Chapter 8

Categorical covariates

In this chapter we discuss how to incorporate categorical covariates into a regression model.

8.1 Incorporating categorical covariates in a regression model

In many applications one is interested to investigate the effect of a covariate, which divides the whole population into several subgroups (not only just in two, as in the case of a binary covariate). For example in the study of allergies in early childhood, one might divide the parents according to their smoking status into four groups, i.e. we consider the covariate X_1 defined as

$$X_1 = \begin{cases} 1 & \text{mother and father are both non-smokers} \\ 2 & \text{mother non-smoker, father smoker} \\ 3 & \text{mother smoker, father non-smoker} \\ 4 & \text{mother and father are both smokers} \end{cases}$$

and we are interested in the differences between the groups with respect to the probabilities

$$\pi(x) = P(Y = 1 | X_1 = x)$$

with

$$Y = \begin{cases} 1 & \text{Child do not suffer from an allergy} \\ 0 & \text{Child suffer from an allergy} \end{cases} .$$

We can of course simply estimate these probabilities by the corresponding relative frequency of children with an allergy in each category as shown in Table 8.1. However, our aim is to assess the differences between the categories, i.e. we are interested in the six pairwise comparisons and a quantification of the corresponding six differences. We can quantify these differences on the logit scale by introducing the parameters

$$\delta_{kl} = \text{logit}\, \pi(l) - \text{logit}\, \pi(k) \ ,$$

Category	$\hat{\pi}(x)$	logit $\hat{\pi}(x)$
1	0.22	-1.295
2	0.33	-0.720
3	0.28	-0.920
4	0.40	-0.425

Table 8.1: The four categories defined by X_1 and the relative frequency $\hat{\pi}$ of children suffering from allergies within each category on the probability and the logit scale.

k	l	$\hat{\delta}_{kl} = \text{logit}\,\hat{\pi}(l) - \text{logit}\,\hat{\pi}(k)$
1	2	0.575
1	3	0.375
1	4	0.870
2	3	-0.200
2	4	0.295
3	4	0.495

Table 8.2: The six pairwise comparisons: Differences in the empirical frequency of children with allergies expressed on the logit scale.

and using the empirical estimates for $\pi(x)$ we can of course obtain estimates for the parameters δ_{kl} as indicated in Table 8.2.

The reader should note that indeed any of the six values δ_{kl} is of subject matter interest in the application. δ_{12} describes the effect of "Father only smoking", δ_{13} describes the effect of "Mother only smoking", δ_{14} describes the effect of "Father and mother smoking", δ_{24} describes the effect of "Mother smoking additional to father smoking", δ_{34} describes the effect of "Father smoking additional to mother smoking", and δ_{23} compares the the effect of "Mother only smoking" with the effect of "Father only smoking".

If we would like to approach the estimation of δ_{kl} in the framework of a regression model, we have a slight complication. This is related to the fact that if we know for example δ_{12} and δ_{23}, we also know $\delta_{13} = \delta_{12} + \delta_{23}$, and if we know δ_{12} and δ_{13} we know $\delta_{23} = \delta_{13} - \delta_{12}$ etc. So although we have six parameters, we have only three "free" parameters in the sense, that if we know for example δ_{12}, δ_{13}, and δ_{14}, then we also know $\delta_{23} = \delta_{13} - \delta_{12}$, $\delta_{24} = \delta_{14} - \delta_{12}$ and $\delta_{34} = \delta_{14} - \delta_{13}$. In a regression model formulation we will find only three "free" parameters, and they are typically chosen as

$$\beta_x^{(1)} = \delta_{1x} \quad \text{for } x = 1, 2, 3, 4$$

with the convention $\beta_1^{(1)} = \delta_{11} = 0$. The logistic model for $\pi(x) = P(Y = 1|X_1 = x)$ reads now

$$\text{logit}\,\pi(x) = \beta_0 + \beta_x^{(1)} \quad \text{with } x = 1, 2, 3, \text{or } 4 \ .$$

We can simply verify that $\beta_x^{(1)}$ is identical to δ_{1x} by noting

$$\delta_{1x} = \text{logit}\,\pi(x) - \text{logit}\,\pi(1) = \beta_0 + \beta_x^{(1)} - \beta_0 = \beta_x^{(1)} \ .$$

If we fit this logistic regression model, we obtain an output like

variable	beta	SE	95%CI	p-value
smokep_2	0.575	0.207	[0.168,0.981]	0.006
smokep_3	0.375	0.203	[-0.024,0.773]	0.065
smokep_4	0.870	0.163	[0.550,1.190]	<0.001

and we can observe that $\hat{\beta}_{1,2}$, $\hat{\beta}_{1,3}$, and $\hat{\beta}_{1,4}$ coincide with the empirical estimates of δ_{12}, δ_{13} and δ_{14} in Table 8.2.

