Homework 3

Sarah Adilijiang

Problem 1

(a)

```
library(MASS)
data(quine)
```

The GLM model for Poisson (count) response here is:
likelihood: $P(Days_i = y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y!},\ y_i = 0, 1, 2, ...$
linear predictor: $\eta_i = \beta_0 + \beta_1 Eth_i + \beta_2 Sex_i + (\beta_3\, to\, \beta_5) Age_i + \beta_6 Lrn_i + \epsilon_i$
link function (logit): $\eta_i = \log \mu_i$

```
# Poisson model (without dispersion)
m.glm = glm(Days~., quine, family=poisson)
summary(m.glm)
```

```
##
## Call:
## glm(formula = Days ~ ., family = poisson, data = quine)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -6.808  -3.065  -1.119   1.818   9.909
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.71538    0.06468  41.980  < 2e-16 ***
## EthN        -0.53360    0.04188 -12.740  < 2e-16 ***
## SexM         0.16160    0.04253   3.799 0.000145 ***
## AgeF1       -0.33390    0.07009  -4.764 1.90e-06 ***
## AgeF2        0.25783    0.06242   4.131 3.62e-05 ***
## AgeF3        0.42769    0.06769   6.319 2.64e-10 ***
## LrnSL        0.34894    0.05204   6.705 2.02e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2073.5  on 145  degrees of freedom
## Residual deviance: 1696.7  on 139  degrees of freedom
## AIC: 2299.2
##
## Number of Fisher Scoring iterations: 5
```

```
# quasi-Poisson model (with dispersion)
m.glm_q = glm(Days~., quine, family=quasipoisson)
summary(m.glm_q)
```

```
##
## Call:
```

```
## glm(formula = Days ~ ., family = quasipoisson, data = quine)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -6.808  -3.065  -1.119   1.818   9.909
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7154     0.2347  11.569  < 2e-16 ***
## EthN         -0.5336     0.1520  -3.511 0.000602 ***
## SexM          0.1616     0.1543   1.047 0.296914
## AgeF1        -0.3339     0.2543  -1.313 0.191413
## AgeF2         0.2578     0.2265   1.138 0.256938
## AgeF3         0.4277     0.2456   1.741 0.083831 .
## LrnSL         0.3489     0.1888   1.848 0.066760 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 13.16691)
##
##     Null deviance: 2073.5  on 145  degrees of freedom
## Residual deviance: 1696.7  on 139  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

Answer:

In the summary of Poisson model, the residual deviance (1696.7) is much larger than the degree of freedom of the model (139), i.e. D >> n-p. And the quasi-Poisson model automatically computes the dispersion parameter, which is 13.16691 >> 1. So the quasi-Poisson model is preferred in this case for the main effects model.

Then let's use the quasi-Poisson model to compare the main effects model with models with two-way interactions. Here F-tests are used for comparing between models and dropping insignificant interaction terms due to dispersion problem.

The linear predictor in the model with all two-way interaction terms is:
linear predictor: $\eta_i = \beta_0 + \beta_1 Eth_i + \beta_2 Sex_i + (\beta_3 \, to \, \beta_5) Age_i + \beta_6 Lrn_i + \beta_7 Eth_i : Sex_i + (\beta_8 \, to \, \beta_{10}) Eth_i : Age_i + \beta_{11} Eth_i : Lrn_i + (\beta_{12} \, to \, \beta_{14}) Sex_i : Age_i + \beta_{15} Sex_i : Lrn_i + (\beta_{16}, \beta_{17}) Age_i : Lrn_i + \epsilon_i$

```
# quasi-Poisson model with all two-way interactions
m.glm_q2 = glm(Days~(Eth+Sex+Age+Lrn)**2, quine, family=quasipoisson)
summary(m.glm_q2)
```

```
##
## Call:
## glm(formula = Days ~ (Eth + Sex + Age + Lrn)^2, family = quasipoisson,
##     data = quine)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.6533  -2.7796  -0.5301   1.5749   8.1955
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  2.93246    0.32178   9.113  1.4e-15 ***
## EthN        -0.17399    0.39733  -0.438  0.66219
## SexM        -0.71452    0.40047  -1.784  0.07676 .
## AgeF1       -0.04270    0.41558  -0.103  0.91833
## AgeF2       -0.08632    0.52931  -0.163  0.87071
## AgeF3       -0.15290    0.38960  -0.392  0.69538
## LrnSL        0.21608    0.47672   0.453  0.65112
## EthN:SexM    0.43902    0.30152   1.456  0.14783
## EthN:AgeF1  -0.92889    0.47997  -1.935  0.05516 .
## EthN:AgeF2  -1.33398    0.44220  -3.017  0.00308 **
## EthN:AgeF3  -0.11242    0.44137  -0.255  0.79935
## EthN:LrnSL   0.26415    0.37260   0.709  0.47965
## SexM:AgeF1  -0.05565    0.53386  -0.104  0.91714
## SexM:AgeF2   1.09942    0.50040   2.197  0.02981 *
## SexM:AgeF3   1.15949    0.45383   2.555  0.01179 *
## SexM:LrnSL   0.04143    0.44920   0.092  0.92666
## AgeF1:LrnSL -0.13019    0.51372  -0.253  0.80035
## AgeF2:LrnSL  0.37340    0.47688   0.783  0.43507
## AgeF3:LrnSL       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 10.72308)
##
##     Null deviance: 2073.5  on 145  degrees of freedom
## Residual deviance: 1368.7  on 128  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
# drop one interaction term from the quasi-Poisson model with two-way interactions
drop1(m.glm_q2, test="F")
```

