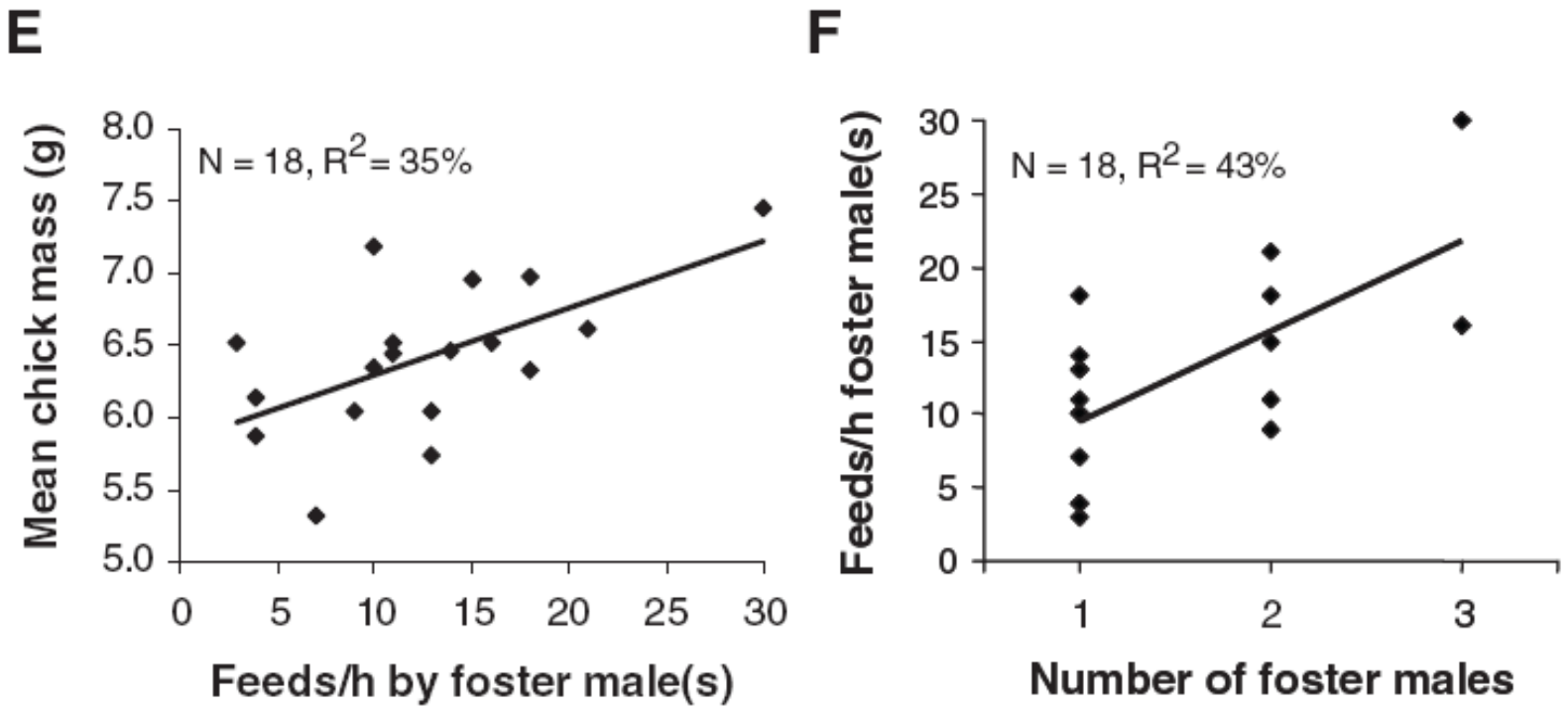
Outline

Chick mass example

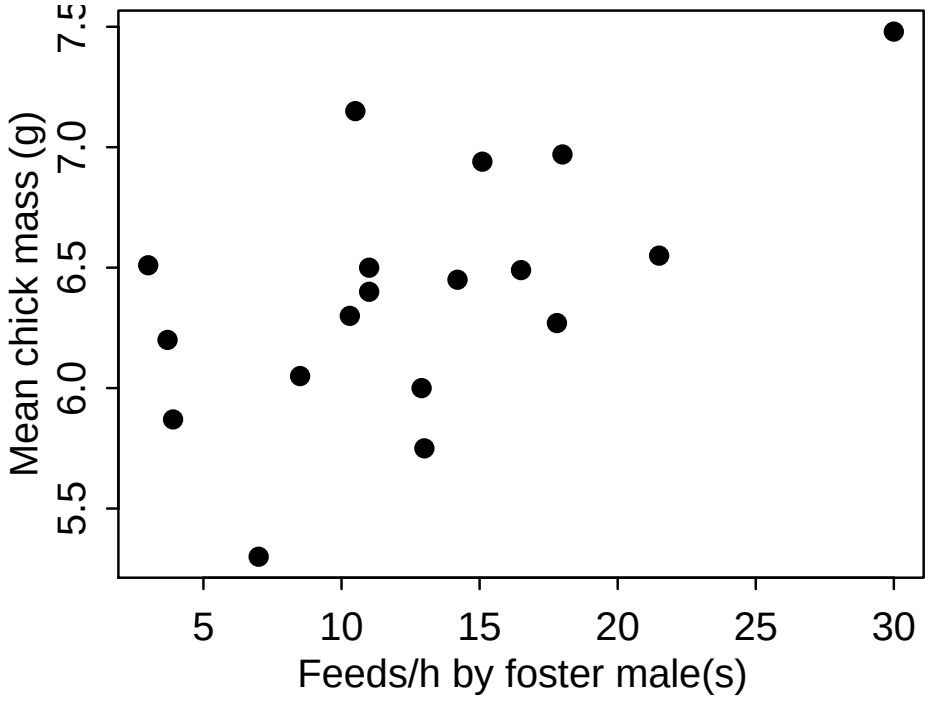
Data collected from 18 nests, superb fairy wrens



Chick mass example

```
> russell
 feed mass
  7.0 5.30
 16.5 6.49
  3.7 6.20
 10.5 7.15
 12.9 6.00
 18.0 6.97
 14.2 6.45
  8.5 6.05
 17.8 6.27
  3.9 5.87
 11.0 6.50
 11.0 6.40
  3.0 6.51
 21.5 6.55
 15.1 6.94
 13.0 5.75
 10.3 6.30
 30.0 7.48
```

Relationship between provisioning rate of foster males and chick mass?



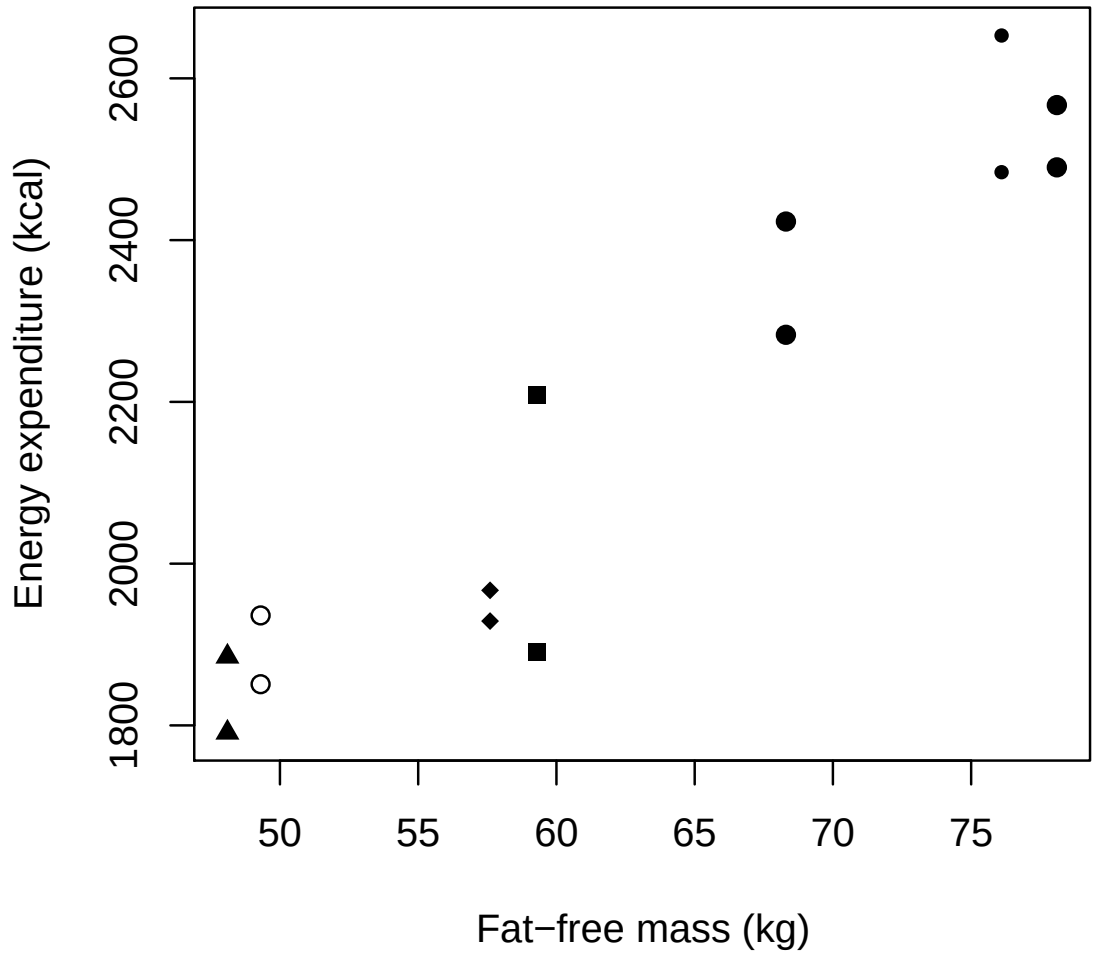
Nutritional requirement and body size

Does nutritional requirement depend on body size? How?

Expt: 7 men, 24-hour energy expenditure (kcal) was measured, in conditions of quiet sedentary activity, repeated twice.

