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.

   TRUE. r_{XY} is a zero vector, so $\hat{\beta}_k^* = 0$ and $\hat{\beta}_k = 0$ for $k = 1, \cdots, p-1$.

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

   TRUE. Because the projection to the column space of the design matrix is still well defined.

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

   TRUE. Since these are defined as ratios of two sum of squares and the changes of scale on the numerator and denominator are canceled out.

   (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.

   TRUE.

   (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.

   TRUE. Since when testing an individual X variable, there may be other correlated X variables in the reduced model, while when testing the regression relation, the reduced model does not contain any X variable.

   (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.

   TRUE. Suppose there are two factors that mainly explain the variation in the data and are statistically significant when tested individually. But now if we throw a bunch of X variables which has no effect on the outcome, which will not increase our SSR but the number of variables increases. The problem of this setting is the loss of degrees of freedom, $MSE = SSE/(n-p)$. If SSR and SSE remains roughly the same, with larger p, MSE becomes larger while $MSR = SSR/(p-1)$, MSR becomes smaller, so F^* will decrease. So there will be 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.

TRUE. r_{XX} matrix is block diagonal. (Another explanation: $R_k^2 = 0$ so $VIF_k = 1$.)

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

FALSE. With other correlated X variables in the model, the regression coefficient of an X variable could be nonzero even when it is uncorrelated with the response variable.

(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.

TRUE. Consider the standardized model, and denote the set of the rest of the X variables by $\tilde{X}$. Then the correlation matrices:

$$\mathbf{r}_{XX} = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{r}_{\tilde{X}\tilde{X}} \end{bmatrix}, \quad \mathbf{r}_{XY} = \begin{bmatrix} \mathbf{0} \\ \mathbf{r}_{\tilde{X}Y \cdot} \end{bmatrix}$$

The fitted standardized regression coefficients:

$$\hat{\boldsymbol{\beta}}^* = \mathbf{r}_{XX}^{-1}\mathbf{r}_{XY} = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{r}_{\tilde{X}\tilde{X}}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{r}_{\tilde{X}Y \cdot 1} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{r}_{\tilde{X}\tilde{X}}^{-1}\mathbf{r}_{\tilde{X}Y} \end{bmatrix}.$$

Note that $\hat{\beta}_1^* = 0$.

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)

$$H_0 : \beta_1 = \beta_{10}, \quad \beta_2 = \beta_{20} \;\; vs. \;\; H_a : \text{not every equality in } H_0 \text{ holds},$$

where β_{10} and β_{20} are two prespecified constants.
Define

$$\tilde{Y}_i = Y_i - \beta_{10}X_{i1} - \beta_{20}X_{i2},$$

then the reduced model is defined as

$$\tilde{Y}_i = \beta_0 + \beta_3 X_{3i} + \epsilon_i.$$

As in lecture note, we define F^*

$$F^* = \frac{\frac{SSE(reduced) - SSE(full)}{df(reduced) - df(full)}}{\frac{SSE(full)}{df(full)}},$$

We would reject the null at level α if $F^* > F(1-\alpha, df(reduced) - df(full), df(full))$

(b)

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

Define a vector $c = (0, 1, -1, 0)$, then H_0 could be written as $c^T\beta = 0$. Then

$$T^* = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{c^T var(\beta) c}}$$

We reject H_0 if $|T*| > t(1 - \alpha/2; n - 4)$.

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.

Proof. Under the standardized model,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^* &= (\mathbf{X}^{*'}\mathbf{X}^*)^{-1}\mathbf{X}^{*'}\mathbf{Y} \\
&= \begin{bmatrix} \frac{1}{n} & \mathbf{0}' \\ \mathbf{0} & \mathbf{I}_{p-1} \end{bmatrix} \begin{bmatrix} n\overline{Y} \\ \sqrt{n-1} s_Y r_{Y1} \\ \vdots \\ \sqrt{n-1} s_Y r_{Y,p-1} \end{bmatrix} \\
&= \begin{bmatrix} \overline{Y} \\ \sqrt{n-1} s_Y r_{Y1} \\ \vdots \\ \sqrt{n-1} s_Y r_{Y,p-1} \end{bmatrix}.
\end{aligned}$$

Note:

$$\hat{\beta}_k = \frac{1}{\sqrt{n-1} s_{X_k}} \hat{\beta}_k^*, \;\; k = 1, 2, \ldots, p-1.$$

Then we have, for $k = 1, 2, \ldots, p-1$,

$$\begin{aligned}
\hat{\beta}_k &= \frac{1}{\sqrt{n-1} s_{X_k}} \hat{\beta}_k^* \\
&= \frac{\sqrt{n-1} s_Y r_{Yk}}{\sqrt{n-1} s_{X_k}} \\
&= \frac{s_Y}{s_{X_k}} r_{Yk}.
\end{aligned}$$

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

(b) Let $\mathcal{I} := \{k : 1 \le k \le p-1, k \ne j\}$. 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.

