Linear Regression

Regression

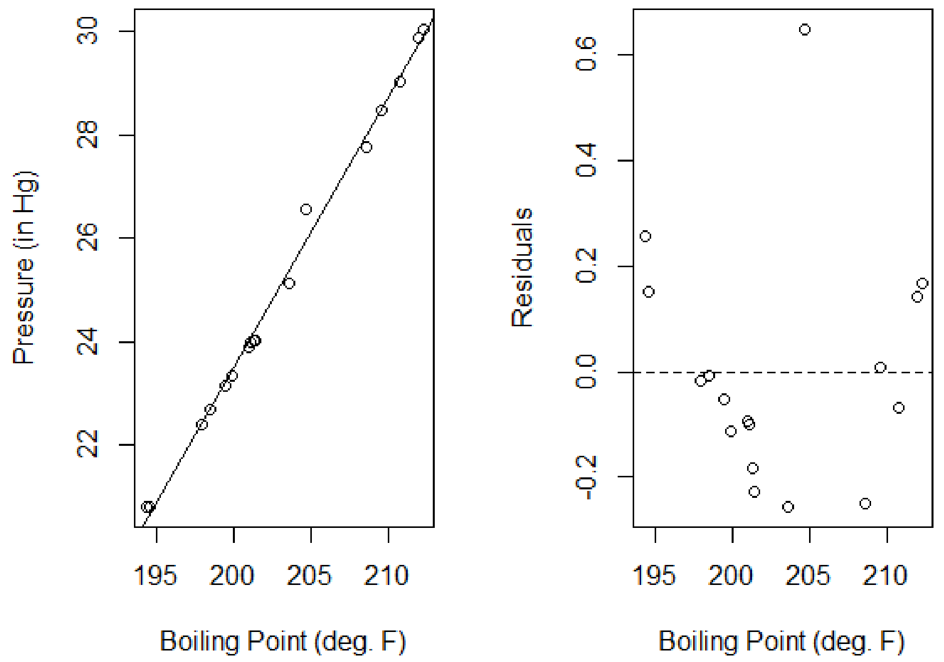
- Model a continuous variable *Y* as a mathematical function of one or more *X* variables

