

MANE 3332.03

Section 1

MANE 3332.03

1 / 29

Lecture 26, April 28

- Topics:
 - Linear Regression (Chapters 11 and 12)
- Assignments:
 - **New: Chapter 9, Case 3 Upper Quiz (assigned 4/29/2025, due 5/1/2025)**
 - **New: Chapter 9, Case 3 Lower Quiz (assigned 4/29/2025, due 5/1/2025)**
 - **New: Chapter 9, Case 3 2-sided Quiz (assigned 4/29/2025, due 5/1/2025)**
 - Technical Report Two (assigned 4/17/2025, due 5/15/2025)
- Attendance
- Questions?

Handouts

- Regression slides
- Regression slides marked

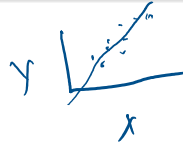
Class Schedule

Tuesday Lecture	Thursday Lecture
4/29: Chapter 11	5/1: Chapter 11
5/6: Review	5/8: Dead Day (no class)

3 Sessions plus final exam

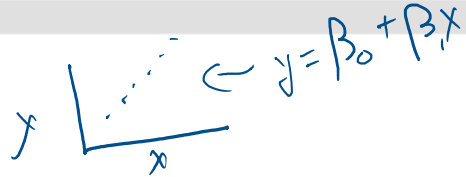
Final Exam: Tuesday May 13, 2025 10:15 am - 12:00 pm

Technical Report Two: Thursday May 15, 2025 11:59 pm



Simple Linear Regression *one x, one y*

- **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



$$y = \beta_0 + \beta_1 x$$

Empirical Models

$$y = f(x)$$

- 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

At: neural networks
 Support Vector Machines (SVM)
 Multiple Adaptive Regression Spline (MARS)

Mechanistic Models

Are based on the physics (science)
 and engineering

- Hooke's law
- Ohm's law

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

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

we want to estimate β_0 & β_1

β_0 - y-intercept
 β_1 - slope

- Assuming ε has zero mean and ^{Constant} variance σ^2

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

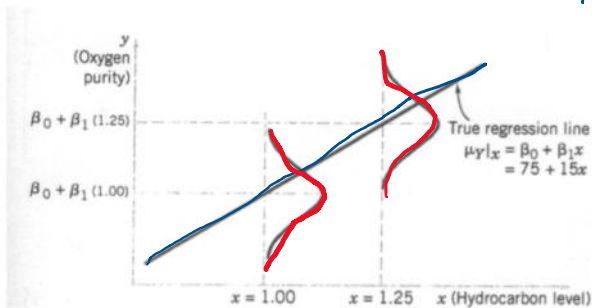
- Examine the graphic shown below

Forget normality

Figure 11-2

Figure 11-2 on page 283

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



Normality is not
required to find
 β_0 & β_1

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 (L) and was developed by Gauss
- Examine figure shown below
- Minimize

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

minimize total square error

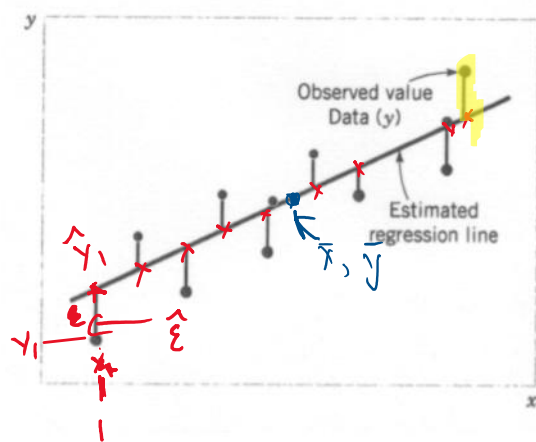
$$L = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

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

$$\frac{dL}{d\beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^{2-1} = 0$$

$$\frac{dL}{d\beta_1} = -2 \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i)^{2-1} = 0$$

Figure 11-3, page 285

**FIGURE 11.3**

Deviations of the data from the estimated regression model.

