HW5_Question6

Wookyeong Song (most of them from Yan-Yu Chen)

2022/11/3

A multiple linear regression case study by R.

You need to submit your codes alongside with the answers, plots, outputs, etc. You are required to use R Markdown: Please submit a .rmd file and its corresponding .html file.

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. (data file: property.txt; 1st 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,
                     col.names =c('Y', paste('X',1:4, sep = "")))
```

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

```
par(mfrow=c(3,2))
hist(property$X1)
hist(property$X2)
hist(property$X3)
hist(property$X4)
hist(property$Y)
summary(property$X1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   4.000   7.864  15.000  20.000
```

```
summary(property$X2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   8.130  10.360   9.688  11.620  14.620
```

```
summary(property$X3)
```

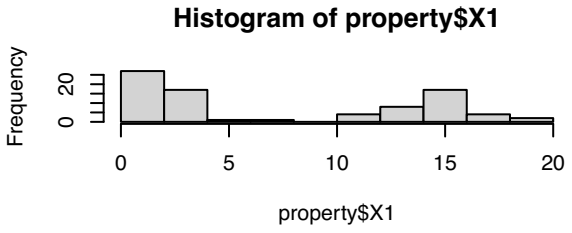
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.03000 0.08099 0.09000 0.73000
```

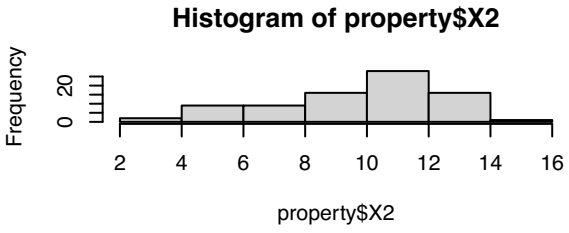
```
summary(property$X4)
```

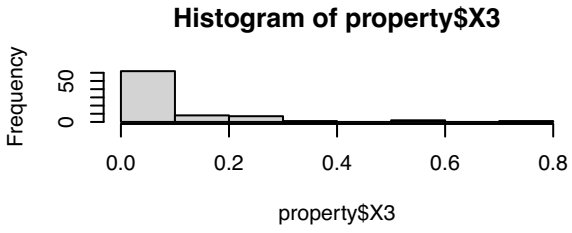
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   27000   70000  129614  160633  236000  484290
```

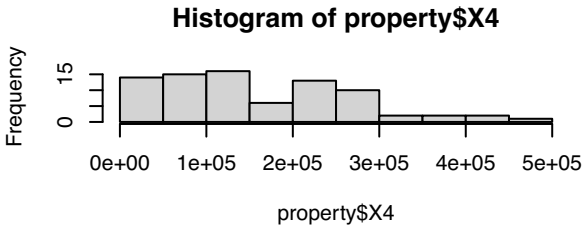
```
summary(property$Y)
```

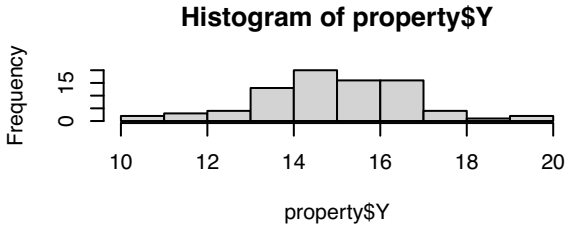
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.50   14.00   15.00   15.14   16.50   19.25
```







