ISyE 3770, Spring 2024
Statistics and Applications

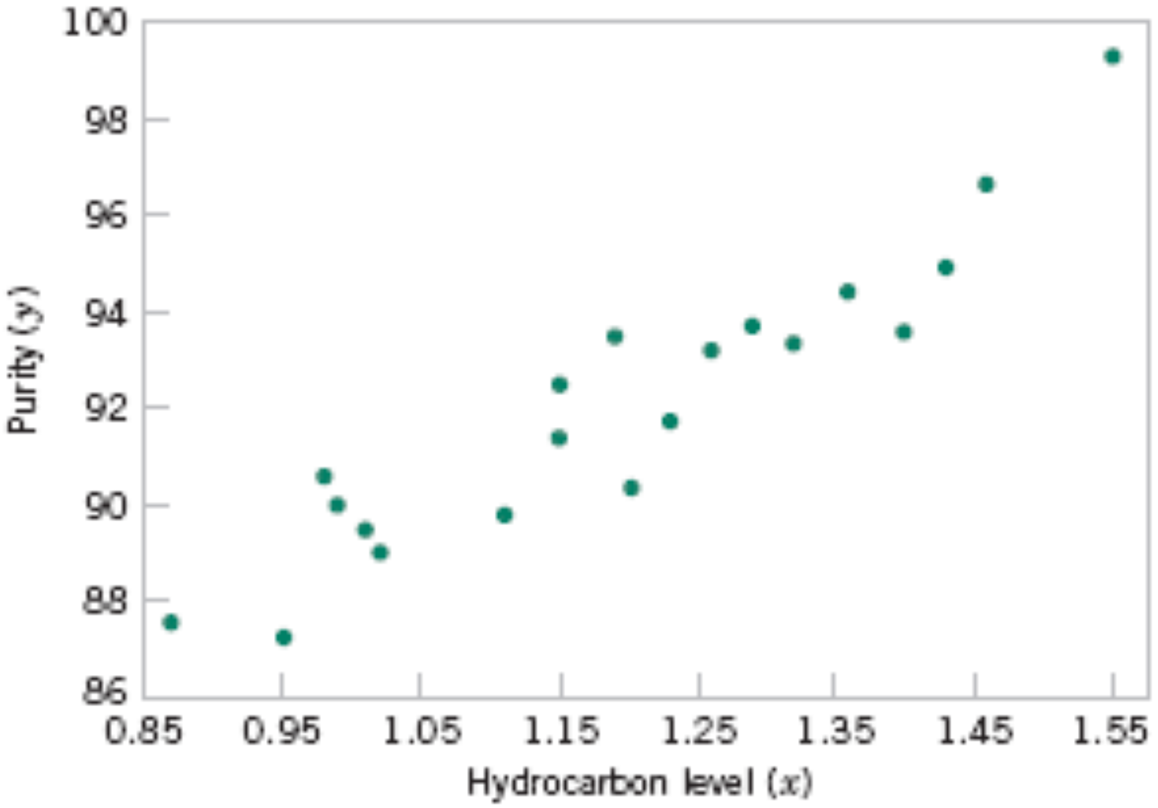
Linear Regression

Instructor: Jie Wang
H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Tech

jwang3163@gatech.edu
Office: ISyE Main 447

Scatter Diagram

• Many problems in engineering and science involve exploring the relationships between two or more variables.
• Regression analysis is a statistical technique that is very useful for these types of problems.



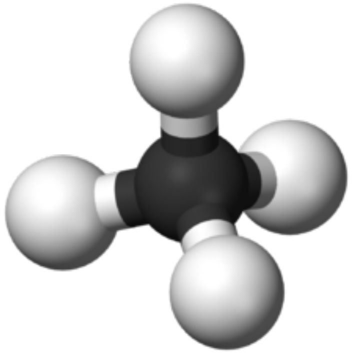
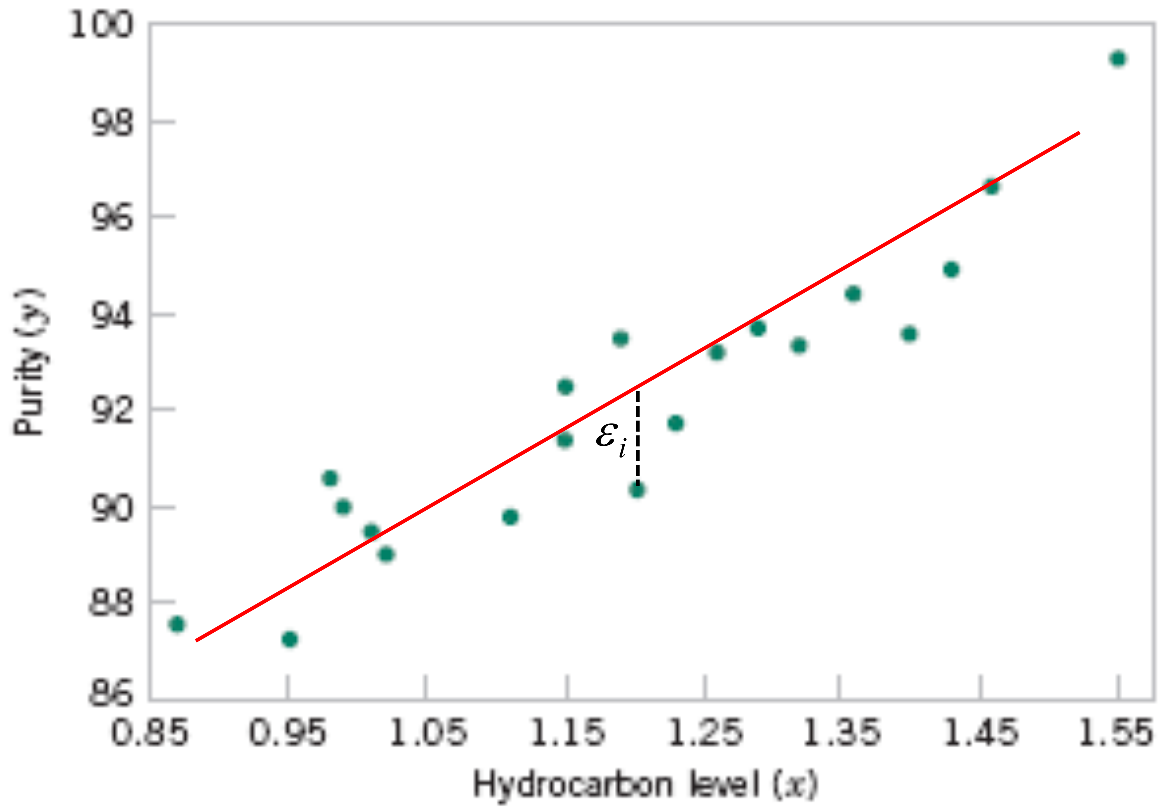


Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

Simple Linear Regression

Based on the scatter diagram, it is probably reasonable to assume that the mean of the random variable Y is related to X by the following simple linear regression model:



Response

Regressor or Predictor

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1,2,\cdots,n$$

Intercept

Slope

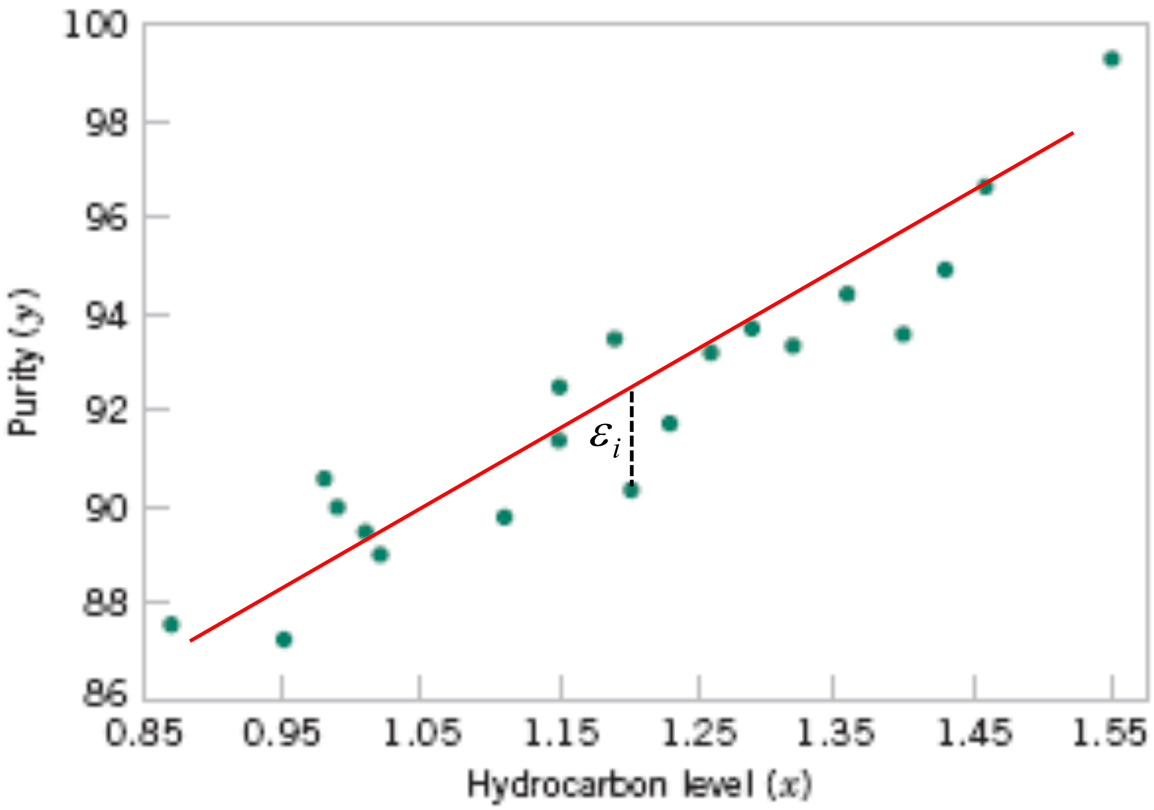
Random error

$$\varepsilon_i \sim N(0,\sigma^2)$$

where the slope and intercept of the line are called regression coefficients.

•The case of simple linear regression considers a single regressor or predictor x and a dependent or response variable Y.

