STA 141 Homework 2 Haozhe Gu 999200555

**The codes and results derived by using these codes constitute my own work.**

**I have consulted the following resources regarding this assignment: NONE**

**Assumptions: Data are independent and randomly selected.**

**Q1: Relationship between categorical variables**

First, we want to find the relationship between variables like Cancer.Incedence, Smoke etc.

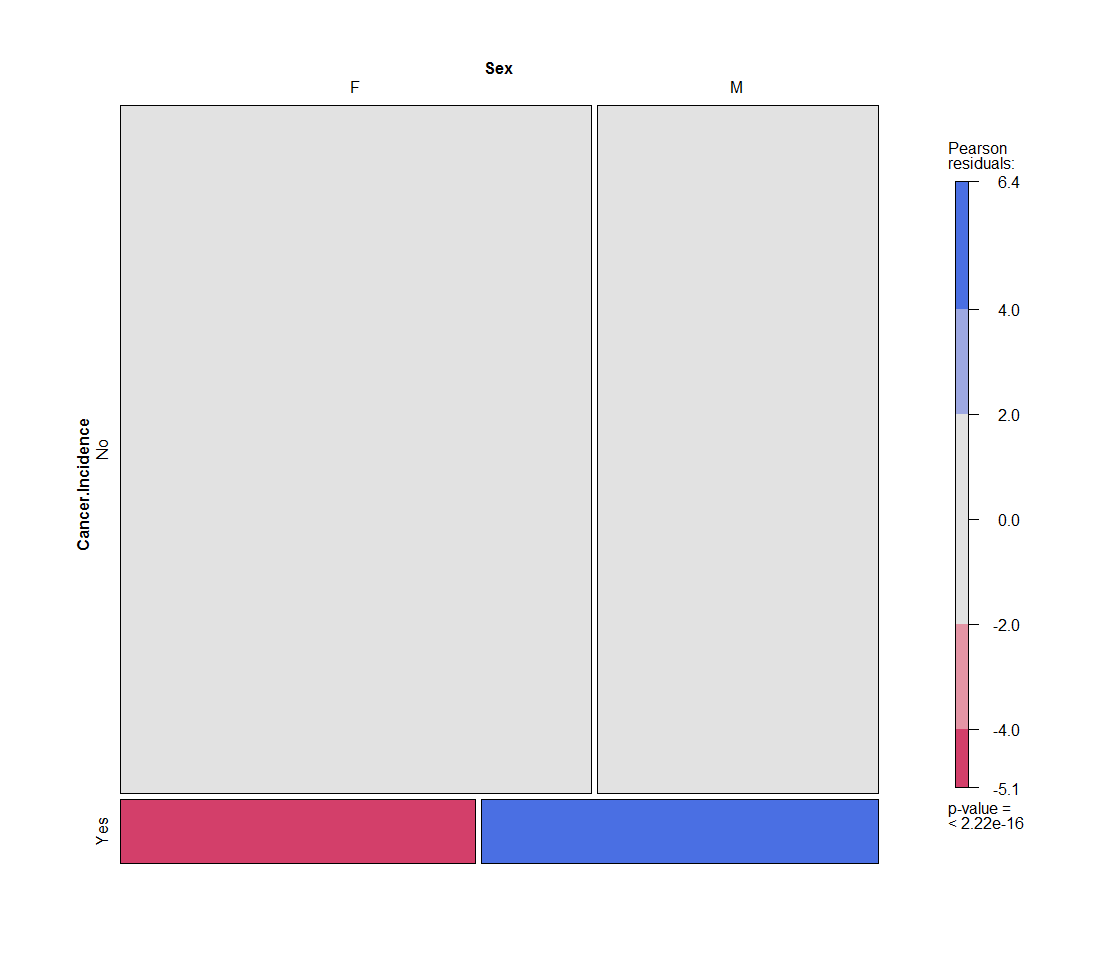
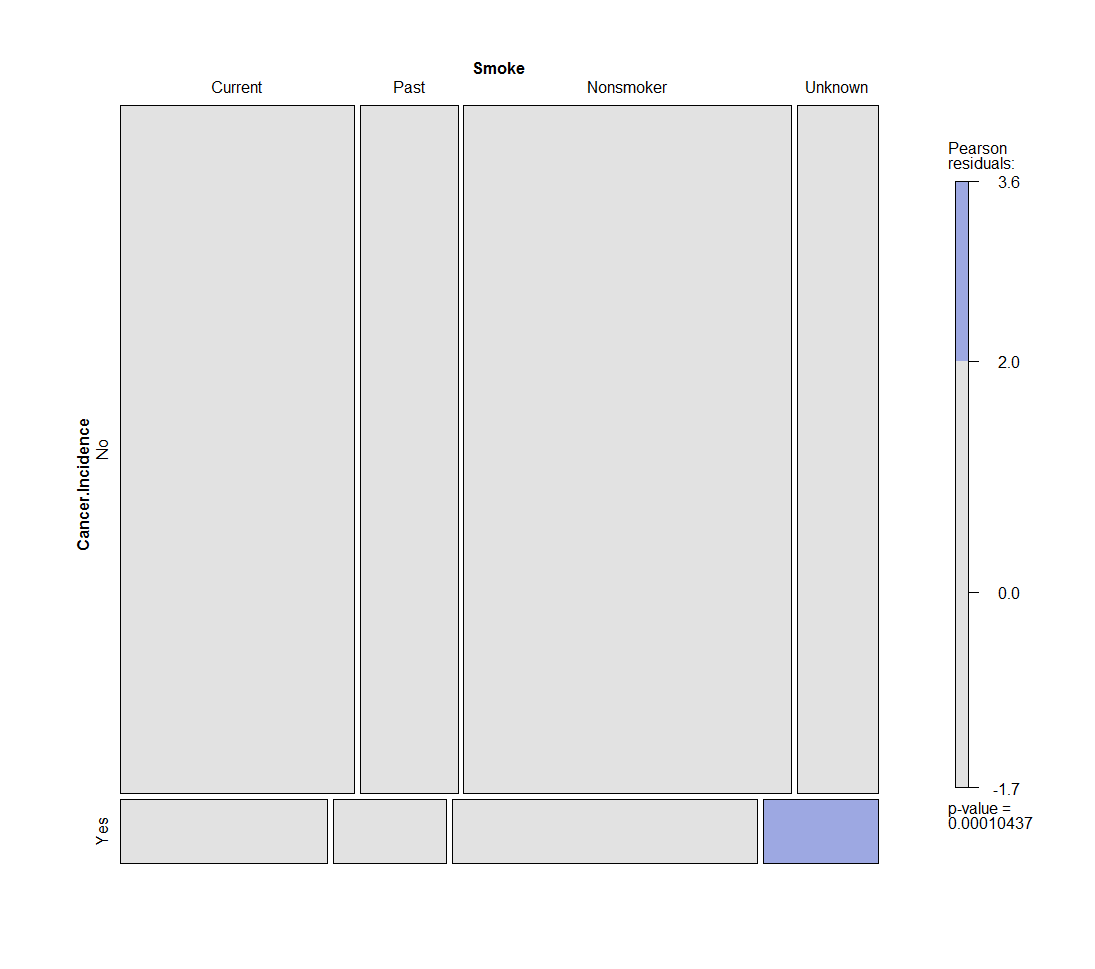


