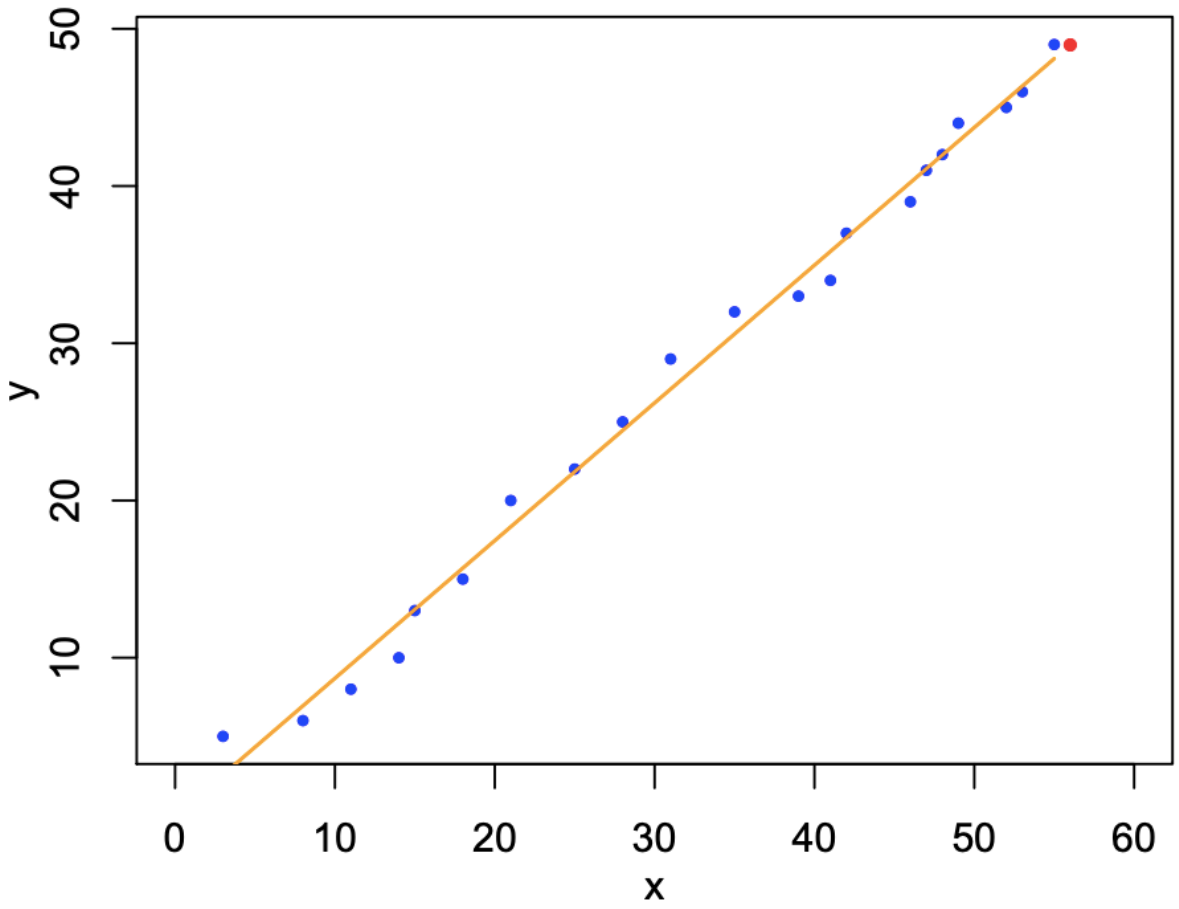


Nile University

Types of Regression Models



Simple linear regression



Stamp cost (cents) vs. time (years since 1960) (Red dot = 49 cents is predicted cost in 2016.) (Actual cost of a stamp dropped from 49 to 47 cents on 4/8/16.)

Linear Regression for Bivariate Data

- Ingredients:
  - Bivariate data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
  - Model: $y_i = f(x_i) + E_i$
  - where f(x) is some function, E_i random error.
- Total squared error:

$$\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

- Model allows us to predict the value of y for any given value of x.
- x is called the independent or predictor variable.
- y is the dependent or response variable.

