Chapter 2

Simple Linear Regression

Summary

SOME CONSIDERATIONS IN THE USE OF REGRESSION

1. Regression models are intended as interpolation equations over the range of the regressor variable(s) used to fit the model.

2. The disposition of the x values plays an important role in the least-squares fit.

3. **Outliers** are observations that differ considerably from the rest of the data. They can seriously disturb the least-squares fit.

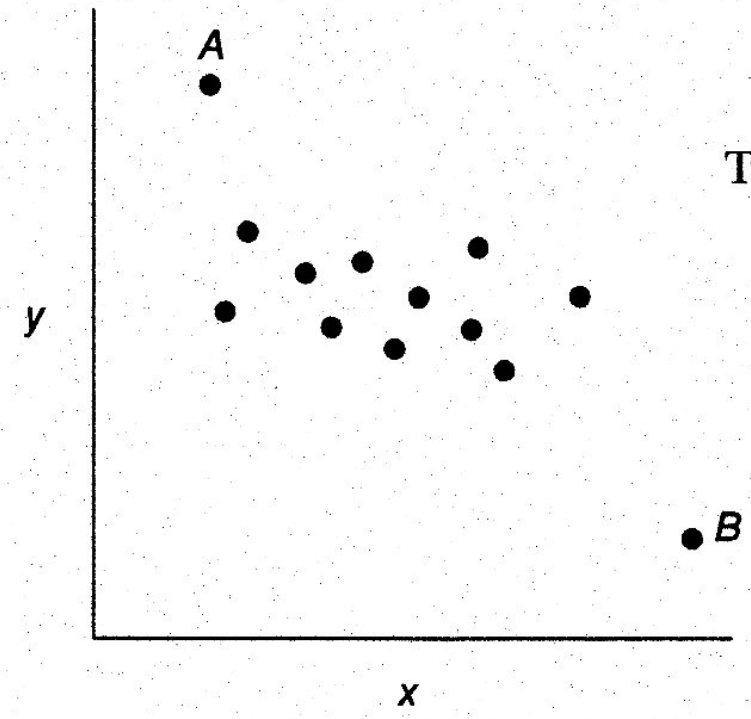
4. strong relationship between two variables, this does not imply that the variables are related in any causal sense. Causality implies necessary correlation.

5. In some applications of regression the value of the regressor variable x required to predict y is unknown.

SOME CONSIDERATIONS IN THE USE OF REGRESSION

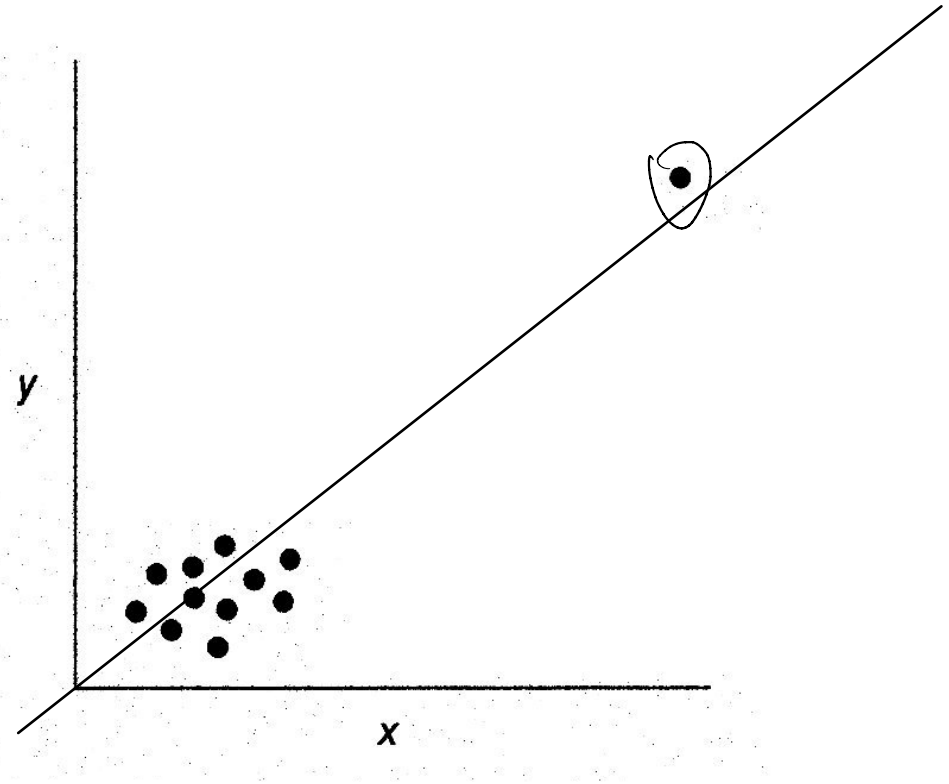
1. Regression models are intended as interpolation equations over the range of the regressor variable(s) used to fit the model.

2. The disposition of the x values plays an important role in the least-squares fit.



Two influential observations

3. **Outliers** are observations that differ considerably from the rest of the data. They can seriously disturb the least-squares fit.