Figure 1.1 Figure 1.2

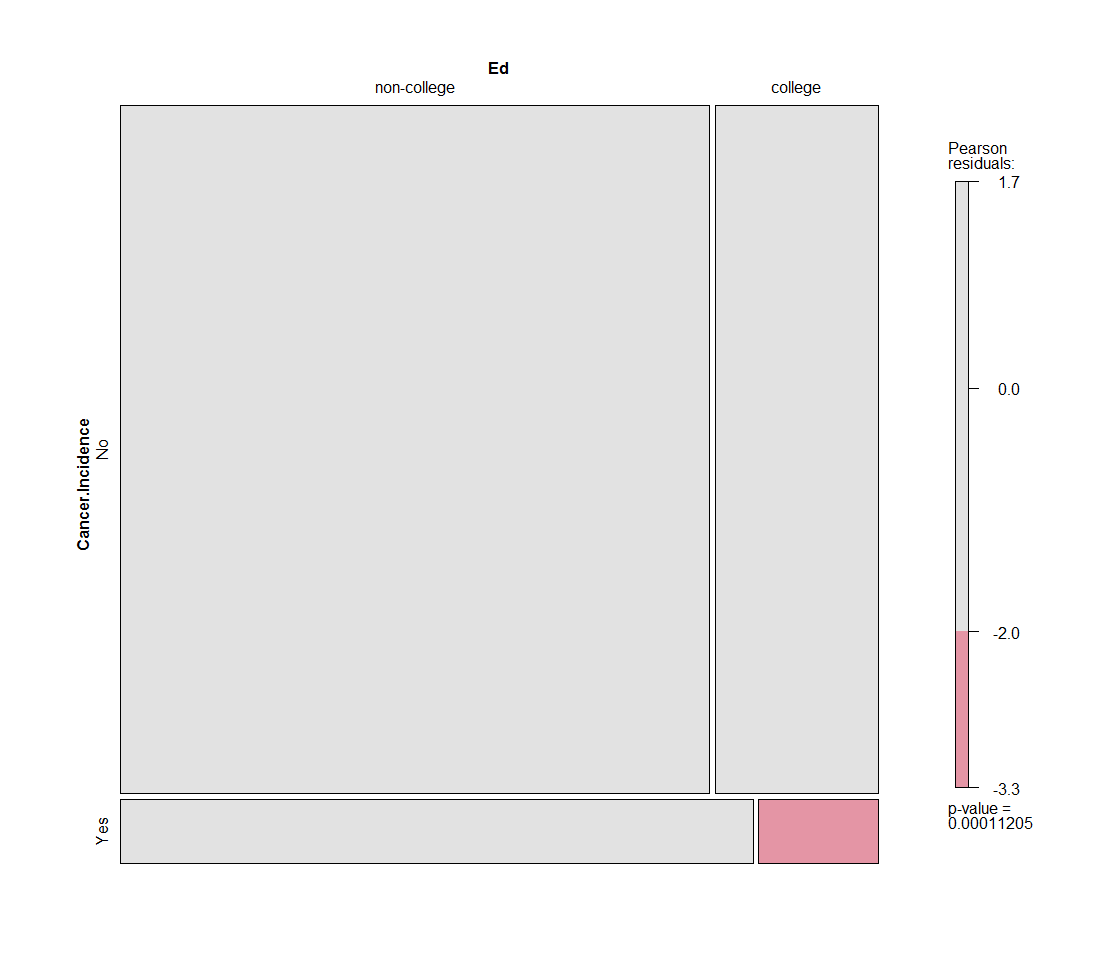
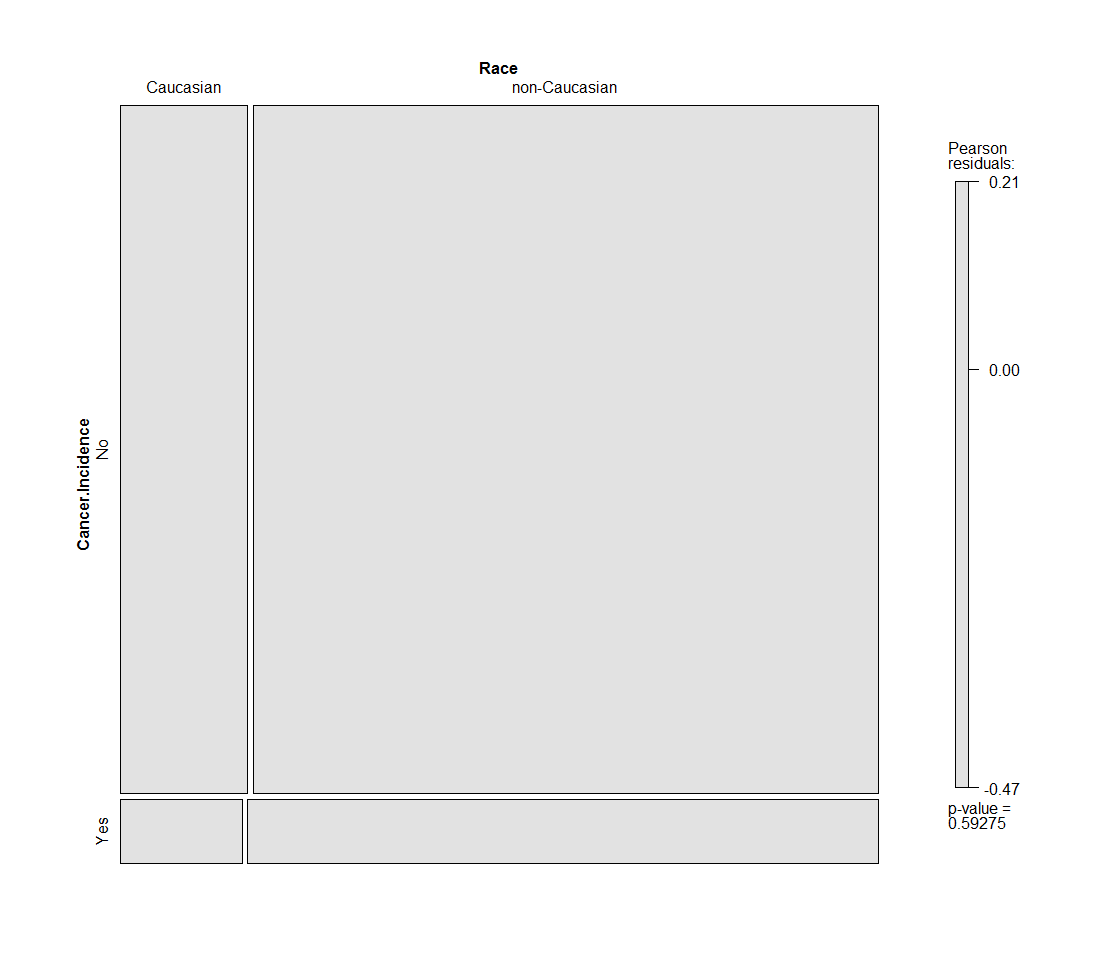
 

