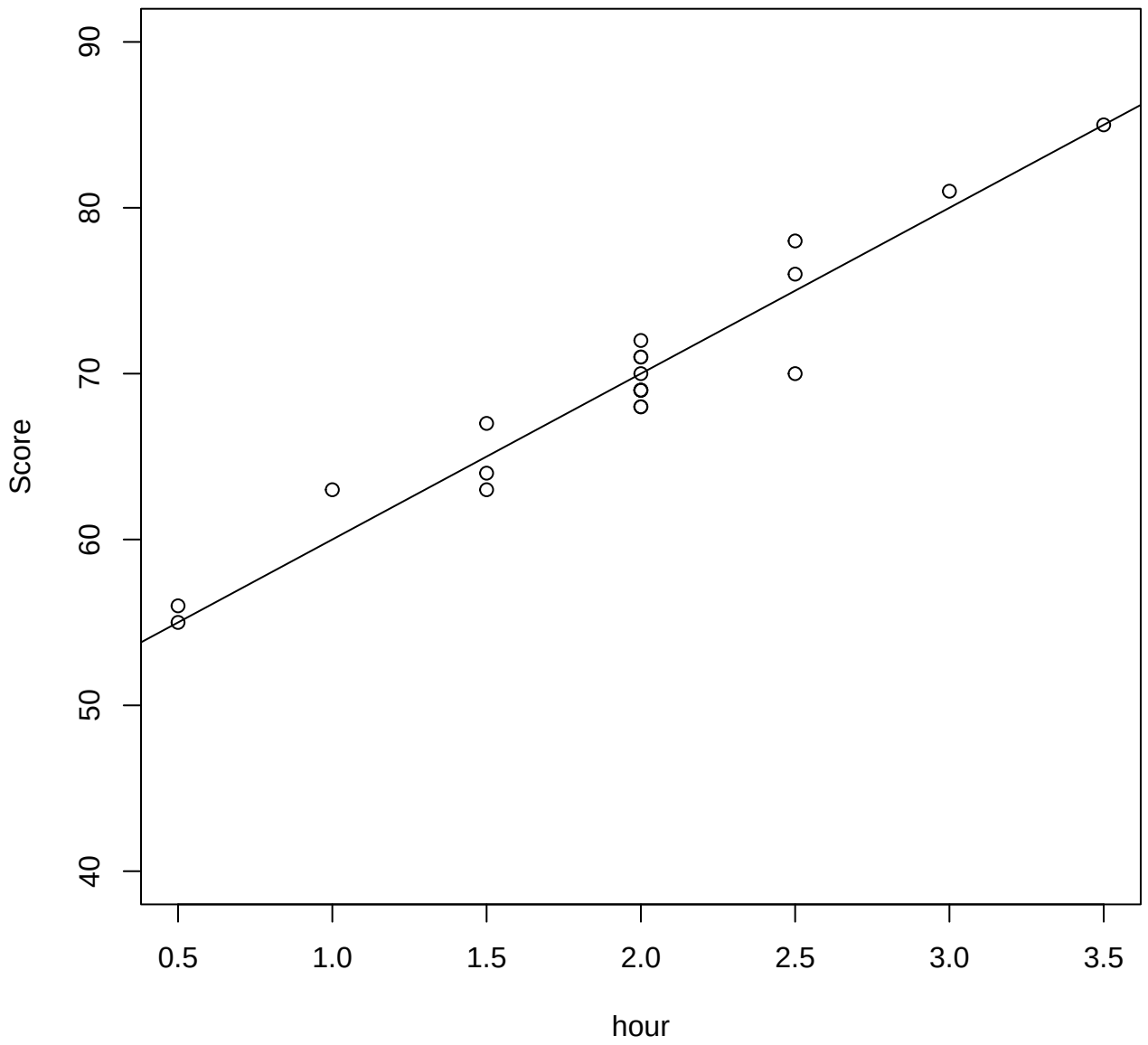
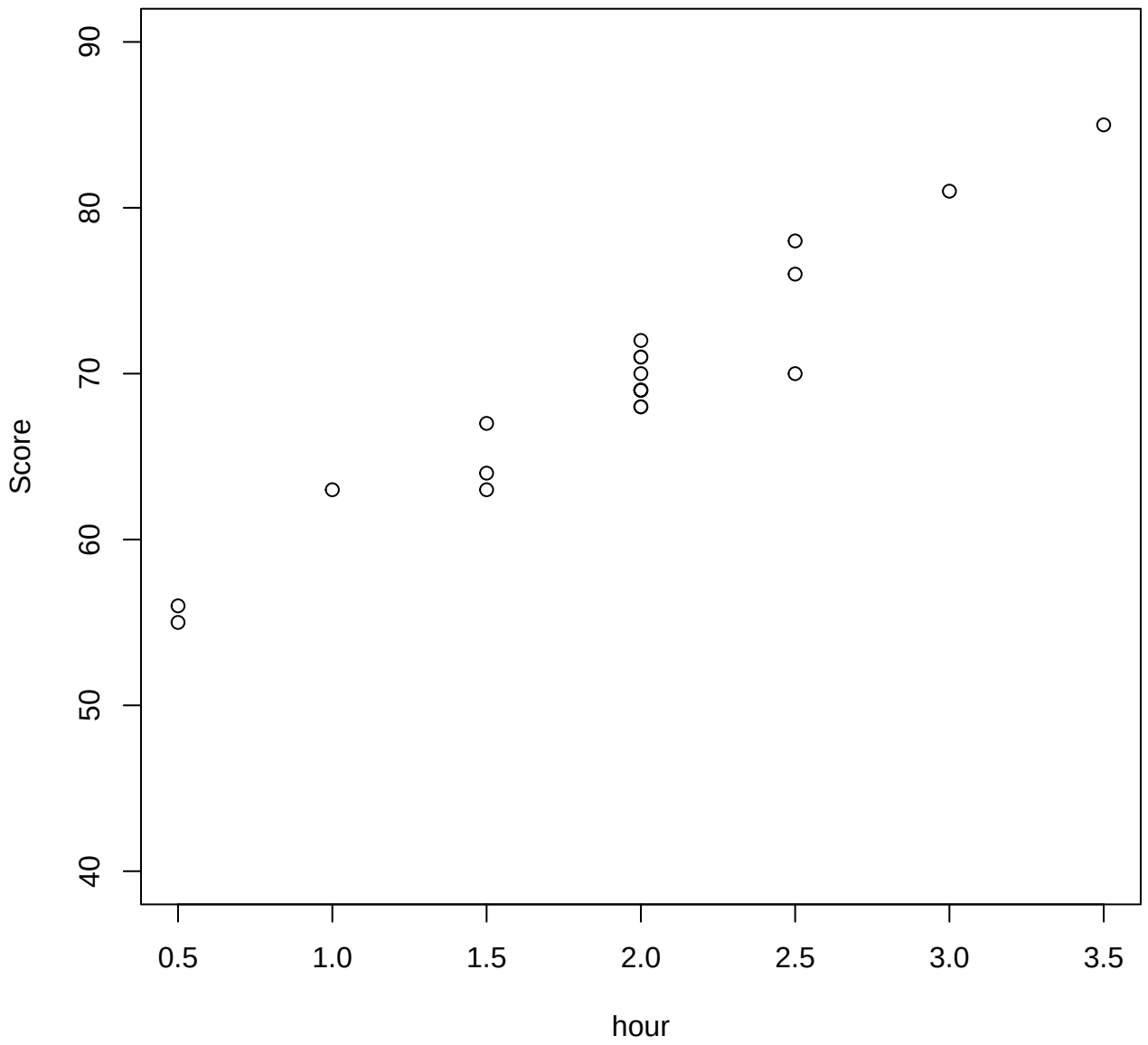
Chapter 2: Simple Linear Regression