Simple linear regression

- Finding the best fitting line
- Bivariate data $(x_1, y_1), \ldots, (x_n, y_n)$.
- Simple linear regression: fit a line to the data

$$y_i = ax_i + b + E_i, where, E_i \sim N(0, \sigma^2)$$

σ is a fixed value

- Total squared error:

$$\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b_i)^2$$

- Goal: Find the values of a and b that give the 'best fitting line'.
- Best fit: (least squares) The values of a and b that minimize the total squared error.
- Suppose that we want to obtain a point estimate (a reasonable value) of a population parameter.

What is linear about linear regression?

- Linear in the parameters a, b, ...

$$y = ax + b$$
$$y = ax + bx + c.$$

- It is not because the curve being fit has to be a straight line –although this is the simplest and most common case.
- Notice: in the board question you had to solve a system of simultaneous linear equations. Fitting a line is called simple linear regression.

Simple Linear Regression

$y_i = a x_i + b + e_i$ where $e_i \in N(0, \sigma^2)$

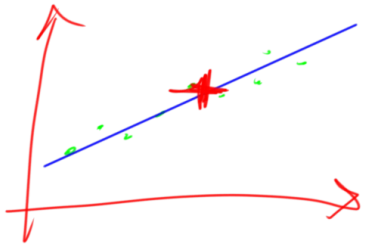
$e_i = y_i - a x_i - b \Rightarrow E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a x_i - b)^2$

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

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

$$= \sum (y_i - a x_i - \bar{y} - a\bar{x})^2$$

$$= \sum [(y_i - \bar{y}) - a(x_i - \bar{x})]^2$$



$$\frac{\partial E}{\partial a} = 2 \sum [(y_i - \bar{y}) - a(x_i - \bar{x})](x_i - \bar{x})(-1)$$

$$\frac{\partial E}{\partial a} = 2 \sum [(y_i - \bar{y}) - a(x_i - \bar{x})](x_i - \bar{x})(-1) = 0$$

$$\sum [(y_i - \bar{y})(x_i - \bar{x}) + a(x_i - \bar{x})^2] = 0$$

$$-\sum (y_i - \bar{y})(x_i - \bar{x}) + a \sum (x_i - \bar{x})^2 = 0$$

$$\Rightarrow a = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{cov(x, y)}{var(x)}$$

$$\Rightarrow b = \bar{y} - a\bar{x}$$

Formulas for simple linear regression

- Model:

$$y_i = ax_i + b + E_i, where, E_i \sim N(0, \sigma^2)$$

- Using calculus and algebra:

$$a = \frac{S_{xy}}{S_{xx}} \; and \; b = \bar{y} - a\bar{x}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$S_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \,, and \; S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- This is just for simple linear regression. For polynomials and other functions you need other formulas.

Formulas for simple linear regression

Bivariate data: (1, 3), (2, 1), (4, 4)

1. Calculate the sample means for x and y.
2. Use the formulas to find a best-fit line in the xy-plane.

$$y_i = ax_i + b$$

$$a = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \; and \; b = \bar{y} - a\bar{x}$$

3. Show the point $(\bar{x}, \bar{y})$ is always on the fitted line.

Measuring the fit

- $y = (y_1, y_2, \ldots, y_n)$ = data values of the response variable.
- $y = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$ = 'fitted values' of the response variable.
- TSS = total sum of squares = total variation.