```
## Single term deletions
##
## Model:
## Days ~ (Eth + Sex + Age + Lrn)^2
##         Df Deviance F value   Pr(>F)
## <none>       1368.7
## Eth:Sex  1   1391.6  2.1432 0.145658
## Eth:Age  3   1497.4  4.0144 0.009077 **
## Eth:Lrn  1   1374.1  0.5105 0.476204
## Sex:Age  3   1518.2  4.6626 0.003982 **
## Sex:Lrn  1   1368.8  0.0085 0.926594
## Age:Lrn  2   1380.3  0.5430 0.582359
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

The interaction term "Sex:Lrn" has a p-value = 0.926594, which is the most insignificant term in the model, so we drop this interaction term first and then continue to drop the next one.

```
# fit a new model removing the "Sex:Lrn" term
m.glm_q3 = glm(Days~Eth+Sex+Age+Lrn+Eth:Sex+Eth:Age+Eth:Lrn+Sex:Age+Age:Lrn,
               quine, family=quasipoisson)
```

```r
# drop a second term from the quasi-Poisson model with two-way interactions
drop1(m.glm_q3, test="F")
```

```
## Single term deletions
##
## Model:
## Days ~ Eth + Sex + Age + Lrn + Eth:Sex + Eth:Age + Eth:Lrn +
##     Sex:Age + Age:Lrn
##         Df Deviance F value   Pr(>F)
## <none>       1368.8
## Eth:Sex  1   1392.0  2.1856 0.141745
## Eth:Age  3   1497.9  4.0584 0.008568 **
## Eth:Lrn  1   1374.4  0.5303 0.467821
## Sex:Age  3   1519.5  4.7365 0.003617 **
## Age:Lrn  2   1387.1  0.8627 0.424431
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

The interaction term "Eth:Lrn" has a p-value = 0.467821, which is the most insignificant term in the model now, so we drop this interaction term and then continue to drop the next one.

```r
# fit a new model removing the "Eth:Lrn" term
m.glm_q4 = glm(Days~Eth+Sex+Age+Lrn+Eth:Sex+Eth:Age+Sex:Age+Age:Lrn,
               quine, family=quasipoisson)

# drop another term
drop1(m.glm_q4, test="F")
```

```
## Single term deletions
##
## Model:
## Days ~ Eth + Sex + Age + Lrn + Eth:Sex + Eth:Age + Sex:Age +
##     Age:Lrn
##         Df Deviance F value   Pr(>F)
## <none>       1374.4
## Eth:Sex  1   1395.2  1.9662 0.163233
## Eth:Age  3   1512.6  4.3572 0.005845 **
## Sex:Age  3   1523.1  4.6899 0.003829 **
## Age:Lrn  2   1390.5  0.7618 0.468891
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

The interaction term of "Age:Lrn" has a p-value = 0.468891, which is the most insignificant term in the model now, so we drop this interaction term and then continue to drop the next one.

```r
# fit a new model removing the "Age:Lrn" term
m.glm_q5 = glm(Days~Eth+Sex+Age+Lrn+Eth:Sex+Eth:Age+Sex:Age,
               quine, family=quasipoisson)

# drop another term
drop1(m.glm_q5, test="F")
```

```
## Single term deletions
```

```
##
## Model:
## Days ~ Eth + Sex + Age + Lrn + Eth:Sex + Eth:Age + Sex:Age
##         Df Deviance F value   Pr(>F)
## <none>       1390.5
## Lrn      1   1460.2  6.6184 0.011199 *
## Eth:Sex  1   1410.0  1.8568 0.175310
## Eth:Age  3   1531.5  4.4611 0.005101 **
## Sex:Age  3   1530.2  4.4218 0.005363 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

The interaction term of "Eth:Sex" has a p-value = 0.175310, which is the most insignificant term in the model now, so we drop this interaction term and then continue to drop the next one.

