EE4211: Data Science for the Internet of Things

Regression

Biplab Sikdar

Agenda

- Examples
- Fitting to bivariate data
- Fitting polynomials
- Fitting to multivariate data
- Errors and overfitting

Regression

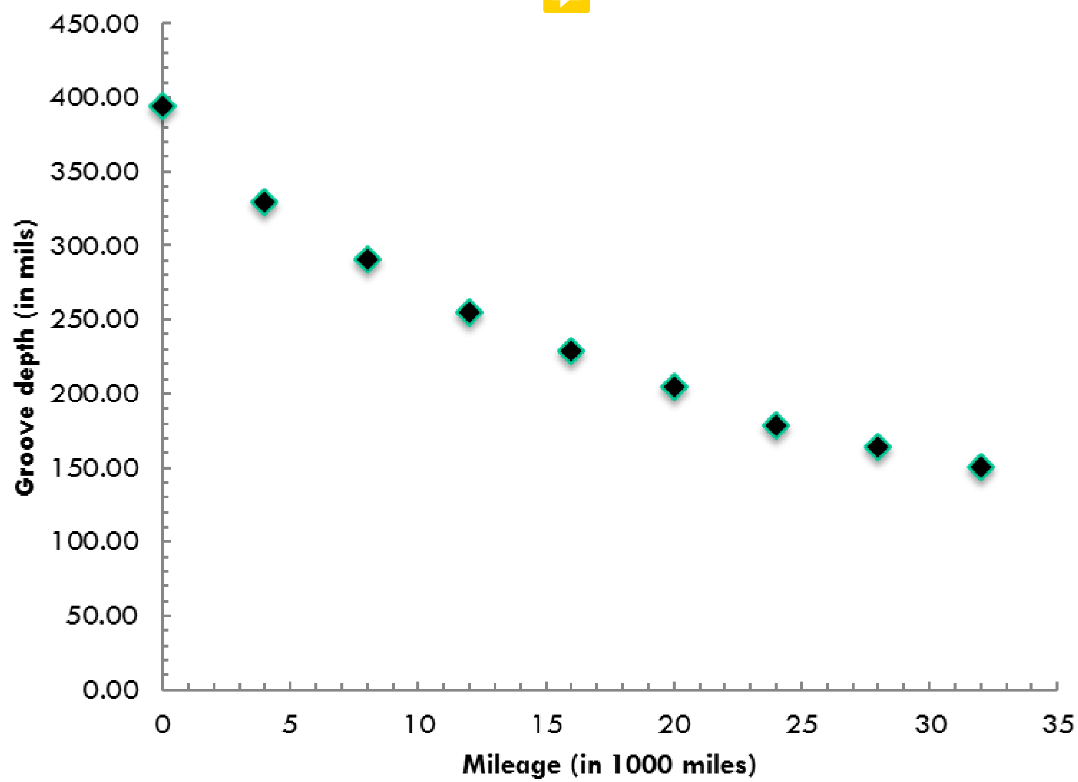
- Regression analysis: describes the relationship between two (or more) variables
- Examples:
  - Income and educational level
  - Demand for electricity and the weather
  - Home sales and interest rates
- Our focus:
  - Gain some understanding of the mechanics (e.g. regression line, regression error)
  - Learn how to setup a regression analysis.
  - Learn how to interpret and use the results.

Example

Example

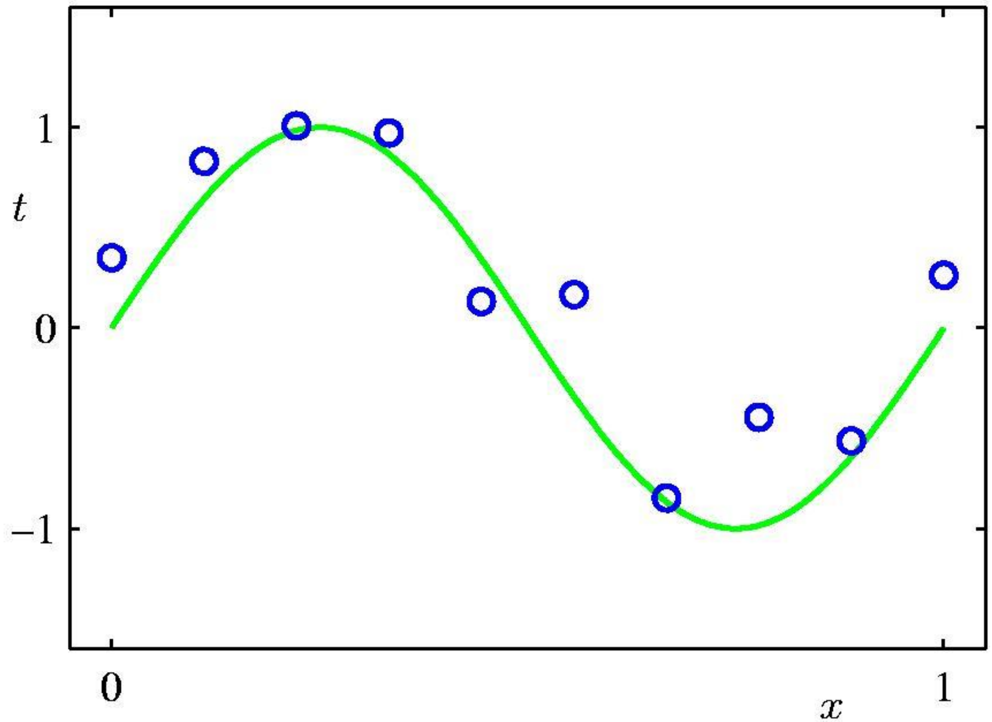
Example

Milage(in 1000 miles)	Groove Depth (in mils)
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



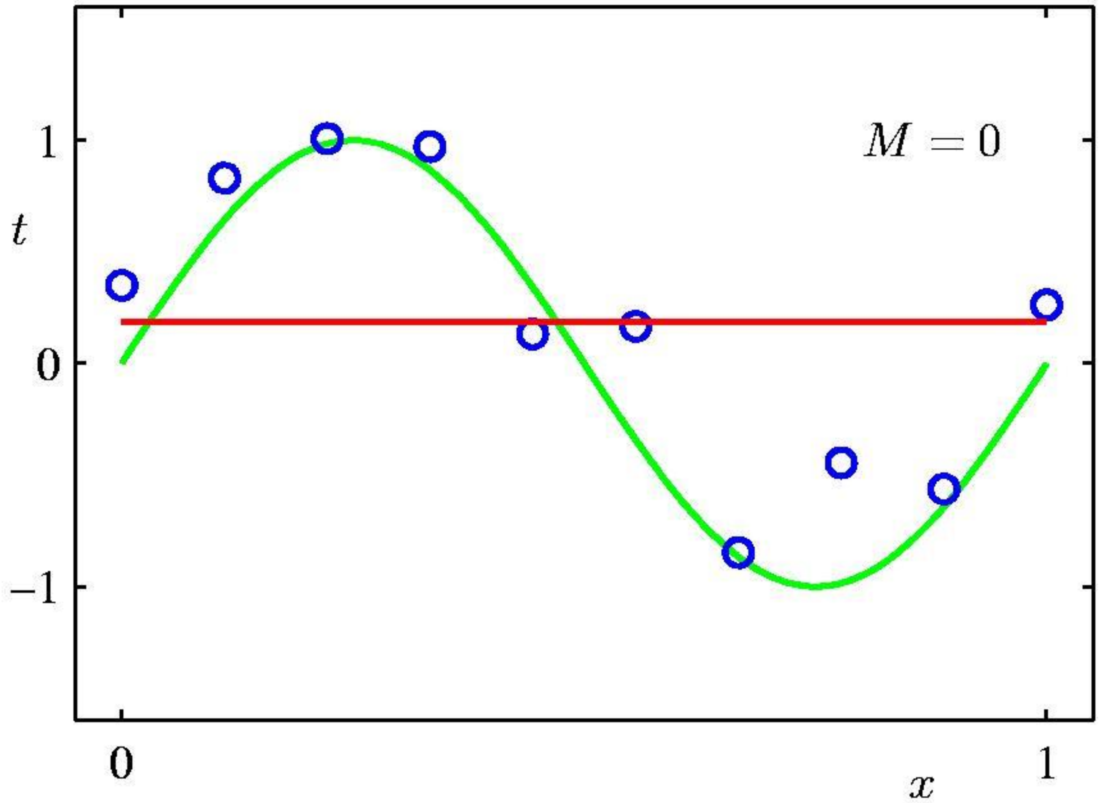
Tire tread wear vs. mileage. From: *Statistics and Data Analysis; Tamhane and Dunlop; Prentice Hall.*