- The model can predict *Y* when only *X* is known

- The basic model is: $Y = \beta_1 + \beta_2 X + \epsilon$

Assumptions

- Linearity
- Normally distributed
- Homogeneity of Variance
- Independence

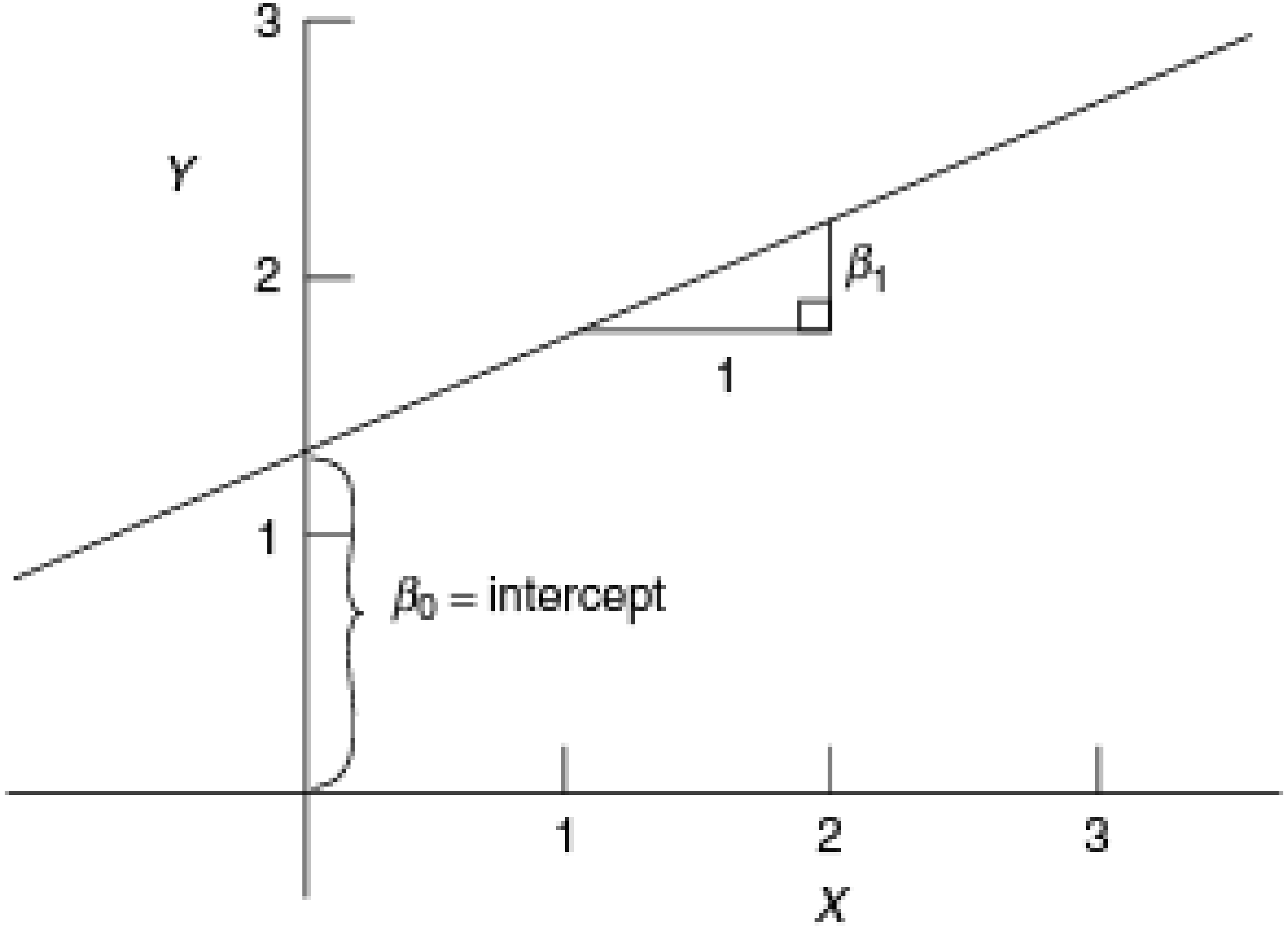


Simple Linear Regression

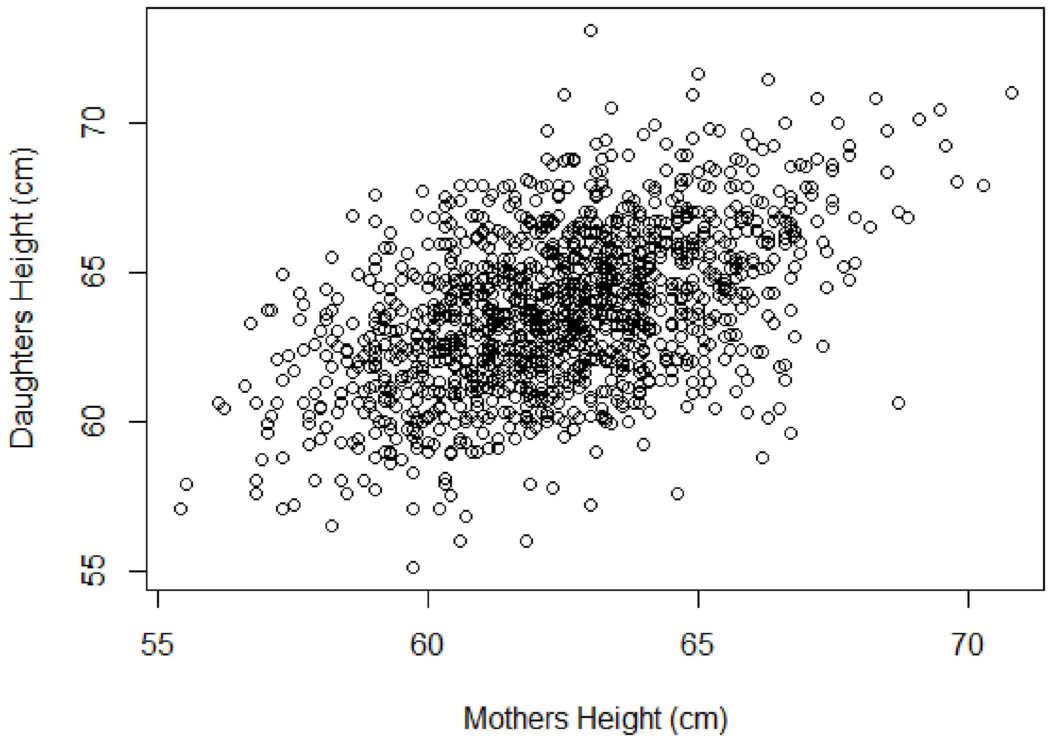
$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

$$Var(Y \mid X = x) = \sigma^2$$

Mechanics of Regression



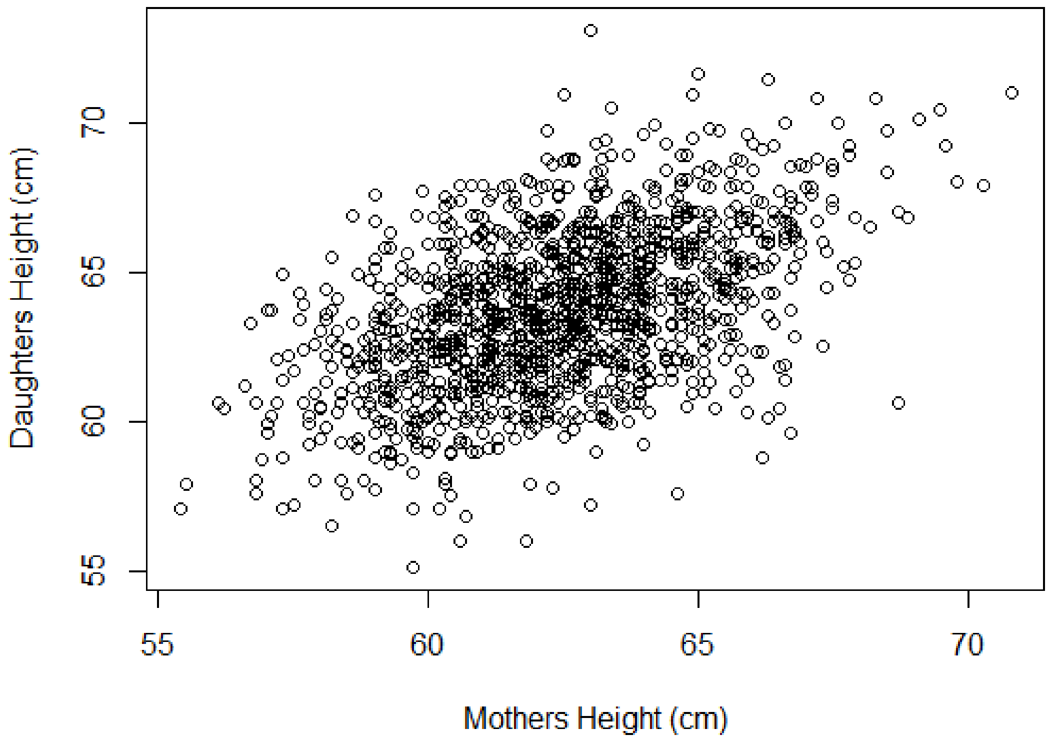
Scatterplots and Regression



E (daughter height | mother height = x)= $\beta_0 + \beta_1 x$

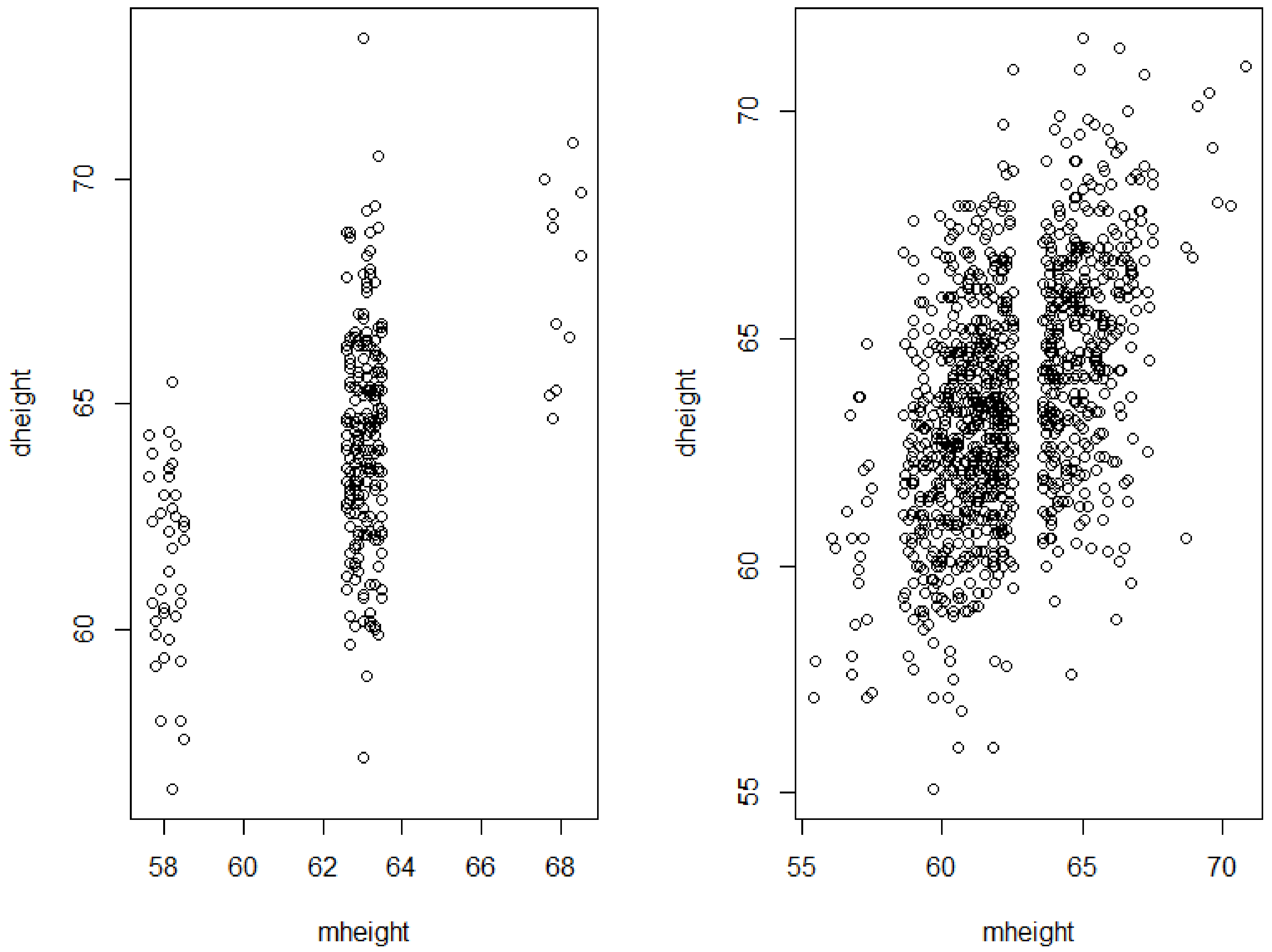
Weisberg, S. (2014). ***Applied Linear Regression, fourth edition***, Hoboken NJ: John Wiley. The **alr4 package** available from CRAN contains the data for the book.

