Part 1:

Notes: We use **charges**, **age**, **bmi** and **smoker** to represent the variables in the report.

(a)Firstly, we plot the histogram of the 4 variables separately to observe the data characteristics. The frequencies of different ages are very close. The histogram of **bmi** is unimodal and normal whereas the histogram of **charges** is unimodal but extremely right-skewed. As **smoker** is a categorical variable, the frequencies only appear at 0 and 1. Therefore, we assume that the data of **age** follows the uniform distribution, the data of **bmi** follows the normal distribution, the data of **smoker** follows the binomial distribution and the data of charges follows the exponential distribution.

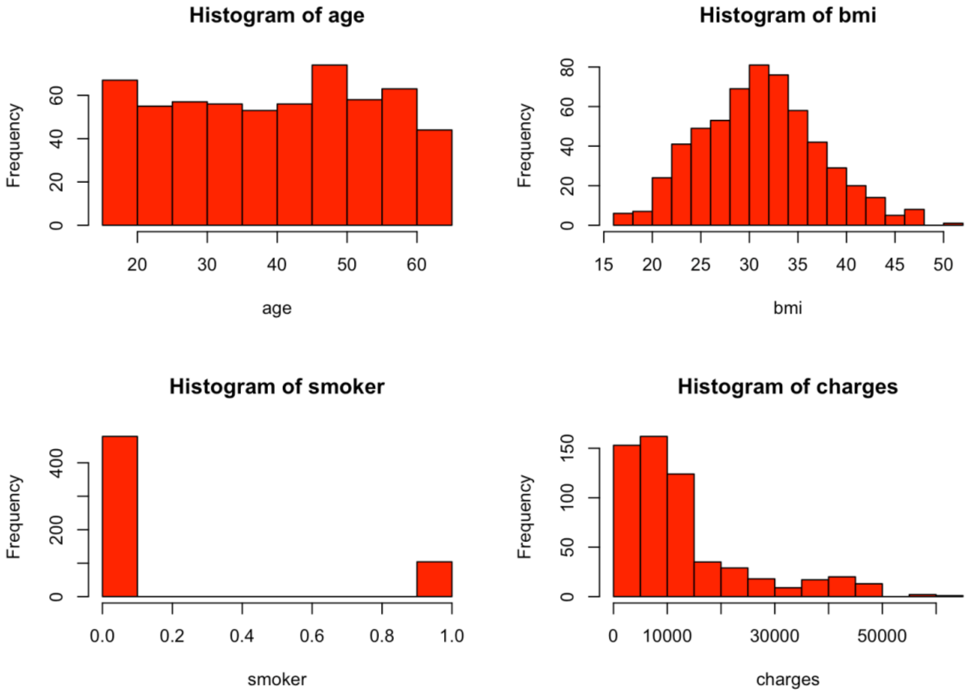

Figure 1: Histograms of 4 variables.

We take the mean and variance of the data as parameters of the distribution. Then we use Q-Q plot to compare these data with the distributions. The data of **age**, **bmi** and **smoker** looks roughly in line with the uniform distribution, the normal distribution and the binomial distribution respectively although there are still some small deviations. However, we can see that the top right of the **charges** data does not fit with the exponential distribution model well, therefore we can only judge that this data feature is severely right-skewed.

