Analysis of Continuous Data

Homework 3

12/11/2015

This homework consists of a SAS exercise, and a theoretical exercise. For the SAS exercise, it is again important that you write your results in the requested report format. If you wish, you may back up the answers to the 'theoretical' questions with simulation results that you summarise. You are however expected to find general results (simulations are not necessary).

Note that this homework should be made individually, even if some consultation along the road with staff and students is allowed, your answers should be personal and written in your own words. They are due on the minerva dropbox by noon on **November 22 before Midnight**. It should be sent through that route to Els Goetghebeur and Emmanuel Abatih. The marks that you will receive for this homework contribute to your final score for Analysis of Continuous Data.

1 A SAS Exercise

Cystic fibrosis (CF) is the most common, classic mendelian autosomal recessive, life-limiting disease among the white population. It is a multisystem disease that results from loss of function in the CF transmembrane conductance regulator (CFTR) gene, classically leading to respiratory tract, gastrointestinal (GI), pancreatic, and reproductive abnormalities. In one study, O'Neill[1] and colleagues measured the Maximal static expiratory pressures

[1] O'Neill S, Leahy F, Pasterkamp H, Tal A. The effects of chronic hyperinflation, nutritional status, and posture on respiratory muscle strength in cystic fibrosis.Am Rev Respir Dis. 1983 Dec;128(6):1051-4.

(PEmax) of 25 patients with CF together with other important covariates. Their goal was to determine factors that influence the PEmax.

The aim of this homework is to build a predictive model for PEmax using the weight (in kg) of the patients based on the data in the attached excel file called "cystfibr.xls". One way to do this is by fitting a simple linear regression model of the form: $Y_i = \alpha + \beta \cdot X_i + \epsilon_i$ satisfying all the usual assumptions. Use the resulting model to answer the following questions:

1. Give the parameter estimates, their standard deviations and their 95% confidence intervals.
2. Give a clear and useful interpretation of the estimated regression coefficient β.
3. Calculate a 95% confidence interval of the regression coefficient β and interpret the interval.
4. Perform a two-sided statistical hypothesis test to test the hypothesis that the regression coefficient β is zero (use significance level 0.05).
5. Assess the assumptions underlying the linear regression model (scope of the model, study of outliers and residuals, linearity of the curve, constancy of the variance, lack-of-fit, ...), and give a detailed discussion on the model quality.
6. Write an executive summary containing your main conclusions from your statistical analysis (max. 1/2 page).
7. **Bonus Question**: In addition to the weight of the patients, their heights (in cm) were also recorded. Is the interaction between weight and height statistically significant at the 5% level? If so interpret this interaction.

2 Theoretical Part

In a clinical study, physicians are interested in the effect of a new treatment on the blood pressure. In a random sample from the population of patients for whom the treatment was prescribed, they observe patients who use the new treatment and patients who do not. On the other hand, a certain gene is known to have an extreme influence on the blood pressure. In this exercise, we aim to find out what happens when the treatment effect is estimated with and without accounting for the gene effect.

Suppose that the blood pressure is modelled correctly by the following population model:

$$Y_i \quad = \quad \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot x_{2i} + \beta_3 x_{1i} \cdot x_{2i} + \varepsilon_i \tag{1}$$

where $\varepsilon_i \sim N(0, \sigma^2)$ $(i = 1, \ldots, n)$, Y_i is the blood pressure for individual i, x_{1i} wether person i uses the new treatment $(x_{1i} = 1)$ or not $(x_{1i} = 0)$ and $x_{2i} = 1 (x_{2i} = 0)$ indicates the presence (absence) of the gene. Assume that there are no further confounders of the effect of x_{1i} on Y_i after adjusting for x_{2i} ; and in the random sample from the population we have:

$$P(X_{2i} = 1 | X_{1i} = 0) = q_0$$
$$P(X_{2i} = 1 | X_{1i} = 1) = q_1$$

and 70% of these patients are taking the treatment.

1. Model (1) represents the true data-generating model, but this is of course unknown to the statistician. Suppose we analyze the data with the model

$$Y_i \quad = \quad \widetilde{\beta}_0 + \widetilde{\beta}_1 \cdot x_{1i} + \widetilde{\beta}_2 \cdot x_{2i} + \widetilde{\varepsilon}_i \tag{2}$$

where $\widetilde{\varepsilon}_i \sim N(0, \widetilde{\sigma}^2)$ $(i = 1, \ldots, n)$.

(a) Using the true data-generating model (1), write down the average causal effect of the treatment on the blood pressure for fixed levels of x_2, separately for patients with the genotype and for patients without the genotype.

(b) Does estimating the effect of x_1 on Y given x_2 based on model (2) result in unbiased estimates of these causal effects ? Explain.

(c) How does the variance (and hence precision) compare. Can you say something about the mean squared errors?

2. Suppose now that we ignore the ”confounding” effect of the gene and analyse the data with the simple model

$$Y_i \quad = \quad \beta_0^* + \beta_1^* \cdot x_{1i} + \varepsilon_i^* \qquad (3)$$

where $\varepsilon_i^* \sim N(0, \sigma^{*2})$ $(i = 1, \ldots, n)$.

   (a) Write the parameter β_1^* in function of the parameters of Model (model1), and interpret this parameter.

   (b) Does estimating β_1^* result in an unbiased estimator of the (unconditional) average causal effect, assuming a randomized design ? Explain

   (c) How does the variance (and hence the precision) of $\hat{\beta}_1^*$ compare with that of $\hat{\beta}_1$ from model (1) ?

3. What happens if $\beta_3 = 0$ in Model (1)? Give a discussion for both Models (2) and (3).