Conditional regression

Contents

1. Introduction 1

2. Linear regression of math on reading 1

3. Conditional regression of math on reading by career 4

4. Centering data to facilitate easier interpretation of results 6

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. Linear regression of math on reading

In this example, we consider an example where the categorical variable has several categories with unequal numbers of observations. Verbal skill is often considered to be the core feature of intelligence in studies of cognitive ability. The differences in this skill over individuals are also strongly associated with individual performance in many mental tasks. However, the correlation between verbal skill and performance may also depend on other personal traits.

The file **Math on Reading by Career.lsf** contains three variables. The data and syntax files can be found in the **MVABOOK examples\Chapter2** folder.

Math on Reading by Career.lsf

	Career	Math	Reading
1	1.00	48.71	15.24
2	1.00	43.49	6.33
3	1.00	44.08	15.00
4	1.00	47.50	23.00
5	1.00	63.88	34.67
6	1.00	45.62	15.43
7	1.00	43.77	12.69
8	1.00	49.49	13.20
9	1.00	42.89	13.94
10	1.00	49.69	8.91
11	1.00	42.23	9.19
12	1.00	56.15	17.19
13	2.00	57.87	29.27
14	2.00	47.02	14.00
15	2.00	38.26	7.45
16	2.00	47.26	10.28
17	2.00	48.75	15.63
18	2.00	46.05	25.82
19	2.00	37.50	3.46
20	2.00	38.08	4.41

The variable Career is coded as follows:

- 1 = trades
- 2 = police or security
- 3 = business management
- 4 = sales
- 5 = military service
- 6 = teacher training
- 7 = industrial operations
- 8 = undecided
- 9 = real estate management

We now fit a simple linear regression to these data, using the mathematics score as outcome and the reading score as single predictor. The model can be expressed as

$$Math_{ij} = \alpha_i + \gamma_i Reading_{ij} + z_{ij}, \qquad i = 1, 2, ...9,\ j = 1, 2, ..., n_i,$$

where $Math_{ij}$ and $Reading_{ij}$ are the scores for student j in career group j and z_{ij} is the error in regression. To fit this model using PRELIS, only two lines of syntax are needed.

Math on Reading by Career.prl

```
System file 'Math on Reading by Career.lsf'
Regress Math on Reading
```

Results are as follows:

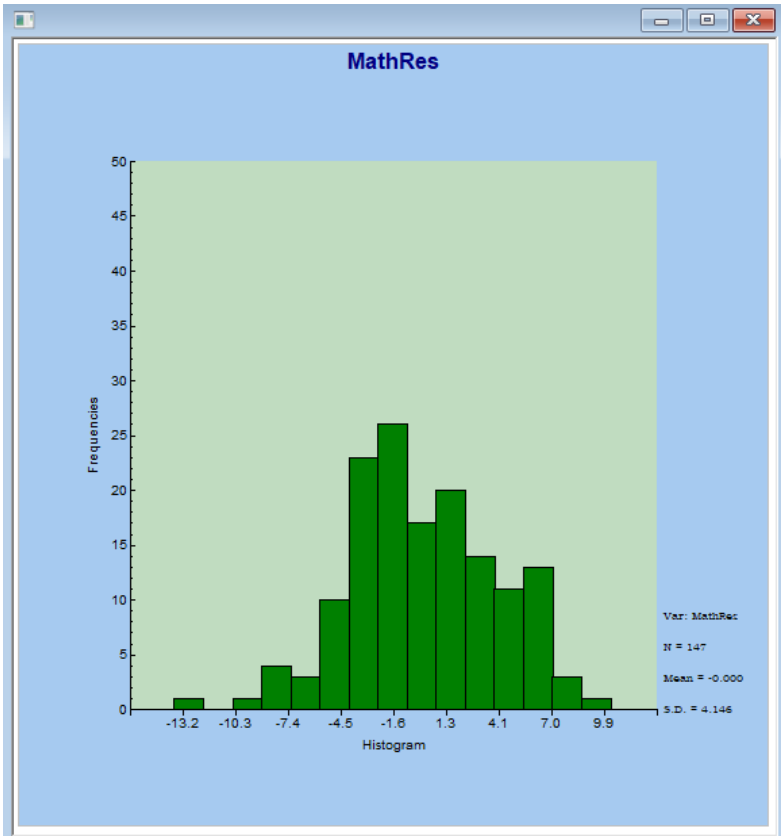
```
Estimated Equations

     Math = 37.969 + 0.640*Reading + Error, R² = 0.693
 Standerr  (0.754)  (0.0354)
 t-values   50.356   18.084
 P-values   0.000    0.000

 Error Variance = 17.311
```