Proof. We have,

$$\begin{aligned}
SSE(X_{\mathcal{I}}^*) - SSE(X_{\mathcal{I}}^*, X_j^*) &= Y^T(I - H(X_{\mathcal{I}}^*))Y - Y^T(I - H(X_{\mathcal{I}}^*, X_j^*))Y \\
&= Y^T(H(X_{\mathcal{I}}^*, X_j^*) - H(X_{\mathcal{I}}^*))Y \\
&= Y^T(n^{-1}11^T + X_{\mathcal{I}}^* X_{\mathcal{I}}^{*T} + X_j^* X_j^{*T} - n^{-1}11^T \\
&\qquad\qquad - X_{\mathcal{I}}^* X_{\mathcal{I}}^{*T})Y \\
&= Y^T X_j^* X_j^{*T} Y
\end{aligned}$$

$$\begin{aligned}
SSR(X_j^*) &= Y^T(H(X_j^*) - J_n)Y \\
&= Y^T(n^{-1}11^T + X_j^* X_j^{*T} - n^{-1}J_n)Y \\
&= Y^T X_j^* X_j^{*T} Y
\end{aligned}$$

Thus,

$$\begin{aligned}
LHS &= SSE(X_{\mathcal{I}}) - SSE(X_{\mathcal{I}}, X_j) \\
&= SSE(X_{\mathcal{I}}^*) - SSE(X_{\mathcal{I}}^*, X_j^*) \\
&= SSR(X_j^*) \\
&= SSTO - SSE(X_j^*) \\
&= SSTO - SSE(X_j) \\
&= RHS
\end{aligned}$$

□

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.)

Proof. For a model with two X variables,

$$r_{XX} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$$

$$r_{XX}^{-1} = \frac{1}{1 - r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{bmatrix}$$

So $VIF_1 = VIF_2 = \frac{1}{1-r_{12}^2}$. □

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?

$SSR = 58.232+5.490+0.448 = 64.17$. $SSE = 27.048$. $SSTO = 64.17 + 27.048 = 91.218$.

(b) Derive the following sum of squares:

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

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

$SSR(X_1) = 58.232$. $SSE(X_1) = SSTO - SSR(X_1) = 32.986$. $SSR(X_2, X_1 \cdot X_2|X_1) = 5.490 + 0.448 = 5.938$.

$SSR(X_1 \cdot X_2|X_1, X_2) = 0.448$. $SSR(X_1, X_2) = 58.232+5.490 = 63.722$. $SSE(X_1, X_2) = 27.048 + 0.448 = 27.496$.

```
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
```

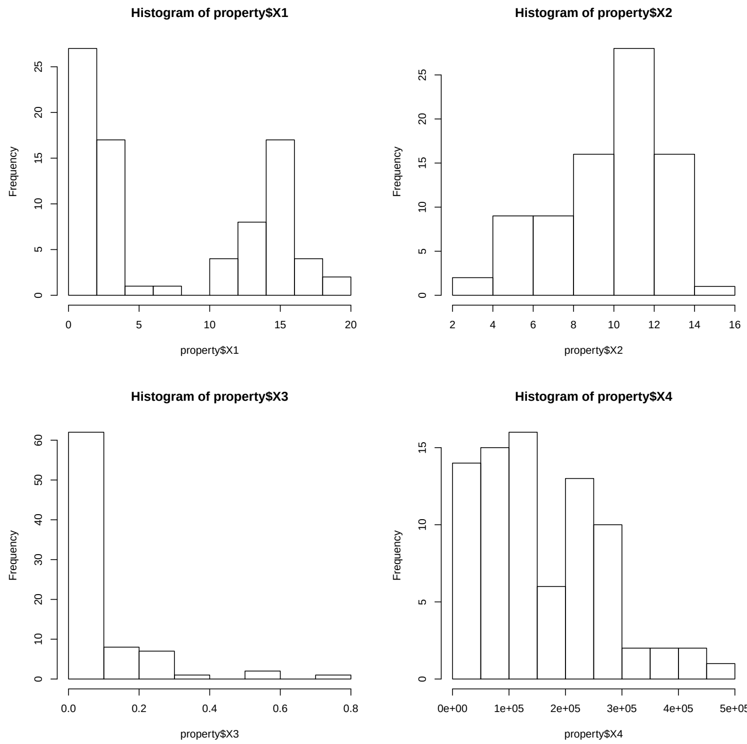
6. **A multiple linear regression case study by R**. You should use R and the *lm()* function and its associated functions (e.g., *summary()*, *anova()*, *confint()*, *predict.lm()*) to do this problem. Please also attach your R codes and plots.

