1. Mixed Effects Model [40pts]

Consider the following mixed effects models,

\[ y_j = X_j \beta + Z_j b_j + e_j, \quad j = 1, \ldots, m. \]

(1)

where \( y_j \) is an \( n_j \times 1 \) vector of observations, \( X_j \) is an \( n_j \times p^* \) design matrix for the fixed effects, \( Z_j \) is an \( n_j \times k \) design matrix for the random effects, \( \beta \) is a \( p^* \times 1 \) vector of fixed effects, \( b_j \) is a \( k \times 1 \) vector of random effects, and \( e_j \) is an \( n_j \times 1 \) vector of error terms. We assume that for any \( j = 1, \ldots, m \), we have

\[ \mathbb{E}[e_j e_j^T] = \sigma^2 I_{n_j}, \quad \mathbb{E}[b_j b_j^T] = \sigma^2 D, \quad \text{and} \quad \mathbb{E}[e_j b_j^T] = 0. \]

Moreover, both the error terms and the random effects are uncorrelated across groups, such that \( \mathbb{E}[e_j e_{j'}^T] = 0 \), and \( \mathbb{E}[b_j b_{j'}^T] = 0 \), for any \( j \neq j' \). Finally, we also assume that this model is balanced in the sense that \( Z_j = Z \), and \( n_j = n \), for every \( j = 1, \ldots, m \); with \( N := mn \).

1. [10pts] One can stack the \( m \) equations in (1), in order to obtain a global formula, \( y = X\beta + Zb + e \) summarizing the entire model. What is the dimension of the corresponding terms, \( y \), \( X \), \( \beta \), \( Z \), \( b \), and \( e \) in this new formula?

In a balanced design, the quantities \( y \), \( X \), \( \beta \), \( Z \), \( b \) and \( e \) are of dimensions \((N \times 1)\), \((N \times p^*)\), \((p^* \times 1)\), \((N \times mk)\), \((mk \times 1)\), and \((N \times 1)\), respectively.

2. [10pts] Next, we make the following two distributional assumptions,

\[ e_j \overset{\text{iid}}{\sim} \text{MVN}_n(0, \sigma^2 I_n), \quad b_j \overset{\text{iid}}{\sim} \text{MVN}_k(0, \sigma^2 D), \]

for every \( j = 1, \ldots, m \). Define \( \eta := Zb + e \) as a new ‘error’ term comprised of the random effects and the classical error terms. Compute the expectation of \( \eta \).
Clearly, it is easy to verify that this term has a null expectation. For clarity of presentation, we drop the dependency of the expectation on $X$, such that $\mathbb{E}[\cdot] := \mathbb{E}[\cdot | X]$,

$$
\mathbb{E}[\eta] = \begin{bmatrix} Z_1 \mathbb{E}[b_1] + \mathbb{E}[e_1] \\ \vdots \\ \vdots \\ Z_m \mathbb{E}[b_m] + \mathbb{E}[e_m] \end{bmatrix} = 0.
$$

3. **[20pts]** Finally, compute the variance of $\eta$. What is the dimension of this variance/covariance matrix?

The variance of $\eta$ can be obtained for any index $j = 1, \ldots, m$,

$$
\text{Var}[\eta_j] = \mathbb{E}[\eta_j \eta_j^T] - \mathbb{E}[\eta_j] \mathbb{E}[\eta_j]^T
$$

$$
= \mathbb{E}[Z_j b_j + e_j] (Z_j b_j + e_j)^T
$$

$$
= \mathbb{E}[Z_j b_j b_j^T Z_j^T + Z_j b_j e_j^T + e_j b_j^T Z_j^T + e_j e_j^T]
$$

$$
= Z_j \mathbb{E}[b_j b_j^T] + Z_j \mathbb{E}[b_j e_j^T] + Z_j \mathbb{E}[e_j b_j^T] + Z_j \mathbb{E}[e_j e_j^T]
$$

$$
= \sigma^2 (I_n + Z_j D Z_j^T),
$$

since both $\mathbb{E}[b_j e_j^T] = 0$ and $\mathbb{E}[e_j b_j^T] = 0$, by assumption. This derivation can also be verified straightforwardly using the properties of the variance operator,

$$
\text{Var}[\eta_j] = \text{Var}[Z_j b_j + e_j] = Z_j \text{Var}[b_j] Z_j^T + \text{Var}[e_j] + 2 Z_j \text{Cov}[b_j, e_j].
$$

The dimension of this variance/covariance matrix is $N \times N$.

2. **Box-Cox Transformation [50pts]**

Given a sample of realizations, $y_1, \ldots, y_n$, recall that $\tilde{y}_i := \psi(y_i; \lambda)$ is defined as the *scaled power* transformation,

$$
\psi(y_i; \lambda) := \begin{cases} 
(y_i^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0, \\
\log(y_i) & \text{if } \lambda = 0;
\end{cases}
$$

for some $\lambda \in \mathbb{R}$. We further assume that $\tilde{y}_i \overset{iid}{\sim} N(x_i^T \beta, \sigma^2)$, for every $i = 1, \ldots, n$, and some given vectors of covariates, $x_i$’s, of order $p^* \times 1$.