Figure 1.3 Figure 1.4

Figure1.1 to Figure 1.4 are pair-wise comparisons of Cancer Incidence with other categorical variables. From the Pearson residuals, Cancer Incidence only have dependence on Sex where when Cancer Incidence is Yes, Female have less incidence than its expected frequency while Male have more incidence than its expected frequency.

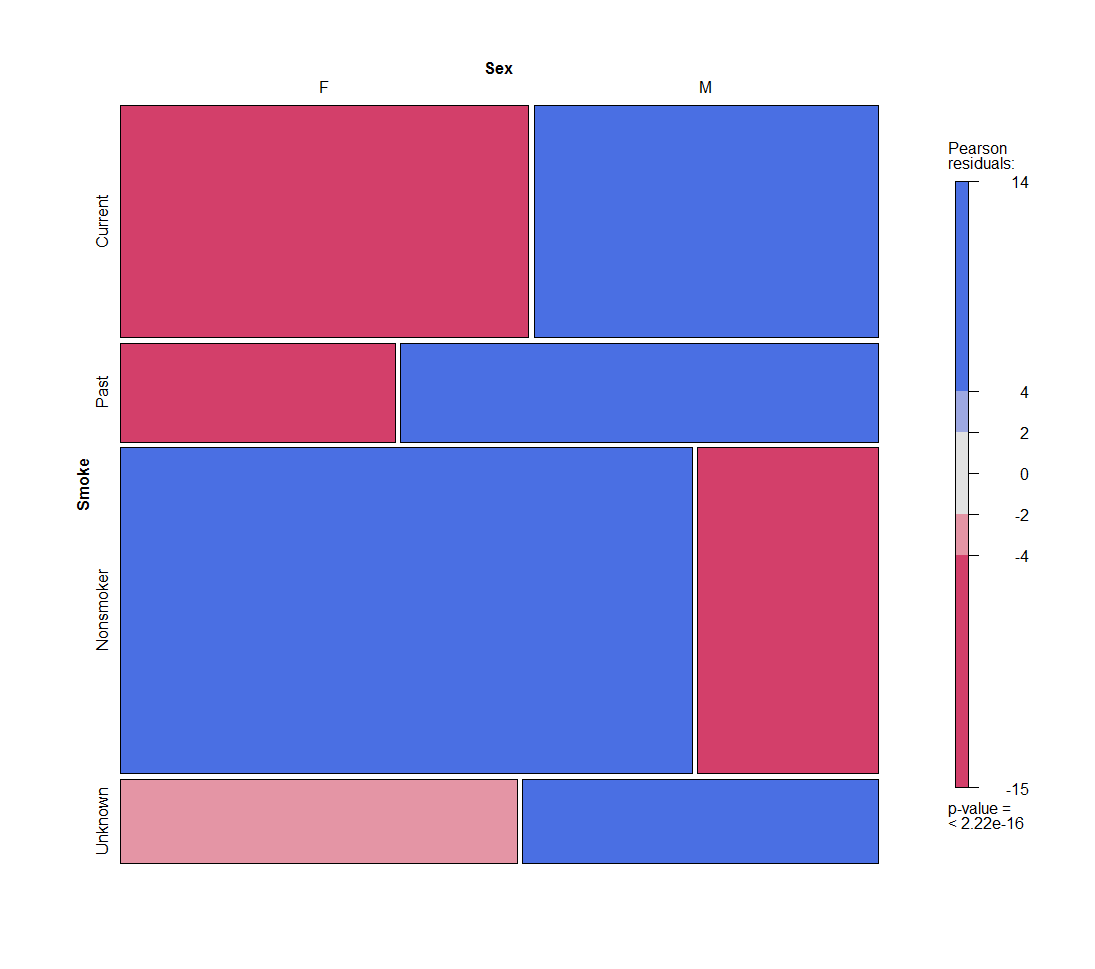
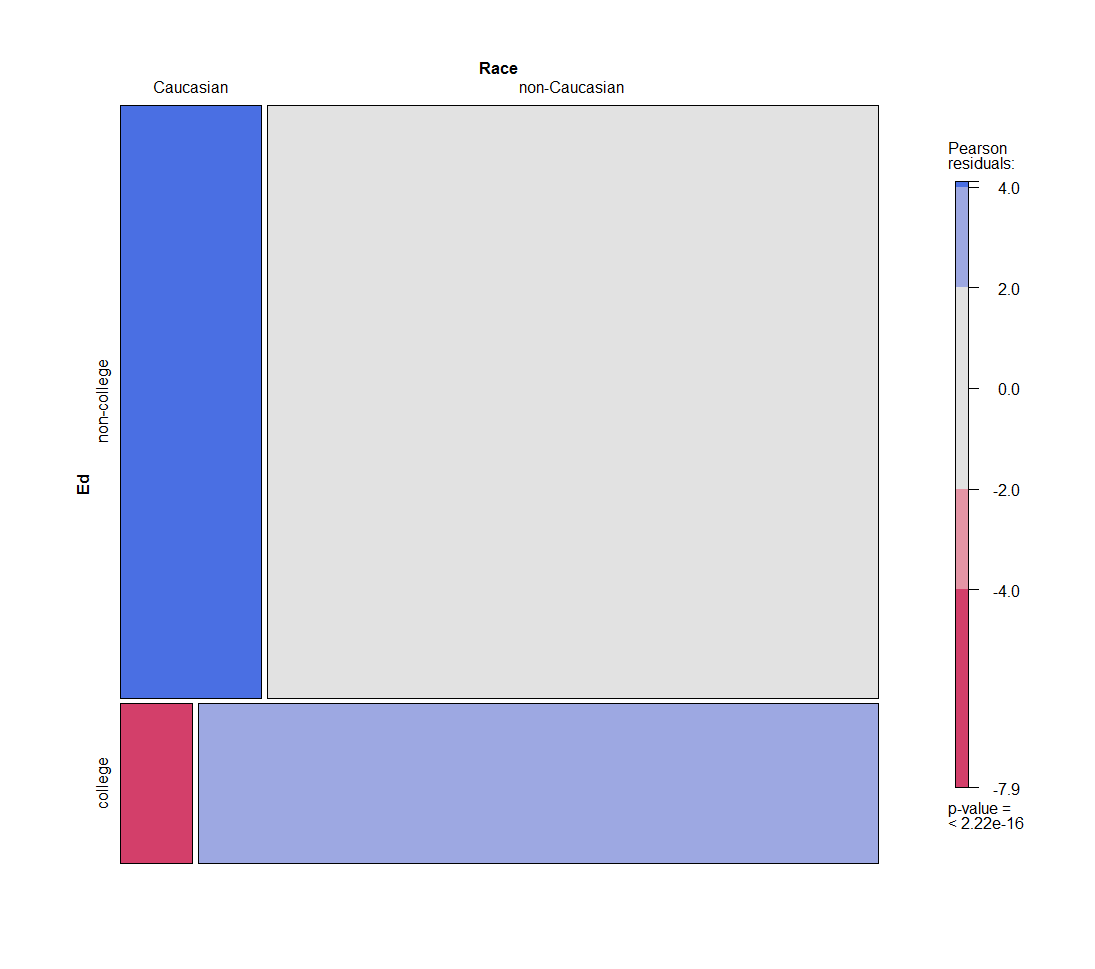
 

