Statistics 206

Homework 6 (Solution)

Due : Nov. 11, 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. Tell true or false of the following statements and briefly explain your answer.

   (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.

ANS. **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.

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

   (c) Taking transformations of the X variables as in the standardized regression model (referred to as correlation transformation) will not change coefficients of multiple determination.

ANS. **TRUE**. The sum of squares, i.e., SSE, $SSTO$ and SSR, won't change, since the fitted values remain the same.

(d) 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.

ANS. **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.

(e) 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.

ANS. **TRUE**. Suppose there are are some 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 even through some factors are statistically significant when tested individually.

(f) 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.

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

(g) 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.

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

$$r_{XX} = \begin{bmatrix} 1 & 0 \\ 0 & r_{\tilde{X}\tilde{X}} \end{bmatrix} \quad r_{XY} = \begin{bmatrix} 0 \\ r_{\tilde{X}Y} \end{bmatrix}$$

The fitted standardized regression coefficients:

$$\hat{\beta}^* = r_{XX}^{-1} r_{XY} = \begin{bmatrix} 1 & 0 \\ 0 & r_{\tilde{X}\tilde{X}}^{-1} \end{bmatrix} \begin{bmatrix} 0 \\ r_{\tilde{X}Y} \end{bmatrix} = \begin{bmatrix} 0 \\ r_{\tilde{X}\tilde{X}}^{-1} r_{\tilde{X}Y} \end{bmatrix}.$$

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

(h) If the coefficient of multiple determination of regressing an X variable to the rest of the X variables is large, then its least-squares estimated regression coefficient tends to have large sampling variability.

ANS. TRUE See the conclusion of Question 4

$$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum_i (X_{ij} - \overline{X}_j)^2} \frac{1}{1 - R_j^2}$$

2. **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 $X_{(-j)} := \{X_k : 1 \le k \le p-1, k \ne j\}$. Show that

$$SSR(X_j | X_{(-j)}) = SSR(X_j),$$

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

Proof. We have,

$$\begin{aligned}
SSE(X^*_{(-j)}) - SSE(X^*_{(-j)}, X^*_j) &= Y^T(I - H(X^*_{(-j)}))Y \\
&\quad - Y^T(I - H(X^*_{(-j)}, X^*_j))Y \\
&= Y^T(H(X^*_{(-j)}, X^*_j) - H(X^*_{(-j)}))Y \\
&= Y^T(n^{-1}11^T + X^*_{(-j)}X^{*T}_{(-j)} + X^*_j X^{*T}_j \\
&\qquad\qquad - n^{-1}11^T - X^*_{(-j)}X^{*T}_{(-j)})Y \\
&= Y^T X^*_j X^{*T}_j Y
\end{aligned}$$

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

Thus,

$$\begin{aligned}
LHS &= SSE(X_{(-j)}) - SSE(X_{(-j)}, X_j) \\
&= SSE(X^*_{(-j)}) - SSE(X^*_{(-j)}, X^*_j) \\
&= SSR(X^*_j) \\
&= SSTO - SSE(X^*_j) \\
&= SSTO - SSE(X_j) \\
&= RHS
\end{aligned}$$

□

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

4. **(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

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

□

For the following question, **See the separate pdf file generated by RMarkdown**.

5. **(Commercial Property Cont'd) Partial coefficients and added-variable plots.** 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)

   (a) Perform regression of the rental rates Y on the four predictors X_1, X_2, X_3, X_4 (Model 1). *Hint: To help answer the subsequent questions, the predictors should enter the model in the order X_1, X_2, X_4, X_3.*

   (b) Based on the R output of Model 1, obtain the fitted regression coefficient of X_3 and calculate the coefficient of partial determination $R^2_{Y3|124}$ and partial correlation $r_{Y3|124}$. Explain what $R^2_{Y3|124}$ measures and interpret the result.

   (c) Draw the added-variable plot for X_3 and make comments based on this plot.

   (d) Regressing the residuals $e(Y|X_1, X_2, X_4)$ to the residuals $e(X_3|X_1, X_2, X_4)$. Compare the fitted regression slope from this regression with the fitted regression coefficient of X_3 from part (b). What do you find?

