Exercise, data Mental

Agresti and Finlay report data from a Florida study investigating the relationship between mental health and several explanatory variables using a random sample of 40 individuals (file *mental.txt*). The outcome of interest is an index of mental impairment that incorporates measures of anxiety and depression. Two predictors are considered: a life-event score that combines the number and severity of various stressful life events and an index of socieconomic status (ses).

1. Simple Regressions

1a) Draw a scatterplot matrix with the three variables and comment briefly on the pairwise relationships between the variables. Compute also the pairwise correlations.

1b) Run a simple linear regression of the mental impairment index on the life events index, intrpret the slope and tests its significance using a t-test.

1c) What proportion of the variation across subjects in the index of mental health is explained by the life events index? How is this proportion related to Pearson's correlation coefficient?

1d) Check the linearity of this relationship by adding a quadratic term on the index of life events. In general it is advisable to center variables to their mean before squaring; this reduces collinearity and simplifies interpretation. Either way you should find that we don't really need a quadratic term.

1e) Regress the index of mental impairment on SES. Calculate Pearson's correlation as a summary of the association. Interpretation of the regression coefficient is limited by the arbitrary nature of both scales. Rerun the regression standardizing both indices and interpret the resulting slope. Compare it with Pearson's r.

2. Multiple Regression

2a) Run a regression of the index of mental impairment on both SES and the index of live events. Interpret briefly the estimate of the coefficient of life events, and compare it with the estimate from the simple linear regression.

2b) Interpret the t-test for the net effect of life events after adjusting for SES.

2c) What proportion of the variation in mental health has been 'explained' by the two variables together? What's the square root of this value?

2d) Compute fitted values for this model and calculate Pearson's correlation between observed and fitted values. Does this number look familiar?

2e) What proportion of the variation left unexplained by SES can be attributed to life events? How is this number related to the partial correlation between mental impairment and life events given SES?

3. Added variable plots Interpret the plots given by function *avPlots()* from package car.

Exercise, data wages

The data file *wages.txt* contains an extract from the Current Population Survey to explore, amomg other things, how hourly wages differ among men and women with similar observed characteristics. The file contains information on each worker on years of education, an indicator for southern states, indicator for females, years of work experience, indicator of union membership, the hourly wages (dollars), age, ethnicity, occupation, sector and an indicator for married respondents.

1. Working with wages

1a) Fit a linear model to explore how hourly wages depend on education, work experience, union membership, region, occupation and gender.

1b) Describe the net effects of education, work experience and union membership on wages.

1c) Describe the gender gap after adjusting for the effects of the other variables in the model and test its significance.

1d) Plot the residuals vs. the fitted values. Any outliers?

2. Working with log(wages)

2a) Fit the same multiple linear model as before but working with $Y = log(wages)$.

2b) Describe the coefficients of education, work experience and union membership in terms of the effects of these variables on wages (not on log wages).

2c) Check whether the returns to work experience are the same for males and females.

3. Regression diagnostics

3a) Plot the residuals. Any outliers?

3b) Check the normality of the residuals.

3c) Did we need to transform the data? Was the logarithm a good transformation? Explore these questions with de Box-Cox transformation.

A note on the log transformation $\lambda = 0$ in the Box-Cox transformation leads to a logarithmic transformation of the data. This transformation may be useful when our data shows a distribution which is skewed to the left (long right tail) and we are interested in restoring symmetry in its distribution. This transformation may also be useful to linearize some relationship that don's seem to be linear at first (modelling relationships).

Monetary amounts (incomes, wages, amount of purchases), variables describing sizes (cities, corporations, electricity usage) recover symmetry when this transformation is applied. It may also help to log transform data with values that range over several orders of magnitude. This is because modelling techniques often have difficulty with very wide data ranges and, such data often comes from multiplicative processes, so log units seem more natural. For instance, if management gives everyone in the department a raise, this is not probably going to be the same amount to everybody but a fixed percentage raise: how much extra money I end up with depends on my initial salary. This is a multiplicative process in which log transforming can make modelling easier.

Log-linear model: $logY_i = \beta_0 + \beta_1 X_i + \epsilon_i$ In the log-linear model, the literal interpretation of the estimated coefficient $\hat{\beta}$ is that a one-unit increase in X will produce and expected increase in $log Y$ of $\hat{\beta}$ units. In terms of Y intself, this means that the expected value of Y is multiplied by $e^{\hat{\beta}}$. So, in terms of effects of changes in X on Y (unlogged):

- Each 1-unit increase in X multiplies the expected value of Y by $e^{\hat{\beta}}$.
- To compute the effects on Y of another change in X than an increase of 1 unit, call this change c, we need to include this c in the exponent, $e^{c\hat{\beta}}$.
- For small values of $\hat{\beta}$, $e^{\hat{\beta}} \approx 1 + \hat{\beta}$. We can use this for a quick interpretation of the coefficients: $100 \cdot \hat{\beta}$ is the expected percentage change in Y for a unit increase in X. For instance, $\hat{\beta} = 0.06$, $e^{0.06} \approx 1.06$, so a 1-unit change in X leads to (approximately) and expected increase in Y of $6\,\%$.