Statistics 206

Homework 5 Solution

Due : Nov. 9, 2015, 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.

   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) 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: $R^2_k = 0$ so $VIF_k = 1$.)

   (g) 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.

   (h) 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 $\hat{\beta}_1^{*} = 0$.

(i) To quantify a qualitative variable with three classes C_1, C_2, C_3, we need the following indicator variables:

$$X_1 = \begin{cases} 1 & if \quad C_1 \\ 0 & if \quad otherwise \end{cases} \quad X_2 = \begin{cases} 1 & if \quad C_2 \\ 0 & if \quad otherwise \end{cases} \quad X_3 = \begin{cases} 1 & if \quad C_3 \\ 0 & if \quad otherwise \end{cases}$$

FALSE. We only need X_1 and X_2. C_3 is represented by both $X_1 = X_2 = 0$. Indeed, $X_1 + X_2 + X_3 \equiv 1$, so three of them are in perfect intercorrelation. If all three are included in a model, the LS estimators will not be defined.

(j) Polynomial regression models with higher-order powers (e.g., higher than the third power) are preferred since they provide better approximations to the regression relation.

FALSE. Polynomial regression models with higher-order powers could be highly variable and hard to generalize.

(k) In interaction regression models, the effect of one variable depends on the value of another variable with which it appears together in a cross-product term.

TRUE.

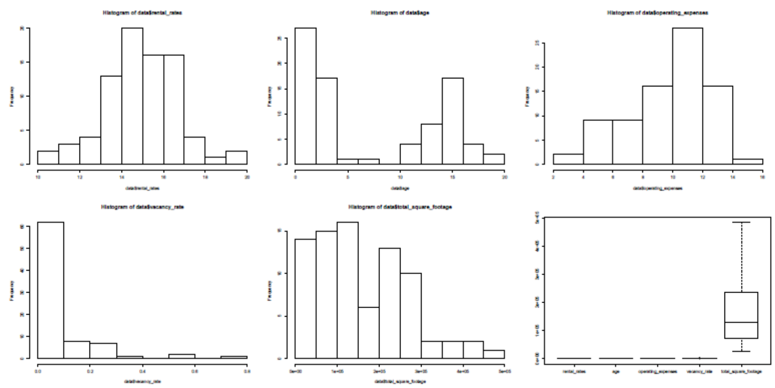
(l) With a qualitative variable, the best way is to fit separate regression models under each of its classes.

FALSE. This usually would not be as efficient as fitting one regression model using indicator variables due to loss of degrees of freedom (since each class will have a smaller sample size and more parameters are being fitted).

2. **(Homework 4, Problem 4 Continued). 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) Draw histogram and boxplot and obtain a summary for each variable. Comment on the distributions of these variables. Do the X variables appear to be in the same scale?



```
> summary(data)
Y                X1               X2               X3
Min.   :10.50    Min.   : 0.000   Min.   : 3.000   Min.   :0.00000
1st Qu.:14.00    1st Qu.: 2.000   1st Qu.: 8.130   1st Qu.:0.00000
Median :15.00    Median : 4.000   Median :10.360   Median :0.03000
Mean   :15.14    Mean   : 7.864   Mean   : 9.688   Mean   :0.08099
3rd Qu.:16.50    3rd Qu.:15.000   3rd Qu.:11.620   3rd Qu.:0.09000
Max.   :19.25    Max.   :20.000   Max.   :14.620   Max.   :0.73000
X4
Min.   : 27000
1st Qu.: 70000
Median :129614
Mean   :160633
3rd Qu.:236000
Max.   :484290
```

X_3 and X_4 are not in the same scale as X_1 and X_2. Rental rates is relatively normal. Age appears to fall into two groups. Operating expenses is slightly skewed to left. Vacancy rate and square-footage skew to right.