```
# fit a new model removing the "Eth:Sex" term
m.glm_q6 = glm(Days~Eth+Sex+Age+Lrn+Eth:Age+Sex:Age,
               quine, family=quasipoisson)

# drop another term
drop1(m.glm_q6, test="F")
```

```
## Single term deletions
##
## Model:
## Days ~ Eth + Sex + Age + Lrn + Eth:Age + Sex:Age
##         Df Deviance F value   Pr(>F)
## <none>       1410.0
## Lrn      1   1479.6  6.5566 0.011566 *
## Eth:Age  3   1559.1  4.6853 0.003827 **
## Sex:Age  3   1542.8  4.1748 0.007336 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

Now all the terms are significant in this model, so we stop dropping more interaction terms and keep this model as our final model. Let's have a look at the model summary.

```
# summary of the final quasi-Poisson model
summary(m.glm_q6)
```

```
##
## Call:
## glm(formula = Days ~ Eth + Sex + Age + Lrn + Eth:Age + Sex:Age,
##     family = quasipoisson, data = quine)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.3260  -2.7319  -0.6531   1.6072   8.9348
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.73976    0.30474   8.991 2.15e-15 ***
## EthN         0.15313    0.33133   0.462  0.64472
## SexM        -0.47567    0.33190  -1.433  0.15415
```

```
## AgeF1       -0.14579    0.36491  -0.400  0.69014
## AgeF2        0.18634    0.37402   0.498  0.61916
## AgeF3       -0.08614    0.38979  -0.221  0.82544
## LrnSL        0.45575    0.18145   2.512  0.01321 *
## EthN:AgeF1  -0.97778    0.45115  -2.167  0.03199 *
## EthN:AgeF2  -1.23650    0.42332  -2.921  0.00410 **
## EthN:AgeF3  -0.18097    0.42083  -0.430  0.66786
## SexM:AgeF1  -0.05498    0.49883  -0.110  0.91240
## SexM:AgeF2   0.89849    0.42000   2.139  0.03424 *
## SexM:AgeF3   1.13962    0.42388   2.689  0.00809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 10.88183)
##
##     Null deviance: 2073.5  on 145  degrees of freedom
## Residual deviance: 1410.1  on 133  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
# compare the final model with original models with and without all the two-way interactions
anova(m.glm_q, m.glm_q6, test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: Days ~ Eth + Sex + Age + Lrn
## Model 2: Days ~ Eth + Sex + Age + Lrn + Eth:Age + Sex:Age
##   Resid. Df Resid. Dev Df Deviance      F   Pr(>F)
## 1       139     1696.7
## 2       133     1410.0  6   286.65 4.3904 0.000443 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m.glm_q6, m.glm_q2, test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: Days ~ Eth + Sex + Age + Lrn + Eth:Age + Sex:Age
## Model 2: Days ~ (Eth + Sex + Age + Lrn)^2
##   Resid. Df Resid. Dev Df Deviance      F Pr(>F)
## 1       133     1410.0
## 2       128     1368.7  5   41.386 0.7719 0.5717
```

Answer:

In the summary of final quasi-Poisson model, the dispersion parameter is 10.88183 >> 1. So the quasi-Poisson model is still preferred than the Poisson model for this model with two-way interaction terms.

Comparing this final quasi-Poisson model with the original main effects quasi-Poisson model without two-way interactions, the p-value of F-test is 0.000443, so the larger final model is preferred.

Comparing this final quasi-Poisson model with the quasi-Poisson model with all the two-way interactions, the p-value of F-test is 0.5717, so the smaller final model is preferred.

Therefore, we agree with keeping this final quasi-Poisson model. And the linear predictor in this final model is:
linear predictor: $\eta_i = \beta_0 + \beta_1 Eth_i + \beta_2 Sex_i + (\beta_3 \, to \, \beta_5) Age_i + \beta_6 Lrn_i + (\beta_7 \, to \, \beta_9) Eth_i : Age_i + (\beta_{10} \, to \, \beta_{12}) Sex_i :$

$Age_i + \epsilon_i$

(b)

Suppose $Z \sim Gamma(k, \theta)$, thus Z has the density $f(z; k, \theta) = \frac{1}{\Gamma(k)\theta^k} z^{k-1} e^{-z/\theta} \quad (z \geq 0,\ k > 0,\ \theta > 0)$

Since $Y \sim Poisson(Z)$, so $f_{Y|Z}(y|z) = \frac{e^{-z} z^y}{y!} = \frac{e^{-z} z^y}{\Gamma(y+1)} \quad (y = 0, 1, 2, ...)$

So the joint distribution of Y and Z is :

$$f_{Y,Z}(y, z) = f_{Y|Z}(y|z) \times f_Z(z) = \frac{e^{-z} z^y}{\Gamma(y+1)} \times \frac{1}{\Gamma(k)\theta^k} z^{k-1} e^{-z/\theta} = \frac{1}{\Gamma(k)\Gamma(y+1)\theta^k} z^{(y+k)-1} e^{-z/(\frac{1}{1+1/\theta})}$$