   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 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. 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.

   ```
   > property=read.table('property.txt',header=FALSE)
   > names(property)=c('Y','X1','X2','X3','X4')
   > par(mfrow=c(2,2))
   > hist(property$X1)
   > hist(property$X2)
   > hist(property$X3)
   > hist(property$X4)
   > summary(property$X1)
   > summary(property$X2)
   > summary(property$X3)
   > summary(property$X4)
   ```

Age and total square footage are discrete variables. The other variables are continuous.

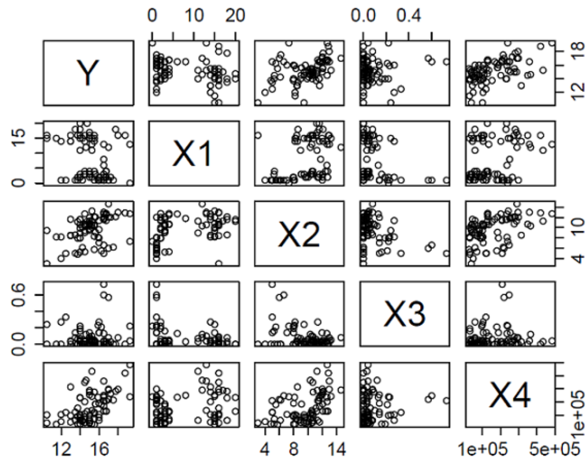



```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.000   2.000   4.000   7.864  15.000  20.000
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 3.000   8.130  10.360   9.688  11.620  14.620
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.00000 0.00000 0.03000 0.08099 0.09000 0.73000
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27000   70000  129600  160600  236000  484300
```

Age is bimodal; "operating expenses" is left-skewed; vacancy rate is right-skewed with lots of zeros; total square footage is right-skewed.

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

```
> pairs(property)
```



No obvious nonlinearity.

```
> cor(property)

Y           X1          X2           X3          X4
Y   1.00000000 -0.2502846  0.4137872  0.06652647 0.53526237
X1 -0.25028456  1.0000000  0.3888264 -0.25266347 0.28858350
X2  0.41378716  0.3888264  1.0000000 -0.37976174 0.44069713
X3  0.06652647 -0.2526635 -0.3797617  1.00000000 0.08061073
X4  0.53526237  0.2885835  0.4406971  0.08061073 1.00000000
```

X_1 and X_3, X_2 and X_3, X_1 and Y are negatively correlated, X_3 and X_4, X_3 and Y are not much correlated, other pairs are moderately positively correlated.

(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?

```
> fit1=lm(Y~X1+X2+X3+X4,data=property)
> summary(fit1)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4, data = property)

Residuals:
Min       1Q  Median      3Q     Max
-3.1872 -0.5911 -0.0910  0.5579  2.9441

Coefficients:
Estimate Std. Error t value Pr(>|t|)
```