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

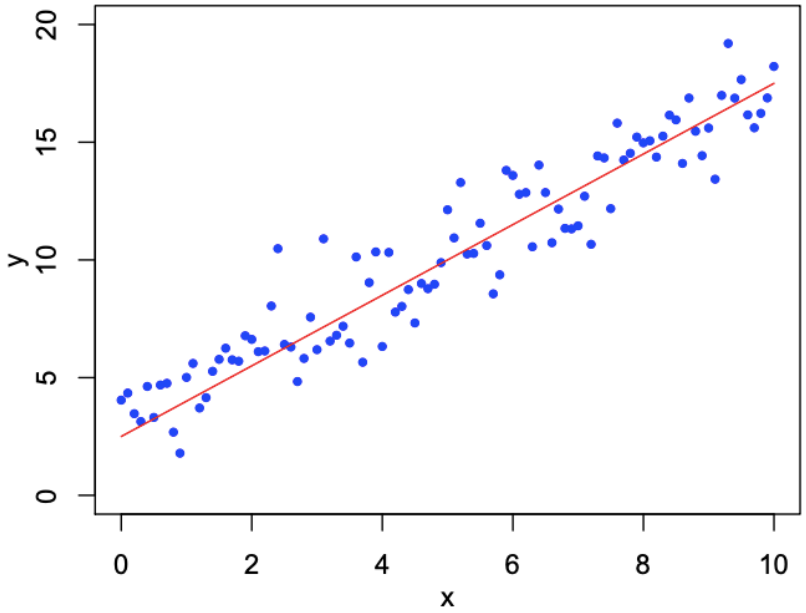
- RSS = residual sum of squares.

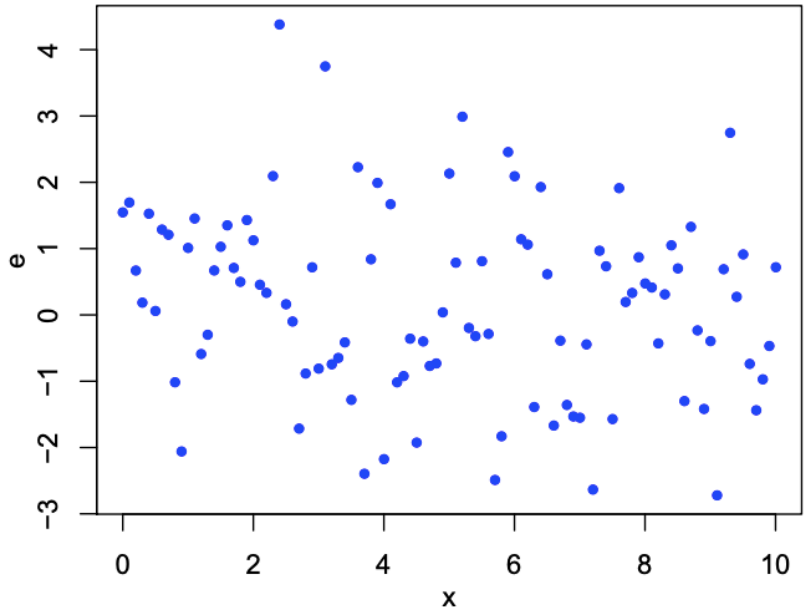
$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

- RSS = unexplained by model squared error (due to random fluctuation)
  RSS/TSS = unexplained fraction of the total error.
- $R^2 = 1 - \text{RSS/TSS}$ is measure of goodness-of-fit
- R^2 is the fraction of the variance of y explained by the model

