Chapter 3

Simple Linear Regression

STAT210/410 Study Plan

Topic	Weeks covered	Readings	Assessment
Topic 1: Simple Linear regression (SLR)	Wk 1	Chapter 3	Online Quiz due 9th March
Topic 2: Multiple Linear Regression (MLR)	Wk2 & 3	Chapter 4	Written Assessment A2 due 23rd March
Topic 3: Model building	Wk 4	Chapter 5	
Topic 4: Variable Screening and regression pitfalls	Wk 5	Chapters 6, 7	
Topic 5: Residual Analysis	Wk 6	Chapter 8	Written Assessment A3 due 13th April
Topic 6 Generalised Linear Models (GLMs)	Wk 9 & 10	Chapter 9	
Topic 7: Principles of Experimental Design	Wk 11	Chapter 11	Written Assessment A4 due 11th May
Topic 8: ANOVA, contrasts	Wk 12 & 13	Chapter 12	
STAT410 ONLY			
ART: Nonparametric Regression		Section 9.9	Written Assessment ART due 18th May

Chapter 3 outline

- Lecture 1:
  - Introduction
  - Linear statistical models
  - Method of least squares

- Lecture 2:
  - Model assumptions, estimator of σ2
  - Inference about the slope

- Lecture 3:
  - Coefficient of Coefficient of determination (R^2), correlation (r)
  - Using the model for estimation & prediction

Lecture 1

Simple Linear Regression

Chapter 3 Outline

Lecture 1:

- Introduction
- Linear statistical models
- Method of least squares

Lecture 2:

- Model assumptions, estimator of σ^2
- Inference about the slope

Lecture 3:

- Coefficient of Coefficient of determination (R^2), correlation (r)
- Using the model for estimation & prediction

Introduction

Simple Linear Regression (SLR):

Modelling the straight line association between two quantitative variables

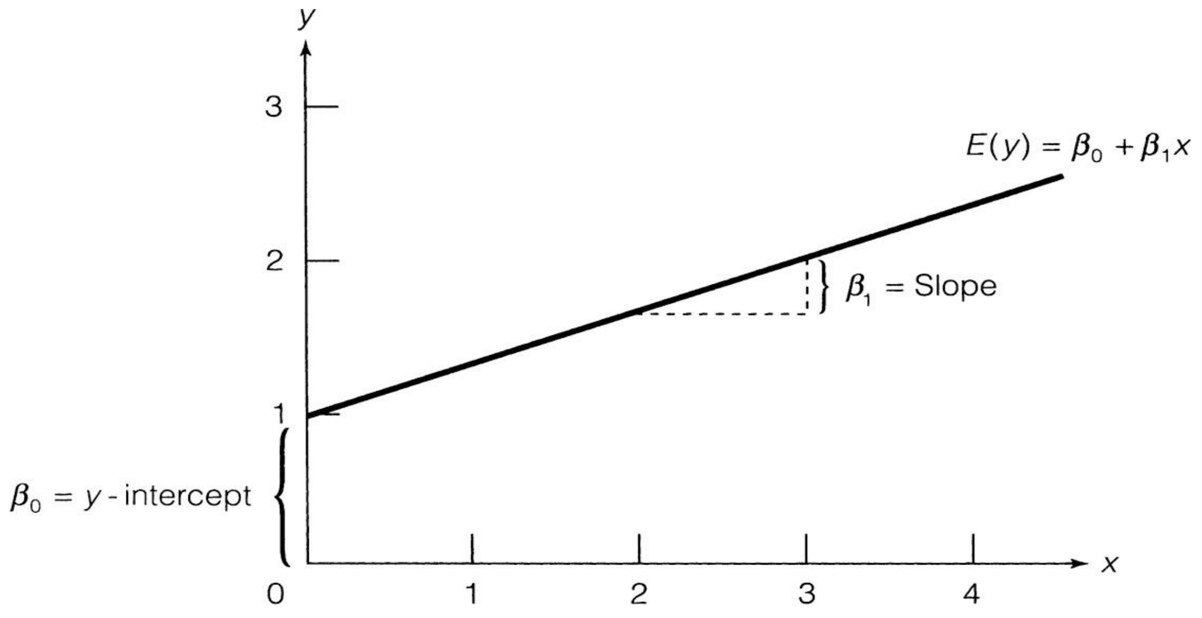


Figure 3.1

A First-Order (Straight-Line) Model

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

where

y = **Dependent** variable (variable to be modeled—sometimes called the **response** variable)

x = Independent variable (variable used as a **predictor** of y)

$E(y) = \beta_0 + \beta_1 x$ = Deterministic component

Also known as the ***systematic*** component

ε = (epsilon) = Random error component

β_0 = (beta zero) = ***y*-intercept** of the line, i.e., point at which the line intercepts or cuts through the y-axis (see Figure 3.1)

β_1 = (beta one) = **Slope** of the line, i.e., amount of increase (or decrease) in the mean of y for every 1-unit increase in x (see Figure 3.1)

Steps in Regression Analysis

Step 1. Hypothesize the form of the model for E(y)

Step 2. Collect the data (sample data)

Step 3. Use the collected data to estimate the unknown parameters in the model

Step 4. Investigate the random error term, checking model assumptions

Step 5. Statistically check the usefulness of the model

Step 6. When satisfied that the model is useful, use it for prediction, estimation and so on.

SLR $\mathbf{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x} + \boldsymbol{\epsilon}$

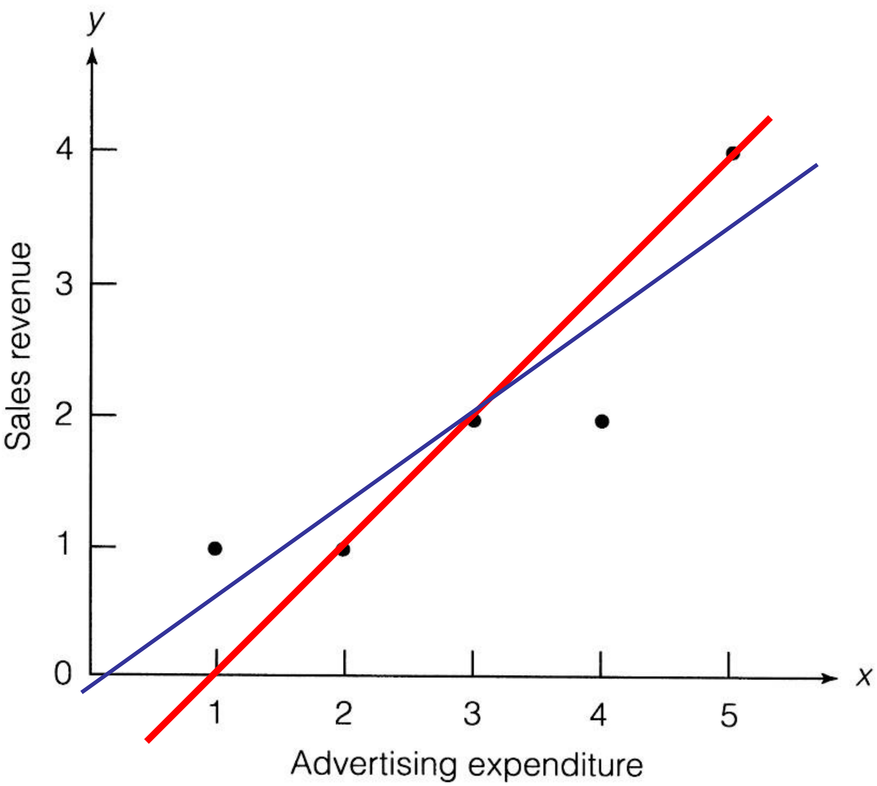
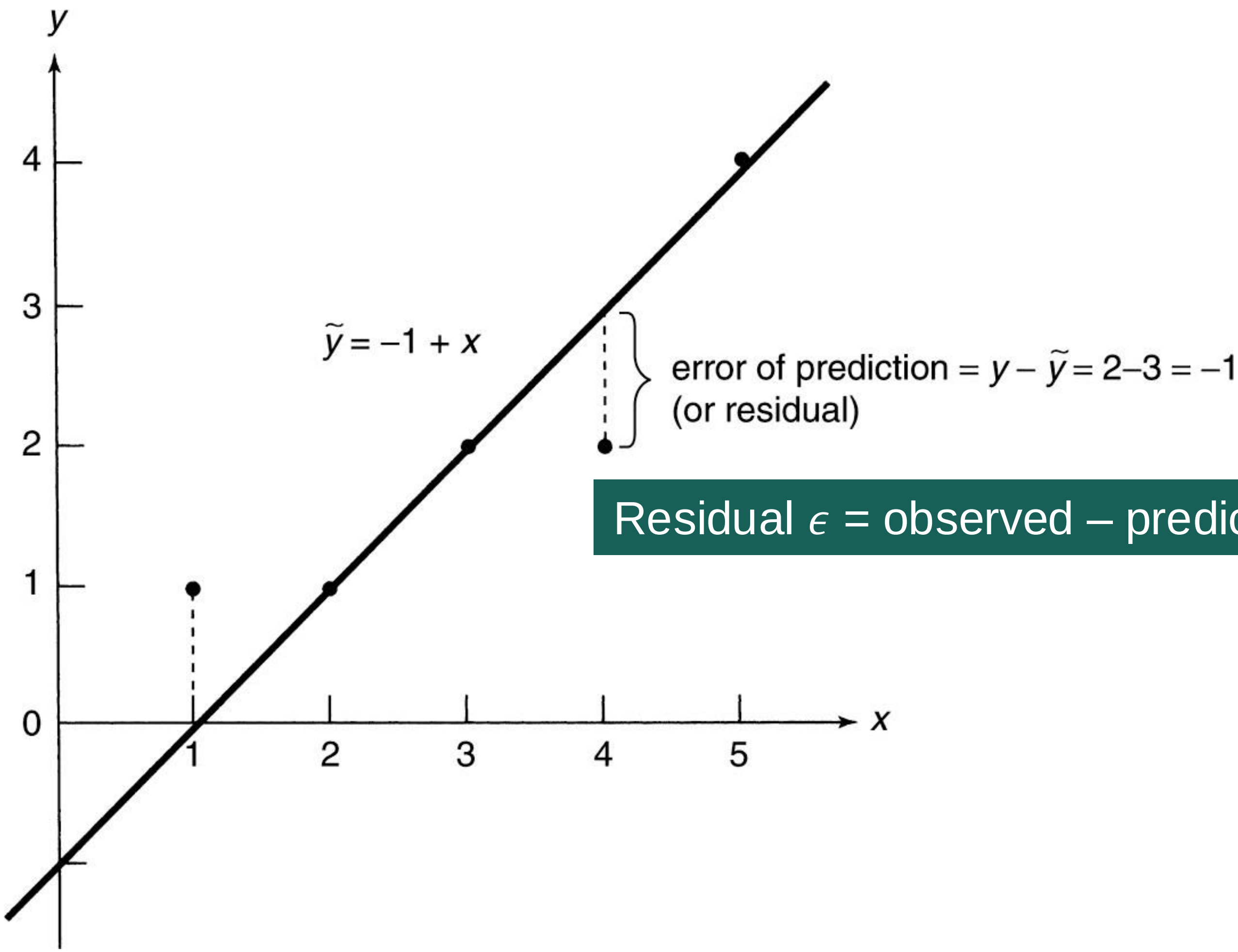


Table 3.1 Appliance store data

Month	Advertising Expenditure x, hundreds of dollars	Sales Revenue y, thousands of dollars
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

Is there an association between the advertising expenditure and the sale revenue?

Figure 3.3 Visual straight-line fit to data in Table 3.1



Residual ϵ = observed – predicted value

Table 3.2 Comparing observed and predicted values for the visual model

x	y	Prediction $\tilde{y} = -1 + x$	Error of prediction $(y - \tilde{y})$	Squared error $(y - \tilde{y})^2$
1	1	0	$(1 - 0) = 1$	1
2	1	1	$(1 - 1) = 0$	0
3	2	2	$(2 - 2) = 0$	0
4	2	3	$(2 - 3) = -1$	1
5	4	4	$(4 - 4) = 0$	0
			Sum of errors (SE) = 0	Sum of squared errors (SSE) = 2

