Working with linear and logistics regression models

Morten Frydenberg ©
Department of Biostatisics, Aarhus Univ, Denmark

Further remarks on logistic regression

- **Diagnostics**: residuals and leverages
- **Test of fit**: The Hosmer-Lemeshow test
- **Enough data**?

General things for regression models:

- **The `lincom` commando**
- **Collinearity** - correlated explanatory variables
- **What model should I use?**
- **Automatic model selection**
- **The consequences of model selection**

Logistic regression models: Diagnostics

In the linear regression we saw some example of statistics:

residuals, **standardized residuals** and **leverage**

which can be used in the **model checking** and search for strange or **influential** data points.

Such statistics can also be defined for the logistic regression model.

But they are much more **difficult to interpret** and **cannot** in general be **recommended**.

Checking the validity of a logistic regression model will mainly be based on **comparing** it with other **models**.

We will return to this later!

Logistic regression models: Test of fit

A common, and to some extend informative, test of fit is the **Hosmer-Lemeshow** test.

Consider the model for obesity from Monday

$$\text{logit}\left(\Pr\left(\text{obese}\right)\right) = \beta_0 + \beta_1 \cdot woman + \beta_2 \cdot \left(age - 45\right)$$

```
Logit estimates                                   Number of obs   =       4690
                                                  LR chi2(2)      =      55.68
                                                  Prob > chi2     =     0.0000
Log likelihood = -1767.7019                       Pseudo R2       =     0.0155
```

obese	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Isex_2	.2743977	.0903385	3.04	0.002	.0973375	.451458
age45	.0344723	.0051354	6.71	0.000	.0244072	.0445374
_cons	-2.147056	.0721981	-29.74	0.000	-2.288561	-2.00555

Significantly better than nothing – but is it good?

Logistic regression models: Test of fit

What about comparing the **estimated prevalence** with the **observed prevalence**?

In the Hosmer-Lemeshow test the data is **divided** into groups (traditionally 10) according to the **estimated** probabilities

and the **observed** and **expected** counts are compared in these groups by a chi-square test.

Most programs, that can fit a logistic regression model, can calculate this test.

In Stata it is done by (**after fitting the model**):

`lfit, group(10) table`

The data is divided into **deciles** after the estimated probabilities.

Logistic regression models: Test of fit

OUTPUT

Logistic model for obese, goodness-of-fit test
(Table collapsed on quantiles of estimated probabilities)

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0841	64	40.9	462	485.1	526
2	0.0953	43	45.5	453	450.5	496
3	0.1045	44	44.6	398	397.4	442
4	0.1112	42	50.3	422	413.7	464
5	0.1217	44	51.4	394	386.6	438
6	0.1332	52	63.0	441	430.0	493
7	0.1456	53	61.7	389	380.3	442
8	0.1592	62	69.8	392	384.2	454
9	0.1834	98	89.9	424	432.1	522
10	0.2407	99	83.8	314	329.2	413

```
      number of observations =      4690
            number of groups =        10
     Hosmer-Lemeshow chi2(8) =         26.01
                 Prob > chi2 =          0.0010
```

One problem: Too many in the tails

Significant difference between observed and expected!

Logistic regression models: Test of fit

`xi: logit obese i.sex*age45`

`lfit, group(10) table`

Logistic model for obese, goodness-of-fit test
(Table collapsed on quantiles of estimated probabilities)

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.0796	36	35.9	466	466.1	502
2	0.1011	42	41.1	406	406.9	448
3	0.1053	49	49.6	429	428.4	478
4	0.1096	50	54.8	458	453.2	508
5	0.1124	52	54.2	436	433.8	488
6	0.1153	51	46.4	355	359.6	406
7	0.1182	52	53.9	410	408.1	462
8	0.1590	76	70.3	428	433.7	504
9	0.2133	96	91.8	391	395.2	487
10	0.3310	97	103.0	310	304.0	407

```
      number of observations =      4690
            number of groups =        10
     Hosmer-Lemeshow chi2(8) =          2.43
                 Prob > chi2 =          0.9650
```

The model 'fits' – when we look at in this way !!!!!!!

