EPIB-621 Solutions 2008 Final Exam

1. (a) The null and alternative hypotheses are:

H_0: No association between age and cancer
H_A: There is an association between age and cancer

Taking observed data minus expected data, squaring each term and summing up, we find $\chi^2 = 83.33$ with one degree of freedom, so that $p \approx 0$. Or can do by the Z-test, with $Z = 9.12$ and same $p \approx 0$ from normal tables. So, we can reject the null hypothesis, and conclude that there is strong evidence of an association between age and cancer.

(b) Plugging into the formula from the course notes for confidence intervals for the difference between two proportions, we find a 95% confidence interval for $p_1 - p_2$ is (-0.19, -0.12). Thus, there is a strong and clinically important association between age and cancer. Note that this is much more information that just knowing the p-value from part (a).

2. (a) Using the simple formula for updating beta densities, in other words, simply adding the observed numbers of successes and failures to the prior α and β prior parameters for each surgeon, we find that the posterior distributions are beta(95, 55) for Surgeon 1, and beta(75, 75) for surgeon 2.

(b) Surgeon 1 was closer, as his prior mean was 70%, versus 30% for Surgeon 2, while the observed rate was 60%.

3. (a) The slope implies that BMI increases by 0.1 units for each year of increase in age, on average in the population from which this sample was taken.

(b) Yes, it is possible. We know that

$$\overline{Y} = \alpha + \beta \times \overline{X}$$

since the regression line always passes through the means of X and Y. We know that $\overline{X} = 50$, $\beta = 0.1$, and $\overline{Y} = 25$. Plugging in these values to the above equation and solving for α gives $\alpha = 20$.

4. (a) Looking at the outputs, we find:

$$OR = \exp(0.38211) = 1.465$$

$$lower\ CI\ limit = \exp(0.38211 - 1.96 \times 0.164) = 1.06$$

$$upper\ CI\ limit = \exp(0.38211 + 1.96 \times 0.164) = 2.02$$

(b) Plugging into the inverse logit function, we have:

$$\frac{\exp\left(12.567 + 25 \times 0.38 - 2003 \times 0.007235 - 0.771\right)}{1 + \exp\left(12.567 + 25 \times 0.38 - 2003 \times 0.007235 - 0.771\right)} = 0.1492$$

So the probability is about 15%.

(c) The intercept provides the probability of an event on the logit scale, when all variables are set equal to zero. Here that implies someone born in the year zero, which is not only very far from the range of the data set, but also very far from the year 2000, for example, which explains the very large intercept. In other words, the intercept needs to be a large value to counteract the the fact that such large values are plugged into the equation for year.

5. (a) Model #7 gives exactly the same values as in Question #4.

(b) There is not much evidence for confounding, the coefficients hardly change from model to model. Any confounding can be expected to be rather small in magnitude.

(c) See question #4, answer is almost exactly the same since there was no evidence of confounding.

6. (a) For active subjects, we have:

$$\frac{\exp\left(-2.5\right)}{1 + \exp\left(-2.5\right)} = 0.076$$

For non-active subjects, we have:

$$\frac{\exp\left(-2.2\right)}{1 + \exp\left(-2.2\right)} = 0.10$$

Since the population contains a 50% - 50% mix of active and non-active subjects, we have overall $\frac{0.10+0.076}{2} = 8.8\%$ would have diabetes.

7. (a) From the table, we find OR = 1.37, 95% CI = (0.85, 2.21). This is a wide interval and includes the null value of 1 as well as values that would be clinically interesting, so it is inconclusive.

(b) For the most part less likely, as evidence leans that way in most results throughout the table, but evidence is less strong for the most severe disease.

(c) In general, adjusting for covariates resulted in a slightly decreased effect compared to the unadjusted results.

8. (a) Looking at the results for mu in the table, we can calculate

$$\frac{\exp\left(1.513\right)}{1 + \exp\left(1.513\right)} = 0.8195$$

(b) Looking at the variable called p.r1.r2, we conclude that region 1 has higher values compared to region 2 with probability close to 1, i.e., virtual certainty.