   (e) Obtain the regression sum of squares from part (d) and compare it with the extra sum of squares $SSR(X_3|X_1, X_2, X_4)$ from the R output of Model 1. What do you find?

   (f) Calculate the correlation coefficient r between the two sets of residuals $e(Y|X_1, X_2, X_4)$ and $e(X_3|X_1, X_2, X_4)$. Compare it with $r_{Y3|124}$. What do you find? What is r^2?

   (g) Regressing Y to the residuals $e(X_3|X_1, X_2, X_4)$. Compare the fitted regression slope from this regression with the fitted regression coefficient of X_3 from part (b). What do you find? Can you provide an explanation?

6. **(Commercial Property Cont'd). Standardized Regression model.** 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) 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) Obtain SSTO, SSE and SSR under the standardized model and compare them with those from the original model. What do you find?

(d) 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?

7. **(Commercial Property Cont'd). Multicollinearity.** 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) Obtain $\mathbf{r}_{XX}^{-1}$ and get the variance inflation 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 $SSR(X_4)$ 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 $SSR(X_2)$ and $SSR(X_2|X_4)$. What do you find? Provide an interpretation for your observations.

8. **(Commercial Property Cont'd) Polynomial Regression.** 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)

   Based on the analysis from Homework 5, the vacancy rate (X_3) is not important in explaining the rental rates (Y) when age (X_1), operating expenses (X_2) and square footage (X_4) are included in the model. So here we will use the latter three variables to build a regression for rental rates.

   (a) Plot rental rates (Y) against the age of property (X_1) and comment on the shape of their relationship.

(b) Fit a polynomial regression model with linear terms for centered age of property ($\tilde{X}_1$), operating expenses (X_2), and square footage (X_4), and a quadratic term for centered age of property ($\tilde{X}_1$). Write down the model equation. Obtain the fitted regression function and also express it in terms of the original age of property X_1. Draw the observations Y against the fitted values $\widehat{Y}$ plot. Does the model provide a good fit?

(c) Compare R^2, R_a^2 of the above model with those of Model 2 from Homework 5 ($Y \sim X_1 + X_2 + X_4$). What do you find?

(d) Test whether or not the quadratic term for centered age of property ($\tilde{X}_1$) may be dropped from the model at level 0.05. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion.

(e) Predict the rental rates for a property with $X_1 = 4, X_2 = 10, X_4 = 80,000$. Construct a 99% prediction interval and compare it with the prediction interval from Model 2 of Homework 5.

HW6 Question 5, 6, 7, and 8

Wookyeong Song (mostly from Yan-Yu Chen)

2022/11/11

5. (Commercial Property Cont'd) Partial coefficients and added-variable plots.

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)

(a) Perform regression of the rental rates Y on the four predictors X_1, X_2, X_3, X_4 (Model 1). *Hint: To help answer the subsequent questions, the predictors should enter the model in the order X_1, X_2, X_4, X_3.*

```
summary(fit <- lm(Y ~ X1 + X2 + X4 + X3, data = property))
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X4 + X3, 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 ***
## X4           7.924e-06  1.385e-06   5.722 1.98e-07 ***
## X3           6.193e-01  1.087e+00   0.570     0.57
## ---
## 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
```

(b) Based on the R output of Model 1, obtain the fitted regression coefficient of X_3 and calculate the coefficient of partial determination $R^2_{Y3|124}$ and partial correlation $r_{Y3|124}$. Explain what $R^2_{Y3|124}$ measures and interpret the result.

$\hat{\beta}_3 = 0.619$ Note that

$$R^2_{Y3|124} = \frac{SSR(X_3|X_1, X_2, X_4)}{SSE(X_1, X_2, X_4)} = \frac{SSR(X_3|X_1, X_2, X_4)}{SSR(X_3|X_1, X_2, X_4) + SSE(X_1, X_2, X_3, X_4)}$$

```
anova.fit<-anova(fit)
anova.fit
```

```
## 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 ***
## X4         1 50.287  50.287 38.9062 2.306e-08 ***
## X3         1  0.420   0.420  0.3248  0.570446
## Residuals 76 98.231   1.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
R_Y3<-anova.fit[4,2]/(anova.fit[4,2]+anova.fit[5,2])
r_Y3<-sqrt(R_Y3) # beta_3 is positive
R_Y3
```

```
## [1] 0.004254889
```

