Chapter 8.
Simple Linear Regression

Regression analysis:

- **regression analysis** is a statistical methodology to estimate the relationship of a response variable to a set of predictor variable.

- 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

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.

Section 8.1.
Simple Linear Regression

Model Assumption

- Specific settings of the **predictor variable (x)** and corresponding values of the **response variable (Y)**

$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$$

Assume :

y_i - observed value of the random variable Y_i on x_i

Simple Linear Regression :

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\quad i = 1,..n$$

where random error are $\varepsilon_i \sim^{i.i.d.} N(0, \sigma^2)$.

Response Y_i : $E(Y_i) = \beta_0 + \beta_1 x_i,\ Var(Y_i) = \sigma^2$

$$Y_i \sim^{ind.} N(\beta_0 + \beta_1 x_i, \sigma^2), i = 1,\ldots,n$$

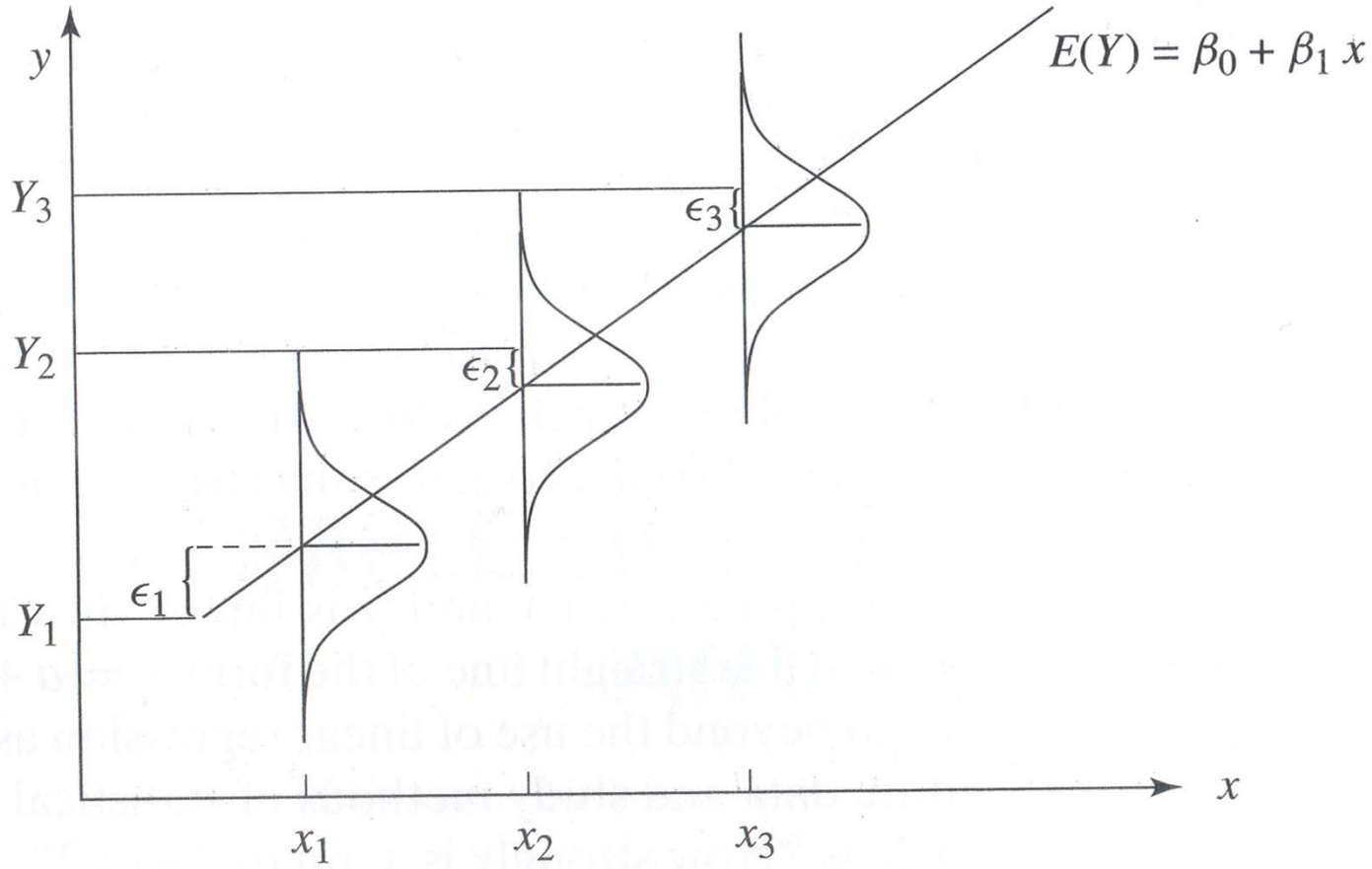


Figure 10.1 Simple Linear Regression Model

Example 1. (Tires Tread Wear vs. Mileage: Scatter Plot)

