Statistical Data Analysis, Lecture 11

dr. Dennis Dobler

Vrije Universiteit Amsterdam

6 May 2020

Topics in this course

1. Summarizing data
2. Exploring distributions
3. Density estimation
4. Bootstrap methods
5. Nonparametric tests
6. Analysis of categorical data
7. Multiple linear regression

Chapter 8: Linear regression analysis

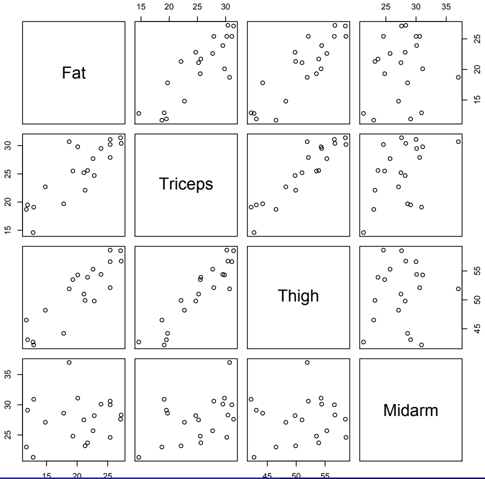
Contents of Chapter 8:

1. The multiple linear regression model
    - parameter estimation
    - selection of explanatory variables
2. Diagnostics
    - plots
    - outliers
    - leverage points
    - influence points
3. Collinearity

multiple linear regression

Idea (1)

Example Consider the following data on bodyfat, and other body measures of 20 females.



The variable Fat is very difficult to measure.

Question Can we predict this variable from one or more of the other variables, which are easy to measure?

Idea (2)

Regression: a response variable (dependent variable) is modelled as a function of explanatory variables (independent variables) and a measurement error.

Linear regression: a response variable is modelled
as a linear function of explanatory variables
plus a measurement error.

Other types of regression: nonlinear regression, generalized linear regression (see the course on Statistical Models)

Multiple linear regression model

The model:

$$\begin{aligned} Y_i &= \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + e_i \\ \mathrm{E} e_i &= 0 \\ \mathrm{E} e_i e_j &= \begin{cases} \sigma^2, & i = j, \\ 0, & i \neq j, \end{cases} \end{aligned}$$

where

- Y_i: i^{th} response observation
- x_{ij}: (known) value of the j^{th} explanatory variable for the i^{th} observation,
- $\beta_0, \beta_1, \ldots, \beta_p$, and σ^2: unknown constants (parameters)
- e_i: unknown stochastic measurement error in i^{th} observation

Model in matrix notation

In matrix notation:

$$\begin{aligned} Y &= X\beta + e \\ \mathrm{E}e &= 0 \\ \mathrm{Cov}(e) &= \sigma^2 I_{n\times n} \end{aligned}$$

with

- $Y = (Y_1, \ldots, Y_n)^T$ the stochastic vector of observations
- $X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$ design matrix, (known) values of the explanatory variables (we assume $\mathrm{rank}(X) = p + 1$)
- $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ the vector of unknown parameters
- σ^2 the unknown variance
- $e = (e_1, \ldots, e_n)^T$ the stochastic vector of measurement errors

Further assumptions

It is common to assume normally distributed errors:
$e_i \sim N(0, \sigma^2) \quad \text{i.i.d.} \qquad i = 1, \ldots, n$

Hence,

$$Y_i \sim N(\beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p, \sigma^2)$$

Note that the Y_i are not identically distributed, since the expectation of Y_i depends on the measured values of the explanatory variables for observation i.

parameter estimation

Least squares approach

In a least squares approach we find $\hat{\beta}$ that minimizes $S(\beta) = \|Y - X\beta\|^2$. This yields the parameter estimator

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

We have

$$\begin{aligned} \mathrm{E}\hat{\beta} &= (X^T X)^{-1} X^T \mathrm{E} Y = (X^T X)^{-1} X^T X \beta = \beta \\ \mathrm{Cov}(\hat{\beta}) &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

The residuals are $R_i = Y_i - \hat{Y}_i$, where $\hat{Y}_i = \hat{\beta}_0 + x_{i1}\hat{\beta}_1 + \cdots + x_{ip}\hat{\beta}_p$, and the residual sum of squares is

$$RSS = S(\hat{\beta}) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \|Y - X\hat{\beta}\|^2$$

Finally, $\hat{\sigma}^2 = \frac{RSS}{n-p-1}$ and $\widehat{\mathrm{Cov}}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1}$.