Mean response



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

Simple Linear Regression

The method of least squares is used to estimate the parameters, β_0 and β_1 by minimizing the sum of the squares of the vertical deviations in Figure 11-3.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n$$

sum of the squares of the error

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

Minimize

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

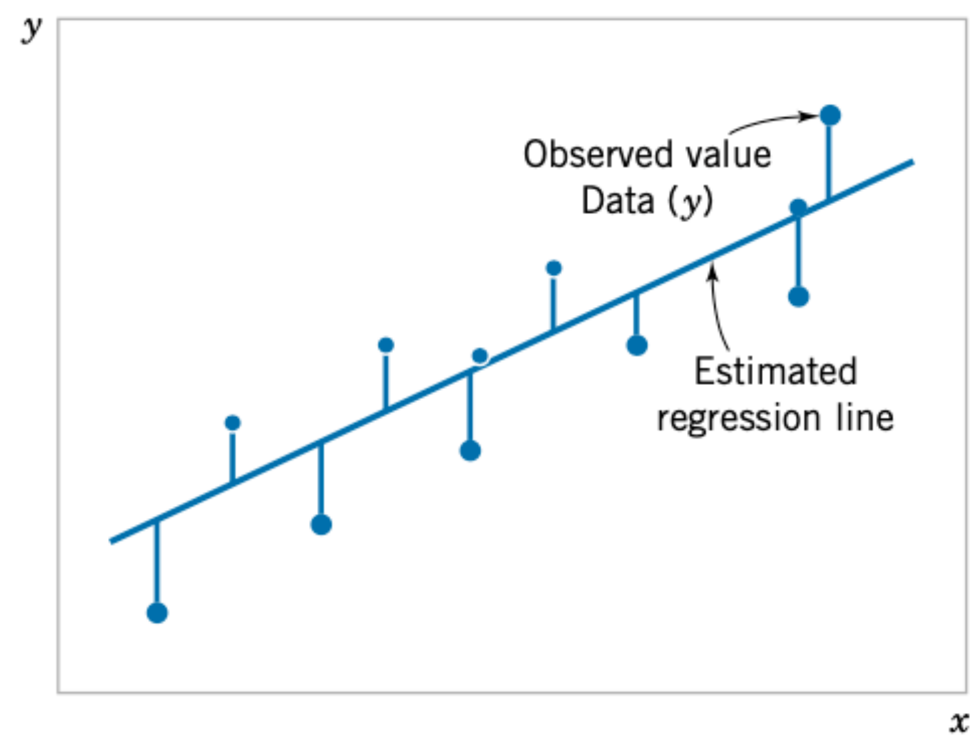


Figure 11-3 Deviations of the data from the estimated regression model.

Least Square Normal Equations

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

Alternative Notation

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \qquad S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

Sxx: sum of the squares of the difference between x and $\bar{x}$

Sxy: sum of the product of the difference between $x - \bar{x}$ and $y - \bar{y}$

$$\left.\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \end{aligned}\right\} \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Fitted (estimated) regression model

Example 1: Gas purity

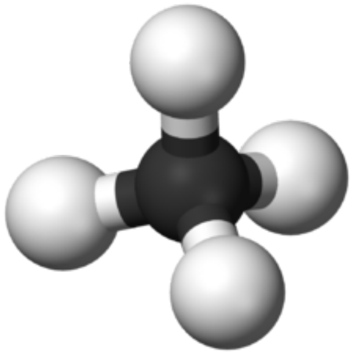
Find the least square estimates of the simple linear regression describing the relationship between Purity (y) and Hydrocarbon Levels (x).
Also, calculate the predicted purity when hydrocarbon level is 1.01. Find the prediction error.

$$n = 20 \quad \sum_{i=1}^{20} x_i = 23.92 \quad \sum_{i=1}^{20} y_i = 1{,}843.21 \quad \bar{x} = 1.1960 \quad \bar{y} = 92.1605$$

$$\sum_{i=1}^{20} y_i^2 = 170{,}044.5321 \quad \sum_{i=1}^{20} x_i^2 = 29.2892 \quad \sum_{i=1}^{20} x_i y_i = 2{,}214.6566$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$



```
data1 <- read.table("Example_1.txt", header=FALSE)
```

Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

Fitted model

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{10.17744}{0.68088} = 14.94748$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 92.1605 - (14.94748)1.196 = 74.28331$$

$$\hat{y} = 74.283 + 14.947x$$

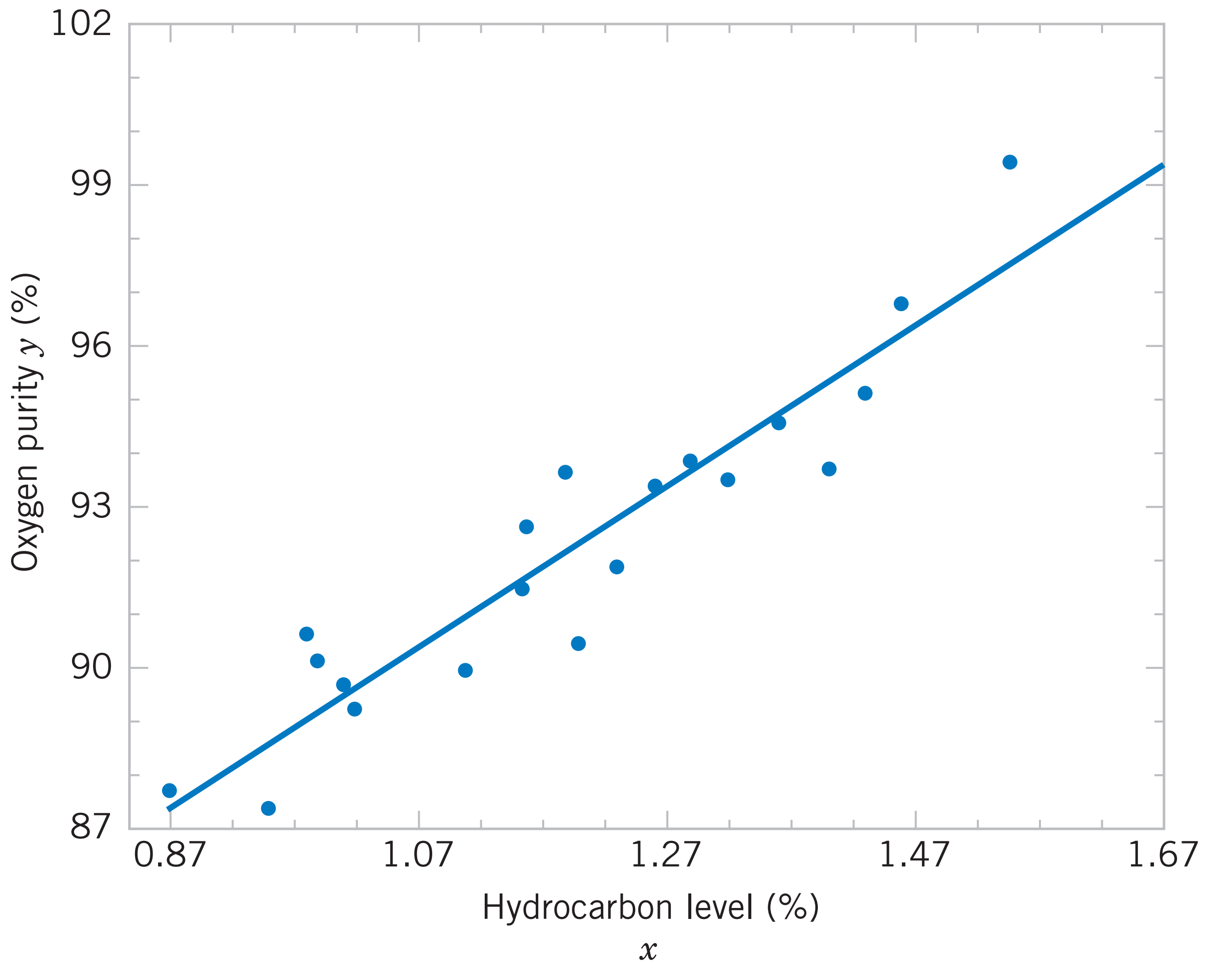


Figure 11-4 Scatter plot of oxygen purity y versus hydrocarbon level x and regression model $\hat{y} = 74.283 + 14.947x$.

Results using R

```
Call:
lm(formula = data1[, 2] ~ data1[, 1])

Residuals:
     Min       1Q   Median       3Q      Max 
-1.83029 -0.73334  0.04497  0.69969  1.96809 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   74.283      1.593   46.62  < 2e-16 ***
data1[, 1]    14.947      1.317   11.35 1.23e-09 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.087 on 18 degrees of freedom
Multiple R-squared:  0.8774, Adjusted R-squared:  0.8706 
F-statistic: 128.9 on 1 and 18 DF,  p-value: 1.227e-09
```

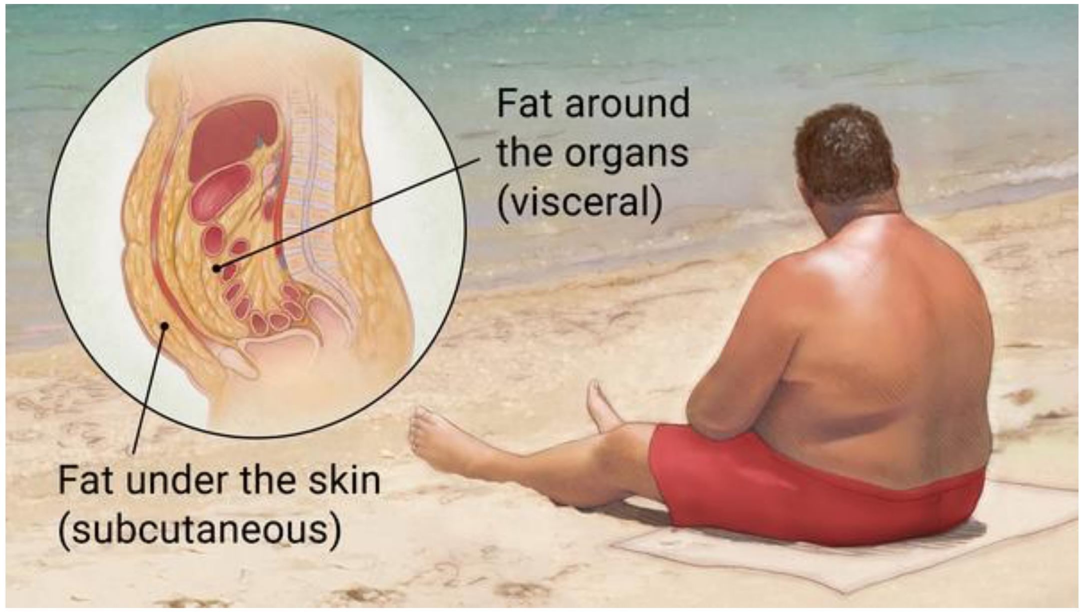
$\hat{y} = \beta_0 + \beta_1 x$

$\beta_0 = 74.283$

$\beta_1 = 14.947$

Example 2: Diabetes and Obesity

- Diabetes and obesity are serious health concerns in the US. Measuring the amount of body fat of a person is one way to monitor body weight control. To measure body fat accurately one needs x-ray machine.
- BMI = mass (kg) / (height(m))^2 is used as a proxy to body fat.
- In a study of 250 men at Brigham Young U, both BMI x and body fat y were measured. The summary statistics are:



$$\sum_{i=1}^{n} x_i = 6322.28 \qquad \sum_{i=1}^{n} x_i^2 = 162674.18$$

$$\sum_{i=1}^{n} y_i = 4757.90 \qquad \sum_{i=1}^{n} y_i^2 = 107679.27$$

$$\sum_{i=1}^{n} x_i\, y_i = 125471.10$$

Fit a linear regression model

Body fat example

$n = 250$

$\sum_{i=1}^{n} x_i = 6322.28$, $\sum_{i=1}^{n} x_i^2 = 162674.18$

$\sum_{i=1}^{n} y_i = 4757.90$, $\sum_{i=1}^{n} y_i^2 = 107679.27$

$\sum_{i=1}^{n} x_i y_i = 125471.10$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 162674.18 - \frac{6322.28^2}{250}$$

$$= 2789.282$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

$$= 125471.10 - \frac{6322.28 \times 4757.90}{250}$$

$$= 5147.996$$

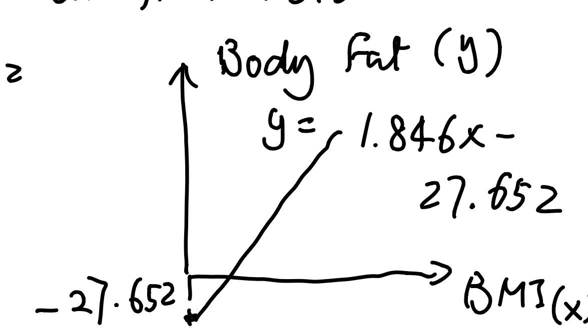
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{6322.28}{250} = 25.28912$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{4757.90}{250} = 19.0316$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{5147.996}{2789.282} = 1.845635 \approx 1.846$$

