Conditional regression

Contents

1. Introduction

Conditional regression refers to a situation commonly encountered in regression problems and refers to the case when there are observed categorical variables in addition to the *y* and *x* variables used in the regression. These categorical variables often represent some form of group membership such as gender, marital status or the like. In such a case one can consider the regression of n such a case one can consider the regression of *y* on **x** for each category of the categorical variable to investigate the extent to which the regression is the same or different across the levels of the categorical variable.

2. Predicting weight by age

In this example, we consider a simple example where we are interested in the relationship between the gestational age and birthweight of 24 children. Gender is the categorical variable of interest in this example. The data are given in the file **birthweight1.lsf** and was originally published in Dobson and Barnett (2008). The data and syntax files can be found in the **MVABOOK examples\Chapter2** folder.

The variables are

- Sex: gender of child, coded 1 for boys and 2 for girls
- Age35: the gestational age in weeks – 35
- Bweight: birth weight recorded in kilograms.

The age variable is record for purposes of interpretation. By subtracting 35 from the observed value, this enables us to interpret the estimated intercept as the difference in birth weight at 35 weeks instead of at age 0, a value that does not occur for any of the children.

birthweight1.lsf

	SEX	AGE35	BWEIGHT
1	1.00	5.00	2.97
2	1.00	3.00	2.80
3	1.00	5.00	3.16
4	1.00	0.00	2.93
5	1.00	1.00	2.63
6	1.00	2.00	2.85
7	1.00	6.00	3.29
8	1.00	5.00	3.47
9	1.00	2.00	2.63
10	1.00	3.00	3.18
11	1.00	5.00	3.42
12	1.00	3.00	2.98
13	2.00	5.00	3.32
14	2.00	1.00	2.73
15	2.00	5.00	2.94
16	2.00	3.00	2.75
17	2.00	7.00	3.21
18	2.00	4.00	2.82
19	2.00	5.00	3.13
20	2.00	2.00	2.54
21	2.00	1.00	2.41
22	2.00	3.00	2.99
23	2.00	4.00	2.88
24	2.00	5.00	3.23

We now fit a conditional regression to these data. The keyword By is used to instruct PRELIS to do the regression by categories of the categorical variable SEX.

birthweight1.prl

```
System File birthweight1.lsf
Regress BWEIGHT on AGE35 by SEX
```

Results are as follows:

```
Estimated Equations

 For SEX = 1, Sample Size = 12:

  BWEIGHT = 2.651  + 0.112*AGE35 + Error, R² = 0.546
 Standerr  (0.122)  (0.0323)
 t-values   21.669   3.466
 P-values   0.000    0.005

 Error Variance = 0.0404
```

```
For SEX = 2, Sample Size = 12:

  BWEIGHT = 2.422  + 0.130*AGE35 + Error, R² = 0.712
 Standerr  (0.108)  (0.0262)
 t-values   22.374   4.978
 P-values   0.000    0.000

 Error Variance = 0.0249


 The following chi-squares test the hypothesis that all
 regression coefficients are zero except the intercept.
 Variable       -2lnL  Chi-square   df  Covariates
 --------  ----------  ----------   --  ----------
  BWEIGHT      -6.649       9.468    1   AGE35
  BWEIGHT     -12.461      14.958    1   AGE35

Analysis of Variance Table
   Regression d.f.    Residual d.f.           F  Covariates
   ---------------    -------------           -  ----------
        0.485    1       0.404   10      12.011  AGE35
        0.616    1       0.249   10      24.782  AGE35
```

For both gender groups the age effect is highly significant. While the estimated intercept for boys is higher than for girls (2.651 vs 2.422), we see that the estimated increase in weight associated with a week's increase in age is higher for girls than for boys (0.130 vs 0.112). Assuming a linear model to be appropriate, the R^2 s indicate a better fit of the model to the data for boys.

3. Testing equal regressions

The model we fitted can be mathematically expressed as

$$BWEIGHT_{ij} = \alpha_i + \gamma_i AGE35_{ij} + z_{ij},\ i = 1,2$$

with $E\left(z_{ij}^2\right) = \sigma_i^2$. The model has a total of six parameters: α_1, α_2, γ_1, γ_2, σ_1^2 and σ_2^2. To test whether the regressions for the two gender groups are equal, we want to test the hypotheses $\alpha_1 = \alpha_2$, $\gamma_1 = \gamma_2$, $\sigma_1^2 = \sigma_2^2$.

