Section 1

MANE 3332

Subsection 1

Regression Slides

Handouts

- Linear Regression Slides

Simple Linear Regression

- **Regression analysis** is a statistical technique for modeling and investigating the relationship between two or more variables.
- Simple linear regression considers the relationship between a single independent variable and a dependent variable
- A good tool to examine the relationship is a scatter diagram

Empirical Models

- An **empirical model** is a model that captures the relationship between regressor inputs and a response variable that is not based upon theoretical knowledge
- There are many types of empirical models
- Discuss wind-powered generator

Simple Linear Regression Model

- A simple linear regression model is shown below

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

where Y is the dependent (or response) variable, x is the independent (or regressor) variable and ε is the random error term

- We can use this model to predict Y for a given value of x

$$E(Y|x) = \mu_{Y|x} = \beta_0 + \beta_1 x$$

- Assuming ε has zero mean and variance σ^2

$$\begin{aligned} E(Y|x) = & \quad E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x + E(\varepsilon) \\ = & \quad \beta_0 + \beta_1 x \\ V(Y|x) = & \quad V(\beta_0 + \beta_1 x + \varepsilon) = V(\beta_0 + \beta_1 x) + V(\varepsilon) \\ = & \quad 0 + \sigma^2 \end{aligned}$$

- Examine the graphic shown below

Figure 11-2

Figure 11-2 on page 283

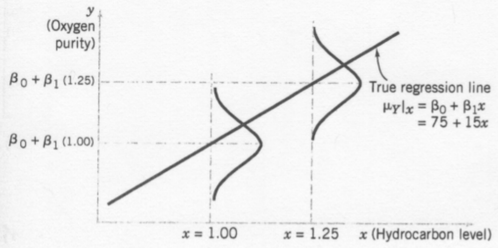


FIGURE 11.2

The distribution of Y for a given value of x for the oxygen purity-hydrocarbon data.

Figure 1: Figure 11-2

Method of Least Squares

- The method use to estimate values for β_0 and β_1 is called least squares and was developed by Gauss
- Examine figure shown below
- Minimize

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- The solution to this problem is called the least squares normal equations
- Examine the graphics shown below

Figure 11-3, page 285

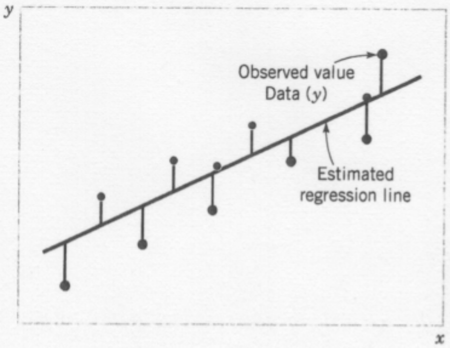


FIGURE 11.3

Deviations of the data from the estimated regression model.

Figure 2: Figure 11-3

Equations 11-7 and 11-8 on page 285

Least Squares Estimates

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11.7}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{11.8}$$

where $\bar{y} = (1/n) \sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n) \sum_{i=1}^{n} x_i$.

Figure 3: Equations

R

- In this course, we will use R to estimate the parameters and calculate sums of squares quantities
- Example Problem

In the accompanying table, x is the tensile force applied to a steel specimen in thousands of pounds, and y is the resulting elongation in thousandths of an inch:

x	1	2	3	4	5	6
y	14	33	40	63	76	85

(a) Graph the data to verify that it is reasonable to assume that the regression of Y on x is linear.

(b) Find the equation of the least squares line, and use it to predict the elongation when the tensile force is 3.5 thousand pounds.

Miller & Freund (2005). Probability & Statistics for Engineers, 7th edition

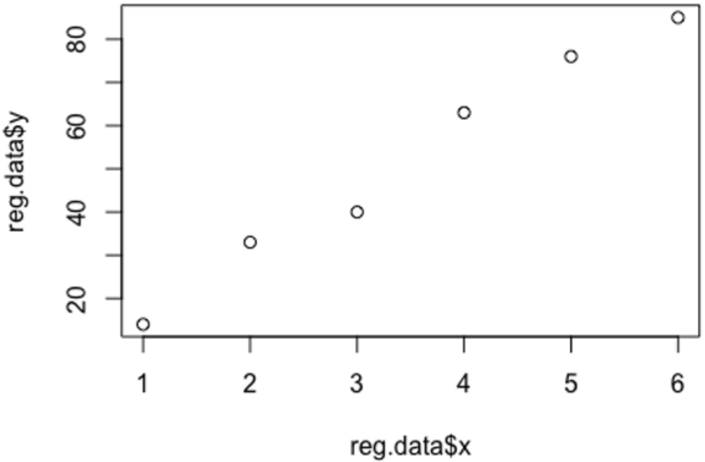
Creating Regression Data in R

```
x<-c(1,2,3,4,5,6)
y<-c(14,33,40,63,76,85)

reg.data <- data.frame(y,x)
summary(reg.data)

##          y               x       
##  Min.   :14.00   Min.   :1.00  
##  1st Qu.:34.75   1st Qu.:2.25  
##  Median :51.50   Median :3.50  
##  Mean   :51.83   Mean   :3.50  
##  3rd Qu.:72.75   3rd Qu.:4.75  
```