(b) 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?

```
> apply(property,2,mean)   #sample mean
Y            X1           X2           X3           X4
1.513889e+01 7.864198e+00 9.688148e+00 8.098765e-02 1.606333e+05
> apply(property,2,sd)   #sample standard deviation
```

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

Correlation transformation:

```
> n=dim(property)[1]
> Y=property$Y
> Y_s=(1/sqrt(n-1))*(Y-mean(Y))/sd(Y)
> X1=property$X1
> X1_s=(1/sqrt(n-1))*(X1-mean(X1))/sd(X1)
> X2=property$X2
> X2_s=(1/sqrt(n-1))*(X2-mean(X2))/sd(X2)
> X3=property$X3
> X3_s=(1/sqrt(n-1))*(X3-mean(X3))/sd(X3)
> X4=property$X4
> X4_s=(1/sqrt(n-1))*(X4-mean(X4))/sd(X4)
> apply(cbind(Y_s,X1_s,X2_s,X3_s,X4_s),2,mean)
Y_s           X1_s           X2_s           X3_s           X4_s
-2.475792e-17 -4.830639e-18  6.347937e-18 -1.403105e-18  1.481986e-17
> apply(cbind(Y_s,X1_s,X2_s,X3_s,X4_s),2,sd)
Y_s      X1_s      X2_s      X3_s      X4_s
0.1118034 0.1118034 0.1118034 0.1118034 0.1118034
```

The sample mean of the transformed variables are 0 and the sample standard deviations of the transformed variables are $1/\sqrt{n-1} = 1/\sqrt{80}$.

(c) 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? 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 4, Problem 4?

$$Y_i^* = \beta_0^* + \beta_1^* X_{i1}^* + \beta_2^* X_{i2}^* + \beta_3^* X_{i3}^* + \beta_4^* X_{i4}^* + \epsilon_i^*, \quad i = 1, 2, \cdots, 81.$$

```
> fit_s=lm(Y_s~X1_s+X2_s+X3_s+X4_s)
> summary(fit_s)

Call:
lm(formula = Y_s ~ X1_s + X2_s + X3_s + X4_s)

Residuals:
Min        1Q    Median        3Q       Max
-0.207223 -0.038429 -0.005914  0.036276  0.191422

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.652e-17  8.213e-03   0.000     1.00
```

```
X1_s        -5.479e-01  8.232e-02  -6.655 3.89e-09 ***
X2_s         4.236e-01  9.490e-02   4.464 2.75e-05 ***
X3_s         4.846e-02  8.504e-02   0.570     0.57
X4_s         5.028e-01  8.786e-02   5.722 1.98e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.07392 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 regression intercept of the standardized model is zero.

Transform back to original model.