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 = 19.0316 - 25.28912 \times 1.846$$

$$= -27.652$$



Continue: body fat vs BMI

- **Use the equation of the fitted line to predict that body fat would be observed, on average, for a man with BMI = 30**

- **Suppose the observed body fat of a man with a BMI of 25 is 25%, find the residual for that observation.**

For a man BMI = 30, use our model, predicted body fat

$$\hat{y} = 1.846x - 27.652$$

$$= 1.846 \times 30 - 27.652$$

$$= 27.728 \quad (\%)$$

For BMI = 25, use our model, predicted body fat

$$\hat{y} = 1.846 \times 25 - 27.652$$

$$= 18.498 \quad (\%)$$

observed y is 25 (%)

under estimate: residual $y - \hat{y} =$

$$25 - 18.498$$

$$= 6.502$$

Example 3: sale price and taxes

$y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$

11-4. An article in *Technometrics* by S. C. Narula and J. F. Wellington ["Prediction, Linear Regression, and a Minimum Sum of Relative Errors" (Vol. 19, 1977)] presents data on the selling price and annual taxes for 24 houses. The data are shown in the following table.

y Sale Price/1000	x Taxes (Local, School), County)/1000	Sale Price/1000	Taxes (Local, School), County)/1000
25.9	4.9176	30.0	5.0500
29.5	5.0208	36.9	8.2464
27.9	4.5429	41.9	6.6969
25.9	4.5573	40.5	7.7841
29.9	5.0597	43.9	9.0384
29.9	3.8910	37.5	5.9894
30.9	5.8980	37.9	7.5422
28.9	5.6039	44.5	8.7951
35.9	5.8282	37.9	6.0831
31.5	5.3003	38.9	8.3607
31.0	6.2712	36.9	8.1400
30.9	5.9592	45.8	9.1416

relating a to b

Variable a: response

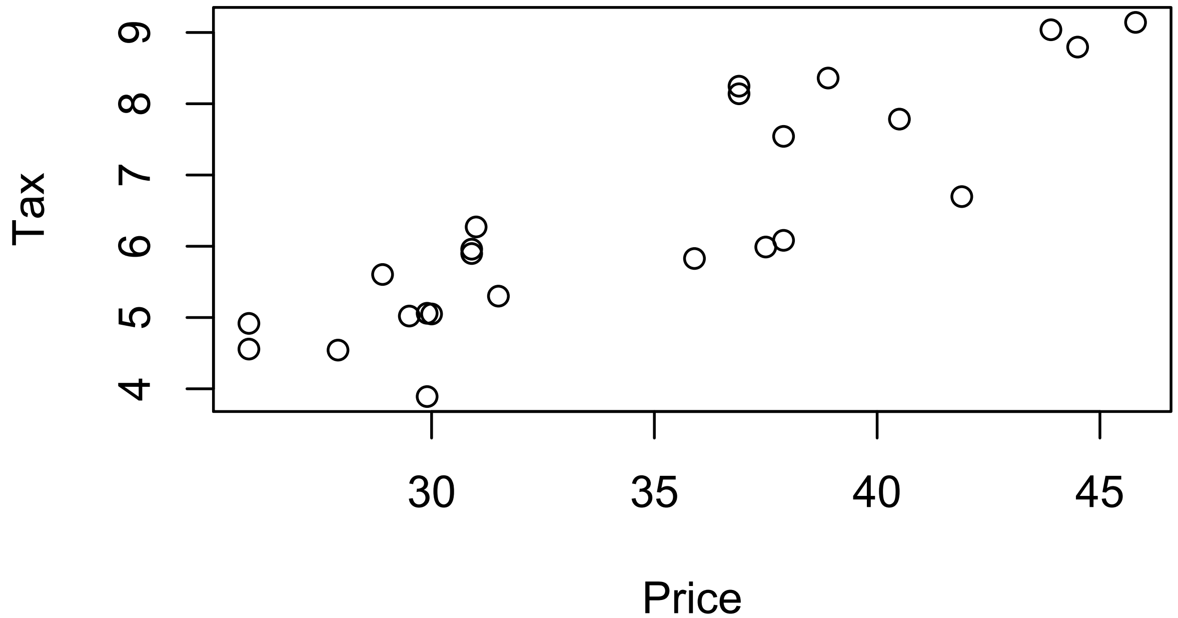
Variable b: regressor/input

(a) Assuming that a simple linear regression model is appropriate, obtain the least squares fit relating selling price to taxes paid. What is the estimate of σ^2?

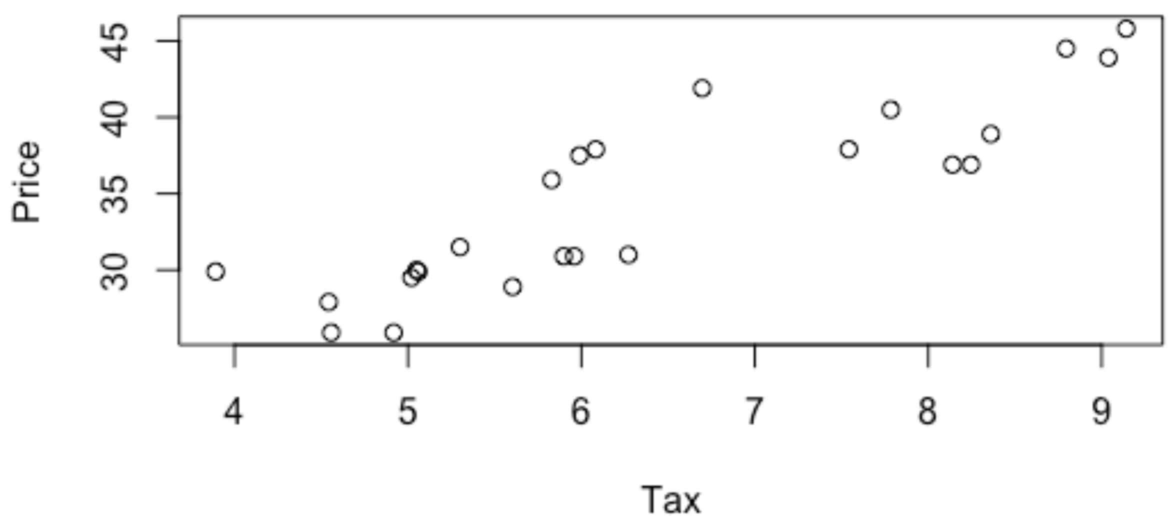
(b) Find the mean selling price given that the taxes paid are $x = 7.50$.

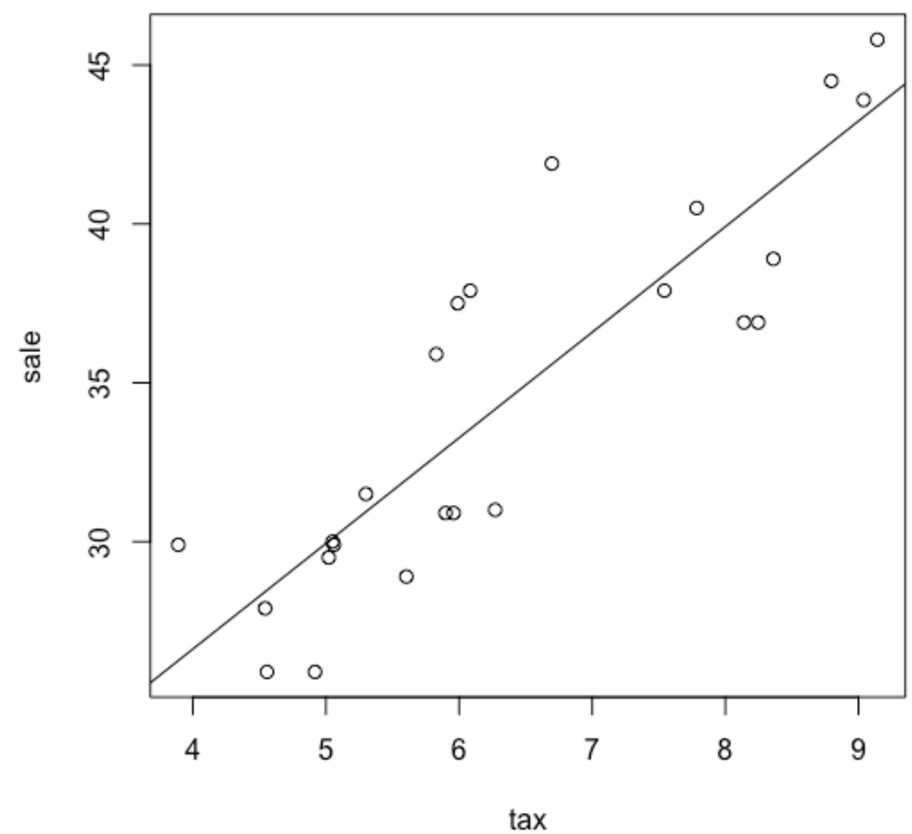
(c) Calculate the fitted value of y corresponding to $x = 5.8980$. Find the corresponding residual.

(d) Calculate the fitted $\hat{y}_i$ for each value of x_i used to fit the model. Then construct a graph of $\hat{y}_i$ versus the corresponding observed value y_i and comment on what this plot would look like if the relationship between y and x was a deterministic (no random error) straight line. Does the plot actually obtained indicate that taxes paid is an effective regressor variable in predicting selling price?