Now the simple comparison of the four categories might be misleading with respect to an evaluation of the effect of smoking, because the effect of smoking might be confounded with other factors. One candidate might be the parental allergy status, because we have previously seen, that at least mothers with an allergy tend to smoke less than mothers without an allergy. To adjust for this, we can consider a second covariate

$$X_2 = \begin{cases} 1 & \text{mother and father have no allergies} \\ 2 & \text{mother not affected by allergies, father affected} \\ 3 & \text{mother affected by allergies, father not affected} \\ 4 & \text{mother and father are both affected} \end{cases}$$

and we can consider the logistic regression model

$$\text{logit}\, \pi(x_1, x_2) = \beta_0 + \beta^{(1)}_{x_1} + \beta^{(2)}_{x_2} \ .$$

The output from fitting this model may look like

variable	beta	SE	95%CI	p-value
smokep_2	0.583	0.209	[0.173,0.993]	0.005
smokep_3	0.513	0.209	[0.103,0.922]	0.014
smokep_4	1.017	0.170	[0.683,1.350]	<0.001
allergyp_2	0.367	0.176	[0.023,0.711]	0.037
allergyp_3	0.523	0.187	[0.157,0.890]	0.005
allergyp_4	0.759	0.191	[0.385,1.133]	<0.001

and we can indeed observe that the adjusted estimates are larger than the unadjusted estimates.

Of course, you can include categorical covariates the same way in the classical linear regression model and you can mix categorical, continuous and binary covariates.

8.2 Some technicalities in using categorical covariates

In our above considerations we have slightly cheated. We have previously said that a logistic regression model looks like

$$\text{logit}\, \pi(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \ . \qquad \text{(A)}$$

Now our model looks like

$$\text{logit}\,\pi(x_1, x_2, \ldots, x_p) = \beta_0 + \beta^{(1)}_{x_1} + \beta^{(2)}_{x_2} + \ldots \beta^{(p)}_{x_p} \quad . \tag{B}$$

The reader should not worry about this. As long as we are interested in a correct interpretation of our parameters, the representation (B) together with the considerations of the previous section is perfect.

However, with respect to the computation of the estimates it can be useful to shift the representation (B) into one which follows the structure of (A), as this allows us to use the same computational procedures to handle categorical covariates as to handle continuous and binary covariates. This can be achieved by introducing for each category an indicator variable. If, for example, covariate X_j has k_j categories, then we define the $k_j - 1$ indicators

$$\tilde{X}_{jk} = \begin{cases} 1 & \text{if } X_j = k \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } k = 2, 3, \ldots, k_j$$

which allows to represent $\beta^{(j)}_{x_j}$ as

$$\beta^{(j)}_2 \tilde{x}_{j,2} + \beta^{(j)}_3 \tilde{x}_{j,3} + \ldots + \beta^{(j)}_{k_1} \tilde{x}_{jk_1}$$

(note that at most one of the indicator variables is 1, hence we pick up this way the correct $\beta^{(j)}_{x_j}$) and we can express (B) as

$$\begin{aligned} &\text{logit}\,\pi(\tilde{x}_{12}, \tilde{x}_{13}, \ldots, \tilde{x}_{1k_1}, \tilde{x}_{22}, \tilde{x}_{23}, \ldots, \tilde{x}_{2k_2}, \ldots, \tilde{x}_{p2}, \tilde{x}_{p3}, \ldots, \tilde{x}_{pk_p}) \\ &\quad = \beta_0 + \beta^{(1)}_2 \tilde{x}_{1,2} + \beta^{(1)}_3 \tilde{x}_{1,3} + \ldots + \beta^{(1)}_{k_1} \tilde{x}_{1k_1} + \beta^{(2)}_2 \tilde{x}_{22} + \beta^{(2)}_3 \tilde{x}_{23} + \ldots + \beta^{(2)}_{k_2} \tilde{x}_{2k_2} + \ldots + \\ &\qquad + \beta^{(p)}_2 \tilde{x}_{p2} + \beta^{(p)}_3 \tilde{x}_{p3} + \ldots + \beta^{(p)}_{k_1} \tilde{x}_{pk_1} \quad . \end{aligned}$$

Most statistical packages do this step internally when handling categorical covariates.