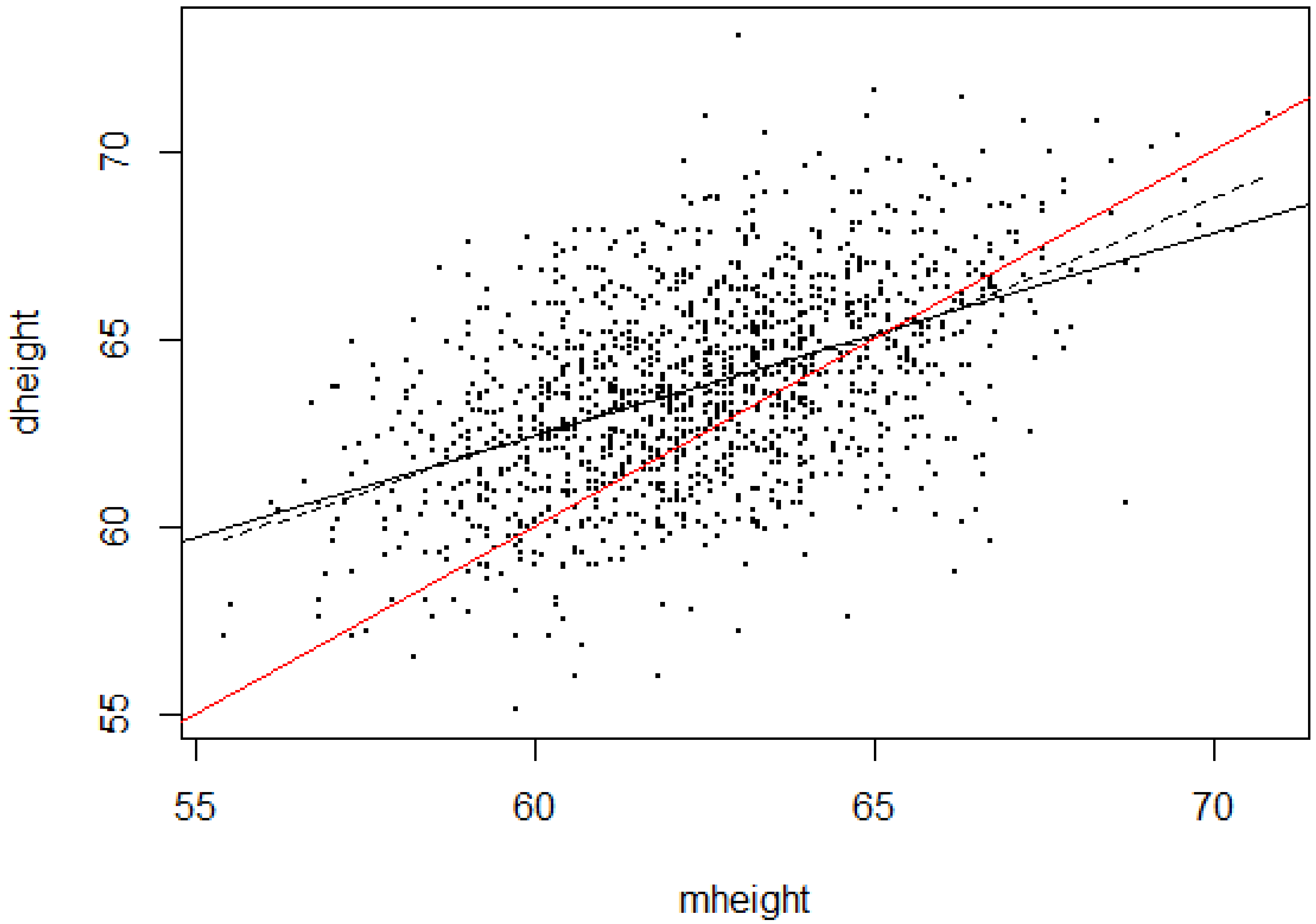
Practice → Draw a regression line



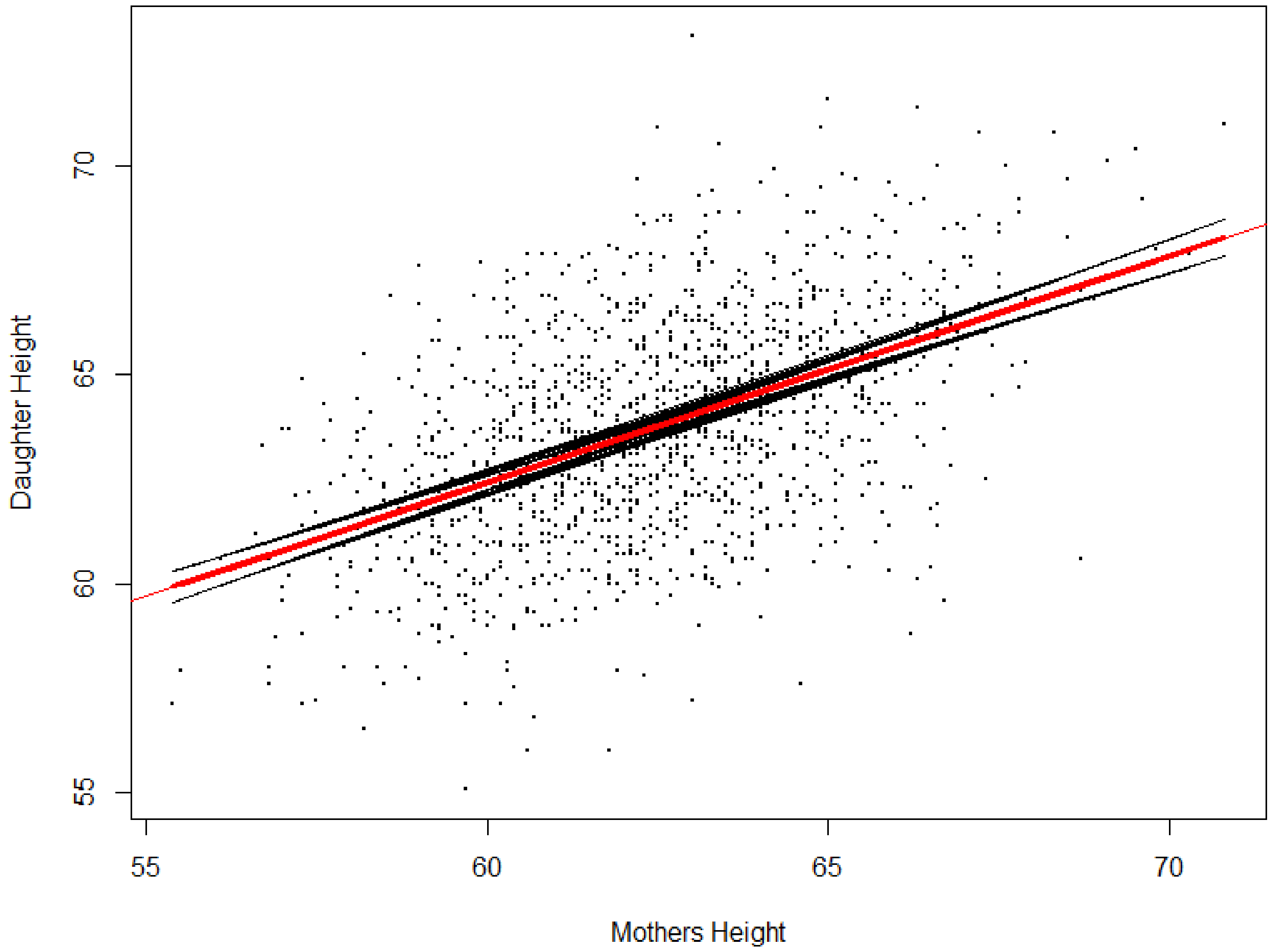
E (daughter height | mother height = x)= $\beta_0 + \beta_1 x$