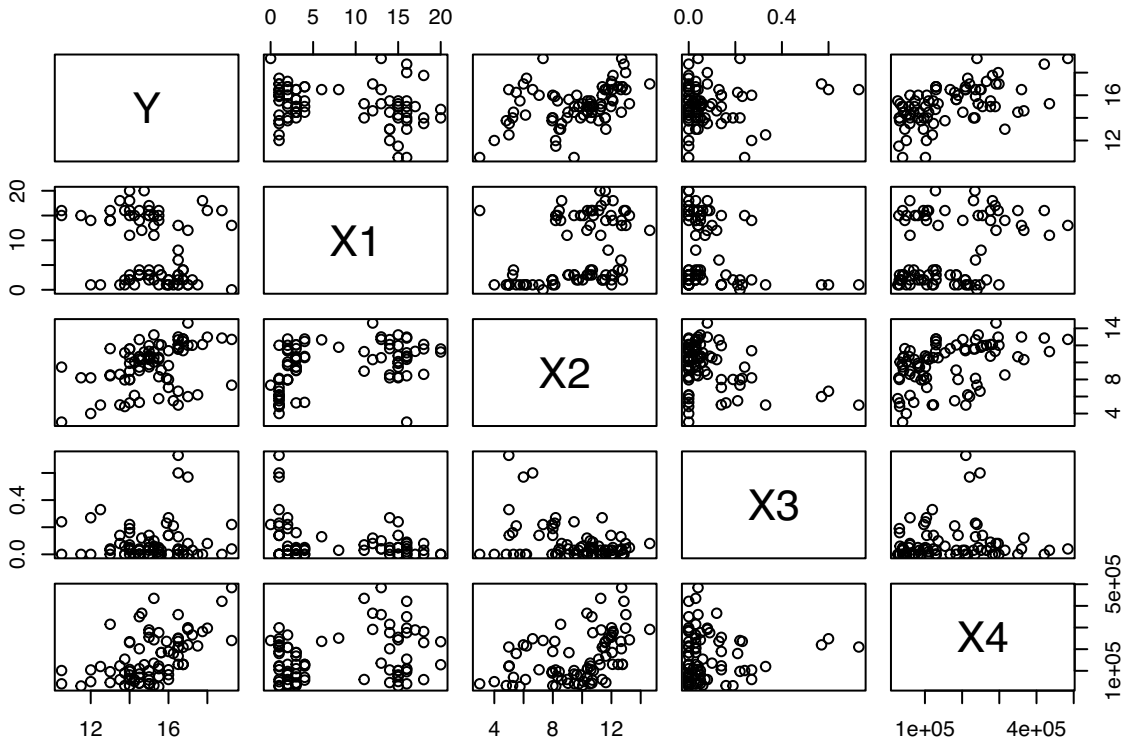




Age is bimodal; "operating expenses" is left-skewed; vacancy rate is right-skewed with lots of zeros; total square footage is right-skewed.

