IB Statistics
Example Sheet 3

Samuel Lee

Question 1

A vector $X \in \mathbb{R}^n$ has an n-variate normal distribution if, for all $t \in \mathbb{R}^n$, $t^T X$ is normally distributed. We have $\mathbb{E}[X_i] = \mu_i$ where X_i and μ_i are the i^{th} elements of the vectors X and μ. We also have $\text{Cov}[X_i, X_j] = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)^T] = \Sigma_{ij} = \Sigma_{ji}$ where Σ_{ij} is the element in row i and column j of the $n \times n$ matrix Σ. Given the above, we say that $X \sim N_n(\mu, \Sigma)$.

From the definition of matrix multiplication, we have

$$[AX]_i = \sum_{j=1}^{n} A_{ij} X_j \implies \mathbb{E}\big[[AX]_i\big] = \sum_{j=1}^{n} A_{ij}\mu_j \implies \mathbb{E}[AX] = A\mu$$

Also, the bilinearity of the covariance operator gives us

$$\begin{aligned}
\text{Cov}\big[[AX]_i, [AX]_j\big] &= \text{Cov}\left[\sum_{k=1}^{n} A_{ik}(X_k - \mu_k), \sum_{h=1}^{n} A_{jh}(X_h - \mu_h)\right] \\
&= \sum_{k=1}^{n}\sum_{h=1}^{n} A_{ik}A_{jh}\text{Cov}[X_k, X_h] \\
&= [A\Sigma A^T]_{ij}
\end{aligned}$$

The last equality can be shown through the definition of matrix multiplication:

$$\begin{aligned}
[\Sigma A^T]_{ij} &= \sum_{h=1}^{n} \Sigma_{ih}[A^T]_{hj} = \sum_{h=1}^{n} \Sigma_{ih}A_{jh} \\
[A\Sigma A^T]_{ij} &= \sum_{k=1}^{n} A_{ik}[\Sigma A^T]_{kj} = \sum_{k=1}^{n} A_{ik}\sum_{h=1}^{n} \Sigma_{kh}A_{jh} = \sum_{k=1}^{n}\sum_{h=1}^{n} A_{ik}\Sigma_{kh}A_{jh}
\end{aligned}$$

Therefore, AX has an m-variate normal distribution with $AX \sim N_m(A\mu, A\Sigma A^T)$.

Question 2

Letting $Y = XZ$ where X and Z are independent,

$$\begin{aligned}\mathrm{Cov}[X, Y] &= \mathbb{E}[(X - \mathbb{E}[X])(XZ - \mathbb{E}[XZ])] \\ &= \mathbb{E}[(X - \mathbb{E}[X])(XZ - \mathbb{E}[X]\mathbb{E}[Z])] \\ &= \mathbb{E}[X^2 Z] - \mathbb{E}[X]^2\mathbb{E}[Z] \\ &= \mathbb{E}[X^2]\mathbb{E}[Z] - \mathbb{E}[X]^2\mathbb{E}[Z] \\ &= \mathrm{Var}[X]\mathbb{E}[Z] = 0 \implies \mathbb{E}[Z] = 0\end{aligned}$$

Letting Z be equal to 1 and -1 with probability $\frac{1}{2}$ each, we see that $\mathbb{E}[Z] = 0 \implies \mathrm{Cov}[X, Y] = 0$. Also, Y is normally distributed since

$$\Pr[Y \le y] = \Pr[XZ \le y] = \frac{1}{2}\Pr[X \le y] + \frac{1}{2}\Pr[X \ge -y] = \Phi(y)$$

given that $\Phi(y) = 1 - \Phi(-y)$. However, we have

$$\Pr[Y \le 1 \mid X = 1] = 1 \ne \Pr[Y \le 1] = \Phi(1)$$

for $x > 0$. Since $\Pr[Y \le y \mid X = x] \ne \Pr[Y \le y]$, the joint probability density function of X and Y cannot be equal to the product of the probability density functions of X and Y:

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) \implies F_{Y|X}(y, x) = \frac{\int_{-\infty}^{y} f(x, t)\mathrm{d}t}{f_X(x)} = \frac{f_X(x)\int_{-\infty}^{y} f_Y(t)\mathrm{d}t}{f_X(x)} = F_Y(y)$$

which means X and Y cannot be independent.

Question 3

The stacked vector $Y = \begin{pmatrix} PX \\ (I - P)X \end{pmatrix}$ has a $2n$-variate normal distribution, since for any $t_1, t_2 \in \mathbb{R}^n$,

$$t^T Y = \begin{pmatrix} t_1 & t_2 \end{pmatrix} \begin{pmatrix} PX \\ (I - P)X \end{pmatrix} = t_1^T PX + t_2^T (I - P)X = [t_1^T P + t_2^T (I - P)]X$$

where the above is normally distributed since X has a multivariate normal distribution. Since Y has a multivariate normal distribution which is fully characterised by its mean and covariance matrix, PX and $(I - P)X$ are independent if their covariance is zero, and we have

$$\mathrm{Cov}[PX, (I - P)X] = \mathbb{E}[PXX^T(I - P)^T] = \sigma^2 P(I - P) = 0$$

which shows the two vectors are independent.