Example (1)

Apply this model to the bodyfat data.

```
> bodyfat=read.table("bodyfat.txt",header=TRUE)
> bodyfat
    Fat Triceps Thigh Midarm
1  11.9    19.5  43.1   29.1
2  22.8    24.7  49.8   28.2
....
20 21.1    25.2  51.0   27.5
> is.data.frame(bodyfat)
[1] TRUE
> is.matrix(bodyfat)
[1] FALSE
```

The variable `bodyfat` is an object of type `dataframe` in *R*, which is default when using `read.table`. You need this type in order to use the function `lm` for fitting linear models to data.

Example (2)

```
> fatlm=lm(Fat~Triceps+Thigh+Midarm,data=bodyfat)
> fatlm

Call:
lm(formula = Fat ~ Triceps + Thigh + Midarm)

Coefficients:
(Intercept)      Triceps        Thigh       Midarm
    117.085        4.334       -2.857       -2.186
```

The first argument in `lm` is a `model formula` like `response` $\sim$ `var1+...+varp`.
R includes an intercept by default. You can switch off the intercept using `response` $\sim$ `var1+...+varp-1`.

The output of the function `lm` is an object of type `linear model`. You can apply several functions to this, e.g. `summary`, `coef`, `residuals`, `fitted`, `vcov` and `confint` (see `help(lm)` in *R*).

Example (3)

```
> summary(fatlm);

Call:
lm(formula = Fat ~ Triceps + Thigh + Midarm)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7263 -1.6111  0.3923  1.4656  4.1277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
Triceps        4.334      3.016   1.437    0.170
Thigh         -2.857      2.582  -1.106    0.285
Midarm        -2.186      1.595  -1.370    0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared: 0.8014,Adjusted R-squared: 0.7641
F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06
```

Example (4)

```
> residuals(fatlm)
         1          2          3          4          5          6
-2.9549896  2.5811589 -2.2866822 -3.0273199  1.1423925 -0.5437185
....
> fitted(fatlm)
       1        2        3        4        5        6        7        8
14.85499 20.21884 20.98668 23.12732 11.75761 22.24372 25.71432 22.27064
....
> vcov(fatlm)
            (Intercept)    Triceps       Thigh      Midarm
(Intercept)   9956.5279 300.197963 -257.382315 -158.670413
Triceps        300.1980   9.093309   -7.779145   -4.788026
Thigh         -257.3823  -7.779145    6.666803    4.094616
Midarm        -158.6704  -4.788026    4.094616    2.545617
> confint(fatlm)
                 2.5 %     97.5 %
(Intercept) -94.444550 328.613940
Triceps      -2.058507  10.726691
Thigh        -8.330476   2.616780
Midarm       -5.568367   1.196247
```

variable selection

A good linear regression model

Not all available explanatory variables have explanatory power.

The goal is to find the best possible model with the smallest number of explanatory variables.
Of course, this is contradictory! Decisions have to be made.

There exists no standard strategy to find the optimal model.

The practical context also plays a role.

We consider several ways of comparing two models.

Determination coefficient

As a global check of the model fit one can compute the determination coefficient. This is a comparison between the models

$$Y = 1\beta_0 + e \qquad \text{and} \qquad Y = X\beta + e.$$

In the first model (the empty model) we have $\hat{\beta} = \overline{Y}$ and the residual sum of squares is

$$SSY = \sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

The determination coefficient is defined as

$$R^2 = \frac{SSY - RSS}{SSY} = 1 - \frac{RSS}{SSY} \qquad (0 \leq R^2 \leq 1)$$

which is the fraction of explained variance (explained by the full model).