From the results we see that there is a strong relationship between reading and mathematics scores (p = 0.000).

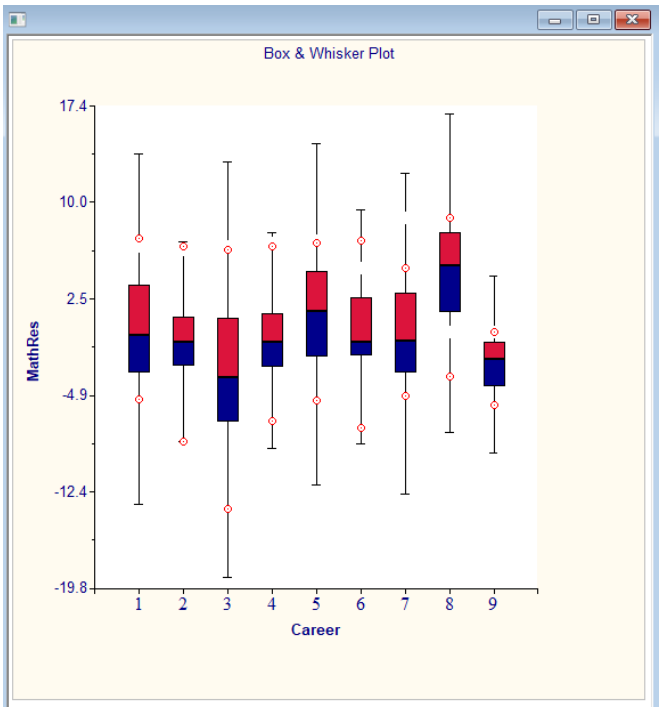
As a first step, we calculate the residuals for this model for different career groups. To do so, we first request the creation and addition of the residuals to the LSF file. This can be accomplished by amending the second line of the PRL file and adding an OU line

```
System file 'Math on Reading by Career.lsf'
Regress Math on Reading Res=MathRes
Output MA = CM RA = 'Math on Reading by Career Res.lsf'
```

A histogram of the residuals obtained for this model is shown below.

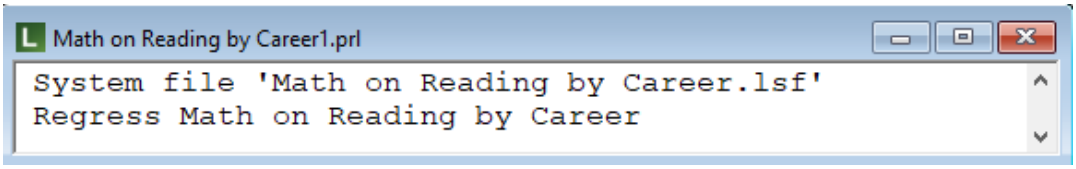


When we make a bivariate plot of the residuals by career choice, we note that career choice may cause differences in the regression of mathematics scores on reading scores. Groups 3 and 8 seem to be the most different, corresponding to business management and “undecided” respectively.



3. Conditional regression of math on reading by career

We now include career choice as predictor in the regression.



Results are shown below. The univariate statistics for career choice show the distribution of respondents over the 9 categories. Business management (category 4) and police or security (category 2) are the most popular choices in this sample.