Question 4

We define the following

$$A = \begin{pmatrix} I_{n_1} & \mathbf{0} \end{pmatrix}$$

where $\mathbf{0}$ is a $n_1 \times n - n_1$ matrix of zeros. We have $AX = X_1$, and the result from Question 1 tells us

$$X_1 = AX \sim N(A\mu, A\Sigma A^T)$$

We can see that $A\mu = \mu_1$, and

$$A\Sigma A^T = \begin{pmatrix} I_{n_1} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} I_{n_1} \\ \mathbf{0}^T \end{pmatrix} = \begin{pmatrix} I_{n_1} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \Sigma_{11} \\ \Sigma_{21} \end{pmatrix} = \Sigma_{11}$$

which means $X_1 \sim N(\mu_1, \Sigma_{11})$.

Question 5

Given a sample $Y = Y_1, \ldots, Y_n$, the likelihood function is

$$L(a, b, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} (Y_i - a - bx_i)^2 \right]$$

and taking logs, we get

$$\ell(a, b, \sigma^2) = \log L(a, b, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - a - bx_i)^2$$

The MLEs satisfy the following first-order conditions:

$$\frac{\partial \ell}{\partial a} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - a - bx_i) = 0 \qquad \implies \hat{a} = \frac{\sum_{i=1}^{n} Y_i - \hat{b} \sum_{i=1}^{n} x_i}{n} = \bar{Y}$$

$$\frac{\partial \ell}{\partial b} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i (Y_i - a - bx_i) = 0 \qquad \implies \hat{b} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y}) x_i}{\sum_{i=1}^{n} x_i^2}$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (Y_i - a - bx_i)^2 = 0 \qquad \implies \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b} x_i)^2$$

The simple linear regression model is equivalent to the general linear regression model with the following design matrix

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

and the following coefficient vector

$$\beta = \begin{pmatrix} a \\ b \end{pmatrix}$$

By brute force, we can find the expression for X^TX:

$$X^TX = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

The determinant of the above is

$$\det(X^TX) = n\sum_{i=1}^n x_i^2 - n^2\bar{x}^2 = n\sum_{i=1}^n (x_i - \bar{x})^2$$

Therefore, the inverse of X^TX is

$$(X^TX)^{-1} = \begin{pmatrix} \frac{1}{n}\frac{\sum x_i^2}{\sum (x_i-\bar{x})^2} & -\frac{\bar{x}}{\sum (x_i-\bar{x})^2} \\ -\frac{\bar{x}}{\sum (x_i-\bar{x})^2} & \frac{1}{\sum (x_i-\bar{x})^2} \end{pmatrix}$$

Again, by brute force, we can find the expression of X^TY:

$$X^TY = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum Y_i x_i \end{pmatrix} = \begin{pmatrix} n\bar{Y} \\ \sum Y_i x_i \end{pmatrix}$$

Finally, we get the expression for $\hat{\beta}$:

$$\begin{aligned} \hat{\beta} = (X^TX)^{-1}X^TY &= \begin{pmatrix} \frac{1}{n}\frac{\sum x_i^2}{\sum (x_i-\bar{x})^2} & -\frac{\bar{x}}{\sum (x_i-\bar{x})^2} \\ -\frac{\bar{x}}{\sum (x_i-\bar{x})^2} & \frac{1}{\sum (x_i-\bar{x})^2} \end{pmatrix} \begin{pmatrix} n\bar{Y} \\ \sum Y_i x_i \end{pmatrix} \\ &= \begin{pmatrix} \frac{\bar{Y}\sum x_i^2 - \bar{x}\sum Y_i x_i}{\sum (x_i-\bar{x})^2} \\ \frac{\sum Y_i x_i - n\bar{Y}\bar{x}}{\sum (x_i-\bar{x})^2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{\bar{Y}\sum x_i^2 - \bar{x}\sum Y_i x_i}{\sum (x_i-\bar{x})^2} \\ \frac{\sum (Y_i-\bar{Y})(x_i-\bar{x})}{\sum (x_i-\bar{x})^2} \end{pmatrix} \\ &= \begin{pmatrix} \bar{Y} - \bar{x}\frac{\sum (Y_i-\bar{Y})(x_i-\bar{x})}{\sum (x_i-\bar{x})^2} \\ \frac{\sum (Y_i-\bar{Y})(x_i-\bar{x})}{\sum (x_i-\bar{x})^2} \end{pmatrix} \end{aligned}$$

where we used $\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x}) = \sum_{i=1}^n Y_i x_i - n\bar{Y}\bar{X}$ by expansion, which also means

$$\bar{Y} - \bar{x}\frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \bar{Y} - \bar{x}\frac{\sum_{i=1}^n Y_i x_i - n\bar{Y}\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\bar{Y}\sum_{i=1}^n x_i^2 - \bar{x}\sum_{i=1}^n Y_i x_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Therefore, our expression for $\hat{\beta}$ means that

$$\begin{aligned} \hat{b} &= \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})} = \frac{S_{XY}}{S_{XX}} \\ \hat{a} &= \bar{Y} - \hat{b}\bar{x} \\ \hat{\sigma}^2 &= n^{-1}\|Y - X\hat{\beta}\|^2 \end{aligned}$$

which is the general formula for the simple linear regression model. With demeaned regressors such that $\bar{x} = 0$, we get the expressions from before.