A point remote in x space

4. strong relationship between two variables, this does not imply that the variables are related in any causal sense. Causality implies necessary correlation.

Year	Mental_defectives	Radio_licenses	First_name_President	Number_letters
1924	8	1350	Calvin	6
1925	8	1960	Calvin	6
1926	9	2270	Calvin	6
1927	10	2483	Calvin	6
1928	11	2730	Calvin	6
1929	11	3091	Calvin	6
1930	12	3647	Herbert	7
1931	16	4620	Herbert	7
1932	18	5497	Herbert	7
1933	19	6260	Herbert	7
1934	20	7012	Franklin	8
1935	21	7618	Franklin	8
1936	22	8131	Franklin	8
1937	23	8593	Franklin	8

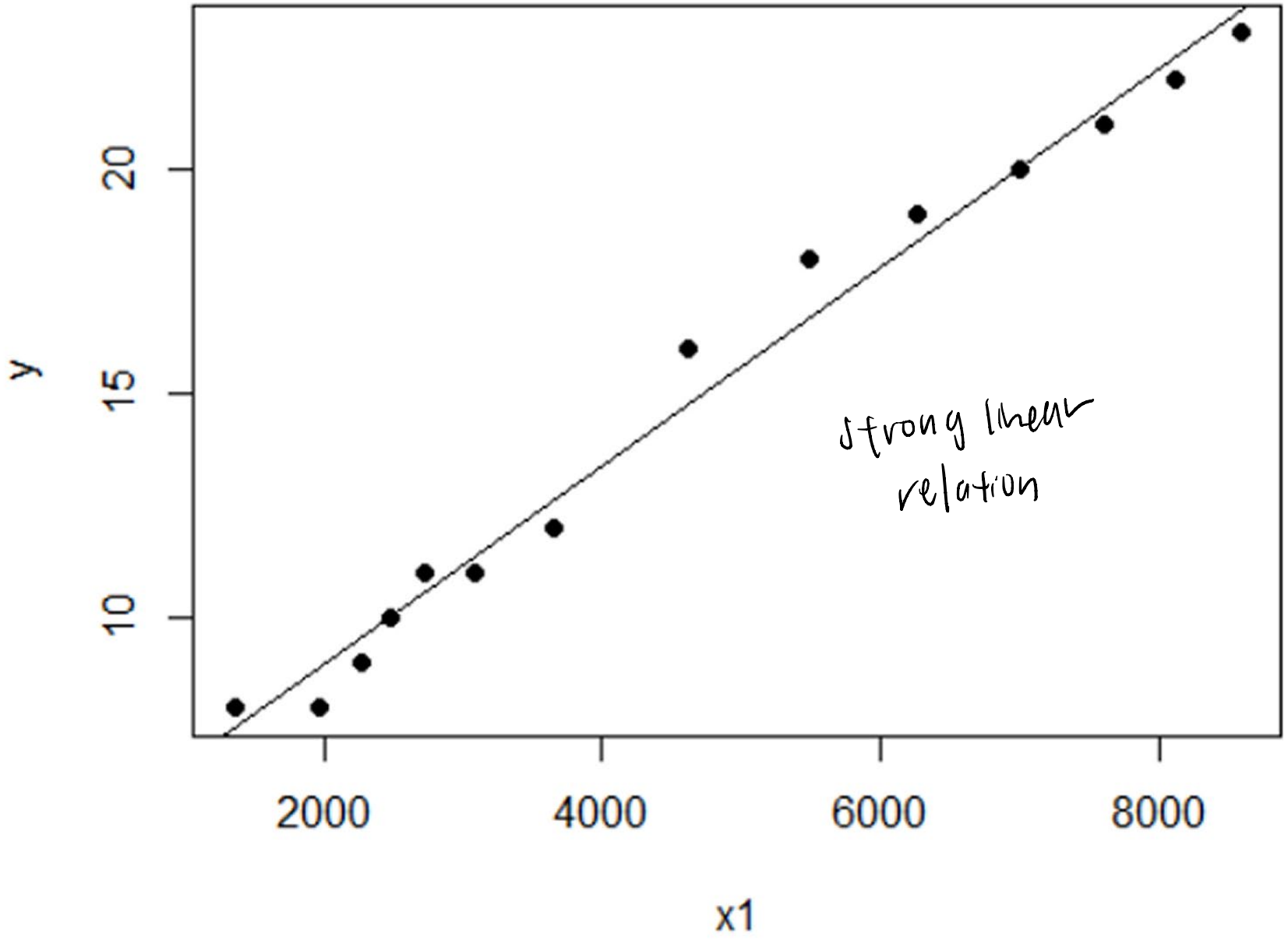
huge R^2 (close to 1)

Year	Number of Certified Mental Defectives per 10,000 of Estimated Population in the U.K (y)	Number of Radio Receiver Licenses Issued (Millions) in the U.K (x_1)	First Name of President of the U.S. (x_2)

```r
# mental_defectives.R
#
library(MASS)
reg.data <- read.table("D:\\desk_top\\teaching\\st3131\\data\\c
reg.data
names(reg.data)
y <- reg.data$Mental_defectives
x1 <- reg.data$Radio_licenses
x2 <- reg.data$Number_letters
plot(x1,y,pch=16)
abline(lm(y~x1))
summary(lm(y~x1))

plot(x2,y,pch=16)
```