Polynomial Curve Fitting

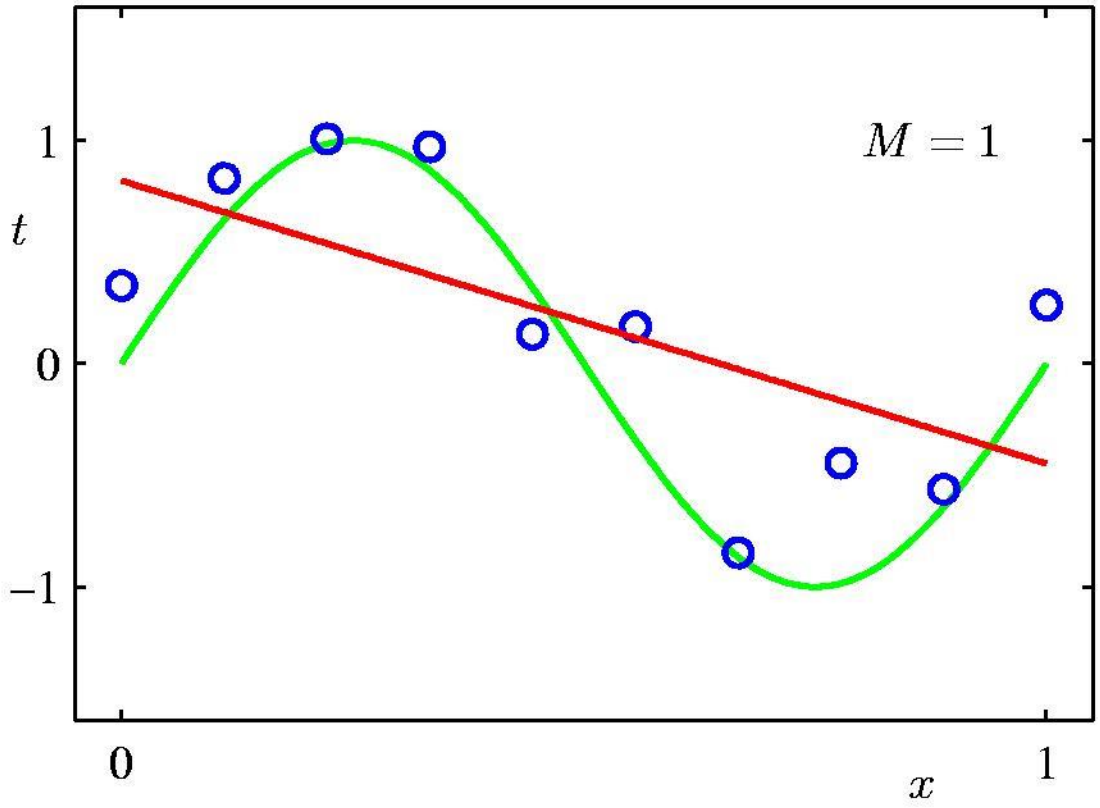


$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

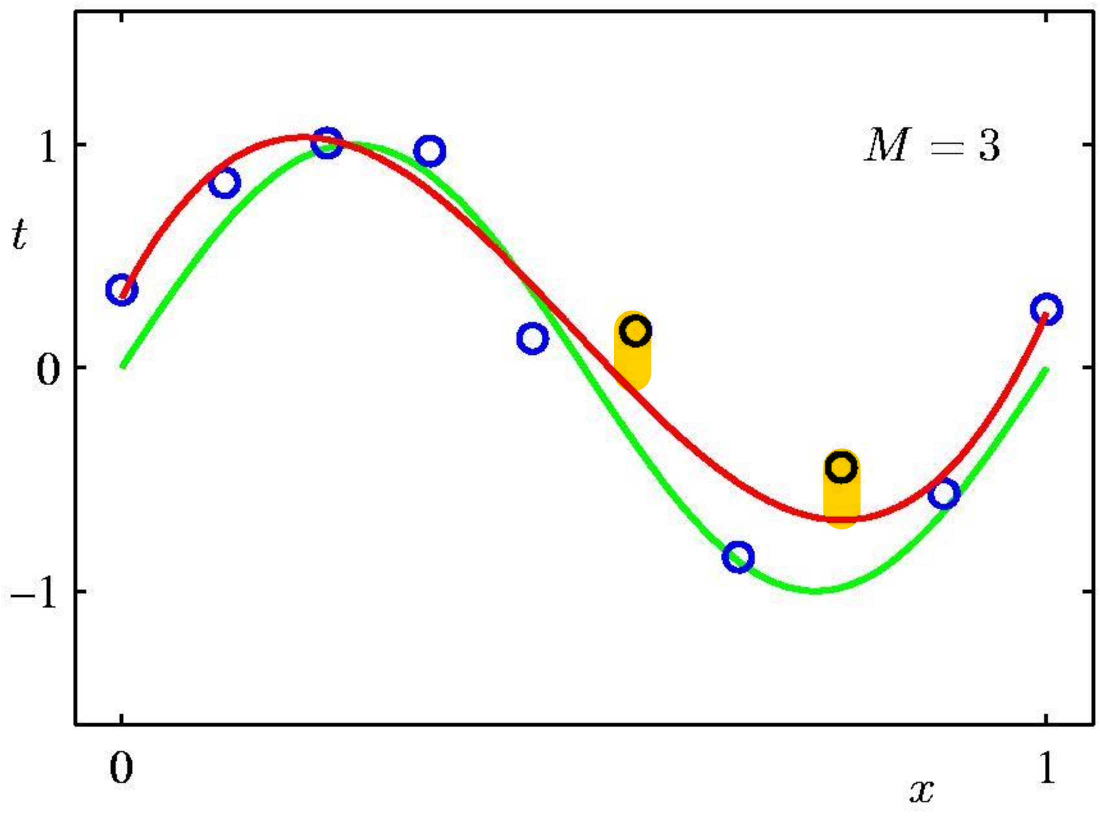
0^{th} Order Polynomial



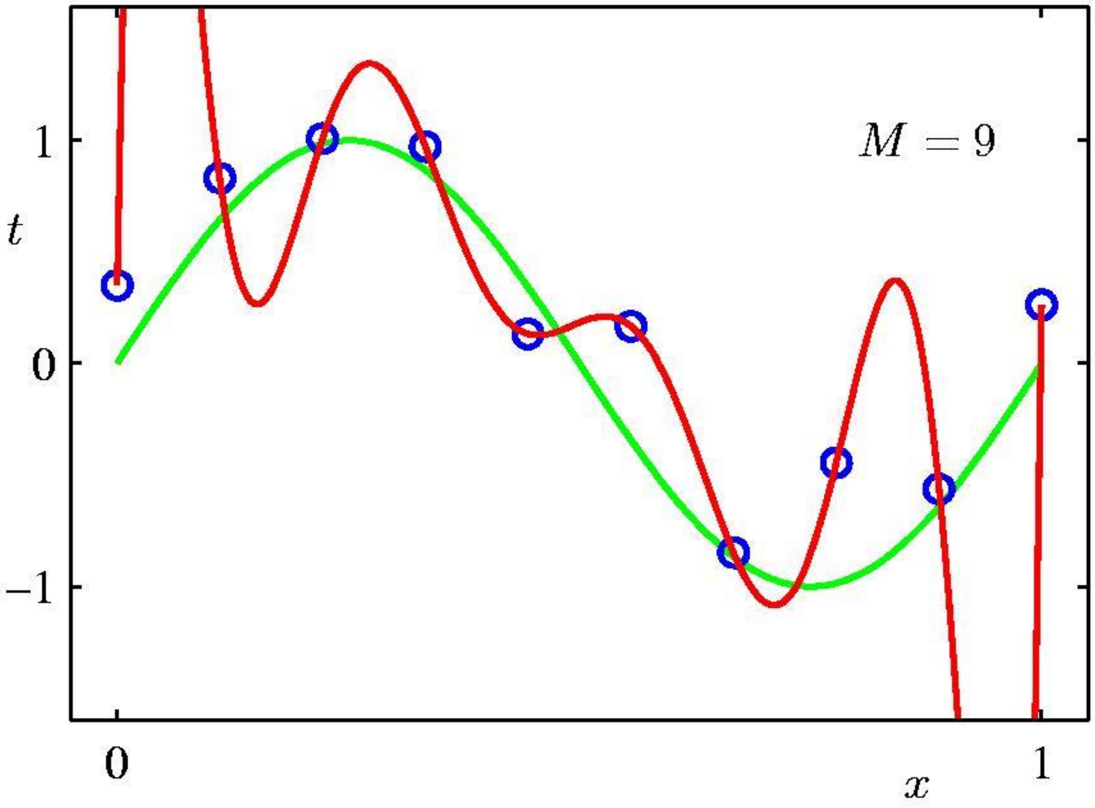
1st Order Polynomial



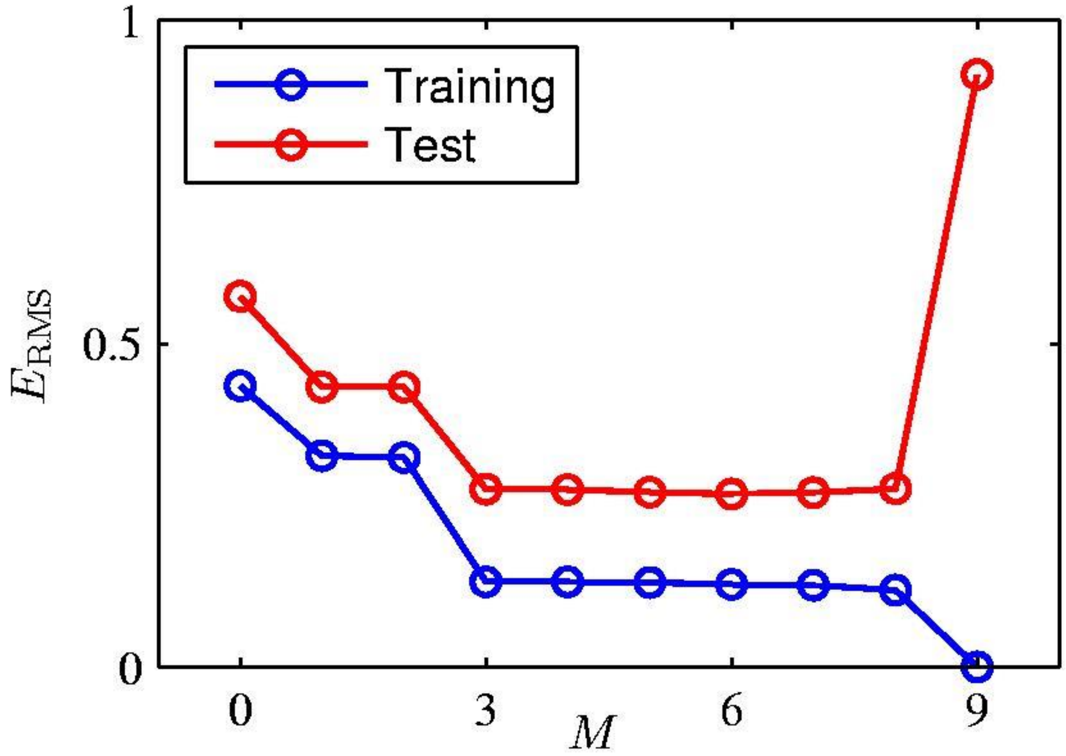
3^{rd} Order Polynomial



9th Order Polynomial



Over-fitting



Root-Mean-Square (RMS) Error

Linear Regression

- Linear Regression:
  - Simple linear regression: $\{y; x\}$
  - Multiple linear regression: $\{y; x_1, x_2, \cdots, x_n\}$
  - Multivariate linear regression: $\{y_1, y_2, \cdots, y_m; x_1, x_2, \cdots, x_n\}$

Linear Regression

- Response/outcome/dependent variable: y
- Predictor/explanatory/independent variable: x
- Example 1: Estimate electricity demand for home cooling (y) from the average daily temperature (x)
- Example 2: Relationship between the head size and body size of a newborn
- Regression analysis: statistical methodology to estimate the relationship between x and y
- Correlation analysis: statistical methodology used to asses the strength of relationship between x and y

Linear Regression