Figure 3.3 Visual straight-line fit to data in Table 3.1

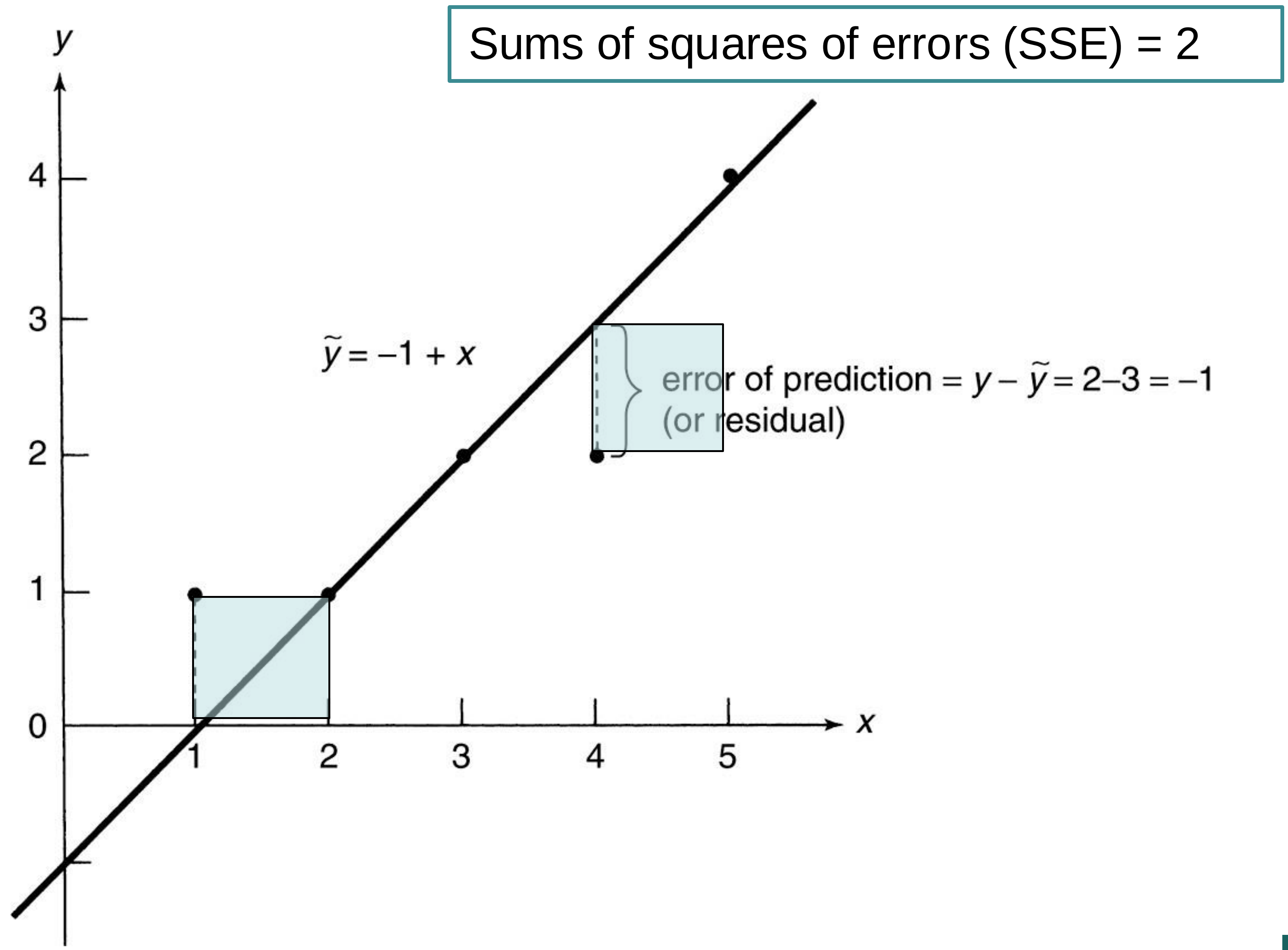


Figure 3.4

Plot of the least squares line

$\hat{y} = -0.1 + 0.7x$

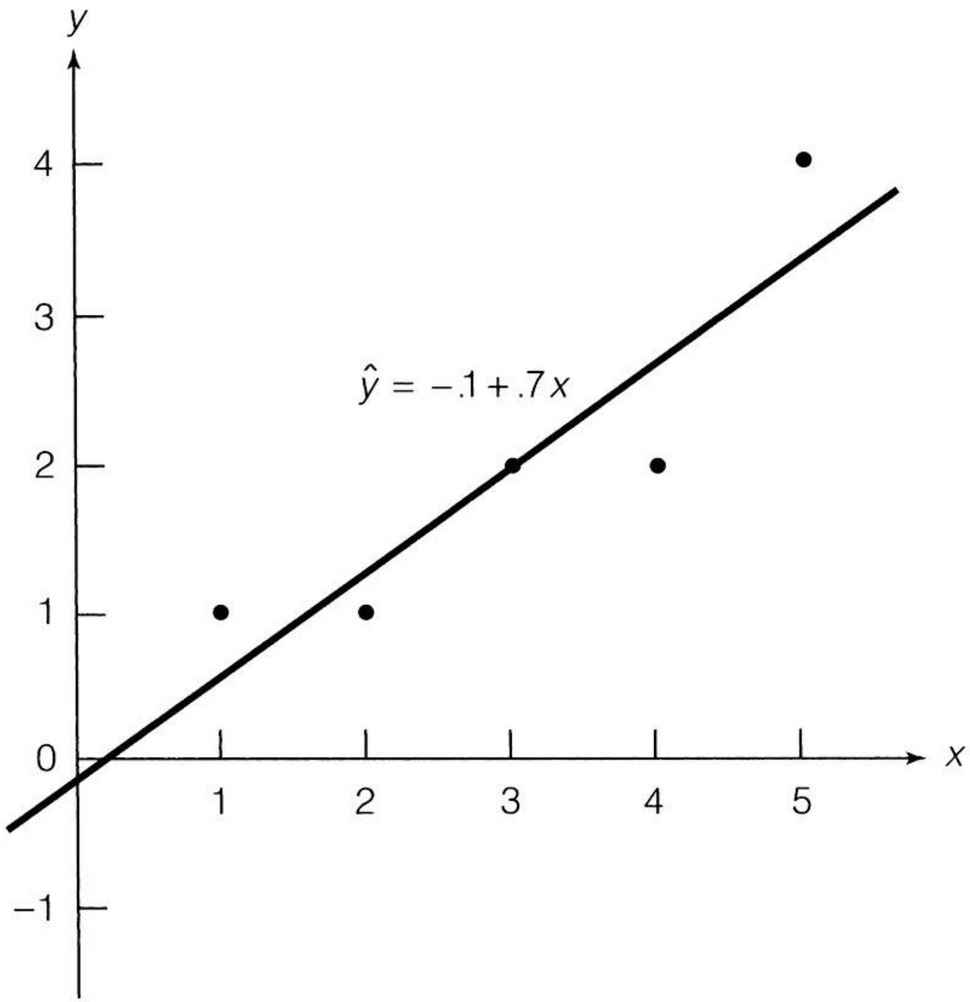
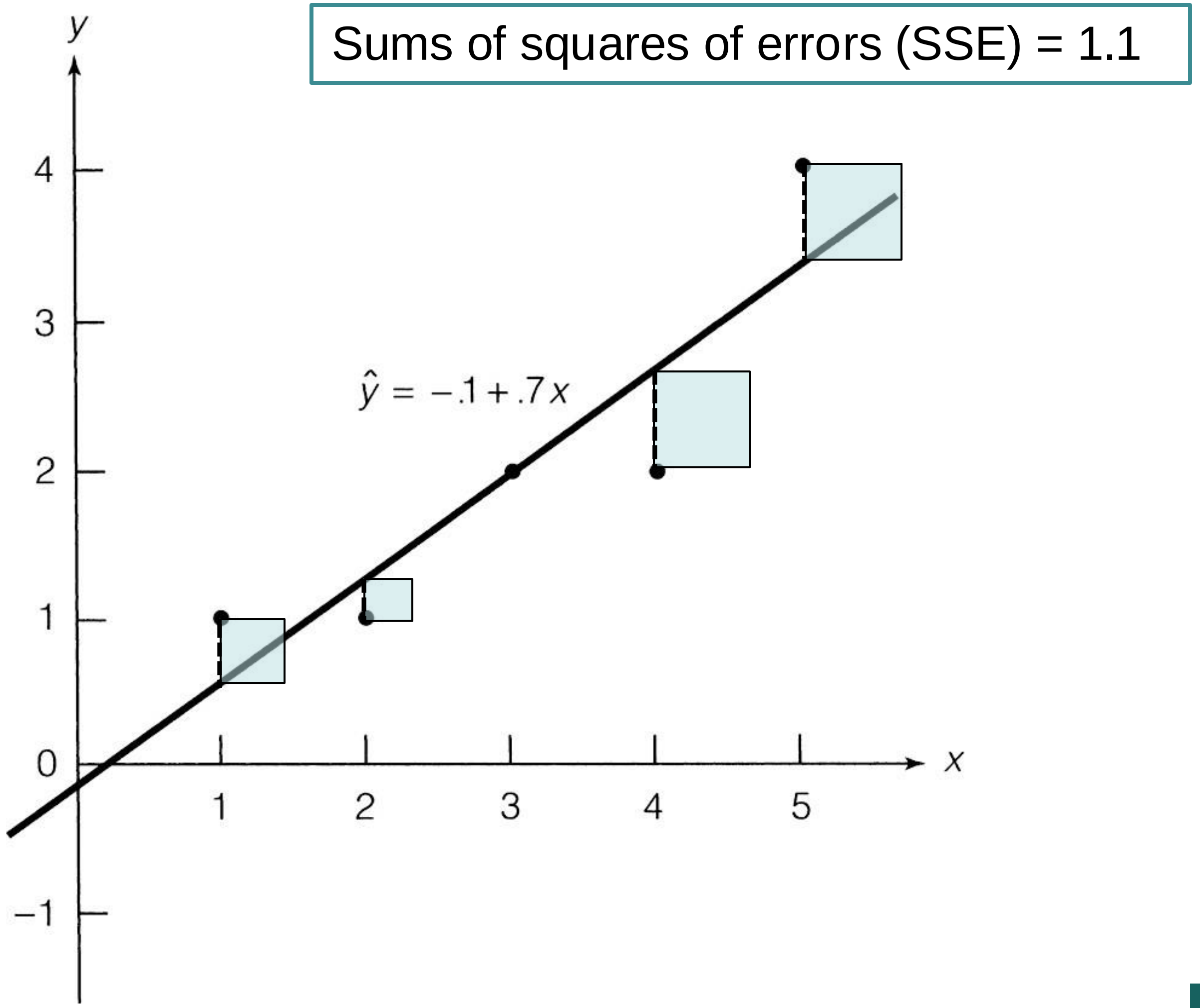


Table 3.4 Comparing observed and predicted values for the least squares model

x	y	Predicted $\hat{y} = -.1 + .7x$	Residual (error) $(y - \hat{y})$	Squared error $(y - \hat{y})^2$
1	1	.6	$(1 - .6) = .4$	.16
2	1	1.3	$(1 - 1.3) = -.3$	.09
3	2	2.0	$(2 - 2.0) = 0$	.00
4	2	2.7	$(2 - 2.7) = -.7$	.49
5	4	3.4	$(4 - 3.4) = .6$	.36
			Sum of errors (SE) $= 0$	SSE $= 1.10$

Figure 3.4 Plot of the least squares line $\hat{y} = -.1 + .7x$



Least squares line

The **least squares line:** $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x$

is the one that satisfies two properties:

1. $SE = \sum(y_i - \widehat{y_i}) = 0$ i.e the sum of the residuals is 0

2. $SSE = \sum(y_i - \widehat{y_i})^2$ is the smallest value.

Notation

The straight-line model: $\mathbf{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x} + \boldsymbol{\epsilon}$

The fitted line using the Least Squares method: $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$

y response (dependent) variable

x explanatory (independent) variable

$\hat{y}$ predictor of some future value of y

$\widehat{\beta_0}$ Point estimate of β_0

$\widehat{\beta_1}$ Point estimate e of β_1

$\widehat{\beta_0}$ and $\widehat{\beta_1}$ are calculated based on collected data $\{x_i , y_i\}$

Steps in Regression Analysis

Step 1. Hypothesize the form of the model for E(y)

Step 2. Collect the data (sample data)

Step 3. Use the collected data to estimate the unknown parameters in the model

Step 4. Investigate the random error term, checking model assumptions

Step 5. Statistically check the usefulness of the model

Step 6. When satisfied that the model is useful, use it for prediction, estimation and so on.

SLR Example

Laetasaric acid is a compound that holds promise for the control of fungus diseases in crops. The data show the results of growing the fungus in various concentrations of the acid.

Question: Is there a relationship between the level of acid and the fungus growth?

Source: Statistics for the Life Sciences, (2nd edn), M. L. Samuels & J. A. Witmer, (1999), p. 512

	acid	fungus
1	0	33.3
2	0	31.0
3	3	29.8
4	3	27.8
5	6	28.0
6	6	29.0
7	10	25.5
8	10	23.8
9	20	12.5
10	20	15.5
11	30	11.7
12	30	10.0

R Commands

\# denotes a comment: non-executable

```
# Read data from a file into a data frame
acid.df <- read.table(“acid.txt”,header=T)
```

Name of object where result is stored, in this case, the data frame

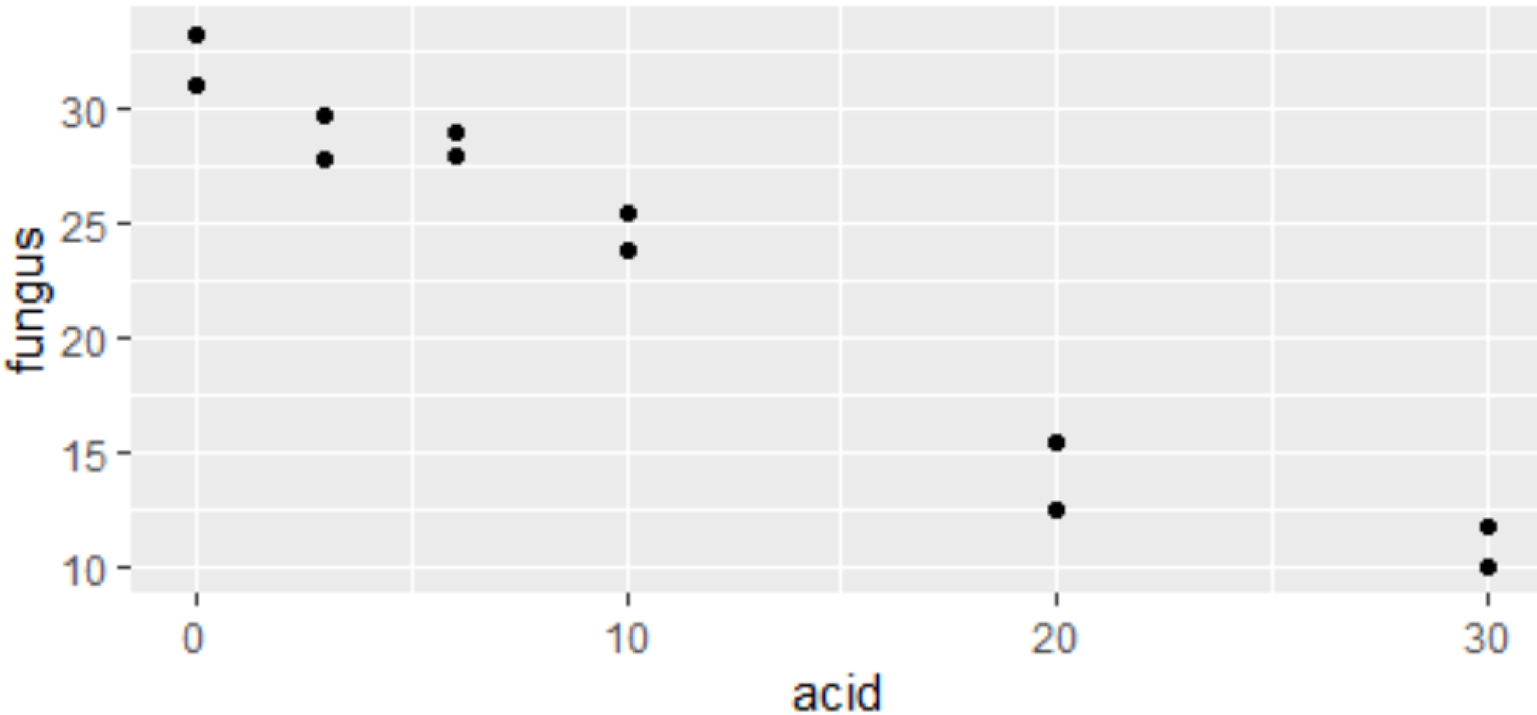
<- assigns result from right to object on left

Name of data file in quotation marks

Reads first row of data file as variable (column) names

```
# Simple scatterplot
library(ggplot2)
ggplot(acid.df, aes(x=acid, y=fungus))+
  geom_point()
```

SLR Example



There appears to be a *negative linear association* between acid concentration and fungus growth.

R Commands (cont)

```
# Fit data and save linear model object
reg.lm <- lm(fungus~acid,data=acid.df)
```

Object which stores results of analysis

General form of model: lm(y~x), y is the response and x is the predictor

Indicate where the data are stored

```
# Display summary of fit:
# regression coefficients
summary(reg.lm)

# Print ANOVA table
anova(reg.lm)
```

R outputs

`summary(reg.lm)`

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

SLR equation: $\widehat{Fungal\ growth}$= 31.78 - 0.75 *acid

Intercept: $\widehat{\beta_0}$

```
Coefficients:
         Estimate    SE       t    Pr(>|t|)
Inter       31.78  0.83   38.17    3.63e-12
acid        -0.75  0.05  -13.98    6.89e-08
```

Slope: $\widehat{\beta_1}$

```
Residual standard error: 1.937 on 10df
Multiple R-Squared: 0.9513,
Adjusted R-squared: 0.9464
F-statistic: 195.3 on 1 and 10 DF,
 p-value: 6.888e-08
```

Lecture 2

Simple Linear Regression

Chapter 3 Outline

Lecture 1:

- Introduction
- Linear statistical models
- Method of least squares

Lecture 2:

- Model assumptions, estimator of σ^2
- Inference about the slope

Lecture 3:

- Coefficient of Coefficient of determination (R^2), correlation (r)
- Using the model for estimation & prediction

Notation

The straight-line model: $\mathbf{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x} + \boldsymbol{\epsilon}$

The fitted line using the Least Squares method: $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}x$

y response (dependent) variable

x explanatory (independent) variable

$\hat{y}$ predictor of some future value of y

$\widehat{\beta_0}$ Point estimate of β_0 ← The intercept

$\widehat{\beta_1}$ Point estimate e of β_1 ← The slope

$\widehat{\beta_0}$ and $\widehat{\beta_1}$ are calculated based on collected data $\{x_i , y_i\}$

Steps in Regression Analysis

Step 1. Hypothesize the form of the model for E(y)

Step 2. Collect the data (sample data)

Step 3. Use the collected data to estimate the unknown parameters in the model

Step 4. Investigate the random error term, checking model assumptions

Step 5. Statistically check the usefulness of the model

Step 6. When satisfied that the model is useful, use it for prediction, estimation and so on.

