MSAN 601
Linear Regression Analysis

Paul Intrevado

Simple Linear Regression

Wednesday 24th August, 2016

00:12

UNIVERSITY OF
SAN FRANCISCO

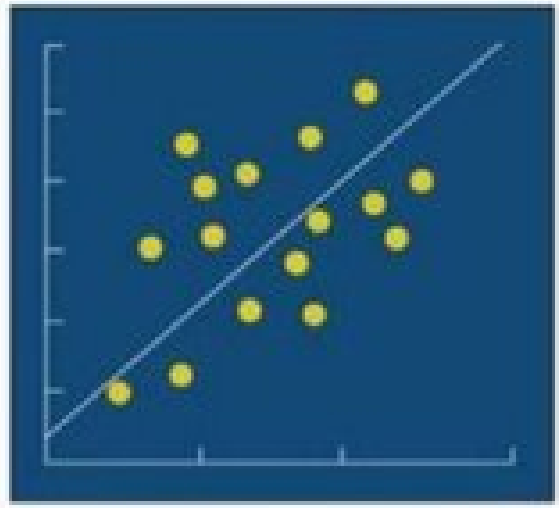
Master of Science
in Analytics

Section 1

Reference Textbooks

APPLIED LINEAR REGRESSION MODELS

FOURTH EDITION



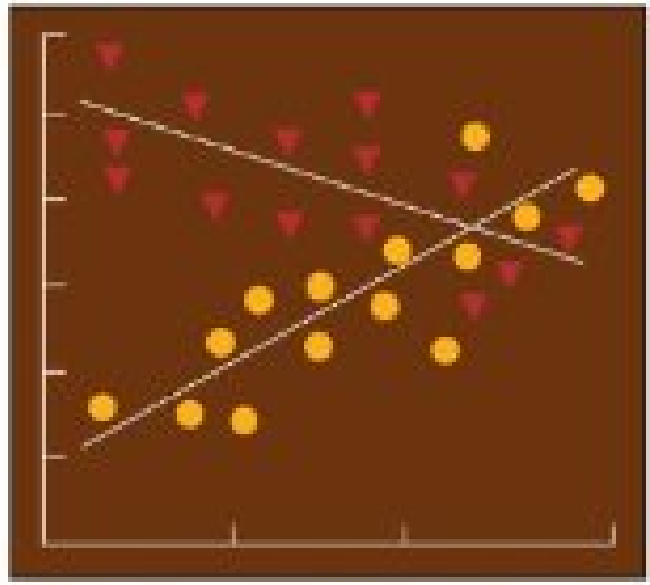
Kutner

Nachtsheim

Neter

APPLIED LINEAR STATISTICAL MODELS

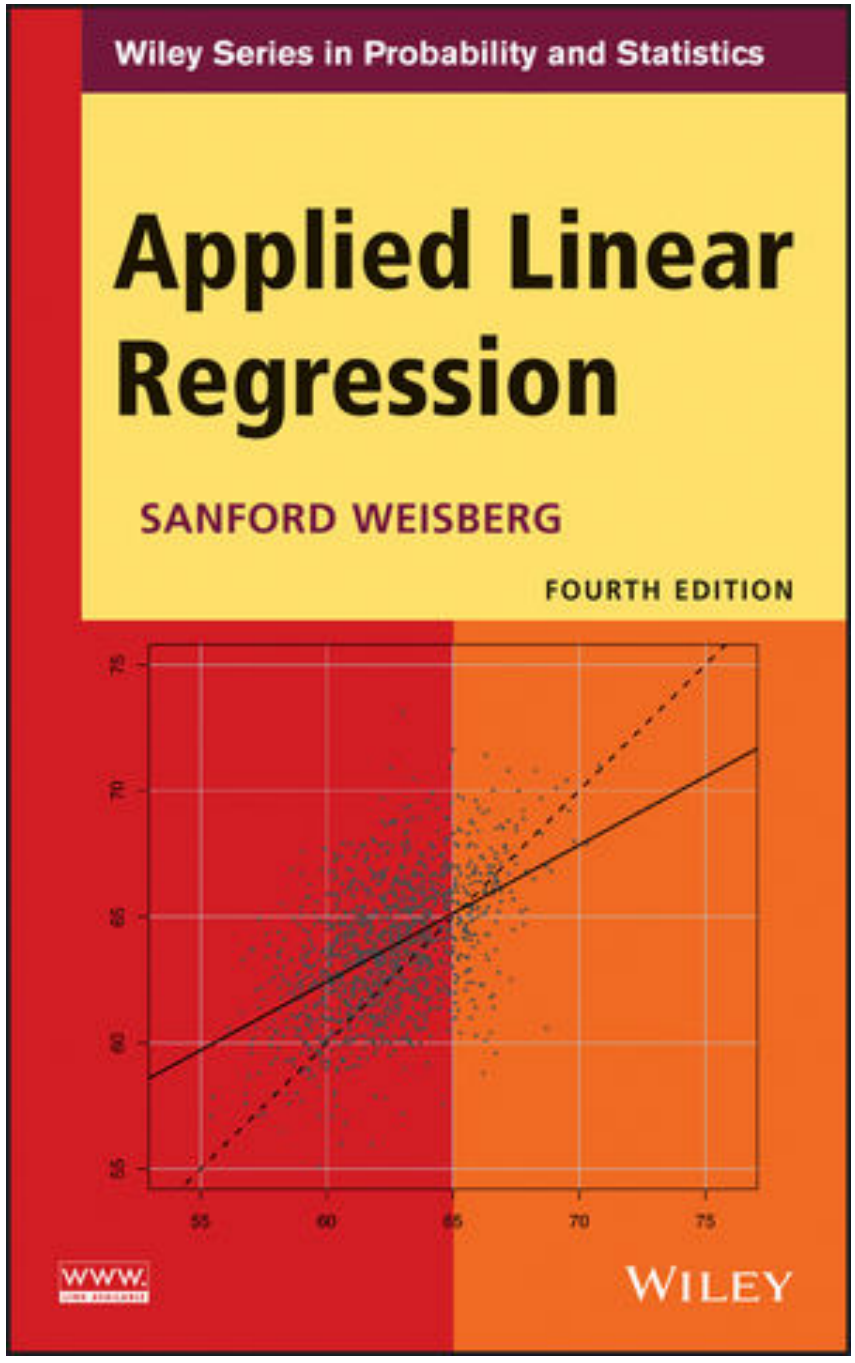
FIFTH EDITION

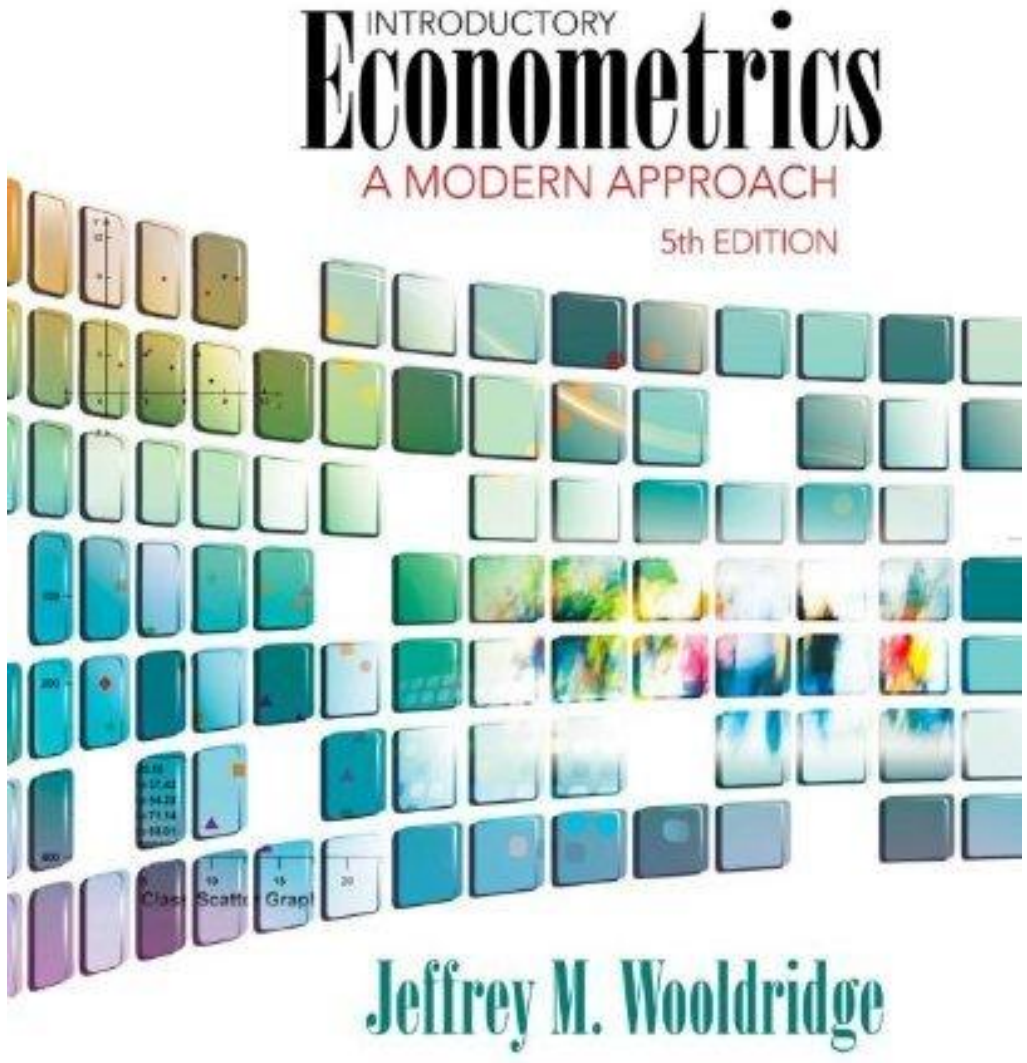


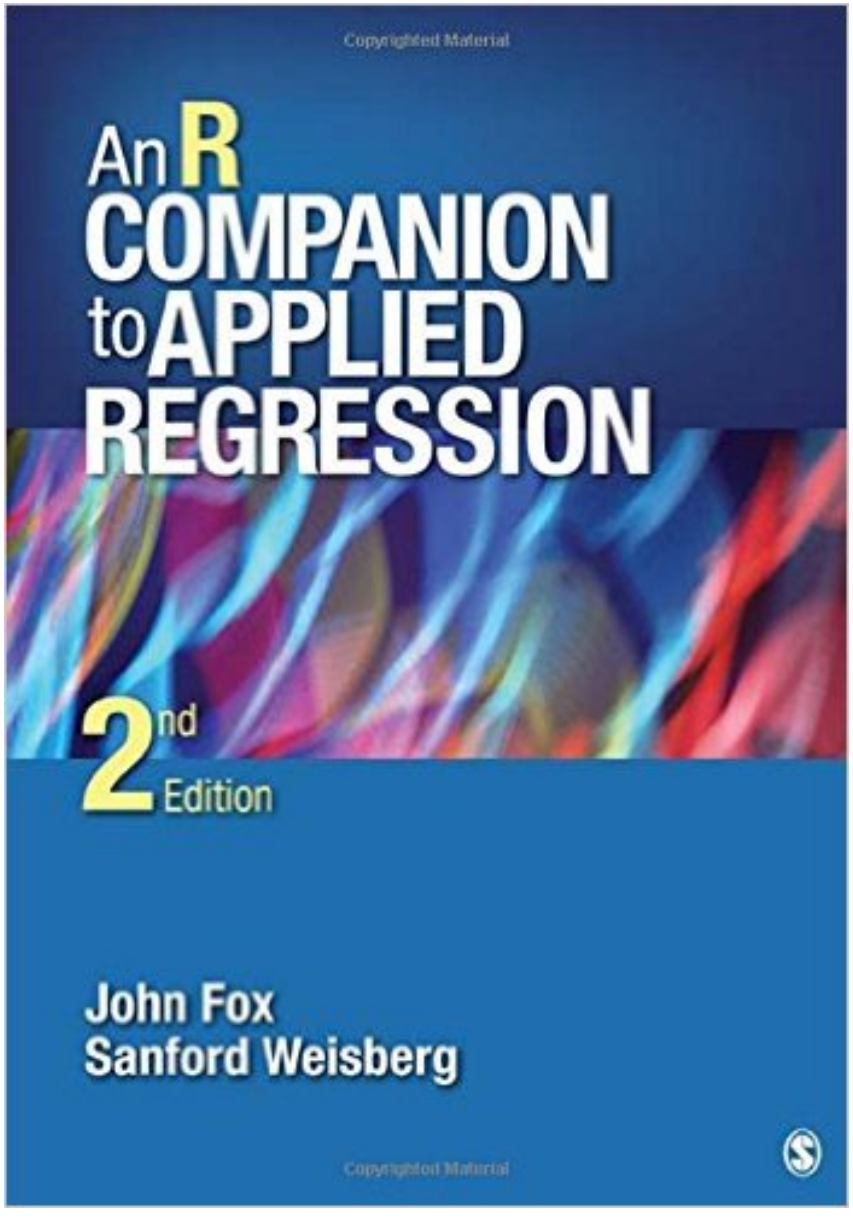
Kutner

Nachtsheim

Neter


Wiley Series in Probability and Statistics
Applied Linear Regression
SANFORD WEISBERG
FOURTH EDITION
55
60
65
70
75
WILEY




Copyrighted Material
An R
COMPANION
to APPLIED
REGRESSION
2nd
Edition
John Fox
Sanford Weisberg
Copyrighted Material

Section 2

Linear Regression with One Predictor Variable

Why *Simple*? Why *Linear*? Why *Regression*?

- Simple: only dealing with two variables, an explanatory variable (X) and a response variable (Y)
- Linear: attempting to quantify a linear relationship, as opposed to a non-linear (curvilinear) relationship
- Regression: method of analysis developed by Sir Francis Galton in the late 1800s
    - when studying the relationship between the heights of parents and their children, he noted children of both shorter and taller parents *regressed* to the group mean height, and the term *regression* has since endured

- Functional relationship, e.g., $y = f(x) = 3x - 9$
  - for each x, function returns a corresponding value of y
- Statistical relationships are imperfect
  - observations for a statistical relationship do not typically fall directly on the curve (line) of the relationship
  - even so, statistical relationship can be of great use, and will form the the basis for linear regression

- Analyze individual variables
  - numerically: mean, median, mode, variance, quantiles, skewness, kurtosis, etc.
  - graphically: histograms, box plots, etc.
- Analyze relationship(s) between variables
  - numerically: covariance, correlation
  - graphically: scatter plots

- Formula for correlation

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \quad (1)$$

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

  where $\bar{x}, \bar{y}$ are sample means, s_x, s_y are sample standard deviations and s_{xy} is the sample covariance between x and y

- The x and y terms in (1) are multiplicative, therefore it doesn't matter which variable you select to input as x and which variable you select to input as y
- SLR is sensitive to which variable you choose as the explanatory variable and which variable you choose at the response: your analysis will change

Section 3

Defining the Simple Linear Regression Model

- A regression model is a formal means of expressing the two essential ingredients of a statistical relation:
  - A tendency of the response variable Y to vary with the predictor variable X in a systematic fashion
  - A scattering of points around the curve of the statical relationship
- These two characteristics are embodied in a regression model by postulating that:
  - There is a probability distribution of Y for each level of X
  - The means of these probability distributions vary in some systematic fashion with X

- The existence of a statistical relationship between the response variable Y and the explanatory variable X does not imply in any way that Y depends *causally* on X
- No matter how strong the relationship between Y and X, no cause and effect pattern is implied by the regression model
- Examples?

- Two variables: the explanatory variable, X, explains—in part or wholly—the response variable Y
- You may encounter other terminology when referring to explanatory and/or response variables:

Y	X
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor variable

Table 1: Terminology for Simple Linear Regression

- The graph of the regression function is called the regression line (or curve)
- The equation that accounts for statistical inexactitude of the relationship in a regression model is as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{2}$$

where

- Y_i is the value of the response variable in the i^{th} trial
- β_0 and β_1 are parameters: β_0 is the y-intercept and β_1 is the slope of the regression function
- X_i is a known constant, the value of the predictor variable
- ε_i is a random error term
- $i = 1, \ldots, n$
  - **n.b.** in Wooldridge 5*e*, this term is labelled u

A Note on the Error Term ε

ε is a random error term with

1. $E[\varepsilon_i|X_i] = 0$ (exogeneity)
2. $\sigma^2[\varepsilon_i|X_i] = \sigma^2$ (constant variance *or* homoscedasticity)
3. $\sigma[\varepsilon_i, \varepsilon_j|X_i] = 0 \ \forall \ i, j; i \neq j$ (no serial correlation)
4. $\sigma[X_i, \varepsilon_i] = 0 \ \forall i$

n.b. there is **no** requirement for the ε_i's to be normally distributed **yet**

Assumption for the Error Term ε

- Without loss of generality, we assume the average value of the error term, $E[\varepsilon] = 0$
- As long as the intercept β_0 is included in (2), nothing is lost by assuming $E[\varepsilon] = 0$. E.g., if the average error, $E[\varepsilon] = 5$, then (2) becomes:

$$E[y] = \beta_0 + \beta_1 x + E[\varepsilon]$$
$$E[y] = \beta_0 + \beta_1 x + 5$$

- To center the mean error $E[\varepsilon]$ about zero, we simply subtract 5, while adding 5 to the intercept β_0, to obtain:

$$E[y] = (\beta_0 + 5) + \beta_1 x + (5 - 5)$$

- The new equation has $E[\varepsilon] = 0$, a y-intercept of $(\beta_0 + 5)$, and an *unchanged* slope of β_1 (the line is shifted up 5 units)

Why Simple? Why Linear? [REVISITED]

- Simple: only one explanatory variable, X
- Linear:
  - linear in the parameters because no parameters appear as exponents or are multiplied or divided by another parameter
  - linear in the explanatory variable because this variable appears only in the first power
  - a model that linear in only the parameters is called a linear model
  - a model that linear in both the parameters and the variables is called a *first-order model*, a reference to the power of variables
- **n.b.** (2) may also be referred to as the two-variable linear regression model or the bivariate linear regression model as it relates two variables, X and Y

- Observe in (2) that the response Y_i is the sum of
  - the constant term $\beta_0 + \beta_1 X_i$, the **systematic part of** of Y_i
  - the random term ε_i, the **unsystematic part of** Y_i

  Hence Y_i is a random variable

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{2}$$

- Taking the expectation of (2) we obtain

$$\begin{aligned} E[Y_i] &= E[\beta_0 + \beta_1 X_i + \varepsilon_i] \\ E[Y_i] &= E[\beta_0] + E[\beta_1 X_i] + E[\varepsilon_i] \\ E[Y_i] &= \beta_0 + \beta_1 E[X_i] + E[\varepsilon_i] \\ E[Y_i] &= \beta_0 + \beta_1 X_i + E[\varepsilon_i] \\ E[Y_i] &= \beta_0 + \beta_1 X_i \end{aligned}$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{2}$$

- Taking the variance of (2) we obtain

$$\begin{aligned}
\sigma^2[Y_i] &= \sigma^2[\beta_0 + \beta_1 X_i + \varepsilon_i] \\
\sigma^2[Y_i] &= \sigma^2[\beta_0] + \sigma^2[\beta_1 X_i] + \sigma^2[\varepsilon_i] \\
\sigma^2[Y_i] &= \sigma^2[\varepsilon_i] = \sigma^2 \\
\sigma^2[Y_i] &= \sigma^2
\end{aligned} \tag{3}$$

- We therefore conclude that the probability distributions of Y have the same variance, regardless of the level of the predictor variable X

- Given the assumption that the error terms are uncorrelated, i.e.,

$$\sigma[\varepsilon_i, \varepsilon_j] = 0 \ \forall \ i, j; i \neq j$$

it follows that the responses Y_i and Y_j are also uncorrelated

$$\sigma[Y_i, Y_j] = 0 \ \forall \ i, j; i \neq j$$

- An electrical distributor is studying the relationship between the number of bids requested by construction contractors for basic lighting equipment during a week (X) and the time required to prepare those bids (Y), and the regression model is determined to be

$$Y_i = 9.5 + 2.1X_i + \varepsilon_i$$

and the regression function is

$$E[Y] = 9.5 + 2.1X$$

- Suppose in the i^{th} week, $X_i = 45$ bids are prepared and the actual number of hours required is $Y_i = 108$, which implies that the residual is $\varepsilon_i = 4$, as $9.5 + 2.1(45) = 104$

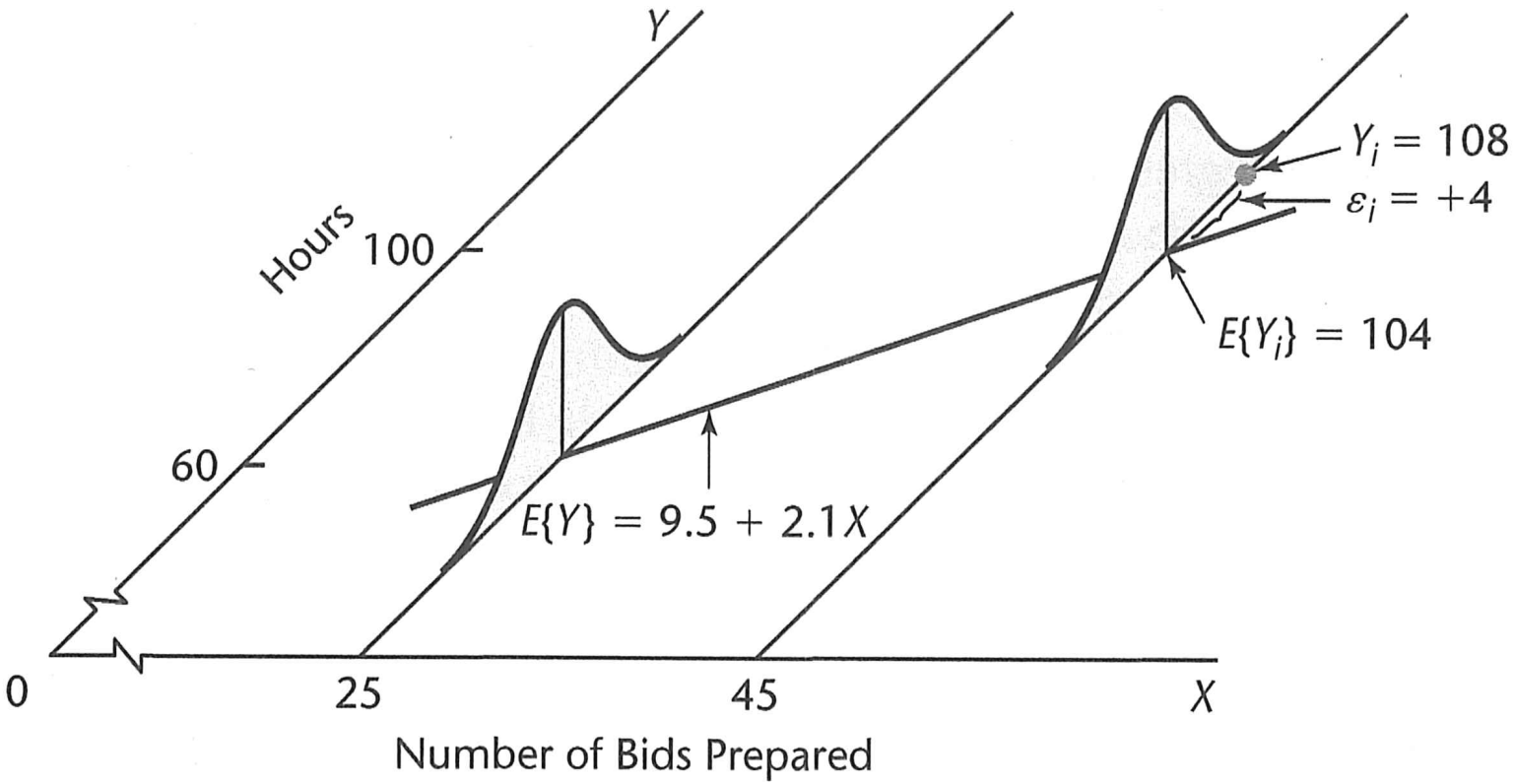


Figure 1: Illustration of an SLR Model

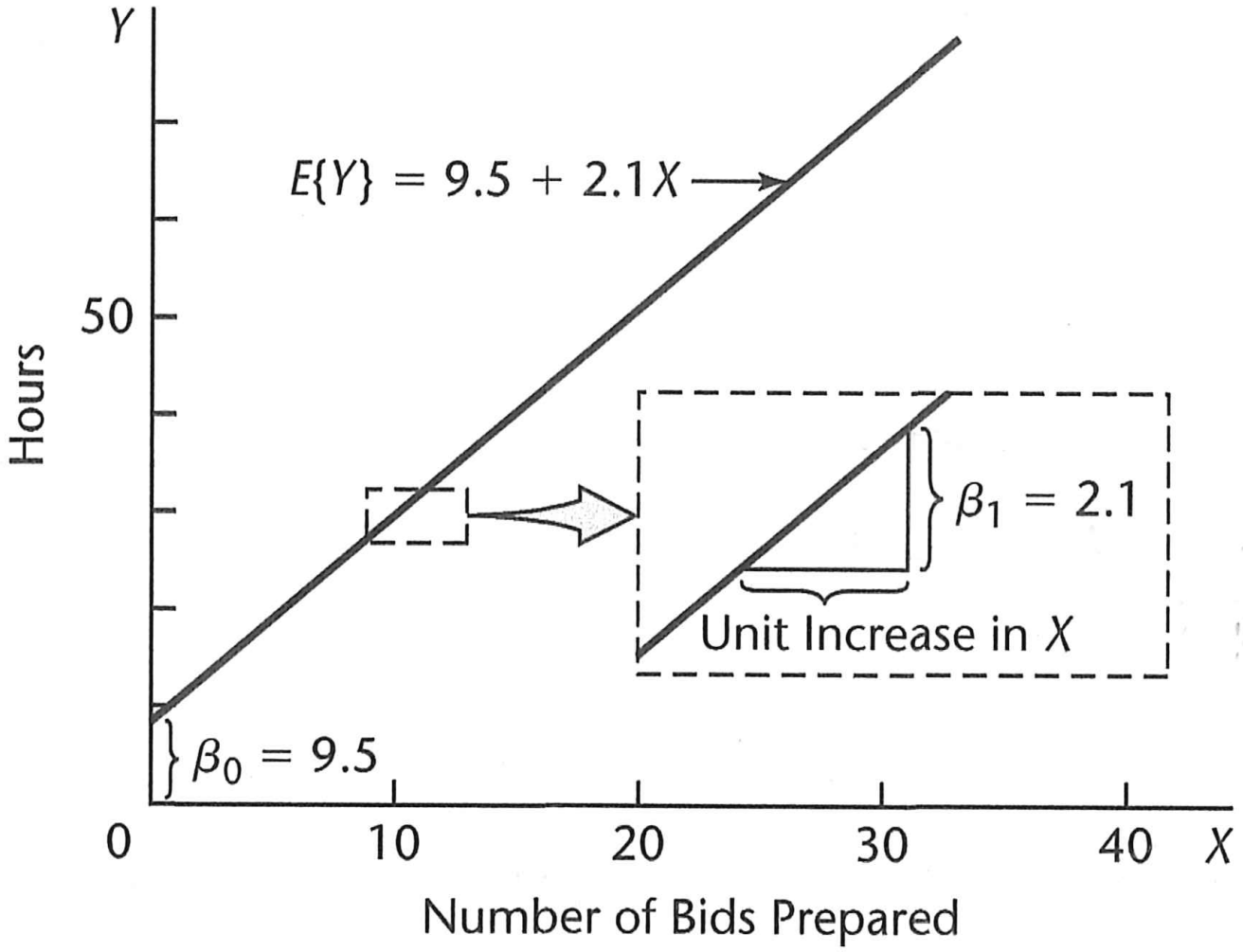


Figure 2: Meaning of the SLR Parameters

- Observational Data
  - difficult to establish cause-and-effect

- Experimental Data
  - randomization balances out the effects of any other variables that affect the response variable