Question 6

The model is linear in $\sin(2\alpha)$ with no intercept. For a linear model $Y_i = \beta X_i + \varepsilon_i$ with no intercept, the log-likelihood function given a sample $Y = Y_1, \ldots, Y_n$ is

$$\ell(\beta, Y) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta X_i)^2$$

The maximum-likelihood estimator of β satisfies the first-order condition

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{n} X_i(Y_i - \beta X_i) = 0 \implies \hat{\beta} = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}$$

In the example given, $\beta = \frac{v^2}{g}$, and $v = \sqrt{\beta g} = f(\beta)$. Since f is an injective function, the maximum-likelihood estimator of v is $f(\hat{\beta})$, which is equal to

$$\hat{v} = f(\hat{\beta}) = f\left(\frac{\sum_{i=1}^{n} m_i \sin(2\alpha_i)}{\sum_{i=1}^{n} \sin^2(2\alpha_i)}\right) \approx f(28155.9564) \approx 525.557$$

Question 7

This is a fixed-effects model with no regressors. Given a sample $Y = Y_{1,1}, \ldots, Y_{1,n_1} \ldots, Y_{I,1}, \ldots Y_{I,n_I}$, the log-likelihood function is

$$\ell(\mu, Y) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \mu_i)^2$$

where $n = \sum_{j=1}^{I} n_i$. The MLEs satisfy the first-order conditions

$$\frac{\partial \ell}{\partial \mu_k} = \frac{1}{\sigma^2}\sum_{j=1}^{n_k}(Y_{kj} - \mu_k) = 0 \qquad \implies \hat{\mu}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} Y_{kj} = \bar{Y}_k$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \mu_i)^2 = 0 \qquad \implies \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$$

As mentioned, this is a fixed-effects model, and we get the same estimates for β and σ^2 by running a regression of Y on a set of dummy variables for every $i \in \{1, \ldots, I\}$. The $n \times I$ design matrix is

$$X = \begin{pmatrix} i_{n_1} & 0_{n_1} & \ldots & 0_{n_1} \\ 0_{n_2} & i_{n_2} & \ldots & 0_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_I} & 0_{n_I} & \ldots & i_{n_I} \end{pmatrix}$$

where $i_{n_i} \in \mathbb{R}^{n_i}$ is a vector of ones and $0_{n_i} \in \mathbb{R}^{n_i}$ is a vector of zeroes. The coefficient vector is

$$\beta = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_I \end{pmatrix}$$

Again, through brute force we get an expression for X^TX:

$$X^TX = \begin{pmatrix} i_{n_1}^T & 0_{n_2}^T & \cdots & 0_{n_I}^T \\ 0_{n_1}^T & i_{n_2}^T & \cdots & 0_{n_I}^T \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_1}^T & 0_{n_2}^T & \cdots & i_{n_I}^T \end{pmatrix} \begin{pmatrix} i_{n_1} & 0_{n_1} & \dots & 0_{n_1} \\ 0_{n_2} & i_{n_2} & \dots & 0_{n_2} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_I} & 0_{n_I} & \dots & i_{n_I} \end{pmatrix} = \begin{pmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_I \end{pmatrix}$$

The inverse is straightforward to calculate:

$$(X^TX)^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{n_I} \end{pmatrix}$$

As before, we get an expression for X^TY:

$$X^TY = \begin{pmatrix} i_{n_1}^T & 0_{n_2}^T & \cdots & 0_{n_I}^T \\ 0_{n_1}^T & i_{n_2}^T & \cdots & 0_{n_I}^T \\ \vdots & \vdots & \ddots & \vdots \\ 0_{n_1}^T & 0_{n_2}^T & \cdots & i_{n_I}^T \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_I \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n_1} Y_{1j} \\ \vdots \\ \sum_{j=1}^{n_I} Y_{Ij} \end{pmatrix}$$

where Y is a stacked vector with $Y_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix}$. Combining these expressions, we get the expression for $\hat{\beta}$

$$\hat{\beta} = (X^TX)^{-1}X^TY = \begin{pmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{n_I} \end{pmatrix} \begin{pmatrix} \sum_{j=1}^{n_1} Y_{1j} \\ \vdots \\ \sum_{j=1}^{n_I} Y_{Ij} \end{pmatrix} = \begin{pmatrix} \frac{1}{n_1}\sum_{j=1}^{n_1} Y_j \\ \vdots \\ \frac{1}{n_I}\sum_{j=1}^{n_I} Y_j \end{pmatrix} = \begin{pmatrix} \bar{Y}_1 \\ \vdots \\ \bar{Y}_I \end{pmatrix}$$

We then end up with $\hat{\sigma}^2$ by plugging in the value of $\hat{\beta}$ into the following

$$\hat{\sigma}^2 = n^{-1}\|Y - X\hat{\beta}\|^2 = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$$

Question 8

This is a normal linear model with heteroskedastic errors. The Generalised Least Squares estimator is equal to the estimated coefficients from a weighted least squares regression of $\Sigma^{-\frac{1}{2}}Y$ on $\Sigma^{-\frac{1}{2}}X$. Pre-multiplying the model equation by $\Sigma^{-\frac{1}{2}}$, we get

$$\Sigma^{-\frac{1}{2}}Y = \Sigma^{-\frac{1}{2}}X\beta + \Sigma^{-\frac{1}{2}}\varepsilon$$

