STP530 HW9

2023-11-06

Commercial properties. A commercial real estate company evaluates vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions.
The data below are taken from 81 suburban commercial properties that are the newest, best located, most attractive, and expensive for five specific geographic areas.
Shown here are the age (X1), operating expenses and taxes (X2), vacancy rates (X3), total square footage (X4), and rental rates (Y).

i:	1	2	3	...	79	80	81
X_{i1}:	1	14	16	...	15	11	14
X_{i2}:	5.02	8.19	3.00	...	11.97	11.27	12.68
X_{i3}:	0.14	0.27	0	...	0.14	0.03	0.03
X_{i4}:	123,000	104,079	39,998	...	254,700	434,746	201,930
Y_i:	13.50	12.00	10.50	...	15.00	15.25	14.50

8.8. Refer to Commercial properties Problems 6.18 and 7.7. The vacancy rate predictor (X3) does not appear to be needed when property age (X1), operating expenses and taxes (X2), and total square of centered footage (X4) are included in the model as predictors of rental rates (Y).

a. The age of the property (X1) appears to exhibit some curvature when plotted against the rental rates (Y). Fit a polynomial regression model with centered property age (XI), the square of centered property age (X1^2), operating expenses and taxes (X2), and total square footage (X4).

V1 = Y = rental rates
V2 = X1 = property age (X1*= X1.centered)
V3 = X2 = operating expenses and taxes
V4 = X3 = vacancy rates
V5 = X4 = total square footage

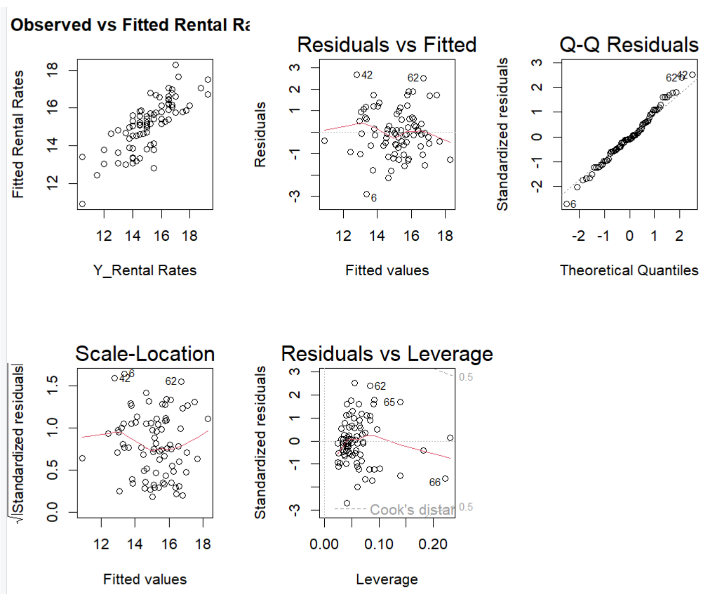
```
> # X1: number of copiers served
> summary(X1_age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.000   4.000   7.864  15.000  20.000
```

```
E{Y} = 10.19 - 0.1818(X1*) + 0.01415(X1*)^2 + 0.314(X2) + 0.000008046(X4)
E{Y} = 10.19 - 0.1818(X1-7.864) + 0.01415(X1-7.864)^2 + 0.314(X2) + 0.000008046(X4)
E{Y} = 12.495128 – 0.404364 (X1) + 0.01415(X1^2) + 0.314(X2) + 0.000008046(X4)
```

```
> # 1) Center X1, to resolve multicollinearity
> mydata$X1.centered <- mydata$X1 - mean(mydata$X1)
> # 2) squared term of the centered X1 (property age)
> mydata$X1.centered.sq <- X1.centered ^ 2
> head(mydata)
     Y X1    X2   X3     X4 X1.centered X1.centered.sq
1 13.5  1  5.02 0.14 123000   -6.864198      47.117208
2 12.0 14  8.19 0.27 104079    6.135802      37.648072
3 10.5 16  3.00 0.00  39998    8.135802      66.191282
4 15.0  4 10.70 0.05  57112   -3.864198      14.932023
5 14.0 11  8.97 0.07  60000    3.135802       9.833257
6 10.5 15  9.45 0.24 101385    7.135802      50.919677
> # 3) Fit the polynominal regression model
> m1 <- lm(Y ~ X1.centered + X1.centered.sq + X2 + X4, data = mydata) #
X3 dropped
> summary(m1)

Call:
lm(formula = Y ~ X1.centered + X1.centered.sq + X2 + X4, data = mydata)

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.centered    -1.818e-01  2.551e-02  -7.125 5.10e-10 ***
X1.centered.sq  1.415e-02  5.821e-03   2.431   0.0174 *
X2              3.140e-01  5.880e-02   5.340 9.33e-07 ***
X4              8.046e-06  1.267e-06   6.351 1.42e-08 ***
---
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
```