Career Frequency Percentage Bar Chart

1	12	8.2	▯▯▯▯▯▯▯▯▯▯▯▯
2	23	15.6	▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯
3	19	12.9	▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯
4	24	16.3	▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯
5	22	15.0	▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯
6	9	6.1	▯▯▯▯▯▯▯▯▯▯
7	12	8.2	▯▯▯▯▯▯▯▯▯▯▯▯
8	17	11.6	▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯
9	9	6.1	▯▯▯▯▯▯▯▯▯▯

Estimated Equations

For Career = 1, Sample Size = 12:

Math = 38.033 + 0.655*Reading + Error, R^2 = 0.591

Standerr	(2.924)	(0.172)
t-values	13.008	3.803
P-values	0.000	0.003

Error Variance = 18.104

For Career = 2, Sample Size = 23:

Math = 36.149 + 0.707*Reading + Error, R^2 = 0.864

Standerr	(1.456)	(0.0613)
t-values	24.820	11.541
P-values	0.000	0.000

Error Variance = 11.127

For Career = 3, Sample Size = 19:

Math = 40.512 + 0.383*Reading + Error, R^2 = 0.376

Standerr	(2.811)	(0.120)
t-values	14.411	3.200
P-values	0.000	0.005

Error Variance = 22.362

For Career = 4, Sample Size = 24:

Math = 36.085 + 0.736*Reading + Error, R^2 = 0.855

Standerr	(1.162)	(0.0646)
t-values	31.062	11.402
P-values	0.000	0.000

Error Variance = 10.161

For Career = 5, Sample Size = 22:

Math = 39.527 + 0.630*Reading + Error, R^2 = 0.581

Standerr	(2.201)	(0.120)
t-values	17.961	5.266
P-values	0.000	0.000

Error Variance = 13.065

For Career = 6, Sample Size = 9:

Math = 34.430 + 0.827*Reading + Error, R^2 = 0.821

Standerr	(2.948)	(0.146)
t-values	11.678	5.661
P-values	0.000	0.000

```
Error Variance = 14.716


For Career = 7, Sample Size = 12:

    Math = 37.439 + 0.653*Reading + Error, R² = 0.752
Standerr  (2.910)  (0.119)
t-values   12.864   5.507
P-values   0.000    0.000

Error Variance = 12.673


For Career = 8, Sample Size = 17:

    Math = 41.436 + 0.676*Reading + Error, R² = 0.755
Standerr  (2.169)  (0.0994)
t-values   19.107   6.799
P-values   0.000    0.000

Error Variance = 15.206


For Career = 9, Sample Size = 9:

    Math = 35.292 + 0.647*Reading + Error, R² = 0.935
Standerr  (1.602)  (0.0644)
t-values   22.027   10.053
P-values   0.000    0.000

Error Variance = 4.357
```

The estimated intercept varies between 34.430 for the career choice 6 (teacher training) and 41.436 (the undecided category). The estimated coefficient for reading varies between 0.383 (business management) and 0.736 (sales).

4. Centering data to facilitate easier interpretation of results

In order to make the intercept easier to interpret (as it stands now, it is the expected mathematics score for a student who scored zero on the reading test), we opt to group mean center the reading score by subtracting the mean for each career group from the reading scores of students who chose that career. By doing so, we change the interpretation of the intercept to be the mean mathematics score when reading is at the mean for each career. The file **Math on Reading with groupcentered Reading.lsf** contains the group mean centered data in the variable Read_gc.

Math on Reading with groupcentered Readin...

	Career	Math	Reading	Read_gc
1	1.00	48.71	15.24	-0.16
2	1.00	43.49	6.33	-9.07
3	1.00	44.08	15.00	-0.40
4	1.00	47.50	23.00	7.60
5	1.00	63.88	34.67	19.27
6	1.00	45.62	15.43	0.03
7	1.00	43.77	12.69	-2.71
8	1.00	49.49	13.20	-2.20
9	1.00	42.89	13.94	-1.46
10	1.00	49.69	8.91	-6.49
11	1.00	42.23	9.19	-6.21
12	1.00	56.15	17.19	1.79
13	2.00	57.87	29.27	8.38
14	2.00	47.02	14.00	-6.89
15	2.00	38.26	7.45	-13.44
16	2.00	47.26	10.28	-10.61
17	2.00	48.75	15.63	-5.26
18	2.00	46.05	25.82	4.93
19	2.00	37.50	3.46	-17.43
20	2.00	38.08	4.41	-16.48