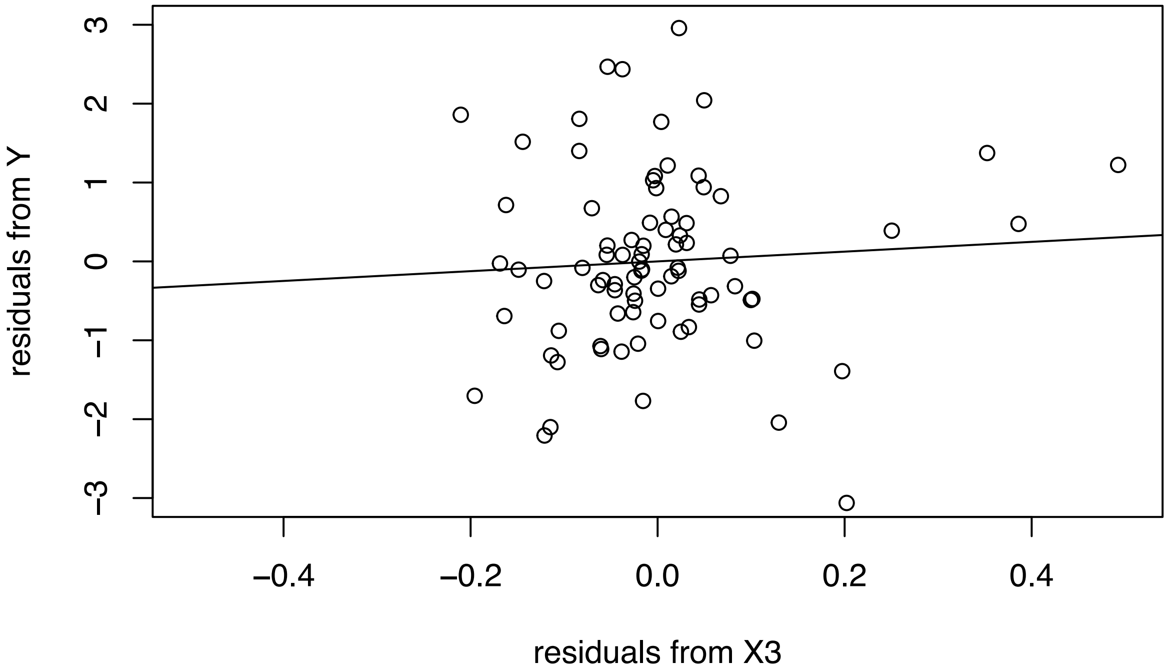
```
r_Y3
```

```
## [1] 0.06522951
```

$R^2_{Y3|124}$ measures the marginal contribution in proportional reduction in SSE by adding X_3 into the model containing X_1, X_2, X_4, and SSE is reduced by 0.43% when X_3 is added to the model.

(c) Draw the added-variable plot for X_3 and make comments based on this plot.

```
fit_Y_X1X2X4 <- lm(Y~X1+X2+X4, data=property)
fit_X3_X1X2X4 <- lm(X3~X1+X2+X4, data=property)
residuals_YX3 <- data.frame(cbind(fit_Y_X1X2X4$residuals,fit_X3_X1X2X4$residuals))
names(residuals_YX3) <- c("Y", "X3")
fit_YX3_X1X2X4 <- lm(Y~X3, data=residuals_YX3)
plot(residuals_YX3[,2], residuals_YX3[,1], xlab="residuals from X3",
     ylab="residuals from Y", xlim=c(-0.5, 0.5), ylim=c(-3, 3))
abline(fit_YX3_X1X2X4)
```



There is no obvious linear relation between $e(Y|X_1, X_2, X_4)$ and $e(X_3|X_1, X_2, X_4)$, and the points seem to concentrate around the origin. So we may conclude that X_3 doesn't add much explaining ability to the model of X_1, X_2, X_4.

