Homework 6

Greg DePaul

2023-03-28

Problem 5 - Commercial Property - 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)

```
my_data <- read.table("property.txt", header=FALSE)
colnames(my_data) <- c('rental_rates','age','operating_expenses','vacancy_rates','square_footage')
fit_1 = lm(rental_rates ~ age + operating_expenses + square_footage + vacancy_rates, data=my_data)
```

(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

$$R^2_{Y_3|X_1X_2X_4}$$

and calculate the coefficient of partial determination

$$r_{Y_3|X_1X_2X_4}$$

and partial correlation. Explain what

$$R^2_{Y_3|X_1X_2X_4}$$

```
summary(fit_1)
```

measures and interpret the result.

```
##
## Call:
## lm(formula = rental_rates ~ age + operating_expenses + square_footage +
##     vacancy_rates, data = my_data)
##
## 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 ***
## age                 -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
## operating_expenses   2.820e-01  6.317e-02   4.464 2.75e-05 ***
## square_footage       7.924e-06  1.385e-06   5.722 1.98e-07 ***
## vacancy_rates        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
```

```
anova(fit_1)
```

```
## Analysis of Variance Table
##
## Response: rental_rates
##                    Df Sum Sq Mean Sq F value    Pr(>F)
## age                 1 14.819  14.819 11.4649  0.001125 **
## operating_expenses  1 72.802  72.802 56.3262 9.699e-11 ***
## square_footage      1 50.287  50.287 38.9062 2.306e-08 ***
## vacancy_rates       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
```

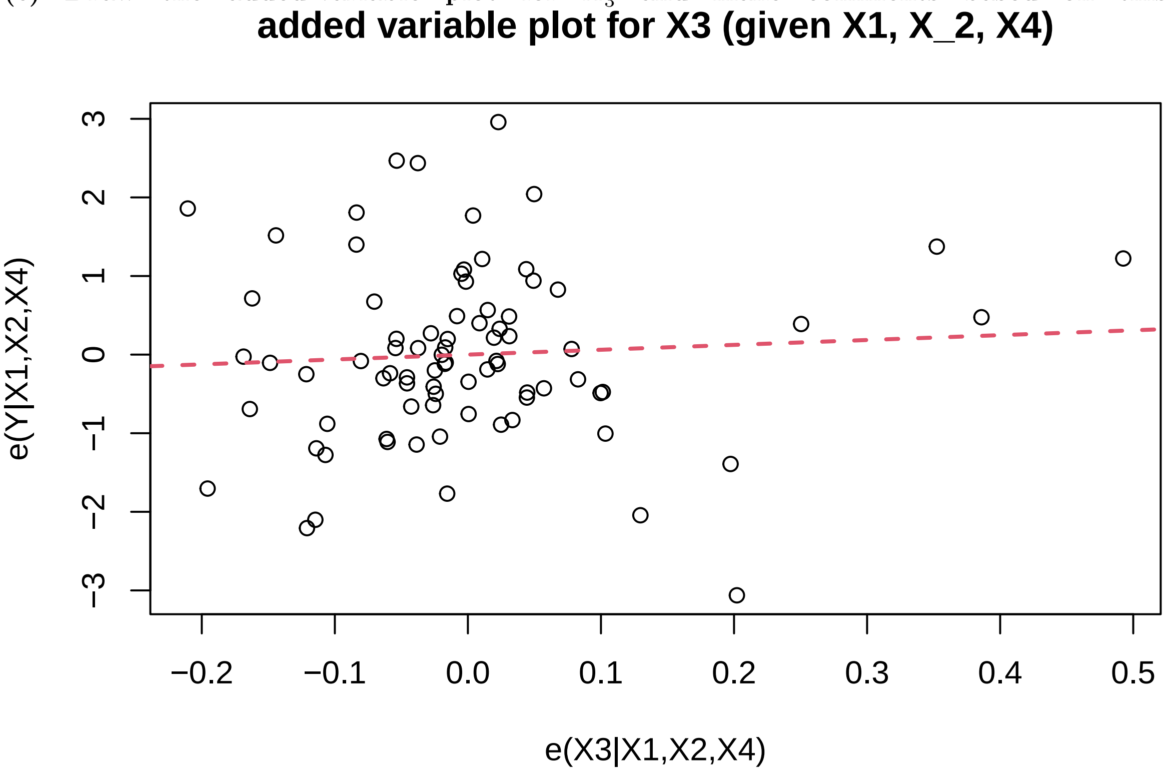
$$R^2_{Y_3|Y_1Y_2Y_4} = \frac{SSR(X_3|X_1, X_2, X_4)}{SSE(X_1, X_2, X_4)} = \frac{0.420}{0.420 + 98.231} = 0.004257433$$