Hence the marginal distribution of Y is:

$$f_Y(y) = \int_0^\infty f_{Y,Z}(y, z)\, dz = \int_0^\infty \frac{1}{\Gamma(k)\Gamma(y+1)\theta^k} z^{(y+k)-1} e^{-z/(\frac{1}{1+1/\theta})}\, dz$$

$$\rightarrow \ = \frac{\Gamma(y+k)(\frac{1}{1+1/\theta})^{y+k}}{\Gamma(k)\Gamma(y+1)\theta^k} \int_0^\infty \frac{1}{\Gamma(y+k)(\frac{1}{1+1/\theta})^{y+k}} z^{(y+k)-1} e^{-z/(\frac{1}{1+1/\theta})}\, dz = \frac{\Gamma(y+k)(\frac{1}{1+1/\theta})^{y+k}}{\Gamma(k)\Gamma(y+1)\theta^k}$$

Therefore, the conditional density of Z given Y is:

$$f_{Z|Y}(z|y) = \frac{f_{Y,Z}(y, z)}{f_Y(y)} = \frac{\frac{1}{\Gamma(k)\Gamma(y+1)\theta^k} z^{(y+k)-1} e^{-z/(\frac{1}{1+1/\theta})}}{\frac{\Gamma(y+k)(\frac{1}{1+1/\theta})^{y+k}}{\Gamma(k)\Gamma(y+1)\theta^k}} = \frac{1}{\Gamma(y+k)(\frac{1}{1+1/\theta})^{y+k}} z^{(y+k)-1} e^{-z/(\frac{1}{1+1/\theta})}$$

where $z \geq 0,\ k' = y + k > 0,\ \theta' = \frac{1}{1+1/\theta} > 0$

So the conditional density of Z given Y is also from the $Gamma$ family: $Z|Y = y \sim Gamma(k', \theta')$

When $Z \sim Gamma(k = \mu\phi,\ \theta = 1/\phi)$, plugging in the values of k and θ, we get that:

$$Z|Y = y \sim Gamma(k' = y + \mu\phi,\ \ \theta' = \frac{1}{1+\phi})$$

(c)

From question (b), we have derived that:

$$f_Y(y) = \frac{\Gamma(y+k)(\frac{1}{1+1/\theta})^{y+k}}{\Gamma(k)\Gamma(y+1)\theta^k} = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}(\frac{1}{1+1/\theta})^y(\frac{1/\theta}{1+1/\theta})^k = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}(\frac{1}{1+1/\theta})^y(1-\frac{1}{1+1/\theta})^k$$

Since $k > 0$ and $\theta > 0$, we can get that $r = k > 0$ and $0 < p = \frac{1}{1+1/\theta} < 1$, so the marginal of Y has a negative binomial distribution: $f(y; r, p) = \frac{\Gamma(y+r)}{\Gamma(r)\Gamma(y+1)} p^y (1-p)^r$, where y is the number of trials until the r^{th} success, and $1 - p$ is the probability of success in the independent trials.

When $Z \sim Gamma(k = \mu\phi,\ \theta = 1/\phi)$, plugging in the values of k and θ, we get that:

$$r = k = \mu\phi, \quad p = \frac{1}{1+1/\theta} = \frac{1}{1+\phi}$$

So the marginal of Y has a negative binomial distribution:

$$f(y; \mu, \phi) = \frac{\Gamma(y+\mu\phi)}{\Gamma(\mu\phi)\Gamma(y+1)}(\frac{1}{1+\phi})^y(1-\frac{1}{1+\phi})^{\mu\phi}$$

and the mean and varaince are:

$$E(Y) = \frac{rp}{1-p} = \frac{\mu\phi/(1+\phi)}{1-1/(1+\phi)} = \mu, \quad Var(Y) = \frac{rp}{(1-p)^2} = \frac{\mu\phi/(1+\phi)}{(1-1/(1+\phi))^2} = \mu(1+\phi)/\phi$$

Therefore, $Var(Y) \neq E(Y)$, which defines a probability model with Poisson observations but a latent *Gamma* variable yielding a variance that is not equal to the mean.

(d)

When $Z \sim Gamma(k=\nu,\ \theta=\mu/\nu)$, just changing the parameter values, we still can derive as shown above that the conditional density of Z given Y is from the *Gamma* family: $Z|Y=y \sim Gamma(k', \theta')$

Plugging in the values of k and θ, we get that:

$$Z|Y=y \sim Gamma(k'=y+k=y+\nu,\ \ \theta' = \frac{1}{1+1/\theta} = \frac{1}{1+\nu/\mu})$$

Also, we can derive as shown above that the marginal of Y has a negative binomial distribution: $f(y;r,p) = \frac{\Gamma(y+r)}{\Gamma(r)\Gamma(y+1)}p^y(1-p)^r$, where y is the number of trials until the r^{th} success, and $1-p$ is the probability of success in the independent trials.

