Statistics 206

Homework 3 Solution

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?

   Here n=10 and we know $p \leq n = 10$. So maximum value of p is 10. The maximum number of X variables is $p - 1 = 9$.

   (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.)?

   No. of main effects:${}^4C_1 = 4$
   No. of 2^{nd} order interactions: ${}^4C_2 = 6$
   No. of 3^{rd} order interactions: ${}^4C_3 = 4$
   No. of 4^{th} order interactions: ${}^4C_4 = 1$
   Hence,total no. of X variables in the interaction model=4+6+4+1=15

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

   The residuals have variance covariance matrix $\sigma^2(I_n - H)$ and hence will not be uncorrelated or have constant variance except for the trivial case when H=0 which we do not consider. The fitted values have variance covariance matrix $\sigma^2 H$ and hence will not be uncorrelated or have constant variance except for the trivial case (which we do not consider) when $H = cI_n$ for some constant c.

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

$$(A\mathbf{Z})_{\mathbf{j}} = \sum_{\mathbf{k=1}}^{\mathbf{n}} \mathbf{a_{jk}Z_k} \ \forall \ \mathbf{j = 1, .., s}$$

$$(E(A\mathbf{Z}))_{\mathbf{j}} = \mathbf{E}((\mathbf{AZ})_{\mathbf{j}}) = \mathbf{E}(\sum_{\mathbf{k=1}}^{\mathbf{n}} \mathbf{a_{jk}Z_k}) = \sum_{\mathbf{k=1}}^{\mathbf{n}} \mathbf{a_{jk}E(Z_k)} = (\mathbf{AE(Z)})_{\mathbf{j}} \ \forall \ \mathbf{j = 1, .., s}$$

   Hence Proved.

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

$$W = A\mathbf{Z}, U = B\mathbf{Z}, C = Cov(W, U), D = A\Sigma B^T$$

$$C_{ij} = Cov(W_i, U_j) = Cov(\sum_{k=1}^{n} a_{ik}\mathbf{Z}_k, \sum_{k=1}^{n} b_{jk}\mathbf{Z}_k) =$$

$$\sum_{k=1}^{n}\sum_{l=1}^{n} a_{ik}b_{jl}Cov(\mathbf{Z}_k, \mathbf{Z}_l) = \sum_{k=1}^{n}\sum_{l=1}^{n} a_{ik}b_{jl}\Sigma_{kl} = D_{ij}\ \forall\ i = 1, .., s, j = 1, .., t$$

Hence Proved.

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

$$(\mathbf{I}_n - \mathbf{H})' = \mathbf{I}_n' - \mathbf{H}' = \mathbf{I}_n - \mathbf{H}$$
$$(\mathbf{I}_n - \mathbf{H})^2 = \mathbf{I}_n^2 - \mathbf{I}_n\mathbf{H} - \mathbf{H}\mathbf{I}_n + \mathbf{H}^2 = \mathbf{I}_n - \mathbf{H}$$

It projects a vector onto the linear subspace of $\mathbf{R}^n$ that is orthogonal to the column space of X. Its rank is $n - p$.

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

$$(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n)' = \mathbf{I}_n' - \frac{1}{n}\mathbf{J}_n' = \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n$$
$$(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n)^2 = \mathbf{I}_n^2 - \mathbf{I}_n\frac{1}{n}\mathbf{J}_n - \frac{1}{n}\mathbf{J}_n\mathbf{I}_n + \frac{1}{n^2}\mathbf{J}_n^2 = \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n$$

It projects a vector onto the linear subspace of $\mathbf{R}^n$ that is orthogonal to the subspace spanned by $\mathbf{1}_n$. Its rank is $n - 1$.

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

$$(\mathbf{H} - \frac{1}{n}\mathbf{J}_n)' = \mathbf{H}' - \frac{1}{n}\mathbf{J}_n' = \mathbf{H} - \frac{1}{n}\mathbf{J}_n$$
$$(\mathbf{H} - \frac{1}{n}\mathbf{J}_n)^2 = \mathbf{H} - \frac{1}{n}\mathbf{J}_n\mathbf{H} - \mathbf{H}\frac{1}{n}\mathbf{J}_n + \frac{1}{n^2}\mathbf{J}_n^2 = \mathbf{H} - \frac{1}{n}\mathbf{J}_n\mathbf{H} - \mathbf{H}\frac{1}{n}\mathbf{J}_n + \frac{1}{n}\mathbf{J}_n = \mathbf{H} - \frac{1}{n}\mathbf{J}_n$$

since $\mathbf{J}_n\mathbf{H} = \mathbf{J}_n$

$\mathbf{J}_n\mathbf{H} = \mathbf{J}_n$ because $\mathbf{H}$ is the projection matrix onto the column space of X and every column of $\mathbf{J}_n$, namely $\mathbf{1}_n$, is in the column space of X.

It projects a vector onto the linear subspace of column space of X that is orthogonal to the subspace spanned by $\mathbf{1}_n$.Its rank is $p - 1$.