13:1 (Top Level)

```
> summary(lm(y~x1))

Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9024 -0.5181 -0.2144  0.4317  1.3014

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.5822231  0.4233140   10.82 1.51e-07 ***
x1          0.0022042  0.0000807   27.31 3.58e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7262 on 12 degrees of freedom
Multiple R-squared:  0.9842,    Adjusted R-squared:  0.9828
F-statistic:   746 on 1 and 12 DF,  p-value: 3.577e-12
```


y
10
15
20
2000
4000
6000
8000
x1
strong linear relation

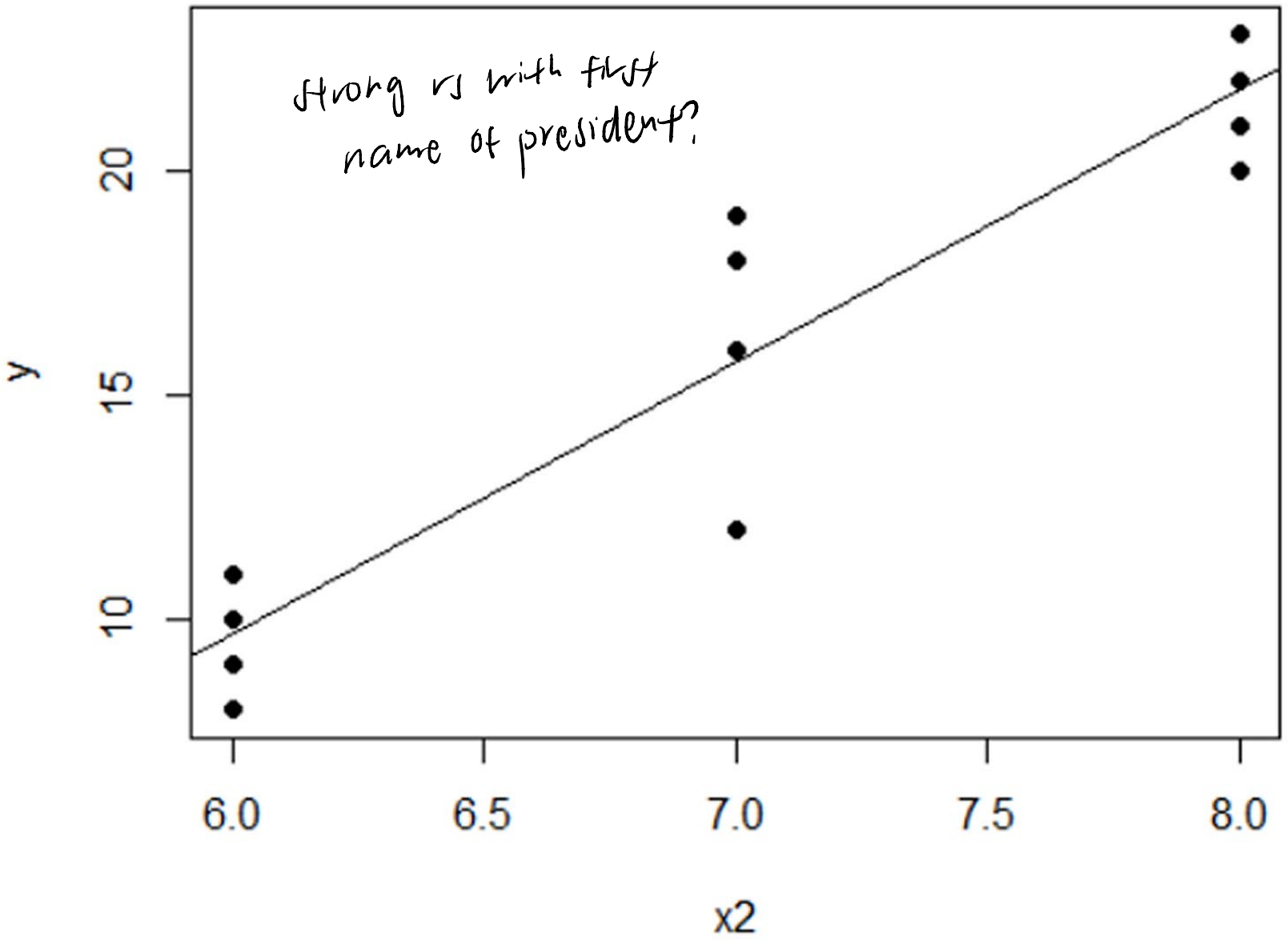
```
14  plot(x2,y,pch=16)
15  abline(lm(y~x2))
16  summary(lm(y~x2))
17
```

13:1 (Top Level)