Fitted Regression Line [R Code]

Using data from the Toluca Company, a refrigeration manufacturer with data collected on lot size (X) and work hours (Y), we plot Y on X as well as the estimated regression equation

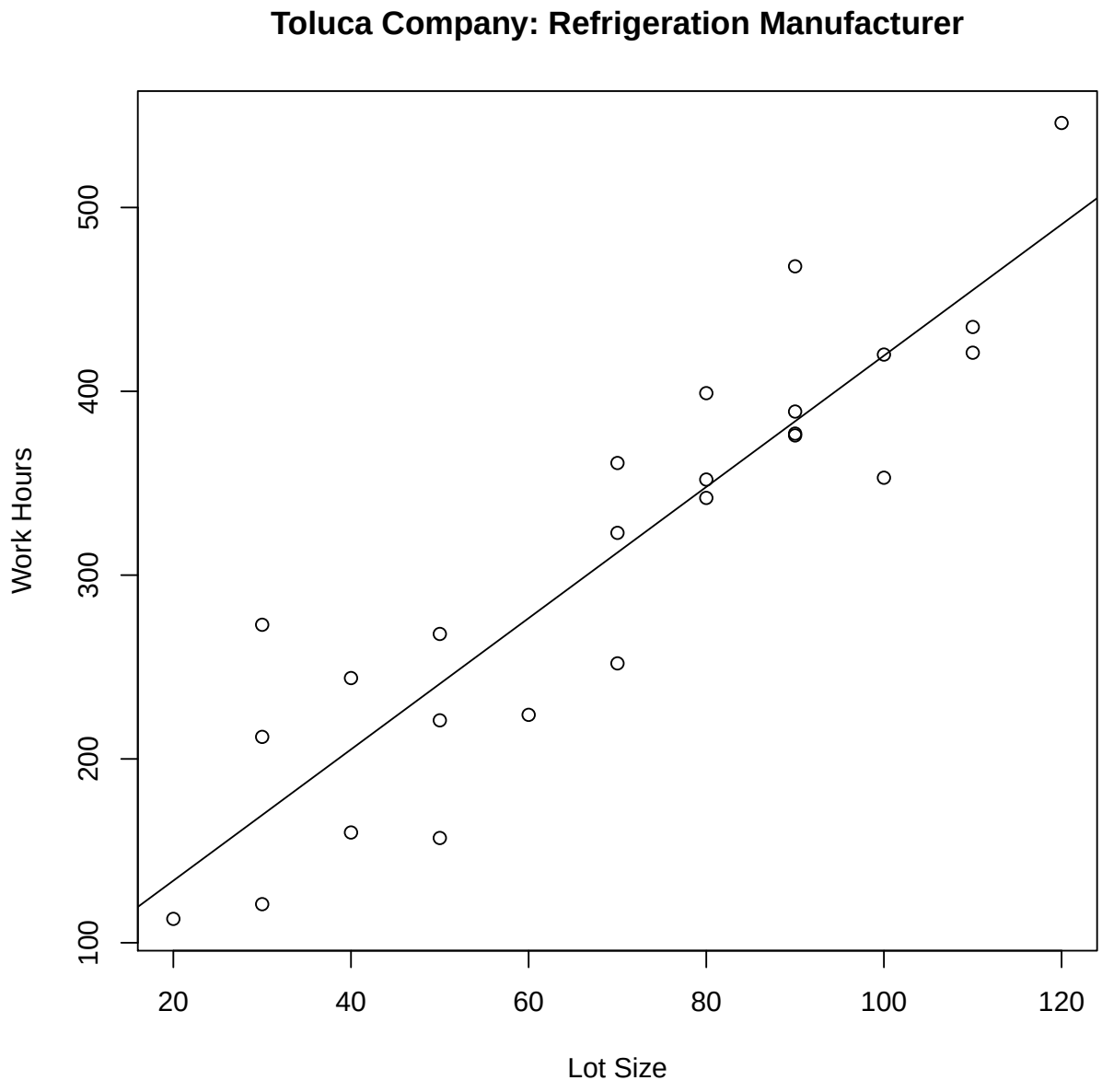
```
> toluca <- read.table("toluca.txt" , sep = "" , header = TRUE)

> lm_toluca <- lm(workHours ~ lotSize , data = toluca)

> plot(toluca$lotSize , toluca$workHours , xlab = "Lot Size" ,
   ylab = "Work Hours" , main = "Toluca Company: Refrigeration Manufacturer")

> abline(lm_toluca)
```

Fitted Regression Line Graph



Running the Regression Model [R Code]

```
> summary(lm_toluca)

Call:
lm(formula = workHours ~ lotSize, data = toluca)

Residuals:
    Min      1Q  Median      3Q     Max
-83.876 -34.088  -5.982  38.826 103.528

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.366     26.177   2.382   0.0259 *
lotSize        3.570      0.347  10.290 4.45e-10 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 48.82 on 23 degrees of freedom
Multiple R-squared:  0.8215, Adjusted R-squared:  0.8138
F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

Section 4

Defining the Ordinary Least Squares Estimates

How to Find Good Estimates of β_0 and β_1

- We are going to choose to estimate β_0 and β_1 such that we minimize the sum of the error terms squared.
  - Why choose to minimize the sum of the error terms squared?
- Recall, for the observations (X_i, Y_i)
  - Y_i is an observed (true) value at X_i
  - $E[Y_i] = \beta_0 + \beta_1 X_i$ what we *expect* Y_i to be at X_i
- The sum of the squared difference between these two terms

$$\sum_{i=1}^{n} (Y_i - E[Y_i])^2 \;=\; \sum_{i=1}^{n} (Y_i - [\beta_0 + \beta_1 X_i])^2$$

  is what we seek to minimize

- Define Q as the the sum of squared differences between Y_i and $E[Y_i]$

$$Q = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2 \tag{4}$$

- We will employ an analytical (calculus-based) approach to determine estimates of β_0 and β_1, b_0 and b_1 respectively, that minimize Q, by taking partial derivates of Q with respect to β_0 and β_1 we obtain

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \tag{5}$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) \tag{6}$$

n.b. Don't forget to verify second-order conditions (second-order partial derivates) verify that we indeed finding a minimum, i.e., we need

$$\frac{\partial^2 Q}{\partial {\beta_0}^2} = 2 > 0$$

and

$$\frac{\partial^2 Q}{\partial {\beta_1}^2} = 2{X_i}^2 > 0$$

to ensure we are finding a minimum

- Set first-order partial derivatives equal to zero and simplify, using b_0 and b_1 to denote the particular values of β_0 and β_1 that minimize Q

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0$$

$$\sum_{i=1}^{n} X_i(Y_i - b_0 - b_1 X_i) = 0$$

- Expanding and rearranging the terms in the above equations, we obtain the **normal equations**

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_i = 0 \tag{7}$$

$$\sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 = 0 \tag{8}$$

- We now need to solve these two equations simultaneously
- Divide

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_i = 0 \tag{7}$$

by n to obtain

$$\frac{1}{n} \sum_{i=1}^{n} Y_i = b_0 + \frac{b_1}{n} \sum_{i=1}^{n} X_i = 0$$

$$\bar{Y} = b_0 + b_1 \bar{X}$$

and solving for b_0 we obtain

$$b_0 = \bar{Y} - b_1 \bar{X} \tag{9}$$

- Now we have an easy way to solve for b_0 once we obtain a value for b_1

- To solve for b_1, substitute the value for b_0 in (9) into (8)

$$\sum_{i=1}^{n} X_i(Y_i - (\bar{Y} - b_1\bar{X}) - X_i b_1) = 0$$

$$\sum_{i=1}^{n} X_i(Y_i - \bar{Y} + b_1\bar{X} - X_i b_1) = 0$$

$$\sum_{i=1}^{n} X_i(Y_i - \bar{Y}) = b_1 \left[\sum_{i=1}^{n} X_i(X_i - \bar{X})\right]$$

$$\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y} = b_1 \left[\sum_{i=1}^{n} X_i^2 - n(\bar{X})^2\right]$$

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = b_1 \left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]$$

$$\text{therefore, provided } \sum_{i=1}^{n}(X_i - \bar{X})^2 > 0$$

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{s_{xy}}{s_{xx}} \qquad (10)$$

- *where* s_{xy} and s_x are the sample covariance between x and y and the sample standard deviation of x respectively
- proofs that
  - $\sum_{i=1}^{n} X_i y_i - n\bar{X}\bar{Y} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$
  - $\sum_{i=1}^{n} X_i^2 - n(\bar{X})^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$

  can be found in Wooldridge 5e, Appendix A.1, equations [A.7] and [A.8]

How to Estimate β_0 and β_1 - Wooldridge Method

- Define the Population Regression Function (PRF)

$$E[y|x] = \beta_0 + \beta_1 x \tag{11}$$

- Firstly, we obtain a random sample of size n from a population

$$\{(x_i, y_i) : i = 1, \ldots, n\}$$

  - you can visualize this as two columns of matched (paired) data, x and y, with n rows, and i serving as a row index
- As this data is obtained from (2), $y = \beta_0 + \beta_1 x + \varepsilon$, we can write

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{12}$$

How to Estimate β_0 and β_1 - Wooldridge Method cont'd

$$Cov(x, \varepsilon) = 0 \tag{13}$$

recall $Cov(a, b) = E[(a - \mu_a)(b - \mu_b)] = E[(a - \mu_a)b]$
$= E[a(b - \mu_b)] = E[ab] - E[a]E[b] = E[ab] - \mu_a\mu_b$

Proof.

$$Cov(x, \varepsilon) = E[x\varepsilon] - E[x]E[\varepsilon]$$

x is constant and can be removed from the expectation

$$Cov(x, \varepsilon) = xE[\varepsilon] - E[x]E[\varepsilon]$$

recalling the assumption that the mean error is zero, $E[\varepsilon] = 0$

$$Cov(x, \varepsilon) = x \times 0 - E[x] \times 0 = 0$$

□

How to Estimate β_0 and β_1 - Wooldridge Method cont'd

- Recall (2) $y = \beta_0 + \beta_1 x + \varepsilon$
- Solve for ε in (2) and substitute into
  1. $E[\varepsilon] = 0$

  $\therefore \; E[y - \beta_0 - x\beta_1] = 0$

  2. $Cov(x, \varepsilon) = E[x\varepsilon] - E[x]E[\varepsilon] = E[x\varepsilon] = 0$

  $\therefore \qquad\qquad\qquad\qquad\qquad\qquad = E[x(y - \beta_0 - x\beta_1)] = 0$
- In sum, we now have two additional equations

$$E[y - \beta_0 - x\beta_1] = 0 \tag{14}$$

$$E[x(y - \beta_0 - x\beta_1)] = 0 \tag{15}$$

How to Estimate β_0 and β_1 - Wooldridge Method cont'd

- Choose estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ to solve the **sample** counterparts of β_0 and β_1 respectively, using the method of moments approach to estimation
  - see Wooldridge 5*e*, Appendix C.4 for details
- Replace population moments in (14) and (15) with sample counterparts to obtain

$$\frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1)}{n} = 0 \tag{16}$$

$$\frac{\sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1)}{n} = 0 \tag{17}$$

How to Estimate β_0 and β_1 - Wooldridge Method cont'd

- Summing over n (for the n rows of sample data we collected from the population), we obtain for (16)

$$\frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - x_i\hat{\beta}_1)}{n} = 0 \tag{16}$$

$$\frac{\sum_{i=1}^{n} y_i}{n} - \frac{\sum_{i=1}^{n} \hat{\beta}_0}{n} - \frac{\sum_{i=1}^{n}(x_i\hat{\beta}_1)}{n} = 0$$

$$\bar{y} - \hat{\beta}_0 - \hat{\beta}_1\bar{x} = 0 \tag{18}$$

- Algebraically rearranging to solve for $\hat{\beta}_0$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \tag{19}$$

- We now have an easy way to compute $\hat{\beta}_0$ once we solve for $\hat{\beta}_1$
  - ...the remainder is identical to the previous derivation

Ordinary Least Squares Estimates of β_0 and β_1

- In sum, the ordinary least squares (OLS) estimates for β_0 and β_1 are

$$b_0 = \bar{Y} - b_1 \bar{X} \tag{9}$$

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{s_{xy}}{s_{xx}} \tag{10}$$

- Wooldridge 5*e* may occasionally have some atypical mathematical notation. We will deviate from this notation when convenient.

1. So far we have deviated from the textbook by labeling our error term ε, where the textbook has labeled it u
2. For ease of exposition, we will also refer to the OLS estimates for for β_0 and β_1 as b_0 and b_1, as opposed to the textbook's representation of them as $\hat{\beta}_0$ and $\hat{\beta}_1$

Textbook Notation	MSAN 601 Notation
u	ε
$\hat{u}$	e
$\hat{\beta}_0$	b_0
$\hat{\beta}_1$	b_1

Table 2: Notational Deviations from Wooldridge 5*e*

Theorem

- *Under the conditions of the regression model* (2)*, the least-squares estimators are unbiased and have a minimum variance among all unbiased linear estimators*

$$E\{b_0\} = \beta_0 \tag{20}$$

$$E\{b_1\} = \beta_1 \tag{21}$$

- *The estimators b_0 and b_1 are more precise, i.e., their sampling distributions are less variable, than any other estimators belonging to the class of unbiased estimators that are linear functions of the observations $Y_1, \ldots, Y_n$*

Proof.

Wooldridge 5e or Kutner 4e/5e

Theorem

- *Under the Gauss-Markov Assumptions, the ordinary least-squares (OLS) estimators are BLUE*
  - B *Best: deals with the* **efficiency** *and* **consistency** *of estimators*
  - L *Linear*
  - U *Unbiased*
  - E *Estimators*
- *Implication: there are no better other linear estimators which we can apply to our sample which provide us with estimates of the population parameters*

Proof.

Wooldridge 5e, Kutner 4e/5e or online □

1. The population process must be linear in the *parameters*, e.g., $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
2. Data from the population, $\{X_i, Y_i\}$, are from a random sample
3. No perfect multicollinearity between independent variables
4. Zero Conditional Mean
   - $E(\varepsilon_i | X_{i1}, ..., X_{ik}) = 0$
   - The expected value of the error is zero, and does not depend on any of the regressors nor on their values
5. Homoscedasticity
   - $V(\varepsilon_i | X_{i1}, ..., X_{ik}) = \sigma^2$
   - The distribution of the variance of the error terms does not depend on any of the regressors or their values
6. No serial correlation, i.e., $cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall \ i \neq j$

- Given sample estimators b_0 and b_1 of the parameters in the regression function $E\{Y\} = \beta_0 + \beta_1 X$, we estimate the regression function as follows

$$\hat{Y} = b_0 + b_1 X \tag{22}$$

  where $\hat{Y}$ is the estimated regression function at the level X of the predictor variable
- We call a **value** of the response variable a **response** and $E\{Y\}$ the **mean response**
- The mean response is the mean of the probability distribution of Y corresponding to a value of the predictor variable X
- $\hat{Y}$ is therefore a point estimate of the mean response corresponding to a value of the predictor variable X

- It can be shown as an extension of the Gauss-Markov theorem that $\hat{Y}$ is an unbiased estimator of $E\{Y\}$, with minimum variance in the class of unbiased linear estimators
- For specific cases (individual data), we will call $\hat{Y}_i$

$$\hat{Y}_i = b_0 + b_1 X_i \qquad i = 1, \ldots, n \tag{23}$$

  the **fitted value** for the i^{th} case
- There is a fitted value for each observation in the sample
- Be careful not to conflate the *fitted value* $\hat{Y}_i$ with the *observed value* Y_i

- There are similarly residuals for each of the n observations in the sample, where a residual is defined as

$$e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 X_i \tag{24}$$

where the i^{th} residual is the difference between the observed value Y_i and the corresponding fitted value $\hat{Y}_i$

n.b. residuals are not the same as errors

- the regression model error term value $\varepsilon_i = Y_i - E\{Y_i\}$
  - this is vertical deviation of Y_i from the unknown true regression line, and is hence unknown (cannot be calculated)
- the residual $e_i = Y_i - \hat{Y}_i$ is the vertical deviation of Y_i from the fitted value $\hat{Y}_i$ on the estimated regression line, and it is known (can be calculated)

A Graphical Example of Residuals

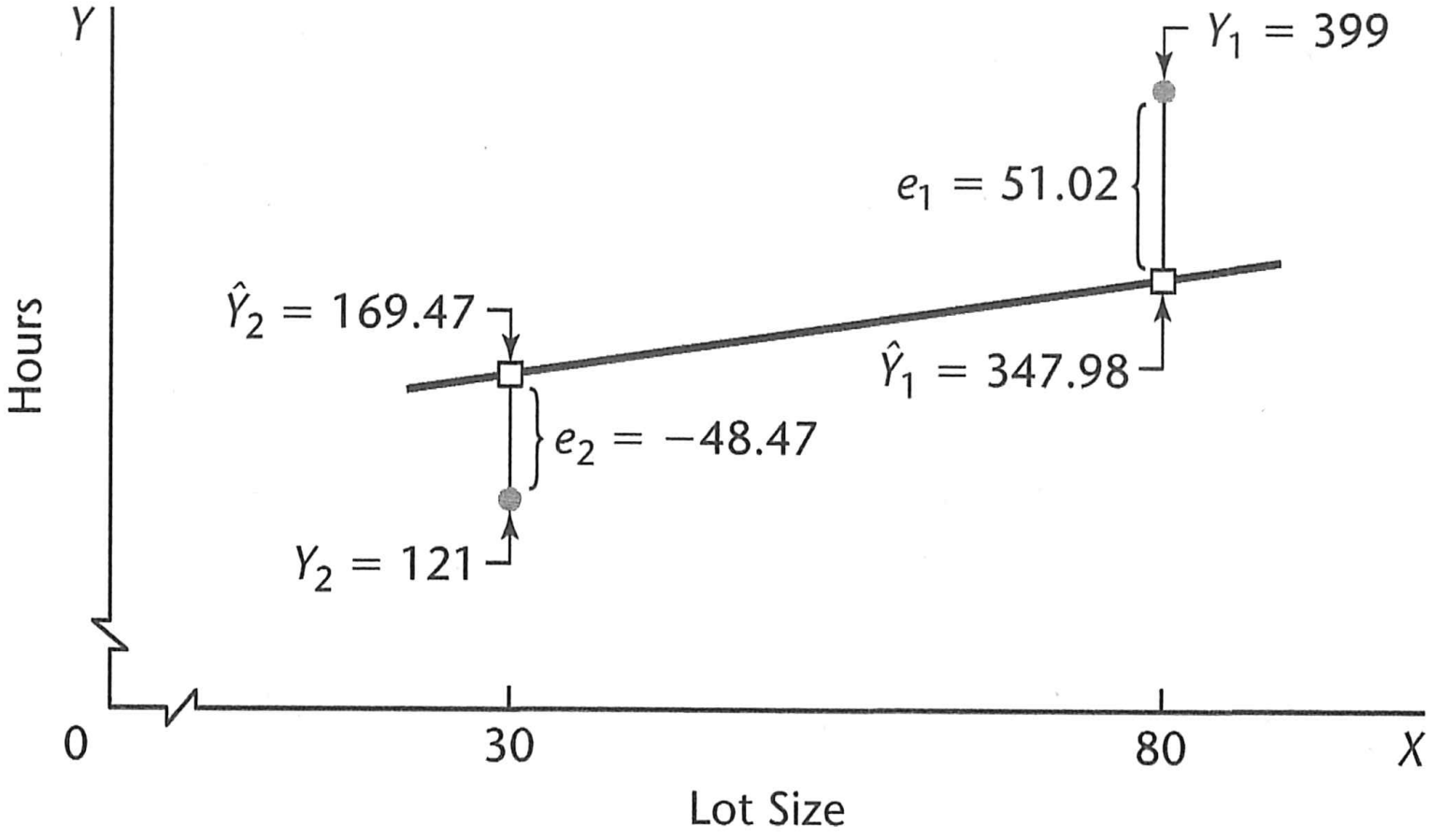


Figure 3: A Graphical Example of Residuals

Using data from the Toluca Company, a refrigeration manufacturer with data collected on lot size (X) and work hours (Y), we plot Y on X as well as the estimated regression equation

```
> residuals(lm_toluca)

          1           2           3           4           5           6  ...
 51.0179798 -48.4719192 -19.8759596  -7.6840404  48.7200000 -52.5779798  ...
```

```
> fitted(lm_toluca)

       1        2        3        4        5        6        7        8  ...
347.9820 169.4719 240.8760 383.6840 312.2800 276.5780 490.7901 347.9820  ...
```

```
> toluca$workHours

 [1] 399 121 221 376 361 224 546 352 353 157 160 252 389 113 435 420 212
[18] 268 377 421 273 468 244 342 323
```

Section 5

Properties of The Fitted Regression Line

Properties of The Fitted Regression Line

1. $$\sum_{i=1}^{n} e_i = 0 \tag{25}$$

   - The sum—and by association, the sample average—of OLS residuals is zero
   - No proof needed. We **chose** our OLS estimates, b_0 and b_1, such that the residuals sum to zero for *any data set*

2. $$\sum_{i=1}^{n} {e_i}^2 \quad \text{is a minimum} \tag{26}$$

   - This was a requirement we imposed for the derivation of the least squares estimators of the regression parameters in (5) and (6)

3. $$s_{xe} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(e_i - \bar{e})}{(n-1)} = 0 \qquad (27)$$

Proof.

$$s_{xe} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(e_i - \bar{e})}{(n-1)}$$

We know $\bar{e} = 0$ from (25), substitute and expand

$$s_{xe} = \frac{\sum_{i=1}^{n} X_i e_i - \bar{X}\sum_{i=1}^{n} e_i}{(n-1)}$$

Similarly, we know that $\sum_{i=1}^{n} e_i = 0$ from (25)

$$s_{xe} = \frac{\sum_{i=1}^{n} X_i e_i}{(n-1)}$$

We previously defined $e_i = Y_i - b_0 - b_1 X_i$ in (24), substitute

$$s_{xe} = \frac{\sum_{i=1}^{n} X_i (Y_i - b_0 - b_1 X_i)}{(n-1)}$$

From (17), we know the numerator equals zero, $\therefore s_{xe} = 0$ □

4. The point $(\bar{x}, \bar{y})$ is always on the OLS regression line.

Proof.

This is exactly what (9) and (19) demonstrates $\square$

- Writing each y_i as the sum of its fitted value and its residual

$$Y_i = \hat{Y}_i + e_i \tag{28}$$

we make the following additional observation

5. $$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i \tag{29}$$

Proof.
Summing (28) we obtain
$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i + \sum_{i=1}^{n} e_i$
We know $\sum_{i=1}^{n} e_i = 0$ from (25)
$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$

□

n.b. An implication of this property is that $\bar{Y} = \bar{\hat{Y}}$

6

$$s_{\hat{Y}e} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})(e_i - \bar{e})}{(n-1)} = 0 \tag{30}$$

Proof.

$s_{\hat{y}e} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})(e_i - \bar{e})/(n-1)$
We know $\bar{e} = 0$ from (25), substitute and expand
$s_{\hat{Y}e} = \left[\sum_{i=1}^{n} \hat{Y}_i e_i - \bar{\hat{Y}} \sum_{i=1}^{n} e_i\right] /(n-1)$
Similarly, we know that $\sum_{i=1}^{n} e_i = 0$ from (25)
$s_{\hat{Y}e} = \sum_{i=1}^{n} \hat{Y}_i e_i/(n-1)$
We previously defined $\hat{Y}_i = b_0 + b_1 X_i$ in (23), substitute
$s_{\hat{Y}e} = \sum_{i=1}^{n}(b_0 + b_1 X_i)e_i/(n-1)$
Expand $s_{\hat{Y}e} = [b_0 \sum_{i=1}^{n} e_i + b_1 \sum_{i=1}^{n} X_i e_i] /(n-1)$
We know $\sum_{i=1}^{n} e_i = 0$ from (25) and that $\sum_{i=1}^{n} x_i e_i = 0$ from Proof of (27), the numerator therefore equals zero, $\therefore s_{\hat{Y}e} = 0$ □

Section 6

Estimation of Error Terms Variance σ^2

Point Estimator of σ^2

- We know that variance σ^2 of a single population is estimated by the sample variance s^2
- To obtain s^2, we square the deviation of an observation Y_i from the estimated mean $\bar{Y}$, $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$, called the sum of squares, which is divided by the degrees of freedom, $n-1$ (where one degree of freedom is lost in estimating the population mean μ by $\bar{Y}$) to obtain the sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$$

  which is an unbiased estimator of the variance σ^2
- The sample variance s^2 is often called a *mean square*, as the sum of squares has been divided by the appropriate degrees of freedom

- Recall from (3), for a regression model, that the variance associated with each observation Y_i is σ^2, which is that same for each error term ε_i
- We now need to calculate a sum of squares, recognizing that that the Y_i now come from different probability distributions with different means that depend upon the level X_i, therefore the deviation of an observation must be calculated around its own estimated mean
- These deviations are the residuals $e_i = Y_i - \hat{Y}_i$, and the sum of squares, denoted **error sum of squares** (SSE) is

$$SSE \equiv \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 \tag{31}$$

n.b. This may also be referred to as residual sum of squares

- The error sum of squares (SSE) has $n-2$ degrees of freedom, as two degrees of freedom are lost by having to estimate β_0 and β_1 with b_0 and b_1 in obtaining the means $\hat{Y}_i$. Hence the appropriate mean square is

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} {e_i}^2}{n-2} \quad (32)$$

  where MSE stands for **error mean square** or residuals mean square

- MSE is an unbiased estimator of σ^2 for the regression model (2), i.e., $E\{MSE\} = \sigma^2$
- An estimator of the standard deviation σ is $s = \sqrt{MSE}$

Manual Calculation of $\sqrt{MSE}$ [R Code]

```
> sqrt(sum((residuals(lm_toluca))^2)/23) # where n=25, therefore n-2=23
[1] 48.82331 # this is s = sqrt(MSE)

> summary(lm_toluca)

Call:
lm(formula = workHours ~ lotSize, data = toluca)

Residuals:
    Min      1Q  Median      3Q     Max
-83.876 -34.088  -5.982  38.826 103.528

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.366     26.177   2.382   0.0259 *
lotSize        3.570      0.347  10.290 4.45e-10 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 48.82 on 23 degrees of freedom
Multiple R-squared:  0.8215, Adjusted R-squared:  0.8138
F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

Section 7

Normal Error Regression Model

- No matter the form of the distribution of the error terms ε_i (and hence of the Y_i), the least squares method provides unbiased point estimators of β_0 and β_1 that have minimum variance among all unbiased linear estimators.
- However, to establish interval estimates and perform tests, we are required to make an assumption about the form of the distribution of the ε_i
- We therefore assume that the error terms ε_i are normally distributed, which greatly simplifies the theory of regression analysis and is justifiable in many empirical settings

Defining the **Normal** Error Regression Model

- The graph of the regression function is called the regression line (or curve)
- The equation that accounts for statistical inexactitude of the relationship in a regression model is as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{33}$$

where

- Y_i is the value of the response variable in the i^{th} trial
- β_0 and β_1 are parameters: β_0 is the y-intercept and β_1 is the slope of the regression function
- X_i is a known constant, the value of the predictor variable
- ε_i are independent $\sim N(0, \sigma^2)$
- $i = 1, \ldots, n$

Notes on the Normal Error Regression Model

1. $\sim N(0, \sigma^2)$ is read as distributed normally with a mean of 0 and a variance of σ^2

Notes on the Normal Error Regression Model

1. $\sim N(0, \sigma^2)$ is read as distributed normally with a mean of 0 and a variance of σ^2
2. The normal error regression model (33) is the same as regression model (2) save for the assumption in (33) of the error terms being normally distributed

Notes on the Normal Error Regression Model

1. $\sim N(0, \sigma^2)$ is read as distributed normally with a mean of 0 and a variance of σ^2
2. The normal error regression model (33) is the same as regression model (2) save for the assumption in (33) of the error terms being normally distributed
3. Given the normal error regression model (33) assumption of normal errors, the assumption of $\sigma[\varepsilon_i, \varepsilon_j] = 0 \ \forall \ i, j; i \neq j$ in (2) now becomes one of independence, i.e., $\varepsilon_i \sim N(0, \sigma^2) iid$
    - The implication here is that the outcome in any one trial has no effect on the error term for any other trial, positive or negative, small or large

Notes on the Normal Error Regression Model

1. $\sim N(0, \sigma^2)$ is read as distributed normally with a mean of 0 and a variance of σ^2
2. The normal error regression model (33) is the same as regression model (2) save for the assumption in (33) of the error terms being normally distributed
3. Given the normal error regression model (33) assumption of normal errors, the assumption of $\sigma[\varepsilon_i, \varepsilon_j] = 0 \ \forall \ i, j; i \neq j$ in (2) now becomes one of independence, i.e., $\varepsilon_i \sim N(0, \sigma^2) iid$
   - The implication here is that the outcome in any one trial has no effect on the error term for any other trial, positive or negative, small or large
4. There are various valid arguments to justify the assumption of normally distributed error terms, the most pragmatic of which is the use of inferential procedures that are based on the t distribution are usually only sensitive to large departures from normality (particularly with respect to skewness), actual confidence coefficients and risks of errors will be close to the levels for exact normality

From Correlation to Independence of the ε_i's

Did you observe

Before we made the assumption of normally distributed errors terms, we assumed $\sigma[\varepsilon_i, \varepsilon_j] = 0 \ \forall \ i, j; i \neq j$

After we made the assumption of normally distributed errors terms, we now claim that the $\varepsilon_i \sim N(0, \sigma^2)$ *iid*, where *iid* is read as *independent and identically distributed*

n.b. The independence assumption is **stronger** than the assumption of $\sigma[\varepsilon_i, \varepsilon_j] = 0$, so how did this happen? The covariance assumption implies independence **if** we assume normally distributed error terms.