$\hat{y}$ - fitted or predicted value from 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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (11.7)$$

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

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

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

Figure 4: Example Problem

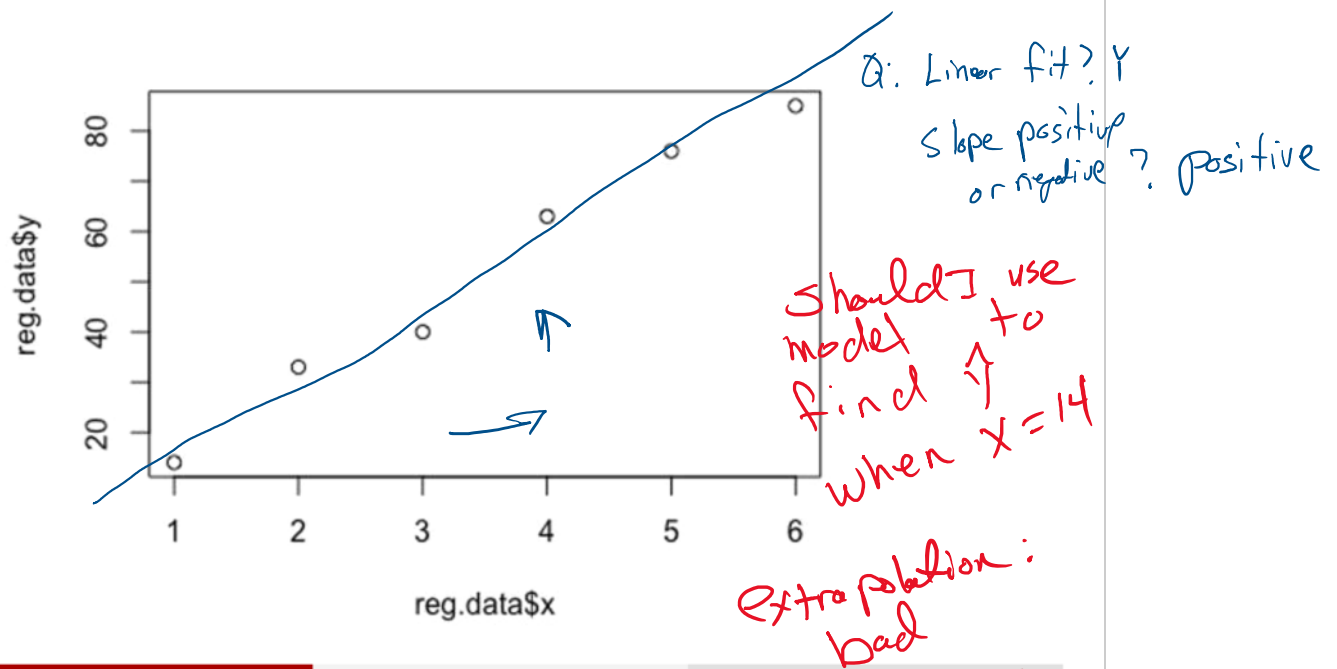
Creating Regression Data in R

```
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)
```



linear model
model

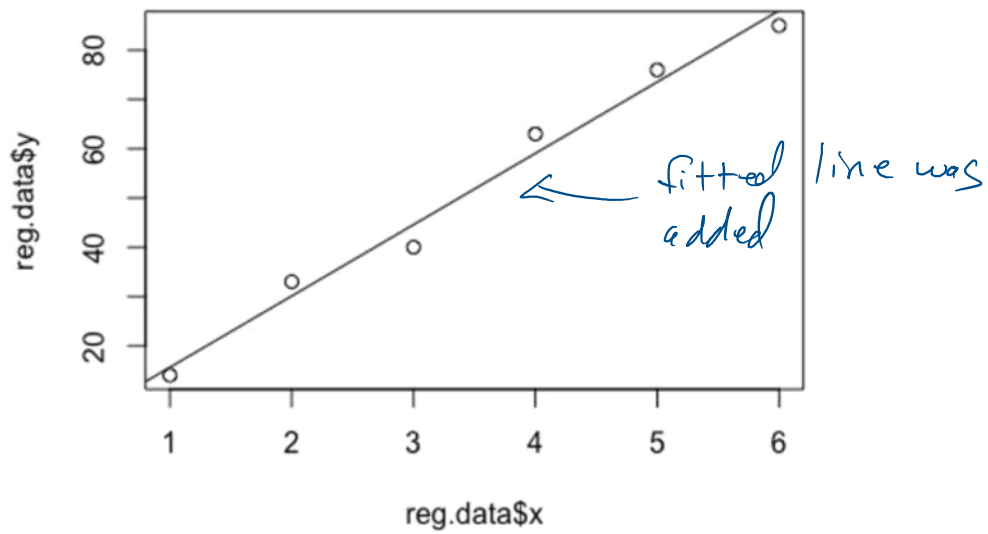
```
reg.model <- lm(y~x, data=reg.data)
summary(reg.model)
```

```
##
## Call:
## lm(formula = y ~ x, data = reg.data)
##
## Residuals: come back
##      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
```