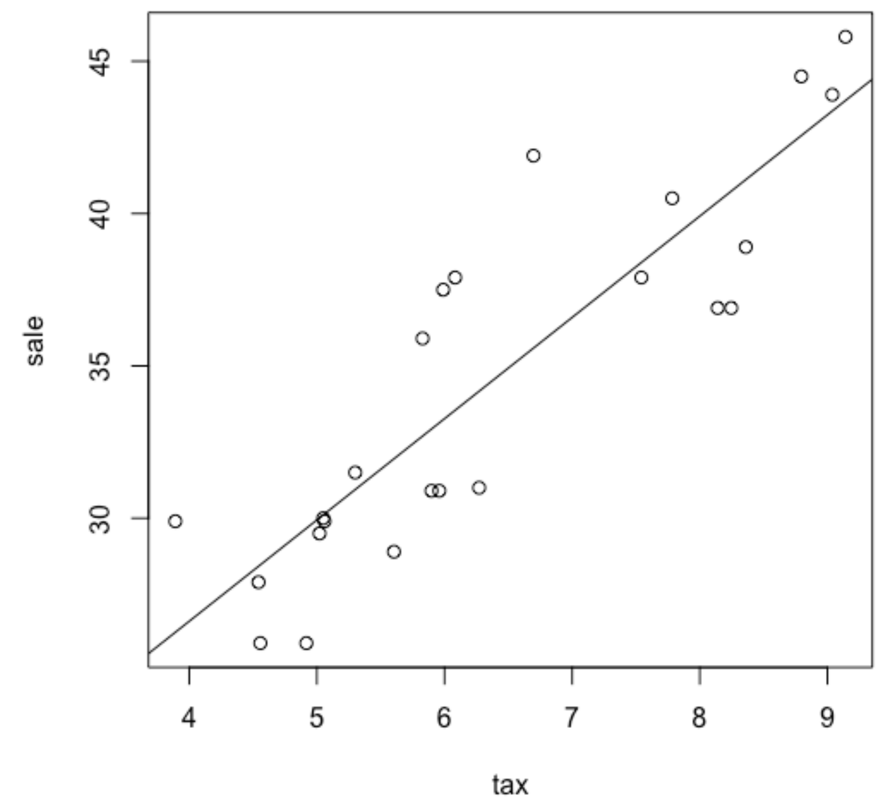



```
data = read.table("house.txt",header=FALSE)

price = data[,1]

tax = data[,2]

plot(tax,price,xlab="Tax",ylab="Price")

abline(13.3202,3.3244)
```





```
model = lm(price~tax)
summary(model)
```




```
Call:
lm(formula = price ~ tax)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8343 -2.3157 -0.3669  1.9787  6.3168 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  13.3202     2.5717   5.179 3.42e-05 ***
tax           3.3244     0.3903   8.518 2.05e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.961 on 22 degrees of freedom
Multiple R-squared:  0.7673,	Adjusted R-squared:  0.7568 
F-statistic: 72.56 on 1 and 22 DF,  p-value: 2.051e-08
```

Regression model: $$\boxed{y} = 3.3244\boxed{x} + 13.3202$$

price tax

Prediction $$x = 7.5, y = 3.3244 \times 7.5 + 13.3202 = 38.2532$$

Model Diagonosis

Adequacy of Regression Model

$$\boxed{y = \beta_0 + \beta_1 x} + \varepsilon$$

$\varepsilon \sim N(0, \sigma^2)$

Analysis of Residual Patterns is useful for checking:

- Independency assumption
- Constant variance

Plot residuals (e_i) against predicted response ($\hat{y}_i$)

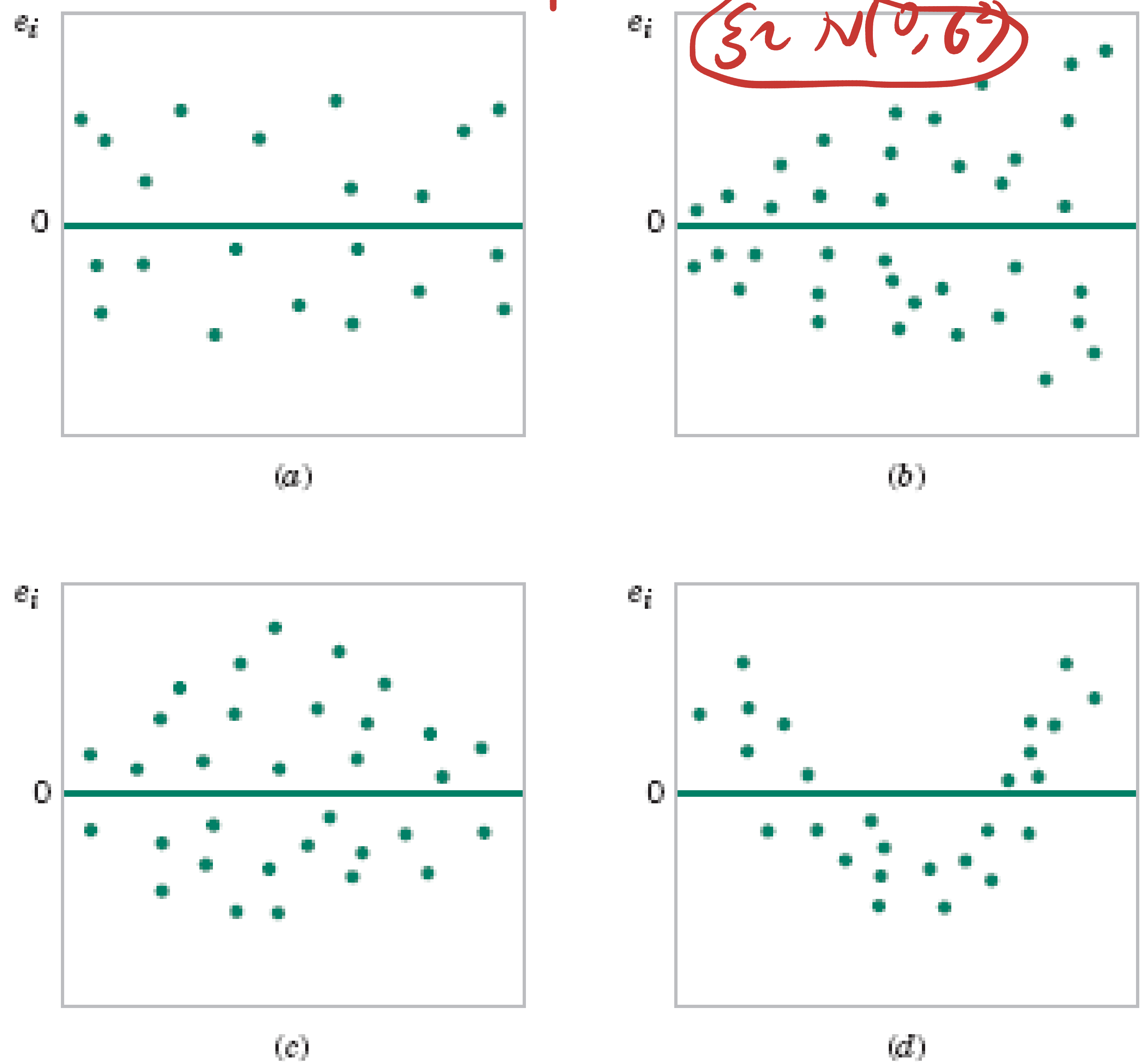
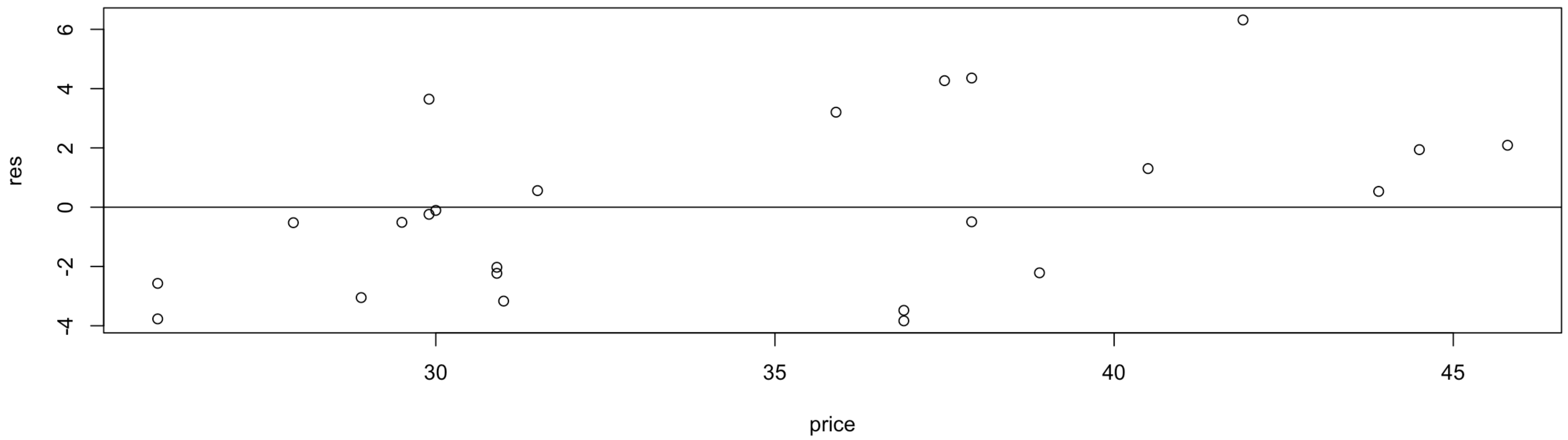


Figure 11-9 Patterns for residual plots. (a) satisfactory, (b) funnel, (c) double bow, (d) nonlinear. [Adapted from Montgomery, Peck, and Vining (2001).]

House example residual plot

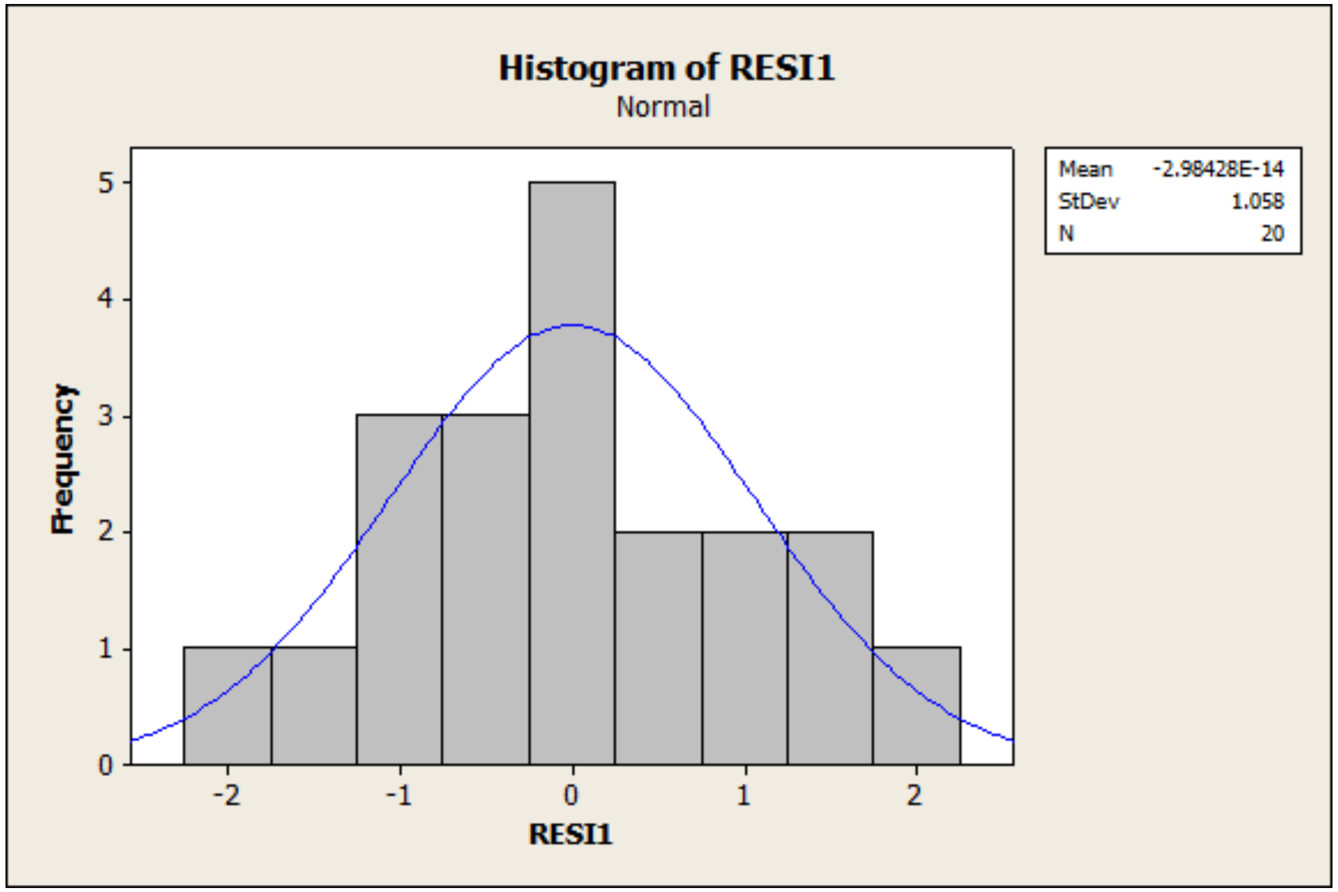
- model = lm(price~tax)
- res = resid(model)
- plot(price, res)
- abline(0, 0)