Homoscedastic

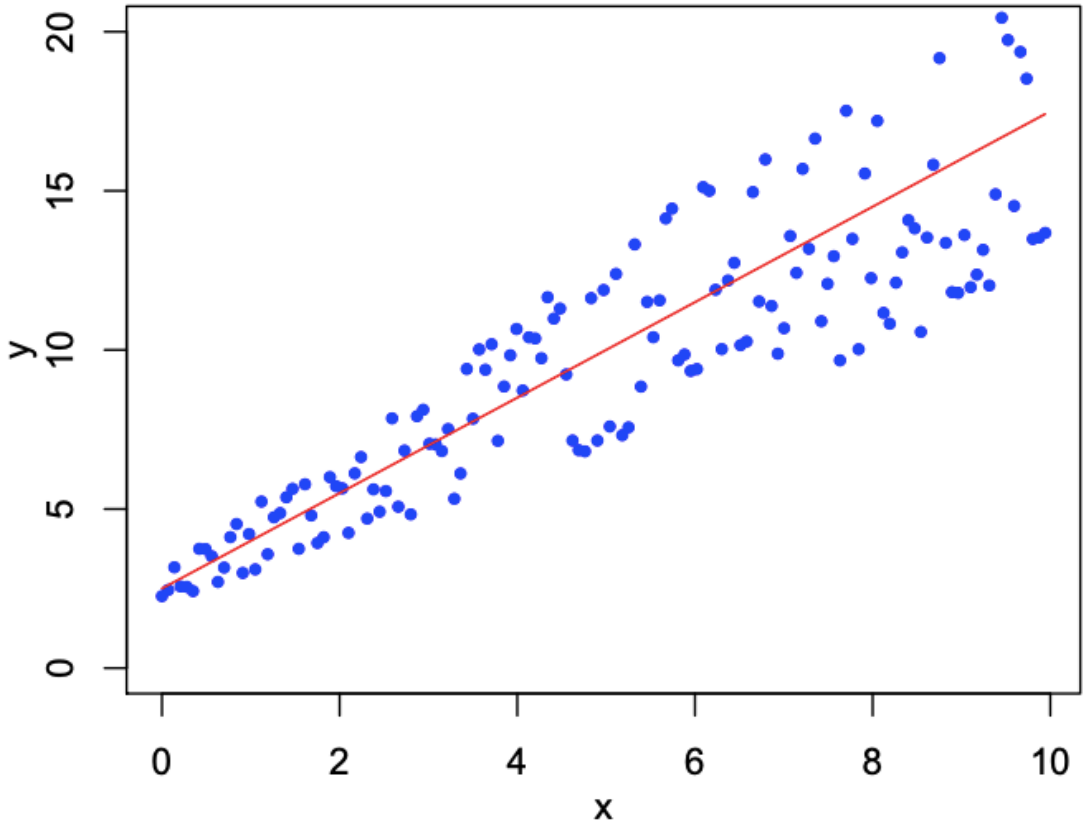
- BIG ASSUMPTIONS: the E_i are independent with the same variance σ^2.



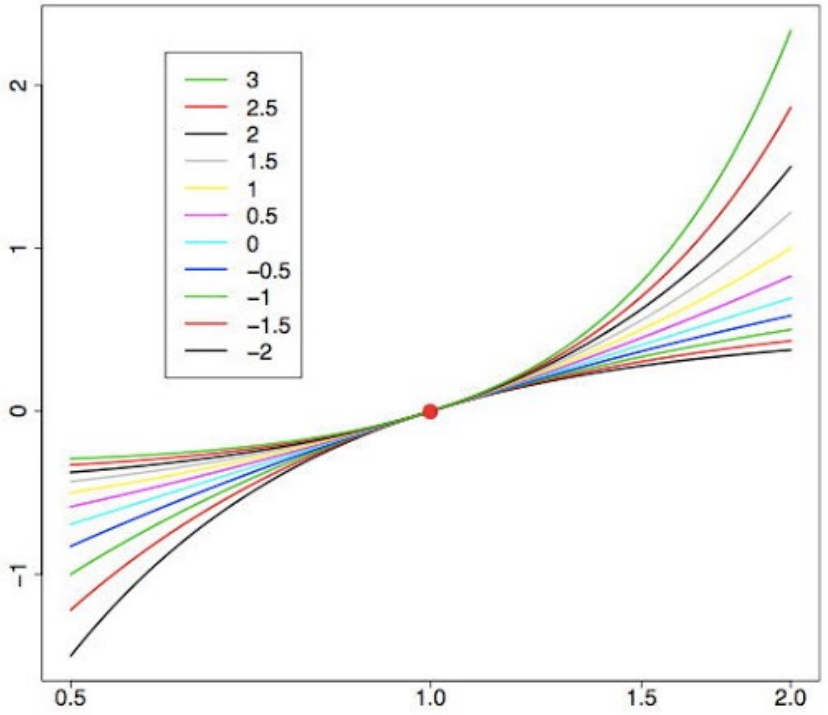


- Regression line (left) and residuals (right).
- Homoscedasticity = uniform spread of errors around regression line.