Weisberg, S. (2014). ***Applied Linear Regression, fourth edition***, Hoboken NJ: John Wiley. The **alr4 package** available from CRAN contains the data for the book.

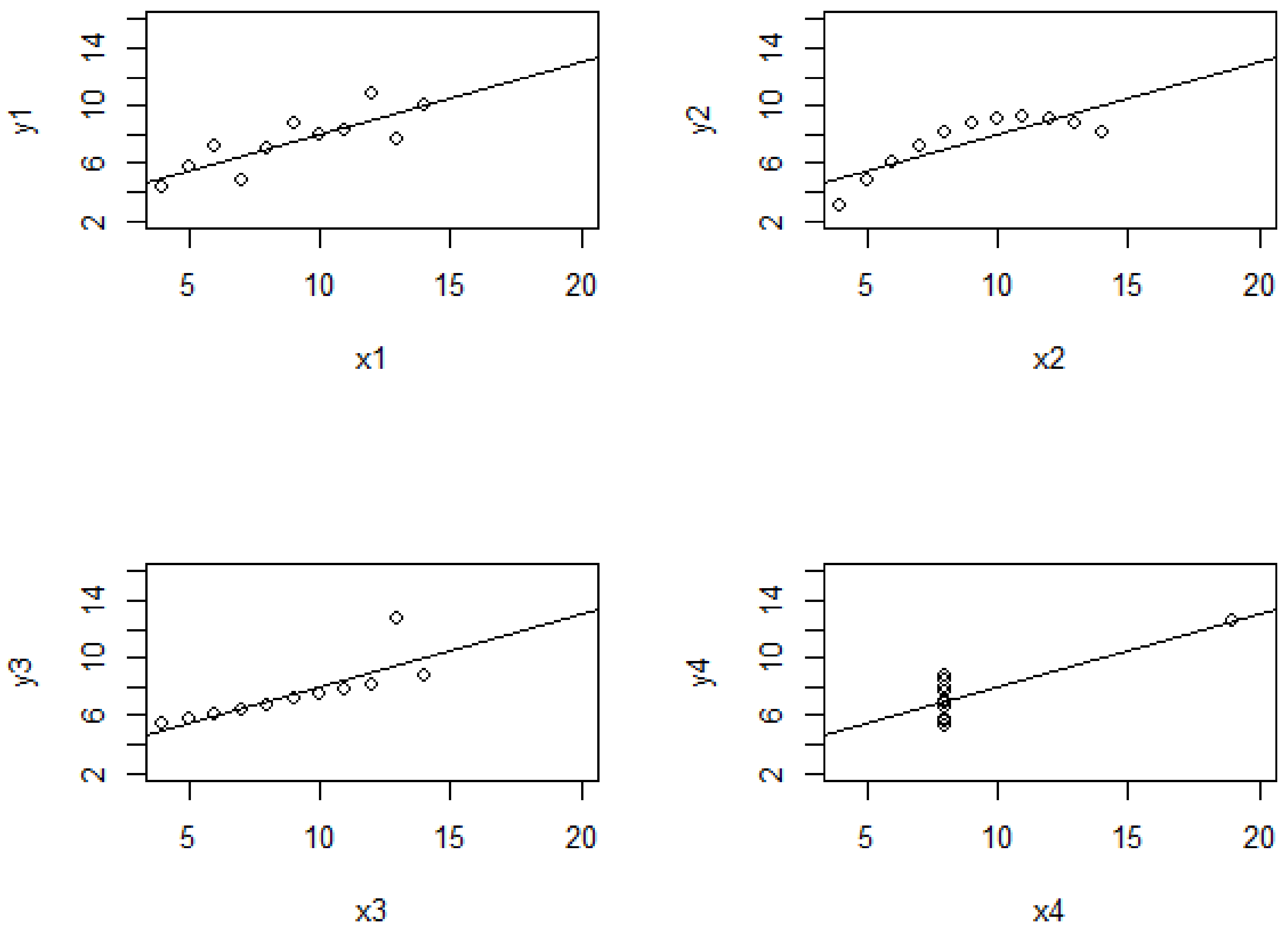

dheight
70
65
60
58
60
62
64
66
68
mheight
dheight
70
65
60
55
55
60
65
70
mheight



