Stats 203V: Introduction to Regression Models and ANOVA Summer 2020

Homework 3

Due date : July 26, 11:59 PM (PT)

For full credit, solutions (.pdf file, either typed or handwritten and scanned) must be uploaded to `Gradescope` by **11:59 pm PT on on Sunday July 26**. Any changes to the Homework will be posted on `Canvas`.

Collaboration on homework problems is fine, but your write up should be your own and the write up should mention the names of your collaborators.

From now on, we request that solutions to questions using `R` be prepared using `R Markdown`, e.g. within `R Studio`. Please convert your `.Rmd` file to a `.pdf` file before uploading.

1. (20+10+10 points) [Prostate cancer surgery dataset `prostate` in `library(faraway)`] Faraway, Exercises 2.4, 3.1 and 4.1. Also, briefly interpret what you see on the plots in your answer of Exercise 2.4.

2. (10+10 points) [Teenage gambling dataset `teengamb` in `library(faraway)`] Faraway, Exercises 3.3 and 4.2.

3. (10 points) [School expenditure and test scores, USA 1994-95, `sat` in `library(faraway)`] Faraway, Exercise 3.4.

4. (10 points) Consider a linear regression model

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t, \quad t = 1, \ldots, T; \tag{1}$$

where in fact a variable has been omitted and we can in fact model

$$\epsilon_t = \beta_2 Z_t + \delta_t, \quad t = 1, \ldots, T; \tag{2}$$

where the series (Z_t) can be assumed independent of the series (δ_t). Suppose that δ_i are i.i.d. with mean zero and also that

$$Z_{t+1} = Z_t + \eta_t, \;\; t = 0, \ldots, T-1, \tag{3}$$

with η_t a sequence of independent variables with common mean $\mu_\eta > 0$, and Z_0 is independent of the series (η_t).

   (a) Show that $\operatorname{cov}(\epsilon_{t+1}, \epsilon_t) = \beta_2^2 \operatorname{Var}(Z_t) > 0$, giving one explanation for positive autocorrelation in the residuals from model (1).

   (b) In the soft drink sales example discussed in Lecture 10, explain why (3) might or might not be a reasonable assumption. The file `SoftDrink.txt` is available in the Lecture 10 folder on `Canvas` if you would like to access the data.

5. (10 points) Let $Y_1, \ldots, Y_n$ be a random sample from $N(\theta, \sigma^2)$. Find the linear unbiased estimate of θ with minimum variance.

   [Hint : Write it as a regression model and then use the result derived in class.]

6. (10 points) Let $Y_1, \ldots, Y_n$ be independent random variables and let $Y_i \sim N(i\theta, i^2\sigma^2)$, for $i = 1, \ldots, n$. Find the weighted least squares estimate and calculate its variance.

   [Hint: Write this down as regression model with non-constant variance and then convert it to a usual constant variance model.]

7. (7+2+8+3 points) Consider the multiple linear regression model $Y = X\beta + \epsilon$, where X is $n \times p$ design matrix with linearly independent columns $X^{(1)}, \ldots, X^{(p)}$, and $\epsilon \sim N_n(0, \sigma^2 I_n)$. Let us write X as

$$X = [X_\omega \;\; X^{(p)}].$$

   (a) Using the formula of determinant of 2×2 block matrix using *Schur's complement*, show that

$$\det(X'X) = \det(X'_\omega X_\omega)\left[X^{(p)\prime}(I_n - H_\omega)X^{(p)}\right],$$

   where $H_\omega = X_\omega(X'_\omega X_\omega)^{-1}X'_\omega$.

   (b) Deduce that

$$\frac{\det(X'_\omega X_\omega)}{\det(X'X)} \geq \frac{1}{X^{(p)\prime}X^{(p)}}.$$

   (c) Hence show that

$$\mathrm{Var}(\hat{\beta}_p) \geq \sigma^2(X^{(p)\prime}X^{(p)})^{-1}.$$

   with equality if and only if $X^{(p)}$ is orthogonal to other columns of X. Here $\hat{\beta}$ denotes the OLS estimator of β.

   (d) Was the normality assumption necessary?