Chapter 2
Linear Regression Modelling

In this section

- Textbook Chapters 11 & 12
- Straight-line regression
- Polynomial regression
- Multiple regression
- Factor predictors
- The Linear Model
- Model selection and checking

Linear regression Modelling

In the previous section, we attempted to explain the variation in an observed response variable by fitting models with one or more factors.

Factors are discrete variables taking different levels; in this section we will now utilize continuous variables that can similarly explain variation in an observed response.

Simple Linear Regression

We will investigate models relating two quantities x and y through equations of the form

$$y = ax + b$$

where a and b are constants (that is, a straight line).

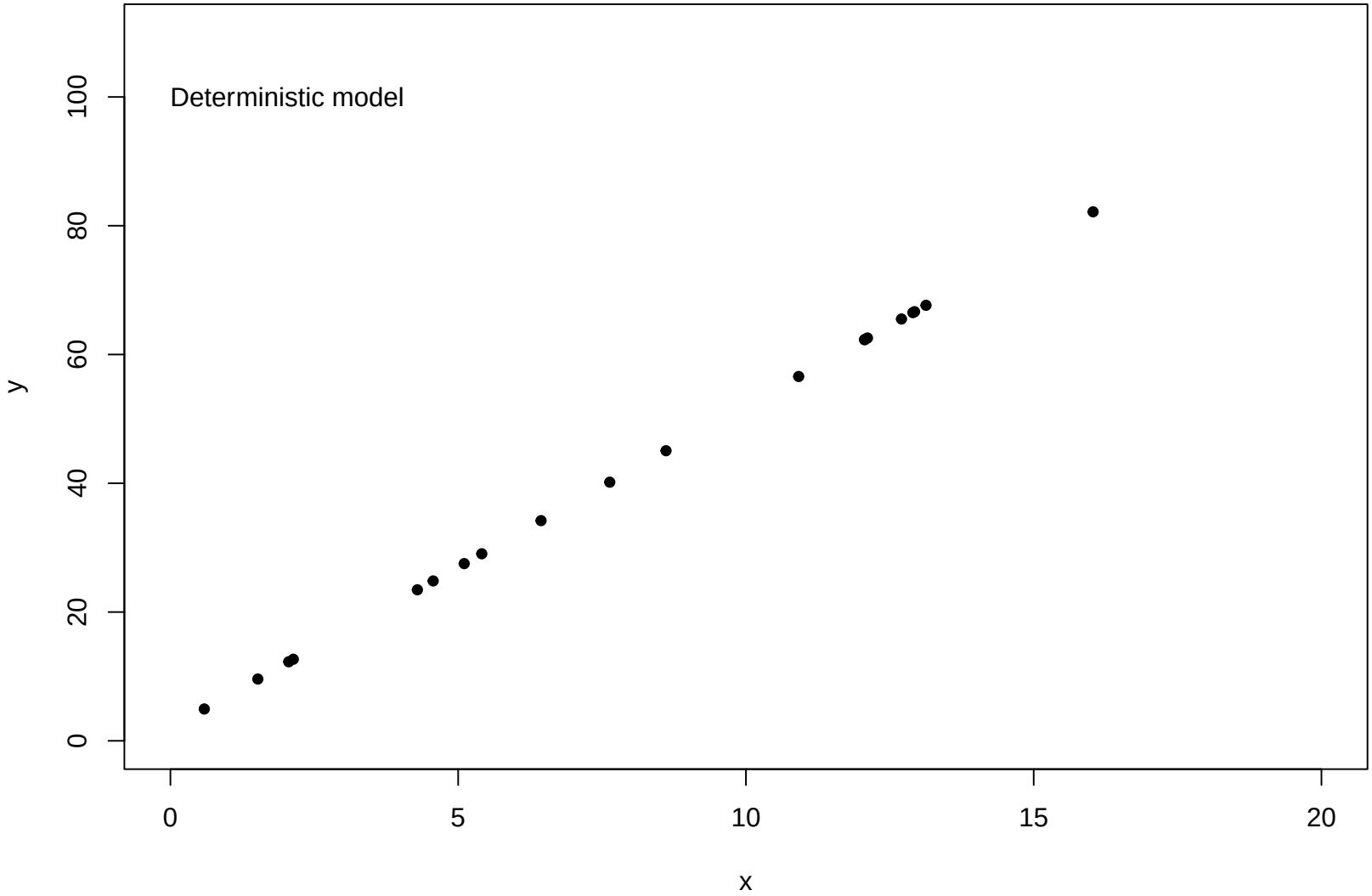
Variables x and y will not be treated exchangeably - we will regard y as being a function of x.

Simple Linear Regression

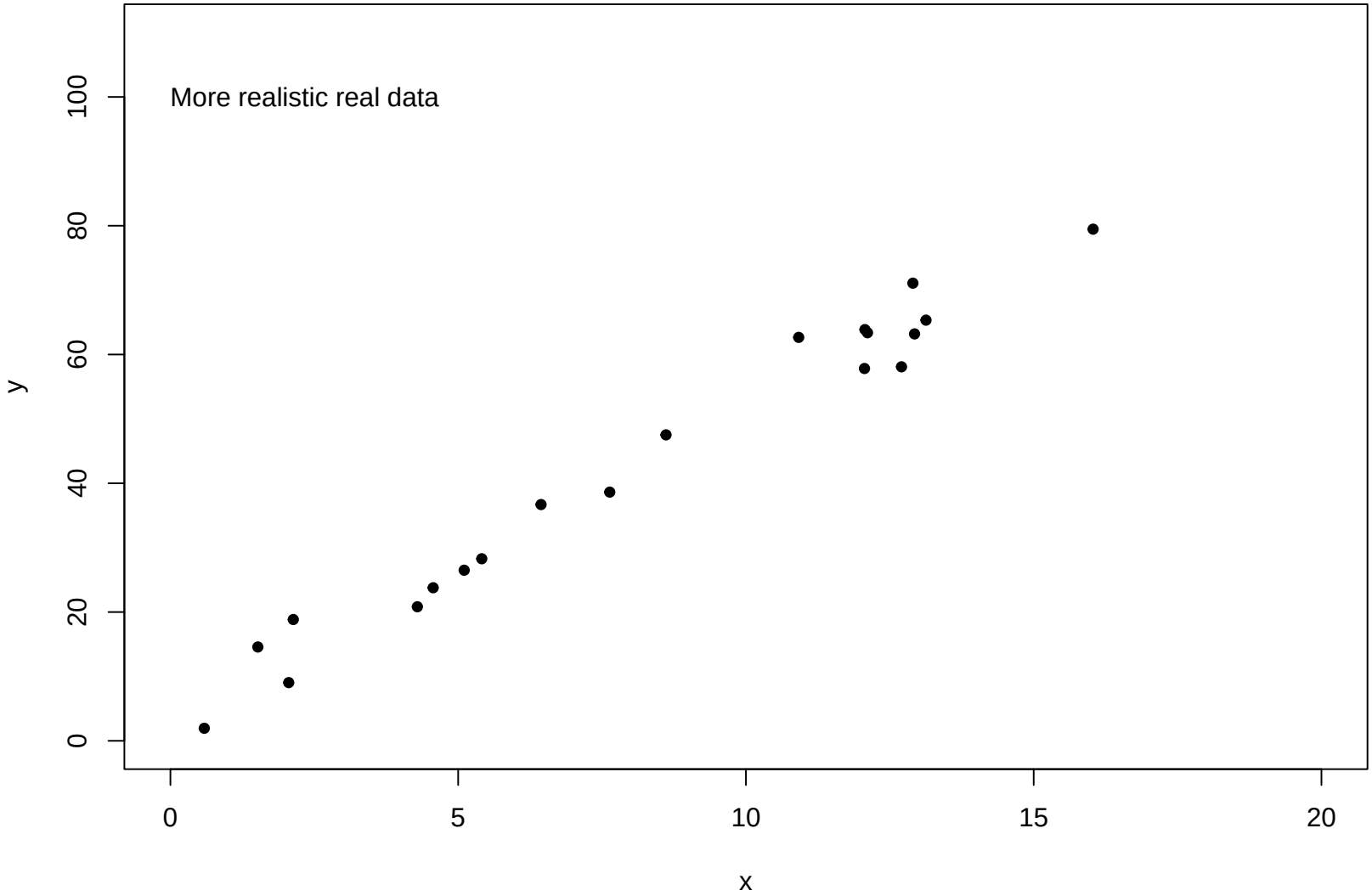
Such models are deterministic, that is, if we know x (and the values of the constants), we can compute y exactly without error.

A more useful model allows for the possibility that the system is not observed perfectly, that is, we do not observe (x, y) pairs that are always consistent with a simple functional relationship.

Simple Linear Regression



Simple Linear Regression



Simple Linear Regression

Example: Pharmacokinetic Model

If a dose of drug is taken at time $x = 0$, the concentration of drug still in the bloodstream at time x is often well-modelled by a simple equation.

Let

- D denote the amount of drug taken at $x = 0$
- x time
- $y^\star$ is the concentration in the bloodstream.

Simple Linear Regression

Example: Pharmacokinetic Model

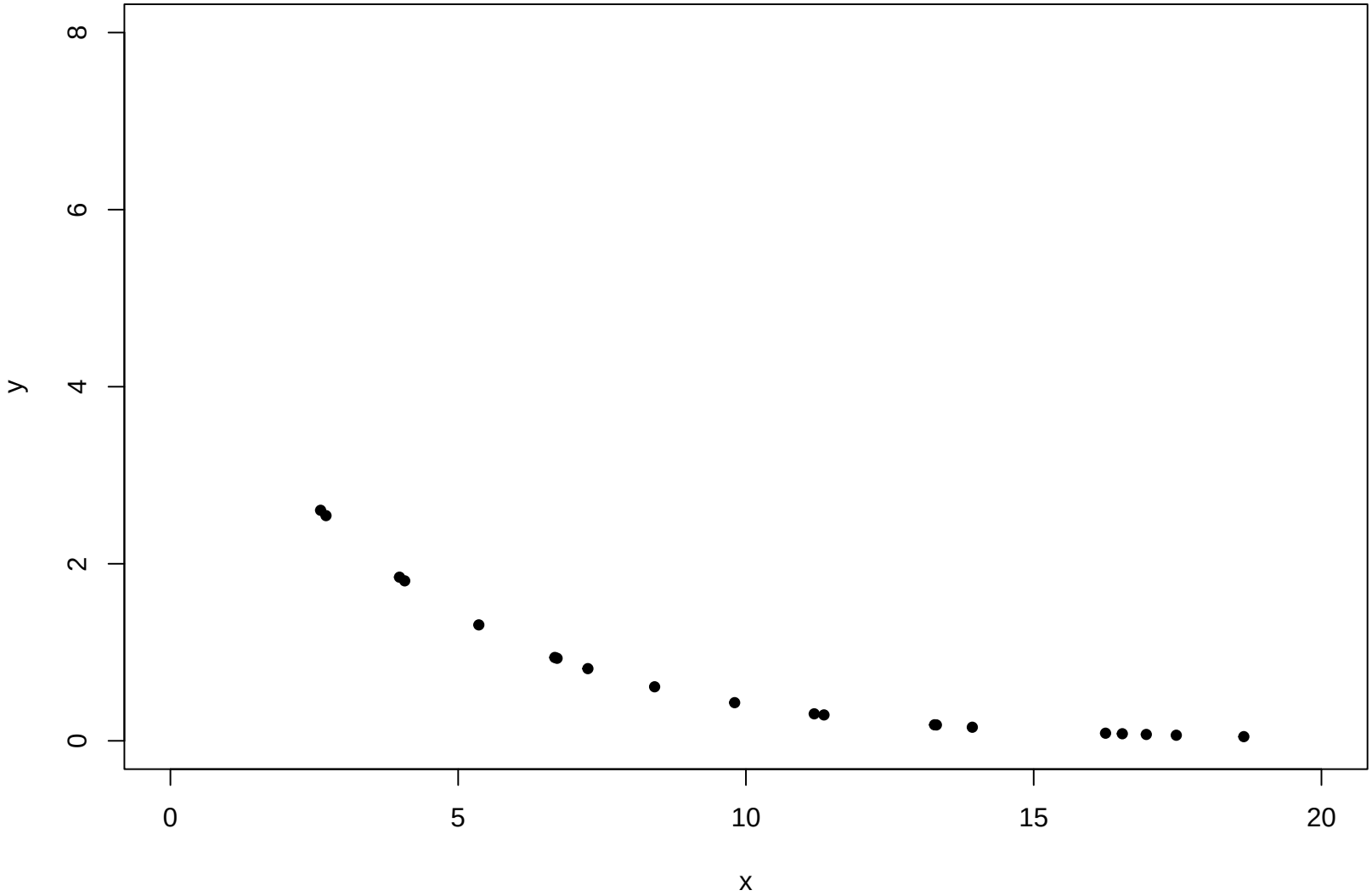
Then

$$y^\star = \frac{D}{V} \exp\{-\lambda x\}$$

where

- λ is the elimination rate
- V is the volume of bloodstream.

Simple Linear Regression



Simple Linear Regression

Example: Pharmacokinetic Model

Taking logs of both sides, setting $y = \log y^\star$, then

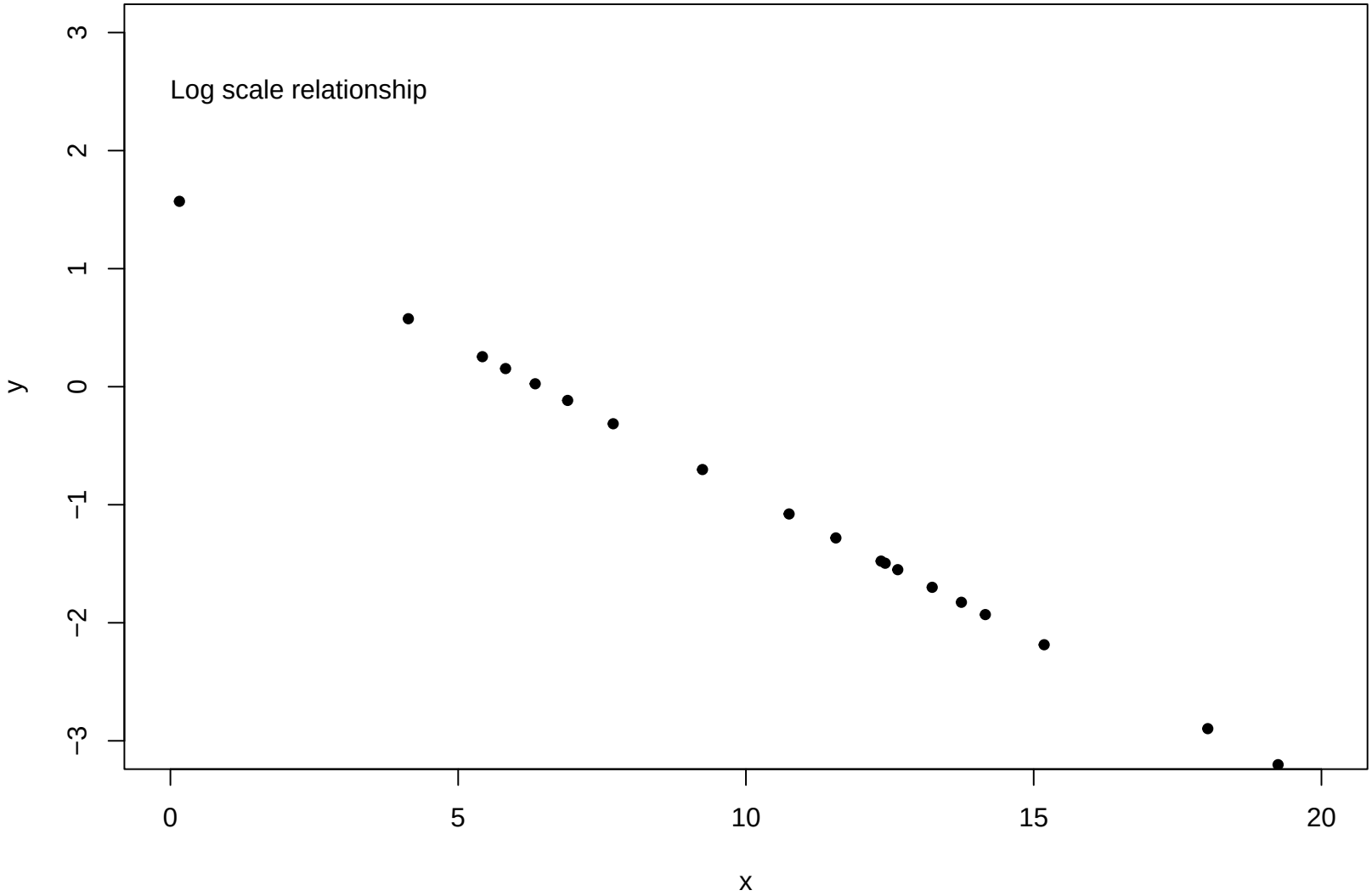
$$y = -\lambda x + \log(D/V) = -\lambda x + (\log D - \log V)$$

that is, $y = ax + b$ where

- $a = -\lambda$
- $b = (\log D - \log V)$

Here $\log$ means natural log $\ln$.

Simple Linear Regression

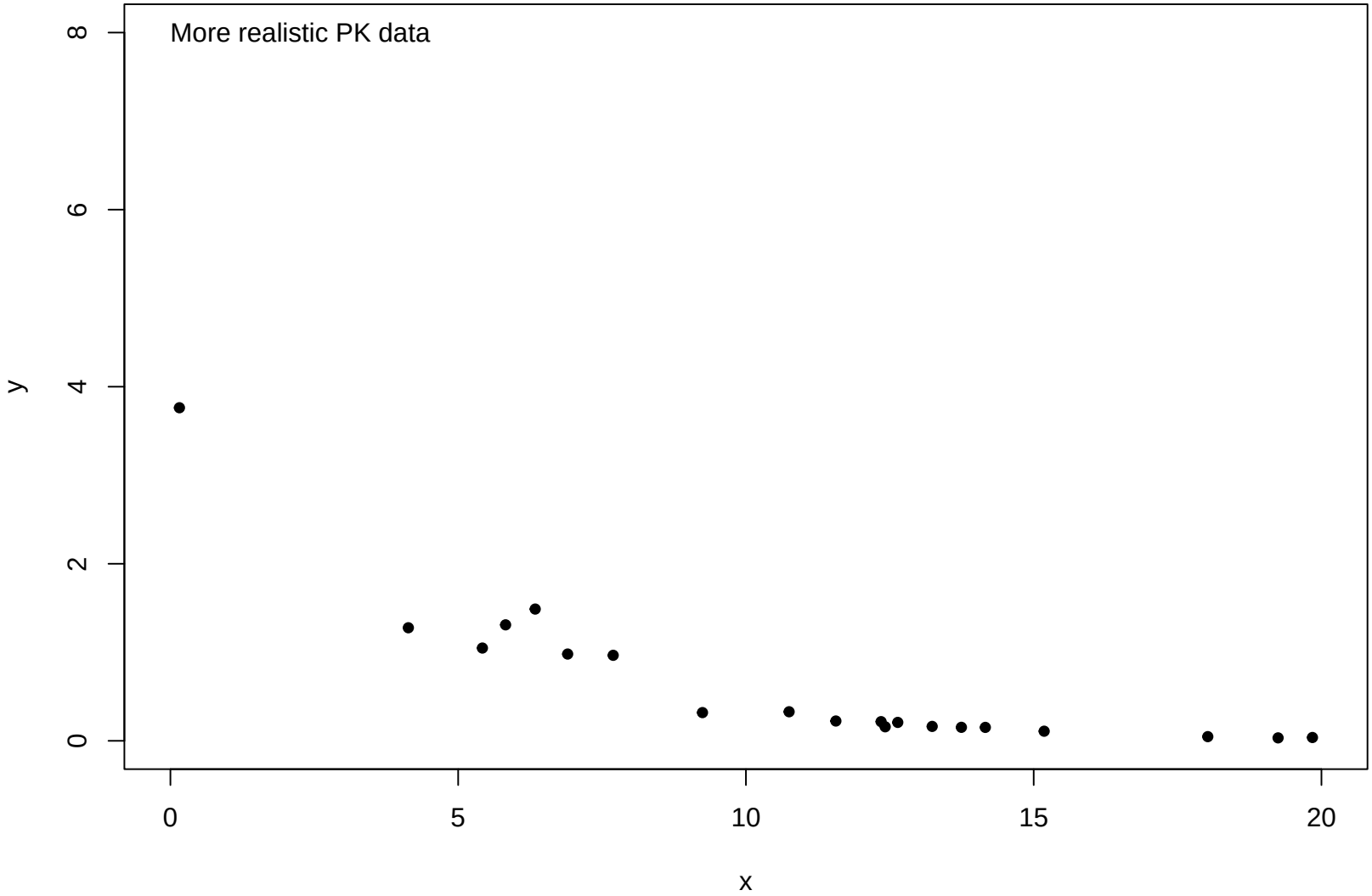


Simple Linear Regression

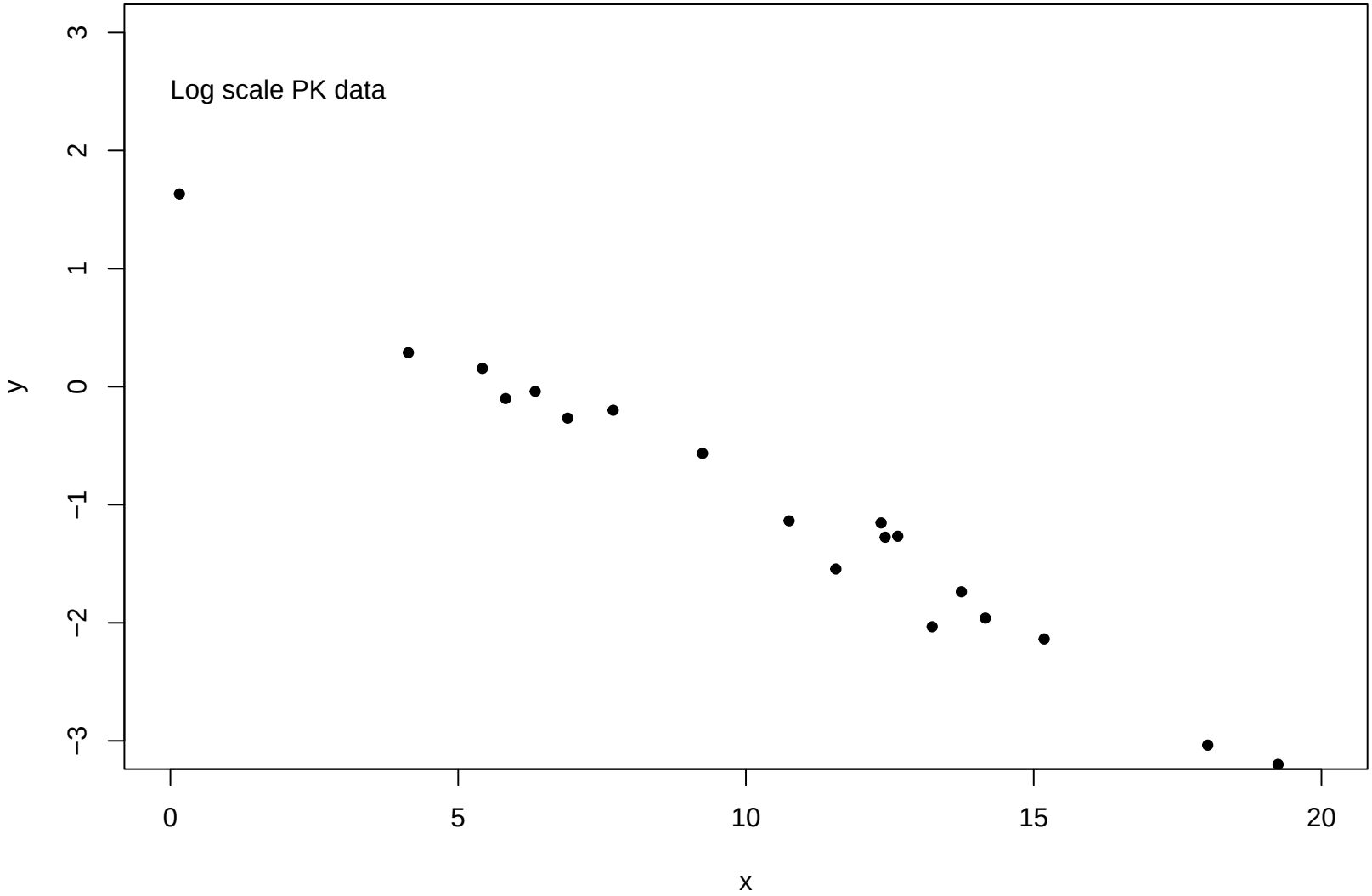
However, in practice, when we measure concentration, we do so with random error.

- We need to acknowledge the presence of random error in the measurement process
- Deterministic models do not permit this

Simple Linear Regression



Simple Linear Regression



Probabilistic Models

In a probabilistic model, we allow for the possibility that y is observed with random error, that is,

$$y = ax + b + ERROR$$

where $ERROR$ is a random term that is present due to imperfect observation of the system due to measurement error or incomplete information.

Probabilistic Models

Note that we do not treat x and y interchangeably;

- x is a fixed quantity that is measured without error
- y is measured with random error.

