

Multiple linear regression 1

Morten Frydenberg ©
Institut for Biostatistik

Why do we need multiple linear regression.

An example

Interpretation of the parameters

The general model

The assumptions.

The parameters.

Estimation.

The distribution of the estimates

Confidence intervals

The F-test

Checking the model

Fitted values, residuals and leverage

Morten Frydenberg

Linear and Logistic regression - Note 2.1

1

Why do we need a multiple regression

The simple linear regression model only models how the dependent variable, y , depends on **one** independent variable (covariate), x_1 .

We are often interested in **how** several independent variables, x_1, x_2, \dots, x_k , influence the dependent variable, y .

Sometimes we want to **adjust** the influence of some of the information, such as age and sex, before we look at the 'effect' of other variables.

Morten Frydenberg

Linear and Logistic regression - Note 2.1

2

A multiple linear regression model

We will here start by considering a **random** subsample consisting of 200 persons from the Framingham data set used in the book.

A **multiple** linear regression model:

$$\ln(sbp) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{woman} + \beta_3 \cdot \ln(bmi) + E$$

Where the **errors**, E , are assumed to be **independent** and **normal** with mean zero and standard deviation σ .

Note, that variable **woman** is a **dummy**/indicator variable, that it is **one** if the person a woman and **zero** if it is a man.

Morten Frydenberg

Linear and Logistic regression - Note 2.1

3

Interpretation of the coefficients 0 - the constant

$$\ln(sbp) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{woman} + \beta_3 \cdot \ln(bmi) + E$$

The first coefficient (the constant term) is the **expected** $\ln(sbp)$ for

a **man** (that is ok!)

$\text{age}=0$??????

$\text{bmi}=1 \text{ kg/m}^2$?????? ($\ln(1)=0$).

As in the simple linear regression this not of any interest.

But again we can control the interpretation, by choosing a relevant reference value for **age** and **bmi**. E.g.

$$\ln(sbp) = \alpha_0 + \beta_1 \cdot (\text{age} - 45) + \beta_2 \cdot \text{woman} + \beta_3 \cdot \ln\left(\frac{\text{bmi}}{25}\right) + E$$

↑
age45

logBMI25

Morten Frydenberg

Linear and Logistic regression - Note 2.1

4

Interpretation of the coefficients 1

$$\ln(sbp) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{woman} + \beta_3 \cdot \ln(bmi) + E$$

The **expected** $\ln(sbp)$ for a **man** with $\text{bmi}=27 \text{ kg/m}^2$ is:

$$\beta_0 + \beta_1 \cdot \text{age} + \beta_3 \cdot \ln(27)$$

The **expected** $\ln(sbp)$ for another **man** with the same **bmi**, but **1.7 year older**:

$$\beta_0 + \beta_1 \cdot (\text{age} + 1.7) + \beta_3 \cdot \ln(27)$$

The difference is: $1.7\beta_1$

We see that this difference

• **does not** depend on the **age** of the first man.

• **does not** depend on the **bmi** as long as it is the same for the two men.

• **would be the same if the two persons were women.**

Morten Frydenberg

Linear and Logistic regression - Note 2.1

5

Interpretation of the coefficients 2

$$\ln(sbp) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{woman} + \beta_3 \cdot \ln(bmi) + E$$

The **expected** $\ln(sbp)$ for a **50 year old man** with $\text{bmi}=27 \text{ kg/m}^2$ is:

$$\beta_0 + \beta_1 \cdot 50 + \beta_3 \cdot \ln(27)$$

The **expected** $\ln(sbp)$ for **woman** with the same **age** and **bmi**

$$\beta_0 + \beta_1 \cdot 50 + \beta_2 + \beta_3 \cdot \ln(27)$$

The difference is: β_2

We see that this difference

• **does not** depend on the **age** as long as it is the same for the two persons.

• **does not** depend on the **bmi** as long as it is the same for the two persons.

