Attendance : 1-C

MANE 3332.01

Lecture 27

Agenda

- Linear Regression Lecture (Chapters 11 and 12)
- Chapter 9, Case 2 2-sided Quiz (assigned 12/2/2025, due 12/4/2025)
- Chapter 9, Case 2 Lower Quiz (assigned 12/2/2025, due 12/4/2025)
- Chapter 9, Case 2 Upper Quiz (assigned 12/2/2025, due 12/4/2025)
- Chapter 9, Case 3 2-sided Practice Problems (assigned 12/2/2025, due 12/4/2025)
- Chapter 9, Case 3 Lower Practice Problems (assigned 12/2/2025, due 12/4/2025)
- Chapter 9, Case 3 Upper Practice Problems (assigned 12/2/2025, due 12/4/2025)
- NEW: Chapter 9, Case 3 2-sided Quiz (assigned 12/4/2025, due 12/9/2025)
- NEW: Chapter 9, Case 3 Lower Quiz (assigned 12/4/2025, due 12/9/2025)
- New: Chapter 9, Case 3 Upper Quiz (assigned 12/4/2025, due 12/9/2025)
- Attendance
- Questions?

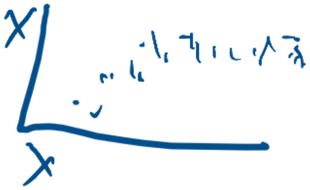
Handouts

- Regression slides
- Regression slides marked

Class Schedule

Week	Tuesday Lecture	Thursday Lecture
14	**12/2** - Chapter 9, Case 3	**12/4** - Linear Regression
15	**12/9** - Review Session	**12/11** - Study Day (no class)

The final exam for MANE 3332.01 is **Thursday December 18, 2025 at 10:15 AM - 12:00 PM**.



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
  x y
- 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

Splines
Support Vector Machines

Neural network

β - beta $\qquad y = a + mx$

ε - epsilon

Simple Linear Regression Model

- A simple linear regression model is shown below

Theoretical model

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

error term

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

ε is not observable

- 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

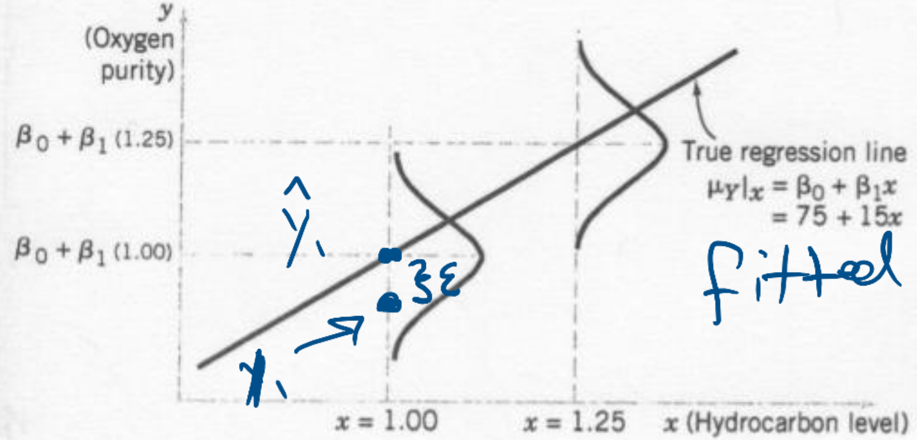
$$\begin{aligned} E(Y|x) &= E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x + E(\varepsilon) \\ &= \beta_0 + \beta_1 x \\ V(Y|x) &= V(\beta_0 + \beta_1 x + \varepsilon) = V(\beta_0 + \beta_1 x) + V(\varepsilon) \\ &= 0 + \sigma^2 \end{aligned}$$

- Examine the graphic shown below

normality is not needed to find estimates for β_0 & β_1

Figure 1

Figure 1



fitted line

FIGURE 11.2

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

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

Figure 11-2

Method of Least Squares

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

- 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

minimized the squared error of the estimate

- Minimize

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

$(y_i - \hat{y}_i)$

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

minimize L

$$\frac{dL}{d\beta_0} \rightarrow \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^{2-1}(-1)$$

$$\frac{dL}{d\beta_1} \rightarrow \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^{2-1}(-x_i)$$