A Basic Probabilistic Model

We model the variation in y as a function of x. We observe data pairs $(x_i, y_i), i = 1, \ldots, n$.

Terminology:

- y - Dependent variable or response variable
- x - Independent variable, or predictor, or covariate

A Basic Probabilistic Model

The model we study takes the form

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

where ϵ is a random error term, specifically

- a random variable with mean zero and finite variance σ^2
- represents the error present in the measurement of the response.

Parameters:

- β_0 - Intercept parameter
- β_1 - Slope parameter

A Basic Probabilistic Model

For example

- $\beta_1 > 0$ - increasing y with increasing x
- $\beta_1 < 0$ - decreasing y with increasing x
- $\beta_1 = 0$ - no relationship between x and y

Note:

$$\mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x}$$

where $\mathbb{E}[Y|x]$ is the expected value of Y for a fixed value of x.

A Basic Probabilistic Model

Recall the notation

- Y - a random variable with a probability distribution
- y - a fixed observed value of the variable Y.

A Basic Probabilistic Model

Fundamental Problem: If we believe the straight-line model with error is correct, how do we find the values of parameters β_0 and β_1.

We only have the observed data $\{(x_i, y_i), i = 1, \ldots, n\}$.

Least Squares Fitting

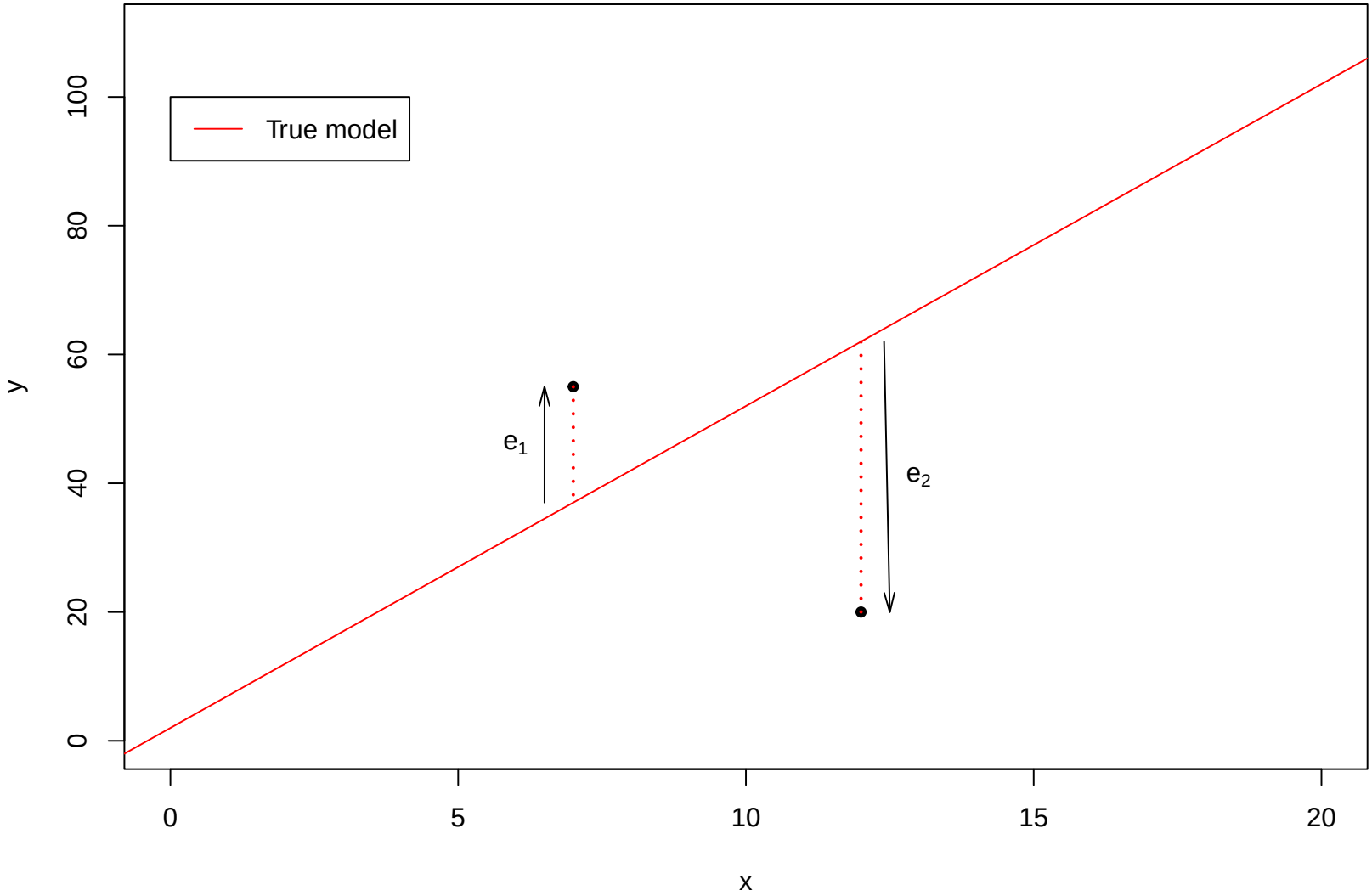
We select the best values of β_0 and β_1 by minimizing the error in fit. For two data pairs (x_1, y_1) and (x_2, y_2), the errors in fit are

$$\begin{aligned} e_1 &= y_1 - (\beta_0 + \beta_1 x_1) \\ e_2 &= y_2 - (\beta_0 + \beta_1 x_2) \end{aligned}$$

respectively.

With two points, we can interpolate the data with a straight line.

Least Squares Fitting



Least Squares Fitting

It is reasonable to choose the best parameters by considering the aggregate errors across the data set: we might inspect

$$\sum_{i=1}^{2} e_i = e_1 + e_2$$

and seek to minimize this by manipulating (β_0, β_1).

But note that, potentially, $e_1 > 0$ and $e_2 < 0$ so there is a possibility that these fitting errors cancel each other out.

Least Squares Fitting

Therefore we look at squared errors (as a large negative error is as bad as a large positive error)

$$
\begin{aligned}
e_1^2 &= (y_1 - (\beta_0 + \beta_1 x_1))^2 \\
e_2^2 &= (y_2 - (\beta_0 + \beta_1 x_2))^2
\end{aligned}
$$

Least Squares Fitting

For n data, we obtain n misfit squared errors

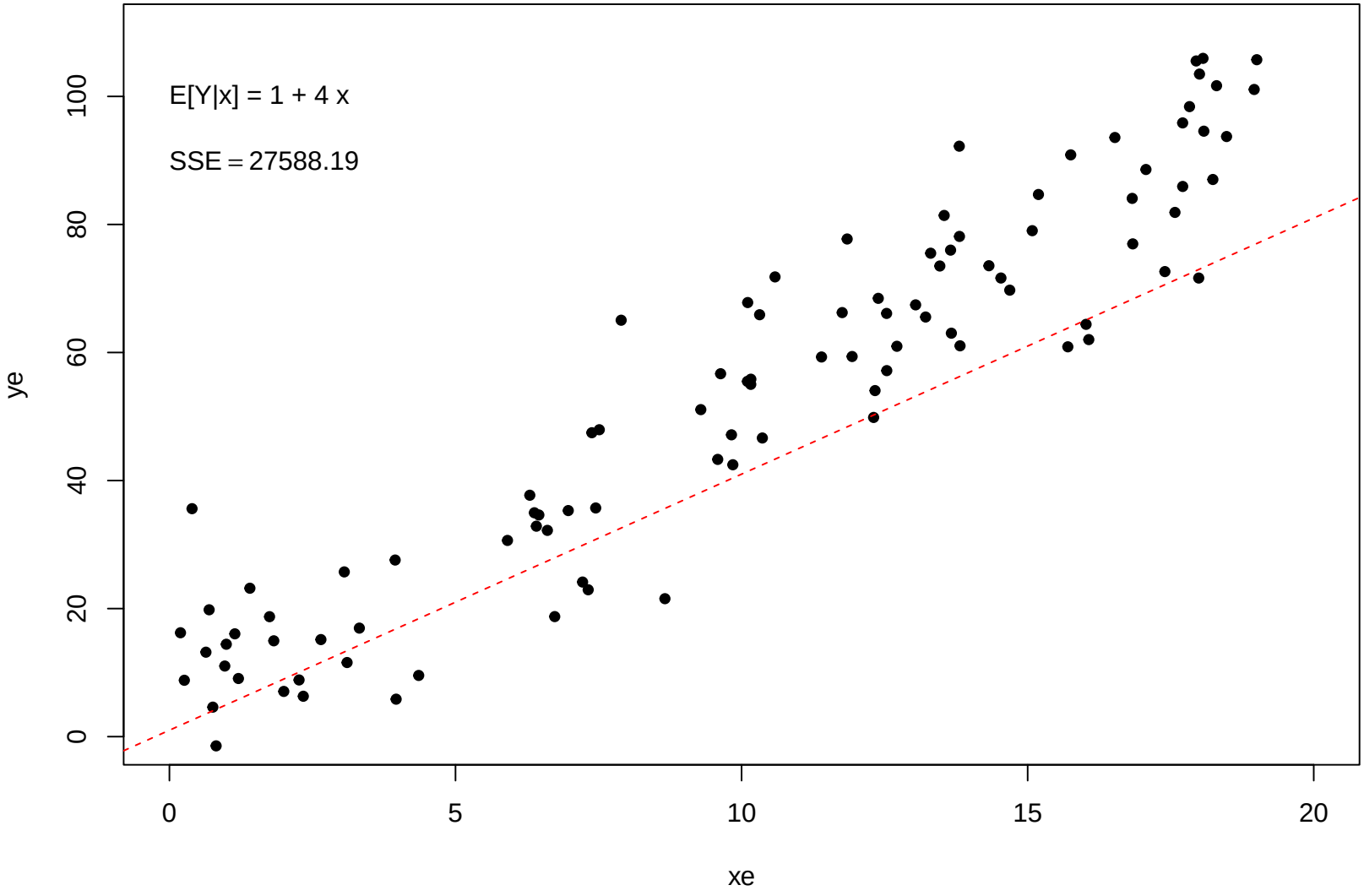
$$e_1^2, \ldots, e_n^2$$

We select β_0 and β_1 as the values of the parameters that minimize SSE, where

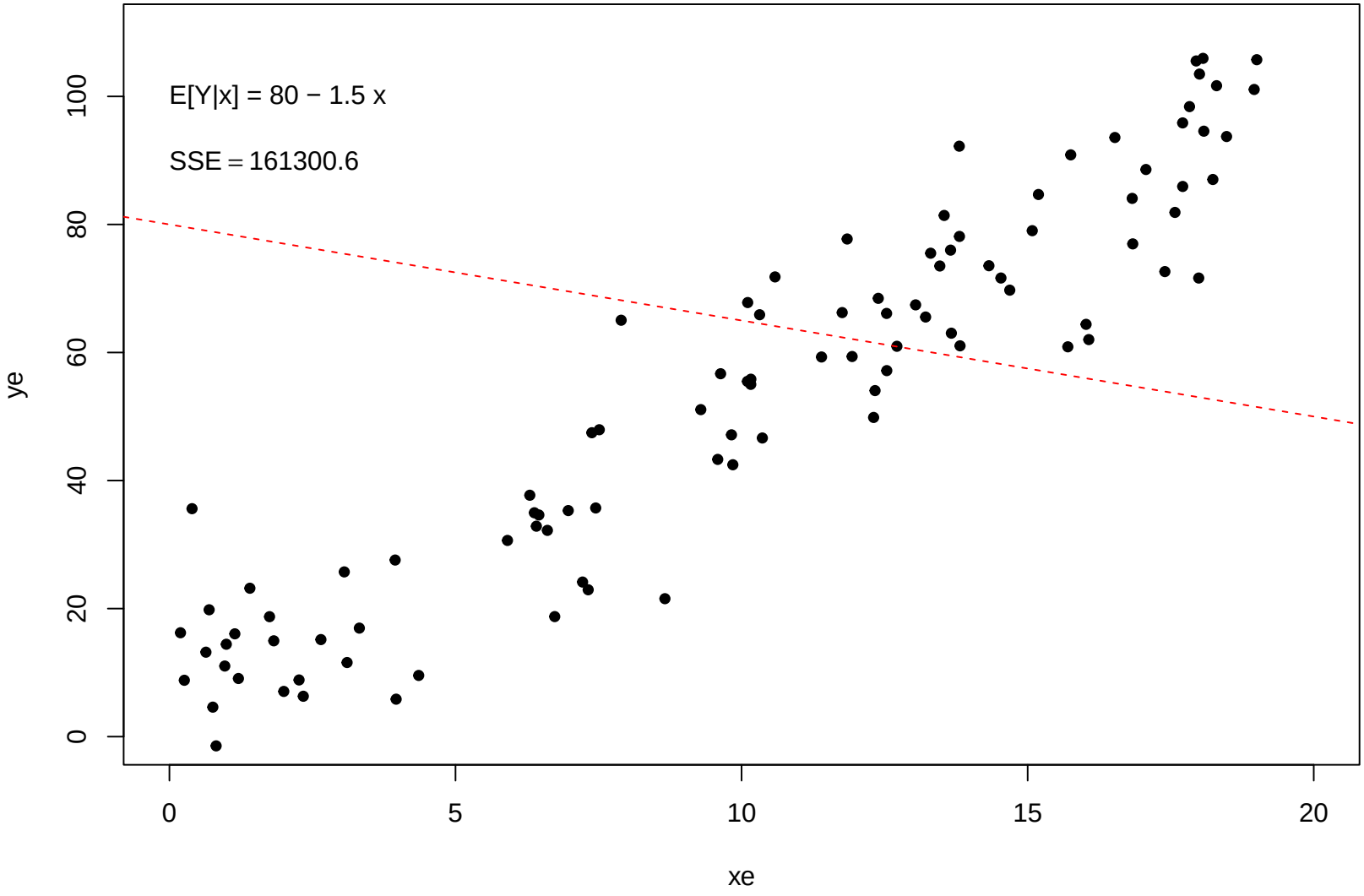
$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

We wish to make the total misfit squared error as small as possible.

Least Squares Fitting



Least Squares Fitting



Least Squares Fitting

SSE - sum of squared errors - is similar to the SSE for ANOVA. We could write

$$SSE = SSE(\beta_0, \beta_1)$$

to show the dependence of SSE on the parameters.

Minimization of $SSE(\beta_0, \beta_1)$ is achieved analytically using calculus.

Least Squares Fitting

It follows that the best-fit values $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are given by

$$\widehat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \qquad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

where

- SS_{xx}:

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

- SS_{xy}:

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

Least Squares Fitting

$\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the least-squares estimates

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

is the least-squares line of best fit. The fitted-values are

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

and the residuals are

$$\widehat{e}_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

Estimating σ^2

Using the LS procedure, we can construct an estimate of the error or residual error variance

Recall that the variance of ϵ is σ^2, and an estimate of σ^2 is

$$\widehat{\sigma}^2 = \frac{SSE(\widehat{\beta}_0, \widehat{\beta}_1)}{n-2} = s^2$$

say.

Estimating σ^2

Note that the denominator $n-2$ is again a degrees of freedom parameter of the form

(Number of data points) - (Number of model parameters)

or $n-p$. Here $p=2$ (β_0 and β_1).

Note also that

$$SSE(\widehat{\beta}_0, \widehat{\beta}_1) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = SS_{yy} - \widehat{\beta}_1 SS_{xy}$$

where

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

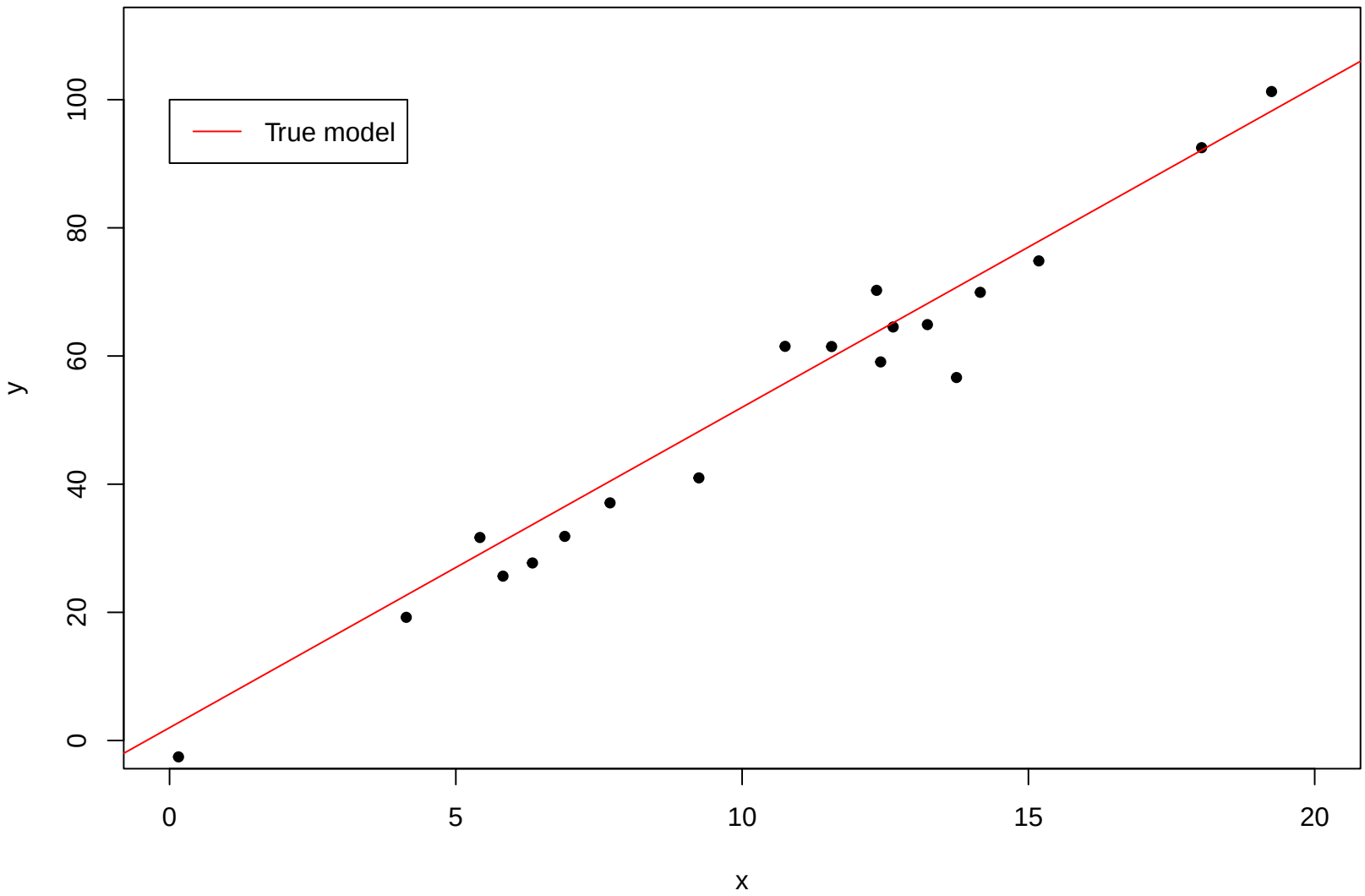
Example

Simulated example: here $n = 20$ and

$$\beta_0 = 2 \qquad \beta_1 = 5 \qquad \sigma^2 = 6^2$$

```
> set.seed(1234)
> n<-20
> y<-2+5*x+rnorm(n)*6
> par(mar=c(4,4,2,2))
> plot(x,y,pch=19,cex=0.8,xlim=range(0,20), ylim=range(0,110))
> abline(2,5,col='red')
> legend(0,100,c('True model'),col='red',lty=1)
```

Example



Example

Estimation:

```
> SS.xx<-sum((x-mean(x))^2)
> SS.xy<-sum((x-mean(x))*(y-mean(y)))
> b1.hat<-SS.xy/SS.xx
> b0.hat<-mean(y)-b1.hat*mean(x)
> sigsq.hat<-sum((y - b0.hat - b1.hat*x)^2)/(n-2)
> c(b0.hat,b1.hat,sigsq.hat)

+ [1] -4.851720  5.488668 32.190393
```

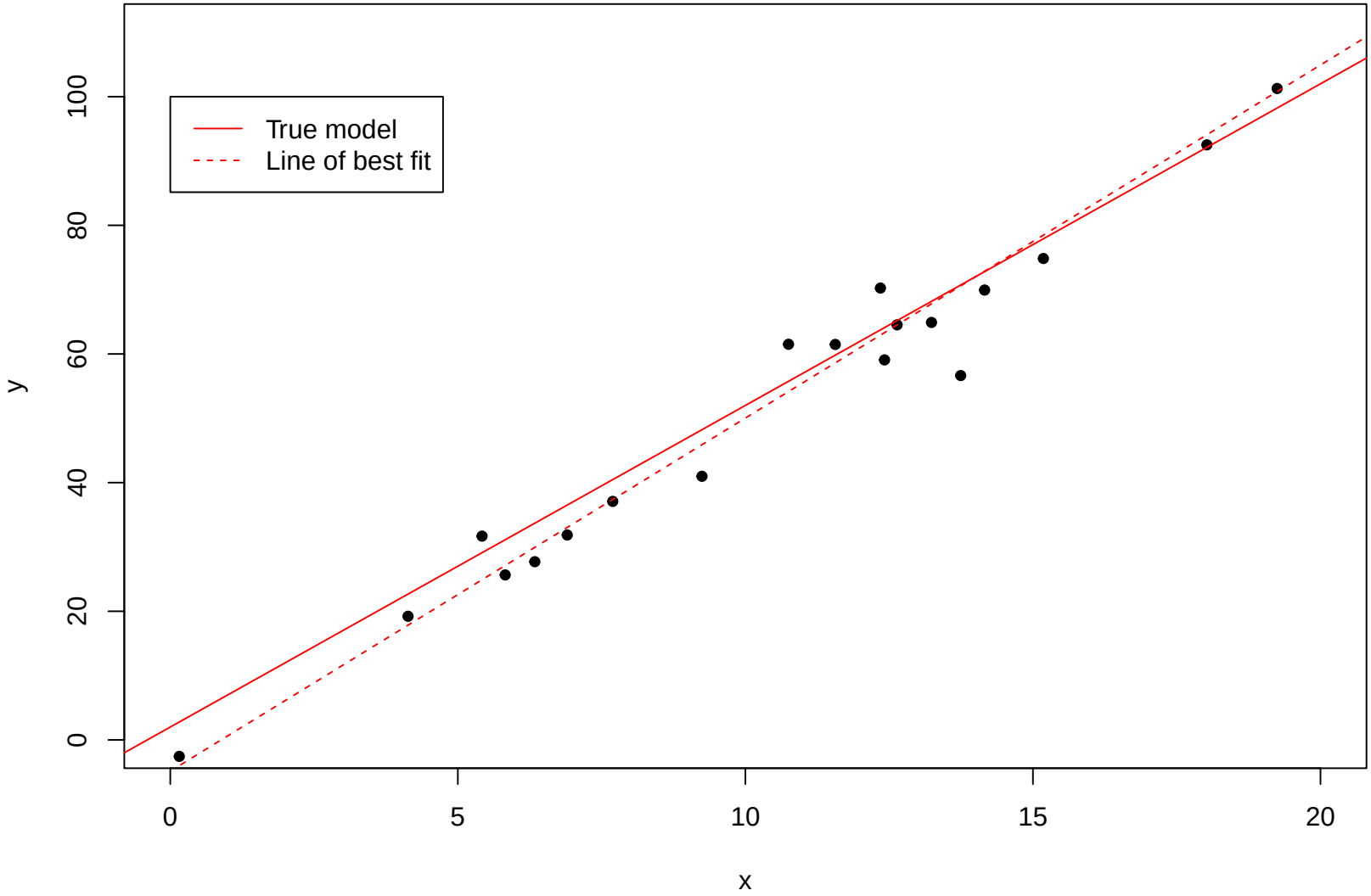
That is

$$\widehat{\beta}_0 = -4.8517 \qquad \widehat{\beta}_1 = 5.4887 \qquad \widehat{\sigma}^2 = 32.1904$$

so that

$$\widehat{\sigma} = 5.6737.$$

Example



Example

Using the `lm` function

```
> fit1<-lm(y~x)
> summary(fit1)

+
+ Call:
+ lm(formula = y ~ x)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -13.9386  -2.9048  -0.7366   1.7927  11.6499
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept)  -4.8517     3.0097  -1.612    0.124
+ x             5.4887     0.2494  22.008 1.84e-14 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 5.674 on 18 degrees of freedom
+ Multiple R-squared:  0.9642, Adjusted R-squared:  0.9622
+ F-statistic: 484.4 on 1 and 18 DF,  p-value: 1.835e-14
```

Model Assumptions for Least-Squares

To utilize least-squares (LS) for the probabilistic model

$$Y = \beta_0 + \beta_1 \mathbf{x} + \epsilon$$

we make the following four assumptions

1. The expected value of ϵ is zero so that

$$\mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x}.$$

2. The variance of ϵ, denoted σ^2, is constant and does not depend on x.

Model Assumptions for Least-Squares

3. The probability distribution of ϵ is a symmetric distribution about zero
   - a stronger assumption is that $\epsilon \sim Normal(0, \sigma^2)$.
4. The errors for two different measured responses at x_1 and x_2 are independent
   - the error ϵ_1 in measuring y_1 at x_1 is independent of the error ϵ_2 in measuring y_2 at x_2.