Figure 1.5 Figure 1.6

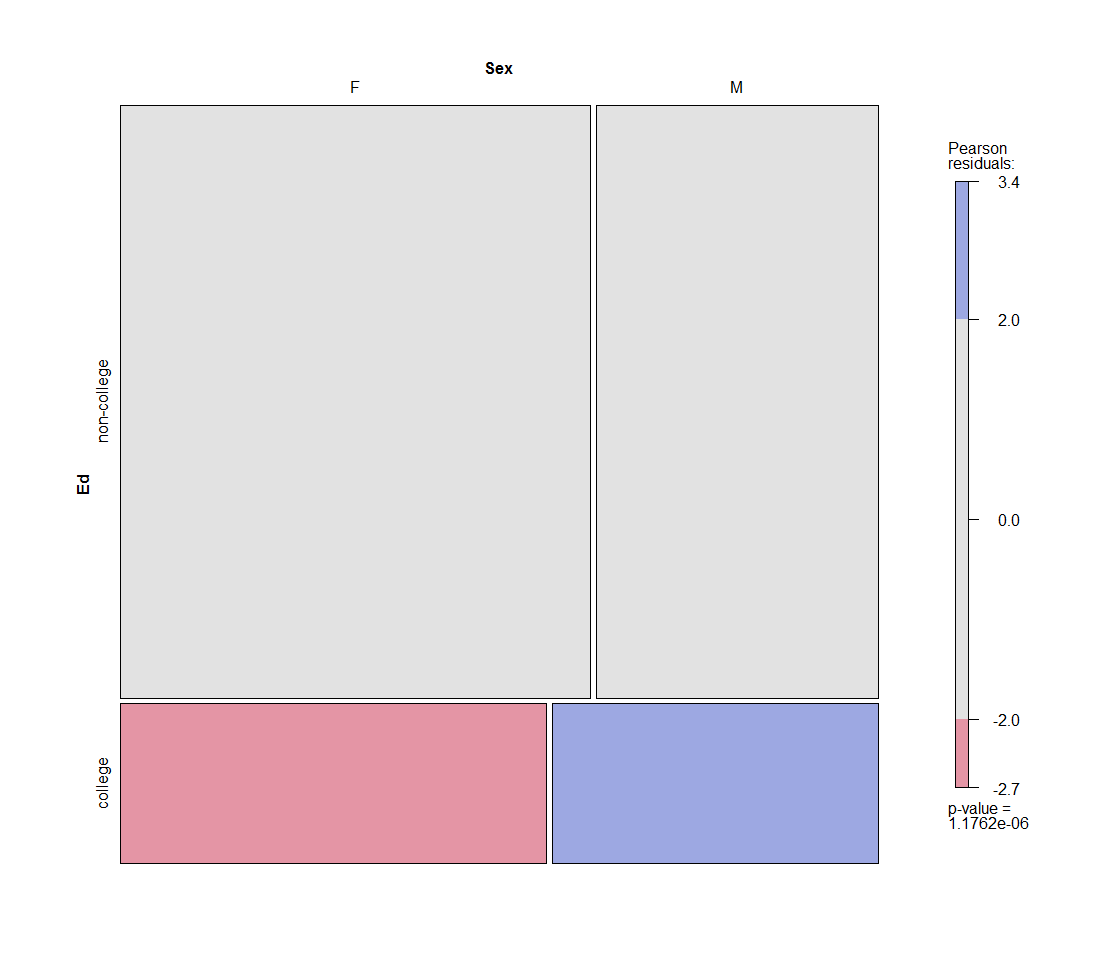


Figure 1.7

Figure 1.5 shows that Female actually have more incidence of Non-smoker than its expected frequency while Male have less incidence of Non-smoker than its expected frequency. However, the situation reversed in terms of Current-smoker and Past-smoker.

Figure 1.6 shows that Caucasian received more incidence of non-college education than its expected frequency; and, they received less incidence of college education than its expected frequency.

Figure 1.7 shows that the association between Education Level and Sex are not strong enough to draw a conclusion.

**Q2: Relationship between continuous variables**

In order to find the relationship between continuous variables, we use a correlation plot in figure 2.1