- One response variable and one explanatory variable
- We denote the explanatory variable as X and response variable as Y
- n pairs of observations $\{y_i; x_i\}$, $i = 1, \cdots, n$
- y_i is the observed values of the random variable Y_i and is related to x_i by:

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

- ϵ_i: random error with $E[\epsilon_i] = 0$ and $VAR[\epsilon_i] = \sigma^2$
- The "true regression line" models the true but unknown mean of Y_i

$$E[Y_i] = \hat{y}_i = \beta_0 + \beta_1 x_i$$

NUS National University of Singapore

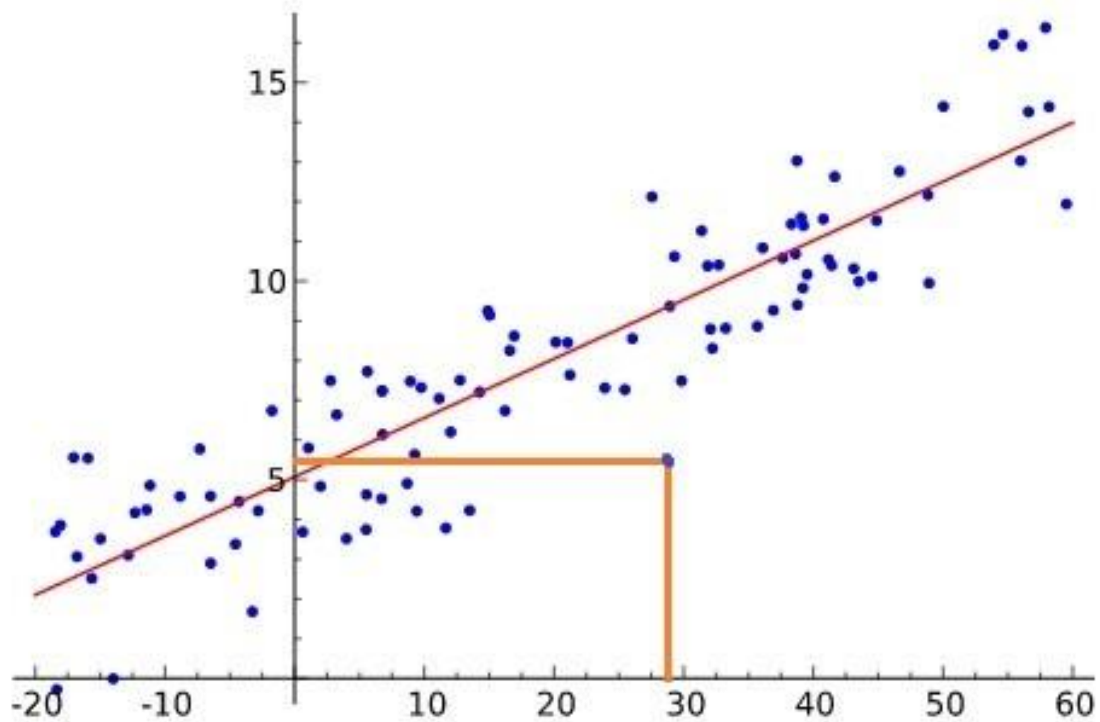
Linear Regression

- The error ϵ_i:
  - Independent and identically distributed
  - Variety of cause:
    - Measurement errors
    - Other variables affecting Y_i not included in the model
  - Assumption $E[\epsilon_i] = 0$: implies there is no systematic bias
  - Usual model for ϵ_i: $\epsilon_i \sim N(0, \sigma^2)$

Linear Regression

- Step 1: Plot the data and inspect for linearity

	X	Y
1	37.70	9.82
2	16.31	5.00
3	28.37	9.27
4	-12.13	2.98
⋮	⋮	⋮
98	9.06	7.34
99	28.54	10.37
100	-17.19	2.33



Linear Regression

- In simple linear regression, the data is represented as:

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

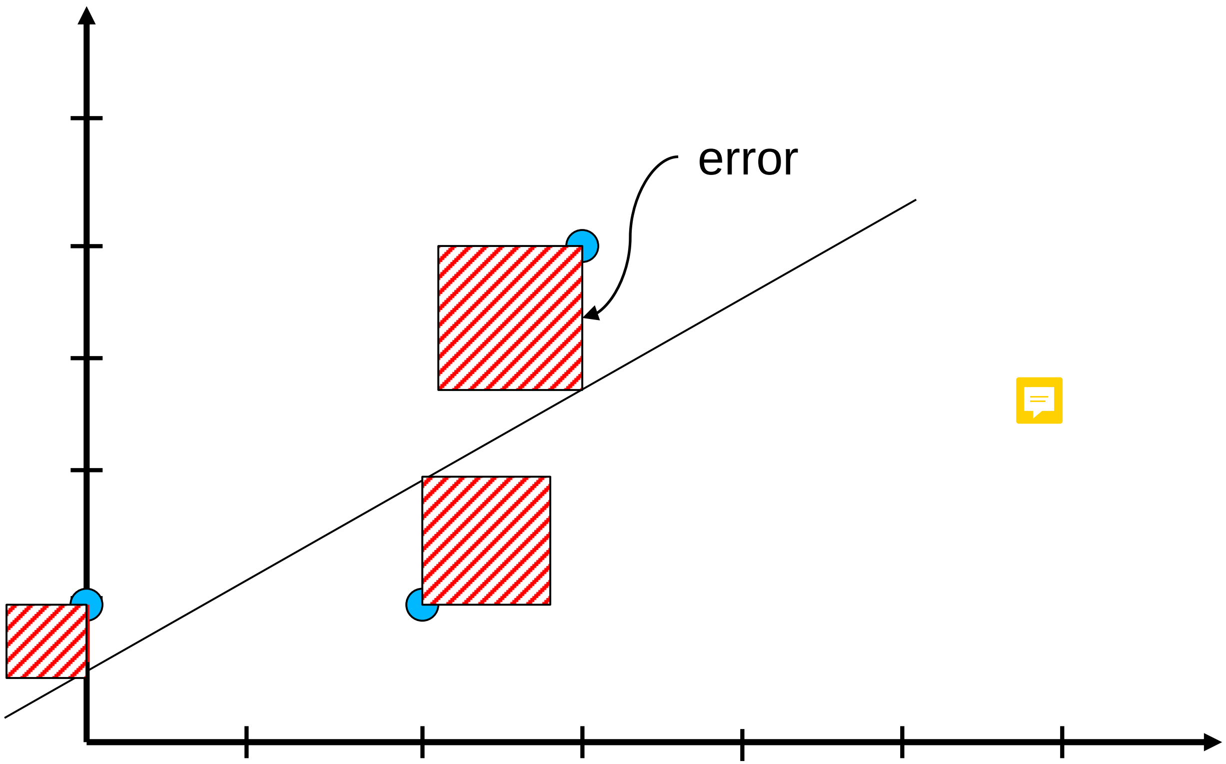
  where $\epsilon_i \sim N(0, \sigma^2)$

- The fitted model:

$$\hat{y} = \beta_0 + \beta_1 x$$

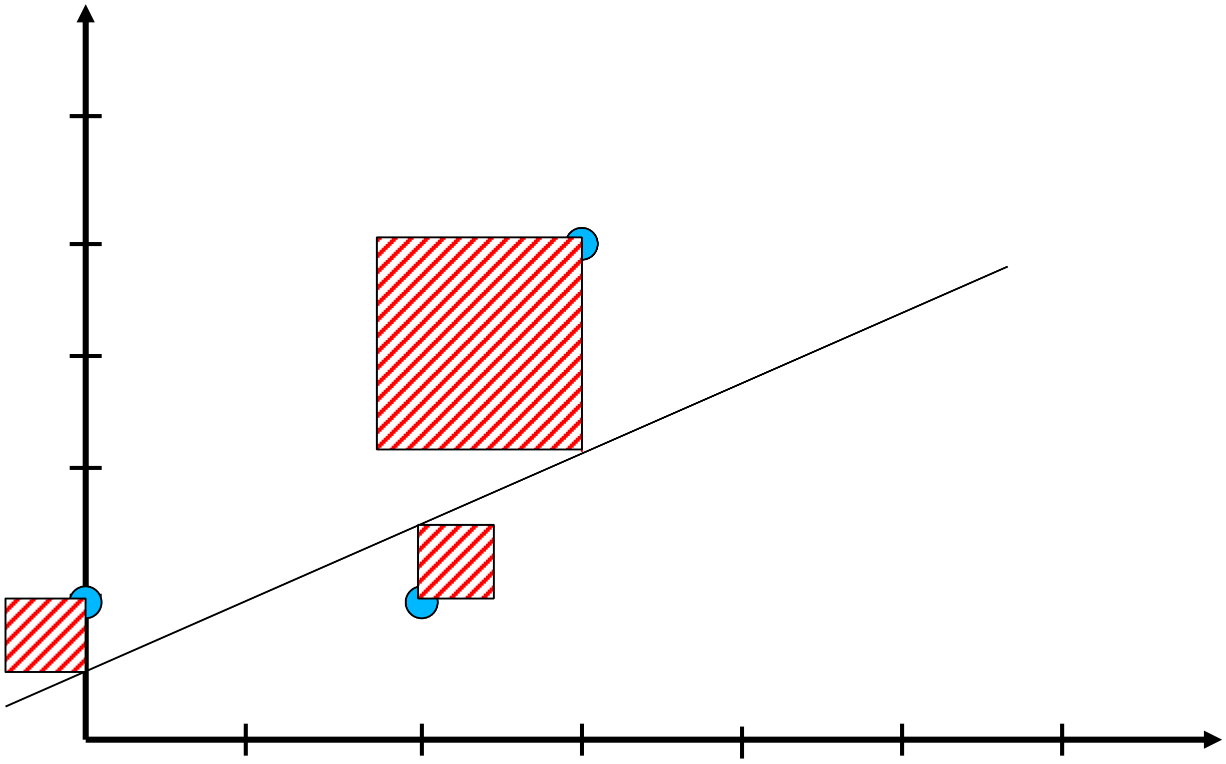
  where
  - β_0: intercept
  - β_1: slope of regression line

Least Squares Fitting



(0,1), (2,1), (3,4)

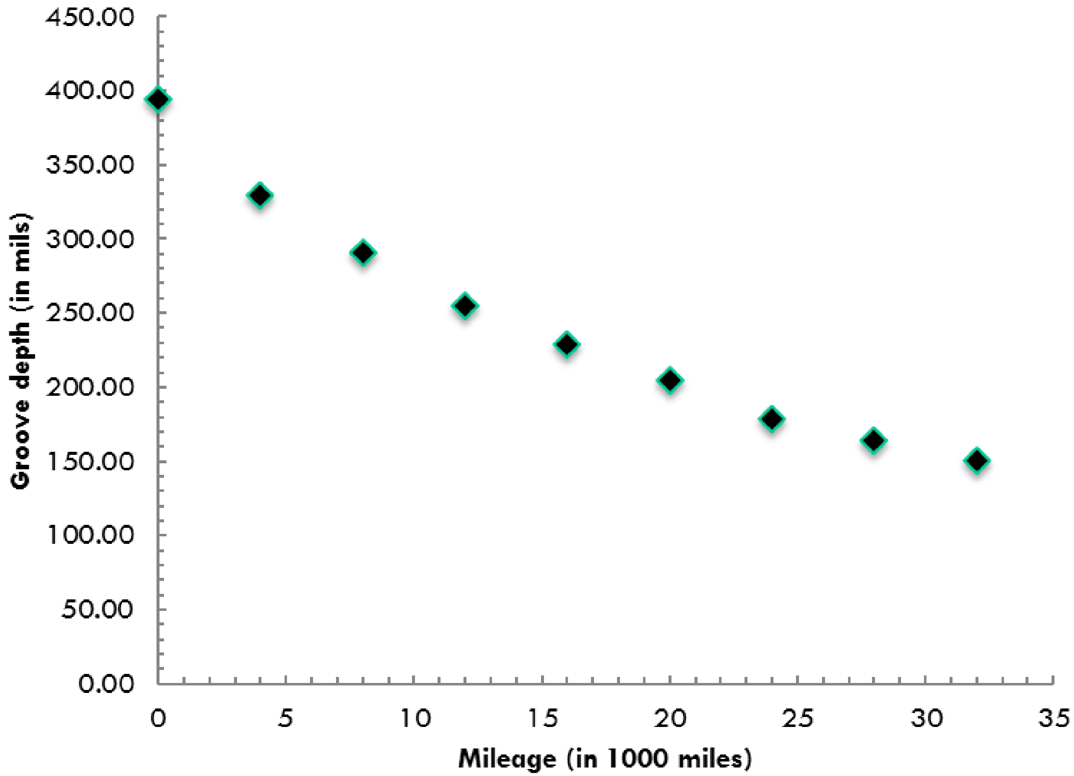
Least Squares Fitting



(0,1), (2,1), (3,4)