(d) Regressing the residuals $e(Y|X_1, X_2, X_4)$ to the residuals $e(X_3|X_1, X_2, X_4)$. Compare the fitted regression slope from this regression with the fitted regression coefficient of X_3 from part (b). What do you find?

```
fit$coef[5]
```

```
##        X3
## 0.6193435
```

```
fit_YX3_X1X2X4$coef[2]
```

```
##        X3
## 0.6193435
```

The fitted regression slope from this regression and the fitted regression coefficient of X_3 from part (b) are the same.

(e) Obtain the regression sum of squares from part (d) and compare it with the extra sum of squares $SSR(X_3|X_1, X_2, X_4)$ from the R output of Model 1. What do you find?

```
anova(fit_YX3_X1X2X4)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value Pr(>F)
## X3         1  0.420 0.41975  0.3376 0.5629
## Residuals 79 98.231 1.24343
```

```
anova(fit_YX3_X1X2X4)[1,2]
```

```
## [1] 0.4197463
```

```
anova(fit)[4,2]
```

```
## [1] 0.4197463
```

The regression sum of squares from part (d) and the extra sum of squares $SSR(X_3|X_1, X_2, X_4)$ from the R output of Model 1 are the same.

(f) Calculate the correlation coefficient r between the two sets of residuals $e(Y|X_1, X_2, X_4)$ and $e(X_3|X_1, X_2, X_4)$. Compare it with $r_{Y3|124}$. What do you find? What is r^2?

```
cor(residuals_YX3$Y, residuals_YX3$X3)
```

```
## [1] 0.06522951
```

The correlation coeffcient r between the two sets of residuals $e(Y|X_1, X_2, X_4)$ and $e(X_3|X_1, X_2, X_4)$ and $r_{Y3|124}$ are equal. r^2 is the coefficient of simple determination, i.e., the R^2 of the simple linear regression.

(g) Regressing Y to the residuals $e(X_3|X_1, X_2, X_4)$. Compare the fitted regression slope from this regression with the fitted regression coefficient of X_3 from part (b). What do you find? Can you provide an explanation?

```
Y_residualsX3 <- data.frame(cbind(property[,1], summary(fit_X3_X1X2X4)$residuals))
names(Y_residualsX3) <- c("Y", "X3")
fit_Y_residualsX3 <- lm(Y~X3, data=Y_residualsX3)
summary(fit_Y_residualsX3)
```

```
##
## Call:
## lm(formula = Y ~ X3, data = Y_residualsX3)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7641 -1.1392 -0.1056  1.1221  4.1630
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.1389     0.1921  78.807   <2e-16 ***
## X3            0.6193     1.6528   0.375    0.709
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.729 on 79 degrees of freedom
## Multiple R-squared:  0.001774,   Adjusted R-squared:  -0.01086
## F-statistic: 0.1404 on 1 and 79 DF,  p-value: 0.7089
```

```r
fit_Y_residualsX3$coef[2]
```

```
##        X3
## 0.6193435
```

The fitted regression slope from this is same as that from part(b).

Denote P_{124} the projection onto the subspace of X_1, X_2, and X_4. Then $(X_3^T(I-P_{124})X_3)^{-1}X_3^T(I-P_{124})Y$ is the same as $(X_3^T(I-P_{124})X_3)^{-1}X_3^T(I-P_{124})(I-P_{124})Y$. That is, it does not matter whether we project Y onto the subspace that orthogonal to (X_1, X_2, X_4) first. Therefore, to regress Y directly on $e(X_3|X_1, X_2, X_4)$ is same as to regress $e(Y|X_1, X_2, X_4)$ on $e(X_3|X_1, X_2, X_4)$.

6. (Commercial Property Cont'd). Standardized Regression model.

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) 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?

```r
colMeans(property)
```

```
##            Y           X1           X2           X3           X4
## 1.513889e+01 7.864198e+00 9.688148e+00 8.098765e-02 1.606333e+05
```

```r
apply(property,2, sd)
```

```
##            Y           X1           X2           X3           X4
## 1.719584e+00 6.632784e+00 2.583169e+00 1.345512e-01 1.090990e+05
```

```r
X_star<-1/sqrt(nrow(property)-1)*apply(property[,-1], 2,
                                FUN=function(x){(x-mean(x))/sd(x)})
```

Check: 1. Sample mean of the transformed variables is zero. 2. Sample sd of the transformed variables is $\frac{1}{\sqrt{n-1}}$.

```
new_sd <- 1/sqrt(nrow(property)-1)
new_sd
```

```
## [1] 0.1118034
```

```
colMeans(X_star)
```

```
##            X1            X2            X3            X4
## -5.680684e-18  7.506427e-18 -6.384746e-18  1.250714e-17
```

```
apply(X_star,2, sd)
```

```
##        X1        X2        X3        X4
## 0.1118034 0.1118034 0.1118034 0.1118034
```

(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?

