Problem 1

a) The model assumes that

$$\ln y_i = \beta_0 + \beta_1 \ln x_i + e_i, \tag{1}$$

where the error terms e_i are independent $N(0, \sigma^2)$ for each observation $i = 1, 2, \ldots, n$. The unknown parameters are β_0, β_1 and σ^2 and these are estimated to $\hat{\beta}_0 = 2.11 \pm 0.09$, $\hat{\beta}_1 = 0.74 \pm 0.03$ and $\hat{\sigma} = 0.6511$.

b) If brain size depends on body size in the above model, then $\beta_1 \neq 0$ (H_1), if not $\beta_1 = 0$ (H_0). The p-value for the test of this H_0 and H_1 is smaller than $2 \cdot 10^{-16}$ (second row under `Coefficients`) and so we can clearly reject H_0 in favour of H_1 at the $\alpha = 0.05$ level of significance.

c) Predicted log brain size for humans becomes

$$\widehat{\ln y} = \hat{\beta}_0 + \hat{\beta}_1 \ln x = 2.11 + 0.74 \cdot 4.127 = 5.1784, \tag{2}$$

and predicted brain size

$$\hat{y} = e^{\widehat{\ln y}} = e^{5.1784} = 177 \tag{3}$$

grams. This is clearly much smaller than the average brains size in humans of 1320 grams. The deviation from the predicted value on the log scale is $log(1320) - 5.17 = 2.00$, that is, more than three times $\sigma = 0.65$ so something unusual has clearly happened during the recent evolutionary history of the human species, see `https://en.wikipedia.org/wiki/Evolution_of_the_brain` for a review of theories.

d) Extending the model to include a quadratic effect of log body mass do not improve model fit, that is, the coefficient for $(\ln x)^2$ is not signifcantly different from zero, so based on this test there is no evidence for non-linearity on the log-log scale in the data.

e) On the log-log scale we have

$$\ln y = \beta_0 + \beta_1 \ln x. \tag{4}$$

Exponentiating both sides yields

$$y = e^{\beta_0 + beta_1 \ln x} = e^{\beta_0} e^{\beta_1 \ln x} = e^{\beta_0} e^{\ln x_1^\beta} = e_0^\beta x_1^\beta, \tag{5}$$

that is, a relationship on the form

$$y = cx^b, \tag{6}$$

where $b = \beta_1$ and $c = e^{\beta_0}$. Since b equals the slope β_0 in (1), an estimate of b is $\hat{b} = 0.762 \pm 0.037$.

If we had direct proportionality between y and x, that is, $y = cx$, we would have $\beta_1 = b = 1$. This null hypothesis $H_0 : \beta_1 = 1$ vs. $H_1 : \beta_1 \neq 1$ can be tested using the test statistic

$$T = \frac{\hat{\beta}_1 - 1}{\widehat{SE\hat{\beta}_1}} \tag{7}$$

which is t-distributed with $n - 2 = 58$ degrees of freedom under this H_0. The observed value becomes $T = (0.7624 - 1)/0.034 = -6.84$ which is clearly smaller than lower critical values of a two-side test with $\alpha = 0.05$, $-t_{0.025,58} \approx -t_{0.025,60} = -2.00$. We can thus reject the null hypothesis of direct proportionality.

Problem 2

a) The unknown parameters are μ, $\alpha_1, \ldots, \alpha_4$, $\beta_1, \ldots, \beta_6$. We impose the constraints that $\alpha_1 = \beta_1 = 0$ to make the model identifiable. With these contraints (so called treatment contrasts), the interpreation of μ becomes the mean within the group defined by the first levels of i and j, α_2 the difference in mean between group $i = 2$ relative to group $i = 1$ and so on.

Focusing on a particular subpopulation, say the age group of 40-54 year olds in Fredericia, the total number of cancer indicidents y out of total number of persons n in the subpopulation will follow a binomial distribtuion if each out of the n person has the same probability p of getting cancer and we have independence between different persons within each subpopulations.

The link function ensures that the model predicts probabilites p on the interval between 0 and 1.

b) Within the age group of 40-54 year olds (the age category $j = 1$), the predicted propability of lung cancer in Horsens (first level $i = 1$ of city) becomes

$$\hat{p} = \frac{1}{1 + \exp(-\hat{\mu})} = \frac{1}{1 + \exp(-(-5.96))} = 0.0025 \tag{8}$$

and within Fredricia ($i = 4$)

$$\hat{p} = \frac{1}{1 + \exp(-(\hat{\mu} + \hat{\alpha}_4))} = \frac{1}{1 + \exp(-(-5.96 + 0.33))} = 0.0035 \tag{9}$$

c) Rewriting the model on the form

$$\frac{p}{1-p} = e^{\mu + \alpha_i + \beta_j} \tag{10}$$

we see that the odds $p/(1-p)$ is always increased by a factor of e^{α_4} estimated to $e^{\hat{\alpha}_4} = e^{0.33} = 1.39$ when comparing an individual in Fredericia $i = 4$ to an individual in Horsens $i = 1$, given that we are comparing individuals in the same age group (same β_j for both individuals), that is, a 39% increase in the odds of lung cancer.

d) Under the null hypothesis of no overdispersion, the deviance D is chi-square distributed with $n - p = 24 - 9 = 15$ degrees of freedom. We reject H_0 if D is larger than upper 0.05 quantile of this distribution, $\chi^2_{0.05,15} = 24.99$. Based on the observed deviance $D = 23.63$ there is thus no evidence for overdispersion in the data.

Variouus forms of positive association between the development of lung cancer between different individuals could generate overdispersion, e.g. genetic relatedness (if lung cancer is partly heritable), missing covariates (other factors influencing lung cancer not included in the model), and a incorrect link function.

e) In the first model comparison we test a model without an effect of city, that is, $H_0 : \alpha_1 = \cdots = \alpha_4 = 0$ versus the alternative hypothesis H_1 that there are differences between any of the cities. Based on the p-value we can not reject the null hypothesis that city has no effect, so conducting the test this way, we find no evidence for differences in the risk of lung cancer among the cities.

The second and third model comparisons deals with a model where we have merged the cities of Horsens, Kolding and Vejle into one category and with Fredericia kept as a second category of the factor `petro`. `anova(glm0,glm1)` tests if the there is a difference between Fredericia relative to the other cities assuming that there is no difference in risk of cancer among the other cities. Assuming no difference among the cities is reasonable a priori given that we expect to see a elevated risk only in Fredericia. Merging the other cities into a single category reduces the model complexity (the number of parameters) and increase the statistical power of the test of interest and conducting the test this way we indeed find a signficant difference between Fredericia and the other cities (p-value= 0.03).

The third test compares the model with Horsens, Kolding and Vejle merged into on category but with a possibly elevated risk in Fredericia against the model with differences among all four cities. This test is non-significant (p-value= 0.88). There is thus nothing in the data indicating that the assumption of model `glm1` of a common risk level in Horsens, Kolding and Vejle is violated.