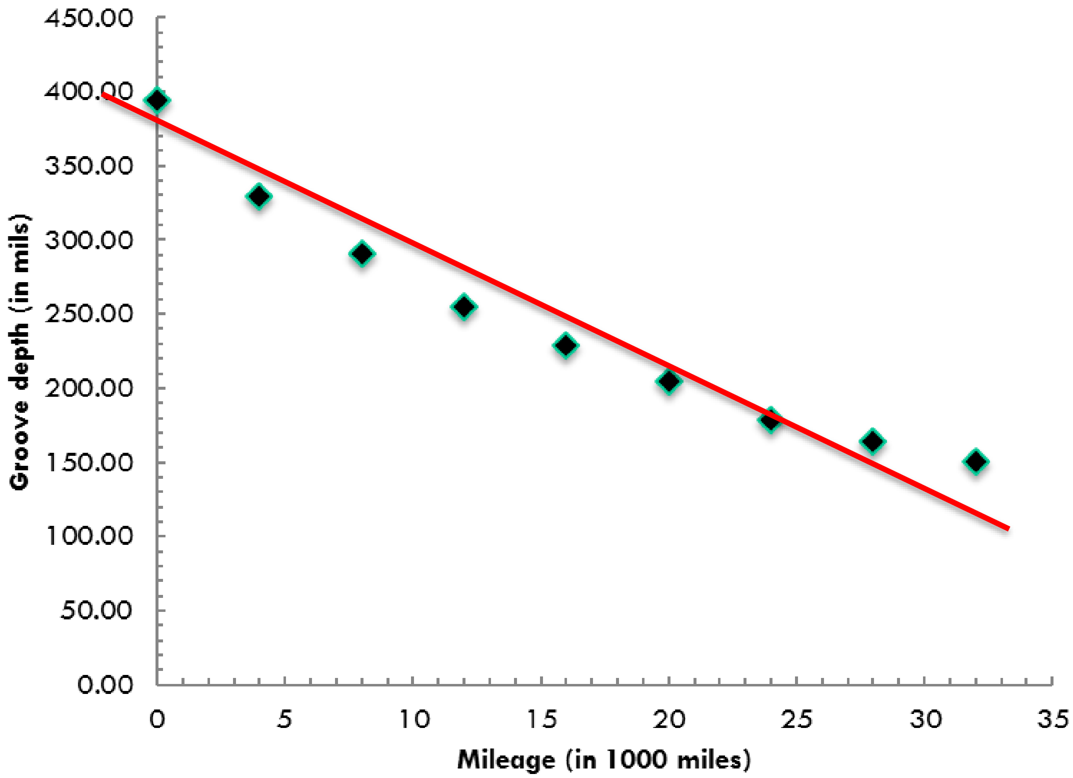
Linear Regression

Milage(in 1000 miles)	Groove Depth (in mils)
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

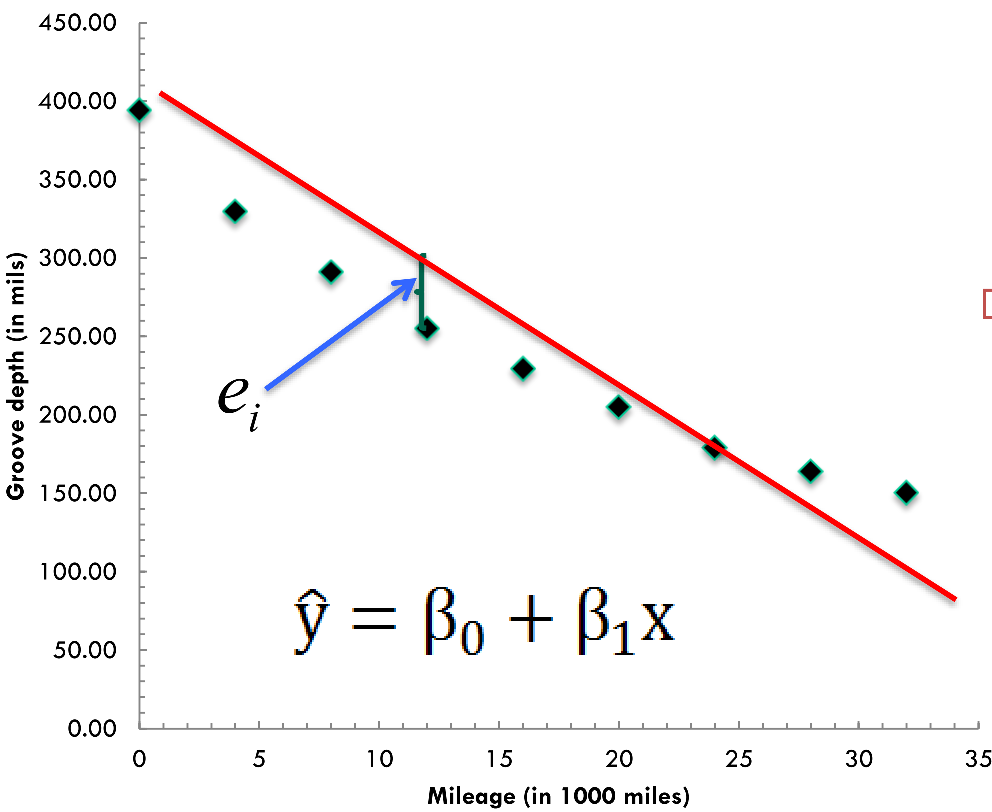


Tire tread wear vs. mileage. From: *Statistics and Data Analysis; Tamhane and Dunlop; Prentice Hall.*

Linear Regression



Least Squares Fitting



- e_i: difference between real data and fitted line

$$e_i = y_i - \hat{y}_i$$
$$= y_i - (\beta_0 + \beta_1 x_i)$$

- Goal: minimize the sum of the square of the error

$$Q = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Least Squares Fitting

- Obtain the values of β_0 and β_1 that minimizes the squared error

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)] = 0$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}x_i[y_i - (\beta_0 + \beta_1 x_i)] = 0$$

$$n\beta_0 + \beta_1\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}y_i$$

$$\Rightarrow$$

$$\beta_0\sum_{i=1}^{n}x_i + \beta_1\sum_{i=1}^{n}x_i^2 = \beta_1\sum_{i=1}^{n}x_i\,y_i$$

NUS National University of Singapore

Least Squares Fitting

$$\hat{\beta}_0 = \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Least Squares Fitting

- To simplify:

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

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=0}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=0}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

$$\Rightarrow \quad \hat{\beta}_0 = \bar{y} + \beta_1 \bar{x}, \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

NUS National University of Singapore

Example

$n = 9$

$\sum x_i = 144, \sum x_i^2 = 3264$

$\sum y_i = 2197.32, \sum y_i^2 = 589887.08$

$\sum x_i y_i = 28167.72$

$\bar{x} = 16, \bar{y} = 244.15$

$S_{xy} = -6989.40$

$S_{xx} = 960$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = -7.281$

$\hat{\beta}_0 = \bar{y} + \beta_1 \bar{x} = 360.64$

$\hat{y} = 360.64 - 7.281x$

Milage(in 1000 miles)	Groove Depth (in mils)
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