Subject #	fat-free mass (kg)	energy expenditure (kcal)	
1	49.3	1851	1936
2	59.3	2209	1891
3	68.3	2283	2423
4	48.1	1885	1791
5	57.6	1929	1967
6	78.1	2490	2567
7	76.1	2484	2653

Nutritional requirement and body size



Find formula to predict energy expenditure (then nutritional requirement) as a function of body size.

Objectives and Regression line

Objectives:

describe the relationship between feeding rate by foster male(s) (x) and chick mass (y), or between mass (x) and energy expenditure (y)

predict mass of a new chick with foster male(s) providing a given feeding rate, or energy expediture on a new day, given the person's fat-free mass.

Main idea of simple linear regression: fit data with a straight line

$$y = b_0 + b_1 x$$

b_0 is the intercept and b_1 is the slope.

Goal: find b_0, b_1 for the best fitting line.

Least squares approach.

Least squares

Find b_0, b_1 that minimize the sum of squares

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$$

y_i is the observed value, $\hat{y}_i = b_0 + \hat{b}_1 x_i$ is the fitted value.

The best fitting line has

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}, \\ b_0 &= \bar{y} - b_1 \bar{x}. \end{aligned}$$

Chick mass example

$\sum_{i=1}^{n} x_i = 227.9$, $\sum x_i^2 = 3658$, $\sum x_i y_i = 1493.7$ g, $(n = 18)$

$\sum y_i = 115.18$ g, $\sum y_i^2 = 741.69$ g^2, thus

$$\bar{x} = \qquad = 12.67 \text{ feeds/h}, \quad \bar{y} = \qquad = 6.40 \text{ g}$$

$$\sum (x_i - \bar{x})^2 = \qquad = 772.43$$

$$b_1 = \qquad = .0458 \text{ g/feed}$$

$$b_0 = \qquad = 5.82 \text{ g}$$

Best fitting line

$y = 5.82 + 0.0458 * x$, or:

chick mass in g $= 5.82$ g $+0.0458*$ # feeds/h

Chick mass example

y = chick mass (response), x =# feeds/h (predictor): do not play the same role. The regression equation is **not symmetric**.

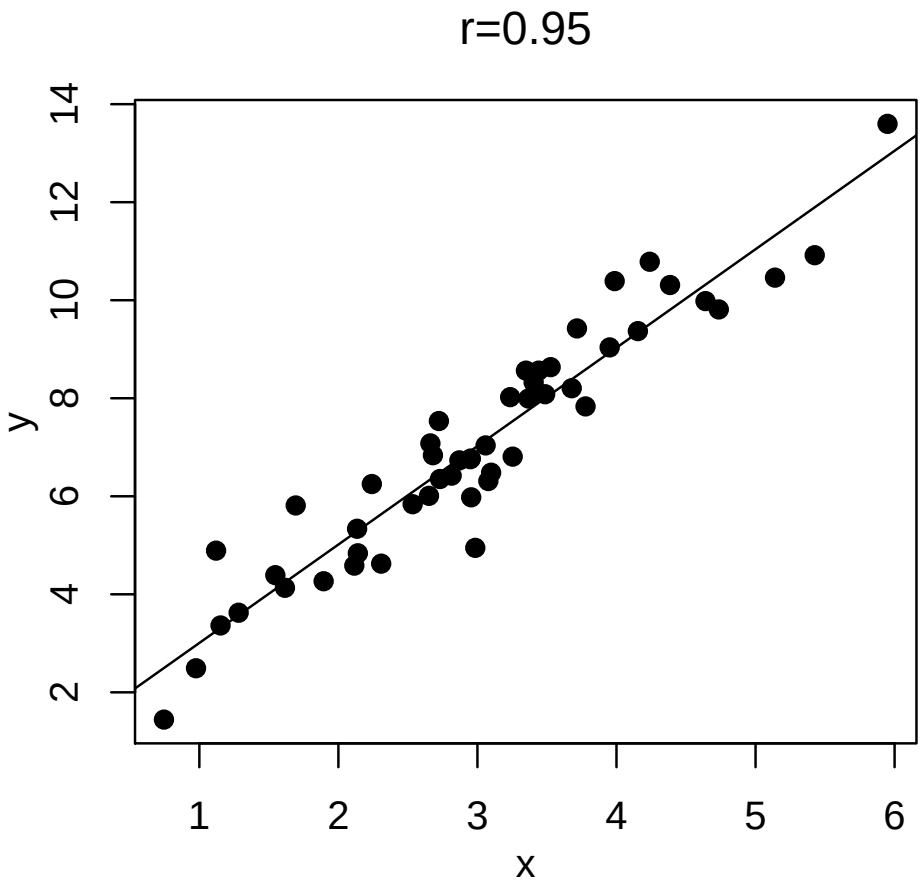
$$y = 5.82 + 0.0458 * x$$

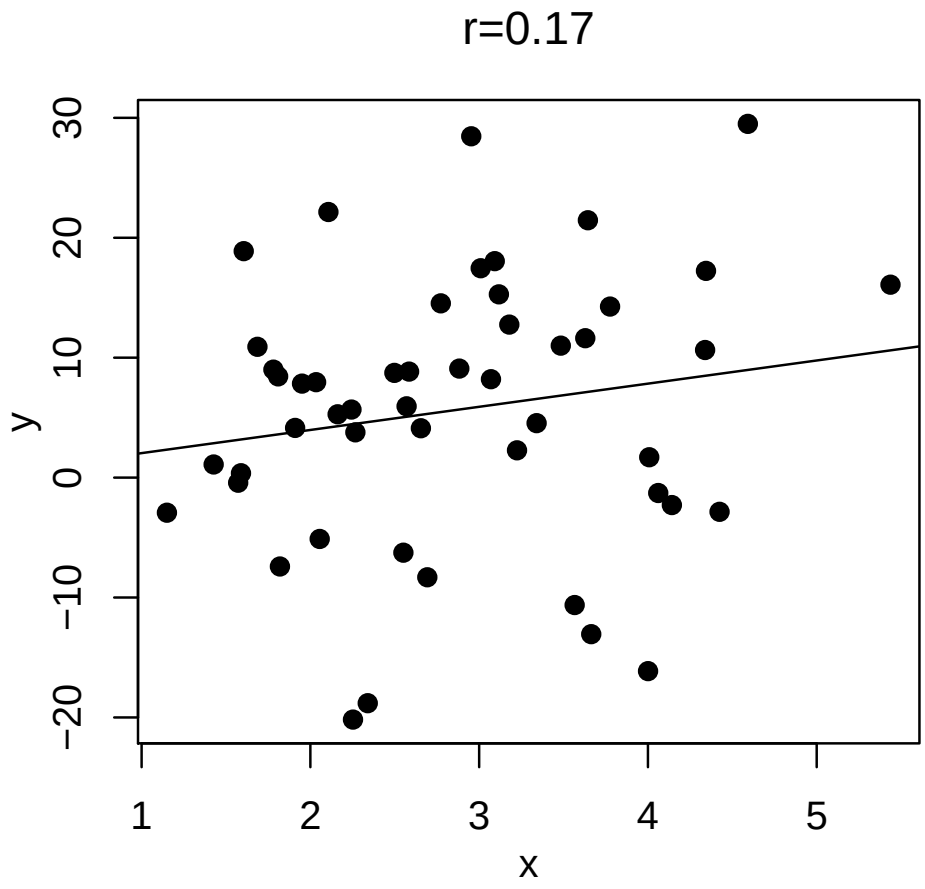
We can now predict y if we know a new x value. Example:

at $x = 20$ feeds/h:	prediction $\hat{y} =$	$= 6.73$ g.
at $x = 0$ feeds/h:	prediction $\hat{y} =$	
at $x = 100$ feeds/h:	prediction $\hat{y} =$	$= 10.4$ g.

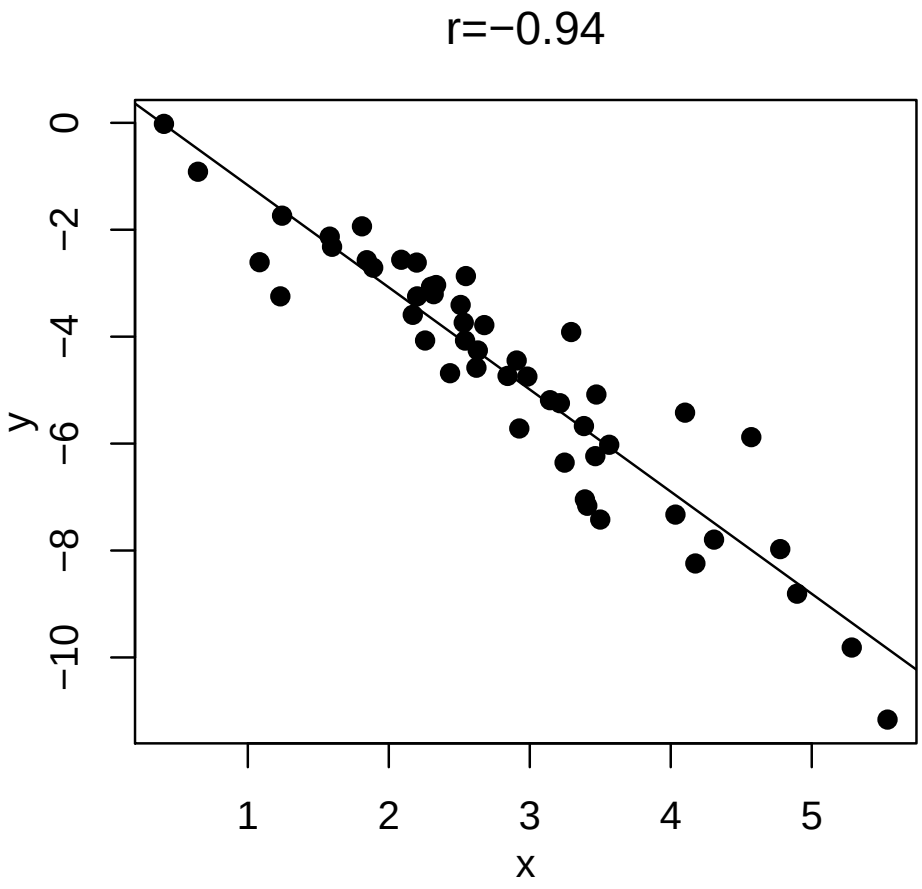
Warning: be very cautious about predictions outside the range of original data (3-30 feeds/h here).

