SSO, Lecture 5

Eduard Belitser

VU Amsterdam

Overview

1. multiple linear regression model and parameters
2. a good model
3. strategies
4. prediction
5. validation

model and parameters regression

Example - Bodyfat data (1)

Data of 20 females between 25 and 30 years old on amount of body fat, triceps skinfold thickness, thigh circumference and midarm circumference.

```
> bodyfat
    Fat Triceps Thigh Midarm
1  11.9    19.5  43.1   29.1
2  22.8    24.7  49.8   28.2
3  18.7    30.7  51.9   37.0
4  20.1    29.8  54.3   31.1
...
19 14.8    22.7  48.2   27.1
20 21.1    25.2  51.0   27.5
```

Body fat is hard to measure, while the other 3 variables are easy to measure.

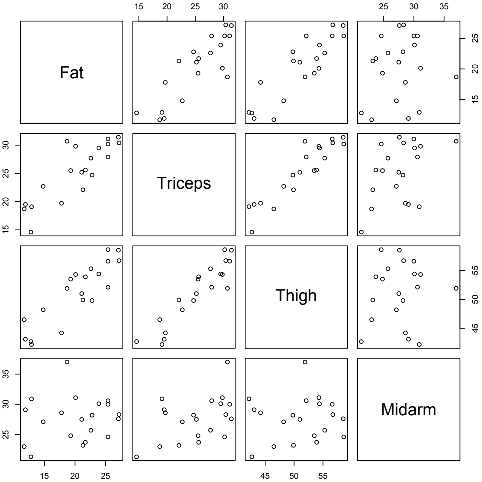
Question Can we predict `Fat` from the other 3 variables?

Example - Bodyfat data (2)

Scatter plots of all pairs of two variables:

```
> pairs(bodyfat)
```

Question Can we predict `Fat` from the other 3 variables?



The multiple linear regression model

The multiple linear regression model (*meervoudig lineair regressiemodel*) is :

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + e$$

with

- Y the dependent variable (response variable)
- $x_1, \ldots, x_k$ the independent variables (explanatory variables, predictor variables)
- $\beta_0, \ldots, \beta_k$ unknown population parameters
- e the stochastic error (fluctuation)

Assumption: the error e has a normal$(0, \sigma^2)$ distribution with unknown variance σ^2.

Note simple linear regression is a special case of multiple linear regression ($k = 1$).

Examples of explanatory variables

Possible explanatory variables (prediction variables):

- all x_i different
$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + e$$
- powers of x_i's
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + e$$
- interactions between x_i's
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e$$

Essential All models are linear in the β_i's, not necessarily in the x_i's.

Estimating parameters

To find the best parameters we minimize the sum of squared differences between the observations and the model:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_k x_{i,k})^2.$$

This yields again the least squares estimates (*kleinste kwadraten schatters*) for the β's. (In Triola: $b_i = \hat{\beta}_i$).

In R: `lm(y~x1+...+xk, data = ...)`

Example - Bodyfat data (3)

Estimating the regression parameters:

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

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

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
...
```

From the output we can find $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$.

The Sum of Squared Errors (SSE)

The Sum of Squared Errors (SSE) is

$$SSE = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \ldots - \hat{\beta}_k x_{i,k})^2.$$

The estimated variance of the errors e_i is

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-k-1}.$$

Example - Bodyfat data (4)

The estimated variance of the bodyfat data:

```
> summary(bodyfatlm)

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

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
```

From this output: $\hat{\sigma} = 2.48$, so $\hat{\sigma}^2 = 6.15$.

Estimated errors

The i^{th} residual (*residu*) is the estimated error of the i^{th} observation is

$$y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \ldots - \hat{\beta}_k x_{i,k}.$$