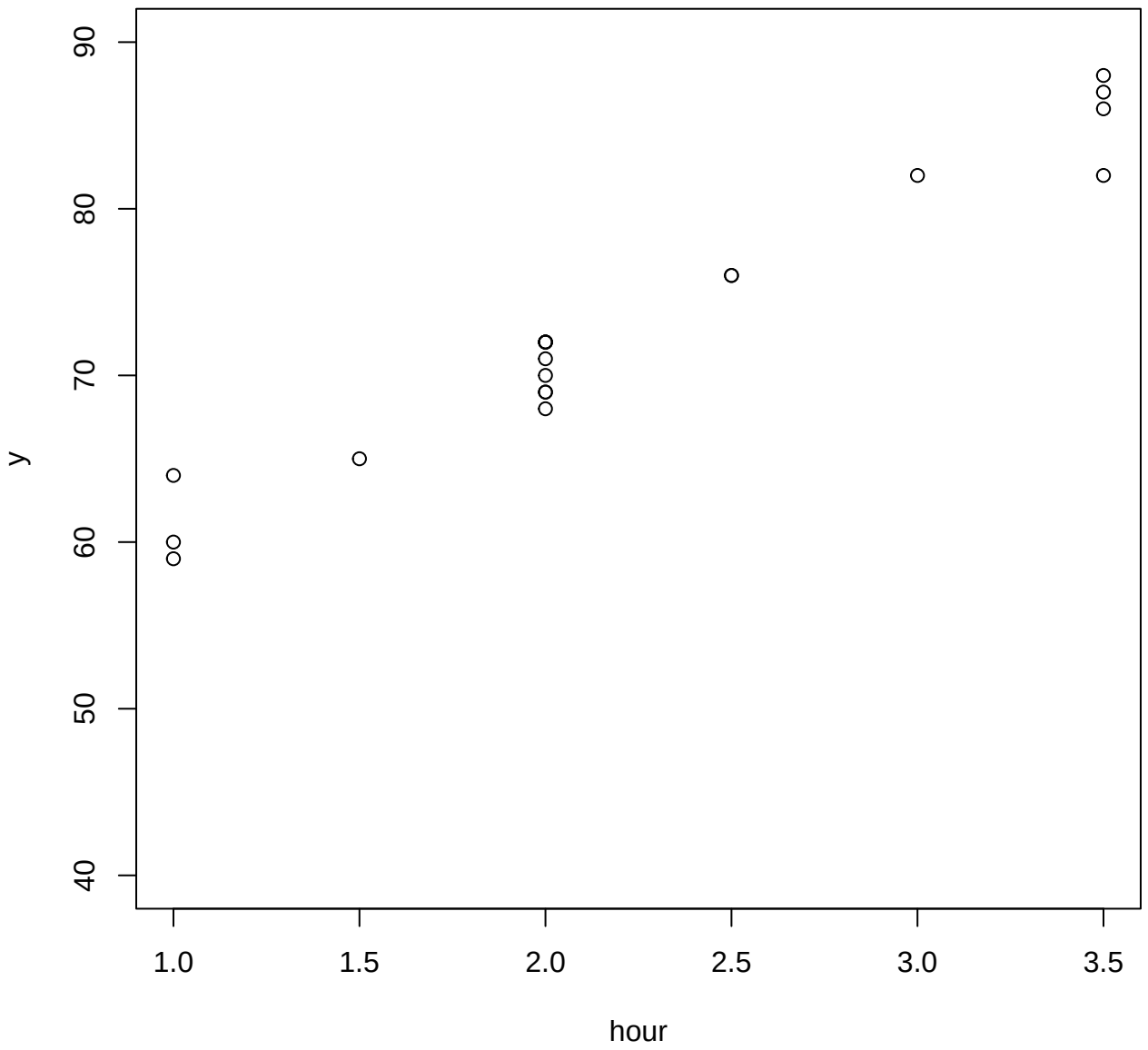
- 2.1 The Model
- 2.2 Parameter Estimation and Estimator Properties
- 2.3 Hypothesis Testing
- 2.4 Confidence Intervals for Parameters
- 2.5 Prediction
- More Detail on Degrees of Freedom; ANOVA
- 2.6 Coefficient of Determination
- 2.12 Maximum Likelihood Estimation; Gauss-Markov Theorem
- 2.11 Regression Through the Origin
- 2.10 Issues Which May Arise - Hazards of Regression