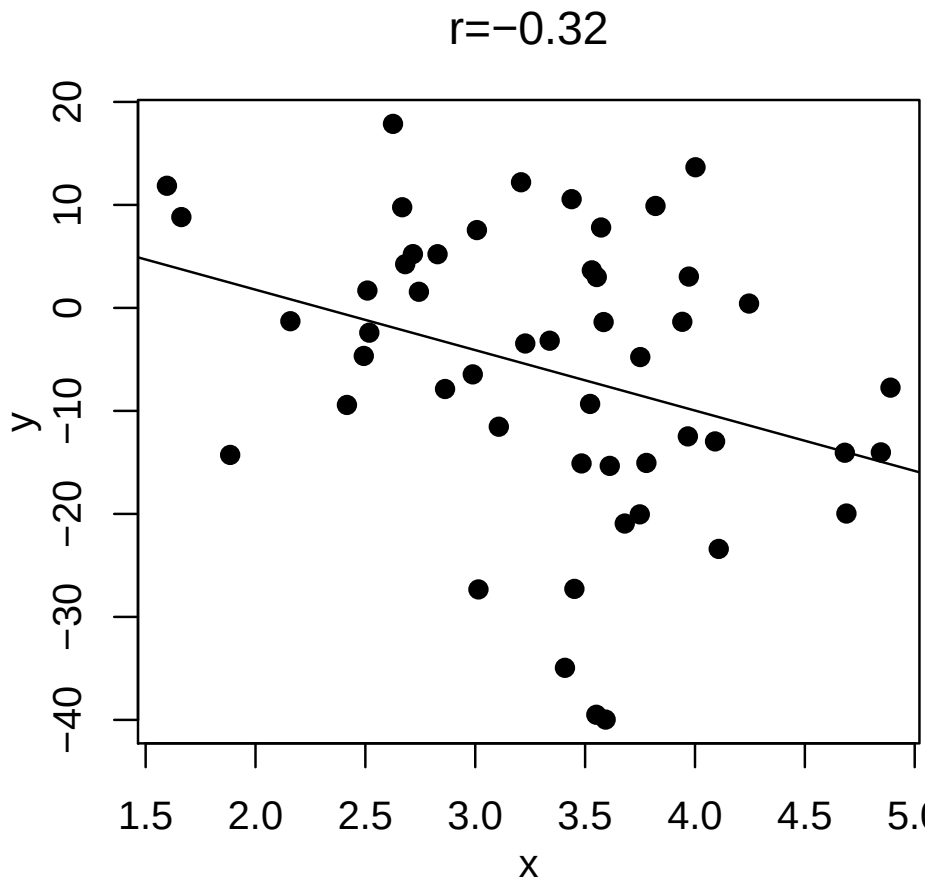
Correlation: examples



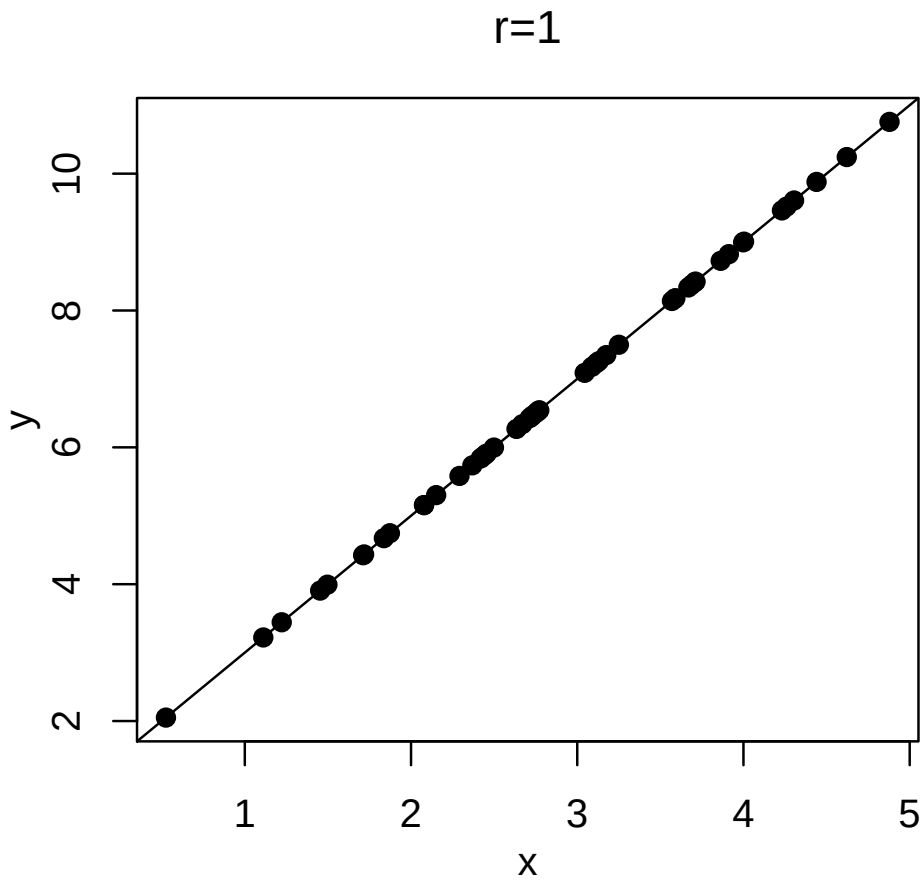


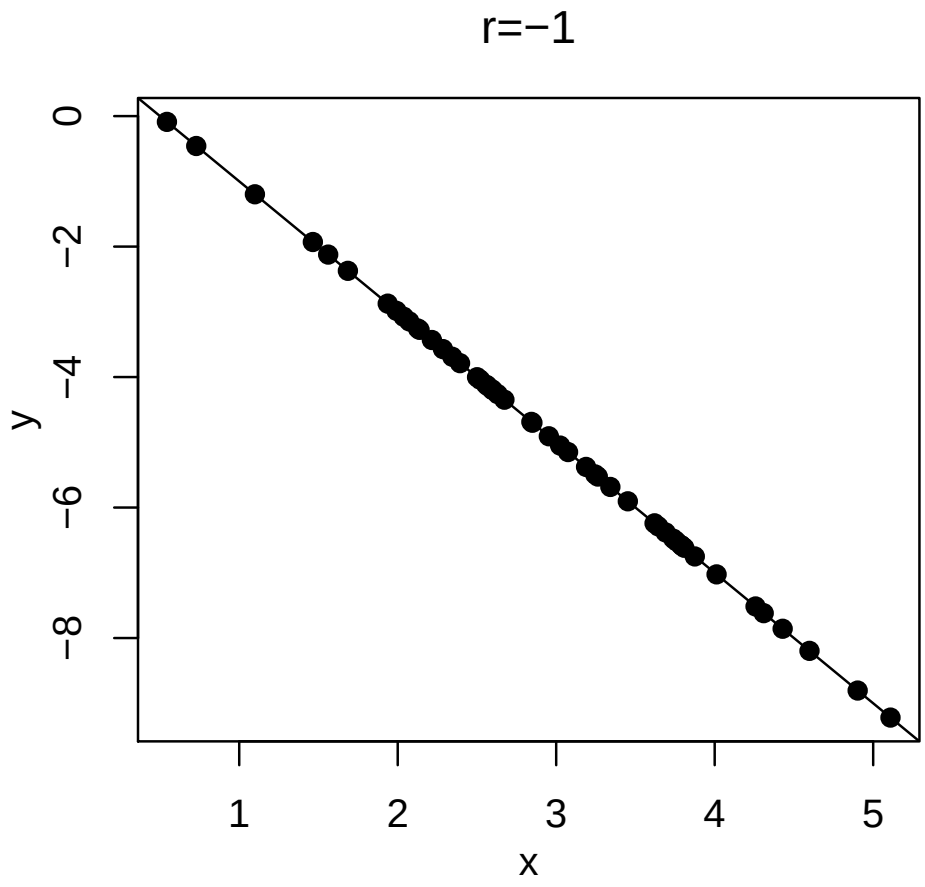
Correlation: examples



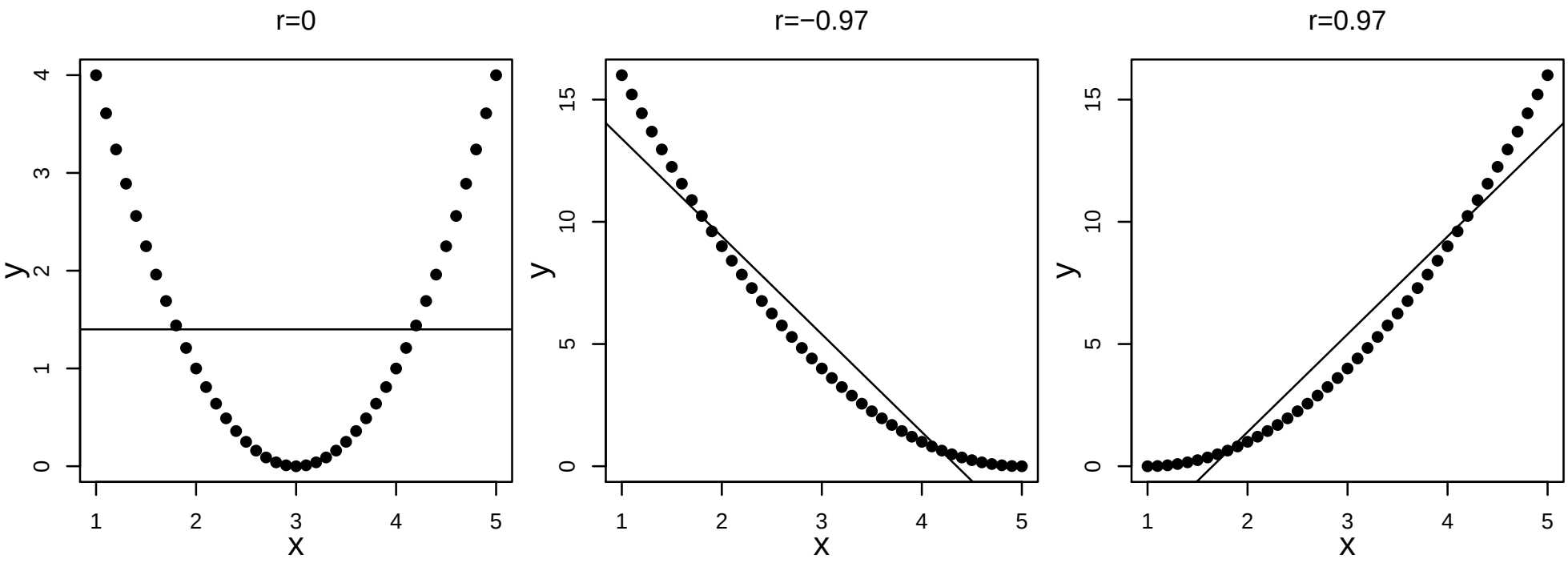


Correlation: examples





Correlation: examples



Correlation coefficient r

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, s_x\, s_y}$$

Notice that $b_1 = r\dfrac{s_y}{s_x}$, or $r = b_1\dfrac{s_x}{s_y}$.

r has no dimension
r always between -1 and 1
symmetric in X and Y.

measures the strength of the *linear* relationship between x and y. Does *not* measure non-linear relationships.