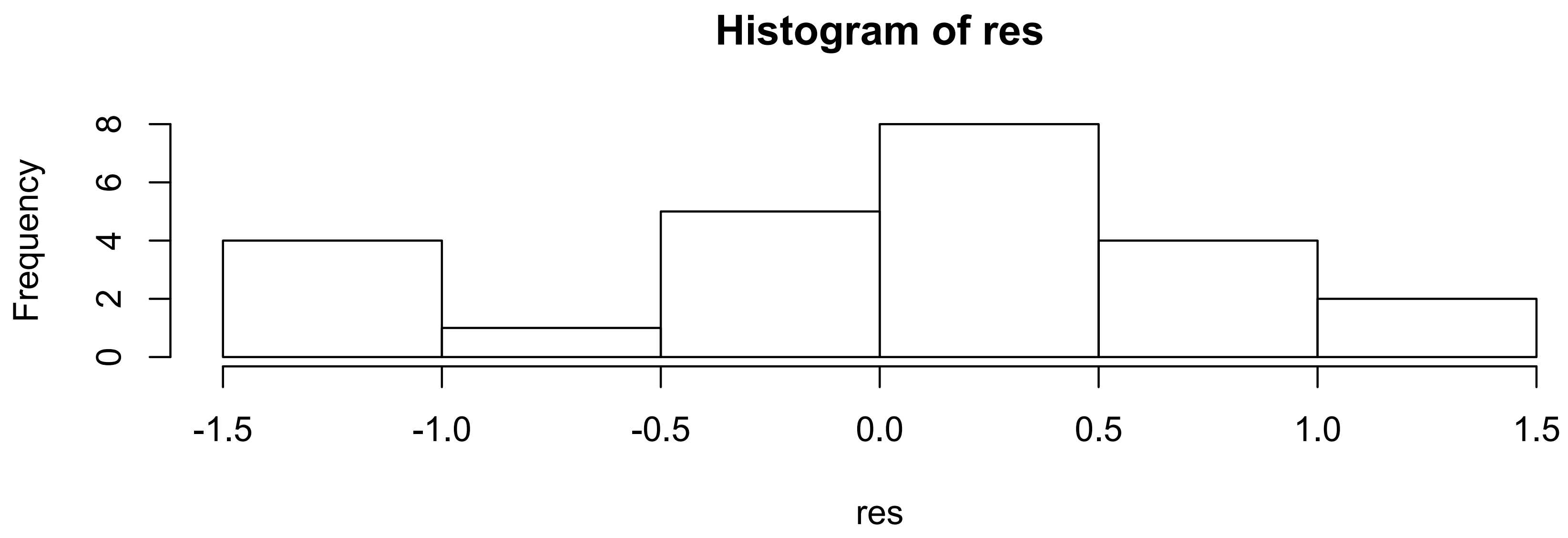


Adequacy of Regression Model

Histogram for residuals:

- Normality assumption



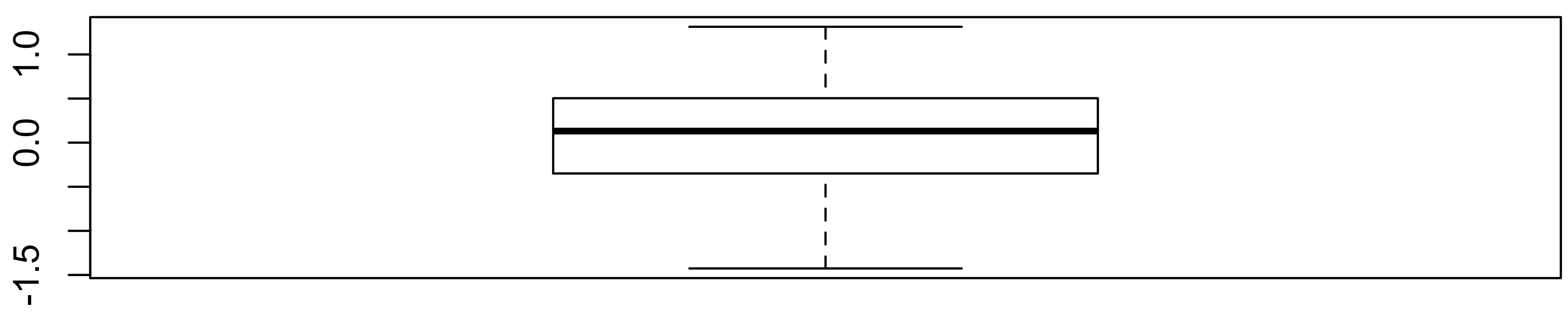


Adequacy of Regression Model

Boxplots:

- It is used to detect observations with large residuals (Outliers)

boxplot(res)



*

Adequacy of Regression Model

Coefficient of Determination (R^2)
R-square statistic

R^2 is called the coefficient of determination and is often used to judge the adequacy of a regression model.

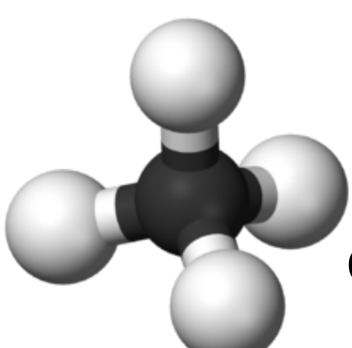
$0 \le R^2 \le 1$;

- We often refer (loosely) to R^2 as the amount of variability in the data explained or accounted for by the regression model.
- It is the <u>square of the correlation coefficient</u> between Y and X

$$R^2 = 1 - \frac{SS_E}{SS_T}$$

$$SS_E = SS_T - \hat{\beta}_1 S_{xy}$$

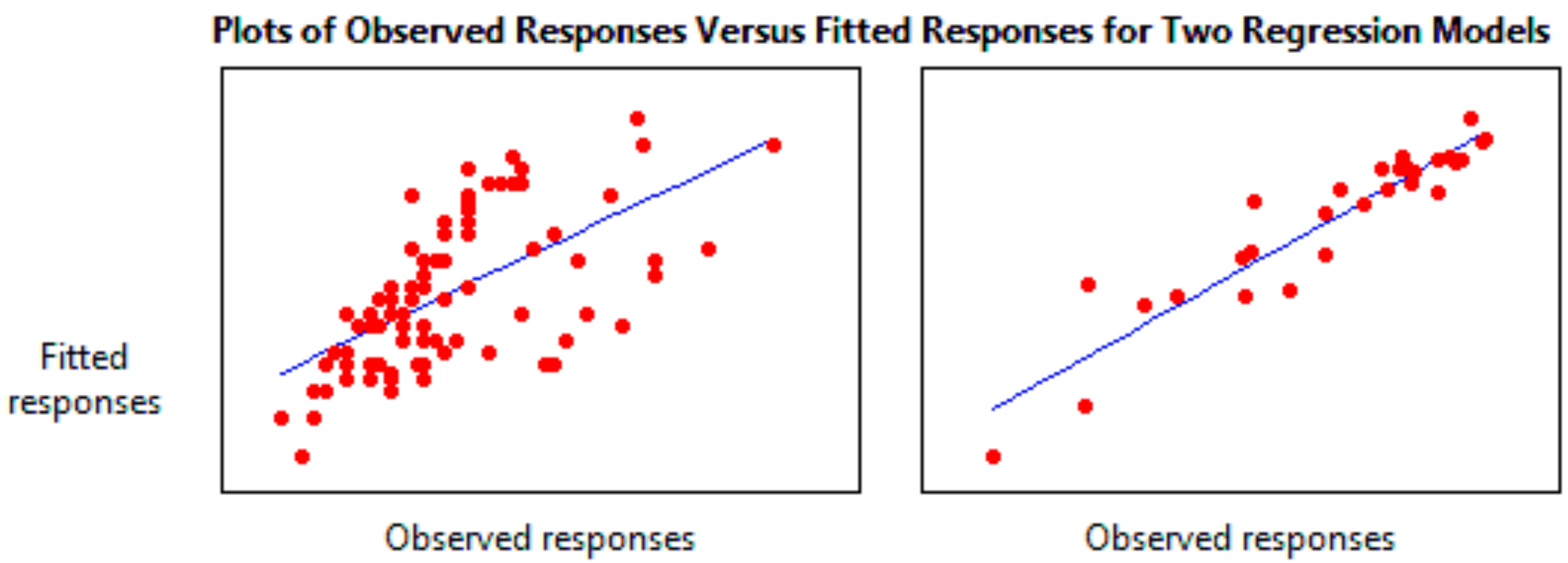
$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$