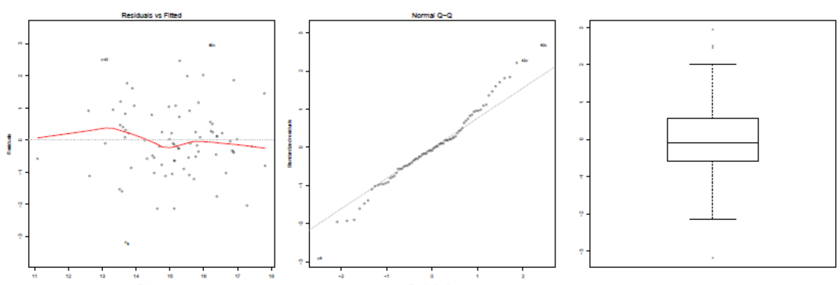
```
(Intercept)  1.220e+01  5.780e-01  21.110  < 2e-16 ***
X1          -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
X2           2.820e-01  6.317e-02   4.464 2.75e-05 ***
X3           6.193e-01  1.087e+00   0.570     0.57
X4           7.924e-06  1.385e-06   5.722 1.98e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.137 on 76 degrees of freedom
Multiple R-squared: 0.5847,Adjusted R-squared: 0.5629
F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

Fitted regression function:

$$Y = 12.2 - 0.142X_1 + 0.282X_2 + 0.619X_3 + 7.92 \times 10^{-6} X_4$$

$MSE = 1.137^2 = 1.293$, $R^2 = 0.5847$, $R_a^2 = 0.5629$.

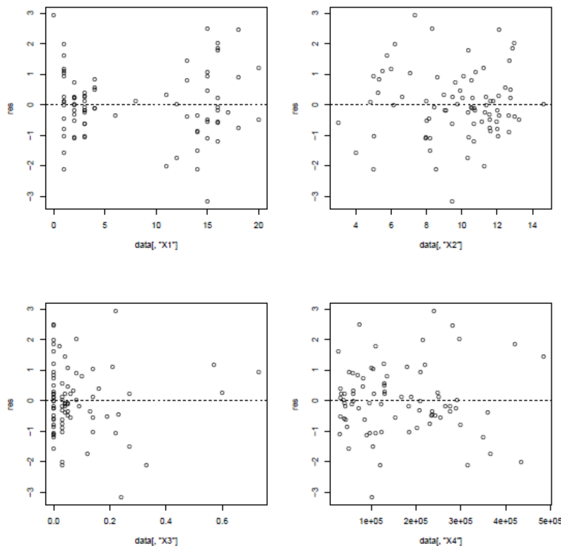
(d) Draw residuals vs. fitted values plot, residuals Normal Q-Q plot and residuals box-plot. Comment on the model assumptions based on these plots. (Hint: for a compact report, please use *par(mfrow)* to create one multiple paneled plot).



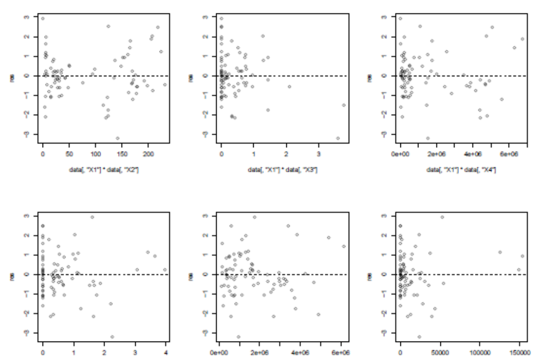
Residuals vs. fitted values plot shows no obvious nonlinearity. Residuals Q-Q plot shows slightly heavy tails. Residuals boxplot shows that most of the residuals are in between -2 and 2 and residual distribution is nearly symmetric.

(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.

Residuals vs. each predictor:



No obvious pattern.
Residuals vs. each two-way interaction (6 in total):



No obvious pattern.

(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.)

- $H_0 : \beta_0 = 0 \;\; vs. \;\; H_a : \beta_0 \neq 0$, $T^* = 21.11$, Under H_0, $T^* \sim t_{(76)}$, $pvalue < 2 \times 10^{-16}$
- $H_0 : \beta_1 = 0 \;\; vs. \;\; H_a : \beta_1 \neq 0$, $T^* = -6.655$, Under H_0, $T^* \sim t_{(76)}$, $pvalue = 3.89 \times 10^{-9}$

- $H_0 : \beta_2 = 0 \;\; vs. \;\; H_a : \beta_2 \neq 0$, $T^* = 4.464$, Under H_0, $T^* \sim t_{(76)}$, $pvalue = 2.75 \times 10^{-5}$
- $H_0 : \beta_3 = 0 \;\; vs. \;\; H_a : \beta_3 \neq 0$, $T^* = 0.57$, Under H_0, $T^* \sim t_{(76)}$, $pvalue = 0.57$
- $H_0 : \beta_4 = 0 \;\; vs. \;\; H_a : \beta_4 \neq 0$, $T^* = 5.722$, Under H_0, $T^* \sim t_{(76)}$, $pvalue = 1.98 \times 10^{-7}$

$\beta_0, \beta_1, \beta_2$ and β_4 are significant and β_3 is not significant. This implies that we could consider dropping X_3 from the model.

(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.