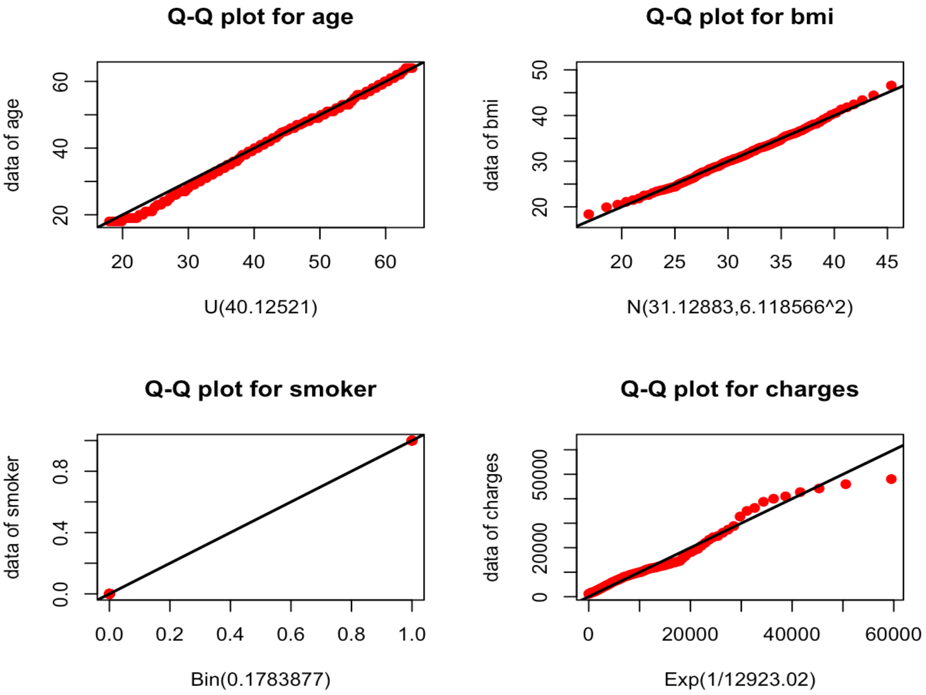

Figure 2: Q-Q plots for 4 variables.

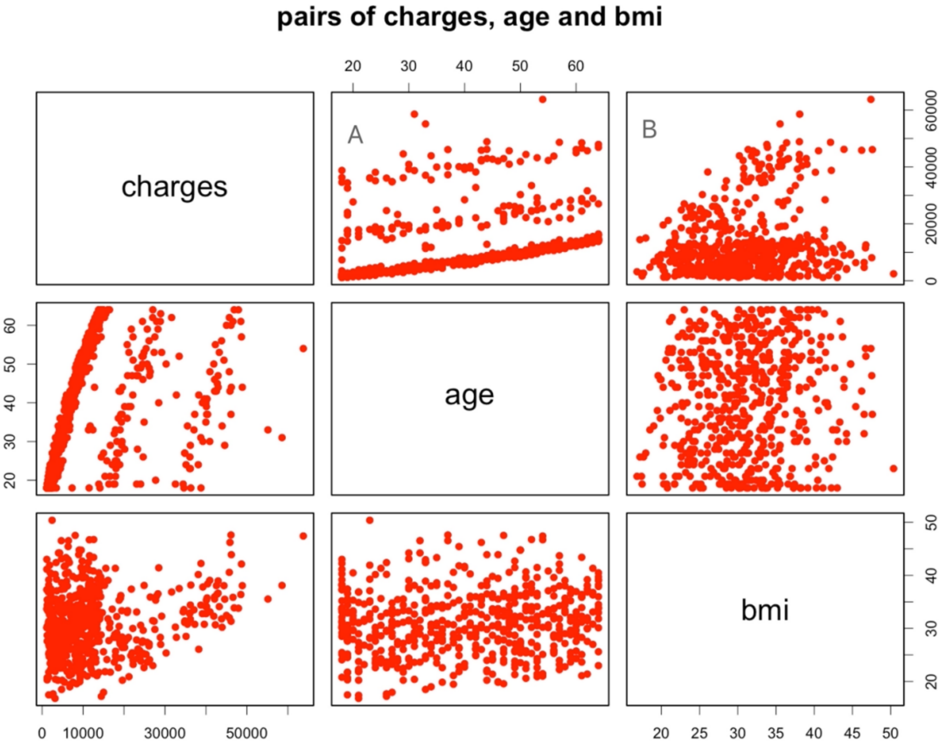

Figure 3: Plot showing the pairwise relationships between the 3 variables.

(b) **Comments:**The two scatter plots(plot A and plot B) in the top row of figure 3 show the pairwise relationship between the response variable **charges** and the explanatory variables **age** and **bmi** respectively. The points in the middle plot (A) of the top row can be roughly divided into upper, middle, lower, three parts and there are the most points concentrated in the lower part. And we can see that the **charges** tend to increase with a fixed slope as **age** increases so we initially guess that there is a linear relationship between **charges** and **age**. In contrast, the distribution of points in the right top plot (B) is more messy. But we can still abstractly divide these points into two part. The first part uniformly distributed in the bottom of the plot and this part also has the most points judged by the intensity. The other part shows that the data of **charges** has an increasing trend as the **bmi** increases. We can guess that this clustering phenomenon may be due to the influence of the categorical variable but we need further confirmation.

