Statistical Modeling and Inference
Assignment IV

Joseph Pritchard - Adelaide University

April 11, 2023

Question I

a

To linearise, perform the following adjustments

$$y = \frac{\alpha x}{\delta + x}$$

$$\frac{1}{y} = \frac{\delta + x}{\alpha x}$$

$$= \frac{\delta}{\alpha}\frac{1}{x} + \frac{1}{\alpha}$$

$$\frac{1}{y} = \frac{1}{\alpha} + \frac{\delta}{\alpha}\frac{1}{x}$$

Writing explicitly as a linear equation

$$y' = \beta_0 + \beta_1 x'$$

where $y' = \frac{1}{y}$, $x' = \frac{1}{x}$, $\beta_0 = \frac{1}{\alpha}$ and $\beta_1 = \frac{\delta}{\alpha}$

b

We may write a linear regression model equation as

$$Y = X\beta$$

where y is our response vector containing y_i response variables, X is the design matrix and $\beta = [\beta_0, \beta_1]^T$ is the vector of model parameters.
We know the least square estimates for β_i are given by

$$[\beta_0, \beta_1]^T = (X^T X)^{-1} X^T Y$$

So to find least square estimates for the model parameters we can write the right hand side quantity in an explicit form. First find $X^T X$

$$X^T X = \begin{bmatrix} 1 & ... & 1 \\ x'_1 & ... & x'_n \end{bmatrix} \begin{bmatrix} 1 & x'_1 \\ . & . \\ . & . \\ . & . \\ 1 & x'_n \end{bmatrix} \tag{1}$$

$$= \begin{bmatrix} n & \Sigma x'_i \\ \Sigma x'_i & \Sigma (x'_i)^2 \end{bmatrix} \tag{2}$$

Where the Sigma symbol is taken to mean a sum over the index i. Now invert this matrix

$$(X^T X)^{-1} = \frac{1}{n\Sigma(x'_i)^2 - (\Sigma x'_i)^2} \begin{bmatrix} \Sigma(x'_i)^2 & -\Sigma x'_i \\ -\Sigma x'_i & n \end{bmatrix} \tag{3}$$

We also need the quantity

$$X^T Y = \begin{bmatrix} 1 & ... & 1 \\ x'_1 & ... & x'_n \end{bmatrix} \begin{bmatrix} y'_1 \\ . \\ . \\ . \\ y'_n \end{bmatrix} \tag{4}$$

$$= \begin{bmatrix} 1 & ... & 1 \\ x'_1 & ... & x'_n \end{bmatrix} \begin{bmatrix} \Sigma y'_i \\ \Sigma y'_i x'_i \end{bmatrix} \tag{5}$$

Combining the above we have

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \frac{1}{n\Sigma(x'_i)^2 - (\Sigma x'_i)^2} \begin{bmatrix} \Sigma(x'_i)^2 & -\Sigma x'_i \\ -\Sigma x'_i & n \end{bmatrix} \begin{bmatrix} \Sigma y'_i \\ \Sigma y'_i x'_i \end{bmatrix} \tag{6}$$

$$= \frac{1}{n\Sigma(x'_i)^2 - (\Sigma x'_i)^2} \begin{bmatrix} \Sigma(x'_i)^2 \Sigma y_j - \Sigma x'_i (\Sigma y'_j x'_j) \\ -n\Sigma x'_i y'_i - \Sigma x'_i \Sigma y'_j \end{bmatrix} \tag{7}$$

where $x'_i = \frac{1}{x_i}$, $y'_i = \frac{1}{y_i}$, $\hat{\beta}_0 = \frac{1}{\hat{\alpha}}$ and $\hat{\beta}_1 = \frac{\hat{\delta}}{\hat{\alpha}}$. To find the estimated value of the parameter $\hat{\delta}$, simply take the ratio of the two model parameters

$$\frac{\hat{\beta}_1}{\hat{\beta}_0} = \frac{\hat{\delta}}{\hat{\alpha}}(\hat{\alpha}) = \hat{\delta}$$

c

Our model, including the error term ϵ, is

$$y' = \beta_0 + \beta_1 x' + \epsilon$$

Back-transforming to our original variables gives

$$\frac{1}{y} = \beta_0 + \beta_1 \frac{1}{x'} + \epsilon$$

$$y = \frac{1}{\frac{1}{\alpha} + \frac{\delta}{\alpha}\frac{1}{x} + \epsilon}$$

$$y = \frac{\alpha x}{x + \delta + \alpha\epsilon}$$

If we instead fir the population growth model directly we would have a model

$$y = \frac{\alpha x}{x + \delta} + \epsilon$$

Notice the error term appears with a different effect on the resulting response variable and so the two models are not equivalent.