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

$$
\begin{aligned}
E(SSTO) &= E\{Y'(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n)Y\} = E\{Tr((\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n)YY')\} \\
&= Tr\{(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n)E(YY')\} = Tr\{(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n)(\sigma^2\mathbf{I}_n + X\beta\beta'X')\} \\
&= (n-1)\sigma^2 + Tr(\beta'X'(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n)X\beta) \\
&= (n-1)\sigma^2 + Tr(\beta'X'((\mathbf{I}_n - \mathbf{H}) + (\mathbf{H} - \frac{1}{n}\mathbf{J}_n))X\beta) \quad \text{by } (b) = (a) + (c) \text{ from problem 2} \\
&= (n-1)\sigma^2 + Tr(\beta'X'(\mathbf{I}_n - \mathbf{H})X\beta) + Tr(\beta'X'(\mathbf{H} - \frac{1}{n}\mathbf{J}_n)X\beta) \\
&= (n-1)\sigma^2 + 0 + \beta_1^2\sum(X_i - \bar{X})^2 \quad \text{by } (\mathbf{I}_n - \mathbf{H})X = 0 \text{ and next part} \\
&= (n-1)\sigma^2 + \beta_1^2\sum(X_i - \bar{X})^2. \\
E(SSR) &= E\{Y'(\mathbf{H} - \frac{1}{n}\mathbf{J}_n)Y\} = E\{Tr((\mathbf{H} - \frac{1}{n}\mathbf{J}_n)YY')\} \\
&= Tr\{(\mathbf{H} - \frac{1}{n}\mathbf{J}_n)E(YY')\} = Tr\{(\mathbf{H} - \frac{1}{n}\mathbf{J}_n)(\sigma^2\mathbf{I}_n + X\beta\beta'X')\} \\
&= (2-1)\sigma^2 + Tr(\beta'X'(\mathbf{H} - \frac{1}{n}\mathbf{J}_n)X\beta) \\
&= \sigma^2 + Tr(\beta'X'X\beta - \beta'X'\frac{1}{n}\mathbf{J}_nX\beta) \quad \text{since } \mathbf{H}X = X \\
&= \sigma^2 + \beta'X'X\beta - \beta'X'\frac{1}{n}\mathbf{J}_nX\beta \\
&= \sigma^2 + (n\beta_0^2 - 2\beta_1\sum X_i + \beta_1^2\sum X_i^2) - (n\beta_0^2 - 2\beta_1\sum X_i + n\beta_1^2(\bar{X}_i)^2) \\
&= \sigma^2 + \beta_1^2\sum(X_i - \bar{X})^2.
\end{aligned}
$$

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 $\hat{\boldsymbol{\beta}}$.

   $$e = (I-H)Y, \quad \widehat{\beta} = (X'X)^{-1}X'Y,$$

   $$Cov(e, \widehat{\beta}) = (I-H)cov(Y)((X'X)^{-1}X)' = \sigma^2(I-H)IX(X'X)^{-1} = \sigma^2(I-H)X(X'X)^{-1} = 0,$$

   since $(I-H)X = X - X = 0$. Therefore $\hat{\boldsymbol{\beta}}$ and the residuals $\mathbf{e}$ are uncorrelated.
   Also, $\hat{Y} = X\hat{\beta}$.
   Hence $Cov(\hat{Y}, e) = Cov(X\hat{\beta}, e) = XCov(\hat{\beta}, e) = X.0 = 0$.
   Therefore $\hat{\mathbf{Y}}$ and the residuals $\mathbf{e}$ are uncorrelated.

   (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.)
   Clearly, $e = (I_n - H)Y$ and $d = (H - \frac{1}{n}J_n)Y$ are jointly normally distributed from

Hint. Also $Cov(e,d) = (I_n - H)Var(Y)(H - \frac{1}{n}J_n) = \sigma^2(H - H^2 - \frac{1}{n}J_n + (\frac{1}{n}J_n)^2) = 0$ as $H^2 = H$ and $(\frac{1}{n}J_n)^2 = \frac{1}{n}J_n$ as they are projection matrices.
Since e and d are jointly normally distributed and uncorrelated, they are independent. Hence, $SSE = e^Te$ and $SSR = d^Td$ being functions of e and d are also independent. From part (a), e and $\hat{\boldsymbol{\beta}}$ are uncorrelated and using Hint they are jointly normal. Hence e and $\hat{\boldsymbol{\beta}}$ are independent and so is $SSE = e^Te$ and $\hat{\boldsymbol{\beta}}$, SSE being a function of e.