Console Terminal

~/

```
Residual standard error: 0.7262 on 12 degrees of freedom
Multiple R-squared:  0.9842,    Adjusted R-squared:  0.9828
F-statistic:   746 on 1 and 12 DF,  p-value: 3.577e-12

> plot(x2,y,pch=16)
> abline(lm(y~x2))
> summary(lm(y~x2))

Call:
lm(formula = y ~ x2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7206 -1.4485  0.2574  1.3015  3.2794

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -26.5882     4.2840  -6.206 4.54e-05 ***
x2            6.0441     0.6202   9.746 4.73e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.933 on 12 degrees of freedom
Multiple R-squared:  0.8878,    Adjusted R-squared:  0.8785
F-statistic: 94.98 on 1 and 12 DF,  p-value: 4.728e-07
```


strong r with first name of president?
y
10
15
20
6.0
6.5
7.0
7.5
8.0
x2

REGRESSION THROUGH THE ORIGIN

No-intercept model $y = \beta_1 x + \varepsilon$ $(y_i, x_i), i = 1, 2, \ldots, n$

Least-squares function (one unknown) $S(\beta_1) = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2$

Normal equation $\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$

Least-squares estimator of the slope $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$

$\hat{\beta}_1$ is unbiased for β_1

The estimator of σ^2

chi square random variable

$$\hat{\sigma}^2 = MS_{\text{Res}} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - \hat{\beta}_1 \sum_{i=1}^{n} y_i x_i}{n-1}$$

100(1 – α) percent CI on β_1 is

$$\hat{\beta}_1 - t_{\alpha/2,n-1}\sqrt{\frac{MS_{\text{Res}}}{\sum_{i=1}^{n} x_i^2}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-1}\sqrt{\frac{MS_{\text{Res}}}{\sum_{i=1}^{n} x_i^2}}$$

100(1 – α) percent CI on $E(y|x_0)$, the mean response at $x = x_0$

$$\hat{\mu}_{y|x_0} - t_{\alpha/2,n-1}\sqrt{\frac{x_0^2 MS_{\text{Res}}}{\sum_{i=1}^{n} x_i^2}} \le E(y|x_0) \le \hat{\mu}_{y|x_0} + t_{\alpha/2,n-1}\sqrt{\frac{x_0^2 MS_{\text{Res}}}{\sum_{i=1}^{n} x_i^2}}$$

100(1 – α) percent prediction interval on a future observation at $x = x_0$

$$\hat{y}_0 - t_{\alpha/2,n-1}\sqrt{MS_{\text{Res}}\left(1+\frac{x_0^2}{\sum_{i=1}^{n} x_i^2}\right)} \le y_0 \le \hat{y}_0 + t_{\alpha/2,n-1}\sqrt{MS_{\text{Res}}\left(1+\frac{x_0^2}{\sum_{i=1}^{n} x_i^2}\right)}$$

Example Shelf-Stocking Data

	Time	Cases
1	10.15	25
2	2.96	6
3	3.00	8
4	6.88	17
5	0.28	2
6	5.06	13
7	9.14	23
8	11.86	30
9	11.69	28
10	6.04	14
11	7.57	19
12	1.74	4
13	9.38	24
14	0.16	1
15	1.84	5

This data set contains

time = time required for a merchandiser to stock a grocery store shelf with a soft drink product

cases = number of cases of products stocked

Shelf-Stocking Data

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} = \frac{1841.98}{4575.00} = 0.4026$$