(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
```

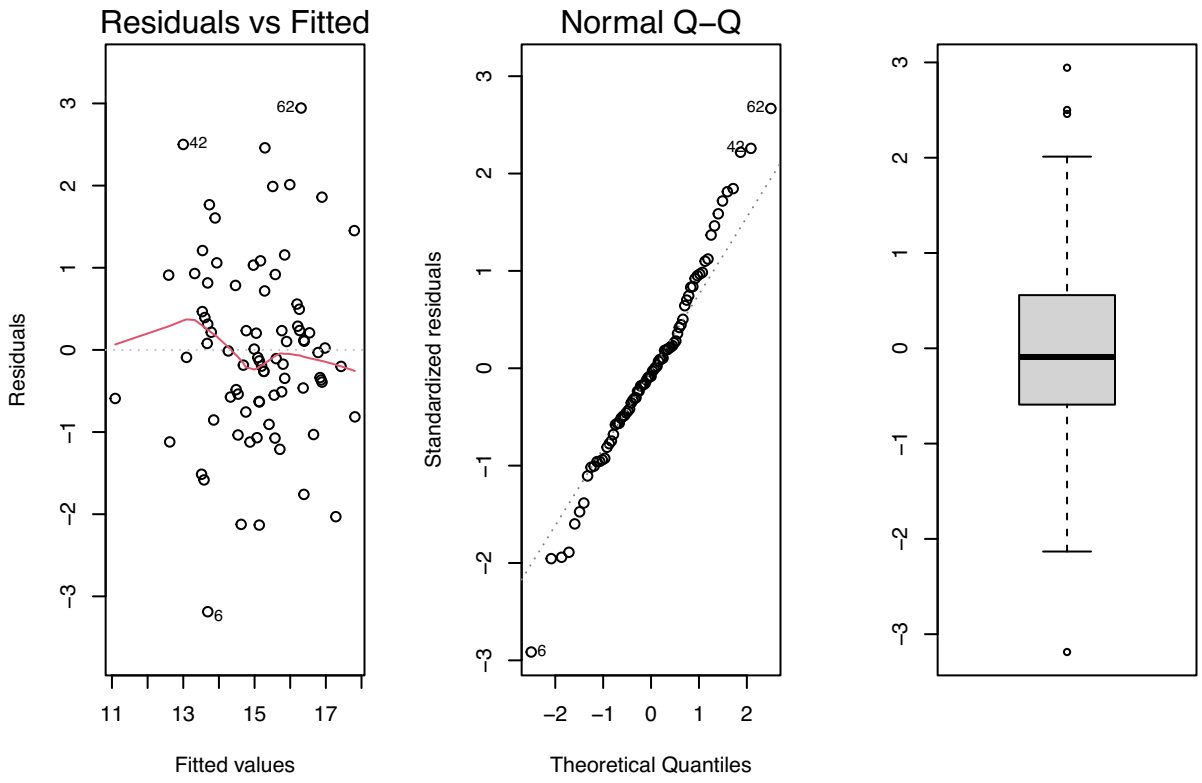
Fitted regression function:

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

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

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