SLR Model Assumptions

SLR Model: $y = \beta_0 + \beta_1 x + \varepsilon$

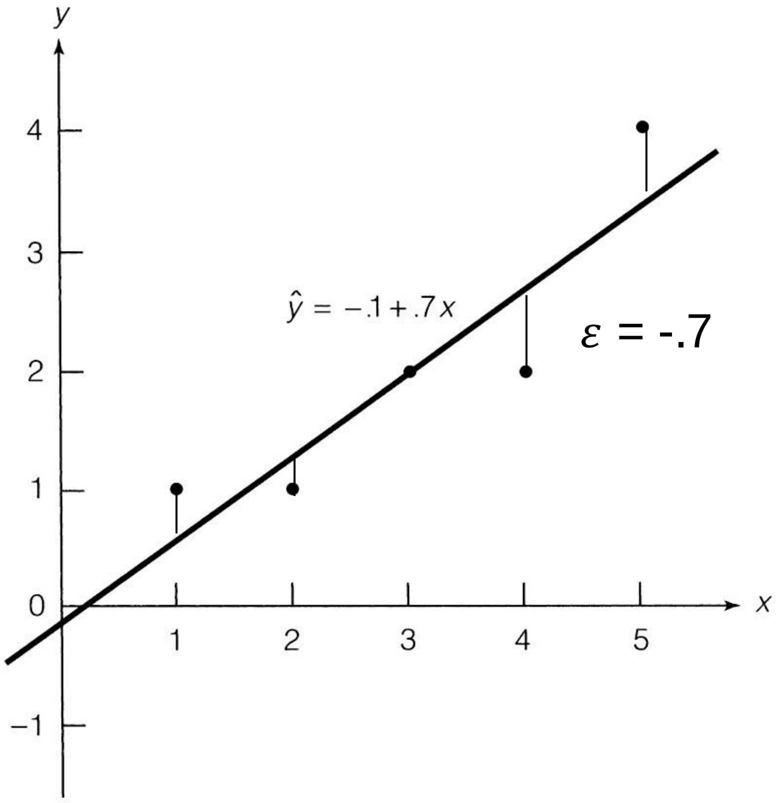
The residuals/errors: $\varepsilon = y - \hat{y}$

e.g. x = 4, y = 2

$$\hat{y} = -0.1 + 0.7 * 4 = 2.7$$

residual:

$e = 2 - 2.7 = -0.7$



The probability distribution of ϵ determines the reliability of the least squares estimators and the utility of the SLR model.

SLR Model Assumptions

SLR Model: $y = \beta_0 + \beta_1 x + \varepsilon$

Assumptions:

1. The residuals are independent
2. The residuals are normally distributed, $\epsilon \sim N(0, \sigma^2)$
   - Mean 0
   - Variance, σ^2, which is constant with regard to X

SLR Model Assumptions

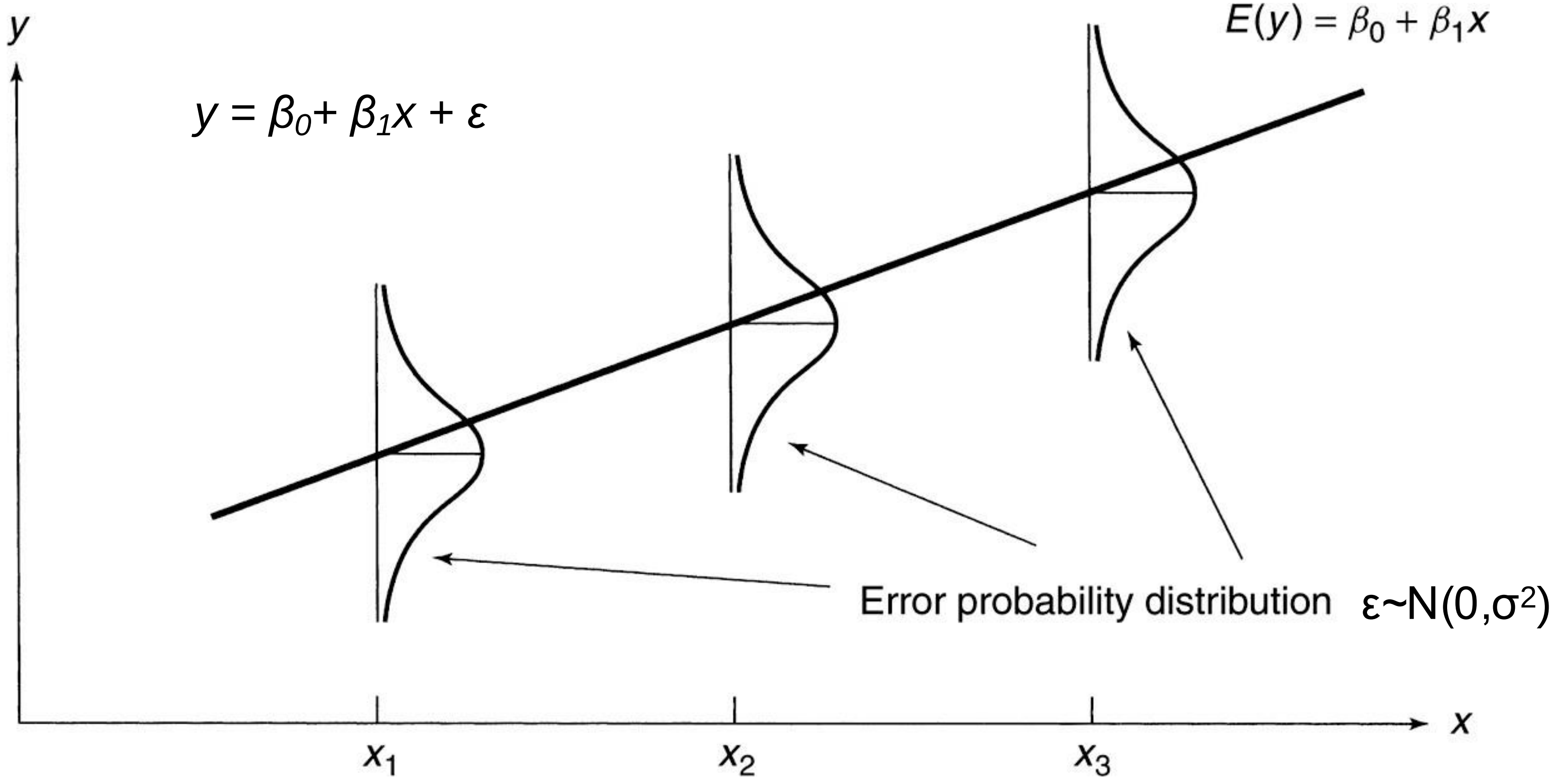
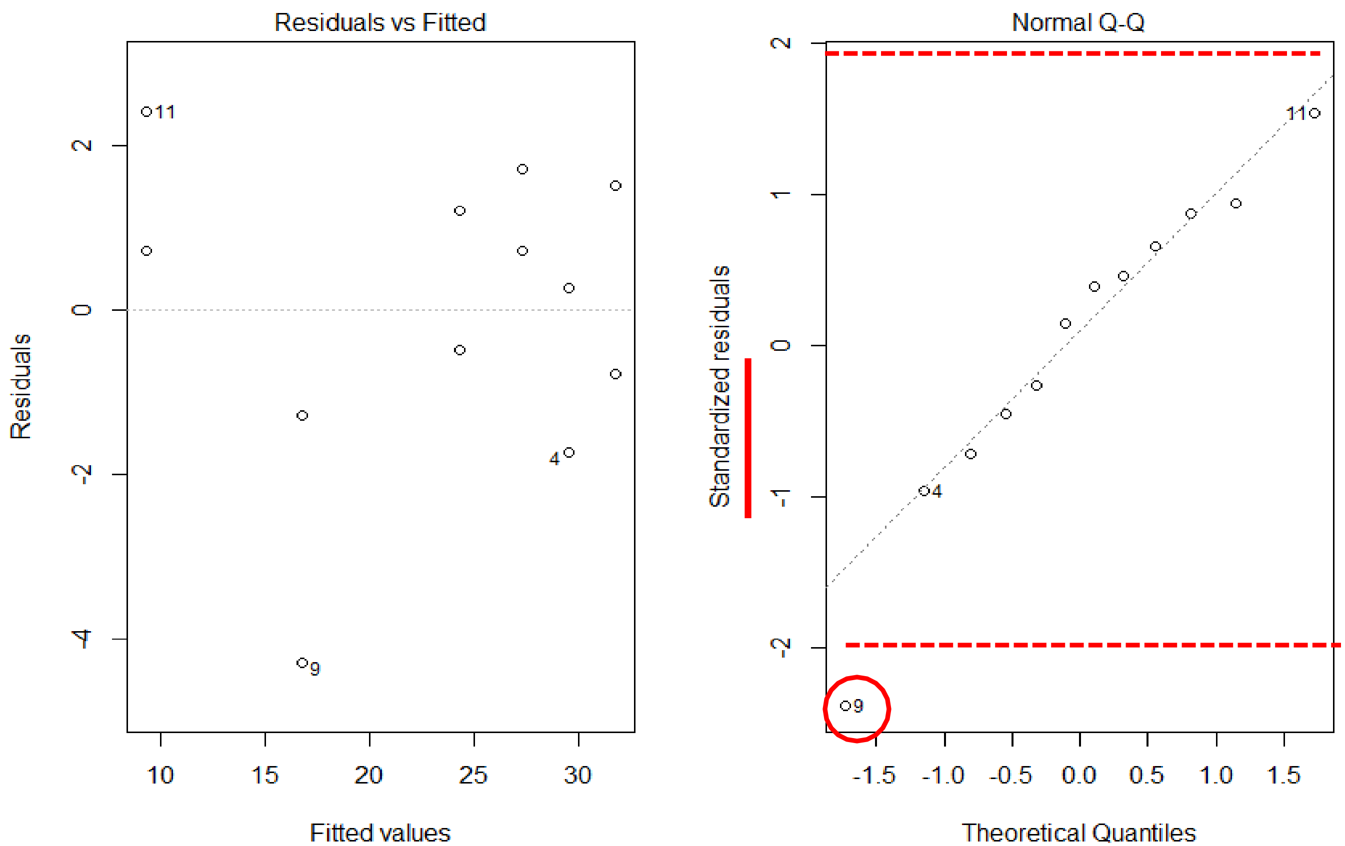


Figure 3.6: The probability distribution of χ

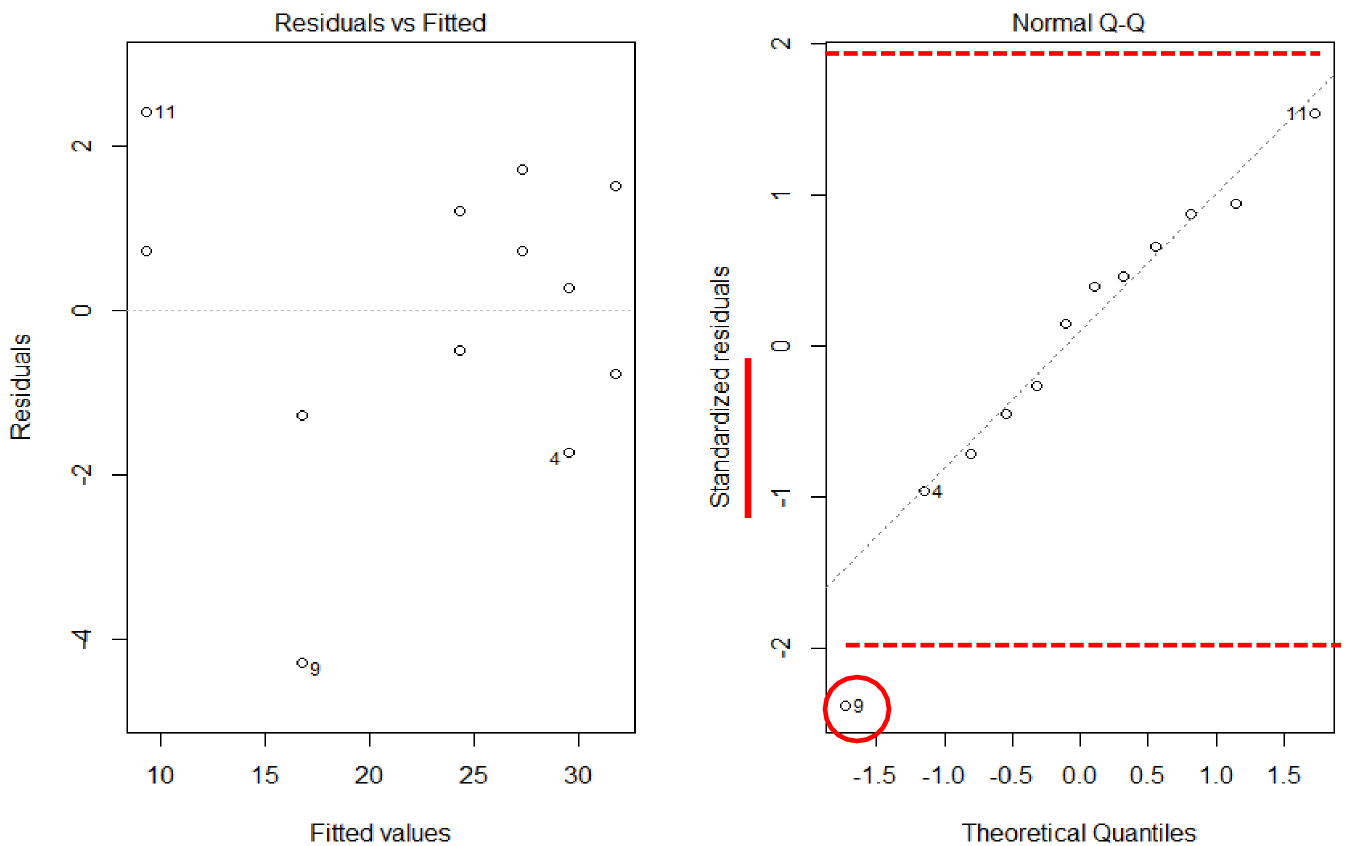
Checking Model Assumptions



Residuals vs fitted values:

The residuals appear to be randomly scattered about 0, which suggests that a straight line model is appropriate and the assumption of constant variance and independent residuals appears valid, apart from one extreme value (obs 9).

Checking Model Assumptions



Normal Q-Q plot:

Most of the points are in the straight line. The normal QQ plot suggests that residuals are approximately normally distributed.

Checking Model Assumptions

Formal test of normality

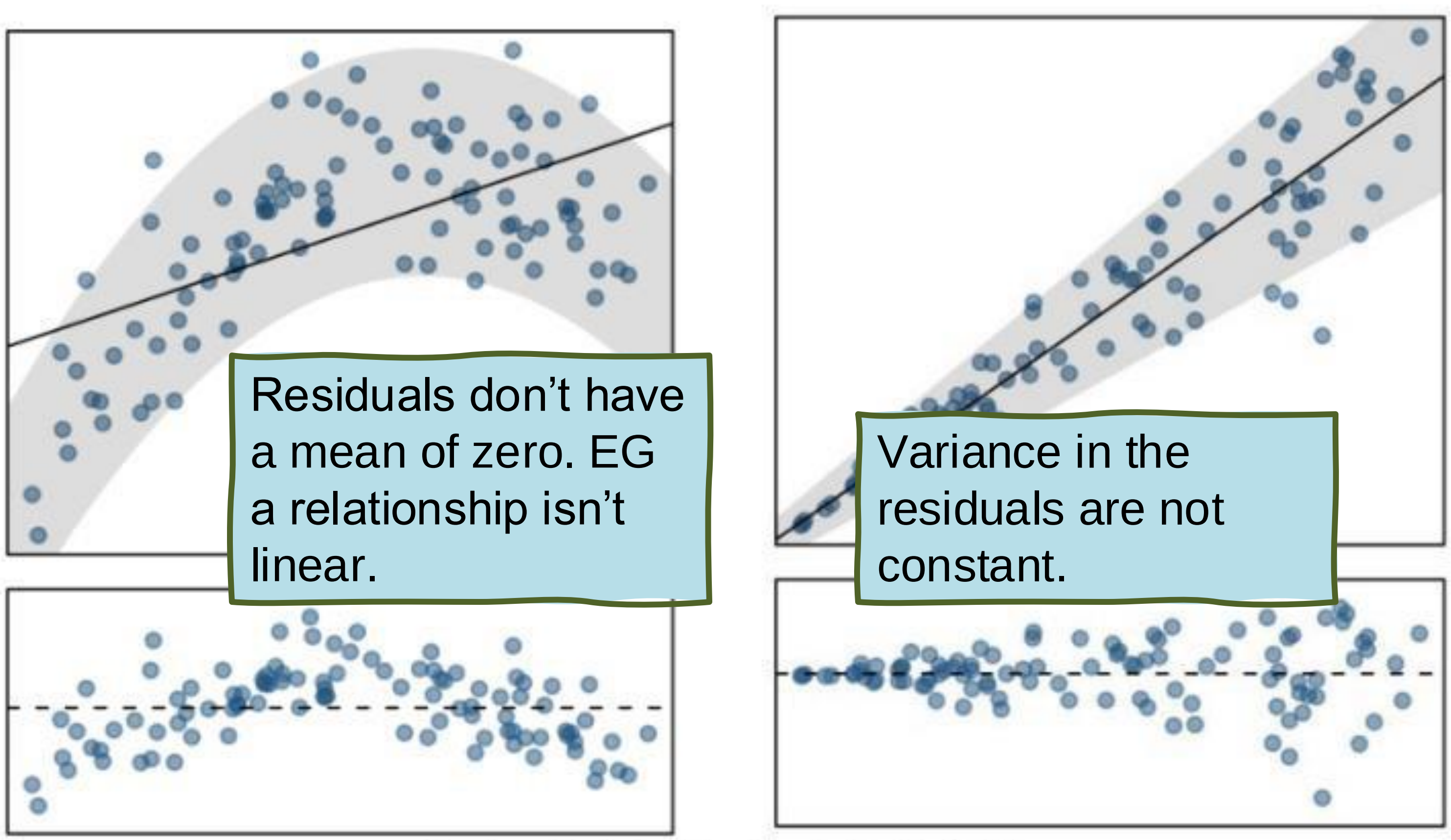
H_0: Residuals are normally distributed

```
> shapiro.test(reg.lm$residuals)

Shapiro-Wilk normality test
data:  acidreg.lm$residuals
W = 0.932, p-value = 0.4047
```

NOTE: If the p-value is <0.05, there is evidence to reject the null hypothesis.

