Question 1

(a) Exploratory Data Analysis (**10 marks**)

The IHD data set contains 788 observations of 9 variables, the ID variable was removed because it is not relevant for modelling purposes. The population under consideration consists of all members of the insurance company who made claims relating to ischemic heart disease from January 1998 to December 1999. The data set is not the result of a designed experiment, and so is considered to be observational. Hence, any relationships identified between the response and predictors cannot be considered causal.

Table 1: IHD Quantitative Variables

Variable	Mean	SD	Min	Q1	Median	Q3	Max
Cost	2799.956	6690.26	0	161.125	507.2	1905.45	52664.9
Age	58.718	6.754	24	55	60	64	70
Interventions	4.707	5.595	0	1	3	6	47
Drugs	0.447	1.064	0	0	0	0	9
EDVisit	3.425	2.637	0	2	3	5	20
Complications	0.057	0.248	0	0	0	0	3
Comorbidities	3.766	5.951	0	0	1	5	60
Duration	164.03	120.916	0	41.75	165.5	281	372

Gender is the single qualitative variable, with 608 observations (77.2%) for Female and 180 observations (22.8%) for Male. It is apparent from the table that many of the quantitative variables have positively skewed distributions, particularly **Cost, Interventions, Drugs, EDVisit** and **Comorbidities**. The response variable in this question is **EDVisit**.

The pairs plot (Figure 1) gives a visualisation of the relationships between the quantitative variables in the IHD data. The positively skewed distributions mentioned earlier are apparent. Notably, however, **Age** is negatively skewed.

EDVisit has a moderate positive correlation with **Drugs** (r = 0.53), and has a weak positive correlation with **Cost** and **Interventions** (r = 0.38 and r = 0.37 respectively).

There is a strong and positive relationship between **Interventions** and **Cost** ($r = 0.73$) which makes sense as interventions cost money so the patients may have higher claims from the insurance. There are also many weak correlations amongst the predictors.

Note: If you separate Male vs Female in the pairs plot, it can be very crowded. So you should always make sure that the plot is readable.