```
par(mfrow=c(1,3))
#layout(matrix(c(1,2,3,3), nrow=2,byrow = F))
plot(fit1,which=c(1,2))
boxplot(fit1$residuals)
```

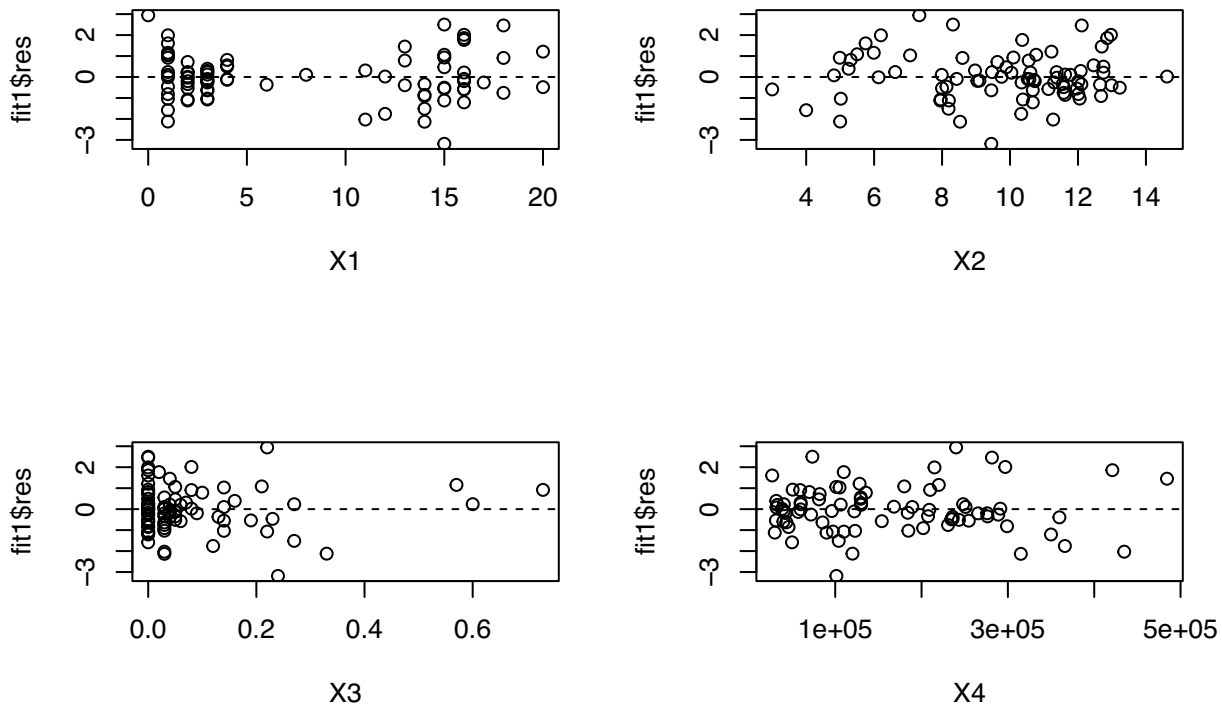


Residuals vs. fitted values plot shows no obvious nonlinearity. Residuals Q-Q plot shows slightly heavy tails. Residuals boxplot shows that most of the residuals are in between -2 and 2 and residual distribution is nearly symmetric.