Plot the Y observations against the fitted values. Does the response function provide a good fit?

There is some scatter, but overall, it appears that this model appears to reasonably fit.



b. Calculate Ra^2. What information does this measure provide?

The Adjusted R-squared value is 0.5927. It indicates that approximately 59.27% of the variability in rental rates (Y) can be explained by the age of the properties, operating expenses and taxes, and total square footage, along with the nonlinear effect of age (since a squared term of age is included), after adjusting for the number of predictors in the model. This means that the model is relatively good at explaining the variability of the rental rates, but there is still around 40.73% of the variability that is not explained by the model.

c. Test whether or not the square of centered property age (x1^2) can be dropped from the model; use a = .05. State the alternatives, decision rule, and conclusion. What is the p-value of the test?

Step1 : Diagnostic

① Linearity ← Y~X2, Y~X4

② no outlier ← residual ~ $\hat{Y}$

③ constant variance ← residual ~ $\hat{Y}$

④ normality of residuals - QQ-plot

⑤ Independence.

$\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

Step2: Hypothesis

$$Y \sim \underbrace{X1.centered}_{X1^*} + \underbrace{X1.centered.sq}_{(X1^*)^2} + X2 + X4$$

Full model : $\hat{Y} = 10.19 - 0.1818(X1^*) + 0.0145(X1^*)^2 + 0.314(X2) + 0.000008046(X4)$

Reduced : $\hat{Y} = 10.19 - 0.1818(X1^*) + \cancel{} \; 0.314(X2) + 0.000008046(X4)$

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_{11} X_1^2 + \beta_2 X2 + \beta_3 X4$$

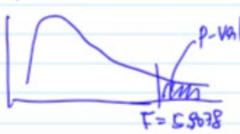
$H_0: \beta_{11} = 0$: should drop X_1^2

$H_1: \beta_{11} \neq 0$: keep X_1^2

Step3 : Test-statistic

$$F_{obs} = \frac{\frac{SSE(R) - SSE(F)}{df(R) - df(F)}}{\frac{SSE(F)}{df(F)}} = 5.9078$$

Step 4: p-value



p-value = 0.01743 < 0.05 = α

Step5: Conclusion.

With p = 0.01743, smaller than α = 0.05, we reject H_0.

We conclude the full model with polynomial term X_1^2 fit significantly better than the reduced model.

We should not drop X_1^2.

```
# Part 4: Test if the square of centered property age (x1^2) can be c
# from the model; use a = .05.
# State the alternatives, decision rule, and conclusion.
# What is the p-value of the test?

# m2: additive, reduced model
m2 <- lm(Y ~ X1.centered + +X2 + X4, data = mydata) # x1^2 dropped
summary(m2)
vif(m2) # good
summary(m2)$r.squared
summary(m1)$r.squared

# Compare models
anova(m2, m1) # p=0.01743, suggested to retain x1^2
rsq::rsq.partial(objF = m1, objR = m2) # 0.07212732, pretty small
```

```
> summary(m2)$r.squared
[1] 0.5829752
> summary(m1)$r.squared
[1] 0.6130541
> # Compare models
> anova(m2, m1) # 4.569e-06, x1^2 should be retained
Analysis of Variance Table

Model 1: Y ~ X1.centered + +X2 + X4
Model 2: Y ~ X1.centered + X1.centered.sq + X2 + X4
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)  
1     77 98.650                              
2     76 91.535  1    7.1154 5.9078 0.01743 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> rsq::rsq.partial(objF = m1, objR = m2) # 0.2764483
$adjustment
[1] FALSE

$variables.full
[1] "X1.centered"    "X1.centered.sq" "X2"             "X4"            

$variables.reduced
[1] "X1.centered" "X2"          "X4"         

$partial.rsq
[1] 0.07212732
```

D. Estimate the mean rental rate when X1 = 8, X2 = 16, and X4 = 250,000; use a 95% confidence interval. Interpret your interval.