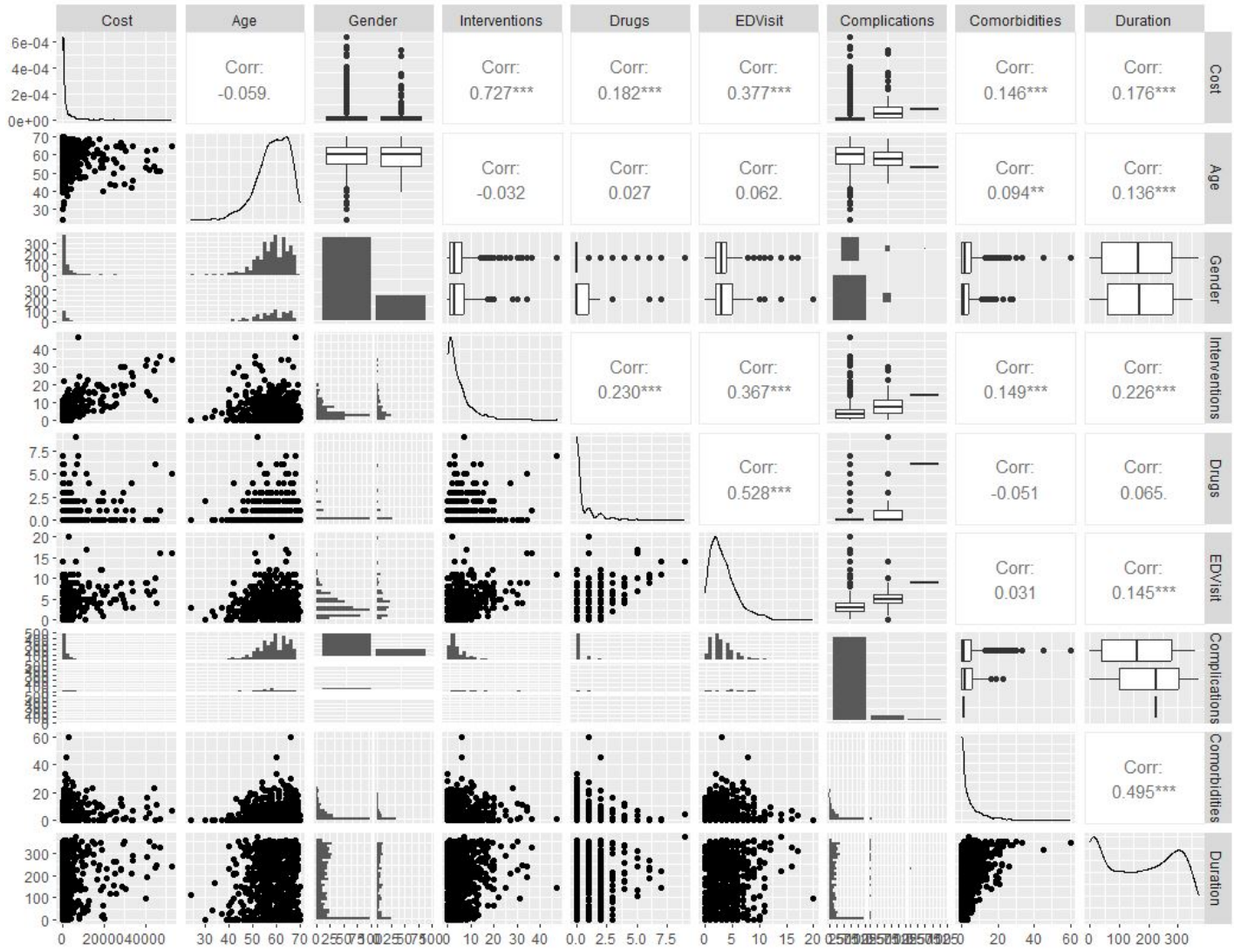

Figure 1: Pairs plot for the IHD dataset

Figure 2: Correlation plot for IHD data

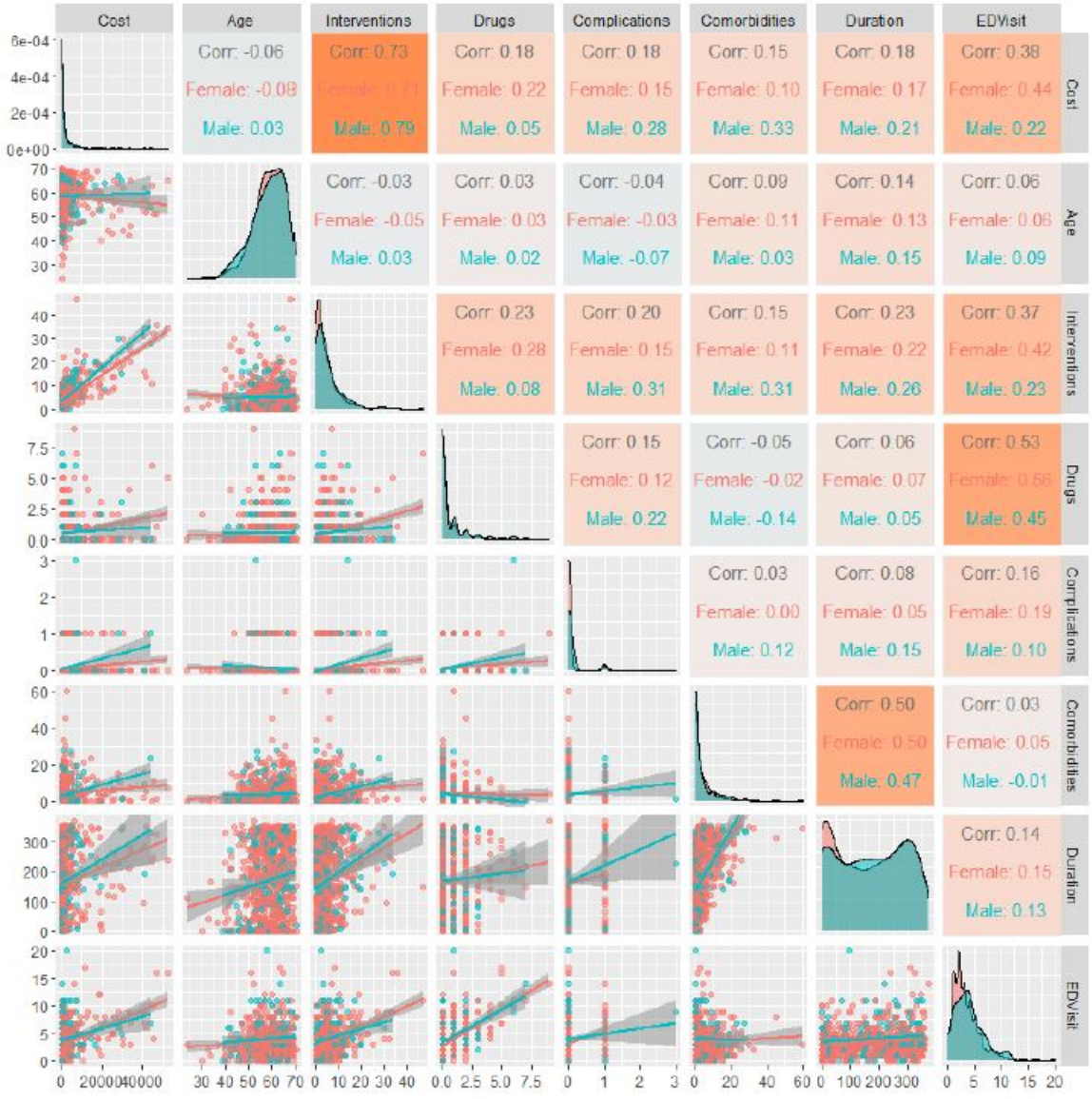

Figure 3: Pairs plot for the IHD dataset, separated by Gender

(b) Model for EDVisit (**30 marks**)

The response variable is **EDVisit**. As we are dealing with counts, these data should be modeled using a GLM with a Poisson error distribution. The first model includes all the available predictors.

```
modP <- glm(EDVisit ~ . - ID, family=poisson, data = IHD)
summary(modP)
```

The residual deviance for the model is 1039 on 778 degrees of freedom, resulting in an extremely low χ^2 value (9.2x 10^{-10}), indicating that the model suffers from overdispersion. In an attempt to overcome the issue of overdispersion, the model was refitted assuming a negative binomial distribution, using the glm.nb command in R:

```
mod.nb1 <- glm.nb(EDVisit ~ . - ID, data = IHD)
```

This model reduces the deviance to 819, with a χ^2 value of 0.14. Since 0.14 is considerably above the 0.05 threshold, the model represents a good fit to the data, although there is still a large amount of unexplained variation.

Three further models were fitted using the negative binomial distribution, all with EDVisit as the response. The predictors for each model are summarised in Table 1.

Table 1: Negative binomial models for EDVisit

Model	Description
mod.nb1	glm.nb(EDVisit $\sim$. - ID, data = HD)
mod.nb2	glm.nb(EDVisit $\sim$ Cost + Age+Gender+Interventions+Drugs, data = IHD)
mod.nb3	glm.nb(EDVisit $\sim$ Cost+Interventions + Gender+Drugs, data = IHD)
mod.nb4	glm.nb(EDVisit $\sim$ Cost*Interventions + Age+Gender+Drugs, data = IHD)

The models were compared using AICc and the IT approach, giving the results in Table 2.

From this we can conclude that **mod.nb4**, the model with **Age, Gender, Drugs, Cost, Interventions** and the **interaction between Cost and Interventions** is the preferred model, with an Akaike weight of 0.65.