Plugging in the values of k and θ, we get that:

$$r = k = \nu, \quad p = \frac{1}{1+1/\theta} = \frac{1}{1+\nu/\mu} = \frac{\mu}{\mu+\nu}$$

So the marginal of Y has a negative binomial distribution:

$$f(y;\mu,\nu) = \frac{\Gamma(y+\nu)}{\Gamma(\nu)\Gamma(y+1)}(\frac{\mu}{\mu+\nu})^y(1-\frac{\mu}{\mu+\nu})^\nu$$

and the mean and varaince are:

$$E(Y) = \frac{rp}{1-p} = \frac{\nu\mu/(\mu+\nu)}{1-\mu/(\mu+\nu)} = \mu, \quad Var(Y) = \frac{rp}{(1-p)^2} = \frac{\nu\mu/(\mu+\nu)}{(1-\mu/(\mu+\nu))^2} = (\mu+\nu)\mu/\nu$$

Therefore, $Var(Y) \neq E(Y)$, which also defines a probability model with Poisson observations but a latent *Gamma* variable yielding a variance that is not equal to the mean.

(e)

From question (d), we get that the marginal of Y has a negative binomial distribution:

$$f(y;\mu,\nu) = \frac{\Gamma(y+\nu)}{\Gamma(\nu)\Gamma(y+1)}(\frac{\mu}{\mu+\nu})^y(1-\frac{\mu}{\mu+\nu})^\nu = \frac{\Gamma(y+\nu)}{\Gamma(\nu)\Gamma(y+1)}\ exp\{y\log(\frac{\mu}{\mu+\nu}) + \nu\log(1-\frac{\mu}{\mu+\nu})\}$$

When ν is known, set $\theta = \log(\frac{\mu}{\mu+\nu})$, $b(\theta) = -\nu\log(1-\frac{\mu}{\mu+\nu}) = -\nu\log(1-e^\theta)$, $a(\phi)=1$, $C(y,\phi) = \frac{\Gamma(y+\nu)}{\Gamma(\nu)\Gamma(y+1)}$, hence the negative binomial distribution has the exponential family form: $f(y;\theta,\phi) = exp\{\frac{y\theta-b(\theta)}{a(\phi)}\}C(y,\phi)$

$\phi = \nu$

So $b'(\theta) = \frac{\nu e^\theta}{1-e^\theta}$, and its inverse function is $b'^{-1}(\theta) = \log(\frac{\theta}{\theta+\nu})$

Thus the canomical link here is $\eta = g(\mu) = b'^{-1}(\mu) = \log(\frac{\mu}{\mu+\nu})$

(f)

```r
# aggregate the data
# and compute the mean and variance of the counts for each combination of factors
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
quine_agg = quine %>% group_by(Eth,Sex,Age,Lrn) %>% summarise(Mean=mean(Days), Var=var(Days))
```

In question (d), we have derived that: $E(Y) = \mu, \quad Var(Y) = (\mu + \nu)\mu/\nu = \mu + \mu^2/\nu$, thus we can fit a model with function: $Var \sim 1 \times Mean + \beta \times Mean^2$, so we can get the value of ν as $\hat{\nu} = 1/\hat{\beta}$

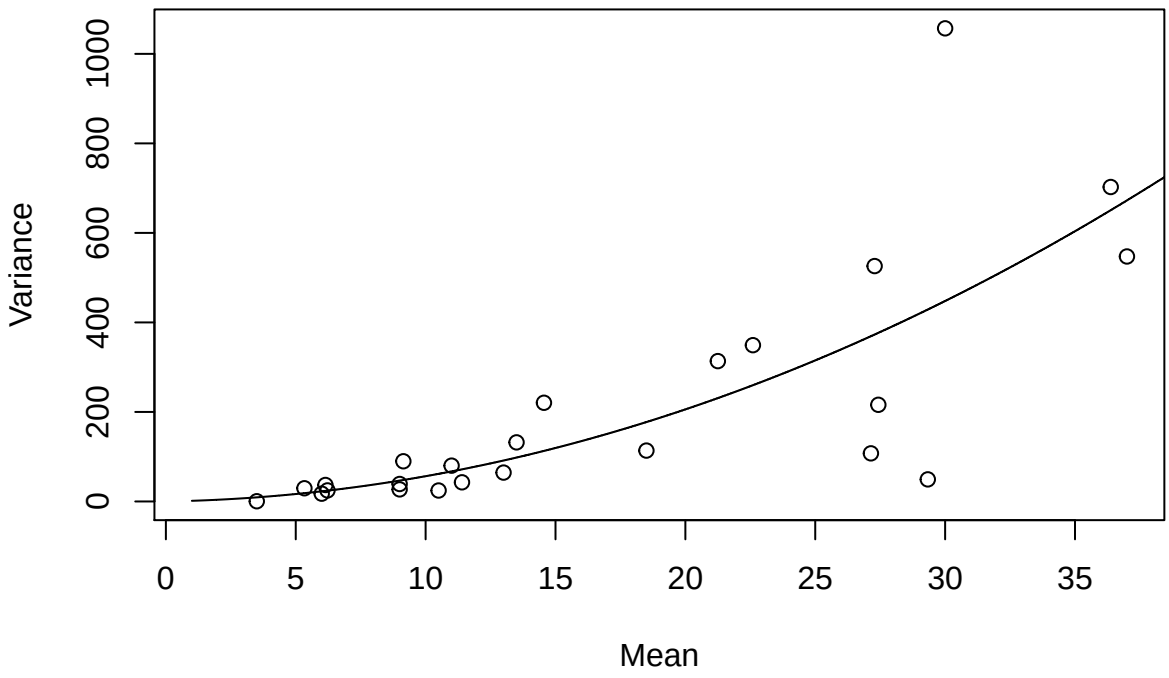
```r
# fit a model: Var ~ Mean + beta * Mean^2
model = lm(Var~offset(Mean)+I(Mean^2)-1, quine_agg)
summary(model)
```

```
##
## Call:
## lm(formula = Var ~ offset(Mean) + I(Mean^2) - 1, data = quine_agg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -379.50  -31.00   -2.41   44.36  609.14
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## I(Mean^2)  0.46429    0.06278   7.396 1.61e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 173 on 23 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7322, Adjusted R-squared:  0.7206
## F-statistic:  62.9 on 1 and 23 DF,  p-value: 4.978e-08
```

```r
# value of v
nu = 1/coef(model); nu
```

```
## I(Mean^2)
##  2.153812
```

```r
# plot Var~Mean
plot(Var~Mean, quine_agg, ylab="Variance")
x=seq(1,40,0.001)
lines(x,predict(model,data.frame(Mean=x)))
```