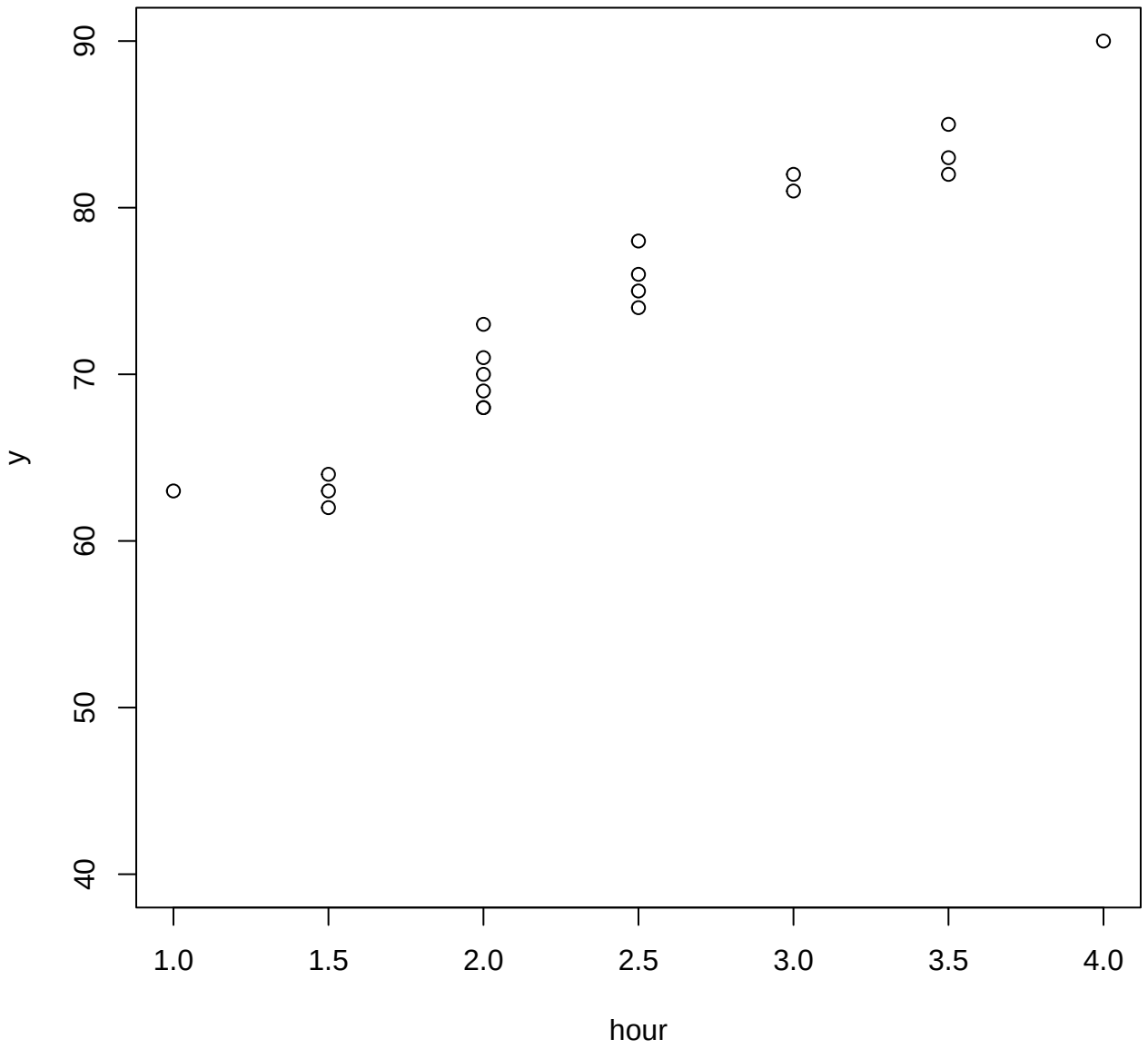
Score and Hours in the Trainning Program

ID	Score	hours
1	76	2.5
2	67	1.5
3	69	2.0
4	55	0.5
5	85	3.5
6	69	2.0
7	64	1.5
8	63	1.0
9	68	2.0
10	70	2.5
11	78	2.5
12	81	3.0
13	70	2.0
14	68	2.0
15	56	0.5
16	63	1.5
17	71	2.0
18	69	2.0
19	72	2.0
20	71	2.0


Score
90
80
70
60
50
40
0.5
1.0
1.5
2.0
2.5
3.0
3.5
hour


Score
90
80
70
60
50
40
0.5
1.0
1.5
2.0
2.5
3.0
3.5
hour


y
90
80
70
60
50
40
1.0
1.5
2.0
2.5
3.0
3.5
hour


y
90
80
70
60
50
40
1.0
1.5
2.0
2.5
3.0
3.5
4.0
hour

2.1 The Model

- Measurement of y (response) changes in a linear fashion with a setting of the variable x (predictor):

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{linear relation}} + \underbrace{\varepsilon}_{\text{noise}}$$

  - The linear relation is deterministic (non-random).
  - The noise or error is random.

- Noise accounts for the variability of the observations about the straight line.
  No noise $\Rightarrow$ relation is deterministic.
  Increased noise $\Rightarrow$ increased variability.

