EPIB-621 Solutions 2011 Midterm Exam

1. (a) This answer will of course vary from researcher to researcher, here is one example: I do not have much idea of this proportion, but doubt it is less than 5%, or higher than 25%, and my best guess would be 15%. So, I will look for a beta density with a mean of 15%, and an SD of 5%, so that, approximately, 95% of my prior density will cover the range from5% to 25%. Plugging $\mu = 0.15$ and $\sigma = 0.05$ into the equations that convert means and SDs into beta coefficients

$$\alpha = -\frac{\mu\ (\sigma^2 + \mu^2 - \mu)}{\sigma^2}$$

and

$$\beta = \frac{(\mu - 1)\,(\sigma^2 + \mu^2 - \mu)}{\sigma^2}$$

one finds $\alpha = 7.5$ and $\beta = 42.5$.

(b) One simply needs to add the number of successes $x = 196$ to α and the number of failures $(n - x) = 804$ to β from part (a), find find the posterior is beta(203.5, 846.5).

(c) Plugging into the formulae for beta density means and variances, we find:

$$\mu = \frac{\alpha}{\alpha + \beta} = \frac{203.5}{203.5 + 846.5} = 0.194$$

and

$$\sigma = \sqrt{\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}} = 0.01219$$

2. (a) FEV seems strongly correlated with Age and with Height (with maybe a quadratic relationship with Height ?). From the boxplots, we can see that FEV seems higher in smokers than in non-smokers.

(b) Age, smoke and Height seem to be correlated. There is potential confounding for these 3 variables.

(c) We can see that all the coefficients in the multiple model change dramatically compared to the simple models (with no overlap in the confidence intervals), which suggest high level of confounding and maybe near collinearity. It is then very difficult to accurately estimate the effects of smoking on lung function without a much higher sample size. Then, I would not use the multiple regression model. I would probably do a stratified analysis in this case.

3. (a) When the age increases by 1 year, the fat percentage increases by 0.5%.
The difference between females and males in fat percentage is 16% (females have 16% more fat than males).

(b) The average fat percentage in a 50 year old male is $20.11 + 0.24 \times 50 - 29.26 + 0.57 \times 50 = 31.35\%$.
The average fat percentage in a 50 year old female is $20.11+0.24\times 50 = 32.11\%$.
Then, the difference in fat percentage between a 50 year old male and a 50 year old female is $31.35 - 32.11 = -0.76\%$.

(c) It is the increase in percentage fat when the age increase by one year in females.

4. (a) We use the model average, giving

$$y = -0.3546 + 2.8882 \times x - 0.4151 \times w + 2.0920 \times z$$

(b) Plugging in to answer in (a), we get:

$$y = -0.3546 + 2.8882 \times 1 - 0.4151 \times 1 + 2.0920 \times (-1) = 0.0265$$

(c) Using the $ols and the $se values from model 2, we calculate

$$1.058403 \pm 1.96 * 0.5541594 = (-2.14455542,\ 0.02774942)$$

or approximately (-2.14, 0.028).

5. (a) False. For example, one could have residuals that are mostly very far from the line (poor fit) but follow a normal distribution.

(b) False. The average BP difference between a 20-30 year old individual who receives treatment A and a 20-30 year old individual who receives treatment B is equal to $\beta_{Age20-30} + \beta_{Age20-30:TreatmentA}$