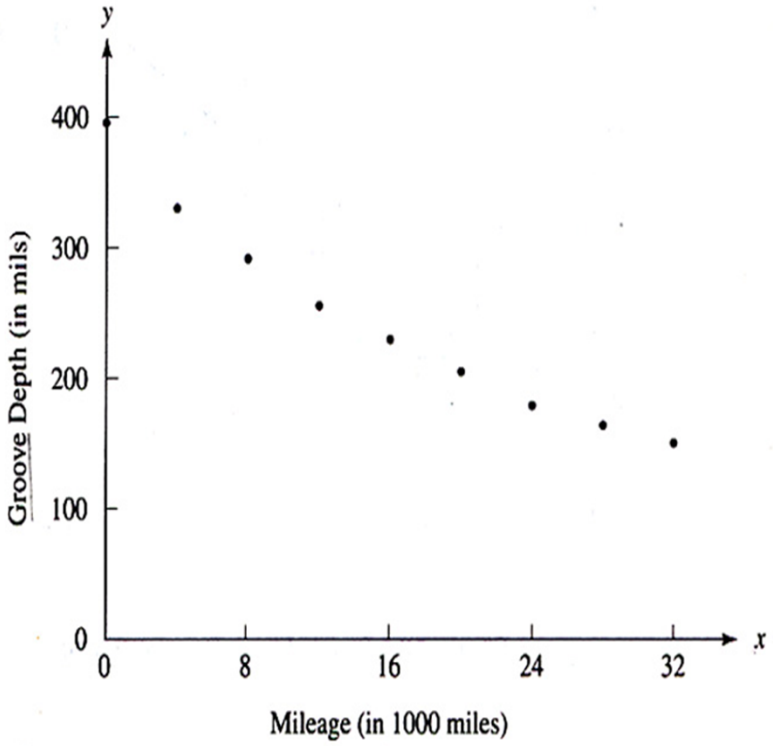


Figure 10.2 Scatter Plot of Groove Depth vs. Mileage

Table 10.1 Mileage and Groove Depth of a Car Tire

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

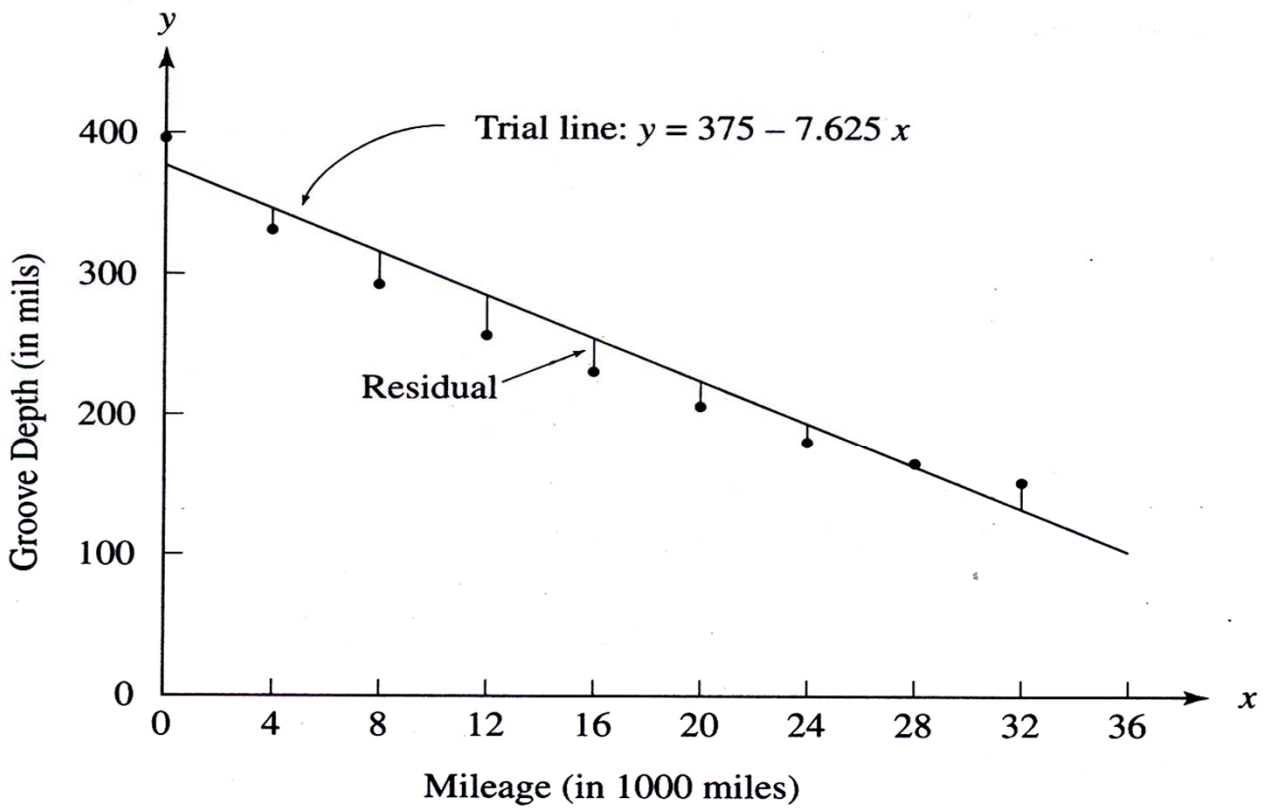


Figure 10.3 Scatter Plot with a Trial Straight Line Fit

Least Square Criterion (residual sum square)

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

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

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

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

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

Setting these partial derivatives equal to zero and simplifying, we get the **Normal Equation**

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

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

- Solve the equations and we get the **least square estimators** of β_0 and β_1.

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

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

Maximum Likelihood Estimators of β_0 and β_1.

- Under normal assumption of the errors, the likelihood function of the parameters, β_0 and β_1, given the (observed) responses is

$$L(\beta_0,\beta_1;Y_1,\ldots,Y_n)=\prod_{i=1}^{n} f_{Y_i}(y_i;\beta_0,\beta_1), \quad [\text{Given } Y_i \sim^{ind} N(\beta_0+\beta_1 x_i,\sigma^2).]$$

$$=\prod_{i=1}^{n}\left(2\pi\sigma^2\right)^{-1/2}\exp\left\{-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}\right\},$$

$$=\left(2\pi\sigma^2\right)\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right\}.$$

Maximum Likelihood Estimators (MLE) of β_0 and β_1 :

$$(\hat{\beta}_0,\hat{\beta}_1)_{MLE}=\max_{\beta_0,\beta_1\in R}\{L(\beta_0,\beta_1)\}$$

$$=\min_{\beta_0,\beta_1\in R}\left\{\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right\}=(\hat{\beta}_0,\hat{\beta}_1)_{LSE}$$

- To simplify expressions of the LS solution, we introduce

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

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

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

$$\text{Coefficient LSE}: \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

- We get The 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 2 (Tire Tread vs. Mileage: Least Square Fit)