Answer:

As shown in the summary, this model fits well and the coefficient is very significant, so we can get that: $\hat{\nu} = 1/\hat{\beta} = 2.153812$, so we fit following model using the negative-binomial family in the glm function with an estimated $\hat{\nu} = 2.153812$.

First, as in question (a), we fit a model with only main effects. The GLM model for Negative Binomial (count) response here is:
likelihood: $P(Days_i = y_i) = \frac{\Gamma(y_i+\nu)}{\Gamma(\nu)\Gamma(y_i+1)}(\frac{\mu_i}{\mu_i+\nu})^{y_i}(1-\frac{\mu_i}{\mu_i+\nu})^{\nu},\ y_i = 0, 1, 2, ...$
linear predictor: $\eta_i = \beta_0 + \beta_1 Eth_i + \beta_2 Sex_i + (\beta_3\, to\, \beta_5) Age_i + \beta_6 Lrn_i + \epsilon_i$
link function (logit): $\eta_i = \log(\frac{\mu_i}{\mu_i+\nu})$

```
# fit a Negative Binomial main effects model with nu=2.153812
library(MASS)
m.glm_nb = glm(Days~., negative.binomial(nu), quine)
summary(m.glm_nb)
```

```
##
## Call:
## glm(formula = Days ~ ., family = negative.binomial(nu), data = quine)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3196  -1.1216  -0.3474   0.4949   2.7924
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  2.88498    0.22710  12.704  < 2e-16 ***
## EthN        -0.56731    0.15241  -3.722 0.000286 ***
## SexM         0.08790    0.15899   0.553 0.581250
## AgeF1       -0.44430    0.23917  -1.858 0.065335 .
## AgeF2        0.09382    0.23438   0.400 0.689566
## AgeF3        0.35987    0.24646   1.460 0.146507
## LrnSL        0.29760    0.18599   1.600 0.111854
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.1538) family taken to be 1.582495)
##
##     Null deviance: 296.80  on 145  degrees of freedom
## Residual deviance: 252.96  on 139  degrees of freedom
## AIC: 1125.4
##
## Number of Fisher Scoring iterations: 7
```

Answer:

In the model summary, dispersion parameter $= 1.582495 > 1$, so F-tests will be used for comparing between models and dropping insignificant interaction terms due to dispersion problem.