Fitted equation $\hat{y} = 0.4026x$

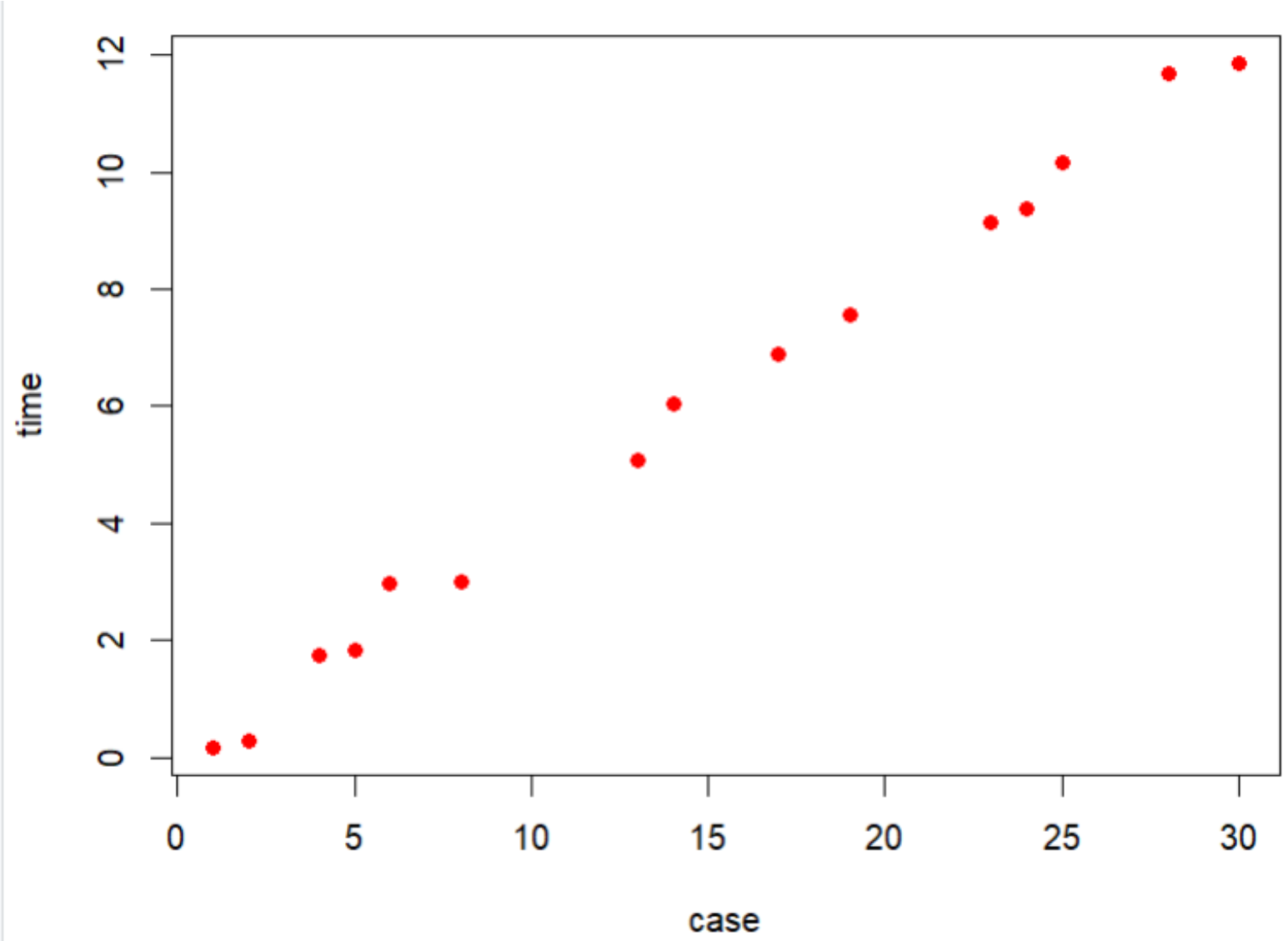
$MS_{\text{Res}} = 0.0893$

$R_0^2 = 0.9883$

H_0: $\beta_1 = 0$ $t_0 = 91.13$

```r
# ch2_shelf_stocking.R
#
library(MASS)
rm(list = ls())
shelf.data <- read.table("D:\\nus_teaching\\st3131\\data\\Shelf_Stocking.csv",
                         header = T, sep=",")
shelf.data

#choose simpler names for the two variables
names(shelf.data) <- c("time", "case")
shelf.data

#attach() function is used to access variables present in the dataframe
attach(shelf.data)
plot(case,time,pch=16,col="red")
fitted.model <- lm(time ~ 0 + case )
summary(fitted.model)
anova(fitted.model)
```