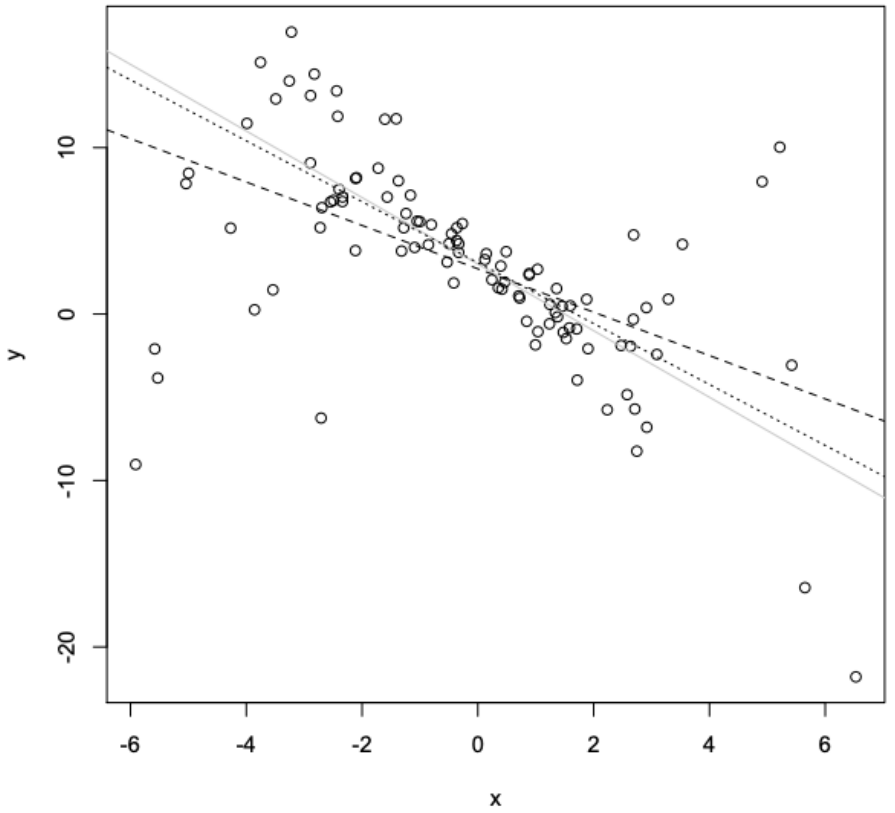
Heteroscedastic



Solution



Transform the dependent variable



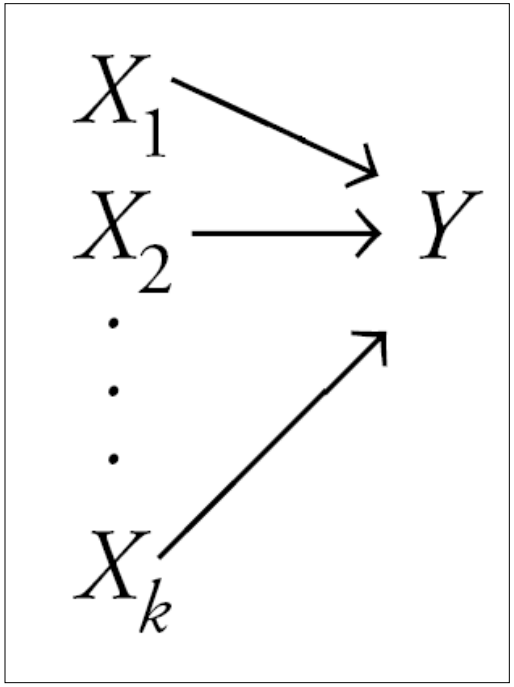

```
fit.wls = lm(y~x, weights=1/(1+0.5*x^2))
abline(fit.wls$coefficients,lty=3)
```

Figure 5: Figure 2, with addition of weighted least squares regression line (dotted).