R: `lm` for Linear Model, and `cor` for CORrelation

```
> russell = read.table("../data/russell.txt", header=T)
> str(russell)
'data.frame':   18 obs. of  2 variables:
 $ feed: num  7 16.5 3.7 10.5 12.9 18 14.2 8.5 17.8 3.9 ...
 $ mass: num  5.3 6.49 6.2 7.15 6 6.97 6.45 6.05 6.27 5.87 ...

> fit = lm(mass~feed, data=russell) # lm = linear model
> fit
Coefficients:
(Intercept)         feed
    5.81935      0.04577

> 5.81935 + 0.04577 * 20
[1] 6.73475

> predict(fit, data.frame(feed=c(0,20,100)) )
        1         2         3
 5.819350  6.734813 10.396665
> cor(russell$mass, russell$feed)
[1] 0.5890986
> with(russell,  cor(mass,feed) )
[1] 0.5890986
```

How about uncertainty?

What is the uncertainty in the fitted line (intercept and slope), and in predicted mass values?

We got: chick mass in g $=$ 5.82 g $+0.0458*$ # feeds/h

Is there real trend? that is, do we have evidence that the slope in the entire population is $\neq 0$? We observe $b_1 = 0.0458$ from our sample.

or prediction: $\hat{y} = 6.40$ g at 20 feeds/h.

Prediction interval instead of a single number 6.40 g?

Linear Regression model

We consider y values as comming from a random variable Y:

The simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

where $e_i \sim$ iid $\mathcal{N}(0, \sigma_e^2)$, $i = 1, \ldots, n$.

Y is called a dependent variable or response variable.

x is called an independent variable or covariate.

e's are called errors.

Assumptions

$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

1. The straight line relationship between y and x is correct: the curve that describes the average trend is straight: not curved.
2. Errors e_i are independent.
3. Errors e_i have homogeneous variance: $\text{var}(e_i) = \sigma_e^2$.
4. Errors e_i have normal distribution: $e_i \sim \mathcal{N}(0, \sigma_e^2)$.

σ_e^2 is sometimes written as σ^2.

Testing the presence of a trend

$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

If $\beta_1 = 0$, then the model becomes $Y_i = \beta_0 + e_i$: the Y values are independent of the X values, no trend up or down for Y when X increases.

We need a test for the null hypothesis $H_0 : \beta_1 = 0$, meaning no linear association between X and Y.

Partitionning the variation

Total variation: dfTot $= n - 1$ and

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

Variation explained by the trend: dfReg $= 1$ and

$$\text{SSReg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = b_1\left[\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)\right]$$

$$= \quad b_1 \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \quad = \quad r^2(n-1)s_y^2$$

Residual variation: dfErr $= n - 2$ and

$$\text{SSErr} \quad = \quad \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \quad (1 - r^2)\sum(y_i - \bar{y})^2 \quad = \quad (1 - r^2)\,(n-1)s_y^2$$

ANOVA for regression, Coefficient of determination

SSTotal $=$ SSReg $+$ SSErr

Source	df	SS	MS	F
Regression		1.62		8.50
Error		3.04	0.19	–
Total		4.66	–	–

Coefficient of determination r^2: proportion of the total variation explained by the linear regression. r = correlation coefficient

$$r^2 = \frac{\text{SS Regression}}{\text{SS Total}}$$

Chicks: we had $r = 0.59$. Indeed $r^2 =$ $= .347 = (.59)^2$.

34.7% of the variation in chick mass is explained by the variation in #feeds/h by foster males.

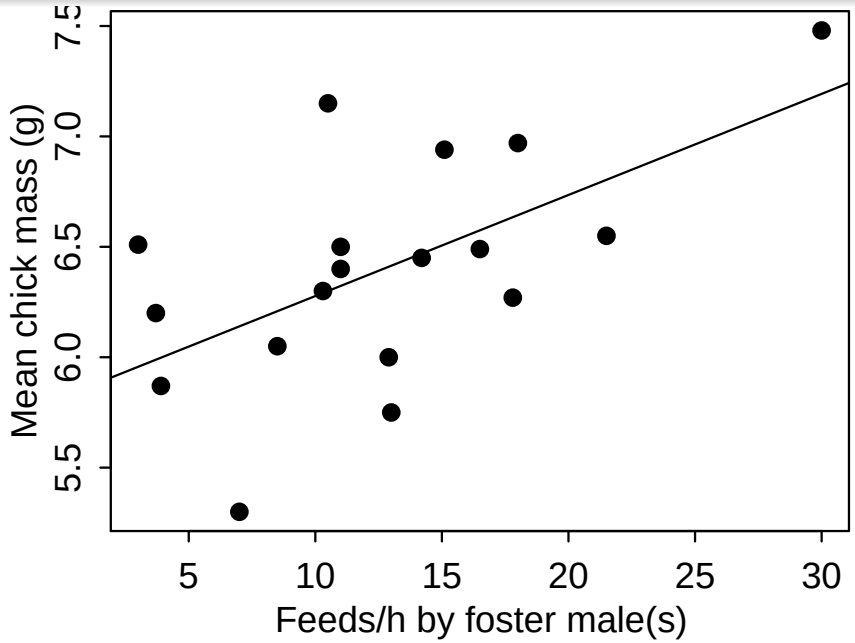
ANOVA for regression, Residual variance

Source	df	SS	MS	F
Regression	1	1.62	1.62	8.50
Error	16	3.04	0.19	–
Total	17	4.66	–	–

Residual variation

Estimate σ_e^2 by s_e^2 = MSErr on df = dfErr = $n - 2$.
It is the typical variation of the data points above and below the regression line.

Here we get $s_e^2 =$
$s_e = 0.436$ g on df=



ANOVA for testing absence of trend, H_0: $b_1 = 0$

Source	df	SS	MS	F
Regression	1	1.62	1.62	8.50
Error	16	3.04	0.19	–
Total	17	4.66	–	–

F-test for the slope

Under the hypothesis of no trend H_0: $\beta_1 = 0$, the F value has an $F_{1,n-2}$ distribution.

$f = 8.50$, compared to F on df$=$ p-value≈ 0.01.

Strong evidence against no trend, i.e. against H_0: $\beta_1 = 0$. Evidence that feeding rate by foster male(s) *positively* influences chick mass ($\beta_1 > 0$).

Relationship p-value for a trend and correlation r

r^2 might be very high and still F-test fails to reject "no trend" (large p-value).

Example: data set with only 2 points.

Or r^2 might be very low and still the test may detect a trend (very small p-value). It all depends on the sample size.

Example (hypothetical): very large data set on $y =$ people's weight regressed on $x =$ average ice cream consumption per week.

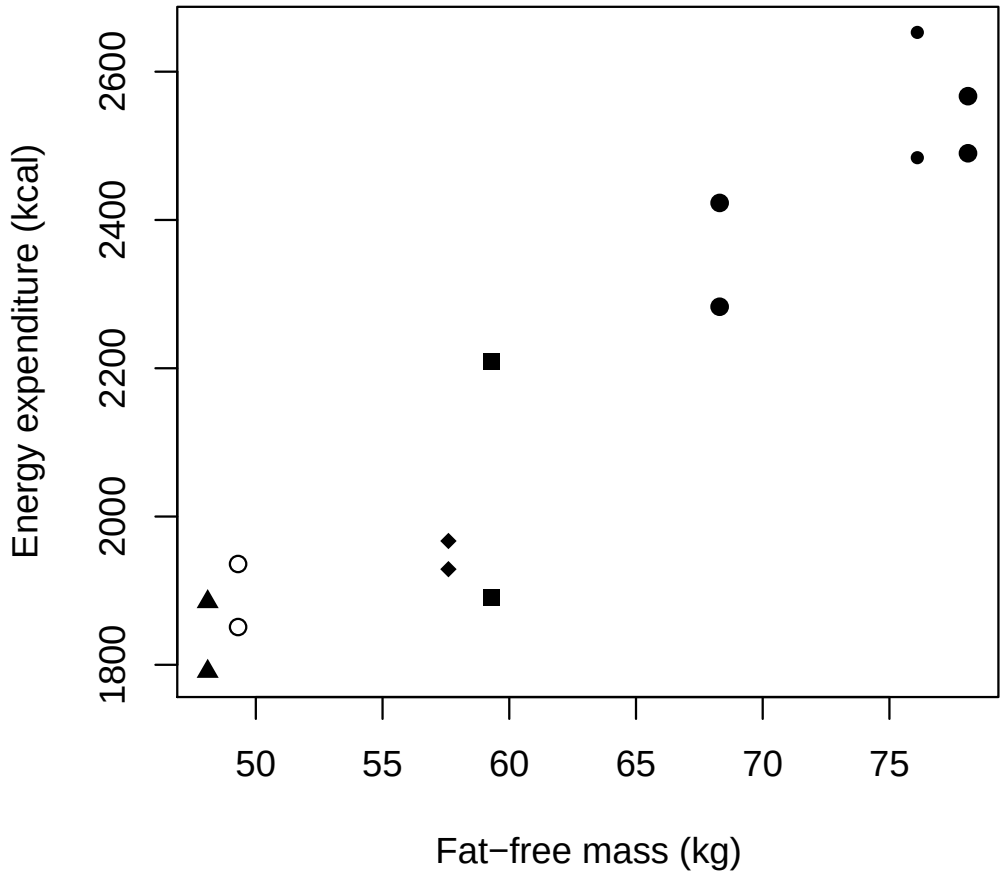
R: lm for Linear Model and anova for the table

```
> russell
   feed mass
1   7.0 5.30
2  16.5 6.49
3   3.7 6.20
...     ...
16 13.0 5.75
17 10.3 6.30
18 30.0 7.48

> fit = lm(mass~feed, data=russell) # lm = linear model
> anova(fit)                        # get ANOVA table

Analysis of Variance Table

Response: mass
          Df  Sum Sq Mean Sq F value  Pr(>F)
feed       1 1.61837 1.61837  8.5037 0.01010 *
Residuals 16 3.04501 0.19031
```

Does nutritional requirement depend on body size?