```
> beta=coefficients(fit1)
> beta  #coefficient estimates in original model
(Intercept)            X1            X2            X3            X4
1.220059e+01 -1.420336e-01  2.820165e-01  6.193435e-01  7.924302e-06
> beta_s=coefficients(fit_s)[2:5]
> beta_s   #coefficient estimates in standardized model (except beta_0)
X1_s        X2_s        X3_s        X4_s
-0.54785261  0.42364683  0.04846136  0.50275715
> multi=sd(Y)/c(sd(X1), sd(X2), sd(X3), sd(X4))  #the multiplier
> product=multi*beta_s
> product
X1_s          X2_s          X3_s          X4_s
-1.420336e-01  2.820165e-01  6.193435e-01  7.924302e-06
> beta_0=mean(Y)-beta[2]*mean(X1)-beta[3]*mean(X2)-beta[4]*mean(X3)-beta[5]*mean(X4)
> beta_0
X1
12.20059
```

We get the same results as those from Homework 4, Problem 4.

(d) Obtain the standard errors of the fitted regression coefficients **of the X variables** in the original model using the standard errors of the fitted standardized regression coefficients. Compare the results with those from the R output of Problem 4 of Homework 4.

We know that $\hat{\beta}_j = \frac{s_Y}{s_{X_j}}\hat{\beta}_j^*$ for j=1,2,3,4 and $\sigma(\hat{\beta}_j) = \frac{s_Y}{s_{X_j}}\sigma(\hat{\beta}_j^*)$. Therefore, the standard errors for $\hat{\beta}_j$, j=1,2,3,4 are:

$\sigma(\hat{\beta}_1) = \frac{1.719584}{6.632784}8.232\text{e}-02 = 2.134\text{e}-02$

$\sigma(\hat{\beta}_2) = \frac{1.719584}{2.583169}9.490\text{e}-02 = 6.317\text{e}-02$

$\sigma(\hat{\beta}_3) = \frac{1.719584}{0.1345512}8.504\text{e}-02 = 1.087\text{e}+00$

$\sigma(\hat{\beta}_4) = \frac{1.719584}{1.090990\text{e}+05}8.786\text{e}-02 = 1.385\text{e}-06$ These results are same as those in

Homework 4, Problem 4.

(e) Obtain SSTO, SSE and SSR under the standardized model and compare them with those from the original model (Problem 4 of Homework 4). How are they related to each other?

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

Response: Y_s
Df  Sum Sq  Mean Sq F value    Pr(>F)
X1_s       1 0.06264 0.062642 11.4649  0.001125 **
X2_s       1 0.30776 0.307756 56.3262 9.699e-11 ***
X3_s       1 0.03543 0.035431  6.4846  0.012904 *
X4_s       1 0.17892 0.178920 32.7464 1.976e-07 ***
Residuals 76 0.41525 0.005464
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

$SSE_s = 0.41525$, $SSR_s = 0.06264+0.30776+0.03543+0.17892 = 0.58475$, $SSTO_s = SSE_s + SSR_s = 1$. From homework 4, problem 4, $SSE = 98.231$, $SSR = 138.327$, and $SSTO = 236.558$ for the original model. In fact, $SSE_s = SSE/SSTO$, $SSR_s = SSR/SSTO$, and $SSTO_s = SSTO/SSTO$.

(f) Calculate R^2, R_a^2 under the standardized model and compare them with R^2, R_a^2 under the original model (Problem 4 of Homework 4). What do you find?

From the summary output in (d), $R^2 = 0.5847$ and $R_a^2 = 0.5629$ under the standardized model. They are the same as in the original model.

3. **(Homework 4, Problem 4 Continued). Multicollinearity.**

(a) Obtain the correlation matrices $\mathbf{r}_{XX}$ of the four X variables and $\mathbf{r}_{XY}$ between the response variable and the four X variables. Confirm that for the standardized model, $\mathbf{X}'\mathbf{X} = \mathbf{r}_{XX}$ and $\mathbf{X}'\mathbf{Y} = \mathbf{r}_{XY}$.

```
> cor(property)[2:5,2:5]   #r_XX
X1          X2          X3          X4
X1  1.0000000  0.3888264 -0.25266347 0.2885835090990\text{e}+05}8.786\text{e}-02=1.385
X2  0.3888264  1.0000000 -0.37976174 0.44069713
X3 -0.2526635 -0.3797617  1.00000000 0.08061073
X4  0.2885835  0.4406971  0.08061073 1.00000000
> cor(property)[2:5,1]   #r_XY
X1          X2          X3          X4
-0.25028456  0.41378716  0.06652647  0.53526237
> X_s=cbind(X1_s,X2_s,X3_s,X4_s)
> t(X_s)%*%X_s   #X'X
X1_s       X2_s        X3_s       X4_s
```

```
X1_s  1.0000000  0.3888264 -0.25266347 0.28858350
X2_s  0.3888264  1.0000000 -0.37976174 0.44069713
X3_s -0.2526635 -0.3797617  1.00000000 0.08061073
X4_s  0.2885835  0.4406971  0.08061073 1.00000000
> t(X_s)%*%Y_s   #X'Y
[,1]
X1_s -0.25028456
X2_s  0.41378716
X3_s  0.06652647
X4_s  0.53526237
```

(b) 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.