when there is no intercept term


time
0
2
4
6
8
10
12
0
5
10
15
20
25
30
case

```
> plot(case,time,pch=16,col="red")
> fitted.model <- lm(time ~ 0 + case )
> summary(fitted.model)

Call:
lm(formula = time ~ 0 + case)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5252 -0.2198 -0.1202  0.1070  0.5443

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
case 0.402619   0.004418   91.13   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2988 on 14 degrees of freedom
Multiple R-squared:  0.9983,    Adjusted R-squared:  0.9982
F-statistic:  8305 on 1 and 14 DF,  p-value: < 2.2e-16

> anova(fitted.model)
Analysis of Variance Table

Response: time
          Df Sum Sq Mean Sq F value    Pr(>F)
case       1 741.62  741.62  8305.2 < 2.2e-16 ***
Residuals 14   1.25    0.09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

pvalue very small (reject H_0)

estimate of variance of epsilon

mean square residual

Maximum likelihood estimation

1. The simple linear regression model for a response variable y and a regressor variable x based on observations (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) can be written as

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

where ϵ_i is a random variable such that $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$, $i = 1, 2, ..., n$ and ϵ_i's are independent.

Assume further that ϵ_i's are normally distributed. The responses y_i's are independently and normally distributed with mean $E(y_i) = \beta_0 + \beta_1 x_i$ and variance $Var(y_i) = \sigma^2, i = 1, 2, ..., n$.

The probability density of y_i is — normally distributed function

$$f(y_i) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2 \right]$$

2. Likelihood function

$$L(y_i, x_i, \beta_0, \beta_i, \sigma^2) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

joint density function

$$= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

3. Maximum-likelihood estimators: $\tilde{\beta}_0$, $\tilde{\beta}_1$, and $\tilde{\sigma}^2$

Maximize $\ln L$

$$\ln L(y_i, x_i, \beta_0, \beta_i, \sigma^2) = -\left(\frac{n}{2}\right)\ln 2\pi - \left(\frac{n}{2}\right)\ln\sigma^2 - \left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

L is a function of 3 unknowns

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

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

$$\left.\frac{\partial \ln L}{\partial \sigma^2}\right|_{\tilde{\beta}_0,\tilde{\beta}_1,\tilde{\sigma}^2} = -\frac{n}{2\tilde{\sigma}^2} + \frac{1}{2\tilde{\sigma}^4}\sum_{i=1}^{n}\left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\right)^2 = 0$$

Maximum-likelihood estimators:

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

$$\tilde{\beta} = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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

LSE: $\hat{\sigma}^2 = \frac{SS_{RES}}{n-2}$

$$\tilde{\sigma}^2 = [(n-1)/n]\hat{\sigma}^2$$

different

The maximum likelihood estimates are the same as the least-squares estimates except for σ^2.

CASE WHERE THE REGRESSOR x IS RANDOM

what if x is rv?

x and y Jointly Distributed

Suppose that x and y are jointly distributed random variables but the form of this joint distribution is unknown. It can be shown that all of our previous regression results hold if the following conditions are satisfied:

1. The conditional distribution of y given x is normal with conditional mean $\beta_0 + \beta_1 x$ and conditional variance σ^2.
2. The x's are independent random variables whose probability distribution does not involve β_0, β_1, and σ^2.

Example: The data set consists of a random sample of n adults, the response y is the weight and the regressor variable x is the height of an adult.

CASE WHERE THE REGRESSOR *x* IS RANDOM

x and *y* Jointly Normally Distributed: Correlation Model

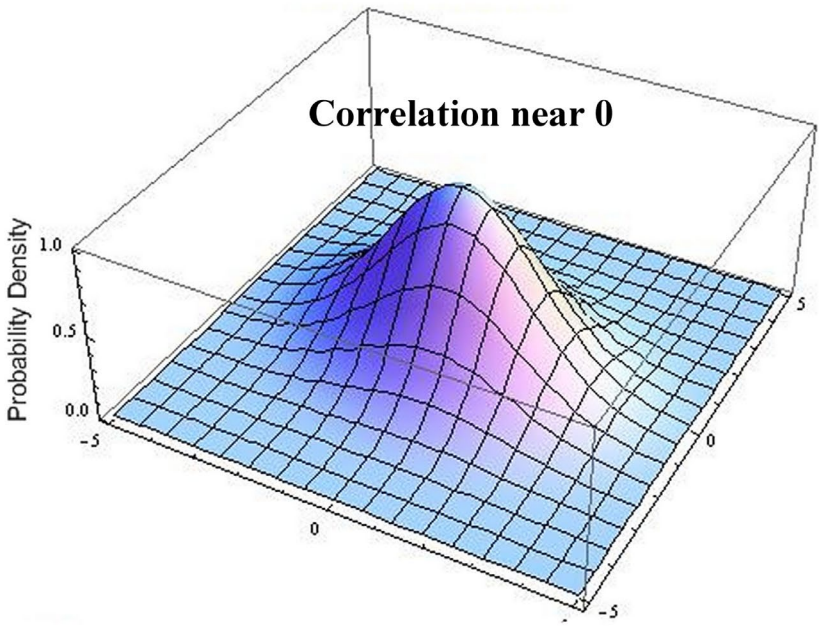