Since $\beta_3 > 0$, then we know we use the positive square root to get:

$$r_{X_3|X_1X_2X_4} = sign(\beta_3)\sqrt{R^2_{X_3|X_1X_2X_4}} = 0.06524901$$

```
eY.124 <- lm(rental_rates ~ age + operating_expenses + square_footage, data=my_data)$residuals
e3.124 <- lm(vacancy_rates ~ age + operating_expenses + square_footage, data=my_data)$residuals
fit_2 <- lm(eY.124~e3.124)
plot(e3.124, eY.124, xlab="e(X3|X1,X2,X4)", ylab="e(Y|X1,X2,X4)", main="added variable plot for X3 (give
abline(fit_2$coefficients, lty=2, lwd=2, col=2)
```

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



```
print(fit_2$coefficients)
```

(d) Regressing the residuals e(Y |X1, X2, X4) to the residuals e(X3|X1, X2, X4). Compare the fitted regression slope from this regression with the fitted regression coefficient of X3 from part (b). What do you find?

```
##   (Intercept)        e3.124
## -5.158855e-17  6.193435e-01
```

We see that the slope found for the added variable plot is identical to that found for the entire model.

```
SSR <- sum((fitted(fit_2) - mean(scale(my_data$rental_rates)))^2)
print(SSR)
```

(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?

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

We see that it's the same value as displayed above!

```
r<- cor(eY.124, e3.124)
print(r)
```

(f) Calculate the correlation coefficient r between the two sets of residuals $e(Y|X1, X2, X4)$ and $e(X3|X1, X2, X4)$. Compare it with $r_{Y3|124}$. What do you find? What is r^2?

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

```
print(r^2)
```

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

It's the same value as derived above for $r_{Y3|124}$.

Lastly,

$$r^2 = 0.004254889$$

which is the same as $R^2_{Y_3|X_1,X_2,X_4}$.

```
temp_fit_1 <- lm(my_data$rental_rates ~ e3.124)
print(temp_fit_1$coefficients)
```

(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?

```
## (Intercept)      e3.124
##  15.1388889   0.6193435
```

It's the same coefficient! This is because we are orthogonalizing the information for X_3 from that of X_1, X_2, and X_4. Therefore, the information in the residuals is equivalent to that remaining in X_3, which explains the same contribution by coefficient.

Problem 6 - Commercial Property - Standardized Regression model.

```
sc.Y = scale(my_data$rental_rates)
print(mean(sc.Y))
```

(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?

```
## [1] -1.706454e-16
```

```
print(sd(sc.Y))
```

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

```
sc.X1 = scale(my_data$age)
print(mean(sc.X1))
```

```
## [1] -2.193033e-17
```

```
print(sd(sc.X1))
```

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

```
sc.X2 = scale(my_data$operating_expenses)
print(mean(sc.X2))
```

```
## [1] -1.949058e-15
```

```
print(sd(sc.X2))
```

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

```
sc.X3 = scale(my_data$vacancy_rates)
print(mean(sc.X3))
```

```
## [1] 4.386066e-17
```

```
print(sd(sc.X3))
```

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

```
sc.X4 = scale(my_data$square_footage)
print(mean(sc.X4))
```

```
## [1] 1.308967e-16
```

```
print(sd(sc.X4))
```

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

```
fit_3 = lm(scale(rental_rates) ~ scale(age) + scale(operating_expenses) + scale(square_footage) + scale
summary(fit_3)
```

(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?

```
##
## Call:
## lm(formula = scale(rental_rates) ~ scale(age) + scale(operating_expenses) +
##     scale(square_footage) + scale(vacancy_rates), data = my_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.85346 -0.34372 -0.05289  0.32446  1.71213
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               5.335e-16  7.346e-02   0.000     1.00
## scale(age)               -5.479e-01  8.232e-02  -6.655 3.89e-09 ***
## scale(operating_expenses)  4.236e-01  9.490e-02   4.464 2.75e-05 ***
## scale(square_footage)      5.028e-01  8.786e-02   5.722 1.98e-07 ***
## scale(vacancy_rates)       4.846e-02  8.504e-02   0.570     0.57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6611 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 standardized first-order regression is:

$$Y = (-3.365e - 16) + -0.5479X_1' + 0.4236X_2' + 0.04846X_3' + 0.5028X_4'$$

The fitted intercept is nearly zero at $-3.365e - 16$.