The re-weighted equation has spherical errors $\Sigma^{-\frac{1}{2}}\varepsilon$, since the covariance matrix of the composite errors is

$$\text{Cov}[\Sigma^{-\frac{1}{2}}\varepsilon] = \mathbb{E}\left[\Sigma^{-\frac{1}{2}}\varepsilon\varepsilon^T\Sigma^{-\frac{1}{2}}\right] = \sigma^2\Sigma^{-\frac{1}{2}}\Sigma\Sigma^{-\frac{1}{2}} = \sigma^2 I$$

where we used $\Sigma^{-\frac{1}{2}^T} = \Sigma^{-\frac{1}{2}}$ since covariance matrices are symmetric. Therefore the re-weighted errors are homoskedastic and uncorrelated, and the Gauss-Markov assumptions are satisfied. This means the weighted least squares estimate of β given the model $\Sigma^{-\frac{1}{2}}Y = \Sigma^{-\frac{1}{2}}X\beta + \Sigma^{-\frac{1}{2}}\varepsilon$ is the best linear unbiased estimator, and this is equal to

$$\hat{\beta}^{WLS} = \left[\left(\Sigma^{-\frac{1}{2}}X\right)^T\Sigma^{-\frac{1}{2}}X\right]^{-1}\left(\Sigma^{-\frac{1}{2}}X\right)^T\Sigma^{-\frac{1}{2}}Y = \left(X^T\Sigma^{-1}X\right)^{-1}X^T\Sigma^{-1}Y$$

But this is precisely the estimate that minimises the following over β:

$$\hat{\beta}^{WLS} = \arg\min_{\beta} \|\Sigma^{-\frac{1}{2}}(Y - X\beta)\|^2 = \arg\min_{\beta}(Y - X\beta)\Sigma^{-1}(Y - X\beta) = \tilde{\beta}^{GLS}$$

again, using the fact that $\Sigma^{-\frac{1}{2}^T} = \Sigma^{-\frac{1}{2}}$. Therefore, the Generalised Least Squares estimator is unbiased and efficient.

Question 9

(a)

The null hypothesis is that conditional on all other $p - 1$ characteristics, characteristic j does not help to predict the health measurement of a patient. Specifically, there is no difference in the means of Y between patients with characteristic j and patients without, conditional on the values of all other characteristics.

The likelihood function given a sample Y is

$$L(\beta, \sigma^2|Y) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(Y_i - x_i\beta)^2\right]$$

where x_i is the i^{th} row of X. Taking logs, we get

$$\ell(\beta, \sigma^2|Y) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|Y - X\beta\|^2$$

This is maximised over β by the minimiser of $\|Y - X\beta\|^2$, which is the OLS estimate $\hat{\beta}$. The MLE of σ^2 satisfies

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\|Y - X\hat{\beta}\|^2 = 0 \implies \hat{\sigma}^2 = \frac{1}{n}\|Y - X\hat{\beta}\|^2$$

Without loss of generality, we assume $j = 1$ (which is the same as swapping the 1^{st} and j^{th} columns in X and the 1^{st} and j^{th} entries of β). This means we can express the model in terms of the partitioned matrices

$$Y = \begin{pmatrix} X_1 & X_2 \end{pmatrix}\begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix} + \varepsilon$$

where X_1 is a $n \times 1$ vector containing the values of the first characteristic, and X_2 is the $n \times (p-1)$ matrix containing the values of all other characteristics. Likewise, β_1 is a scalar and $\beta_2 \in \mathbb{R}^{p-1}$. Under H_0, the true model is assumed to be

$$Y = \begin{pmatrix} X_1 & X_2 \end{pmatrix} \begin{pmatrix} 0 \\ \beta_2 \end{pmatrix} + \varepsilon = X_2\beta_2 + \varepsilon$$

And likewise, the derivation above implies the likelihood function is maximised by

$$\tilde{\beta} = \begin{pmatrix} 0 \\ (X_2^T X_2)^{-1} X_2^T Y \end{pmatrix} = \begin{pmatrix} 0 \\ \tilde{\beta}_2 \end{pmatrix}$$
$$\tilde{\sigma}^2 = \frac{1}{n} \| Y - X_2 \tilde{\beta}_2 \|^2$$

Therefore, the likelihood ratio is

$$\Lambda(H_0, H_1) = \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left[-\frac{1}{2\hat{\sigma}^2} (Y_i - x_i\hat{\beta})^2 \right]}{\prod_{i=1}^n \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \exp\left[-\frac{1}{2\tilde{\sigma}^2} (Y_i - x_i\tilde{\beta})^2 \right]}$$

We can simplify this as such

$$\begin{aligned} \Lambda(H_0, H_1) &= \left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \right)^{\frac{n}{2}} \frac{\exp\left(-\frac{1}{2\hat{\sigma}^2} \| Y - X\hat{\beta} \|^2 \right)}{\exp\left(-\frac{1}{2\tilde{\sigma}^2} \| Y - X_2\tilde{\beta}_2 \|^2 \right)} \\ &= \left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \right)^{\frac{n}{2}} \frac{\exp\left(-\frac{1}{2} \right)}{\exp\left(-\frac{1}{2} \right)} \\ &= \left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \right)^{\frac{n}{2}} = \left(\frac{\| Y - X\tilde{\beta} \|^2}{\| Y - X\hat{\beta} \|^2} \right)^{\frac{n}{2}} = \left(1 + \frac{\| Y - X\tilde{\beta} \|^2 - \| Y - X\hat{\beta} \|^2}{\| Y - X\hat{\beta} \|^2} \right)^{\frac{n}{2}} \end{aligned}$$

The likelihood ratio is increasing in $\frac{RSS_0 - RSS}{RSS}$, where RSS_0 and RSS are the residual sum of squares under the restricted and unrestricted model. With some linear algebra we can find that under H_0 we have $RSS_0 - RSS \sim \sigma^2 \chi_1^2$ and $RSS \sim \sigma^2 \chi_{n-p}^2$, and the two are independent.