(c) We fit the multiple linear regression model, including all three explanatory variables together with all interaction terms at first, the fitted model is

$$\widehat{charges} = 259 \times age - 20527.31\, I(smoker = yes) + 47.06 \times bmi - 17.44 \times age\, I(smoker = yes) + 1471.54 \times bmi\, I(smoker = yes) - 3275.97$$

and summary output includes the following.

```
Call:

lm(formula = charges ~ (age * smoker + bmi * smoker))

Coefficients:

                   Estimate     Std. Error        t value              Pr(>|t|)

(Intercept)         -3275.97      1241.48        -2.639              0.00855 **

age                   259.00        15.43         16.785              < 2e-16 ***

smokeryes          -20527.31     2990.75          -6.864             1.74e-11***

bmi                    47.06        36.20          1.300               0.19404

age:smokeryes      -17.44          39.00          -0.447             0.65498

smoker:bmi        1471.54          83.46          17.633             < 2e-16 ***


Residual standard error: 4786 on 577 degrees of freedom
```

```
Multiple R-squared:  0.8363, Adjusted R-squared:  0.8348

F-statistic: 589.3 on 5 and 577 DF,  p-value: < 2.2e-16
```

The coefficient of determination is $R^2 = 83.63\%$ which indicates that the regression explains 83.63% of the variation in **charges**. And the F-test for this fit has a highly significant p-value, $< 2.2 \times 10^{-16}$, indicating that this model is a significantly better fit than the null model. This means these explanatory variables together explain a significant amount of the variation in the response variable **charges.**

In the Coefficients part, we can see that the coefficients of variables **age**, **smoker**, **smoker:bmi** all have highly significant p-values which are < 2e-16, 1.74e-11, and < 2e-16. But the coefficients of **bmi** and **age:smoker** have no significant p-values which are 0.19404 and 0.65498.This suggests that we should fit a model excluding these two terms. But considering the highly significant p-values of **smoker:bmi**, we need to retain the **bmi** term in the model.

Next we look at the analysis of the variance results to investigate further.

```
                Df   Sum Sq      Mean Sq    F value       Pr(>F)

smokeryes        1   5.0572e+10  5.0572e+10 2207.7487     <2e-16 ***

age              1   7.5452e+09  7.5452e+09  329.3900     <2e-16 ***

smoker:bmi       1   7.1219e+09  7.1219e+09  310.9109     <2e-16 ***

bmi              1   2.2545e+09  2.2545e+09   98.4224     <2e-16 ***

age:smokeryes 1      5.7155e+06  5.7155e+06     0.2495     0.6176

Residuals      577   1.3217e+10 2.2907e+07
```

We reorder the explanatory variables by SS from largest to smallest value. We can see that **smoker** has a highly significant p-value, $< 2.2 \times 10^{-16}$; in the presence of **smoker**, **age** has a very significant p-value, $< 2.2 \times 10^{-16}$; and in the presence of **smoker** and **age**, **smoker:bmi** also has a highly significant p-value, $< 2.2 \times 10^{-16}$; in the presence of **smoker**, **age** and **smoker:bmi**, **bmi** has a highly significant p-value, $< 2.2 \times 10^{-16}$ as well. But in the presence of the first four terms, **age:smoker** has a no significant p-value equal to 0.6176 with a small F value equal to 0.2495. So we are supposed to remove the **age:smoker** term out of the model but retain the **bmi** term.

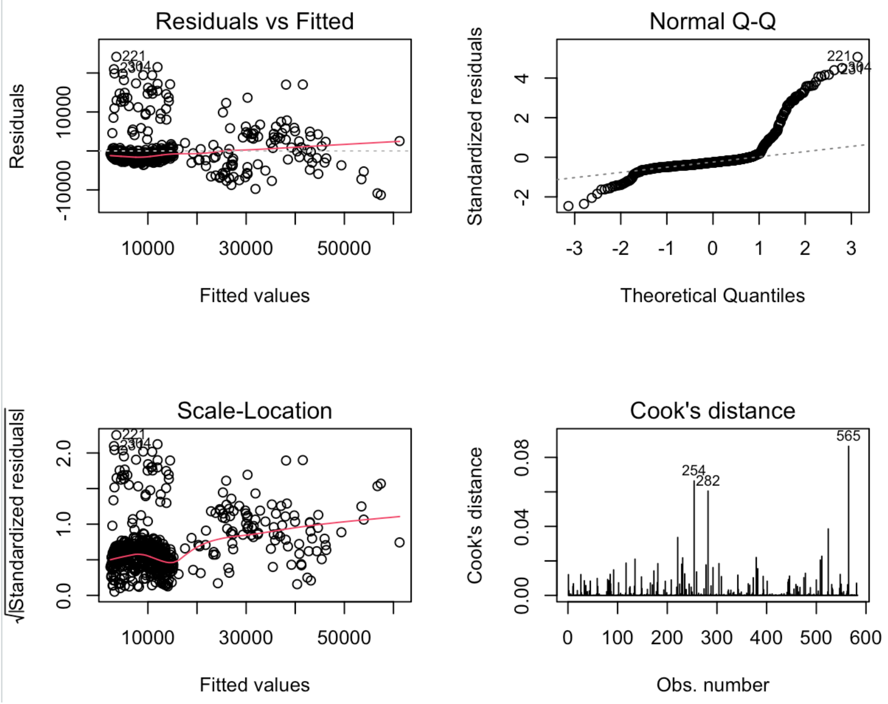

Figure 4: Plots of residuals for the regression of charges on (age, smoker, bmi, age:smoker, bmi:smoker)

