Simple Linear Regression

Instructor: Junghye Lee

Department of Industrial Engineering
junghyelee@unist.ac.kr

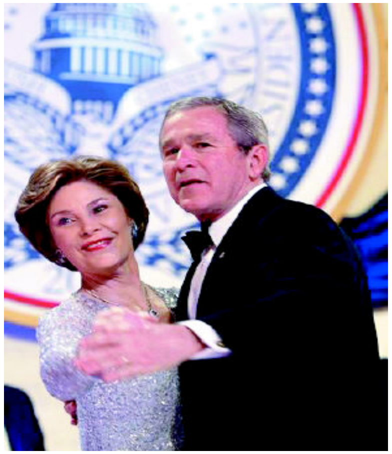
Contents

Introduction

- Example

David Beckham: 1.83m
Victoria Beckham: 1.68m

Brad Pitt: 1.83m
Angelina Jolie: 1.70m



George Bush :1.81m
Laura Bush: ?

- To predict height of the wife in a couple, based on the husband's height
  - **Response** (outcome or dependent) **variable** (Y): height of the wife
  - **Predictor** (explanatory or independent) **variable** (X): height of the husband

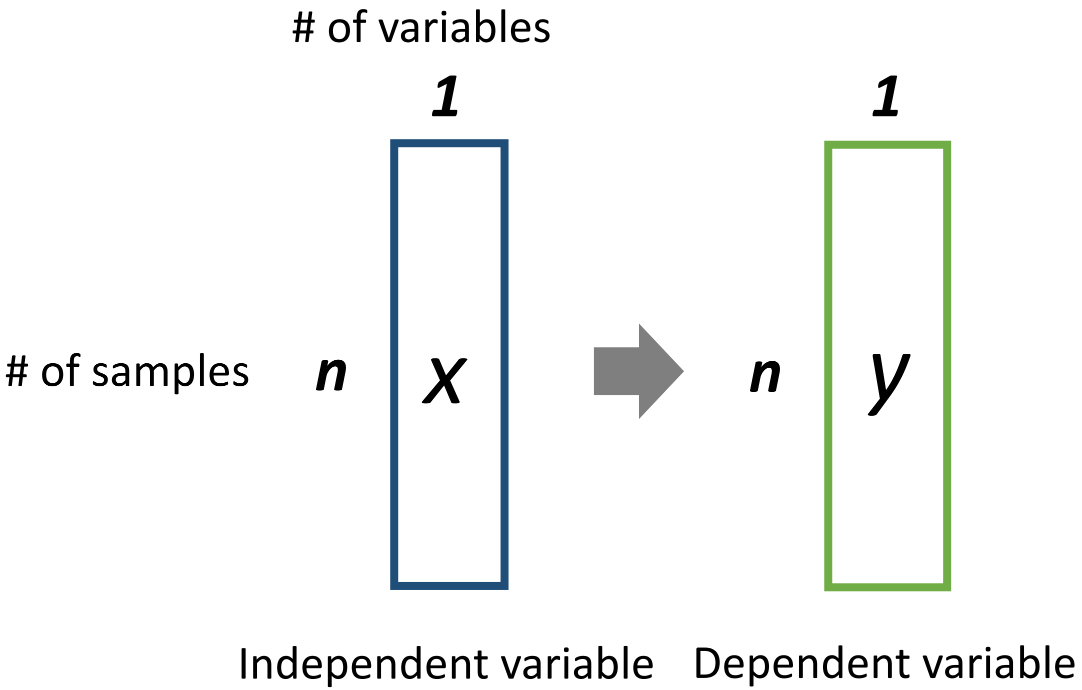
Regression Analysis

- **Regression analysis** is a statistical methodology to estimate the relationship of a response variable (i.e., dependent variable) to a set of predictor variables (i.e., independent, explanatory variables, factors).

- When there is just one predictor variable, we will use **simple linear regression**. When there are two or more predictor variables, we use **multiple linear regression**.