Firstly, given a symmetric idempotent $n \times n$ matrix A and a standard normal vector $Z \in \mathbb{R}^n$, we have

$$\| AZ \|^2 = Z^T A^T A Z = Z^T A Z = Z^T U \Lambda U^T Z$$

where A can be diagonalised to get $A = U\Lambda U^T$ since A is symmetric. $U^T Z$ is also a standard normal vector, since

$$\mathrm{Cov}[U^T Z] = \mathbb{E}[U^T Z Z^T U] = U^T \sigma^2 I_n U = \sigma^2 I_n$$

Therefore,

$$\| AZ \|^2 = \sum_{i=1}^n \lambda_i w_i^2$$

where w_i is the i^{th} element of $U^T Z$. This is the sum of $\mathrm{rank}(A)$ independent squared standard normal variables, which has a χ^2 distribution with $\mathrm{rank}(A)$ degrees of freedom.

Now we can show that $RSS_0 - RSS \sim \sigma^2\chi^2_1$ and $RSS \sim \sigma^2\chi^2_{n-p}$ under H_0, where the two are independent. Letting $P = X(X^TX)^{-1}X^T$ and $P_2 = X_2(X_2^TX_2)^{-1}X_2^T$, We have

$$RSS_0 - RSS = \|(I-P_2)Y\|^2 - \|(I-P)Y\|^2 = Y^T(I-P_2)Y - Y^T(I-P)Y = Y^T(P-P_2)Y = \|(P-P_2)Y\|^2$$

since $I - P_2$ and $I - P$ are symmetric and idempotent.

The matrix $(P - P_2)$ is also symmetric and idempotent, since

$$(P-P_2)^T = P^T - P_2^T = P - P_2$$
$$(P-P_2)(P-P_2) = P^2 - PP_2 - P_2P + P_2^2 = P - 2P_2 + P_2 = P - P_2$$

where we used $PP_2 = P_2P = P_2$ since P_2 projects onto the column span of X_2, and is preserved under P which projects onto the column span of $\begin{pmatrix} X_1 & X_2 \end{pmatrix}$. Under H_0, $Y = X_2\beta_2$, and

$$\|(P-P_2)Y\|^2 = \|(P-P_2)(X_2\beta_2+\varepsilon)\|^2 = \|(P-P_2)\varepsilon\|^2 = \sigma^2\left\|(P-P_2)\frac{\varepsilon}{\sigma}\right\|^2 \sim \sigma^2\chi^2_1$$

using the result above with $(P-P_2)$ being symmetric and idempotent, $\frac{\varepsilon}{\sigma}$ being a standard normal vector, and $\text{rank}(P - P_2) = \text{tr}(P - P_2) = p - (p - 1) = 1$. Doing the same steps, we get

$$RSS = \|(I-P)Y\|^2 \sim \sigma^2\chi^2_{n-p}$$

Now we have to show $RSS_0 - RSS$ and RSS are independent. We have

$$\begin{pmatrix}(P-P_2)\varepsilon \\ (I-P)\varepsilon\end{pmatrix} \sim N_{2n}\left(0, \sigma^2\begin{pmatrix}(P-P_2)(P-P_2)^T & (P-P_2)(I-P)^T \\ (I-P)(P-P_2)^T & (I-P)(I-P)^T\end{pmatrix}\right) \equiv N_{2n}\left(0, \sigma^2\begin{pmatrix}P-P_2 & 0 \\ 0 & I-P\end{pmatrix}\right)$$

which shows that $(P - P_2)\varepsilon$ and $(I - P)\varepsilon$ are independent, and since $RSS_0 - RSS$ and RSS are functions of the former two, they are also independent.

After all that, we can show that

$$\frac{(RSS_0 - RSS)/1}{RSS/(n-p)} \sim \frac{\sigma^2\chi^2_1/1}{\sigma^2\chi^2_{n-p}/(n-p)} \equiv \frac{\chi^2_1/1}{\chi^2_{n-p}/(n-p)} \equiv F_{1,n-p} \equiv t^2_{n-p}$$

So we can use either the critical value from the F-distribution or the two-sided critical values from the t-distribution to get the test of exact size α. We reject the null hypothesis if the test statistic $\frac{(RSS_0-RSS)/1}{RSS/(n-p)}$ is greater than $F_{1,n-p}(\alpha)$ where $F_{1,n-p}(\alpha)$ is the upper $100\alpha\%$ critical value for the $F_{1,n-p}$ distribution.

(b)

The test result suggests that conditional on all other attributes, attribute j^* is associated with, and helps to predict, the health measurement. The strength of the evidence for this grows in the number of characteristics p, since the residual sum of squares must decrease as covariates are added to the model. If characteristic j^* remains significant after most of the variation in the data has been 'explained' by adding many covariates, it means that there is variation in j^* that is associated strongly enough with the health measurement and is sufficiently uncorrelated with all other characteristics.

However, if it is the magnitude of the association we are concerned with, that is, how much variation in the health measurement is associated with j^*, then the test result is weaker as the number of characteristics grows, for the same reason that there is less variation to explain after many characteristics have been added to the model.