Weighted regression

Multiple linear regression

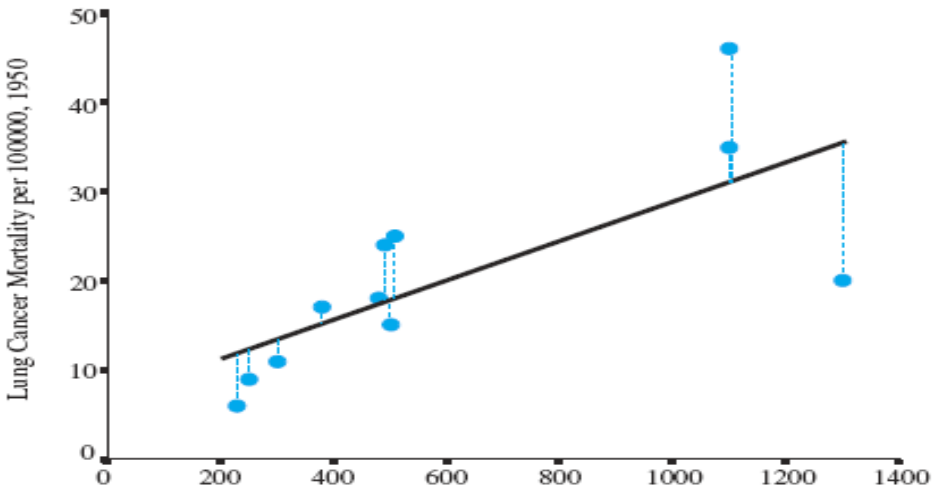
- Multiple regression simultaneously considers the influence of multiple explanatory variables on a response variable Y
- The intent is to look at the independent effect of each variable while "adjusting out" the influence of potential confounders



Multiple linear regression

- A simple regression model (one independent variable) fits a regression line in 2-dimensional space

- A multiple regression model with two explanatory variables fits a regression plane in 3-dimensional space



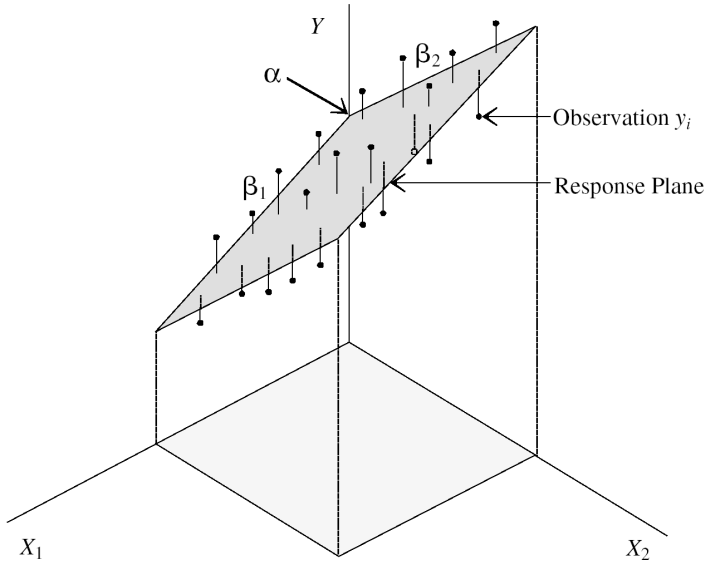


FIGURE 15.1 Three-dimensional response plane.

Multiple linear regression

- Linear Regression: finding the best fitting polynomial
- Bivariate data: $(x_1, y_1), \ldots, (x_n, y_n)$.
- Linear regression: fit a parabola to the data

$$y_i = ax_i^2 + bx_i + c + E_i, \qquad where\ E_i \sim N(0, \sigma^2)$$

and where **σ** is a fixed value, the same for all data points.

- Total squared error:

$$\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - ax_i^2 - bx_i - c)^2$$

- Goal: Find the values of a, b, c that give the 'best fitting parabola'.
- Best fit: (least squares) The values of a, b, c that minimize the total squared error.
- Can also fit higher order polynomials.

Overfitting a polynomial

- Increasing the degree of the polynomial increases R^2
- Increasing the degree of the polynomial increases the complexity of the model.
- The optimal degree is a tradeoff between goodness of fit and complexity.
- If all data points lie on the fitted curve, then $y = \hat{y}_n$ and $R^2 = 1$.

- Outliers and other troubles Question: Can one point change the regression line significantly?

Multiple Linear Regression

- Multivariate data: ($x_{i,1}$, $x_{i,2}$, . . . , $x_{i,m}$, y_i) (n data points: i = 1, . . . , n)
- Model:

$$\hat{y}_i = a_1 x_{i,1} + a_2 x_{i,2} + \cdots + a_m x_{i,m}$$

- $x_{i,j}$ are the explanatory (or predictor) variables.