Table 2: Model selection results for EDVisit analysis

```
Model selection based on AICc:

         K AICc Delta_AICc AICcWt Cum.Wt    LL
mod.nb4  8 3235       0.00   0.65   0.65 -1609
mod.nb2  7 3237       2.22   0.21   0.86 -1611
mod.nb1 11 3239       3.68   0.10   0.96 -1608
mod.nb3  6 3240       5.60   0.04   1.00 -1614
```

We should check the residual deviance for the preferred model. The residual deviance is 815.27 on 781 df, which is not significant (p=0.19). The model fit appears adequate.

From the summary table (Table 3), the final model is:

$$\begin{aligned} EDVisit_i &\sim Possion(\mu_i) \\ \log_e(\mu_i) &= \beta_0 + \beta_1 Cost + \beta_2 Interventions + \beta_3 Age + \beta_4 Male \\ &\quad + \beta_5 Drugs + \beta_6\ Cost : Interventions \\ &= 0.45 + 2.69 \times 10^{-5} Cost + 0.016 Interventions + 0.0076 Age + 0.192 Male \\ &\quad + 0.21 Drugs - 5.38 \times 10^{-7} \beta_6\ Cost : Interventions \end{aligned}$$

Here μ is the mean number of EDVisit.

The interaction term **Cost:Interventions** is significant (p-value = 0.038), so we don't interpret the coefficients for Cost ($\beta_1 = 2.69 \times 10^{-5}$) and Interventions ($\beta_2 = 0.016$). *The coefficient for the interaction term is negative, -5.38×10^{-7}, suggesting that the higher the Cost, the lesser the effects of Interventions on the number of EDVisit.*

The expected number of EDVisit differs significantly between Male and Female, given the same values for Cost, Interventions and Drugs (p-value = 0.00017). The expected EDVisit is 21% higher in Male compared to the Female group ($\beta_4 = 0.192$, $exp(0.192) = 1.212$).

For each additional drug prescribed, the expected number of EDVisit increases by 23.4% ($\beta_5 = 0.21$, $exp(0.21) = 1.234$).

Table 3: Table of coefficients for the preferred model

```
Coefficients:
                    Estimate Std. Error  z value Pr(>|z|)
(Intercept)         4.48e-01   2.01e-01    2.23  0.02598
Cost                2.69e-05   6.14e-06    4.37  1.2e-05
Interventions       1.62e-02   5.24e-03    3.09  0.00201
Age                 7.63e-03   3.37e-03    2.27  0.02343
Gender1             1.92e-01   5.10e-02    3.77  0.00017
Drugs               2.10e-01   1.63e-02   12.93  < 2e-16
Cost:Interventions -5.38e-07   2.60e-07   -2.07  0.03855
```

(c) Predicted Values for EDVisit Versus Actual Values (**5 marks**)

The plot of Predicted versus Actual values for EDVisit (Figure 4) demonstrates that the model for predicting EDVisit using **Cost, Interventions, Age, Gender, Drugs and the interaction term Cost:Intervention** maybe useful for some purpose, however there is a great variability in the predictions.

The variance in the prediction errors is also increasing as the actual values increase. This produces the fan-shaped distribution shown in the plot. Further investigation is required to improve the model accuracy.