d

i

We know the residual variance, S_e, obeys the relation

$$S_e^2 = \frac{||Y - X\hat{\beta}||}{n - p}$$

where n is the sample size and p is the number of model parameters.
When calculating variance, we divide by the number of degrees of freedom in the model when standardising and so the residual variance has degrees of freedom $n - p$. The R output provided states the degrees of freedom for the residual variance is 13 and we know the model to have 2 parameters thus the sample size is

$$n = d + p = 13 + 2 = 15$$

where d is the number of degrees of freedom.

ii

Response vector is

$$Y = \begin{bmatrix} 1 & \frac{1}{y_1} \\ 1 & \frac{1}{y_2} \\ 1 & \frac{1}{y_3} \\ 1 & \frac{1}{y_4} \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{489} \\ 1 & \frac{1}{476} \\ 1 & \frac{1}{513} \\ 1 & \frac{1}{382} \end{bmatrix} \tag{8}$$

Design vector is

$$X = \begin{bmatrix} 1 & \frac{1}{x_1} \\ 1 & \frac{1}{x_2} \\ 1 & \frac{1}{x_3} \\ 1 & \frac{1}{x_4} \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{4} \\ 1 & \frac{1}{3} \\ 1 & \frac{1}{7} \\ 1 & \frac{1}{1} \end{bmatrix} \tag{9}$$

iii

iv

The confidence interval for the parameters of the linear regression model, taken from the notes, is

$$\lambda^T \hat{\beta} = t_{n-p,\frac{\alpha}{2}} S_e \sqrt{\lambda^T (X^T X)^{-1} \lambda}$$

setting λ equal to $(1, 0)$ we have the confidence interval for β_0. We know S_e from the R output to be 0.0002495, we calculated n and p previously, *alpha* is the critical value of our confidence interval - in this case $\alpha = 0.1$. The only quantity we now need is the first entry of the matrix $X^T X$. In linear regression, the first column of the design matrix act to pick out the intercept value and thus are all unit value. As a result, the quantity $X^T X$ is simply the sample size n. The confidence interval is then

$$\begin{aligned} &\beta_0 \pm t_{13}(0.0002495)\sqrt{\frac{1}{15}} \\ &= \beta_0 \pm (1.770933)(0.0002495)(0.2581989) \\ &= \beta_0 \pm 0.0001140846 \end{aligned}$$

v

Question II

a

The likelihood function of our model is

$$l = \prod_1^n f(x_i, \theta)$$

were x_i are our predictor variables and θ is the vector of parameters. In our case we have

$$l = \prod_1^n (\beta log(x_i) + \epsilon_i)$$

where

$$\epsilon_i = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{x_i^2}{2\sigma^2}\right)$$

The log likelihood function is given by the log of this function

$$L = \sum_1^n log\left(\beta log(x_i) + \epsilon_i\right)$$

where for brevity the gaussian term has been left as ϵ_i.

b

The score vector is the derivative of the log likelihood function with respect to each of the model parameters β and σ^2. Begin by noting

$$\frac{d\sigma}{d\sigma^2} = (\frac{d\sigma^2}{d\sigma})^{-1} = \frac{1}{2\sigma}$$

Thus

$$\frac{dL}{d\sigma^2} = \frac{d\sigma}{d\sigma^2}\frac{dL}{d\sigma} = \frac{1}{2\sigma}\sum_1^n \frac{d}{d\sigma}(log(\beta log(x_i) + \epsilon_i))$$

$$= \frac{1}{2\sigma}\sum_1^n \left(\frac{1}{\beta log(x_i) + \epsilon_i}\frac{d}{d\sigma}\epsilon_i\right)$$

Evaluating the gaussian derivative gives

$$\frac{d}{d\sigma}\epsilon_i = \frac{d}{d\sigma}\frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{x_i^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)\left(-\sigma^{-2}\right) exp\left(-\frac{x_i^2}{2}\sigma^2\right) + \left(-x_i^2\sigma\right) exp\left(-\frac{x_i^2}{2}\sigma^2\right)\left(\frac{1}{\sqrt{2\pi}}\right)\sigma^{-1}$$

Combining the above

$$\frac{dL}{d\sigma^2} = \frac{1}{2\sigma}\sum_1^n \left(\frac{1}{\beta log(x_i) + \epsilon_i}\left(\left(\frac{1}{\sqrt{2\pi}}\right)\left(-\sigma^{-2}\right) exp\left(-\frac{x_i^2}{2}\sigma^2\right) + \left(-x_i^2\sigma\right) exp\left(-\frac{x_i^2}{2}\sigma^2\right)\left(\frac{1}{\sqrt{2\pi}}\right)\sigma^{-1}\right)\right)$$

