Statistics 206

Homework 6

Due : Nov. 13, 2019, In Class

1. Tell true or false of the following statements.

   (a) If the response variable is uncorrelated with all X variables in the model, then the least-squares estimated regression coefficients of the X variables are all zero.

   (b) Even when the X variables are perfectly correlated, we might still get a good fit of the data.

   (c) Taking correlation transformation of the variables will not change coefficients of multiple determination.

   (d) If all the X variables are uncorrelated, then the magnitude and the sign of a standardized regression coefficient reflect the comparative importance and direction of effect, respectively, of the corresponding X variable, in terms of explaining the response variable.

   (e) In a regression model, it is possible that none of the X variables is statistically significant when being tested individually, while there is a significant regression relation between the response variable and the set of X variables as a whole.

   (f) In a regression model, it is possible that some of the X variables are statistically significant when being tested individually, while there is no significant regression relation between the response variable and the set of X variables as a whole.

   (g) If an X variable is uncorrelated with the rest of the X variables, then in the standardized model, the variance of its least-squares estimated regression coefficient equals to the error variance.

   (h) If an X variable is uncorrelated with the response variable, then its least-squares estimated regression coefficient must be zero.

   (i) If an X variable is uncorrelated with the response variable and also is uncorrelated with the rest of the X variables, then its least-squares estimated regression coefficient must be zero.

2. Consider a general linear model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \cdots, n.$$

   Describe how you would test:

   (a)
$$H_0 : \beta_1 = \beta_{10}, \quad \beta_2 = \beta_{20} \quad vs. \quad H_a : \text{not every equality in } H_0 \text{ holds},$$
   where β_{10} and β_{20} are two prespecified constants.

(b)

$$H_0 : \beta_1 = \beta_2 \ \ vs. \ \ H_a : \beta_1 \neq \beta_2$$

3. **Uncorrelated X variables**. When $X_1, \cdots, X_{p-1}$ are uncorrelated, show the following results. (Hint: Show these results under the standardized regression model and then transform them back to the original model.)

   (a) The fitted regression coefficients of regressing Y on $(X_1, \cdots, X_{p-1})$ equal to the fitted regression coefficients of regressing Y on each individual X_j $(j = 1, \cdots, p-1)$ alone.

   (b) Let $\mathcal{I} := \{k : 1 \leq k \leq p-1, k \neq j\}$. Show that

   $$SSR(X_j | X_{\mathcal{I}}) = SSR(X_j),$$

   where $SSR(X_j)$ denotes the regression sum of squares when regressing Y on X_j alone.

4. **Variance Inflation Factor for models with 2 X variables.** Show that for a model with two X variables, X_1 and X_2, the variance inflation factors are

$$VIF_1 = VIF_2 = \frac{1}{1 - R_1^2} = \frac{1}{1 - R_2^2}.$$

(Hint: Note $R_1^2 = R_2^2 = r_{12}^2$, where r_{12} is the sample correlation coefficient between X_1 and X_2.)

5. **Multiple regression (cont'd).** 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) What are the regression sum of squares and error sum of squares of this model? What is SSTO?

   (b) Derive the following sum of squares:

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

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

```
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
```

6. **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 canvas under Files/Homework/property.txt; The first column is Y, followed by X_1, X_2, X_3, X_4.)

   (a) Read data into R. What is the type of each variable? Draw plots to depict the distribution of each variable and obtain summary statistics for each variable. Comment

on the distributions of these variables.

(b) Draw the scatter plot matrix and obtain the correlation matrix. What do you observe?

(c) 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?

(d) 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).

(e) 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.

(f) 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: Use R outputs.)

(g) 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.

(h) 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?

(i) 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.

(j) 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?

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

7. **(Commericial Property Cont'd). Standardized Regression model.** 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 Files/Homework/property.txt; The first column is Y, followed by X_1, X_2, X_3, X_4.)

(a) Calculate the sample mean and sample standard deviation of each variable. Perform the correlation transformation. What are sample means and sample standard deviations of the transformed variables?

(b) Write down the model equation for the the standardized first-order regression model with all four transformed X variables and fit this model. What is the fitted regression intercept?

(c) Transform the fitted standardized regression coefficients back to the fitted regression coefficients of the original model. Do you get the same results as those from Homework 5?

(d) Obtain the standard errors of the fitted regression coefficients (for X variables) of the original model using the standard errors of the fitted standardized regression coefficients. Compare the results with those from the R output of Problem 5.

(e) Obtain SSTO, SSE and SSR under the standardized model and compare them with those from the original model. What do you find?

(f) Calculate R^2, R_a^2 under the standardized model and compare them with R^2, R_a^2 under the original model. What do you find?

8. **(Commericial Property Cont'd). Multicollinearity.**

(a) Obtain $\mathbf{r}_{XX}^{-1}$ and get the variance inflator factors VIF_k $(k = 1, 2, 3, 4)$. Obtain R_k^2 by regressing X_k to $\{X_j : 1 \leq j \neq k \leq 4\}$ $(k = 1, 2, 3, 4)$. Confirm that

$$VIF_k = \frac{1}{1 - R_k^2}, \quad k = 1, 2, 3, 4.$$

Comment on the degree of multicollinearity in this data.

(b) Fit the regression model for relating Y to X_4 and fit the regression model for relating Y to X_3, X_4. Compare the estimated regression coefficients of X_4 in these two models. What do you find? Calculate $SSR(_4)$ and $SSR(X_4|X_3)$. What do you find? Provide an interpretation for your observations.

(c) Fit the regression model for relating Y to X_2 and fit the regression model for relating Y to X_2, X_4. Compare the estimated regression coefficients of X_2 in these two models. What do you find? Calculate $SSR(_2)$ and $SSR(X_2|X_4)$. What do you find? Provide an interpretation for your observations.

9. **(Optional Problem) Variance Inflation Factor.** Use the formula for the inverse of a partitioned matrix to show:

$$r_{XX}^{-1}(k, k) = \frac{1}{1 - R_k^2},$$

i.e., the kth diagonal element of the inverse correlation matrix equals to $\frac{1}{1-R_k^2}$, where R_k^2 is the coefficient of multiple determination by regressing X_k to the rest of the X variables.

Hints: (i) Assume all X variables are standardized by the correlation transformation; (ii) You only need to prove this for $k = 1$ because you can permute the rows and columns of r_{XX} and r_{XY} to get the result for other k; (iii) Apply the inverse formula below with $A = r_{XX}$ and $A_{11} = r_{11}$, i.e., the first diagonal element of r_{XX}.

Inverse of a partitioned matrix. Suppose A is a $(p+q) \times (p+q)$ square matrix $(p, q \geq 1)$:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where A_{11} is a $p \times p$ square matrix and A_{22} is a $q \times q$ square matrix. Suppose A_{11} and A_{22} are invertible. Then A is invertible and

$$A^{-1} = \begin{bmatrix} \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1} & -\left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1} A_{12}A_{22}^{-1} \\ -\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1} A_{21}A_{11}^{-1} & \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1} \end{bmatrix}$$