(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.

Residuals vs. each predictor:

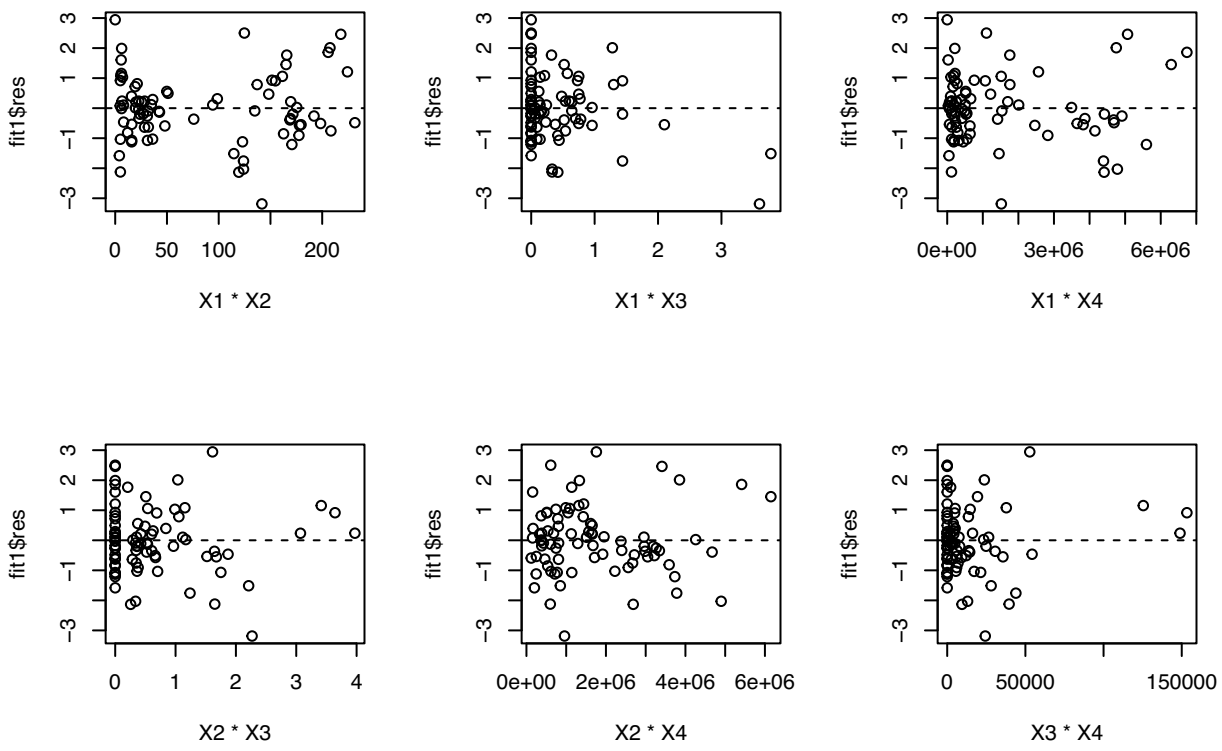
```
par(mfrow=c(2,2))
with(property,{
plot(X1,fit1$res)
abline(h=0,lty=2)
plot(X2,fit1$res)
abline(h=0,lty=2)
plot(X3,fit1$res)
abline(h=0,lty=2)
plot(X4,fit1$res)
abline(h=0,lty=2)
})
```



No obvious pattern.

Residuals vs. each two-way interaction (6 in total):