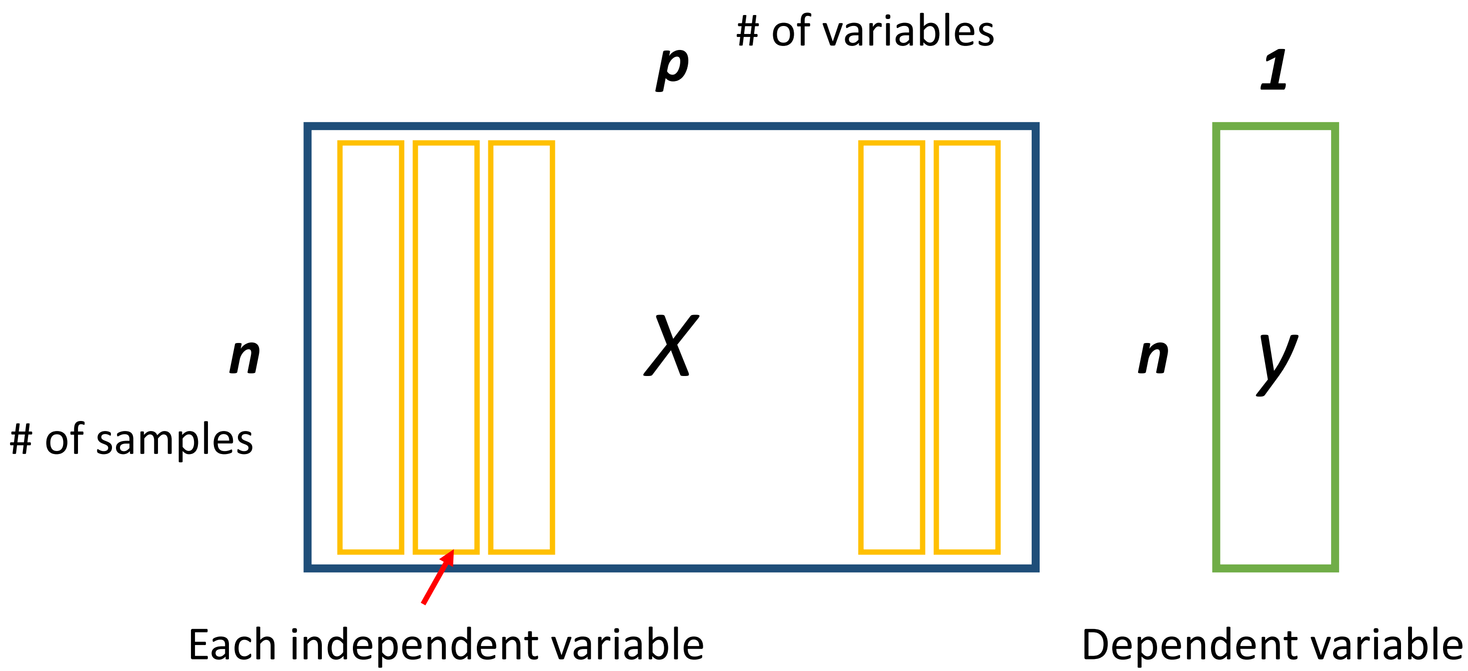
- When it is not clear which variable represents a response and which is a predictor, **correlation analysis** is used to study the strength of the relationship.

Simple Linear Regression



- Why simple? Because the number of ***"independent variable"*** is one.
- Dependent variable should be continuous according to the definition of regression, but independent variable can be any type.
- However, we assume ***x*** is continuous here.

Multiple Linear Regression



History

- The earliest form of linear regression was the method of least squares, which was published by *Legendre* in 1805, and by *Gauss* in 1809.

- The method was extended by *Francis Galton* in the 19th century to describe a biological phenomenon.

- This work was extended by *Karl Pearson* and *Udny Yule* to a more general statistical context around 20th century.

Probabilistic Model

- We denote the n observed values of the **predictor variable** X **as**

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

- We denote the corresponding n observed values of the **response variable** Y as

$$y_1, y_2, \dots, y_n$$

- In summary, we have paired dataset $D = \{x_i, y_i\}_{i=1}^{n}$.

Notations of the Simple Linear Regression

y_i: Observed value of the **random variable** Y_i depends on x_i

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad (i = 1,2, \dots, n)$$

(x_i is given so not a random variable.)

