Heart Disease data -

This dataset contains information on the number of heart disease-related deaths in a sample of doctors, classified by age group and smoking status (smoker/non-smoker). The dataset is structured as follows:

- **Column 1: Observation number** (Variable V1 = 1 to 10)
- **Column 2: Age** (Variable V2 = 1, 2, 3, 4, 5)
  Age groups are defined as follows:
  - 1 = 35-44 years
  - 2 = 45-54 years
  - 3 = 55-64 years
  - 4 = 65-74 years
  - 5 = 75-84 years
- **Column 3: Deaths** (Variable V3 = Number of Deaths)
  This is a count variable representing the number of deaths.
- **Column 4: Person-Years at Risk** (Variable V4 = Aggregate years)
  This is a count variable representing the total person-years at risk for each observation.
- **Column 5: Smoker** (Variable V5 = 1, 2)
  This is a factor variable where:
  - 1 = Smoker
  - 2 = Non-Smoker

The AIC values for the models without interaction terms are presented in the table.

Model	AIC Value
Deaths ~ 1	701.34
Deaths ~ Smoker	277.13
Deaths ~ Smoker + Age	218.66
Deaths ~ Smoker + Age + Age^2	77.72
Deaths ~ Smoker + Age + Age^2 + log(`Person-Years`)	77.11
Deaths ~ Age + Age^2 + log(`Person-Years`)	75.81
Deaths ~ Smoker + Age + Age^2 + log(`Person-Years`) + Smoker*Age	68.52

The model with the lowest AIC value, excluding interaction terms, was Deaths ~ Age + Age^2 + log('Person-Years'). In this model, Age^2 represents the squared term of Age, capturing the potential non-linear relationship between age and heart disease mortality, where the effect of age on deaths may accelerate in older age groups. Including the quadratic term allows the model to account for this non-linearity without adding unnecessary complexity. log('Person-Years') refers to the logarithm of Person-Years at risk. No offset was included. It is important to note that, based on this model, the variable Smoker was not statistically significant.

When considering all the interaction terms, the model with least AIC value is shown in **Figure A**.

```
Step:  AIC=68.52
Deaths ~ Smoker + Age + Age2 + log(`Person-Years`) + Smoker:Age

                          Df Deviance    AIC
<none>                         1.4517 68.520
- log(`Person-Years`)      1   3.8239 68.892
+ Age:log(`Person-Years`)  1   1.2054 70.273
+ Age:Age2                 1   1.4509 70.519
- Smoker:Age               1  12.0462 77.114
- Age2                     1  15.0483 80.116
```

Figure A: Interaction model with AIC

As depicted in **Figure A**, interaction term between Smoker and Age is statistically significant. For non-smokers, the interaction of Smoker and Age is 0, while for smokers, it equals the Age variable itself. This suggests that the rate of deaths increases at a differential rate for smokers compared to non-smokers. The reason why Age^2 is significant is that it accounts for the non-linearity in the rate of increase in deaths as age progresses. The AIC value of this model is 68.52, which is lower than the model without interaction terms (75.81). So, this model is considered the best fit for the current data. The scale parameter ϕ is assumed 1.

Figure B provides additional insights on the model. As the summary describes, log('Person-Years') is not statistically significant compared to the other predictor variables. However, it is important to examine diagnostic plots before drawing any final conclusions. The deviance for the fitted model is 1.4517 with 4 degrees of freedom.

```
Call:
glm(formula = Deaths ~ Smoker + Age + Age2 + Smoker * Age + log(`Person-Years`),
    family = poisson)

Deviance Residuals:
       1         2         3         4         5         6         7         8         9
 0.32165  -0.30774   0.03825   0.09349  -0.04309  -0.68101   0.20454   0.56966  -0.61900
      10
 0.16801

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -6.96620    5.57962  -1.249 0.211844
Smoker2            -1.77851    0.87658  -2.029 0.042466 *
Age                 2.07973    0.18322  11.351  < 2e-16 ***
Age2               -0.22104    0.06116  -3.614 0.000302 ***
log(`Person-Years`)  0.78366    0.50565   1.550 0.121193
Smoker2:Age         0.31294    0.09935   3.150 0.001633 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 644.2690  on 9  degrees of freedom
Residual deviance:   1.4517  on 4  degrees of freedom
AIC: 68.52

Number of Fisher Scoring iterations: 4
```