dheight
70
65
60
55
55
60
65
70
mheight



Daughter Height
Mothers Height
55
60
65
70
55
60
65
70

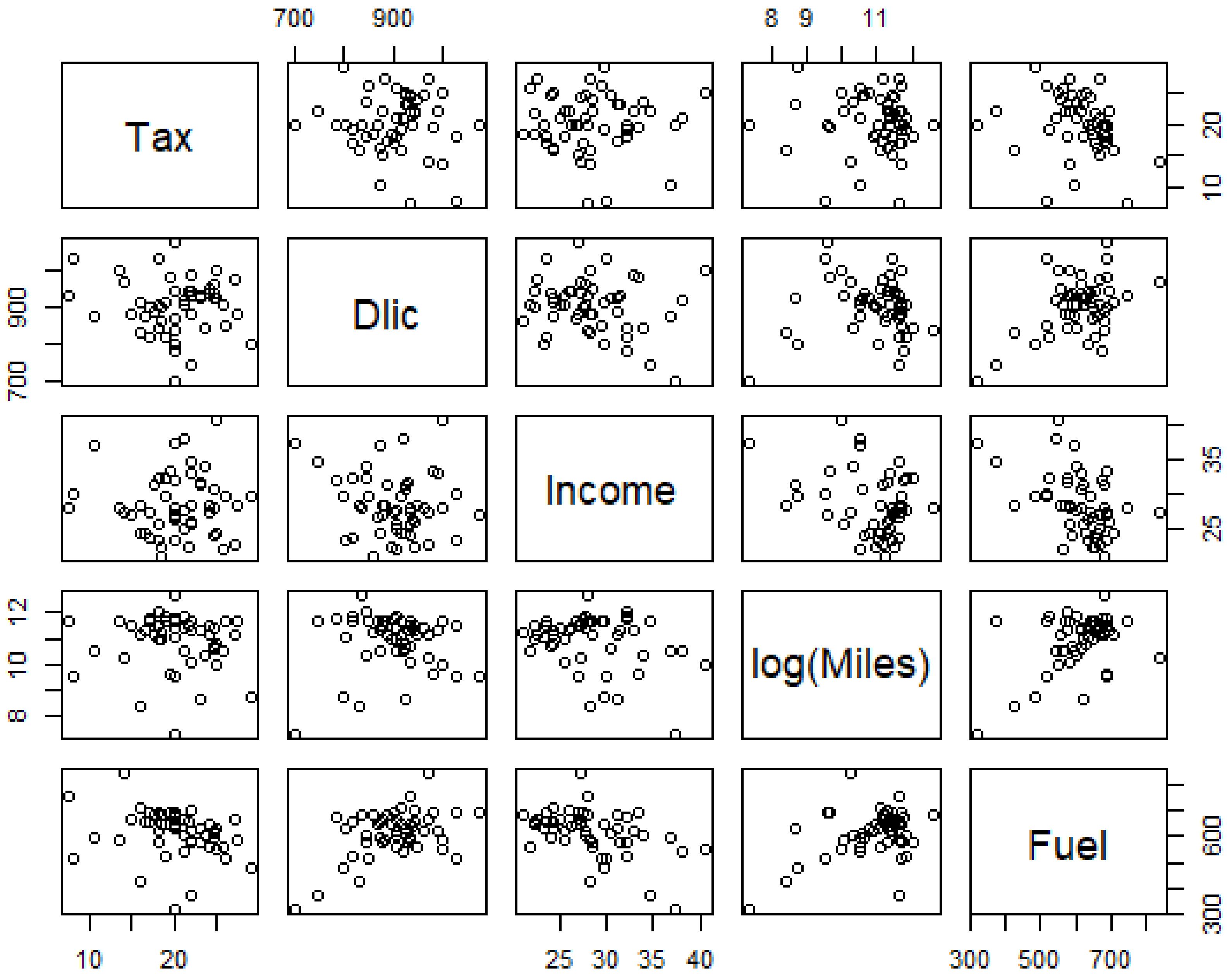


y1
x1
y2
x2
y3
x3
y4
x4
2
6
10
14
5
10
15
20

Table 1.1 Variables in the Fuel Consumption Data[a]

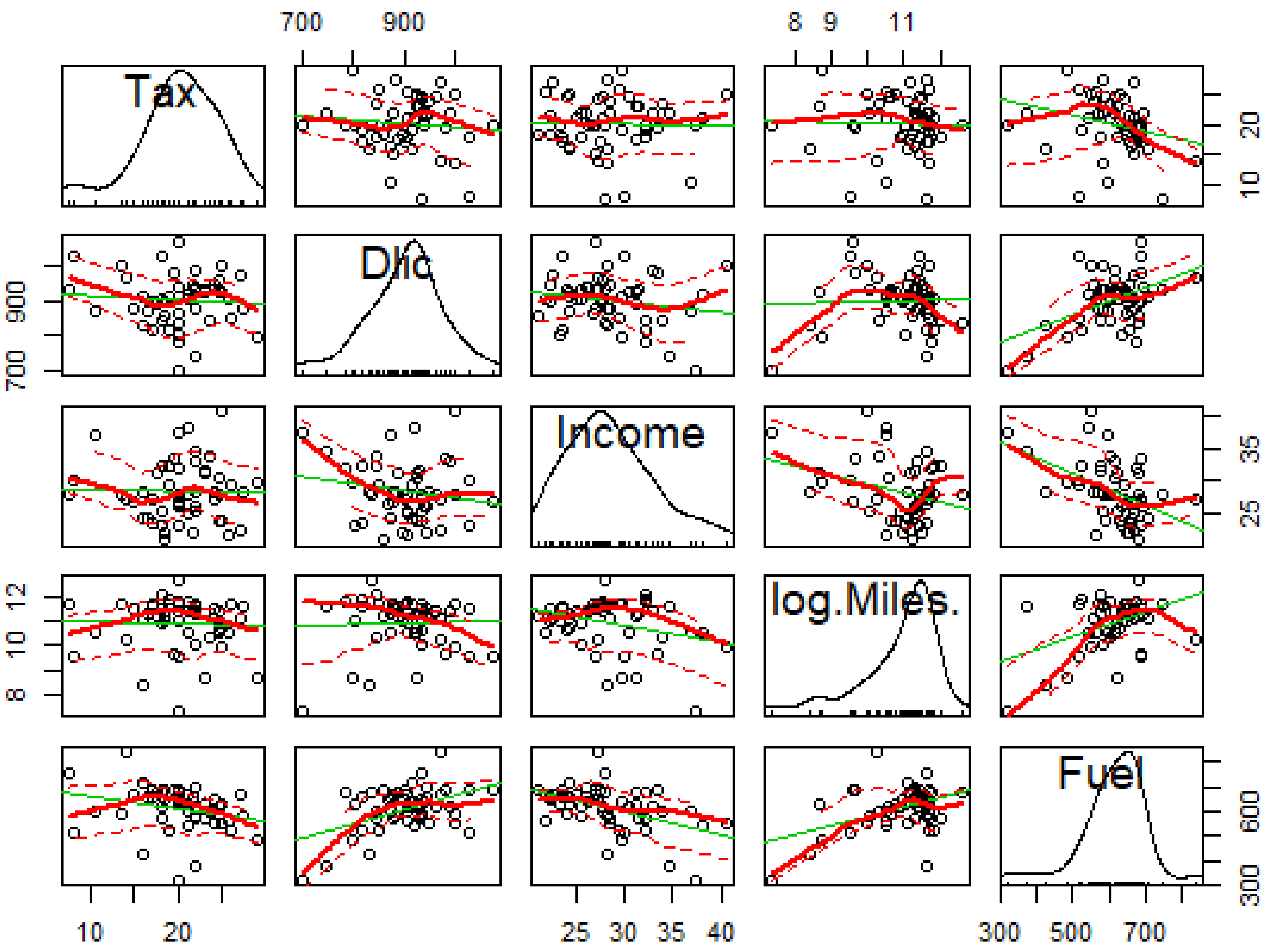
Drivers	Number of licensed drivers in the state
FuelC	Gasoline sold for road use, thousands of gallons
Income	Per person personal income for the year 2000, in thousands of dollars
Miles	Miles of Federal-aid highway miles in the state
Pop	2001 population age 16 and over
Tax	Gasoline state tax rate, cents per gallon
Fuel	1000 × Fuelc/Pop
Dlic	1000 × Drivers/Pop
log(Miles)	Natural logarithm of Miles

[a]All data are for 2001, unless otherwise noted. The last three variables do not appear in the data file, but are computed from the previous variables, as described in the text.



