Statistics 206

Homework 3

Due : October 19, 2015, In Class

1. Answer the following questions with regard to the general linear regression model and explain your answer.

   (a) What is the maximum number of X variables that can be included in a general linear regression model used to fit a data set with 10 cases?

   (b) With 4 predictors, how many X variables are there in the interaction model with all main effects and all interaction terms (2nd order, 3rd order, etc.)?

   (c) Are the residuals uncorrelated? Do they have constant variance? How about the fitted values?

2. $\mathbf{Z}$ is an n-dimensional random vector with expectation $\mathbf{E}(\mathbf{Z})$ and variance-covariance matrix:

$$\mathbf{Var}(\mathbf{Z}) = \mathbf{Cov}(\mathbf{Z}, \mathbf{Z}) = \Sigma.$$

   A is an $s \times n$ nonrandom matrix and B is a $t \times n$ nonrandom matrix. Show the following:

   (a) $\mathbf{E}(A\mathbf{Z}) = A\mathbf{E}(\mathbf{Z})$.

   (b) $\mathbf{Cov}(A\mathbf{Z}, B\mathbf{Z}) = A\Sigma B^T$. So in particular, $\mathbf{Var}((A\mathbf{Z}) = A\Sigma A^T$.

3. **Projection matrices**. Show the following are projection matrices, i.e., being symmetric and idempotent. Which linear subspace each of these matrices projects to? What are the ranks of these matrices? You can take $\mathbf{H}$ as the hat matrix from a simple linear regression model with n cases (where the X values are not all equal).

   (a) $\mathbf{I}_n - \mathbf{H}$

   (b) $\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n$

   (c) $\mathbf{H} - \frac{1}{n}\mathbf{J}_n$

4. Derive $E(SSTO)$ and $E(SSR)$ under the simple linear regression model using matrix algebra.

5. Under the general linear regression model, show that:

   (a) The residuals vector $\mathbf{e}$ is uncorrelated with the fitted values vector $\hat{\mathbf{Y}}$ and the LS estimator $\hat{\boldsymbol{\beta}}$.

   (b) With Normality assumption on the error terms, SSE is independent with SSR and the LS estimator $\hat{\boldsymbol{\beta}}$. (*Hint:* If $\mathbf{Z}$ is a multivariate Normal random vector, then $A\mathbf{Z}$ and $B\mathbf{Z}$ are jointly normally distributed.)

6. **Multiple linear regression by matrix algebra in R**. Consider the following data set with 5 cases, one response variable Y and two predictor variables X_1, X_2.

```
case     Y     X1       X2
1       -0.97  -0.63  -0.82
2       2.51  0.18   0.49
3      -0.19  -0.84  0.74
4       6.53  1.60   0.58
5       1.00  0.33  -0.31
```

Consider the first-order model for the following questions.

(a) Write down the model equations and the coefficient vector $\boldsymbol{\beta}$. Write down the design matrix and the response vector.

(b) In R, create the design matrix $\mathbf{X}$ and the response vector $\mathbf{Y}$. Calculate $\mathbf{X}'\mathbf{X}$, $\mathbf{X}'\mathbf{Y}$ and $(\mathbf{X}'\mathbf{X})^{-1}$. Copy your results here.

(c) Obtain the least-squares estimators $\hat{\boldsymbol{\beta}}$. Copy your results here.

(d) Obtain the hat matrix $\mathbf{H}$ and copy it here. What are $rank(\mathbf{H})$ and $rank(\mathbf{I} - \mathbf{H})$? (Hint: You may use *rankMatrix()* in library *Matrix*)

(e) Obtain the fitted values, the residuals, SSE and MSE. What should be the degrees of freedom of SSE? Copy your results here. You may use the following codes (with suitable modification) for SS:

```
> sum((Y-mean(Y))^2)
> sum((Y-Yhat)^2)
> sum((Yhat-mean(Y))^2)
```

Consider the nonadditive model with interaction between X_1 and X_2 for the following questions.

(f) Write down the model equations and the coefficient vector $\boldsymbol{\beta}$.

(g) Specify the design matrix and the response vector. Obtain the hat matrix $\mathbf{H}$. Find $rank(\mathbf{H})$ and $rank(\mathbf{I} - \mathbf{H})$. Compare the ranks with those from part (d), what do you observe?

(h) Obtain the least-squares estimators $\hat{\boldsymbol{\beta}}$. Copy your results here.

(i) Obtain the fitted values, the residuals, SSE and MSE. What should be the degrees of freedom of SSE? Copy your results here.

(j) Which model appears to fit the data better?

7. For each of the following regression models, indicate whether it can be expressed as a general linear regression model. If so, indicate which transformations and/or new variables need to be introduced.

(a) $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 \log X_{i2} + \beta_3 X_{i1}^2 + \epsilon_i$.

(b) $Y_i = \epsilon_i \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^2)$. $(\epsilon_i > 0)$

(c) $Y_i = \beta_0 \exp(\beta_1 X_{i1}) + \epsilon_i$.

(d) $Y_i = \{1 + \exp(\beta_0 + \beta_1 X_{i1} + \epsilon_i)\}^{-1}$.

8. (**Optional Problem**) Under the simple linear regression model with Normal errors, derive the sampling distributions for SSR and $SSTO$ when $\beta_1 = 0$.