Hypothesis Testing

In the model where

$$\mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x}$$

it is of interest to test the hypothesis

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_a &: \beta_1 \neq 0 \end{aligned}$$

H_0 implies that there is no systematic influence of x on the value of Y. This is a two-sided alternative.

Hypothesis Testing

To test H_0 vs H_a we us the test statistic

$$T = \frac{\widehat{\beta}_1}{\text{e.s.e}(\widehat{\beta}_1)}$$

where $\text{e.s.e}(\widehat{\beta}_1)$ is the Estimated Standard Error of $\widehat{\beta}_1$, computed as

$$\text{e.s.e}(\widehat{\beta}_1) = \frac{s}{\sqrt{SS_{xx}}}$$

where s is the estimate of σ defined previously.

Hypothesis Testing

If H_0 is true, and in fact $\beta_1 = 0$, then

$$T = \frac{\widehat{\beta}_1}{s/\sqrt{SS_{xx}}} \sim t_{n-2}$$

that is, the t-distribution with $n - 2$ degrees of freedom.

We can carry out a significance test at level α in the usual way (use a p-value, or construct the rejection region).

Hypothesis Testing

We might also consider a one-sided test:

- If $H_a : \beta_1 \neq 0$, we use the two-sided rejection region, with critical values
$$C_R = \pm t_{\alpha/2,n-2}$$
- If $H_a : \beta_1 > 0$, we use the one-sided rejection region, with critical value
$$C_R = t_{\alpha,n-2}$$
- If $H_a : \beta_1 < 0$, we use the one-sided rejection region, with critical value
$$C_R = -t_{\alpha,n-2}$$

Hypothesis Testing

To test

$$\begin{aligned} H_0 &: \beta_1 = b \\ H_a &: \beta_1 \neq b \end{aligned}$$

for any b, the test statistic is

$$T = \frac{\widehat{\beta}_1 - b}{s/\sqrt{SS_{xx}}}$$

(for example, $b = 1$ may be of interest). If H_0 is true

$$T \sim t_{n-2}.$$

Hypothesis Testing

`lm` computes the key quantities. Using the `summary` function

```
> summary(fit1)$coef

+              Estimate Std. Error   t value     Pr(>|t|)
+ (Intercept) -4.851720  3.0096733 -1.612042 1.243479e-01
+ x            5.488668  0.2493914 22.008246 1.835225e-14
```

- first column: estimates
- second column: estimated standard errors
- third column: test statistic where null hypothesis is that parameter is zero
- fourth column: two-sided p-value for the test

Confidence Interval

A $100(1-\alpha)\%$ confidence interval for β_1 is

$$\widehat{\beta}_1 \pm t_{\alpha/2,n-2} \times \text{e.s.e.}(\widehat{\beta}_1)$$

where

$$\begin{aligned} t_{\alpha/2,n-2} &: \text{ upper } \alpha/2 \text{ point of } t_{n-2} \text{ distn} \\ \text{e.s.e.}(\widehat{\beta}_1) &: \text{ estimated standard error of } \widehat{\beta}_1 \end{aligned}$$

Confidence Interval

Note: we could perform a similar analysis for β_0, but this is generally of less interest.

The only quantity that needs attention is the estimated standard error of $\widehat{\beta}_0$. It can be shown that

$$\text{e.s.e.}(\widehat{\beta}_0) = \sqrt{\frac{1}{n}\left(1 + \frac{n\overline{x}^2}{SS_{xx}}\right)}$$

Confidence Interval

The `confint` function computes the confidence intervals:

```
> confint(fit1)

+                  2.5 %   97.5 %
+ (Intercept) -11.174810 1.471369
+ x             4.964716 6.012620
```

That is, the 95% CIs for β_0 and β_1 are

$$(-11.174810, 1.471369) \qquad \text{and} \qquad (4.964716, 6.012620)$$

respectively.

The Coefficient of Correlation

To measure the strength of association between the two variables x and y we can use the

Pearson Product Moment Coefficient Of Correlation

or correlation coefficient which measures the strength of the linear relationship between x and y.

The Coefficient of Correlation

The correlation coefficient, r, is defined by

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

where

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 \qquad SS_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

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

The Coefficient of Correlation

In R the `cor` function can be used

```
> cor(x,y)

+ [1] 0.9819212

> SS.yy<-sum((y-mean(y))^2)
>
> SS.xy/sqrt(SS.xx*SS.yy)

+ [1] 0.9819212
```

The Coefficient of Correlation

Note: $-1 \leqslant r \leqslant 1$.

- If r is close to 1, there is a strong linear relationship between x and y where y increases with x.
- If r is close to -1, there is a strong linear relationship between x and y where y decreases with x.

Note: In the model

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

$\beta_1 = 0 \implies r \approx 0$, so tests for $\beta_1 = 0$ can also be used to deduce a lack of correlation between the variables.

The Coefficient of Correlation

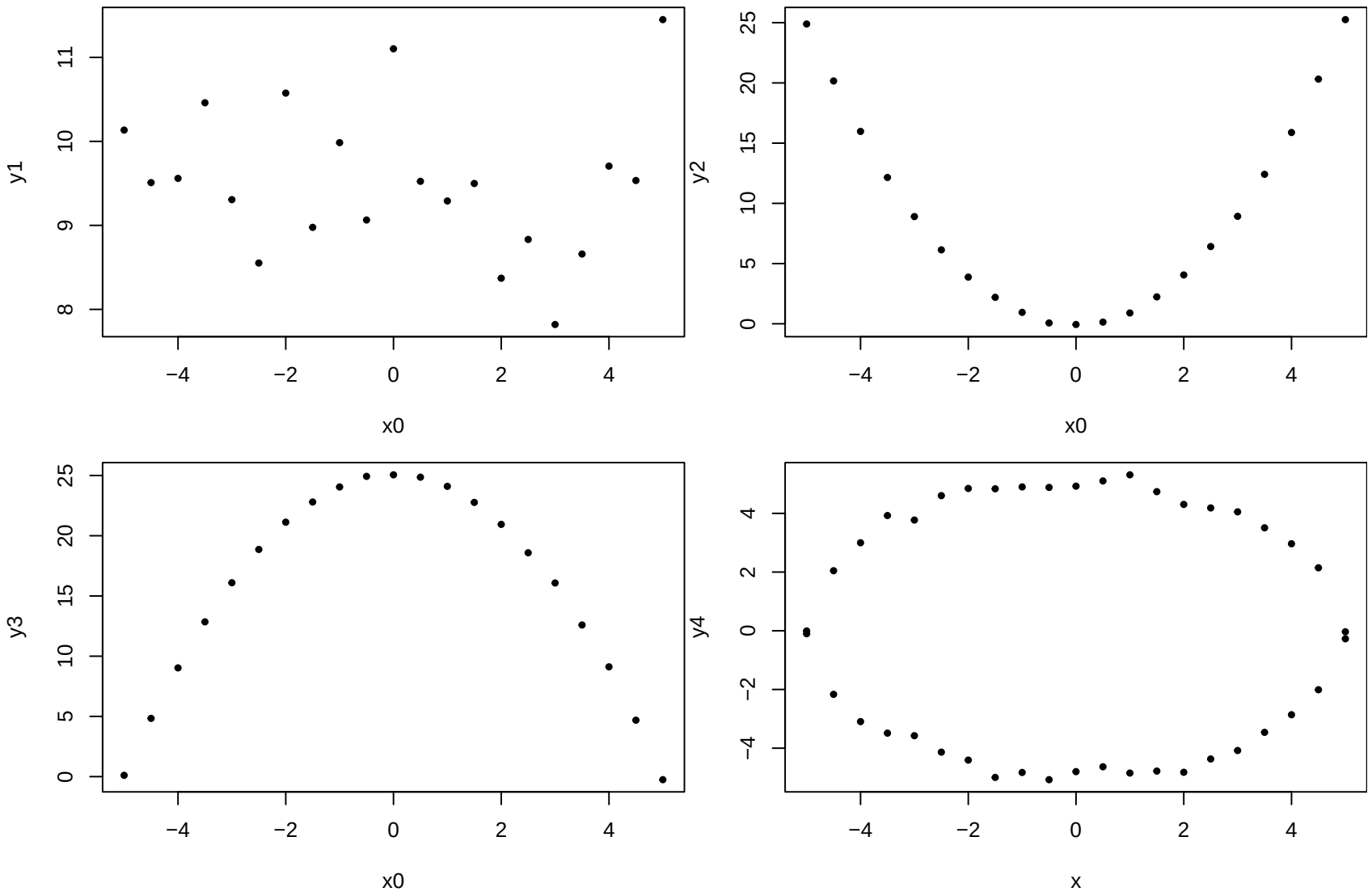
Notes:

1. A strong linear relationship is not necessarily a causal relationship, that is, just because $r \approx 1$ does not mean that x causes changes in y (we may have a spurious correlation).

2. Just because $r \approx 0$ does not mean that that x and y are unrelated, merely that they are uncorrelated.

   That is, it is possible to construct examples where x and y have a strong functional relationship, but where $r = 0$.

The Coefficient of Correlation



The Coefficient of Correlation

All of these data sets have $r \approx 0$.

- top left: no systematic relationship
- top right: almost deterministic parabolic relationship
- bottom left: almost deterministic parabolic relationship
- bottom right: data lie on a circle

Testing Correlation

We use ρ to denote the true correlation between X and Y.

We can test the hypothesis that $\rho = 0$ (that is, that X and Y are uncorrelated) using the statistic r. For testing

$$\begin{aligned} H_0 &: \rho = 0 \\ H_a &: \rho \neq 0 \end{aligned}$$

we can use the test statistic

$$T = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}}$$

If H_0 is true, then approximately

$$T \sim t_{n-2}.$$

Testing Correlation

Alternately, we could use

$$Z = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right)$$

and then, if H_0 is true, we have that approximately

$$Z \sim N \left(\frac{1}{2} \log \left(\frac{1+\rho}{1-\rho} \right), \frac{1}{n-3} \right)$$

when $\rho = 0$, so that (approximately)

$$\sqrt{n-3}\, Z \sim Normal(0,1)$$

R^2 statistic

A related quantity is the

Coefficient of Determination

or R^2 statistic

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

R^2 statistic

Note that the total variation in y is recorded via

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

and the random variation is recorded via

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

R^2 statistic

Therefore the variation explained by the linear regression is

$$SSR = SS_{yy} - SSE \qquad \text{as} \qquad SS_{yy} = SSR + SSE$$

Thus

$$r^2 = \frac{SSR}{SS_{yy}} = \frac{\text{Variation explained by Regression}}{\text{Total Variation}}$$

r^2 is a measure of model adequacy, that is, if $r^2 \approx 1$, then the linear model is a good fit.

R^2 statistic

In `lm`, the quantity reported as `Multiple R-squared` is the R^2 statistic.

```
> summary(fit1)


+
+ Call:
+ lm(formula = y ~ x)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -13.9386  -2.9048  -0.7366   1.7927  11.6499
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept)  -4.8517     3.0097  -1.612    0.124
+ x             5.4887     0.2494  22.008 1.84e-14 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 5.674 on 18 degrees of freedom
+ Multiple R-squared:  0.9642, Adjusted R-squared:  0.9622
+ F-statistic: 484.4 on 1 and 18 DF,  p-value: 1.835e-14
```

R^2 statistic

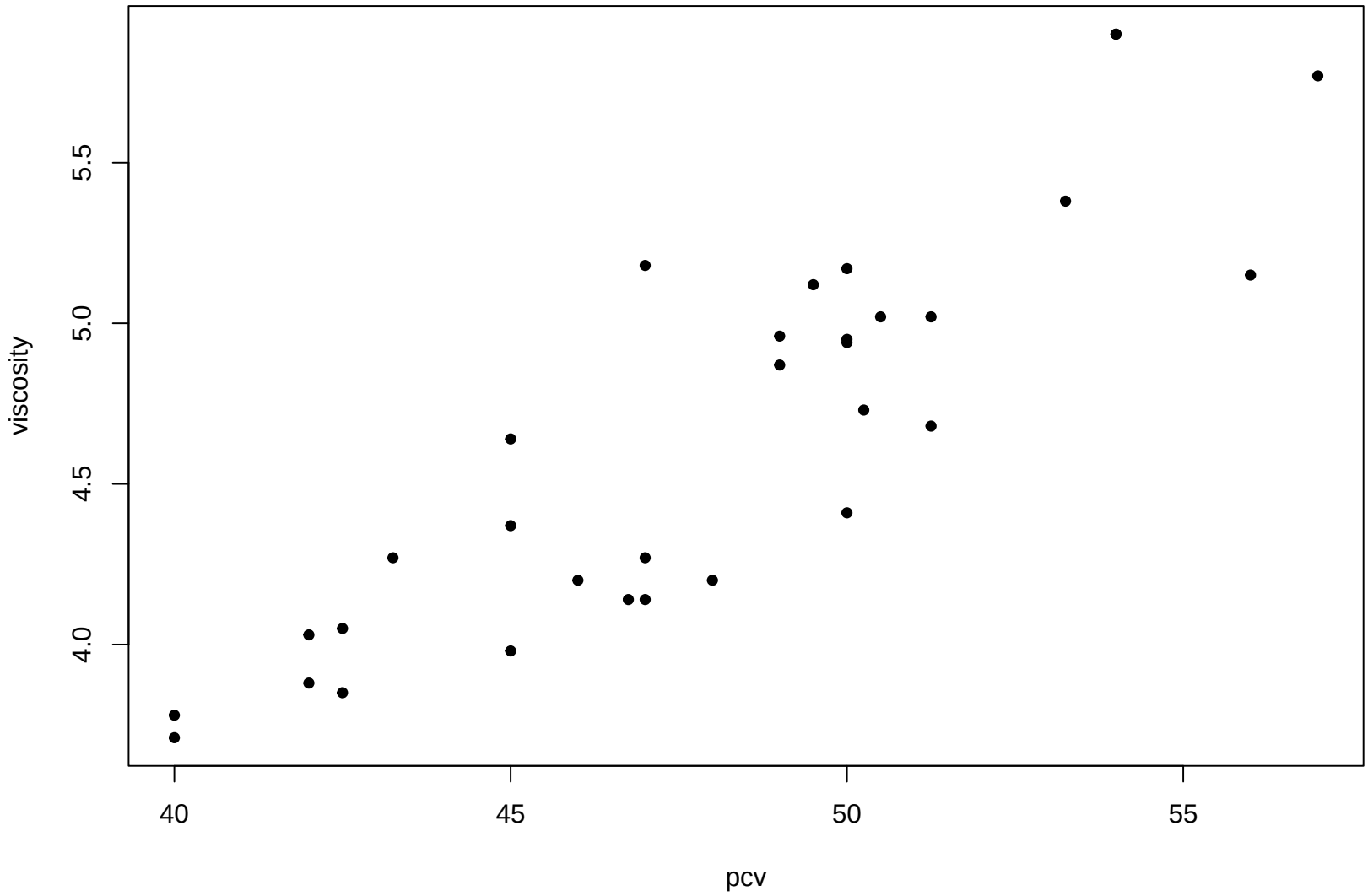
```
> cor(x,y)^2

+ [1] 0.9641693

> summary(fit1)$r.squared

+ [1] 0.9641693
```

R^2 statistic

Example: Blood viscosity vs PCV

```
> load('204DataSets.rda')
> head(Viscosity)

+   Patient viscosity  pcv plasfib plaspro
+ 1       1      3.71 40.0     344    6.27
+ 2       2      3.78 40.0     330    4.86
+ 3       3      3.85 42.5     280    5.09
+ 4       4      3.88 42.0     418    6.79
+ 5       5      3.98 45.0     774    6.40
+ 6       6      4.03 42.0     388    5.48

> par(mar=c(4,4,2,2))
> plot(viscosity~pcv,data=Viscosity,pch=19,cex=0.8)
```

R^2 statistic



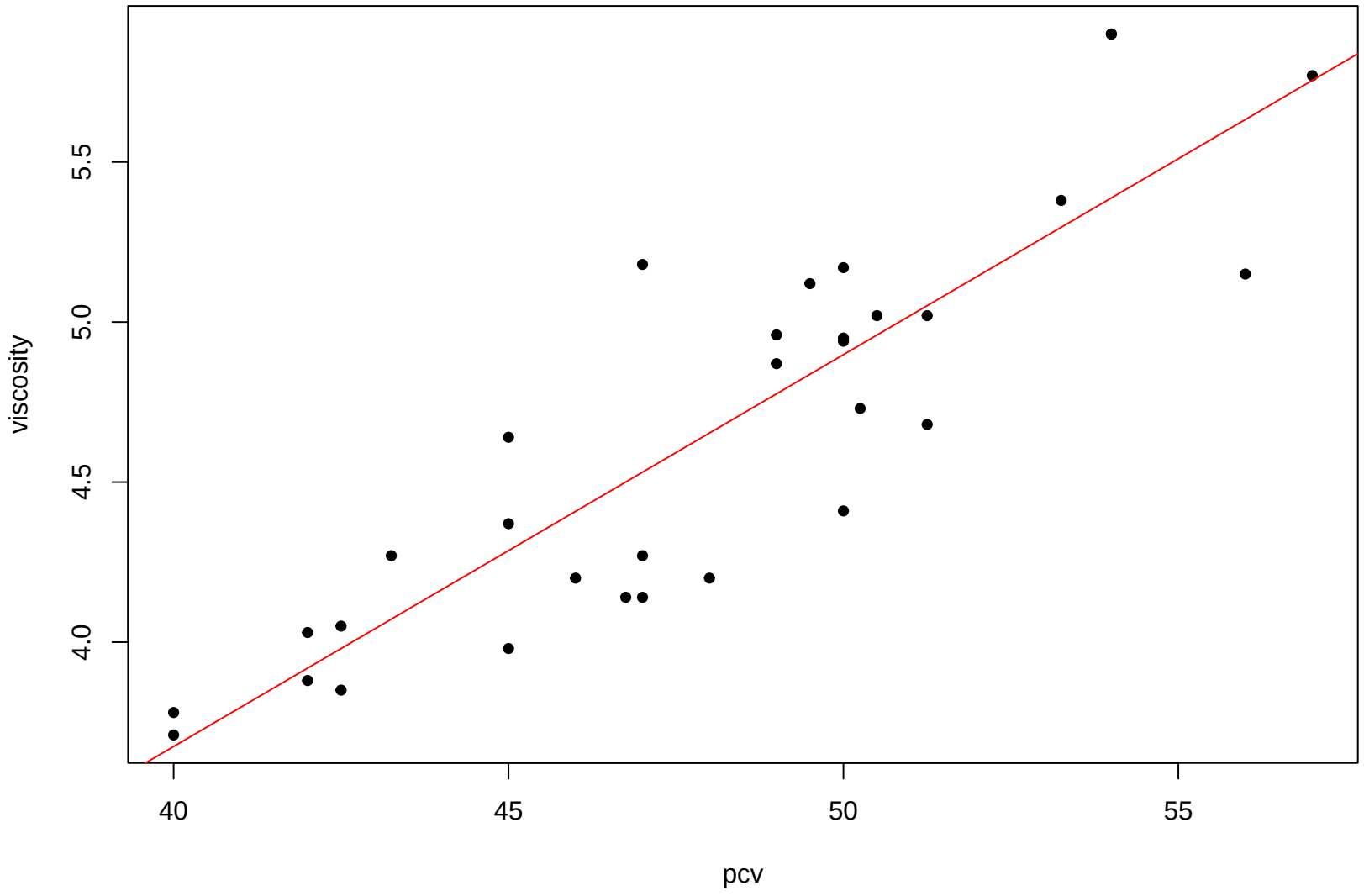
R^2 statistic

```
> fit.vis<-lm(viscosity~pcv,data=Viscosity)
> summary(fit.vis)

+
+ Call:
+ lm(formula = viscosity ~ pcv, data = Viscosity)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.48814 -0.22152  0.04686  0.12955  0.64915
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept) -1.22336    0.58422  -2.094   0.0448 *
+ pcv          0.12243    0.01214  10.088 3.73e-11 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.3012 on 30 degrees of freedom
+ Multiple R-squared:  0.7723, Adjusted R-squared:  0.7647
+ F-statistic: 101.8 on 1 and 30 DF,  p-value: 3.731e-11
```

R^2 statistic

```
> cor(Viscosity$pcv,Viscosity$viscosity)

+ [1] 0.878818

> cor(Viscosity$pcv,Viscosity$viscosity)^2

+ [1] 0.7723211

> par(mar=c(4,4,2,2))
> plot(viscosity~pcv,data=Viscosity,pch=19,cex=0.8)
> abline(coef(fit.vis),col='red')
```

R^2 statistic



R^2 statistic

Example: Blood Viscosity vs PCV

We have

- $n = 32$
- $r = 0.879$
- $R^2 = r^2 = (0.879)^2 = 0.772$

Test of $\rho = 0$:

$$T_{obs} = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}} = 10.087$$

We compare with a $t_{n-2} \equiv t_{30}$ distribution; the p-value is 3.73×10^{-11}, so there is strong evidence that $\rho \neq 0$.

Prediction

After the straight-line model is fitted, it can be used for forecasting or prediction. That is, given a new x value we can predict the corresponding y.

At any value x_p, the prediction $\widehat{y}_p$ is

$$\widehat{y}_p = \widehat{\beta}_0 + \widehat{\beta}_1 x_p$$

This is the best predictor of Y at this x value.

Prediction

We can also compute the standard error of this prediction; if the value of the random error variance σ^2 is known, then

$$\text{s.e.}(\widehat{y}_p) = \sigma\sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

If σ is unknown, we estimate σ by $\widehat{\sigma} = s$ as defined previously

$$s^2 = \frac{SSE(\widehat{\beta}_0, \widehat{\beta}_1)}{n-2}$$

so that

$$\text{e.s.e.}(\widehat{y}_p) = s\sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}}$$

Prediction

This prediction is the expected value of y at $x = x_p$. That is, we have worked out

$$Var[\widehat{Y}_p | x_p] = Var[\widehat{\beta}_0 + \widehat{\beta}_1 x_p | x_p]$$

to compute the s.e. for $\widehat{Y}_p$.

But we can actually predict an error corrupted version of $\widehat{Y}_p$, $\widehat{Y}_p^\star$ say, where

$$\widehat{Y}_p^\star = \widehat{Y}_p + \epsilon_p$$

where ϵ_p is a new random error.

Prediction

But

$$
\begin{aligned}
Var[\widehat{Y}_p^\star | x_p] &= Var[\widehat{Y}_p + \epsilon_p | x_p] \\
&= Var[\widehat{Y}_p | x_p] + Var[\epsilon_p | x_p] \\
&= Var[\widehat{Y}_p | x_p] + \sigma^2
\end{aligned}
$$

that is, there is an extra piece of variation due to ϵ_p.