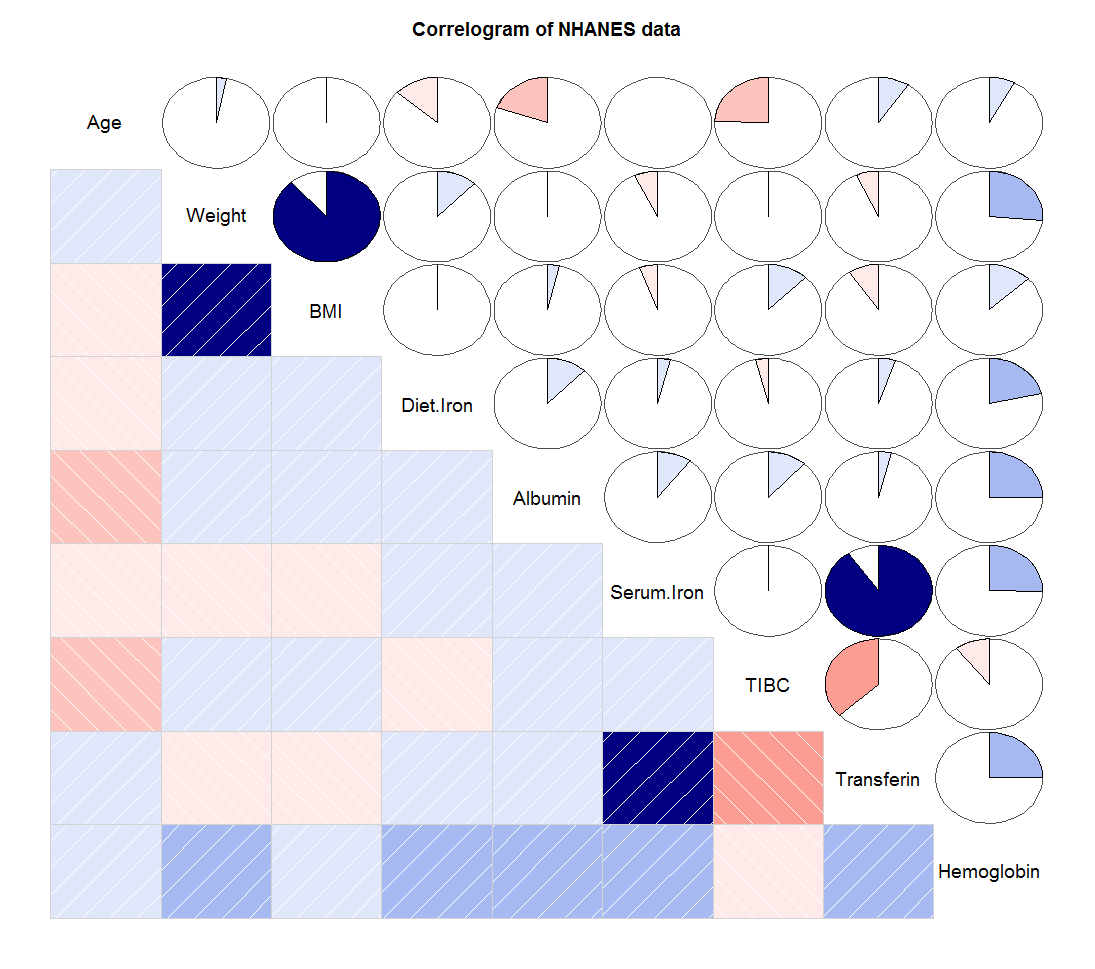


Figure 2.1

Figure1.7 shows that there is strong positive relationship between Weight & BMI and Serum.Iron & Transferin. Other variables are not related closely enough to