```
SSE_standardized <- sum((fitted(fit_3) - scale(my_data$rental_rates))^2)
SSR_standardized <- sum((fitted(fit_3) - mean(scale(my_data$rental_rates)))^2)
```

```
SSTO_standardized <- SSE_standardized + SSR_standardized

SSE_original <- sum((fitted(fit_1) - my_data$rental_rates)^2)
SSR_original  <- sum((fitted(fit_1) - mean(my_data$rental_rates))^2)
SSTO_original <- SSE_original + SSR_original

print(c(SSE_original, SSE_standardized))
```

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

```
## [1] 98.23059 33.22003
```

```
print(c(SSR_original, SSR_standardized))
```

```
## [1] 138.32691  46.77997
```

```
print(c(SSTO_original, SSTO_standardized))
```

```
## [1] 236.5575  80.0000
```

All of these values decrease, even though we haven't really changed the model dramatically.

(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?

$$\text{original} : R^2 = 0.5847, R_a^2 = 0.5629$$

$$\text{standardized} : R^2 = 0.5847, R_a^2 = 0.5629$$

These should be the same values.

Problem 7 - Commercial Property - Multicollinearity.

Obtain 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

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

```
correlation_matrix_including_y <- cor(my_data)
correlation_matrix <- correlation_matrix_including_y[2:5,2:5]
inverse_correlation_matrix <- solve(correlation_matrix)
print(inverse_correlation_matrix)
```

Comment on the degree of multicollinearity in this data.

```
##                           age operating_expenses vacancy_rates square_footage
## age                 1.2403482         -0.2870567     0.2244927     -0.2495354
## operating_expenses -0.2870567          1.6482246     0.6092380     -0.6926391
## vacancy_rates       0.2244927          0.6092380     1.3235525     -0.4399669
## square_footage     -0.2495354         -0.6926391    -0.4399669      1.4127219
```

```
VIF <- diag(inverse_correlation_matrix)
print(VIF)
```

```
##                age operating_expenses      vacancy_rates     square_footage
##           1.240348           1.648225           1.323552           1.412722
```

```
temp_fit_1 = lm(age ~ operating_expenses + vacancy_rates + square_footage, data=my_data)
R_1_squared = summary(temp_fit_1)$r.squared
print(R_1_squared)
```

```
## [1] 0.1937747
```

```
temp_fit_2 = lm(operating_expenses~ age + vacancy_rates + square_footage, data=my_data)
R_2_squared = summary(temp_fit_2)$r.squared
print(R_2_squared)
```

```
## [1] 0.3932866
```

```
temp_fit_3 = lm(vacancy_rates ~ operating_expenses + age + square_footage, data=my_data)
R_3_squared = summary(temp_fit_3)$r.squared
print(R_3_squared)
```

```
## [1] 0.2444576
```

```
temp_fit_4 = lm(square_footage ~ operating_expenses + vacancy_rates + age, data=my_data)
R_4_squared = summary(temp_fit_4)$r.squared
print(R_4_squared)
```

```
## [1] 0.2921466
```

We can verify the relationship between the coefficient of partial determination and VIF:

```
print( abs(VIF[1] - 1/(1 - R_1_squared)) < 0.000000000000001 )
```

```
##  age
## TRUE
```

```
print( abs(VIF[2] - 1/(1 - R_2_squared)) < 0.000000000000001 )
```

```
## operating_expenses
##               TRUE
```

```
print( abs(VIF[3] - 1/(1 - R_3_squared)) < 0.000000000000001 )
```

```
## vacancy_rates
##          TRUE
```

```
print( abs(VIF[4] - 1/(1 - R_4_squared)) < 0.000000000000001 )
```

```
## square_footage
##           TRUE
```

```

temp_fit_1 = lm(rental_rates ~ square_footage, data=my_data)
print(temp_fit_1$coefficients)
```

(b) Fit the regression model for relating Y to X4 and fit the regression model for relating Y to X3,X4. Compare the estimated regression coefficients of X4 in these two models. What do you find? Calculate SSR(X4) and SSR(X4|X3). What do you find? Provide an interpretation for your observations.