Find the equation of the line for the tire tread wear data

$$\sum x_i = 144, \sum y_i = 2197.32, \sum x_i^2 = 3264, \sum y_i^2 = 589,887.08, \sum x_i y_i = 28,167.72$$

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

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = 28,167.72 - \frac{1}{9}(144 * 2197.32) = -6989.40$$

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

The slope and intercept estimates are

$$\hat{\beta}_1 = \frac{-6989.40}{960} = -7.281 \ \ and \ \ \hat{\beta}_0 = 244.15 + 7.281*16 = 360.64$$

Therefore, the equation of the LS line is

$$y = 360.64 - 7.281x.$$

Conclusion: 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 that $y = 360.64 - 7.281*25 = 178.62$ mils

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

Goodness of Fit of the LS Line

Coefficient of Determination and Correlation

$$\text{Fitted line}: \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (i = 1, 2, ...n)$$

- The residuals: $e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right), i = 1, ..., n$ are used to evaluate the goodness of fit of the LS line.
- Decomposition of Sum Squares:

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

SST = SSR + SSE

Note: Sum of Squares of Total (SST)
Sum of Squares of Regression (SSR)
Sum of Squares of Errors (SSE)

Coefficient of Determination

- Define: $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SSR},$$
- The ratio is called the coefficient of determination. It explains the percentage of the variation in response variable (Y) is accounted for by linear regression on the explanatory variable (x).
- It can be shown that R^2 is the square of the sample linear correlation coefficient (r), i.e.

$$0 \le R^2 = r^2 = \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \right)^2 \le 1.$$

Example 3(Tire Tread Wear vs. Mileage: Coefficient of Determination and Correlation)

- For the tire tread wear data, calculate R^2 using the results from **Example 10.2** we have

$$SST = S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i^2) = 589,887.08 - \frac{1}{9}(2197.32)^2 = 53,418.73$$

- Calculate $SSR = SST - SSE = 53,418.73 - 2531.53 = 50,887.20$

- Therefore $r^2 = \frac{50,887.20}{53,418.73} = 0.953 \ \ and \ \ r = -\sqrt{0.953} = -0.976$

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.

Estimation of σ^2

An unbiased estimate of σ^2 is given by

$$MSE = \frac{SSE}{n-2}$$

Example 4. (Tire Tread Wear Vs. Mileage: Estimate of σ^2

Find the estimate of for the tread wear data using the results from Example 10.3 we have *SSE=2351.3* and *n-2=7*,therefore

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{2351.53}{7} = 361.65$$

which has df=7. The estimate of σ is

$$\hat{\sigma} = \sqrt{MSE} = \sqrt{361.65} = 19.02$$

Statistical Inference on β_0 and β_1

Point estimators: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$

Sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{\beta}0 \sim N\left(\beta 0, \frac{\sum xi^2}{nSxx}\sigma^2\right), \; SE\left(\hat{\beta}0\right) = \sqrt{\frac{\sum xi^2}{nSxx} MSE}$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{Sxx}\right), \; SE\left(\hat{\beta}_1\right) = \sqrt{\frac{MSE}{Sxx}}$$

where $\hat{\sigma}^2 = MSE$, the mean square error.

Statistical Inference on β_0 and β_1, Con't

Sampling distribution (parameter-free):

$$\frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \sim t(n-2) \qquad \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t(n-2)$$

Confidence Interval's for β_0 and β_1 :

$$\hat{\beta}_0 \pm t_{\frac{\alpha}{2}}(n-2) \cdot SE(\hat{\beta}_0),$$

$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}}(n-2) \cdot SE(\hat{\beta}_1).$$

Testing hypothesis on β_0 and β_1

Hypothesis test: $H_0 : \beta_1 = \beta_1^0 \quad H_a : \beta_1 \neq \beta_1^0$

or $H_0 : \beta_1 = 0 \quad H_a : \beta_1 \neq 0$

-- Test statistic:

$$T = \frac{\hat{\beta}_1 - \beta_1^0}{SE(\hat{\beta}_1)}$$

-- At the significance level α, we reject H_0 in favor of H_a iff $|T| > t_{\alpha/2}(n-2)$

-- Can be used to show whether there is a linear relationship between x and Y.

Analysis of Variance (ANOVA), Con't

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

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

$$\frac{MSR}{MSE}=\frac{SSR}{MSE}=\frac{\hat{\beta}_1^{\,2}Sxx}{MSE}=\left(\frac{\hat{\beta}_1}{\sqrt{MSE/Sxx}}\right)^2$$

$$\frac{MSR}{MSE}=\left(\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}\right)^2=T^2\sim F(1,n-2)\ \text{under}\ H_0:\beta_1=0.$$

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	$MSR=\frac{SSR}{1}$	$F=\frac{MSR}{MSE}$
Error	SSE	n - 2	$MSE=\frac{SSE}{n-2}$	
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	
Total	53,418.73	8		

Regression Diagnostics

Checking for Model Assumptions

- Checking for Linearity
- Checking for Constant Variance
- Checking for Normality
- Checking for Independence

