Statistics 206

Homework 5

NOT Due

1. Tell true or false of the following statements.

   (a) The multiple coefficient of determination R^2 is always larger/not-smaller for models with more X variables.

   (b) If all the regression coefficients associated with the X variables are estimated to be zero, then $R^2 = 0$.

   (c) The adjusted multiple coefficient of determination R_a^2 may decrease when adding additional X variables into the model.

   (d) Models with larger R^2 is always preferred.

   (e) If the response vector is a linear combination of the columns of the design matrix $\mathbf{X}$, then the coefficient of multiple determination $R^2 = 1$.

2. **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) Calculate the trace of $\mathbf{H}$ and compare it with $rank(\mathbf{H})$ from part (d). What do you find?

   (f) 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?

3. Under the general linear regression model, show that the residuals are uncorrelated with the fitted values and the estimated regression coefficients.

4. Under the multiple regression model (with X variables $X_1, \cdots, X_{p-1}$), show the following.

   (a) The LS estimator of the regression intercept is:

   $$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \cdots - \hat{\beta}_{p-1} \overline{X}_{p-1},$$

   where $\hat{\beta}_k$ is the LS estimator of β_k, and $\overline{X}_k = \frac{1}{n} \sum_{i=1}^{n} X_{ik}$ $(k = 1, \cdots, p-1)$.

   (Hint: Plug in $\hat{\beta}_1, \cdots, \hat{\beta}_{p-1}$ to the least squares criterion function $Q(\cdot)$ and solve for b_0 that minimizes that function.)

   (b) SSE and the coefficient of multiple determination R^2 remain the same if we first center all the variables and then fit the regression model.

   (Hint: Use part (a) and the fact that SSE is the minimal value achieved by the Least Squares criterion function.)

5. **Multiple regression: read R output.** The following data set has 30 cases, one response variable Y and two predictor variables X_1, X_2.

```
case Y    X1      X2
1   2.86  0.36  2.14
2  -0.50  0.66  0.74
3   3.24  0.66  1.91
4   0.44 -0.52 -0.41
5   0.04 -0.68  0.45
...  ...   ...   ...
29  2.60  0.84 -0.49
30  0.98 -0.11  2.41
```

Consider fitting the nonadditive model with interaction between X_1 and X_2. (R output is given at the end.)

(a) Write down the first 4 rows of the design matrix $\mathbf{X}$.

(b) What is the error sum of squares of this model?

(c) We want to conduct prediction at $X_1 = 0, X_2 = 0$ and it is given that

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.087 & -0.014 & -0.035 & -0.004 \\ -0.014 & 0.115 & -0.012 & -0.052 \\ -0.035 & -0.012 & 0.057 & -0.014 \\ -0.004 & -0.052 & -0.014 & 0.050 \end{bmatrix}.$$

What is the predicted value? What is the prediction standard error? Construct a 95% prediction interval. (Note: You can try this question after next Wed.'s lecture; For now consider estimating the mean response.)

```
Call:
lm(formula = Y ~ X1 + X2 + X1:X2, data = data)

Residuals:
Min      1Q  Median      3Q     Max
-2.8660 -0.2055  0.1754  0.5436  2.0143

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9918     0.3006   3.299 0.002817 **
X1            1.5424     0.3455   4.464 0.000138 ***
X2            0.5799     0.2427   2.389 0.024433 *
X1:X2        -0.1491     0.2271  -0.657 0.517215
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.02 on 26 degrees of freedom
Multiple R-squared: 0.7035,     Adjusted R-squared: 0.6693
F-statistic: 20.56 on 3 and 26 DF,  p-value: 4.879e-07
```

6. (**Optional Problem.**) Under the Normal error model (with X variables $X_1, \cdots, X_{p-1}$), show that if $\beta_1 = \cdots = \beta_{p-1} = 0$, then $SSR \sim \sigma^2 \chi^2_{(p-1)}$ and $SSTO \sim \sigma^2 \chi^2_{(n-1)}$.

(Hint: use eigen- decomposition and the fact that $(\mathbf{H}_n - \frac{1}{n}\mathbf{J}_n)\mathbf{1}_n = \mathbf{0}$*.)*

7. **(Optional Problem). Expectation and covariance of quadratic forms.** Let $\mathbf{y}$ be a d-dimensional random vector with $E(\mathbf{y}) = \boldsymbol{\mu}$ and $Var(\mathbf{y}) = \boldsymbol{\Sigma}$. Let $\mathbf{A}$ and $\mathbf{B}$ be $d \times d$ symmetric matrices. Show the following:

   (a) $E(\mathbf{y}'\mathbf{A}\mathbf{y}) = tr(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}$.

   (b) Assume that $\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$Cov(\mathbf{y}'\mathbf{A}\mathbf{y}, \mathbf{y}'\mathbf{B}\mathbf{y}) = 2tr(\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}\boldsymbol{\Sigma}) + 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}\boldsymbol{\mu}.$$

   Specifically:

$$Var(\mathbf{y}'\mathbf{A}\mathbf{y}) = 2tr((\mathbf{A}\boldsymbol{\Sigma})^2) + 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\mu}.$$

   (Hint: (i) First show the above for $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}_d$; (ii) Use the fact: $X \sim N(0, \sigma^2)$, then $E(X^3) = 0$ and $E(X^4) = 3\sigma^4$.)

   (c) Use part (a) to derive $E(SSE)$ and $E(SSR)$.