```
##    (Intercept) square_footage
##   1.378368e+01   8.436639e-06
```

```
temp_fit_2 = lm(rental_rates ~ vacancy_rates + square_footage, data=my_data)
print(temp_fit_2$coefficients)
```

```
##    (Intercept)  vacancy_rates square_footage
##   1.376413e+01   3.007359e-01   8.406741e-06
```

We see that the coefficients are:

$$\beta_{reduced} = 8.436639e - 06$$

$$\beta_{with X_3} = 8.406741e - 06$$

```
SSR_X_4 <- sum((fitted(temp_fit_1) - mean(scale(my_data$rental_rates)))^2)
SSR_X_3_X_4 <- sum((fitted(temp_fit_2) - mean(scale(my_data$rental_rates)))^2)

SSR_X_4_given_X_3 <- SSR_X_3_X_4 - SSR_X_4

print(SSR_X_4)
```

```
## [1] 18631.84
```

```
print(SSR_X_4_given_X_3)
```

```
## [1] 0.130138
```

```

print(temp_fit_1$coefficients)
```

(c) Fit the regression model for relating Y to X2 and fit the regression model for relating Y to X2,X4. Compare the estimated regression coefficients of X2 in these two models. What do you find? Calculate SSR(2) and SSR(X2|X4). What do you find? Provide an interpretation for your observations.

```
##    (Intercept) square_footage
##   1.378368e+01   8.436639e-06
```

```
temp_fit_2 = lm(rental_rates ~ operating_expenses + square_footage, data=my_data)
print(temp_fit_2$coefficients)
```

```
##        (Intercept) operating_expenses     square_footage
##       1.260617e+01       1.469682e-01       6.903097e-06
```

```
SSR_X_4 <- sum((fitted(temp_fit_1) - mean(scale(my_data$rental_rates)))^2)
SSR_X_2_X_4 <- sum((fitted(temp_fit_2) - mean(scale(my_data$rental_rates)))^2)

SSR_X_4_given_X_2 <- SSR_X_2_X_4 - SSR_X_4

print(SSR_X_4)
```

```
## [1] 18631.84
```

```
print(SSR_X_4_given_X_2)
```

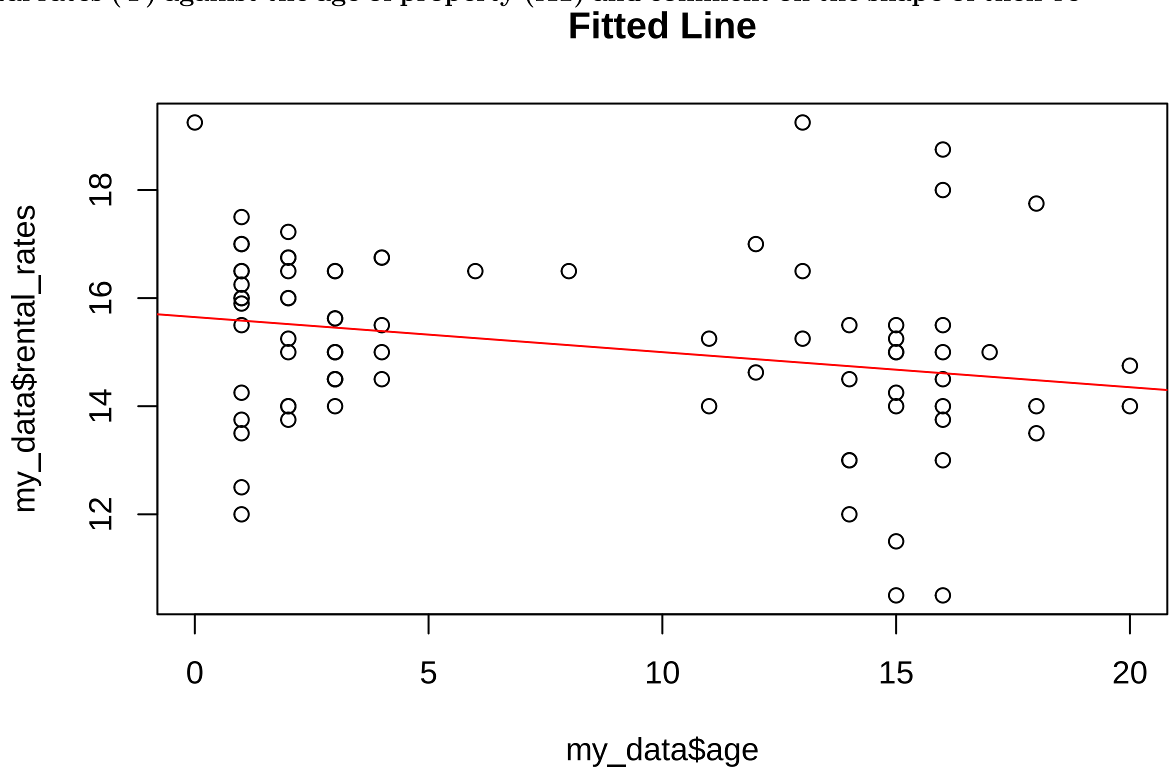
```
## [1] 9.290987
```

Problem 8 - Commercial Property - Polynomial Regression.

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.