Example

```
> summary(fatlm);

Call:
lm(formula = Fat ~ Triceps + Thigh + Midarm)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7263 -1.6111  0.3923  1.4656  4.1277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
Triceps        4.334      3.016   1.437    0.170
Thigh         -2.857      2.582  -1.106    0.285
Midarm        -2.186      1.595  -1.370    0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared:  0.8014, Adjusted R-squared:  0.7641
F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06
> RSS=sum(residuals(fatlm)^2); SSY=sum((Fat-mean(Fat))^2);
> (SSY-RSS)/SSY
[1] 0.8013586
```

Overall F-test

A high R^2-value indicates a good fit (roughly). The overall F-test provides a statistical test in order to judge what is high.

We test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ using

$$F = \frac{(n-p-1)(SSY - RSS)}{p \; RSS}$$

which has the $F_{p,n-p-1}$ distribution under H_0 if the errors are normally distributed.

H_0 is rejected for large values of F, since a large difference between SSY and RSS is an indication for H_1 being true.

Example

```
> summary(fatlm);

Call:
lm(formula = Fat ~ Triceps + Thigh + Midarm)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7263 -1.6111  0.3923  1.4656  4.1277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
Triceps        4.334      3.016   1.437    0.170
Thigh         -2.857      2.582  -1.106    0.285
Midarm        -2.186      1.595  -1.370    0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared: 0.8014,Adjusted R-squared: 0.7641
F-statistic:  21.52 on 3 and 16 DF, p-value:  7.343e-06
> (20-3-1)*(SSY-RSS)/(3*RSS)
[1] 21.51571
```

Partial F-test

A partial F-test can be used to test whether, in addition to the variables $X_1, \ldots, X_p$, one or more of the variables $X_{p+1}, \ldots, X_q$ should also be included in the model. We test

H_0: $\beta_{p+1} = \cdots = \beta_q = 0$; $\beta_0, \beta_1, \ldots, \beta_p$ arbitrary,
H_1: $\beta_j \neq 0$ for some j, $p+1 \leq j \leq q$; $\beta_0, \beta_1, \ldots, \beta_p$ arbitrary.

We can again use an F-test, comparing sums of residuals:

$$F^{p,q} = \frac{(n-q-1)(RSS_p - RSS_q)}{(q-p)RSS_q}.$$

where RSS_p is the residual sum of squares for the model under H_0 and RSS_q is the residual sum of squares for the model that includes $X_1, \ldots, X_q$ (i.e. not under H_0).

$F^{p,q}$ has under H_0 the $F_{q-p,n-q-1}$-distribution if the errors are normally distributed. H_0 is rejected for large values of $F^{p,q}$.

Example

```
> fatlm2=lm(Fat~Triceps)
> fatlm2

Call:
lm(formula = Fat ~ Triceps)

Coefficients:
(Intercept)      Triceps
    -1.4961       0.8572

> anova(fatlm,fatlm2)
Analysis of Variance Table

Model 1: Fat ~ Triceps + Thigh + Midarm
Model 2: Fat ~ Triceps
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     16  98.405
2     18 143.120 -2   -44.715 3.6352 0.04995 *
```

Here we see that the model including all 3 variables is significantly better than including only `Triceps`.

t-test

A t-test can be used to test whether the single variable X_k should be included in the model. We test

H_0: $\beta_k = 0$; β_j arbitrary for $j \neq k$,
H_1: $\beta_k \neq 0$; β_j arbitrary for $j \neq k$.

We use the following test statistic:

$$T_k = \frac{\hat{\beta}_k}{\sqrt{\widehat{\mathrm{Cov}}(\hat{\beta})_{kk}}}$$

which follows a t_{n-p-1}-distribution under H_0.

H_0 is rejected for $|T_k| > t_{n-p-1;1-\alpha/2}$.

This test is equivalent to the partial F-test, applied to only X_k.