Thus

$$
\text{e.s.e.}(\widehat{y}_p^\star) = s\sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{SS_{xx}}} > \text{e.s.e.}(\widehat{y}_p)
$$

Prediction Intervals

A $100(1-\alpha)\%$ prediction interval for the mean value at $x = x_p$ is

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

whereas for an individual new value (predicted with error) at $x = x_p$ is

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

Prediction Intervals

Using the predict function:

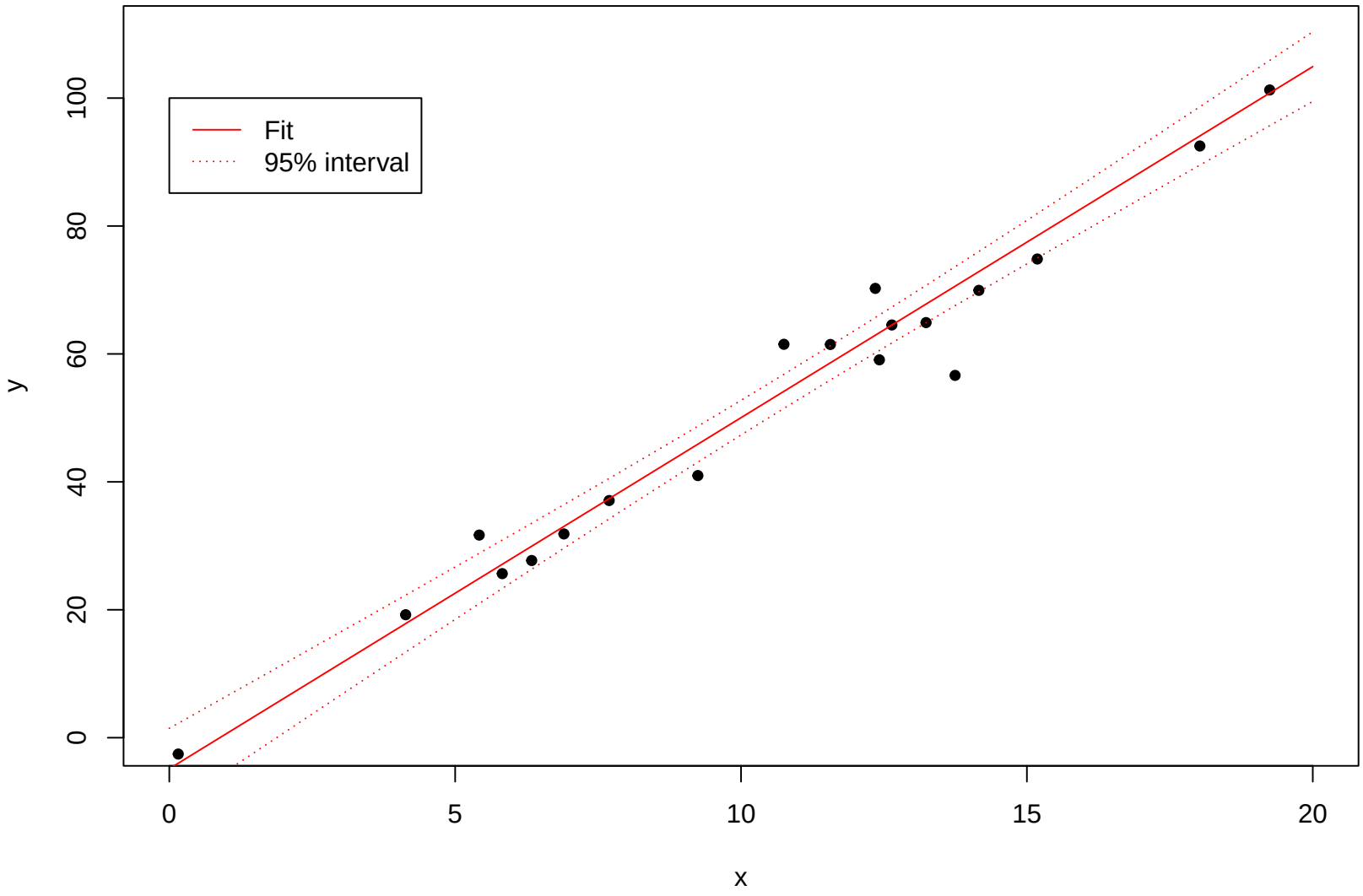
```
> xp<-seq(0,20,by=0.01)
> int.mean<-predict(fit1,newdata=data.frame(x=xp),interval='confidence', se.fit=TRUE)
> int.obs<-predict(fit1,newdata=data.frame(x=xp),interval='prediction', se.fit=TRUE)
> names(int.mean)

+ [1] "fit"            "se.fit"         "df"             "residual.scale"

> head(int.mean$fit)

+         fit       lwr      upr
+ 1 -4.851720 -11.17481 1.471369
+ 2 -4.796834 -11.11517 1.521504
+ 3 -4.741947 -11.05554 1.571641
+ 4 -4.687060 -10.99590 1.621778
+ 5 -4.632174 -10.93626 1.671916
+ 6 -4.577287 -10.87663 1.722055

> head(int.mean$se.fit)

+        1        2        3        4        5        6
+ 3.009673 3.007412 3.005151 3.002890 3.000630 2.998370
```

Prediction Intervals

Prediction interval for mean:

```
> par(mar=c(4,4,2,2))
> plot(x,y,pch=19,cex=0.8,xlim=range(0,20), ylim=range(0,110))
> lines(xp,int.mean$fit[,1],col='red')
> lines(xp,int.mean$fit[,2],col='red',lty=3)    #Lower boundary of interval
> lines(xp,int.mean$fit[,3],col='red',lty=3)    #Upper boundary of interval
> legend(0,100,c('Fit','95% interval'),col='red',lty=c(1,3))
```

Prediction Intervals

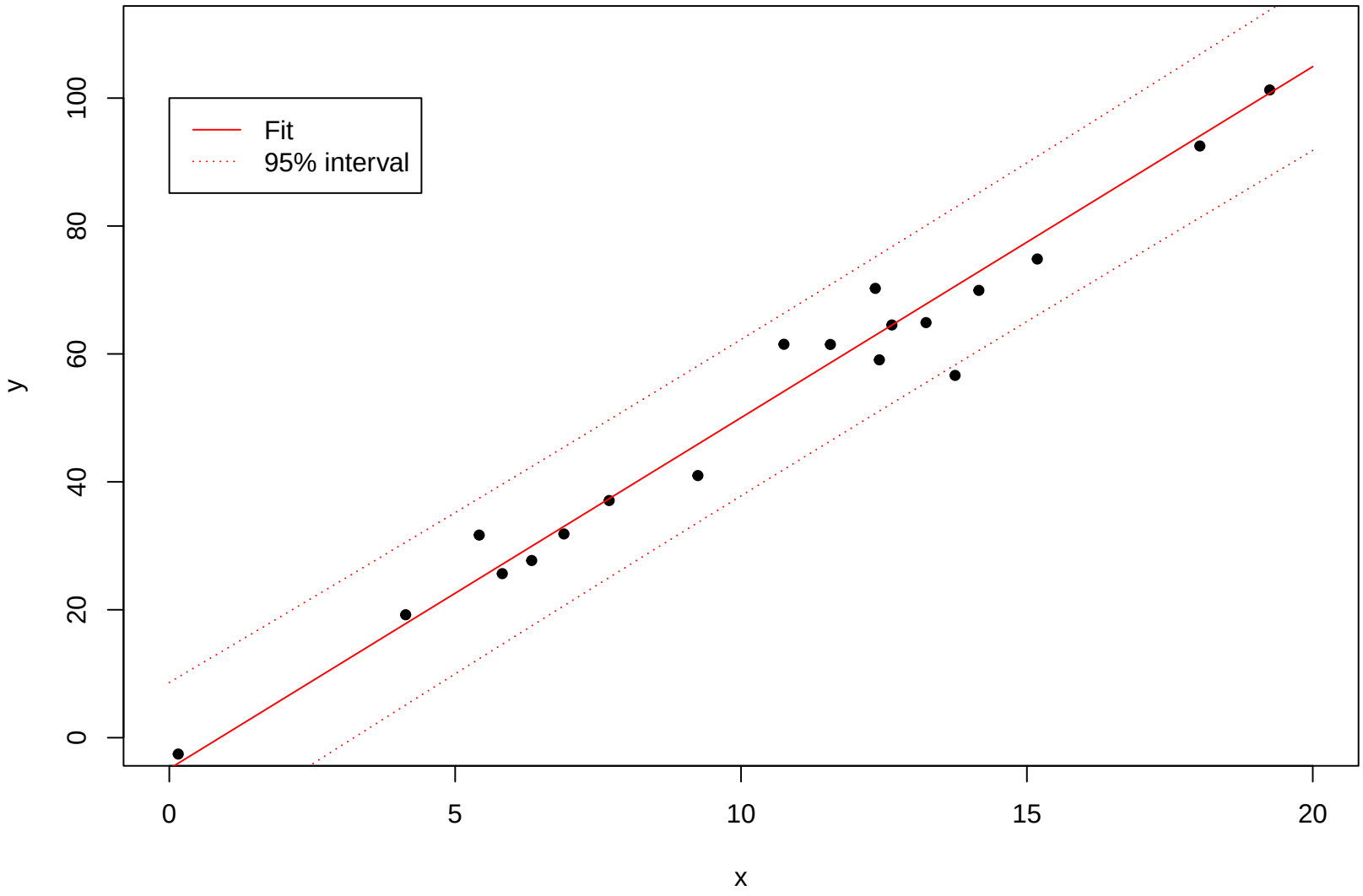


Prediction Intervals

Prediction interval for individual observation:

```
> par(mar=c(4,4,2,2))
> plot(x,y,pch=19,cex=0.8,xlim=range(0,20), ylim=range(0,110))
> lines(xp,int.obs$fit[,1],col='red')
> lines(xp,int.obs$fit[,2],col='red',lty=3)    #Lower boundary of interval
> lines(xp,int.obs$fit[,3],col='red',lty=3)    #Upper boundary of interval
> legend(0,100,c('Fit','95% interval'),col='red',lty=c(1,3))
```

Prediction Intervals



ANOVA F-test in Regression

An ANOVA F-test can be constructed to test overall (global) fit of the linear regression model.

The decomposition of sums of squares for regression takes the form

$$SS = SSR + SSE$$

where

- $SS = SS_{yy}$: overall or total sum of squares
- SSR: sum of squares due to Regression
- SSE: sum of squares due to Error

ANOVA F-test in Regression

$$
\begin{aligned}
SS &= \sum_{i=1}^{n} (y_i - \overline{y})^2 \\
SSR &= \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 \\
SSE &= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2
\end{aligned}
$$

where

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \qquad i = 1, \ldots, n$$

ANOVA F-test in Regression

Degrees of Freedom

- TOTAL: $n - 1$
- REGRESSION: 1
  - there is one parameter that explains the influence of x on y: β_1.
- ERROR: $n - 2$
  - error d.f. is $n - p$
  - here $p = 2$ records the total number of parameters in the mean model, that is, β_0 and β_1.

The ANOVA Table

SOURCE	DF	SS	MS	F
REGRESSION	1	SSR	$MSR = \dfrac{SSR}{1}$	$F = \dfrac{MSR}{MSE}$
ERROR	$n-2$	SSE	$MSE = \dfrac{SSE}{(n-2)}$	
TOTAL	$n-1$	SS		

The ANOVA Table

The test of the hypothesis

$$\begin{aligned} H_0 &: \mathbb{E}[Y|\mathbf{x}] = \beta_0 \\ H_a &: \mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x} \end{aligned}$$

can be completed by using the test statistic

$$F = \frac{MSR}{MSE}$$

If H_0 is true

$$F \sim F_{1,n-2}.$$

The ANOVA Table

```
> summary(aov(lm(y~x)))

+             Df Sum Sq Mean Sq F value   Pr(>F)
+ x            1  15592   15592   484.4 1.84e-14 ***
+ Residuals   18    579      32
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> library(car)
> Anova(lm(y~x))

+ Anova Table (Type II tests)
+
+ Response: y
+            Sum Sq Df F value    Pr(>F)
+ x         15591.8  1  484.36 1.835e-14 ***
+ Residuals   579.4 18
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA Table

```
> library(supernova)
> supernova(lm(y~x))

+  Analysis of Variance Table (Type III SS)
+  Model: y ~ x
+
+                                SS df        MS       F   PRE     p
+  ----- --------------- | --------- -- --------- ------- ----- -----
+  Model (error reduced) | 15591.832  1 15591.832 484.363 .9642 .0000
+  Error (from model)    |   579.427 18    32.190
+  ----- --------------- | --------- -- --------- ------- ----- -----
+  Total (empty model)   | 16171.259 19   851.119

> supernova(lm(y~x), type=2)

+  Analysis of Variance Table (Type II SS)
+  Model: y ~ x
+
+                                SS df        MS       F   PRE     p
+  ----- --------------- | --------- -- --------- ------- ----- -----
+  Model (error reduced) | 15591.832  1 15591.832 484.363 .9642 .0000
+  Error (from model)    |   579.427 18    32.190
+  ----- --------------- | --------- -- --------- ------- ----- -----
+  Total (empty model)   | 16171.259 19   851.119
```

The ANOVA Table

This is just like the ANOVA in the one-way layout (CRD) with n groups, but where

$$\mu_i = \beta_0 + \beta_1 \mathbf{x}_i$$

That is, the group means are structured, that is, we have a formula relating the μ_i quantities.

The ANOVA Table

Consider $r = 3$ replicates at x each of the values $(1, 2, 3, 4)$ in a regression, and say

$$Y_{ij} = 2 + x_i + \epsilon_{ij}$$

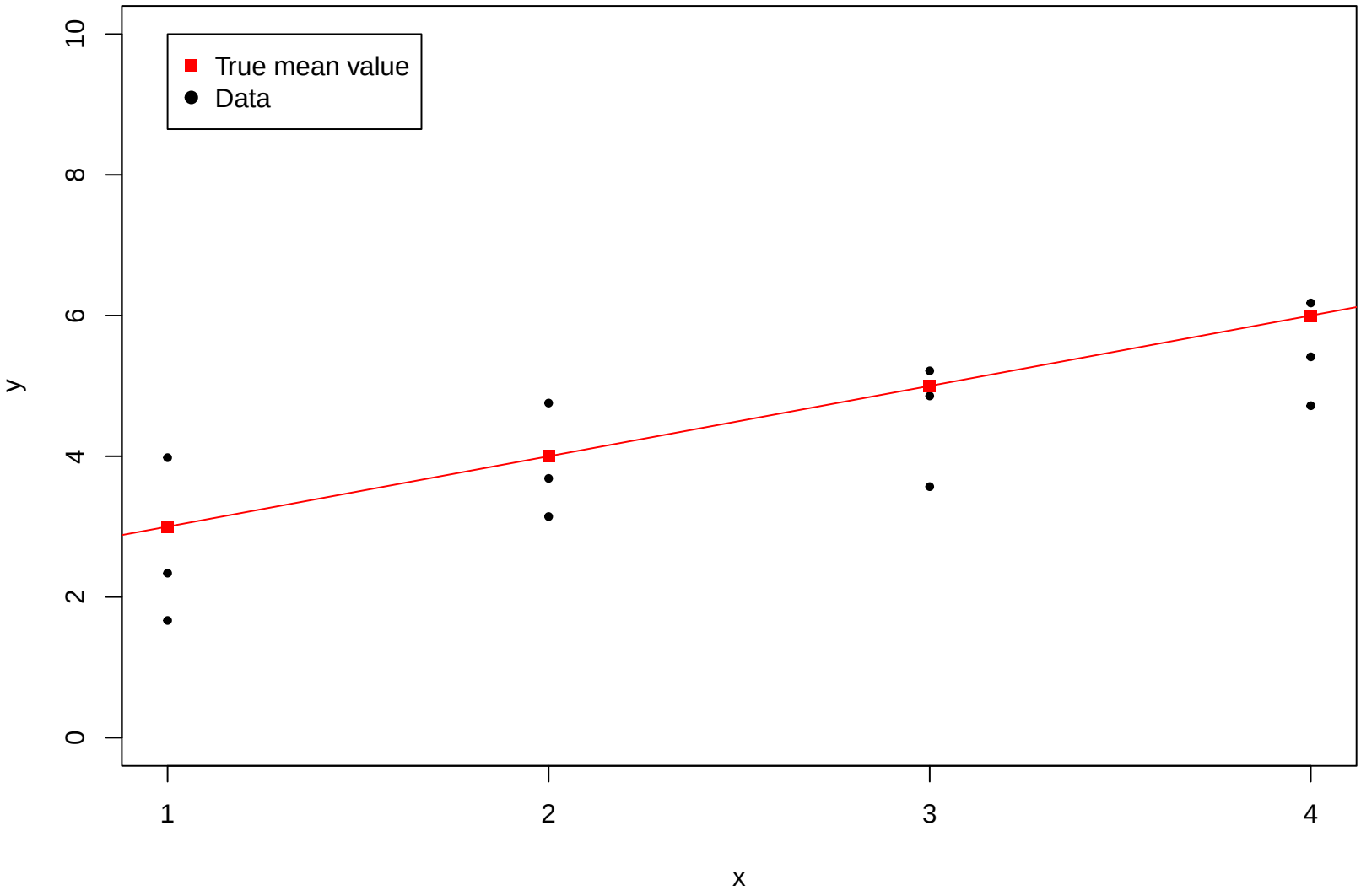
for $i = 1, 2, 3, 4$ and $j = 1, 2, 3$. Then the group means at

$$x_1 = 1 \qquad x_2 = 2 \qquad x_3 = 3 \qquad x_4 = 4$$

are

$$\mu_1 = 2 + x_1 = 3 \quad \mu_2 = 2 + x_2 = 4 \quad \mu_3 = 2 + x_3 = 5 \quad \mu_4 = 2 + x_4 = 6.$$

The ANOVA Table



The ANOVA Table

In a standard one-way layout, we would have four means that are treated as separate parameters

$$\mu_1, \mu_2, \mu_3, \mu_4$$

but in the regression case we have the formula

$$\mu_i = \beta_0 + \beta_1 \mathbf{x}_i$$

so the means are specified using only two parameters (β_0, β_1).

The ANOVA Table

Blood viscosity data:

```
> fit.vis<-lm(viscosity~pcv, data=Viscosity)
> summary(fit.vis)


+
+ Call:
+ lm(formula = viscosity ~ pcv, data = Viscosity)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.48814 -0.22152  0.04686  0.12955  0.64915
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept) -1.22336    0.58422  -2.094   0.0448 *
+ pcv          0.12243    0.01214  10.088 3.73e-11 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.3012 on 30 degrees of freedom
+ Multiple R-squared:  0.7723, Adjusted R-squared:  0.7647
+ F-statistic: 101.8 on 1 and 30 DF,  p-value: 3.731e-11
```

The ANOVA Table

```
> summary(aov(fit.vis))

+             Df Sum Sq Mean Sq F value   Pr(>F)
+ pcv          1  9.230   9.230   101.8 3.73e-11 ***
+ Residuals   30  2.721   0.091
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the p-values in the

- t-test of the slope coefficient
- ANOVA test

are identical. This is because the tests are equivalent, and

$$T_{obs}^2 = F_{obs}.$$

Checking the Local Fit

A plot of the residuals

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

can reveal model inadequacies. We should observe that in plots of

- $\widehat{e}$ vertical vs x horizontal
- $\widehat{e}$ vertical vs $\widehat{y}$ horizontal

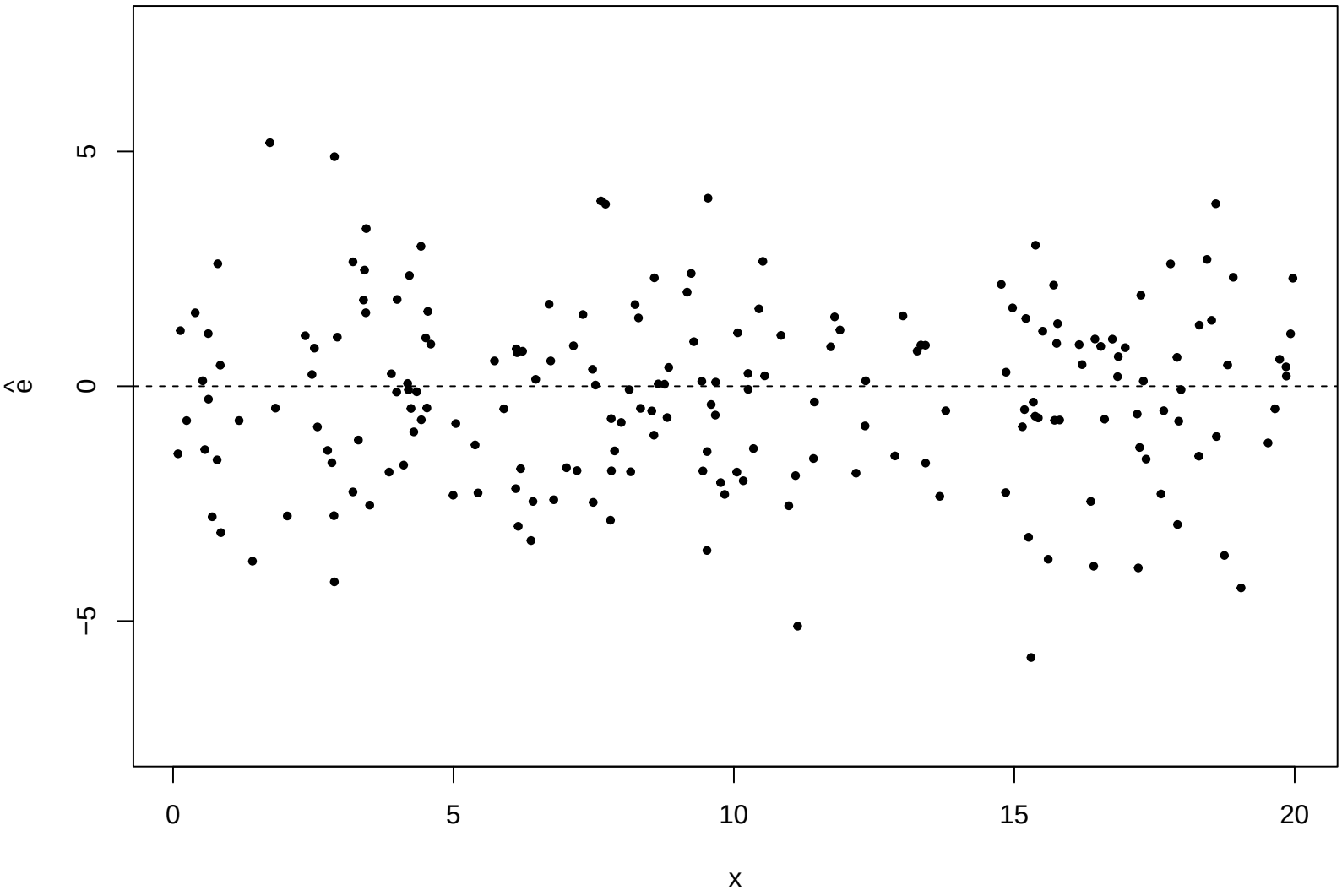
there is no discernible pattern

Checking the Local Fit

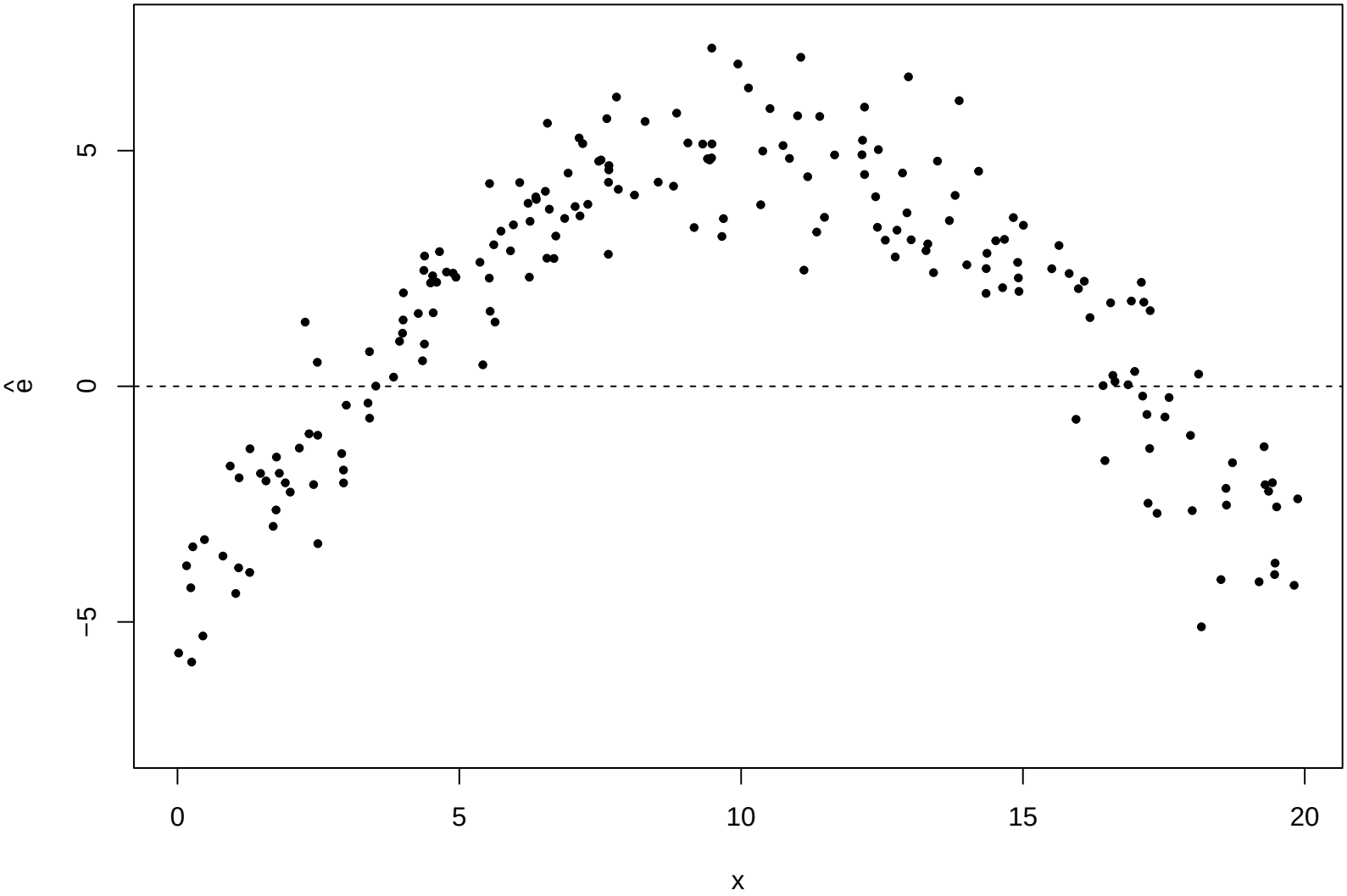
Specifically these plots should

- be vertically centered at zero for every horizontal value
- have the same variability around zero for every horizontal value