y and x are jointly distributed according to the **bivariate normal distribution**.

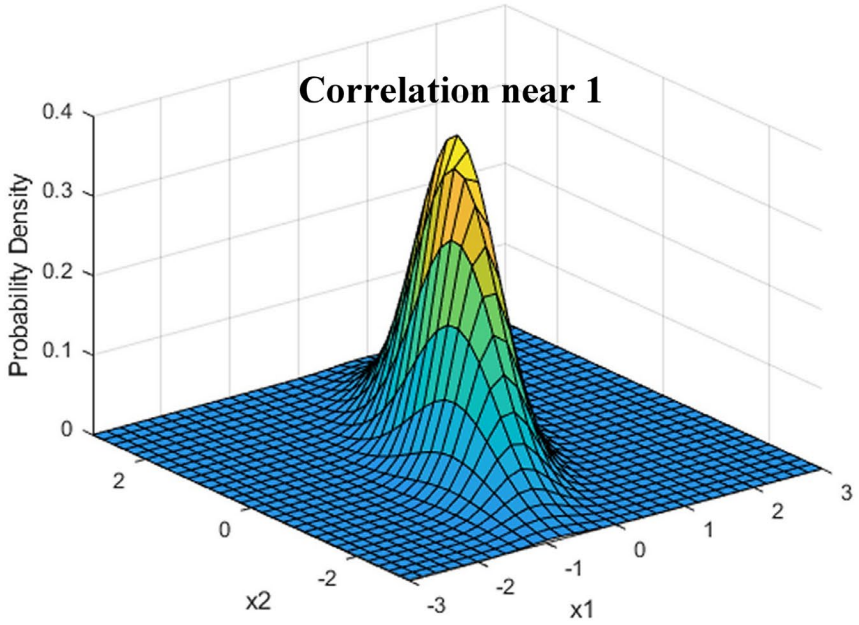
$$f(y,x)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$
$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y-\mu_1}{\sigma_1}\right)^2+\left(\frac{x-\mu_2}{\sigma_2}\right)^2-2\rho\left(\frac{y-\mu_1}{\sigma_1}\right)\left(\frac{x-\mu_2}{\sigma_2}\right)\right]\right\}$$

correlation coefficient between y and x

$$\rho=\frac{E(y-\mu_1)(x-\mu_2)}{\sigma_1\sigma_2}=\frac{\sigma_{12}}{\sigma_1\sigma_2}$$

σ_{12} is the **covariance** of y and x.


Correlation near 0
Probability Density
1.0
0.5
0.0
-5
0
5
0
-5


Correlation near 1
Probability Density
0.4
0.3
0.2
0.1
0
2
0
-2
x2
-3
-2
-1
0
1
2
3
x1

Maximum-likelihood estimators

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

The **conditional distribution** of y for a given value of x is

$$f(y|x) = \frac{1}{\sqrt{2\pi}\sigma_{1.2}} \exp\left[-\frac{1}{2}\left(\frac{y-\beta_0-\beta_1 x}{\sigma_{1.2}}\right)^2\right]$$

$$\beta_0 = \mu_1 - \mu_2 \rho \frac{\sigma_1}{\sigma_2}$$

$$\beta_1 = \frac{\sigma_1}{\sigma_2}\rho$$

$$\sigma_{1.2}^2 = \sigma_1^2(1-\rho^2)$$

The conditional distribution of y given x is normal with mean $E(y|x) = \beta_0 + \beta_1 x$ and variance $\sigma_{1.2}^2$.

Sample correlation coefficient

1. $$r = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2 \right]^{1/2}} = \frac{S_{xy}}{[S_{xx} SS_T]^{1/2}}$$

2. $$\hat{\beta}_1 = \left(\frac{SS_T}{S_{xx}} \right)^{1/2} r$$

between -1 and 1

perfectly negative r

perfectly positive r

3. $$r^2 = \hat{\beta}_1^2 \frac{S_{xx}}{SS_T} = \frac{\hat{\beta}_1 S_{xy}}{SS_T} = \frac{SS_R}{SS_T} = R^2$$

direct connection

Hypothesis testing of population correlation coefficient

Case 1: $H_0: \rho = 0, \quad H_1: \rho \neq 0$

The test statistic $t_0 = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ follows the t distribution with

$n - 2$ degrees of freedom if $H_0: \rho = 0$ is true.

Reject the null hypothesis if $|t_0| > t_{\alpha/2,\, n-2}$

Case 2: $H_0: \rho = \rho_0, \quad H_1: \rho \neq \rho_0$

For moderately large samples $n \geq 25$, $Z = \operatorname{arctanh} r = \frac{1}{2} \ln \frac{1+r}{1-r}$

is approximately normally distributed with mean

$\mu_Z = \operatorname{arctanh} \rho = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$ and variance $\sigma_Z^2 = (n-3)^{-1}$.

Test statistic $Z_0 = (\operatorname{arctanh} r - \operatorname{arctanh} \rho_0)(n-3)^{1/2}$

Reject $H_0: \rho = \rho_0$ if $|Z_0| > Z_{\alpha/2}$

100(1 – α) percent **confidence interval for** ρ

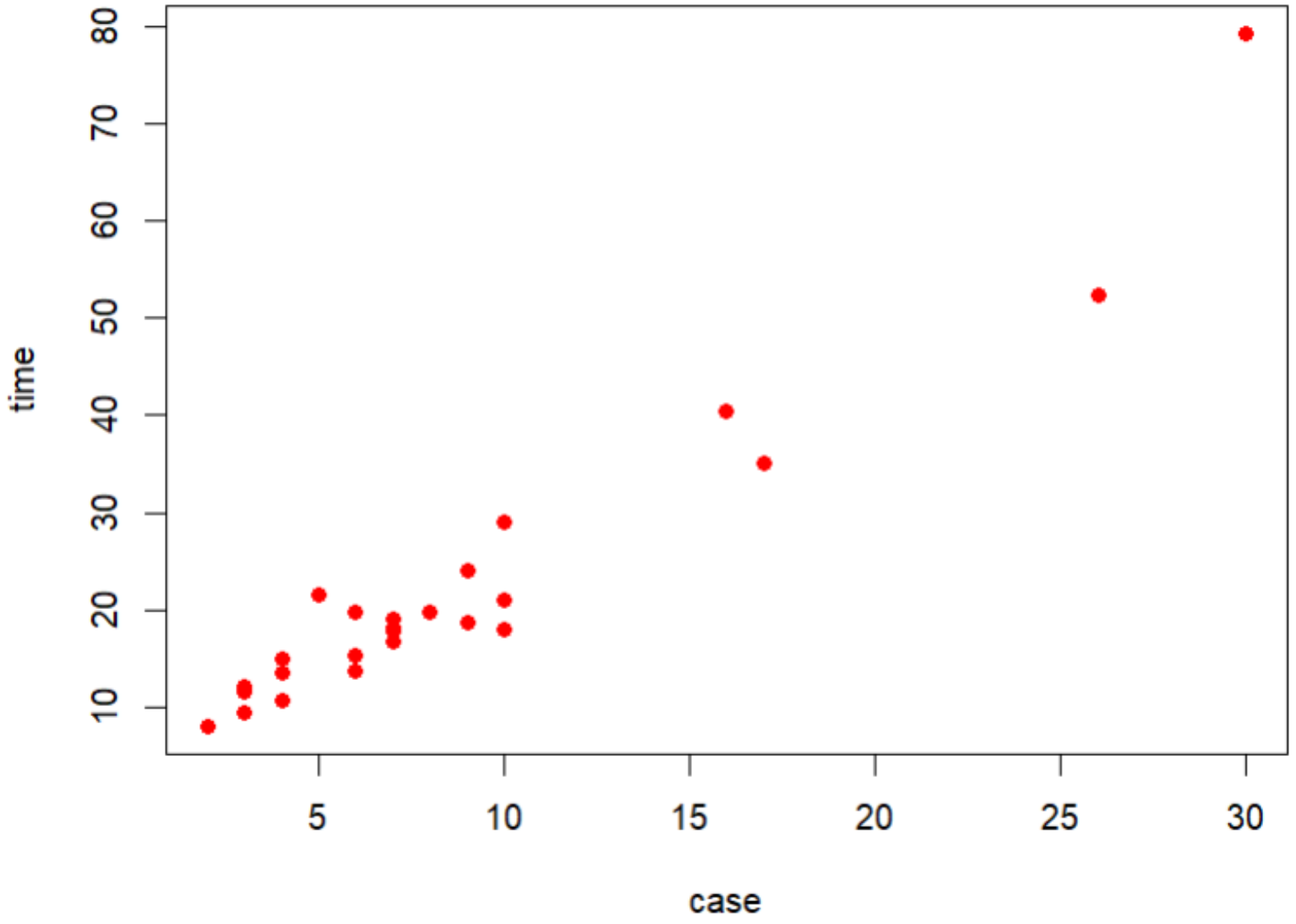
$$\tanh\left(\operatorname{arctanh} r - \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right) \leq \rho \leq \tanh\left(\operatorname{arctanh} r + \frac{Z_{\alpha/2}}{\sqrt{n-3}}\right)$$

where $\tanh u = (e^u - e^{-u})/(e^u + e^{-u})$.

The delivery time data

The response y is the amount of time in minutes required by a route driver to stock the vending machines with beverage products and do minor maintenance of the machines in an outlet. The two regressor variables are (i) number of cases of products stocked (x_1), and (ii) the distance walked in feet (x_2).

	time	case	dist
1	16.68	7	560
2	11.50	3	220
3	12.03	3	340
4	14.88	4	80
5	13.75	6	150
6	18.11	7	330
7	8.00	2	110
8	17.83	7	210
9	79.24	30	1460
10	21.50	5	605
11	40.33	16	688
12	21.00	10	215
13	13.50	4	255
14	19.75	6	462
15	24.00	9	448
16	29.00	10	776
17	15.35	6	200
18	19.00	7	132
19	9.50	3	36
20	35.10	17	770
21	17.90	10	140
22	52.32	26	810
23	18.75	9	450
24	19.83	8	635
25	10.75	4	150


time
80
70
60
50
40
30
20
10
5
10
15
20
25
30
case

Example Delivery Time Data

$H_0: \rho = 0, \quad H_1: \rho \neq 0$

$$r = \frac{S_{xy}}{[S_{xx} SS_T]^{1/2}} = \frac{2473.3440}{[(1136.5600)(5784.5426)]^{1/2}} = 0.9646$$