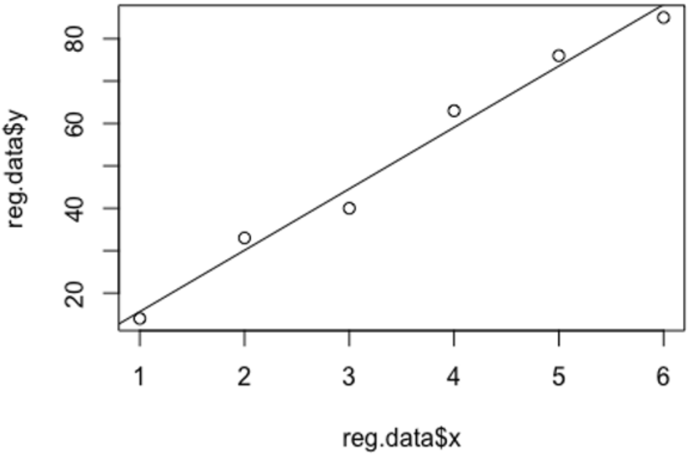
```
plot(reg.data$x,reg.data$y)
```



```
reg.model <- lm(y~x,data=reg.data)
summary(reg.model)

##
## Call:
## lm(formula = y ~ x, data = reg.data)
##
## Residuals:
##      1      2      3      4      5      6
## -1.619  2.895 -4.590  3.924  2.438 -3.048
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.1333     3.6859   0.307 0.773825
## x            14.4857     0.9465  15.305 0.000106 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.959 on 4 degrees of freedom
## Multiple R-squared:  0.9832, Adjusted R-squared:  0.979
## F-statistic: 234.2 on 1 and 4 DF,  p-value: 0.0001063
```

```
plot(reg.data$x,reg.data$y)
abline(reg.model)
```



Hypothesis Test

- It is possible to perform a hypothesis involving the slope parameter, β_1

$$\begin{aligned} H_0 : \beta_1 = & \quad \beta_{1,0} \\ H_1 : \beta_1 \neq & \quad \beta_{1,0} \end{aligned}$$

where $\beta_{1,0}$ is a constant (often 0).

- Requires the assumption that $\varepsilon \sim \mathrm{NID}(0, \sigma^2)$
- The test statistic is a t-random variable

$$t_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$$

- A similar test can be formed for β_0

```
reg.model <- lm(y~x,data=reg.data)
summary(reg.model)

##
## Call:
## lm(formula = y ~ x, data = reg.data)
##
## Residuals:
##      1      2      3      4      5      6
## -1.619  2.895 -4.590  3.924  2.438 -3.048
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.1333     3.6859   0.307 0.773825
## x            14.4857     0.9465  15.305 0.000106 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.959 on 4 degrees of freedom
## Multiple R-squared:  0.9832, Adjusted R-squared:  0.979
## F-statistic: 234.2 on 1 and 4 DF,  p-value: 0.0001063
```

Examining Model Adequacy

- Two major concerns
  - Does the model provide an adequate explanation of the data?
  - Are the model assumptions satisfied?

Residual Analysis

- The residuals are defined to be

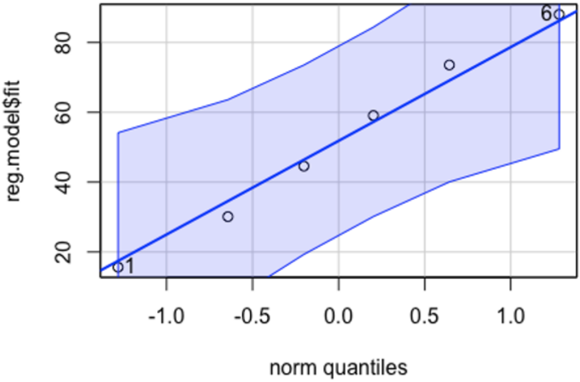
$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x$$

- Examine normality assumption by generating a normal probability plot of residuals

```
library(car)

## Loading required package: carData

qqPlot(reg.model$fit)
```