claim a strong relationship. Figure 2.2 and Figure 2.3 shows the linear relationship of these two sets after log-transformation.

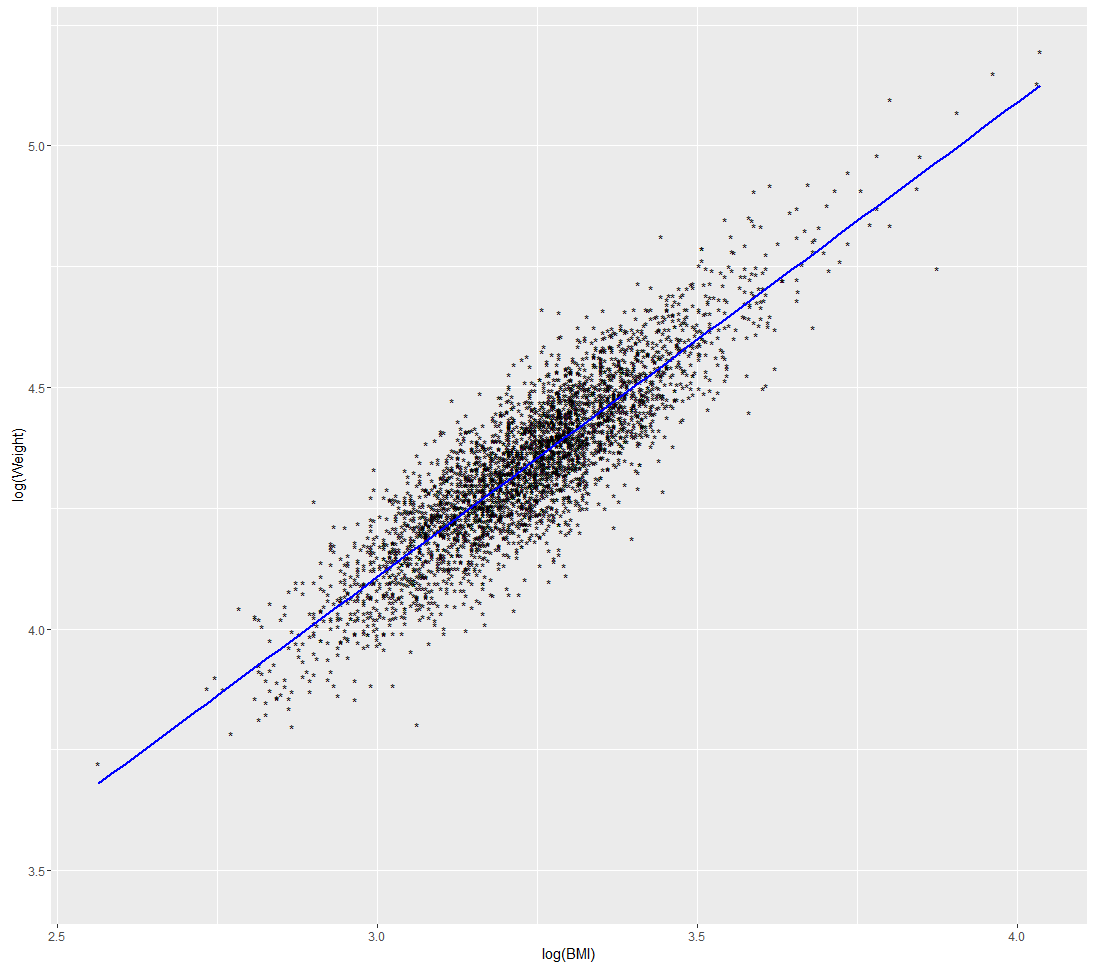
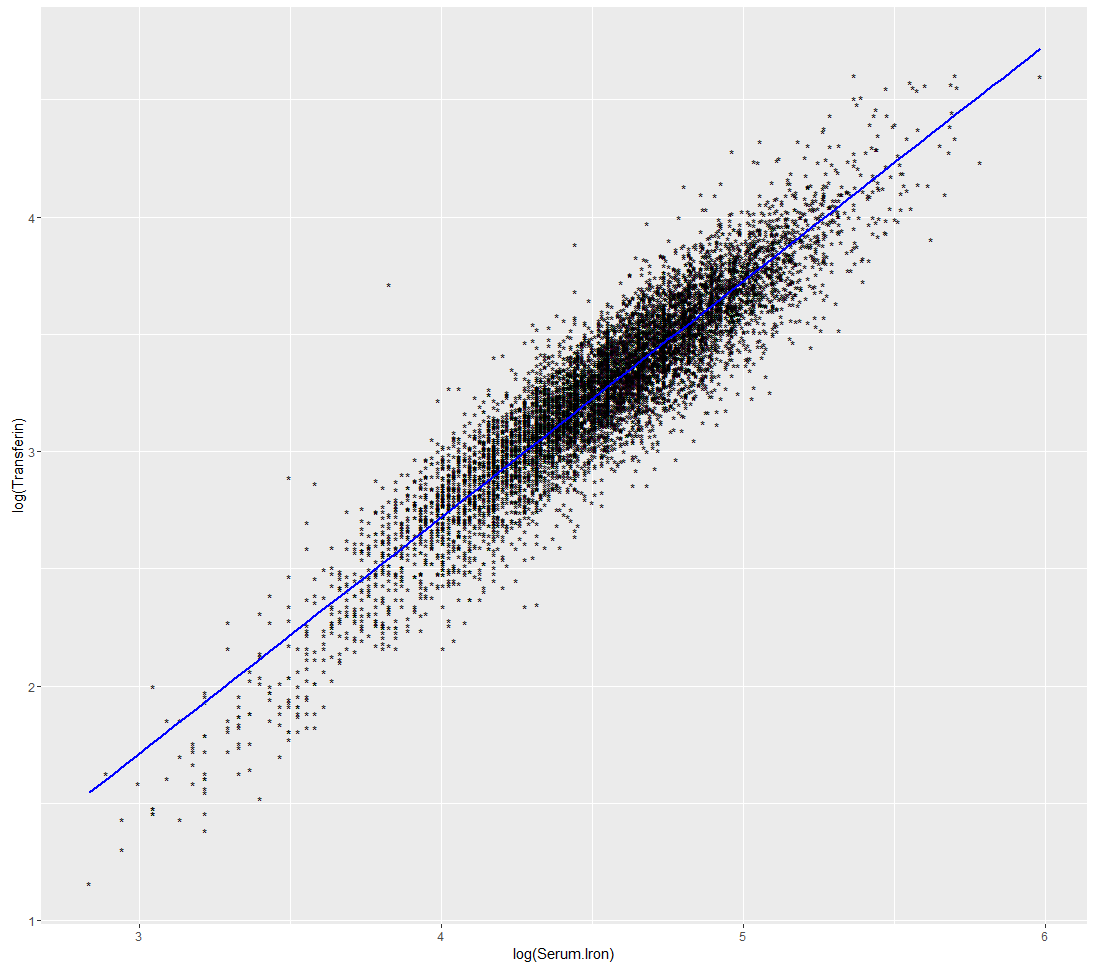
 