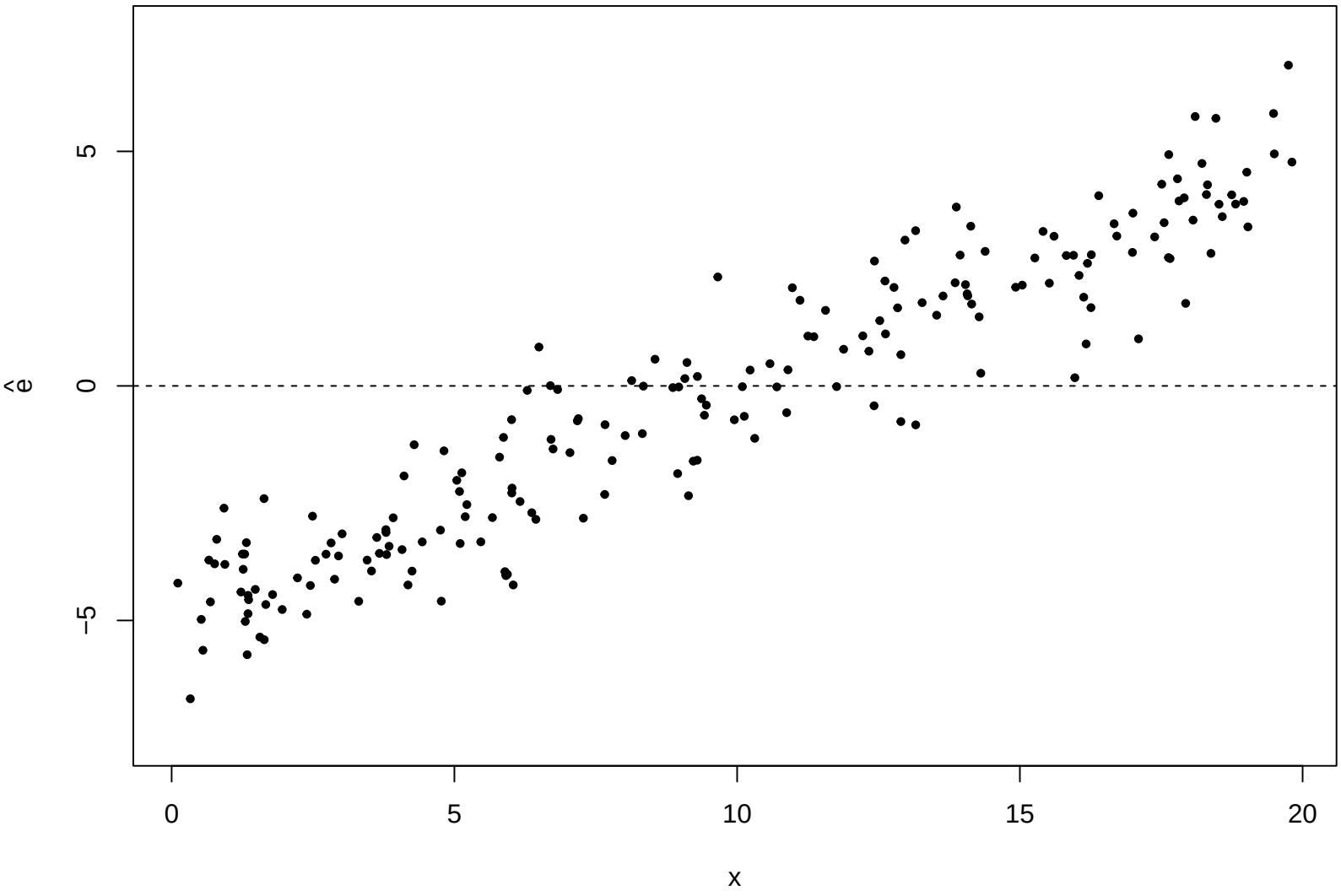
Good fit: no residual pattern



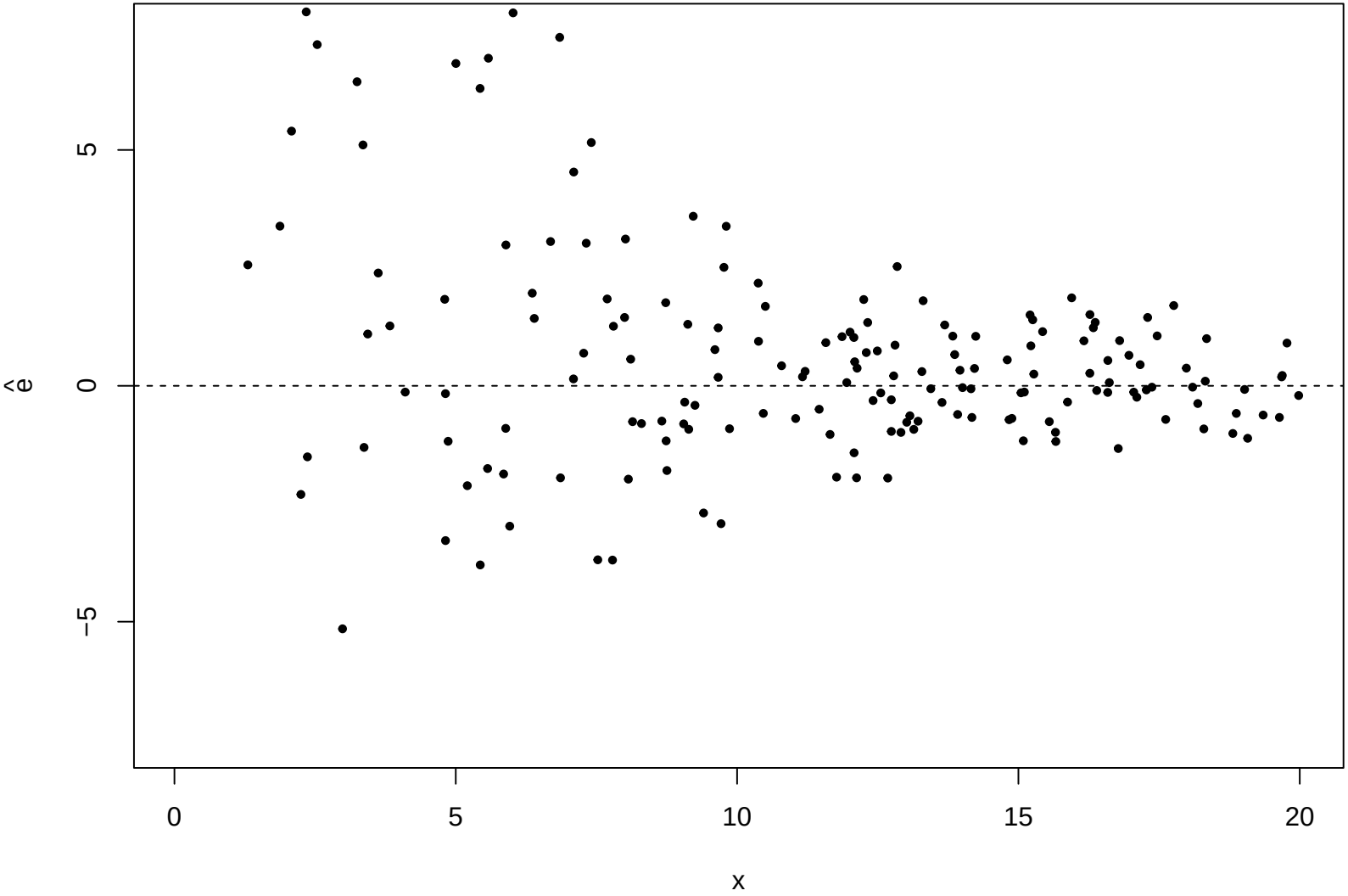
Pood fit: $\hat{e}$ not mean zero for all x



Pood fit: $\hat{e}$ not mean zero for all x



Pood fit: $\hat{e}$ not constant variance



Residual plots: notes

(i) It is rarely possible to make a definitive conclusion from inspecting a plot

(ii) A residual plot with which does not have mean zero for $\hat{e}$ for all values indicates that the proposed regression model is likely incorrect. That is, we assumed that

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 x$$

but in fact

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 x + h(x)$$

is the true (data generating) model, for some $h(x)$.

R^2 and adjusted R^2

Model assessment can be made using either the R^2 statistic

$$R^2 = 1 - \frac{SSE}{SS}$$

and the adjusted R^2 statistic

$$R^2_{adj} = 1 - \frac{SSE/EDF}{SS/TDF}$$

where

- EDF = error degrees of freedom = $n - 2$
- TDF = total degrees of freedom = $n - 1$

Polynomial Regression

In many practical situations, the simple straight-line

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

is not appropriate. Instead, a model including powers of x

$$x^2, x^3, \ldots, x^k$$

should be considered. For example

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x^j = \beta_0 + \beta_1 x + \cdots + \beta_k x^k$$

Polynomial Regression

The Polynomial Regression Model

$$Y = \beta_0 + \beta_1 \mathbf{x} + \cdots + \beta_k \mathbf{x}^k + \epsilon$$

where ϵ is a random error term as before can be used to model data.

Polynomial Regression

Two immediate questions:

1. How to carry out inference ?
   - estimation
   - testing
   - prediction
2. How to choose k ?

We begin by addressing 1. The estimation of parameters can be again carried out using the principle of least squares provided that the model assumptions listed before are valid.

$k = 2$: quadratic regression

We choose the three parameters $\beta = (\beta_0, \beta_1, \beta_2)^\top$ to minimize the sum of squared errors

$$SSE(\beta) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2$$

that is the fitted values for parameters β are

$$\widehat{y}_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

$k = 2$: quadratic regression

The estimates $\widehat{\beta}$ can be found to minimize SSE using calculus techniques (differentiating with respect to the elements of β) to give the minimum SSE

$$SSE(\beta) = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i - \widehat{\beta}_2 x_i^2)^2$$

$k = 2$: quadratic regression

We can also compute

- the estimated standard errors, which allow tests of parameters to be carried out, and confidence intervals calculated.

- prediction intervals.

- the estimate of the residual error variance σ^2, which is

$$\widehat{\sigma}^2 = \frac{SSE(\widehat{\beta})}{n-p}$$

  Here p is the number of parameters estimated, three, so we divide by $n - 3$.

$k = 2$: quadratic regression

- residuals
  - can be used to assess the fit of the model.
  - the residuals should be patternless if the model fit is good.

- R^2 and Adjusted R^2 statistics
  - used to assess the global fit of the model.
  - used to compare the quality of fit with other models.

Example: Pressure data

Example: Hooker's Pressure Data

This is a standard data set in R: from the help file

> *Forbes, J. (1957). Further experiments and remarks on the measurement of heights by boiling point of water. Transactions of the Royal Society of Edinburgh,* ***21****, 235-243.*
>
> *In his original paper, Forbes provided additional data collected by the botanist Dr. Joseph Hooker on temperatures and boiling points measured often at higher altitudes in the Himalaya Mountains.*

Example: Pressure data

Example: Hooker's Pressure Data

For the Hooker pressure data, a quadratic polynomial ($k = 2$) might be suitable.

$$Y = \beta_0 + \beta_1 \mathbf{x} + \beta_2 \mathbf{x}^2 + \epsilon$$

We need to estimate β_0, β_1 and β_2 for these data to see if the model fits better than the straight line model we fitted previously.

It transpires that the quadratic model produces a set of residuals that are patternless, which the straight line model when fitted does not.

Example: Pressure data

In the data set there are $n = 31$ observations.

- `pressure` is the x variable, representing the measured air pressure, measured in inches of Mercury
- `temp` is the y variable, representing the measured boiling temperature, measured in degrees F.

```
> head(Hooker)

+    temp pressure
+ 1 210.8   29.211
+ 2 210.2   28.559
+ 3 208.4   27.972
+ 4 202.5   24.697
+ 5 200.6   23.726
+ 6 200.1   23.369

> nrow(Hooker)

+ [1] 31
```

Example: Pressure data

```
> summary(lm(temp~pressure+I(pressure^2), data=Hooker))

+
+ Call:
+ lm(formula = temp ~ pressure + I(pressure^2), data = Hooker)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.79906 -0.26314 -0.01578  0.25139  0.73891
+
+ Coefficients:
+                 Estimate Std. Error t value Pr(>|t|)
+ (Intercept)   126.701623   2.112363  59.981  < 2e-16 ***
+ pressure        4.157627   0.199069  20.885  < 2e-16 ***
+ I(pressure^2)  -0.043754   0.004552  -9.612 2.29e-10 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.3956 on 28 degrees of freedom
+ Multiple R-squared:  0.9981, Adjusted R-squared:  0.998
+ F-statistic:  7307 on 2 and 28 DF,  p-value: < 2.2e-16
```

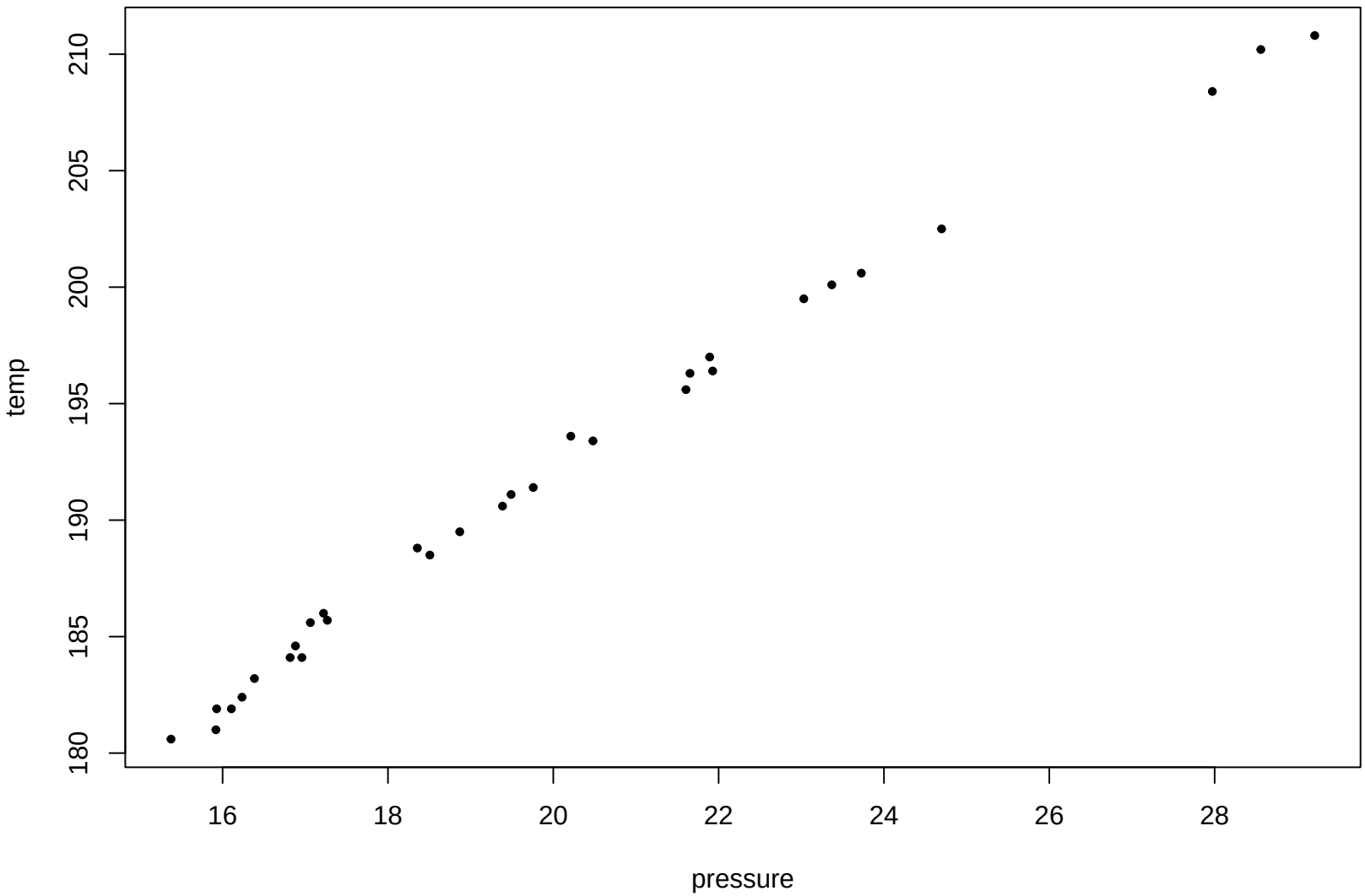
We need to use `I(pressure^2)` to compute the quadratic term in the formula.

Example: Pressure data

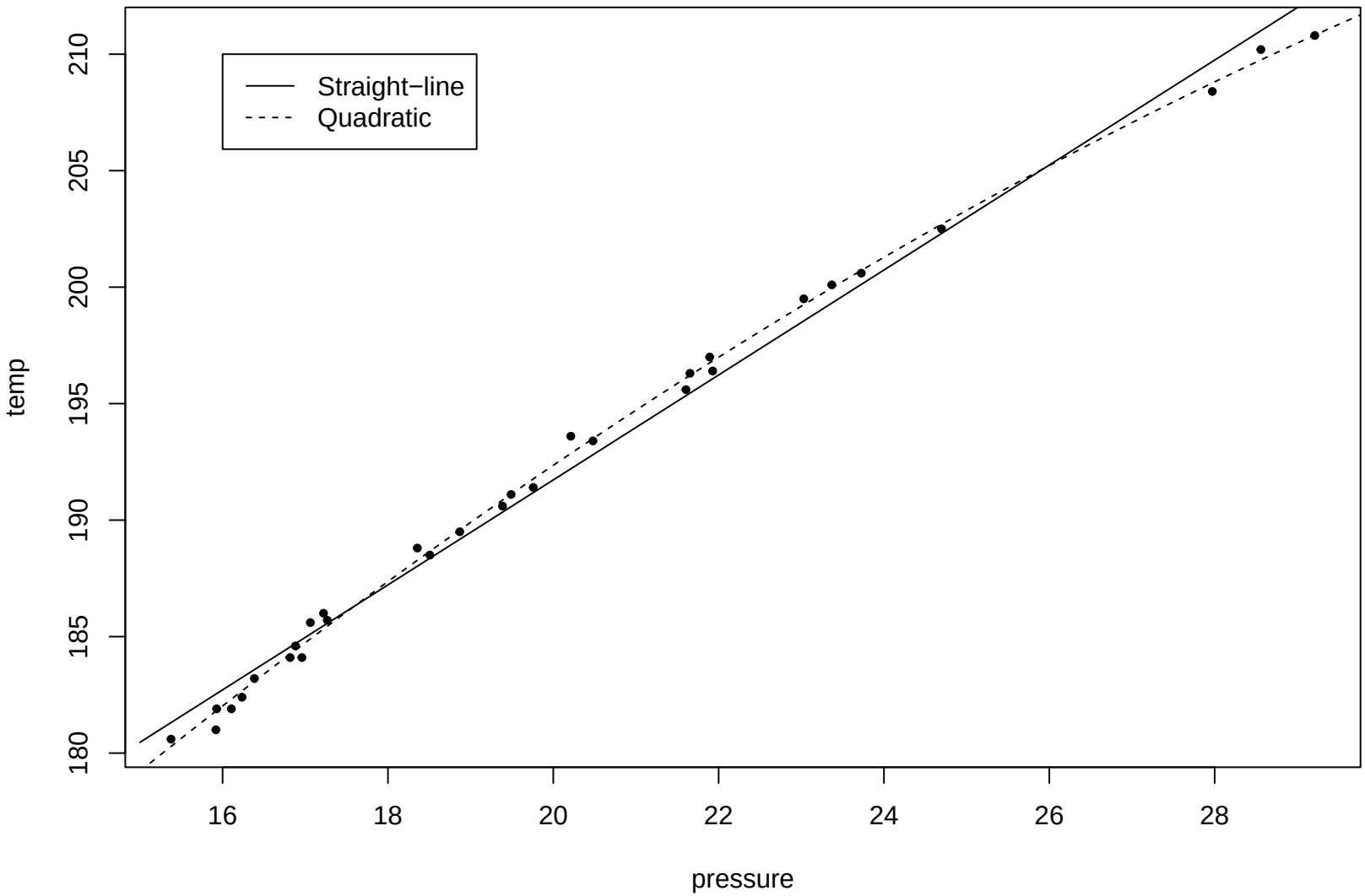
We might also consider a straight-line fit:

```
> summary(lm(temp~pressure, data=Hooker))

+
+ Call:
+ lm(formula = temp ~ pressure, data = Hooker)
+
+ Residuals:
+     Min      1Q  Median      3Q     Max
+ -1.6735 -0.6805  0.2203  0.5296  1.3976
+
+ Coefficients:
+              Estimate Std. Error t value Pr(>|t|)
+ (Intercept) 146.67290    0.77641  188.91   <2e-16 ***
+ pressure      2.25260    0.03809   59.14   <2e-16 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.806 on 29 degrees of freedom
+ Multiple R-squared:  0.9918, Adjusted R-squared:  0.9915
+ F-statistic:  3498 on 1 and 29 DF,  p-value: < 2.2e-16
```

Example: Pressure data



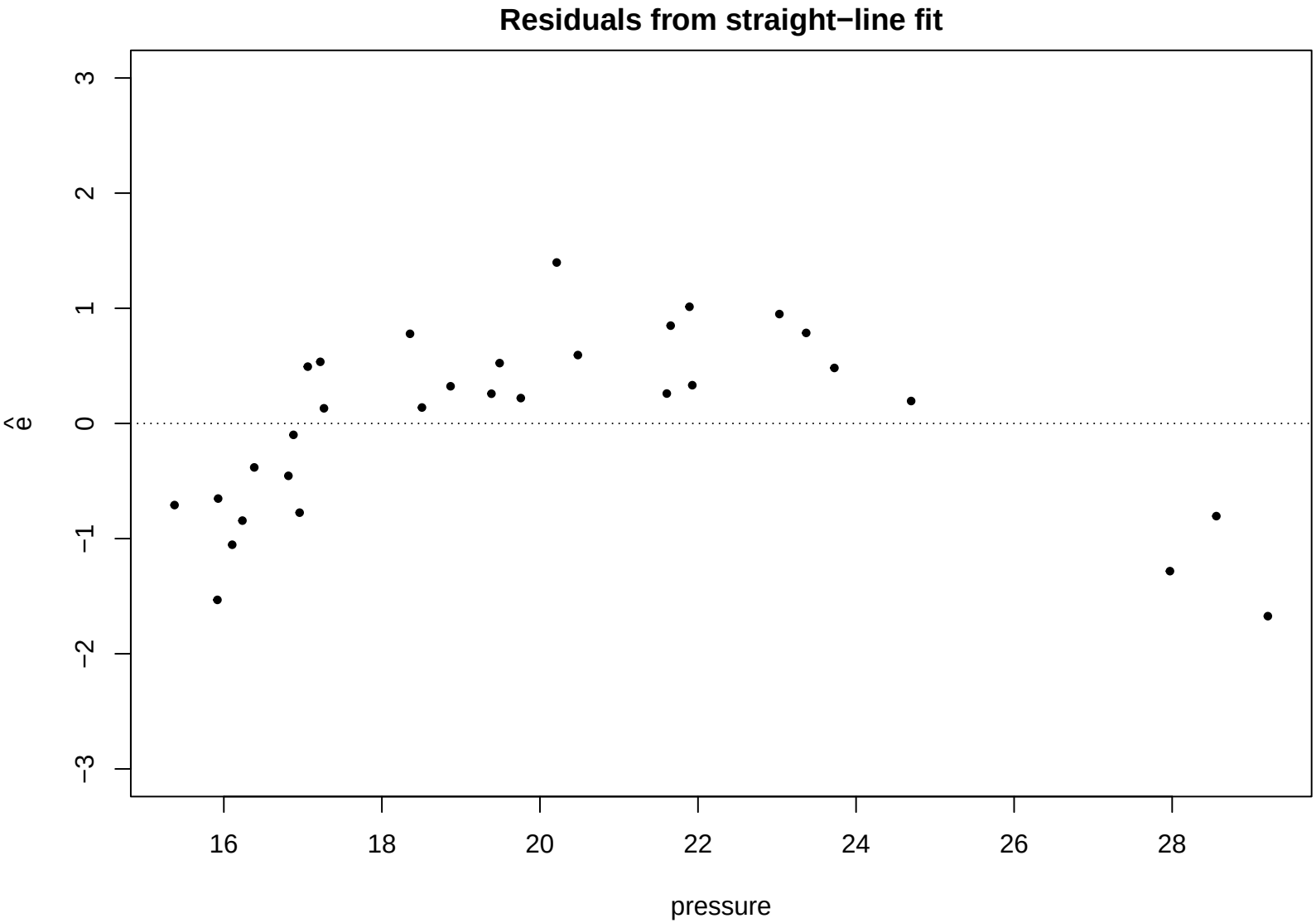
Example: Pressure data



Example: Pressure data

```
> fit.lin<-lm(temp~pressure, data=Hooker)
> fit.quad<-lm(temp~pressure+I(pressure^2), data=Hooker)
> ehat.lin<-residuals(fit.lin)
> ehat.quad<-residuals(fit.quad)
> par(mar=c(4,5,2,2))
> plot(Hooker$pressure, ehat.lin, ylim=range(-3,3),
+         pch=19, cex=0.6, xlab='pressure', ylab=expression(hat(e)))
> abline(h=0,lty=3)
> title('Residuals from straight-line fit')
```