```
> anova(fit1)
Analysis of Variance Table

Response: Y
Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 14.819  14.819 11.4649  0.001125 **
X2         1 72.802  72.802 56.3262 9.699e-11 ***
X3         1  8.381   8.381  6.4846  0.012904 *
X4         1 42.325  42.325 32.7464 1.976e-07 ***
Residuals 76 98.231   1.293
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Source of Variation	SS	d.f.	MS	F^*
Regression	$SSR = 138.327$	4	$MSR = 34.58175$	$F^* = 26.75543$
Error	$SSE = 98.231$	76	$MSE = 1.292513$	
Total	$SSTO = 236.558$	80		

Note $SSR = 14.819 + 72.802 + 8.381 + 42.325 = 138.27$, and $SSTO = SSR + SSE = 236.558$.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \;\; vs.$$

$$H_a : \text{ not all } \beta_k \; (k = 1, 2, 3, 4) \text{ equal zero.}$$

$$F^* = \frac{MSR}{MSE} = 26.75543$$

Under H_0, $F^* \sim F_{4,76}$.

```
> qf(0.99,4,76)
[1] 3.57652
```

Since $F^* = 26.75543 > 3.58 = F(0.99; 4, 76)$, reject H_0 and conclude that there is regression relation between Y and the set of X variables $\{X_1, X_2, X_3, X_4\}$.

(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?

We consider Model 2 because β_3 is not significant (from part (e)).

```
> fit2=lm(Y~X1+X2+X4,data=property)
> summary(fit2)

Call:
lm(formula = Y ~ X1 + X2 + X4, data = property)

Residuals:
Min      1Q  Median      3Q     Max
-3.0620 -0.6437 -0.1013  0.5672  2.9583

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.237e+01  4.928e-01  25.100  < 2e-16 ***
X1          -1.442e-01  2.092e-02  -6.891 1.33e-09 ***
X2           2.672e-01  5.729e-02   4.663 1.29e-05 ***
X4           8.178e-06  1.305e-06   6.265 1.97e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.132 on 77 degrees of freedom
Multiple R-squared: 0.583,Adjusted R-squared: 0.5667
F-statistic: 35.88 on 3 and 77 DF,  p-value: 1.295e-14
```

Fitted regression function:

$$Y = 12.37 - 0.1442X_1 + 0.2672X_2 + 8.178 \times 10^{-6} X_4$$

$MSE = 1.132^2 = 1.281$, $R^2 = 0.583$, $R_a^2 = 0.5667$. Compared to Model 1, MSE is a little bit smaller(1.281 vs. 1.293), R^2 is a little bit smaller(0.583 vs. 0.5847) but R_a^2 is a little bit larger(0.5667 vs. 0.5629).

(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.

The standard errors of the regression coefficient estimates are smaller under Model 2. For $\hat{\beta}_1$, 2.134×10^{-2}(Model 1)$>$ 2.092×10^{-2}(Model 2); For $\hat{\beta}_2$, 6.317×10^{-2}(Model

1)> 5.729×10^{-2}(Model 2); For $\hat{\beta}_4$, 1.385×10^{-6}(Model1)> 1.305×10^{-6}(Model 2)
95% confidence interval under Model 2:

```
> cf2=confint(fit2,parm=c('X1','X2','X4'),level=.95)
> cf2
2.5 %         97.5 %
X1 -1.858219e-01 -1.025074e-01
X2  1.530784e-01  3.812557e-01
X4  5.578873e-06  1.077755e-05
```

If these intervals were constructed under Model 1, their width would be wider since the standard errors of the regression coefficient estimates are larger in Model 1, as well as the multipliers (t-percentiles) due to less degrees of freedom under Model 1. We can check this in R:

```
> cf1=confint(fit1,parm=c('X1','X2','X4'),level=.95)
> cf1          # confidence interval under Model 1
2.5 %         97.5 %
X1 -1.845411e-01 -9.952615e-02
X2  1.561979e-01  4.078352e-01
X4  5.166283e-06  1.068232e-05
> cf2[,2]-cf2[,1]          # width under Model 2
X1           X2           X4
8.331458e-02 2.281773e-01 5.198675e-06
> cf1[,2]-cf1[,1]          # width under Model 1
X1           X2           X4
8.501498e-02 2.516373e-01 5.516038e-06
```