Logistic regression models: Do you have enough data?

All inference in logistic regression models are based on asymptotics , i.e. **assuming that you have a lot of data** !

Rule of thumb:
You should have at least **10 events** per variable (parameter) in the model.

A large standard error typical indicates that you have to little information concerning the variable and that the **estimate and standard error are not valid**.

Lower your **ambitions** or get **more data** !

A exact methods exists, but only one (**expensive**) program can do it.

And it will give also wide confidence intervals.

The lincom command after logit or regress

Consider the model:

$$\text{logit}(\Pr(\text{obese})) = \beta_0 + \beta_1 \cdot woman + \beta_2 \cdot (age - 45)$$

```
   obese |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
_Isex_2  |   .2743977   .0903385     3.04   0.002     .0973375     .451458
  age45  |   .0344723   .0051354     6.71   0.000     .0244072    .0445374
  _cons  |  -2.147056   .0721981   -29.74   0.000    -2.288561    -2.00555
------------------------------------------------------------------------------
```

Here men are reference.

If we want to find the log odds for a 45 year old women we can calculate by hand $-2.147+0.274=-1.873$

But what about confidence interval?

We could change the reference to women and fit the model once more.
But.......

The lincom command after logit or regress

$$\text{logit}(\Pr(\text{obese})) = \beta_0 + \beta_1 \cdot woman + \beta_2 \cdot (age - 45)$$

Stata has a command that can be used for this: "lincom"

```
lincom _cons+_Isex

 ( 1)  _Isex_2 + _cons = 0
------------------------------------------------------------------------
obese |      Coef.   Std. Err.       z    P>|z|      [95% Conf. Interval]
------+-----------------------------------------------------------------
  (1) |  -1.8726     .05813    -32.21   0.000    -1.986602   -1.758714
------------------------------------------------------------------------
```

You can add ", or" to get odds/odds ratios.

```
lincom _cons+_Isex,or

( 1)  _Isex_2 + _cons = 0
------------------------------------------------------------------------------
  obese | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------+---------------------------------------------------------------------
    (1) |   .1537145   .0089363   -32.21   0.000     .1371606     .172266
------------------------------------------------------------------------------
```


The lincom command after logit or regress

$$\text{logit}(\Pr(\text{obese})) = \beta_0 + \beta_1 \cdot woman + \beta_2 \cdot (age - 45)$$

Some examples:

Odds for a 42 year old woman:

```
lincom _cons+_Isex-age45*3,or
 ( 1)  _Isex_2 - 3 age45 + _cons = 0
------------------------------------------------------------------------------
   obese | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
     (1) |   .1386122   .0088678   -30.89   0.000     .1222772    .1571295
------------------------------------------------------------------------------
```

Odds ratio for 4.5 age difference:

```
lincom age45*4.5,or
 ( 1)  4.5 age45 = 0
------------------------------------------------------------------------------
   obese | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
     (1) |   1.167804   .0269869     6.71   0.000     1.116091    1.221914
------------------------------------------------------------------------------
```


