550.413 Assignment 3

Fall 2016

Instruction: This assignment consists of 4 problems. The assignment is due on **Wednesday, November 2, 2016** at 3pm, in class. If you cannot make it to class, please leave the assignment under the door at Whitehead Hall 306E and email the course instructor. If possible, please type up your assignments, preferably using LaTeX.

Problem 1: (10pts)

Let $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ be a linear model where $\mathbf{X}$ is of size $n \times p$ and the error terms $\boldsymbol{\epsilon}$ are independent, normally distributed with mean 0 and variance σ^2. Suppose furthermore that the columns of $\mathbf{X}$ can be partitioned as

$$\mathbf{X} = \begin{bmatrix} \mathbf{W} & \mathbf{Z} \end{bmatrix}$$

where $\mathbf{W}$ is of size $n \times q$ and is of full-column rank and $\mathbf{Z}$ is of size $n \times (p-q)$ and is of full-column rank, for some q satisfying $1 \leq q \leq p$, and that $\mathbf{W}^T\mathbf{Z} = \mathbf{0}$.

We now partition $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$ where $\boldsymbol{\beta}_1$ is of size $q \times 1$ and $\boldsymbol{\beta}_2$ is of size $(p-q) \times 1$. Let $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix}$ be the least square estimate of $\boldsymbol{\beta}$.

(a) Show that $\hat{\boldsymbol{\beta}}_1 = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\boldsymbol{y}$ and $\hat{\boldsymbol{\beta}}_2 = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{y}$.

(b) Show that $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ are independent.

(c) Let $\boldsymbol{a}$ be a $q \times 1$ vector and $\boldsymbol{b}$ be a $(q-p) \times 1$ vector. Let (l_1, u_1) and (l_2, u_2) be the individual 95% confidence intervals for $\boldsymbol{a}^T\boldsymbol{\beta}_1$ and $\boldsymbol{b}^T\boldsymbol{\beta}_2$ based on $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$, respectively. Is the confidence interval (l_1, u_1) independent of the confidence interval (l_2, u_2) ? Justify your answer.

Problem 2: (10pts)

Let $\mathbf{W}$ be a $n \times p$ matrix and $\mathbf{X}$ be a $n \times q$ matrix and that $\mathcal{C}(\mathbf{W}) \subseteq \mathcal{C}(\mathbf{X})$. Denote by $\mathbf{P_X}$ and $\mathbf{P_W}$ the symmetric idempotent matrices projecting onto $\mathcal{C}(\mathbf{X})$ and

$\mathcal{C}(\mathbf{W})$, respectively. Show that $\mathbf{P_X} - \mathbf{P_W}$ is the symmetric orthogonal projection onto $\mathcal{C}((\mathbf{I} - \mathbf{P_W})\mathbf{X})$.

You can do it by arguing as follows.

- First show that $\mathbf{P_X} - \mathbf{P_W}$ is idempotent. Hint: $\mathbf{P_X P_W} \boldsymbol{z} = \mathbf{P_W} \boldsymbol{z}$ for all $\boldsymbol{z}$; in addition $\mathbf{P_W P_X} = (\mathbf{P_X P_W})^\top$.
- Next, show that for any $\boldsymbol{z}$, $(\mathbf{P_X} - \mathbf{P_W})\boldsymbol{z} \in \mathcal{C}((\mathbf{I} - \mathbf{P_W})\mathbf{X})$. Hint: Since $\mathcal{C}(\mathbf{X})$ and $\mathcal{N}(\mathbf{X}^\top)$ are orthogonal complements, any vector $\boldsymbol{z} \in \mathbb{R}^n$ can be written as $\boldsymbol{z} = \mathbf{X}\boldsymbol{v} + \boldsymbol{w}$ for some vectors $\boldsymbol{v}$ and some $\boldsymbol{w} \in \mathcal{N}(\mathbf{X}^\top)$; what is the relationship between $\mathcal{N}(\mathbf{W}^\top)$ and $\mathcal{N}(\mathbf{X}^\top)$?.
- Finally, show that if $\boldsymbol{z} \in \mathcal{C}((\mathbf{I} - \mathbf{P_W})\mathbf{X})$ then $(\mathbf{P_X} - \mathbf{P_W})\boldsymbol{z} = 0$.

Problem 3: (20pts)

The `kidiq.dta` dataset is available from the url `http://www.stat.columbia.edu/~gelman/arm/examples/child.iq/kidiq.dta` accompanying the book "Data Analysis using Regression and Multilevel/Hierarchical Models" by Gelman and Hill. The dataset contains observations from a sample of 434 children. The variables include the child cognitive test scores at age 3 or 4, whether the mother finishes high school (coded as 1) or not (coded as 0), mother's IQ, age of mother at child's birth, and whether the mother work or not in the first three years of child's life. More specifically, the variable `mom.work` takes on the value

- `mom.work` = 1 if mother did not work in first three years of child's life
- `mom.work` = 2 if mother worked in second or third year of child's life
- `mom.work` = 3 if mother worked part-time in first year of child's life
- `mom.work` = 4 if mother worked full-time in first year of child's life

After downloading the `kidiq.dta` file you can read the data into `R` using the following snippet of code

```
library("foreign")
iq.data <- read.dta("kidiq.dta")
```

Using this dataset, answer the following questions.

(a) Perform a regression with `kid_score` as the response variable and the remaining variable except `mom_hs` as predictor variables.

(b) Provide a quick discussion regarding the coefficients for the predictor variables. What do they say ?

(c) Using the model in part [(a)], test the hypothesis that the predictor variables `mom_work` and `mom_age` is associated with the response variable. When do you recommend mothers should give birth ? What are your assumption for making this recommendation ?

(d) What happens when you add `mom_hs` as a predictor variable to the model in part (a) ? Have your conclusion about the timing of birth changed ?

(e) Using the model in part (d), perform some diagnostics, e.g., check the constant variance assumption, normality of errors. Look for outliers, influential points, and points with high leverage.

(f) Consider augmenting the model in part (d) with one whose predictor variables include interactions between say `mom.hs` and `mom_age` or interactions between say `mom.work` and `mom_age`. Write down the "formula" for the resulting model and discuss how it differs from the "formula" for the model in part (d). Test the hypothesis that the interaction term in the augmented model is not significant.

Problem 4: (20pts)

The link `http://www.amstat.org/publications/jse/v16n3/kuiper.xls` is a dataset collected from the Kelly Blue Book for several hundred 2005 used GM cars. Do something with this data.

This is meant to be an open-ended question. For some ideas of the kind of analysis one can attempt, see the article `http://www.amstat.org/publications/jse/v16n3/datasets.kuiper.html`.