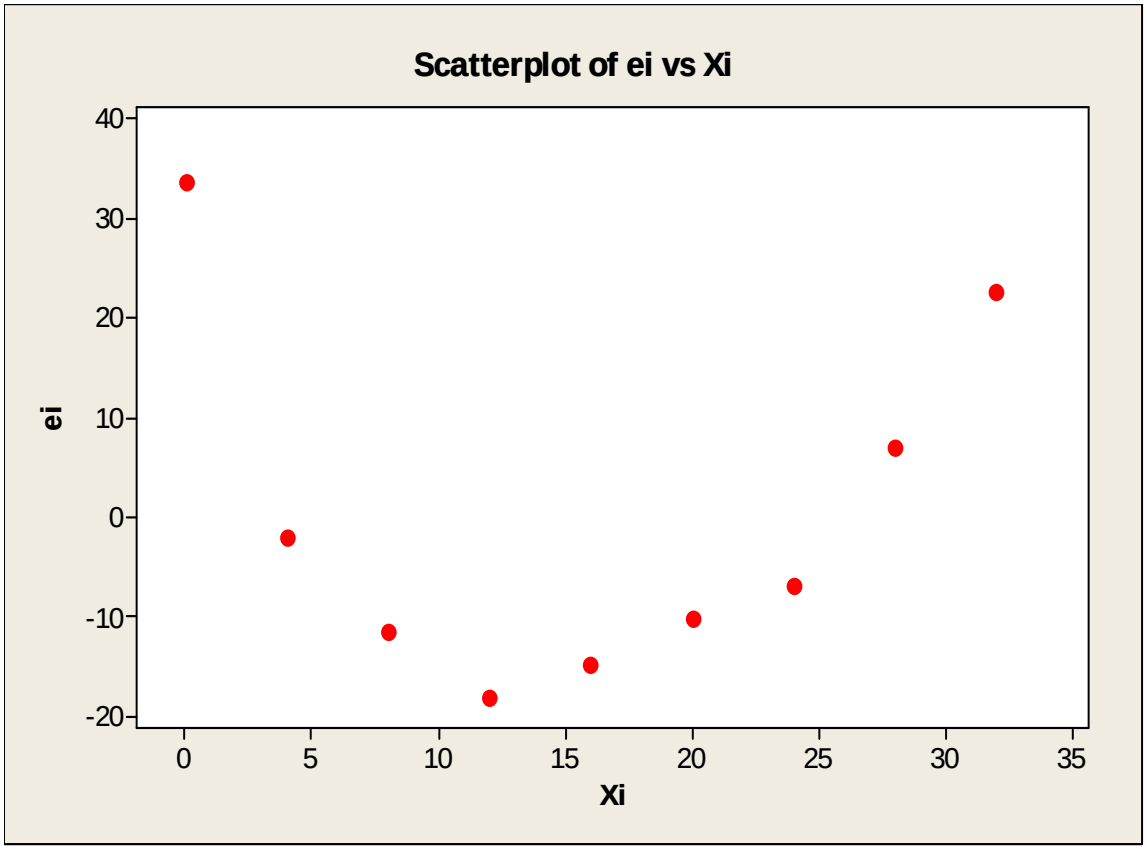
Checking for Linearity

i	Xi	Yi	^ Yi	ei
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

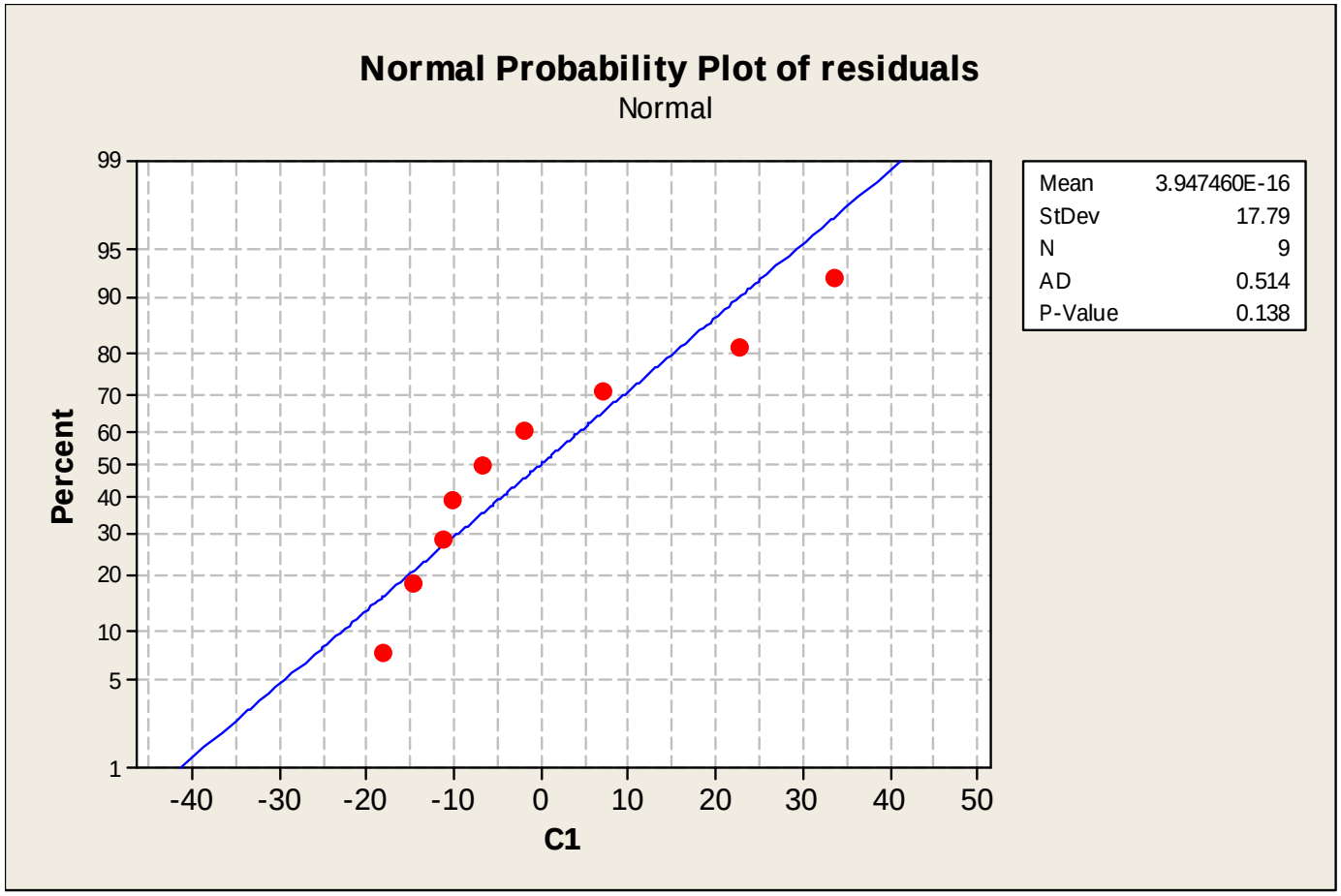
Xi = Mileage