```
par(mfrow=c(2,3))
with(property,{
plot(X1*X2,fit1$res)
abline(h=0,lty=2)
plot(X1*X3,fit1$res)
abline(h=0,lty=2)
plot(X1*X4,fit1$res)
abline(h=0,lty=2)
plot(X2*X3,fit1$res)
abline(h=0,lty=2)
plot(X2*X4,fit1$res)
abline(h=0,lty=2)
plot(X3*X4,fit1$res)
abline(h=0,lty=2)
})
```



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?

We read the output of `summary(fit1)` in (c).

- $H_0 : \beta_0 = 0 \;\; vs. \;\; H_a : \beta_0 \neq 0$, $T^* = 21.11$, 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. 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 $SSTO = SSR + SSE = 236.558$.

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

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

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

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

Since $F^* = 26.75543 > 3.577 = F(0.99; 4, 76)$, 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 (f)).

```
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)$> 2.092 \times 10^{-2}$(Model 2); For $\hat{\beta}_2$, 6.317×10^{-2}(Model 1)$> 5.729 \times 10^{-2}$(Model 2); For $\hat{\beta}_4$, 1.385×10^{-6}(Model1)$> 1.305 \times 10^{-6}$(Model 2)

95% confidence interval under Model 2:

```
confint(fit2,parm=c('X1','X2','X4'),level=.95)
```

```
##             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.

(j) Consider a property with the following characteristics: $X_1 = 4$, $X2 = 10$, $X_3 = 0.1$, $X_4 = 80000$. Construct 99% prediction intervals under Model 1 and Model 2, respectively. Compare these two sets of intervals, what do you find?

99% prediction interval under Model 1:

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

Statistics 206

Homework 5 (Solution)

Due : Nov. 3, 2022, 11:59PM

Instructions:

- You should upload homeworkX files on canvas (under "Assignments/hwX") before its due date.
- Your homework may be prepared by a word processor (e.g., Latex) or through handwriting.
- For handwritten homework, you should either scan or take photos of your homework: Please make sure the pages are clearly numbered and are in order and the scans/photos are complete and clear; Check before submitting.
- Please name the files following the format: "FirstName-LastName-HwX". If there are several files, you can use "-Questions1-5", "-Questions6", etc., to distinguish them. E.g., "Jie-Peng-Hw1-Questions1-5.pdf", "Jie-Peng-Hw1-Questions6.rmd".
- Your name should be clearly shown on the submitted files: By putting on your name, you also acknowledge that you are the person who did and prepared the submitted homework.
- **Optional Problems** are more advanced and are not counted towards the grade.
- Showing/sharing/uploading homework or solutions outside of this class is prohibited.

1. 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.)

   Proof. By the hint, taking partial derivative of $Q(\cdot)$ w.r.t. b_0 we have

   $$-\sum_{i=1}^n (Y_i - b_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_{p-1} X_{i,p-1}) = 0$$

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

□

(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.)

Proof. By part (a), if we center all the variables, the only change to the least squares criterion function is adding a constant term and taking out a constant term which is equal to $\hat{\beta}_0$, but now $\hat{\beta}_0^*$ for the new setup of centered data is 0, i.e., nothing really changed in the criterion function. Hence the minimal value achieved by the least squares criterion function SSE remains unchanged. And also the $SSTO$ is unchanged, therefore the coefficient of multiple determination would remain the same as well. □

2. **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 $\mathbf{X}$.

ANS.
$$\begin{vmatrix} 1 & 0.36 & 2.14 & 0.7704 \\ 1 & 0.66 & 0.74 & 0.4884 \\ 1 & 0.66 & 1.91 & 1.2606 \\ 1 & -0.52 & -0.41 & 0.2132 \\ \cdots & \cdots & \cdots & \cdots \end{vmatrix}$$

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

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

(c) Derive the following sum of squares:

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

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

ANS. $SSR(X_1) = 58.232$
$SSE(X_1) = SSTO - SSR(X_1) = 32.986$
$SSR(X_2|X_1) = 5.490$
$SSR(X_2, X_1 \cdot X_2|X_1) = 5.490 + 0.448 = 5.938$
$SSR(X_1 \cdot X_2|X_1, X_2) = 0.448$
$SSR(X_1, X_2) = 58.232 + 5.490 = 63.722$
$SSE(X_1, X_2) = 27.048 + 0.448 = 27.496$

(d) 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.

ANS. The predicted value when $X_1 = 0, X_2 = 0$ is
$\hat{Y_h} = X_h'\hat{\boldsymbol{\beta}} = 0.9918 + 0 + 0 + 0 = 0.9918$ where

$$X_h' = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}, \quad \hat{\boldsymbol{\beta}}' = \begin{bmatrix} 0.9918 & 1.5424 & 0.5799 & -0.1491 \end{bmatrix}.$$

$s(pred) = \sqrt{MSE\left[1 + X_h'(X'X)^{-1}X_h\right]} = 1.02 \times \sqrt{(1 + 0.087)} = 1.063.$
A 95% confidence prediction interval is $0.9918 \pm t(0.975; 30 - 4) \times 1.063 = 0.9918 \pm 2.056 \times (1.063) = (-1.19, 3.18)$.

(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.

ANS. The full model: $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + \epsilon$
The reduced model: $Y = \beta_0 + \beta_1X_1 + \epsilon$
$H_0 : \beta_2 = \beta_3 = 0$, $H_A : \beta_2 \neq 0$ or $\beta_3 \neq 0$.
Test statistic:

$$F^* = \frac{\frac{SSE(reduced) - SSE(full)}{df(full) - df(reduced)}}{\frac{SSE(full)}{df(full)}} = \frac{SSR(X_2, X_1 \cdot X_2|X_1)/2}{MSE(full)}$$

Under H_0, the test statistic should follow the $F_{2,26}$ distribution. We reject H_0 if $F^* > F_{2,26}(0.99) = 5.526$.

$$F^*_{obs} = \frac{5.938/2}{1.040} = 2.855 < 5.526$$

. Therefore, we do not reject H_0. We can drop both X_2 and X_1X_2 from the model.