```
> rxx_inv=solve(t(X_s)%*%X_s)
> diag(rxx_inv)   #VIF_k
X1_s     X2_s     X3_s     X4_s
1.240348 1.648225 1.323552 1.412722

> fit3=lm(X1~X2+X3+X4)
> Rs1=summary(fit3)$r.squared
> 1/(1-Rs1)   #VIF_1
[1] 1.240348
> fit4=lm(X2~X1+X3+X4)
> Rs2=summary(fit4)$r.squared
> 1/(1-Rs2)   #VIF_2
[1] 1.648225
> fit5=lm(X3~X1+X2+X4)
> Rs3=summary(fit5)$r.squared
> 1/(1-Rs3)   #VIF_3
[1] 1.323552
> fit6=lm(X4~X1+X2+X3)
> Rs4=summary(fit6)$r.squared
> 1/(1-Rs4)   #VIF_4
[1] 1.412722
```

The largest VIF is 1.65, so there is no severe multicollinearity in this data.

(c) 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.

```
> fit1=lm(Y~X4)
> fit1$coefficients[2]
X4
8.436639e-06
> fit2=lm(Y~X3+X4)
> fit2$coefficients[3]
X4
8.406741e-06
```

They are pretty close.

```
> anova(fit1)
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

> anova(fit2)
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
```

$SSR(X_4) = 67.775$ and $SSR(X_4|X_3) = 66.858$, they are very close.
X_3 is not much correlated with X_4 (correlation coefficient between X_3, X_4 is 0.08), so the effect of having X_3 in the model is very small on the regression coefficient and SSR of X_4.

(d) 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.

```
> fit4=lm(Y~X2)
> fit4$coefficients[2]
X2
0.2754531
> fit5=lm(Y~X4+X2)
```

```
> fit5$coefficients[3]
X2
0.1469682
```

The estimated regression coefficient of X_2 is about half in the model with both X_2 and X_4 compared to that in the model with only X_2.

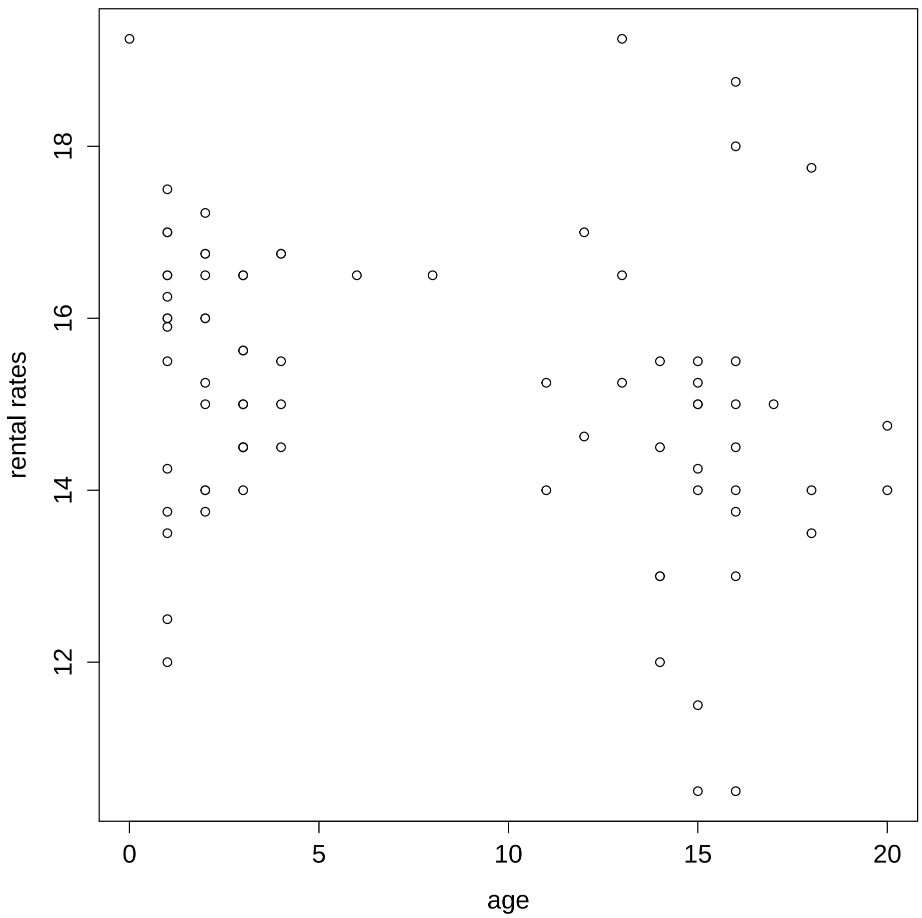
```
> anova(fit4)
Analysis of Variance Table

