Homwork 5

Tyler Poelking

3/6/2017

R Markdown

Question 1:

a) One of the included covariates in this model is whether or not the person is a male or a female. This covariate strongly influences who is able to get breast cancer and who is able to get prostate cancer. People with breast cancer are practically always girls and people with prostate cancer, of course, are always boys. This phenomenon can be clearly observed in the data, too, since no Female has prostate cancer and no male has breast cancer. Because of this, if the person has one of these cancers, say, breast cancer for example, the fact that it is a female is a given, so using gender as a covariate in the model would be meaningless. In this model, it would be difficult to model these counts of 0, since the P_{ijk} will equal 0 and it is not possible to take the logit of 0, since it is log(p/(1-p)) and NaN.

b) Assume that $\{Y_{ijk}\}$ is a set of independent Poisson(U_{ijk}) random variables where $P_{ijk} = (u_{ijk}/n)$ is modeled by log $p_{ijk} = (x\hat{}t)_{ijk}$ * B. In this model, i = 1,2,3 refers to cancer type, where 1 = Colorectal, 2 = Pancreas, and 3 = Lung. j = 1,2,3 refers to region, where 1 = Newfoundland, 2 = Ontario, and 3 = Quebec. Lastly, k = 1,2 refers to gender, where 1 = Female and 2 = Male.

Model 2 is equivalent to $\log(P_{ijk}) = \log(n) + B_0 + B_1 CL_p + B_2 CP_p + B_3 RO_p + B_4 RQ_p + B_5 GM_p + B_6 CLiGm_p + B_7 CPiGM_p + B_8 ROiGM_p + B_9 RQiGM_p$. Here, Cancer type Colorectal, Region Newfoundland, and Gender Female are the baselines. That is, they are accounted for in the intercept coefficient. $CL_p = 1$ when cancer type is lung cancer, $CP_p = 1$ when cancer type is Pancreas. They both = 0 when person p has cancer type of colorectal. $RO_i = 1$ when the region for person p is Ontario, $RQ_p = 1$ when the region for person p is Quebec. They both = 0 when the regon for person p is Newfoundland. The coefficients in this model represent the estimated log odds change in cancer death rate between the baseline values and whichever values the coefficents represent. For example, B_1 represents the estimated log odds change in a person with lung cancer from a female with colorectal cancer in Newfoundland.

This model is different from Model 1 because it includes interaction effects between Region and Gender as well as between Cancer Type and Gender. Model one only includes the main effects of the factor variables. In other words, Model two takes into consideration the cancer type in combination with the gender. It also takes into consideration the region in combination with the gender. Model 1 accounted for these only seperately.

c) Fixing region and gender, we estimate that the probability of a cancer patient having lung cancer is eˆB1 = eˆ0.95628= 2.601999 times more likely than the cancer patient having colorectal cancer. A 95% confidence interval for is [eˆL, eˆU], where L = 0.95268 - 1.96*0.01752 and U = 0.95268 + 1.96*0.01752. Taking e to the U and e to the L, we get[eˆL, eˆU] = [2.505131, 2.683224]. The other coefficients can be interpreted in a similar manner. For example, we estimate that the probability of a cancer patient having pancreas cancer is eˆB2 = eˆ-0.82573= 0.4379152 times more likely than the cancer patient having colorectal cancer. B3 and B4 are similar in that, fixing Cancer Type and Gender, eˆB3 or eˆB4 equal the estimated multiplicative change in the cancer rate when going from the Newfoundland region to their corresponding regions (B3 for Ontario. B4 for Quebec). Lastly fixing Cancer Type and Region, eˆB5 equals the estimated multiplicative change in the cancer rate when going from the a femail to a male.

Based off the coefficient summary table and it's Pr(>|z|) values, every covariate adds significantly to the explaination of the log odds of death rate. This is also true when calculating p-vals for each covariate in the analysis of Deviance Table on page 4 of the homework (using 1-pchisq(x=ChangeDevianceResids, df=change in df)) for each iteration of the table. In both cases, all p-values were near zero. This gives evidence that Model 1 is a decent model to use. Model 1 supports the notion that the log odds rate of cancer is significantly

different between patients with colorectal cancer and patients with either lung or pancreas cancer. It also supports the notion that the log odds rate of cancer is significantly different between patients in Newfoundland and patients in either Ontario or Quebec. Lastly, it supports the notion that the log odds rate of cancer is significantly different between female and male patients. Model 1 does not, however, include any interaction effects between these factor variables, so it does not support the claim, for example, that the effect a person living in Ontario has on odds of death rate is differs depending on whether that person is a girl or boy.

d)

```
#Change in deviance and when adding both 2way terms
ChangeDev = 182.7+24
df = 4

#Significance test. Calculate P val.
1-pchisq(ChangeDev,df)
```

```
## [1] 0
```

Based on the above test, the two interaction terms considered in Model 2 are jointly significant. The p-value produced is near zero.

e) A good fitting model will have a deviance residual plot with points randomly centered around zero and with constant variance. We also know that a poisson distribution with a large mean will be an approximately normal distribution. The deviance residual plots under both Model 1 have constant spread around zero for smaller fitted values, but for mid-range fitted values, the points stay below zero, and for high fitted values, the points stay above zero. Also, for any given cancer in the Dev. Residuals over Cancer Type plot in Model 1, males tend to be either above zero while females are below, or vise versa. In other words, there is not consistent spread for a specific gender across Cancer Type. These two negative characteristics do not appear in Model 2. Instead, Model 2's plots have points randomly centered around zero and point with constant variance. However, Model 1 fits the normal qq plot better, but this difference in marginal. Both follow the qqplot relatively well, which is what we'd exepct in a good model. In both model's residual plots, the variance of corresponding Regions is the same. For instance, the variance in residuals corresponding to Ontario are larger than residuals corresponding to Newfoundland in both Model 1 and Model 2. The main differentiating factors involve the first two plots mentioned above and the qq-plot. Taking these into account, I'd selesct Model 2 over model one, since the differences in the z score residual is minimal.