MATH 6020 – HW 1

Due back on Saturday, 13th February [1].

-1- Generate a data set for the following model that you can replicate.

$$Y_i = \beta_0 + \beta_1 z_i + \epsilon_i, \text{ for } i = 1, \ldots, n$$

Let, $\mathbb{E}\{\epsilon_i\} = 0$, $\mathbb{V}\text{ar}\{\epsilon_i\} = \sigma^2$, and $\mathbb{C}\text{ov}\{\epsilon_i, \epsilon_j\} = 0 \,\forall\, i \neq j$.

-a- Write the design matrix $\mathbf{Z}$.

-b- Estimate the parameters for the following models and compute the R^2 for each:

$$\begin{aligned}
&\text{Model 1} : y_i = \beta_0 + \beta_1 z_i + \epsilon_i \\
&\text{Model 2} : y_i - \bar{y} = \beta_0 + \beta_1 z_i + \epsilon_i \\
&\text{Model 3} : y_i - \bar{y} = \beta_1 z_i + \epsilon_i \\
&\text{Model 4} : c y_i = \beta_0 + \beta_1 z_i + \epsilon_i \text{ for } c \in \mathbb{R}
\end{aligned}$$

Note that parameter estimates in each of the cases may not be the same (despite the common notation).

-2- Show that, $\mathbf{P}\,\mathbf{X}'\,\mathbf{X} = \mathbf{Q}\,\mathbf{X}'\,\mathbf{X} \implies \mathbf{P}\,\mathbf{X}' = \mathbf{Q}\,\mathbf{X}'$ for any conforming matrices $\mathbf{P}, \mathbf{Q}$, and $\mathbf{X}$. Use this to show that $(\mathbf{X}'\,\mathbf{X})^{-}\,\mathbf{X}'$ is a generalized inverse of $\mathbf{X}$.

-3- Consider the model: $\mathbf{Y} = \mathbf{Z}\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathbb{E}\{\boldsymbol{\epsilon}\} = 0$, $\mathbb{V}\text{ar}\{\boldsymbol{\epsilon}\} = \sigma^2 I_n$.

-a- Let $\hat{\mathbf{Y}}$ be the vector of predicted response variables based on least squares estimation of the model. Let $\hat{\boldsymbol{\epsilon}}$ be the estimated error vector. Show that $\mathbb{C}\text{ov}\{\hat{\mathbf{Y}}, \hat{\boldsymbol{\epsilon}}\} = \mathbf{O}$.

-b- (Part IV: Result 7.4) Derive the MLE of $\boldsymbol{\beta}$ and σ^2 under the assumption that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$. Note that for: $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \Sigma)$,

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu})\right\}$$

where determinant$(\Sigma) = |\Sigma|$.

-4- Consider the random vector $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma = \sigma^2 I_n)$. Derive the distribution of $\mathbf{C}\,\mathbf{X}$ (for $\mathbf{C}$ being a non-singluar matrix of constants) either through MGFs or through density transformations of $\mathbf{X}$.

-5- (Refer class notes from Jan 21.) We defined $\mathbf{q}_j = \frac{1}{\sqrt{\lambda_j}}\mathbf{Z}\,\mathbf{e}_j$ where $\mathbf{e}_j$ is an element of the set of orthonormal eigenvectors of $\mathbf{Z}'\,\mathbf{Z}$ and λ_j its corresponding eigenvalue. Show that $\mathbf{q}_j'\,\mathbf{q}_j = 1$, thus showing that its inverse is also 1.

[1] HW Version: 2021-02-10 at 09:29

-6- In `R` use the command `?glm` to bring up the help file on the `glm` function. Read through it to understand the different arguments you can pass to the `glm` function and what type of output it produces. For inputs you mostly need to pay attention to the `formula`, `family`, and `data` paramters. The `formula` parameter works the same as it does in the `lm` function. The `family` parameter tells you what members of the exponential family the `glm` function use for modeling the distribution of the response variable. Among the outputs pay particular attention to `coefficients`, `residuals`, `deviance`, and to a lesser extent `aic`.

-7- Use dataset simulated in question -1-.

- -a- Find parameter values for the distribution of $\hat{\boldsymbol{\beta}}$, and $\hat{\boldsymbol{\epsilon}}$ based on the assumption that $\boldsymbol{\epsilon}$ is distributed normally. (Use software for matrix computations.)
- -b- Estimate these parameters through least squares and maximum likelihood.
- -c- Plot the surface of the bivariate normal density of $\hat{\boldsymbol{\beta}}$ using least squares parameter estimates from part -b-. The `rgl` package in `R` is one option to use.

-8- Use dataset titled HW1-Prob.csv. Interpret you answers in case question below.

- -a- Plot a scatterplot of the dependent variable with respect to each of the independent variables.
- -b- Use `R` to estimate the least squares regression equation: $Y_i = \beta_0 + \beta_1 z_{1i} + \beta_2 z_{2i} + \epsilon_i$, for $i = 1, \ldots, n$
- -c- Construct simultaneous 95% confidence intervals for each element of $\boldsymbol{\beta}$.
- -d- Perform a likelihood ratio test for β_2 at a significance level of α.
- -e- Let $\mathbf{z}_0 = [1, 7, 8]'$. Construct a $100(1-\alpha)\%$ confidence interval for $\mathrm{E}\{Y_0 | \, \mathbf{z}_0\}$ assuming that Y_0 is observed.
- -f- For the same vector, $\mathbf{z}_0$, construct a $100(1-\alpha)\%$ prediction interval for Y_0 when it is not observed.
- -e- Check model adequacy by your choice of residual plots. What do you conclude?