Oxygen purity example: $R^2 = SS_R/SS_T = 152.13/173.38 = 0.877$

The model accounts for 87.7% of the variability in the data

Interpretation



$R^2 = 0.38$

$R^2 = 0.874$

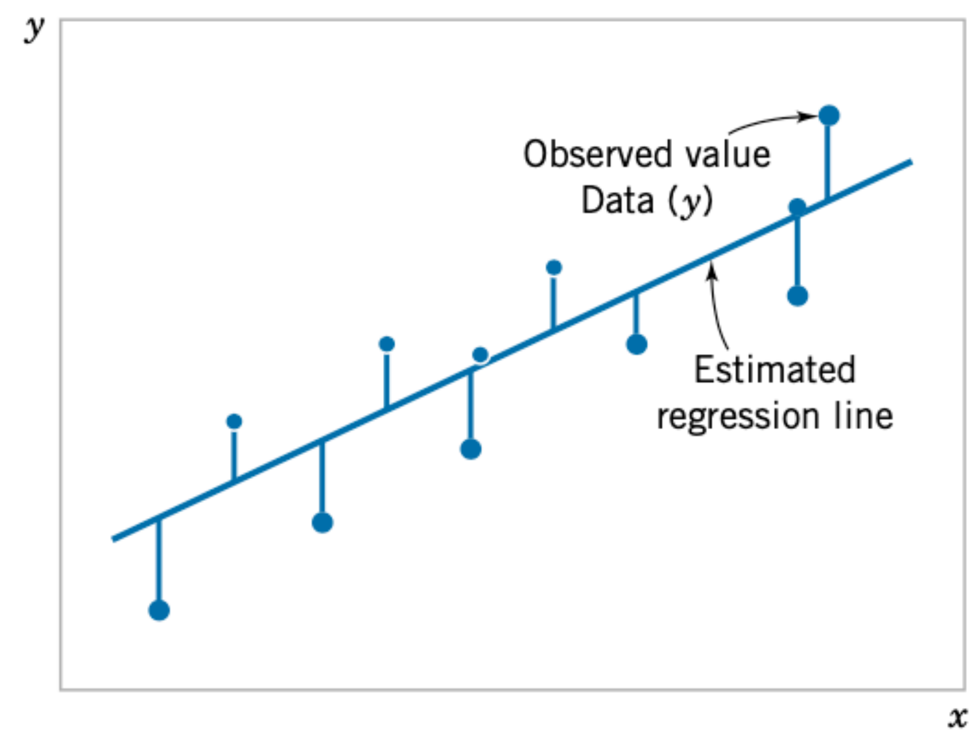
Estimation of Variance (σ^2)

The **<u>error sum of squares</u>** is

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SS_E = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

$$E(SS_E) = (n-2)\sigma^2.$$



An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} \qquad (11\text{-}13)$$

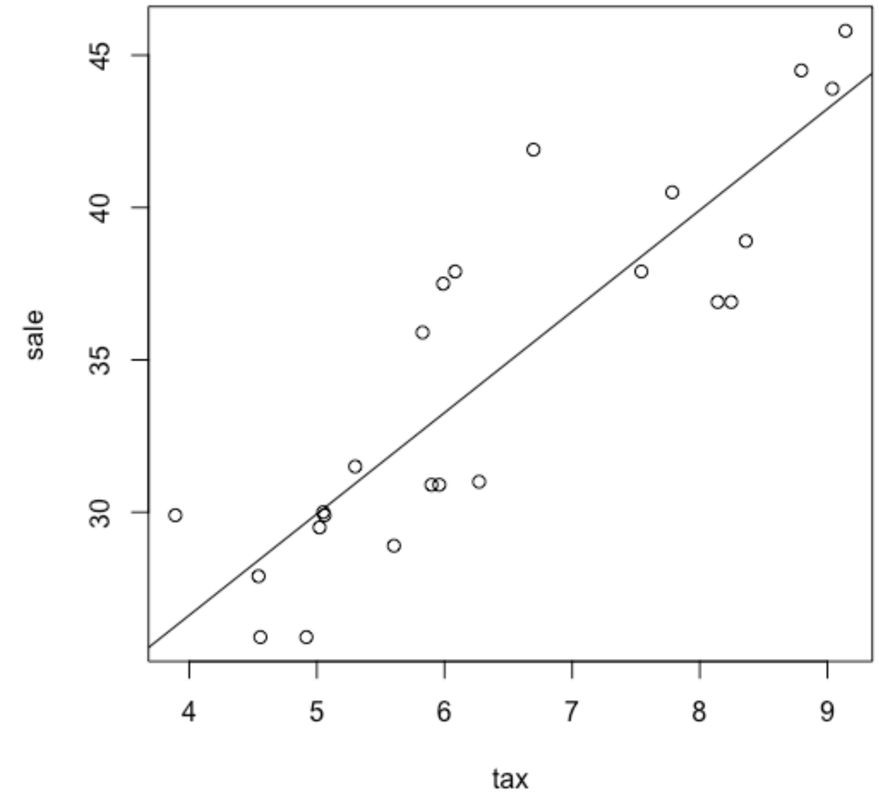
where SS_E can be easily computed using (easier formula)

$$SS_E = \boxed{SS_T} - \hat{\beta}_1 S_{xy} \qquad SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

Total sum of square for y

```
model = lm(price~tax)
summary(model)

Call:
lm(formula = price ~ tax)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8343 -2.3157 -0.3669  1.9787  6.3168

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.3202     2.5717   5.179 3.42e-05 ***
tax           3.3244     0.3903   8.518 2.05e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.961 on 22 degrees of freedom
Multiple R-squared:  0.7673,	Adjusted R-squared:  0.7568
F-statistic: 72.56 on 1 and 22 DF,  p-value: 2.051e-08
```

6



Confidence interval

Confidence Interval for Regression Coefficients

confint(model)

Mean and variance of the slope estimator $E(\hat{\beta}_1) = \beta_1$ $V(\hat{\beta}_1) = \dfrac{\sigma^2}{S_{xx}}$

Use this to find standard error

Under the assumption that the observations are normally and independently distributed, a $100(1-\alpha)\%$ **confidence interval on the slope** β_1 in simple linear regression is

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \qquad (11\text{-}29)$$

Mean and variance of the intercept estimator $E(\hat{\beta}_0) = \beta_0$ and $V(\hat{\beta}_0) = \sigma^2\left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right]$

Similarly, a $100(1-\alpha)\%$ **confidence interval on the intercept** β_0 is

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} \qquad (11\text{-}30)$$

$$\operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2\bar{x}/S_{xx}.$$

Confidence Interval for Slope

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n} y_i(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

Mean and variance of the slope estimator $E(\hat{\beta}_1) = \beta_1 \qquad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$

Under the assumption that the observations are normally and independently distributed, a $100(1-\alpha)\%$ **confidence interval on the slope** β_1 in simple linear regression is

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \qquad (11\text{-}29)$$

The width of confidence interval indicates the overall quality of regression line.

Confidence Interval for Intercept

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n} y_i(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

Mean and variance of the intercept estimator $E(\hat{\beta}_0) = \beta_0$ and $V(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$

Similarly, a $100(1 - \alpha)\%$ **confidence interval on the intercept** β_0 is

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} \qquad (11\text{-}30)$$

The width of confidence interval indicates the overall quality of regression line.

Gas purity example

EXAMPLE 11-4 Oxygen Purity Confidence Interval on the Slope

We will find a 95% confidence interval on the slope of the regression line using the data in Example 11-1. Recall that $\hat{\beta}_1 = 14.947$, $S_{xx} = 0.68088$, and $\hat{\sigma}^2 = 1.18$ (see Table 11-2). Then, from Equation 11-29 we find

$$\hat{\beta}_1 - t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

or

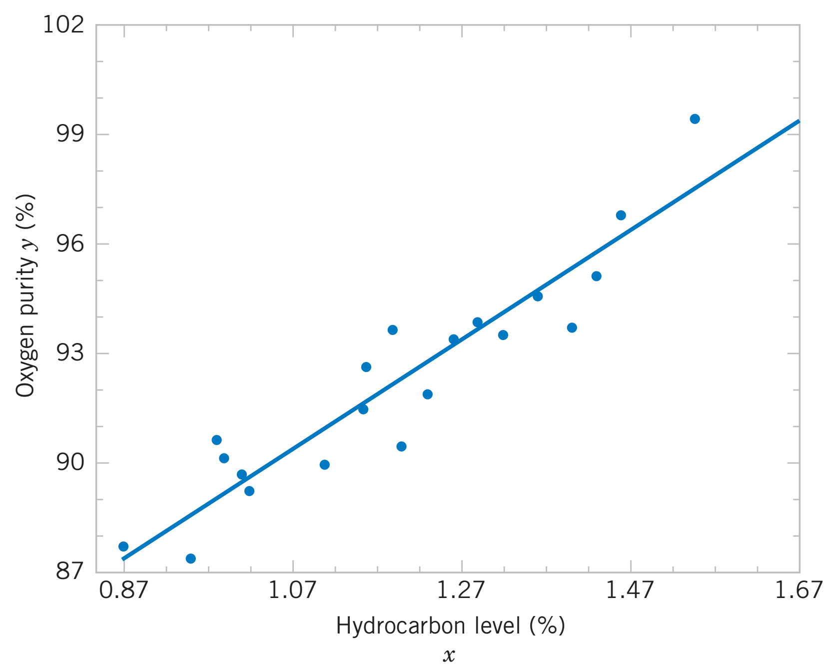
$$14.947 - 2.101\sqrt{\frac{1.18}{0.68088}} \le \beta_1 \le 14.947 + 2.101\sqrt{\frac{1.18}{0.68088}}$$

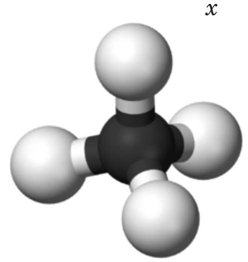
This simplifies to

$$12.181 \le \beta_1 \le 17.713$$

Practical Interpretation: This CI does not include zero, so there is strong evidence (at $\alpha = 0.05$) that the slope is not zero. The CI is reasonably narrow (± 2.766) because the error variance is fairly small.

Figure 11-4 Scatter plot of oxygen purity y versus hydrocarbon level x and regression model $\hat{y} = 74.283 + 14.947x$.





confint(model)

```
              2.5 %   97.5 %
(Intercept) 70.93555 77.63108
data[, 1]   12.18107 17.71389
```

Prediction interval

- Predicting response for **new** observation
- The **new** observation is **independent** of data used to build linear regression model

```
data <- read.table("Example_1.txt")
> x <- data[,1]
> y <- data[,2]
> model <- lm(y~x)
> predict(model, data.frame(x = 1), interval=c("prediction"))
```

A $100(1-\alpha)$ % **prediction interval on a future observation** Y_0 at the value x_0 is given by

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \le Y_0 \le \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]} \tag{11-33}$$

The value $\hat{y}_0$ is computed from the regression model $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Gas Purity Example

EXAMPLE 11-6 Oxygen Purity Prediction Interval

To illustrate the construction of a prediction interval, suppose we use the data in Example 11-1 and find a 95% prediction interval on the next observation of oxygen purity at $x_0 = 1.00\%$. Using Equation 11-33 and recalling from Example 11-5 that $\hat{y}_0 = 89.23$, we find that the prediction interval is

$$89.23 - 2.101\sqrt{1.18\left[1 + \frac{1}{20} + \frac{(1.00 - 1.1960)^2}{0.68088}\right]}$$

$$\le Y_0 \le 89.23 + 2.101\sqrt{1.18\left[1 + \frac{1}{20} + \frac{(1.00 - 1.1960)^2}{0.68088}\right]}$$

which simplifies to

$$86.83 \le y_0 \le 91.63$$

This is a reasonably narrow prediction interval.

Minitab will also calculate prediction intervals. Refer to the output in Table 11-2. The 95% PI on the future observation at $x_0 = 1.00$ is shown in the display.

By repeating the foregoing calculations at different levels of x_0, we may obtain the 95% prediction intervals shown graphically as the lower and upper lines about the fitted regression model in Fig. 11-8. Notice that this graph also shows the 95% confidence limits on $\mu_{Y|x_0}$ calculated in Example 11-5. It illustrates that the prediction limits are always wider than the confidence limits.