Checking Model Assumptions

What condition is this linear model violating?



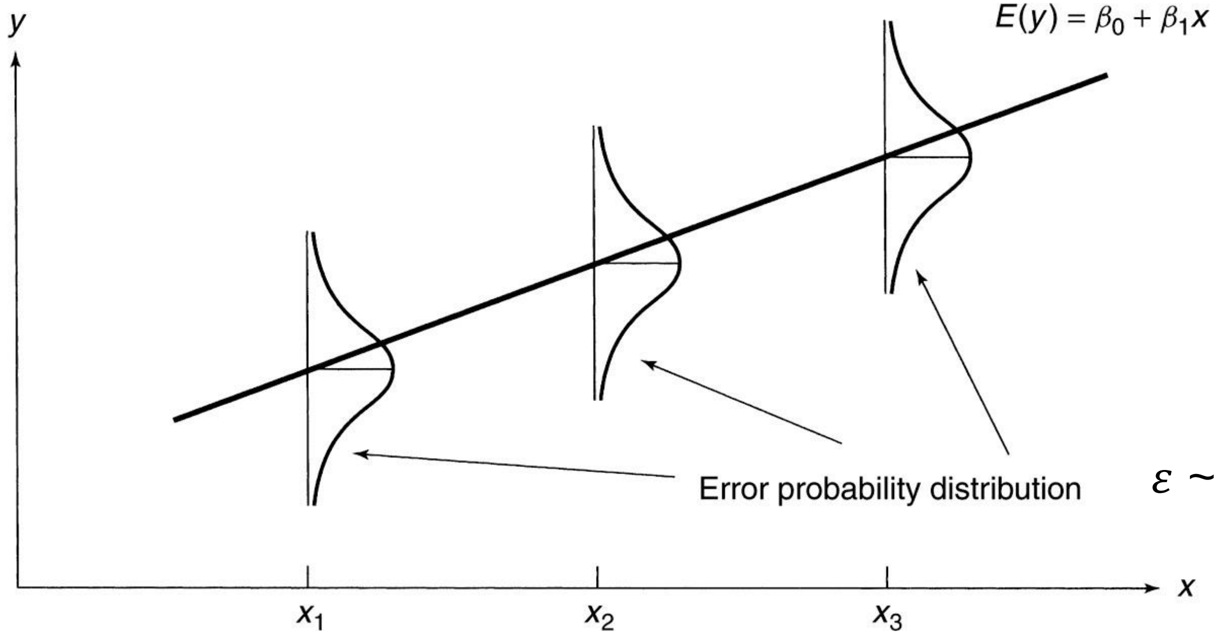
Residuals don't have a mean of zero. EG a relationship isn't linear.

Variance in the residuals are not constant.

OpenIntro Statistics. Diez, Barr & Rundel (2015).

Estimating σ^2

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



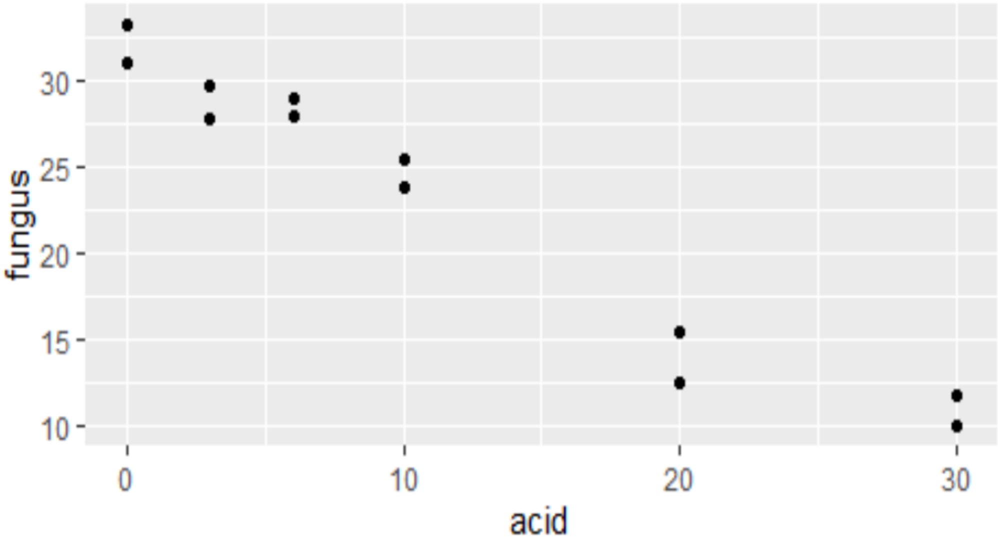
$\varepsilon \sim N(0,\sigma^2)$

Large variance $\sigma^2 \leftrightarrow$ greater variability in the random errors ϵ

$\rightarrow$ greater errors in the estimation of β_0, β_1

$\rightarrow$ unreliable prediction of $\hat{y}$

SLR Example



Source: Statistics for the Life Sciences, (2nd edn), M. L. Samuels & J. A. Witmer, (1999), p. 512

	acid	fungus
1	0	33.3
2	0	31.0
3	3	29.8
4	3	27.8
5	6	28.0
6	6	29.0
7	10	25.5
8	10	23.8
9	20	12.5
10	20	15.5
11	30	11.7
12	30	10.0

Steps in Regression Analysis

Step 1. Hypothesize the form of the model for E(y)

Step 2. Collect the data (sample data)

Step 3. Use the collected data to estimate the unknown parameters in the model

Step 4. Investigate the random error term, checking model assumptions

Step 5. Statistically check the usefulness of the model

Step 6. When satisfied that the model is useful, use it for prediction, estimation and so on.

Estimating σ^2

```
summary(reg.lm)
################
Coefficients:
          Estimate    SE        t     Pr(>|t|)
Inter       31.78   0.83   38.17    3.63e-12
acid        -0.75   0.05 -13.98     6.89e-08
```

Fungal growth = 31.78 -0.75acid

Intercept: $\widehat{\beta_0}$

slope: $\widehat{\beta_1}$

```
Residual standard error: 1.937 on 10df
Multiple R-Squared: 0.9513,
Adjusted R-squared: 0.9464
F-statistic: 195.3 on 1 and 10 DF,
 p-value: 6.888e-08
```

s = residual standard error
= estimate of σ

Estimating σ^2

```
anova(reg.lm)
##########################
Analysis of Variance Table
Response: density
       Df  SumSq MeanSq Fvalue    Pr(>F)
acid    1 732.64 732.64  195.31 6.8e-08
Resids 10  37.51   3.75
```

s^2 = MSE (MS residuals)
= estimate of σ^2

NB: $1.937^2 = 3.75$

Estimating σ^2

- The residuals $\varepsilon \sim N(0, \sigma^2)$ with σ^2 unknown
- σ^2 can be estimated using
  - *The residual standard error*, s, from the summary table
  - The *MSE*, s^2, from the ANOVA table

Inference about β_1

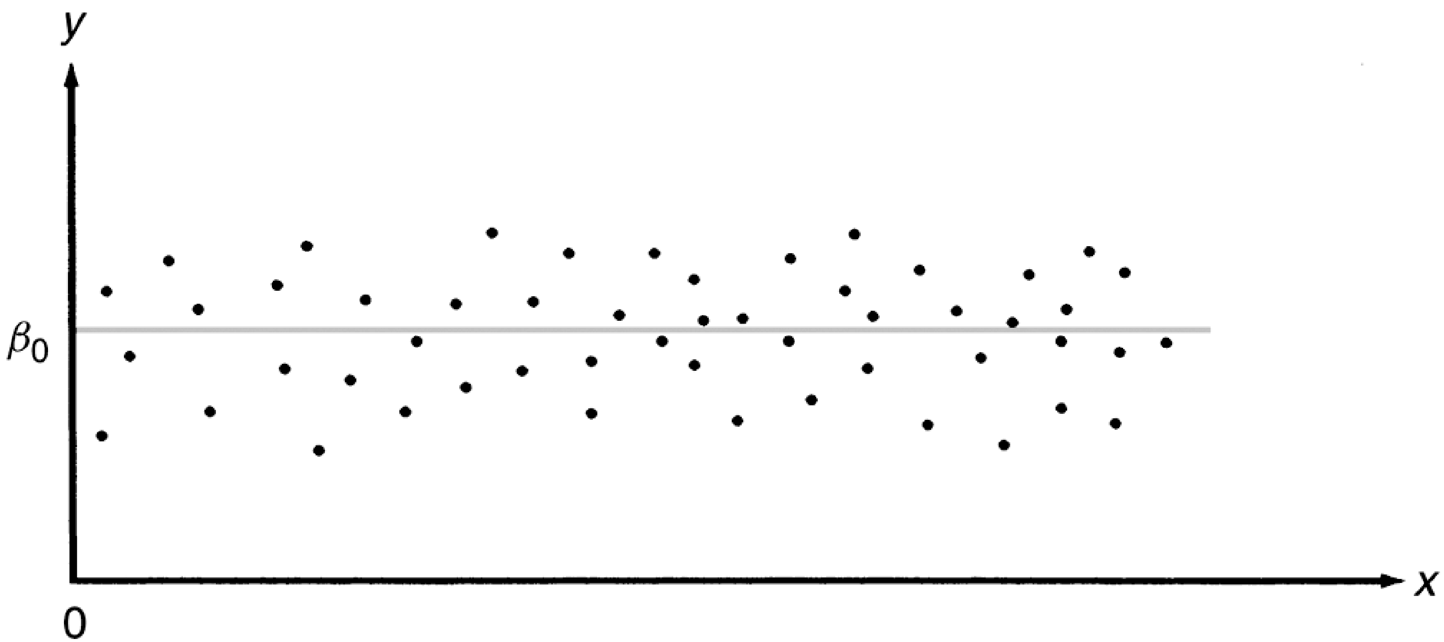


Figure 3.8 Graphing the model with $\beta_1 = 0$, $y = \beta_0 + \varepsilon$

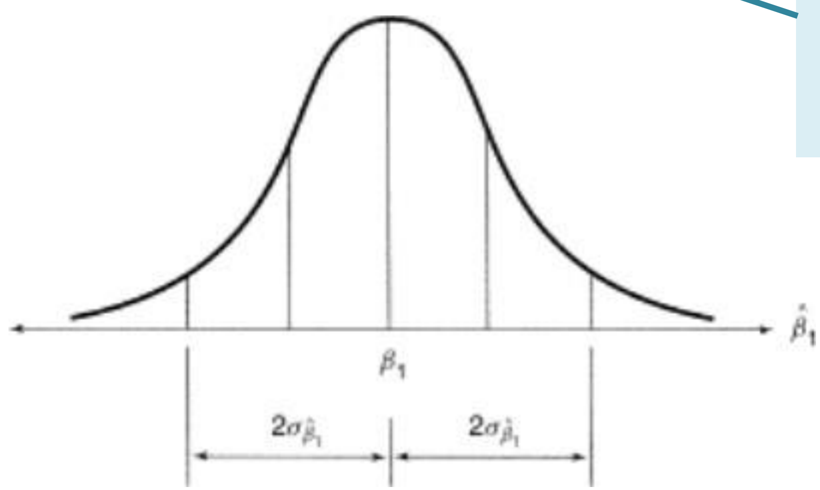
Inference about β_1

Sampling Distribution of $\hat{\beta}_1$

If we make the four assumptions about ε (see Section 3.4), then the sampling distribution of $\hat{\beta}_1$, the least squares estimator of the slope, will be a normal distribution with mean β_1 (the true slope) and standard deviation

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SS_{xx}}} \text{ (See Figure 3.9.)}$$

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



Inference about β_1

Test H_0: $\beta_1=0$

```
summary(reg.lm)
################
Coefficients:
           Estimate   SE       t     Pr(>|t|)
Inter         31.78  0.83   38.17    3.63e-12
acid          -0.75  0.05  -13.98    6.89e-08
Residual standard error: 1.937 on 10df
Multiple R-Squared: 0.9513,
Adjusted R-squared: 0.9464
F-statistic: 195.3 on 1 and 10 DF,
 p-value: 6.888e-08
```

Fungal growth = 31.78 -0.75acid

Intercept: $\widehat{\beta_0}$

slope: $\widehat{\beta_1}$

s($\widehat{\beta_1}$)

Test H_0: $\boldsymbol{\beta_1=0}$

Inference about β_1

1. H_0: $\beta_1=0$ i.e. there is no linear association between fungal growth and acid concentration

2. Test statistic: $t = \frac{\widehat{\beta_1}}{s(\beta_1)} = \frac{-0.75}{0.05} \approx -14$

3. df = n-2 = 12-2 =10

4. p-value for two-sided alternative (H_0:$\beta_1 \neq 0$), using R

```
>2*pt(-14, df=10, 0.025)
[1] 6.2544e-08
```

(p-value = 6.3×10^{-8})

5. *Conclusion*: p-value << 0.05, reject H_0. There is a (negative) linear association between fungal growth and acid concentration