```
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


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
```

3. 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:

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

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

ANS. 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 $F^* > F(1-\alpha, df(reduced) - df(full), df(full))$

4. **(Optional Problem). Sampling distributions of SS.** 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 of projection matrix and the fact that $(\mathbf{H}_n - \frac{1}{n}\mathbf{J}_n)\mathbf{1}_n = \mathbf{0}$*)*

ANS. When $\beta_1 = \cdots = \beta_{p-1} = 0$, $\boldsymbol{X\beta} = \beta_0\mathbf{1}_n$.

$$\begin{aligned} SSR &= Y^{'}(H - \frac{1}{n}J_n)Y \\ &= d'd \end{aligned}$$

Where $d = (H - \frac{1}{n}J_n)Y = (H - \frac{1}{n}J_n)(\beta_0 1_n + \epsilon) = (H - \frac{1}{n}J_n)\epsilon$, since $(\boldsymbol{H} - \frac{1}{n}\boldsymbol{J}_n)\mathbf{1}_n = \mathbf{0}_n$. Thus,

$$SSR = \epsilon'(H - \frac{1}{n}J_n)\epsilon$$

Let $z = Q\epsilon$, then

$$SSR = \epsilon'\mathbf{Q}'\mathbf{\Lambda}\mathbf{Q}\epsilon = z'\mathbf{\Lambda}z = \sum_1^{p-1} z_i^2$$

$$E(z) = E(Q\epsilon) = 0_n, Var(z) = var(Q\epsilon) = Q'var(\epsilon)Q = \sigma^2 I_n$$

Under normal error model, z_i are $iid\ N(0, \sigma^2)$. Thus, $SSR \sim \sigma^2\chi^2_{p-1}$

Similar for SSTO.

$$\begin{aligned} SSTO &= \boldsymbol{Y}'(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}_n)\boldsymbol{Y} \\ &= (\beta_0 1_n + \epsilon)'(I_n - \frac{1}{n}J_n)(\beta_0 1_n + \epsilon) \\ &= \epsilon'(I_n - \frac{1}{n}J_n)\epsilon \\ &= \epsilon'\mathbf{Q}'\mathbf{\Lambda}\mathbf{Q}\epsilon \\ &= (Q\epsilon)'\mathbf{\Lambda}(Q\epsilon) \\ &= \sum_1^{n-1} z_i^2 \end{aligned}$$

Under normal error model, z_i are iid $N(0, \sigma^2)$. Thus, $SSTO \sim \sigma^2 \chi^2_{n-1}$

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

   Proof.

$$\begin{aligned} E(\mathbf{y}'\mathbf{A}\mathbf{y}) &= E(tr(\mathbf{y}'\mathbf{A}\mathbf{y})) \\ &= E(tr(\mathbf{A}\mathbf{y}\mathbf{y}')) \\ &= tr(E(\mathbf{A}\mathbf{y}\mathbf{y}')) \\ &= tr(\mathbf{A}E(\mathbf{y}\mathbf{y}')) \\ &= tr(\mathbf{A}\boldsymbol{\Sigma} + \mathbf{A}\boldsymbol{\mu}\boldsymbol{\mu}') \\ &= tr(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} \end{aligned}$$

   □

   (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 $E(X^4) = 3\sigma^4$.)

   Proof. We can represent $\mathbf{y}$ as $\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, then

$$\begin{aligned} \mathbf{y}'\mathbf{A}\mathbf{y} &= (\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{z})'\mathbf{A}(\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{z}) \\ &= (\boldsymbol{\mu}' + \mathbf{z}'\boldsymbol{\Sigma}^{1/2})\mathbf{A}(\boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{z}) \\ &= \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} + 2\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}^{1/2}\mathbf{z} + \mathbf{z}'\boldsymbol{\Sigma}^{1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2}\mathbf{z}. \end{aligned}$$

   Thus,