Example

Again the output is in summary:

```
> summary(fatlm)

Call:
lm(formula = Fat ~ Triceps + Thigh + Midarm)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7263 -1.6111  0.3923  1.4656  4.1277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
Triceps        4.334      3.016   1.437    0.170
Thigh         -2.857      2.582  -1.106    0.285
Midarm        -2.186      1.595  -1.370    0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared: 0.8014,Adjusted R-squared: 0.7641
F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06
```

Partial correlation

In order to judge whether the variable X_k is useful in addition to the other variables we can look at

- the part of Y that cannot be explained by the other variables
- the part of X_k that cannot be explained by the other variables

We can perform this check by computing the linear correlation between

- $R_Y(X_{-k})$: the residuals of Y regressed on the other variables
- $R_{X_k}(X_{-k})$: the residuals of X_k regressed on the other variables

This is called the partial correlation $\rho(X_k, Y)$ between X_k and Y.

Its aim is similar to that of the t-test for β_k.

Example

This has to be done manually in *R*:

```
> attach(bodyfat)
> RYXK=residuals(lm(Fat~Triceps+Midarm))
> RXKXK=residuals(lm(Thigh~Triceps+Midarm))
> cor(RYXK,RXKXK)
[1] -0.2665991
```

When this partial correlation is far from 0, it indicates that the variable should be included.

(Look up the function `attach` and `detach` in *R*.)

Two strategies for finding a good model

In practice we need a strategy for building a model.

The step up method:

1. start with the empty model $Y = 1\beta_0 + e$
2. add the variable that yields the maximum increase in R^2
3. if the added variable is significant (t-test), go back to step 2.

The step down method:

1. start with the full model $Y = X\beta + e$
2. test all variables in a t-test
3. if the largest p-value is larger than 0.05, remove the corresponding variable and go back to step 2

Example — step up (1)

We apply the step up strategy to the bodyfat data:

```
> summary(lm(Fat~Triceps))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4961     3.3192  -0.451    0.658
Triceps       0.8572     0.1288   6.656 3.02e-06 ***
Multiple R-squared: 0.7111,Adjusted R-squared: 0.695
> summary(lm(Fat~Thigh))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.6345     5.6574  -4.178 0.000566 ***
Thigh         0.8565     0.1100   7.786  3.6e-07 ***
Multiple R-squared: 0.771,Adjusted R-squared: 0.7583
> summary(lm(Fat~Midarm))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.6868     9.0959   1.615    0.124
Midarm        0.1994     0.3266   0.611    0.549
Multiple R-squared: 0.02029,Adjusted R-squared: -0.03414
```

The first variable to add is `Thigh`.

Example — step up (2)

The second step:

```
> summary(lm(Fat~Thigh+Triceps))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.1742     8.3606  -2.293   0.0348 *
Thigh         0.6594     0.2912   2.265   0.0369 *
Triceps       0.2224     0.3034   0.733   0.4737
Multiple R-squared: 0.7781,Adjusted R-squared: 0.7519
> summary(lm(Fat~Thigh+Midarm))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.99695    6.99732  -3.715  0.00172 **
Thigh         0.85088    0.11245   7.567 7.72e-07 ***
Midarm        0.09603    0.16139   0.595  0.55968
Multiple R-squared: 0.7757,Adjusted R-squared: 0.7493
```

Both `Tricpes` and `Midarm` are not significant when added.

Resulting model: `Fat = -23.6345 + 0.8565*Thigh + error` with $R^2 = 0.771$ and $\hat{\sigma} = 2.51$.

Example — step down (1)

We now apply the step down strategy to the bodyfat data:

```
> summary(lm(Fat~Triceps+Thigh+Midarm))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
Triceps        4.334      3.016   1.437    0.170
Thigh         -2.857      2.582  -1.106    0.285
Midarm        -2.186      1.595  -1.370    0.190

Multiple R-squared: 0.8014,Adjusted R-squared: 0.7641
```

We see that none of the variables is significant. The first variable to remove is `Thigh`, which has the highest p-value.