Using simulation to see effects of noise

Experiment with this R program:

```
simple.sim <- function(intercept=0,slope=1,x=seq(1,10),
              sigma=1, ...)
{
    noise <- rnorm(length(x),sd = sigma)
    y <- intercept + slope*x + noise
    title1 <- bquote(sigma == .(sigma))
    plot(x,y,pch=16,main=title1, ...)
    abline(intercept,slope,col=4,lwd=2)
}
```

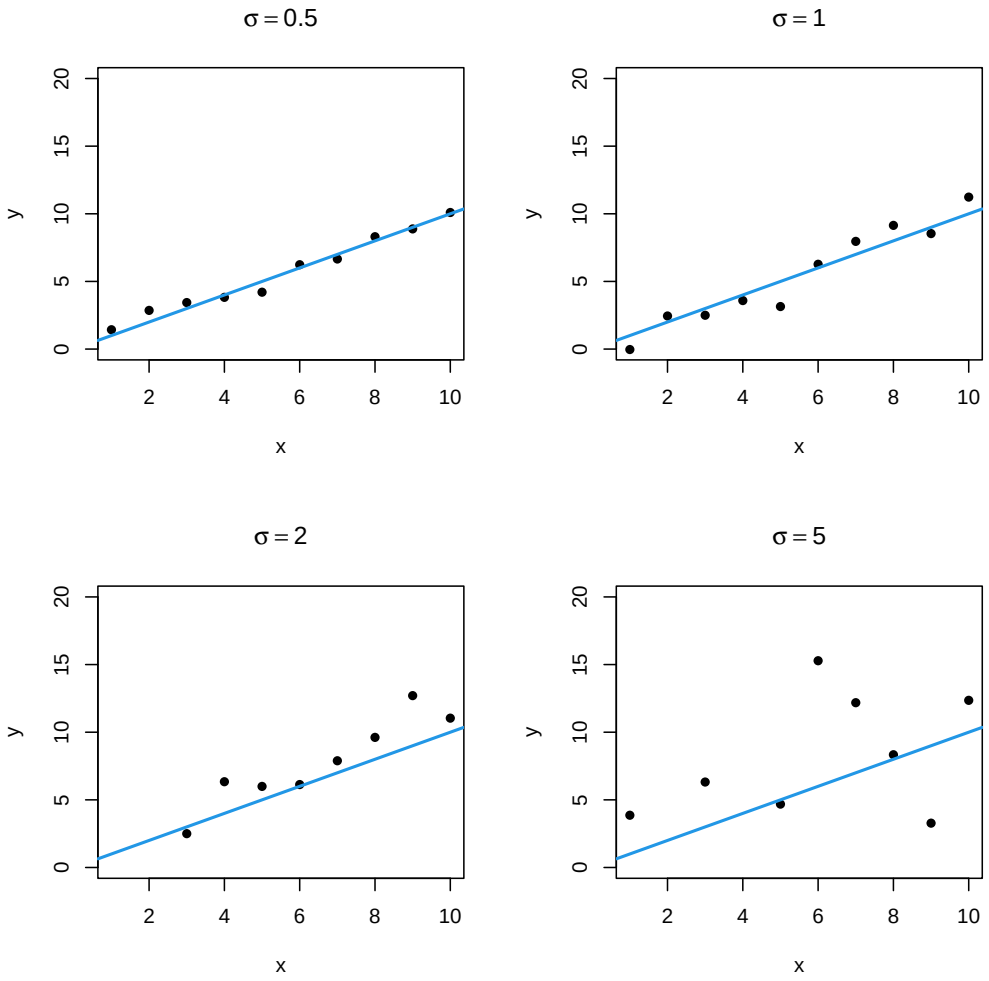
Using simulation to see effects of noise

```
simple.sim(sigma=.5,
    ylim=c(0,20))

simple.sim(sigma=1,
    ylim=c(0,20))

simple.sim(sigma=2,
    ylim=c(0,20))

simple.sim(sigma=5,
    ylim=c(0, 20))
```



The Setup

- Assumptions:
  1. $E[y|x] = \beta_0 + \beta_1 x$.
  2. $\text{Var}(y|x) = \text{Var}(\beta_0 + \beta_1 x + \varepsilon|x) = \sigma^2$.
- Data: Suppose data $y_1, y_2, \ldots, y_n$ are obtained at settings $x_1, x_2, \ldots, x_n$, respectively. Then the model on the data is

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

(ε_i i.i.d. N(0, σ^2) and $E[y_i|x_i] = \beta_0 + \beta_1 x_i$.)

Either

1. the x's are fixed values and measured without error (controlled experiment)

   OR
2. the analysis is conditional on the observed values of x (observational study).

Parameter Estimation, Fitted Values and Residuals

1. Least Squares Estimation

   Distributional assumptions are not required

2. Maximum Likelihood Estimation (covered later)

   Distributional assumptions are required

Least Squares Estimation

- Assumptions:

  1. $E[\varepsilon_i] = 0$
  2. Var$(\varepsilon_i) = \sigma^2$
  3. ε_i's are independent.

- Note that normality is not required.