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
> anova(fit5)
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
```

$SSR(X_2) = 40.503$ and $SSR(X_2|X_4) = 9.291$. When X_4 is already in the model, the marginal contribution from adding X_2 is relatively small since X_2 and X_4 are moderately correlated (correlation coefficient between them is 0.44) and both are moderately correlated with Y.

4. **(Homework 4, Problem 4 Continued). Polynomial Regression.** 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.

   Based on the analysis from Homework 4, Problem 4, 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.



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:

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

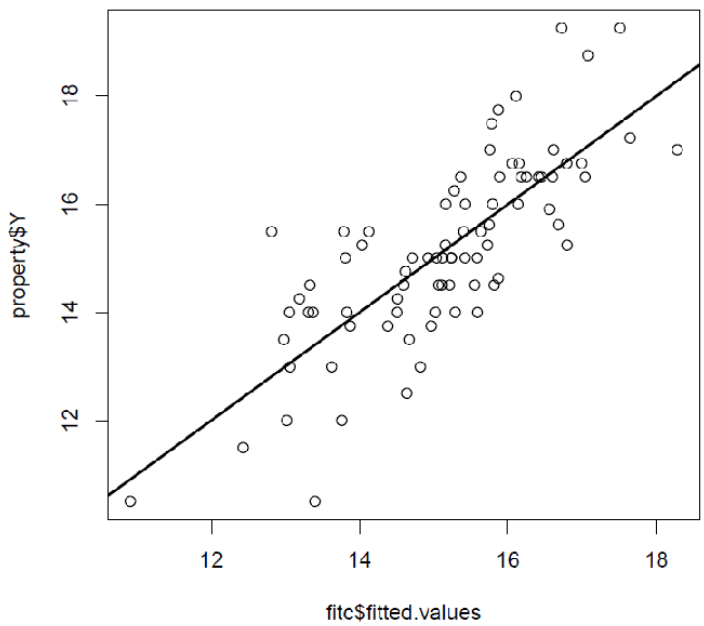
```
> summary(fitc)
```

```
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
```

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 4, Problem 4 ($Y \sim X_1 + X_2 + X_4$). What do you find?

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

Response: Y
Df Sum Sq Mean Sq F value    Pr(>F)
X1          1 14.819  14.819 12.3036 0.0007627 ***
X2          1 72.802  72.802 60.4463 2.968e-11 ***
X4          1 50.287  50.287 41.7522 8.907e-09 ***
I(X1^2)     1  7.115   7.115  5.9078 0.0174321 *
Residuals 76 91.535   1.204
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1
```

For the above model:
$R^2 = 0.6131$, $R_a^2 = 0.5927$. For Model 2 from homework 4:
$R^2 = 0.583$, $R_a^2 = 0.5667$. So the model here has a better fit of the data than Model 2 of homework 4.