Example — step down (2)

The second step:

```
> summary(lm(Fat~Triceps+Midarm))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.7916     4.4883   1.513   0.1486
Triceps       1.0006     0.1282   7.803 5.12e-07 ***
Midarm       -0.4314     0.1766  -2.443   0.0258 *

Residual standard error: 2.496 on 17 degrees of freedom
Multiple R-squared: 0.7862,Adjusted R-squared: 0.761
```

All remaining variables are significant.

Resulting model:

`Fat = 6.7916 + 1.0006*Triceps -0.4314*Midarm + error`

with $R^2 = 0.7862$ and $\hat{\sigma} = 2.496$.

Example — final model

Now we are left with two different models.

Model 1: $(R^2 = 0.771, \hat{\sigma} = 2.51)$
`Fat = -23.6345 + 0.8565*Thigh + error`

Model 2: $(R^2 = 0.7862, \hat{\sigma} = 2.496)$
`Fat = 6.7916 + 1.0006*Triceps -0.4314*Midarm + error`

Question Which one do we prefer, and why?

Model 1 is preferred, because it has less variables, a comparable estimate of error variance, and an only slightly lower value of R^2.

diagnostics

The need for diagnostics

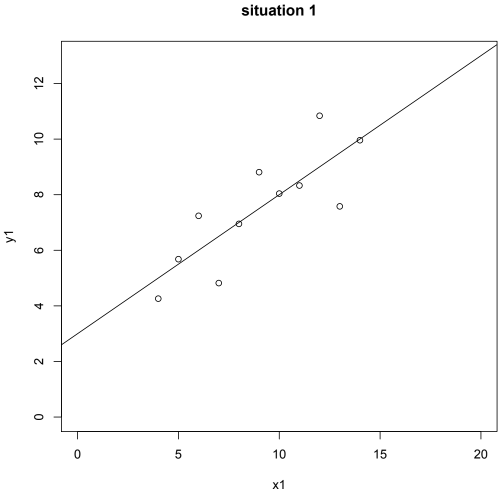
The model checks so far do not check the model assumptions, i.e. the linearity of the relation and the normality of the errors.

For that, we need diagnostic tools, both graphical and numerical checks.

In the following 4 examples of artificial data the fitted model is `y = 3.0 + 0.5*x + error` and $\hat{\sigma}^2 = 1.5$ and $R^2 = 0.67$.

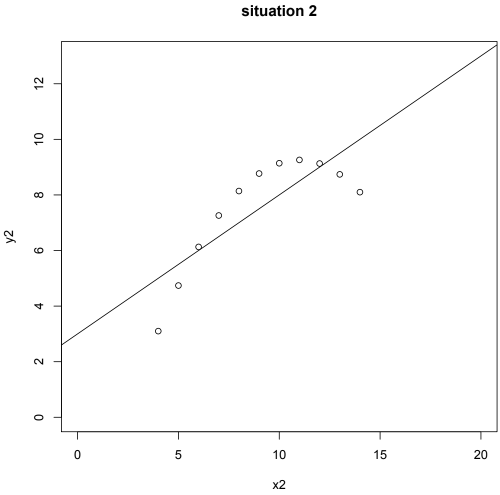
The differences between the 4 situations illustrates the need for diagnostic tools, apart from looking at R^2 and $\hat{\sigma}$.

Situation 1



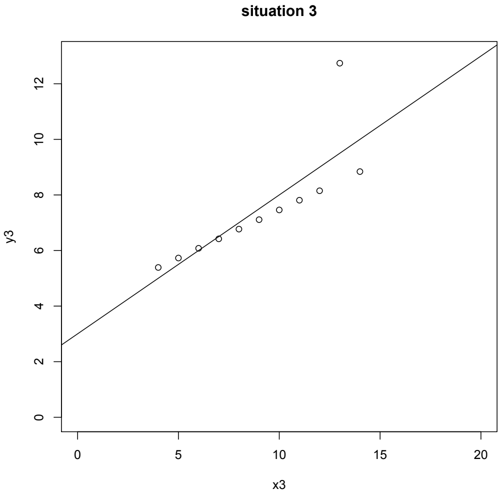
Looks ok.