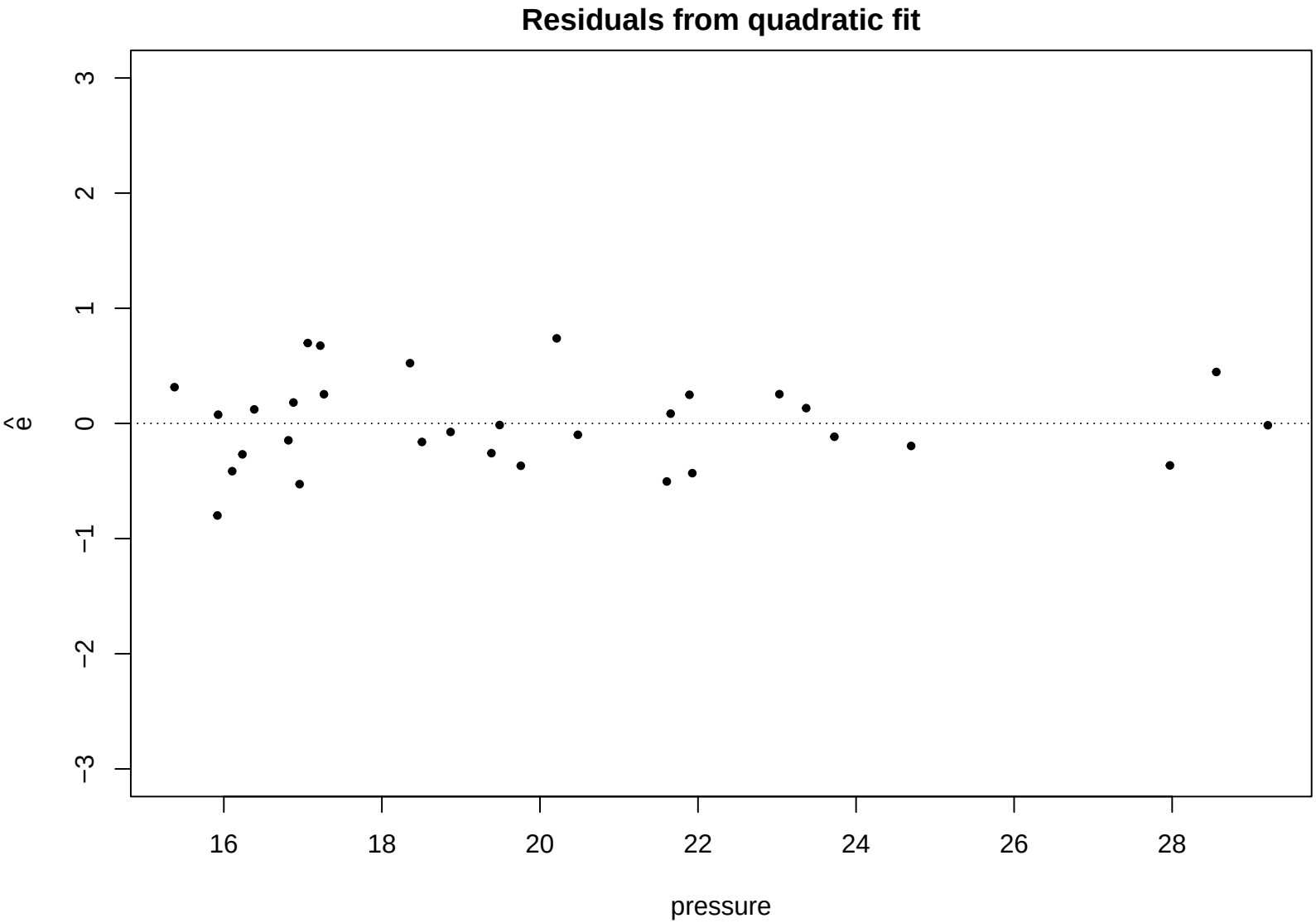
Example: Pressure data



Example: Pressure data

```
> par(mar=c(4,5,2,2))
> plot(Hooker$pressure, ehat.quad, ylim=range(-3,3),
+         pch=19, cex=0.6, xlab='pressure', ylab=expression(hat(e)))
> abline(h=0,lty=3)
> title('Residuals from quadratic fit')
```

Example: Pressure data



Example: Pressure data

It is common to use the standardized residuals

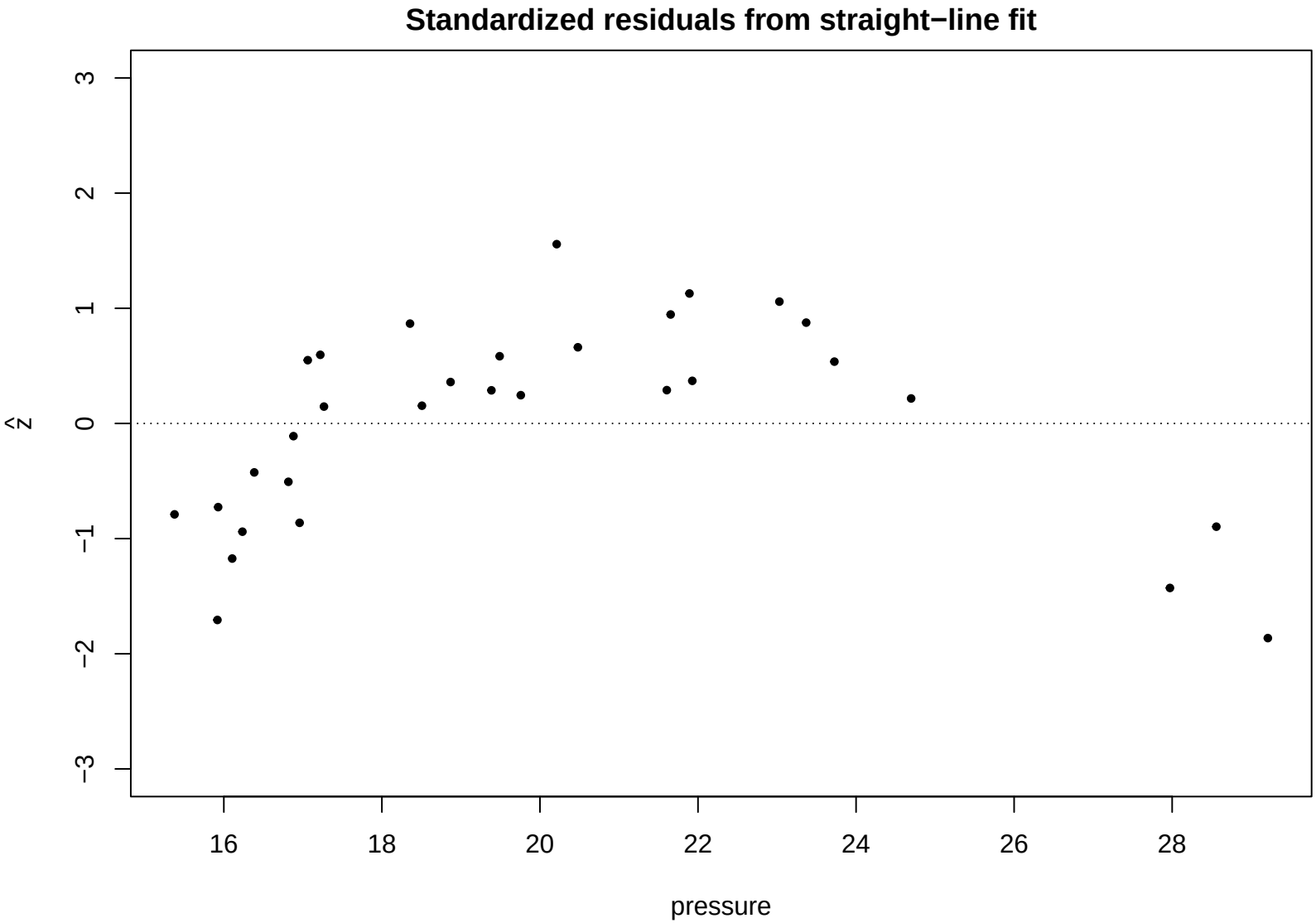
$$\widehat{z}_i = \frac{\widehat{e}_i}{\widehat{\sigma}} = \frac{y_i - \widehat{y}_i}{\widehat{\sigma}}$$

where $\widehat{\sigma}^2$ is the estimate of σ^2 defined previously, as

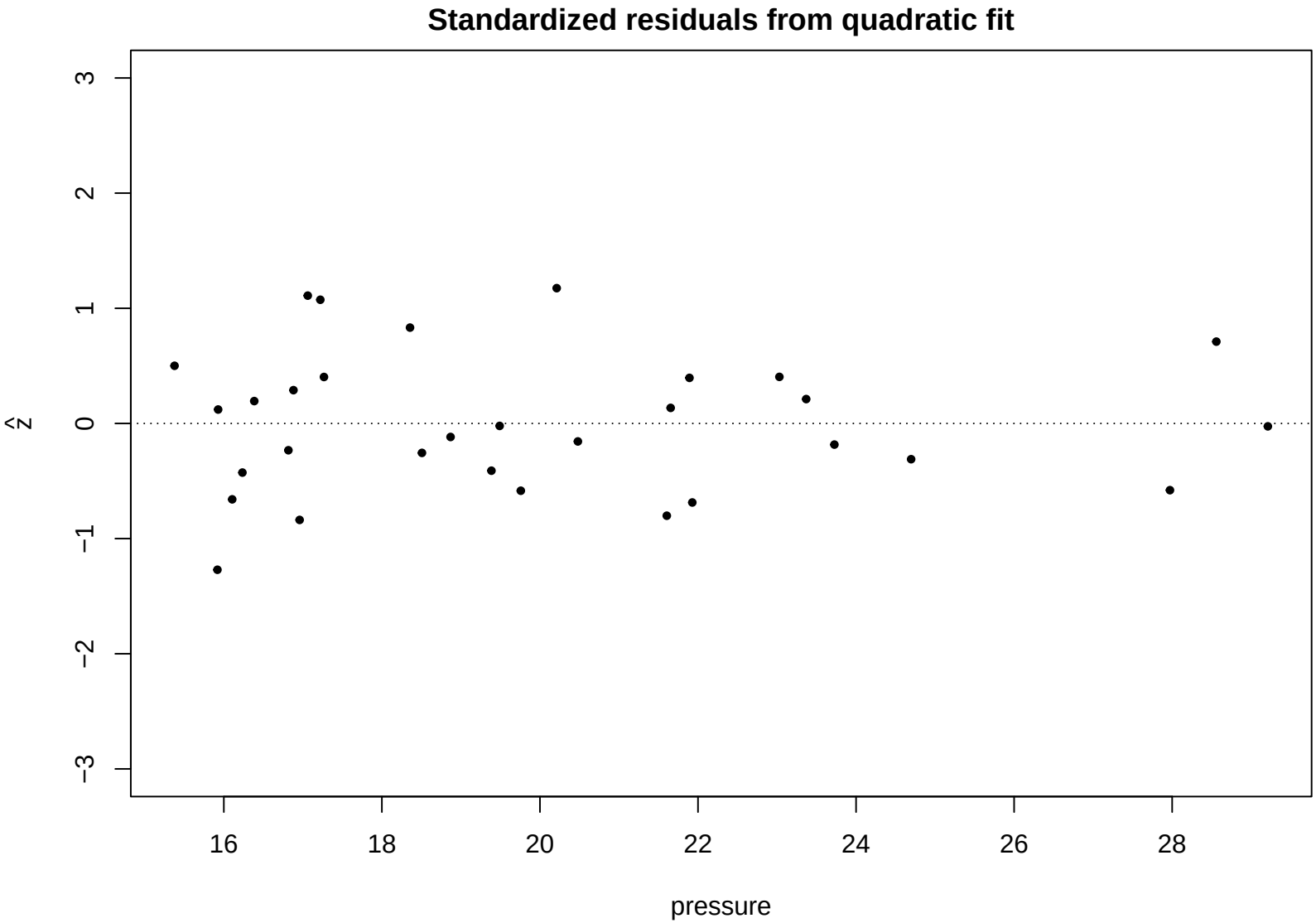
$$\mathrm{Var}[\widehat{z}_i] \approx 1$$

It is easier to compare $\widehat{e}_i$ across different models when inspecting residuals.

Example: Pressure data



Example: Pressure data



Example: Pressure data

Note:

```
> summary(fit.lin)$coef   #Straight-line fit estimates

+               Estimate Std. Error   t value     Pr(>|t|)
+ (Intercept) 146.672897 0.77641329 188.91085 2.266565e-46
+ pressure      2.252596 0.03808725  59.14306 8.634599e-32

> summary(fit.quad)$coef  #Quadratic fit estimates

+                   Estimate  Std. Error   t value     Pr(>|t|)
+ (Intercept)   126.70162290 2.112363414 59.980978 4.024869e-31
+ pressure        4.15762721 0.199069236 20.885333 1.298517e-18
+ I(pressure^2)  -0.04375373 0.004551914 -9.612162 2.287907e-10
```

The coefficients of `pressure` are appreciably different in the two models.

Example: Pressure data

Even though we describe the two models using the same parameter symbols

$$\begin{aligned} \text{Straight-line} \quad \mathbb{E}[Y|x] &= \beta_0 + \beta_1 x \\ \text{Quadratic} \quad \mathbb{E}[Y|x] &= \beta_0 + \beta_1 x + \beta_2 x^2 \end{aligned}$$

in fact, each term can be only be interpreted within the model it sits.

The Linear Model

Note: Although the model based on

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

is not linear in x, it is linear in the parameters. Because of this, we term this a linear model.

It is this fact that makes the least-squares solutions easy to find.

The Linear Model

This model is no more difficult to fit than the model

$$y = \beta_0 + \beta_1 \frac{x}{1+x} + \beta_2 (1 - e^{-x})$$

say - it is still a linear in the parameters model. It is in the general class of models

$$y = \beta_0 + \beta_1 g_1(x) + \beta_2 g_2(x)$$

where $g_1(x)$ and $g_2(x)$ are general functions of x.

The Linear Model

In fact, any model of the form

$$y = \sum_{j=0}^{k} \beta_j g_j(\mathrm{x}) + \epsilon$$

can be fitted routinely using least-squares; if we know x, then we can simply compute

$$g_0(\mathrm{x}), g_1(\mathrm{x}), \ldots, g_k(\mathrm{x})$$

and plug those values into the formula..

The Linear Model

Example: Harmonic regression: gene expression data

Measure the gene expression for 43 genes across a 48 hour cycle.

- gene expression measured on the log scale
- data are standardized to have mean zero.

The Linear Model

Example: Harmonic regression: gene expression data

Let

$$
\begin{aligned}
g_0(x) &= 1 \\
g_j(x) &= \begin{cases} \cos(\lambda_j x) & j \text{ odd} \\ \sin(\lambda_j x) & j \text{ even} \end{cases}
\end{aligned}
$$

where k is an even number, $k = 2p$ say.

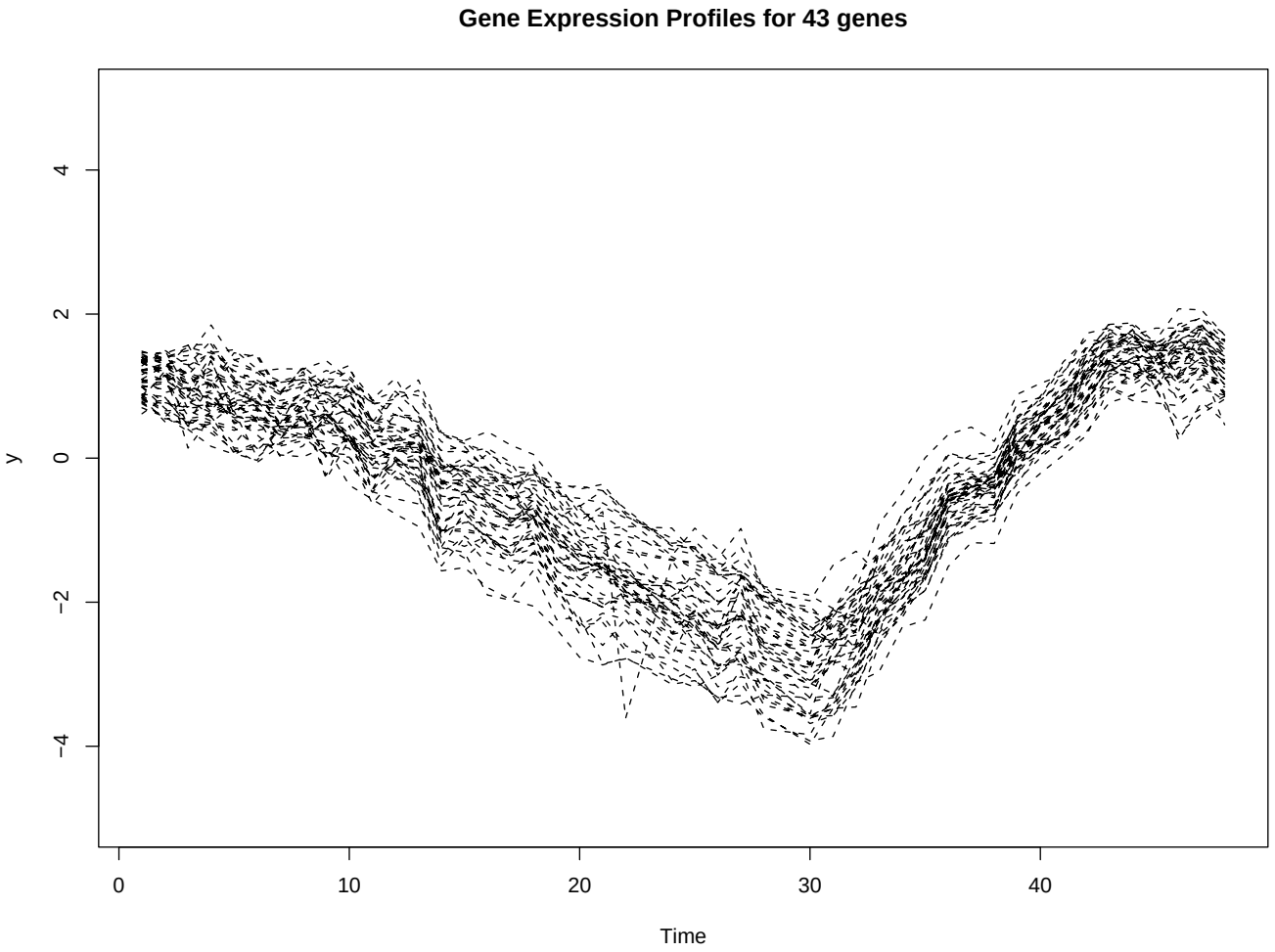
$\lambda_j, j = 1, 2, \ldots, p$ are constants

$$\lambda_1 < \lambda_2 < \cdots < \lambda_p$$

For fixed x, $\cos(\lambda_j x)$ and $\sin(\lambda_j x)$ are also fixed, known values.

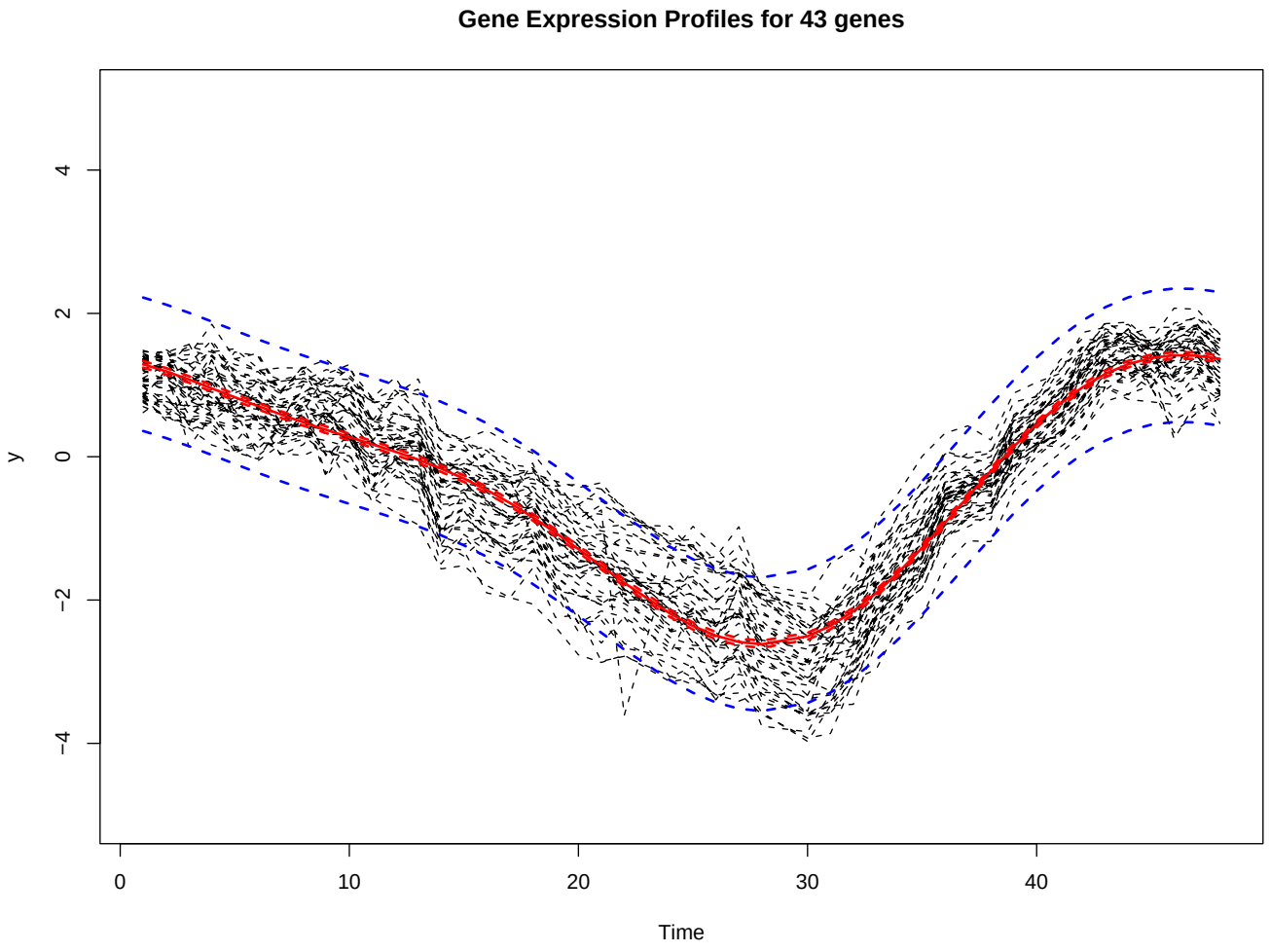
The Linear Model

Gene expression data



The Linear Model

Harmonic Regression Fit with $p = 2$.



The Linear Model

Calculations are straightforward because the system of equations based on the derivatives

$$\frac{\partial}{\partial \beta_j}\{SSE(\beta)\} = 0 \qquad j = 0, 1, \ldots, k$$

can always be solved routinely, so we can always find $\widehat{\beta}$.

In the general model, simple formulae for

- $\widehat{\beta}$
- e.s.e.$(\widehat{\beta})$
- $\widehat{\sigma}^2$

can be found using a matrix formulation.

Factor predictors

The one-way layout ANOVA can be formulated in the form of the general model.

Recall

- k independent groups
- means $\mu_1, \ldots, \mu_k$
- y_{ij} is the jth observation in the ith group

Factor predictors

Let

$$\begin{aligned} \beta_0 &= \mu_1 \\ \beta_i &= \mu_{i+1} - \mu_1 \qquad i = 1, \ldots, k-1. \end{aligned}$$

Define new predictor $x_{ij}(t)$ where

$$x_{ij}(t) = \begin{cases} 1 & \text{if } t = i \\ 0 & \text{if } t \neq i \end{cases}$$

That is, $x_{ij}(t)$ is one if t matches i (the group label) and is zero otherwise.

Factor predictors

Using the linear regression formulation, we may write

$$Y_{ij} = \beta_0 + \sum_{t=1}^{k-1} \beta_t x_{ij}(t+1) + \epsilon_{ij}.$$

For any i, j, $x_{ij}(t)$ is non-zero for only one value of t, when $t + 1 = i$.

Therefore, only one term contributes to the summation; the other terms are all zero.

Factor predictors

That is

$$\mathbb{E}[Y_{ij}|x] = \begin{cases} \beta_0 & i = 1 \\ \beta_0 + \beta_1 & i = 2 \\ \beta_0 + \beta_2 & i = 3 \\ \vdots & \\ \beta_0 + \beta_{k-1} & i = k \end{cases}$$

The parameters $\beta_0, \beta_1, \ldots, \beta_{k-1}$ can be estimated using least-squares.

Factor predictors

We term these $x_{ij}(t)$ quantities dummy variables, and the corresponding original variables factor predictors or factors.

This clarifies the link between

ANOVA and Linear Modelling

- they are essentially the same model formulation.

This link extends to all ANOVA models.