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.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \qquad i = 1, \dots, 5$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} -0.97 \\ 2.51 \\ -0.19 \\ 6.53 \\ 1.00 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & -0.63 & -0.82 \\ 1 & 0.18 & 0.49 \\ 1 & -0.84 & 0.74 \\ 1 & 1.60 & 0.58 \\ 1 & 0.33 & -0.31 \end{bmatrix}$$

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

```
> t(X)%*%X
           X1     X2
   5.00 0.6400 0.6800
```

```
X1 0.64 3.8038 0.8089
X2 0.68 0.8089 1.8926
> t(X)%*%Y
      [,1]
    8.8800
X1 12.0005
X2  5.3621
> solve(t(X)%*%X)
                           X1          X2
    0.21184719 -0.02140278 -0.06696786
X1 -0.02140278  0.29134054 -0.11682948
X2 -0.06696786 -0.11682948  0.60236791
```

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

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1.265271 \\ 2.679724 \\ 1.233270 \end{bmatrix}$$

(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*)

```
> H = X%*%solve(t(X)%*%X)%*%t(X) # hat matrix
> H
            [,1]       [,2]        [,3]       [,4]        [,5]
[1,]  0.74859901 0.02181768  0.01132102 -0.1770289  0.39529119
[2,]  0.02181768 0.27197293  0.35049579  0.2534024  0.10231125
[3,]  0.01132102 0.35049579  0.82936038 -0.1072487 -0.08392853
[4,] -0.17702890 0.25340235 -0.10724866  0.7973084  0.23356681
[5,]  0.39529119 0.10231125 -0.08392853  0.2335668  0.35275928
```

$rank(\mathbf{H}) = n - p = 5 - 2 = 3$, $rank(\mathbf{I} - \mathbf{H}) = p = 2$.

(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)
```

```
> sum((Y-mean(Y))^2) # for SSTO
[1] 35.14712
> sum((Y-Yhat)^2)# for SSE
[1] 0.91145
> sum((Yhat-mean(Y))^2)# for SSR
[1] 34.23567
```

```
> Yhat
             [,1]
[1,] -1.43423719
[2,]  2.35192330
[3,] -0.07307774
[4,]  6.26812586
[5,]  1.76726576
> residuals
           [,1]
[1,]  0.4642372
[2,]  0.1580767
[3,] -0.1169223
[4,]  0.2618741
[5,] -0.7672658
> SSE
        [,1]
[1,] 0.91145
> MSE
         [,1]
[1,] 0.455725
```

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

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

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i, \qquad i = 1, \dots, 5$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

(i) 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?

$$\mathbf{X} = \begin{bmatrix} 1 & -0.63 & -0.82 & 0.5166 = (-0.63) \times (-0.82) \\ 1 & 0.18 & 0.49 & 0.0882 = 0.18 \times 0.49 \\ 1 & -0.84 & 0.74 & -0.6216 = (-0.84) \times 0.74 \\ 1 & 1.60 & 0.58 & 0.9280 = 1.60 \times 0.58 \\ 1 & 0.33 & -0.31 & -0.1023 = 0.33 \times (-0.31) \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} -0.97 \\ 2.51 \\ -0.19 \\ 6.53 \\ 1.00 \end{bmatrix}$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$rank(\mathbf{H}) = 4$, $rank(\mathbf{I} - \mathbf{H}) = 1$. $rank(\mathbf{H})$ is one more and $rank(\mathbf{I} - \mathbf{H})$ is one less, compared to part(d).

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

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1.051738 \\ 1.987286 \\ 1.804233 \\ 1.387774 \end{bmatrix}$$

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

$$\widehat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix} -0.963 \\ 2.416 \\ -0.145 \\ 6.566 \\ 1.006 \end{bmatrix} \quad \mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = \begin{bmatrix} 0.007 \\ 0.094 \\ -0.045 \\ 0.036 \\ -0.006 \end{bmatrix}$$

$SSE = \mathbf{e}'\mathbf{e} = 0.01223284$, $MSE = \frac{SSE}{d.f.(SSE)} = \frac{0.01223284}{5-4} = 0.01223284$.

(l) Which model fits the data better?

The second model fits the data better since it has a much smaller SSE therefore much larger R^2 ($R^2 = 1 - SSE/SSTO$ and SSTO is the same).

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

Yes. Define $\tilde{X}_{i2} = \log X_{i2}$, $X_{i3} = X_{i1}^2$,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 \log \tilde{X}_{i2} + \beta_3 X_{i3} + \epsilon_i.$$

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

Yes.

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

define $\tilde{Y}_i = \log(Y_i)$, $\tilde{X}_{i2} = X_{i2}^2$, and $\tilde{\epsilon}_i = \log(\epsilon_i)$,

$$\tilde{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 \tilde{X}_{i2} + \tilde{\epsilon}_i.$$

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

No.

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

Yes.

$$\log(1/Y_i - 1) = \beta_0 + \beta_1 X_{i1} + \epsilon_i,$$

define $\tilde{Y}_i = \log(1/Y_i - 1)$,

$$\tilde{Y}_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i.$$

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

$$\hat{\beta_1} \sim N(0, \frac{\sigma^2}{s_{xx}}) \quad \text{when} \quad \beta_1 = 0, \quad \text{where} \quad s_{xx} = \sum_{i=1}^{n} (X_i - \bar{X})^2. \tag{1}$$

$$\therefore \frac{\sqrt{s_{xx}}\hat{\beta}_1}{\sigma} \sim N(0,1) \tag{2}$$

$$\text{From homework 1,} \quad SSR = s_{xx}\hat{\beta}_1^{\,2} \sim \sigma^2\chi_1^2 \tag{3}$$

$$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\epsilon_i - \bar{\epsilon})^2 \sim \sigma^2\chi_{n-1}^2 \quad \text{from properties of normal distribution.} \tag{4}$$