Let's examine two ε_i's, and denote them X and Y. Assuming the error terms are $\varepsilon_i \sim N(0, \sigma^2)$, we have the following normal probability density functions (pdf's) for each

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{34}$$

and

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2}\right) \tag{35}$$

[1] Thanks to Harry O'Reilly (MSAN '16) for this proof

From Correlation to Independence of the ε_i's cont'd

We also know the trivial statistical result that if a joint density is equal to the product of the individual densities, i.e.,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \tag{36}$$

this is the definition of independence. Therefore if we can prove the joint density is the product of the individual densities, we will have proven independence. Taking the product of (34) and (35) we obtain

$$f_X(x)f_Y(y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{37}$$

[1] Thanks to Harry O'Reilly (MSAN '16) for this proof

From Correlation to Independence of the ε_i's cont'd

We now consider the bivariate normal probability density function

$$f_{X,Y}(x,y) = \frac{1}{2\pi}|\mathbf{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \qquad (38)$$

where

- $\mathbf{x}^T = \begin{bmatrix} x & y \end{bmatrix}$, the random vector
- $\boldsymbol{\mu}^T = \begin{bmatrix} 0 & 0 \end{bmatrix}$, the mean vector
- $\mathbf{\Sigma} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$, the covariance matrix

n.b. The off-diagonal terms of the covariance matrix are zero because we have assumed uncorrelated error terms

[1] Thanks to Harry O'Reilly (MSAN '16) for this proof

Expanding and simplifying the bivariate normal pdf

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{1}{2}\begin{bmatrix} x & y \end{bmatrix}\begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}\right) \quad (39)$$

$$= \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \quad (40)$$

demonstrating that the bivariate normal pdf (40) is equal to the product of two individual pdf's (37).

Given our assumption of uncorrelated error terms, we conclude, based on the assumption of normally distributed error terms, that the error terms are also independent.

[1] Thanks to Harry O'Reilly (MSAN '16) for this proof

Section 8

Inferences Concerning β_1

- We are frequently interested in drawing inferences about β_1, the slope of the regression line in the normal error regression model

$$Y_i = \beta_0 + \beta_1 X_i \tag{41}$$

- Of particular interest is a test to evaluate whether or not $\beta_1 = 0$, i.e., to examine whether or not there is any linear association between the dependent variable Y and the independent variable X
- These hypothesis tests take the form

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

- If it is concluded that β_1 is not statistically significantly different from zero, the implication is

$$E\{Y\} = \beta_0 + (0)X = \beta_0$$

Normal Error Regression Model when $\beta_1 = 0$

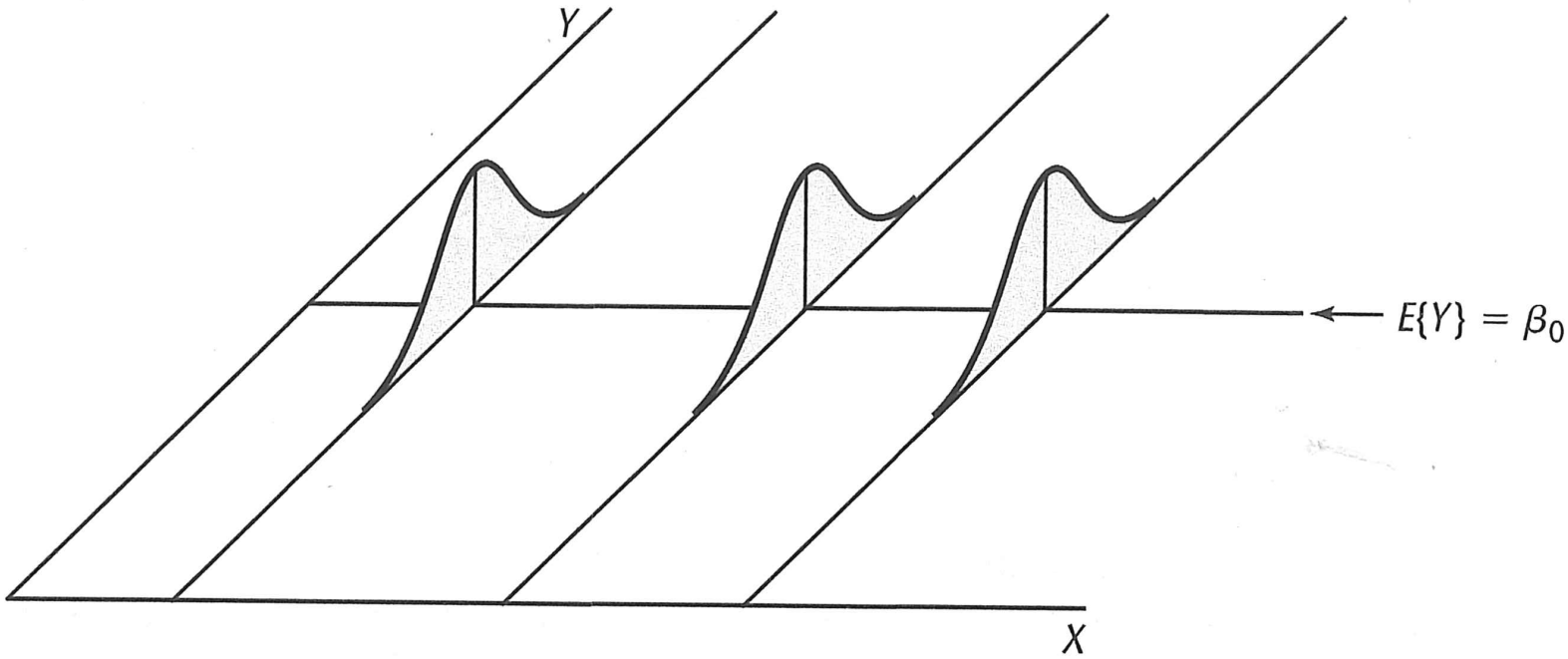


Figure 4: Normal Error Regression Model when $\beta_1 = 0$

Normal Error Regression Model when $\beta_1 = 0$ cont'd

- The condition that $\beta_1 = 0$ implies something even stronger than that of no association between X and Y

Normal Error Regression Model when $\beta_1 = 0$ cont'd

- The condition that $\beta_1 = 0$ implies something even stronger than that of no association between X and Y
- Since all probability distributions of Y are normal with constant variance σ^2, the condition of $\beta_1 = 0$ implies that all probability distributions at any level of X are identical, i.e., the have the same mean and variance

n.b. Without the condition $\beta_1 = 0$, all probability distributions of Y were normal with variance σ^2, but their means were dependent on the value of X, i.e., they were not all the same

- Recall that the point estimator b_1 is

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{42}$$

- The sampling distribution of b_1 refers to the different values of b_1 that would be obtained with repeated sampling when the levels of the predictor variable X are held constant from sample to sample
- For the normal error regression model (41), the sampling distribution of b_1 is normal, with mean and variance

$$E\{b_1\} = \beta_1 \tag{43}$$

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{44}$$

Estimated Variance of b_1: $s^2\{b_1\}$

We can estimate the variance of the sampling distribution of b_1:

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

by replacing the parameter σ^2 with MSE, the unbiased estimator of σ^2

$$s^2\{b_1\} = \frac{MSE}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- We can express b_1 as a linear combination of the Y_i

$$b_1 = \sum_{i=1}^{n} k_i^{(\beta_1)} Y_i \tag{45}$$

where

$$k_i^{(\beta_1)} = \frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{46}$$

n.b. The $k_i^{(\beta_1)}$ are a function of the X_i, which are constants, hence our conclusion that b_1 is a linear combination of the Y_i

Proof.

$$b_1 = \sum_{i=1}^{n} k_i^{(\beta_1)} Y_i = \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} Y_i(X_i - \bar{X}) - \bar{Y}\sum_{i=1}^{n}(X_i - \bar{X})$$

$$= \sum_{i=1}^{n} Y_i(X_i - \bar{X}) - \bar{Y}(n\bar{X} - n\bar{X})$$

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} Y_i(X_i - \bar{X}) \tag{47}$$

$$b_1 = \sum_{i=1}^{n} k_i^{(\beta_1)} Y_i = \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- **Property 1**

$$\sum_{i=1}^{n} k_i^{(\beta_1)} = 0 \tag{48}$$

Proof.

We need only examine the numerator of (46)

$\sum_{i=1}^{n}(X_i - \bar{X})$

$\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \bar{X}$

$n\bar{X} - n\bar{X} = 0$

$\therefore \sum_{i=1}^{n} k_i^{(\beta_1)} = 0$ □

- **Property 2**

$$\sum_{i=1}^{n} X_i k_i^{(\beta_1)} = 1 \tag{49}$$

Proof.

Expand both the numerator and denominator

$$\begin{aligned}
\sum_{i=1}^{n} X_i k_i^{(\beta_1)} = \frac{\sum_{i=1}^{n} X_i(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} &= \frac{\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i \bar{X}}{\sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n} X_i \bar{X} + \sum_{i=1}^{n} \bar{X}^2} \\
&= \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{\sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + n\bar{X}^2} \\
&= \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} \\
&= 1
\end{aligned}$$

□

- **Property 3**

$$\sum_{i=1}^{n} {k_i}^{2(\beta_1)} = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{50}$$

Proof.

$$\sum_{i=1}^{n} {k_i}^{2(\beta_1)} = \sum_{i=1}^{n} \left[\frac{(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right]^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2}$$

$$= \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

□

Normality Given b_1 is a linear combination of the Y_i (see 45), and the Y_i are independently, normally distributed according to (41), such a linear combination results in a normally distributed random variable.

- This is formally stated in Appendix (A.40) in Kutner et al. 4e/5e, and a proof can be found in any basic mathematical statistics book

Mean b_1 is an unbiased point estimator of β_1

Proof.

$$E\{b_1\} = E\left\{\sum_{i=1}^{n} k_i^{(\beta_1)} Y_i\right\} = \sum_{i=1}^{n} k_i^{(\beta_1)} E\{Y_i\} = \sum_{i=1}^{n} k_i^{(\beta_1)} E\{\beta_0 + \beta_1 X_i\}$$

$$= \beta_0 \sum_{i=1}^{n} k_i^{(\beta_1)} + \beta_1 \sum_{i=1}^{n} k_i^{(\beta_1)} X_i$$

$$= \beta_1$$

from (48) and (49) □

- A proof that b_1 has a minimum variance among all unbiased linear estimators can be found in Kutner et al. 4e/5e in §2.1

Variance

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{51}$$

Proof.

$$\begin{aligned}
\sigma^2\{b_1\} = \sigma^2\left\{\sum_{i=1}^{n} k_i{}^{(\beta_1)} Y_i\right\} &= \sum_{i=1}^{n} k_i{}^{2(\beta_1)} \sigma^2\{Y_i\} \\
&= \sum_{i=1}^{n} k_i{}^{2(\beta_1)} \sigma^2 \\
&= \sigma^2 \sum_{i=1}^{n} k_i{}^{2(\beta_1)} \\
&= \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}
\end{aligned}$$

from (50)

Sampling Distribution of $(b_1 - \beta_1)/s\{b_1\}$

- Given b_1 is normally distributed, we know that the standardized statistic $(b_1 - \beta_1)/\sigma\{b_1\}$ is a standard normal variable
- We need to estimate $\sigma\{b_1\}$ by $s\{b_1\}$, and are therefore interested in the distribution of the statistic $(b_1 - \beta_1)/s\{b_1\}$

n.b. When a statistic is standardized but the denominator is an estimated standard deviation rather than the true standard deviation, it is called a *studentized statistic*

- It can be shown that the studentized statistic

$$(b_1 - \beta_1)/s\{b_1\} \sim t_{(n-2)} \tag{52}$$

for regression model (41)

  - see Kutner et al. 4e/5e §2.1 for proof

- We lose 2 degrees of freedom because we estimate β_0 and β_1 with b_0 and b_1

- As the t distribution is symmetric around its mean of 0, it follows that

$$t_{(\alpha/2;n-2)} = -t_{(1-\alpha/2;n-2)}$$

- Since $(b_1 - \beta_1)/s\{b_1\}$ follows a t distribution, we can make the following statement

$$P\{b_1 - t_{(1-\alpha/2;n-2)}s\{b_1\} \leq \beta_1 \leq P\{b_1 + t_{(1-\alpha/2;n-2)}s\{b_1\}\} = 1-\alpha$$

which holds for all possible values of β_1, the $1-\alpha$ confidence limits for β_1 are

$$b_1 \pm t_{(1-\alpha/2;n-2)}s\{b_1\} \tag{53}$$

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

$$t^* = \frac{b_1}{s\{b_1\}}$$

If $|t^*| \leq t_{(1-\alpha/2;n-2)}$, do not reject H_0

If $|t^*| > t_{(1-\alpha/2;n-2)}$, reject H_0

- In Kutner et al. 4e/5e, they say "conclude H_0" and "conclude H_a", which some statisticians may find offensive

$$H_0 : \beta_1 \leq 0$$
$$H_a : \beta_1 > 0$$

$$t^* = \frac{b_1}{s\{b_1\}}$$

If $t^* \leq t_{(1-\alpha;n-2)}$, do not reject H_0

If $t^* > t_{(1-\alpha;n-2)}$, reject H_0

n.b. Type I error here is $1 - \alpha$ for the ones sided test, as opposed to $1 - \alpha/2$ for the two-sided test

Testing β_1 for a Non-Zero Value

$$H_0 : \beta_1 = \delta$$
$$H_a : \beta_1 \neq \delta$$

$$t^* = \frac{b_1 - \delta}{s\{b_1\}}$$

If $|t^*| \leq t_{(1-\alpha/2;n-2)}$, do not reject H_0

If $|t^*| > t_{(1-\alpha/2;n-2)}$, reject H_0

n.b. When computing t^*, note the numerator now accounts for δ

Hypothesis Test for $H_0 : \beta_1 = 0$ [R Code]

```
> qt(.975,23) # inverse cumulative probability t-distribution function
[1] 2.068658

> 3.570/0.347 # b_1/s{b_1} = t value
[1] 10.28818

> summary(lm_toluca)
Call:
lm(formula = workHours ~ lotSize, data = toluca)

Residuals:
    Min      1Q  Median      3Q     Max
-83.876 -34.088  -5.982  38.826 103.528

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.366     26.177   2.382   0.0259 *
lotSize        3.570      0.347  10.290 4.45e-10 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 48.82 on 23 degrees of freedom
Multiple R-squared:  0.8215, Adjusted R-squared:  0.8138
F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

Confidence Interval for β_1 [R Code]

```
> 3.57 + qt(.975,23)*0.347 # b_1 + t_(0.975,23)*s{b_1}
[1] 4.287824


> 3.57 - qt(.975,23)*0.347 # b_1 - t_(0.975,23)*s{b_1}
[1] 2.852176

# can change alpha using toggle <level=0.**> where ** = 1-alpha
# in <confint> command, default is alpha = 0.05 <=> <level=0.95>
> confint(lm_toluca)
                2.5 %     97.5 %
(Intercept) 8.213711 116.518006
lotSize     2.852435   4.287969


# this includes the estimated coefficients in the output
> cbind(Estimate=coef(lm_toluca),confint(lm_toluca))
             Estimate    2.5 %     97.5 %
(Intercept) 62.365859 8.213711 116.518006
lotSize      3.570202 2.852435   4.287969
```

Section 9

Inferences Concerning β_0

Sampling Distribution of b_0

- Recall that the point estimator b_0 is

$$b_0 = \bar{Y} - b_1 \bar{X} \tag{54}$$

- The sampling distribution of b_0 refers to the different values of b_0 that would be obtained with repeated sampling when the levels of the predictor variable X are held constant from sample to sample
- For the normal error regression model (41), the sampling distribution of b_0 is normal, with mean and variance

$$E\{b_0\} = \beta_0 \tag{55}$$

$$\sigma^2\{b_0\} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right] \tag{56}$$

Estimated Variance of b_0: $s^2\{b_0\}$

We can estimate the variance of the sampling distribution of b_0:

$$\sigma^2\{b_0\} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right]$$

by replacing the parameter σ^2 with MSE, the unbiased estimator of σ^2

$$s^2\{b_0\} = MSE \left[\frac{1}{n} + \frac{\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right]$$

- We can express b_0 as a linear combination of the Y_i

$$b_0 = \sum_{i=1}^{n} k_i^{(\beta_0)} Y_i \tag{57}$$

where

$$k_i^{(\beta_0)} = \frac{1}{n} - \frac{\bar{X}(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{58}$$

n.b. The $k_i^{(\beta_0)}$ are a function of the X_i, which are constants, hence our conclusion that b_0 is a linear combination of the Y_i

Proof.

$$b_0 = \sum_{i=1}^{n} k_i^{(\beta_0)} Y_i = \frac{\sum_{i=1}^{n} Y_i}{n} - \frac{\bar{X} \sum_{i=1}^{n} Y_i (X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Recall from (47): $\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} Y_i (X_i - \bar{X})$

$$b_0 = \sum_{i=1}^{n} k_i^{(\beta_0)} Y_i = \bar{Y} - \bar{X} \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$b_0 = \sum_{i=1}^{n} k_i^{(\beta_0)} Y_i = \bar{Y} - \bar{X} b_1$$

- **Property 1**

$$\sum_{i=1}^{n} k_i^{(\beta_0)} = 1 \tag{59}$$

Proof.

$$\begin{aligned}
\sum_{i=1}^{n} k_i^{(\beta_0)} &= \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{\bar{X}(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \right] \\
&= \sum_{i=1}^{n} \frac{1}{n} - \frac{\bar{X} \sum_{i=1}^{n}(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
&= 1 - \frac{\bar{X}(n\bar{X} - n\bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\
\sum_{i=1}^{n} k_i^{(\beta_0)} &= 1
\end{aligned}$$

- **Property 2**

$$\sum_{i=1}^{n} X_i k_i^{(\beta_0)} = 0 \tag{60}$$

Proof.

$$\begin{aligned}
\sum_{i=1}^{n} X_i k_i^{(\beta_0)} &= \sum_{i=1}^{n} \left[\frac{X_i}{n} - \frac{\bar{X} X_i (X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right] \\
&= \sum_{i=1}^{n} \frac{X_i}{n} - \bar{X} \frac{\sum_{i=1}^{n} X_i (X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \\
&= \bar{X} - \bar{X} \frac{\sum_{i=1}^{n} X_i (X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})} \\
&= \bar{X} - \bar{X} \frac{\sum_{i=1}^{n} X_i (X_i - \bar{X})}{\sum_{i=1}^{n} X_i (X_i - \bar{X}) - \bar{X} \sum_{i=1}^{n} (X_i - \bar{X})} \\
&= \bar{X} - \bar{X} \frac{\sum_{i=1}^{n} X_i (X_i - \bar{X})}{\sum_{i=1}^{n} X_i (X_i - \bar{X}) - \bar{X} (n\bar{X} - n\bar{X})} \\
&= \bar{X} - \bar{X} \\
&= 0
\end{aligned}$$

□

Normality Given b_0 is a linear combination of the Y_i, and the Y_i are independently, normally distributed according to (41), such a linear combination results in a normally distributed random variable.

- This is formally stated in Appendix (A.40) in Kutner et al. 4e/5e, and a proof can be found in any basic mathematical statistics book

Mean b_0 is an unbiased point estimator of β_0

Proof.

$$E\{b_0\} = E\left\{\sum_{i=1}^{n} k_i^{(\beta_0)} Y_i\right\} = \sum_{i=1}^{n} k_i^{(\beta_0)} E\{Y_i\} = \sum_{i=1}^{n} k_i^{(\beta_0)} E\{\beta_0 + \beta_1 X_i\}$$

$$= \beta_0 \sum_{i=1}^{n} k_i^{(\beta_0)} + \beta_1 \sum_{i=1}^{n} k_i^{(\beta_0)} X_i$$

$$= \beta_0$$

from (59) and (60) □

Variance

$$\sigma^2\{b_0\} = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] \tag{61}$$

Proof.

$$\sigma^2\{b_0\} = \sigma^2\left\{\sum_{i=1}^{n} k_i{}^{(\beta_0)} Y_i\right\} = \sum_{i=1}^{n} k_i{}^{2(\beta_0)}\sigma^2\{Y_i\}$$

$$= \sum_{i=1}^{n} k_i{}^{2(\beta_0)}\sigma^2$$

$$= \sigma^2 \sum_{i=1}^{n} k_i{}^{2(\beta_0)} = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

□

n.b. The algebra in the final equality of the proof is left for the student to explore

Sampling Distribution of $(b_0 - \beta_0)/s\{b_0\}$

- Given b_0 is normally distributed, we know that the standardized statistic $(b_0 - \beta_0)/\sigma\{b_0\}$ is a standard normal variable
- We need to estimate $\sigma\{b_0\}$ by $s\{b_0\}$, and are therefore interested in the distribution of the statistic $(b_0 - \beta_0)/s\{b_0\}$

$$(b_0 - \beta_0)/s\{b_0\} \sim t_{(n-2)} \tag{62}$$

for regression model (41)

- We lose 2 degrees of freedom because we estimate β_0 and β_1 with b_0 and b_1

- As the t distribution is symmetric around its mean of 0, it follows that

$$t_{(\alpha/2;n-2)} = -t_{(1-\alpha/2;n-2)}$$

- Since $(b_0 - \beta_0)/s\{b_0\}$ follows a t distribution, we can make the following statement

$$P\{b_0 - t_{(1-\alpha/2;n-2)}s\{b_0\} \leq \beta_0 \leq P\{b_0 + t_{(1-\alpha/2;n-2)}s\{b_0\}\} = 1-\alpha$$

which holds for all possible values of β_0, the $1-\alpha$ confidence limits for β_0 are

$$b_0 \pm t_{(1-\alpha/2;n-2)}s\{b_0\} \tag{63}$$

Considerations for Inferences on β_0 and β_1

1. Departures from Normality
   - If the probability distributions of Y are not exactly normal but do not depart seriously, the sampling distributions of b_0 and b_1 will be approximately normal, and the use of the t distribution will provide approximately the specified confidence coefficient or level of significance

Considerations for Inferences on β_0 and β_1

1. Departures from Normality
   - If the probability distributions of Y are not exactly normal but do not depart seriously, the sampling distributions of b_0 and b_1 will be approximately normal, and the use of the t distribution will provide approximately the specified confidence coefficient or level of significance
2. Spacing of the X Levels
   - The variances of b_0 and b_1, $\sigma^2\{b_0\}$ and $\sigma^2\{b_1\}$ respectively, are affected by the spacing of the X levels in the data (for given n and σ)
   - As the spacing in the X levels grows, $\sum_{i=1}^{n}(X_i - \bar{X})^2$ grows, shrinking the variance

Section 10

Interval Estimation of $E\{Y_h\}$

- A common objective in regression analysis is to estimate the mean for one or more probability distributions of Y
- Let X_h denote the level X for which we wish to estimate the mean response
  - X_h may be a value which occurred in the sample, or it may be some other value of the predictor variable *within the scope of the model*
- The mean response when $X = X_h$ is denoted by $E\{Y_h\}$
- The point estimator $\hat{Y}_h$ of $E\{Y_h\}$ is

$$\hat{Y}_h = b_0 + b_1 X_h \tag{64}$$

Sampling Distribution of $\hat{Y}_h$

- The sampling distribution of $\hat{Y}_h$ refers to the different values of $\hat{Y}_h$ that would be obtained if repeated samples were selected, each holding the levels of the predictor variable X constant, and calculating $\hat{Y}_h$ for each sample
- For the normal error regression model (41), the sampling distribution of $\hat{Y}_h$ is normal, with mean and variance

$$E\{\hat{Y}_h\} = E\{Y_h\} \tag{65}$$

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] \tag{66}$$

Estimated Variance of $\hat{Y}_h$: $s^2\{\hat{Y}_h\}$

We can estimate the variance of the sampling distribution of $\hat{Y}_h$:

$$\sigma^2\{\hat{Y}_h\} = \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

by replacing the parameter σ^2 with MSE, the unbiased estimator of σ^2

$$s^2\{\hat{Y}_h\} = MSE \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

Normality Given $\hat{Y}_h$ is a linear combination of b_0 and b_1, and both b_0 and b_1 are linear combinations of the Y_i, and the $\hat{Y}_h$ are therefore normally distributed

Mean $\hat{Y}_h$ is an unbiased point estimator of $E\{Y_h\}$

Proof.

$$E\{\hat{Y}_h\} = E[b_0 + b_1 X_h] = E\{b_0\} + X_h E\{b_1\} = \beta_0 + \beta_1 X_h \quad (67)$$

□

Variance

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] \tag{68}$$

- This is a two-step proof: first we show that $\bar{Y}$ and b_1 are uncorrelated, then we obtain the variance of $\hat{Y}_h$

Lemma

Observing that $\bar{Y} = \sum \frac{1}{n} Y_i$ *and that* $b_1 = \sum k_i Y_i$, *we compute*

$$\sigma\{\bar{Y}, b_1\} = \sigma\{\frac{1}{n} Y_i, k_i Y_i\} = \sum \frac{1}{n} k_i \sigma^2\{Y_i\} = \frac{\sigma^2}{n} \sum k_i = 0$$

where $\sum k_i^{(\beta_1)} = 0$ from (48).

n.b. The proof of

$$\sigma\left\{\sum a_i Y_i, c_i Y_i\right\} = \sum a_i c_i \sigma^2\{Y_i\}$$

for independent Y_i, can be found in Kutner et al. 4e/5e Appendix A.3 Equation (A.32)

Variance

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

Proof.

$$\begin{aligned}
\sigma^2\{\hat{Y}_h\} &= \sigma^2\{\bar{Y} + b_1(X_h - \bar{X})\} \\
&= \sigma^2\{\bar{Y}\} + (X_h - \bar{X})^2\sigma^2\{b_1\} \\
&\textit{from } (44) \\
&= \frac{\sigma^2\{Y_i\}}{n} + (X_h - \bar{X})^2\left[\frac{\sigma^2}{\sum(X_i - \bar{X})^2}\right] \\
&= \frac{\sigma^2}{n} + (X_h - \bar{X})^2\left[\frac{\sigma^2}{\sum(X_i - \bar{X})^2}\right] \\
&= \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]
\end{aligned}$$

n.b. The proof that $\hat{Y}_h = \bar{Y} + b_1(X_h - \bar{X})$ can be found in Kutner et al. 4e/5e, Chapter 1, Equations (1.5) and (1.15)

Sampling Distribution of $(\hat{Y}_h - E\{Y_h\})/s\{\hat{Y}_h\}$

$$\frac{\hat{Y}_h - E\{Y_h\}}{s\{\hat{Y}_h\}} \sim t_{(n-2)} \tag{69}$$

for regression model (41)

$$\hat{Y}_h \pm t_{(1-\alpha/2;n-2)} s\{\hat{Y}_h\} \tag{70}$$

Final Notes on $E\{Y_h\}$

1. Since the X_i are known constants in regression model (41), the interpretation of confidence intervals and risks of error in inferences on the mean response is in terms of taking repeated samples in which the X observations are at the same levels in the actual study
2. The variance of $\hat{Y}_h$ is smallest when $X_h = \bar{X}$

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$$

3. The confidence limits are for a mean response $E\{Y_h\}$ are not sensitive to moderate departures from the assumption of normally distributed error terms, and if the sample size is sufficiently large, the limits are not even sensitive to large departures from normality
4. Confidence limits apply when a **single** mean response is to be estimated

Confidence Interval Estimation for $\hat{Y}_h$ [R Code]

