Throughout, vectors are denoted by bold letters, e.g., $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$.

A1 Consider the linear regression model M_0

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where X is an $n \times p$ design matrix of rank p and $H = X(X^\top X)^{-1}X^\top$ is the hat matrix. Consider also the model M_1, viz.

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{z}\gamma + \boldsymbol{\varepsilon}^* = X^* \begin{pmatrix} \boldsymbol{\beta} \\ \gamma \end{pmatrix} + \boldsymbol{\varepsilon}^*,$$

where $\boldsymbol{z}$ is an $n \times 1$ vector corresponding to one additional predictor and $\gamma \in \mathbb{R}$ is an additional parameter. Let $H^* = X^*(X^{*\top}X^*)^{-1}X^{*\top}$ denote the hat matrix corresponding to M_1.

(a) Let I_n denote the $n \times n$ identity matrix in $\mathbb{R}^n$. Verify that $I_n - H$ is symmetric and idempotent.

(b) Prove that

$$\left(X^\top X - \frac{X^\top \boldsymbol{z}\boldsymbol{z}^\top X}{\boldsymbol{z}^\top \boldsymbol{z}} \right)^{-1} = (X^\top X)^{-1} + \frac{(X^\top X)^{-1}X^\top \boldsymbol{z}\boldsymbol{z}^\top X(X^\top X)^{-1}}{\boldsymbol{z}^\top (I_n - H)\boldsymbol{z}}.$$

(c) Use part (b) to show that

$$H^* = H + \frac{(I_n - H)\boldsymbol{z}\boldsymbol{z}^\top (I_n - H)}{\boldsymbol{z}^\top (I_n - H)\boldsymbol{z}}.$$

(d) Let $\boldsymbol{R}_0 = (I_n - H)\boldsymbol{Y}$ denote the vector of residuals in model M_0, set $\boldsymbol{z}^* = (I_n - H)\boldsymbol{z}$ and define

$$\varrho = \frac{\boldsymbol{z}^{*\top}\boldsymbol{R}_0}{\sqrt{(\boldsymbol{z}^{*\top}\boldsymbol{z}^*)(\boldsymbol{R}_0^\top \boldsymbol{R}_0)}}$$

to be the uncentered correlation coefficient between $\boldsymbol{R}_0$ and $\boldsymbol{z}^*$. Show that

$$\varrho^2 = \frac{\mathrm{SSE}_0 - \mathrm{SSE}_1}{\mathrm{SSE}_0},$$

where SSE_i denotes the sum of squares due to residual error in model M_i, $i = 0, 1$.

A2 Consider the Negative Binomial distribution with parameters $\mu > 0$ and $\theta_Z > 0$; the corresponding probability mass function is given by

$$f(y; \mu, \theta_z) = \frac{\Gamma(y + \theta_z)}{\Gamma(y+1)\Gamma(\theta_z)} \left(\frac{\theta_z}{\mu + \theta_z} \right)^{\theta_z} \left(\frac{\mu}{\mu + \theta_z} \right)^{y}, \quad y = 0, 1, \ldots,$$

where $\Gamma(\cdot)$ denotes the Gamma function. *Assume throughout that θ_Z, the "number of successes until the experiment is stopped", is known.*

(a) Show that the Negative Binomial family is an exponential dispersion family when θ_z is known. Identify all functions appearing in the general formula of an exponential dispersion family.

(b) Compute the mean and variance of the Negative Binomial distribution and identify the mean-variance relationship when θ_z is known.

(c) Determine the canonical link of a Negative Binomial GLM with known θ_z. Comment on the suitability of this link for modeling.

A3 `R` **exercise.** Load the data set `mammals`, viz.

```
library(MASS)
data(mammals)
attach(mammals)
head(mammals)

##                     body brain
## Arctic fox         3.385  44.5
## Owl monkey         0.480  15.5
## Mountain beaver    1.350   8.1
## Cow              465.000 423.0
## Grey wolf         36.330 119.5
## Goat              27.660 115.0
```

This data set comprises average `body` weight in kg and `brain` weight in g for 62 species of mammals. It is of interest to find how the brain weight depends on the body weight.

(a) Find a suitable linear regression model for the data. Transform the explanatory variable and/or the response if appropriate. Comment on the fit and *interpret the model*.

(b) Fit the following two models:

```
m1 <- lm(brain~body)
m2 <- glm(brain~body,family=gaussian(link="identity"))
```

Compare `summary(m1)` and `summary(m2)`: (i) compare the estimated coefficients and their standard errors; (ii) calculate the estimate of σ^2 using `m1` and relate it to the estimated dispersion parameter reported in the summary of `m2`; (iii) relate the `Null deviance` and the `Residual deviance` in the summary of `m2` to the total sum of squares, residual sum of squares and regression-explained sum of squares in `m1`; (iv) find a way to calculate the F-statistic in the summary of `m1` using only the summary of `m2`.

(c) Fit the gamma GLM with the log-link to the data, viz.

```
m3 <- glm(brain~log(body),family=Gamma(link="log"))
```

Explore `summary(m3)`, report the estimated coefficients and comment on their significance. Think of an interpretation.

(d) Fit the gamma GLM as in part (c), say `m4`, but now using the reciprocal link (called `inverse` in `R`). Would you prefer this model or the model fitted in part (c)? Provide a brief explanation (think of model fit but also of model interpretation), and plot the data along with the fitted regression curves (that is, estimated average brain weight using models `m3` and `m4`).

(e) Compare the model `m3` from part (c) to the best model you found in part (a).

(f) The average body weight of a male polar bear is 450 kg. Calculate his average brain weight predicted using your chosen model in part (a), as well as the models `m3` and `m4`. Which model prediction do you trust the most and why?