Then fit a Negative Binomial model with all the two-way interaction terms. The linear predictor in this model is:
linear predictor: $\eta_i = \beta_0 + \beta_1 Eth_i + \beta_2 Sex_i + (\beta_3 \, to \, \beta_5) Age_i + \beta_6 Lrn_i + \beta_7 Eth_i : Sex_i + (\beta_8 \, to \, \beta_{10}) Eth_i : Age_i + \beta_{11} Eth_i : Lrn_i + (\beta_{12} \, to \, \beta_{14}) Sex_i : Age_i + \beta_{15} Sex_i : Lrn_i + (\beta_{16}, \beta_{17}) Age_i : Lrn_i + \epsilon_i$

```r
# Negative Binomial model with all two-way interactions
m.glm_nb2 = glm(Days~(Eth+Sex+Age+Lrn)**2, negative.binomial(nu), quine)
summary(m.glm_nb2)
```

```
##
## Call:
## glm(formula = Days ~ (Eth + Sex + Age + Lrn)^2, family = negative.binomial(nu),
##     data = quine)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4046  -0.9397  -0.2890   0.4861   2.3901
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.99755    0.33901   8.842 6.34e-15 ***
## EthN        -0.23960    0.39473  -0.607  0.54493
## SexM        -0.76911    0.38368  -2.005  0.04712 *
## AgeF1       -0.02351    0.41906  -0.056  0.95534
## AgeF2       -0.52451    0.54964  -0.954  0.34174
## AgeF3       -0.25272    0.40800  -0.619  0.53675
## LrnSL        0.38003    0.48835   0.778  0.43790
## EthN:SexM    0.36338    0.29723   1.223  0.22374
## EthN:AgeF1  -0.70869    0.44119  -1.606  0.11067
## EthN:AgeF2  -1.23735    0.43359  -2.854  0.00504 **
## EthN:AgeF3   0.03939    0.45208   0.087  0.93071
## EthN:LrnSL   0.07251    0.34421   0.211  0.83348
```

```
## SexM:AgeF1   0.01718    0.47947   0.036  0.97147
## SexM:AgeF2   1.53047    0.51870   2.951  0.00377 **
## SexM:AgeF3   1.24907    0.45871   2.723  0.00737 **
## SexM:LrnSL   0.07437    0.41446   0.179  0.85788
## AgeF1:LrnSL -0.41893    0.48488  -0.864  0.38920
## AgeF2:LrnSL  0.51217    0.48963   1.046  0.29752
## AgeF3:LrnSL       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.1538) family taken to be 1.319882)
##
##     Null deviance: 296.8  on 145  degrees of freedom
## Residual deviance: 209.1  on 128  degrees of freedom
## AIC: 1103.6
##
## Number of Fisher Scoring iterations: 10
```

```
# drop one interaction term from the Negative Binomial model with two-way interactions
drop1(m.glm_nb2, test="F")
```

```
## Single term deletions
##
## Model:
## Days ~ (Eth + Sex + Age + Lrn)^2
##         Df Deviance    AIC F value   Pr(>F)
## <none>       209.10 1103.6
## Eth:Sex  1   210.99 1103.0  1.1584 0.283822
## Eth:Age  3   222.56 1107.8  2.7462 0.045698 *
## Eth:Lrn  1   209.15 1101.6  0.0326 0.856941
## Sex:Age  3   233.42 1116.0  4.9631 0.002721 **
## Sex:Lrn  1   209.13 1101.6  0.0194 0.889507
## Age:Lrn  2   213.01 1102.5  1.1979 0.305176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

The interaction term "Sex:Lrn" has a p-value = 0.889507, which is the most insignificant term in the model, so we drop this interaction term first and then continue to drop the next one.

```
# fit a new model removing the "Sex:Lrn" term
m.glm_nb3 = glm(Days~Eth+Sex+Age+Lrn+Eth:Sex+Eth:Age+Eth:Lrn+Sex:Age+Age:Lrn,
                negative.binomial(nu), quine)

# drop a second term from the Negative Binomial model with two-way interactions
drop1(m.glm_nb3, test="F")
```

```
## Single term deletions
##
## Model:
## Days ~ Eth + Sex + Age + Lrn + Eth:Sex + Eth:Age + Eth:Lrn +
##     Sex:Age + Age:Lrn
##         Df Deviance    AIC F value   Pr(>F)
## <none>       209.13 1101.6
## Eth:Sex  1   211.10 1101.1  1.2111 0.273164
## Eth:Age  3   222.56 1105.8  2.7607 0.044827 *
```

```
## Eth:Lrn  1   209.18 1099.6  0.0300 0.862774
## Sex:Age  3   233.45 1114.1  4.9992 0.002593 **
## Age:Lrn  2   215.30 1102.3  1.9019 0.153442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

The interaction term "Eth:Lrn" has a p-value = 0.862774, which is the most insignificant term in the model now, so we drop this interaction term and then continue to drop the next one.

```
# fit a new model removing the "Eth:Lrn" term
m.glm_nb4 = glm(Days~Eth+Sex+Age+Lrn+Eth:Sex+Eth:Age+Sex:Age+Age:Lrn,
                negative.binomial(nu), quine)

# drop another term
drop1(m.glm_nb4, test="F")
```

```
## Single term deletions
##
## Model:
## Days ~ Eth + Sex + Age + Lrn + Eth:Sex + Eth:Age + Sex:Age +
##     Age:Lrn
##         Df Deviance    AIC F value   Pr(>F)
## <none>       209.18 1099.6
## Eth:Sex  1   211.10 1099.1  1.1902 0.277312
## Eth:Age  3   225.79 1106.4  3.4417 0.018793 *
## Sex:Age  3   233.45 1112.3  5.0267 0.002498 **
## Age:Lrn  2   215.32 1100.4  1.9087 0.152412
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

The interaction term of "Eth:Sex" has a p-value = 0.277312, which is the most insignificant term in the model now, so we drop this interaction term and then continue to drop the next one.