Yi = Groove Depth

$Y=\beta_0 + \beta_1 x$

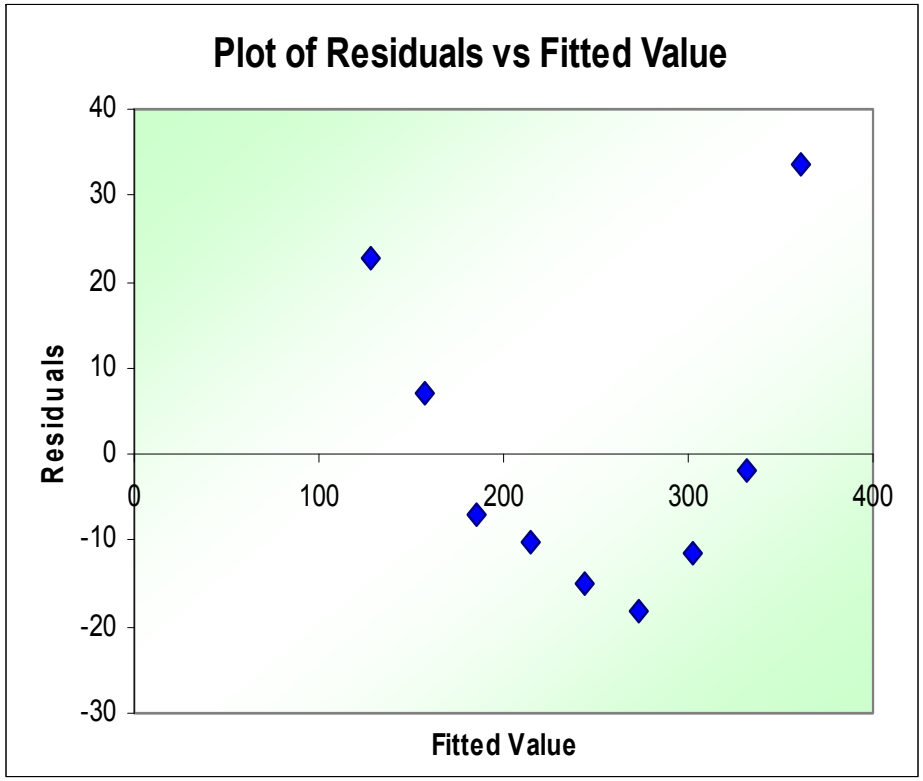
Yi _hat = fitted value **ei =residual**



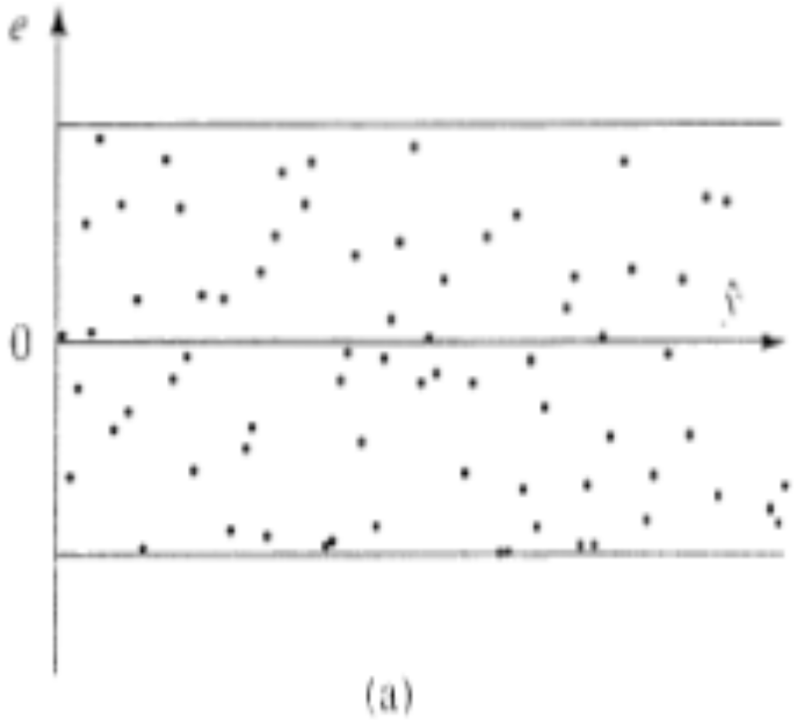
Checking for Normality