ϵ_i: random error with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$

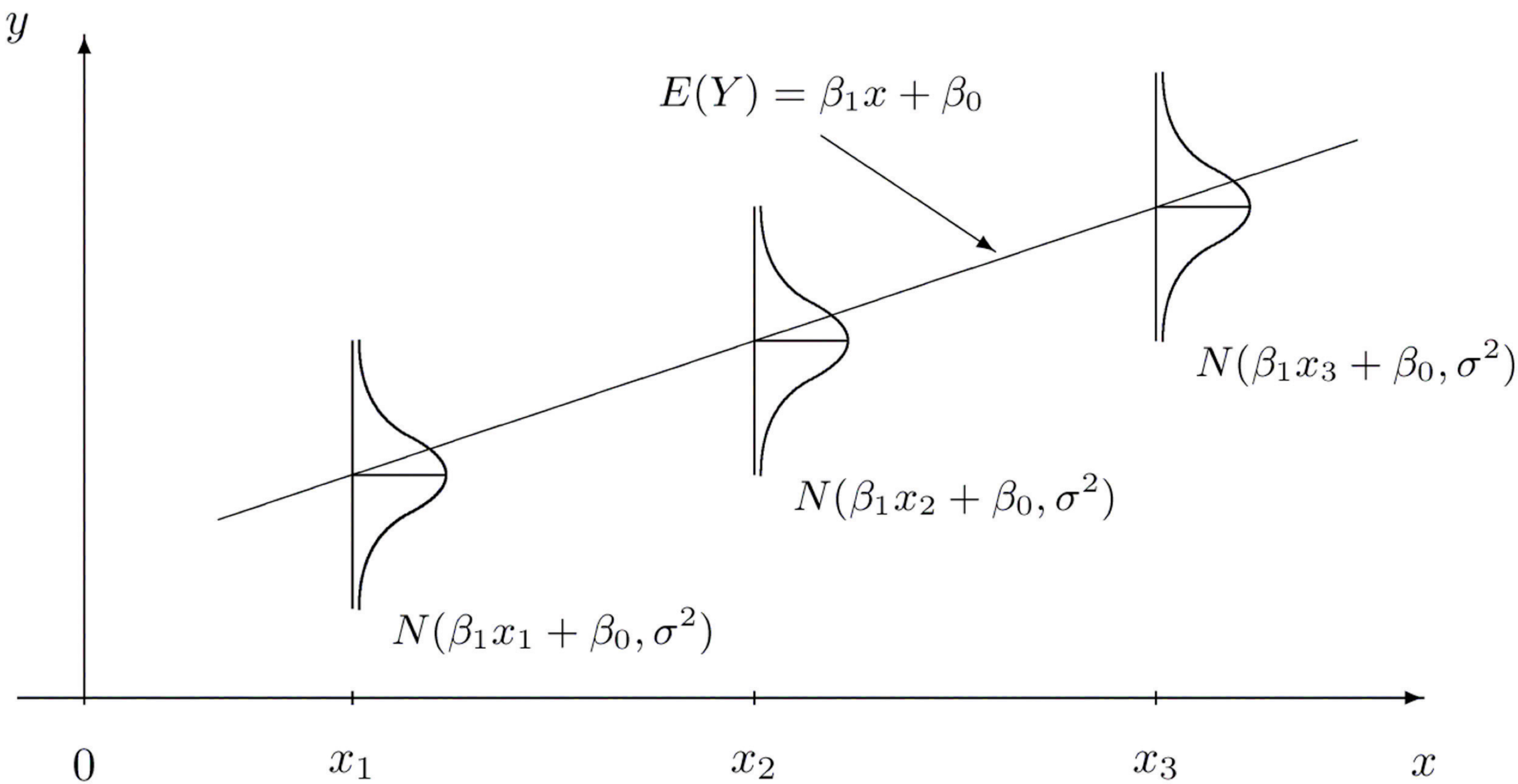
Unknown Mean of Y_i $\longrightarrow$ $E(Y_i) = \mu_i = \beta_0 + \beta_1 x_i$

True Regression Line

Unknown Slope

Unknown Intercept

Simple Linear Regression



4 Assumptions for Statistical Inference

For Y_i

- Linear function of the predictor variable
- Have a common variance σ^2, same for all values of x.

For ϵ_i

- Normally distributed
- Independent

Comments

- Linear not in x, but in the parameters β_0 and β_1
- Predictor variable is not set as predetermined fixed values, is random along with Y.
- The model can be considered as a conditional model.

