Homework 2 | Part 2

Problem 2: Given vectors height, girth, and volume that reflects those traits on $n = 31$ columns at a historical building. Here girth is the circumference of the base of the column.

→ The following was run: fit1 ← lm(Height ~ Girth)

a) Write down the model & assumptions of model fit1, clearly define all notation

$$\text{fit1} = y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad , \quad i = 1, 2, \ldots, 31$$

where..
- β_0 = the intercept / predicted height when girth = 0
- β_1 = how much the estimated height will increase/decrease per 1 unit increase/decrease in x_i (girth)
- ϵ_i = the random error term
- y_i = predicted height of the ith column
- x_i = girth of the ith column

Assumptions of the linear model:
- errors ϵ_i are assumed to be i.i.d & obs. are independent
- homoscedasticity, the errors have constant variance
- errors $\epsilon_i \sim N(0, \sigma^2)$
- Linearity, the relationship between X & mean of Y is linear

b) Write down the model & assumptions for fit3 in matrix form.

$$\text{fit3} = Y = X\beta + \epsilon$$

where...
- Y = the vector of observed response values (heights)
- X = the design matrix containing the predictor features (girth & height)
- β = the vector of parameters including the intercept & slope for each parameter
- ϵ = the vector of errors

↳ expanded...

$$\begin{pmatrix} height_1 \\ \vdots \\ height_{31} \end{pmatrix} = \begin{pmatrix} 1 & girth_1 & Volume_1 \\ \vdots & \vdots & \vdots \\ 1 & girth_{31} & Volume_{31} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{31} \end{pmatrix} \quad i = 1, \ldots, 31$$

Assumptions of the linear model:

- errors ϵ_i are assumed to be i.i.d & obs. are independent
- homoskedasticity, the errors have constant variance
- errors $\epsilon_i \sim N(0, \sigma^2)$
- Linearity, the relationship between X & mean of Y is linear
- No perfect multi-collinearity between predictors

c) Is fit 1 nested in fit 3? Answer yes or no.

Yes.

d) Now write carefully (using clearly defined stat. formula, dist. notation, etc. how you would test the null hypothesis vs. alternative hypothesis...

H_0: model in fit 1

H_a: model in fit 3

... at significance level $\alpha = 0.01$. Clearly define the test statistic you are using, its dist. under the null hyp., and decision rule testing at $\alpha = 0.01$.

$H_0: \beta_2 = 0 \qquad H_a: \beta_2 \neq 0$

↳ Use an F-test

$$F = \frac{RSS_1 - RSS_3 / 3-2}{RSS_3 / n-3}$$

under H_0

↳ $F \sim F_{1,28}$

$F_{1,28,.99}$

$F^* = 7.636$

Decision: Reject the null if $F > F^*$ and fail to reject H_0 o.w.

Problem 3: For homoskedastic (constant variance) uncorrelated Gaussian linear reg., you wrote models as...

$$Y = X\beta + \epsilon \quad \text{where} \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

a) If X is $n \times p$ where $p = 6$, $n > p$, $n > p$, and $\text{rank}(X) = 5$,

No, we cannot estimate the parameter vector β uniquely by OLS since a rank of 5 indicates that one of the columns of X is linearly dependent, and since it is less than parameters p the matrix is not invertible. Therefore, the linear system does not have a unique solution.

b) What link function does the model use?

$\eta_i = g(\mu_i)$, $\mu_i = X_i^T \beta$ which implies the link is the identity func.

c) What is the natural exponential family parameter ϕ in terms of notation above.

$$f(y; \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right)$$

ϕ corresponds to the variance σ^2 of the error terms

d) What is the distribution of the ith element of random vector Y, where X_i^T is the ith row of matrix X

$$Y_i \sim (X_i^T \beta, \sigma^2)$$