Expt: 7 men, 24-hour energy expenditure (kcal) measured twice, in conditions of quiet sedentary activity.

	fat-free mass (kg)	energy expenditure (kcal)	
1	49.3	1851	1936
2	59.3	2209	1891
3	68.3	2283	2423
4	48.1	1885	1791
5	57.6	1929	1967
6	78.1	2490	2567
7	76.1	2484	2653



Does nutritional requirement depend on body size?

$x =$ mass, $e1, e2$= first and second energy expenditure measurements. Using $\bar{x} = 62.4$ kg, $\sum(x_i - \bar{x})^2 = 877.74$,
$\bar{e1} = 2161.6$ kcal, $\bar{e2} = 2175.4$ kcal,
$\sum(e1_i - \bar{e1})^2 = 455855.7$, $\sum(x_i - \bar{x})(e1_i - \bar{e1}) = 19282.8$,
$\sum(e2_i - \bar{e2})^2 = 772147.7$, $\sum(x_i - \bar{x})(e2_i - \bar{e2}) = 24667.1$

first measurements:

$$b_1 = \frac{19282.8}{\sqrt{877.74}} = 21.97 \text{ kcal/kg}$$
$$b_0 = 2161.6 - 21.97 * 62.4 = 790.7 \text{ kcal}$$

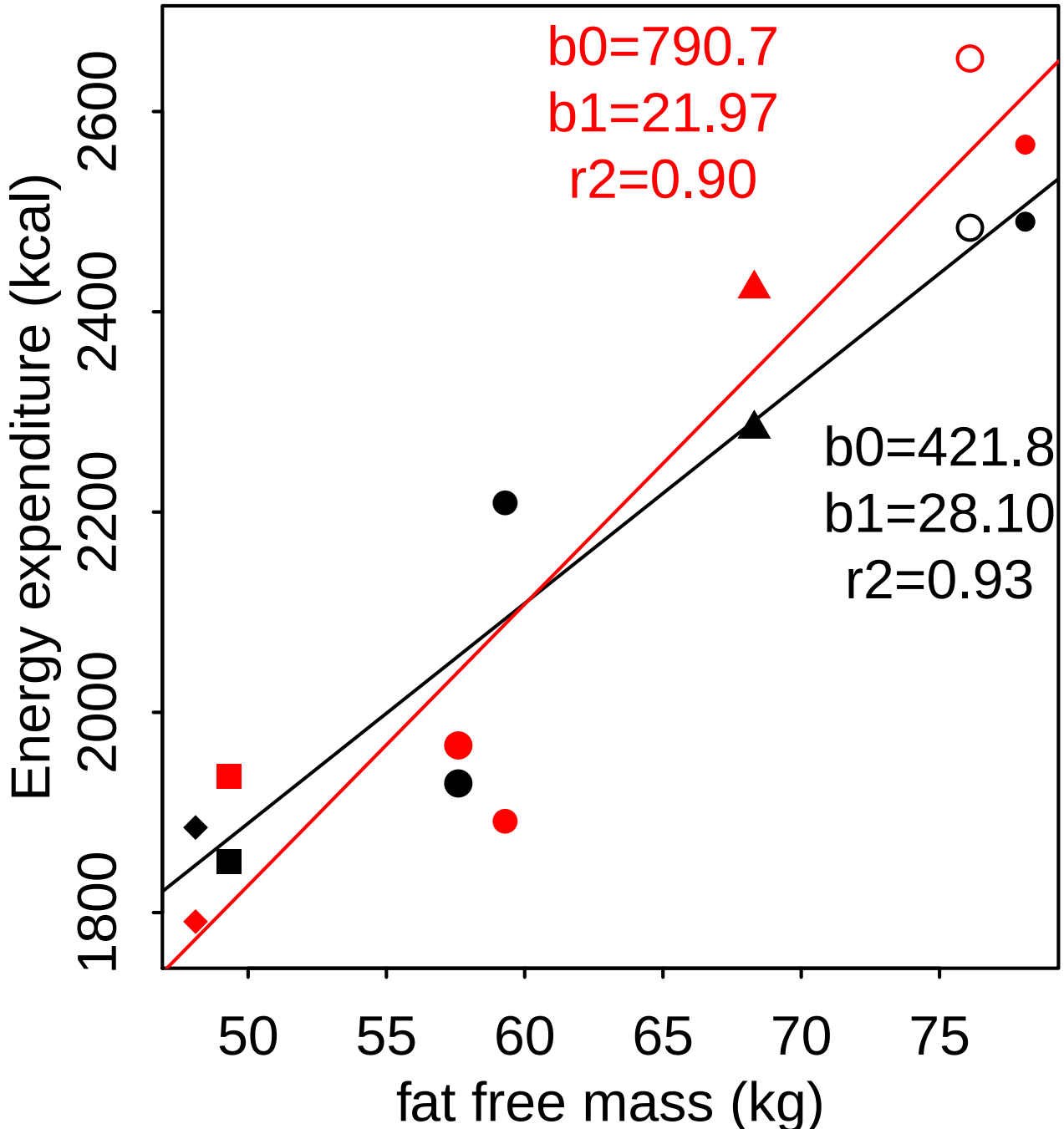
correlation:
$$r = \frac{19282.8}{\sqrt{877.74 * 455855.7}} = 0.964$$

second measurements:

$$b_1 = \frac{24667.1}{\sqrt{877.74}} = 28.10 \text{ kcal/kg}$$
$$b_0 = 2161.6 - 21.97 * 62.4 = 421.8 \text{ kcal}$$

correlation:
$$r = \frac{24667.1}{\sqrt{877.74 * 772147.7}} = 0.947$$

Regression lines



graph with R: scatterplot and regression lines

```
fit1 = lm(e1~mass, data=dat) # fit the regression lines first
fit2 = lm(e2~mass, data=dat)

# next: plot both sets of data, one after the other
plot(e1~mass, data=dat, pch=15:21, ylim=c(1780,2670),
 ylab="Energy expenditure (kcal)", xlab="fat free mass (kg)")

points(e2~mass, data=dat, pch=15:21, col=2)

# then add the lines, taken from the regression fits above
abline(fit1)
abline(fit2, col=2)

# adding some text next:
text(x=c(65,75), y=c(2600,2200), col=2:1,
 c("b0=790.7\nb1=21.97\nr2=0.90","b0=421.8\nb1=28.10\nr2=0.93"))
```

R commands and AVOVA table

```
> dat
  Subject mass   e1   e2
1       1 49.3 1851 1936
2       2 59.3 2209 1891
3       3 68.3 2283 2423
4       4 48.1 1885 1791
5       5 57.6 1929 1967
6       6 78.1 2490 2567
7       7 76.1 2484 2653

> fit1 = lm(e1~mass, data=dat)
> anova(fit1)
Response: e1
          Df Sum Sq Mean Sq F value    Pr(>F)
mass       1 423618  423618  65.702 0.0004635 ***
Residuals  5  32238    6448

> fit2 = lm(e2~mass, data=dat)
> anova(fit2)
Response: e2
          Df Sum Sq Mean Sq F value   Pr(>F)
mass       1 693219  693219  43.914 0.001178 **
Residuals  5  78929   15786
```

T-test for testing the absence of trend

Slope b_1 is a random variable: varies from experiment to expt.

Standard error of the slope

$$\text{SE}_{b_1} = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{s_e}{\sqrt{n-1}\, s_x}$$

This tells us how close b_1 is from β_1, typically.

Chicks: We had $s_e = 0.436$ g and
$s_x = \sqrt{772.43/(18-1)} = 6.74$ feeds/h.
The slope was estimated at $b_1 = .0458$ g/feed. SE of this estimate:

$$\text{SE}_{b_1} = \qquad\qquad\qquad\qquad = .0157$$

T-test for testing the absence of trend

1. H_0: no trend, i.e. horizontal slope: $\beta_1 = 0$.
2. Test statistic: $t = \dfrac{b_1}{\text{SE}_{b_1}}$
3. Under H_0, this t-value has a t-distribution with df = $n - 2$ (was dfErr in the ANOVA table).
   p-value $= \mathbb{P}\{T_{n-2} > t \text{ or } < -t\}$.

Chicks: $t = \dfrac{.0458}{.0157} = 2.916$. Against t distribution with df$=$ we get p-value≈ 0.01. Strong evidence for $b_1 \neq 0$.

Note that $t^2 = 2.916^2 = 8.503 = f$.
F-test or t-test for $\beta_1 = 0$: **same p-value, same conclusion.**

(Same conclusion also as test of H_0: $r = 0$ from Chapter 16, which we don't cover.)

Confidence intervals for population slope β_1

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

$$b_1 \pm t_{\alpha/2,n-2}\,\mathrm{SE}_{b_1}.$$

Chicks: for 95% confidence we use multiplier $t = 2.11$ and get

$$.0458 \pm 2.11 * .0157$$

i.e. $.0458 \pm .0331$ or $(.0125, .0790)$ g/feed.

