Statistics 206

Homework 4

Due : October 26, 2015, In Class

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. 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.)

3. **Multiple regression.** 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) 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.

(c) What are the regression sum of squares and error sum of squares of this model? What is SSTO?

(d) Derive the following sum of squares:

$$SSR(X_1),\quad SSE(X_1),\quad SSR(X_2|X_1),\quad SSR(X_2, X_1 \cdot X_2|X_1),$$

$$SSR(X_1 \cdot X_2|X_1, X_2),\quad SSR(X_1, X_2),\quad SSE(X_1, X_2).$$

(e) Test whether both X_2 and the interaction term X_1X_2 can be dropped out of the model at level 0.01. Write down the full model and the reduced model. State the null and alternative hypotheses, test statistic and its null distribution, decision rule and the conclusion.

```
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
```

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 58.232  58.232 55.9752 6.067e-08 ***
X2         1  5.490   5.490  5.2775    0.0299 *
X1:X2      1  0.448   0.448  0.4311    0.5172
Residuals 26 27.048   1.040
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

4. **A multiple linear regression case study by R**. You should use R and the *lm()* function and its associated functions (e.g., *summary(), anova(), confint(), predict.lm()*) to do this problem. Please also attach your R codes and plots.

   A commercial real estate company evaluates age (X_1), operating expenses (X_2, in thousand dollar), vacancy rate (X_3), total square footage (X_4) and rental rates (Y, in thousand dollar) for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data are taken from 81 *suburban commercial properties.* (The data is on smartsite under Resources/Homework/property.txt; The first column is Y, followed by X_1, X_2, X_3, X_4.)

   (a) Read data into R. Draw the scatter plot matrix and obtain the correlation matrix. What do you observe?

   (b) Perform regression of the rental rates Y on the four predictors X_1, X_2, X_3, X_4 (Model 1). What are the Least-squares estimators? Write down the fitted regression function. What are MSE, R^2 and R_a^2?

   (c) Draw residuals vs. fitted values plot, residuals Normal Q-Q plot and residuals boxplot. Comment on the model assumptions based on these plots. (Hint: for a compact report, please use *par(mfrow)* to create one multiple paneled plot).

   (d) Draw residuals vs. each predictor variable plots, and residuals vs. each two-way interaction term plots. How many two-way interaction terms are there? Analyze your plots and summarize your findings. (Hint: again, use the *par(mfrow)* to create a few multiple paneled plots; Otherwise, there will be too many pages of plots!).

   (e) For each regression coefficient, test whether it is zero or not (under the Normal error model) at level 0.01. State the null and alternative hypotheses, the test statistic, its null distribution and the pvalue. Which regression coefficient(s) is (are) significant, which is/are not? What is the implication? (Hint: You can find relevant information from the R *summary()* output.)

   (f) Obtain SSTO, SSR, SSE and their degrees of freedom. Summarize these into an ANOVA table. Test whether there is a regression relation at $\alpha = 0.01$. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and your conclusion. (Hint: You can find relevant information from the R *anova()* output.)

(g) You now decide to fit a different model by regressing the rental rates Y on three predictors X_1, X_2, X_4 (Model 2). Why would you make such a decision? Get the Least-squares estimators and write down the fitted regression function. What are MSE, R^2 and R_a^2? How do these numbers compare with those from Model 1?

(h) Compare the standard errors of the regression coefficient estimates for X_1, X_2, X_4 under Model 2 with those under Model 1. What do you find? Construct 95% confidence intervals for regression coefficients for X_1, X_2, X_4 under Model 2. If these intervals were constructed under Model 1, how would their widths compare with the widths of the intervals you just constructed, i.e., being wider or narrower? Justify your answer.

(i) Consider a property with the following characteristics: $X_1 = 4, X_2 = 10, X_3 = 0.1, X_4 = 80,000$. Construct 99% prediction intervals under Model 1 and Model 2, respectively. Compare these two sets of intervals, what do you find?

(j) Which of the two Models you would prefer and why?

5. **(Optional Problem.)** Show the following with regard to the multiple coefficient of determination R^2.

(a)

$$R^2 = \widehat{Corr}^2(Y, \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_{p-1} X_{p-1}) = \widehat{Corr}^2(Y, \widehat{Y}).$$

(Hint: By 2 (b), we can assume that all the variables are centered and then show $R^2 = \frac{(\mathbf{Y}^T \mathbf{X}\hat{\boldsymbol{\beta}})^2}{(\mathbf{Y}^T\mathbf{Y})(\mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{X}\hat{\boldsymbol{\beta}})}$*.)*

(b)

$$R^2 = \max_{b_1, \cdots, b_{p-1}} \widehat{Corr}^2(Y, b_1 X_1 + \cdots + b_{p-1} X_{p-1}).$$

(Hint: Show that $(b_1, \cdots, b_{p-1})$ *that maximize the function* $\widehat{Corr}^2(Y, b_1 X_1 + \cdots + b_{p-1} X_{p-1})$ *are the LS estimators* $(\hat{\beta}_1, \cdots, \hat{\beta}_{p-1})$ *and then use part (a).)*