```
> predict(lm_toluca,data.frame(lotSize=33),interval="confidence")

       fit     lwr     upr
1 180.1825 146.816 213.549

# use <level=0.xx> to set alpha, default is alpha = 0.05

# may choose to generate multiple CIs as follows

> predict(lm_toluca,data.frame(lotSize=c(33,99)),interval="confidence")
       fit      lwr      upr
1 180.1825 146.8160 213.5490
2 415.8159 386.8106 444.8211

# *** the X_h value(s) passed to predict must be a data frame and MUST have
# the same name as the predictor variable in your lm object, e.g.,
# in the toluca example, the data frame must be named lotSize

# don't forget the <interval="confidence"> toggle
```

Section 11

Prediction of a New Observation

Prediction of A New Observation

- We now consider the prediction of a new **observation** Y corresponding to a given level of X of the predictor variable
- The new observation on Y to be predicted is viewed as the result of a new trial, independent of the trials on which the regression analysis is based
- We denote the level of X for the trial as X_h and the new observation on Y as $Y_{h(new)}$
- We assume (41) applies

What's the Difference between $\hat{Y}_h$ and $\hat{Y}_{h(new)}$?

1. $\hat{Y}_{h(new)}$, the estimate of $Y_{h(new)}$, is the prediction of an **individual outcome** drawn from the distribution of Y

2. $\hat{Y}_h$, the estimate of $E\{Y_h\}$, is an estimate of the **mean** of the distribution of Y

Prediction of $Y_{h(new)}$

- Although the mean of the distribution of Y is, as usual, estimated by $\hat{Y}_h$, and the variance of the distribution of Y estimated by MSE, we must take care to observe the nuance of predicting individual outcomes

Prediction of $Y_{h(new)}$

- Although the mean of the distribution of Y is, as usual, estimated by $\hat{Y}_h$, and the variance of the distribution of Y estimated by MSE, we must take care to observe the nuance of predicting individual outcomes
- When computing a confidence interval for $E\{Y_h\}$ using (72), we obtain an upper limit and lower limit, implying that the mean of the distribution of Y, and by association, the distribution itself, can be located as far left or as far right as depicted in Figure 5

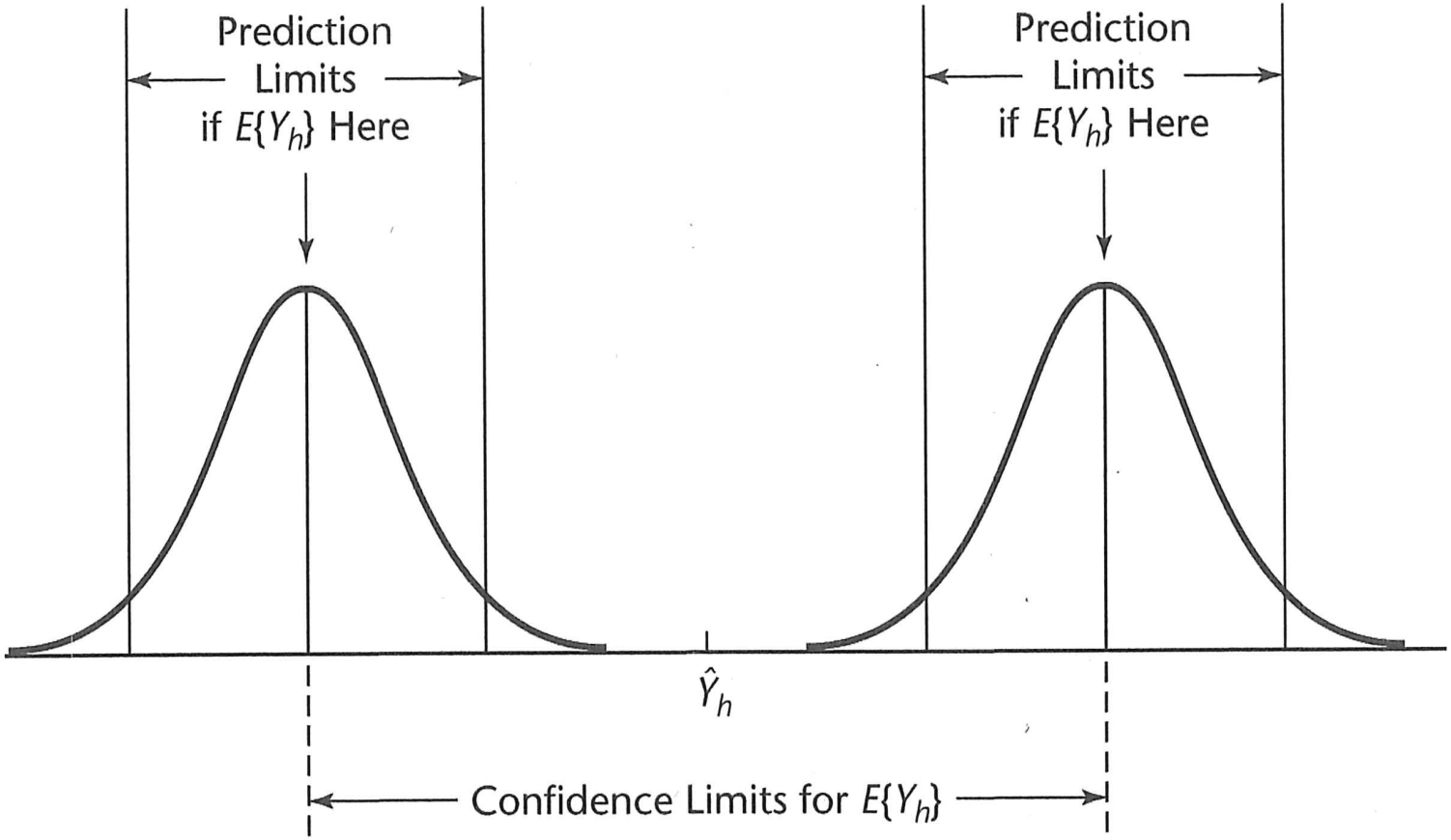


Figure 5: Prediction of $Y_{h(new)}$

- Although the mean of the distribution of Y is, as usual, estimated by $\hat{Y}_h$, and the variance of the distribution of Y estimated by MSE, we must take care to observe the nuance of predicting individual outcomes
- When computing a confidence interval for $E\{Y_h\}$ using (72), we obtain an upper limit and lower limit, implying that the mean of the distribution of Y, and by association, the distribution itself, can be located as far left or as far right as depicted in Figure 5
- Since we cannot be certain of the location of the distribution of Y, prediction limits for $Y_{h(new)}$ must account for both
  1. Variation in possible location of the distribution of Y
  2. Variation within the probability distribution of Y

Sampling Distribution of $(Y_{h(new)} - \hat{Y}_h)/s\{pred\}$

$$\frac{Y_{h(new)} - \hat{Y}_h}{s\{pred\}} \sim t_{(n-2)} \tag{71}$$

for regression model (41)

$$\hat{Y}_h \pm t_{(1-\alpha/2;n-2)} s\{pred\} \tag{72}$$

Variance: $\sigma^2\{pred\}$

- We denote the variance of the prediction error by $\sigma^2\{pred\}$, defined as the variance of the difference between $Y_{h(new)}$ and $\hat{Y}_h$

$$\sigma^2\{pred\} = \sigma^2\{Y_{h(new)} - \hat{Y}_h\}$$

- From Kutner et al. 4e/5e, Appendix A.3, Equation (A.31b), we know for independent Y_i

$$\sigma^2\{Y_1 - Y_2\} = \sigma^2\{Y_1\} + \sigma^2\{Y_2\}$$

therefore

$$\begin{aligned} \sigma^2\{pred\} &= \sigma^2\{Y_{h(new)} - \hat{Y}_h\} \\ &= \sigma^2\{Y_{h(new)}\} + \sigma^2\{\hat{Y}_h\} \\ \sigma^2\{pred\} &= \sigma^2 + \sigma^2\{\hat{Y}_h\} \qquad (73) \\ & \qquad (74) \end{aligned}$$

where $\sigma^2\{\hat{Y}_h\}$ is given by (68)

- Substituting (68) into (73), we obtain

$$\sigma^2\{pred\} = \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$
$$= \sigma^2\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

- An unbiased estimator of $\sigma^2\{pred\}$ is

$$s^2\{pred\} = MSE + s^2\{\hat{Y}_h\}$$
$$= MSE\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right] \quad (75)$$

Final Notes on Prediction Intervals

1. Equation (75) indicates that the prediction intervals wider the further X_h is from $\bar{X}$. The reason for this is that the estimate of the mean $\hat{Y}_h$ is less precise as X_h is located further from $\bar{X}$
2. The prediction limits—unlike confidence limits—for a mean response $E\{Y_h\}$ are sensitive to departures from normality—diagnostic measures will follow in subsequent chapters to address this potential issue
3. Prediction and confidence intervals differ conceptually
   - A confidence interval is an inference on a **parameter**
   - A prediction interval is a statement about the value of a **random variable**

Prediction Interval Estimation for $\hat{Y}_{h(new)}$ [R Code]

```
> predict(lm_toluca,data.frame(lotSize=33),interval="prediction")

       fit      lwr      upr
1 180.1825 73.81494 286.5501

# use <level=0.xx> to set alpha, default is alpha = 0.05

# may choose to generate multiple PIs as follows

> predict(lm_toluca,data.frame(lotSize=c(33,99)),interval="prediction")
       fit       lwr      upr
1 180.1825  73.81494 286.5501
2 415.8159 310.73473 520.8970

# compared with CI calculations, PIs are wider

# *** the X_h value(s) passed to predict must be a data frame and MUST have
# the same name as the predictor variable in your lm object, e.g.,
# in the toluca example, the data frame must be named lotSize

# don't forget the <interval="prediction"> toggle
```

Section 12

Confidence Bands for Regression Lines

- Although we will not address the topic in this course, I encourage you to read about Working-Hotelling **Confidence Bands for Regression Lines**, which can be found in Kutner et al. 4e/5e in §2.6

Section 13

Analysis of Variance Approach to Regression Analysis

- The Analysis of Variance (ANOVA) approach is based on partitioning the sums of squares and degrees of freedom associated with the response variable Y

Total Sum of Squares (SSTO)

- The variance of a single population is estimated by the sample variance s^2

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} \tag{76}$$

where the numerator is called the sum of squares, i.e., the squared sum of deviations between observations (Y_i) and the mean ($\bar{Y}$)

- The denominator is divided by the degrees of freedom (dof), $n-1$, where one dof is lost by estimating μ_y, the true population mean, with $\bar{Y}$, the sample mean
- We will define the **Total Sum of Squares** ($SSTO$) as the numerator of (76)

$$SSTO \equiv \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \tag{77}$$

Error/Residual Sum of Squares (SSE)

- We will define deviations around the fitted regression line as the **Error or Residual Sum of Squares** (SSE)

$$SSE \equiv \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 \tag{78}$$

- if all Y_i observations fall on the fitted regression line $SSE = 0$
- the greater the variation of the Y_i observations around the fitted regression line, the greater SSE

Regression Sum of Squares (SSR)

- We will define deviations of fitted regression values around the mean as the **Regression Sum of Squares** (SSR)

$$SSR \equiv \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \tag{79}$$

- SSR may be considered a measure of that part of the variability of Y_i that is associated with the regression line
- The larger SSR is in relation to $SSTO$, the greater the effect of the regression relation in accounting for the total variation in the Y_i observations

Graphical Representation of Deviations

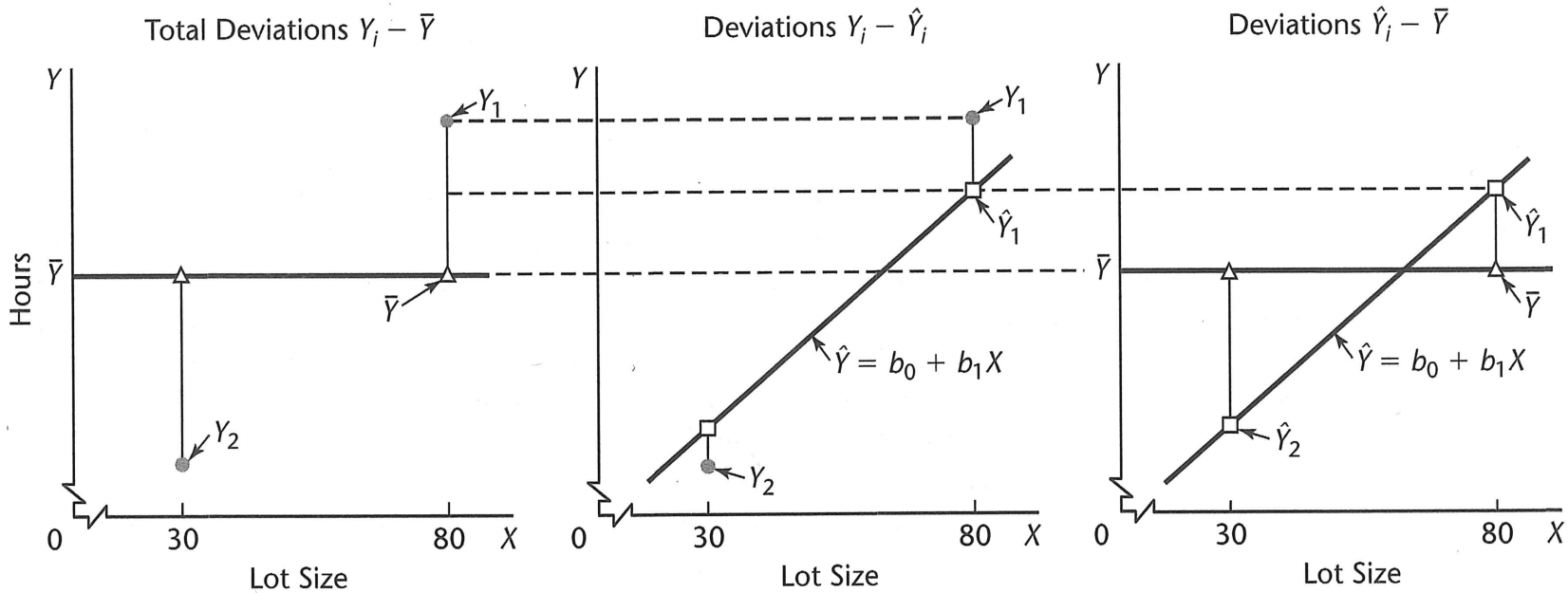


Figure 6: Squaring deviations, from left to right: SSTO, SSE and SSR

- Different textbooks use varying notation when referencing the different sums of square (SoS) terms

	Total SoS	Regression SoS	Residual SoS
	$\sum(Y_i - \bar{Y})^2$	$\sum(\hat{Y}_i - \bar{Y})^2$	$\sum(Y_i - \hat{Y}_i)^2$
Kutner 4e/5e	SSTO	SSR	SSE
Ramanathan 5e	TSS	RSS	ESS
Weisberg 4e	SYY	SSReg	RSS
Wooldridge 5e	SST	SSE	SSR

Table 3: A Note on Terminology

- In MSAN 601, we will use Kutner's 4e/5e notation
  - **n.b.** Wooldridge 5e uses the inverse of Kutner's 5e notation for *SSR* and *SSE*; take care not to confuse the two

Relationship Among Deviations

$$\underbrace{Y_i - \bar{Y}}_{\text{Total deviation}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\substack{\text{Deviation of fitted regression value around mean}}} + \underbrace{Y_i - \hat{Y}_i}_{\substack{\text{Deviation around fitted regression line}}}$$

$$SSTO = SSR + SSE \tag{80}$$

Proof.

$\sum_{i=1}^{n} \left[Y_i - \bar{Y}\right]^2 = \sum_{i=1}^{n} \left[Y_i - \bar{Y}\right]^2$

add and subtract $\hat{Y}_i$ from RHS

$\sum_{i=1}^{n} \left[Y_i - \bar{Y}\right]^2 = \sum_{i=1}^{n} \left[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{y})\right]^2$

recall $e_i = Y_i - \hat{Y}_i$ from (78) and replace in RHS

$\sum_{i=1}^{n} \left[Y_i - \bar{Y}\right]^2 = \sum_{i=1}^{n} \left[e_i + (\hat{Y}_i - \bar{Y})\right]^2$

expand RHS

$\sum_{i=1}^{n} \left[Y_i - \bar{Y}\right]^2 = \sum_{i=1}^{n} e_i^2 + 2\sum_{i=1}^{n} e_i(\hat{Y}_i - \bar{Y}) + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$

$SSE \equiv \sum_{i=1}^{n} e_i^2$, $SSR \equiv \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$,

$SSTO \equiv \sum_{i=1}^{n} \left[Y_i - \bar{Y}\right]^2$

$SSTO = SSE + 2\sum_{i=1}^{n} e_i(\hat{Y}_i - \bar{Y}) + SSR$

expand the remaining term

$SSTO = SSE + 2\sum_{i=1}^{n} e_i\hat{Y}_i - 2\bar{Y}\sum_{i=1}^{n} e_i + SSR$

recalll $\sum_{i=1}^{n} e_i = 0$, $\sum_{i=1}^{n} e_i\hat{Y}_i = 0$

$\therefore SSTO = SSE + SSR$ □

Breakdown of the Degrees of Freedom

1. $SSTO : (Y_i - \bar{Y})^2$
   - $n - 1$ *dof*
   - one degree of freedom lost estimating μ_Y with $\bar{Y}$
2. $SSE : (Y_i - \hat{Y}_i)^2$
   - $n - 2$ *dof*
   - two degrees of freedom lost estimating β_0,β_1 with b_0,b_1
3. $SSR : (\hat{Y}_i - \bar{Y})^2$
   - 1 *dof*
   - two degrees of freedom are associated with the regression line, corresponding to the slope and the intercept
   - one of those two degrees of freedom are lost estimating $\bar{Y}$

- A sum of squares divided by its associated degrees of freedom is called a *mean square*, abbreviated *MS*

1. $MSE = SSE/(n-2)$
2. $MSR = SSR/1$

n.b. Mean squares are **not** additive, i.e.,

$$\frac{SSTO}{n-1} \neq \frac{SSR}{1} + \frac{SSE}{n-2}$$

Source of Variation	SS	df	MS
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$	

Figure 7: SLR ANOVA Table

- In order to make inferences bases on the ANOVA approach, we need to know the expected value of each of the mean squares
  1. $E\{MSE\} = \sigma^2$
  2. $E\{MSR\} = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$
- Proofs of both expectations can be found in Kutner et al. 4e/5e, Chapter 2, §2.7, Equation (2.58)

Expected Mean Squares: Implications

1. The mean of the sampling distribution of MSE is σ^2, whether or not X and Y are linearly related, i.e., whether or not $\beta_1 = 0$

Expected Mean Squares: Implications

1. The mean of the sampling distribution of *MSE* is σ^2, whether or not X and Y are linearly related, i.e., whether or not $\beta_1 = 0$
2. The mean of the sampling distribution of *MSR* is also σ^2 when $\beta_1 = 0$; hence when $\beta_1 = 0$, the sampling distributions of *MSE* and *MSR* are located identically, and *MSE* and *MSR* are often of the same order of magnitude

Expected Mean Squares: Implications

1. The mean of the sampling distribution of *MSE* is σ^2, whether or not X and Y are linearly related, i.e., whether or not $\beta_1 = 0$
2. The mean of the sampling distribution of *MSR* is also σ^2 when $\beta_1 = 0$; hence when $\beta_1 = 0$, the sampling distributions of *MSE* and *MSR* are located identically, and *MSE* and *MSR* are often of the same order of magnitude
3. When $\beta_1 \neq 0$, the mean of the sampling distribution of *MSR* is greater since

$$\beta_1^2 \sum (X_i - \bar{X})^2$$

is a positive term

Expected Mean Squares: Implications

1. The mean of the sampling distribution of *MSE* is σ^2, whether or not X and Y are linearly related, i.e., whether or not $\beta_1 = 0$
2. The mean of the sampling distribution of *MSR* is also σ^2 when $\beta_1 = 0$; hence when $\beta_1 = 0$, the sampling distributions of *MSE* and *MSR* are located identically, and *MSE* and *MSR* are often of the same order of magnitude
3. When $\beta_1 \neq 0$, the mean of the sampling distribution of *MSR* is greater since

$$\beta_1^2 \sum (X_i - \bar{X})^2$$

is a positive term

- This suggests a comparison of *MSE* and *MSR* would be a good proxy to evaluate whether or not $\beta_1 = 0$
  1. if $MSR \approx MSE$, then we can assume $\beta_1 \approx 0$
  2. if $MSR >> MSE$, then we can assume $\beta_1 \neq 0$

A New Hypothesis Tests for β_1

- ANOVA provides us with a rich environment to test regression models
- In the SLR case, we can test for the following:

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

where the test statistic, denoted by F^*, compares MSR and MSE in the following fashion

$$F^* = \frac{MSR}{MSE}$$

where large values of F^* support H_a and values of F^* close to 1 support H_0

A New Hypothesis Tests for β_1

- ANOVA provides us with a rich environment to test regression models
- In the SLR case, we can test for the following:

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

where the test statistic, denoted by F^*, compares MSR and MSE in the following fashion

$$F^* = \frac{MSR}{MSE}$$

where large values of F^* support H_a and values of F^* close to 1 support H_0

n.b. **This is an upper-tail F-test**

- Employing *Cochran's Theorem*—see Kutner et al. 4e/5e Chapter 2, §2.7, (2.61)— we can establish that if H_0 holds, F^* follows the $F_{(1,n-2)}$ distribution
- Therefore, for the aforementioned hypothesis test, the decision rule is, when the risk of Type I error is to be controlled at α

$$\textit{If } F^* \leq F(1-\alpha; 1; n-2), \text{do not reject } H_0$$
$$\textit{If } F^* > F(1-\alpha; 1; n-2), \textit{reject } H_0$$

- The R-squared of the regression or the **Coefficient of Determination** is a measure of the goodness-of-fit of the regression line, $\hat{Y}_i$'s, to the Y_i's

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \tag{81}$$

- R^2 is the ratio of the proportionate reduction of total variation associated with the use of the predictor variable X
- Thus the larger R^2 is, the more total variation of Y is rereduced by introducing the predictor variable X
- It is interpreted as the fraction of the sample variation in Y that is explained by the variation in X

$$0 \leq R^2 \leq 1$$

- $R^2 = r_{xy}^2 = r_{y\hat{y}}^2$, where r is the correlation

The Limiting Values of R^2

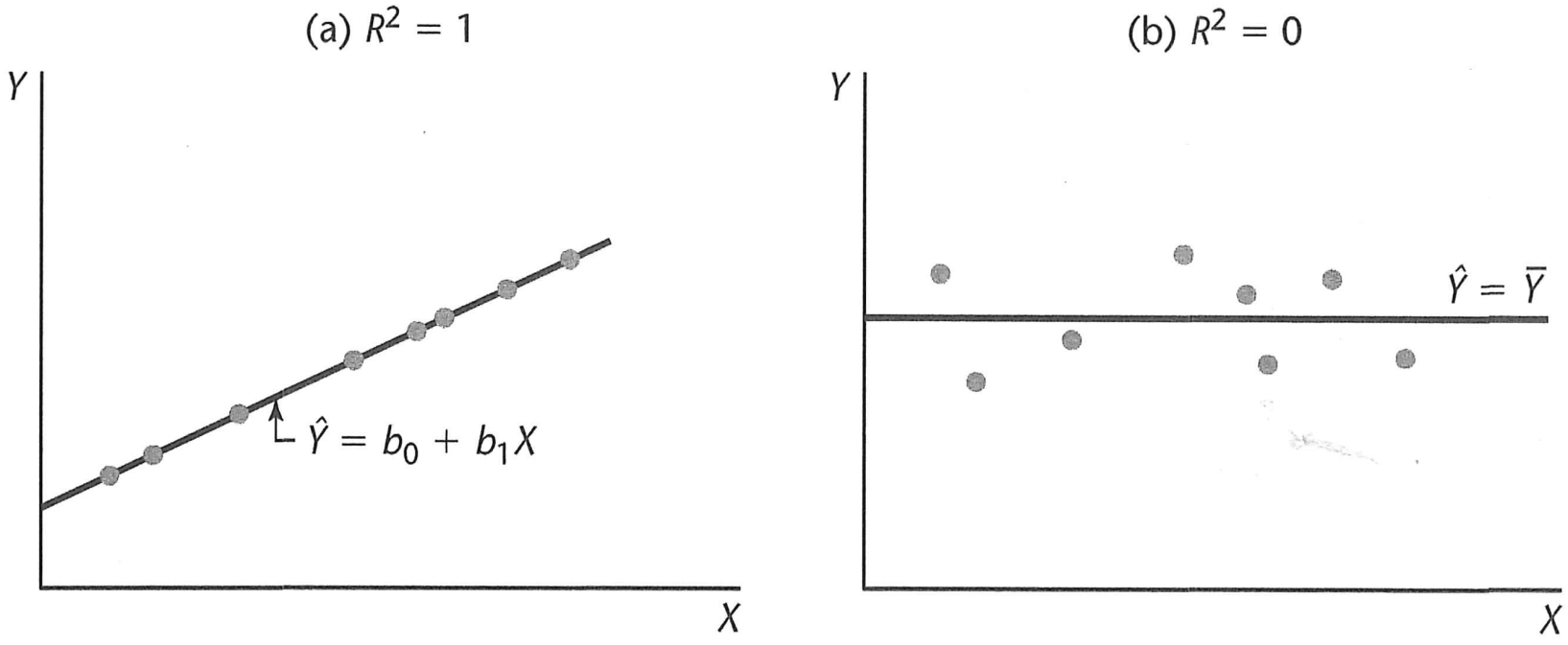


Figure 8: The Limiting Values of R^2

The Limitations of R^2

1. High R^2 **does not** imply that good predictions can be made

The Limitations of R^2

1. High R^2 **does not** imply that good predictions can be made
2. High R^2 **does not** imply that the estimated regression line is a good fit

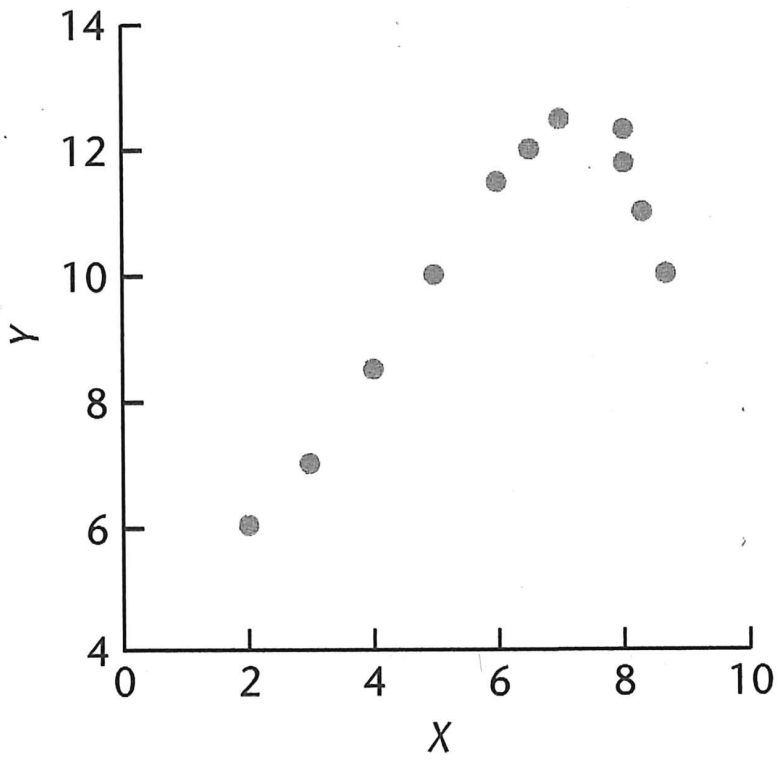
The Limitations of R^2

1. High R^2 **does not** imply that good predictions can be made
2. High R^2 **does not** imply that the estimated regression line is a good fit
3. $R^2 \approx 0$ **does not** imply that X and Y are not related

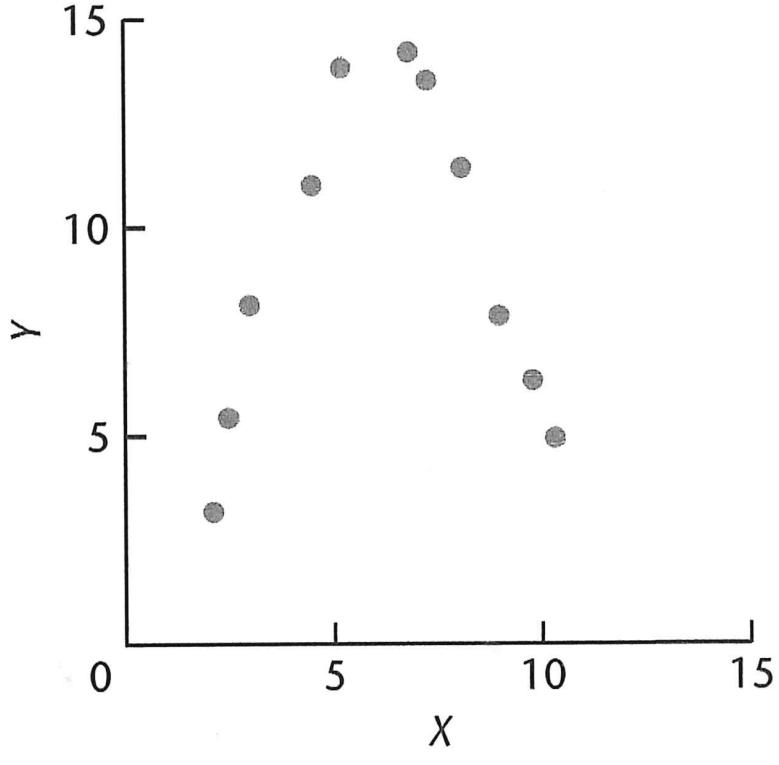
The Limitations of R^2

1. High R^2 **does not** imply that good predictions can be made
2. High R^2 **does not** imply that the estimated regression line is a good fit
3. $R^2 \approx 0$ **does not** imply that X and Y are not related

Scatter Plot with $R^2 = .69$
Linear regression is not a good fit



Scatter Plot with $R^2 = .02$
Strong relation between X and Y



Section 14

Running the Regression Model in R

Fitted Regression Line [R Code]

Using data from the Toluca Company, a refrigeration manufacturer with data collected on lot size (X) and work hours (Y), we plot Y on X as well as the estimated regression equation

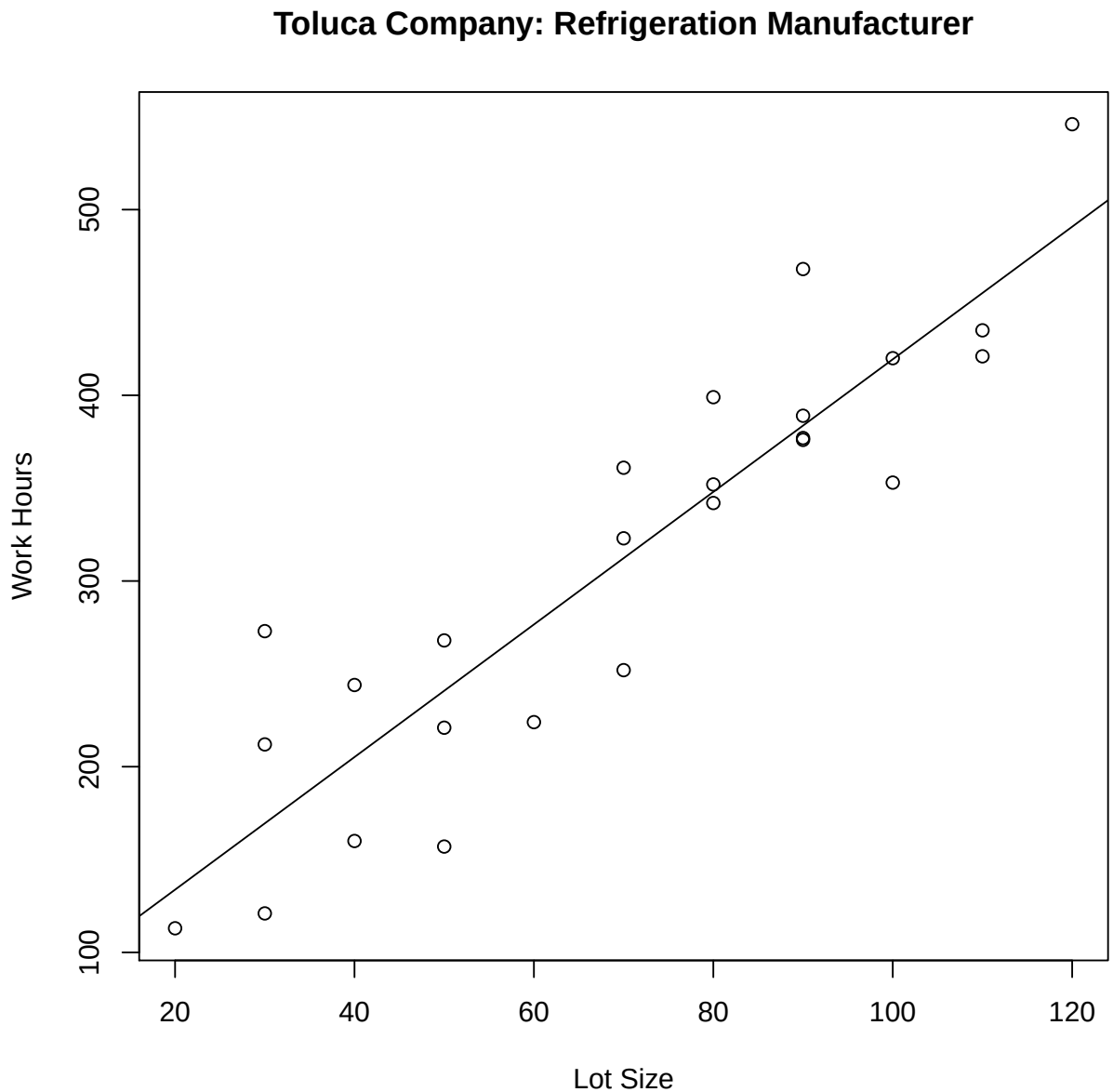
```
> toluca <- read.table("toluca.txt" , sep="" , header=TRUE)

