Preliminary Examination
LINEAR MODELS

Answer all questions and show all work.
Q1, Q2, and Q4 are 20 points each, and Q3 is 10 points.

1. For any non-zero $\mathbf{w} \in \mathcal{R}^n$. Consider the column space $\mathcal{C}(\mathbf{w})$. Let $\mathbf{P}_w$ denote the $n \times n$ projection matrix to $\mathcal{C}(\mathbf{w})$.

   a. For any vector $\mathbf{v} \in \mathcal{R}^n$, prove

   $$\mathbf{P}_w\mathbf{v} = \left(\frac{\mathbf{v}'\mathbf{w}}{\mathbf{w}'\mathbf{w}}\right)\mathbf{w}$$

   b. In regression, let $\mathbf{X} = (\mathbf{1} \,|\, \mathbf{x}_1 \,|\, x_2 \,|\, \cdots \,|\mathbf{x}_{r-1} \,|\, \mathbf{x}_r)$ and $\mathbf{X}_{r-1} = (\mathbf{1} \,|\, \mathbf{x}_1 \,|\, x_2 \,|\, \cdots \,|\mathbf{x}_{r-1})$. Further, let

   $$\mathbf{z}_r = \mathbf{x}_r - \mathbf{P}_{X_{r-1}}\mathbf{x}_r = (\mathbf{I} - \mathbf{P}_{X_{r-1}})\mathbf{x}_r.$$

   Prove that for any $\mathbf{v} \in \mathcal{C}(\mathbf{X}_{r-1})$, $\mathbf{v} \perp \mathbf{z}_r$.

   c. Show that $\mathbf{P}_X - \mathbf{P}_{X_{r-1}} = \mathbf{P}_{z_r}$. That is, prove that $\mathbf{P}_X - \mathbf{P}_{X_{r-1}}$ is symmetric and idempotent, and that $(\mathbf{P}_X - \mathbf{P}_{X_{r-1}})\mathbf{v} = \mathbf{v}$ for any $\mathbf{v} \in \mathcal{C}(\mathbf{z}_r)$.

   d. Use (a)-(c) to prove that

   $$\hat{\mathbf{Y}} = \mathbf{P}_{X_{r-1}}\mathbf{Y} + \left(\frac{\mathbf{e}'_{r-1}\mathbf{z}_r}{\mathbf{z}'_r\mathbf{z}_r}\right)\mathbf{z}_r$$

   where $\mathbf{e}_{r-1} = (\mathbf{I} - \mathbf{P}_{X_{r-1}})\mathbf{Y}$.

   e. Explain why it is clear that the multiplier of $\mathbf{z}_r$ in part (d) is equal to $\hat{\beta}_r$, the ordinary least squares estimate of the regression coefficient β_r in the full original regression.

2. Three groups of n observations are fitted using the following fixed-effects model:

   $$Y_{ij} = \mu + \theta_i + \epsilon_{ij},$$

   where $\{\epsilon_{ij}\}$ are independent and identically distributed $N(0, \sigma^2)$ random variables, for $i = 1, 2, 3$ and $j = 1, \ldots, n$. To avoid identifiability issues, we set $\sum_{i=1}^{3} \theta_i = 0$ and remove θ_3 from the above formulation, that is, the parameters in our model are μ, θ_1, θ_2, and σ^2.

   a. Specify the design matrix and compute the *correlation* between the least squares estimators (LSE) $\hat{\theta}_1$ and $\hat{\theta}_2$.

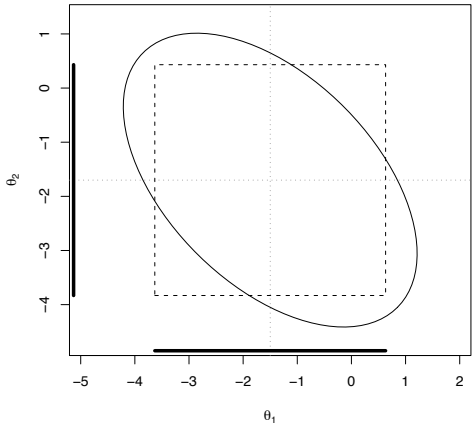
b. Give the expressions for the estimators $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\sigma}^2$. Construct $100(1-\alpha)\%$ *individual* confidence intervals for θ_1 and θ_2 based on these estimates, *respectively*.

c. Show that a $100(1-\alpha)\%$ joint confidence region for θ_1 and θ_2 can be specified by

$$(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})' \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \leq \frac{2\hat{\sigma}^2}{n} F(\alpha; 2, 3(n-1))$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)'$, and $F(\alpha, 2, 3(n-1))$ is the $1-\alpha$ quantile of an $F_{2,3(n-1)}$ distribution.