Figure B: Summary of interaction term model

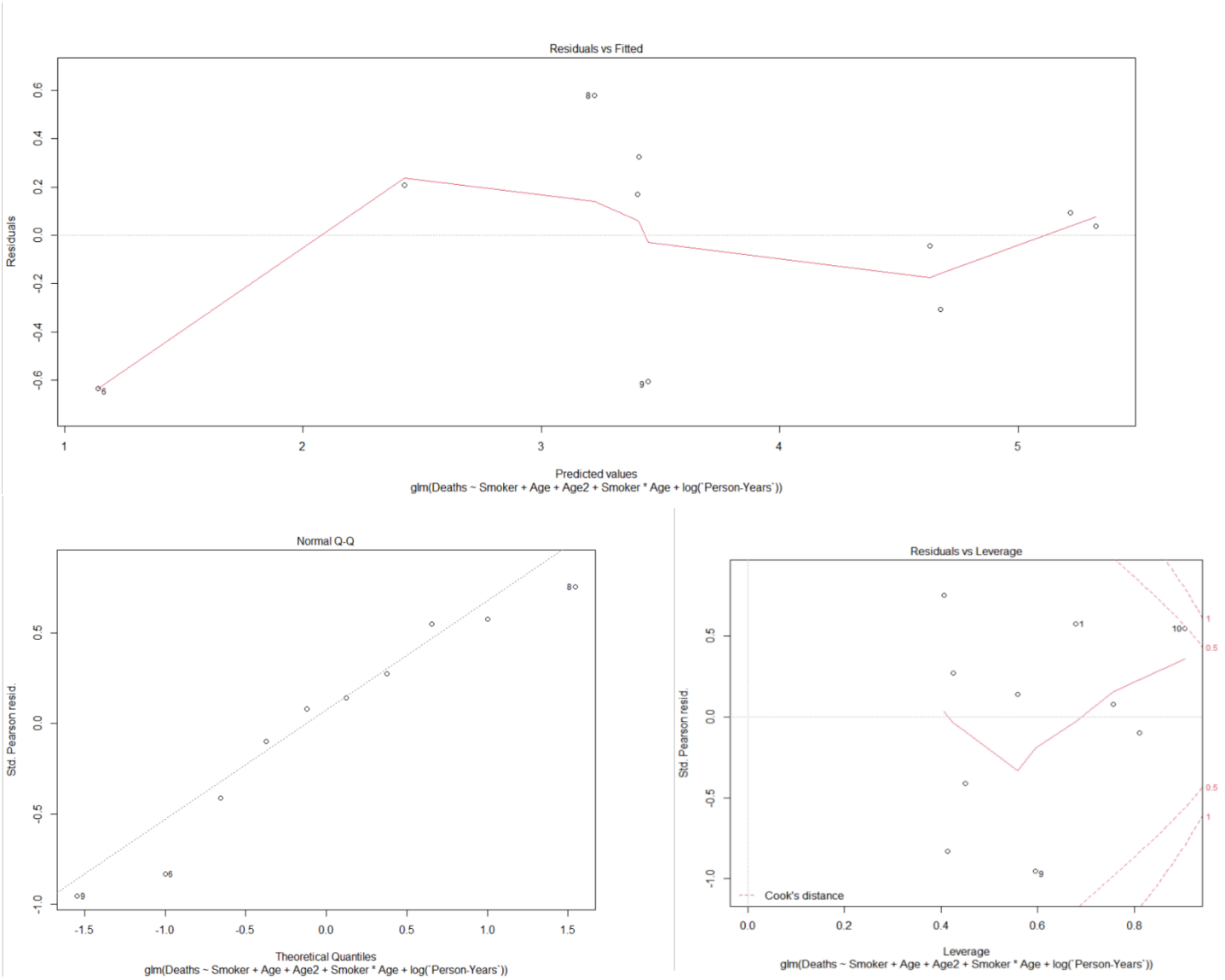

Figure C: Diagnostic plots

Diagnostic plots of Figure C provide additional insight.

- In the Residuals vs Fitted plot, there is only a minor variation in the residuals. Ideally, the deviance residuals should not display any patterns or non-constant variance. Given the relatively constant variation observed, there are no strong indications of non-linearity or heteroscedasticity, suggesting that the model fits well.
- The Normal Q-Q plot shows that most of the observations lie close to the straight line, which suggests that the residuals follow a normal distribution. This indicates that the normality assumption is reasonable, and the model errors are normally distributed.
- In the Residuals vs Leverage plot, we observe that observation 10 has a relatively high leverage value, indicating that this point has a significant influence on the model. While this does not necessarily mean that observation 10 is problematic, it may warrant further investigation to determine whether it disproportionately impacts the model's fit.