The standardized model equation is $Y_i = \beta_0^* + \beta_1^* X_{i1}^* + \beta_2^* X_{i2}^* + \beta_3^* X_{i3}^* + \beta_4^* X_{i4}^*$, $i = 1, 2, 3, \ldots, 81$. Then we fit the standardized model and present the regression results

```
property_star <- data.frame(cbind(property[,1], X_star))
names(property_star) <- c('Y','X1','X2','X3','X4')
fit_star <- lm(Y~X1+X2+X3+X4, data=property_star)
summary(fit_star)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = property_star)
##
## 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)  15.1389     0.1263 119.845  < 2e-16 ***
## X1           -8.4262     1.2662  -6.655 3.89e-09 ***
## X2            6.5159     1.4596   4.464 2.75e-05 ***
## X3            0.7454     1.3079   0.570     0.57
## X4            7.7326     1.3513   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
```

The fitted model is $Y = 15.1389 - 8.4262X_1^* + 6.5159X_2^* + 0.7454X_3^* + 7.7326X_4^*$. The fitted regression intercept is 15.1389.

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

The original model:

```
anova(lm(Y~X1+X2+X3+X4 ,data=property)) # original model
```

```
## 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
```

```
anova(fit_star) # standardized model
```

```
## 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
```

Note that they are exactly the same.

```
sum(anova.fit[,2]) # SSTO
```

```
## [1] 236.5575
```

```
anova.fit[5,2] # SSE
```

```
## [1] 98.23059
```

```
sum(anova.fit[,2])-anova.fit[5,2] # SSR
```

```
## [1] 138.3269
```

(d) 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?

From the `R` output in 5(a) and 6(b), R^2, R_a^2 under the original model are $0.58447, 0.5629$ respectively. R^2, R_a^2 under the standardized model are also $0.58447, 0.5629$ respectively.

7. (Commercial Property Cont'd). Multicollinearity

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) Obtain $\mathbf{r}_{XX}^{-1}$ and get the variance inflation 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.

```
r_inverse<-solve(t(X_star)%*% X_star)
diag(r_inverse)
```

```
##       X1       X2       X3       X4
## 1.240348 1.648225 1.323552 1.412722
```

```
r_sq_1 <- summary(lm(formula=X1~X2+X3+X4, data=property))$r.squared
r_sq_2 <- summary(lm(formula=X2~X1+X3+X4, data=property))$r.squared
r_sq_3 <- summary(lm(formula=X3~X1+X2+X4, data=property))$r.squared
r_sq_4 <- summary(lm(formula=X4~X1+X2+X3, data=property))$r.squared
1/(1-c(r_sq_1,r_sq_2,r_sq_3,r_sq_4))
```

```
## [1] 1.240348 1.648225 1.323552 1.412722
```

The same results from two methods confirm $VIF_k = \frac{1}{1-R_k^2}$, $k = 1, 2, 3, 4$. All four VIF values are a little bit higher than 1 and far less than 10, so we can conclude that there is not much multicollinearity in the model.

(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 $SSR(X_4)$ and $SSR(X_4|X_3)$. What do you find? Provide an interpretation for your observations.

```
fit_X4 <- lm(Y~X4, data=property)
fit_X4$coef[2]
```

```
##           X4
## 8.436639e-06
```

```
fit_X3X4 <- lm(Y~X3+X4, data=property)
fit_X3X4$coef[3]
```

```
##           X4
## 8.406741e-06
```