Tax
Dlic
Income
log(Miles)
Fuel
700
900
8
9
11
20
10
900
700
35
25
12
10
8
600
300
10
20
25
30
35
40
300
500
700

Scatterplot matrix



Ordinary Least Squares (OLS)

- Lots of methods to create the optimum relationship between variables
- OLS is the most common method
- Here parameters are estimated to minimize the residual sum of squares

Quantities needed to calculate OLS regression line

- Mean of X
- Mean of Y
- Variance of X
- Variance of Y
- Covariance of XY

- Slope = Covariance(XY)/Variance(X)
- Intercept= mean(y) – slope*mean(x)

Variance

- Measure of the spread of data

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

- var= Sum(Observation – mean)^2 / number of observations

- A special case of covariance

Covariance

- Variation of two variables with each other
- Sum(x- mean(x) *(y-mean(y))/sample size
- We can combine the variance and covariance to get a standardized measure of relationship

$$\operatorname{cov}(X, Y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}.$$

Correlation

- Standard measure of relationship between two variables

$$r = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2}\sqrt{\sum(Y-\overline{Y})^2}}$$

- Not causative

- Covariance (X,Y)/ sqrt[Var(X)*Var(Y)]

$$r_{xy} \qquad s_{xy}/(\mathrm{SD}_x\mathrm{SD}_y)$$

OLS estimation-provides estimates of the parameters not actual values

Fitted Values

$$\hat{y}_i = \hat{\mathrm{E}}(Y|X = x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Residuals

$$\hat{e}_i = y_i - \hat{\mathrm{E}}(Y|X = x_i) = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \quad i = 1, \ldots, n$$

Equation for statistical errors

$$e_i = y_i - (\beta_0 + \beta_1 x_i) \quad i = 1, \ldots, n$$

Residual Sum of Squares

- Residual sum of squares (RSS)
- Sum of squared residuals (SSR)
- Sum of squared errors of prediction (SSE)

- Deviations of predicted from actual empirical values of data

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

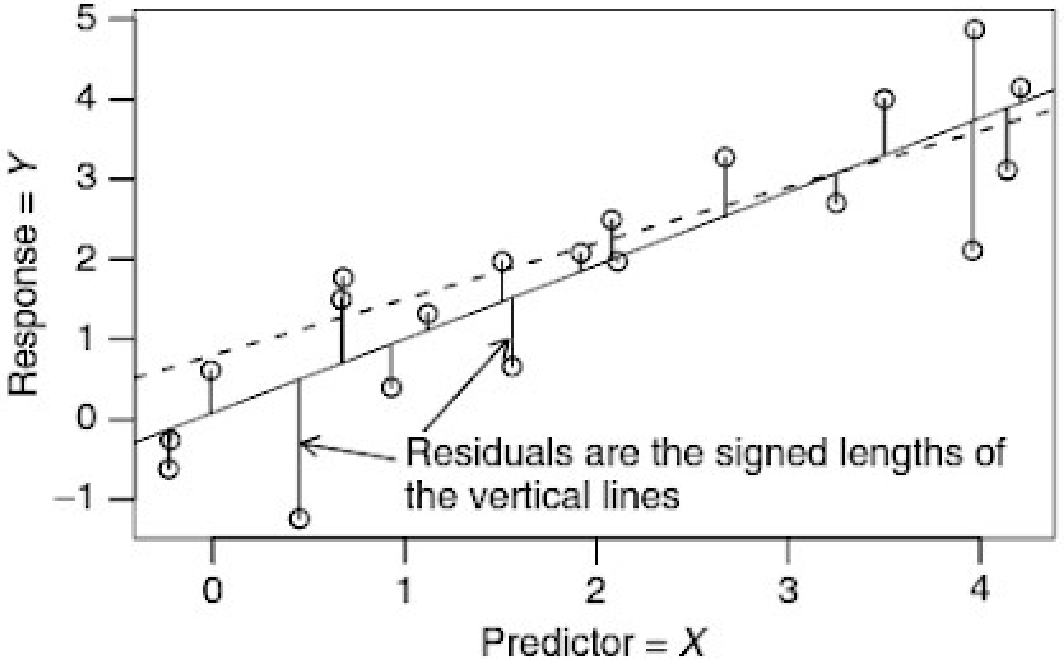
$$\frac{\partial \mathrm{RSS}(\beta_0, \beta_1)}{\beta_0} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial \mathrm{RSS}(\beta_0, \beta_1)}{\beta_1} = -2\sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\beta_0 n + \beta_1 \sum x_i = \sum y_i$$

$$\beta_0 \sum x_i + \beta_1 \sum x_i^2 = \sum x_i y_i$$

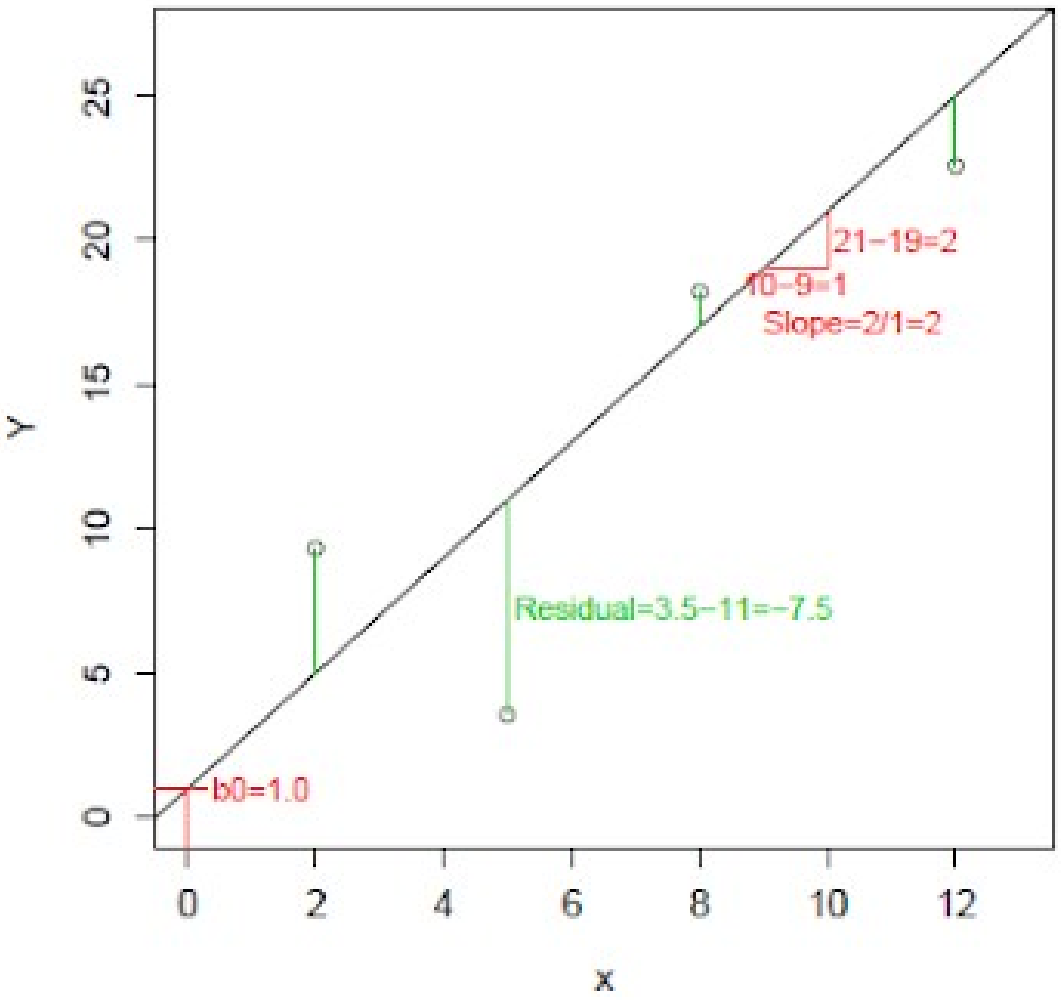
Residual Sum of Squares



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

$$\hat{\beta}_1 = \frac{\mathrm{SXY}}{\mathrm{SXX}} = r_{xy}\frac{\mathrm{SD}_y}{\mathrm{SD}_x} = r_{xy}\left(\frac{\mathrm{SYY}}{\mathrm{SXX}}\right)^{1/2}$$