```
temp_fit_1 <- lm(rental_rates ~ age, data=my_data)
plot(my_data$age, my_data$rental_rates, main = "Fitted Line")
abline(temp_fit_1, col='red')
```

(a) Plot rental rates (Y) against the age of property (X1) and comment on the shape of their relationship.



```
centered_age <- my_data$age - mean(my_data$age)
operating_expenses <- my_data$operating_expenses
square_footage <- my_data$square_footage
rental_rates <- my_data$rental_rates

fit_4 <- lm(rental_rates ~ centered_age + operating_expenses + square_footage + I(centered_age^2))

print(fit_4$coefficients)
```

(b) Fit a polynomial regression model with linear terms for centered age of property (X_1) operating expenses (X_2), and square footage (X_4), and a quadratic term for centered age of property (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 $\hat{Y}$ plot. Does the model provide a good fit?

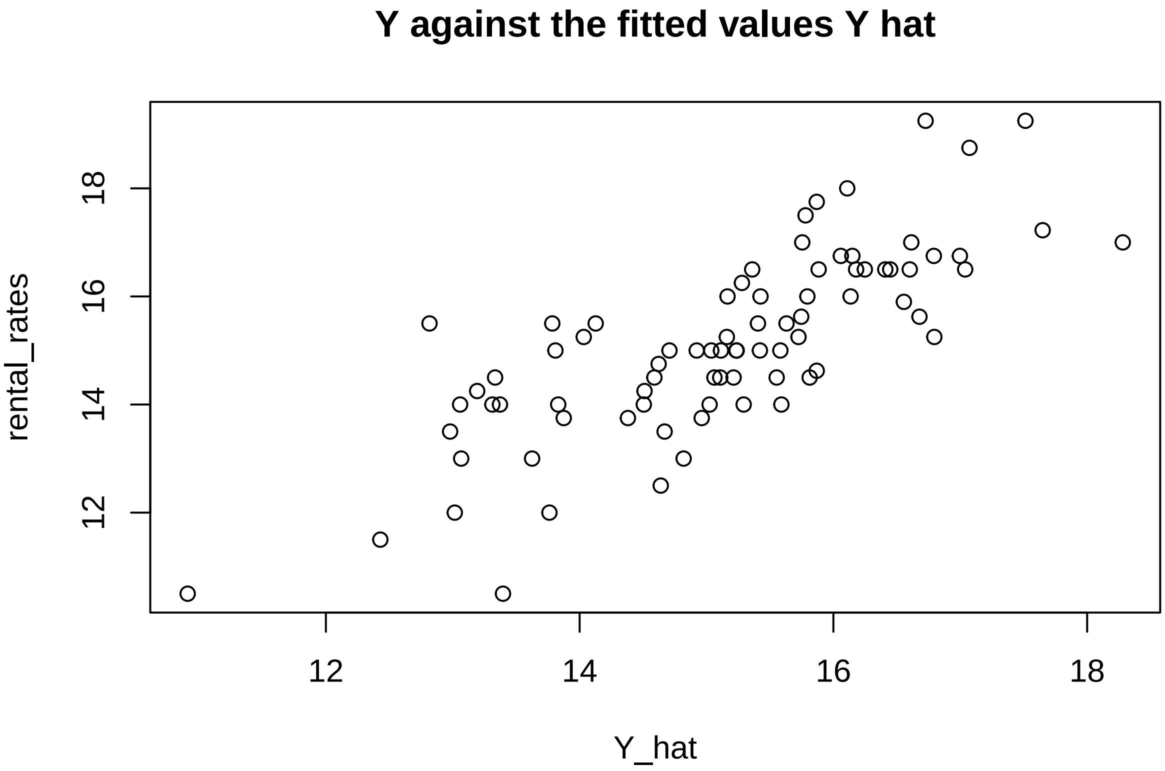
```
##        (Intercept)       centered_age operating_expenses     square_footage
##       1.018934e+01      -1.817749e-01       3.140313e-01       8.045878e-06
##  I(centered_age^2)
##       1.414773e-02
```