Inference about β_1

Testing: $\beta_1 = 0$

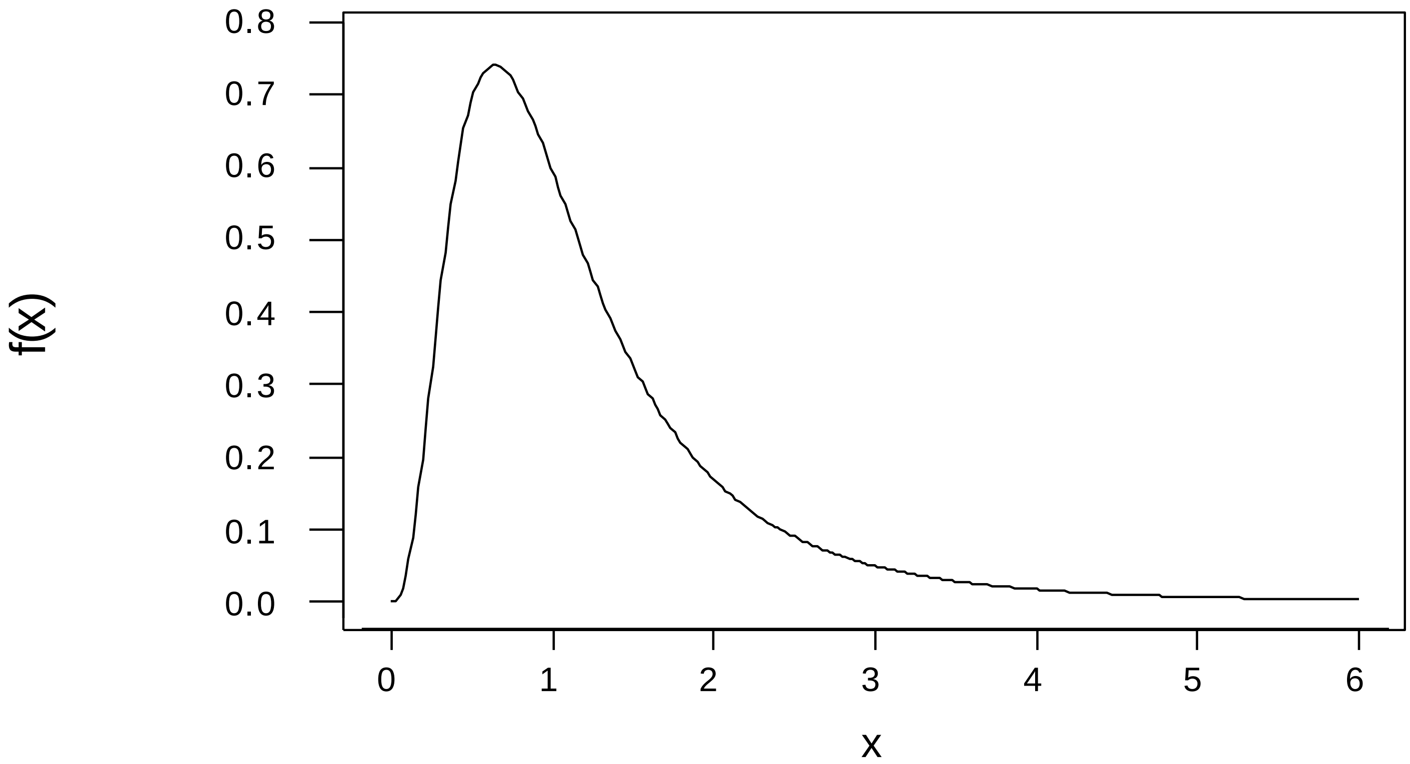
```
> anova(reg.lm)
Analysis of Variance Table
Response: density
        Df   SumSq MeanSq  F value    Pr(>F)
acid     1 732.64 732.64  195.31     6.8e-08
Resids  10  37.51   3.75
```

F= $\frac{Regression\ MS}{Error\ MS} = \frac{732.64}{3.75} = 195.37$, Degrees of freedom: 1 and 10

```
> 1-pf(195.31,1,10)
[1] 6.9e-08
```

Graph of F distribution

Distribution depends on two parameters: degrees of freedom of numerator (between group variance) and denominator (within group variance)



Confidence Intervals for regression coefficients

CI for a parameter:

For 1 µg/mL increase in acid concentration, fungal growth will *decrease* by 0.75mm, on average. The *margin of error* is 0.12 mm. This is stated with 95% confidence.

95% CI for slope, $\boldsymbol{\beta_1}$:

$-0.75 \pm 2.23 \times 0.05 \quad = -0.75 \pm 0.12$

$= (-0.87, -0.63)$

Q: Give a practical interpretation of the slope and the 95% CI

Confidence Intervals for regression coefficients

CI for a parameter:

estimate $\pm$ t ×se(estimate)

We are 95% confident that for 1 µg/mL increase in acid concentration, fungal growth will *decrease* by between 0.63 and 0.87mm.

95% CI for slope, β_1:

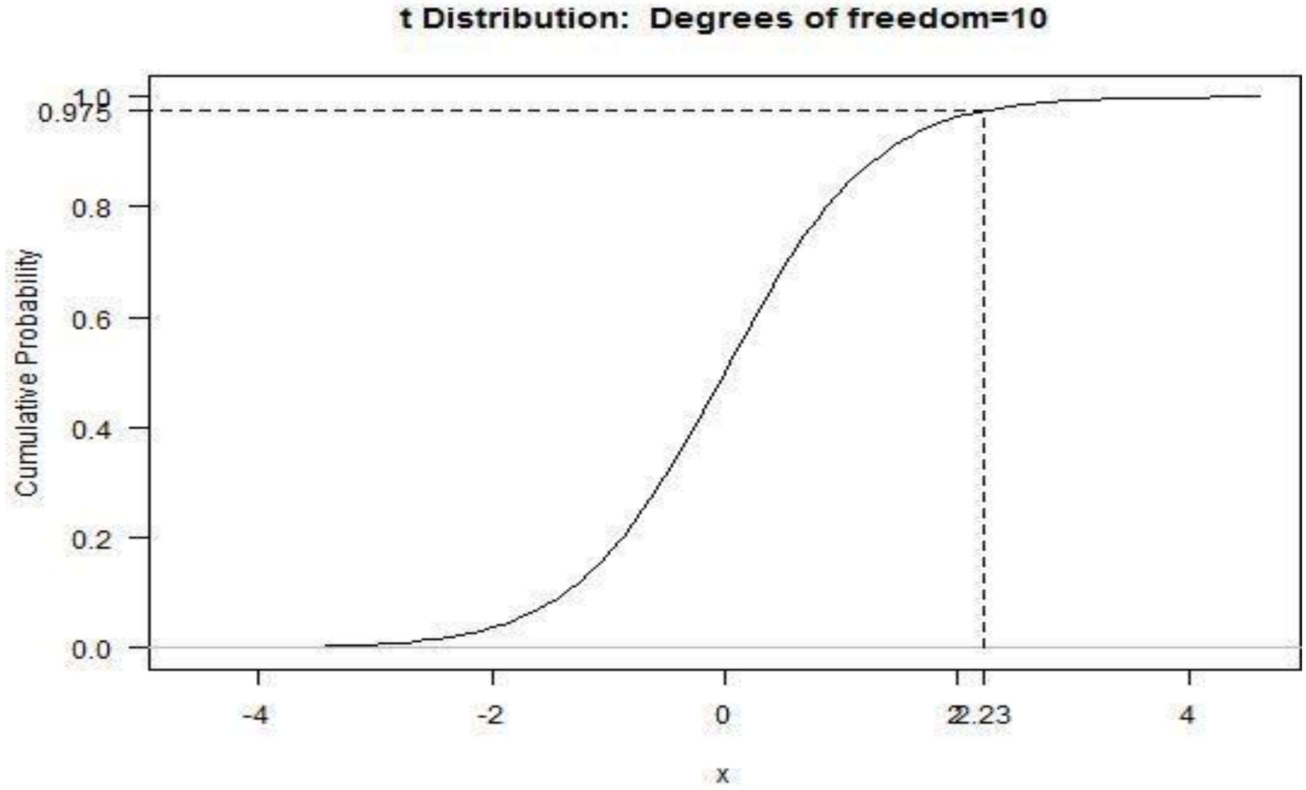
-0.75 $\pm$ **2.23** x 0.05

= -0.75 $\pm$ 0.12

= (-0.87, -0.63)

Q: Give a practical interpretation of the slope and the 95% CI

Distribution depends on one parameter: degrees of freedom




```
qt(df=10,p=0.975)
[1] 2.23
qt(df=10,p=0.025)
[1] -2.23
```

Confidence Intervals for regression coefficients

```
confint(reg.lm, level=0.95)

             2.5 %    97.5 %
(Intercept) 29.93    33.64
acid        -0.87    -0.63
```

F Test

Relationship between F and t

$$F_{1,n-2} = t_{n-2}^{2}$$

Why two tests?

The t-test evaluates the components of the model, the f-test evaluates the model.

Summary

- SLR Model assumptions
  - Residuals are independent
  - Residuals $\epsilon \sim N(0, \sigma^2)$ with mean 0, and constant variance with regard to X

- σ^2 can be estimated using
  - The residual standard error, s,
  - The MSE, s^2

- Inference about the slope, 95% CI

Lecture 3

Simple Linear Regression

Chapter 3 Outline

Lecture 1:

- Introduction
- Linear statistical models
- Method of least squares

Lecture 2:

- Model assumptions, estimator of σ^2
- Inference about the slope

Lecture 3:

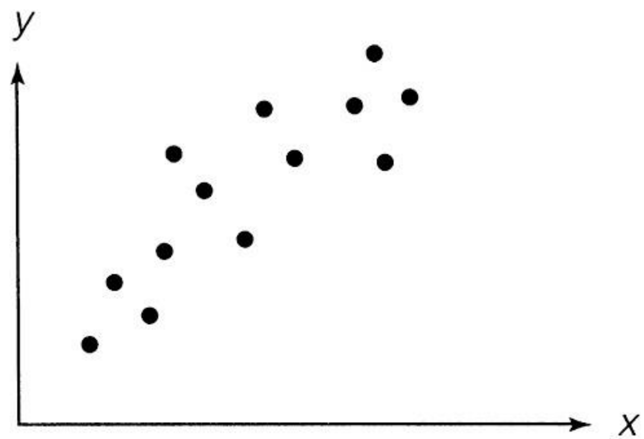
- Coefficient of determination (R^2), correlation (r)
- Using the model for estimation & prediction

Correlation Coefficient, *r*

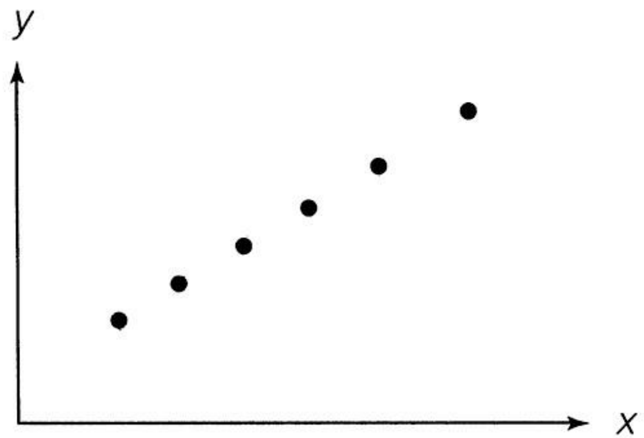
A measure of the *direction* and *strength* of the *linear* association between two *quantitative* variables.

- $-1 \le r \le 1$
- $r = \pm 1$, **perfect** *linear* association
- $r = 0$, **no** *linear* association