(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?

```
> newX=data.frame(X1=4,X2=10,X3=0.1,X4=80000)
```

99% prediction interval under Model 1:

```
> predict.lm(fit1, newX, interval="prediction", level=0.99, se.fit=TRUE)
$fit
fit     lwr      upr
1 15.1485 12.1027 18.19429

$se.fit
[1] 0.1908982

$df
[1] 76
```

```
$residual.scale
[1] 1.136885
```

$$s(pred) = \sqrt{0.1908982^2 + 1.293} = 1.153$$

99% prediction interval under Model 2:

```
> predict.lm(fit2, newX, interval="prediction", level=0.99, se.fit=TRUE)
$fit
fit      lwr      upr
1 15.11985 12.09134 18.14836

$se.fit
[1] 0.1833524

$df
[1] 77

$residual.scale
[1] 1.131889
```

$$s(pred) = \sqrt{0.1833524^2 + 1.281} = 1.147$$

The width of the interval is narrower and the standard error is smaller under Model 2.

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

We prefer Model 2 because it is simpler (less X variables) and essentially the same goodness of fit (R^2 similar to that of Model 1). It also has smaller standard errors and more degrees of freedom, resulting in narrower confidence intervals and prediction intervals.

7. **(Commercial 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 Resources/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?

See the separate pdf file generated by RMarkdown.

8. **(Commercial 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.

See the separate pdf file generated by RMarkdown.

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

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

Proof. Assume X has been standardized since it does not change r_{XX} and R_k^2. We define:

$$\mathbf{X}_{(-1)} = \begin{bmatrix} X_{12} & \dots & X_{1,p-1} \\ X_{22} & \dots & X_{2,p-1} \\ \vdots & \vdots & \vdots \\ X_{n2} & \dots & X_{n,p-1} \end{bmatrix}, \mathbf{X}_1 = \begin{bmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{n1} \end{bmatrix}.$$

Hence,

$$\begin{aligned} r_{XX}^{-1}(1,1) &= (r_{11} - r_{1\mathbf{X}_{(-1)}} r^{-1}_{\mathbf{X}_{(-1)}\mathbf{X}_{(-1)}} r_{\mathbf{X}_{(-1)}1})^{-1} \\ &= (r_{11} - [r_{1\mathbf{X}_{(-1)}} r^{-1}_{\mathbf{X}_{(-1)}\mathbf{X}_{(-1)}}] r_{\mathbf{X}_{(-1)}\mathbf{X}_{(-1)}} [r^{-1}_{\mathbf{X}_{(-1)}\mathbf{X}_{(-1)}} r_{\mathbf{X}_{(-1)}1}])^{-1} \\ &= (1 - \hat{\beta}'_{1\mathbf{X}_{(-1)}} \mathbf{X}'_{(-1)} \mathbf{X}_{(-1)} \hat{\beta}_{1\mathbf{X}_{(-1)}})^{-1}, \end{aligned}$$

where $\hat{\beta}_{1\mathbf{X}_{(-1)}}$ is the regression coefficients of X_1 on $X_2, \dots, X_{p-1}$ except the intercept. In fact, the intercept should be zero, since all X variables are standardized with mean zero. On the other hand, (it would be more straightforward if we can write everything in explicit matrix form)

$$\begin{aligned} R_1^2 = \frac{SSR}{SSTO} &= \frac{\hat{\beta}'_{1\mathbf{X}_{(-1)}} \mathbf{X}'_{(-1)} \mathbf{X}_{(-1)} \hat{\beta}_{1\mathbf{X}_{(-1)}}}{1} \\ &= \hat{\beta}'_{1\mathbf{X}_{(-1)}} \mathbf{X}'_{(-1)} \mathbf{X}_{(-1)} \hat{\beta}_{1\mathbf{X}_{(-1)}}. \end{aligned}$$

Therefore

$$VIF_1 = r_{XX}^{-1}(1,1) = \frac{1}{1 - R_1^2}.$$

□