It is an unlucky consequence of reducing the parameters δ_{jk} of interest to a few free parameters $\beta^{(j)}_k$ that in the output of most computer programs we only find standard errors, confidence intervals and p-values for the free parameters $\beta^{(j)}_k$ corresponding to the interest parameters δ_{1k}, but no inference on the remaining parameters δ_{2k} , δ_{3k}, ... In the example of the previous section there was some interest in the parameter δ_{23} describing the difference in the effect of maternal and paternal smoking. We can of course compute an estimate for this parameter (after adjustment for the allergy status) as

$$\hat{\delta}_{23} = \hat{\beta}^{(1)}_3 - \hat{\beta}^{(1)}_2 = 0.51 - 0.58 = -0.07 \quad .$$

But we have no chance to find a confidence interval or p-value for this parameter from the output, because the standard error of $\hat{\delta}_{23}$ is not just a function of the standard errors of $\hat{\beta}^{(1)}_3$ and $\hat{\beta}^{(1)}_2$.

So we have to convince the statistical package to produce standard errors, confidence intervals and p-values for $\hat{\delta}_{23}$. The better packages allow to require this directly by certain options or

additional procedures. Some packages only allow to change the category we use as a reference category. In our examples so far we have used 1 as reference category, such that $\beta_k^{(j)}$ refers to the difference between category k and category 1 in covariate j. However, this is completely arbitrary. You can chose any other category c as reference category, such that $\beta_k^{(j)}$ refers to the difference between category k and category c, and especially $\beta_c^{(j)} = 0$. So to obtain inference on $\hat{\delta}_{23}$, we have just to convince our program to use 2 as reference category, such that $\beta_3^{(j)} = \delta_{23}$. The real lousy programs require that the user changes the coding of the covariates or to define indicator variables by hand.

Remark: We would like to emphasize, that the assignment of numbers 1, 2, 3, ... to the categories of a categorical covariate is completely arbitrary. You can use any other numbers or letters or what you want to label the categories. Most statistical programs use numbers, and we do it in this course for convenience, too. However, these numbers have only the task to distinguish the categories and no further meaning.

8.3 Testing the effect of a categorical covariate

If it is the aim of a regression analysis to show that a certain covariate has an influence on the outcome, then we can in the case of a binary or continuous covariate just try to reject the null hypothesis $H_0 : \beta_j = 0$. In the case of a categorical covariate we have to reject the null hypothesis

$$H_0 : \delta_{kl} = 0 \text{ for all } k, l$$

in order to be able to state that the covariate has an influence. We can of course re-express this null hypotheses in the free parameters of our regression model as

$$H_0 : \beta_k^{(j)} = 0 \text{ for all } k \ .$$

Most statistical packages allow to perform such a test, either using the the Wald test principle or the likelihood ratio test principle, which give usually similar results. (For more details see Appendix C.3.)

The standard output of a regression model includes typically p-values for the single parameters, but the reader should be aware of that there is no simple relation between these p-values and the corresponding overall p-value for the test of $H_0 : \delta_{kl} = 0$ for all k, l, and it may happen that the results may look rather conflicting. Table 8.3 and Table 8.4 illustrate two typical conflicts. In the first case the p-values of the two regression coefficients do not look very impressive. However, the overall p-value is less than 0.05. This happens because the reference category *sitting* is actually intermediate between the two other categories: Subjects with low physical work load have fewer sick days than subjects with an occupation implying typical sitting as the working position, and subjects with high physical work load have more sick days. So in this table we have omitted the highest of the three pairwise differences, namely that between high and low physical work load with a $\hat{\beta}$ of 3.4 and a p-value of 0.008. In the second example (Table 8.4)

we compare fourteen regions in Denmark with respect to the hospitalisation costs per patients hospitalised. At first sight, 4 of the 13 regression coefficients are – at the 5% level – significant different from 0, suggesting, that we have some evidence for differences between the regions. However, the overall p-value is only 0.25. Here we have to realise that we can build actually 91 pairs among the fourteen regions, and that our selection is somewhat "biased", because it turns out that the reference region Copenhagen is the most expensive: All other regions are less expensive, as indicated by the negative regression coefficients. Hence the presented results tend to include big pairwise differences, and in any case the largest of all pairwise differences. It is more appropriate to relate the 4 significant differences to the 91 differences over all, and this is much less impressive, as we have to expect about 4.5 significant differences among 91 differences by chance, even if their are no true differences between the regions.