Method

- Minimize

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

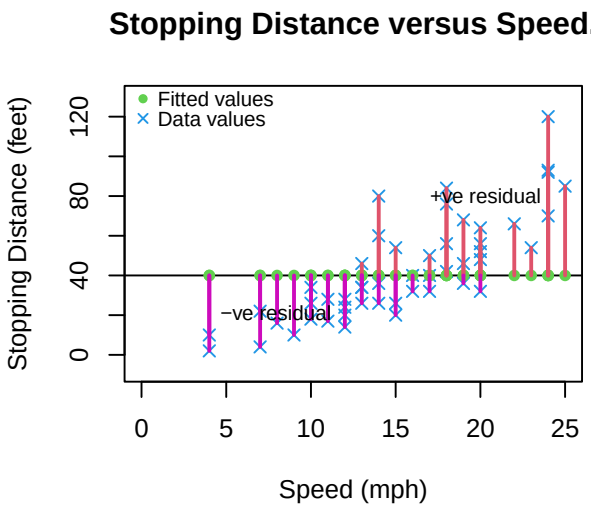
with respect to the parameters or regression coefficients β_0 and β_1: $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

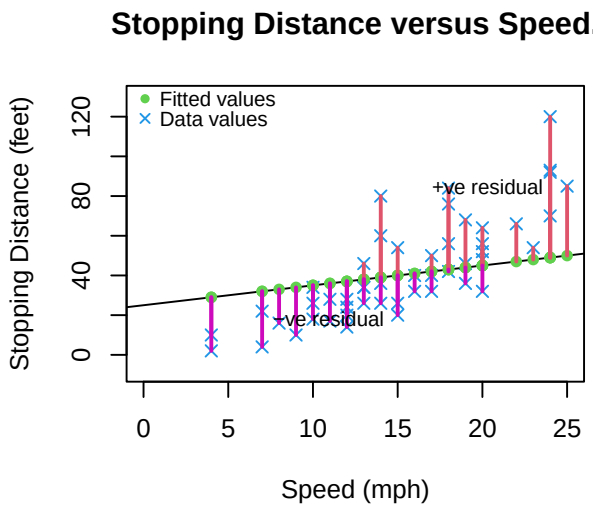
- Justification: We want the fitted line to pass as close to all of the points as possible.

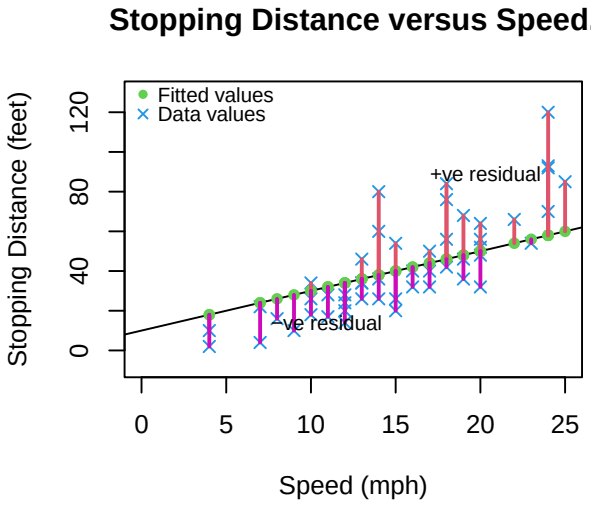
- Aim: small Residuals (observed - fitted response values):

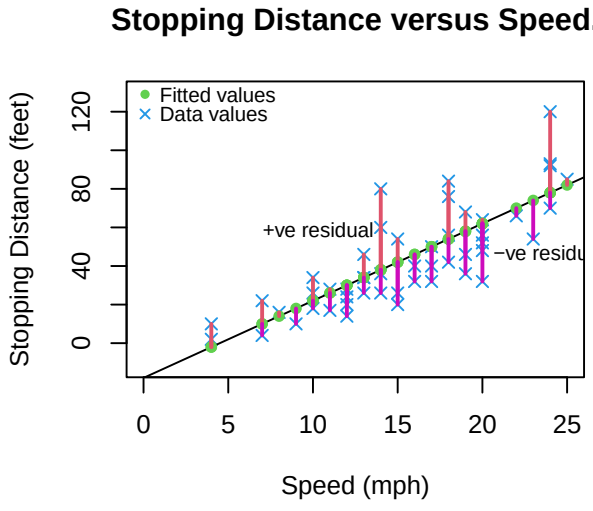
$$e_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$$

Look at the following plots:









Red bars: lengths of residuals

Data above fitted line ⇝ +ve residuals

Data below fitted line ⇝ -ve residuals

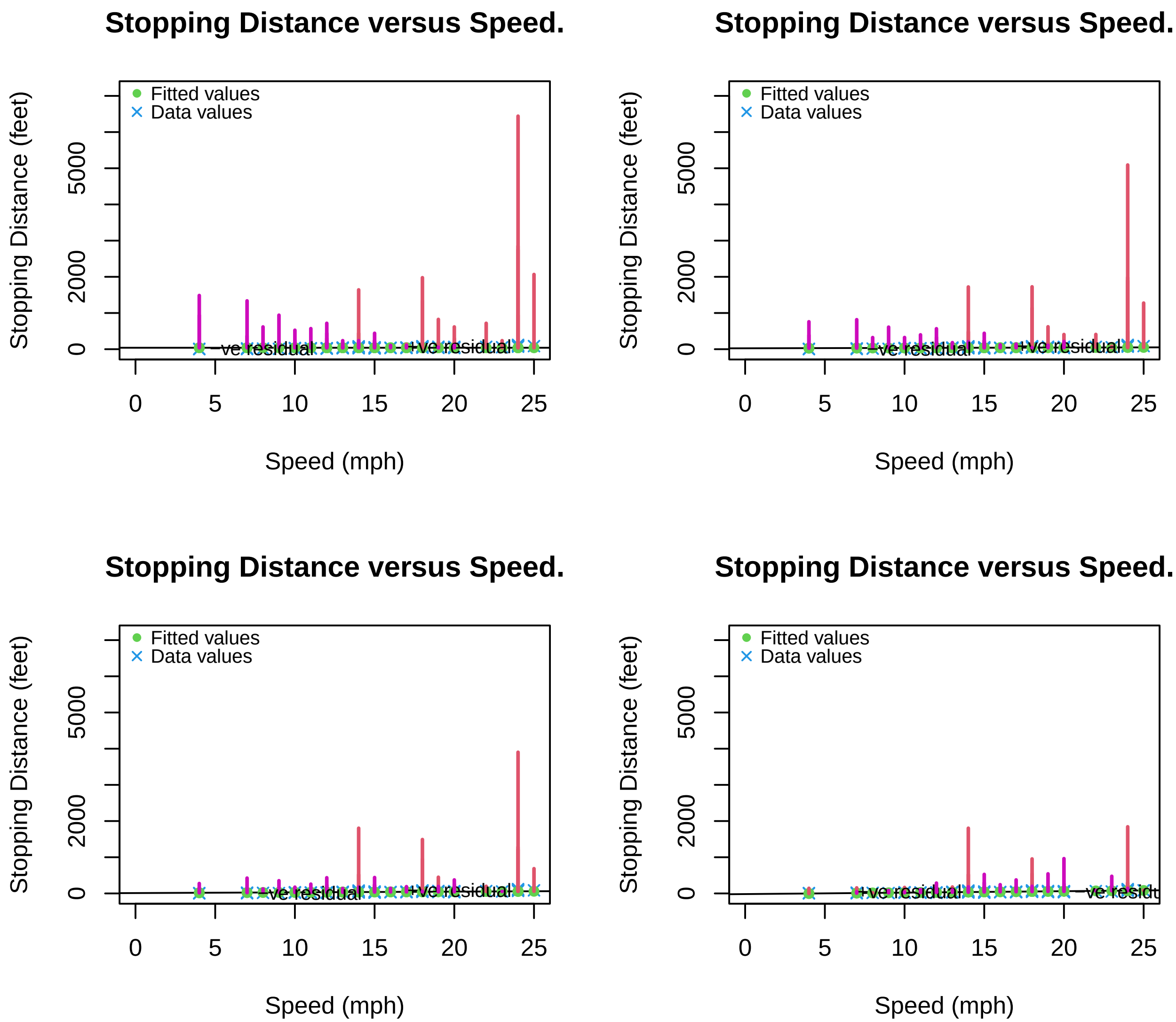
Observations

The first three lines do not pass as close to the plotted points as the fourth, even though the sum of the residuals is about the same in all four cases.

Negative residuals cancel out positive residuals.

The Key: minimize squared residuals

Look at the following plots:



Red bars: lengths of squared residuals

Data above fitted line $\leadsto$ +ve residuals^2

Data below fitted line $\leadsto$ +ve residuals^2

$\Rightarrow$ Minimizing sum of squared residuals $\leadsto$ small residuals.

Regression Coefficient Estimators

The minimizers of

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

are

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

and

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

where

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

and

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

Calculator Formulas

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

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

Example to Illustrate Estimation Procedure

- First few records of the `cars` data:

```
head(cars)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

- The total number of observations in the data set is 50. i.e. $n = 50$.

Hand Calculation to Illustrate Estimation Procedure

(y = dist, x = speed)

- $\sum_{i=1}^{50} x_i = 4 + 4 + \cdots 25 = 770$
- $\bar{x} = \frac{770}{50} = 15.4$
- $\sum_{i=1}^{50} y_i = 2 + 10 + \cdots 85 = 2149$
- $\bar{y} = \frac{2149}{50} = 42.98$

Hand Calculation to Illustrate Estimation Procedure

- $\sum_{i=1}^{50} x_i^2 = 4^2 + 4^2 + \cdots + 25^2 = 1.323 \times 10^4$
- $\sum_{i=1}^{50} y_i^2 = 2^2 + 10^2 + \cdots + 85^2 = 1.249 \times 10^5$
- $\sum_{i=1}^{50} x_i y_i = (4)(2) + \cdots (25)(85) = 3.848 \times 10^4$

Hand Calculation to Illustrate Estimation Procedure

- $S_{xx} = 1.323 \times 10^4 - \frac{(770)^2}{50} = 1370$
- $S_{xy} = 3.848 \times 10^4 - \frac{(770)(2149)}{50} = 5387.4$
- $S_{yy} = 1.249 \times 10^5 - \frac{(2149)^2}{50} = 3.254 \times 10^4$

Hand Calculation to Illustrate Estimation Procedure

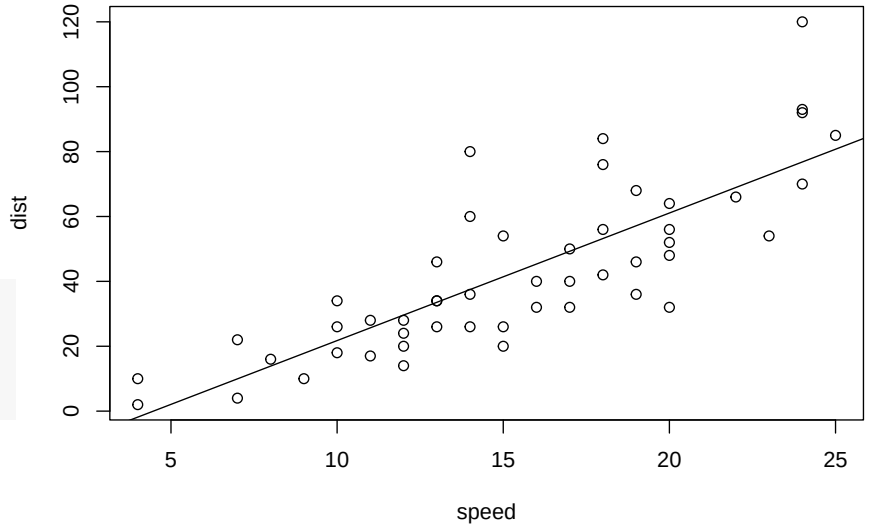
- $\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{5387.4}{1370} = 3.932$

- $\widehat{\beta}_0 = \bar{y} - \beta_1 \bar{x} = 42.98 - 3.932(15.4) = -17.579$

  $\rightsquigarrow$ **Fitted Line:** $\widehat{y} = -17.579 + 3.932x$

Code to plot data with overlaid fitted line:

```
plot(cars)
abline(beta0, beta1)
```



HomeMade R Intercept and Slope Estimators

```
ls.est
## function (data)
## {
##     x <- data[, 1]
##     y <- data[, 2]
##     S <- function(x, y) {
##         sum((x - mean(x)) * (y - mean(y)))
##     }
##     Sxy <- S(x, y)
##     Sxx <- S(x, x)
##     b1 <- Sxy/Sxx
##     b0 <- mean(y) - mean(x) * b1
##     list(ests = c(b0 = b0, b1 = b1), data = data)
## }
```

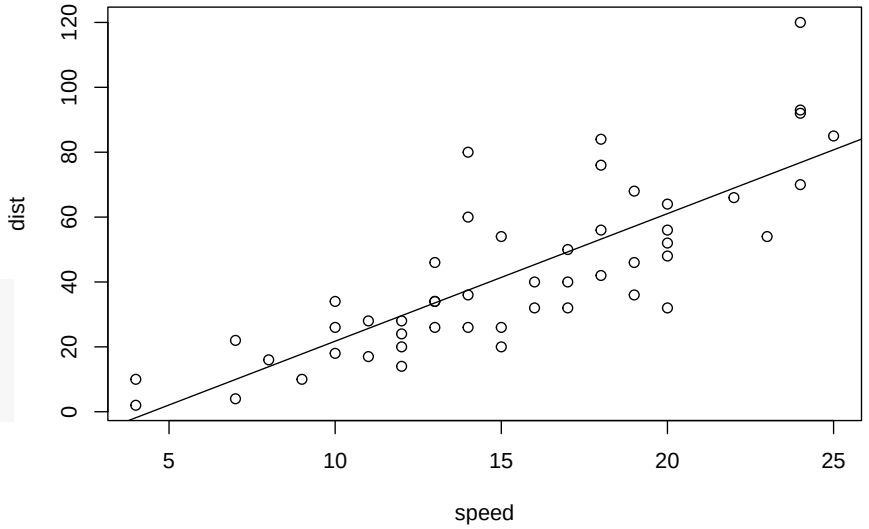
R commands (home-made version)

```
source("ls.est.R")
cars.obj <- ls.est(cars) # estimate regression parameters
cars.obj$ests      # intercept and slope estimate


##      b0      b1
## -17.58    3.93
```

Code to plot data with overlaid fitted line:

```
plot(cars)
abline(cars.obj$ests)
```



Residuals

$$\widehat{\varepsilon}_i = e_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$$

$$= y_i - \widehat{y}_i$$

$$= y_i - \bar{y} - \widehat{\beta}_1 (x_i - \bar{x})$$

Residuals in R:

```
source("res.R")
res


## function (ls.object)
## {
##     b0 <- ls.object$ests["b0"]
##     b1 <- ls.object$ests["b1"]
##     x <- ls.object$data[, 1]
##     y <- ls.object$data[, 2]
##     resids <- y - b0 - b1 * x
##     resids
## }
```

Residuals in R (Example):

```
res(cars.obj)[1:8]     # first 8 residuals

## [1]  3.85 11.85 -5.95 12.05  2.12 -7.81
## [7] -3.74  4.26

sum(res(cars.obj))   # observe: sum of residuals is 0

## [1] 3.69e-13
```

Estimation of σ^2

- The residual sum of squares or error sum of squares is given by

$$SSE = \sum_{i=1}^{n} e_i^2 = S_{yy} - \widehat{\beta}_1 S_{xy}$$

  Note: $SSE = SS_{\text{Res}}$ and $S_{yy} = SS_{\text{T}}$

- An unbiased estimator for the error variance is

$$\widehat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{\left[S_{yy} - \widehat{\beta}_1 S_{xy}\right]}{n-2}$$

  Note: $MSE = MS_{\text{Res}}$

- $\widehat{\sigma}$ = Residual Standard Error = $\sqrt{\text{MSE}}$

Estimation of σ^2

The crude way to calculate the MSE is by summing the squares of the residuals and dividing by $n - 2$:

```
sum(res(cars.obj)^2)/(nrow(cars)-2)   #   (MSE)


## [1] 237
```

The next slide shows how to obtain the variance estimate without calculating all of the residuals.

Hand Calculation to Illustrate Variance Estimation Procedure

$$\widehat{\sigma}^2 = \frac{1}{n-2}(S_{yy} - \widehat{\beta}_1 S_{xy})$$

$$= \frac{1}{48}(3.254 \times 10^4 - 3.932(5387.4)) = 236.532 = \text{MSE}$$

Example Summary: the fitted regression line relating stopping distance (y) to speed (x) is

$$\widehat{y} = -17.579 + 3.932x$$

The error variance is estimated as MSE $= 236.532$.

The root mean square error is $\widehat{\sigma} = 15.38$.

R commands (built-in version)

```
cars.lm <- lm(dist ~ speed, data = cars)
summary(cars.lm)$coefficients

##             Estimate Std. Error t value
## (Intercept)   -17.58      6.758   -2.60
## speed           3.93      0.416    9.46
##             Pr(>|t|)
## (Intercept) 1.23e-02
## speed       1.49e-12
```

Using Extractor Functions

Partial Output:

```
coef(cars.lm)
## (Intercept)       speed
##      -17.58        3.93
```

```
summary(cars.lm)$sigma
## [1] 15.4
```

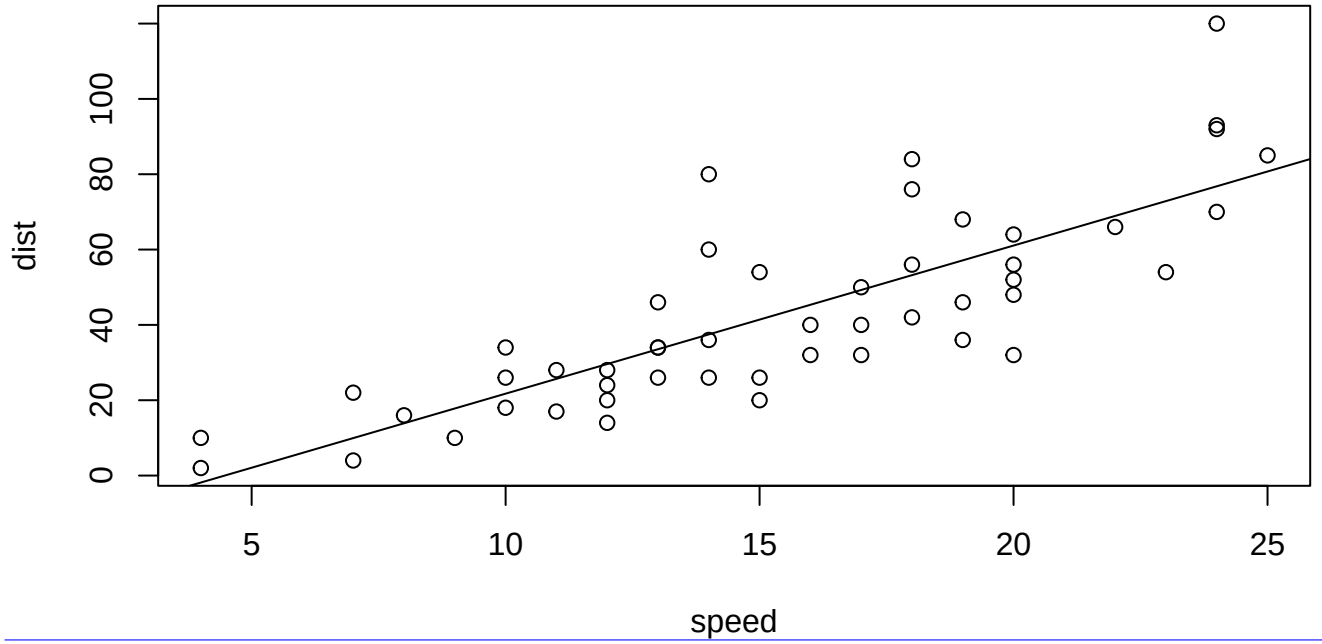
From the output,

- the slope estimate is $\widehat{\beta}_1 = 3.932$.
- the intercept estimate is $\widehat{\beta}_0 = -17.579$.

- Estimate of σ: the Residual standard error is 15.38 which is the square root of the MSE: $\widehat{\sigma}^2 = 236.532$.

Other R commands

- fitted values: `predict(cars.lm)`
- residuals: `resid(cars.lm)`
- diagnostic plots: `plot(cars.lm)` (these include a plot of the residuals against the fitted values and a normal probability plot of the residuals)
- Also `plot(cars); abline(cars.lm)` (this gives a plot of the data with the fitted line overlaid)



Consequences of Least-Squares

1. $e_i = y_i - \bar{y} - \widehat{\beta}_1(x_i - \bar{x})$
   (follows from intercept formula)
2. $\sum_{i=1}^{n} e_i = 0$ (follows from 1.)
3. $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i$
   (follows from 2.)
4. The regression line passes through the centroid $(\bar{x}, \bar{y})$ (follows from intercept formula)
5. $\sum x_i e_i = 0$
   (set partial derivative of $S(\beta_0, \beta_1)$ wrt β_1 to 0)
6. $\sum \widehat{y}_i e_i = 0$
   (follows from 2. and 5.)

Properties of Least Squares Estimators

- $E[\widehat{\beta}_1] = \beta_1$
- $E[\widehat{\beta}_0] = \beta_0$
- $\text{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$
- $\text{Var}(\widehat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$

Standard Error Estimators

- $\widehat{\text{Var}}(\widehat{\beta}_1) = \dfrac{\text{MSE}}{S_{xx}}$

  so the standard error (s.e.) of $\widehat{\beta}_1$ is estimated by

$$\sqrt{\frac{\text{MSE}}{S_{xx}}}$$

- `cars` e.g.: $\text{MSE} = 236.532$, $S_{xx} = 1370$

  $\Rightarrow$ s.e. of $\widehat{\beta}_1$ is $\sqrt{236.532/1370} = 0.416$

- $\widehat{\text{Var}}(\widehat{\beta}_0) = \text{MSE}\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right)$

  so the standard error (s.e.) of $\widehat{\beta}_0$ is estimated by

$$\sqrt{\text{MSE}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

- `cars` e.g.:

  $\bar{x} = 15.4$, $n = 50$

  $\Rightarrow$ s.e. of $\widehat{\beta}_0$ is $\sqrt{236.532\left(\frac{1}{50} + \frac{15.4^2}{1370^2}\right)} = 6.758$

Distributions of $\widehat{\beta}_1$ and $\widehat{\beta}_0$ (assuming normal responses)

- y_i is $N(\beta_0 + \beta_1 x_i, \sigma^2)$, and

  $\widehat{\beta}_1 = \sum_{i=1}^{n} a_i y_i \qquad (a_i = \frac{x_i - \bar{x}}{S_{xx}})$

  $\Rightarrow \widehat{\beta}_1$ is $N(\beta_1, \frac{\sigma^2}{S_{xx}})$.

  Also, $\frac{SSE}{\sigma^2}$ is χ^2_{n-2} (independent of $\widehat{\beta}_1$) so

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{xx}}} \sim t_{n-2}$$

  $\widehat{\beta}_0 = \sum_{i=1}^{n} c_i y_i \qquad (c_i = \frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{S_{xx}})$

  $\Rightarrow \widehat{\beta}_0$ is $N(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right))$ and

$$\frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\text{MSE}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}$$