➢Checking for Constant Variance



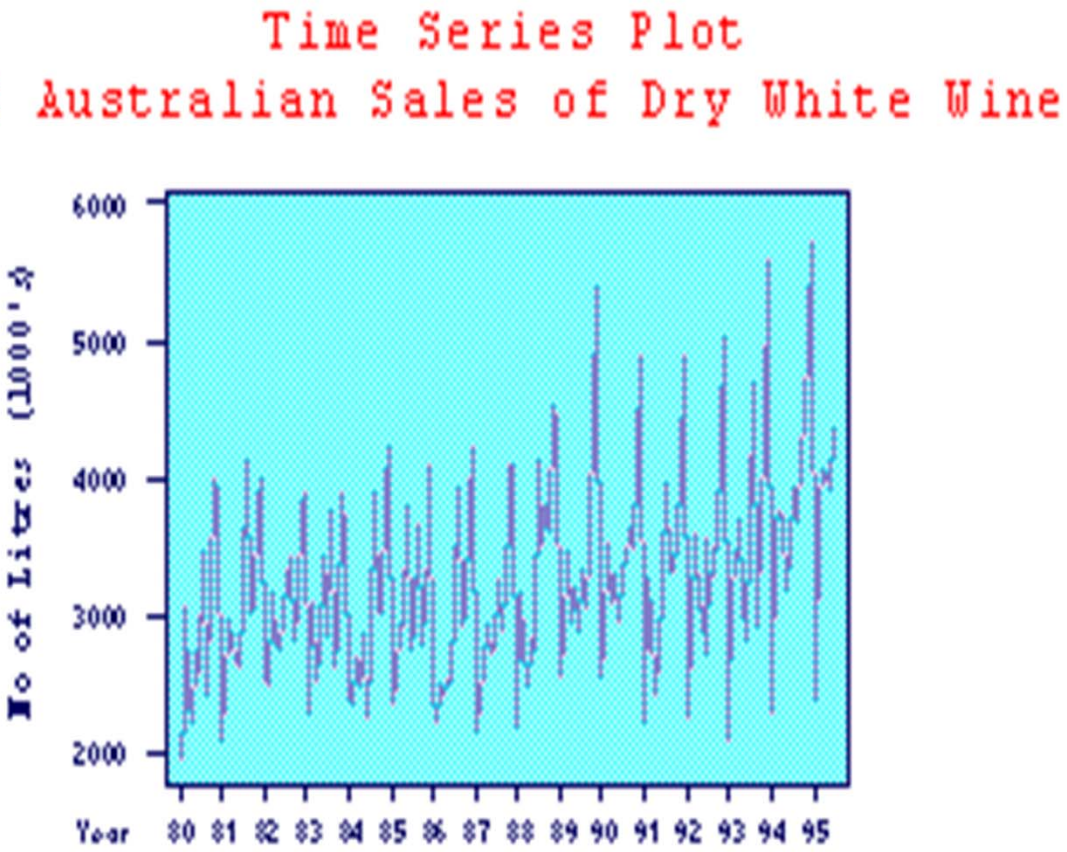
Var(Y) is not constant.



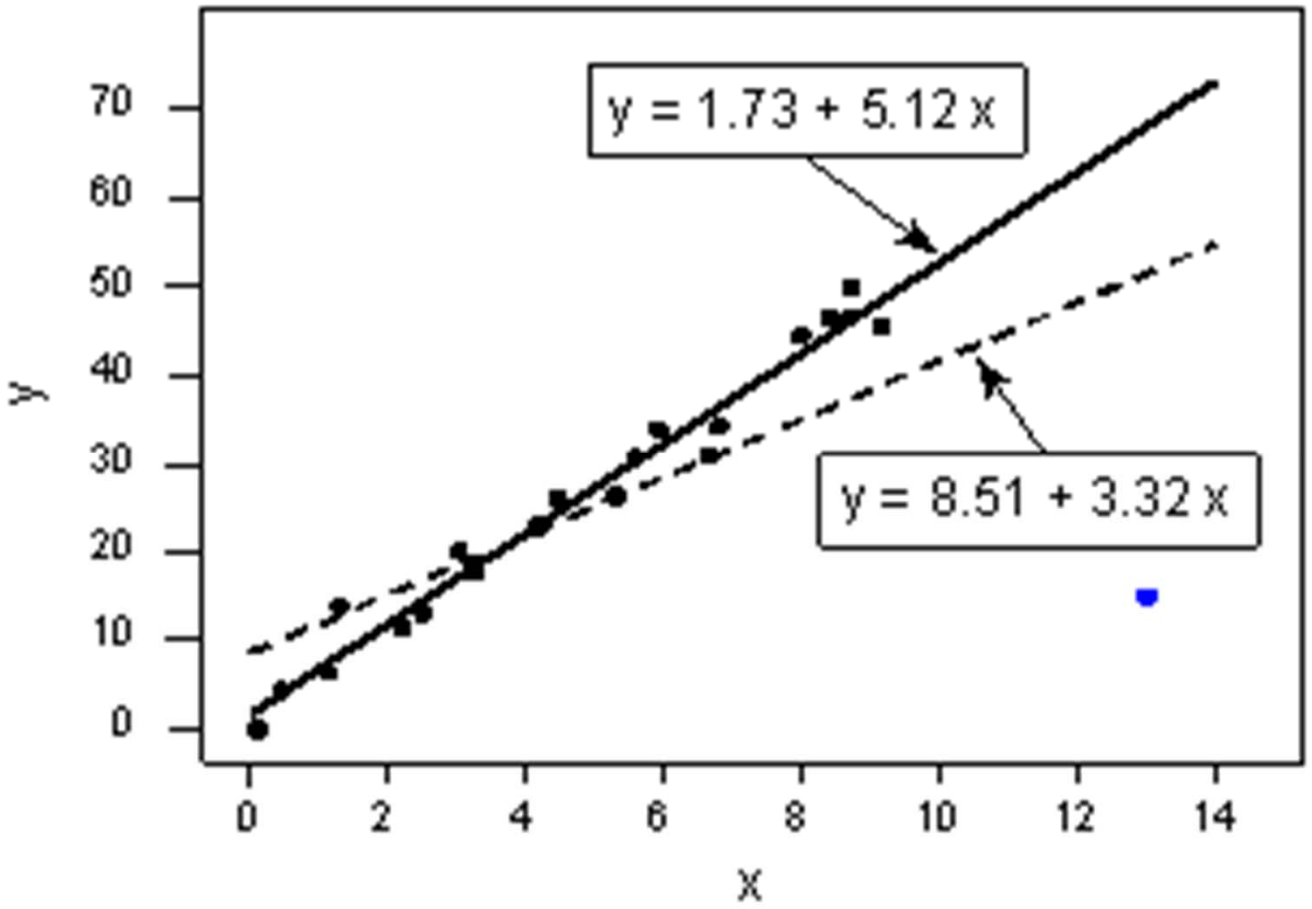
A sample residual plots when Var(Y) is constant.