> lm_toluca <- lm(workHours ~ lotSize , data = toluca)

> plot(toluca$lotSize , toluca$workHours , xlab="Lot Size" ,
   ylab="Work Hours" , main="Toluca Company: Refrigeration Manufacturer")

> abline(lm_toluca)
```

Fitted Regression Line Graph



Running the Regression Model [R Code]

```
> summary(lm_toluca)

Call:
lm(formula = workHours ~ lotSize, data = toluca)

Residuals:
    Min      1Q  Median      3Q     Max
-83.876 -34.088  -5.982  38.826 103.528

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.366     26.177   2.382   0.0259 *
lotSize        3.570      0.347  10.290 4.45e-10 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 48.82 on 23 degrees of freedom
Multiple R-squared:  0.8215, Adjusted R-squared:  0.8138
F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

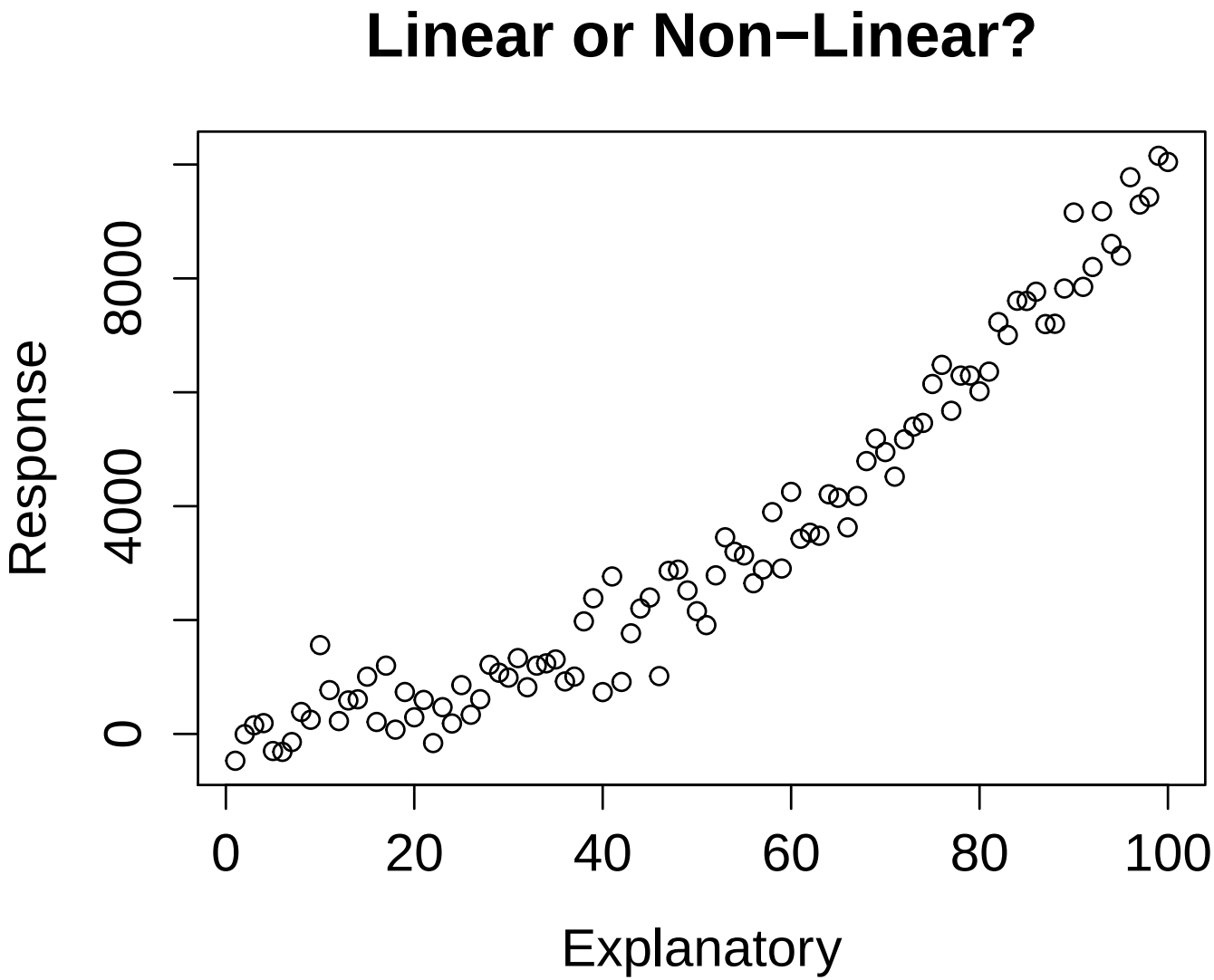
How to Best Analyze a Regression Model

- You've certainly noticed by now that when calling

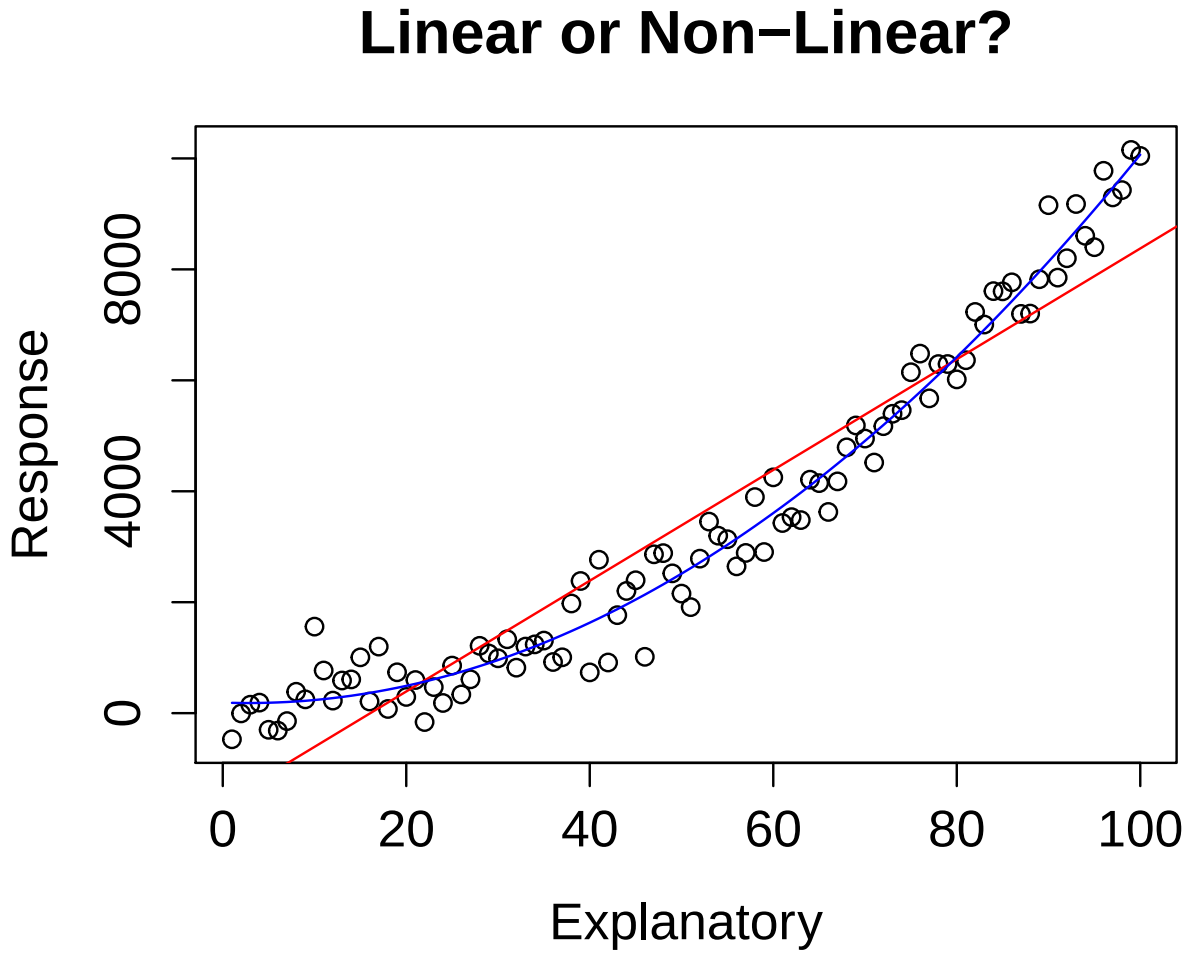
```
> summary(<lm_object>)
```

  in R, you get a mess of information
- Some of this information will be useful to us as modelers and analysts, and some of the information is for more advanced users
- **In an effort to evaluate the quality of the model**, the information we want to concern ourselves with is
  1. Generate a scatter plot, plot the estimated regression function and verify the linearity of the relationship(s)
  2. the F-statistic and its associated p-value
  3. multiple R^2 for simple linear regression (SLR)
  4. multiple *and* adjusted R^2 for multiple linear regression (MLR)
  5. the estimated coefficient(s) and their associated p-values regression

** this scatter is NOT based on the Toluca data**



** this scatter is NOT based on the Toluca data**



Linear $R^2 = 0.9102$

Quadratic $R^2 = 0.9768$

The F-statistic and its p-value [Model-Level Analysis]

The **first** statistic we should examine when assessing a linear regression model is the F-statistic

```
> summary(lm_toluca)$fstatistic

   value    numdf    dendf
105.8757   1.0000  23.0000
```

- A large F-statistic is good, but we also want to have a significant p-value, i.e., p-value that is less than our predetermined level of statistical significance α
- Observe that the above R command provides us with the F-statistic but not the associated p-value
- To get all the information, we will need to examine the Analysis of Variance (ANOVA) table

n.b. $n \equiv \#$ of observations (rows of data)

```
> anova(lm_toluca)
Analysis of Variance Table

Response: workHours
           Df Sum Sq Mean Sq F value    Pr(>F)
lotSize     1 252378  252378  105.88 4.449e-10 ***
Residuals 23  54825    2384
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

Source of Variation	*SS*	*df*	*MS*
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$	

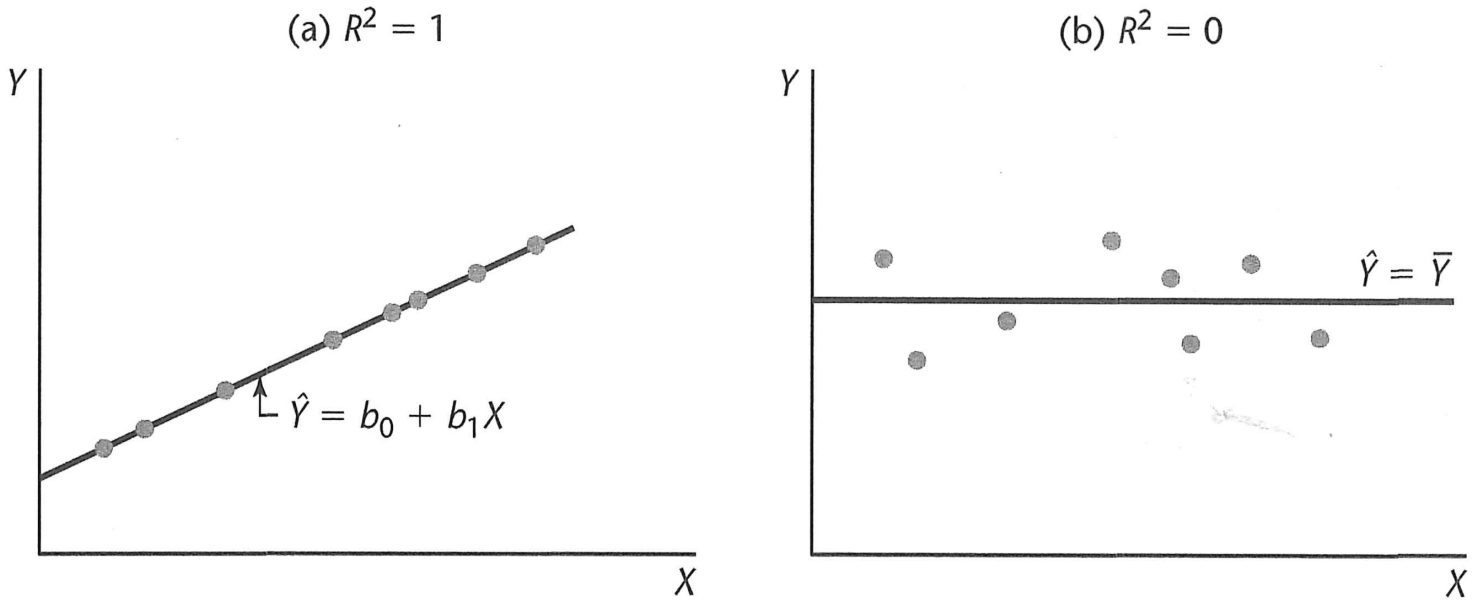
Figure 9: SLR ANOVA Table

The Coefficient of Determination (R^2) [Model-Level Analysis]

- R^2 is one (of many) indicators of the quality of the model
- *Ceteris paribus*, larger R^2 is better than smaller R^2
- **n.b.** a large R^2—in and of itself—is not necessarily imply that the estimated regression line is a good fit

```
> summary(lm_toluca)$r.squared
[1] 0.8215335
```

- Recall, $0 \leq R^2 \leq 1$



- The R-squared of the regression or the **Coefficient of Determination** is a measure of the goodness-of-fit of the regression line, $\hat{Y}_i$'s, to the Y_i's

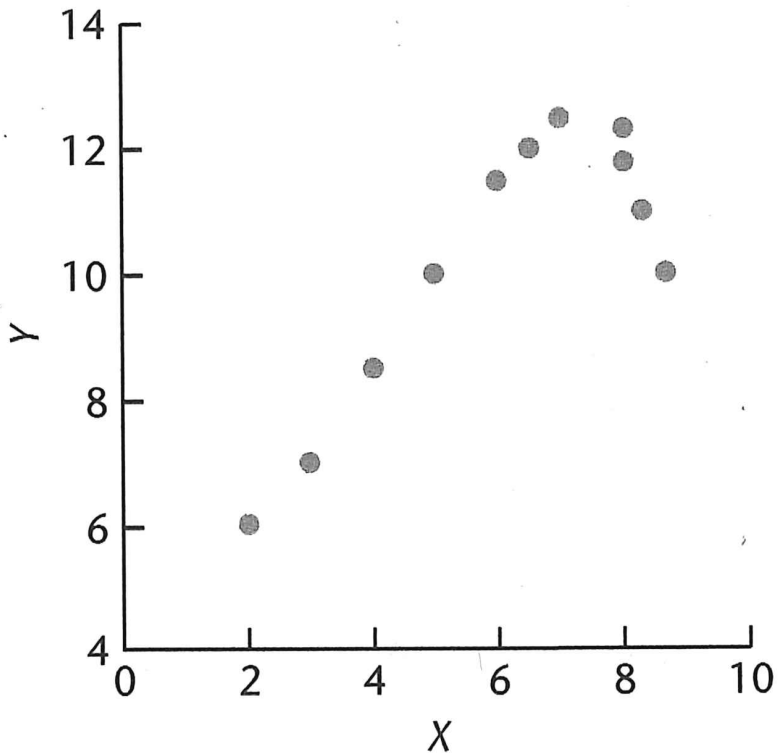
$$R^2 = \frac{SSR}{SSTO} = \frac{SSTO - SSE}{SSTO} = \frac{SSTO}{SSTO} - \frac{SSE}{SSTO} = 1 - \frac{SSE}{SSTO}$$

- R^2 is the ratio of the proportionate reduction of total variation associated with the use of the predictor variable X
- Thus the larger R^2 is, the more total variation of Y is reduced by introducing the predictor variable X
- It is interpreted as the fraction of the sample variation in Y that is explained by the variation in X

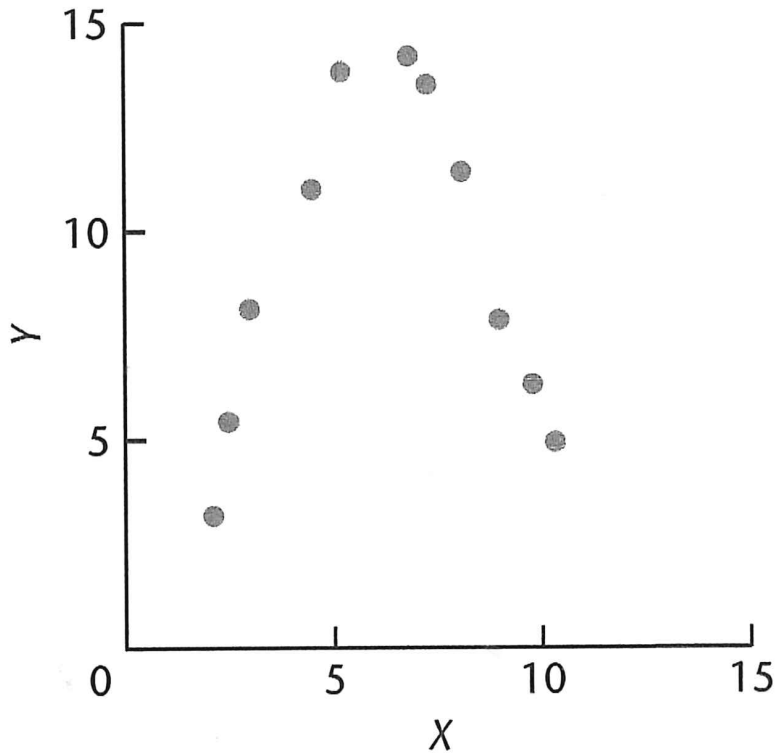
The Limitations of R^2 [Model-Level Analysis]

1. High R^2 **does not** imply that the estimated regression line is a good fit
2. $R^2 \approx 0$ **does not** imply that X and Y are not related

Scatter Plot with $R^2 = .69$
Linear regression is not a good fit



Scatter Plot with $R^2 = .02$
Strong relation between X and Y



Estimated Coefficients and their p-values [Variable-Level Analysis]

To exclusively obtain the coefficient(s) and their associated p-values in R

```
> summary(lm_toluca)$coefficients

             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 62.365859 26.1774339  2.382428 2.585094e-02
lotSize      3.570202  0.3469722 10.289592 4.448828e-10
```

- In the large majority of cases, the `(Intercept)` term provides no insight and requires no interpretation, as we seldom collect data around $X = 0$
- The magnitude (size) of the coefficients *does not* provides us with any information about their relevance to the SLR model
- the p-value(s), if *less* than our predetermined level of statistical significance α, indicate whether or not predictor variable(s) are statistically significant in the model
  - **n.b.** do not eliminate the *intercept* from the model even if it is statistically insignificant

Equivalence of t- and F-statistics in SLR

- An interesting mathematical fact is that in simple linear regression, if we take the t-statistic value for the singular explanatory variable (not the intercept, the other one) and square it, i.e., t^2, this will equal the value of the F-statistic

$$t^2 = F$$

E.g. For Toluca, the t-statistic for `lotSize` is 10.289, which, when squared, is 105.86, exactly the value of the F-statistic (accounting for rounding)

- This fact is always true for all SLR models
- **n.b.** This *only* applies in SLR, *NOT* MLR

Retrieving Residuals & Fitted Values [R Code]

```
> residuals(lm_toluca)
          1           2           3           4           5           6  ...
 51.0179798 -48.4719192 -19.8759596  -7.6840404  48.7200000 -52.5779798  ...
```