Testing the intercept: t-test for H_0: $\beta_0 = 0$

Standard error of the intercept

b_0 has a normal distribution around the true β_0 with

$$\mathrm{SE}_{b_0} = s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

A $(1 - \alpha)$ confidence interval for β_0 is $b_0 \pm t_{\alpha/2, n-2}\, \mathrm{SE}_{b_0}$.

Chicks: we had $b_0 = 5.82$g. What is its standard error?

$$\mathrm{SE}_{b_0} = 0.436 * \sqrt{\frac{1}{18} + \frac{12.67^2}{772.43}} = .224 \text{ g}$$

For 95% confidence we use multiplier $t = 2.11$ again:
confidence interval for true intercept β_0:

$$5.82 \pm 2.11 * .224$$

i.e. 5.82 ± 0.473 or $(5.34, 6.30)$ g (intercept at 0 feeds/h).

Testing the intercept: t-test for H_0: $\beta_0 = 0$

1. H_0: 0 intercept, i.e. $\beta_0 = 0$, i.e. mean 0 at $x = 0$.
2. Test statistic: $t = \dfrac{b_0}{\mathrm{SE}_{b_0}}$
3. Under H_0, this t-value has a t-distribution with df = $n - 2$ (was dfErr in the ANOVA table).
   p-value $= \mathbb{P}\{T_{n-2} > t \text{ or } < -t\}$.

Chicks: $t =$ $= 26.0$ and p-value $\ll .001$ (df=). It makes sense!

Warning: many times this is **not** of interest.
Chicks: intercept $=$ average body mass of chicks fed by 2 parents only. Has a clear meaning, but not of real interest.
Energy expenditure: intercept $=$ average kcal of men with fat-free body mass of 0. Definitely not of interest.

R functions: `lm` (**l**inear **m**odel), `summary`

```
> fit = lm(mass~feed, data=russell)
> summary(fit)
Call:
lm(formula = mass ~ feed, data = russell)

Residuals:
      Min        1Q    Median        3Q       Max
-0.839762 -0.229710 -0.005071  0.268414  0.850032

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.81935    0.22376  26.007 1.61e-14 ***
feed         0.04577    0.01570   2.916   0.0101 *
---
Signif. codes:  0’***’0.001’**’0.01’*’0.05’.’0.1’ ’1

Residual standard error: 0.4362 on 16 degrees of freedom
Multiple R-Squared: 0.347,   Adjusted R-squared: 0.3062
F-statistic: 8.504 on 1 and 16 DF,  p-value: 0.01010
```

For t-test: summary()

```
> dat
  Subject mass   e1   e2
1       1 49.3 1851 1936
2       2 59.3 2209 1891    > fit1 = lm(e1~mass, data=dat)
3       3 68.3 2283 2423    > fit2 = lm(e2~mass, data=dat)
...                  ...
> summary(fit1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   790.73     171.82   4.602 0.005830 **
mass           21.97       2.71   8.106 0.000463 ***

Residual standard error: 80.3 on 5 degrees of freedom
Multiple R-squared: 0.9293,     Adjusted R-squared: 0.9151
F-statistic:  65.7 on 1 and 5 DF,  p-value: 0.0004635

> summary(fit2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  421.803    268.855   1.569  0.17746
mass          28.103      4.241   6.627  0.00118 **

Residual standard error: 125.6 on 5 degrees of freedom
Multiple R-squared: 0.8978,     Adjusted R-squared: 0.8773
F-statistic: 43.91 on 1 and 5 DF,  p-value: 0.001178
```

Assessing assumptions: Diagnostic plots

Recall the assumptions:

1. The straight line relationship is correct.
2. Errors e_i are independent.
3. Errors e_i have homogeneous variance: $\text{var}(e_i) = \sigma_e^2$.
4. Errors e_i have normal distribution: $e_i \sim \mathcal{N}(0, \sigma_e^2)$.

We check these assumptions by examining the residuals

$$r_i = y_i - \hat{y}_i$$

residuals are the deviations from the regression line, or "errors".

Normal quantile plot of the residuals, and
residual plot: r_i versus $\hat{y}_i$.

Residuals: to check assumptions

```
> russell
   feed mass
1   7.0 5.30
2  16.5 6.49
3   3.7 6.20
4  10.5 7.15
5  12.9 6.00
6  18.0 6.97
7  14.2 6.45
...    ...
16 13.0 5.75
17 10.3 6.30
18 30.0 7.48
```

```
> fit = lm(mass~feed, data=russell)

> russell$predicted = fitted(fit)
  # gets fitted values from fitted line,
  # and uses them to create new column
  # named 'predicted' in data 'russell'.

> russell$residual  = residuals(fit)
  # now getting residuals and use them
  # to create a new column named
  # residuals' in data 'russell'.
```

```
> russell
   feed mass predicted residual
1   7.0 5.30      6.14 -0.84
2  16.5 6.49      6.57 -0.08
3   3.7 6.20      5.99  0.21
4  10.5 7.15      6.30  0.85
5  12.9 6.00      6.41 -0.41
6  18.0 6.97      6.64  0.33
7  14.2 6.45      6.47 -0.02
...                   ...
18 30.0 7.48      7.19  0.29
```

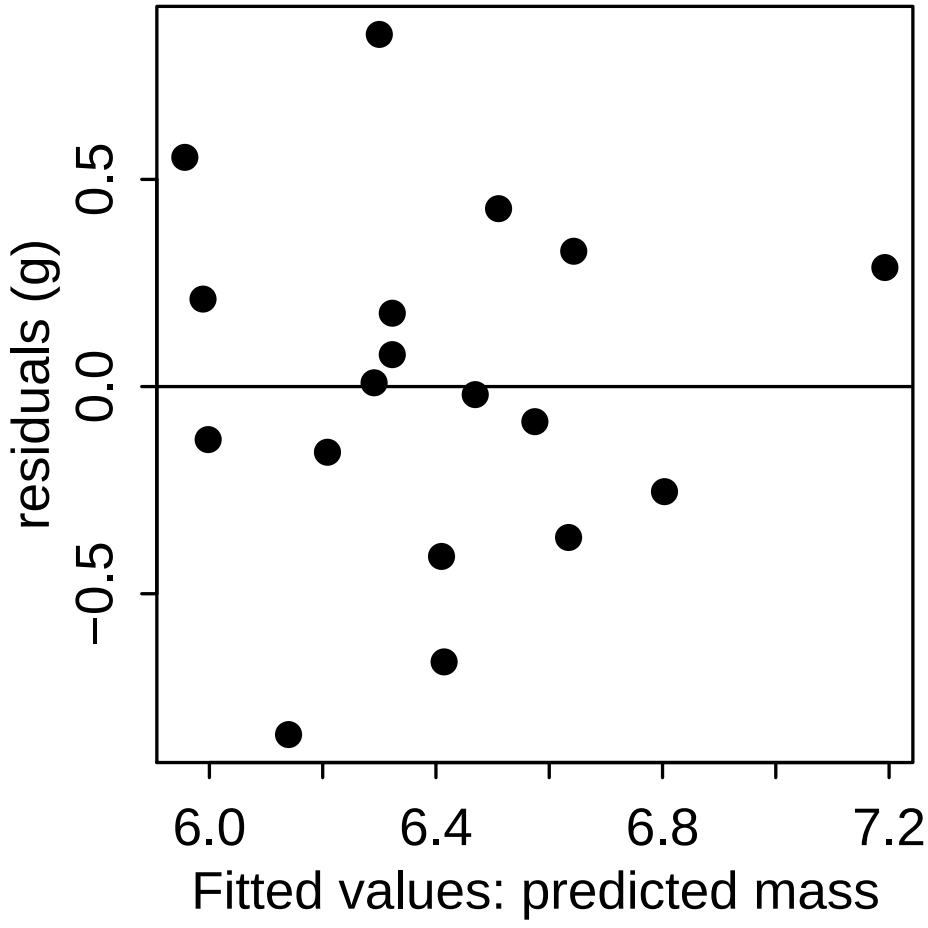
Residuals: to check assumptions

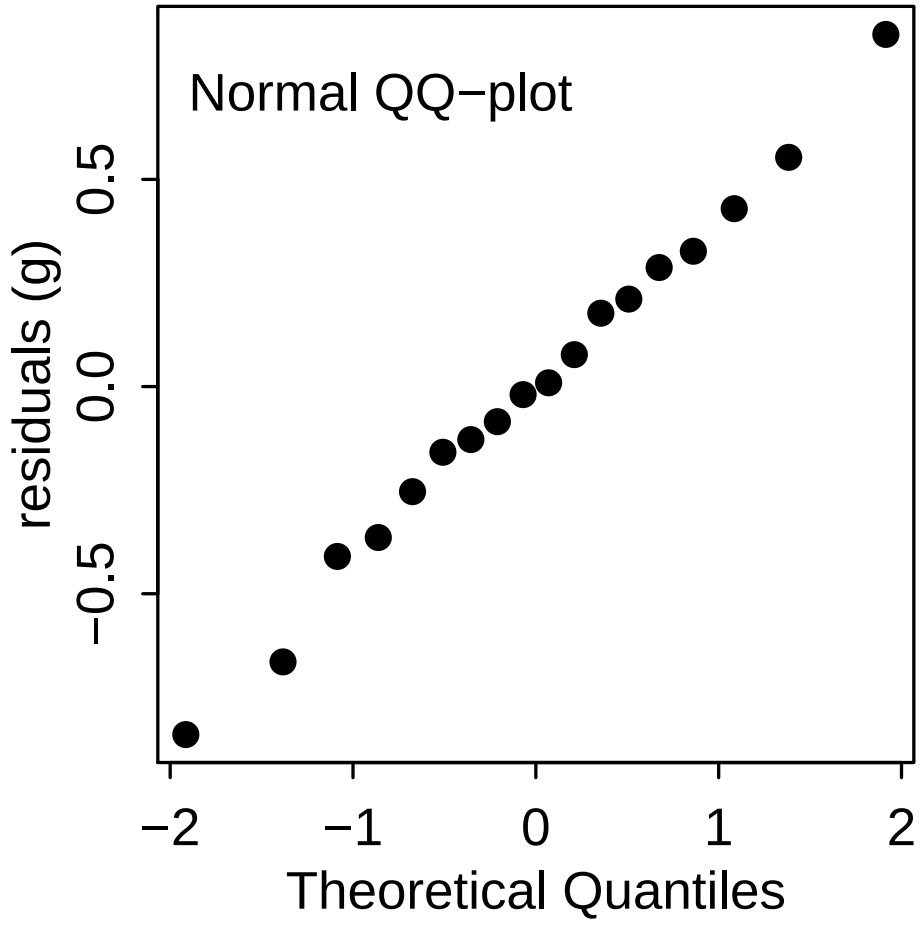
Normal quantile plot of the residuals, and
residual plot: r_i (y-axis) versus $\hat{y}_i$ (x axis).