category	beta	CI	p-value	overall p-value
sitting (reference)				
low physical work load	-1.8	[-4.4,0.7]	0.17	} 0.031
high physical work load	1.6	[-0.9,4.3]	0.21	

Table 8.3: Regression analysis of number of sick days/year in dependence on the type of occupation.

category	beta	CI	p-value	overall p-value
Copenhagen Region (reference)				
Frederiksborg Region	-0.38	[-0.94, 0.17]	0.179	
Roskilde Region	-0.60	[-1.15, -0.04]	0.034	
West Zealand Region	-0.44	[-1.00, 0.11]	0.115	
Storstrm Region	-0.05	[-0.60, 0.51]	0.864	
Bornholm Region	-0.33	[-0.89, 0.23]	0.244	
Funen Region	-0.64	[-1.20, -0.09]	0.023	
South Jutland Region	-0.49	[-1.04, 0.07]	0.085	} 0.25
Ribe Region	-0.09	[-0.64, 0.46]	0.754	
Vejle Region	-0.57	[-1.13, -0.02]	0.043	
Ringkjbing Region	-0.20	[-0.75, 0.36]	0.496	
Aarhus Region	-0.22	[-0.77, 0.34]	0.442	
Viborg Region	-0.26	[-0.82, 0.29]	0.351	
North Jutland Region	-0.59	[-1.14, -0.03]	0.038	

Table 8.4: Regressions analysis of the hospitalisation costs (in 1000 USD) of each patient hospitalised in Denmark in 1998-2002 in dependence on the region, adjusted for age and sex.

8.4 The handling of categorical covariates in Stata

The two categorical covariates used in Section 8.1 can be found in the dataset `allergy3`.

```
. use allergy3, clear

. list in 1/10

     +---------------------------------------+
     | childnr   allergyc   smokep   allergyp |
     |---------------------------------------|
  1. |       1          0        1          1 |
  2. |       2          1        1          1 |
  3. |       3          0        1          3 |
  4. |       4          1        1          3 |
  5. |       5          0        1          2 |
     |---------------------------------------|
  6. |       6          1        1          2 |
  7. |       7          0        1          4 |
  8. |       8          1        1          4 |
  9. |       9          0        3          1 |
 10. |      10          1        3          1 |
     +---------------------------------------+
```

We have besides the binary outcome variable `allergyc` the two categorical covariates `smokep` and `allergyp`. The coding of these covariates has been explained in Section 8.1.

To compute the unadjusted effects for the different categories of the parental allergy status on the allergy status of the child we can use

```
. xi: logit allergyc i.smokep
i.smokep          _Ismokep_1-4        (naturally coded; _Ismokep_1 omitted)

Iteration 0:   log likelihood = -693.45852
Iteration 1:   log likelihood = -678.39626
Iteration 2:   log likelihood = -678.29045
Iteration 3:   log likelihood = -678.29044

Logistic regression                               Number of obs   =       1125
                                                  LR chi2(3)      =      30.34
                                                  Prob > chi2     =     0.0000
Log likelihood = -678.29044                       Pseudo R2       =     0.0219

------------------------------------------------------------------------------
    allergyc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Ismokep_2 |   .5746724   .2072584     2.77   0.006     .1684535    .9808914
```

```
  _Ismokep_3 |   .3748066    .203354     1.84   0.065      -.02376    .7733731
  _Ismokep_4 |    .870092   .1630818     5.34   0.000     .5504575    1.189727
       _cons |  -1.294727   .1261929   -10.26   0.000    -1.542061   -1.047393
------------------------------------------------------------------------------
```

The `xi:` construct tells Stata, that we would like to include categorical covariates in our model, and the `i.smokep` tells Stata, that we would like to use `smokep` as a categorical covariate. If we would just say `smokep`, Stata would use this variable as a continuous one. Stata has chosen the category with the smallest value as a reference category. This is the standard strategy of Stata, called "natural coding". This first line of the output saying "`naturally coded`" reminds the user on this.

To adjust our effect estimates for the effect of the allergy status of the parents, we add the covariate `allergyp` to the model. Since this is also a categorical covariate, we have to use again the `i.`-notation.