Question 10

Under the null hypothesis we assume equality of means, so we have $H_0 : \mu_1 = \ldots = \mu_I = \mu_0$. Under the model prescribed by H_0 the maximum likelihood estimate of μ_0 is simply the estimate from running a regression of Y_{ij} with an intercept and no regressors, so $\hat{\mu}_0 = \bar{Y}$. Likewise we have $\hat{\sigma}_0^2 = n^{-1}\|Y - \hat{\mu}_0 X\|^2 = n^{-1}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2$. Therefore, using our MLEs from Question 7,

$$\begin{aligned}
\Lambda(H_0, H_1) &= \frac{\prod_{i=1}^{I}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi\hat{\sigma}_1^2}}\exp\left[-\frac{1}{2\hat{\sigma}_1^2}(Y_{ij} - \bar{Y}_i)^2\right]}{\prod_{i=1}^{I}\prod_{j=1}^{n_i}\frac{1}{\sqrt{2\pi\hat{\sigma}_0^2}}\exp\left[-\frac{1}{2\hat{\sigma}_0^2}(Y_{ij} - \bar{Y})^2\right]} \\
&= \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}\right)^{\frac{n}{2}} \\
&= \left(\frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2}\right)^{\frac{n}{2}}
\end{aligned}$$

We have

$$\begin{aligned}
\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y})^2 &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i + \bar{Y}_i - \bar{Y})^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 + 2\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)(\bar{Y}_i - \bar{Y}) + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\bar{Y}_i - \bar{Y})^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 + 2\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)(\bar{Y}_i - \bar{Y}) + \sum_{i=1}^{I} n_i(\bar{Y}_i - \bar{Y})^2
\end{aligned}$$

The middle term disappears since

$$\begin{aligned}
\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)(\bar{Y}_i - \bar{Y}) &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij}\bar{Y}_i - Y_{ij}\bar{Y} - \bar{Y}_i^2 + \bar{Y}_i\bar{Y}) \\
&= \sum_{i=1}^{I}\bar{Y}_i\sum_{j=1}^{n_i}Y_{ij} - \bar{Y}\sum_{i=1}^{I}\sum_{j=1}^{n_i}Y_{ij} - \sum_{i=1}^{I}\sum_{j=1}^{n_i}\bar{Y}_i^2 + \bar{Y}\sum_{i=1}^{I}\sum_{j=1}^{n_i}\bar{Y}_i \\
&= \sum_{i=1}^{I} n_i\bar{Y}_i^2 - \bar{Y}\sum_{i=1}^{I} n_i\bar{Y}_i - \sum_{i=1}^{I} n_i\bar{Y}_i^2 + \bar{Y}\sum_{i=1}^{I} n_i\bar{Y}_i = 0
\end{aligned}$$

Therefore,

$$\Lambda(H_0, H_1) = \left(\frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{I} n_i(\bar{Y}_i - \bar{Y})^2}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2}\right)^{\frac{n}{2}} = \left(1 + \frac{\sum_{i=1}^{I} n_i(\bar{Y}_i - \bar{Y})^2}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2}\right)^{\frac{n}{2}}$$

Therefore, the likelihood ratio is monotonic in

$$T = \frac{\frac{1}{I-1}\sum_{i=1}^{I} n_i(\bar{Y}_i - \bar{Y})^2}{\frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2}$$

and we reject the null hypothesis given that T is large enough. It remains to be shown that T follows an $F_{I-1,n-I}$ distribution. We denote the response vector by $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_I \end{pmatrix}$ where Y_i is the $n_i \times 1$ vector with j^{th} entry equal to Y_{ij}, and the design matrix by $X = \begin{pmatrix} X_1 \\ \vdots \\ X_I \end{pmatrix}$ and X_i is the $n_i \times I$ matrix with entries in column i all equal to one and zero otherwise. Letting $P = X(X^TX)^{-1}X^T$ be the projection matrix onto the column span of X, we get

$$PY = \begin{pmatrix} \bar{Y}_{1,n_1} \\ \vdots \\ \bar{Y}_{I,n_I} \end{pmatrix}$$

where $\bar{Y}_{i,n_i}$ is the $n_i \times 1$ vector with values all equal to $\bar{Y}_i$. Likewise, letting $P_0 = X_0(X_0^TX_0)^{-1}X_0^T$ be the projection matrix onto the column span of $X_0 = \mathbf{1}_n$ where $\mathbf{1}_n$ is a $n \times 1$ vector of ones, we get

$$P_0Y = \begin{pmatrix} \bar{Y} \\ \vdots \\ \bar{Y} \end{pmatrix}$$

where P_0 is $n \times 1$. Then,

$$\|(P - P_0)Y\|^2 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{I} n_i (\bar{Y}_i - \bar{Y})^2$$

Likewise,

$$\|(I_n - P)Y\|^2 = \sum_{I=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

From the expressions of P and P_0, we can see that they are symmetric and idempotent. Furthermore, under H_0, we have $Y = \mu_0 X_0$, and

$$\|(P - P_0)Y\|^2 = \|(P - P_0)(\mu_0 X_0 + \varepsilon)\|^2 = \|(P - P_0)\varepsilon\|^2$$
$$\|(I_n - P)Y\|^2 = \|(I_n - P)(\mu_0 X_0 + \varepsilon)\|^2 = \|(I_n - P)\varepsilon\|^2$$

since P_0 projects onto the column span of X_0 and preserves X_0, while P also preserves X_0 since P projects onto the column span of X, and the columns of X sum to $X_0 = \mathbf{1}_n$. Therefore, from our previous result,

$$\|(P - P_0)\varepsilon\|^2 = \sigma^2 \left\| (P - P_0)\frac{\varepsilon}{\sigma} \right\|^2 \sim \sigma^2 \chi^2_{I-1}$$
$$\|(I_n - P)\varepsilon\|^2 = \sigma^2 \left\| (I_n - P)\frac{\varepsilon}{\sigma} \right\|^2 \sim \sigma^2 \chi^2_{n-I}$$

since we have $\text{rank}(P - P_0) = \text{tr}(P - P_0) = \text{tr}(P) - \text{tr}(P_0) = I - 1$, where we used the fact that $\text{tr}(P) = \text{tr}(X(X^TX)^{-1}X^T) = \text{tr}(X^TX(X^TX)^{-1}) = \text{tr}(I_I) = I$, and similarly for P_0 and I_n.