Figure 11-3, page 285

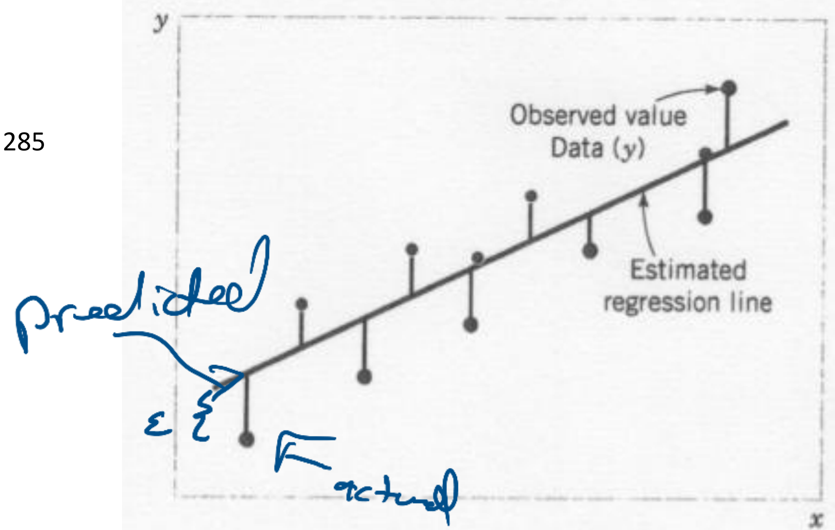


FIGURE 11.3

Deviations of the data from the estimated regression model.

Figure 11-3

Equations 11-7 and 11-8 on page

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

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

Example Problem

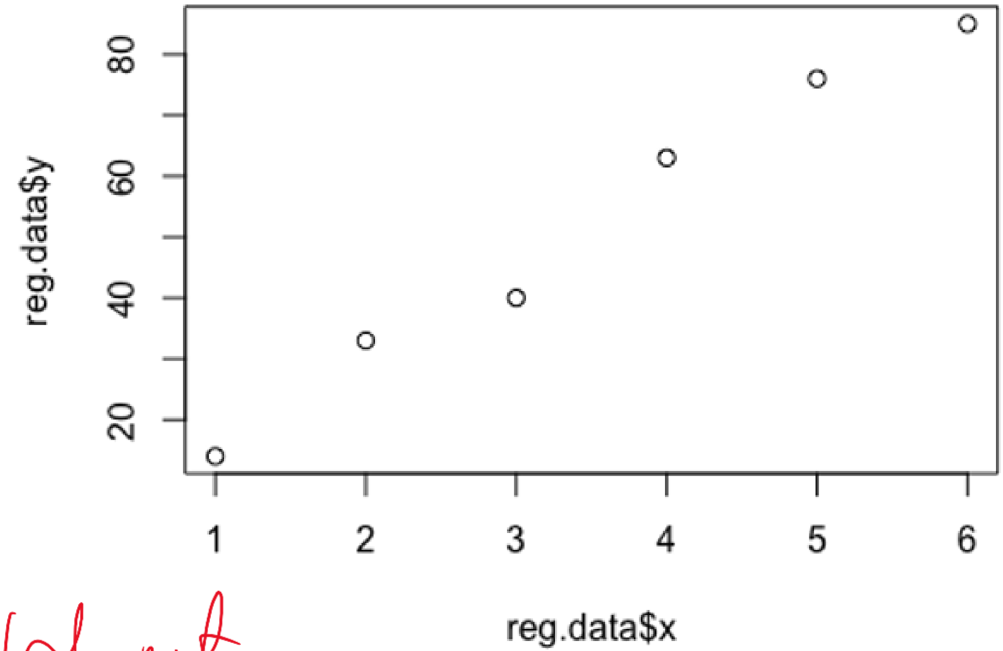
Creating Regression Data in R

```
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
##  Max.   :85.00   Max.   :6.00
```

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



You should not extrapolate linear regression results

```r
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
```

$y = 1.1333 + 14.4857x$

β

p-value

β_0

β_1

Residual

$$y = \beta_0 + \beta_1 X + \varepsilon$$ (theoretical model)

fomd $y = \hat{\beta}_0 + \hat{\beta}_1 X$ (fitted model)