Collinearity

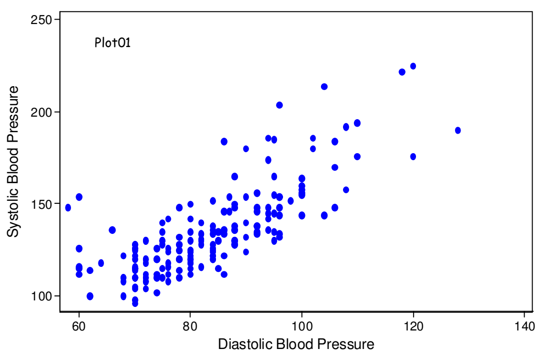
Consider a subsample of the serum cholesterol data set and the **three** models:

```
model 0:   regress logscl sex sbp dbp
model 1:   regress logscl sex     dbp
model 2:   regress logscl sex sbp
```

```
----------------------------------------------------
    Variable |   model0       model1       model2
-------------+--------------------------------------
         sbp |  .00126448                 .0014988    <- Estimate
             |  .00087992                 .0005548    <- Se
             |     0.1524                   0.0075    <- p
         dbp |  .00056517    .00239702
             |  .00164485     .0010424
             |     0.7315       0.0226
         sex |  .02080574    .02446746    .0197773
             |  .02636149    .02631111   .02613048
             |     0.4310       0.3536      0.4501
       _cons |  5.1444085    5.1555212   5.1615877
             |  .09912234    .09909537   .08539118
             |     0.0000       0.0000      0.0000
-------------+--------------------------------------
           N |        194          194         194
----------------------------------------------------
                                      legend: b/se/p
```

Each BP-measure is statistical significant, when the other is removed!

Collinearity



SBP and DBP are **highly positively correlated** that will lead to **highly negatively correlated estimates**!!!

Collinearity

This can be seen by listing the **correlation between the estimates**.
In Stata by the command: `vce, cor`

```
regress logscl sbp dbp sex
vce,cor
             |      sbp       dbp       sex     _cons
-------------+----------------------------------------
         sbp |   1.0000
         dbp |  -0.7750    1.0000
         sex |  -0.0967    0.1135    1.0000
       _cons |  -0.0780   -0.5044   -0.4665    1.0000
```

If two estimates are highly correlated, it indicates that it is very difficult to estimate the **"independent effect"** of the each of the two variables.

Often it is even **nonsense** to try to do it!

Often it see better to try to **reformulate the problem**.

Collinearity

One way to work around the problem of colinearity is to '**ortogonalize**' it:

Create two new variable:

one measures the **blood pressure**

and another that measure the **difference** in systolic and diastolic blood pressure.

Some **candidates**:

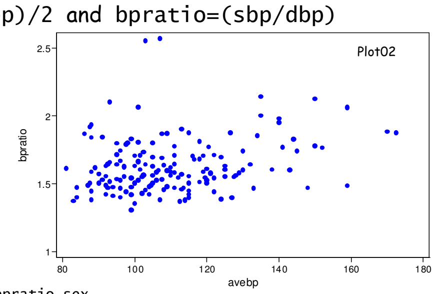
`(sbp+dbp)/2`	and	`(sbp-dbp)`
`(sbp+dbp)/2`	and	`(sbp/dbp)`
`ln(sbp*dbp)/2`	and	`ln(sbp/dbp)`

We will here consider the second pair.

Collinearity

`avebp=(sbp+dbp)/2 and bpratio=(sbp/dbp)`

Only weakly associated




```
regress logscl avebp bpratio sex
vce,cor
             |    avebp   bpratio       sex     _cons
-------------+----------------------------------------
       avebp |   1.0000
     bpratio |  -0.2456    1.0000
         sex |   0.0382   -0.1041    1.0000
       _cons |  -0.4542   -0.6874   -0.2585    1.0000
```


Collinearity

The serum cholesterol data set and the **three** models:

model 0: `regress logscl sex avebp bpratio`
model 1: `regress logscl sex avebp`
model 2: `regress logscl sex bpratio`

```
------------------------------------------------------
    Variable |   model0       model1       model2
-------------+----------------------------------------
       avebp |  .00198973    .00206564
             |   .0007887    .00076285
             |     0.0125       0.0074
     bpratio |  .02769662                 .07148118
             |  .07067134                 .06946246
             |     0.6956                    0.3048
         sex |  .02060675    .02168128    .01806662
             |  .02632924      .026128    .02667689
             |     0.4348       0.4077       0.4991
       _cons |  5.1003417    5.1351912    5.2485724
             |  .12936418    .09374803    .11685799
             |     0.0000       0.0000       0.0000
-------------+----------------------------------------
           N |        194          194          194
------------------------------------------------------
                                       legend: b/se/p
```

Blood pressure seems to play a role,

The ratio between SBP and DBP might not.