To obtain the odds ratio adjusted for maternal smoking we can use

```
. xi: logit allergyc i.smokep i.allergyp
i.smokep          _Ismokep_1-4        (naturally coded; _Ismokep_1 omitted)
i.allergyp        _Iallergyp_1-4      (naturally coded; _Iallergyp_1 omitted)

Iteration 0:   log likelihood = -693.45852
Iteration 1:   log likelihood = -669.34423
Iteration 2:   log likelihood = -669.10843
Iteration 3:   log likelihood = -669.10835
Iteration 4:   log likelihood = -669.10835

Logistic regression                               Number of obs   =       1125
                                                  LR chi2(6)      =      48.70
                                                  Prob > chi2     =     0.0000
Log likelihood = -669.10835                       Pseudo R2       =     0.0351

------------------------------------------------------------------------------
    allergyc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Ismokep_2 |   .5829577   .2090785     2.79   0.005     .1731714     .992744
  _Ismokep_3 |    .512587   .2089995     2.45   0.014     .1029555    .9222186
  _Ismokep_4 |   1.016536   .1702369     5.97   0.000     .6828781    1.350194
_Iallergyp_2 |   .3668414   .1755227     2.09   0.037     .0228233    .7108595
_Iallergyp_3 |   .5231563   .1870375     2.80   0.005     .1565696     .889743
_Iallergyp_4 |   .7586479   .1907552     3.98   0.000     .3847746    1.132521
       _cons |  -1.692582   .1639953   -10.32   0.000    -2.014007   -1.371157
```

The `xi:` construct does nothing else then to add the indicator variables described in Section 8.2. These covariates remain in the dataset, so we can take a look on them:

```
. list smokep _Ismokep_* in 1/10

     +---------------------------------------+
     | smokep   _Ismok~2   _Ismok~3   _Ismok~4 |
     |---------------------------------------|
  1. |      1          0          0          0 |
  2. |      1          0          0          0 |
  3. |      1          0          0          0 |
  4. |      1          0          0          0 |
  5. |      1          0          0          0 |
     |---------------------------------------|
  6. |      1          0          0          0 |
  7. |      1          0          0          0 |
  8. |      1          0          0          0 |
  9. |      3          0          1          0 |
 10. |      3          0          1          0 |
     +---------------------------------------+

. list allergyp _Iallergyp* in 1/10

     +-----------------------------------------+
     | allergyp   _Ialle~2   _Ialle~3   _Ialle~4 |
     |-----------------------------------------|
  1. |        1          0          0          0 |
  2. |        1          0          0          0 |
  3. |        3          0          1          0 |
  4. |        3          0          1          0 |
  5. |        2          1          0          0 |
     |-----------------------------------------|
  6. |        2          1          0          0 |
  7. |        4          0          0          1 |
  8. |        4          0          0          1 |
  9. |        1          0          0          0 |
 10. |        1          0          0          0 |
     +-----------------------------------------+
```

We can see that Stata has created indicator variables for the categories 2, 3, and 4 of the covariates `smokep` and `allergyp`. These variables start with an `_I`, followed by the original name, followed by an `_` and finished by the number of the category. Since `list` tends to abbreviate long variable names, you can see here the full names:

```
. list smokep _Ismokep_* in 1/5, abbreviate(16)

     +-------------------------------------------+
     | smokep   _Ismokep_2   _Ismokep_3   _Ismok |
     |-------------------------------------------|
  1. |      1            0            0         0 |
  2. |      1            0            0         0 |
  3. |      1            0            0         0 |
  4. |      1            0            0         0 |
  5. |      1            0            0         0 |
     +-------------------------------------------+

. list allergyp _Iallergyp* in 1/5, abbreviate(16)

     +-----------------------------------------------------+
     | allergyp   _Iallergyp_2   _Iallergyp_3   _Iallergyp |
     |-----------------------------------------------------|
  1. |        1              0              0             0 |
  2. |        1              0              0             0 |
  3. |        3              0              1             0 |
  4. |        3              0              1             0 |
  5. |        2              1              0             0 |
     +-----------------------------------------------------+
```

These names appear in the output of the `logit` command, as Stata uses these dummy indicator variables as covariates as explained in Section 8.2. Note that the suffices `_2`, `_3` and `_4` do not refer directly to the numbers use in coding the covariates, but just to the ordering of these values. If you have a covariate `xyz` coded as 0, 1, 2, and 3, Stata will still create the dummies `_Ixyz_2`, `_Ixyz_3` and `_Ixyz_4`, with for example `_Ixyz_2` referring to category 1 of `xyz`.