Then we will look at the residuals plots (figure 4). Both the “Residuals vs Fitted” plot and the “Scale-Location” plot show the problem with heteroscedasticity. We can see many dots in the lower left corner of the graph which may be caused by taking too much data of **charges** around 10000. Overall, it generally shows that the fluctuation range of residuals decreases with the increase of **charges**. The Q-Q plot shows that residuals are roughly in the line with the normal distribution in the middle part but some large deviation at both ends, especially at the right end. The Cook’s distance shows some influential points which are point 254, point 282 and point 565. But the values of them are all below the threshold 1 and probably even smaller than 0.1 which makes it is acceptable for the model.

Then we fit the multiple linear regression model of **charges** on terms age, **bmi**, **smoker** and **bmi:smoker**. Then we get the fitted model is

$$\widehat{charges} = 256.27 \times age - 21160.24\ \times I(smoker = yes) + 47.81 \times bmi + 1469.54 \times bmi\ \times I(smoker = yes) - 3189.61$$

and the summary output includes the following.

```
Call:

lm(formula = charges ~ age + bmi * smoker)

Coefficients:

                         Estimate Std. Error t value Pr(>|t|)

              (Intercept)     -3189.61       1225.51    -2.603    0.00949 **

                age             256.27          14.16    18.096    < 2e-16 ***

                 bmi             47.81          36.13      1.323    0.18627

           smokeryes       -21160.24      2632.65    -8.038 5.19e-15 ***

     bmi:smokeryes       1469.54        83.28    17.646    < 2e-16 ***


Residual standard error: 4783 on 578 degrees of freedom

Multiple R-squared:   0.8362,  Adjusted R-squared:   0.8351

F-statistic: 737.7 on 4 and 578 DF,   p-value: < 2.2e-16
```

The coefficient of determination is $R^2 = 83.62\%$ which indicates that the regression explains 83.62% of the variation in **charges**. It is a little less than the previous model but the difference small enough to be ignored. And the F-test for this fit has a highly significant p-value, $< 2.2 \times 10^{-16}$ as well, indicating that this model is a significantly better fit than the null model. This means that these explanatory variables together explain a significant amount of the variation in the response variable **charges** similarly with the previous model.

Then we look at the analysis of the variance results to investigate further.

```
                Df       Sum Sq         Mean Sq        F value        Pr(>F)

smoker          1       5.0817e+10    5.0817e+10      2221.542    < 2.2e-16 ***

age             1       7.5452e+09    7.5452e+09      329.847     < 2.2e-16 ***

bmi:smoker      1       7.1230e+09    7.1230e+09      311.392     < 2.2e-16 ***

bmi             1       2.0090e+09    2.0090e+09       87.826     < 2.2e-16 ***

Residuals       578     1.3222e+10    2.2875e+07
```

We can see that **smoker** has a very significant p-value, $< 2.2 \times 10^{-16}$; in the presence of **smoker**, **age** has a highly significant p-value, $< 2.2 \times 10^{-16}$; and in the presence of **smoker** and **age**, **bmi:smoker** also has a highly significant p-value, $< 2.2 \times 10^{-16}$; in the presence of **smoker**, **age** and **bmi:smoker**, **bmi** has a highly significant p-value, $< 2.2 \times 10^{-16}$ as well. This suggests that we should retain all of the 4 terms in the model.