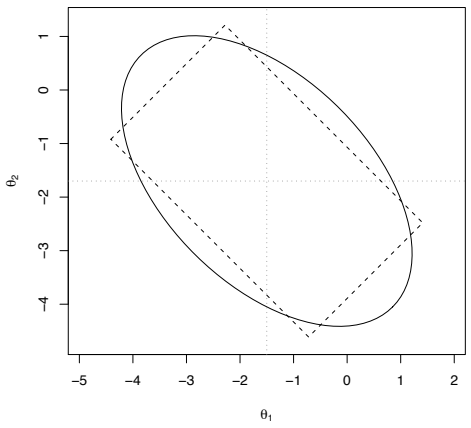
d. The confidence regions from parts (b) and (c) are pictured below: The bold lines mark the separate confidence intervals, the dashed square is the (Cartesian) product of these intervals, and the ellipsoid represents the joint confidence region. What would be your conclusions from this figure using: (i) the separate confidence intervals and (ii) the joint confidence region, respectively? Explain how and why your conclusions from (i) and (ii) are not the same. If you were to test the hypothesis of no difference across groups, which confidence intervals/region is preferred?



e. Not happy with the separate confidence intervals above, you decide to reparametrize the model using

$$\gamma_1 = \frac{\theta_1 + \theta_2}{2}$$
$$\gamma_2 = \frac{\theta_1 - \theta_2}{2}$$

The product of $100(1-\alpha)\%$ confidence intervals for γ_1 and γ_2 are represented by the dashed rectangle in the figure next page. If you now decide to test the hypothesis of no difference across groups using these new confidence intervals, what is your conclusion? Is it the same as your conclusion from (i) in part (d)? Why? How about your conclusion from (ii) in part (d)? Why?



3. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be mutually independent with $\mathbf{x}_j$ distributed as $N_p(\mu_j, \boldsymbol{\Sigma})$.

   a. Prove that $\mathbf{V} = \sum_{j=1}^{n} c_j \mathbf{x}_j$ is distributed as $N_p\left(\sum_{j=1}^{n} c_j \mu_j, \left(\sum_{j=1}^{n} c_j^2\right) \boldsymbol{\Sigma}\right)$.

   b. Prove that $\mathbf{V}$ and $\mathbf{W} = \sum_{j=1}^{n} b_j \mathbf{X}_j$ are jointly multivariate normal with the covariance matrix
   $$\begin{bmatrix} \left(\sum_{j=1}^{n} c_j^2\right) \boldsymbol{\Sigma} & \left(\sum_{j=1}^{n} c_j b_j\right) \boldsymbol{\Sigma} \\ \left(\sum_{j=1}^{n} c_j b_j\right) \boldsymbol{\Sigma} & \left(\sum_{j=1}^{n} b_j^2\right) \boldsymbol{\Sigma} \end{bmatrix}$$

   c. Give a necessary and sufficient condition for $\mathbf{V}$ and $\mathbf{W}$ to be independent, and explain why.

4. Suppose that in an experimental study you suspect that many observations were tainted by a technician and now you want to test them *jointly* for being outliers. To this end, you organize the suspected observations as the last q observations from a total of n and adopt the following *mean shift outlier model* (MSOM) on these observations:

$$\begin{aligned} Y_1 &= \mathbf{x}_1' \boldsymbol{\beta} + \varepsilon_1 \\ &\vdots \\ Y_{n-q} &= \mathbf{x}_{n-q}' \boldsymbol{\beta} + \varepsilon_{n-q} \\ Y_{n-q+1} &= \mathbf{x}_{n-q+1}' \boldsymbol{\beta} + \delta_1 + \varepsilon_{n-q+1} \\ &\vdots \\ Y_n &= \mathbf{x}_n' \boldsymbol{\beta} + \delta_q + \varepsilon_n \end{aligned}$$

This model can be specified in matrix form as follows:

$$\underbrace{\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{X}_2 & \mathbf{I}_q \end{bmatrix}}_{\mathbf{X}} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\delta} \end{bmatrix} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_q]'$, $E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$, and $\text{var}(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2\mathbf{I}_n$ (as usual). After some algebra, we can show that

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} (\mathbf{X}_1'\mathbf{X}_1)^{-1} & -(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_2' \\ -\mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1} & \mathbf{I}_q + \mathbf{X}_2(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_2' \end{bmatrix}.$$

Now consider $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}}$, the LSE for $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ under this model. Let $\hat{\beta}_1$ denote the LSE for $\boldsymbol{\beta}$ when regressing only $\mathbf{Y}_1$ on $\mathbf{X}_1$, that is, when ignoring the last q observations.

a Show that (i) $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_1$ and (ii) $\hat{\boldsymbol{\delta}} = \mathbf{Y}_2 - \mathbf{X}_2\hat{\boldsymbol{\beta}}_1$, that is, the LSE for $\boldsymbol{\delta}$ is the difference between the (removed) observed values and the fitted values for $\mathbf{X}_2$ in the model without the last q suspected observations.

b. Show that the last q observations are perfectly fit by the MSOM: $\hat{\mathbf{Y}}_2 = \mathbf{X}_2\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\delta}} = \mathbf{Y}_2$. What can you say about the relation between the LSE $\hat{\sigma}^2$ for σ^2 under the MSOM and the LSE $\hat{\sigma}_1^2$ for σ^2 under the model without the last q observations?

c. Find the hat matrix for the MSOM and comment on the leverage for the suspected data points in light of the results from part (b).

d. Conduct a joint outlier test by testing $\delta_1 = \cdots = \delta_q = 0$. State the test statistic and its distribution under the null.