$$\hat{\varepsilon}_i = e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

actual value

fitted ~~model~~ value

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

$\hat{\beta}_0 = 1.1333$

$\hat{\beta}_1 = 14.49$

$x_1 = 1, \; y_1 = 14$

$$e_1 = y_1 - (\hat{\beta}_0 + \hat{\beta}_1 x_1)$$

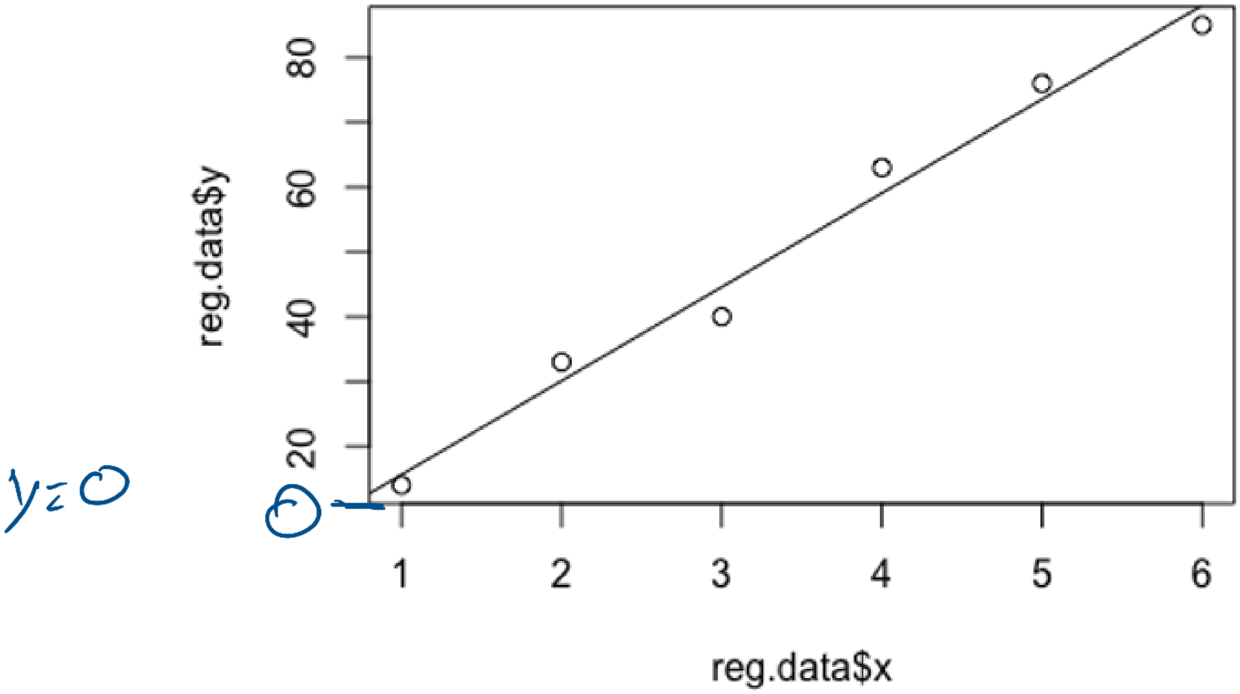
$$= 14 - (1.13333 + 14.49(1))$$

actual

$$= 14 - 15.6233 \leftarrow \text{pred}$$

$$= -1.6233$$

```
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

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

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

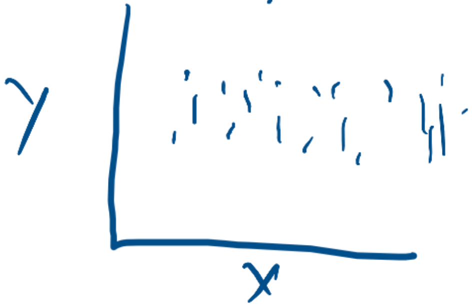
- Requires the assumption that $\varepsilon \sim \mathrm{NID}(0, \sigma^2)$
- The test statistic is a t-random variable

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

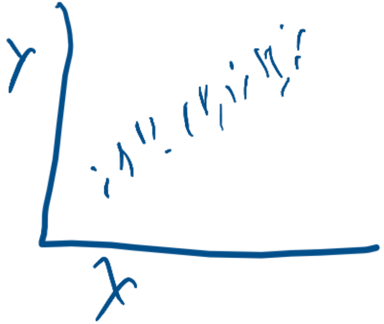
- A similar test can be formed for β_0

Normal and independently Distributed with constant Variance

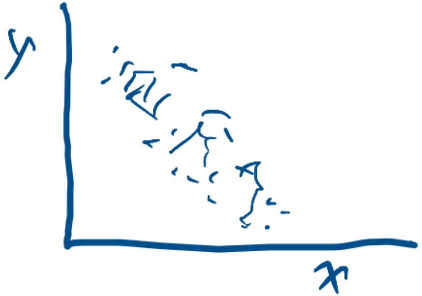
$\beta_1 = 0$



$\beta_1 > 0$



$\beta_1 < 0$ $\quad y = \beta_0 + \beta_1 x$



```
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?

model is linear
What is the shape of the data?