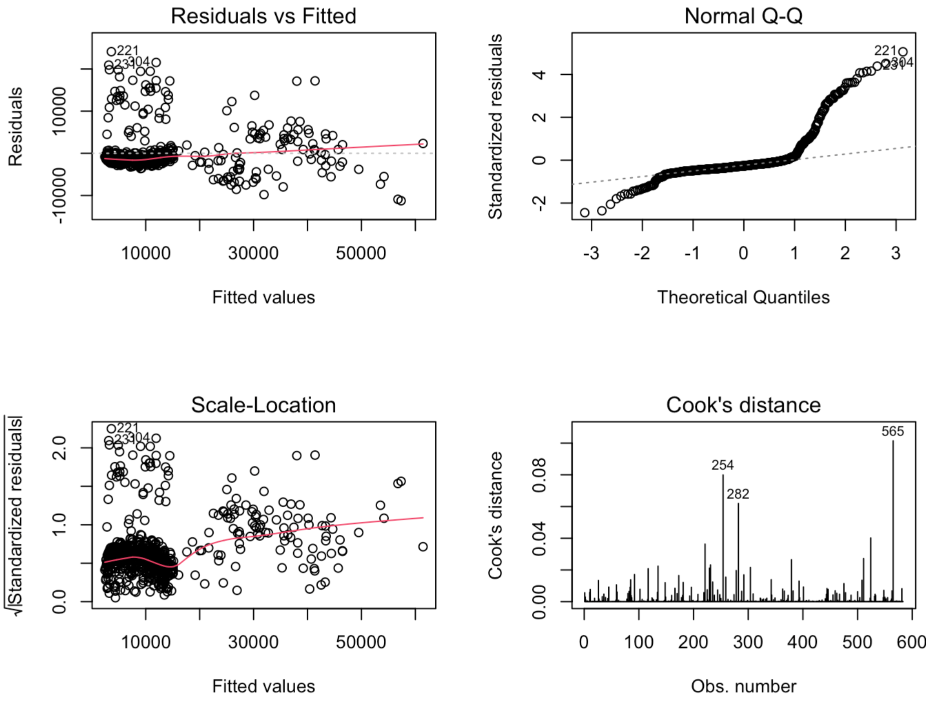

Figure 5: Plots of residuals for the regression of charges on (age, smoker, bmi, bmi:smoker)

Now we will look at the residuals plots (figure 5).It is obvious that they are similar with the residuals plots of the first model. The “Residuals vs Fitted” plot and the “Scale-Location” plot both show the heteroscedasticity of residuals. The Q-Q plot indicates that the middle part of residuals are well-modeled by the normal distribution but there are still large deviations at both ends. As with the first model, the deviations of the right end is larger than the left end. And the Cook’s distance also shows 3 influential points same as before, point 254, point 282 and point 565 but they are all much smaller than the threshold 1 so that we do not need to remove them.

Conclusion: We compare two fitted models from multiple aspects. They have the similar coefficient of determination which are 83.63% and 83.62%. It indicates that these models can explain the similar variation in **charges**. However,from the perspective of residuals analysis, the two have very similar residuals conditions. Homoscedasticity and normality are not well accepted, so the fitted models do not offer excellent linear regressions of **charges** on the 3 explanatory variables with interaction terms. The terms of the multiple linear regression model of **charges** on terms **age**, **bmi**, **smoker** and **bmi:smoker** are all significant but the **age:smoker** term of the first multiple linear regression model is not significant. The second model is more concise, so we think the second model is better. So our fitted model indicates that an increase of 1 year in age of primary beneficiary corresponds to an increase of ＄256.27 in individual medical costs billed by health insurance, an increase of 1 unit in body mass index corresponds to an increase of ＄1517.35 (＄47.81+＄1469.54) in individual medical costs billed by health insurance and another cost of smoker if they are smokers. An increase of 1 in body mass index only corresponds to an increase of ＄47.81 in individual medical costs billed by health insurance and there is no cost of smoker if they are not smokers.

(d) We compute the associated 95% confidence interval which is from 19509.67 to 22623.3 for the mean medical costs of smoking policyholders aged 44 staying in Asia-Pacific region with body mass index of 22.5. We choose the model with terms **age**, **bmi**, **smoker** and **bmi:smoker**. This model is reliable for the policyholders. Because all the information values are within the range of the data on which we built the model. But we think the confidence interval is not reasonably reliable for this model. Because we use the mean and the standard error of the response variable to estimate the confidence interval based on normal distribution. The fitted model we picked provides significant linear relationships to the response variable **charges** with the explanatory terms, but the plots of residuals shows the homoscedasticity and the less normality. We know that the response variable values are random because of the randomness in the error term. So we think the disadvantages of the fitted model shown by the plots of residuals may have impact on the standard error used in the confidence interval prediction. It probably suggests reduction of the prediction accuracy. Therefore, we think this confidence interval from 19509.67 to 22623.3 is not very reliable.

Part 2:

Notes: We use **Risk**, **Age**, **Totchol**, **Glucose** and **Gender** to represent the variables in the report.

(a) We produce box plots (figure 6) to explore the relationship between the response variable **Risk** and each of the explanatory variables **Gender**, **Age**,**Totchol** and **Glucose** because **Risk** is a categorical variable.