- Example: Height and Weight of the children

  Height (X) – given

  Weight (Y) – predict

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

Conditional expectation of Y given $X = x$

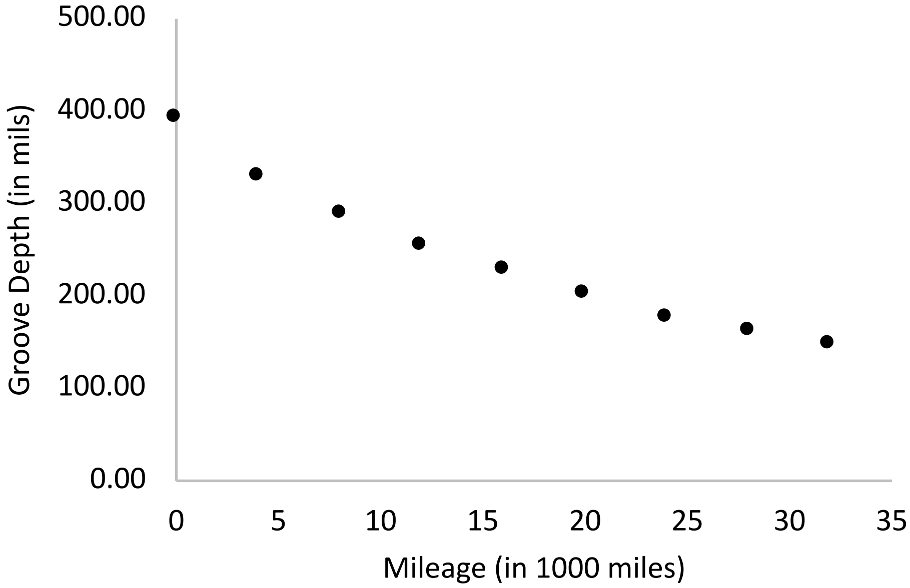
2 Fitting the Simple Linear Regression Model

Example

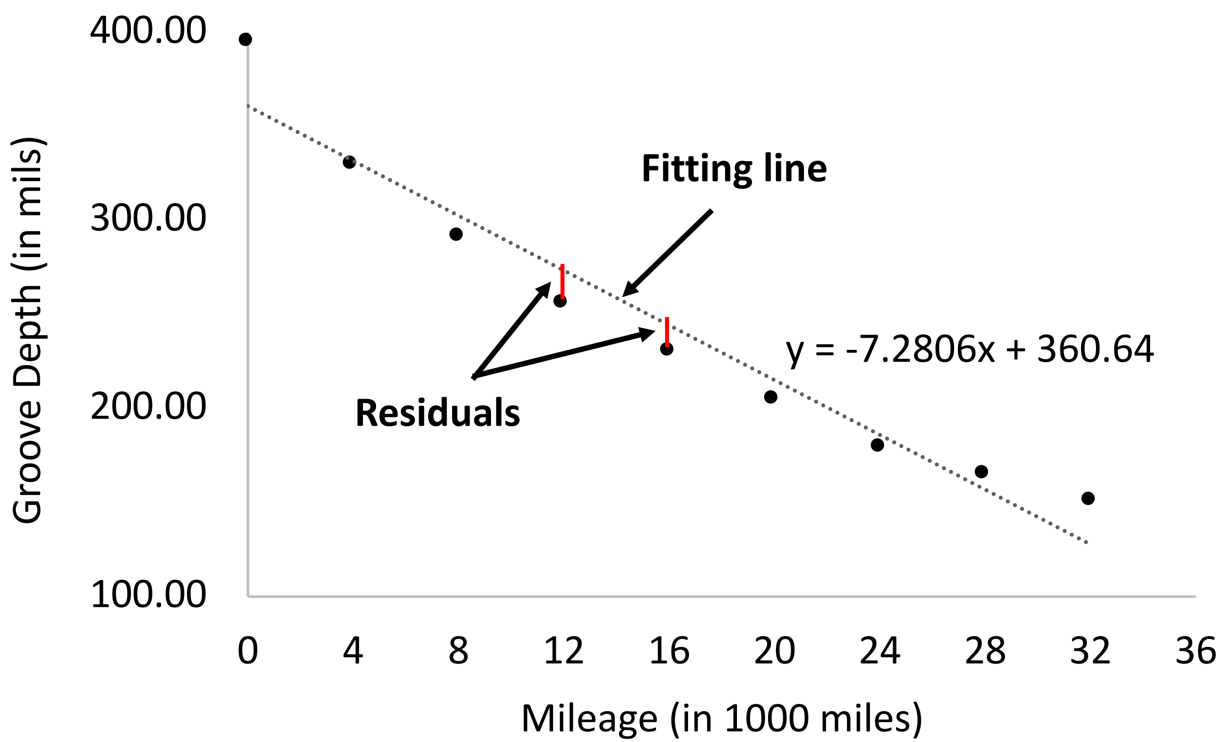
- Tires Tread Wear vs. Mileage

(Statistics and Data Analysis; Tamhane and Dunlop; Prentice Hall)