```
# fit a new model removing the "Eth:Sex" term
m.glm_nb5 = glm(Days~Eth+Sex+Age+Lrn+Eth:Age+Sex:Age+Age:Lrn,
                negative.binomial(nu), quine)

# drop another term
drop1(m.glm_nb5, test="F")
```

```
## Single term deletions
##
## Model:
## Days ~ Eth + Sex + Age + Lrn + Eth:Age + Sex:Age + Age:Lrn
##         Df Deviance    AIC F value   Pr(>F)
## <none>       211.10 1099.6
## Eth:Age  3   227.90 1105.9  3.4754 0.017980 *
## Sex:Age  3   234.74 1110.9  4.8902 0.002963 **
## Age:Lrn  2   217.02 1099.9  1.8383 0.163161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

The interaction term of "Age:Lrn" has a p-value = 0.163161, which is the most insignificant term in the model now, so we drop this interaction term and then continue to drop the next one.

```
# fit a new model removing the "Age:Lrn" term
m.glm_nb6 = glm(Days~Eth+Sex+Age+Lrn+Eth:Age+Sex:Age,
                negative.binomial(nu), quine)

# drop another term
drop1(m.glm_nb6, test="F")
```

```
## Single term deletions
##
## Model:
## Days ~ Eth + Sex + Age + Lrn + Eth:Age + Sex:Age
##         Df Deviance    AIC F value   Pr(>F)
## <none>       217.02 1101.5
## Lrn      1   224.18 1104.7  4.3899 0.038048 *
## Eth:Age  3   233.48 1107.5  3.3627 0.020720 *
## Sex:Age  3   236.68 1109.8  4.0168 0.008977 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

Now all the terms are significant in this model, so we stop dropping more interaction terms and keep this model as our final Negative Binomial model. Let's have a look at the model summary.

```
# summary of the final Negative Binomial model
summary(m.glm_nb6)
```

```
##
## Call:
## glm(formula = Days ~ Eth + Sex + Age + Lrn + Eth:Age + Sex:Age,
##     family = negative.binomial(nu), data = quine)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3430  -1.0406  -0.2574   0.4526   2.6411
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.86123    0.31589   9.058 1.47e-15 ***
## EthN         0.01757    0.33045   0.053  0.95768
## SexM        -0.54747    0.34065  -1.607  0.11040
## AgeF1       -0.26732    0.38162  -0.700  0.48485
## AgeF2       -0.03004    0.41409  -0.073  0.94227
## AgeF3       -0.20174    0.40280  -0.501  0.61730
## LrnSL        0.43687    0.18568   2.353  0.02010 *
## EthN:AgeF1  -0.76660    0.42224  -1.816  0.07169 .
## EthN:AgeF2  -1.21915    0.42744  -2.852  0.00504 **
## EthN:AgeF3  -0.05733    0.44239  -0.130  0.89709
## SexM:AgeF1   0.02829    0.44775   0.063  0.94971
## SexM:AgeF2   1.20059    0.45450   2.642  0.00924 **
## SexM:AgeF3   1.21183    0.45136   2.685  0.00818 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for Negative Binomial(2.1538) family taken to be 1.369464)
##
##     Null deviance: 296.80  on 145  degrees of freedom
## Residual deviance: 217.02  on 133  degrees of freedom
## AIC: 1101.5
##
## Number of Fisher Scoring iterations: 7
```

```
# compare the final model with original models with and without all the two-way interactions
anova(m.glm_nb, m.glm_nb6, test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: Days ~ Eth + Sex + Age + Lrn
## Model 2: Days ~ Eth + Sex + Age + Lrn + Eth:Age + Sex:Age
##   Resid. Df Resid. Dev Df Deviance     F   Pr(>F)
## 1       139     252.96
## 2       133     217.02  6    35.94 4.374 0.000459 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(m.glm_nb6, m.glm_nb2, test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: Days ~ Eth + Sex + Age + Lrn + Eth:Age + Sex:Age
## Model 2: Days ~ (Eth + Sex + Age + Lrn)^2
##   Resid. Df Resid. Dev Df Deviance      F Pr(>F)
## 1       133     217.02
## 2       128     209.10  5   7.9201 1.2001 0.3129
```

Answer:

Comparing this final Negative Binomial model with the original main effects Negative Binomial model without two-way interactions, the p-value of F-test is 0.000459, so the larger final model is preferred.

Comparing this final Negative Binomial model with the Negative Binomial model with all the two-way interactions, the p-value of F-test is 0.3129, so the smaller final model is preferred.

Therefore, we agree with keeping this final Negative Binomial model. And the linear predictor in this final model is:
linear predictor: $\eta_i = \beta_0 + \beta_1 Eth_i + \beta_2 Sex_i + (\beta_3 \, to \, \beta_5) Age_i + \beta_6 Lrn_i + (\beta_7 \, to \, \beta_9) Eth_i : Age_i + (\beta_{10} \, to \, \beta_{12}) Sex_i : Age_i + \epsilon_i$

Finally, let's compare the final Negative Binomial model with the above final quasi-Poisson model in question (a). The selected two-way interaction predictors are the same in the linear predictor equations in both models. However, the residual deviance of the Negative Binomial model is much smaller than that of the quasi-Poisson model, hence the dispersion parameter of Negative Binomial model is much smaller as well. Therefore, the Negative Binomial model is preferred in this case.