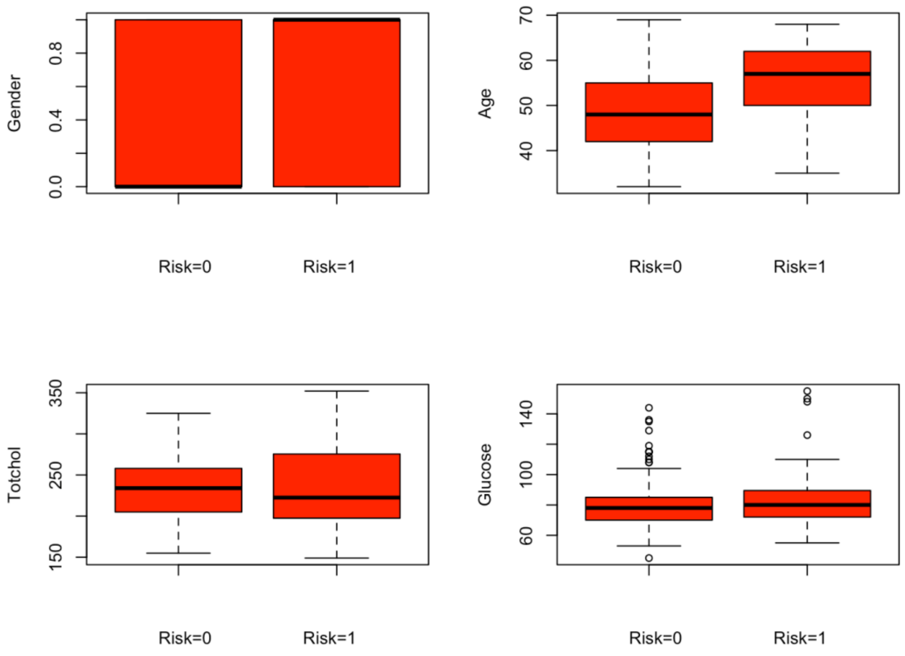

Figure 6: Box plots of response variable and explanatory variables

The first plot shows the relationship between **Risk** and **Gender**. More **Gender** data is concentrated on female when **Risk**=0 and more **Gender** data is concentrated on male when **Risk**=1. It indicates that males have more risks than females.

The second plot shows the relationship between **Risk** and **Age**. When **Risk**=0, the distribution of **Age** is wider than the distribution of **Age** when **Risk**=1. But the center of former is smaller. So we think the older people may have more risks.

The third plot shows the relationship between **Risk** and **Totchol**. The data of **Totchol** has a wider distribution and a smaller median when **Risk**=1.

The last plot shows the relationship between **Risk** and **Glucose**. The values of distribution range when **Risk**=1 is slightly larger than the range when **Risk**=0. But the difference is not obvious. Besides, the outliers when **Risk**=1 are larger than the outliers when **Risk**=0 which may indicates that the extremely large **Glucose** may result the **Risk**.

(b) To investigate if the risk of 10-year coronary heart disease is dependent on patient's gender, we carry out an appropriate hypothesis test of independence.

H_0: The risk of 10-year coronary heart disease is independent on patient's gender.

H_1: The risk of 10-year coronary heart disease is dependent on patient's gender.

The contingency table of frequencies of observations is following.

Frequency	Male	Female	Total
Risk	37	15	52
No risk	104	161	265
Total	141	176	317

Form 1: Contingency table of frequencies of observations.

The contingency table of expected frequencies is below.

Frequency	Male	Female	Total
Risk	$7332/_{317}$	$9152/_{317}$	$16483/_{317}$
No risk	$37365/_{317}$	$46640/_{317}$	$84005/_{317}$
Total	$44697/_{317}$	$55792/_{317}$	317

Form 2: Contingency table of expected frequencies.

Degree of freedom is 1 ((2-1)×(2-1)). Then we perform a X^2 test to get the statistic value X^2 equals to 17.922 and p-value is 2.301×10^{-5} without Yate's continuity correction. Besides, we calculate the X^2 value equals to 16.653 and p-value is 4.487 × 10^{-5} with Yate's continuity correction. We can see these two results are similarly significant because there is not observations smaller than 5. So the Yate's continuity correction is not necessary.

Then we set α(=0.01) as the significance level. Obviously the p-value is smaller than the significance level, so there is strong evidence against the null hypothesis. Therefore, we think the risk of 10-year coronary heart disease is dependent on patient's gender.