```
> fitted(lm_toluca)
       1        2        3        4        5        6        7        8  ...
347.9820 169.4719 240.8760 383.6840 312.2800 276.5780 490.7901 347.9820  ...
```

Section 15

Diagnostics for Residuals

Diagnostics for Residuals

- Direct diagnostic plots of the dependent variable (Y) are often not particularly useful in regression analysis as the response value should be a function of the level of the predictor variable
- Rather, diagnostics for the response variable are carried out indirectly through an examination of the residuals

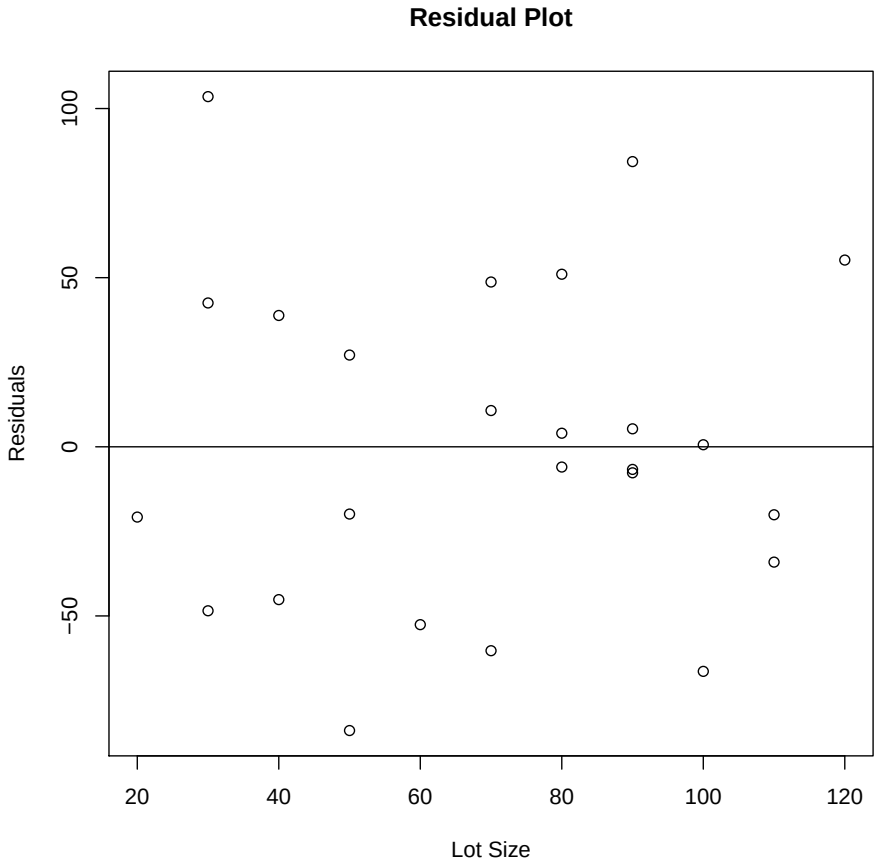
- Direct diagnostic plots of the dependent variable (Y) are often not particularly useful in regression analysis as the response value should be a function of the level of the predictor variable
- Rather, diagnostics for the response variable are carried out indirectly through an examination of the residuals
- We will consider the use of residuals for examining four types of departures from the normal error simple linear regression model
  1. The error terms do not have constant variance
  2. The error terms are not independent
  3. The model fits all but one or a few outlier observations
  4. The error terms are not normally distributed
  5. One or several important predictor variables have been omitted from the model

Subsection 1

Nonconstancy of Error Variance (Heteroskedasticity)

Residual Plot [R Code]

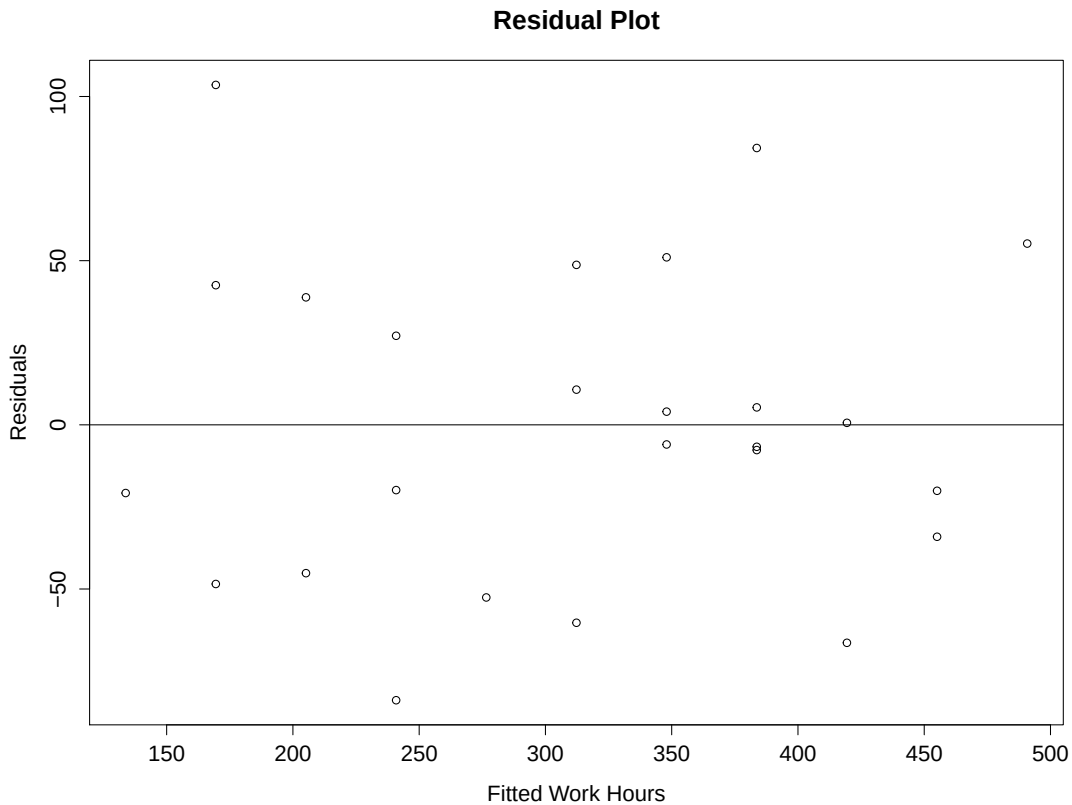
Plot the residuals (e_i) against the predictor variable X, hoping to find a random distribution of e_i's for all values of X




```
> plot(toluca$lotSize, residuals(lm_toluca),xlab="Lot Size",
   ylab="Residuals", main="Residual Plot")
> abline(h=0)
```

Residual Plot [R Code]

You can also plot residuals (e_i) against the fitted values ($\hat{Y}_i$), generating the exact same residual plot as when e_i were plotted against X_i's, albeit with a scaled x-axis. Why?

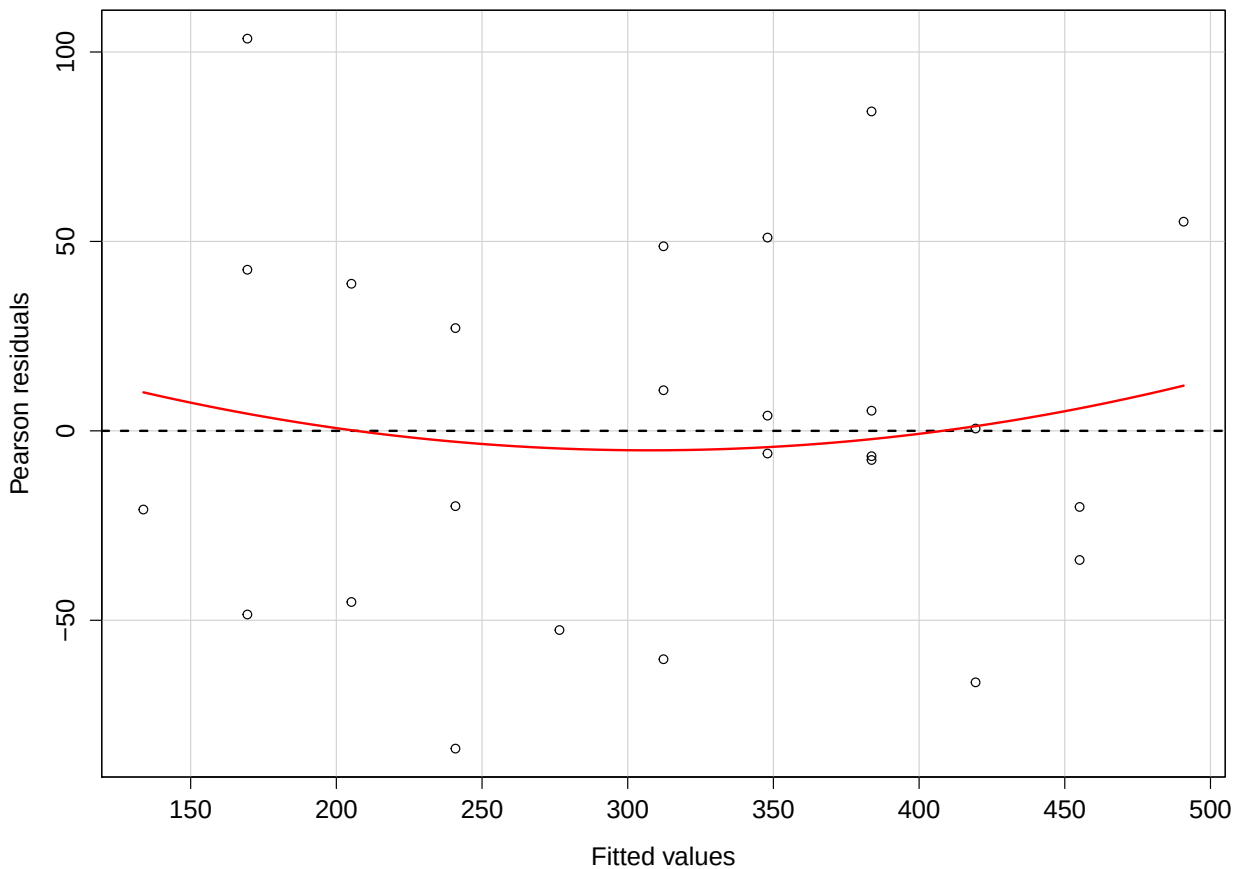



```
> plot(fitted(lm_toluca), residuals(lm_toluca),xlab="Fitted Work Hours",
   ylab="Residuals", main="Residual Plot")
> abline(h=0)
```

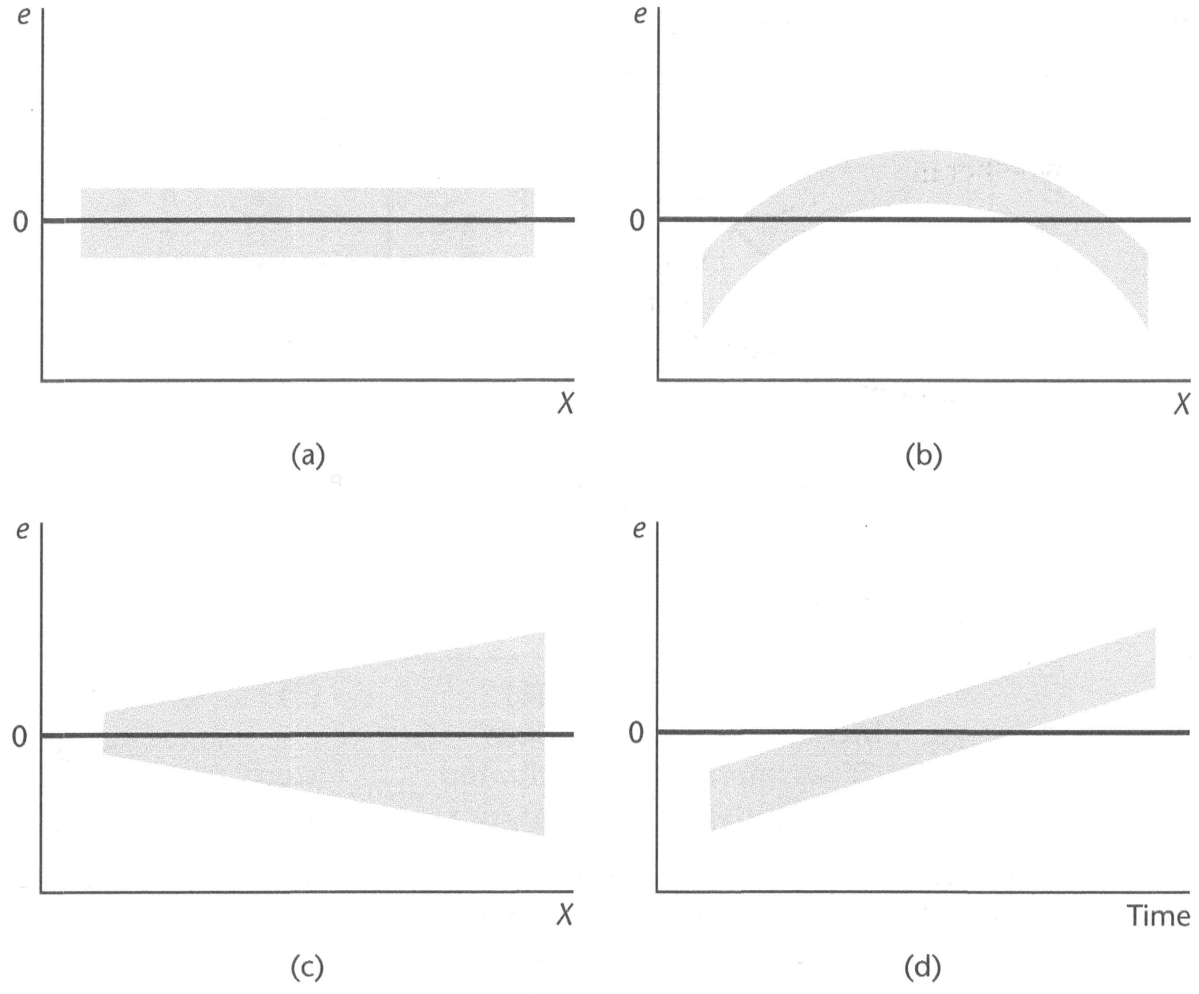
Using the `car` package, you can and should use the following shortcut

```
> residualPlot(lm_toluca)
```

to generate residual plots



Plots of residuals (e_i) against predictor variables (X_i) is useful in discerning whether or not we can establish homoskedasticity, i.e., constant error variance ($V[\varepsilon_i] = \sigma^2$)



Brown-Forsythe Test for Homoskedasticity

- The BF test works well in the context of linear regression in detecting 'megaphone-shaped' heteroskedasticity, i.e., smaller variance at smaller levels of the independent variable and larger variance at larger levels of the independent variable
  - It will detect an inverted megaphone-shaped non constant error variance equally well
- To conduct the BF test, we divide data into two groups based on X, so that one group consists of of cases where the X level is comparatively low and one group where the X level is comparatively high
- If error variance is increasing or decreasing across the entire set of data, the residuals in one group will exhibit larger variability than the other
- Equivalently, the absolute deviations of the residuals around their group mean will tend to be larger for one group than the other group

Brown-Forsythe Test for Homoskedasticity cont'd

- In order to enhance the robustness of the test, the BF utilizes the absolute deviations of the residuals around the **median** for each group
- Although the distribution of the absolute deviations of the residuals is usually not normal, it can be shown that the t^* test statistic approximately follows the t distribution when the variance of the error terms is constant and the sample sizes of both groups are not extremely small

n.b. We will use e_{i1} to denote the i^{th} residual from the first group and e_{i2} to denote the i^{th} residual from the first group, denote the sample size of each group by n_1 and n_2 respectively (where $n = n_1 + n_2$) and the medians of each group by $\widetilde{e}_1$ and $\widetilde{e}_2$ respectively

Brown-Forsythe Test for Homoskedasticity cont'd

- The Brown-Forsythe test employs the observations absolute deviations, denoted d_{i1} and d_{i2} respectively, of the residuals around their group median

$$d_{i1} = |e_{i1} - \widetilde{e}_1| \qquad d_{i2} = |e_{i2} - \widetilde{e}_2|$$

- We define the two-sample t test statistic for the BF test as

$$t^*_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(1-\alpha/2, n-2)}$$

where the pooled variance is

$$s^2 = \frac{\sum(d_{i1} - \bar{d}_1)^2 + \sum(d_{i2} - \bar{d}_2)^2}{n-2}$$

and the hypothesis test is

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_a : \sigma_1^2 \neq \sigma_2^2$$

Brown-Forsythe Test for Homoskedasticity [R Code]

n.b. The Brown-Forsythe test is a modification of the Levene test, where the former measures absolute deviations of the residuals against the **median** group residual whereas the latter measures the absolute deviations of the residuals against the **mean** group residual

```
> install.packages("car")
> library("car")
```

- A Levene test in R calls the function `leveneTest` with the parameter `center=mean`
- A Brown-Forsythe test in R calls the function `leveneTest` with the parameter `center=median`

- If you try to use call a BF test in R on an `lm` object with numeric predictor variables, you get the following error

```
> leveneTest(lm_toluca, center = median)

Error in leveneTest.formula(formula(y), data = model.frame(y), ...) :
Levene's test is not appropriate with quantitative explanatory variables.
```

Brown-Forsythe Test for Homoskedasticity [R Code] cont'd

- The inherent problem is that R will not create the two groups that you need to test against each other automatically
- You can certainly do this on your own, create an additional column, and subsequently code each observation with a 1 or a 2 to indicate group affiliation—ensuring the column is coded as a factor and not as numeric
- There are no hard and fast rules for grouping, and this can be quickly become a subjective process; an easy grouping is simply to split the total observations in half

n.b. The Levene/Brown-Forsythe test is not limited to 2 groupings; the tests can accommodate k groupings (and the definition of the hypothesis test will change accordingly)

Breusch-Pagan Test for Homoskedasticity

- This test assumes that error terms are independent and normally distributed and that the variance of the error term ε_i, denoted by σ_i^2, is related to the level of X in the following way

$$log_e(\sigma_i^2) = \gamma_0 + \gamma_1 X \tag{82}$$

  where (82) implies that σ_i^2 either increases or decreases with the level of X, depending on the sign of γ_1
- The hypothesis test for constant variance is

$$H_0 : \gamma_1 = 0 \quad \textit{(homoskedastic residuals)}$$
$$H_a : \gamma_1 \neq 0 \quad \textit{(heteroskedastic residuals)}$$

Breusch-Pagan Test for Homoskedasticity cont'd

- Obtaining the test statistic is a multi-step process
  1. Regress Y_i on X_i
  2. Regress the squared residuals e_i^2 form the previous regression against X_i
  3. Compute the test statistic X^2_{BP}

$$X^2_{BP} = \frac{SSR_{(e^2X)}/2}{(SSE_{(XY)}/n)^2}$$

which, if the null hypothesis holds and n is reasonably large, is $\sim \chi^2_{(1-\alpha/2,1)}$

```
> install.packages("lmtest")
> library("lmtest")
> bptest(lm_toluca)

   studentized Breusch-Pagan test

data:  lm_toluca
BP = 1.1326, df = 1, p-value = 0.2872
```

Breusch-Pagan Test for Homoskedasticity cont'd

- There are many, many variations on BP test, e.g., you will notice that the version implemented in the `lmtest` package uses studentized residuals instead of actual residuals
- Moreover, you will notice that if you use what is the Breusch-Pagan test from the `car` package, the function call is `ncvTest`...why?
- Breusch & Pagan published their paper in 1979, and Cook and Weisberg published an extended/different version in 1983...and Weisberg wrote the `car` package

```
> install.packages("car")
> library("car")
> ncvTest(lm_toluca)

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.8209192    Df = 1     p = 0.3649116
```

Subsection 2

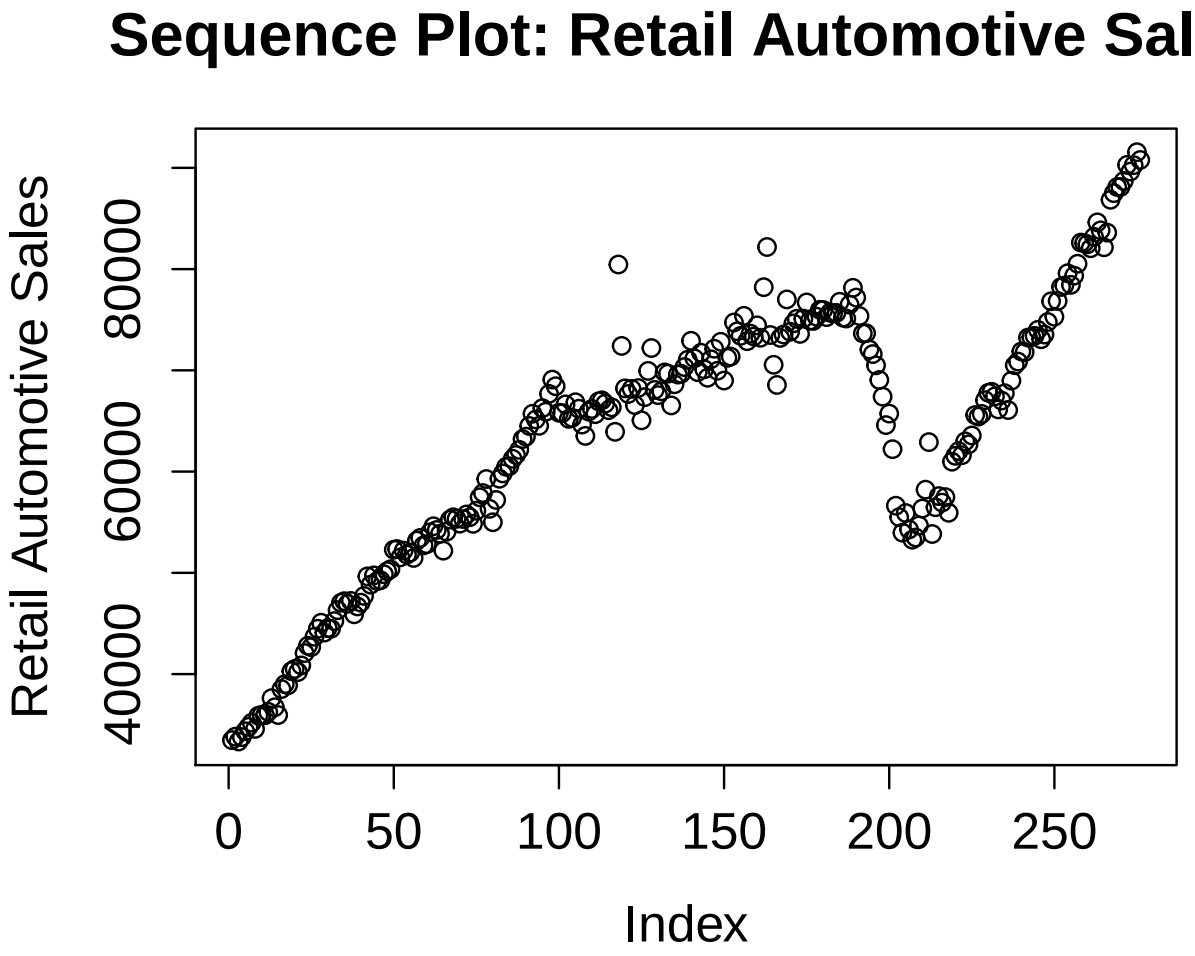
Nonindependence of Error Terms

Nonindependence of Error Terms

- Create a residual plot (e_i on X_i) and observe whether or not a pattern exists as X increases
- If data are obtained in a time sequence or in some other type of sequence, e.g., adjacent geographic areas, it is advisable to prepare a sequence plot
  - A sequence plot does not have the predictor variable X on the x-axis, but rather the sequence in which the data was collected
  - Sequence plots help us identify correlation of patterns over the sequence, be it linear or cyclical trends
  - It is not uncommon for the nonindependence of error terms to be indicative of endogeneity
  - For sequence plots, plot residuals (e_i) on the y-axis against the sequence (on the x-axis)

```
> plot(toluca_SLR$residuals,xlab="Sequence",ylab="Residuals",
    main="Sequence Plot")
```

Sequence Plot for Retail Automotive Sales




```
> plot(retailAutoSales$retailSales,ylab="Retail Automotive Sales",
main="Sequence Plot: Retail Automotive Sales")
```

Subsection 3

Outliers

- Outliers are extreme observations. Residual outliers can be identified with residual plots (against X or $\hat{Y}_h$), as well as dot plots and box plots
- A semi-studentized residual plot can help put our error terms in context
- What is a semi-studentized residual?
  1. To standardize a random variable, we subtract its population mean and divide by its population standard deviation, e.g., $(X - \mu_x)/\sigma_x$ (note here that the use of X does not imply the predictor variable)
  2. Since we do not have population parameters, we use point estimates, and the resulting standardized variables are now studentized variables as the follow the t, not the normal, distribution
  3. Why semi-studentized? Coming up soon...

Standardized & Studentized Residuals

To standardize a residual, compute

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