Hypothesis Test

Hypothesis Test on Regression Parameters

Slope:

Suppose we wish to test

$$H_0\colon \beta_1 = \beta_{1,0}$$
$$H_1\colon \beta_1 \neq \beta_{1,0}$$

An appropriate test statistic would be

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / S_{xx}}}$$

Intercept:

Suppose we wish to test

$$H_0\colon \beta_0 = \beta_{0,0}$$
$$H_1\colon \beta_0 \neq \beta_{0,0}$$

An appropriate test statistic would be

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}}$$

We would reject the null hypothesis if

$$|t_0| > t_{\alpha/2, n-2}$$

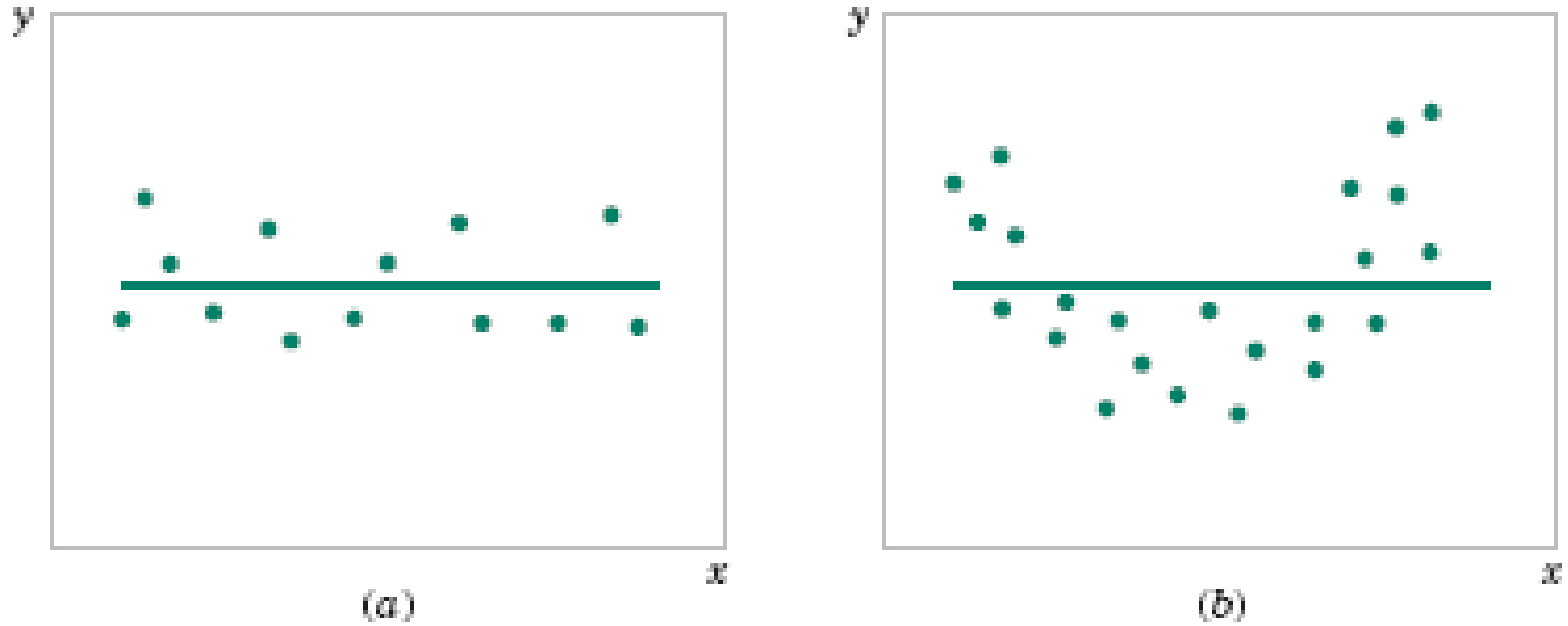
Confidence Intervals can also be used to test the above hypotheses.

Hypothesis Test on Slope

An important special case of the hypotheses on the slope is

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

These hypotheses relate to the significance of regression.
Failure to reject H_0 is equivalent to concluding that there is no linear relationship between *x* and *Y*.



Continue house price

```
Call:
lm(formula = tax ~ price)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.4262 -0.3310  0.1312  0.4967  1.3135 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -1.5844     0.9514  -1.665     0.11    
price         0.2308     0.0271   8.518 2.05e-08 ***
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7802 on 22 degrees of freedom
Multiple R-squared:  0.7673,	Adjusted R-squared:  0.7568 
F-statistic: 72.56 on 1 and 22 DF,  p-value: 2.051e-08
```

Regression model: $\underbrace{\boxed{y}}_{\text{tax}} = 0.2308\underbrace{\boxed{x}}_{\text{price}} - 1.5844$

Deal with non-linearity

Regression on transformed variables

- **Deal with non-linearity: Sometimes visual inspections, or prior knowledge, tells us that there are some non-linear factors in regression model**
- **Examples:**

$$Y = \beta_0 e^{\beta_1 x} \epsilon \qquad \boxed{\ln Y} = \boxed{\ln \beta_0} + \beta_1 x + \ln \epsilon$$

$$Y = \beta_0 + \beta_1 \boxed{\left(\frac{1}{x}\right)} + \epsilon$$

$$y = \beta_0 + \beta_1 x + \beta_2 \boxed{x^2} + \epsilon$$

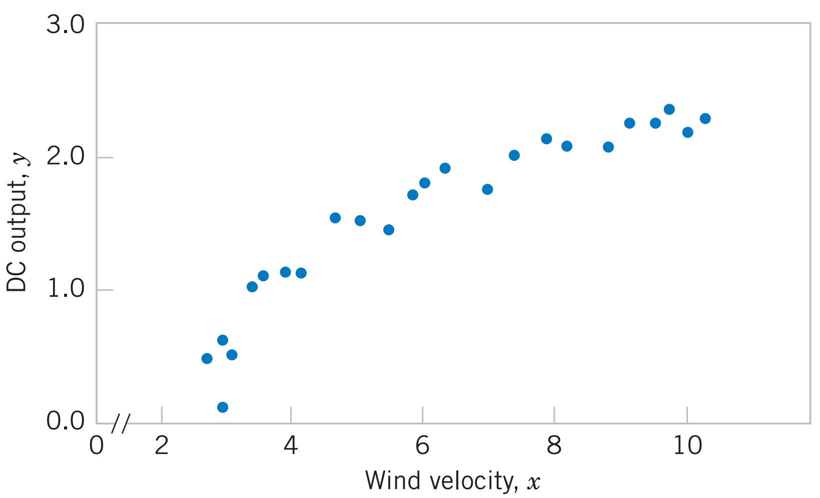


Figure 11-14 Plot of DC output y versus wind velocity x for the windmill data.

Example 4: Wind-mill power

A research engineer is investigating the use of a windmill to generate electricity and has collected data on the DC output from this windmill and the corresponding wind velocity. The data are plotted in Figure 11-14 and listed in Table 11-5 (p.439).

Observation Number, i	Wind Velocity (mph), x_i	DC Output, y_i
1	5.00	1.582
2	6.00	1.822
3	3.40	1.057
4	2.70	0.500
5	10.00	2.236
6	9.70	2.386
7	9.55	2.294
8	3.05	0.558
9	8.15	2.166
10	6.20	1.866
11	2.90	0.653
12	6.35	1.930
13	4.60	1.562
14	5.80	1.737
15	7.40	2.088
16	3.60	1.137
17	7.85	2.179
18	8.80	2.112
19	7.00	1.800
20	5.45	1.501
21	9.10	2.303
22	10.20	2.310
23	4.10	1.194
24	3.95	1.144
25	2.45	0.123

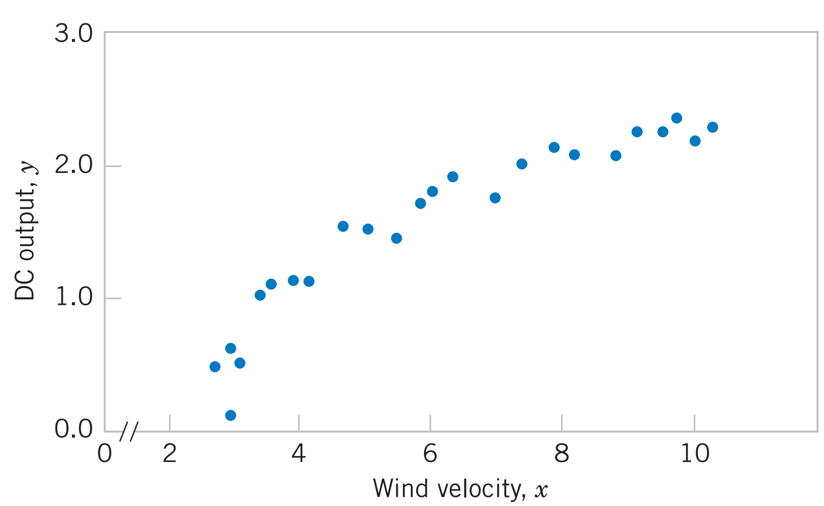


Figure 11-14 Plot of DC output y versus wind velocity x for the windmill data.

Try fitting a linear model?

- **Result of fitting linear regression model**

$$\hat{y} = 0.1309 + 0.2411x$$

The summary statistics for this model are $R^2 = 0.8745$, $MS_E = \hat{\sigma}^2 = 0.0557$, and $F_0 = 160.26$ (the P-value is <0.0001).

- **Residual plot indicates the linear relationship does not capture all the information in the wind-speed variable.**

Residual

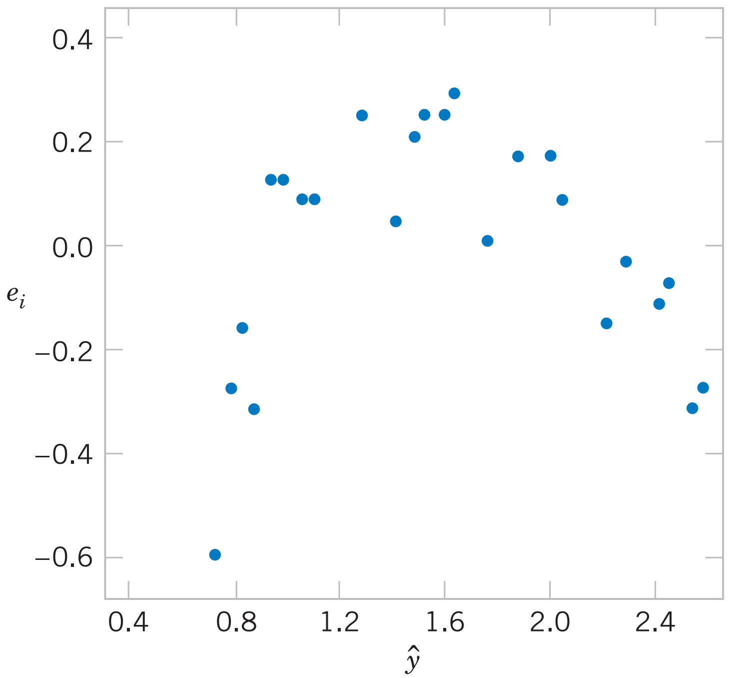


Figure 11-15 Plot of residuals e_i versus fitted values $\hat{y}_i$ for the windmill data.