```
> residuals(bodyfatlm)
         1          2          3          4
-2.9549896  2.5811589 -2.2866822 -3.0273199
         5          6          7          8
 1.1423925 -0.5437185  1.3856834  3.1293594
         9         10         11         12
 1.7051817 -1.2483822  0.8044445  2.2076913
        13         14         15         16
-3.3094005  4.1276946  0.9880521  0.1725323
        17         18         19         20
-0.3736041 -1.3859022 -3.7262800  0.6120883
```

a good model

When is a model good?

Not all available explanatory variables have explanatory power.

The goal is to find the best possible model with the smallest number of explanatory variables.

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.

Coefficient of determination

The multiple coefficient of determination (*meervoudige determinatiecoëfficiënt*) R^2 compares the models

$$Y = \beta_0 + e \qquad \text{and} \qquad Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + e.$$

Left we get $\hat{\beta}_0 = \overline{y}$ with sum of squares

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

The coefficient of determination R^2 is

$$R^2 = \frac{SS_{yy} - SSE}{SS_{yy}} \qquad (0 \leq R^2 \leq 1).$$

This is the proportion of explained variance.

R^2 yields a global check on the multiple linear regression model.
The higher R^2 the more variation the model explains.

Note If $k = 1$, we have $R^2 = r^2$.

Example - Bodyfat data (5)

```
> summary(bodyfatlm)
...
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
> SSE=sum(residuals(bodyfatlm)^2)
> SSYY=sum((bodyfat$Fat-mean(bodyfat$Fat))^2)
> (SSYY-SSE)/SSYY
[1] 0.8013586
```

For this data set the multiple linear regression model explains 80% of the variation. That is quite a lot.

Question When is R^2 high (enough)?

Testing the full multiple linear regression model (1)

In simple linear regression we compare

$$Y = \beta_0 + e \qquad \text{and} \qquad Y = \beta_0 + \beta_1 x + e.$$

If $H_0 : \beta_1 = 0$ is rejected (see t-test of last week) a simple linear regression model is useful, since x has *significant* explanatory power in a linear model.

In multiple linear regression we compare

$$Y = \beta_0 + e \qquad \text{and} \qquad Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + e.$$

Now we test $H_0 : \beta_1 = \ldots = \beta_k = 0$. If this H_0 is rejected, multiple linear regression is useful, since $x_1, \ldots, x_k$ together have *significant* explanatory power in a linear model.

Test for $H_0 : \beta_1 = \ldots = \beta_k = 0$

Setting A multivariate data set with response variable Y and explanatory variables $X_1, \ldots, X_k$. We test the β_i's in the multiple linear regression model: $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + e$.

Hypotheses $H_0 : \beta_1 = \ldots = \beta_k = 0$ versus H_1 : at least one $\beta_i \neq 0$.

Test statistic

$$T = \frac{R^2/k}{(1 - R^2)/(n - (k+1))}$$

The larger R^2, the larger T.

Distribution of T under H_0 F-distribution with k and $n - (k+1)$ degrees of freedom (exact)

Assumption the errors follow a normal distribution

p-value The p-value is always right-sided: $p_{right} = P(T > t)$. We only reject H_0 if R^2 is large, i.e. if T is large.

In R The p-value is in the last line of `summary(lm(y~x))`.

Example - Bodyfat data (6)

The output of the overall F-test of the bodyfat data:

```
> summary(bodyfatlm)
...
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
```

The p-value in the overall test is 0.0000073. Hence for this data the F-test rejects $H_0 : \beta_1 = \ldots = \beta_k = 0$. At least one of the β_i's is not equal to 0.

Testing individual explanatory variables

Not all available explanatory variables have explanatory power.

From all explanatory variables, we need to find relevant explanatory variables.

Therefore we test $H_0 : \beta_i = 0$ for all β_i in the model.

Test for $H_0 : \beta_i = 0$

Setting A multivariate data set with response variable Y and explanatory variables $X_1, \ldots, X_k$. We test $H_0 : \beta_i = 0$ in the multiple linear regression model: $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + e$.

Hypotheses $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$.

Test statistic

$$T = T = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}} \sim t_{n-(k+1)}$$

Distribution of T under H_0 t-distribution with $n - (k + 1)$ degrees of freedom (exact)

Assumption the errors follow a normal distribution

p-value Usually the two-sided p-value is considered

In R The p-value is in the column `Pr(>|t|)` in the output of `summary(lm(y~x))`.

Example - Bodyfat data (7)

The p-values of the individual explanatory variables in the bodyfat data:

```
> summary(bodyfatlm)

Call:
lm(formula = Fat ~ Triceps + Thigh + Midarm, data = bodyfat)
...
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
```

From this output: none of the β_i's is significant. So none of the explanatory variables separately explains a signficant part, but all together they explain 80%!

Confidence intervals for β_i's

The $(1-\alpha)$ confidence interval for the β_i's are

$$\beta_i = \hat{\beta}_i \pm t_{\alpha/2} s_{\hat{\beta}_i}.$$

In R use `confint(lm(y~x))`