```
> install.packages("MASS")
> library("MASS")
> plot(stdres(toluca_SLR),xlab="Lot Size",ylab="Standardized
    Residuals",main="Standardized  Residual Plot")
> abline(h=0)
```

To studentize a residual, compute

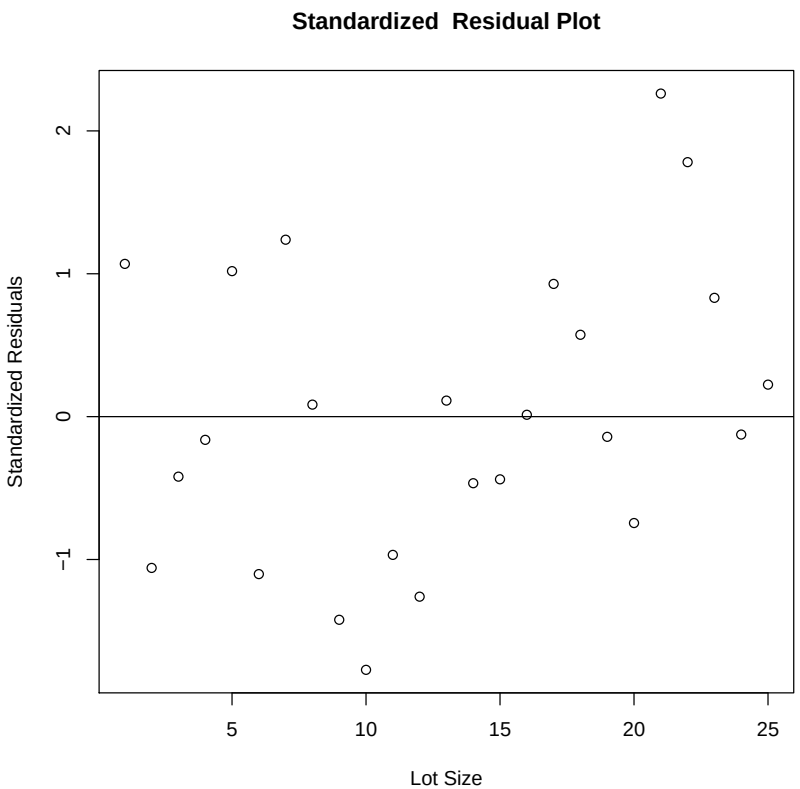
$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE_{(-i)}}} = \frac{e_i}{\sqrt{MSE_{(-i)}}}$$

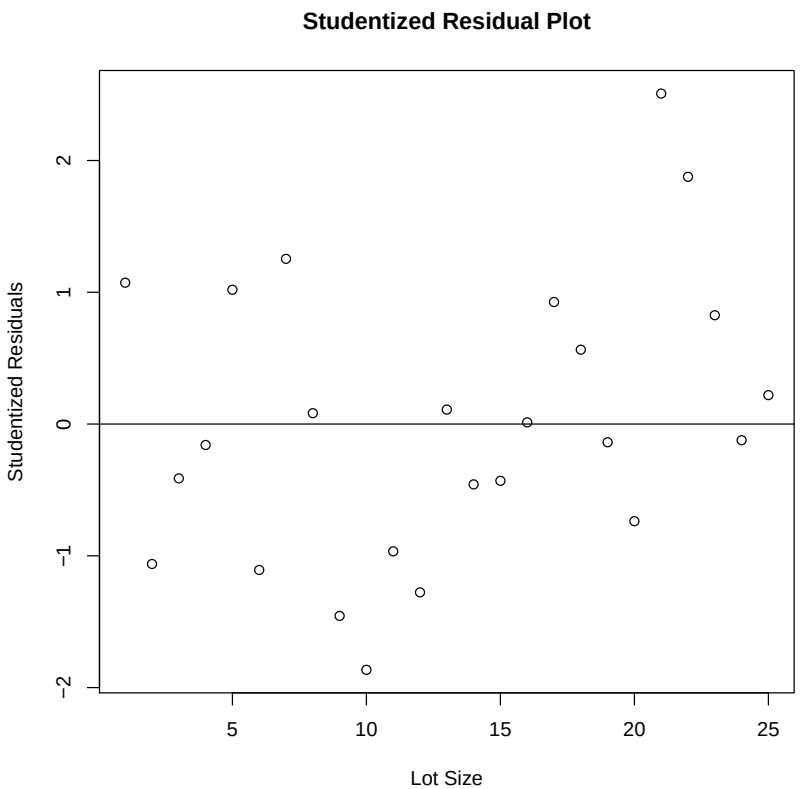
```
%% note studres is also from the MASS package
> plot(studres(toluca_SLR),xlab="Lot Size",ylab="Studentized
    Residuals",main="Studentized Residual Plot")
> abline(h=0)
```

(studentized residuals are also referred to as Jackknifed Residuals)

Standardized & Studentized Residual Graphs

Graphs on left and right are **not the same**





Outliers

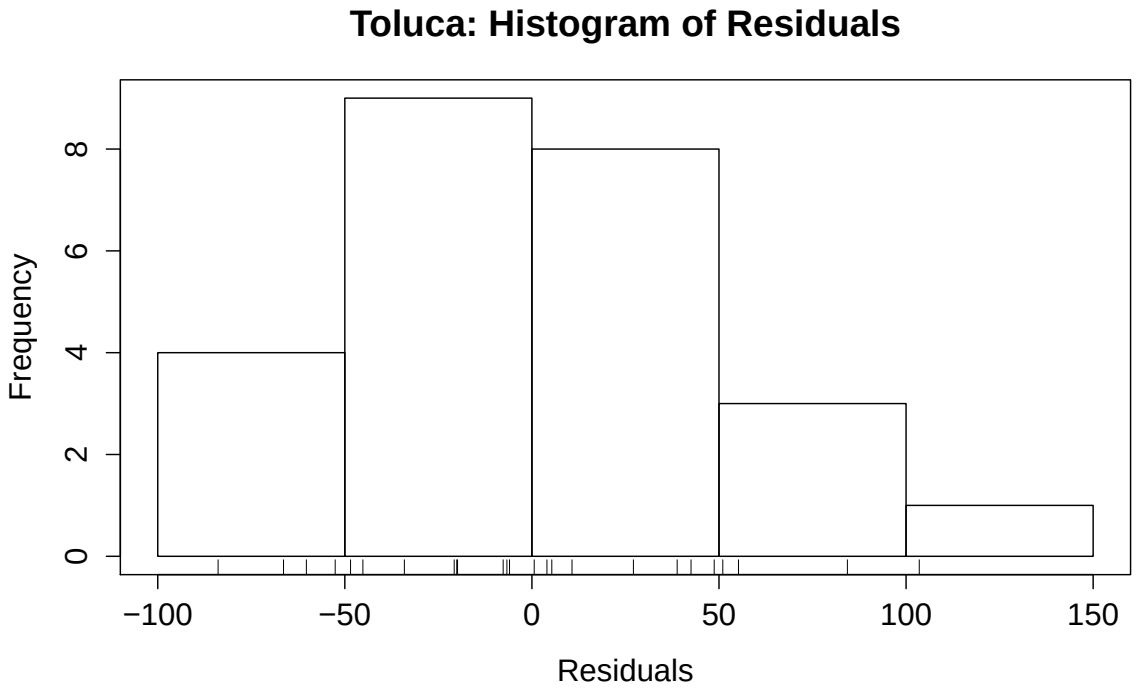
- Outliers can create great difficulty for a number of reasons
- Outliers can cause a fitted regression line to be significantly altered in an attempt to minimize the sums of squared (SS) deviations, with the SS for the singular outliers influencing the fitted regression line disproportionately
- Alternatively, outliers can be indicative of endogeneity (omitted predictor variables)
- **ONLY** discard an outlier if there is evidence that it represents an error in recording, a miscalculation, malfunctioning equipment, or the like
    - **You cannot throw away outliers just because they are inconvenient**

Subsection 4

Non-Normality of Error Terms

Histogram of the Residuals [R Code]

Generate a histogram and rug plot to examine the distribution of the residuals




```
> hist(residuals(lm_toluca),xlab="Residuals",main="Toluca: Histogram
  of Residuals")

> rug(residuals(lm_toluca))

> box()
```

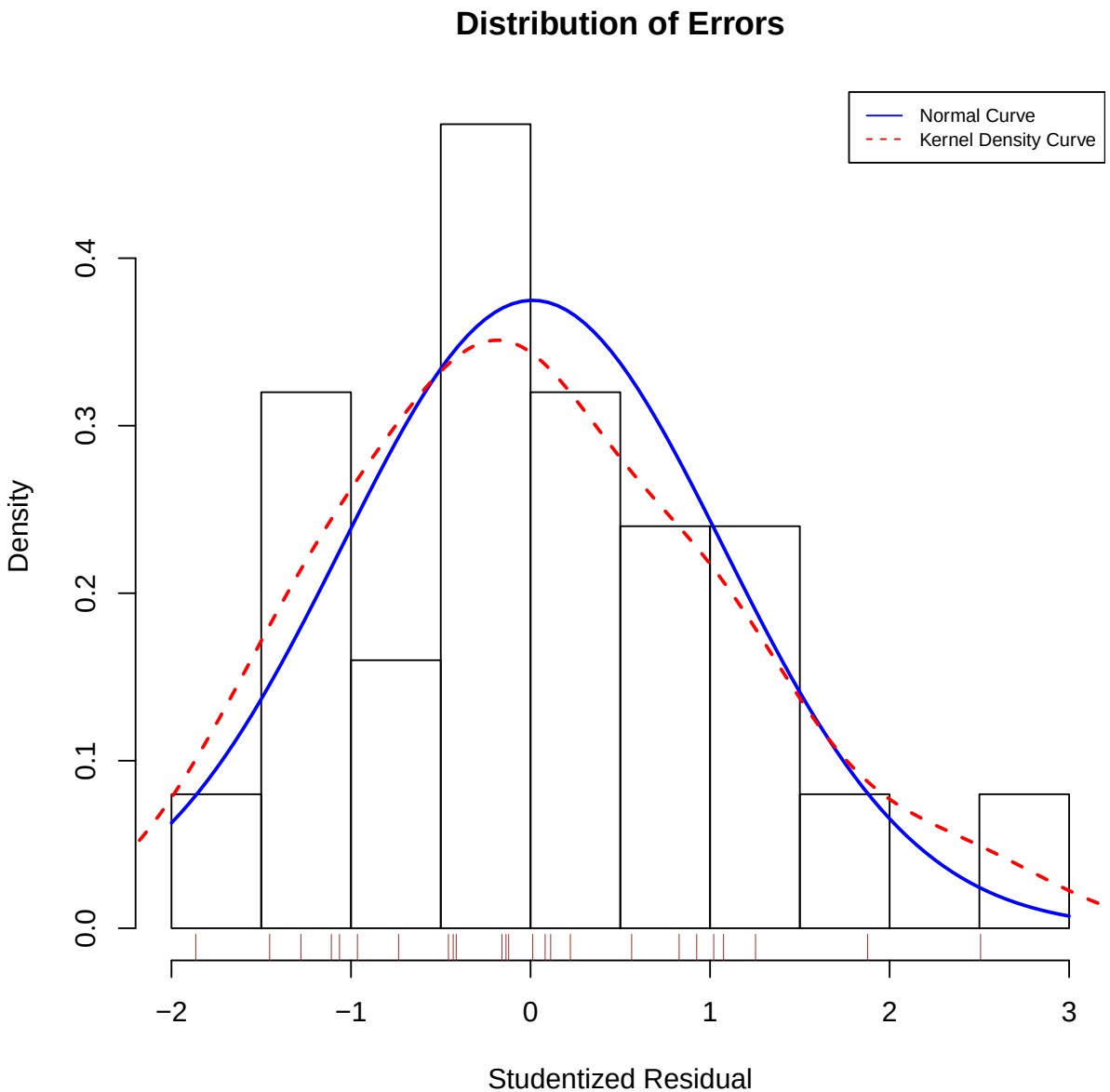
To evaluate the normality of error terms, the simplest visual check is a histogram

```
%% this is a nice little bit of code taken from "R in Action"
%% by Robert Kabacoff

> residplot <- function(fit, nbreaks=10) {
    z <- rstudent(fit)
    hist(z, breaks=nbreaks, freq=FALSE,
    xlab="Studentized Residual",
    main="Distribution of Errors")
    rug(jitter(z), col="brown")
    curve(dnorm(x, mean=mean(z), sd=sd(z)),
        add=TRUE, col="blue", lwd=2)
    lines(density(z)$x, density(z)$y,
        col="red", lwd=2, lty=2)
    legend("topright",
        legend = c( "Normal Curve", "Kernel Density Curve"),
        lty=1:2, col=c("blue","red"), cex=.7)
}

> residplot(lm_toluca)
```

Histogram, Kernel Density & Rug Plots



Shapiro-Wilk Test for Normality

- A more quantitative approach to test for normality is to employ a formalized statistical hypothesis test, controlling for Type I error
- Although we will use the Shapiro-Wilk test, other very common, high-quality tests of normality exist, including the Anderson-Darling test and the Kolmogorov-Smirnov test
- The Shapiro-Wilk hypothesis test is

$$H_0 : \text{normally distributed data}$$

$$H_a : \text{non-normally distributed data}$$

where a p-value $< \alpha$ is a rejection of the null hypothesis

```
> shapiro.test(lm_tolucaR$residuals)

Shapiro-Wilk normality test

data:  lm_toluca$residuals
W = 0.9789, p-value = 0.8626
```

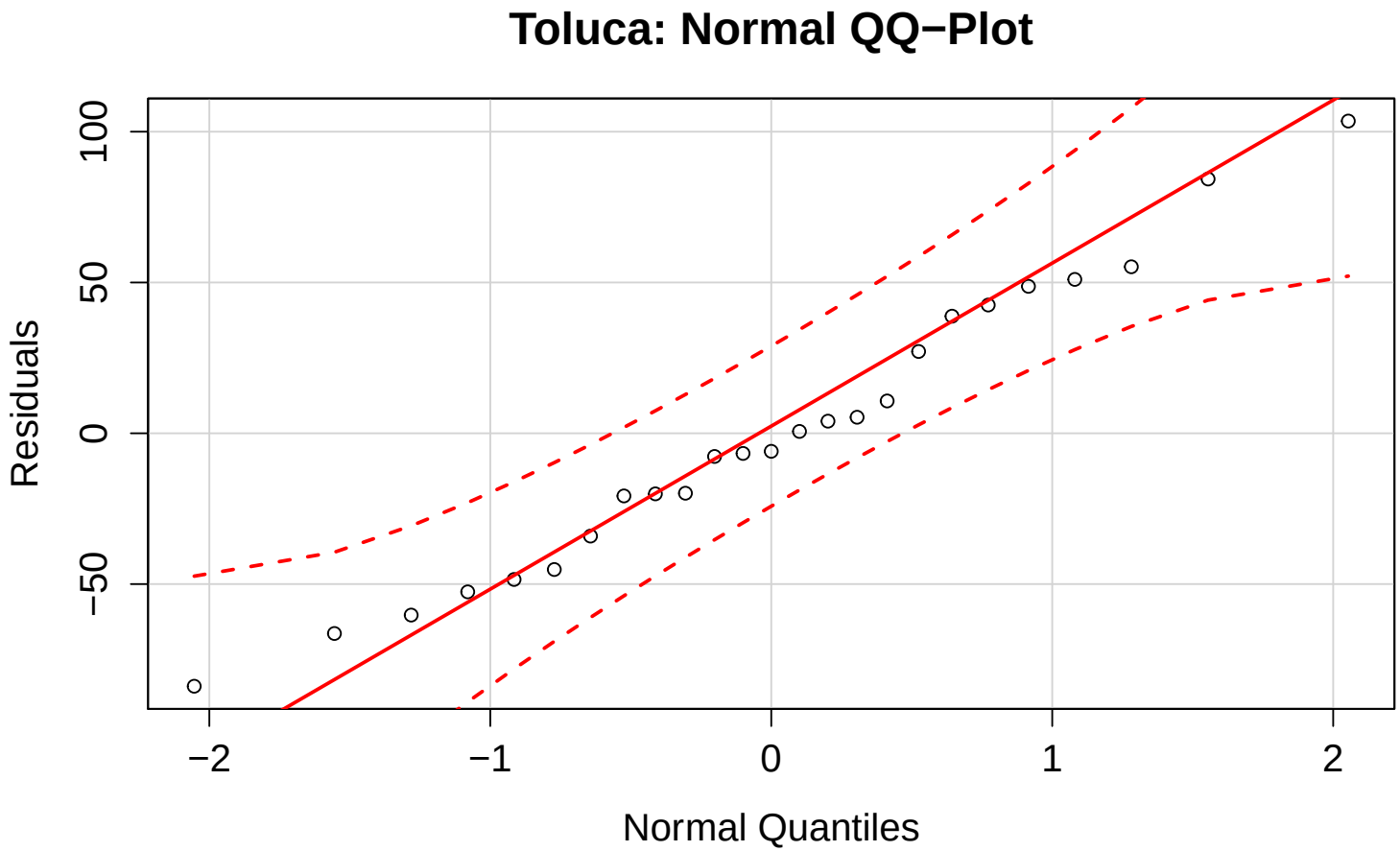
Quantile-Quantile (QQ) Plots

- A more advanced visual check for normality includes quantile-quantile (qq) plots (or the analogous probability-probability (pp) plot)
- A qq plot will plot a theoretical distribution against an empirical distribution
  - if the theoretical and empirical quantiles closely match, we can conclude that the empirical distribution represents the underlying theoretical distribution well
  - if the empirical quantiles deviate significantly from the theoretical quantiles, we conclude that the empirical distribution **does not** represent the underlying theoretical distribution

```
> install.packages("car")
> library("car")
> qqPlot(lm_toluca)
```

QQ-Plot of Residuals [R Code]

Generate a histogram and rug plot to examine the distribution of the residuals




```
> qqPlot(residuals(lm_toluca),main="Toluca: Normal QQ-Plot",
  xlab="Normal Quantiles",ylab="Residuals")
```

QQ Plots in the car Package

- When using the qqPlot function from the car package, you can feed the function a single lm argument

```
> qqPlot(lm_toluca)
```

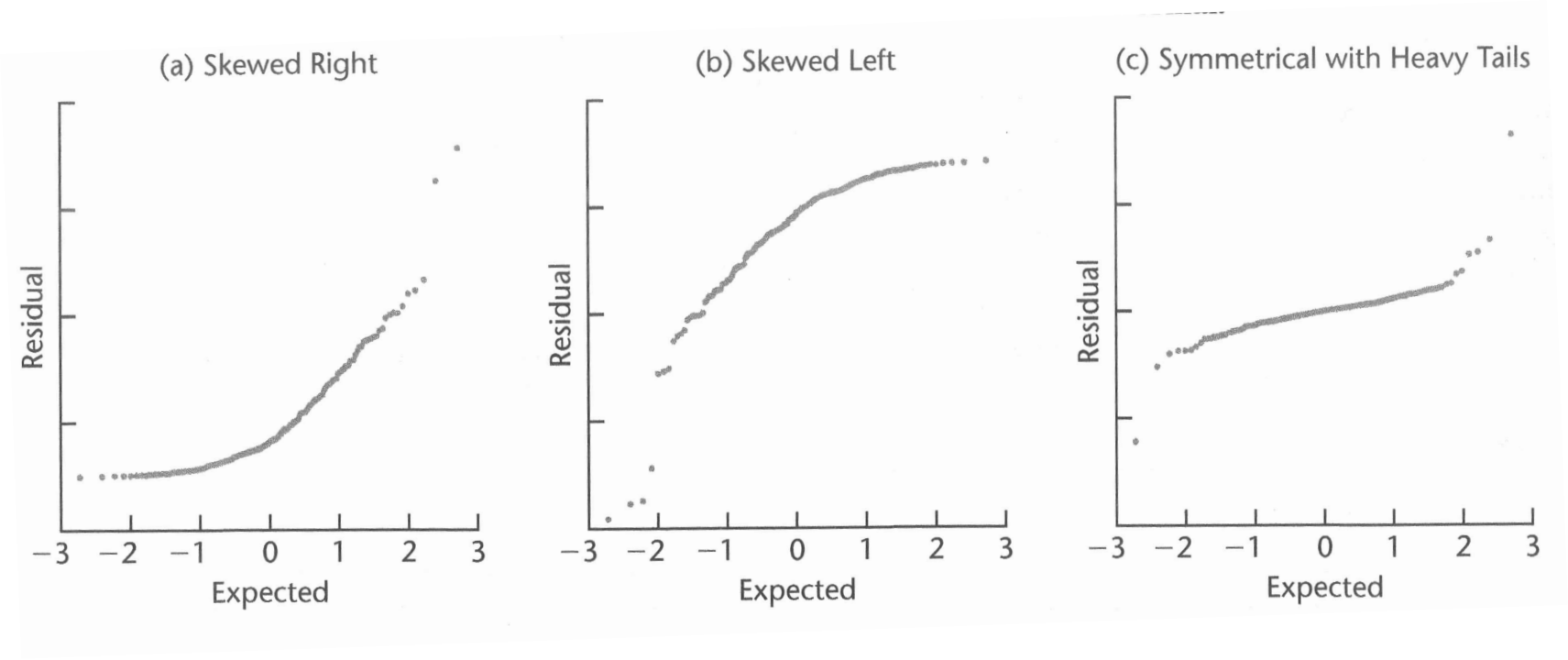
- If doing so, be aware that the qqPlot function extracts the **studentized residuals**—distributed according to the t distribution—from the lm object and plots them against theoretical t quantiles
- The underlying theoretical distribution can also be specified

```
% incorrect
> qqPlot(lm_toluca,distribution="norm")
```

  but doing so in this fashion will result in an incorrect qq plot, as the qqPlot function then plots studentized residuals (t-distributed) against an underlying normal theoretical distribution
- If you are going to specify the underlying theoretical distribution, ensure that you are plotting the distribution-matching residuals