Checking the Goodness of Fit

- Residuals: $e_i = y_i - \hat{y}_i$
- Least squares fitting minimized "error sum of squares": $Q = \sum(y_i - \hat{y}_i)^2$
- Is this good enough?
  - Compare with benchmarks
  - One possible benchmark:

$$Y_i = \beta_0 + \epsilon_i$$

  - Corresponding $Q_{min} = \sum(y_i - \bar{y})^2 = S_{yy}$
  - Referred to as SST: total sum of squares

Checking the Goodness of Fit

- SST: total sum of squares
- SSR: regression sum of squares
- SSE: error sum of squares

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

$$= \sum \left((\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \right)^2$$

$$= \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{SSE} + \underbrace{2 \sum (\hat{y}_i - \bar{y})(y_i - \hat{y}_i)}_{=0}$$

Checking the Goodness of Fit

- SSR: regression sum of squares
  - Represents the variation in y that is accounted for by the regression on x
- SST: measures the variability of y_is around $\bar{y}$
- Coefficient of determination:

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{SSR}{SSR + SSE}$$

represents the proportion of variation in y that is accounted for by the regression on x

Example

$$SST = S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 53418.73$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 2531.53$$

$$SSR = SST - SSE = 50887.20$$

$$r^2 = \frac{50887.20}{53418.73} = 0.953$$

- 95.3% of the variation in the tread wear is accounted for by linear regression on mileage (strongly linear relationship)

Prediction of Future Observations

- Common use of a regression model: predict the value of the response variable Y when the predictor variable x is set at a specific value x^*

$$\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

- A $100(1-\alpha)\%$ confidence interval of the prediction

$$\left[\hat{Y}^* \pm t_{n-2,\alpha/2}\sqrt{MSE}\sqrt{1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}\right]$$

where:

- $t_{n-2,\alpha/2}$: t-distribution with $n-2$ degrees of freedom
- $MSE = \frac{SSE}{n-2}$

Example

- Compute a 95% confidence interval for the groove depth of a tire with a mileage of 25000
  - $MSE = \frac{SSE}{n-2} = \frac{2531.53}{9-2} = 361.65$
  - $t_{n-2,\alpha/2} = 2.365$
  - $\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^* = 178.62$
  - $\left[\hat{Y}^* \pm t_{n-2,\alpha/2}\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}\right]$
  - $\left[178.62 \pm 2.365\sqrt{361.65}\sqrt{1 + \frac{1}{9} + \frac{(25-16)^2}{960}}\right]$

$= [129.44, 227.80]$

NUS National University of Singapore

T-distribution (Student-T Distribution)

Statistician William Sealy Gosset, known as "Student"

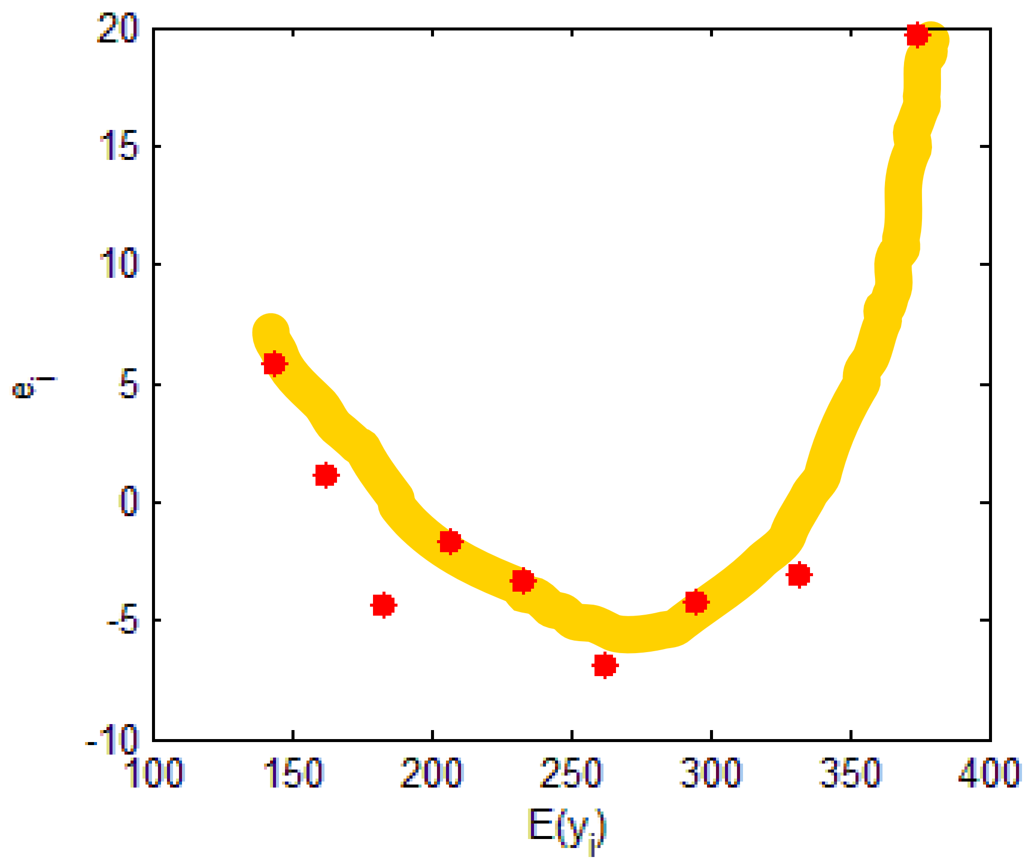
Regression Diagnostics

- **Residual plots:** error $e_i = y_i - \hat{y}_i$ versus $\hat{y}_i$ plots

i	x_i	y_i	$\hat{y}_i$	e_i
1	0	394.33	360.64	33.69
2	4	329.50	331.51	-2.01
3	8	291.00	302.39	-11.39
4	12	255.17	273.27	-18.10
5	16	229.33	244.15	-14.82
6	20	204.83	215.02	-10.19
7	24	179.00	185.90	-6.90
8	28	163.83	156.78	7.05
9	32	150.33	127.66	22.67