Normal quantile plot of residuals: needs to be linear enough to indicate no (or little) departure from normality.

Residual plot: needs to show a **random scatter** with no evident pattern. Patterns may indicate problems such as a curved relationship, or nonhomogeneous variance, or outliers.

Assessing assumptions: Diagnostic plots



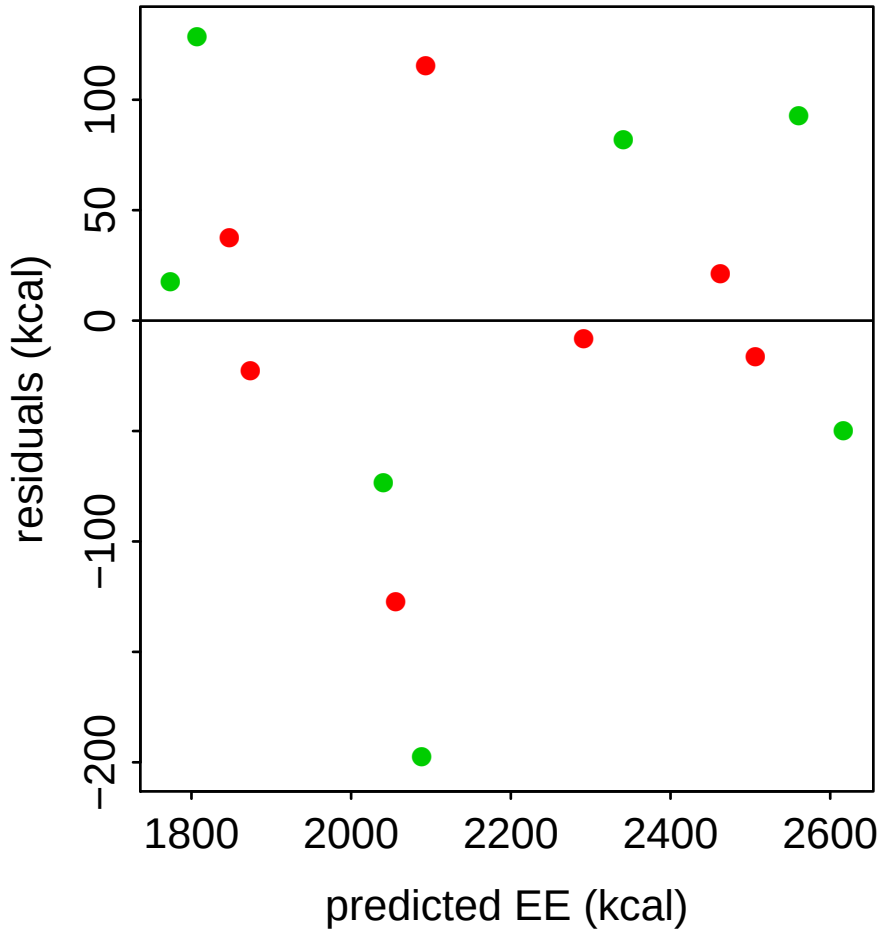


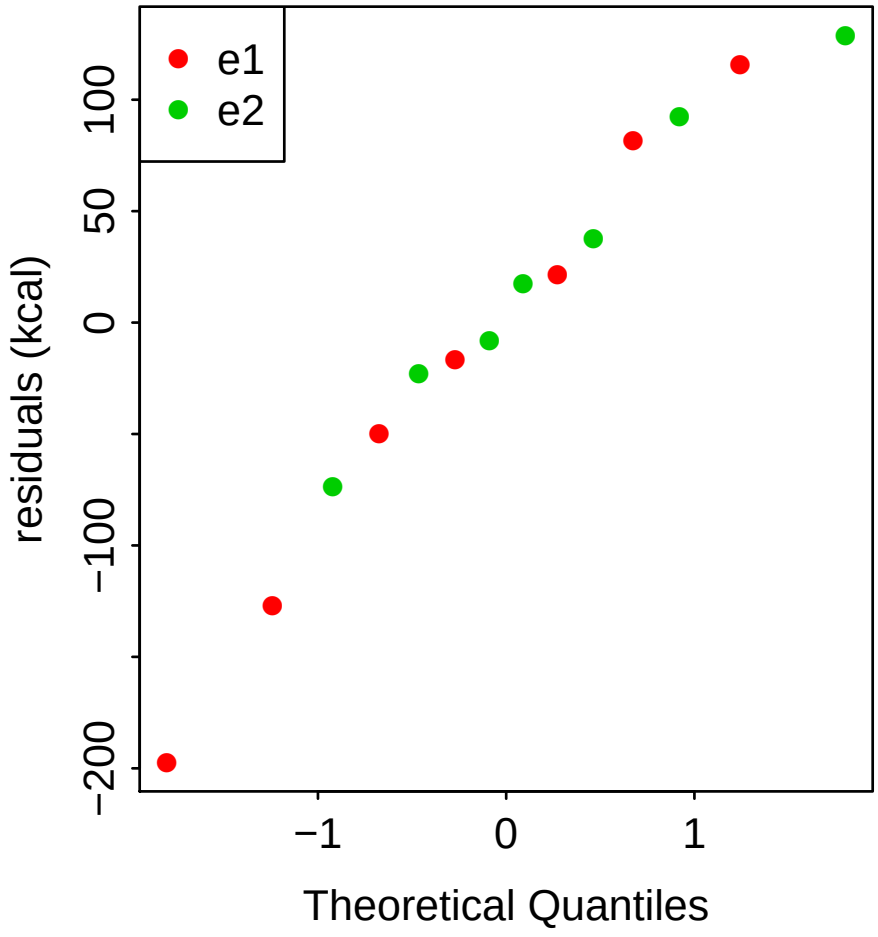
R: diagnotic plots of residuals

```
# first fit the regression line to the data,
# will later use the result to get plots.
fit = lm(mass~feed, data=russell)

# normal quantile plot of residuals to check normality:
qqnorm(residuals(fit))

# residual plot to check homogeneous variance
# and check absence of curve:
plot(fitted(fit), residuals(fit))
abline(h=0)                          # adds horizontal line

# automatically makes 4 plots including the 2 above:
layout( matrix(1:4, nrow=2, ncol=2))
plot(fit)
```

Diagnostic plots for energy expenditure

First and second measurements were fitted separately, but residuals plotted together here.





assumptions of homogeneous variance and normal distribution seem to be met.

Prediction of new values

For a given value x^* of interest, we might want to predict the new value of Y at x^*. We would use

$$\hat{Y} = b_0 + b_1 x^*$$

Suppose new chick is provided with $x^* = 20$ feeds/h by foster males. Then the predicted new chick mass is

$$\hat{y} = 5.82 + 0.0458 * 20 = 6.73 \text{ g}$$

What is the prediction error? Error because

- the regression line $y = b_0 + b_1 x$ is not exactly true,
- the line only describes the **average** trend, i.e. the **average** y given a known x. There is true variation around the line.

Prediction of a new value

$$\hat{y}_{\text{pred}} = b_0 + b_1 x^*$$

The prediction error is typically:

Standard error

$$\text{SE}_{\hat{y}_{\text{pred}}} = s_e \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2} + 1}$$

A $(1 - \alpha)$ prediction interval for the new value is

$$\hat{y}_{\text{pred}} \pm t_{\alpha/2, n-2} \, \text{SE}_{\hat{y}_{\text{pred}}}$$