Collinearity

Look out for it:

- systolic and diastolic blood pressure
- 24 hour blood pressure and 'clinical' blood pressure
- weight and height
- age and parity
- age and time since menopause
- BMI and skinfold measure
- age , birth cohort and calendar time
- volume and concentration
- ……

Remember you will need **a huge amount** of data to disentangle the effects of correlated explanatory variables

Which model should I use?

This a hard question!

The first thing to remember is that all models are **approximations**!

The "true" , the "best" or the "correct" model **does not exist!**

The **quality** of a model depends on what you want to use it for.

So the first thing to clarify is:
What is the **purpose** of your analysis – what is the **aim** of your data collection?

Different purposes – different models!!!!!

When you have found out what you want, you still have an **infinity** of models to chose between.

Which model should I use?

The choice is always a choice between **complicated** and **less complicated** models.

Complicated models are often better models, in the sense that they are **better approximations** to the truth.

But complicated models can be:

Very hard to **estimate** – many parameters.

Very hard to **understand**.

Very hard to **communicate**.

So in these senses they are **not so good** models.

Which model should I use?

Less complicated models are often not as good models, in the sense that they are **not so good approximations** to the truth.

But less complicated models can be:

Easy to **estimate** – few parameters.

Easy to **understand**

Easy to **communicate**

So in these senses they are **better** models.

The first thing to remember is that all models are **approximations**!

Statistical significance has nothing to do with the quality of the model!

Which model should I use?

You can often divide the explanatory variables **into groups**:

1: Variables of **primary interest- main exposure.**

2: Variables of **less interest** – variables you want to **adjust** for.

A good model will try to introduce the **first** group in an **interpretable** way into model.

- You want to **known** "how they work".

E.g. if you specifically are interested on the "effect" of age you should model age in a **understandable** way.

Still you have to look out for collinearity.

Which model should I use?

The **second** type of variables can be introduces any way you like.

It can be very complicated – you do not care- as long as they do the job - that is, **adjust sufficiently**.

If you are not interested in age in itself – you just want to adjust – then age can be introduce in a complicated/weird way, e.g. a fourth order polynomial.

In **general**:

Models with many parameters need more data to obtain precise estimate.

Again few data – lower your ambition !

Automatic model selection

Some programs (even Stata) have programmed algorithms for **automatic model selection**!

That is, procedures that will find the "best" model to answer you question with out knowing what **you want**, **know** or anything else about the **problem**!

It is very rarely of any interest, especially if you have **little data.**

There are in general three types of such algorithms:

Backward selection : You specify a **start model** and the procedure will reduce the model by **removing** variables from the model until nothing can be removed.

The **criteria** for removing variables are typically **high p-values**.

Automatic model selection

Forward selection : You specify a **start model** together with a **list** of variables that might be included in a model. The procedure will build the model by **adding variable** from the list to the model until nothing can be added.

The **criteria** for adding variables is typically **low p-values.**

Best subset selection: You specify a **list** of variables that might be included in a model and **number** of variables you want in the model. The procedure will then search among the possible models and find the "best".

The **criteria** is the typically the **highest likelihood** or related statistics.

Automatic model selection

Some comments:

- These procedures **do not know anything about the subject**.
- They will not consider **transformation** of the variables.
- or **interaction**.
- They will chose **arbitrarily** between explanatory variables that are highly correlated.

Model selection and some implications

Even when you do not use an automatic model selection procedures :The **final** model is selected!

That is, you have spend some time **working** with the model you present!

You might choose only to include **statistical significant** variables in the model.

You might group **two levels** of a explanatory variable **into one level** if there is no statistical significant difference between the two levels.

The implications of this selection:

- The **estimates** tend to be too far **away from null**.
- The **standard errors** are too **small**.
- The **CI's** are to **narrow** and the **p-values** too **small**.