Factor predictor example

Milk protein data example:

```
> summary(aov(protein~Diet,data=Milk))

+               Df Sum Sq Mean Sq F value Pr(>F)
+ Diet           2  10.61   5.303   51.85 <2e-16 ***
+ Residuals   1334 136.43   0.102
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> str(Milk)

+ 'data.frame': 1337 obs. of  4 variables:
+  $ protein: num  3.63 3.57 3.47 3.65 3.89 3.73 3.77 3.9 3.78 3.82 ...
+  $ Time   : num  1 2 3 4 5 6 7 8 9 10 ...
+  $ Cow    : Factor w/ 55 levels "B01","B02","B03",..: 1 1 1 1 1 1 1 1 1 1 ...
+  $ Diet   : Factor w/ 3 levels "barley","barley+lupins",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Factor predictor example

```
> summary(lm(protein~Diet,data=Milk))

+
+ Call:
+ lm(formula = protein ~ Diet, data = Milk)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.89193 -0.21969 -0.01238  0.19807  1.16031
+
+ Coefficients:
+                   Estimate Std. Error t value   Pr(>|t|)
+ (Intercept)        3.53193    0.01551 227.681    < 2e-16 ***
+ Dietbarley+lupins -0.10223    0.02153  -4.749 0.00000227 ***
+ Dietlupins        -0.21955    0.02160 -10.166    < 2e-16 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.3198 on 1334 degrees of freedom
+ Multiple R-squared:  0.07213, Adjusted R-squared:  0.07074
+ F-statistic: 51.85 on 2 and 1334 DF,  p-value: < 2.2e-16
```

Factor predictor example

```
> y<-Milk$protein                        #Response
> x1<-Milk$Diet == 'barley+lupins'       #New predictor for group 2
> x2<-Milk$Diet == 'lupins'              #New predictor for group 3
```

The two variables `x1` and `x2` record the groups corresponding to the factor levels for `Diet`

`barley+lupins` `lupins`

respectively. We then use the `R` formula

$$\texttt{y} \sim \texttt{x1} + \texttt{x2}$$

to represent the summation in the model formula.

Factor predictor example

```
> summary(lm(y~x1+x2))

+
+ Call:
+ lm(formula = y ~ x1 + x2)
+
+ Residuals:
+      Min       1Q   Median       3Q      Max
+ -0.89193 -0.21969 -0.01238  0.19807  1.16031
+
+ Coefficients:
+             Estimate Std. Error t value   Pr(>|t|)
+ (Intercept)  3.53193    0.01551 227.681    < 2e-16 ***
+ x1TRUE      -0.10223    0.02153  -4.749 0.00000227 ***
+ x2TRUE      -0.21955    0.02160 -10.166    < 2e-16 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 0.3198 on 1334 degrees of freedom
+ Multiple R-squared:  0.07213, Adjusted R-squared:  0.07074
+ F-statistic: 51.85 on 2 and 1334 DF,  p-value: < 2.2e-16
```

Factor predictor example

That is, the results of the two analyses are identical.

Note that the baseline group – in this case, the `Diet` value `barley` – is not used to define a new predictor; in the `R` formula

```
lm(y~x1+x2)
```

the intercept β_0 is assumed to be present, and corresponds to the baseline group. The parameters

$$\beta_1, \beta_2$$

represents contrasts or differences from baseline.

Multiple Linear Regression

Multiple linear regression models model the variation in response as a function of more than one independent variable.

Suppose we have variables

$$X_1, X_2, \ldots, X_k$$

recording different features of the experimental units. We wish to model response Y as a function of $X_1, X_2, \ldots, X_k$.

Multiple Linear Regression

Consider the model for data point i

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

where x_{ij} is the measured value of covariate j on experimental unit i.

Multiple Linear Regression

That is

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i$$

The first two terms on the right hand side

$$\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

are the deterministic components representing the expected value of Y for those x values, and ϵ_i is the random component.

Multiple Linear Regression

Example: $k = 2$

$$Y_i = \beta_0 + \beta_1 \mathbf{x}_{i1} + \beta_2 \mathbf{x}_{i2} + \epsilon_i$$

which is a three parameter model with two predictors.

- β_0 is the intercept
- β_1 represents the contribution of predictor X_1 when X_2 is included in the model
- β_2 represents the contribution of predictor X_2 when X_1 is included in the model

Multiple Linear Regression

We can also include interaction terms

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12}(x_{i1} \,.\, x_{i2}) + \epsilon_i$$

where

- The first two terms in x_{i1} and x_{i2} are main effects
- The third term in $(x_{i1} \,.\, x_{i2})$ is an interaction
  - β_{12} represents the amount that X_2 modifies the effect of X_1 when both predictors are included in the model (compared to the case when $X_2 = 0$.

This is a four parameter model.

Multiple Linear Regression

For two continuous predictors X_1 and X_2, the interaction term is obtained by multiplying the two numerical values together

- That is, we compute the product $x_{i1}.x_{i2}$, and include this as a new predictor
- More generally, interactions require more complicated calculations. We will write

$$X_1 : X_2$$

to represent the interaction between X_1 and X_2. This matches R notation.

Multiple Linear Regression Examples

- Multiple regression: Viscosity Example
- Factor Regression
- Interaction
- Residuals

Subgroup analysis

Subgroup analysis is a form of interaction modelling.

Suppose we have

- one factor predictor
- one continuous predictor (or covariate)

where the factor predictor interacts with the covariate so as to modify the slope across the subgroups.

Subgroup analysis

We describe the models using the notation introduced for ANOVA; consider case of

- a single covariate X_1
- a single binary factor predictor X_2

Subgroup analysis

Possible models:

MODEL 0	One horizontal straight line	1
MODEL 1	Two parallel horizontal straight lines	X_2
MODEL 2	One straight line, non-zero slope	X_1
MODEL 3	Two parallel straight lines, with non-zero slope	$X_1 + X_2$
MODEL 4	Two non-parallel straight lines	$X_1 + X_2 + X_1 : X_2$

Subgroup analysis

Note: Always be on the lookout for lurking subgroups (subgroups determined by the levels of an unnoticed factor predictor)

Inferences can change radically when the lurking factor is included in the model

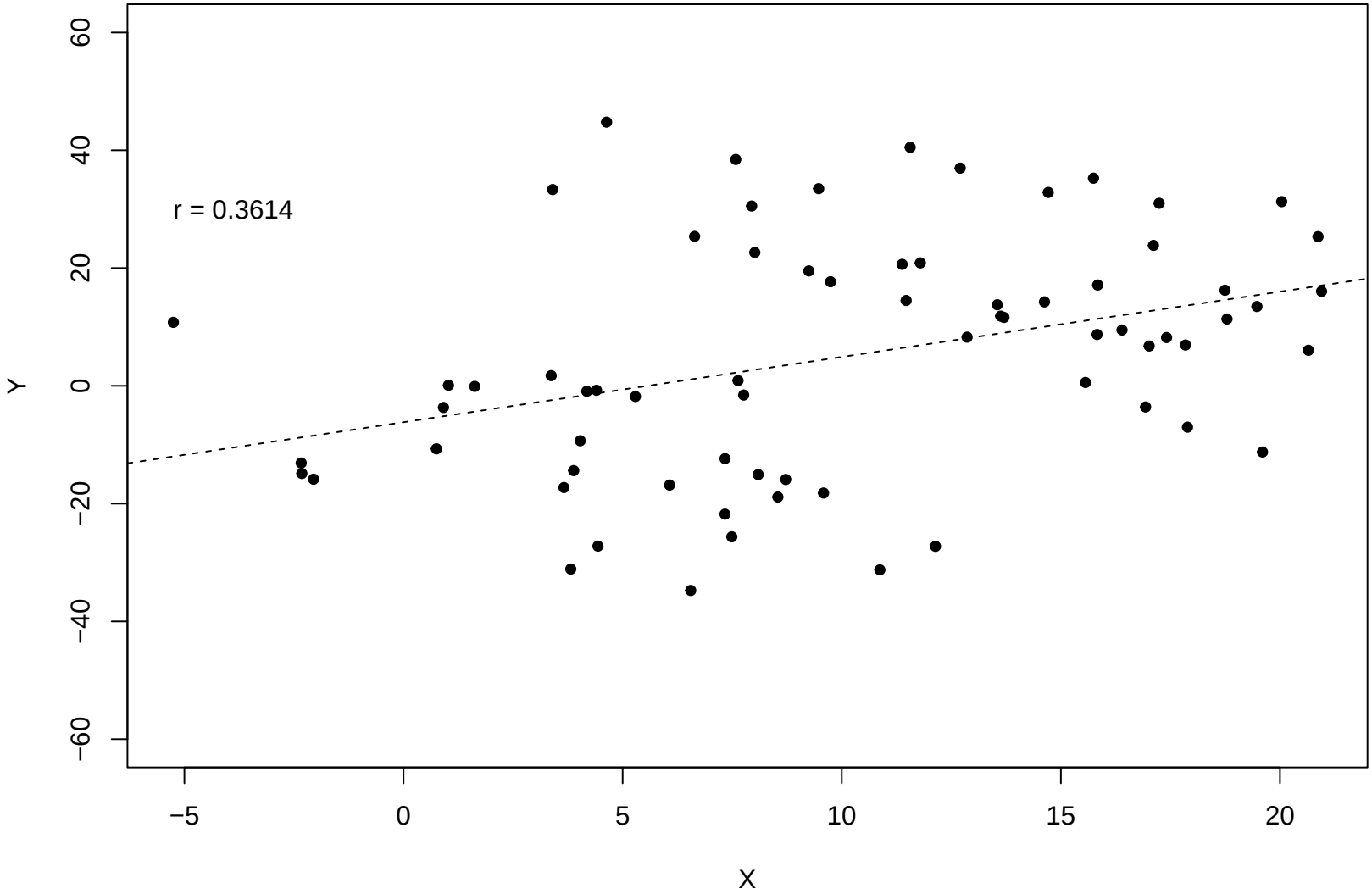
- positive association can be converted into negative association with the continuous covariate.

Subgroup analysis

For example, for factor predictor X_2 taking two levels and continuous covariate X_1.

When the pooled data are examined, a positive association between Y and X_1 is revealed.

Subgroup analysis



Subgroup analysis

If we fit the model

$$\mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x}_1$$

the estimate of the slope is positive: $\widehat{\beta}_1 = 1.1081$.

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


+
+ Call:
+ lm(formula = y ~ x1)
+
+ Residuals:
+     Min      1Q  Median      3Q     Max
+ -37.117 -12.062  -1.197  12.550  45.813
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept)  -6.1727     4.1555  -1.485  0.14206
+ x1            1.1081     0.3467   3.196  0.00211 **
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 18.77 on 68 degrees of freedom
+ Multiple R-squared:  0.1306, Adjusted R-squared:  0.1178
+ F-statistic: 10.21 on 1 and 68 DF,  p-value: 0.002114
```

Subgroup analysis

When the pooled data are separated into subgroups,

- $X_2 = 1$ ($\bullet$) and
- $X_2 = 2$ ($+$).

a negative association between Y and X_1 in each subgroup is revealed.

Subgroup analysis

Here we fit the model

$$\mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_{22} + \beta_{12} \mathbf{x}_1 . \mathbf{x}_{22}$$

where

$$\mathbf{x}_{22} = \begin{cases} 1 & \text{for data in Group 2 } (X_2 = 2) \\ 0 & \text{for data in Group 1 } (X_2 = 1) \end{cases}$$

Subgroup analysis

That is, for

- data in Group 1, the model is

$$\mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x}_1$$

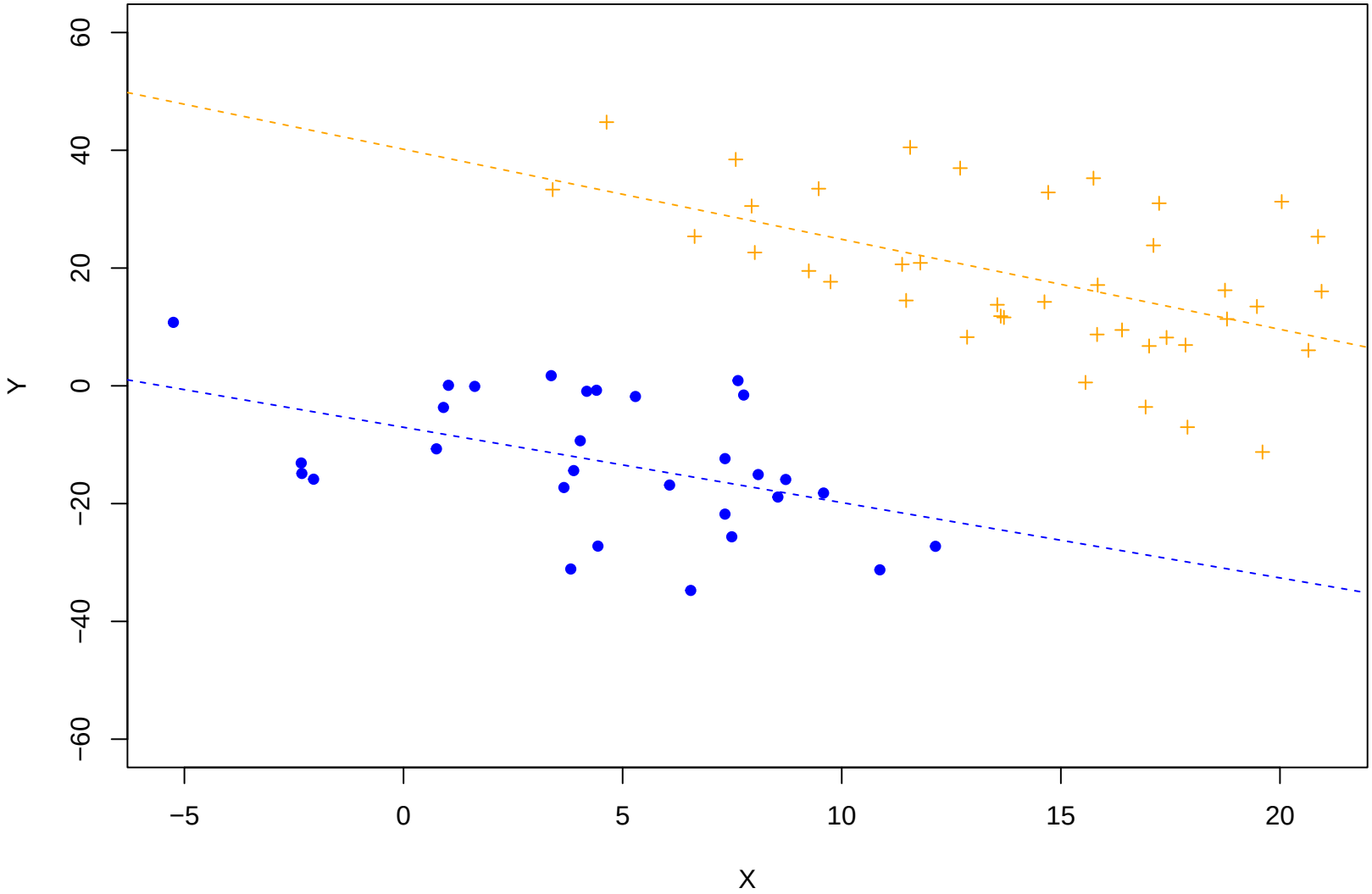
- data in Group 2, the model is

$$\mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 + \beta_{12} \mathbf{x}_1$$

  more explicitly, a straight line with

$$\text{Intercept: } \beta_0 + \beta_2 \qquad \text{Slope: } \beta_1 + \beta_{12}$$

Subgroup analysis



Subgroup analysis

```
+
+ Call:
+ lm(formula = y ~ x1 + x2 + x1:x2)
+
+ Residuals:
+     Min      1Q  Median      3Q     Max
+ -21.443  -7.361  -1.844   8.879  21.742
+
+ Coefficients:
+             Estimate Std. Error t value Pr(>|t|)
+ (Intercept)  -7.0401     3.0019  -2.345   0.0220 *
+ x1           -1.2787     0.4876  -2.622   0.0108 *
+ x22          47.2124     6.4117   7.363 3.64e-10 ***
+ x1:x22       -0.2502     0.6179  -0.405   0.6868
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
+
+ Residual standard error: 10.97 on 66 degrees of freedom
+ Multiple R-squared:  0.7118, Adjusted R-squared:  0.6988
+ F-statistic: 54.35 on 3 and 66 DF,  p-value: < 2.2e-16
```

Subgroup analysis

If we increase X_1, this

- decreases response in subgroup 1
- decreases response in subgroup 2

but appears to increase respons eoverall.

This is known as Simpson's Paradox in Regression. It illustrates that pooling data over subgroups must be carried out with care !

- you must fit the factor predictor in the model if you suspect subgroup differences exist.

Subgroup analysis

In the example, the problem arises due to dependence between X_1 and X_2; all the group with $X_2 = 0$ have low values of X_1, whereas all the group with $X_2 = 1$ have high values of X_1

Dependence between covariates and factor predictors makes model fitting and results interpretation complicated.

Recap

Recap: we can build general models

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i$$

to explain the variation of y in terms of covariates and factor predictors $x_1, \ldots, x_k$.

- Simple Linear Regression
- Polynomial Regression
- Multiple Regression
- Factor Predictor Regression
- Interaction Models

Recap

We can fit each of these models easily using least-squares to obtain

- estimates $\widehat{\beta} = (\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k)^\top$
- standard errors
- goodness of fit measures R^2 and Adjusted R^2
- residuals for model checking
- predictions

Interpreting $\widehat{\beta}_j$

$\widehat{\beta}_j$ can be interpreted as the amount of increase in response y when x_j increases by one unit when the other predictors

$$x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$$

are held fixed.

We can test the hypothesis

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_0 &: \beta_j \neq 0 \end{aligned}$$

using the usual hypothesis testing approach.

Interpreting $\widehat{\beta}_j$

Test statistic:

$$T_{obs,j} = \frac{\widehat{\beta}_j}{\text{e.s.e.}(\widehat{\beta}_j)} = \frac{\text{ESTIMATE}}{\text{STANDARD ERROR}}$$

If H_0 is true,

$$T_{obs,j} \sim t_{n-k-1}$$

as we are estimating $k+1$ parameters overall.

Note: In multiple regression, when testing each of

$$\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$$

we should strictly use a multiple testing correction (as in post-hoc tests in ANOVA)

Model Checking

We can fit models with different numbers of predictors, and

- assess whether any individual covariate is influential in the model (look at t-statistics).
- assess whether there is any explanatory power in the variables combined (look at ANOVA statistics)

ANOVA table for multiple regression

SOURCE	DF	SS	MS	F
REGRESSION	k	SSR	MSR	$F = \dfrac{MSR}{MSE}$
ERROR	$n-k-1$	SSE	MSE	
TOTAL	$n-1$	SS		

where

$$MSR = \frac{SSR}{k} \qquad MSE = \frac{SSE}{n-k-1}$$

the F-statistic is

$$F = \frac{MSR}{MSE}$$

ANOVA table for multiple regression

If H_0 is true

$$F \sim F_{k,n-k-1}.$$

Here

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_a &: \text{At least one } \beta_j \neq 0 \end{aligned}$$

- The model for H_0 has one parameter β_0.
- The model for H_a has $k+1$ parameters

$$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$$

ANOVA table for multiple regression

- Therefore the number of extra parameters for model H_a is

$$(k+1) - 1 = k$$

To obtain model H_0 from model H_a we constrain k parameters to be zero.

ANOVA table for multiple regression

Because we can constrain model H_a by setting some parameters to zero to obtain model H_0, we say that

Model H_0 is nested inside Model H_a

The number, k, of constraints $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ explains why the ANOVA table Regression degrees of freedom is k

- the multiple regression brings in k extra parameters.

ANOVA table for multiple regression

In addition, we can use the R^2 or Adjusted R^2 statistic to check overall model adequacy

$$R^2 = 1 - \frac{SSE}{SS_{yy}} = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{SSR}{SS}$$

which is equal to

$$\frac{\text{VARIATION EXPLAINED BY THE REGRESSION}}{\text{TOTAL VARIATION}}$$

Also

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SS/(n-1)}$$

$R^2 > 0.7$ implies that the model is a good fit, that is, most of the variation observed is explained by the regression model.

ANOVA table for multiple regression

We can now fit completely general models in the form of the Linear Model; if y is the response, and $x_1, \ldots, x_k$ are the covariates or factor predictors, we can include combinations of

- Polynomial Main Effects : $x_j, x_j^2, x_j^3, \ldots$
- Two-way Interactions: $x_{j_1}.x_{j_2}$
- Three-way Interactions: $x_{j_1}.x_{j_2}.x_{j_3}$

etc.

Dummy Variables

Recall that we can fit a factor predictor using dummy variables.

For example, if factor predictor X has L levels, we pick one as the baseline level, and then create $L-1$ new binary predictors $X_1, \ldots, X_{L-1}$, where, for $l = 1, \ldots, L-1$

$$X_l = \begin{cases} 1 & \text{whenever } X = l \\ 0 & \text{otherwise} \end{cases}$$

We can then include $X_1, \ldots, X_{L-1}$ in the regression model, along with an intercept which represents the mean level for the baseline group.

Dummy Variables

Example: $L = 4$

We pick $X = 1$ as the baseline group, and define X_1, X_2 and X_3 as follows:

X	X_1	X_2	X_3
3	0	1	0
1	0	0	0
3	0	1	0
4	0	0	1
2	1	0	0
2	1	0	0

See textbook Section 12.7.

Stepwise Model Selection

We seek a method that allows us to compare nested models.

Suppose we want to compare

$$\begin{aligned} \text{MODEL 1} \quad &: \quad \mathbb{E}[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2 \\ \text{MODEL 2} \quad &: \quad \mathbb{E}[Y|x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \end{aligned}$$

Model 1 is nested inside Model 2 as if we set $\beta_3 = 0$ in Model 2, we get Model 1.

Stepwise Model Selection

If

$$\begin{aligned} \text{MODEL 1} \quad &: \quad \mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 \\ \text{MODEL 2} \quad &: \quad \mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_{12}(\mathbf{x}_1 . \mathbf{x}_2) \end{aligned}$$

we can set $\beta_{12} = 0$ in Model 2 to obtain Model 1, so again the models are nested.

We can set up a hypothesis test to assess whether the simplification of Model 2 to Model 1 (by setting one or more parameters equal to zero) is justified by the data.

ANOVA tests for Comparing Nested Models

Terminology:

- Complete Model

$$\mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_k \mathbf{x}_k$$

- Reduced Model

$$\mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 \mathbf{x}_1 + \cdots + \beta_g \mathbf{x}_g$$

where $g < k$. The reduced model is obtained from the complete model by setting

$$\beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$$

ANOVA tests for Comparing Nested Models

The reduced model is nested inside the complete model.

We wish to test the hypothesis

$$\begin{aligned} H_0 &: \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0 \\ H_a &: \text{At least one of these } \beta_j \neq 0 \end{aligned}$$

We can test this hypothesis by fitting both models, and combining the results; we focus on the sums of squares quantities.

Model comparison method

1. Fit the complete model and obtain the sum of squared errors, SSE_C, available from the ANOVA table.
2. Fit the reduced model and obtain the sum of squared errors, SSE_R, available from the ANOVA table.
3. Form the test statistic

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)}$$

Model comparison method

If H_0 is true, then $F \sim F_{k-g,n-k-1}$.

- $k - g$ is the number of parameters we set equal to zero when moving from complete to reduced model.
- Using this F statistic, we can assess whether there is evidence to support the reduced model over the complete model

Model comparison method

Complete Model ANOVA table:

SOURCE	DF	SS	MS	F
COMPLETE MODEL	k	SSR_C	MSR_C	F_C
ERROR$_C$	$n-k-1$	SSE_C	MSE_C	
TOTAL	$n-1$	SS		

Model comparison method

Reduced Model ANOVA table:

SOURCE	DF	SS	MS	F
REDUCED MODEL	g	SSR_R	MSR_R	F_R
ERROR$_R$	$n-g-1$	SSE_R	MSE_R	
TOTAL	$n-1$	SS		

Model comparison method

The result holds for comparing any two nested models where the standard model assumptions hold:

- ϵ uncorrelated
- ϵ independent of $x_1, \ldots, x_k$
- ϵ has constant variance
- $\epsilon \sim Normal(0, \sigma^2)$

Model comparison method

We cannot use this approach to compare non-nested models; for example

$$\begin{aligned} \text{MODEL 1} \quad &: \quad Y = \beta_0 + \beta_1 x_1 + \epsilon \\ \text{MODEL 2} \quad &: \quad Y = \beta_0 + \beta_2 x_2 + \epsilon \end{aligned}$$

cannot be compared using the F-test.

The statistic

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)} = \frac{①/②}{③/④}$$

① - $SSE_R - SSE_C$: this is the improvement in fit when the reduced model is extended to the complete model

② - $k - g$: this is the number of extra parameters needed to extend the reduced model to the complete model

③ - SSE_C

④ - $n - k - 1$

The statistic

In the numerator

$$① = SSE_R - SSE_C$$

is always positive. This is because we constrain the Complete model to obtain the Reduced model, so the Reduced model is less flexible to fit to the data.

The statistic

In the denominator,

$$\frac{③}{④} = \frac{SSE_C}{n-k-1}$$

is the best guess we have at the true value of σ^2, that is, the estimate of σ^2 constructed using as much information as possible, once the effects of

$$x_1, \ldots, x_k$$

have been accounted for.

Example: Hooker's data

Example: Hooker's Data

We consider the two models:

$$\begin{aligned} \text{MODEL 1} \quad &: \quad Y = \beta_0 + \beta_1 x + \epsilon \\ \text{MODEL 2} \quad &: \quad Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \end{aligned}$$

Here

- MODEL 1: Reduced Model
- MODEL 2: Complete Model

$k = 2$, $g = 1$.

IS THE QUADRATIC TERM NEEDED ?