```
> confint(bodyfatlm)
                 2.5 %     97.5 %
(Intercept) -94.444550 328.613940
Triceps      -2.058507  10.726691
Thigh        -8.330476   2.616780
Midarm       -5.568367   1.196247
> confint(bodyfatlm,level=0.9)
                    5 %        95 %
(Intercept) -57.1237737 291.2931633
Triceps      -0.9306401   9.5988241
Thigh        -7.3647462   1.6510504
Midarm       -4.9716159   0.5994954
```

strategies

How to find the relevant predictors?

The goal is to find the best possible model (high R^2) with the smallest number of explanatory variables.

Since more explanatory variables always explain more, we can consider the R^2 adjusted for the number k of explanatory variables:

$$R^2_{adjusted} = 1 - \frac{n-1}{n-(k+1)}(1-R^2).$$

The goal is to maximize R^2 with as few as possible explanatory variables, and $R^2_{adjusted}$ helps to choose between models with different amounts of variables. Note that the interpretation of $R^2_{adjusted}$ is not fraction of explained variance anymore.

We consider two strategies to find the optimal model.

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 $Y = \beta_0 + e$
2. add the variable that yields the maximum increase in $R^2_{adjusted}$
3. if the added variable is significant (t-test), go back to step 2.

The step down method:

1. start with the full model $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + e$
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

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`.

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 $R^2_{adjusted} = 0.76$ and $\hat{\sigma} = 2.51$.

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.

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 $R^2_{adjusted} = 0.76$ and $\hat{\sigma} = 2.496$.

Up or down?

Now we are left with two different models.

Model 1: $(R^2_{adjusted} = 0.76, \hat{\sigma} = 2.51)$

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

Model 2: $(R^2_{adjusted} = 0.76, \hat{\sigma} = 2.496)$

```
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 a comparable value of $R^2_{adjusted}$.

prediction

The predicted value

Once all $\hat{\beta}_i$'s are known, one can predict the y-value for a (new) measurement of the k explanatory variables (x_j's):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots \hat{\beta}_k x_k$$

For the x-values in the data set, these $\hat{y}$-values are found by

```
> fitted(bodyfatlm)
       1        2        3        4        5
14.85499 20.21884 20.98668 23.12732 11.75761
...
      16       17       18       19       20
23.72747 22.97360 26.78590 18.52628 20.48791
```

Confidence and prediction intervals

Two types of intervals for y for given x-values:

- confidence interval for y: an interval for the mean Y-value for given x-values
- prediction interval for y: an interval for an individual Y-observation for given x-values (this interval is larger!)

Confidence is for the population mean, whereas prediction is for an individual observation.

In R `predict(lm(y~x1+...+xk),newxdata,interval=...,level=...)`

Example - Bodyfat data (7)

Prediction intervals for the body fat data for new data can be found by

- designing a `data.frame` with the new x-values
- applying `predict` to this `data.frame`.

```
> newxdata=data.frame(Triceps=24.5,Thigh=51.3,Midarm=28.7)
> predict(bodyfatlm,newxdata,interval="prediction")
       fit      lwr      upr
1 13.97372 3.053481 24.89396
> predict(bodyfatlm,newxdata,interval="prediction",level=0.95)
       fit      lwr      upr
1 13.97372 3.053481 24.89396
> predict(bodyfatlm,newxdata,interval="confidence",level=0.95)
       fit      lwr      upr
1 13.97372 4.402296 23.54515
```

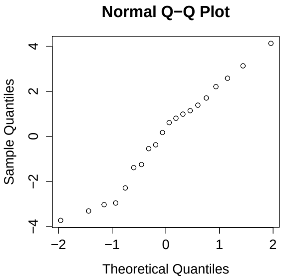
The prediction interval is indeed larger!

validating the model

Validation of normality

As in the case of simple linear regression, one needs to check the normality assumption in a *QQ*-plot of the residuals.

```
> qqnorm(residuals(bodyfatlm))
```



More scatter plots for validation

Next week we will investigate more scatter plots as validation of the model.

to finish

To wrap up

Today we discussed

- multiple linear regression model and parameters
- a good model
- strategies
- prediction
- validation

Next time several problems in multiple linear regression and ANOVA