Morten Frydenberg

Linear and Logistic regression - Note 2.1

6

Interpretation of the coefficients 3

$$\ln(\text{sbp}) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{woman} + \beta_3 \cdot \ln(\text{bmi}) + E$$

The **expected** $\ln(\text{sbp})$ for a **woman** who is 50 year old:

$$\beta_0 + \beta_1 \cdot 50 + \beta_2 + \beta_3 \cdot \ln(\text{bmi})$$

The **expected** $\ln(\text{sbp})$ for another **woman** with the same age, but with a **bmi** which is 10% higher:

$$\beta_0 + \beta_1 \cdot 50 + \beta_2 + \beta_3 \cdot \ln(1.1 \cdot \text{bmi})$$

The difference $\beta_3 \cdot [\ln(1.1 \cdot \text{bmi}) - \ln(\text{bmi})] = \beta_3 \cdot \ln(1.1)$

We see that this difference

• **does not** depend on the **bmi** of the first woman.

• **does not** depend on the **age** as long as it is the same for the two women.

• **would be the same if the two persons were men.**

Morten Frydenberg

Linear and Logistic regression - Note 2.1

7

Interpretation of the coefficients 4

$$\ln(\text{sbp}) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{woman} + \beta_3 \cdot \ln(\text{bmi}) + E$$

$$\beta_3 \cdot [\ln(1.1 \cdot \text{bmi}) - \ln(\text{bmi})] = \beta_3 \cdot \ln(1.1)$$

As the **bmi** is introduced on the **log-scale**, then "differences" of this variable is measured relatively.

So comparing a pair of persons how **only differ** in **bmi**.

One having **bmi**=25 kg/m² and the other **bmi**=27 kg/m².

Then the **expected difference** in $\ln(\text{sbp})$ is

$$\beta_3 \cdot \ln\left(\frac{27}{25}\right) = \beta_3 \cdot 0.077$$

If the **bmi**'s were 21 kg/m² and

23 kg/m², then the **expected** difference in $\ln(\text{sbp})$ would be

$$\beta_3 \cdot \ln\left(\frac{23}{21}\right) = \beta_3 \cdot 0.091$$

Morten Frydenberg

Linear and Logistic regression - Note 2.1

8

Interpretation of the coefficients 5

$$\ln(\text{sbp}) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{woman} + \beta_3 \cdot \ln(\text{bmi}) + E$$

Taking the **exponential** we get:

$$\text{sbp} = \gamma_0 \cdot \gamma_1^{\text{age}} \cdot \gamma_2^{\text{woman}} \cdot \text{bmi}^{\beta_3} \cdot \exp(E)$$

where $\gamma_0 = \exp(\beta_0)$, $\gamma_1 = \exp(\beta_1)$ and $\gamma_2 = \exp(\beta_2)$

That is a non-linear model on the **sbp** scale!

The error is **multiplicative**.

As **medians** are preserved by the exponential transformation then the estimates telling of **effect on the median sbp**.

An example: The age and bmi adjusted median is a factor γ higher for man than for women.

Morten Frydenberg

Linear and Logistic regression - Note 2.1

9

The multiple linear regression in general

Y the **dependent** variable

(x_1, x_2, \dots, x_k) the **independent** variables.

$$Y = \beta_0 + \sum_{p=1}^k \beta_p \cdot x_p + E \quad E \sim N(0, \sigma^2)$$

This model is based on the **assumptions**:

1. The **expected** value of **Y** is $\beta_0 + \sum_{p=1}^k \beta_p \cdot x_p$
2. The **unexplained** random deviations are **independent**.
3. The unexplained random deviations have the **same distributions**.
4. This distribution is **normal**.

Morten Frydenberg

Linear and Logistic regression - Note 2.1

10

The multiple linear regression in general

$$Y = \beta_0 + \sum_{p=1}^k \beta_p \cdot x_p + E \quad E \sim N(0, \sigma^2)$$

We see that the assumptions fall in **two parts**:

The **first concerning** the systematic part

and the three other which focus on the error, the unexplained random variation.

Before we turn to how one can check some of the assumptions we will take a closer look at the first assumption.