We now regress the mathematics scores on the centered reading scores. The model for each career group is now of the form

$$Math_{ij} = \alpha_i + \gamma_i(Reading_{ij} - Reading_{i.}) + z_{ij}, \qquad i = 1,2,...9,\ j = 1,2,...,n_i,$$

Math on Reading by Career2.prl

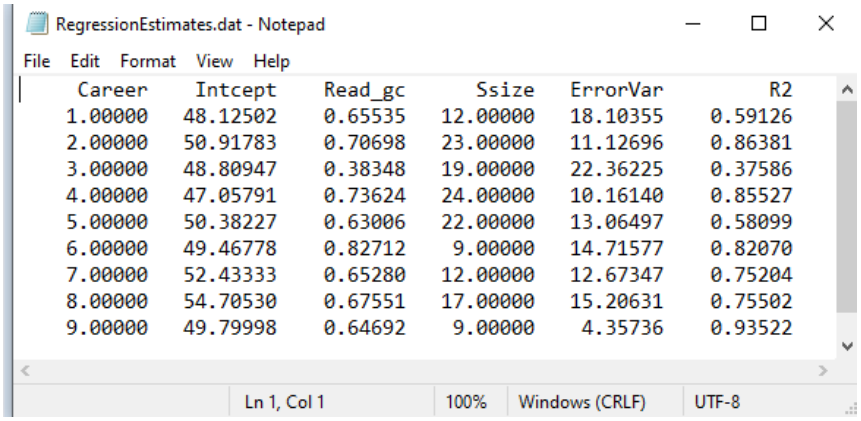
```
System File 'Math on Reading with groupcentered Reading.lsf'
Regress Math on Read_gc by Career
Output YE=RegressionEstimates.dat
```

Results are shown below.

```
Math = 48.125 + 0.655*Read_gc + Error, R² = 0.591
Math = 50.918 + 0.707*Read_gc + Error, R² = 0.864
Math = 48.809 + 0.383*Read_gc + Error, R² = 0.376
Math = 47.058 + 0.736*Read_gc + Error, R² = 0.855
Math = 50.382 + 0.630*Read_gc + Error, R² = 0.581
Math = 49.468 + 0.827*Read_gc + Error, R² = 0.821
Math = 52.433 + 0.653*Read_gc + Error, R² = 0.752
Math = 54.705 + 0.676*Read_gc + Error, R² = 0.755
Math = 49.800 + 0.647*Read_gc + Error, R² = 0.935
```

The estimated regression equations are shown above. Each intercept now represents the expected mathematics score for a student with reading score equal to the mean for the career group he/she is in. The estimated coefficients for Read_gc represent the expected increase in mathematics score associated with an increase of 1 in the centered reading score.

The regression estimates are also summarized in the file RegressionEstimates.dat created during this run.


RegressionEstimates.dat - Notepad
File Edit Format View Help
Career Intcept Read_gc Ssize ErrorVar R2
1.00000 48.12502 0.65535 12.00000 18.10355 0.59126
2.00000 50.91783 0.70698 23.00000 11.12696 0.86381
3.00000 48.80947 0.38348 19.00000 22.36225 0.37586
4.00000 47.05791 0.73624 24.00000 10.16140 0.85527
5.00000 50.38227 0.63006 22.00000 13.06497 0.58099
6.00000 49.46778 0.82712 9.00000 14.71577 0.82070
7.00000 52.43333 0.65280 12.00000 12.67347 0.75204
8.00000 54.70530 0.67551 17.00000 15.20631 0.75502
9.00000 49.79998 0.64692 9.00000 4.35736 0.93522
Ln 1, Col 1
100%
Windows (CRLF)
UTF-8