```
print(mean(my_data$age))
```

```
## [1] 7.864198
```

```
Y_hat <- fitted(fit_4)

plot(Y_hat, rental_rates, main = "Y against the fitted values Y hat")
```



We see that

$$\hat{Y} = 10.18934 - 0.1817749(X_1 - 7.864198) + 0.3140313X_2 + (8.045878e-06)X_4 + (1.414773e-02)(X_1 - 7.864198)^2$$

Also, we see that we have a a decent model since out Y versus $\hat{Y}$ is close to an identity function.

```
summary(fit_4)
```

(c) Compare R2,Ra2 of the above model with those of Model 2 from Homework 5 (Y X1 + X2 + X4). What do you find?

```
##
## Call:
## lm(formula = rental_rates ~ centered_age + operating_expenses +
##     square_footage + I(centered_age^2))
##
## 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 ***
## centered_age       -1.818e-01  2.551e-02  -7.125 5.10e-10 ***
## operating_expenses  3.140e-01  5.880e-02   5.340 9.33e-07 ***
## square_footage      8.046e-06  1.267e-06   6.351 1.42e-08 ***
## I(centered_age^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
```

We see that

$$R^2 = 0.6131 \quad R_a^2 = 0.5927$$

whereas on homework 5, we have

$$\text{Homework 5: } R^2 = 0.583 \quad R_a^2 = 0.5667$$

We see that these coefficients of determination increase.

(d) Test whether or not the quadratic term for centered age of property (X1) 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. We develop the following hypothesis test:

- $H_0 : \beta_4 = 0$
- $H_A : \beta_4 \neq 0$
- Test Statistic: $F^* = \frac{\frac{SSE(R)-SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}}$
- Null distribution of F^* is $F_{p-1,n-p}(\alpha) = F_{4,76}(\alpha)$

```r
alpha = 0.05
p <- 5
n <- 81
SSE_F <- sum((fitted(fit_4) - rental_rates)^2)
fit_5 = lm(rental_rates ~ centered_age + operating_expenses + square_footage)
SSE_R <- sum((fitted(fit_5) - rental_rates)^2)

F_stat <- ((SSE_R - SSE_F) / 1)/(SSE_F / (n - p))

F_crit <- qf(1 - alpha, p - 1, n-p, lower.tail = TRUE)
print(F_stat > F_crit)
```

```
## [1] TRUE
```

We have enough evidence to reject the null hypothesis. Therefore, we choose not to drop the quadratic term for centered age.

```r
#(Type your code in the space below)
newX = data.frame(centered_age = 4 - mean(my_data$age), operating_expenses = 10, vacancy_rates = 0.1, s
print(predict(fit_4, newX, interval='confidence', level=0.99))
```

(e) Predict the rental rates for a property with X1 = 4, X2 = 10, X4 = 80, 000. Construct a 99% prediction interval and compare it with the prediction interval from Model 2 of Homework 5.

```
##        fit      lwr      upr
## 1 14.88699 14.35344 15.42055
```

The confidence interval is smaller than that of the previous homework!