Situation 2



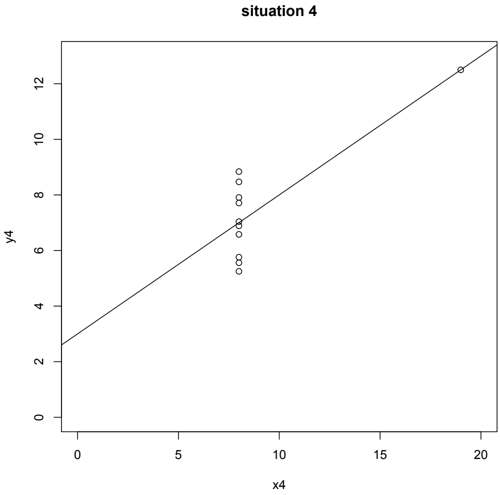
What is the problem? No linear relation between X and Y.

Situation 3



What is the problem? Outlying point in Y.

Situation 4



What is the problem? Outlying point in X.

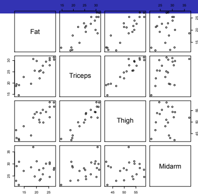
Diagnostic plots

To check the model quality look at

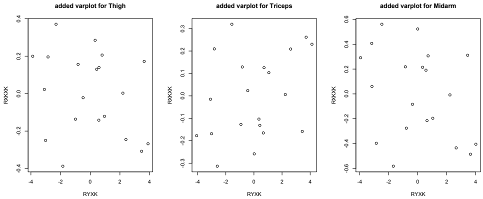
1. scatter plot: plot Y against each X_k separately (this yields overall picture, and shows outlying values.)
2. added variable plot: plot $R_Y(X_{-k})$ against $R_{X_k}(X_{-k})$ for each k (this shows how much X_k contributes in addition to the other variables.)
3. scatter plot: plot residuals against each X_k in the model separately (look at pattern (curved?) and spread.)
4. scatter plot: plot residuals against each X_k not in the model separately (look at pattern — linear? then include!.)
5. scatter plot: plot residuals against Y (look at spread.)
6. normal QQ-plot of the residuals (check normality assumption.)

Example (1)

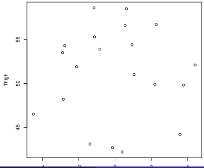
1. scatter plot of Y against each X_k separately (this yields overall picture, and shows outlying values.)



2. added variable plot of $R_Y(X_{-k})$ against $R_{X_k}(X_{-k})$ for each X_k (this shows how much X_k contributes in addition to the other variables.)

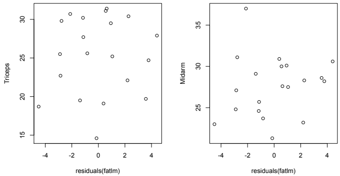


3. scatter plot of residuals against each X_k in the model separately (look at pattern (curved?) and spread.)

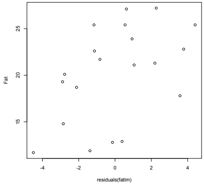


Example (2)

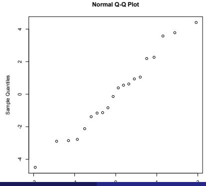
4. scatter plot of residuals against each X_k not in the model separately (look at pattern — linear? then include!.)



5. scatter plot of residuals against Y (look at spread.)



6. normal QQ-plot of the residuals (check normality assumption.)



Conlcusion of example

None of the plots shows outlying values, specific patterns or anything else that indicates that our assumptions are wrong.

Therefore, we stay with the model

```
Fat = -23.6345 + 0.8565*Thigh + error
```

with $\hat{\sigma}^2 = 6.30$ and $R^2 = 0.771$.

to finish

To wrap up

Today we discussed

1. The multiple linear regression model
   - parameter estimation
   - selection of explanatory variables
2. Diagnostics
   - plots
   - outliers
   - leverage points
   - influence points
3. Collinearity

Next week last lecture on linear regression