Mileage (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



Least Squares (LS) Fit



Fitting line $y = \beta_0 + \beta_1 x$ **Residual** $r_i = y_i - (\beta_0 + \beta_1 x_i) \quad (i = 1,2,\ldots,n)$

Objective function $Q = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$ **Sum of squared error (SSE)**

LS Estimate

The "best" fitting straight line in the sense of minimizing Q: LS estimate

- One way to find the LS estimate $\hat{\beta}_0$ and $\hat{\beta}_1$

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

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

- Setting these partial derivatives equal to zero and simplifying, we get

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

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

LS Estimate

- Solve the equations and we get

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

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

LS Estimate

- To simplify, we introduce

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

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

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

- The resulting equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is known as the least squares line, which is an estimate of the true regression line.

Example - Tires Tread Wear vs. Mileage

- Find the equation of the line for the tire tread wear data and we have

$$\sum_{i=1}^{n} x_i = 144, \quad \sum_{i=1}^{n} y_i = 2197.32,$$
$$\sum_{i=1}^{n} x_i^2 = 3264, \quad \sum_{i=1}^{n} y_i^2 = 589887.08,$$
$$\sum_{i=1}^{n} x_i\, y_i = 28167.72$$

and $n = 9$. From these we calculate $\bar{x} = 16, \bar{y} = 244.15,$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$
$$= 28167.72 - \frac{1}{9}(144 * 2197.32) = -6989.40$$
$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 3264 - \frac{1}{9}(144)^2 = 960$$

Example - Tires Tread Wear vs. Mileage

- The slope and intercept estimates are

$$\hat{\beta}_1 = \frac{-6989.40}{960} = -7.281 \text{ and}$$

$$\hat{\beta}_0 = 244.15 + 7.291 * 16 = 360.64$$

- Therefore, the equation of the LS line is

$$y = -7.281x + 360.64$$

There is a loss of 7.281 mils in the tire groove depth for every 1000 miles of driving.

- Given a particular $x = 25$, we can find

$$y = -7.281 * 25 + 360.64 = 178.62 \text{ mils}$$

which means the mean groove depth for all tires driven for 25,000 mils is estimated to be 178.62 miles.

Goodness of Fit of the LS Line

- Coefficient of Determination and Correlation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \ \ (i = 1,2, \dots, n)$$

- The residuals:

$$\epsilon_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) \ \ (i = 1,2, \dots, n)$$

are used to evaluate the goodness of fit of the LS line.

Goodness of Fit of the LS Line

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\textbf{SSR}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\textbf{SSE}} +$$

$$\underbrace{2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{\mathbf{0}}$$

- We define:

$$SST = SSR + SSE$$

- The ratio:

$$\boxed{\textbf{Coefficient of determination}} \leftarrow R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Note: Total sum of squares (SST)

  Regression sum of squares (SSR)

  Error sum of squares (SSE)

Example - Tires Tread Wear vs. Mileage

- For the tire tread wear data, calculate R^2 using the results from example, and we have

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

$$= 589887.08 - \frac{1}{9}(2197.32)^2 = 53418.73$$

- Next calculate

$$SSR = SST - SSE = 53418.73 - 2531.53 = 50887.20$$

- Therefore

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

Example - Tires Tread Wear vs. Mileage

- The Pearson correlation is

$$r = -\sqrt{0.953} = -0.976$$

where the sign of r follows from the sign of $\hat{\beta}_1 = -7.281$ since 95.3% of the variation in tread wear is accounted for by linear regression on mileage, the relationship between the two is strongly linear with a negative slope.

- Consider the linear model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

where ϵ_i is drawn from a normal population with mean 0 and standard deviation σ, the likelihood function for Y is:

$$L = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[\frac{-\sum(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right]$$

Maximum Likelihood Estimators (MLE)

- Thus, the log-likelihood for the data is:

$$\log L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \sum \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}$$

- Solving

$$\frac{\partial \log L}{\partial \beta_0} = 0, \frac{\partial \log L}{\partial \beta_1} = 0, \frac{\partial \log L}{\partial \sigma^2} = 0$$

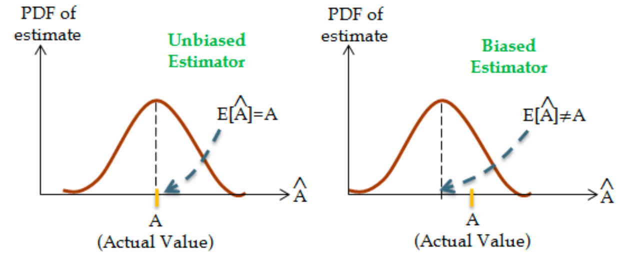
- We obtain the MLEs of the three unknown model parameters $\beta_0, \beta_1, \sigma^2$