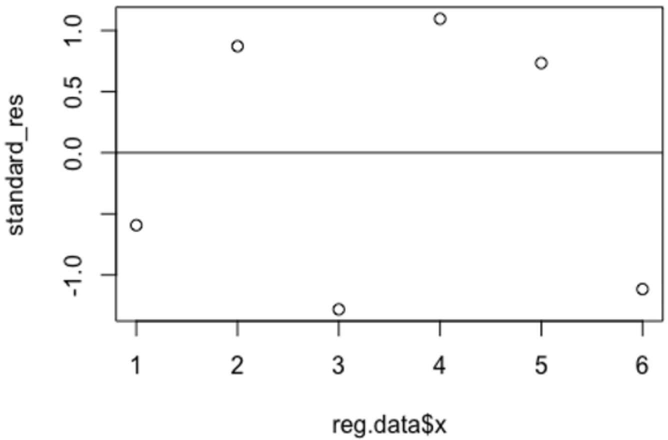



```
## [1] 1 6
```

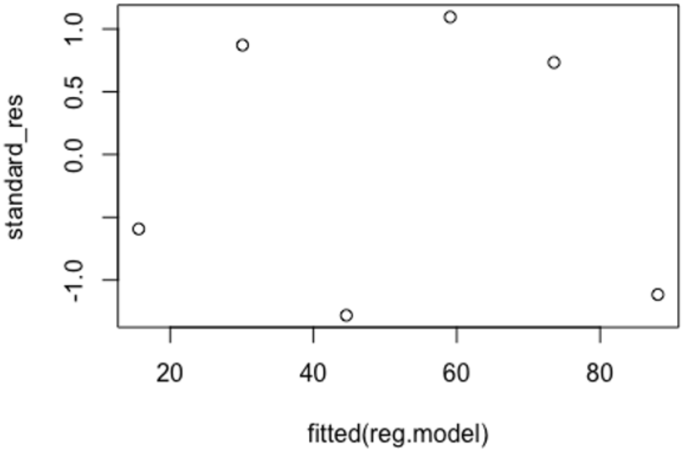
Residual Analysis - Constant Variance

- Examine the assumption of constant variance by plotting residuals versus fitted values and residuals vs x
- Examine if additional terms are required (such as quadratic) by examining residuals vs x
- Residuals are often standardized

```
standard_res <- rstandard(reg.model)
plot(reg.data$x,standard_res)
abline(0,0)
```



```
plot(fitted(reg.model),standard_res)
```



Lack of Fit Test

- If there are repeated observations (identical values of x) a lack of fit test can be performed

$$\begin{aligned} H_0 &: \quad \text{The model is correct} \\ H_1 &: \quad \text{The model is NOT correct} \end{aligned}$$

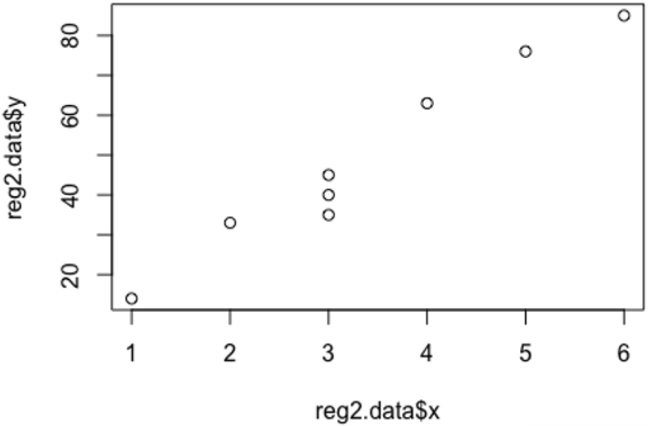
- The repeated observations allows the SS_E error term to be partitioned

$$SS_E = SS_{PE} + SS_{LOF}$$

- The test statistic is

$$F_0 = \frac{MS_{LOF}}{MS_{PE}}$$

```
x2<-c(1,2,3,4,5,6,3,3)
y2<-c(14,33,40,63,76,85,35,45)
reg2.data <- data.frame(y2,x2)
plot(reg2.data$x,reg2.data$y)
```



```
reg2.model <- lm(y2~x2,data=reg2.data)
summary(reg2.model)
```

```
##
## Call:
## lm(formula = y2 ~ x2, data = reg2.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3706 -2.6469  0.8077  3.5315  4.9510
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.6643     4.2719  -0.156    0.882
## x2           14.6783     1.1573  12.683 1.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.893 on 6 degrees of freedom
## Multiple R-squared:  0.964,  Adjusted R-squared:  0.958
## F-statistic: 160.9 on 1 and 6 DF,  p-value: 1.473e-05
```

```
anovaPE(reg2.model)
```

```
##              Df Sum Sq Mean Sq  F value   Pr(>F)
## x2            1 3851.2  3851.2 154.0490 0.006429 **
## Lack of Fit   4   93.7    23.4   0.9365 0.574984
## Pure Error    2   50.0    25.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```