➢Checking for Independence

- Does not apply for Simple Linear Regression Model
- Only apply for time series data



Outlier and Influential Points



Do the two samples yield different results when testing H_0: $\beta_1 = 0$? We obtain the following output when the **blue data point is included**:

```
The regression equation is y = 8.50 + 3.32 x

Predictor      Coef      SE Coef         T         P
Constant      8.505        4.222      2.01     0.058
x            3.3198       0.6862      4.84     0.000

S = 10.45         R-Sq = 55.2%      R-Sq(adj) = 52.8%
```

and the following output when the **blue data point is excluded**:

```
The regression equation is y = 1.73 + 5.12 x

Predictor      Coef      SE Coef         T         P
Constant      1.732        1.121      1.55     0.140
x            5.1169       0.2003     25.55     0.000

S = 2.592         R-Sq = 97.3%      R-Sq(adj) = 97.2%
```

Checking for Outliers & Influential Observations

- What is OUTLIER
- Why checking for outliers is important?
- Mathematical definition

- How to deal with them?

- Investigate (Data errors? Rare events? Can be corrected?)
- Ways to accommodate outliers

1. Non Parametric Methods (robust to outliers)
2. Data Transformations
3. Deletion (or report model results both with and without the outliers or influential observations to see how much they change)

Data Transformations

Reason

- To achieve linearity
- To achieve homogeneity of variance
- To achieve normality or symmetry about the regression equation

- **Method of Linearizing Transformation**

➢ Use mathematical operation, e.g. square root, power, log, exponential, etc.

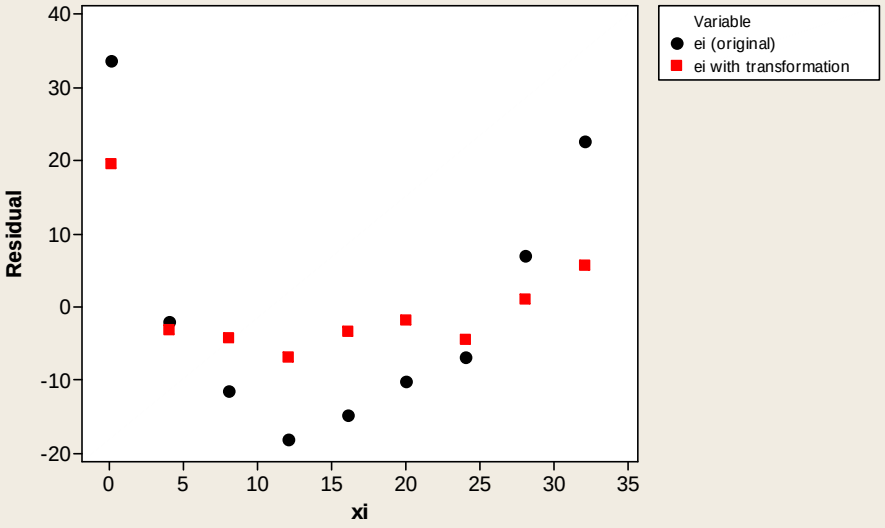
➢ Only one variable needs to be transformed in the simple linear regression.

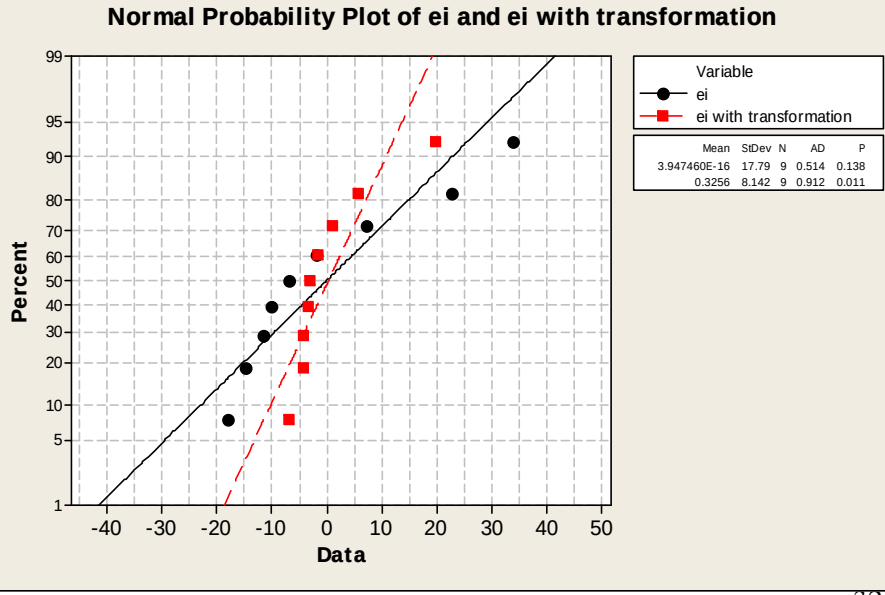
Which one? Predictor or Response? Why?

Exponential transformation on $Y = \alpha \exp(-\beta x)$

$\Leftrightarrow \log Y = \log \alpha - \beta x$

Xi	Yi	$\widehat{\log Y_i}$	$\exp(\widehat{\log Y_i})$	Ei
0	394.33	5.926	374.64	19.69
4	329.50	5.807	332.58	-3.08
8	291.00	5.688	295.24	-4.24
12	255.17	5.569	262.09	-6.92
16	229.33	5.450	232.67	-3.34
20	204.83	5.331	206.54	-1.71
24	179.00	5.211	183.36	-4.36
28	163.83	5.092	162.77	1.06
32	150.33	4.973	144.50	5.83





Residual Check

- Model checking by residual plots:
  1. residual vs fitted value --- e_i vs Y_i
  2. residual vs explanatory variable ---- e_i vs x_i
  3. residual vs lag of residual --- e_i vs e_{i-1}
- Transformation
  1. Boxcox transformation for skewed distribution
  2. log transformation
  3. square root transformation
- Correlation of residuals
  1. correlation coefficient
  2. Durbin-Watson Statistic (series)