$y = \beta_0 + \beta_1 x + \epsilon$
 (intercept) $\rightarrow$ info on $\hat{\beta}_0$
 $\hat{\beta}_0 = 1.1333$

$x \rightarrow$ info on $\hat{\beta}_1$
 Fitted model
 $\hat{y} = 1.1333 + 14.4857x$

```
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

$$H_0 : \beta_1 = \beta_{1,0} = 0$$

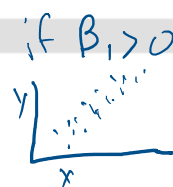
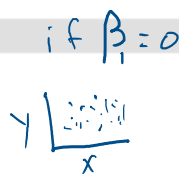
$$H_1 : \beta_1 \neq \beta_{1,0}$$

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

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

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

- A similar test can be formed for β_0



~~NID~~
→ independent
identical

Normal and
independent

```
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
```

$$e_i = y_i - \hat{y}_i \quad \text{note } x_i = 1$$

$$= 14 - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$= 14 - (1.1333 + 14.4857(1))$$

$$= -1.619$$

To fit model $(\hat{\beta}_0, \hat{\beta}_1)$ → No assumptions
 To test $H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$ → Requires IID

Examining Model Adequacy

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

Linear model is the data linear



Assumption checking

- 1) is linear model correct
- 2) ϵ is $NID(0, \sigma^2)$

Formal model $y = \beta_0 + \beta_1 x + \varepsilon \rightarrow \text{unobservable}$

$$\hat{\varepsilon} = y - \hat{y}_{\text{predicted}}$$

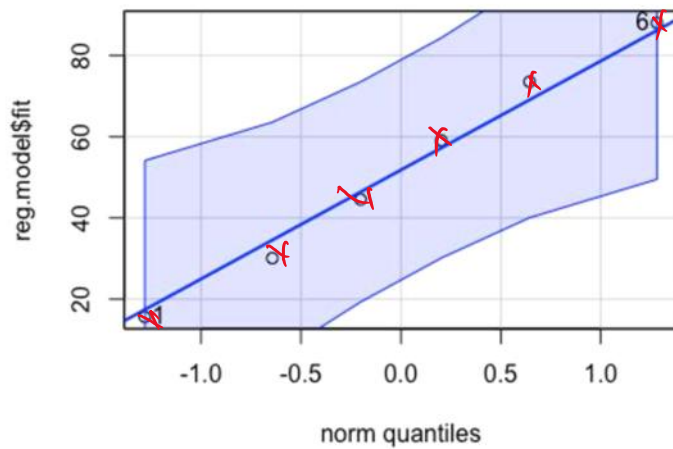
Residual Analysis

- 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
```

H_0 : Data is from normal

H_1 : Data is Not ~~Normal~~

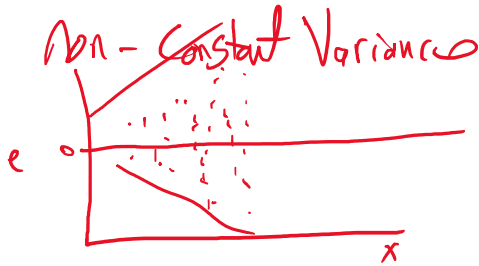
Reject H_0 if one or more points are outside bound.