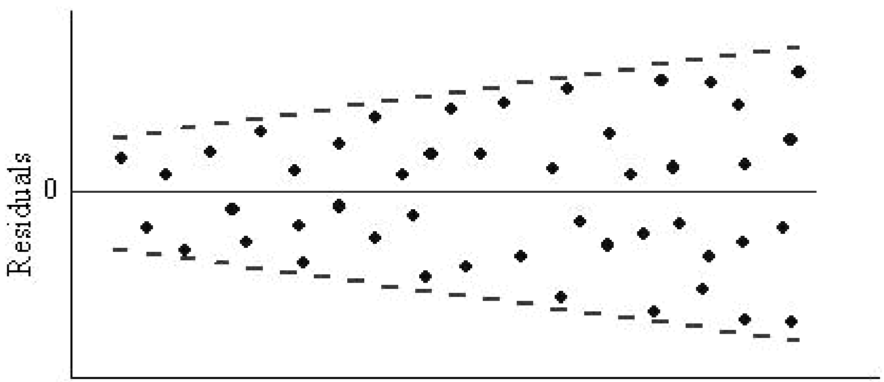
Regression Diagnostics

- **Residual plots:** error $e_i = y_i - \hat{y}_i$ versus $\hat{y}_i$ plots

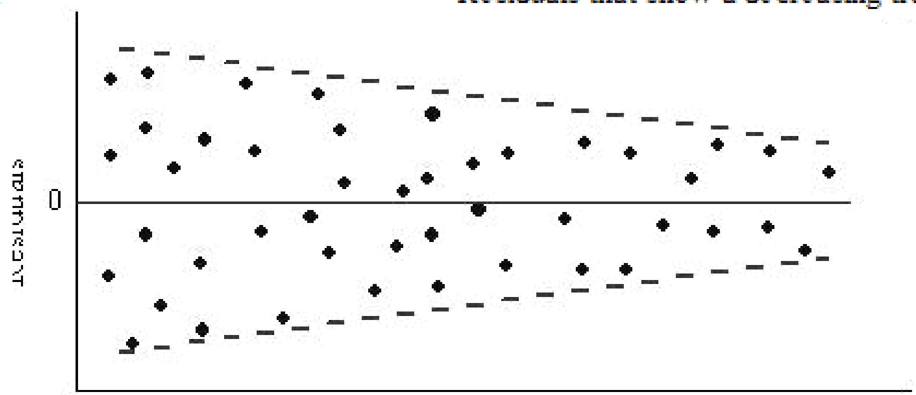


Regression Diagnostics

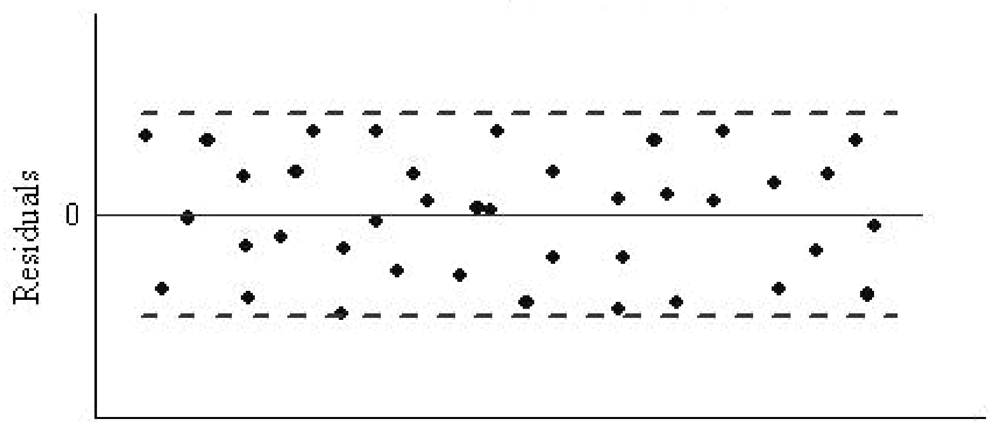
Residuals that show an increasing trend



Residuals that show a decreasing trend

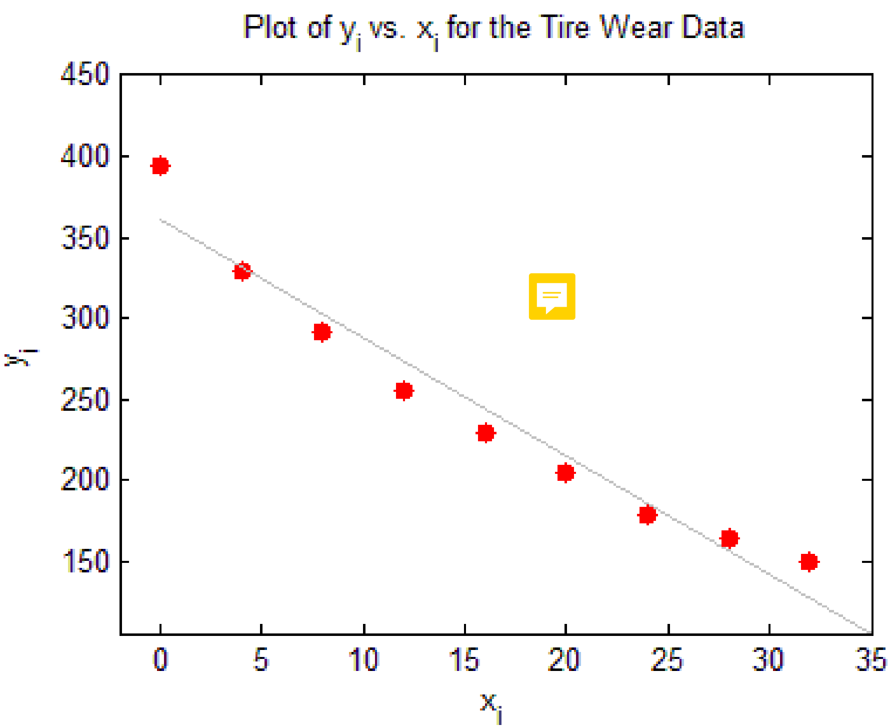


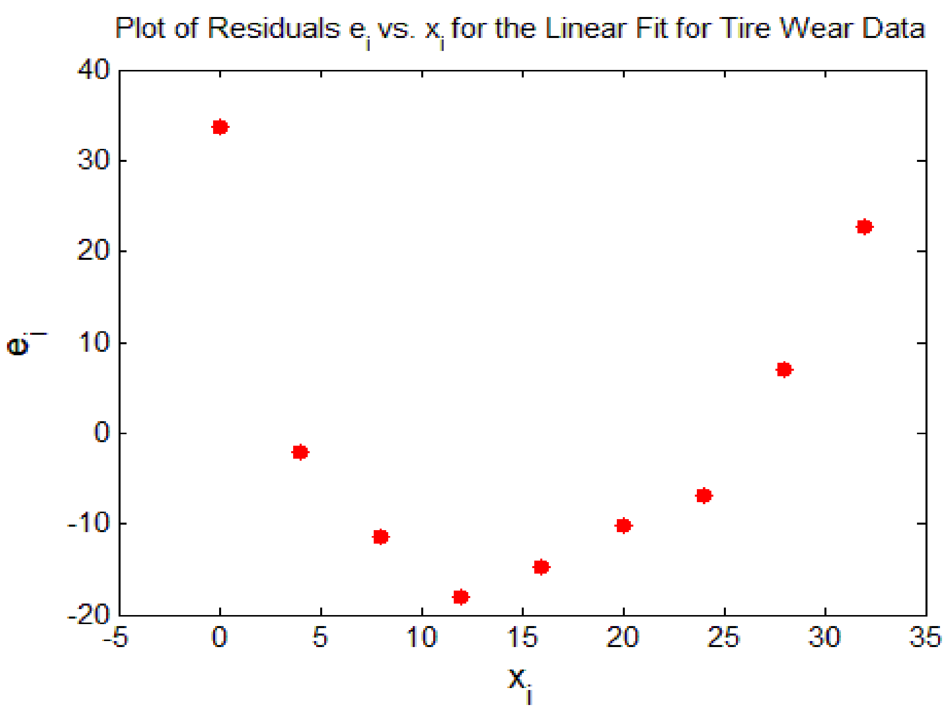
Constant variance



Regression Diagnostics

- Check for linearity: error $e_i = y_i - \hat{y}_i$ versus $\mathrm{x_i}$ plots

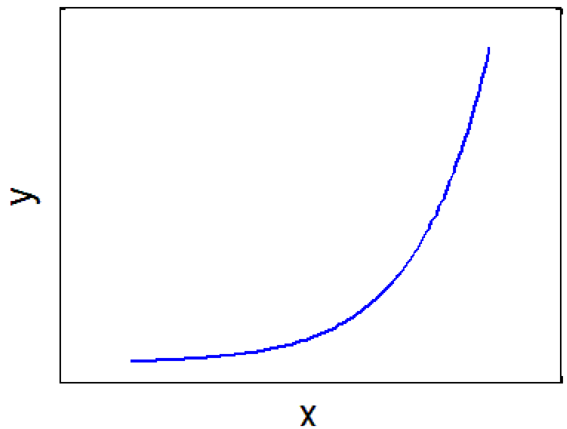


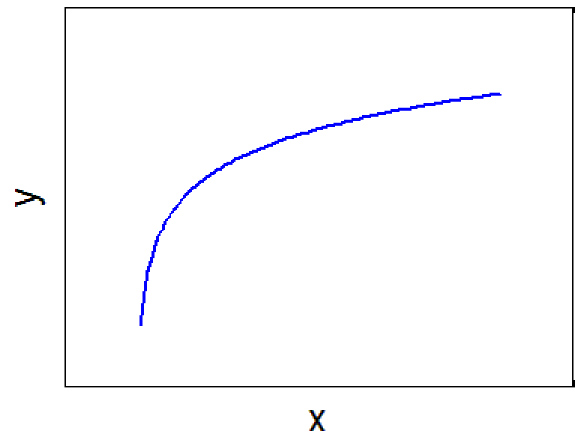


Regression Diagnostics

- Data transformations: Linearizing transformations

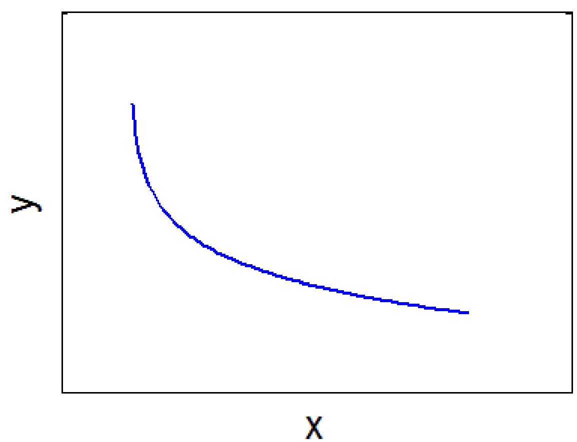
x	y
x^2	y
x^3	y
x	logy
x	1/y

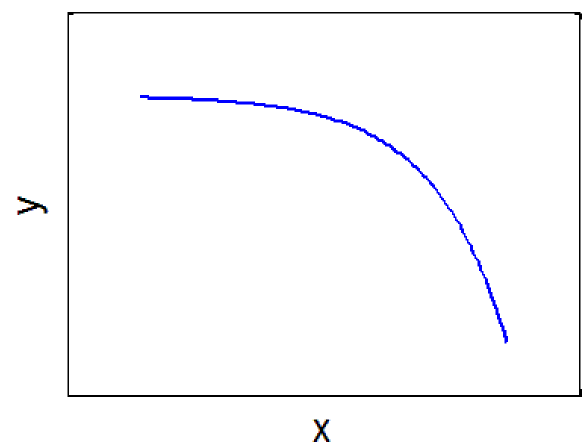




x	y
logx	y
-1/x	y^2
x	y^3
x	y

x	y
logx	y
-1/x	y
x	logy
x	-1/y





x	y
x^2	y
x^3	y
x	y^2
x	y^3

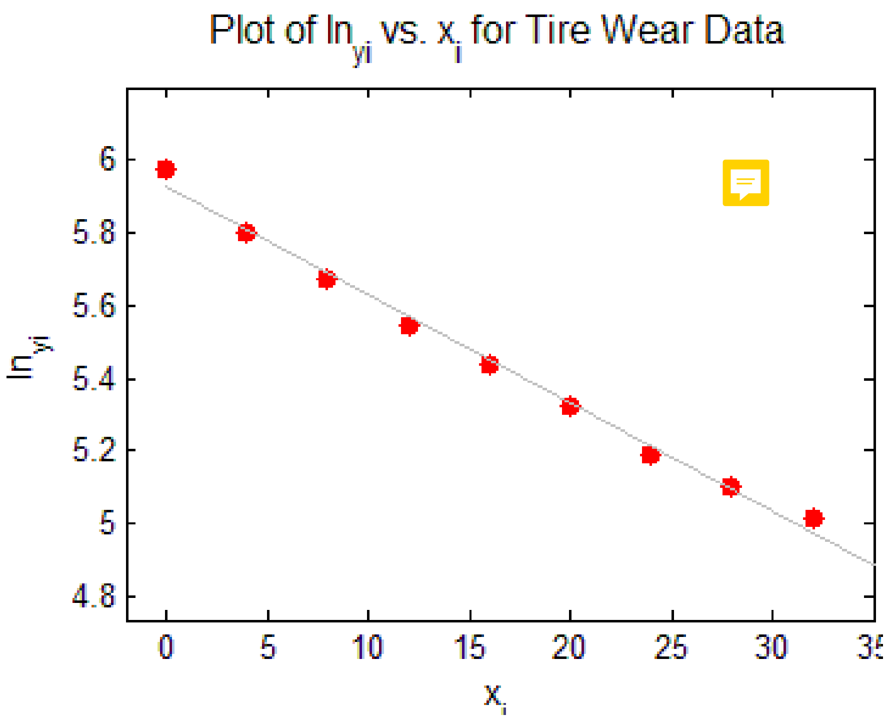
Regression Diagnostics

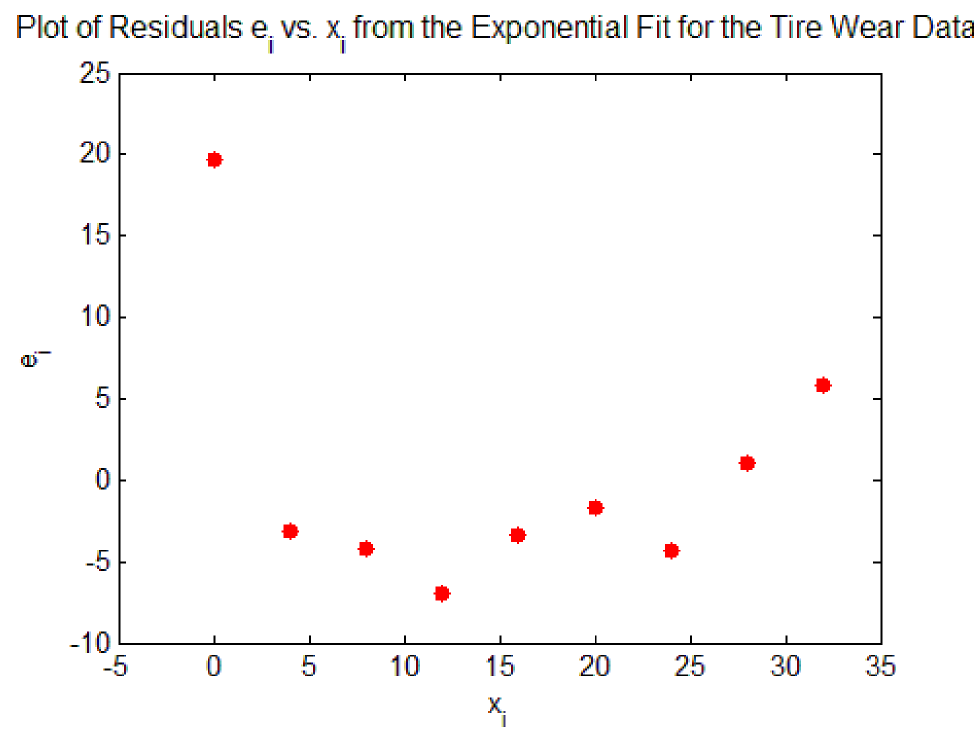
- **Check for linearity:** error $e_i = y_i - \hat{y}_i$ versus $\mathrm{x_i}$ plots

i	x_i	y_i	$\widehat{\ln(y_i)}$	$\hat{y}_i$	e_i
1	0	394.33	5.926	374.64	19.69
2	4	329.50	5.807	332.58	−3.08
3	8	291.00	5.688	295.24	−4.24
4	12	255.17	5.569	262.09	−6.92
5	16	229.33	5.450	232.67	−3.34
6	20	204.83	5.331	206.54	−1.71
7	24	179.00	5.211	183.36	−4.36
8	28	163.83	5.092	162.77	1.06
9	32	150.33	4.973	144.50	5.83

Regression Diagnostics

- Check for linearity: error $e_i = y_i - \hat{y}_i$ versus x_i plots





Multivariate Regression

- We have explored problems with one response variable and one explanatory variable
- Sometimes a straight line is not adequate and quadratic or cubic model is needed