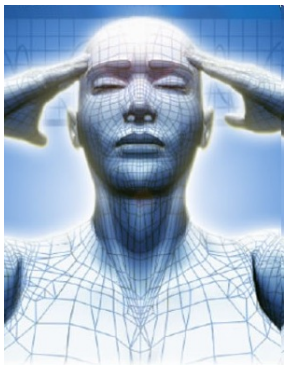
- y_i is the response variable.

- The total squared error is

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a_1 x_{i,1} - a_2 x_{i,2} - \cdots - a_m x_{i,m}))^2$$

Multiple Linear Regression

- Intercept **α** predicts where the regression plane crosses the Y axis
- Slope for variable X_1 ($\boldsymbol{\beta}_1$) predicts the change in Y per unit X_1 holding X_2 constant
- The slope for variable X_2 ($\boldsymbol{\beta}_2$) predicts the change in Y per unit X_2 holding X_1 constant
- A multiple regression model with k independent variables fits a regression "surface" in k + 1 dimensional space (cannot be visualized)



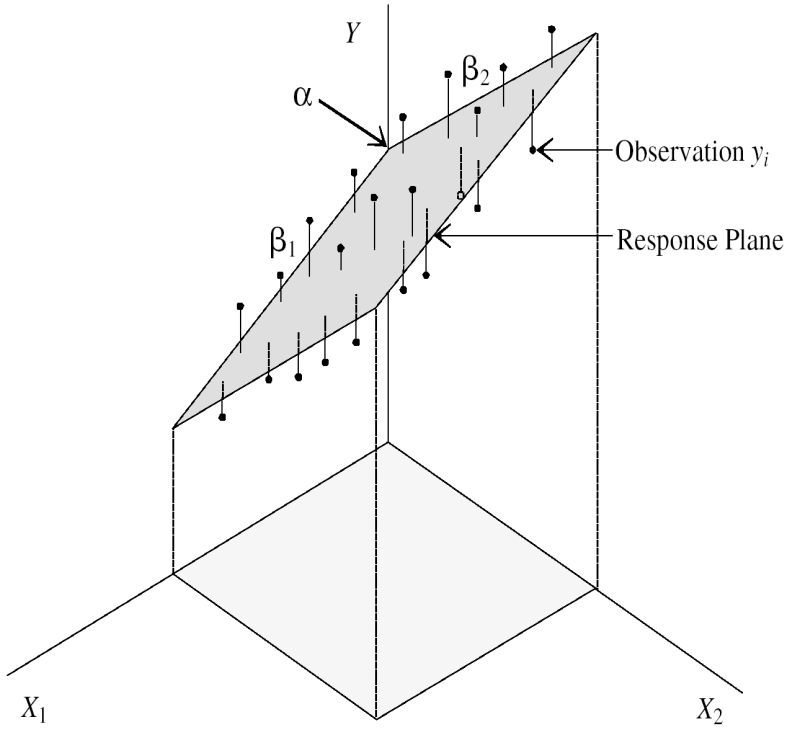


FIGURE 15.1 Three-dimensional response plane.

Multiple Linear Regression

- Multivariate data: ($x_{i,1}$, $x_{i,2}$, . . . , $x_{i,m}$, y_i) (n data points: i = 1, . . . , n)
- Model:

$$\hat{y}_i = a_1 x_{i,1} + a_2 x_{i,2} + \cdots + a_m x_{i,m}$$

- $x_{i,j}$ are the explanatory (or predictor) variables.

- y_i is the response variable.

- The total squared error is

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a_1 x_{i,1} - a_2 x_{i,2} - \cdots - a_m x_{i,m}))^2$$

Probabilistic Model

- y_i: the observed value of the random variable(r.v.) Y_i
- Y_i: depends on fixed predictor values $x_{i1}, x_{i2}, \ldots, x_{ik}$

$$\mathrm{Y_i} = \beta_0 + \beta_1 \mathrm{x}_{i1} + \beta_2 \mathrm{x_{i2}} + \cdots + \beta_\mathrm{k} \mathrm{x_{ik}} + \varepsilon_\mathrm{i} \quad ,\mathrm{i}=1,2,3,\ldots,\mathrm{n}$$