Lastly, $\|(P - P_0)\varepsilon\|^2$ and $\|(I_n - P)\varepsilon\|^2$ are independent, since

$$\begin{pmatrix}(P - P_0)\varepsilon \\ (I_n - P)\varepsilon\end{pmatrix} \sim N_{2n}\left(0, \sigma^2\begin{pmatrix}(P - P_0)^2 & (P - P_0)(I_n - P) \\ (I_n - P)(P - P_0) & (I_n - P)^2\end{pmatrix}\right) \equiv N_{2n}\left(0, \sigma^2\begin{pmatrix}P - P_0 & 0 \\ 0 & I_n - P\end{pmatrix}\right)$$

Therefore, the arguments in Question 9(a) apply, and T has an $F_{I-1,n-I}$ distribution.

Question 11

We have

$$\hat{\beta} = (X^TX)^{-1}X^TY = (X^TX)^{-1}X^T(X\beta + \varepsilon) = \beta + (X^TX)^{-1}X^T\varepsilon$$

so $(X^TX)^{\frac{1}{2}}(\hat{\beta} - \beta) = (X^TX)^{-\frac{1}{2}}X^T\varepsilon$. This is a linear function of ε, and therefore has a multivariate normal distribution with mean zero and the following covariance matrix

$$\text{Cov}[(X^TX)^{\frac{1}{2}}(\hat{\beta} - \beta)] = \mathbb{E}[(X^TX)^{-\frac{1}{2}}X^T\varepsilon\varepsilon^TX(X^TX)^{-\frac{1}{2}}] = (X^TX)^{-\frac{1}{2}}X^T\sigma^2I_nX(X^TX)^{-\frac{1}{2}} = \sigma^2I_p$$

Therefore, for the quadratic form we have

$$(\hat{\beta} - \beta)^TX^TX(\hat{\beta} - \beta) = \|(X^TX)^{\frac{1}{2}}(\hat{\beta} - \beta)\|^2 = \sigma^2\left\|\frac{(X^TX)^{\frac{1}{2}}(\hat{\beta} - \beta)}{\sigma}\right\|^2 \sim \sigma^2\chi^2_p$$

Using the same definitions as before, $PY = X\hat{\beta}$, which means $\hat{\beta} = (X^TX)^{-1}X^TPY$ depends on the data through PY. Likewise, $\hat{\sigma}^2 = n^{-1}\|(I - P)Y\|^2$ which means $\hat{\sigma}^2$ depends on the data through $(I - P)Y$. Using similar arguments from before, we have

$$\hat{\sigma}^2 = \frac{\sigma^2}{n}\left\|(I_n - P)\frac{\varepsilon}{\sigma}\right\|^2 \sim \frac{\sigma^2}{n}\chi^2_{n-p}$$

We can show that $\hat{\beta}$ and $\hat{\sigma}^2$ are independent using arguments we've already used multiple times, so we can note that

$$T(\beta) = \frac{(\hat{\beta} - \beta)^TX^TX(\hat{\beta} - \beta)/p}{n\hat{\sigma}^2/(n - p)} \sim F_{p,n-p}$$

And for our $100(1 - \alpha)\%$ confidence set, we can just take

$$C(Y) = \{b \in \mathbb{R}^p : T(b) \leq F_{p,n-p}(\alpha)\}$$

since

$$\Pr[\beta \in C(Y)] = \Pr[T \leq F_{p,n-p}(\alpha)] = 1 - \alpha$$

The shape of this confidence set is an ellipsoid centred around $\hat{\beta}$.

Question 12

From R, we run a regression of `weights` on a set of dummy variables for `feed` with no intercept, and we get

$$X^TX = \begin{pmatrix} & \text{Casein} & \text{Horsebean} & \text{Linseed} & \text{Meatmeal} & \text{Soybean} & \text{Sunflower} \\ \text{Casein} & 12 & 0 & 0 & 0 & 0 & 0 \\ \text{Horsebean} & 0 & 10 & 0 & 0 & 0 & 0 \\ \text{Linseed} & 0 & 0 & 12 & 0 & 0 & 0 \\ \text{Meatmeal} & 0 & 0 & 0 & 11 & 0 & 0 \\ \text{Soybean} & 0 & 0 & 0 & 0 & 14 & 0 \\ \text{Sunflower} & 0 & 0 & 0 & 0 & 0 & 12 \end{pmatrix}$$

$$\hat{\beta} = \begin{pmatrix} \text{Feed} & \text{Estimate} \\ \text{Casein} & 323.583 \\ \text{Horsebean} & 160.200 \\ \text{Linseed} & 218.750 \\ \text{Meatmeal} & 276.909 \\ \text{Soybean} & 246.429 \\ \text{Sunflower} & 328.917 \end{pmatrix}$$

where we include labels for clarity. From the above, we have

$$\begin{aligned}(\hat{\beta}-\beta)^TX^TX(\hat{\beta}-\beta) = {} & 12(323.58-\beta_1)^2 + 10(160.20-\beta_2)^2 + 12(218.75-\beta_3)^2 \\ & + 11(276.91-\beta_4)^2 + 14(246.43-\beta_5)^2 + 12(328.92-\beta_6)^2\end{aligned}$$