(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 pvalue=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 4, Problem 4.

```
> newX=data.frame(X1=4-mean(property$X1),X2=10,X4=80000)
> predict.lm(fitc, 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 4, problem 4, 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:

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

5. **Polynomial Regression**. Write down model equations for the following models.

(a) A third-order polynomial regression model with one predictor.

Model equation:

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + \beta_2 \tilde{X}_i^2 + \beta_3 \tilde{X}_i^3 + \epsilon_i, \quad i = 1, 2, .., n$$

where $\tilde{X}_i = X_i - \bar{X}$.

(b) A second-order polynomial regression model with K predictors.

Model equation:

$$Y_i = \beta_0 + \sum_{k=1}^{K} \beta_k \tilde{X}_{ik} + \sum_{k=1}^{K} \beta_{kk} \tilde{X}_{ik}^2 + \sum_{1 \le k \ne k' \le K} \beta_{kk'} \tilde{X}_{ik} \tilde{X}_{ik'} + \epsilon_i, i = 1, 2, .., n$$

where $\tilde{X}_{ik} = X_{ik} - \bar{X}_k$.

6. **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.

ANS. After standardization, the fitted simple linear regression coefficient of regressing Y^* on each X_j^* alone is $\tilde{\beta}_j^* = r_{Yj}, j = 1, .., p-1$. The fitted regression coefficients of regressing Y^* on $(X_1^*, \cdots, X_{p-1}^*)$ are $\hat{\beta}^* = r_{XX}^{-1} r_{XY} = r_{XY}$ since $r_{XX} = I_{p-1}$ as the X^* variables are uncorrelated and standardized. Hence $\hat{\beta}_j^* = \tilde{\beta}_j^*$ and so the result holds for the standardized model.

For the original model,the fitted simple linear regression coefficients of regressing Y on X_j alone is $\tilde{\beta}_j = \frac{s_Y}{s_{X_j}} r_{Yj}, j = 1, .., p-1$. Transforming the results from the standardized model,the fitted regression coefficients of regressing Y on $(X_1, \cdots, X_{p-1})$ turns out to be $\hat{\beta}_j = \frac{s_Y}{s_{X_j}} \hat{\beta}_j^* = \frac{s_Y}{s_{X_j}} r_{Yj}, j = 1, .., p-1$. Hence the result holds for the original model as $\tilde{\beta}_j = \hat{\beta}_j, j = 1, 2, .., p-1$ which also leads to $\tilde{\beta}_0 = \hat{\beta}_0$.

(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.

ANS. From lecture notes 9, under the standardized regression model

$$SSE(X_{\mathcal{I}}^*, X_j^*) = SSE(X_{\mathcal{I}}^*) - SSR(X_j^*)$$

Therefore,

$$SSR(X_j^*) = SSE(X_{\mathcal{I}}^*) - SSE(X_{\mathcal{I}}^*, X_j^*) = SSR(X_j^* | X_{\mathcal{I}}^*).$$

For the original model,

$$SSR(X_j) = (n-1)s_Y^2 SSR(X_j^*) = (n-1)s_Y^2 SSE(X_{\mathcal{I}}^*) - (n-1)s_Y^2 SSE(X_{\mathcal{I}}^*, X_j^*)$$
$$= SSE(X_{\mathcal{I}}) - SSE(X_{\mathcal{I}}, X_j) = SSR(X_j|X_{\mathcal{I}})$$

Therefore the result holds for original model too.

7. **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

$$VIF_1 = VIF_2 = \frac{1}{1 - r_{12}^2},$$

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

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

8. **(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

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

ANS. From previous homeworks we know that standardization does not change coefficient of multiple determination and the correlations between Y and X, and the correlations among the X variables. Here lets assume that all X variables are standardized by the correlation transformation. By the inverse formula above, let Z denote $X_2, \cdots, X_p$

$$r_{XX}^{-1}(1,1) = (r_{11} - r_{1Z} r_{ZZ}^{-1} r_{Z1})^{-1} = (1 - r_{1Z} r_{ZZ}^{-1} r_{ZZ} r_{ZZ}^{-1} r_{Z1})^{-1} = (1 - \beta_{1Z}' Z' Z \beta_{1Z})^{-1} = \frac{1}{1 - R_1^2}$$

where $\beta_{1Z} = r_{ZZ}^{-1} r_{Z1}$ is the regression coefficient when X_1 is regressed on Z, and $r_{ZZ} = Z'Z$. And in the standardized case, $R_1^2 = SSR/SSTO = SSR/1 = \beta_{1Z}' Z' Z \beta_{1Z}$.