- $\beta_0, \beta_1, \ldots, \beta_k$ unknown model parameters
- **n** is the number of observations
- $\varepsilon_i \overset{\text{i.i.d}}{\sim} \mathrm{N}(0, \sigma^2)$

Fitting the Model

LS provides estimates of the unknown model parameters, $\beta_0, \beta_1, \ldots, \beta_k$ which minimizes Q

$$Q = \sum_{i=1}^{n} [yi - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik})]^2$$

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} [\mathrm{y_i} - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})] = 0$$

$$\frac{\partial Q}{\partial \beta_j} = -2 \sum_{i=1}^{n} [\mathrm{y_i} - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})] x_{\mathrm{ij}} = 0$$

(j=1,2,…,k)

Multiple Regression Model in Matrix Notation

- $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i,$
  $i = 1,2, \ldots, n$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad and \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- $Y_i's \rightarrow r.v.'s, \; y_i's \rightarrow observed\ values, \; \epsilon_i's \rightarrow$
  $random\ errors$

Multiple Regression Model in Matrix Notation

- $X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$
- $X \rightarrow predictor\ variables$
- The first column of X

  $\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ denotes the constant term β_0

- Finally let

- $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$ and $\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$

Multiple Regression Model in Matrix Notation

- Formula

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

becomes

$$Y = X\beta + \epsilon$$

- Simultaneously, the linear equation

$$\beta_0 n + \beta_1 \sum_{i=1}^{n} x_{i1} + \cdots + \beta_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

are changed to

$$X'X\beta = X'y$$

Solve this equation respect to β and we get

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

(if the inverse of the matrix $X'X$ exists.)

Tire tread wear vs. mileage

- The table gives the measurements on the groove of one tire after every 4000 miles.
- Our Goal: to build a model to find the relation between the mileage and groove depth of the tire.

Mileage (in 1000 miles)	Groove Depth (in mils) (0.001 inch)
0	394.33
4	329.50
8	291.00
12	255.17
16	229.33
20	204.83
24	179.00
28	163.83
32	150.33

Data example:

Input mile depth

Sqmile=mile*mile

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

Tire Wear Data: Quadratic Fit

- We will do tire wear example again in this part using the matrix approach.
- For the quadratic model to be fitted

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 4 & 16 \\ 1 & 8 & 64 \\ 1 & 12 & 144 \\ 1 & 16 & 256 \\ 1 & 20 & 400 \\ 1 & 24 & 576 \\ 1 & 28 & 784 \\ 1 & 32 & 1024 \end{bmatrix} \quad and \quad y = \begin{bmatrix} 394.33 \\ 329.50 \\ 291.00 \\ 255.17 \\ 229.33 \\ 204.83 \\ 179.00 \\ 163.83 \\ 150.33 \end{bmatrix}$$

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

$$X'X = \begin{bmatrix} 9 & 144 & 3264 \\ 144 & 3264 & 82{,}944 \\ 3264 & 82{,}944 & 2{,}245{,}632 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 0.6606 & -0.0773 & 0.0019 \\ -0.0773 & 0.0140 & -0.0004 \\ 0.0019 & -0.0004 & 0.0000 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X'X)^{-1}X'y =$$

$$\begin{bmatrix} 0.6606 & -0.0773 & 0.0019 \\ -0.0773 & 0.0140 & -0.0004 \\ 0.0019 & -0.0004 & 0.0000 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 4 & 8 & 12 & 16 & 20 & 24 & 28 & 32 \\ 0 & 16 & 64 & 144 & 256 & 400 & 576 & 784 & 1024 \end{bmatrix} \begin{bmatrix} 394.33 \\ 329.50 \\ 291.00 \\ 255.17 \\ 229.33 \\ 204.83 \\ 179.00 \\ 163.83 \\ 150.33 \end{bmatrix}$$

$$= \begin{bmatrix} 386.265 \\ -12.722 \\ 0.172 \end{bmatrix}$$

Therefore, the LS quadratic model is

$$\hat{y} = 386.265 - 12.772x + 0.172x^2.$$

This model has the same result as obtained in Example

Nile University