The **expected** value of **Y** is $\beta_0 + \sum_{p=1}^k \beta_p \cdot x_p$

Morten Frydenberg

Linear and Logistic regression - Note 2.1

11

The assumption of linearity

The **expected** value of **Y** is $\beta_0 + \sum_{p=1}^k \beta_p \cdot x_p$

This is based on three (sub) assumptions:

- a. **Additivity:** The contribution from each of the independent variables are **added**.
- b. **Proportionality:** The contribution from independent variables is **proportional** to its value (with a factor β)
- c. **No effectmodification:** The contribution from one independent variables is **the same** whatever the values are for the other.

Morten Frydenberg

Linear and Logistic regression - Note 2.1

12

The assumption of linearity

The **expected** value of Y is $\beta_0 + \sum_{p=1}^k \beta_p \cdot x_p$

If one consider two persons who differ with

Δx_1 in x_1 , Δx_2 in x_2 ... and Δx_k in x_k

then difference in the **expected** value of Y is :

$$\sum_{p=1}^k \beta_p \cdot \Delta x_p$$

Again we see that the contribution for each of the explanatory variables:

are added,
are proportional to the difference
and does not dependent of the difference in the other

Morten Frydenberg

Linear and Logistic regression - Note 2.1

13

Estimation

It is almost impossible to find the estimates by hand, but easy if you use a computer.

In STATA: `regress lnSBP age45 woman lnBMI25`

(Note first we have to generate `lnSBP`, `age45` `woman` and `lnBMI25`)

Source	SS	df	MS	Number of obs	200
Model	1.05572698	3	.351908994	F(3, 196)	16.46
Residual	4.18969066	196	.021375973	Prob > F	= 0.0000
Total	5.24541764	199	.026358883	R-squared	= 0.2013
				Adj R-squared	= 0.1890
				Root MSE	= .14621

lnSBP	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
woman	.0036329	.0208905	0.17	0.862	-.0375662 .0448319
age45	.0065384	.0012844	5.09	0.000	.0040053 .0090715
lnBMI25	.2583399	.0758295	3.41	0.001	.1087934 .4078864
_cons	4.856592	.0154266	314.82	0.000	4.826169 4.887016

Morten Frydenberg

Linear and Logistic regression - Note 2.1

14

Estimation

The last part of the output:

No CI for σ !
It can be calculated by hand

			$\hat{\sigma}$	Root MSE = .14621
lnSBP	Coef.	Std. Err.	t	P> t [95% Conf. Interval]
woman	.0036329	.0208905	0.17	0.862
age45	.0065384	.0012844	5.09	0.000
lnBMI25	.2583399	.0758295	3.41	0.001
_cons	4.856592	.0154266	314.82	0.000

the $\hat{\beta}$'s

the se's

The CI's

Test for $\beta_2 = 0$

The hypothesis: "no difference in $\ln(sbp)$ between men and women **adjusted** for age and bmi"

Morten Frydenberg

Linear and Logistic regression - Note 2.1

15

The distribution of the estimates

It can be shown that the **estimates of the coefficients** have **normal distributions**, with **means** equal to the **true values**.

The formulas for the standard deviation of the estimates are **complicated**, but they estimated by the **standard errors** given in the output.

The estimated standard deviation of the errors is given by:

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n-k-1} \chi^2(n-k-1)$$

The number of parameters are $k+1$

Which gives the confidence interval:

$$95\% \text{ CI for } \sigma: \hat{\sigma} \cdot \sqrt{\frac{n-k-1}{\chi^2_{n-k-1}(0.975)}} \leq \sigma \leq \hat{\sigma} \cdot \sqrt{\frac{n-k-1}{\chi^2_{n-k-1}(0.025)}}$$

Morten Frydenberg

Linear and Logistic regression - Note 2.1

16

Confidence intervals

Just like in the simple regression we get:
(expect we have $n+k-1$ degrees of freedom).

Exact 95% confidence intervals, CI's, for β_p is found from the estimates and standard errors

$$95\% \text{ CI for } \beta_p: \hat{\beta}_p \pm t_{n-k-1}^{0.975} \cdot \text{se}(\hat{\beta}_p)$$

Where $t_{n-k-1}^{0.975}$ is the upper 97.5 percentile in the t-distribution $n-k-1$ degrees of freedom.

These confidence intervals are found in the output.