Example: Hooker's data

Example: Hooker's Data: straight-line

```
> supernova(lm(temp~pressure, data=Hooker), type=2)

+  Analysis of Variance Table (Type II SS)
+  Model: temp ~ pressure
+
+                                SS df       MS        F   PRE     p
+  ----- --------------- | -------- -- -------- -------- ----- -----
+  Model (error reduced) | 2272.474  1 2272.474 3497.902 .9918 .0000
+  Error (from model)    |   18.840 29    0.650
+  ----- --------------- | -------- -- -------- -------- ----- -----
+  Total (empty model)   | 2291.315 30   76.377
```

Example: Hooker's data

Example: Hooker's Data: quadratic

```
> supernova(lm(temp~pressure+I(pressure^2), data=Hooker), type=2)

+  Analysis of Variance Table (Type II SS)
+  Model: temp ~ pressure + I(pressure^2)
+
+                                        SS df       MS        F   PRE     p
+  ------------- --------------- | -------- -- -------- -------- ----- -----
+          Model (error reduced) | 2286.933  2 1143.467 7306.975 .9981 .0000
+       pressure                 |   68.260  1   68.260  436.197 .9397 .0000
+  I(pressure^2)                 |   14.459  1   14.459   92.394 .7674 .0000
+          Error (from model)    |    4.382 28    0.156
+  ------------- --------------- | -------- -- -------- -------- ----- -----
+          Total (empty model)   | 2291.315 30   76.377
```

Example: Hooker's data

Example: Hooker's Data

COMPLETE MODEL	SSR_C	2286.933
	SSE_C	4.382
REDUCED MODEL	SSR_R	2272.474
	SSE_R	18.840

with $n = 31, k = 2, g = 1$, so $k - g = 1$ and $n - k - 1 = 28$.

Example: Hooker's data

Example: Hooker's Data

Hence

$$
\begin{aligned}
F_{obs} &= \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)} \\
&= \frac{(18.840 - 4.382)/1}{4.382/28} \\
&= 92.383
\end{aligned}
$$

Example: Hooker's data

Example: Hooker's Data

We compare F with the

$$F_{k-g,n-k-1} \equiv F_{1,28}$$

distribution. The upper $\alpha = 0.05$ point is

$$F_{0.05,1,28} = 4.20$$

Thus

$$92.383 = F_{obs} > F_{0.05,1,28} = 4.20$$

and H_0 is rejected in favour of H_a.

The quadratic model fits better than the straight-line model.

Example: Hooker's data

From the original ANOVA tables, we already know that Model 1 and Model 2 both fit better than the null model

$$\begin{aligned} \text{MODEL 0} \qquad \mathbb{E}[Y|\mathbf{x}] &= \beta_0 \\ Y &= \beta_0 + \epsilon \end{aligned}$$

where there is no dependence on x.

We have now also confirmed that Model 2 fits better than Model 1.

Diabetes data

Example: Diabetes Data

Factor Predictor: group (X_2)
Continuous Covariate: loggluf (X_1)
Response: logglut (Y)

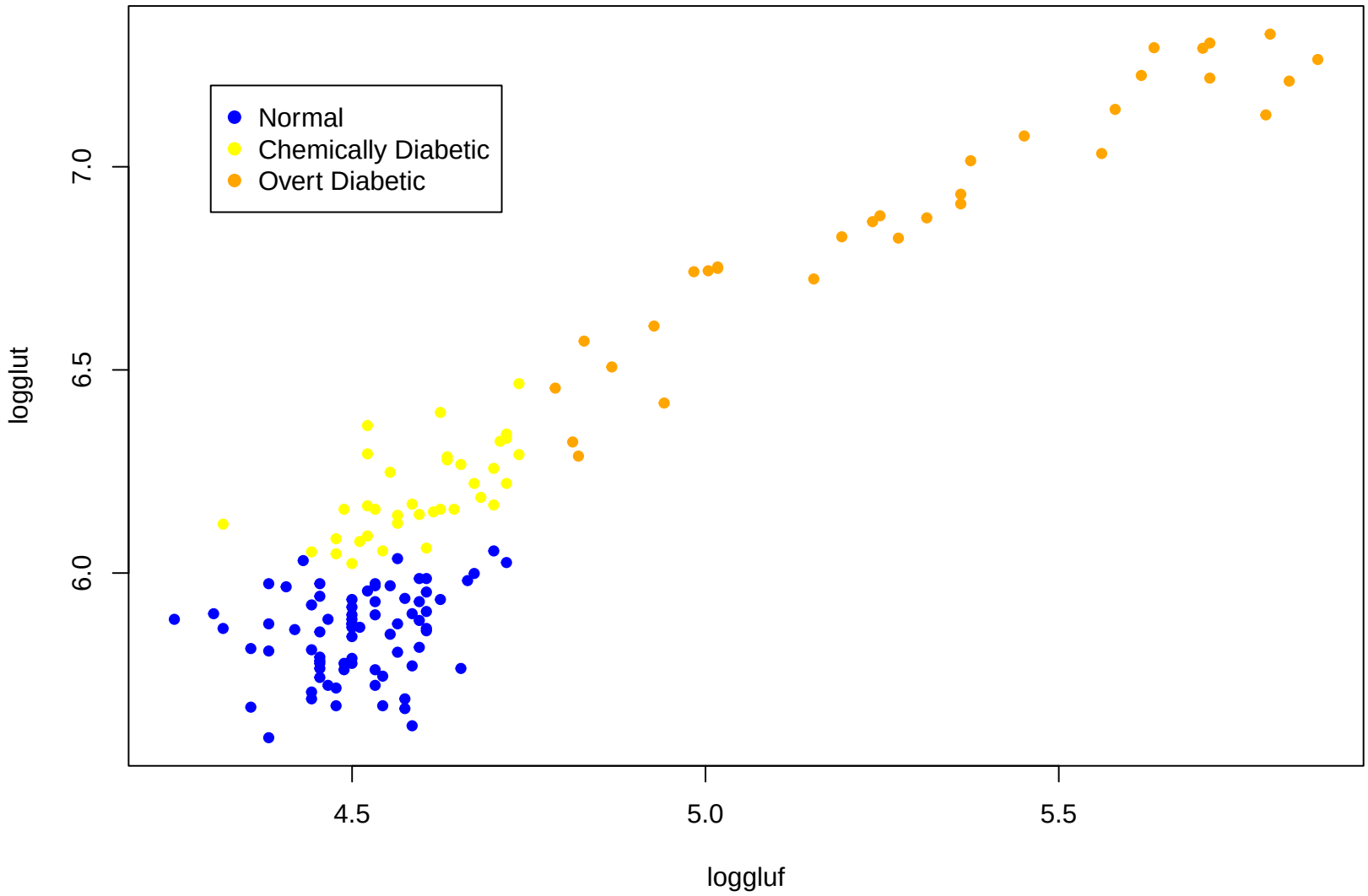
We have five models to confirm:

$$
\begin{aligned}
\text{MODEL 0} &: 1 \\
\text{MODEL 1} &: X_2 \\
\text{MODEL 2} &: X_1 \\
\text{MODEL 3} &: X_1 + X_2 \\
\text{MODEL 4} &: X_1 + X_2 + X_1 : X_2
\end{aligned}
$$

Diabetes data

```
> str(Diabetes)

+ 'data.frame': 144 obs. of  15 variables:
+  $ id          : num  1 2 3 4 5 6 7 8 9 10 ...
+  $ relwt       : num  0.81 0.95 0.94 1.04 1 0.76 0.91 1.1 0.99 0.78 ...
+  $ glufast     : num  80 97 105 90 90 86 100 85 97 97 ...
+  $ glutest     : num  356 289 319 356 323 381 350 301 379 296 ...
+  $ instest     : num  124 117 143 199 240 157 221 186 142 131 ...
+  $ sspg        : num  55 76 105 108 143 165 119 105 98 94 ...
+  $ group       : Factor w/ 3 levels "Overt Diabetic",..: 3 3 3 3 3 3 3 3 3 3 ...
+  $ loggluf     : num  4.38 4.57 4.65 4.5 4.5 ...
+  $ logglut     : num  5.87 5.67 5.77 5.87 5.78 ...
+  $ normal      : num  1 1 1 1 1 1 1 1 1 1 ...
+  $ overt       : num  0 0 0 0 0 0 0 0 0 0 ...
+  $ chemical    : num  0 0 0 0 0 0 0 0 0 0 ...
+  $ normalgluf  : num  4.38 4.57 4.65 4.5 4.5 ...
+  $ overtgluf   : num  0 0 0 0 0 0 0 0 0 0 ...
+  $ chemicalgluf: num  0 0 0 0 0 0 0 0 0 0 ...

> table(Diabetes$group)

+
+      Overt Diabetic Chemically Diabetic              Normal
+                  32                  36                  76
```

Diabetes data



Diabetes data

Example: Diabetes Data

MODEL 4 is the most complex model with 6 parameters

$$\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}, \beta_{30}, \beta_{31}$$

MODEL 4:

$$\mathbb{E}[Y|\mathbf{x}] = \begin{cases} \beta_{10} + \beta_{11}\mathbf{x}_1 & \text{GROUP 1} \\ \beta_{20} + \beta_{21}\mathbf{x}_1 & \text{GROUP 2} \\ \beta_{30} + \beta_{31}\mathbf{x}_1 & \text{GROUP 3} \end{cases}$$

All of the other models are nested inside Model 4; we can obtain them all by setting parameters equal to zero.

Diabetes data

Example: Diabetes Data

In the default parameterization:

β_{10}, β_{11}	Group 1 Intercept and Slope
$\beta_{20} = \beta_{10} + \delta_{20}$ $\beta_{30} = \beta_{10} + \delta_{30}$	Changes in the Intercepts for groups 2 and 3 are δ_{20} and δ_{30}
$\beta_{21} = \beta_{11} + \delta_{21}$ $\beta_{31} = \beta_{11} + \delta_{31}$	Changes in the Slopes fpr groups 2 and 3 are δ_{21} and δ_{31}

Thus the six new parameters are

$$\beta_{10}, \beta_{11}, \delta_{20}, \delta_{30}, \delta_{21}, \delta_{31}$$

Diabetes data

MODEL 0 $\quad \beta_{11} = 0$

$\delta_{20} = \delta_{30} = \delta_{21} = \delta_{31} = 0$

MODEL 1 $\quad \beta_{11} = \delta_{21} = \delta_{31} = 0$

MODEL 2 $\quad \delta_{20} = \delta_{30} = \delta_{21} = \delta_{31} = 0$

MODEL 3 $\quad \delta_{21} = \delta_{31} = 0$

Note: $\beta_{11} = 0 \Longrightarrow \delta_{21} = \delta_{31} = 0$, as X_1 is not included in the model.

Counting Parameters

Whenever we remove a

- continuous covariate, we set one parameter to zero.
- factor predictor with L levels, we set $L - 1$ parameters to zero.
- two-way interaction between the variables, we set

$$1 \times (L - 1) = L - 1$$

  parameters to zero.

Counting Parameters

Models 0,1,2,3 are nested inside Model 4.

Two approaches to finding the best model are used:

1. Start with Model 0 and try to add terms that improve the model fit (Forward Selection)
2. Start with Model 4 and try to remove terms that improve the model fit (Backward Selection)

Note:

- Models 0,1 and 2 are nested inside Model 3.
- Model 0 is nested inside Models 1 and 2.

Therefore we can begin with Model 4, or Model 3 or Model 1 or 2, and simplify to a nested model.

Example: Diabetes data

```
> fit0<-lm(logglut~1, data=Diabetes)                          #Model 0
> fit1<-lm(logglut~group, data=Diabetes)                      #Model 1
> fit2<-lm(logglut~loggluf, data=Diabetes)                    #Model 2
> fit3<-lm(logglut~loggluf+group, data=Diabetes)              #Model 3
> fit4<-lm(logglut~loggluf+group+loggluf:group, data=Diabetes)  #Model 4
```

SSE quantities:

```
> SSE.0<-sum(residuals(fit0)^2)
> SSE.1<-sum(residuals(fit1)^2)
> SSE.2<-sum(residuals(fit2)^2)
> SSE.3<-sum(residuals(fit3)^2)
> SSE.4<-sum(residuals(fit4)^2)
> c(SSE.0,SSE.1,SSE.2,SSE.3,SSE.4)

+ [1] 28.504333  4.159984  3.737941  1.472320  1.317650
```

Example: Diabetes data

Example: Diabetes Data

Here $n = 144$.

Model	Description	SSE	p
0	1	28.504	1
1	X_2	4.160	3
2	X_1	3.738	2
3	$X_1 + X_2$	1.472	4
4	$X_1 + X_2 + X_1 : X_2$	1.318	6

p is the number of non-zero parameters; k or g is always $p-1$ in the following analysis.

Example: Diabetes data

Backward Selection:

Complete Model : Model 4
Reduced Model : Model 3

Here $k = 5, g = 3$ so $k - g = 2$, and

$$n - k - 1 = 144 - 5 - 1 = 138.$$

We have

$$F_{obs} = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)} = \frac{(1.472 - 1.318)/2}{1.318/138} = 8.062$$

Example: Diabetes data

We compare this with the

$$F_{k-g,n-k-1} = F_{2,138}$$

distribution; we have $F_{0.05,2,138} = 3.061$, so we

$$\text{Reject } H_0 \text{ at } \alpha = 0.05.$$

Example: Diabetes data

That is, Model 4

$$X_1 + X_2 + X_1 : X_2$$

fits significantly better than Model 3

$$X_1 + X_2.$$

We cannot simplify the complete model to the reduced model without the loss of significant explanatory power.

The Interaction is Necessary in the Model

Backward selection stops here; we cannot simplify further from the complete model.

Example: Diabetes data

```
> anova(fit3, fit4, test='F')

+ Analysis of Variance Table
+
+ Model 1: logglut ~ loggluf + group
+ Model 2: logglut ~ loggluf + group + loggluf:group
+   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
+ 1    140 1.4723
+ 2    138 1.3176  2   0.15467 8.0994 0.0004721 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> qf(0.95, 2, 138)

+ [1] 3.061716
```

Example: Diabetes data

Forward Selection: we start with Model 0 and build up.

Model 1 vs Model 0 $\quad F_{obs} = 412.568$

Model 2 vs Model 0 $\quad F_{obs} = 940.846$

It seems that Model 2 is the better improvement, so we try the selection path

$$\text{Model 0} \longrightarrow \text{Model 2} \longrightarrow \text{Model 3} \longrightarrow \text{Model 4}$$

Example: Diabetes data

Model	SSE	$SSE_R - SSE_C$
0	28.504	-
2	3.738	24.766
3	1.472	2.266
4	1.318	0.154

We work down the table, $28.504 - 3.738 = 24.766$ etc.

Example: Diabetes data

Comparison	k	g	SSE_C	$SSE_R - SSE_C$	F_{obs}
2 vs 0	1	0	3.738	24.766	940.82
3 vs 2	3	1	1.472	2.266	107.76
4 vs 3	5	3	1.318	0.154	8.06

Recall that $n = 144$, and

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)}$$

Under each H_0,

$$F \sim F_{k-g,n-k-1}.$$

Example: Diabetes data

- $F_{0.05,1,142} \simeq 3.92 < 940.82$
  Therefore Model 0 is NOT an adequate simplification of Model 2
- $F_{0.05,2,140} \simeq 3.07 < 107.76$
  Therefore Model 2 is NOT an adequate simplification of Model 3
- $F_{0.05,2,138} \simeq 3.07 < 8.06$
  Therefore Model 3 is NOT an adequate simplification of Model 4

All of the null hypotheses are rejected.

Example: Diabetes data

By both forward and backward selection, we select Model 4

$$X_1 + X_2 + X_1 : X_2$$

as the most appropriate model.

Here the convention is not to correct for multiple testing (we don't know how many tests we are going to do), although a correction could be used.

F-tests for Unbalanced Designs

Example: Potato Damage Data

The damage to potato plants caused by cold temperatures is to be studied.

In this experimental study, three binary factor predictors were used

- we label them A, B and C rather than X_1, X_2, X_3 to recall the link with Factorial Designs in ANOVA.

F-tests for Unbalanced Designs

Each factor takes two levels:

Example: Potato Damage Data

	Factor	Levels
A	Potato Variety	0- Variety 1, 1- Variety 2
B	Acclimatization Routine	0- Room Temp, 1- Cold Room
C	Preparation Treatment	0- -4C, 1- -8C

Thus we have a $2 \times 2 \times 2$ three-way factorial design.

F-tests for Unbalanced Designs

However, the design is unbalanced; we have different numbers of replicates in each of the 8 factor-level combinations.

- we cannot use conventional 3-way ANOVA; the lack of balance means that the stated p-valuesm ay be misleading if we perform a standard ANOVA.

- we are forced to use the F-test model comparison approach.

F-tests for Unbalanced Designs

```
> str(Potato)

+ 'data.frame': 75 obs. of  5 variables:
+  $ potato: Factor w/ 2 levels "Variety 1","Variety 2": 1 1 1 1 1 1 1 1 1 1 ...
+  $ regime: Factor w/ 2 levels "Room Temp","Cold Room": 1 1 1 1 1 1 1 1 1 1 ...
+  $ temp  : Factor w/ 2 levels " -4 C"," -8 C": 1 1 1 1 1 2 2 2 2 2 ...
+  $ photo : num  4.64 5.79 29.35 18.42 2.26 ...
+  $ leak  : num  2.25 4.34 4.25 6.14 2.38 16.3 5.24 3.25 0.88 3.98 ...

> ftable(xtabs(~potato+regime+temp,data=Potato))

+                      temp  -4 C  -8 C
+ potato    regime
+ Variety 1 Room Temp          5     5
+           Cold Room         12    13
+ Variety 2 Room Temp         13    13
+           Cold Room          7     7
```

There are unequal numbers of replicates in the 8 treatments.

F-tests for Unbalanced Designs

```
> supernova(lm(leak~potato*regime*temp,data=Potato),type=2)

+  Analysis of Variance Table (Type II SS)
+  Model: leak ~ potato * regime * temp
+
+                                              SS df       MS      F   PRE     p
+  ----------------- --------------- | --------- -- -------- ------ ----- -----
+               Model (error reduced) |  8842.339  7 1263.191 17.033 .6402 .0000
+              potato                 |  1503.499  1 1503.499 20.273 .2323 .0000
+              regime                 |  1973.358  1 1973.358 26.609 .2843 .0000
+                temp                 |  1225.714  1 1225.714 16.527 .1979 .0001
+       potato:regime                 |  2093.162  1 2093.162 28.224 .2964 .0000
+         potato:temp                 |    35.447  1   35.447  0.478 .0071 .4917
+         regime:temp                 |     6.791  1    6.791  0.092 .0014 .7631
+  potato:regime:temp                 |    89.198  1   89.198  1.203 .0176 .2767
+               Error (from model)    |  4968.876 67   74.162
+  ----------------- --------------- | --------- -- -------- ------ ----- -----
+               Total (empty model)   | 13811.215 74  186.638
```

The ANOVA F-tests are not reliable due to the unbalanced design, so we must use the nested model F-tests.

F-tests for Unbalanced Designs

Here the most complex model can be written

$$A + B + C + A : B + A : C + B : C + A : B : C$$

that is,

- all main effects (terms 1,2 and 3)
- all two-way interactions (terms 4,5 and 6)
- all three-way interactions (term 7)

We may write this model

$$A * B * C$$

which is termed the full factorial model.

Counting the numbers of parameters

Term	Parameters	
A	$(a-1)$	1
B	$(b-1)$	1
C	$(c-1)$	1
$A:B$	$(a-1)(b-1)$	1
$A:C$	$(a-1)(c-1)$	1
$B:C$	$(b-1)(c-1)$	1
$A:B:C$	$(a-1)(b-1)(c-1)$	1
Total		7

where $a = b = c = 2$.

We have 7 parameters in total (excluding the baseline mean) when all terms are considered, so

$$k = 7$$

Counting the numbers of parameters

In the following tables columns are:

- Complete Model
- Reduced Model
- SSE_C
- SSE_R
- k
- g
- F_{obs} (test statistic)
- $F_{0.05,k-g,n-k-1}$

We denote the critical value by F_α and check whether $F > F_\alpha$.

Potato Damage Data: ANOVA F-Tests

We compare four models: M_{R_1}, M_{R_2} and M_{R_3} are nested within the complete model M_C.

$$
\begin{aligned}
M_C &: A + B + C + A:B + A:C + B:C + A:B:C \\
M_{R_1} &: A + B + C + A:B \\
M_{R_2} &: A + B + C \\
M_{R_3} &: A + B + A:B
\end{aligned}
$$

Potato Damage Data: ANOVA F-Tests

C	R	SSE_C	SSE_R	k	g	F_{obs}	F_α
M_C	M_{R_1}	4968.876	5093.746	7	4	0.561	2.76
M_{R_1}	M_{R_2}	5093.746	7183.674	4	3	28.721	3.92
M_{R_1}	M_{R_3}	5093.746	6319.640	4	3	16.846	3.92

Potato Damage Data: ANOVA F-Tests

Note: The quoted F_α values are approximate as the textbook does not tabulate all F distributions. We take $\alpha = 0.05$

Conclusions

Taking the comparisons in order:

1. M_C *vs* M_{R_1} : $F < F_\alpha$. This result is not significant: Model M_{R_1} is an adequate simplification of Model M_C, and we choose M_{R_1} over M_C.

   The model M_{R_1} now becomes the complete model.

2. M_{R_1} *vs* M_{R_2} : $F > F_\alpha$. This result is significant: Model M_{R_2} is not an adequate simplification of Model M_{R_1}

3. M_{R_1} *vs* M_{R_3} : $F > F_\alpha$. This result is significant: Model M_{R_3} is not an adequate simplification of Model M_{R_1}

Conclusions

Thus the final model is

$$A + B + C + A : B$$

i.e. all main effects, plus the interaction between potato variety and acclimatization routine.

We cannot simplify this model further without significant loss in terms of goodness of fit.

Note: $R^2 = 0.631$ and $R^2_{adj} = 0.610$, so we have a reasonable fit.

Example: Task Distraction Data

Example: Task Distraction Data

In an experimental study, the number of errors made in performing a specified task was recorded. The experiment investigated the influence of various predictors on the numbers of errors made.

There are two factor predictors (A, B) and one continuous covariate (X).

We have a balanced design with 15 people (replicates) in each factor-level subgroup.

Example: Task Distraction Data

Example: Task Distraction Data

A	Group	1 : Non-smoker
		2 : Delayed smoker
		3 : Active smoker
B	Task	1 : Pattern Recognition
		2 : Cognitive Task
		3 : Driving Simulation
X	Distraction Level	

Example: Task Distraction Data

We compare four models with the complete model

Complete Model : $A * B * X$

$$A + B + X + A : B + A : X + B : X + A : B : X$$

Number of parameters

Term	Parameters		Tot.
A	$(a-1)$	$= 3 - 1$	2
B	$(b-1)$	$= 3 - 1$	2
X	(1)		1
$A : B$	$(a-1)(b-1)$	$= 2 \times 2$	4
$A : X$	$(a-1)(1)$	$= 2 \times 1$	2
$B : X$	$(b-1)(1)$	$= 2 \times 1$	2
$A : B : X$	$(a-1)(b-1)(c-1)$	$= 2 \times 2 \times 1$	4
Total			17

Example: Task Distraction Data

We consider the following sequence of models:

- Reduced Model 1: M_{R_1}

$$A + B + X + A : X + B : X$$

- Reduced Model 2: M_{R_2}

$$A + B + X + B : X$$

- Reduced Model 3: M_{R_3}

$$B + X + B : X$$

- Reduced Model 4: M_{R_4}

$$B + X$$

Task Distraction Data: ANOVA F-Tests

$$
\begin{aligned}
M_C &: A + B + X + A:B + A:X + B:X + A:B:X \\
M_{R_1} &: A + B + X + A:X + B:X \\
M_{R_2} &: A + B + X + B:X \\
M_{R_3} &: B + X + B:X \\
M_{R_4} &: B + X
\end{aligned}
$$

C	R	SSE_C	SSE_R	k	g	F	F_α
M_C	M_{R_1}	5660.010	7627.479	17	9	5.084	2.02
M_{R_1}	M_{R_2}	7627.479	7971.274	9	7	2.817	3.07
M_{R_2}	M_{R_3}	7971.274	8404.654	7	5	3.452	3.07
M_{R_3}	M_{R_4}	8404.654	11154.715	5	3	21.105	3.07

Conclusions

Taking the comparisons in order:

1. M_C *vs* M_{R_1} : $F > F_\alpha$. This result is significant: Model M_{R_1} is not an adequate simplification of Model M_C

2. M_{R_1} *vs* M_{R_2} : $F < F_\alpha$. This result is not significant: Model M_{R_2} is an adequate simplification of Model M_{R_1}

3. M_{R_2} *vs* M_{R_3} : $F > F_\alpha$. This result is significant: Model M_{R_3} is not an adequate simplification of Model M_{R_2}

4. M_{R_3} *vs* M_{R_4} : $F > F_\alpha$. This result is significant: Model M_{R_4} is not an adequate simplification of Model M_{R_3}

Follow-up Analysis

In a follow-up analysis, it transpires that the model

$$A + B + X + A : B + A : X + B : X$$

ie selected.

Note: $R^2 = 0.863$ and $R^2_{adj} = 0.831$, so we have a good fit.

Note: we must take great care with the sequence of models.

Pitfalls of Regression Modelling

Five issues to bear in mind in ANOVA, Regression and General Linear Modelling.

1. Model assumptions
2. Data transformations
3. Model selection
4. Multicollinearity
5. Predicting beyond the range of the covariates

See textbook, section 12.12.