(c) For risk is a binomial response, we use logit link as the default link function to fit a generalized linear model for **Risk** as the response variable and **Age**, **Totchol**,

Glocose and **Gender** as the as explanatory variables with no interaction term, summary output includes the following.

```
glm(formula = Risk ~ Age + Totchol + Glucose + Gender, family = binomial)
                   Estimate     Std. Error      z value      Pr(>|z|)

(Intercept)     -8.278878     1.637318       -5.056      4.27e-07 ***

Age               0.094916      0.020693        4.587      4.50e-06 ***

Totchol           0.001019      0.004165        0.245      0.806734

Glucose          0.008647      0.009564        0.904      0.365954

GenderM        1.273111      0.351864        3.618      0.000297 ***

      Null deviance: 282.96   on 316   degrees of freedom

Residual deviance: 237.82   on 312   degrees of freedom

AIC: 247.82
```

Now we look at the individual hypothesis tests of the 4 explanatory variables. The p-values of terms, **Age** and **Gender**, are all highly significant so that we consider to retain them in the model. But the p-value of the **Totchol** term is 0.806734 and the **Glucose** term is 0.365954 which are not significant at 0.1 level. Therefore, we try to remove the insignificant terms to fit some other models. Then we get the following residuals and AIC values.

Terms present	Deviance	df	AIC
Age+Totchol+Glucose+Gender	237.82	312	247.82
Age+Glucose+Gender	237.88	313	245.88
Age+Totchol+Gender	238.63	313	246.63
Age+Gender	238.78	314	244.78

Form 3: Values of deviance, df and AIC of fitted models.

Analysis of deviance:

We compare the second model removed the **Totchol** term with the first model by analyzing the change of deviance, 237.88-237.82=0.06. We calculate the p-value,

$\Pr(\chi^2_{313-312} > 237.88 - 237.82)$ equals to 0.8064959. It indicates that removing the **Totchol** term, the reduction in goodness-of-fit is not significant at the 10% level.

Comparing **Age**+**Totchol**+**Glucose**+**Gender** model with **Age**+**Totchol**+**Gender** model, p-value is $\Pr(\chi^2_{313-312} > 238.63 - 237.82) = 0.3681203$ which means removing the **Glucose** term does not result in a highly significant reduction of goodness-of-fit.

Comparing **Age**+**Totchol**+**Glucose**+**Gender** model with **Age**+**Gender** model, p-value is $\Pr(\chi^2_{314-312} > 238.78 - 237.82) = 0.6187834$ which means removing the **Glucose** term and the **Totchol** term does not result in a highly significant reduction of goodness-of-fit.

We can judge from the analysis of deviance that the **Totcol** term and **Glucose** term should be removed from the model, the **Age** term and **Gender** term should be retained in the model.

Then we consider the AIC values. The smaller the AIC values is, the model has a better goodness-of-fit. We can see the model with terms **Age** + **Gender** has the smallest AIC value. And compared with the model with terms **Age** + **Totchol** + **Glucose** +**Gender**, there is a difference of 244.78-247.82=-3.04 of AIC value which is much smaller than the threshold value of 2. It represents that the model with terms **Age** + **Gender** provides substantial improvement compared with the model with terms **Age** + **Totchol** + **Glucose** +**Gender**. So we prefer the model with explanatory variables **Age** and **Gender**.

Then we test the model fit for this model by the following hypothesis.

H_0: The model fit is adequate.

H_1: The model fit is not adequate.

We calculate the p-value is $\Pr(\chi^2_{314} > 238.78)$ =0.9994319. It is large so that this model is not significantly worse fit than the maximal model at 10% level.

Now, we compare the model with explanatory variables **Age** and **Gender** with the null model. The p-value, $\Pr(\chi^2_{316-314} > 282.96 - 238.78)$ equals to 2.549382 $\times 10^{-10}$ which indicates a very strong evidence that this model fits better than the null model.

We denote $D_{x,y}$ is the number of patients who have a 10-year risk of coronary heart disease, $E_{x,y}$ is the total number of patients, x is the number of the age of the patient and y is the kind of **Gender** where y=1 when **Gender** is male and y=0 when **Gender** is female. $q_{x,y}$ represents the probability a patient has a 10-year risk of coronary heart disease. Then we write the model as the following.

$$D_{x,y} \sim Binomial(E_{x,y}, q_{x,y})$$

$$\text{Where } \begin{cases} q_{x,1} = \frac{e^{0.09750 \times x + 1.28039 - 7.46925}}{1 + e^{0.09750 \times x + 1.28039 - 7.46925}} \\ q_{x,0} = \frac{e^{0.09750 \times x - 7.46925}}{1 + e^{0.09750 \times x - 7.46925}} \end{cases}$$

Or we can write as

$$q_{x,y} = \frac{e^{0.09750 \times x + 1.28039 * I(Gender = Male) - 7.46925}}{1 + e^{0.09750 \times x + 1.28039 * I(Gender = Male) - 7.46925}}$$

Now we see the figure 6. We find that the deviance residuals have a linear relationship with both **Age** and fitted values. The different **Gender** has also impact on the deviance residuals. As a result, we don't think it's a well fitted model. But it is the best one in the models without interaction terms.