Note that if n is large then this percentile is close to 1.96 and one can use the **approximate confidence intervals**:

$$\text{Approx. } 95\% \text{ CI for } \beta_p: \hat{\beta}_p \pm 1.96 \cdot \text{se}(\hat{\beta}_p)$$

Morten Frydenberg

Linear and Logistic regression - Note 2.1

17

The ANOVA table and the F-test

The first part of the output:

An **analysis of variance** table dividing the variation in y in two components: explained by the **model** (i.e. the 3 variables) and the **residual** (the rest)

Source	SS	df	MS	Number of obs	200
Model	1.05572698	3	.351908994	F(3, 196)	16.46
Residual	4.18969066	196	.021375973	Prob > F	= 0.0000
Total	5.24541764	199	.026358883	R-squared	= 0.2013

Source	SS	df	MS	Number of obs	200
Model	1.05572698	3	.351908994	F(3, 196)	16.46
Residual	4.18969066	196	.021375973	Prob > F	= 0.0000
Total	5.24541764	199	.026358883	R-squared	= 0.2013

A F-test testing the hypothesis: "all (except β_0) is zero."

Here the test is highly significant: The model explains a statistically significant part of the variation in y !

Morten Frydenberg

Linear and Logistic regression - Note 2.1

18

The F-test and R-squared

The F-test calculated as: $F = \frac{0.35519}{0.02138} = 16.16$

Source	SS	df	MS
Model	1.05572698	3	0.351908994
Residual	4.18969066	196	0.021375973
Total	5.24541764	199	0.026358883

Number of obs = 200
 F(3, 196) = 16.46
 Prob > F = 0.0000
 R-squared = 0.2013
 Adj R-squared = 0.1890
 Root MSE = .14621

And under the hypothesis it follows an F-distribution with 3 and 196 degrees of freedom.

The R-squared is the amount of the total variation explained by the model (=1.0557/5.2454).

As this will increase if we include more variables in the model one can look at the adjusted R-squared.

Predicted values, residuals and leverages

$$Y = \beta_0 + \sum_{p=1}^k \beta_p \cdot x_p + \epsilon \quad \epsilon \sim N(0, \sigma^2)$$

As in the simple linear regression one can find predicted values, residuals, leverages and standardized residuals:

Predicted value: $\hat{y}_i = \hat{\beta}_0 + \sum_{p=1}^k \hat{\beta}_p \cdot x_{pi}$

Residual: $r_i = y_i - \hat{y}_i = y_i - \sum_{p=1}^k \hat{\beta}_p \cdot x_{pi}$

Leverage: $h_i = \text{a complicated formula}$

Standard-Residual: $z_i = \frac{r_i}{\hat{\sigma} \sqrt{1 - h_i}}$

Leverage

Although the formula for leverage is complicated, the interpretation of leverage is the same:

A high leverage indicates that the data point has extreme values of the explanatory variables and hence a high influence on the estimates.

Checking the model: Independent errors ?

Assumption no. 2: the errors should be independent, is mainly checked by considering how the data was collected.

The assumption is violated if

- some of the persons are relatives (and some are not) and the dependent variable have some genetic component.
- some of the persons were measured using one instrument and others with another.
- in general if the persons were sampled in clusters.

Checking the model 1:

As model is much more complicated than the simple linear regression checking the model is also complicated

Again **assumption no. 2: the errors should be independent**, is mainly checked by considering how the data was collected.

The **distribution of the error** is checked by the same type of plot as for the simple linear regression.

- Histogram and qq-plot of the residuals.
- Plots of residuals versus fitted
- Plots of residuals versus each of the explanatory variables.

Data points that stick out is found by identifying points with large leverages and/or large residuals.

Checking the model 2:

One should also try to check the validity of the linearity assumption that is the assumption of **additivity**, **proportionality** and **no effect modification** (no interaction).

It can be done by:

- Introducing an explanatory variable in a **different scale**, e.g. adding age^2 or $\log(age)$...
- Introducing the explanatory variable as a **categorical** variable instead e.g. use age and divide into **agegroups** instead of age in years.
- Introducing **interaction** between some of the explanatory variables.
-