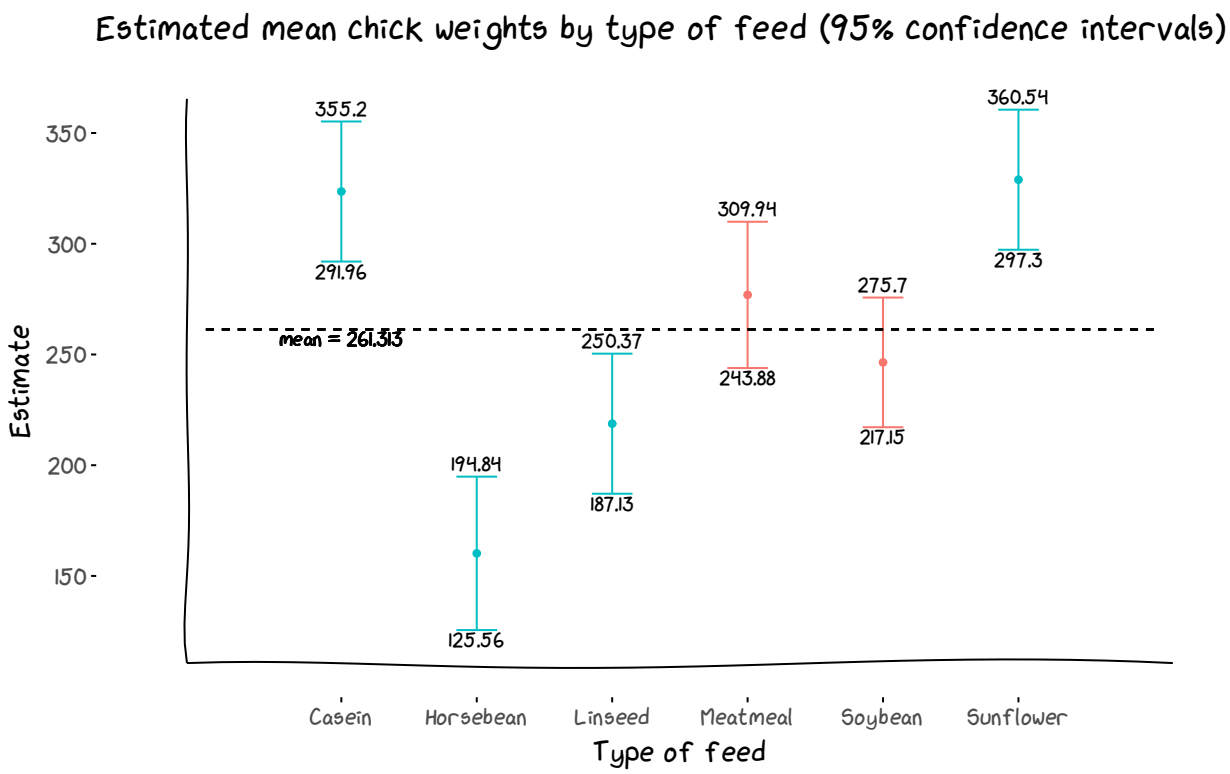
From R, we have $\hat{\sigma}^2 = n^{-1}\|RSS\|^2 = 2754.31$, which means

$$\tilde{\sigma}^2 = n\hat{\sigma}^2/(n-p) = (71-6)^{-1}\|RSS\|^2 = 3008.554$$

and the 95% confidence region is given by values of $\beta_1, \ldots, \beta_6$ such that

$$\begin{aligned}&12(323.58-\beta_1)^2 + 10(160.20-\beta_2)^2 + 12(218.75-\beta_3)^2 \\ &\quad + 11(276.91-\beta_4)^2 + 14(246.43-\beta_5)^2 + 12(328.92-\beta_6)^2 && \leq p\tilde{\sigma}^2F_{p,n-p}(\alpha) \\ & && = 6(3008.554)(2.242) \\ & && = 40471.068\end{aligned}$$

The vector $\beta = 0$ is outside of this confidence region, which means that $\hat{\beta}$ is jointly significant from zero at the 5% level. To carry out the individual tests of significance, we can use the fact that $\frac{\hat{\beta}_j-\beta_j}{\tilde{\sigma}[(X^TX)^{-1}]_{jj}^{\frac{1}{2}}} \sim t_{n-p}$ by taking the j^{th} element of $\hat{\beta}-\beta$ since we've shown before that the latter has a multivariate normal distribution with covariance matrix $\sigma^2(X^TX)^{-1}$, and $\hat{\beta}$ and $\tilde{\sigma}^2$ are independent. The standard errors can be calculated by R, so for the confidence intervals we just have to take $[\hat{\beta}_j - se(\hat{\beta}_j)t_{n-p}(\alpha/2), \hat{\beta}_j + se(\hat{\beta}_j)t_{n-p}(\alpha/2)]$ where *se* denotes the standard error. We have $t_{n-p}(\alpha/2) = t_{71-6=65}(0.025) = 1.997$, so we can calculate the confidence intervals directly, and we get the following:



We can see from the above that the 95% confidence intervals for the estimated means of chicks fed on casein, horsebean, linseed, and sunflower-based feeds exclude the estimated value of the unconditional mean (which is the same as the estimate from the null model). This also says that the estimated means for chicks fed on linseed and meatmeal-based feeds are individually (not jointly) significantly different from 261.313, but this does not mean that they are significantly different from the estimate in the null model; the unconditional mean is unknown and its estimate has its own sampling distribution, and to test for differences this must be taken into account rather than taking the estimated value 261.313 as a deterministic constant.