Figure 2.2 Figure 2.3

**Q3: Relationship between continuous variables conditioned on categorical variables**

* Weight conditioned on Smoke and Sex & Weight conditioned on Smoke and Education Level (Figure 3.1, Figure 3.2)

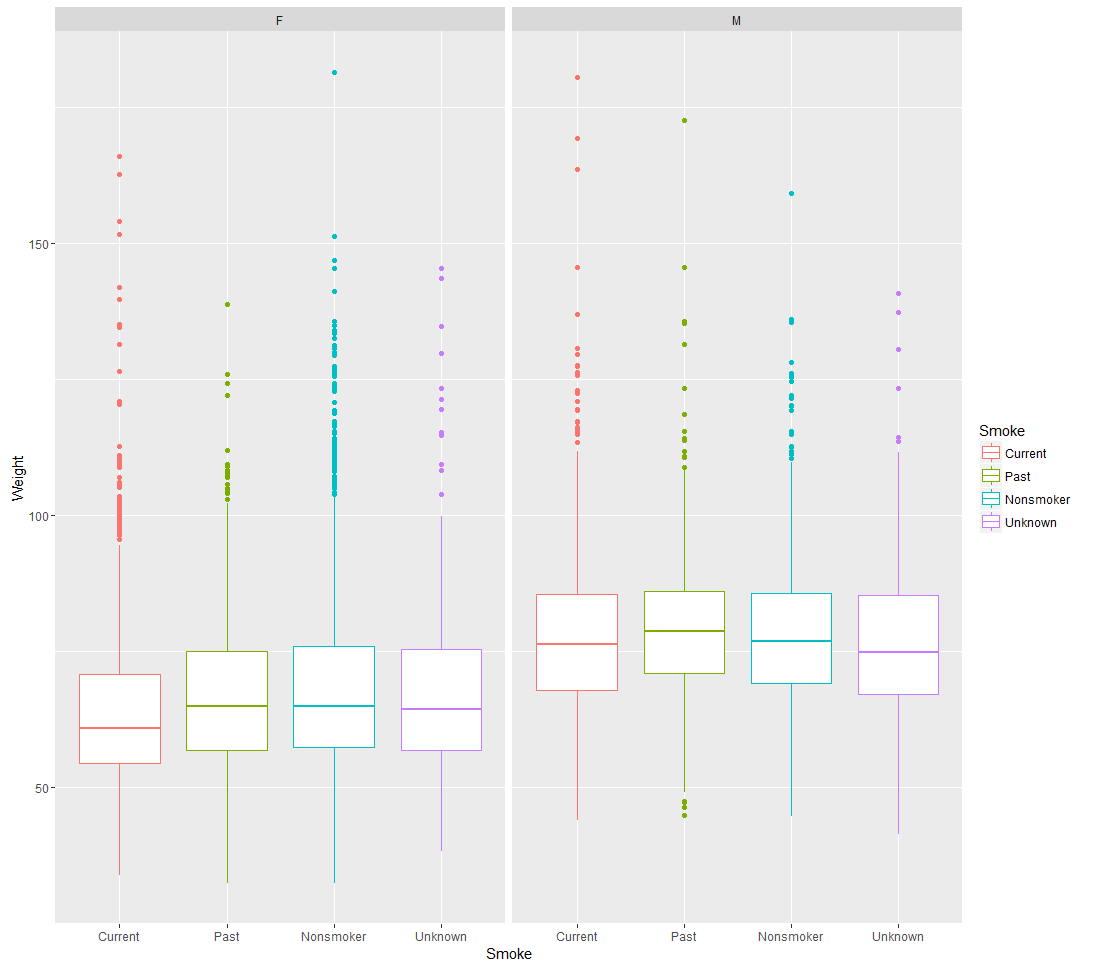
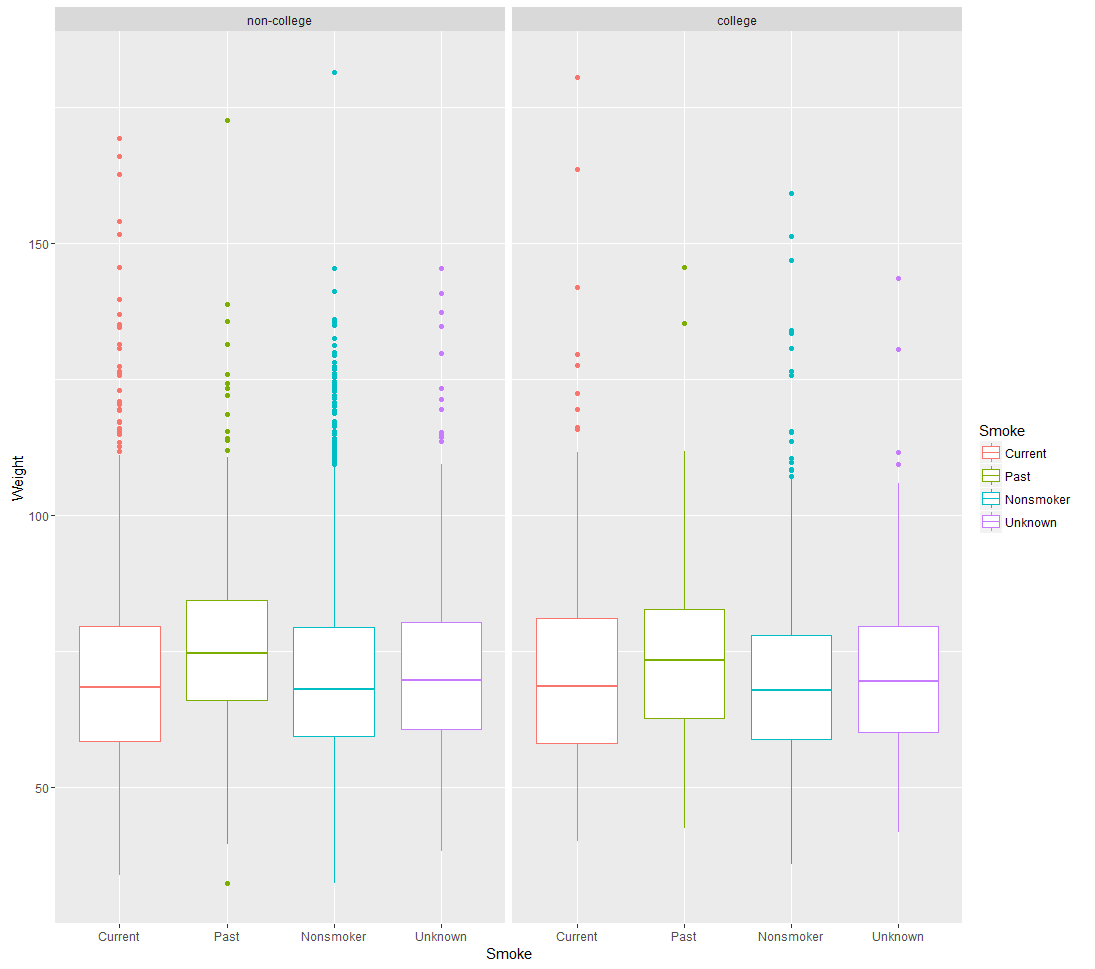
 