Focusing on the model selected, let's assess whether it is better to treat log(`Person Years at Risk`) as an offset or as a covariate. When log(`Person Years at Risk`) is treated as an offset, we obtain the model shown in **Figure D**.

```
Call:
glm(formula = Deaths ~ Smoker + Age + Age2 + Smoker * Age, family = poisson,
    offset = log(`Person-Years`))

Deviance Residuals:
       1         2         3         4         5         6         7         8         9        10
 0.43820  -0.27329  -0.15265   0.23393  -0.05700  -0.83049   0.13404   0.64107  -0.41058  -0.01275

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.35079    0.28764 -32.509  < 2e-16 ***
Smoker2     -1.44097    0.37220  -3.872 0.000108 ***
Age          2.06893    0.18170  11.386  < 2e-16 ***
Age2        -0.19768    0.02737  -7.223 5.08e-13 ***
Smoker2:Age  0.30755    0.09704   3.169 0.001528 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 935.0673  on 9  degrees of freedom
Residual deviance:   1.6354  on 5  degrees of freedom
AIC: 66.703

Number of Fisher Scoring iterations: 4
```

Figure D – Offset model

The residual deviance of this offset model is 1.6354 with 5 degrees of freedom. In contrast, the residual deviance of the model treating **log(Person Years)** as a covariate was 1.4517 with 4 degrees of freedom.

Let:

1. M_0 denote the model with log(`Person Years at Risk`) included as offset
2. M_1 denote the model with log(`Person Years at Risk`) as covariate.

The deviance values for the two models are:

1. $D(M_0) = 1.6354$ on 5 degree of freedom
2. $D(M_1) = 1.4517$ on 4 degree of freedom

The difference in deviance between these two models is: 1.6354 - 1.4517 = 0.1837. This difference is on 1 degree of freedom.

Let α denote the coefficient of log(`Person-Years at risk`) and let β denote parameter containing the free parameter in model M_0.

We now test the following hypotheses:

H_0: $\alpha = 1$ and β is unrestricted.

H_a: α,β are both unrestricted.

To test this, we refer 0.1837 to a χ_1^2 (chi-squared with one degree of freedom). The resulting p-value is **0.6682**. As the p-value is large, we do not have enough evidence to reject null hypothesis.

Therefore, log(`Person-Years`) should be treated as an offset rather than a covariate.

Now, let's consider **negative binomial regression models** to account for any potential overdispersion in the data.

The AIC value for the negative binomial regression model without interaction terms, where log(`Person-Years`) is treated as a covariate, is 77.81. This is identical to the AIC value of the corresponding Poisson regression model (77.81), suggesting that both models fit the data similarly in the absence of overdispersion.

The best-fitting model, including interaction terms, is the same as in the Poisson regression case:

Deaths ~ Smoker + Age + Age2 + Smoker * Age + log(`Person-Years`)

The summary of the negative binomial model is presented in Figure E. The AIC value for this model is 70.547, which is only slightly higher than the AIC for the Poisson model (68.52).

```
Call:
glm.nb(formula = Deaths ~ Smoker + Age + Age2 + Smoker * Age +
    log(`Person-Years`), control = glm.control(maxit = 500),
    init.theta = 1300096560000, link = log)

Deviance Residuals:
       1         2         3         4         5         6         7         8         9        10
 0.32170  -0.30769   0.03981   0.09512  -0.04146  -0.68106   0.20513   0.56947  -0.61877   0.16772

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -6.96620    5.57962  -1.249 0.211844
Smoker2            -1.77851    0.87658  -2.029 0.042466 *
Age                 2.07973    0.18322  11.351  < 2e-16 ***
Age2               -0.22104    0.06116  -3.614 0.000302 ***
log(`Person-Years`)  0.78366    0.50565   1.550 0.121193
Smoker2:Age         0.31294    0.09935   3.150 0.001633 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.300097e+12) family taken to be 1)

    Null deviance: 644.2710  on 9  degrees of freedom
Residual deviance:   1.4517  on 4  degrees of freedom
AIC: 70.547

Number of Fisher Scoring iterations: 1
```

Figure E: Negative binomial regression

Upon comparing the summary statistics (i.e., coefficient estimates, standard errors, z-values, and p-values) between the Poisson and negative binomial models (**Figures B and E**), we observe that they are very similar. This indicates that the two models yield comparable parameter estimates.

The key difference lies in the handling of overdispersion. While negative binomial regression is often used to account for overdispersion, it seems that overdispersion was not a significant issue in the Poisson model, as reflected by the minimal differences in AIC values between the two models.