Hypothesis testing
- residuals are normal
- residuals are independent
- residuals have constant variance
- no pattern in residuals

all assumption checks are done with residual

Residual Analysis

- The residuals are defined to be

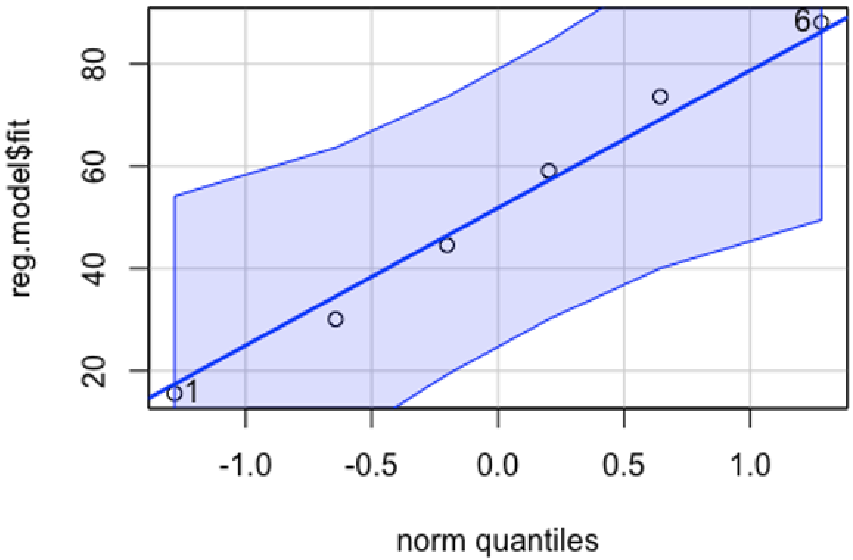
$$\hat{\varepsilon}_i = \quad e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x$$

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

```
library(car)

## Loading required package: carData

qqPlot(reg.model$fit)
```