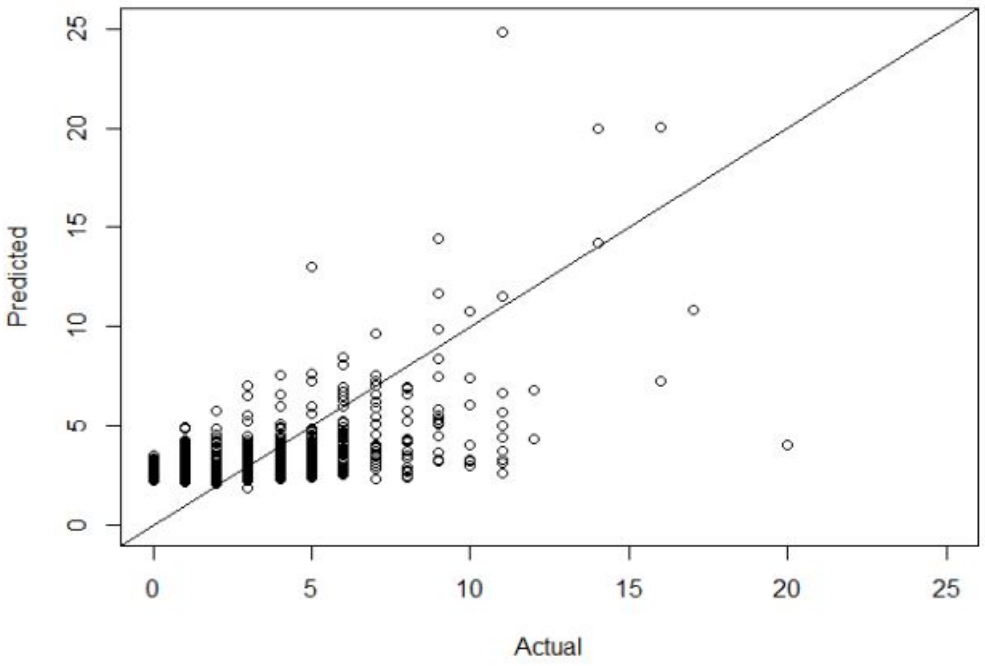

Figure 4: EDVisit Predicted Vs Actual Values using the final model

Question 2 (**45 marks**)

For this question, a new binary variable, FreqV, was created

- FreqV = 1 if EDVisit ≥ 4
- FreqV = 0 otherwise

Three models were developed and compared on the basis of their sensitivity (true positive rate, TPR), specificity (true negative rate, TNR) and overall correction rate. *Note: you can also use the IT approach to compare these models as well.*

- Model 1: glm(FreqV $\sim$. - ID - EDVisit, family = binomial)
- Model 2: glm(FreqV $\sim$ Cost + I($Cost^2$) + Age + Gender + Drugs, family = binomial)
- Model 3: glm(FreqV $\sim$ sqrt(Cost) + sqrt(Drugs) + Gender, family = binomial)

The prediction results for the three models are listed in Table 4. All three models show very similar prediction accuracy.

(*Note: In practice, you should use cross validation technique (STAT330/430) i.e. splitting the data into training and test sets. You build the models using the train test, then evaluate the models using the test set.*)

Table 4: Comparison of prediction results for the three models

Model	TPR	TNR	Overall correction rate
Model 1	0.48	0.90	0.730
Model 2	0.50	0.89	0.736
Model 3	0.52	0.88	0.735

You can aslo compare these models using the IT approach (Table 5). According to the AICc, model 2 was ranked first, followed by model 3, and lastly model 1 (Table 5). Model 2 has an AICc weight of 0.65, which means it has a 65% chance of being the best model out of these three models tested. So Model 2 is chosen as the final model.

(*Note: However, delta AICc is less than 2 for model 3, and there is not much difference in prediction accuracy between model 2 and model 3. So you can also select Model 3 is the preferable model.*)

Table 5: Model selection results for FreqV analysis

```
Model selection based on AICc:

        K  AICc Delta_AICc AICcWt Cum.Wt     LL
model2  6 858.8       0.00   0.65   0.65 -423.3
model3  4 860.5       1.69   0.28   0.93 -426.2
model1 11 863.3       4.48   0.07   1.00 -420.5
```

The summary of the coefficients for Model 2 are listed in Table 6.

Table 6: Model selection results for FreqV analysis

```
Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -3.02e+00   7.81e-01   -3.87  0.00011
Cost         2.28e-04   4.01e-05    5.69  1.3e-08
I(Cost^2)   -3.57e-09   1.20e-09   -2.97  0.00301
Age          2.76e-02   1.29e-02    2.13  0.03319
Gender1      6.18e-01   1.93e-01    3.20  0.00136
Drugs        1.05e+00   1.31e-01    8.04  8.8e-16
```

Residual deviance of the final model is 846.68 on 782 degrees of freedom gives a non-significant p-value of 0.054. This means that the model provides an adequate fit for this dataset. However, this p-value is closed to the threshold. The predictions for this model are unlikely to be very accurate, so this model may be useful for some cases, but should probably not be relied upon. The ROC Curve plot (Figure 5) shows that 3 models for predicting the chances of ED visits being greater or equal to 4 is not too bad, but improvement is needed to increase the model sensitivity.