If we now want to know the additional effect of father's smoking on the top of mother's smoking, we have to compare the category 4 = *mother and father are both smokers* with category 2 = *mother non-smoker, father smoker*. We can do this using Stata's `lincom` command:

```
. lincom _Ismokep_4-_Ismokep_2

 ( 1)  - [allergyc]_Ismokep_2 + [allergyc]_Ismokep_4 = 0

------------------------------------------------------------------------------
    allergyc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .4335785   .2001199     2.17   0.030     .0413506    .8258063
------------------------------------------------------------------------------
```

such that we obtain as the estimate of this difference $\hat{\delta}_{42} = \hat{\beta}_1^{(4)} - \hat{\beta}_1^{(2)} = 1.02 - 0.58 = 0.43$ together with a standard error and a confidence interval.

To test the null hypothesis that the smoking status of the parents is not associated with the allergy status of the child (taking the allergy status of the parents into account), we can use Stata's `test` or `testparm` command to obtain the overall p-value:

```
. test _Ismokep_2 _Ismokep_3 _Ismokep_4

 ( 1)  [allergyc]_Ismokep_2 = 0
 ( 2)  [allergyc]_Ismokep_3 = 0
 ( 3)  [allergyc]_Ismokep_4 = 0

           chi2(  3) =   36.12
         Prob > chi2 =    0.0000

. testparm _Ismokep*

 ( 1)  [allergyc]_Ismokep_2 = 0
 ( 2)  [allergyc]_Ismokep_3 = 0
 ( 3)  [allergyc]_Ismokep_4 = 0

           chi2(  3) =   36.12
         Prob > chi2 =    0.0000
```

Both commands do exactly the same, namely to perform a Wald test on the null hypotheses, that all the three regression coefficients are 0. `test` requires to specify all parameters, whereas `testparm` allows to use the * as a wildcard. In our example we can see that the effect of paternal smoking is highly significant: the p-value is less than 0.0001. (Stata shows the value 0.0000, but this means nothing else but that the p-value is so small that it cannot be represented by four decimal digits. Hence the correct way to report this results is to write "p<0.0001".)

To perform a likelihood ratio test, we have to perform the same steps as in the case of testing a single parameter: It requires to fit the full model, to save the result under a chosen name. e.g. "A" using the `estimates store` command, to fit the model without the covariate to be tested and to compare the likelihood of this model with that of the full model saved as "A" using the `lrtest` command:

```
. xi: logit allergyc i.smokep i.allergyp
i.smokep          _Ismokep_1-4        (naturally coded; _Ismokep_1 omitted)
i.allergyp        _Iallergyp_1-4      (naturally coded; _Iallergyp_1 omitted)

Iteration 0:   log likelihood = -693.45852
Iteration 1:   log likelihood = -669.34423
Iteration 2:   log likelihood = -669.10843
```

```
Iteration 3:   log likelihood = -669.10835
Iteration 4:   log likelihood = -669.10835

Logistic regression                                     Number of obs   =       1125
                                                        LR chi2(6)      =      48.70
                                                        Prob > chi2     =     0.0000
Log likelihood = -669.10835                             Pseudo R2       =     0.0351

------------------------------------------------------------------------------
    allergyc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Ismokep_2 |   .5829577   .2090785     2.79   0.005     .1731714     .992744
  _Ismokep_3 |    .512587   .2089995     2.45   0.014     .1029555    .9222186
  _Ismokep_4 |   1.016536   .1702369     5.97   0.000     .6828781    1.350194
_Iallergyp_2 |   .3668414   .1755227     2.09   0.037     .0228233    .7108595
_Iallergyp_3 |   .5231563   .1870375     2.80   0.005     .1565696     .889743
_Iallergyp_4 |   .7586479   .1907552     3.98   0.000     .3847746    1.132521
       _cons |  -1.692582   .1639953   -10.32   0.000    -2.014007   -1.371157
------------------------------------------------------------------------------

. estimates store A

. xi: logit allergyc i.allergyp
i.allergyp        _Iallergyp_1-4      (naturally coded; _Iallergyp_1 omitted)

Iteration 0:   log likelihood = -693.45852
Iteration 1:   log likelihood = -688.01197
Iteration 2:   log likelihood = -687.99853
Iteration 3:   log likelihood = -687.99853

Logistic regression                                     Number of obs   =       1125
                                                        LR chi2(3)      =      10.92
                                                        Prob > chi2     =     0.0122
Log likelihood = -687.99853                             Pseudo R2       =     0.0079

------------------------------------------------------------------------------
    allergyc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Iallergyp_2 |   .3579979   .1726649     2.07   0.038     .0195809    .6964149
_Iallergyp_3 |   .3292157   .1790193     1.84   0.066    -.0216556    .6800871
_Iallergyp_4 |   .5549968   .1820931     3.05   0.002      .198101    .9118927
       _cons |  -1.057454   .1042384   -10.14   0.000    -1.261758   -.8531508
------------------------------------------------------------------------------

. lrtest A

Likelihood-ratio test                                 LR chi2(3)  =     37.78
```

```
(Assumption: . nested in A)                    Prob > chi2 =    0.0000
```

We can again observe that the effect of `smoking` is highly significant.

