EPIB-621 Solutions 2007 Midterm Exam

1. (a) Intercept is -19,640,000, or about -20 million vehicles sold in the year 0. This obviously has no substantive interpretation for this problem, but the intercept is still important in correctly positioning the regression line.

(b) Slope is 9,936 vehicles per year, meaning that on average, between the years 1997 and 2006, about 10,000 more SUVs were sold each year, compared to the previous year.

(c) Using the approximate limits from the normal distribution, we calculate:

$$9935 \pm 1.96 \times 274.4 \rightarrow (9398.2, 10473.8)$$

A more accurate answer would have used the t distribution with 8 degrees of freedom, substituting 2.306 rather than using the approximate 1.96.

(d) Yes, all assumptions seem satisfied. Model seems very close to linear, and there is no evidence of non-normal residuals (but hard to tell with just 10 points), and variance seems constant throughout the range.

2. (a) Which variables are included in the best model? Which variables are included in the second best model?

Best model includes x_1 and x_2 (posterior probability = 0.809), while second best includes x_1, x_2, and x_3 (posterior probability = 0.191). All other models were much worse, these two probabilities add up to almost the total probability mass of one.

(b) The Bayesian Model Average provides the optimal model for future predictions. One could obtain this either directly from the output (column marked "EV", which stands for expected value, or average), or one can calculate it directly as follows:

$$\begin{aligned}
\alpha^* &= .809 \times (-0.05266) + .191 \times (-0.4973) = -0.05210 \\
\beta_1^* &= .809 \times (1.129) + .191 \times (1.2537) = 1.15284 \\
\beta_2^* &= .809 \times (0.97082) + .191 \times (1.24299) = 1.02283 \\
\beta_3^* &= .809 \times (0) + .191 \times (-0.13272) = -0.02536
\end{aligned}$$

(c) Evidence for confounding occurs when beta coefficients change as other variables enter or exit a model, and/or when the se estimates become inflated when new variables enter the model. We have strong evidence here for confounding of x_3 with both x_1 and x_2, but x_1 and x_2 do not seem to confound each other.

Compare models [4] and [7] with model [1], and notice that both point estimates remain virtually identical for both x_1 and x_2, while se's considerably decrease when both variables are in the model together. This shows that x_1 and x_2 are both useful predictors, and do not confound each other.

On the other hand, compare models [4] and [7] with models [3] and [6], respectively. In both cases, point estimates for x_1 and x_2 change substantially, while se's are either stable or increase. Finally, compare models [1] and [2], and note that the se's considerably increase (almost double) when x_3 is added to the model.

3. Slope is 0.088, with CI (0.033, 0.145). This shows a positive effect, with even the lower limit well above 0. There is likely a clinically important effect, although to really be able to say this, we would need to know more about the particular scales being used.

(b) $R^2 = 0.2798$ says that about 28% of the total variability in logFAT is explained by the model being used.

(c) Plugging into the linear model, we get

$$exp(logFAT) = exp(1.86375 + 0.02796 \times 40 + 0.0880 \times 5) = 30.76$$

(d) No, while zero is part of the CI, it is very wide about zero, so any single point within this interval is very unlikely to be the exact correct value. So despite the null hypothesis of $\alpha = 0$ not being rejected, the chances of $\alpha = 0$ exactly is minuscule.

4. The p-value would be lower for a larger sample size. The p-value is created by a statistic of the form

$$t = \frac{x}{s/\sqrt{n}} = \frac{x/\sqrt{n}}{s}$$

which gets larger as n increases, which leads to a smaller p.

5. (a) Intercept of 110 provides the QoL value for females at age 0, not meaningful here as data range was 20 to 75, but needed to position the line for females.

Slope for age in females is given by the 1.1 term in front of age, showing that females lose 1.1 points on the QoL scale, on average, per year, over the range from 20 to 75 years old.

The coefficient of 5 for the sex term in the equation shows that the intercept for males is five points higher than that for females, i.e., intercept for males is 115, again, as in females, not directly meaningful.

The interaction term of -0.3 shows that the estimated age slope for males is -1.4, i.e., 0.3 points lower than for females, so that males lose 1.4 points

per year that they age, over the range from 20 to 75 years old. This same coefficient also shows that the average difference between males and females, which starts at five points at age zero, decreases by 0.3 points per year.

(b) Directly plugging into the equation, we get 73 points on the QoL scale.

(c)