Conclusion: 10 points are outside, fail to reject H_0 → Normal

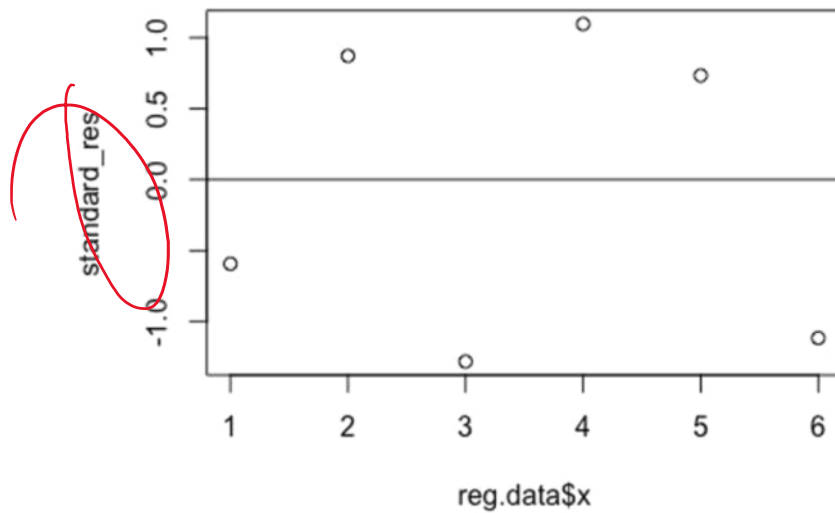
Residual Analysis - Constant Variance



- 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

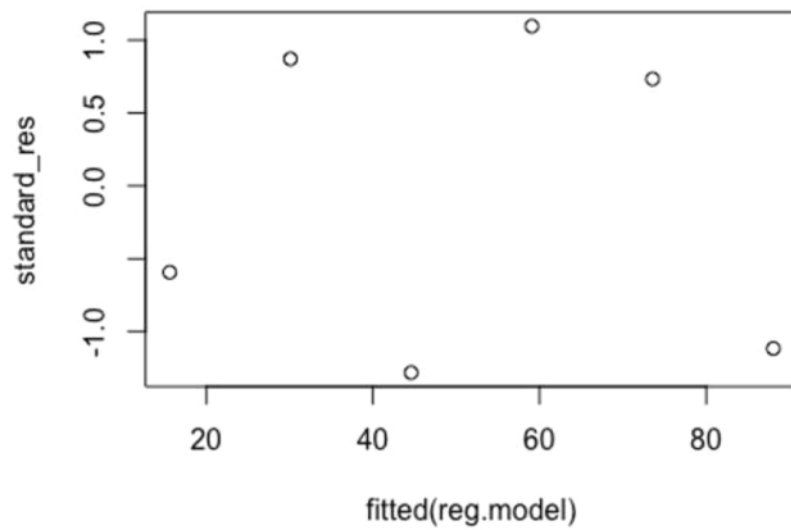


```
standard_res <- rstandard(reg.model)
plot(reg.data$x, standard_res)
abline(0,0)
```



residuals
have variance σ^2
* standerdized Residual
 $\frac{\sigma}{\sqrt{n}}$

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



Good
residual
plot
no errors

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 lack of fit
 H_1 : The model is NOT correct, lack of fit

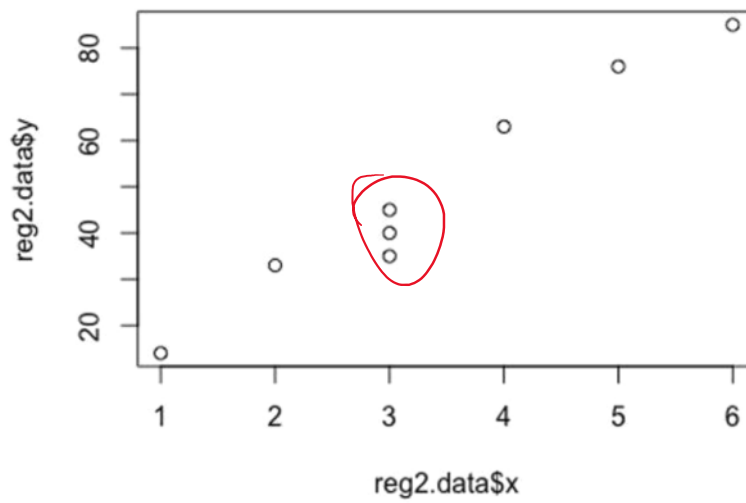
- The repeated observations allows the SS_E error term to be partitioned

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

- 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 0.006428 **
## 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

```

Analysis of Variance (ANOVA)

p-value

Is there lack of fit using $\alpha = .05$?

is $p\text{-value} < \alpha$? so is $.575 < .05$? No

there is
no lack of
fit

what is y when $X = 3.5$?

$$\hat{y} = 1.1333 + 14.4857(3.5)$$

$$= \hat{\beta}_0 + \hat{\beta}_1 X$$