Figure 3.12 a & b Interpreting r

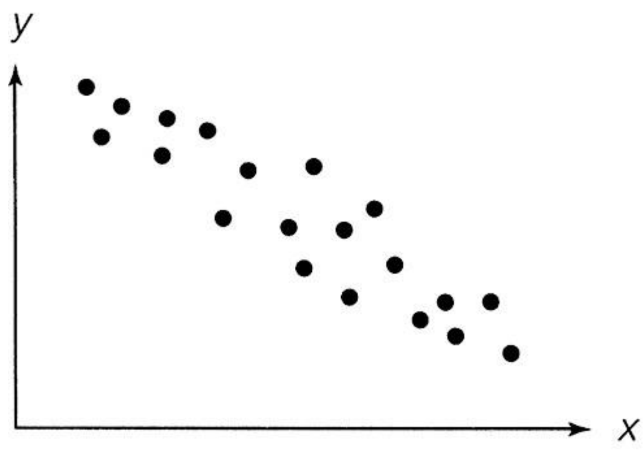


(a) Positive *r*: *y* increases as *x* increases

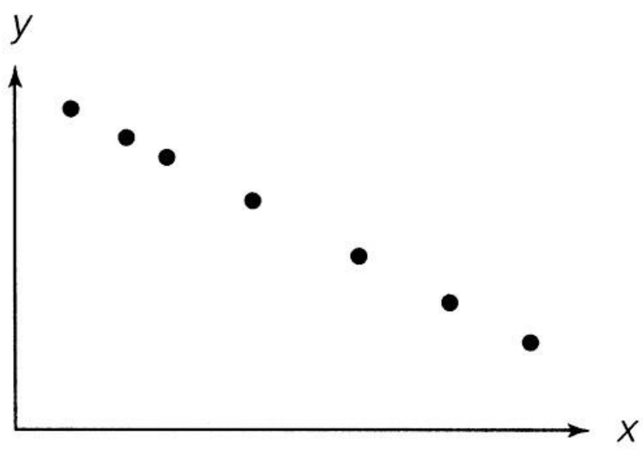


(b) $r = 1$: a perfect positive linear relationship between *y* and *x*

Figure 3.12 c & d Interpreting r

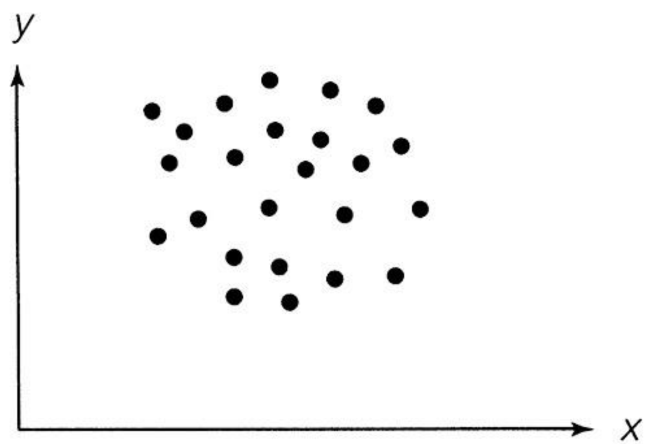


(c) Negative r: y decreases as x increases

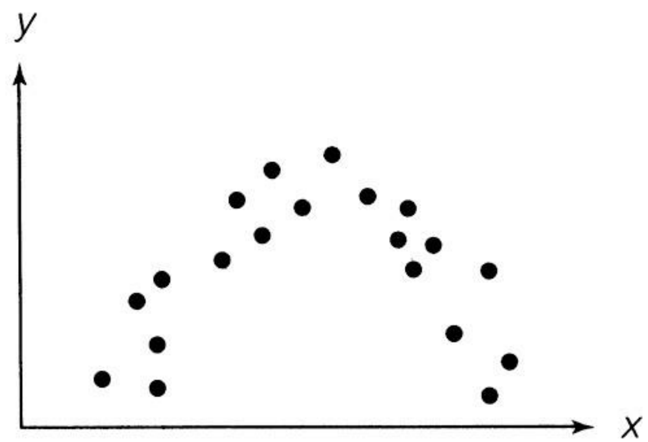


(d) $r = -1$: a perfect negative linear relationship between y and x

Figure 3.12 e & f Interpreting r



(e) *r* near zero: little or no linear relationship between *y* and *x*



(f) *r* near zero: little or no linear relationship between *y* and *x*

Correlation does NOT imply Causation

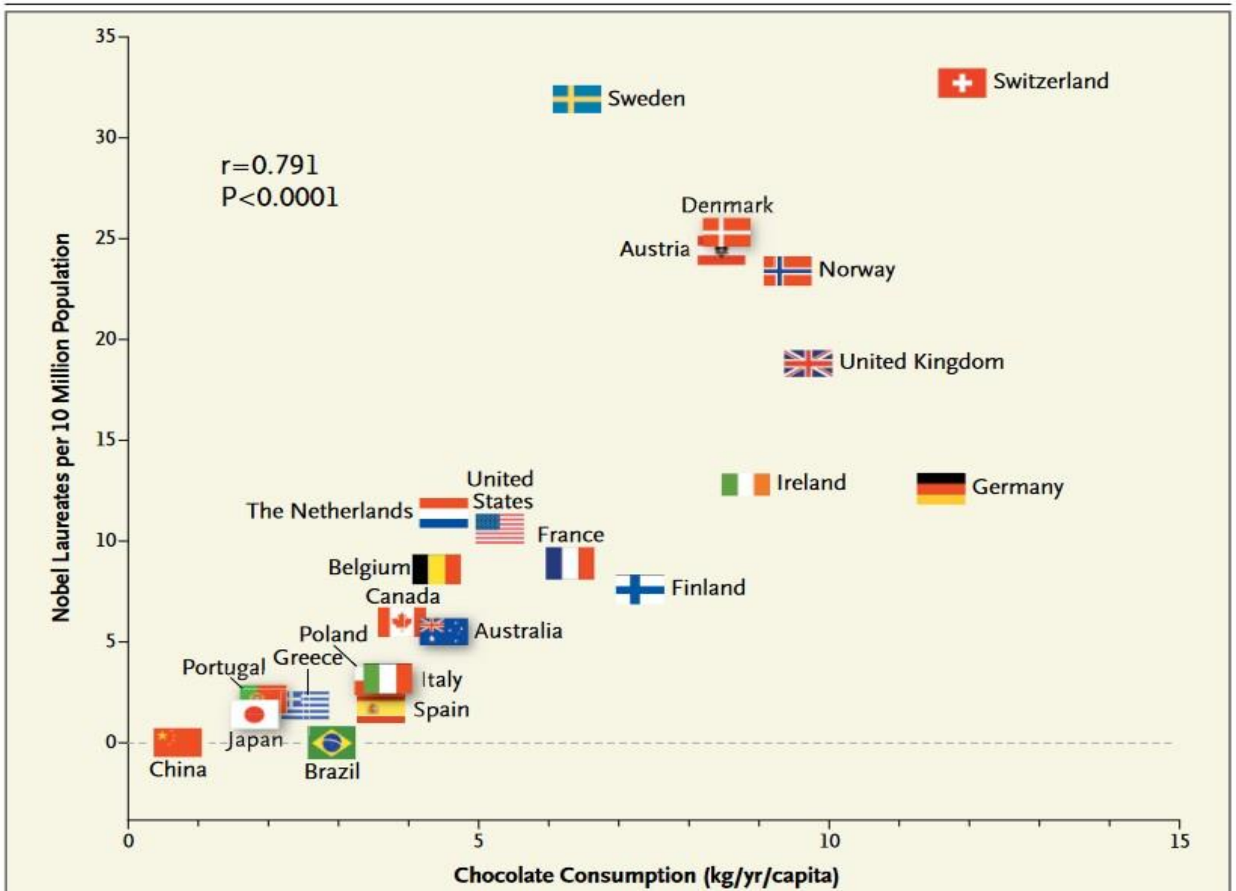
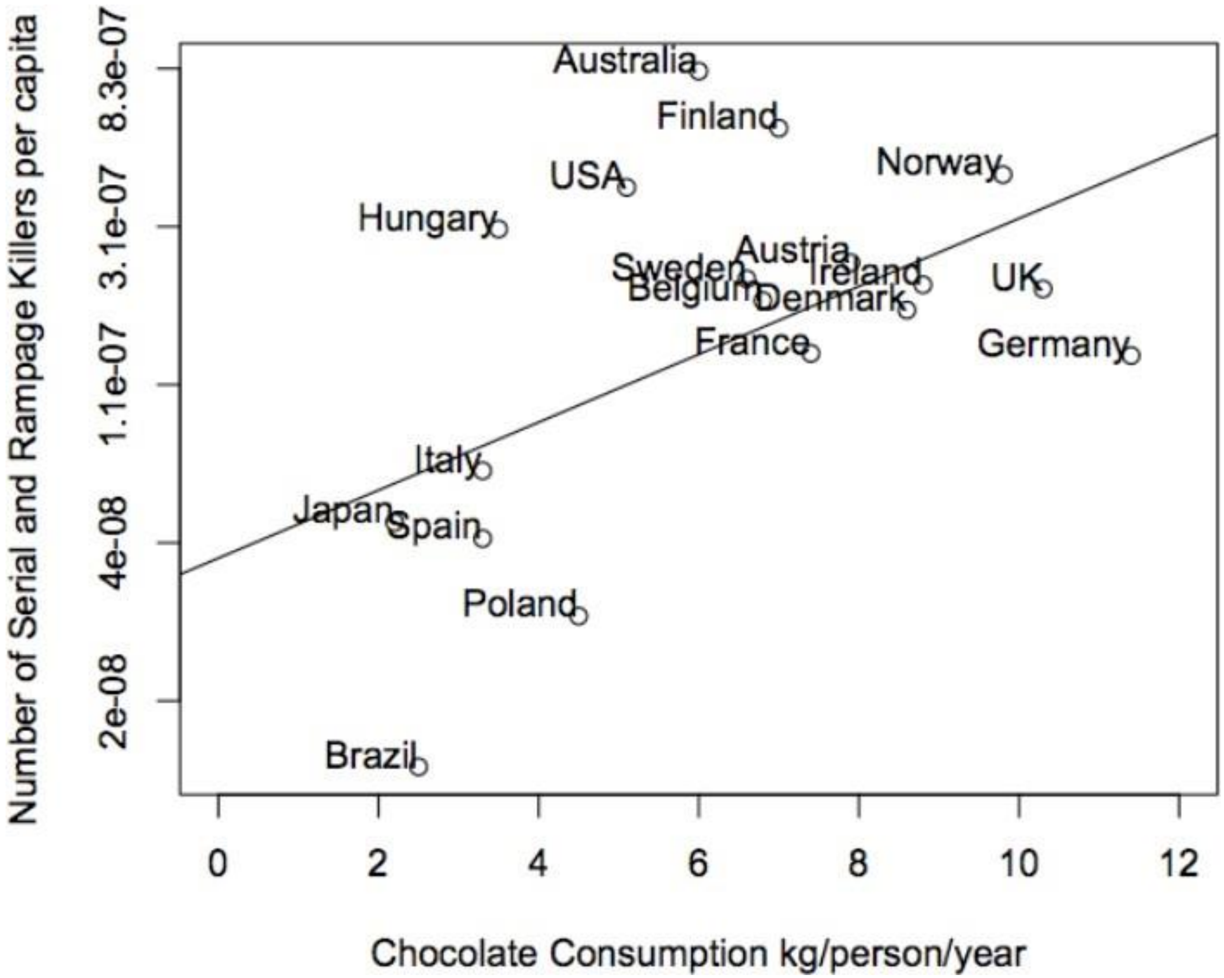


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Messerli 2012*. Chocolate Consumption, Cognitive Function, and Nobel Laureates*. The New England Journal of Medicine https://www.nejm.org/doi/full/10.1056/NEJMon1211064

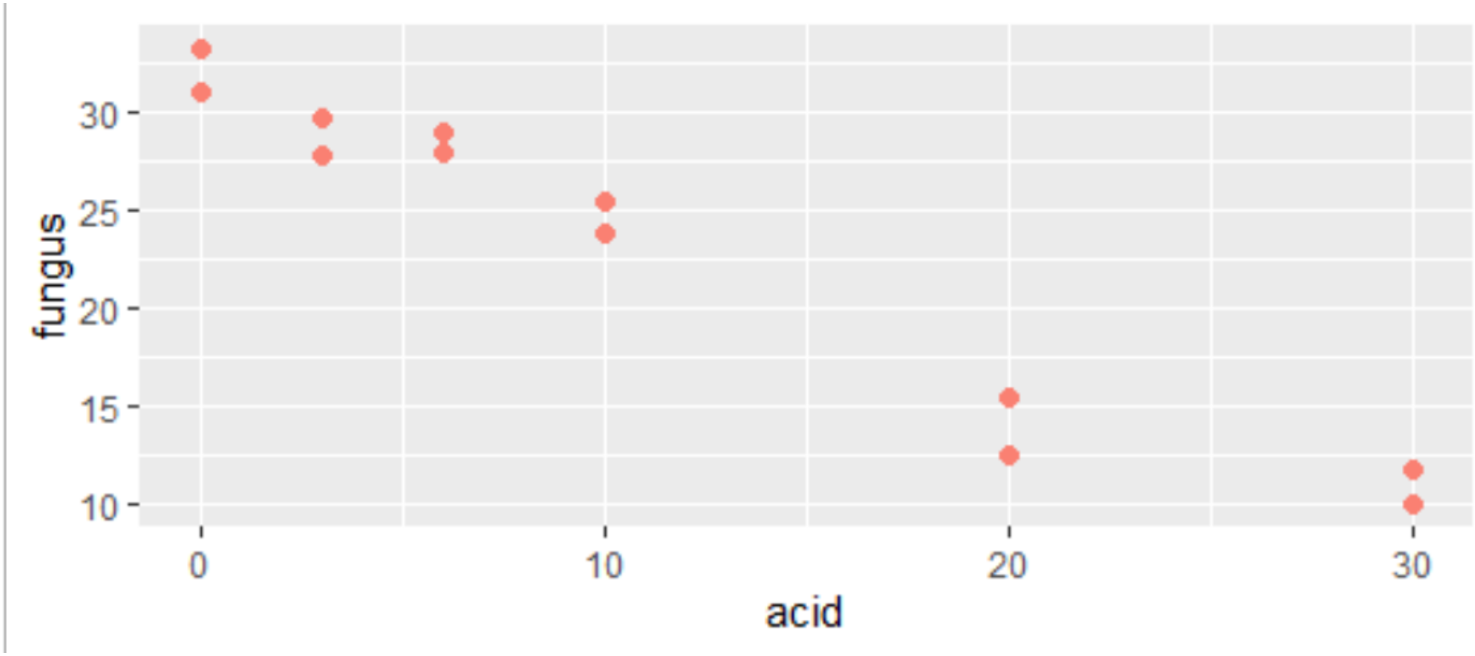
Correlation does NOT imply Causation



Winters & Roberts. Chocolate Consumption, Traffic Accidents and Serial Killers.
http://replicatedtypo.com/wp-content/uploads/2012/11/ChocolateSerialKillers_WintersRoberts.pdf

Example:

Relationship between fungal growth and acid concentration.



Q: Estimate the value of the correlation coefficient

Coefficient of Determination

```
summary(reg.lm)
###############
Coefficients:
           Estimate   SE       t    Pr(>|t|)
Inter         31.78  0.83  38.17    3.63e-12
acid          -0.75  0.05 -13.98    6.89e-08

Residual standard error: 1.937 on 10df
Multiple R-Squared: 0.9513,
Adjusted R-squared: 0.9464
F-statistic: 195.3 on 1 and 10 DF,
 p-value: 6.888e-08
```

Coefficient of Determination, R^2

Correlation and Determination

- The correlation, r, between fungal growth and acid concentration is:

```
>cor(acid.df$fungus,acid.df$acid)
 [1] -0.975
```

There is a __strong__ ($|r| \approx 1$), __negative__ ($r < 0$) __linear__ association between fungal growth and acid concentration

- The value of R^2 is $(-0.975)^2 = 0.9513$

95% of the variability in fungal growth is explained by the __linear association__ with acid concentration

Steps in Regression Analysis

Step 1. Hypothesize the form of the model for E(y)

Step 2. Collect the data (sample data)

Step 3. Use the collected data to estimate the unknown parameters in the model

Step 4. Investigate the random error term, checking model assumptions

Step 5. Statistically check the usefulness of the model

Step 6. When satisfied that the model is useful, use it for prediction, estimation and so on.

A 100(1 – α)% Confidence Interval for the Mean Value of y for $x = x_p$

$$\hat{y} \pm t_{\alpha/2} \text{ (Estimated standard deviation of } \hat{y}\text{)}$$

or

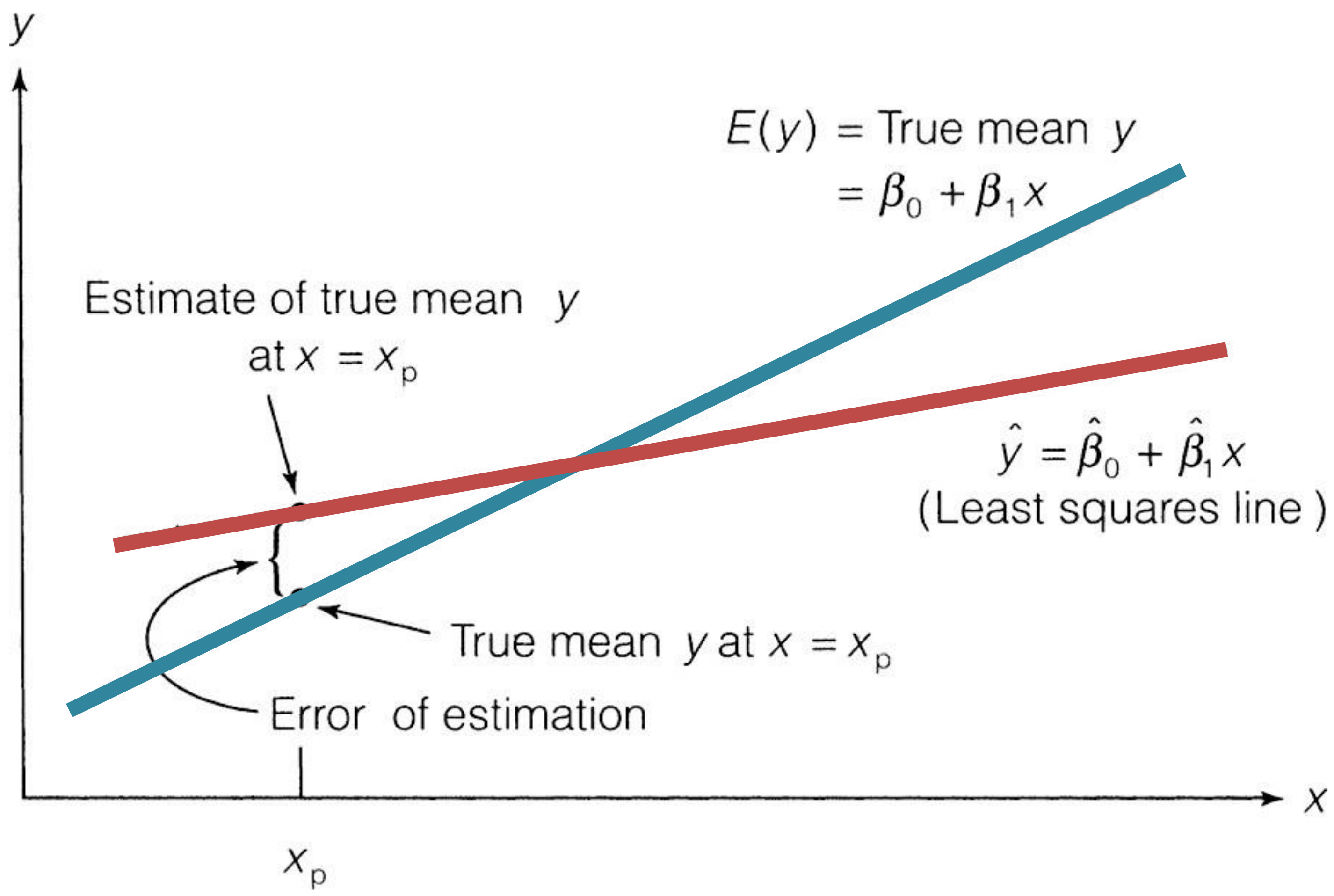
$$\hat{y} \pm (t_{\alpha/2}) s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

where $t_{\alpha/2}$ is based on $(n - 2)$ df

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

NB: You are not expected to calculate this interval "by hand" using the formula.

