Assignment 8, ST2304

Problem 1 Download the data set

```
read.table("https://www.math.ntnu.no/~jarlet/statmod/biochemists.dat")
```

This data set contains a sample of 915 biochemistry graduate students for which the following variables are observed.

'art': count of articles produced during last 3 years of Ph.D.
'fem': factor indicating gender of student, with levels Men and Women
'mar': factor indicating marital status of student, with levels Single and Married
'kid5': number of children aged 5 or younger
'phd': prestige of Ph.D. department
'ment': count of articles produced by Ph.D. mentor during last 3 years

1. Describe a process which would make the distribution of number of articles produced by each student Poisson distributed. What assumptions does this process involve?

2. Fit a poisson regression to the data using number of articles produced as the response variable under the assumption that there is no overdispersion (choose `family=poission`). Based on these assumptions, should any of the explanatory variables in the model be dropped?

3. Is there any indication in the data of overdispersion? Compute the p-value and the critical value of the hypothesis test associated with this question.

4. Then redo the analysis in point 2 using `family=quasipoisson`. What is the estimate of the scale parameter added to the new model?

5. How does the standard errors of the parameter estimates change? Why is it reasonable that the standard errors change the way they do? You may want to base your argument on an analogy with simple linear regression for which the variance of the estimator of the slope $\mathrm{Var}\,\hat{\beta} = \sigma^2 / \sum(x_i - \bar{x})^2$ (see Løvås, p. 276).

6. Based on the quasipoisson model, which explanatory variables should be included? Is it reasonable that fewer explanatory variables now appear to have a significant effect? Which of the two alternative models to you trust the most?

7. What are possible explanations for the observed overdispersion in the data?