A second try

- **As wind speed increases, output (y) approach to an upper limit (consist with physics of windmill operation)**

$$y = \beta_0 + \beta_1\left(\frac{1}{x}\right) + \epsilon$$

Raw data

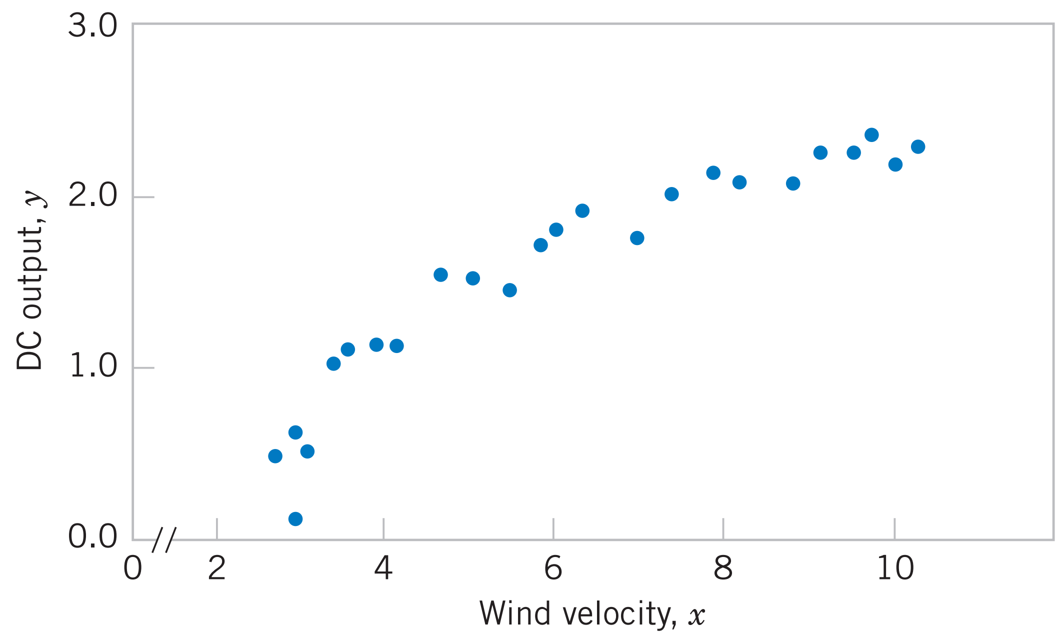


Figure 11-14 Plot of DC output y versus wind velocity x for the windmill data.

Transformed data

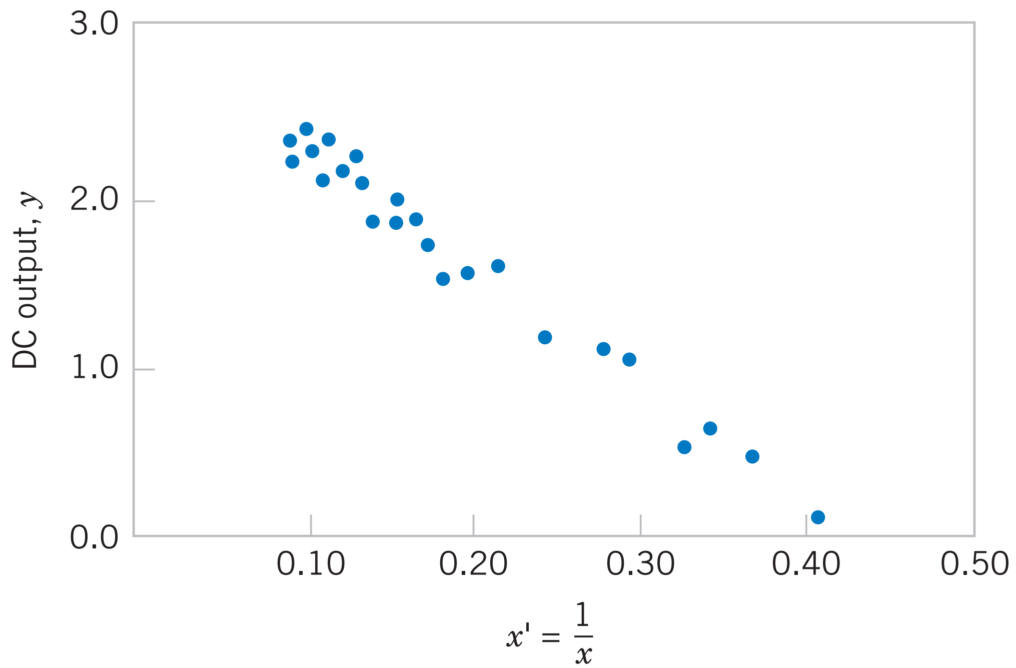


Figure 11-16 Plot of DC output versus $x' = 1/x$ for the windmill data.

$$x' = 1/x.$$

$$\hat{y} = 2.9789 - 6.9345x'$$

The summary statistics for this model are $R^2 = 0.9800$, $MS_E = \hat{\sigma}^2 = 0.0089$, and $F_0 = 1128.43$ (the P value is <0.0001).

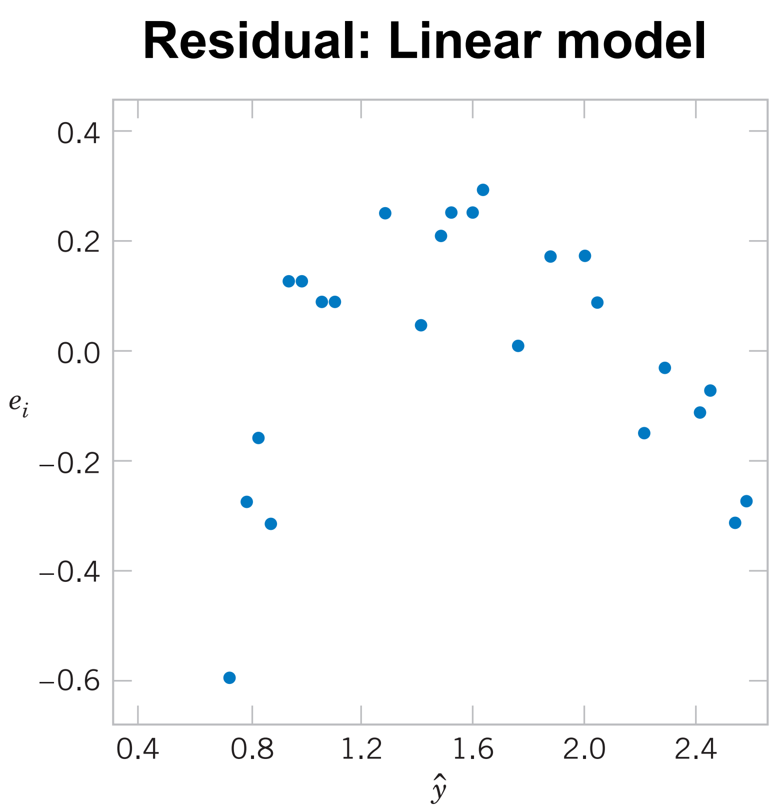


Figure 11-15 Plot of residuals e_i versus fitted values $\hat{y}_i$ for the windmill data.

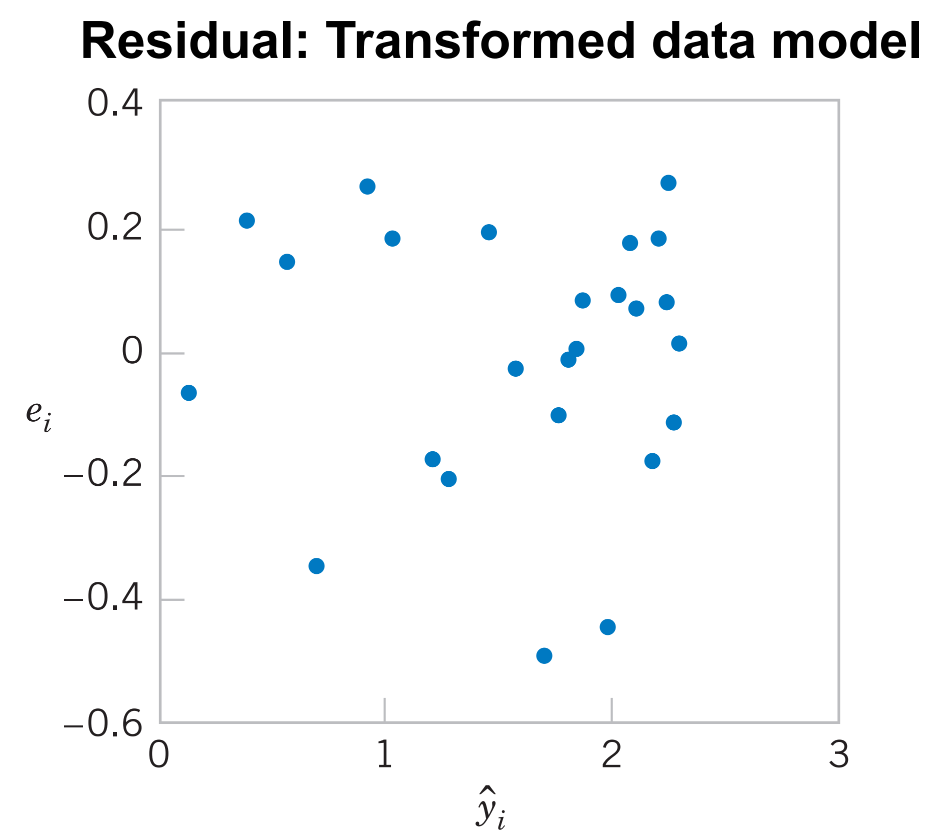


Figure 11-17 Plot of residuals versus fitted values $\hat{y}_i$ for the transformed model for the windmill data.

Summary

- Simple linear regression (one predictor)
- Method-of-least-square to find coefficient
- Model diagnosis
  - Residual diagnosis: plot, normal plot, histogram
  - R-score
  - Confidence interval (slope, intercept, prediction)
  - Hypothesis test (significance of linear model)
- Deal with non-linearity

More example

11-6. The following table presents the highway gasoline mileage performance and engine displacement for Daimler-Chrysler vehicles for model year 2005 (source: U.S. Environmental Protection Agency).

(a) Fit a simple linear model relating highway miles per gallon (y) to engine displacement (x) in cubic inches using least squares.

(b) Find an estimate of the mean highway gasoline mileage performance for a car with 150 cubic inches engine displacement.

(c) Obtain the fitted value of y and the corresponding residual for a car, the Neon, with an engine displacement of **348** cubic inches.

Carline	Engine Displacement (in^3)	MPG (highway)
300C/SRT-8	215	30.8
CARAVAN 2WD	201	32.5
CROSSFIRE ROADSTER	196	35.4
DAKOTA PICKUP 2WD	226	28.1
DAKOTA PICKUP 4WD	226	24.4
DURANGO 2WD	348	24.1
GRAND CHEROKEE 2WD	226	28.5
GRAND CHEROKEE 4WD	348	24.2
LIBERTY/CHEROKEE 2WD	148	32.8