```r

# ch2_delivery_time.R
install.packages("scatterplot3d")
install.packages("psychometric")
library(scatterplot3d)
library(psychometric)
library(MASS)

rm(list = ls())
dat <- read.csv("D:\\nus_teaching\\st3131\\data\\Delivery_Time.csv",
                header = T, sep=",")
dat
obs <- dat[,1]
time <- dat[,2]
case <- dat[,3]
dist <- dat[,4]

plot(case, time, pch=16, col="red")
corr <- cor(case,time)
corr
t0 <- corr*sqrt(length(case)-2)/sqrt(1-corr^2)
t0
pvalue <- 2*pt(t0,df=23, lower.tail=FALSE)
pvalue
CIr(corr,n=length(case), level=0.95)

```

```
> corr <- cor(case,time)
> corr
[1] 0.9646146
> t0 <- corr*sqrt(length(case)-2)/sqrt(1-corr^2)
> t0
[1] 17.54555
> pvalue <- 2*pt(t0,df=23, lower.tail=FALSE)
> pvalue
[1] 8.220192e-15
> CIr(corr,n=length(case), level=0.95)
[1] 0.9202275 0.9845031
```

The p-value is very small,
reject the null hypothesis.

Appendix

Extractor functions for the result of `lm()`

`summary()`	returns summary information about the regression
`plot()`	makes diagnostic plots
`coef()`	returns the coefficients
`residuals()`	returns the residuals (can be abbreviated `resid()`)
`fitted()`	returns fitted values, $\widehat{y}_i$
`deviance()`	returns RSS
`predict()`	performs predictions
`anova()`	finds various sums of squares

The End