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



b0=1.0
Residual=3.5−11=−7.5
21−19=2
10−9=1
Slope=2/1=2
x
Y
0
2
4
6
8
10
12
0
5
10
15
20
25

Coefficient of Determination

- How much variation does your predictor explain?
- SSreg=SYY-RSS
- R^2=SSreg/SYY=1-RSS/SYY
- R^2 is a scale-free one number summary of the strength of the relationship between x and y
- Adjusted R^2 is computed by accounting the df within the experiment

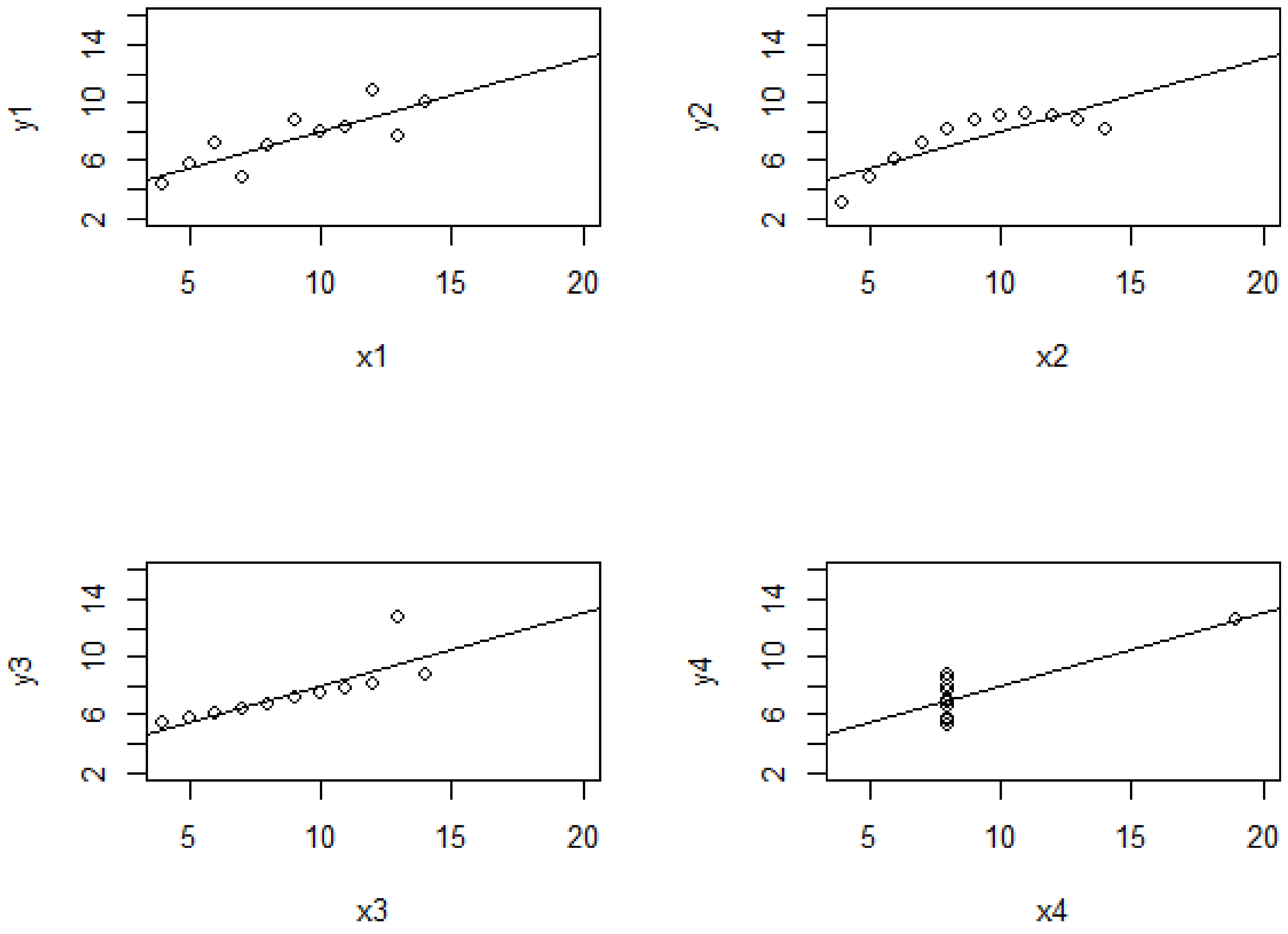
Table 2.1 Definitions of Symbols[a]

Quantity	Definition	Description
$\bar{x}$	$\sum x_i/n$	Sample average of x
$\bar{y}$	$\sum y_i/n$	Sample average of y
$\mathtt{SXX}$	$\sum(x_i-\bar{x})^2=\sum(x_i-\bar{x})x_i$	Sum of squares for the xs
SD_x^2	$\mathtt{SXX}/(n-1)$	Sample variance of the xs
SD_x	$\sqrt{\mathtt{SXX}/(n-1)}$	Sample standard deviation of the xs
$\mathtt{SYY}$	$\sum(y_i-\bar{y})^2=\sum(y_i-\bar{y})y_i$	Sum of squares for the ys
SD_y^2	$\mathtt{SYY}/(n-1)$	Sample variance of the ys
SD_y	$\sqrt{\mathtt{SYY}/(n-1)}$	Sample standard deviation of the ys
$\mathtt{SXY}$	$\sum(x_i-\bar{x})(y_i-\bar{y})=\sum(x_i-\bar{x})y_i$	Sum of cross-products
s_{xy}	$\mathtt{SXY}/(n-1)$	Sample covariance
r_{xy}	$s_{xy}/(\mathrm{SD}_x\mathrm{SD}_y)$	Sample correlation

[a]In each equation, the symbol Σ means to add over all n values or pairs of values in the data.

Estimating variance of our OLS line

- The variance is defined as the RSS divided by the residual df
- Residual df is n- number of parameters estimated
  - Slope and intercept
  - N-2
- This is the residual mean square
- The square root of the residual mean square is standard error of the regression
  - This is in the same units as the response variable


y1
x1
y2
x2
y3
x3
y4
x4
2
6
10
14
5
10
15
20

Confidence intervals

- Intercept
  - B0±t(a,n-2)*se(B0|X)
- seB0|X
  - Var(B0}X)=σ^2 (1/n + mean(x)^2/SXX)
  - Var(B1}X)=σ^2 (1/SXX)
  - se(B1|X)=sqrt(σ^2 (1/n + mean(x)^2/SXX))
- Lines can be fit based on the adjusted equations