The estimated regression coefficients of X4 in these two models are almost the same.

```r
anova(fit_X4)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X4         1  67.775  67.775  31.723 2.628e-07 ***
## Residuals 79 168.782   2.136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(fit_X3X4)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X3         1   1.047   1.047  0.4842    0.4886
## X4         1  66.858  66.858 30.9213 3.626e-07 ***
## Residuals 78 168.652   2.162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(fit_X4)[1,2] # SSR(X4)
```

```
## [1] 67.7751
```

```r
anova(fit_X3X4)[2,2] # SSR(X4|X3)
```

```
## [1] 66.85829
```

$SSR(X_4) = 67.7751$ and $SSR(X_4|X_3) = 66.8583$, they are quite similar. This is expected, since the correlation matrix shows that there is almost no correlation between X_3 and X_4, the marginal effect of adding X_4 into the model which already has X_3 is very closed to the explaining ability of X_4 alone.

(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 $SSR(X_2)$ and $SSR(X_2|X_4)$. What do you find? Provide an interpretation for your observations.

```r
fit_X2 <- lm(Y~X2, data=property)
fit_X2$coef[2]
```

```
##        X2
## 0.2754531
```

```r
fit_X4X2 <- lm(Y~X4+X2, data=property)
fit_X4X2$coef[3]
```

```
##        X2
## 0.1469682
```

Two estimated regression coefficients of X_2 are quite different.

```r
anova(fit_X2)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X2         1  40.503  40.503  16.321 0.0001231 ***
## Residuals 79 196.054   2.482
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(fit_X4X2)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X4         1  67.775  67.775 33.1457 1.611e-07 ***
## X2         1   9.291   9.291  4.5438   0.03619 *
## Residuals 78 159.491   2.045
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(fit_X2)[1,2] # SSR(X2)
```

```
## [1] 40.50333
```

```r
anova(fit_X4X2)[2,2] # SSR(X2|X4)
```

```
## [1] 9.290987
```

$SSR(X_2) = 40.5033 > SSR(X_2|X_4) = 9.2910$. The correlation matrix shows that there X_2 and X_4 are moderately correlated, so the marginal effect of adding X_2 into the model which already has X_4 is expected to less effective compared to the explaining ability of X_2 alone.

8. (Commercial Property Cont'd) Polynomial Regression.

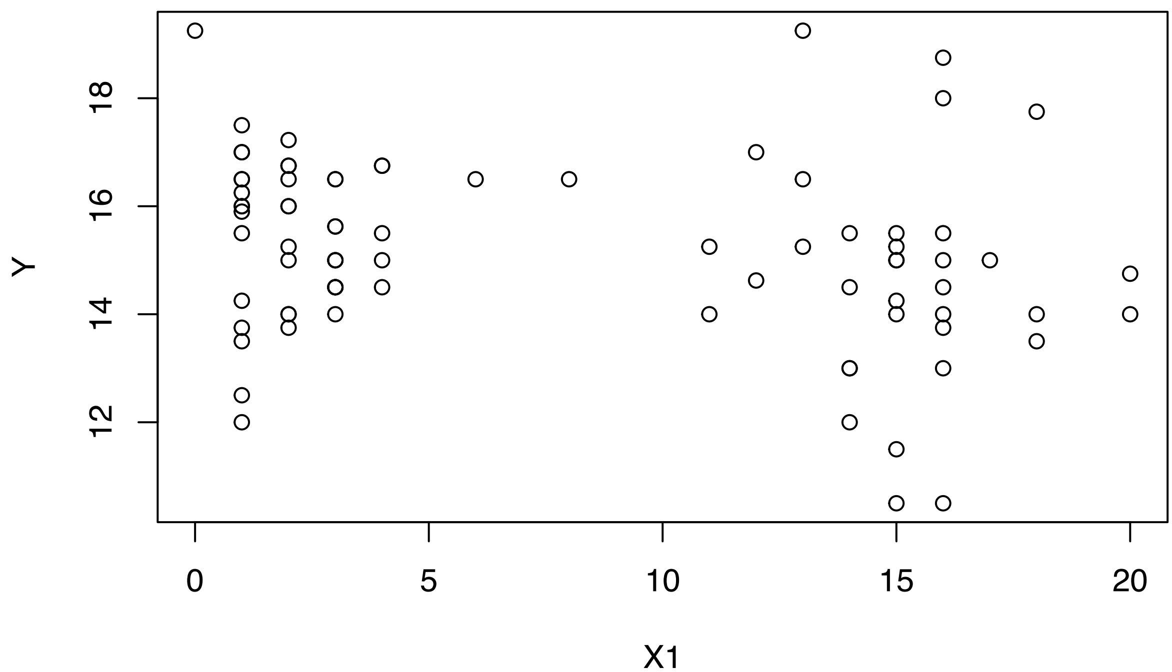
You need to submit your codes alongside with the answers, plots, outputs, etc. You are required to use R Markdown: Please submit a .rmd file **and** its corresponding .html file.