Remark: You can use the `xi:` construct and the `i.`-notation also for string variables. In this case the category which is the first in alphabetic order is used as the references group. For numeric variables the category with the smallest number is used. You can change this behaviour. (Type `help xi` for more information about the `xi` construct.)

8.5 Presenting results of a regression analysis based on categorical covariates in a table

In most publications, results of a regression analysis are summarised in a table. As long as one is working with binary or continuous covariates only, it is straightforward to present the estimated regressions coefficient, confidence intervals and p-values. However, working with a categorical covariate, we have regression coefficients, estimates of pairwise differences, confidence intervals and p-values for all of them and the overall p-values. So we can rise the question which of these numbers we should present.

Let us extend our example of the study on allergies in early childhood by considering the following five covariates:

- *Allergystatus of parents:* A categorical covariate with the four categories mentioned above.
- *Social class:* A categorical covariate with the categories I, II and III.
- *Region:* A categorical covariate with the two categories *rural* and *urban*.
- *Breast feeding:* A binary variable (*yes* or *no*).
- *Age of mother at birth:* A continuous covariate measured in years.

Table 8.5 shows results of the regression analysis in a way, you can typically find in the medical literature. Table 8.6 shows the results in a manner, which I personally find much more fair and useful. Let us discuss the differences.

- In the first table the results for *Allergy of parents* are presented relative to a reference category chosen. This way we are not able to judge all pairwise differences, for example we cannot obtain a confidence interval for the differences between the categories *Father only* and *Mother only*. Since for a categorical covariate like *Allergy of parents* we are typically

covariate	$\hat{\beta}$	95% CI	p-value
Allergy of parents (Reference: None)			
mother only	0.43	[0.11,0.76]	.0085
father only	0.30	[-0.04,0.63]	.083
both	0.86	[0.51,1.20]	<0.0001
Social class (Reference: I)			
II	0.13	[-0.17,0.42]	.41
III	0.51	[0.20,0.81]	.0012
Region	0.30	[0.06,0.54]	.016
Breast feeding	0.36	[0.12,0.61]	.0036
Age of mother at birth	0.00	[-0.01,0.02]	.76

Table 8.5: Results of a logistic regression analysis of the risk of developing an allergy in early childhood in dependence of the parental allergy status, social class, region, breast feeding and age of the mother

covariate	$\hat{\beta}$	95% CI	p-value
Allergy of parents			<0.0001
mother vs none	0.43	[0.11,0.76]	
father vs nome	0.30	[-0.04,0.63]	
both vs none	0.86	[0.51,1.20]	
father vs mother	-0.14	[-0.51,0.24]	
both vs father	0.42	[0.03,0.81]	
both vs mother	0.56	[0.16,0.96]	
Social class			.0032
middle vs. low	0.13	[-0.17,0.42]	
high vs. middle	0.38	[0.09,0.67]	
Region			.016
urban vs rural	0.30	[0.06,0.54]	
Breast feeding	0.36	[0.12,0.61]	.0036
Age of mother at birth	0.00	[-0.01,0.02]	.76

Table 8.6: Results of a logistic regression analysis of the risk of developing an allergy in early childhood in dependence of the parental allergy status, social class, region, breast feeding and age of the mother

interested in all pairwise differences, we should include effect estimates and confidence intervals for all differences, as done in the second table.

There is also some danger, that by representing only the difference to one reference category, we give an overoptimistic impression about the differences among the categories. Since we typical chose as a reference category a category with a low or high risk, we tend to present big differences. This is illustrated by our example: The first table selects the biggest, the third biggest and the fifth biggest difference and omits the smallest, third smallest and

fifth smallest. So on average the three differences in the first table are bigger than the six differences in the second table.