Figure 3.23 Error of estimating the mean value of *y* for a given value of *x*



Extra random error term ($\boldsymbol{\sigma^2}$)

A 100(1 – α)% Prediction Interval for an Individual y for $x = x_p$

$$\hat{y} \pm t_{\alpha/2}\,[\text{Estimated standard deviation of}(y - \hat{y})]$$

or

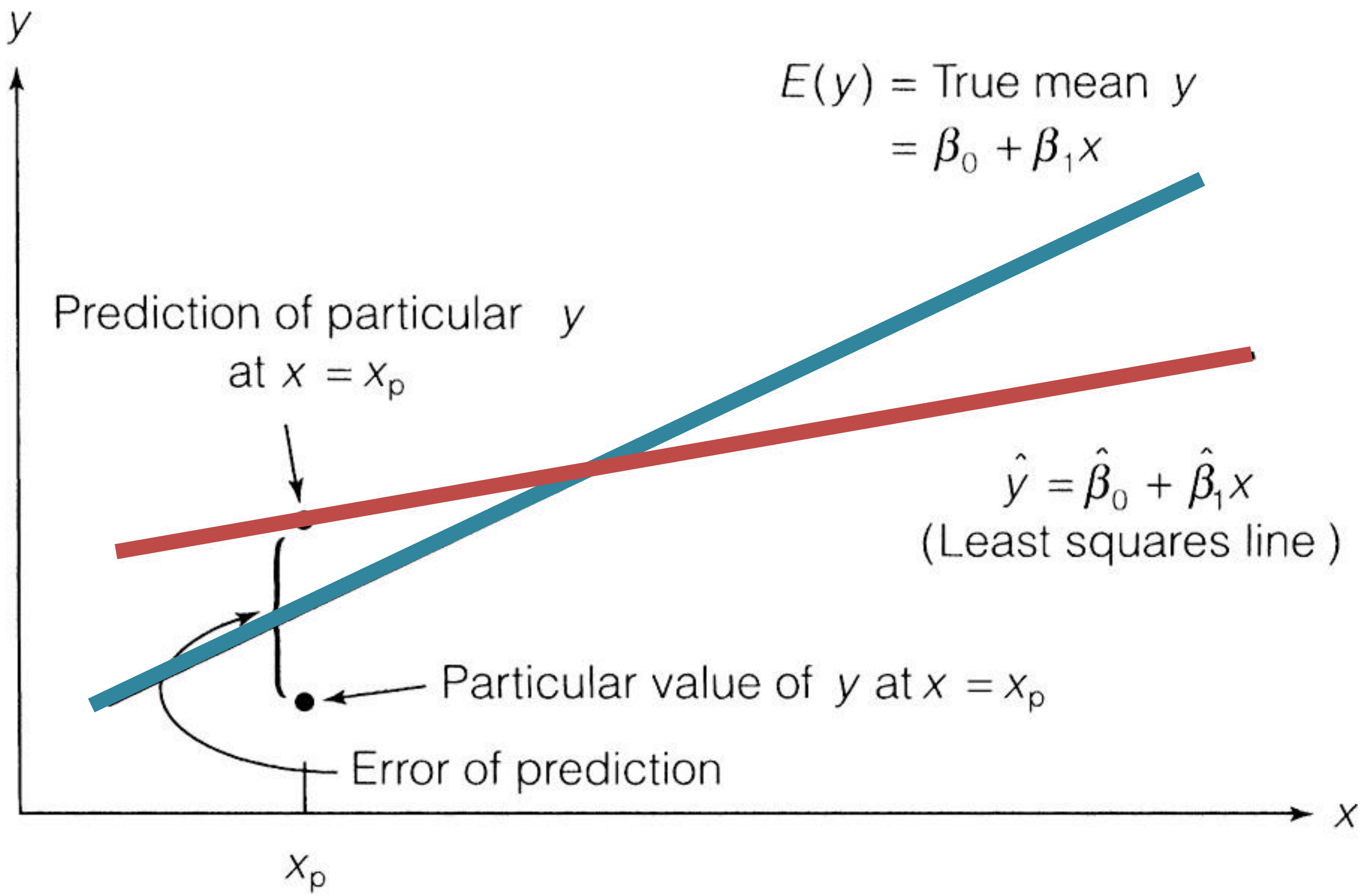
$$\hat{y} \pm (t_{\alpha/2})s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

where $t_{\alpha/2}$ is based on $(n - 2)$ df

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

NB: You are not expected to calculate this interval "by hand" using the formula.

Figure 3.24 Error of predicting a future value of *y* for a given value of *x*



***Confidence* Interval:**
Predicting a *mean* response

```
# predict the mean fungal growth and the
  95% CI when acid conc. = 15

predict(reg.lm, new=data.frame(acid=15),
       interval="confidence",level=0.95)
```

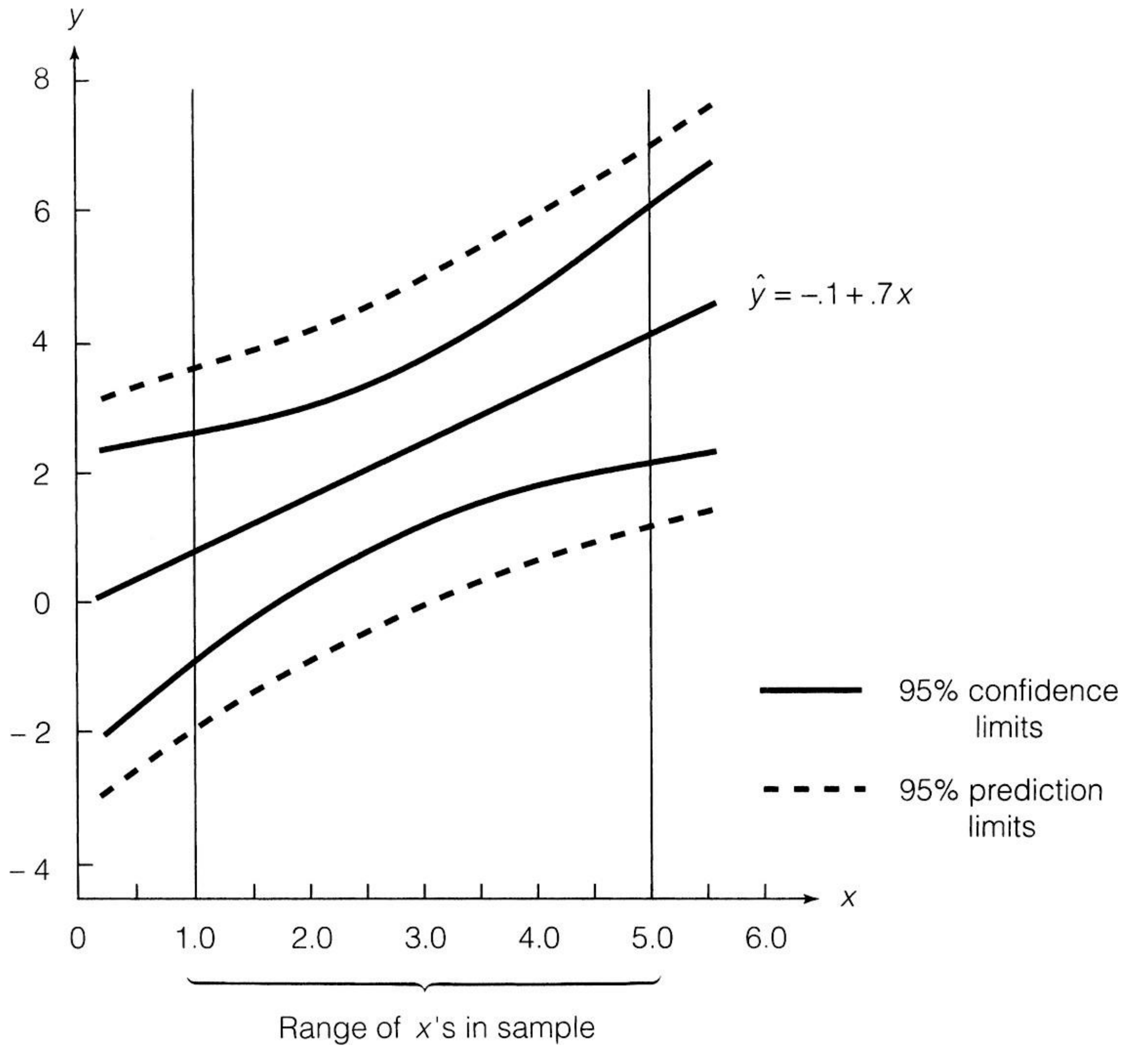
fit	**lwr**	**upr**
20.53	19.22	21.85

Prediction **Interval:** Predicting an *individual* response

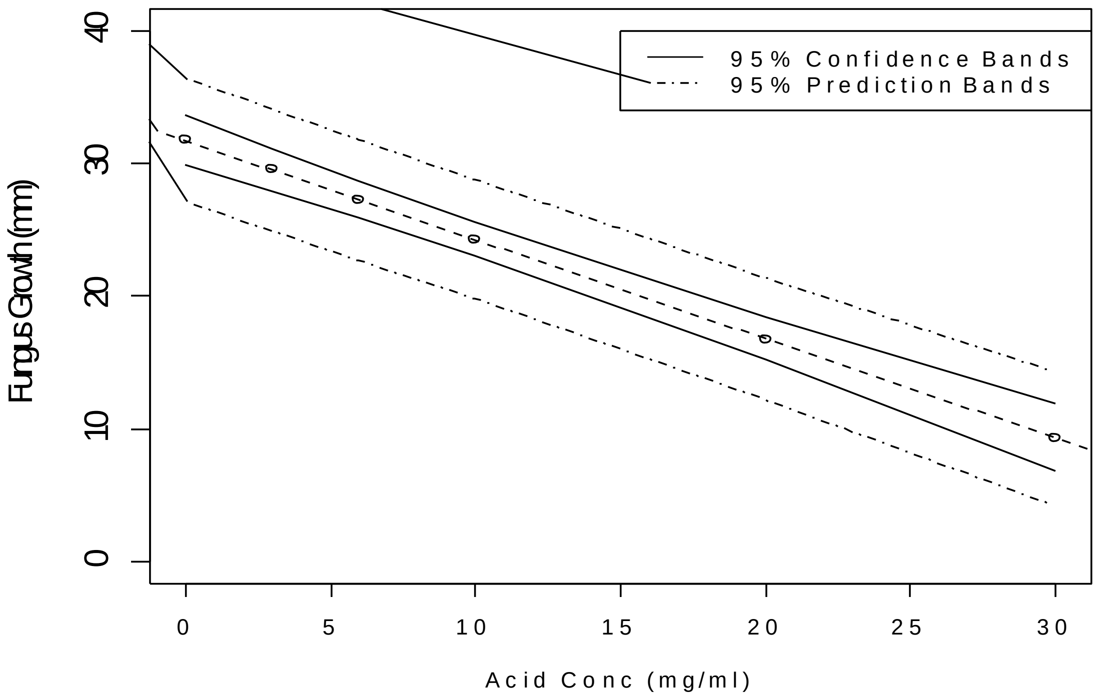
```
# predict the fungal growth and the
  95% PI for an individual plant
  when acid conc. = 15

predict(reg.lm,new=data.frame(acid=15),
 interval="predict",level=0.95)
```

	fit	**lwr**	**upr**
	20.53	16.02	25.04

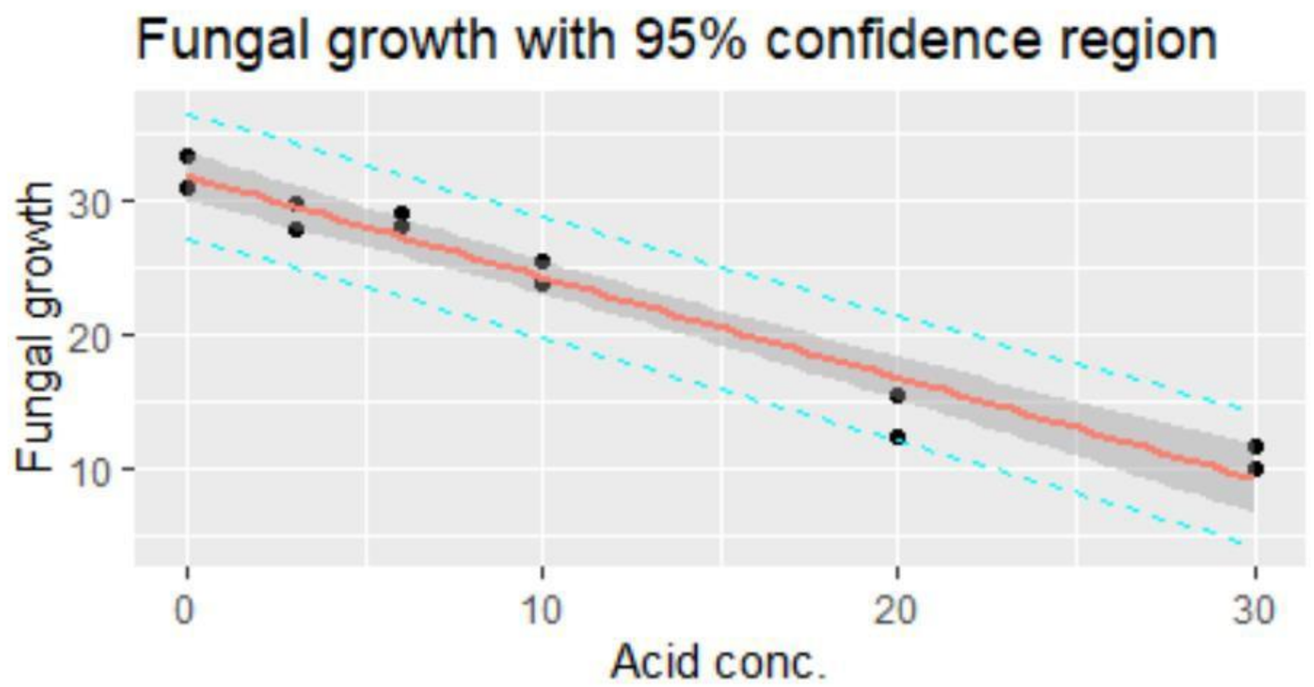
Figure 3.25 Comparison of widths of 95% confidence and prediction intervals



Confidence and Prediction bands for fungal growth example



```
#add CI and PI to scatteplot
fungus.var <- predict(reg.lm, interval="prediction")
new.df <- cbind(acid.df, fungus.var)
ggplot(data=new.df, aes(x=acid, y=fungus)) +
geom_point() +
  geom_smooth(method="lm", color="salmon", se=TRUE) +
  geom_line(aes(y=lwr), color = "cyan", linetype = "dashed")+
  geom_line(aes(y=upr), color = "cyan", linetype = "dashed")+
  labs(title="Fungal growth with 95% confidence region",
      x="Acid conc.", y="Fungal growth")
```



Finding the Residual for a prediction

What if the observed value of Fungus growth was 20mm when acid concentration was 15 µg/mL?

Given our predicted value, we can find the residual value:

Residual = observed – expected

For this example, our observed was 20mm, but our predicted (expected) was 20.53mm, so:

Residual = 20-20.53 = 0.53

Confidence interval vs prediction interval

<u>Confidence interval:</u>

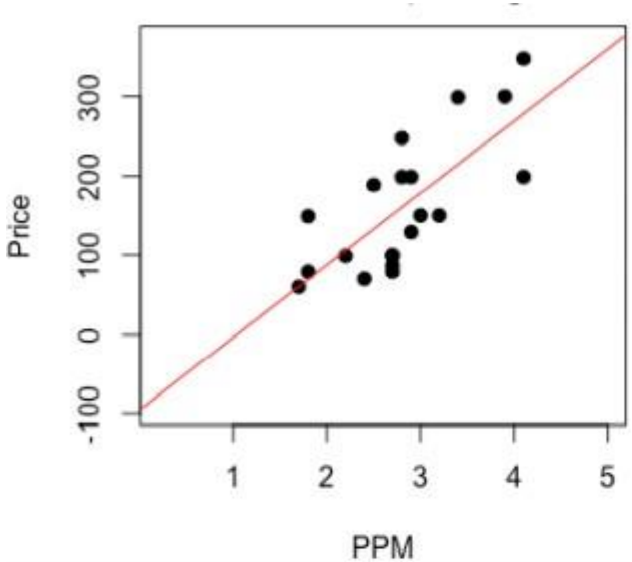
- Use when predicting a mean response
- Mean plus or minus the margin of error (incl. error of estimation)
- Always narrower than a prediction interval

<u>Prediction interval:</u>

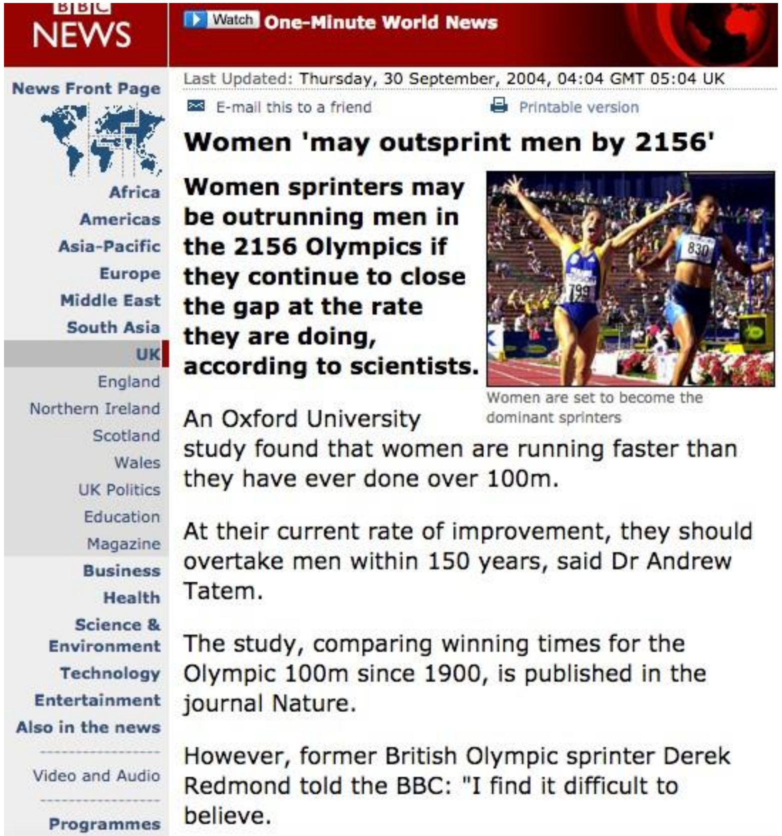
- Use when predicting an individual response
- Mean plus or minus the margin of error (incl. error of estimation and error of prediction)
- Always wider than a confidence interval

Extrapolation

- *Extrapolation* beyond the "scope of the model" occurs when we make predictions for *x* that is not in the range of the sample data used to determine the estimated regression equation.
- Sometimes the intercept might be an extrapolation



Extrapolation


BBC NEWS

Watch One-Minute World News

News Front Page

Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
England
Northern Ireland
Scotland
Wales
UK Politics
Education
Magazine
Business
Health
Science & Environment
Technology
Entertainment
Also in the news

Video and Audio

Programmes

Last Updated: Thursday, 30 September, 2004, 04:04 GMT 05:04 UK

E-mail this to a friend | Printable version

Women 'may outsprint men by 2156'

Women sprinters may be outrunning men in the 2156 Olympics if they continue to close the gap at the rate they are doing, according to scientists.