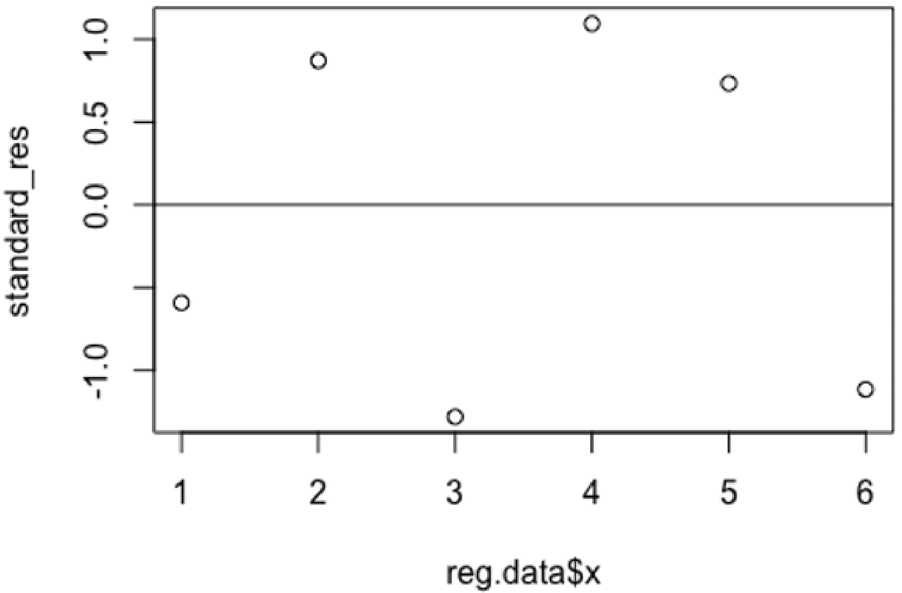

```
## [1] 1 6
```

Residual Analysis - Constant Variance

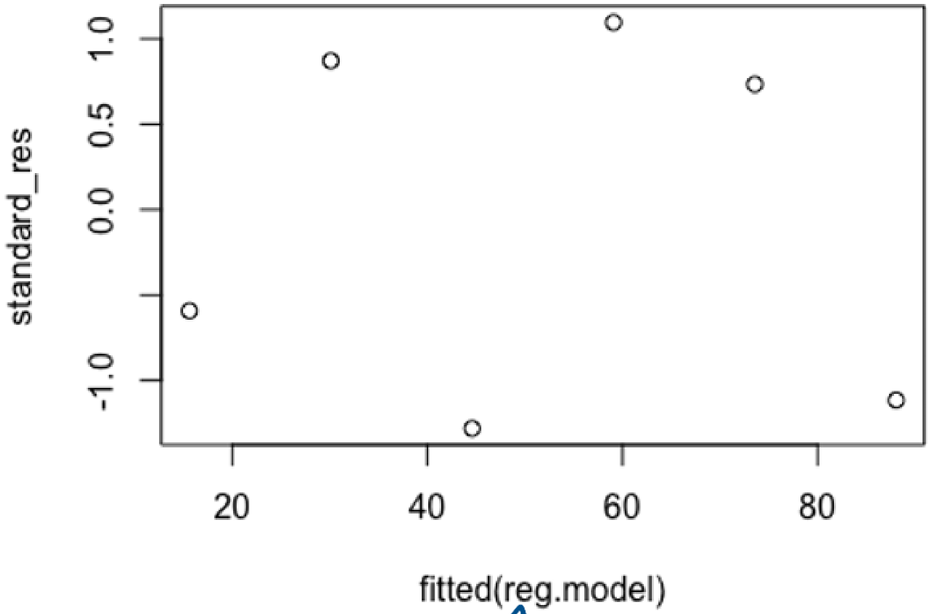
- Examine the assumption of constant variance by plotting residuals versus fitted values and residuals vs x
- Examine if additional terms are required (such as quadratic) by examining residuals vs x
- Residuals are often standardized

$$\frac{e_i}{\sigma_{e_i}}$$

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

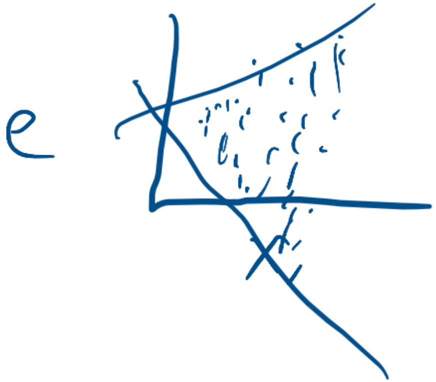


```
plot(fitted(reg.model),standard_res)
```



$\hat{y}$

Non-Constant Variance



same x, differently

Lack of Fit Test

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

H_0 : The model is correct no LOF

H_1 : The model is NOT correct LOF in model

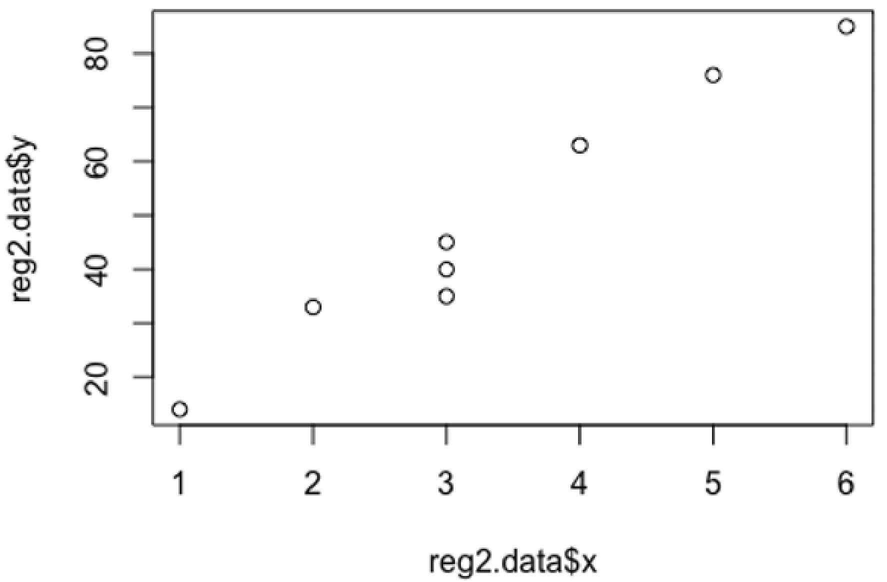
- The repeated observations allows the SS_E error term to be partitioned

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

- 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
```

if $\alpha = .05$

is p-value $< \alpha$

? no → Fail to reject Ho

no LOF in model