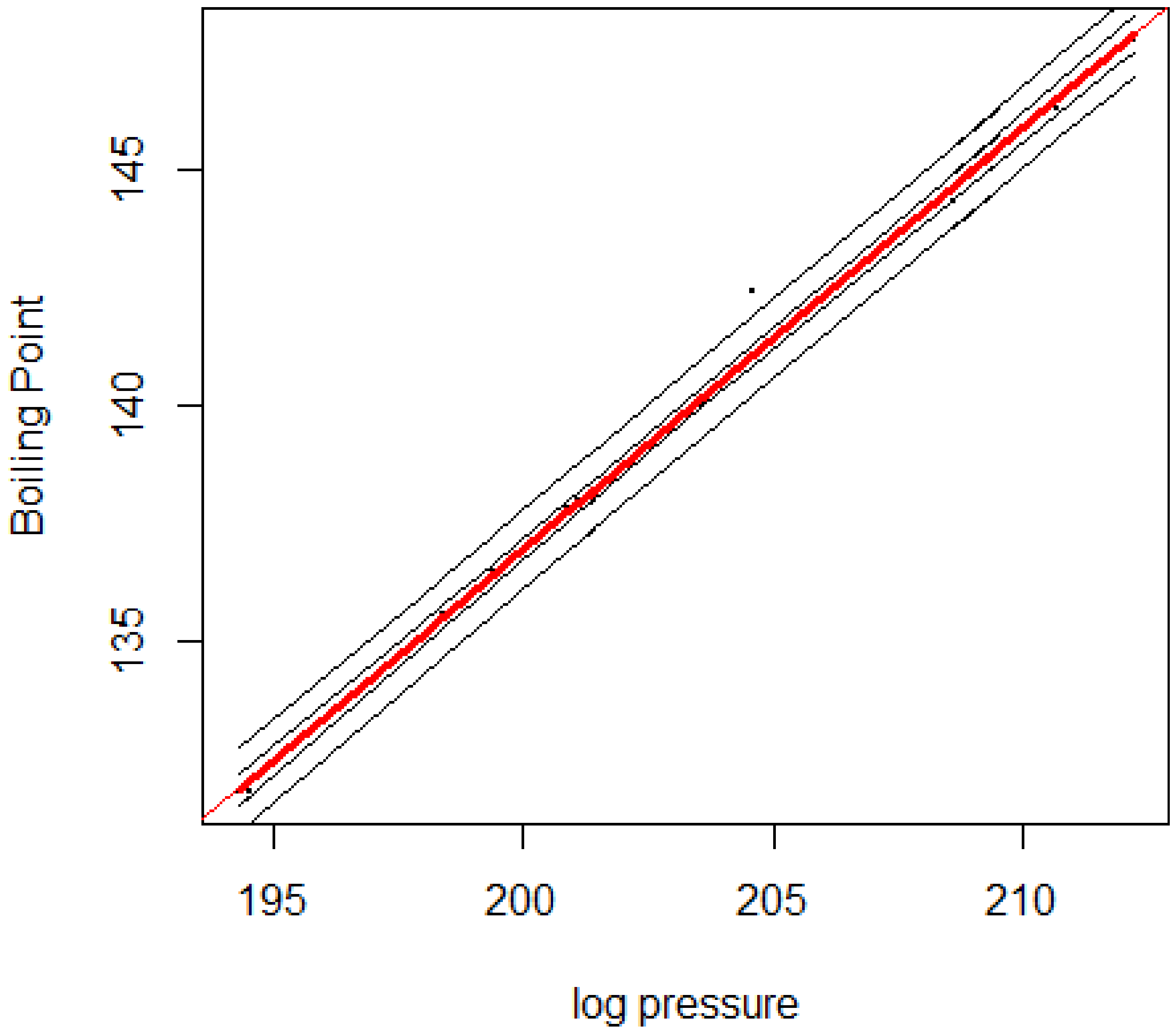
Prediction

- We assume that the mean function is representative of the value we would like to predict

- The standard error of the prediction is

- The se of the prediction is larger than the confidence interval of the fitted line

$$\mathrm{sepred}(\tilde{y}_* | x_*) = \sigma \left(1 + \frac{1}{n} + \frac{(x_* - \overline{x})^2}{\mathrm{SXX}} \right)^{1/2}$$



Boiling Point
135
140
145
195
200
205
210
log pressure

```
> summary(m1) # model summary

Call:
lm(formula = lpres ~ bp, data = Forbes)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.32220 -0.14473 -0.06664  0.02184  1.35978 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -42.13778    3.34020  -12.62 2.18e-09 ***
bp            0.89549    0.01645   54.43  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.379 on 15 degrees of freedom
Multiple R-squared:  0.995,     Adjusted R-squared:  0.9946 
F-statistic:  2963 on 1 and 15 DF,  p-value: < 2.2e-16
```

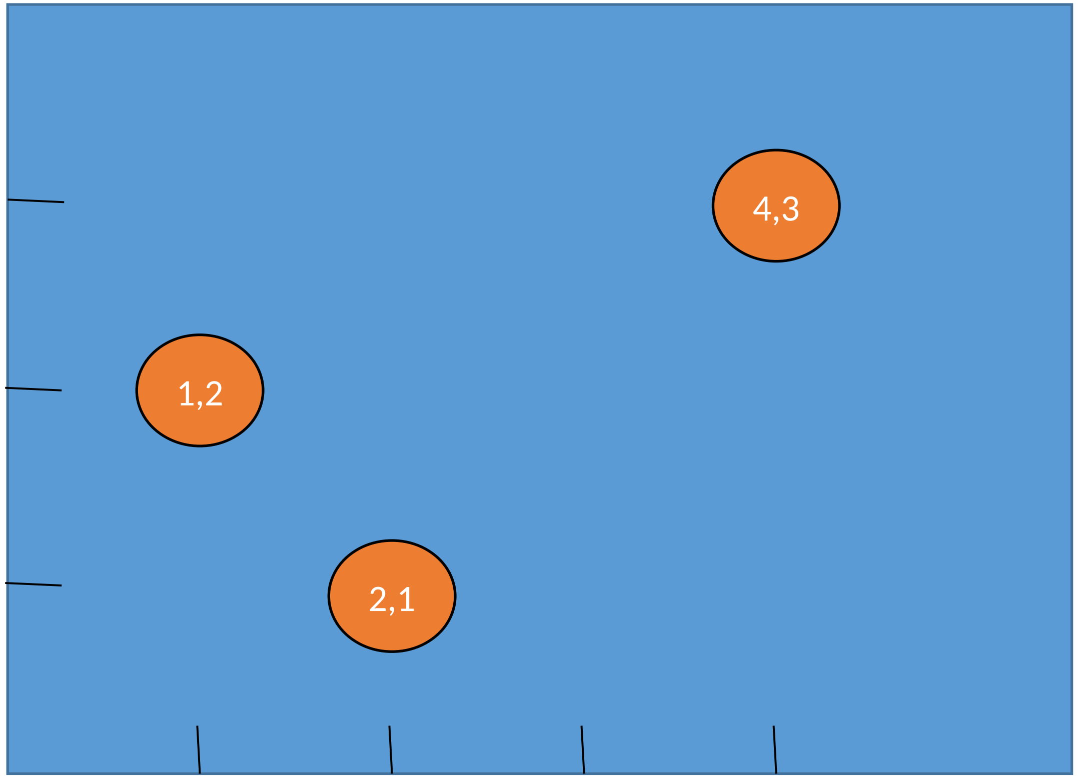
Questions?

Regression vs. ANOVA

- Regression is a more general form of ANOVA which is a general form of a t-test
- While ANOVA is focused on categorical variables Regression is focused on continuous variation
- Linear regression models the relationship between an outcome variable and a set of predictor variables

Calculations for Regression

- Data
- ybar = mxbar + b
- X= (1,2,4)
- Y=(2,1,3)
- M =(Xbar-Xybar)/ (xbar)^2-X^{2}bar



Calculations for Regression

- ybar = mxbar + b
- Xbar= 1+2+4/3=7/3
- Ybar=2+1+3/3=2
- Xybar=(1*2)+(2*1)(4*3)=16/3
- X^2bar=$1^2+2^2+4^2$=21/3=7
- M=(7/3)x(2)-(16/3) / $(7/3)^2$ -7
- Y=3/7x + 1