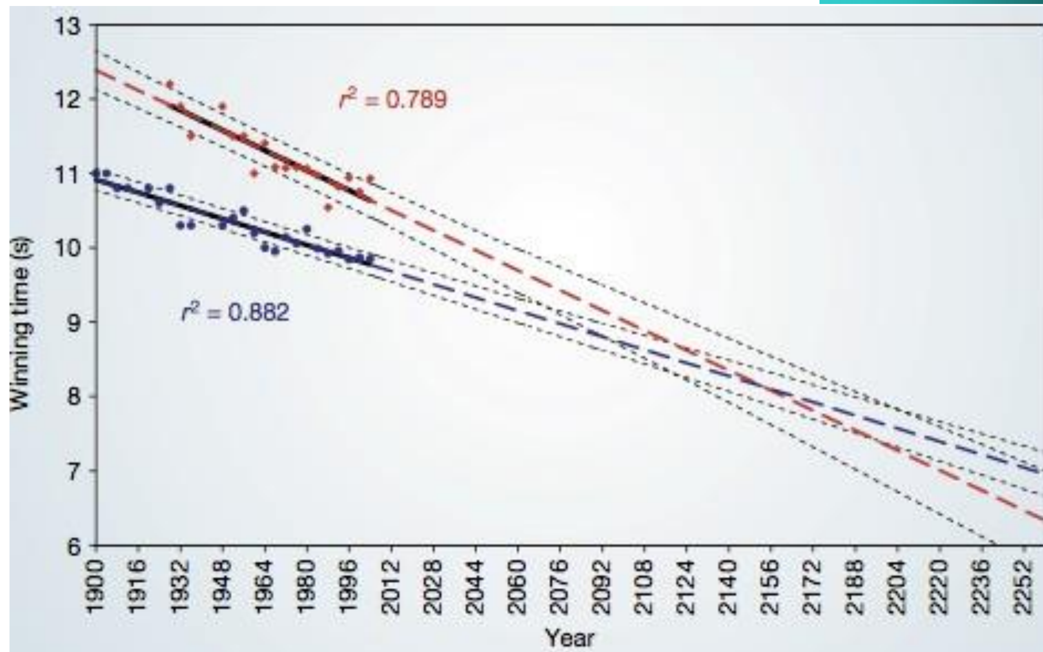
Women are set to become the dominant sprinters

An Oxford University study found that women are running faster than they have ever done over 100m.

At their current rate of improvement, they should overtake men within 150 years, said Dr Andrew Tatem.

The study, comparing winning times for the Olympic 100m since 1900, is published in the journal Nature.

However, former British Olympic sprinter Derek Redmond told the BBC: "I find it difficult to believe.



The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning womens 100-metre sprint time of 8.079s will be faster than the mens at 8.098s.

Tatem, A. J., Guerra, C. A., Atkinson, P. M., & Hay, S. I. (2004). Momentous sprint at the 2156 olympics? Nature, 431(7008), 525.

Recap Chapter 3: SLR

- Estimate intercept and slope, and their standard errors
- Conduct significance tests concerning slope.
- Calculate CI for the slope.
- Correlation coefficient, r, and coefficient of determination, R^2
- Estimate σ and σ^2

Recap Chapter 3: SLR

- Calculation of *CI* for mean response.
- Calculation of *prediction* intervals for individual response.
- Check assumptions - residual plots.
- Give informative interpretation of the regression analysis, relating to the context of the problem
- Refer to pp. 155 – 156 of the text for a summary

Exercises SLR

- Complete the Week 1 workshop
- Complete the exercises in *the SLR Worksheet*
- Additional Exercise: Ex 75, p. 160.
  - Enter data into Excel, save as text file.
  - Create and run a script file in R that
    - Imports and plots the data
    - Runs a regression analysis
    - Produces residuals plots
  - Interpret the results

NB: part solutions given at end of chapter 3 in the text

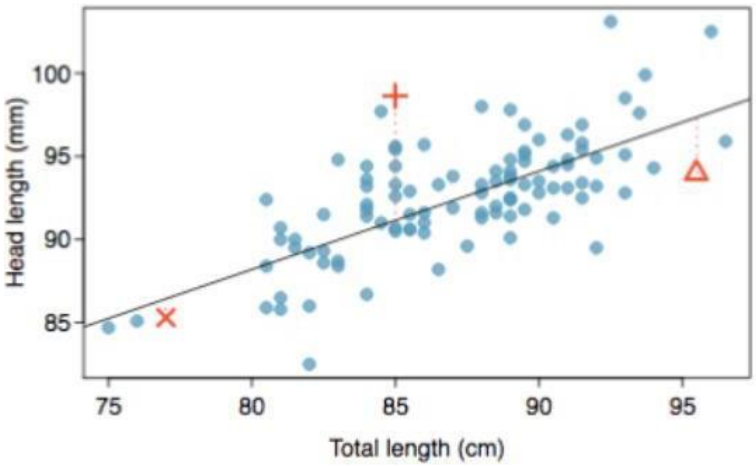
If using 8^{th} edition:
Ex 3.84 on P. 474

Overview of Linear Statistical Models

Simple linear regression

$$Y = \beta_0 + \beta_1 x + \epsilon$$

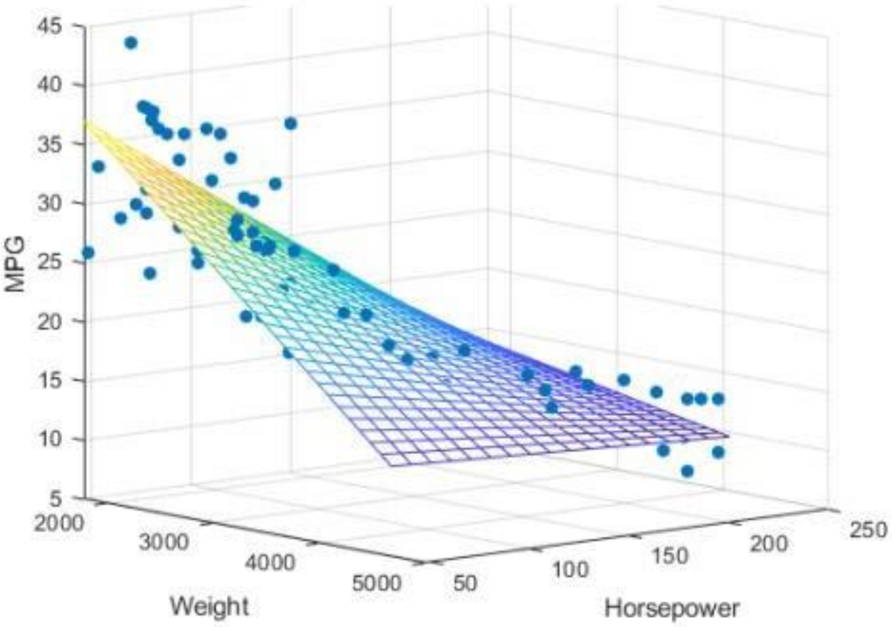
(Chapter 3)



Multiple linear regression

$$Y = \beta_0 + \beta_1 x + \cdots + \beta_k x_k + \epsilon$$

(Chapter 4)



Visual for k = 2
https://www.mathworks.com/help/stats/regress.html

Overview of Linear Statistical Models

polynomial regression

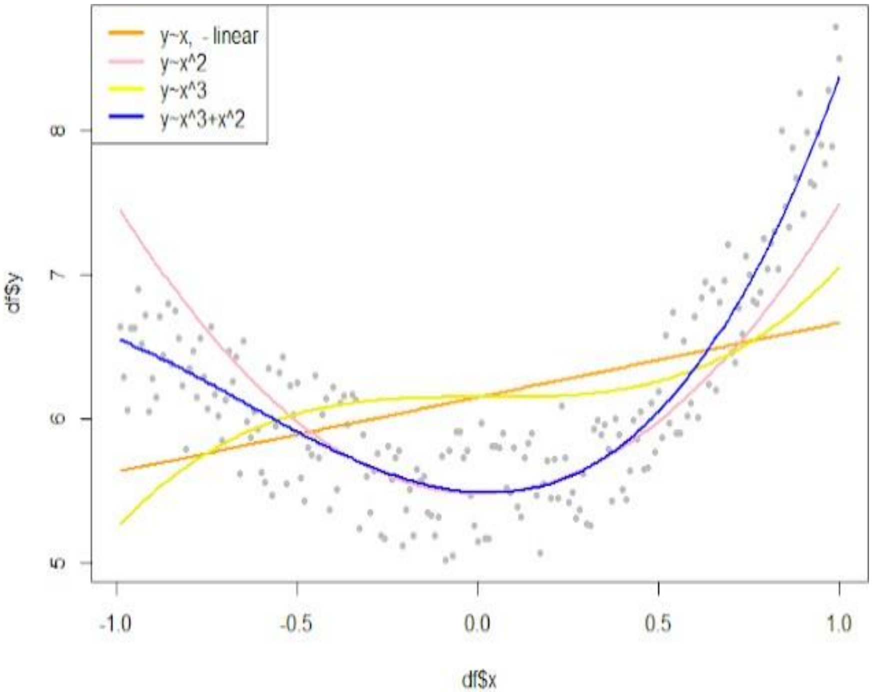
$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

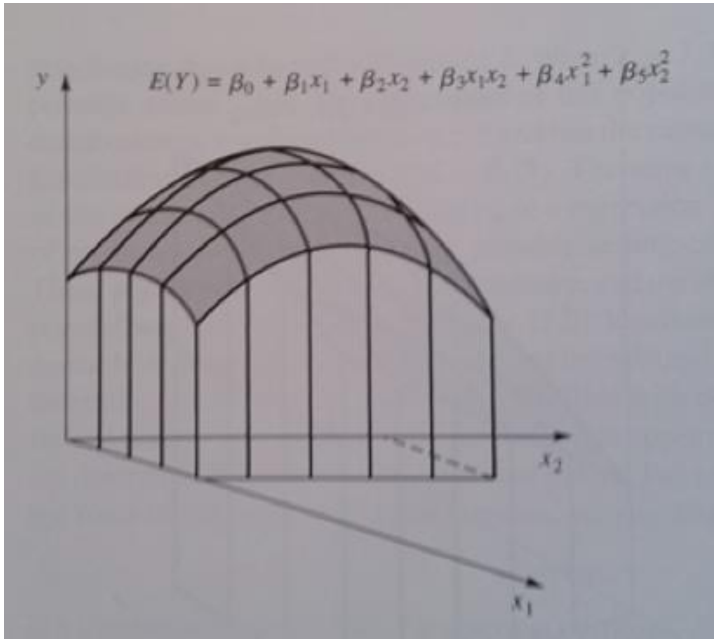
(Chapter 5)

polynomial with interaction

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \epsilon$$

(Chapter 5)

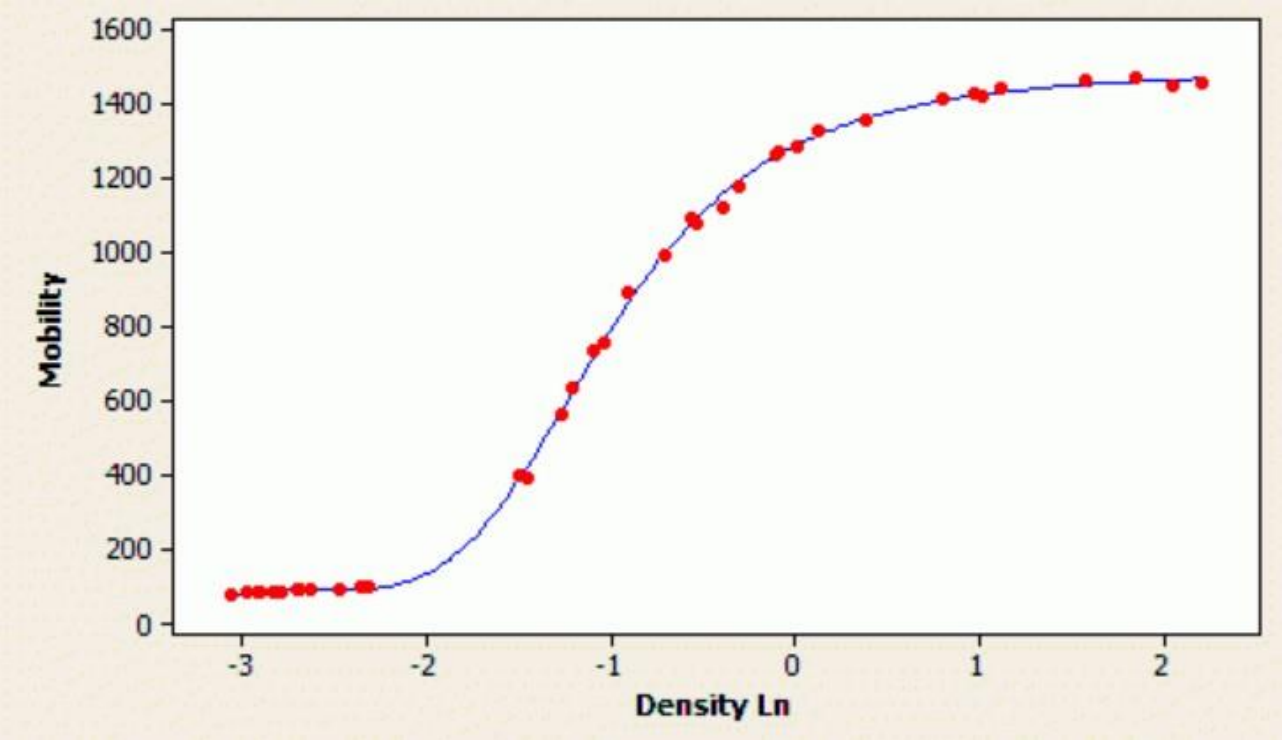




https://www.datatechnotes.com/2018/02/polynomial-regression-curve-fitting-in-r.html

Example of Nonlinear Regression

$$Y = \frac{\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3}{\beta_4 + \beta_5 x + \beta_6 x^2 + \beta_7 x^3}$$ (not covered in STAT210/410)



https://statisticsbyjim.com/regression/difference-between-linear-nonlinear-regression-models/