1. **[10pts]** Use the change of variable formula in order to write the integral of the pdf of the transformed variables, $\tilde{y}_i$’s, with respect to the pdf of the $y_i$’s.

Using the standard change of variable formula, this implies that the integral for any pdf $p(\tilde{y}_i)$ of the random variable, $\tilde{y}_i$, can be written as

$$
\int_{\psi(\mathbb{R}^+)} p(\tilde{y}_i) d\tilde{y}_i = \int_{\mathbb{R}^+} p(\psi(y_i; \lambda)) \frac{d\psi}{dy_i} dy_i.
$$
2. **[20pts]** Derive the Jacobian, denoted $J(y; \lambda)$, of the scaled power transformation, $\psi: \mathbb{R}^+ \rightarrow \mathbb{R}$, for all the $y_i$'s. The Jacobian of the transformation for all the $y_i$'s, is given by

$$J(y; \lambda) = \prod_{i=1}^{n} \frac{d \psi_i}{dy_i}.$$

By straightforward differentiation, this gives

$$J(y; \lambda) = \prod_{i=1}^{n} \frac{d}{dy_i} \left( \frac{(y_i^\lambda - 1)}{\lambda} \right) = \prod_{i=1}^{n} y_i^{\lambda-1},$$

3. **[20pts]** Assuming that the transformed observed values are normally distributed, the likelihood for the resulting vector of realizations is given by

$$L(\beta, \sigma^2, \lambda) = \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left\{ -\frac{1}{2\sigma^2} (\bar{y} - X\beta)^T(\bar{y} - X\beta) \right\} J(y; \lambda).$$

What are the MLE estimators for $\beta$ and $\sigma^2$ in this context? How do they depend on the value $\lambda$? The solution is immediately given by:

$$\hat{\beta}(\lambda) = (X^T X)^{-1} X^T \bar{y},$$

$$\hat{\sigma}^2(\lambda) = \frac{1}{n} (\bar{y} - X\hat{\beta}(\lambda))^T (\bar{y} - X\hat{\beta}(\lambda)).$$

where note that these estimates are a function of $\lambda$, through the transformed values, $\tilde{y}_i$'s.

3. **Ridge Regression [50pts]**

Given some data set $(y_i, x_i), i = 1, \ldots, n$, the generalized Tikhonov regularization is formulated as follows,

$$\text{RSS}(\beta; \lambda) := \left\{ \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \sum_{k=1}^{p} q_{jk} \beta_j \beta_k \right\},$$

with $q_{jk} = q_{kj}$, and where the regularization parameter is constrained to be non-negative, such that $\lambda \geq 0$. Here, we have removed the intercept, such that $\beta$ is a vector of dimension $p \times 1$.

1. **[10pts]** Using matrix algebra, the right-hand side of equation (3) can be expressed as

$$\text{RSS}(\beta; \lambda) := (y - X\beta)^T (y - X\beta) + \lambda \beta^T Q \beta.$$

Explain why this criterion can be described as a generalization of ridge regression.

The criterion to minimize in ridge regression is given by

$$\text{RSS}(\beta; \lambda) := (y - X\beta)^T (y - X\beta) + \lambda \beta^T I \beta.$$

Thus, we obtain the ridge expression as a special case of the generalized Tikhonov, by setting $Q := I$. 

Department of Mathematics and Statistics, Boston University
2. [10pts] Assume that the matrix $(X^T X + \lambda Q)$ is non-singular. Then, compute $\hat{\beta}$, the minimizer of the regularized RSS criterion in equation (3), for a fixed $\lambda$.

Taking the derivative and setting it to zero, we have

$$\frac{\partial}{\partial \beta} \left((y - X\beta)^T(y - X\beta) + \lambda \beta^T Q \beta\right) = 0,$$

which simplifies to

$$-2X^T y + 2X^T X\beta + 2\lambda Q\beta = 0.$$

Factoring $\beta$, we then obtain,

$$(X^T X + \lambda Q)\beta = X^T y.$$

By assumption, we know that $(X^T X + \lambda Q)$ is invertible, and therefore,

$$\hat{\beta} := (X^T X + \lambda Q)^{-1} X^T y.$$

3. [10pts] What is the hat matrix for this model, such that we have $\hat{y} = H_{\lambda} y$, for a fixed subscript, $\lambda$.

Immediately, we have

$$\hat{y} = X\hat{\beta} = X(X^T X + \lambda Q)^{-1} X^T y,$$

and therefore,

$$H_{\lambda} = X(X^T X + \lambda Q)^{-1} X^T.$$

4. [20pts] Assume that $Q := I$. Then, compute the trace of the hat matrix found in the previous question, $H_{\lambda}$, where we are again fixing $\lambda$ to a particular value.

