Data Analysis in the Health Sciences

Final Exam 2011 – Solution

Question 1

(a) From the R output, intercept is 0.51 and slope is -1.48×10^{-6}. 95% confidence intervals can be computed as

$$Estimate \pm 1.96 se(estimate)$$

A 95% CI for the intercept is $0.51 \pm 1.96 \times 0.00245 = (0.505, 0.514)$.
Similarly, a 95% CI for the slope parameter is $-1.48 \times 10^{-6} \pm 1.96 \times 0.0000476$ =(-9.47e-05, 9.18e-05).

(b) The confidence interval for the slope parameters includes zero and the end points of the interval are not of clinical interest as well. We conclude that the pollutant has no effect on the proportion of male births.

Question 2

(a) Using the BIC criterion, we would select the model including License+Network+Gas.

(b) Using the AIC criterion, we would select the model including License+Network+Gas+Gender.

(c) Using the adjusted R squared criterion, we would select the model including License+Network+Gas+Gender.

(d) The R^2 value represents the proportion of variance in the outcome explained by the variables in the model. The more variables we add, the more variance we explain (adding a variable in the model can't reduce the variance explained: if it has no effect on the outcome, the variance explained will be unchanged but will never decrease). This is why the R^2 value always increases when more variables are added to a model.

Question 3

(a) The estimated intercept is -0.82. This means that the probability of infection when the c-section is not planned, when the mother has no risk factors and when antibiotics were not given prior to the c-section is $e^{-0.82}/(1+e^{-0.82}) = 0.305$.

(b) Odds ratios are equal to the exponential of the coefficients.

- Planned: $OR = e^{-1.072} = 0.34$. The OR for infection comparing women who had a planed c-section with women who didn't (baseline) is 0.34.
- Antibio: $OR = e^{-3.25} = 0.038$. The OR for infection comparing women who received antibiotics prior to the c-section with women who didn't (baseline) is 0.038.
- Risk: $OR = e^{2.029} = 7.61$. The OR for infection comparing women who are at risk with women who aren't (baseline) is 7.61.

(c)

- OR for planned among women who are not at risk is $e^{15.17}$=3,874,782
- OR for planned among women who are at risk is $e^{15.17-17.027}$=0.15
- OR for risk among women whose c-section wasn't planned is $e^{18.38}$=96,013,561
- OR for risk among women whose c-section was planed is $e^{18.38-17.027}$=3.86

Question 4

(a) In the logistic regression model,

$$P(\text{Myocardial infarction among non-smokers}) = \frac{e^{\beta_0}}{1+e^{\beta_0}},$$

or in an equivalent way:

$$logit(P(\text{Myocardial infarction among non-smokers})) = \beta_0.$$

From the table, we can compute P(Myocardial infarction among non-smokers)=90/436=0.2064. Then, $\beta_0 = \ln(0.206/0.794) = -1.34$.

(b) The coefficient β_1 of the logistic regression model represents the log OR of the 2 by 2 table. From the table, we can compute OR=$172 \times 346/(173 \times$

90)=3.82. Then, $\beta_1 = \ln(3.82) = 1.34$.

Question 5

From the results, it seems that energy, height, weight and fat are not associated with the success of follow up in this study. The number of years at risk seems to have a positive impact on follow up as the OR confidence interval for a 10 years difference in the number of years at risk is (1.21,3.39). Furthermore, subjects who had a CHD event at baseline tend to not be followed up with an OR of (0.17,0.64), which is not surprising since a number of these subjects may have died since data was first collected. In light of these results, data doesn't seem to be missing completely at random.

Question 6

(a) Overall, the fit of the model seems reasonable.

(b) Within this group of 25 subjects, the observed proportion of outcomes (y=1) is 0.4, which corresponds to a total of 10 subjects. In this group, the number of subjects with the outcome (y=1) follows a Binomial distribution $\mathcal{B}(n = 25, p)$, where p can be estimated by the observed value $\hat{p} = 0.4$. As a result, the variance of the number of subjects with the outcome in this group is $n\hat{p}(1 - \hat{p}) = 25 \times 0.4 \times 0.6 = 6$ subjects, which corresponds to a standard deviation of 2.44 subjects. The 95% confidence interval for the number of subjects with the outcome in this group is $10 \pm 1.96 \times 2.45 = (5.2, 14.8)$ subjects. Translating these number into proportions (i.e. dividing by 25), the 95% CI for the proportion of subjects with the outcome is (0.20,0.58). This means that the observed proportion 0.4 is within the bounds of chance.

Question 7

(a) The stroke rates (on the logic scale) are measured by the "alpha" coefficients in the model and the rate differences between populations are measured by the "alpha12", "alpha13",... coefficients. As one can see, the confidence intervals of the differences between populations all include zero. However, we note that there are some differences over to 0.3 (on the logic scale), which might be important. Results are inconclusive.

(b) The effectiveness of the medication in the five regions is measured by the "beta" coefficients and the differences in effectiveness between regions is

measured by the "beta12", "beta13",... coefficients. As in (a), the confidence intervals of the differences between populations all include zero. However, we see even larger differences compared to question (a) so results are also inconclusive.

Question 8

Study from researcher A: an advantage of this study is that patients from the hypertension clinic (i.e. patients at risk of CHD events) are included in the sample, which will probably lead to a high number of cases (CHD events) during the 1 year follow-up and therefore to better estimates of the effect of blood pressure (BP) on CHD events. However, the conclusions of this study will be applicable only within the "at-risk" population. Furthermore, measuring BP only once on subjects is a limitation, as measurement error will almost certainly occur.

Study from researcher B: Measuring BP at three different times is certainly an advantage, as measurement error will be less of an issue compared to researcher A. This researcher includes more patients in his sample. However, since the sample is drawn from a pool of normal blood pressure subjects, this is supposed to compensate for the small number of CHD events that will be observed after the 1 year follow up. It is not clear if 200 subject will be sufficient to observe enough CHD events in order to obtain reliable estimates of BP effects.