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)

Based on the analysis from Homework 5, the vacancy rate (X_3) is not important in explaining the rental rates (Y) when age (X_1), operating expenses (X_2) and square footage (X_4) are included in the model. So here we will use the latter three variables to build a regression for rental rates.

(a) Plot rental rates (Y) against the age of property (X_1) and comment on the shape of their relationship.

```
property<-read.table("property.txt",header = FALSE,
                     col.names = c("Y", paste('X', 1:4, sep='')))
with(property, plot(X1, Y))
```



The age of property (X_1) exhibits some curvilinear relation when plotted against the rental rates (Y).

(b) Fit a polynomial regression model with linear terms for centered age of property ($\tilde{X}_1$), operating expenses (X_2), and square footage (X_4), and a quadratic term for centered age of property ($\tilde{X}_1$). Write down the model equation. Obtain the fitted regression function and also express it in terms of the original age of property X_1. Draw the observations Y against the fitted values $\widehat{Y}$ plot. Does the model provide a good fit?

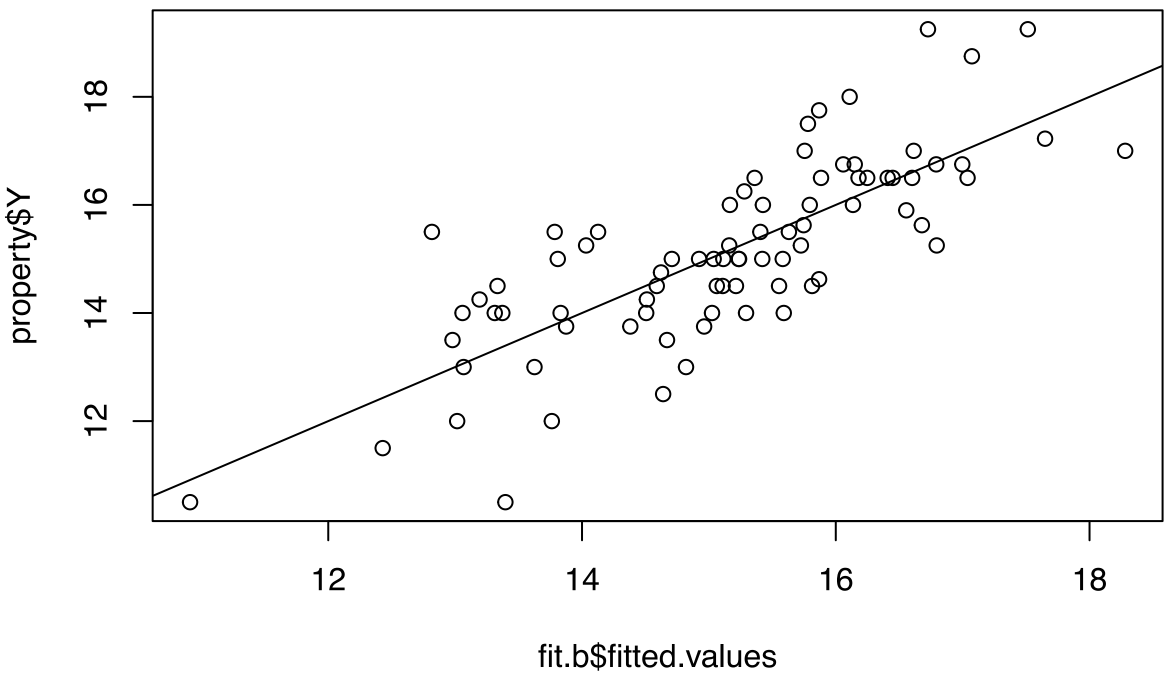
Model equation:

$$Y_i = \beta_0 + \beta_1 \tilde{X}_{i1} + \beta_2 X_{i2} + \beta_3 X_{i4} + \beta_4 \tilde{X}_{i1}^2 + \epsilon_i, \quad i = 1, \cdots, 81$$

```r
property.c<-property[,-4]
property.c$X1<-property$X1-mean(property$X1)
fit.b<-lm(Y~X1+X2+X4+I(X1^2),
          data = property.c)
summary(fit.b)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X4 + I(X1^2), data = property.c)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89596 -0.62547 -0.08907  0.62793  2.68309
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.019e+01  6.709e-01  15.188  < 2e-16 ***
## X1          -1.818e-01  2.551e-02  -7.125 5.10e-10 ***
## X2           3.140e-01  5.880e-02   5.340 9.33e-07 ***
## X4           8.046e-06  1.267e-06   6.351 1.42e-08 ***
## I(X1^2)      1.415e-02  5.821e-03   2.431   0.0174 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.097 on 76 degrees of freedom
## Multiple R-squared:  0.6131, Adjusted R-squared:  0.5927
## F-statistic:  30.1 on 4 and 76 DF,  p-value: 5.203e-15
```

```r
plot(fit.b$fitted.values, property$Y)
abline(a=0,b=1)
```



Fitted regression function:

$$\begin{aligned}\widehat{Y} &= 10.19 - 0.1818\tilde{X}_1 + 0.314X_2 + 8.046 \times 10^{-6}X_4 + 0.01415\tilde{X}_1^2 \\ &= 10.19 - 0.1818(X_1 - 7.8642) + 0.314X_2 + 8.046 \times 10^{-6}X_4 + 0.01415(X_1 - 7.8642)^2.\end{aligned}$$

The model provides a fairly good fit.

(c) Compare R^2, R_a^2 of the above model with those of Model 2 from Homework 5 ($Y \sim X_1 + X_2 + X_4$). What do you find?

For the above model: $R^2 = 0.6131$, $R_a^2 = 0.5927.\backslash$ For Model 2 from homework 5: $R^2 = 0.583$, $R_a^2 = 0.5667$. So the model here has a better fit of the data than Model 2 of homework 5.

(d) Test whether or not the quadratic term for centered age of property ($\tilde{X}_1$) may be dropped from the model at level 0.05. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and the conclusion.

Null and alternative hypotheses:

$$H_0 : \beta_4 = 0, \ vs. \ H_a : \beta_4 \neq 0$$

Test statistic:

$$T^* = \frac{\hat{\beta}_4}{se(\hat{\beta}_4)} = 2.431$$

Under H_0, $T^* \sim t_{76}$, Since $2.431 > 1.99 = t(0.975; 76)$(or p-value=$0.0174 < 0.05$), reject H_0 and conclude that the quadratic term for centered age of property can not be dropped.

(e) Predict the rental rates for a property with $X_1 = 4, X_2 = 10, X_4 = 80,000$. Construct a 99% prediction interval and compare it with the prediction interval from Model 2 of Homework 5.

```
newX<-data.frame(X1=4-mean(property$X1),X2=10,X4=80000)
predict.lm(fit.b, newX, interval="prediction", level=0.99, se.fit=TRUE)
```

```
## $fit
##        fit      lwr      upr
## 1 14.88699 11.93875 17.83524
##
## $se.fit
## [1] 0.201945
##
## $df
## [1] 76
##
## $residual.scale
## [1] 1.097455
```

Recall from homework 5, the prediction interval given by Model 2 is (12.09134, 18.14836) with the fitted value (center of the interval) being 15.11985. The current interval is likely to be less biased due to the inclusion of the quadratic term.

Moreover, the above prediction interval is slightly narrower than the one from Model 2, which is due to smaller MSE of the current model (1.204 here vs. 1.281 of Model 2). The SE of the fitted value is actually slightly larger in the current Model compared with that of Model 2 (0.201945 vs. 0.1833524). But this is more than compensated for by the smaller MSE for the prediction SE:

$$s(pred) = \sqrt{s^2(fitted) + MSE}$$

.