Advanced

- The MLEs of the model parameters β_0 and β_1 are the same as the LSEs – both unbiased

- The MLE of the error variance, however, is biased:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \epsilon_i^2}{n} = \frac{SSE}{n}$$



Unbiased Estimator of σ^2

- An unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n} \epsilon_i^2}{n-2} = \frac{SSE}{n-2}$$

- Find the estimate of σ^2 for the tread wear data using the results from Example
- We have $SSE = 2351.3$ and $n - 2 = 7$, therefore

$$s^2 = \frac{2351.53}{7} = 361.65$$

  which has 7 d.f..
- The estimate of σ is $s = \sqrt{361.65} = 19.02$ miles.

3 Statistical Inference on Coefficients

Statistical Inference on $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$

- Under the normal error assumption
- Point estimators: $\hat{\beta}_0$ and $\hat{\beta}_1$
- Sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{\beta}_0 \sim N\left(\boxed{\beta_0}, \boxed{\sigma^2 \frac{\sum x_i^2}{nS_{xx}}}\right), \qquad SE(\hat{\beta}_0) = s\sqrt{\frac{\sum x_i^2}{nS_{xx}}}$$

$$\hat{\beta}_1 \sim N\left(\boxed{\beta_1}, \boxed{\frac{\sigma^2}{S_{xx}}}\right), \qquad SE(\hat{\beta}_1) = \frac{s}{\sqrt{S_{xx}}}$$

for your homework

Statistical Inference on β_0 and β_1

- Pivotal Quantities (P.Q.'s):

$$\frac{\hat{\beta}_0 - \beta_0}{SE\left(\hat{\beta}_0\right)} \sim t_{n-2}, \qquad \frac{\hat{\beta}_1 - \beta_1}{SE\left(\hat{\beta}_1\right)} \sim t_{n-2}$$

- Confidence Intervals (CI's):

$$\hat{\beta}_0 \pm t_{n-2,\frac{\alpha}{2}} SE\left(\hat{\beta}_0\right), \qquad \hat{\beta}_1 \pm t_{n-2,\frac{\alpha}{2}} SE\left(\hat{\beta}_1\right)$$

Statistical Inference on $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$

- Hypothesis tests:

  General form: $H_0: \beta_1 = c, H_a: \beta_1 \neq c$

  Our interest:

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

- At the significance level α, we reject H_0 in favor of H_a if and only if $|t_0| \geq t_{n-2,\frac{\alpha}{2}}$

- **The first test is used to show whether there is a linear relationship between x and y.**

Analysis of Variance (ANOVA)

- Another test to show whether there is a linear relationship between x and y

$$H_0: \beta_1 = 0, \qquad H_a: \beta_1 \neq 0$$

- Mean square: a sum of squares divided by its d.f.

$$MSR = \frac{SSR}{1},\ MSE = \frac{SSE}{n-2}$$

$$\frac{MSR}{MSE} = \frac{SSR}{s^2} = \frac{\widehat{\beta}_1^2 S_{xx}}{s^2} = \left(\frac{\widehat{\beta}_1}{s/\sqrt{S_{xx}}}\right)^2 = \left(\frac{\widehat{\beta}_1}{SE(\widehat{\beta}_1)}\right)^2 = t_0^2 \sim F_{1,n-2}$$

  - $SSR = \Sigma_i(\hat{y}_i - \bar{y})^2$
  - How can we represent $\hat{y}_i$, $\bar{y}$ with $\hat{\beta}_0$ and $\hat{\beta}_1$?

- It tests the model (simple linear regression) significance.

Analysis of Variance (ANOVA)

- ANOVA Table

Source of Variation (Source)	Sum of Squares (SS)	Degrees of Freedom (d.f.)	Mean Square (MS)	F
Regression	SSR	1	MSR = SSR/1	MSR/MSE
Error	SSE	n – 2	MSE = SSE/n-2	
Total	SST	n – 1		

- Example

Source	SS	d.f.	MS	F
Regression	50,887.20	1	50,887.20	140.71
Error	2531.53	7	361.25	vs. $F_{1,7}$
Total	53,418.73	8		significance level α

Questions?