Figure 3.1 Figure 3.2

From Figure3.1, Female Current-smoker seems to have lower average Weight, while the average Weight of Male seems to be similar across smoker type. Generally, Males have higher Weight than Female which is a common sense.

From Figure3.2, Education Level seems to have no effect on the Weight. However, generally, post-smoker has a higher average weight across smoker type of smoker.

* Age conditioned on Smoke and Sex & Age conditioned on Smoke and Education Level

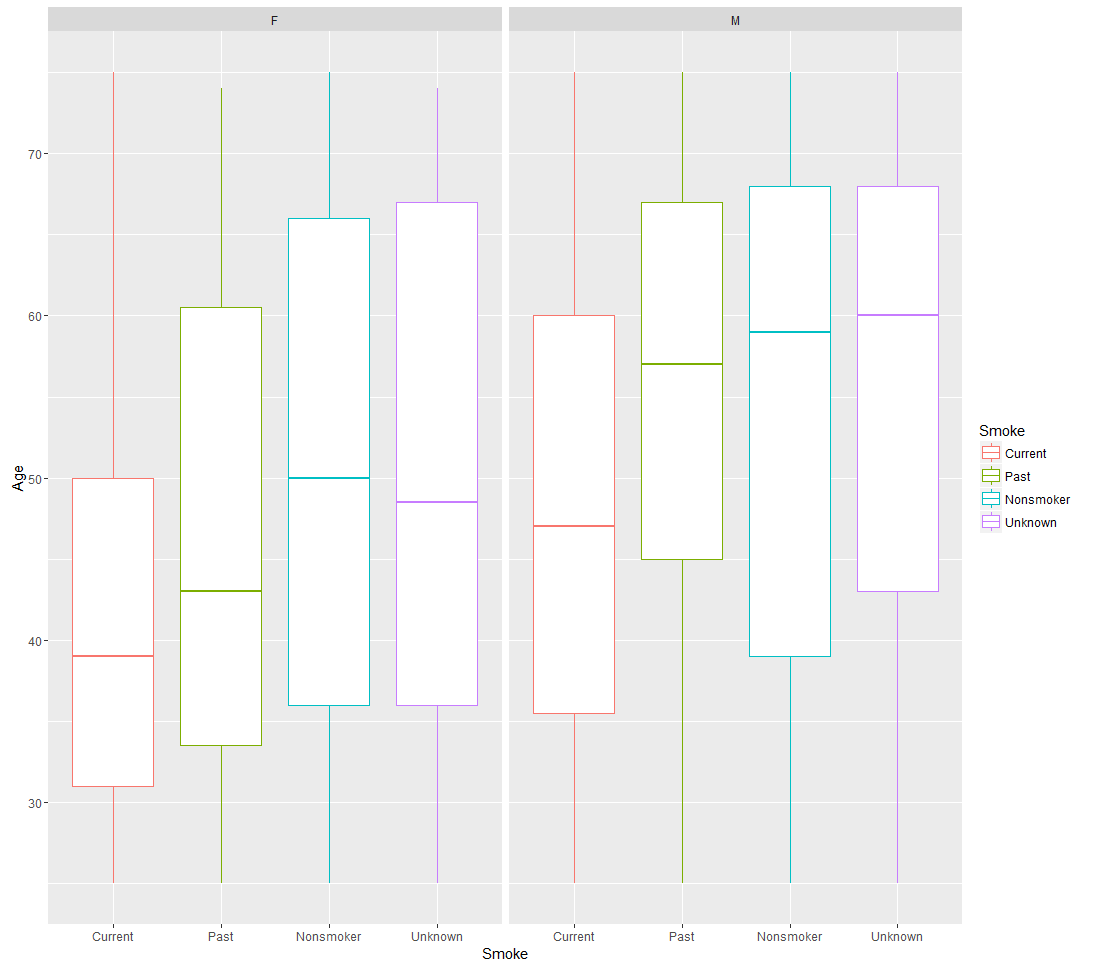
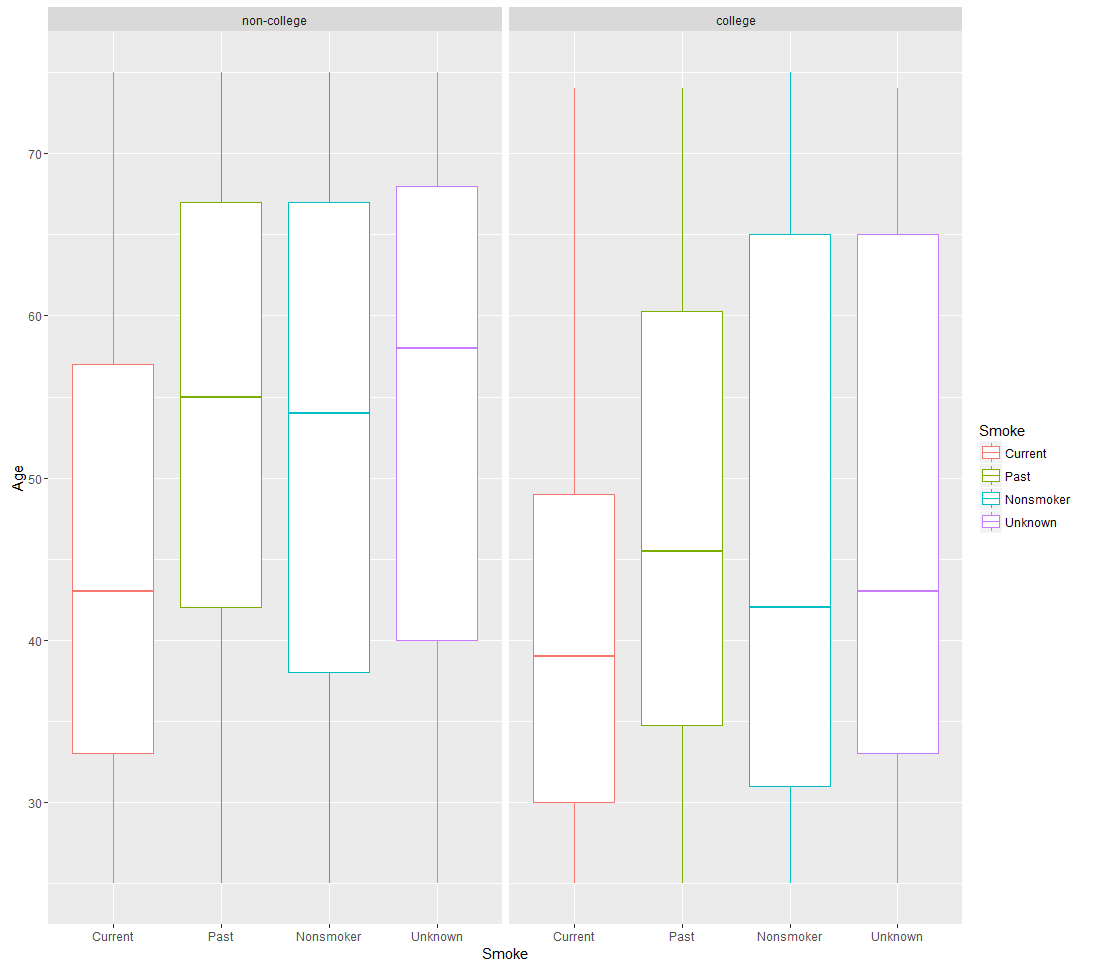
 

Figure 3.3 Figure 3.4

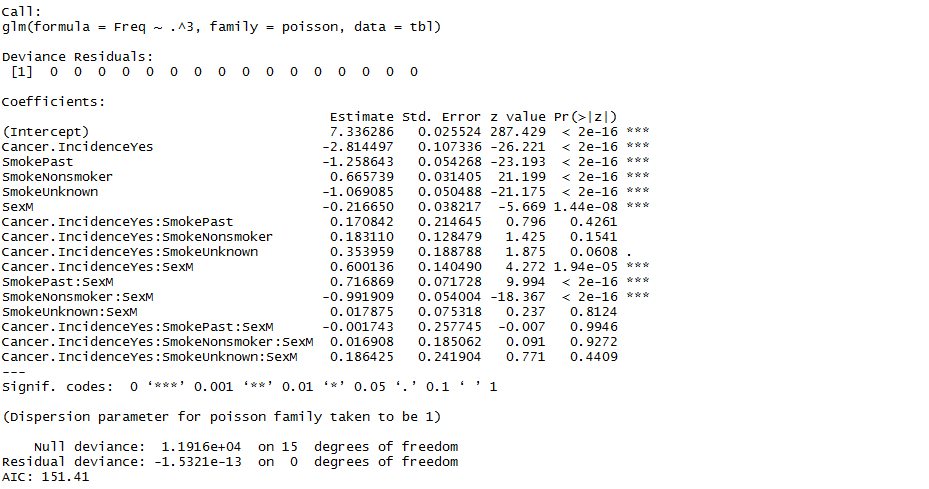
From Figure 3.3, both Current-smoker and Past-smoker for Female have lower average age than Male does. This implies Female has younger group of Current-smoker and quit smoking earlier than Male does.

From Figure 3.4, both Current-smoker and Past-smoker for people with college education have lower average than people with Non-college education does. This implies people with college education has younger group of Current-smoker and quit smoking earlier than people with Non-college education does.

**Q4: Fitting Log-Linear Model to Investigate Association**

In order to investigate specific associations between categorical variables and Cancer Incidence, I fit log-linear model to variable Cancer Incidence, Smoke and Sex.

The model’s assumption is data are independent and randomly selected. Observation frequency are normally distributed (using CLT)



The only significant odd ratio including Cancer Incidence is Cancer.IncidenceYes:SexM. This odd ratio has 95% CI



This implies that at any level of Smoke, Male with Cancer is at least 1.383720 odder than Female with Cancer. Which means Male are more likely to have cancer incidence.

**Appendix**