The β case is much simpler

$$\frac{dL}{d\beta} = \frac{d}{d\beta}\sum_{1}^{n} log(\beta log(x_i) + \epsilon_i)$$

$$= \sum_{1}^{n} \frac{log(x_i)}{\beta log(x_i) + \epsilon_i}$$

c

The maximum likelihood estimations for the parameters β and σ are found by setting the score vector entries to zero and solving for the parameter values. With the above calculated analytic expressions for the score vector entries this problem seems highly complex and a solution has not yet been found.

d

The fisher information matrix is the negative of the expectation value of the matrix of mixed partial derivatives of the log likelihood function with respect to the model parameters

$$I_{\theta(ij)} = -E\left[\frac{d}{d\theta_i}\frac{dl}{d\theta_j}\right]$$

Question III

a

The resulting scatter plot is shown in figure I.

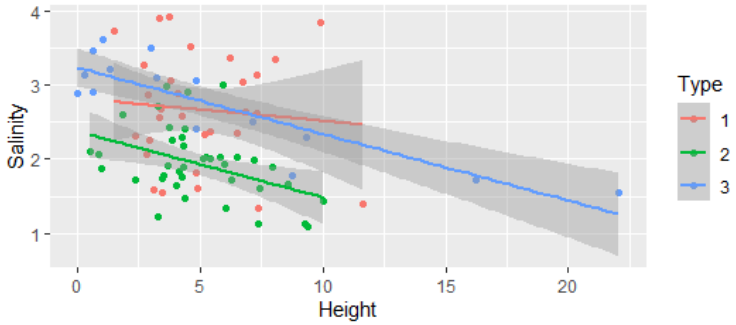


Figure I

From the produced plot, the data appears negative and linear for Type II and Type III crops. The error bars here are relatively narrow such that we may have reasonable confidence in a relationship. A negative and linear relationship also

appears present for type I crop but due to large error bars it is not possible to confidently state this assumption is not erroneous.

b

It is important that the predictor variable Type is specified as a Factor of the model. If this is not done, R will attempt to fit a slope line to this variable, whilst we want Type to modify our current intercept and slope line of Salinity.

The three linear models are fitted in R (Appendix). The results for the separate regression model are

$$y_1 = 5.7874 - (0.2902)x$$
$$y_2 = 2.9467 - (1.6772)x$$
$$y_3 = 22.4455 - (7.9971)x$$

where the y subscript labels the crop type.

c

The null hypothesis is that the interaction coefficient is zero. The alternate hypothesis is that the interaction coefficient is non-zero. The test statistic we use is:

$$t = \frac{\lambda^T \hat{\beta}}{S_e \sqrt{\lambda^T \left(X^T X\right)^{-1} \lambda}}$$

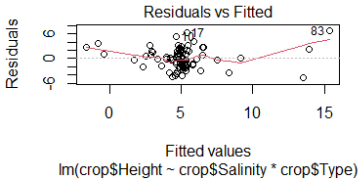
We have two interaction terms to test - Salinity with each of Type I and Type II crop. The p value for Type II is 0.100612. To deem an interaction term statistically significant we must see the p value fall above the 5 percent significance level. This level is not reached in this case so the interaction term is statistically insignificant.
For the Type III interaction case we have a p value of 9.56^{-10}, which is well above the 5 percent significance level and so this term is statistically significant.

d

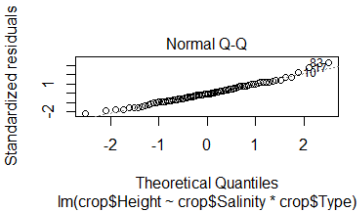
A better fit is given by a lower BIC value. In this case, we see this is achieved by model 3, separate regression, with a BIC score of 418.3894..

e

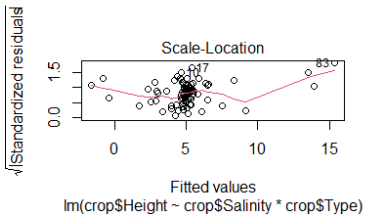
To assess the assumptions of linear regression we analyse the plots below.
Linearity - The data points in the Residuals vs Fitted plot begin to deviate upwards of the central horizontal line for larger fitted values, indicating non-linearity



(a) Figure 4



(b) Figure 5



(c) Figure 6

Normality - The residuals follow a straight line across the range of theoretical quantiles however deviate above the linear slope for high quantile values, indicating the residuals are NOT normally distributed

Equal Variance - Homoscedasticity - We again witness a deviation above the horizontal line. For higher fitter values, indicating unequal variance