A model with all six parameters of interested can be written as follows

$$BWEIGHT_{ij} = \mu + \gamma_1 d_{ij} + \gamma_2 AGE35_{ij} + \gamma_3 d_{ij} AGE35_{ij} + z_{ij},$$

where the indicator variable $d_{ij} = 1$ if $i = 1$ (boys) and $d_{ij} = 0$ if $i = 2$ (girls). When the two expressions above are considered, we note that $\alpha_2 = \mu$, $\gamma_1 = \gamma_2 + \gamma_3$, $\gamma_2 = \gamma_2$ so that $\alpha_1 = \alpha_2 \Leftrightarrow \gamma_1 = 0$ and $\gamma_1 = \gamma_2 \Leftrightarrow \gamma_3 = 0$. The hypotheses of interest can thus be tested by estimating the second of the two equations.

To do this, we created a new data file **birthweight2.lsf** as shown below. In this file, the variables BOYS and BOYSAGE correspond to the variables d_{ij} and $d_{ij} AGE35_{ij}$ respectively.

birthweight2.lsf

	AGE35	BWEIGHT	BOYS	BOYSAGE
1	5.00	2.97	1.00	5.00
2	3.00	2.80	1.00	3.00
3	5.00	3.16	1.00	5.00
4	0.00	2.93	1.00	0.00
5	1.00	2.63	1.00	1.00
6	2.00	2.85	1.00	2.00
7	6.00	3.29	1.00	6.00
8	5.00	3.47	1.00	5.00
9	2.00	2.63	1.00	2.00
10	3.00	3.18	1.00	3.00
11	5.00	3.42	1.00	5.00
12	3.00	2.98	1.00	3.00
13	5.00	3.32	0.00	0.00
14	1.00	2.73	0.00	0.00
15	5.00	2.94	0.00	0.00
16	3.00	2.75	0.00	0.00
17	7.00	3.21	0.00	0.00
18	4.00	2.82	0.00	0.00
19	5.00	3.13	0.00	0.00
20	2.00	2.54	0.00	0.00
21	1.00	2.41	0.00	0.00
22	3.00	2.99	0.00	0.00
23	4.00	2.88	0.00	0.00
24	5.00	3.23	0.00	0.00

We now regress BWEIGHT on BOYS, AGE35 and BOYSAGE using **birthweight2.prl**.

birthweight2.prl

```
System File birthweight2.lsf
Regress BWEIGHT on BOYS AGE35 BOYSAGE
```

The estimates associated with BOYS and BOYSAGE are not statistically significant. This seems to suggest that the intercept for the two gender groups may be the same. Although there is a one-to-one correspondence between the parameters of the two models, the second model assumes that $\sigma_1^2 = \sigma_2^2$. To test this hypothesis, one can use the previously obtained chi-squares and use this as a chi-square with one degree of freedom. When this is done, a chi-square of 0.657 is obtained, which suggests that the hypothesis of equal error variances may hold as well. The estimate of the common error variance is 0.0326. Our final conclusion is that the regression equation is the same for boys and girls.

```
Estimated Equations

  BWEIGHT = 2.422  + 0.228*BOYS + 0.130*AGE35 - 0.0184*BOYSAGE
 Standerr  (0.124)  (0.166)      (0.0300)      (0.0418)
 t-values   19.537   1.378        4.347        -0.441
 P-values   0.000    0.183        0.000         0.664

             + Error, R² = 0.643


 Error Variance = 0.0326
```

The following chi-squares test the hypothesis that all regression coefficients are zero except the intercept.

Variable	-2lnL	Chi-square	df	Covariates
BWEIGHT	-18.414	24.751	3	BOYS AGE35 BOYSAGE

Analysis of Variance Table

Regression d.f.		Residual d.f.		F	Covariates
1.177	3	0.652	20	12.032	BOYS AGE35 BOYSAGE