Here, we need to use the SVD of $X$ given by $UDV^T$, and the eigendecomposition of $X^T X$ given by $VD^2 V^T$. Note that

$$X^T X = VDU^T UXV^T = VD^2 V^T,$$

and

$$(X^T X)^{-1} = V D^{-2} V^T.$$

Then, the hat matrix can be expressed as follows,

$$H_{\lambda} = UDV^T (VD^2 V^T + \lambda I)^{-1} VDU^T$$

$$= UD V (D^2 + \lambda I)^{-1} V^T VDU^T$$

$$= UD (D^2 + \lambda I)^{-1} D U^T.$$

since $VD^2 V^T$ and $\lambda I$ commute and are therefore simultaneously diagonalizable. Finally, since the trace of a matrix is equal to the sum of its eigenvalues, it readily follows that

$$\text{tr}(H_{\lambda}) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda},$$

where $d_j$ is the $j^{th}$ diagonal entry of $D$. 

Department of Mathematics and Statistics, Boston University 4
4. Variable Selection [60pts]

The general formula for the Akaike Information Criterion (AIC) for some model of the data, denoted by the pdf \( p(y|\theta) \) with the dimension of \( \theta \) being \( d \times 1 \), is given by

\[
AIC := -2 \log p(y|\hat{\theta}) + 2d, \tag{4}
\]

where \( \hat{\theta} \) is the MLE of this model, for the vector of observations, \( y \).

1. [20pts] Compute the AIC for a regression model with normal error, using the formula in equation (4).

   In this model, \( \beta \) is a \( p^* \times 1 \) vector. Your answer should be given up to a proportional constant.

   \[
y_i \overset{\text{iid}}{\sim} N(x_i^T \beta, \sigma^2), \quad i = 1, \ldots, n.
\]

   The log-likelihood for this model is given by

   \[
   \log L(\beta, \sigma^2; y) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2.
   \]

   Evaluating this formula at the MLEs for \( \beta \) and \( \sigma^2 \), we obtain

   \[
   \log L(\hat{\beta}, \hat{\sigma}^2; y) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} \sum_{i=1}^{n} (y_i - x_i^T \hat{\beta})^2 = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\hat{\sigma}^2) - \frac{n\hat{\sigma}^2}{2\hat{\sigma}^2}.
   \]

   Thus, up to a constant term, this gives

   \[
   \log L(\hat{\beta}, \hat{\sigma}^2; y) = -\frac{n}{2} \log(\hat{\sigma}^2) + C = -\frac{n}{2} \log \left( \frac{1}{n} \text{RSS}(\hat{\beta}) \right) + C.
   \]

   Plugging this formula into the definition of the AIC, we then have

   \[
   AIC = n \log \left( \frac{\text{RSS}(\hat{\beta})}{n} \right) + 2(p^* + 1),
   \]

   where \( d = p^* + 1 \), since we have also estimated \( \sigma^2 \). The answer \( d = p^* \) is also acceptable.

2. [20pts] You have been asked to analyze a historical data set describing the prevalence of highway accidents. You have to find a small subset of covariates that would provide a parsimonious prediction of the observed number of car accidents. You are tackling this problem by using a backward-stepwise variable selection procedure. Thus, you have started with a model that includes all of the candidate variables. At each step of this selection procedure, you remove a single candidate variable. This is continued until no further improvement is possible. Consider the following step in this protocol.

   \[
   \text{Step: AIC=-74.21} \quad \log2(\text{rate}) \sim \log2(\text{len}) + \log2(\text{ADT}) + \log2(\text{trks}) + \log2(\text{sigs1}) + \log2(\text{slim}) + \log2(\text{hwy})
   \]

   \[
   \begin{array}{|c|c|c|c|}
   \hline
   \text{Df} & \text{Sum of Sq} & \text{RSS} & \text{AIC} \\
   \hline
   \text{- log2(trks)} & 1 & 0.14288 & 3.8097 & -74.714 \\
   \text{<none>} & & 3.6668 & -74.205 \\
   \hline
   \end{array}
   \]

   Department of Mathematics and Statistics, Boston University
The choice of the next variable to remove depends on our selection criterion:

a) If using the AIC as a selection criterion, which variable shall we remove, if any?
b) Instead, if using the RSS as a selection criterion, which variable shall we remove, if any?

- Using the AIC, we should remove log2(TRKS).
- Using the RSS, we should not remove any variable.

3. Similarly, consider the R output for the next step in this backward selection procedure:

```
Step:  AIC=-74.71
log2(rate) ~ log2(len) + log2(ADT) + log2(sigs1) + slim + hwy

Df  Sum of Sq     RSS       AIC
<none>         3.8097 -74.714
- log2(ADT)    1  0.28821  4.0979 -73.870
- hwy          3  1.68565  5.4954 -66.427
- slim         1  1.15948  4.9692 -66.352
- log2(sigs1)  1  1.56367  5.3734 -63.302
```

Here again, the choice of the variable to be removed depends on the selection criterion:

a) If using the AIC as a selection criterion, which variable shall we remove, if any?
b) Instead, if using the RSS as a selection criterion, which variable shall we remove, if any?

- Using the AIC, we should not remove anything.
- Using the RSS, we should not remove anything.