Solution of assignment 8, ST2304

Problem 1

1. A poisson process on the interval between 0 and 3 years for each student. This assumes amongst others that

   - The rate at which articles are produced is constant during those 3 years,
   - The time to produce one article is independent of the time to produce the next,
   - This assumes that the expected number of articles produced is the same for all students within each group,
   - You can not produce (finish) two articles at the same time

2.
```
Call:
glm(formula = art ~ fem + mar + kid5 + phd + ment, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5672  -1.5398  -0.3660   0.5722   5.4467

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.459860   0.093335   4.927 8.35e-07 ***
femWomen    -0.224594   0.054613  -4.112 3.92e-05 ***
marSingle   -0.155243   0.061374  -2.529   0.0114 *
kid5        -0.184883   0.040127  -4.607 4.08e-06 ***
phd          0.012823   0.026397   0.486   0.6271
ment         0.025543   0.002006  12.733  < 2e-16 ***
---


(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1817.4  on 914  degrees of freedom
Residual deviance: 1634.4  on 909  degrees of freedom
AIC: 3314.1

Number of Fisher Scoring iterations: 5
```

   In the summary table we see that the variable `phd`, the prestige of the PhD department, is not significant and can thus be removed. We can see the same from `drop1()`.

3. Overdispersion implies that the variance is larger than expected; under the poisson distribution the variance is assumed to be equal to the mean. We test for overdisperion by testing the null hypothesis that there is no overdispersion against the alternative hypothesis that there is overdispersion. Under the null hypothesis, the residual deviance of the model has a chi-square distribution with $n-p$ degrees of freedom. From the summary table we see that we have `Residual deviance: 1634.4 on 909 degrees of freedom`. We can find the probability to find this value or larger under the null hypothesis using `pchisq(1634.6, df=910,lower.tail=F)`, which is 5.775682e-44. We thus reject H_0, and conclude that there is no overdispersion.

We can also see this from the critical value, `qchisq(.95,df=910)`, which is 981.29. The observed value (1634.4) is larger than the critical value, thus we reject H_0.

4.
```
Call:
glm(formula = art ~ fem + mar + kid5 + phd + ment, family = quasipoisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5672  -1.5398  -0.3660   0.5722   5.4467

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.459860   0.126227   3.643 0.000285 ***
femWomen    -0.224594   0.073860  -3.041 0.002427 **
marSingle   -0.155243   0.083003  -1.870 0.061759 .
kid5        -0.184883   0.054268  -3.407 0.000686 ***
phd          0.012823   0.035700   0.359 0.719544
ment         0.025543   0.002713   9.415  < 2e-16 ***
---

(Dispersion parameter for quasipoisson family taken to be 1.829006)

    Null deviance: 1817.4  on 914  degrees of freedom
Residual deviance: 1634.4  on 909  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

   The estimate of the scale parameter (called Dispersion parameter in the output in R) is 1.829006. Thus, the variance is 1.83 times the mean of the distribution.

5. When comparing the summary output from the model assuming no overdispersion (under 1.) and with overdipsersion (4.) we see that the latter has larger standard errors of the parameter estimates $\hat{\beta}$ (but the same estimates). This is reasonable, since the estimate of the variance $\hat{\sigma}$ has increased, and the variance of $\hat{\beta}$ is given by var $\hat{\beta} = \sigma^2 / \sum (x_i - \bar{x})^2$

6. Using `summary()` and/or `drop1()`, we find that the variables `phd` (the prestige of the PhD department) and `mar` (marital status of the student) are not significant. The latter was significant in the model assuming no overdispersion.

   It makes sense that fewer variables now have a significant effect; because the standard errors are larger (see 5.), the confidence intervals (estimate +- 1.96S.E.) are wider. Under the poisson model the confidence interval of `mar` did not include zero, but under the quassi-poisson model it does include zero. Under the poisson model we thus had a false rejection of H0: the estimate is equal to zero, and wrongly concluded that there is an effect of marital status.

7. In general, we can get overdispersion if the assumptions of the poisson process (see 1.) are violated, so that the process is not truly a poisson process. For example,

   - The time spent on subsequent papers is often not independent, e.g. you may spend a lot of time on the first paper but less on a subequent paper on the same topic, or you may write 2 papers based on the same experiment / field work.

- Not all differences between students are accounted for in the model; for example research topic and amount of lab work involved may influence the number of papers produced, or how hard students work.