$$\begin{aligned} &Cov(\mathbf{y}'\mathbf{A}\mathbf{y}, \mathbf{y}'\mathbf{B}\mathbf{y}) \\ &= Cov(2\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}^{1/2}\mathbf{z} + \mathbf{z}'\boldsymbol{\Sigma}^{1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2}\mathbf{z}, 2\boldsymbol{\mu}'\mathbf{B}\boldsymbol{\Sigma}^{1/2}\mathbf{z} + \mathbf{z}'\boldsymbol{\Sigma}^{1/2}\mathbf{B}\boldsymbol{\Sigma}^{1/2}\mathbf{z}) \\ &= 4Cov(\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}^{1/2}\mathbf{z}, \boldsymbol{\mu}'\mathbf{B}\boldsymbol{\Sigma}^{1/2}\mathbf{z}) + Cov(2\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}^{1/2}\mathbf{z}, \mathbf{z}'\boldsymbol{\Sigma}^{1/2}\mathbf{B}\boldsymbol{\Sigma}^{1/2}\mathbf{z}) + \\ &Cov(\mathbf{z}'\boldsymbol{\Sigma}^{1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2}\mathbf{z}, 2\boldsymbol{\mu}'\mathbf{B}\boldsymbol{\Sigma}^{1/2}\mathbf{z}) + Cov(\mathbf{z}'\boldsymbol{\Sigma}^{1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2}\mathbf{z}, \mathbf{z}'\boldsymbol{\Sigma}^{1/2}\mathbf{B}\boldsymbol{\Sigma}^{1/2}\mathbf{z})) \end{aligned}$$

   Since $E(z_i^3) = 0$, we have

$$Cov(2\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}^{1/2}\mathbf{z}, \mathbf{z}'\boldsymbol{\Sigma}^{1/2}\mathbf{B}\boldsymbol{\Sigma}^{1/2}\mathbf{z}) = Cov(\mathbf{z}'\boldsymbol{\Sigma}^{1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2}\mathbf{z}, 2\boldsymbol{\mu}'\mathbf{B}\boldsymbol{\Sigma}^{1/2}\mathbf{z}) = 0.$$

Besides, we have

$$4Cov(\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}^{1/2}\mathbf{z}, \boldsymbol{\mu}'\mathbf{B}\boldsymbol{\Sigma}^{1/2}\mathbf{z}) = 4\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}\boldsymbol{\mu}.$$

Let $\mathbf{z}'\boldsymbol{\Sigma}^{1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2}\mathbf{z} = \sum_{i,j} a_{i,j}z_iz_j$, and $\mathbf{z}'\boldsymbol{\Sigma}^{1/2}\mathbf{B}\boldsymbol{\Sigma}^{1/2}\mathbf{z} = \sum_{k,l} b_{k,l}z_kz_l$, then

$$\begin{aligned}
&Cov(\mathbf{z}'\boldsymbol{\Sigma}^{1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2}\mathbf{z}, \mathbf{z}'\boldsymbol{\Sigma}^{1/2}\mathbf{B}\boldsymbol{\Sigma}^{1/2}\mathbf{z}) = \sum_{i,j,k,l} a_{i,j}b_{k,l}Cov(z_iz_j, z_kz_l) \\
&= 2\sum_i a_{i,i}b_{i,i} + \sum_{i\neq j} a_{i,j}b_{i,j} + \sum_{i\neq j} a_{i,j}b_{j,i} = 2tr(\boldsymbol{\Sigma}^{1/2}\mathbf{A}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}\mathbf{B}\boldsymbol{\Sigma}^{1/2}) \\
&= 2tr(\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}\boldsymbol{\Sigma}).
\end{aligned}$$

□

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

Proof.

$$\begin{aligned}
E(SSE) &= E(Y'(\mathbf{I}-\mathbf{H})Y) \\
&= tr(\mathbf{I}-\mathbf{H})\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I}-\mathbf{H})\mathbf{X}\boldsymbol{\beta} \\
&= (n-p)\sigma^2
\end{aligned}$$

$$\begin{aligned}
E(SSR) &= E(Y'(\mathbf{H}-\frac{1}{n}\mathbf{J})Y) \\
&= tr(\mathbf{H}-\frac{1}{n}\mathbf{J})\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{H}-\frac{1}{n}\mathbf{J})\mathbf{X}\boldsymbol{\beta} \\
&= (p-1)\sigma^2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\frac{1}{n}\mathbf{J}\mathbf{X}\boldsymbol{\beta}
\end{aligned}$$

□

6. **A multiple linear regression case study by R**. You need to submit your codes alongside with the answers, plots, outputs, etc. You are required to use R Markdown: Please submit a .rmd file **and** its corresponding .html file.

   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. (data file: property.txt; 1st column – Y, followed by X_1, X_2, X_3, X_4)

ANS. See solution to Question 6.