EPIB-621 Solutions 2019 Midterm Exam

1. (a) We plug the data into usual the formula for confidence intervals for binomial proportions, that is:

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n} = 60/200 \pm 1.96 \times \sqrt{(60/200 \times ((1-60)/200)/200)}$$

This gives a CI of (0.236, 0.364).

(b) A reasonable choice of low information prior density would be beta(1,1), which is equivalent to a uniform density. Other choices with low values of beta density parameters is also reasonable, such as beta(0.5, 0.5), or beta(0.1, 0.1), etc.

(c) Combining the data with the prior of beta(1,1) gives a posterior density of beta(1+60,200-60+1) = beta(61, 141)

(d) We need to plug into the formula that calculates beta density means and SDs from the α and β parameters. Thus we calculate:

$$\mu = \frac{\alpha}{\alpha+\beta} = 61/(61+141) = 61/202 = 0.302$$

and

$$\sigma = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} = \sqrt{\frac{61 \times 141}{(61+141)^2(61+141+1)}} = 0.032$$

2. (a) The intercept α provides the outcome when the independent variable X is zero. Here that means we expect a response time of five minutes for someone living zero km from the city center, that is, for someone living in the city center itself.

(b) Plugging into the regression formula, we calculate:

$$Y = \alpha + \beta_1 X + \beta_2 X^2 = 5 + 3 \times 10 - 0.05 * 10^2 = 5 + 30 - 5 = 30.$$

Thus it would take 30 minutes to respond, on average.

(c) The reviewer is forgetting that any regression equation applies only over the range of the data, which here is 30 km. Over this range, there are no negative values predicted for response times. The fact that negative response times are predicted for values outside this range, in fact over twice the range, does not invalidate the model over its intended range of operation. Thus the reviewer is not correct.

3. (a) There are 496 degrees of freedom listed in the R output, and 4 parameters estimated in the model (three beta coefficients and one intercept), so the sample size must have been $496 + 4 = 500$.

(b) Looking at the table of coefficient estimates, we see the estimated coefficient for age is 0.46982, with standard error of 0.02060. Therefore a 95% CI is calculated as

$$0.46982 \pm 1.96 \times 0.02060 = (0.429444, 0.510196) \approx (0.43, 0.51)$$

Since there is an interaction term in the model, we need to be careful about the interpretation: Only for those subjects who are not taking the drug, we estimate that for each one year increase in age, the outcome Y increases on average by an estimated 0.47, with the true increase 95% likely to be at least 0.43, but lower than 0.51. While this is a small increase per year, for changes in age of approximately 6 or 7 years, the average change is clinically important, with older subjects tending to do better than younger subjects.

(c) Since both subjects are on the drug the main effect of the drug will "cancel out" when calculating the difference, but we need to consider the interaction term between drug and age. Thus we add the two coefficients to find the overall effect per year, and multiply by the 20 year difference:

$$20 \times (0.46982 + 0.06204) = 10.6372 \approx 10.6$$

Thus, on average, we expect a 10.6 difference in the outcome Y comparing two subjects aged 20 years apart, both taking the drug.

4. (a) Looking at the large changes in coefficients between models across all three variables, we can see there is strong confounding between all three variables, age, weight and sex. For example, comparing model 1 to model 3, we see that when sex is added to a model already containing age and weight, the age and weight coefficients substantially decrease (e.g., age changes from -0.011878870 to -0.006495951, losing almost half of its value compared to when sex is not in the model, etc.). We also note that the SEs change in a similar fashion.

(b) In model 1 the beta coefficient for age is -0.011878870, and the SE is 0.0020023123. The CI can thus be calculated as:

$$-0.011878870 \pm 1.96 \times 0.0020023123 = (-0.0158034, -0.007954338) \approx (-0.016, -0.008)$$

Therefore, for each increase of one year in age, BMD decreases by about 0.012, with 95% CI indicating 95% certainty that the true value is somewhere between a 0.008 and a 0.016 decrease. This is highly clinically important, given that even using the lower limit of the CI of 0.008, there is a clinically important decrease in BMD every two to three years, on average.

5. (a) This is false. The p-value indicates that we have not seen a particularly rare event if the null hypothesis is exactly correct. However, it is possible that the sample size from the study is small, and the confidence interval would not rule out potentially important effects. In other words, despite the large p-value, it could be that the study is inconclusive.

(b) This is false, because one can add polynomial terms or transform either the outcome or independent variable to be able to examine non-linear effects within the context of a linear regression model.