- Sometimes there are more than one predictor variables and their simultaneous effect needs to be modeled
- n pairs of observations $\{y_i; x_{i1}, x_{i2}, \cdots, x_{ik}\}, i = 1, \cdots, n$
- Multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

- Linear in β and not necessarily x's: $x_1 = x, x_2 = x^2, x_k = x^k$

Multivariate Regression

- Least squares fit:

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

- Taking the partial derivatives and equating to zero:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum [y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})] = 0$$

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

for $j = 1, 2, \cdots, k$

Multivariate Regression

- After simplification (for $j = 1,2,\cdots,k$):

$$n\beta_0 + \beta_1 \sum x_{i1} + \cdots + \beta_k \sum x_{ik} = \sum y_i$$

$$n\beta_0 \sum x_{ij} + \beta_1 \sum x_{i1}x_{ij} + \cdots + \beta_k \sum x_{ik}x_{ij} = \sum y_i\, x_{ij}$$

- These have to be solved simultaneously for $\beta_1, \beta_2, \cdots, \beta_k$

Multivariate Regression

- Matrix form:

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

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \widehat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

Multivariate Regression

- Regression model:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

- Simultaneous linear equations whose solution gives the least square estimates:

$$\boldsymbol{X'X\beta} = \boldsymbol{X'y}$$

- Regression parameters:

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

Acknowledgements

- A number of the slides in this lecture are based on material from various sources:
  - Wei Zhu
  - Ajit Tamahane
  - Dorothy Dunlop