```
% correct
> qqPlot(lm_toluca$residuals,distribution="norm")
```

Interpreting QQ Plots



Subsection 5

Endogeneity

- If at all possible, residuals should also be plotted against variables that are omitted from the model that may have important effects on the response
- In the context of regression endogeneity refers to omitted variable that could otherwise be significant in the model, i.e., they can have a significant impact on the response variable Y

Section 16

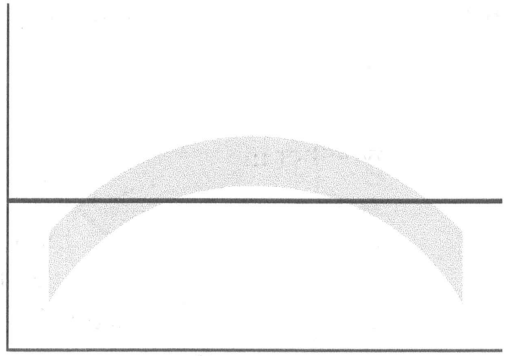
Remedial Measures

What to do when SLR isn't Appropriate?

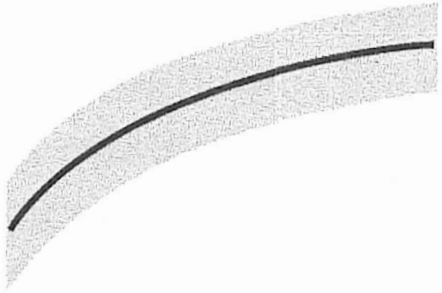
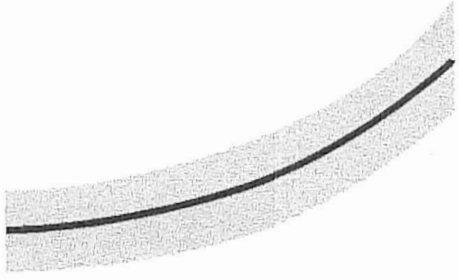
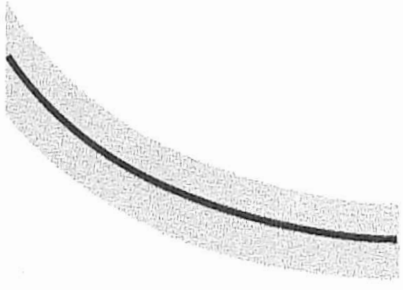
- When an SLR model is not appropriate for a data set, you have two options

1. Abandon the model and develop a more appropriate model
   - This may result in a more appropriate model, albeit more complex, with potential difficulties in estimating the parameters
2. Employ a transformation on the data so that the regression model is appropriate for the transformed data
   - Results in simple estimation methods with a simpler model and potentially fewer parameters, which is very useful when you have a limited number of data points
   - Transformations may obscure the fundamental interconnections between variables, although at other times it may make things clearer

- Transformations can be employed to (approximately) linearize a non-linear model



- To linearize a nonlinear regression relation (above) when the distribution of the error terms is reasonably close to a normal distribution and the error terms have approximately constant variance, we want to first transform the independent variable X
- In this situation
  - Transforming the dependent variable Y in this case may change the distribution of error terms and/or may lead to heteroskedastic error terms

Prototype Regression Pattern	Transformations of X	
	$X' = \log_{10} X$	$X' = \sqrt{X}$
	$X' = X^2$	$X' = \exp(X)$
	$X' = 1/X$	$X' = \exp(-X)$

- At times it may also be helpful to introduce a constant into the transformation
- E.g., if some of the X values are close to zero and a reciprocal transformation ($X' = \frac{1}{X}$), we can shift the origin by employing the transformation

$$X' = \frac{1}{X + k}$$

where k is an appropriately chosen constant

Heteroskedasticity

- Transformations of the **dependent** variable are often helpful

Nonnormality of Error Terms

- Lack of normality and nonconstant error variance often go hand in hand
- Often times, the same transformation that help stabilize the variance also help in normalizing the error terms
- The sequential approach should be as follows
  1. stabilize error variance with a transformation (if necessary)
  2. correct for departures from normality if it remains an issue

Heteroskedasticity & Nonnormality of Error Terms cont'd

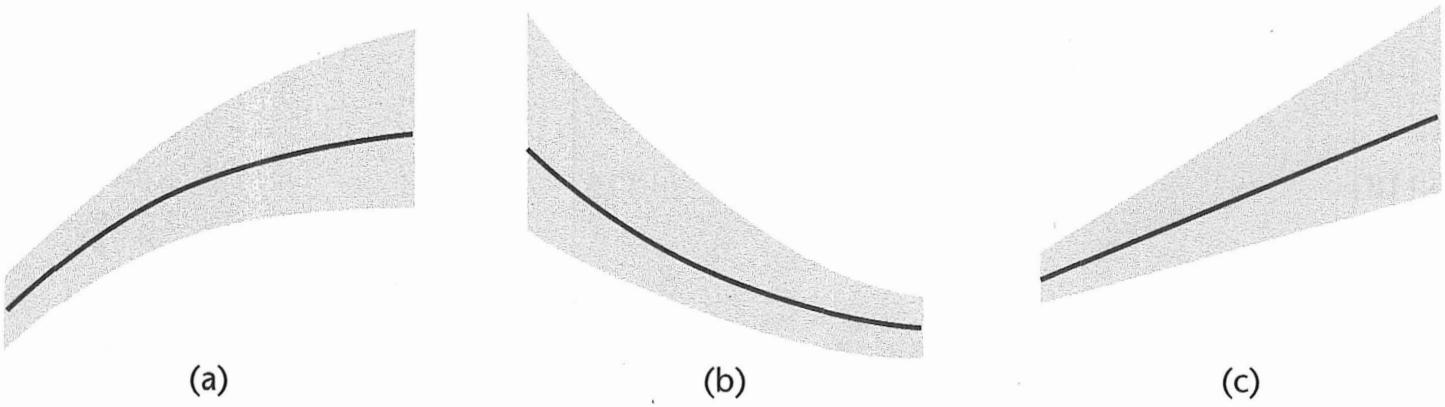
- Transformations on Y are useful since the shapes and spread of the distributions of Y need to be changed
- A transformation on Y may be sufficient to correct the model, whereas other times, a simultaneous transformation on X may be required
- Frequently, the non normality and heteroskedasticity take the form of increasing skewness and increasing variability on the distribution of the error terms as the mean response $E\{Y\}$ increases, e.g., regressing yearly household vacation expenditures (Y) on household income (X)
- It may be desirable to introduce a constant into a transformation of Y_i, such as when Y may be negative, e.g., including a constant to ensure positive Y values for a logarithmic transformation would take the form

$$Y' = log_{10}(Y + k)$$

where k is an appropriately chosen constant

Transformations for Nonnormality and Heteroskedasticity

Prototype Regression Pattern



Transformations on Y

$Y' = \sqrt{Y}$

$Y' = \log_{10} Y$

$Y' = 1/Y$

Note: A simultaneous transformation on X may also be helpful or necessary.

When the Only Problem is Heteroskedasticity

- When the only issue is heteroskedasticity but the scatter is linear, there is a high likelihood that a simultaneous transformations on Y and X will be necessary
- While transforming Y to stabilize variance, the linear relationship will be fundamentally changed to a curvilinear relationship, therefore a simultaneous transformation on X may be required to restore linearity

- It can be difficult based solely on diagnostic plots to know what transformation is best to use
- An automated procedure called the Box-Cox transformation automatically identifies a transformation from the family of power transformations on Y
- The family of power transformations is of the form

$$Y' = Y^{\lambda}$$

  where λ is a power determined by the Box-Cox procedure
- The normal error regression model therefore becomes:

$$Y_i^{\lambda} = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- The Box-Cox procedure uses the method of maximum likelihood to estimate λ, as well as the other parameters β_0, β_1 and σ^2
- In this way, Box-Cox identifies $\hat{\lambda}$ an estimator of λ to employ in the power transformation
- A simplified method for obtaining $\hat{\lambda}$ involves a numerical search in a range of candidate λ values

1. For a given value of λ, standardize the transformed response variable $Y' = Y^{\lambda}$—so the magnitude of the error sum of squares SSE does not depend on the value of λ—in the following manner

$$W_i = \begin{cases} K_1(Y_i^{\lambda} - 1) & \lambda \neq 0 \\ K_2(log_e Y_i) & \lambda = 0 \end{cases}$$

where

$$K_1 = \frac{1}{\lambda K_2^{(\lambda-1)}}$$

$$K_2 = \left(\prod_{i=1}^{n} Y_i\right)^{1/n}$$

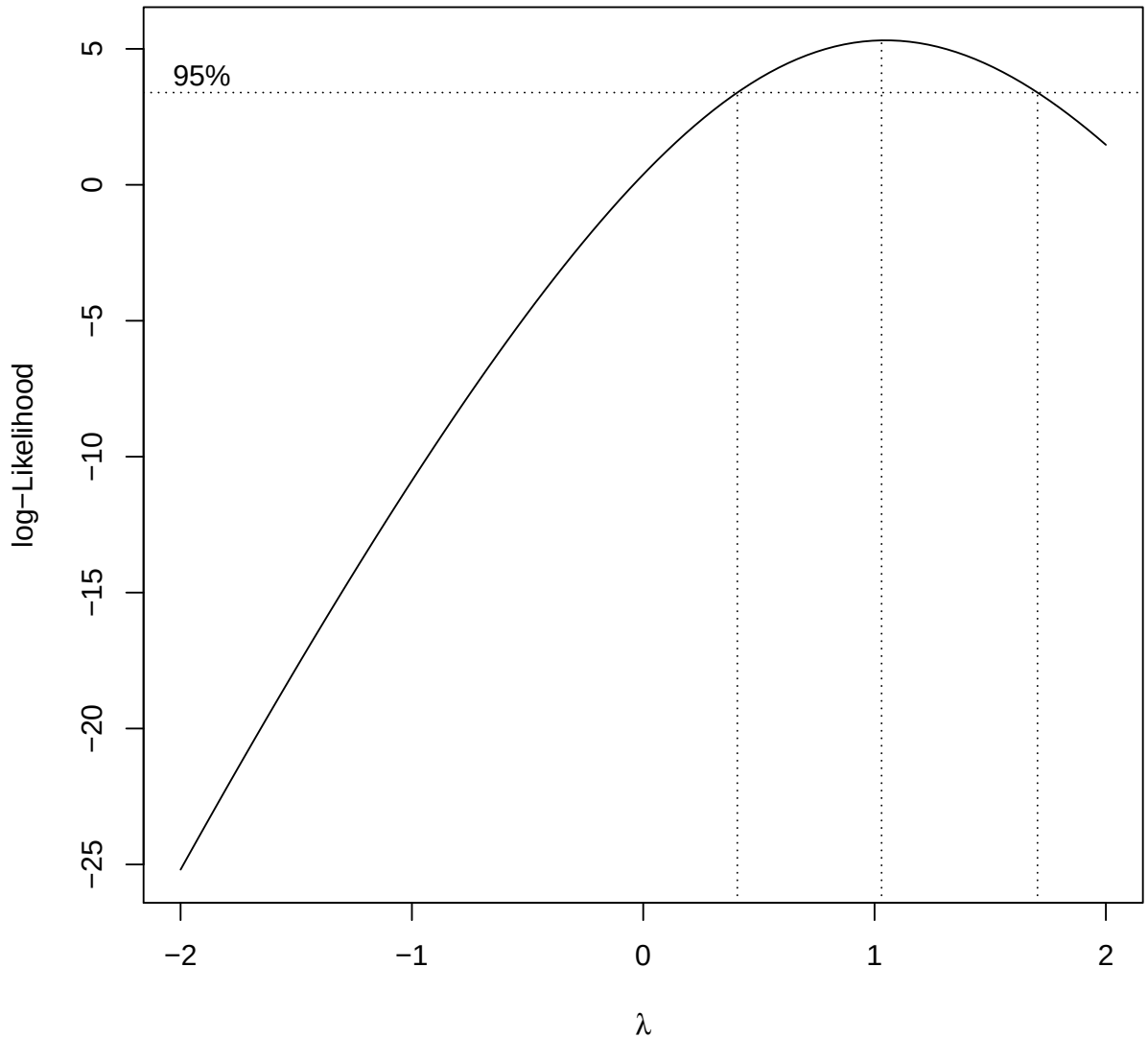
n.b. K_2 is the geometric mean of the Y_i observations

The Box-Cox Numerical Search Procedure cont'd

2. Regress W_i on X and obtain an SSE_λ
3. It can be shown that the maximum likelihood estimate $\hat{\lambda}$ is that value of λ for which SSE_λ is a minimum
4. Repeat steps 1 and 2 for multiple values of λ and identify the λ value minimizing SSE_λ, and use this value of λ as your power transformation

The Box-Cox Numerical Search Procedure [R Code]

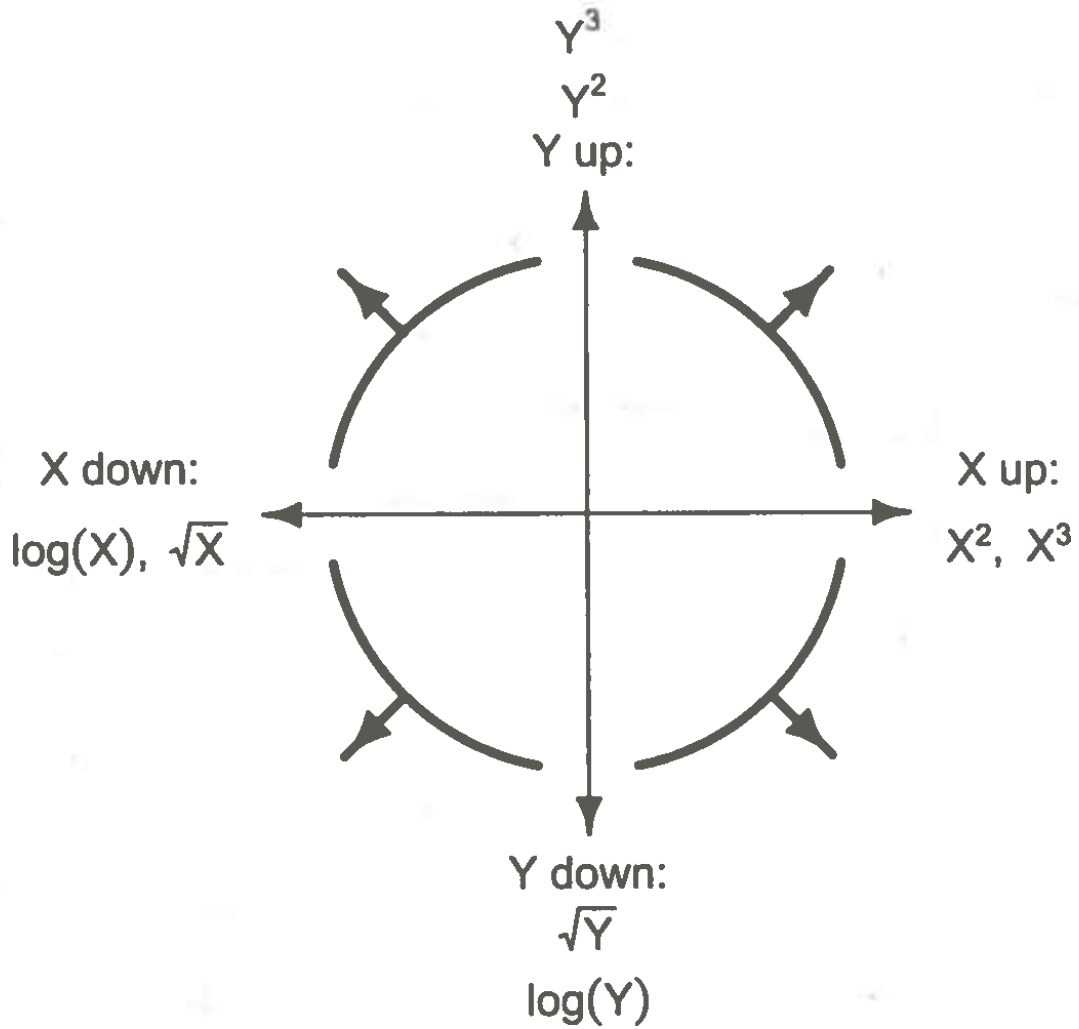
```
> install.packages("MASS")
> library("MASS")
> boxcox(lm_toluca)
```



Final Notes on Transformations

1. *a priori* transformations may be intelligently employed, e.g., regressing demand Y on commodity price X, economists favor logarithmic transforms as the slope of the regression line for the transformed variables then measures the price elasticity of demand
2. After transforming a variable (or variables), don't forget to reexamine residual plots, etc., to ensure that the model conforms to all SLR assumptions
3. When using transforms, the estimators b_0 and b_1 have least squares properties with respect to the transformed observations, not the original observations
4. When using Box-Cox, it is often recommended to round λ for easier interpretation, e.g., employing a $\lambda = 0.5$ if the Box-Cox procedure suggests a transform of $\hat{\lambda} = 0.41$

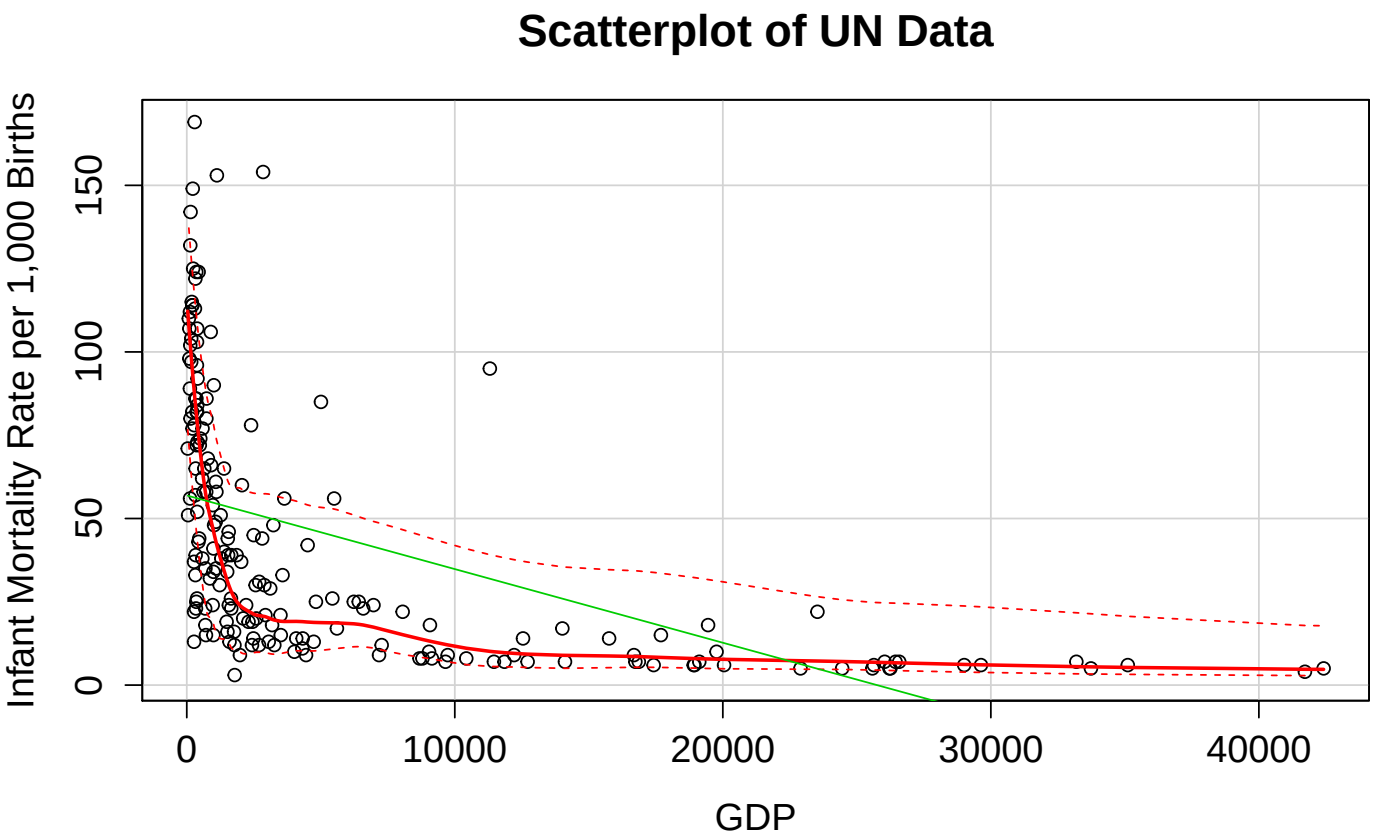
Mosteller & Tukey's Bulging Rule for Transformations



Subsection 1

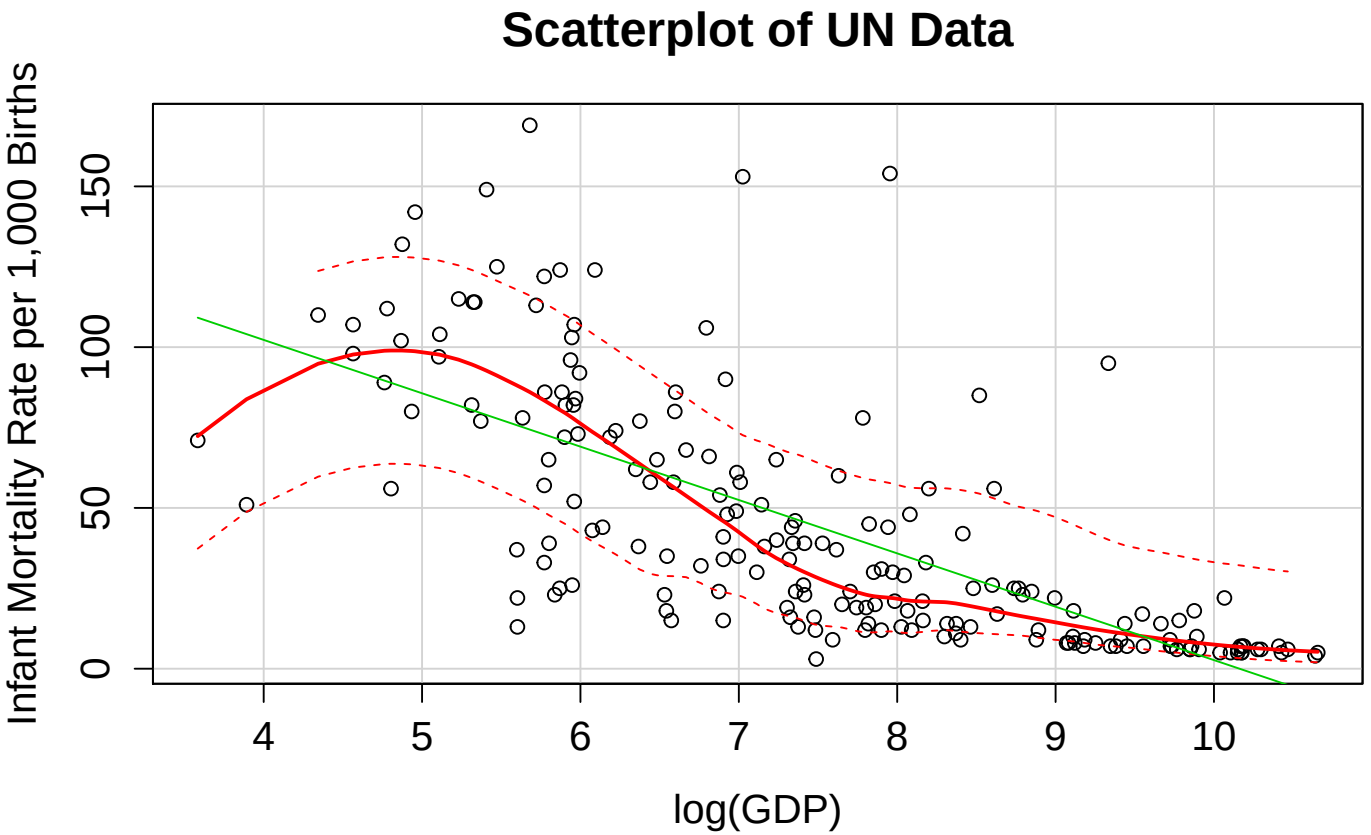
An Example

Using the `UN` data frame in the `car` package, plot `infant.mortality` on `gdp`



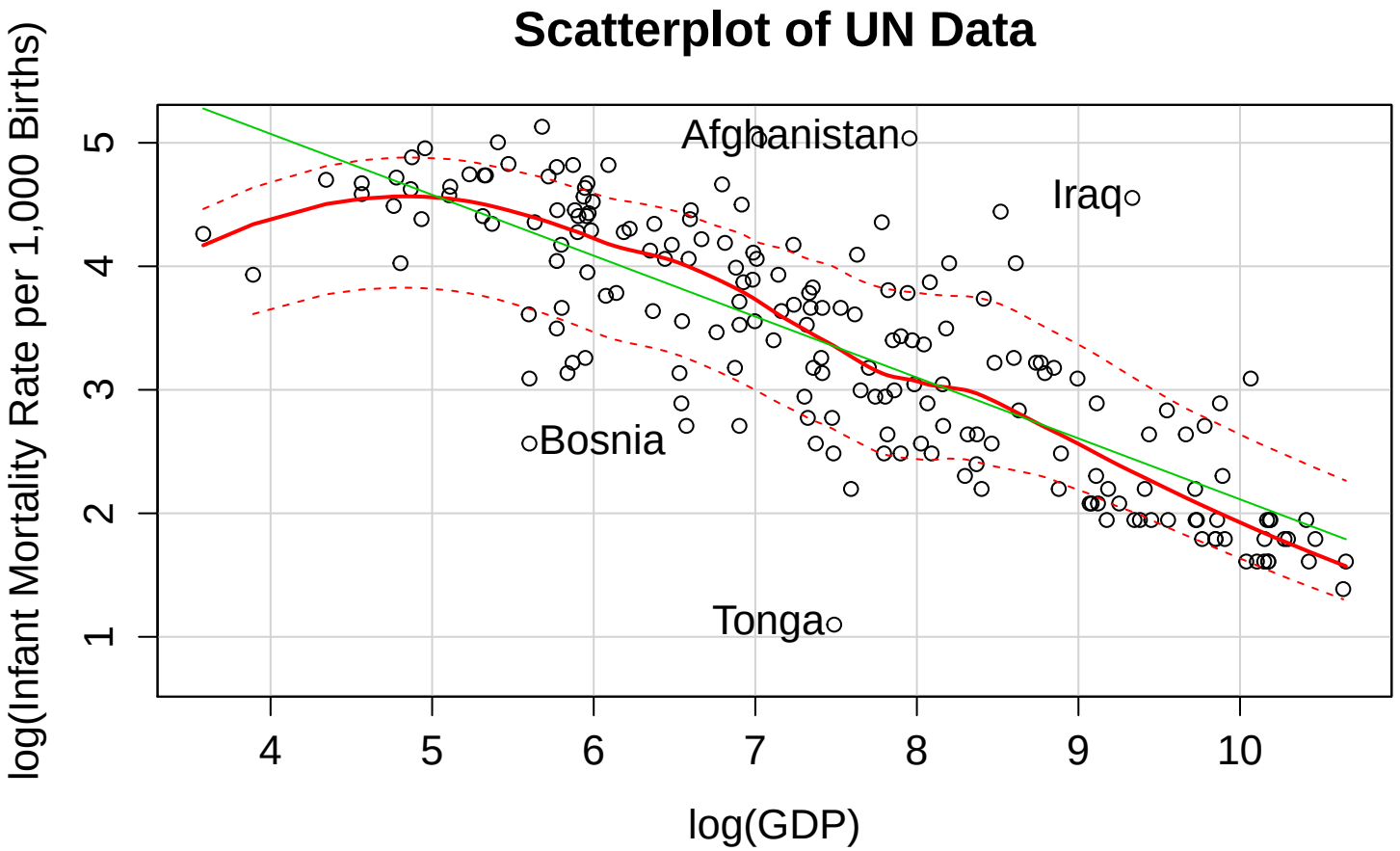

```
> scatterplot(infant.mortality~gdp,data=UN,main="Scatterplot
   of UN Data",xlab="GDP",ylab="Infant Mortality Rate
   per 1,000 Births",boxplot=FALSE)
```

Observing the non-linearity in scatter, we transform the independent variable



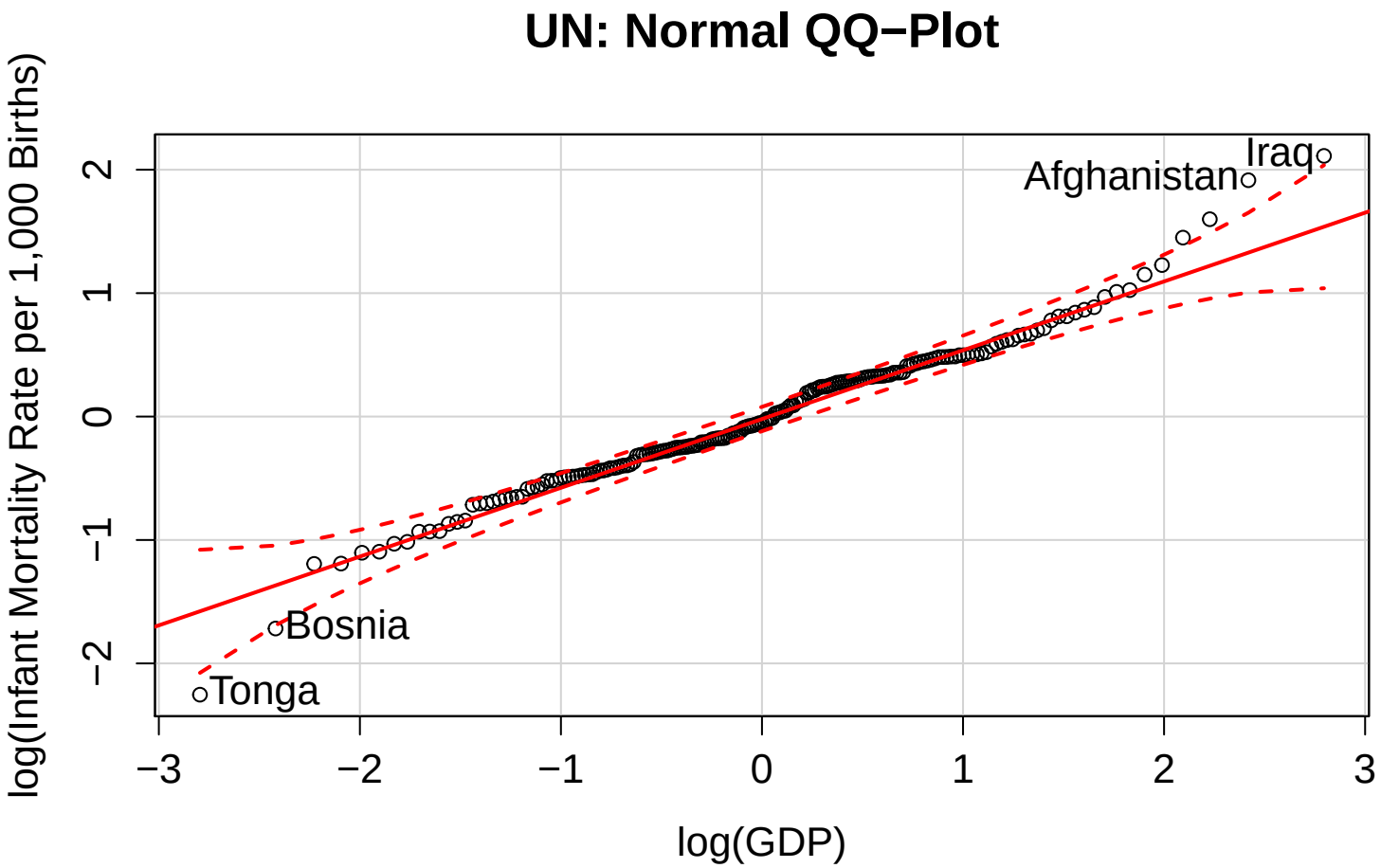

```
> scatterplot(infant.mortality~log(gdp),data=UN,main="Scatterplot
  of UN Data",xlab="log(GDP)",ylab="Infant Mortality Rate
  per 1,000 Births",boxplot=FALSE)
```

Observing heteroskedasticity, we transform the dependent variable




```
> scatterplot(log(infant.mortality)~log(gdp),data=UN,main="Scatterplot
  of UN Data",xlab="log(GDP)",ylab="log(Infant Mortality Rate
  per 1,000 Births)",boxplot=FALSE,id.n=4)
```

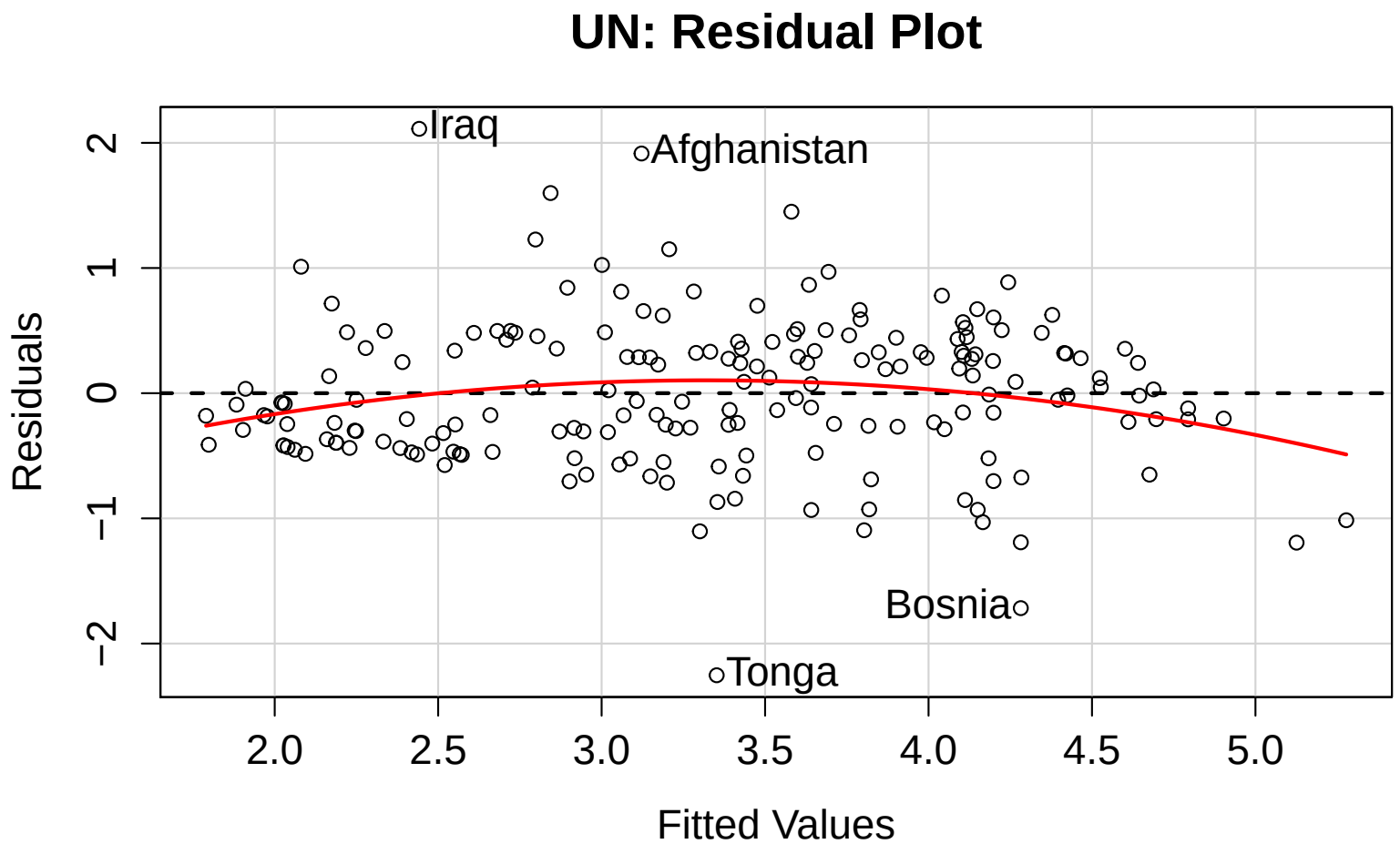
Verify the normality of the error terms.




```
> lm_UN <- lm(log(infant.mortality)~log(gdp),data=UN)

> qqPlot(residuals(lm_UN),main="UN: Normal QQ-Plot", xlab="log(GDP)",
   ylab="log(Infant Mortality Rate per 1,000 Births)",id.n=4)
```

Verify the heteroskedasticity of the error terms.




```
> residualPlot(lm_UN,xlab="Fitted Values",ylab="Residuals",
    main="UN: Residual Plot",id.n=4)
```

One last question

- If we compare countries that differ in `GDP` by 1%, by what % would `infant.mortality` change?

```
>  coef(lm_UN)
(Intercept)    log(gdp)
  7.0452008  -0.4932026
```

- The fitted regression equation is therefore

$$log(infant.mortality) = 7.045 - 0.493 \times log(gdp)$$

- Taking the exponent of each side we obtain

$$infant.mortality = \exp^{(7.045 - 0.493 \times log(gdp))}$$

$$infant.mortality = \exp^{7.045}\, gdp^{-0.493}$$

- If we increase gdp by 1%

$$infant.mortality = \exp^{7.045}(1.01 \times gdp)^{-0.493}$$

$$infant.mortality = \exp^{7.045}\, 1.01^{-0.493} gdp^{-0.493}$$

$$infant.mortality = \exp^{7.045}(0.995) gdp^{-0.493}$$

- Therefore, an increase in the gdp by 1% will result in decrease in `infant.mortality` by 0.5%
- In this type of log-log relationship, economists call β_1 an *elasticity*