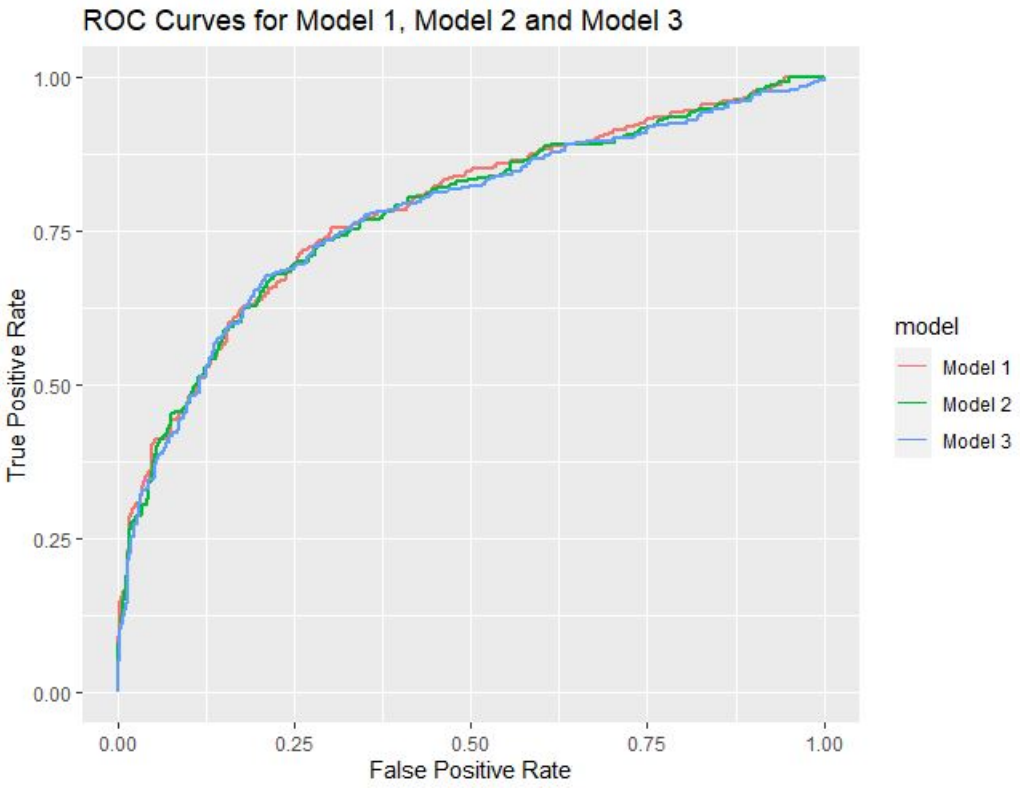

Figure 5: The ROC Curve of 3 models tested for FreqV

The Final model

The linear predictor is:

$$\begin{aligned} log(\frac{p}{1-p}) = \hat{\eta} &= -3.02 + 2.28 \times 10^{-4} Cost - 3.57 \times 10^{-9} Cost^2 \\ &\quad +0.027 Age + 0.62 Male + 1.05 Drugs \end{aligned}$$

Therefore, for a female with Cost and Drugs held the same, for every additional year in age, the odds of ED visits greater or equal to 4 increases by 3%.

Also, for a patient with Cost, Age and Drugs held the same, the odds of ED visits greater or equal to 4 increases by 86% if the patient is a male.

An example of prediction rule develop from the fitted equation.

50 year old female patients with a cost claim of 2800 are likely to have or more 4 visits to the emergency department if they have at least one drug prescribed. (fixed Age = 50, Cost=2800, solve for Drugs).