There are slight differences in the deviance residuals between the Poisson and negative binomial models, which could explain the small difference in AIC values. However, the residual deviance for both models rounds to **1.4517**, indicating a good fit for both approaches.

Now, when considering **log(Person-Years)** as an offset for the negative binomial regression model, we obtain a model with an AIC value of **396**, which is significantly higher than **70.547** (the AIC of the previous model without the offset). The likely reason for this large AIC is the extremely high starting theta value. Despite the difference in AIC, the residual deviance remains like the Poisson regression model at **1.6354**.

Based on the previous deviance test for the Poisson model, we can still consider **log(Person-Years)** as an offset rather than a covariate. Notably, the coefficient estimates, standard errors, z-values, and p-values for the negative binomial model are nearly identical to those in the Poisson regression model with the offset. Comparing **Figure F** (negative binomial with offset) and **Figure D** (Poisson with offset), we observe these similarities, with the main difference being in the deviance residuals. In the negative binomial model, most of the deviance residuals for individual observations are **0.000**, which is not the case in the Poisson offset model.

```
glm.nb(formula = Deaths ~ Smoker + Age + Age2 + Smoker * Age +
    offset(log(`Person-Years`)), control = glm.control(maxit = 500),
    init.theta = 9.228881326e+15, link = log)

Deviance Residuals:
      1       2       3       4       5       6       7       8       9      10
 0.9613  0.0000  0.0000  0.0000  0.0000 -1.3995  0.9668  0.0000  0.0000 -1.9891

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.35079    0.28764 -32.509  < 2e-16 ***
Smoker2     -1.44097    0.37220  -3.872 0.000108 ***
Age          2.06893    0.18170  11.386  < 2e-16 ***
Age2        -0.19768    0.02737  -7.223 5.08e-13 ***
Smoker2:Age  0.30755    0.09704   3.169 0.001528 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(9.228881e+15) family taken to be 1)

    Null deviance: 941.2150  on 9  degrees of freedom
Residual deviance:   1.6354  on 5  degrees of freedom
AIC: 396

Number of Fisher Scoring iterations: 1


              Theta:  9.228881e+15
          Std. Err.:  5.09117e+14

 2 x log-likelihood:  -384
```

Figure F: Negative binomial model with offset

From a broader perspective, as seen by comparing **Figures B, E, D, and F**, we can conclude that negative binomial models share a lot in common with their corresponding Poisson regression models, particularly in terms of coefficient estimates, standard errors, z-values, and residual deviance. However, the primary differences lie in the AIC values and the deviance residuals of individual observations. This may suggest that overdispersion is not an issue in the Poisson models, which explains why the residual deviance is nearly the same between the two modeling approaches.

Additionally, the large theta values in both negative binomial models push the variance function closer to **μ**, which is characteristic of the Poisson general linear model. Negative binomial models are typically preferred when there is clear evidence of overdispersion, but that does not seem to be the case in this dataset, as indicated by the residual deviance similarity across models.\

But until now, we assume that scale parameter $\phi = 1$.

Let's use Wald test.

The scale parameter ϕ can be estimated by phihat.

Here,

$$\hat{\phi} = X_P^2 = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

The variance function for the Poisson model, the $V(\hat{\mu}_i)$ = μ

The null and alternative hypotheses for the Wald test are:

H_0: $\phi = 1$

H_1: Scale parameter is not 1

From the model we estimate,

$\hat{\phi}$ = 0.232

We know that standard error of phihat is,

$$SE_{\hat{\phi}} = \frac{1}{n-p}\sqrt{\sum_{i=1}^{n}(\hat{A}_i - \hat{\phi})^2}$$

From that we get,

Standard error = 0.1

So, coefficient of scale parameter is 0.232 with se=0.1, so the z-value for testing the null hypothesis that the coefficient is 1 is

= (0.232 – 1) / 0.1 = -7.68

Given this z-score, we calculate the p-value, which is approximately 8.016762e-15 , a very small value that is highly significant.

Since the p-value is much smaller than the 0.05 threshold, we reject the null hypothesis that ϕ=1.

This indicates that the scale parameter significantly differs from 1, suggesting overdispersion in the model.

For reference,

```
fv = fitted.values(model2.glm)
V1 = ddata$Deaths
phihat = sum((V1-fv)*(V1-fv)/(fv))/6
se=sqrt(sum(((V1-fv)*(V1-fv)/(fv)-phihat)**2))/6
pnorm(-7.679)
```