Prediction of a new value

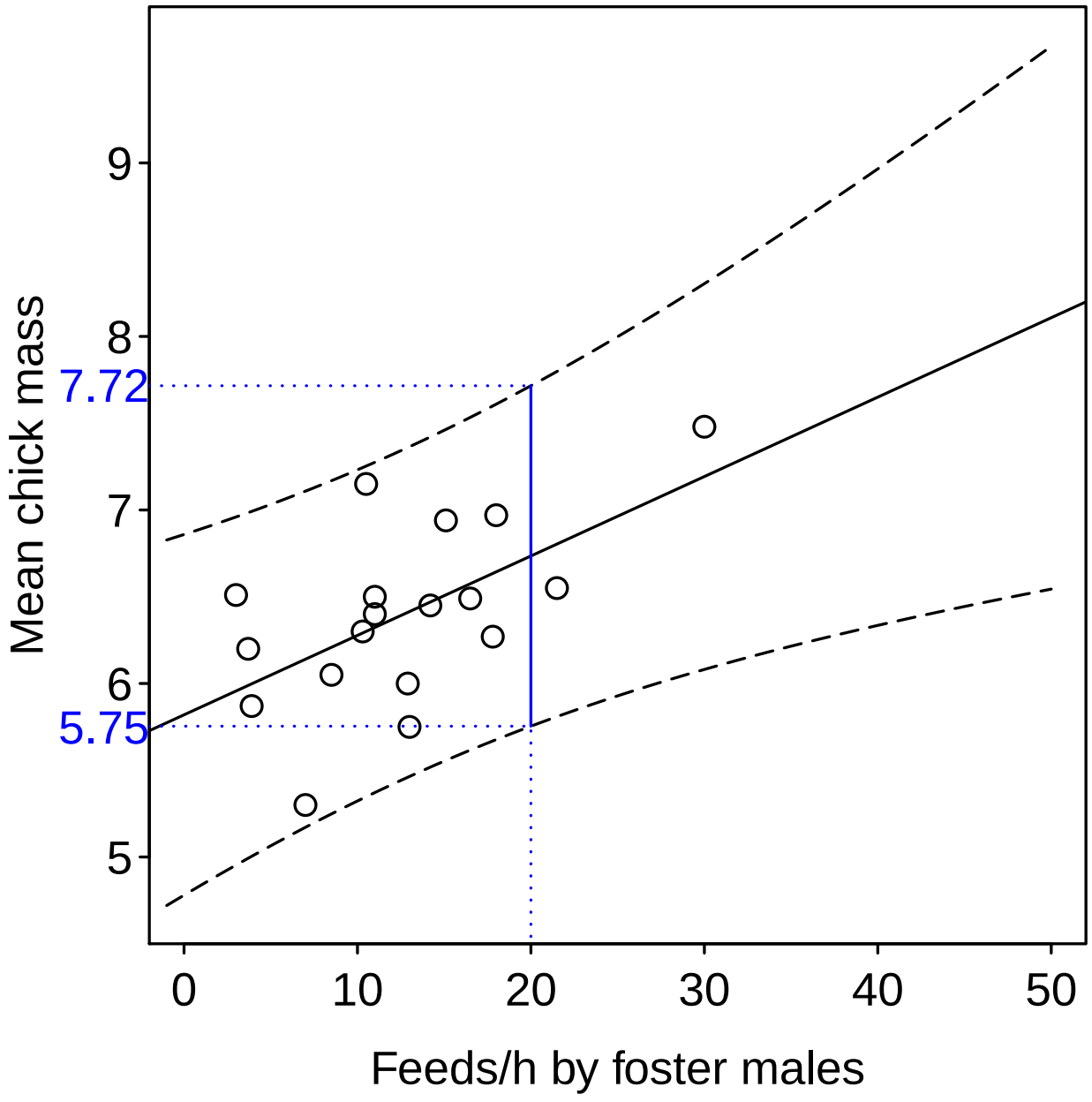
Chicks: at $x^* = 20$ feeds/h

$$\text{SE}_{\hat{y}_{\text{pred}}} = \qquad\qquad = .463$$

A 95% prediction interval for the new value is
i.e. $(5.75, 7.72)$ g

How about predicting Y at $x^* = 100$ feeds/h?
Caution against extrapolation!

Prediction intervals for new values



dotted lines: 95% prediction intevals for new values.

R: predict

```
> fit = lm(mass~feed, data=russell)
> predict(fit, newdata=data.frame(feed=20) )
       1
6.734813

> predict(fit, newdata=data.frame(feed=20),
                interval="prediction")
       fit       lwr      upr
1 6.734813 5.753784 7.715842

> predict(fit, newdata=data.frame(feed=c(20,100)),
                interval="prediction")
        fit       lwr       upr
1  6.734813 5.753784  7.715842
2 10.396665 7.339051 13.454278
```

Prediction for energy expenditure

How can we deal with the fact that we have 2 measurements on each subject?

1. Use each set of measurements separately. Can we combine them to obtain a single prediction?
2. Combine both data sets to a sample size of 14 measurements?

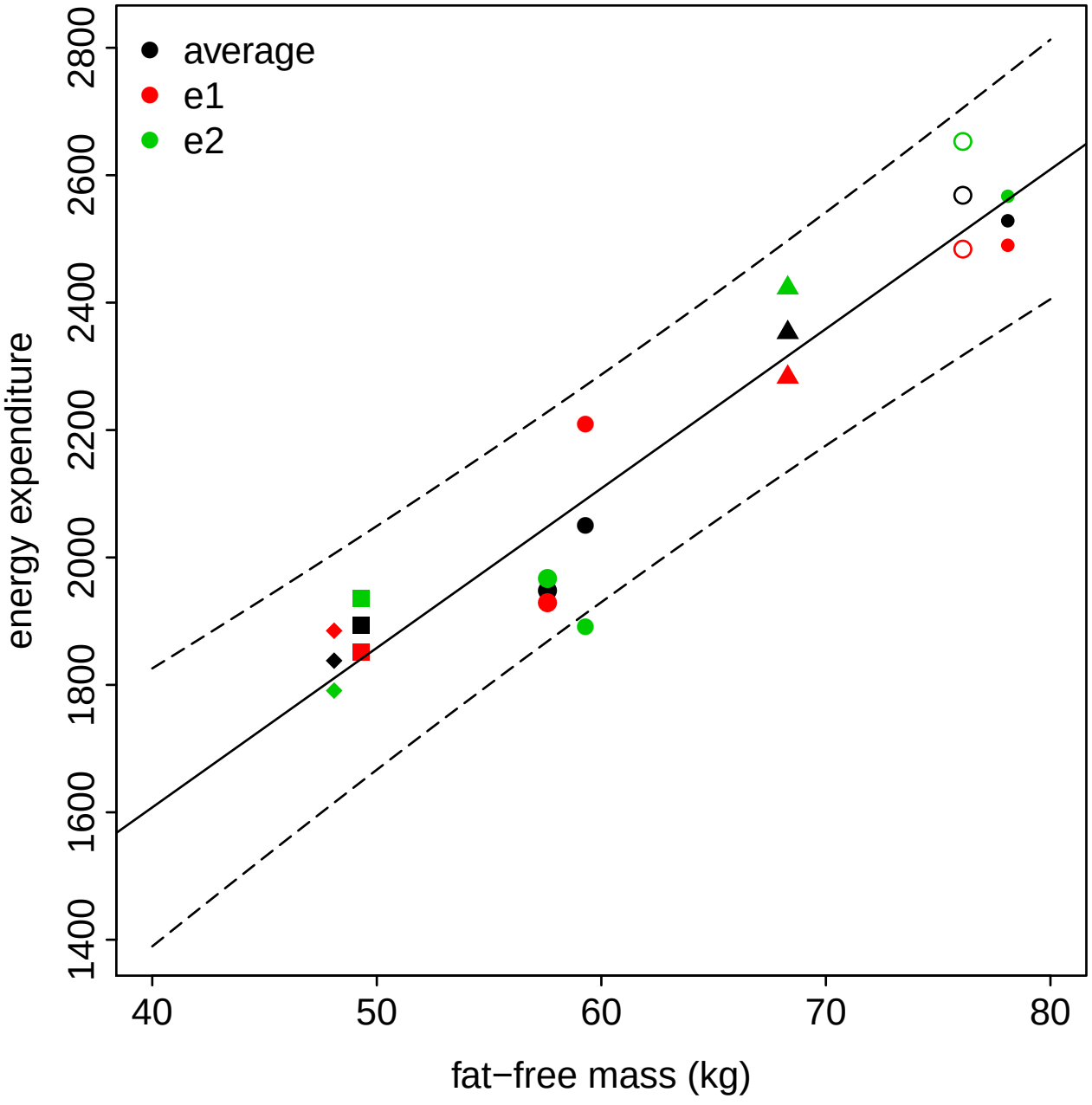
Energy expenditure: averaging the 2 values

```
> dat
  Subject mass   e1   e2
1       1 49.3 1851 1936
2       2 59.3 2209 1891
3       3 68.3 2283 2423
4       4 48.1 1885 1791
5       5 57.6 1929 1967
6       6 78.1 2490 2567
7       7 76.1 2484 2653

> dat$ave = (dat$e1+ dat$e2)/2
> dat
  Subject mass   e1   e2    ave
1       1 49.3 1851 1936 1893.5
2       2 59.3 2209 1891 2050.0
3       3 68.3 2283 2423 2353.0
4       4 48.1 1885 1791 1838.0
5       5 57.6 1929 1967 1948.0
6       6 78.1 2490 2567 2528.5
7       7 76.1 2484 2653 2568.5
```

Regression on the average values

```
> fit.ave = lm(ave~mass, data=dat)
> anova(fit.ave)
          Df Sum Sq Mean Sq F value    Pr(>F)
mass       1 550161  550161  131.01 8.916e-05 ***
Residuals  5  20997    4199

> summary(fit.ave)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  606.264    138.668   4.372  0.00721 **
mass          25.036      2.187  11.446 8.92e-05 ***

Multiple R-squared: 0.9632,     Adjusted R-squared: 0.9559
F-statistic:   131 on 1 and 5 DF,  p-value: 8.916e-05

> predict(fit.ave, newdata=data.frame(mass=c(65,100)),
          interval="prediction")
   fit  lwr  upr
1 2234 2055 2412
2 3110 2833 3386
```

Prediction for energy expenditure



R: graph with prediction curves

```
# first create a range of new body mass data:
> xnew  = seq(40,80,by=.5)
> xnew
 [1] 40.0 40.5 41.0 41.5 42.0 42.5 43.0 43.5 44.0 44.5 45.0 45.5 46.0
[16] 47.5 48.0 48.5 49.0 49.5 50.0 50.5 51.0 51.5 52.0 52.5 53.0 53.5
...
[76] 77.5 78.0 78.5 79.0 79.5 80.0

# then get the predicted energy expenditure at these mass values:
> ypred = predict(fit.ave,data.frame(mass=xnew),interval="prediction")
> ypred
    fit  lwr  upr
1  1608 1390 1826
2  1620 1404 1837
3  1633 1418 1848
4  1645 1432 1859
5  1658 1446 1870
...
78 2572 2372 2771
79 2584 2383 2785
80 2597 2394 2799
81 2609 2405 2813
```

R: graph with prediction curves

```
# prepare: to save as pdf, and to adjust margin sizes:
pdf("bodysize_energy3.pdf",height=5,width=5)
par(mar=c(3.1,3.1,.2,.2), mgp=c(1.7,.3,0), tck=-.01)

# plot the averages, then add both measurements:
plot(ave~mass, data=dat,xlab="fat-free mass (kg)",pch=15:21,
     ylab="energy expenditure", xlim=c(40,80),ylim=c(1400,2810))
points(e1~mass, data=dat, pch=15:21, col=2)
points(e2~mass, data=dat, pch=15:21, col=3)

# add regression line:
abline(fit.ave)

# add prediction curves:
lines(xnew,ypred[,"lwr"], lty=5)
lines(xnew,ypred[,"upr"], lty=5)

# add legend to identify the meaning of colors:
legend("topleft",pch=16,col=1:3,legend=c("average","e1","e2"),bty="n")

# close the plotting device to "finish" off the pdf file:
dev.off()
```