– In the second table the results for *Social class* are presented by only two of the three pairwise differences. This seems to be on first sight a contradiction to the recommendation just given. However, social class is is covariate with ordered categories (or an ordinal covariate). In such a situation the interest is typically in the differences between neighboring categories, and hence the presentation in the second table is more appropriate. (If one feels that the differences between III and I is also of scientific interest, one can of course add this difference.)
The argument of the overoptimistic impression applies also in the case of ordered categories, and it is even more relevant here, because we often chose the lowest category as reference, and if the risk is increasing with the categories, we run exactly in the problem discussed above.
There exist also a further argument to prefer the presentation of the second table. In the case of ordered categories, the scientific interest lies often in the question whether the difference between the second and the third category is bigger or smaller than the difference between the first and the second category. This can we directly judge in the second table where we can compare 0.38 with 0.13. One may argue, that we can of course just subtract the two estimates 0.51 and 0.13 of the first table to obtain the difference between the third and second category. However, if one present results as odds ratios, this task becomes more difficult. (See Exercise 8.7 later.)

– Table 8.6 is much more parsimonious with p-values than Table 8.5. It only reports the overall p-value for each covariate, and not the p-values for (some) pairwise differences. Such a parsimonious use of p-values is usually wise, because the more p-values we present, the bigger becomes the danger of an misinterpretation, as discussed in Section 5.4. The presentation of the first table invites the reader to focus on the small p-values, i.e. invites to a "hunting for p-values" and might mislead the reader with respect to judging our overall evidence for the effect of a covariate, as discussed in Section 8.3. The presentation of the second table guides the reader to focus on the question, whether there is any difference among the categories with respect to the effect on the outcome of interest. And this is typically the first and main question if we work with a categorial covariate. If there is a priori a certain interest in establishing a particular difference between two categories, then one can of course decide to add a p-value for this difference or to provide this p-value in the text of the article.
The reader should also remember our considerations in Section 8.3 that it is hard to deduce an overall p-value from the pairwise p-values, as discussed in the previous section. So one cannot argue that p-values for pairwise differences as shown in the first table can serve as a substitute for the overall p-values in the second table.

– The presentation of the results for the covariate *Region* in the first table are insufficient, because we can only see that there is a difference, but we cannot see, whether the risk is higher in urban or rural areas.
This illustrates that is useful to distinguish between binary covariates (like *Breastfeeding*),

where the name of the variable implies the meaning of the categories 0 and 1, and a dichotomous covariate (like *Region*), where we have just two categories. Since most statistical packages treat the second situation by creating an indicator variable and hence mimicking the situation of a binary variable, it is the responsibility of the user to ensure a correct presentation of the results.
A similar example can be seen by looking on the covariate *Social class*. In the first table, the reader can only guess, whether social class I means low or high social level. The second table adds important information by explicitly labeling the classes. This should remind the reader, that the numbers we attach to categories are arbitrary, and that the user of a statistical package is responsible for attaching a meaning to these numbers.

In summary, the reader should recognise that in presenting results of a regression model in a paper the author has a choice. This choice should be led by the aim to present the results in a fair, useful and understandable manner. The reader should find the results he or she is (or should be) interested in, and nothing more. p-values should be used with care avoiding a misleading focus on selected small p-values.

8.6 Exercise *Physical occupation and back pain*

The dataset `backpain` includes data from an epidemiological cohort study[1] on the occurrence of back pain. You can find information on the age, sex and social class of the subjects, their physical occupation at the begin of the study and whether they suffer from back pain at baseline (variable `b0`) and 5 years later (variable `b5`). Use the `codebook` command to become more familiar with the coding of the variables.

a) What can we conclude about differences between the four types of physical occupation at baseline with respect to the back pain status five years later, if we adjust for age and sex?
b) What can we conclude about the effect of age and sex? Try to express the sex difference as an odds ratio.
c) One of the aims of the study have been to establish that high physical occupation is more dangerous than moderate physical occupation with respect to the development of back pain. What can we conclude about this difference?
d) What happens if we adjust for the social group?
e) Repeat the analysis a) now considering back pain at baseline as the outcome of interest. Can you explain the differences in the results?
f) How would you analyse this data, if the title of the intended paper is *The influence of the type of physical occupation on the development of back pain?*

[1] This study was analysed by Jan Hartvigsen as part of his PhD project. The results are published in ... I am in debt to Jan for his kind permission to use this dataset.

8.7 Exercise *Odds ratios and categorical covariates*

A study similar to that of the last exercise reports the results as odds ratios adjusted for age and sex:

sitting (reference)	
low	0.83
moderate	1.04
heavy	1.93

What is the odds ratio between *heavy* and *low*?