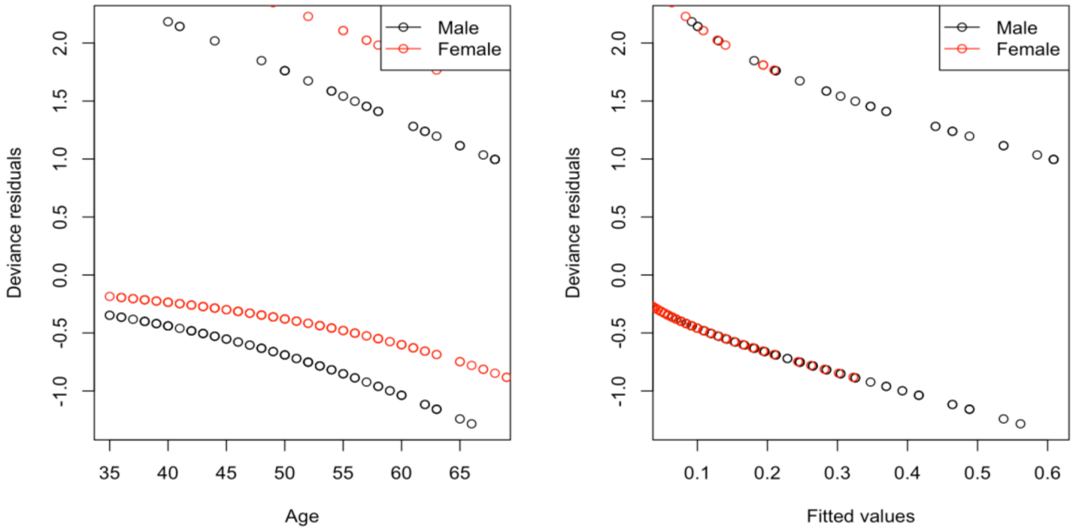

Figure 6: Deviance residuals plotted against Age and fitted values

Conclusion: The above analysis suggest that the model with terms **Age** + **Gender** is the best one to describe the data we have. In case of no interaction term, the terms **Age** and **Gender** have significant effect on the model with all explanatory variables. Then the analysis of deviance and the comparing of AIC values indicate that the model with terms **Age** + **Gender** has a reasonably best goodness-of-fit and a lowest AIC value in the models removed not significant terms. And the p-values of comparing it with the **Age** + **Totchol** + **Glucose** +**Gender** model and the maximal model show that removing the terms **Totchol** and **Glucose** does not results a significant reduction in the goodness-of-fit. The results compared with the maximal model and the null model also prove that the model fits well. However, the plot of deviance residuals shows some disadvantages of this model, such as the linear relationship between the deviance residuals and **Age**, and the influence of **Gender**. We think it may be caused by characteristics of the data but it still indicates that the model fit exist some unknown problems to improve.

Further discussion: We did not consider the interaction term in the model which may results deficiency. The explanatory variable **Gender** is a categorical variable so that we are supposed to do some analysis of models with interaction terms to improve the model's universality. And in real life, we usually think that the total cholesterol level and the glucose level affect our physical health but we rejected them in this model analysis. We want more data of **Totchol**, **Glucose** in order to increase the reliability of the fitted model. Besides, we can also add some transformation into the model to solve the problems shown by the plots of deviance residuals.

(d) We prefer to use the generalized linear model with age of the patient and gender as the explanatory variables to estimate the response variable risk for Michael. We set the age equals to 83 and gender is male to get his risk of 10-year coronary heart disease is 0.8703279.

Comment: The estimated values of the risk is not reliable and slightly insufficient. Firstly, one possible problem is that Micheal is 83 years old but the maximum data of **Age** is 69. This model may be not suitable for predicting risk out of the range of the data we have. Considering the impact of **Glucose** and **Totchol** may lead to improvement of the estimation. As the further discussion said in the part (c), the fitted model well described the impact of **Age** and **Gender** on **Risk** but ignored the impact of **Glucose** and **Totchol**. Besides, interaction terms and more data may possibly improve the accuracy of estimates as well but we need to do more analysis about it.