Durbin-Watson Statistic

Lag 1 autocorrel ation coefficien t

$$Corr\,(y_i, y_{i-1}) \cong \frac{\sum_{t=2}^{n} e_{t-1}e_t}{\sum_{t=1}^{n-1} e_t^2} = r_1 \sim N\left(0, \frac{1}{n}\right).$$

Lag k autocorrel ation coefficien t

$$Corr\,(y_i, y_{i-k}) \cong r_k$$

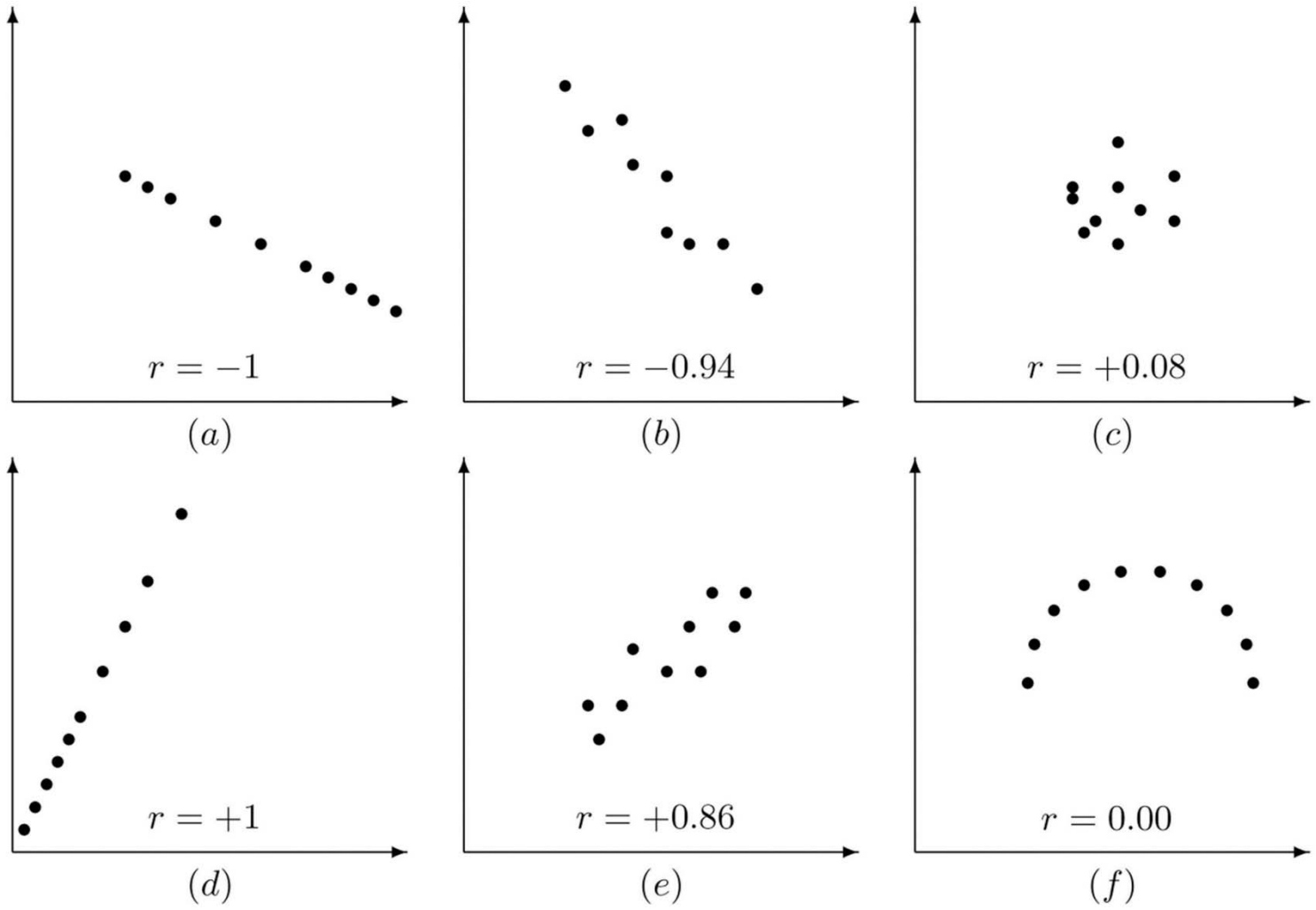
Durbin – Watson Statistic :

$$D = \frac{\sum_{t=2}^{n} (e_{t-1} - e_t)^2}{\sum_{t=1}^{n-1} e_t^2} \cong 2 - 2r_1$$

Correlation Analysis

- **Correlation**: a measurement of how closely two variables share a linear relationship.
  - $$\rho = \text{corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$
- Useful when it is not possible to determine which variable is the predictor and which is the response.
  - Health vs wealth. Which is predictor? Which is response?

Linear Correlation Coefficient



Derivation of T

are these equivalent ?

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \stackrel{?}{=} \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

substitute :

$$r = \hat{\beta}_1 \frac{s_x}{s_y} = \hat{\beta}_1 \sqrt{\frac{S_{xx}}{S_{yy}}} = \hat{\beta}_1 \sqrt{\frac{S_{xx}}{SST}}$$

$$1 - r^2 = \frac{SSE}{SST} = \frac{(n-2)s^2}{SST}$$

then :

$$\mathrm{t} = \hat{\beta}_1 \sqrt{\frac{S_{xx}}{SST}} \sqrt{\frac{(n-2)SST}{(n-2)s^2}} = \frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

∴ yes, they are equivalent .

- Therefore, we can use *t* as a statistic for testing against the null hypothesis

  $H_0: \beta_1 = 0$

- Equivalently, we can test against

  $H_0: \rho = 0$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2)$$