95% confidence interval = (14.20138, 16.18148)

Interpretation: We are 95% sure that the mean Y (Rental Rates) for all observations with X1 (property age) = 8, thus X1.centered = 0.1358025, X2 (operating expenses and taxes) = 16, and X4 (total square footage) = 250,000 falls between 14.20138 and 16.18148.

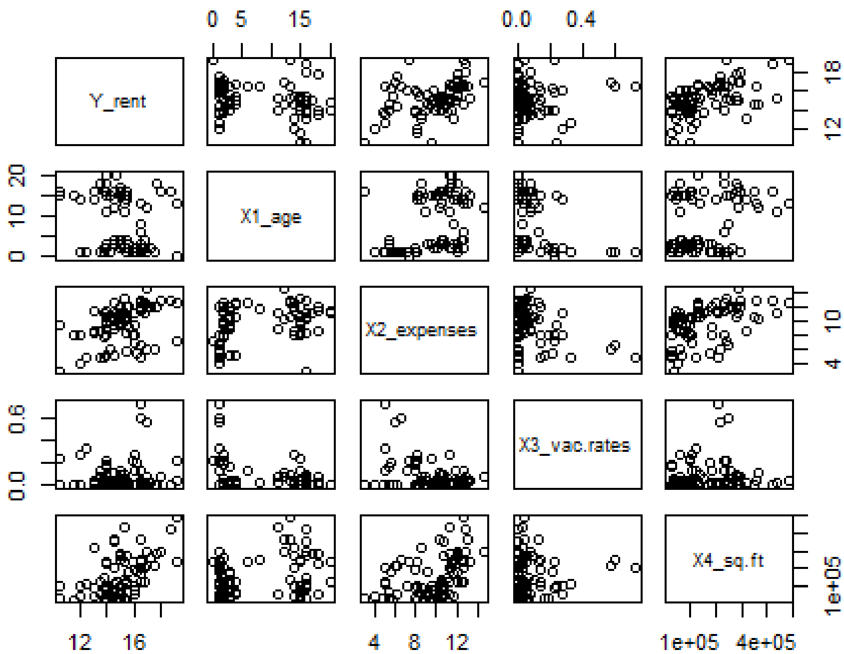
```
# Estimate the mean rental rate (CI) when X1 = 8, X2 = 16, and X4 = 250,000;
# use a 95%  confidence interval. Interpret your interval.

8 - mean(X1)
# First transform the new X1 values
new_data <- data.frame(X1.centered= 8 - mean(X1),
                      X1.centered.sq =(8 - mean(X1))^2,
                      X2 = 16,
                      X4 = 250,000)

# 95% confidence interval of E{Y}
predict(m1, newdata=new_data,interval="confidence", level=.95)
# Interpretation: We are 95% sure that the mean Y (Rental Rates) for
# all observations with X1 (property age) = 8, thus X1.centered = 0.1358025,
# X2 (operating expenses and taxes) = 16, and
# X4 (total square footage) = 250,000 falls between 14.20138 and 16.18148.
```

```
> b0.star <- 10.19
> b1.star <- -0.1818
> b11.star <- 0.01415
> X1.bar <- 7.864198 #mean(X1)
> b0 <- b0.star - b1.star*(X1.bar) + b11.star*(X1.bar^2) # 12.49483
> b1 <- b1.star -2* (b11.star)*(X1.bar) # -0.4043568
> b11 <- b11.star # 0.01415
> b0; b1; b11
[1] 12.49483
[1] -0.4043568
[1] 0.01415
> 8 - mean(X1)
[1] 0.1358025
> # First transform the new X1 values
> new_data <- data.frame(X1.centered= 8 - mean(X1),
+                        X1.centered.sq =(8 - mean(X1))^2,
+                        X2 = 16,
+                        X4 = 250,000)
> # 95% confidence interval of E{Y}
> predict(m1, newdata=new_data,interval="confidence", level=.95)
       fit      lwr      upr
1 15.19143 14.20138 16.18148
```

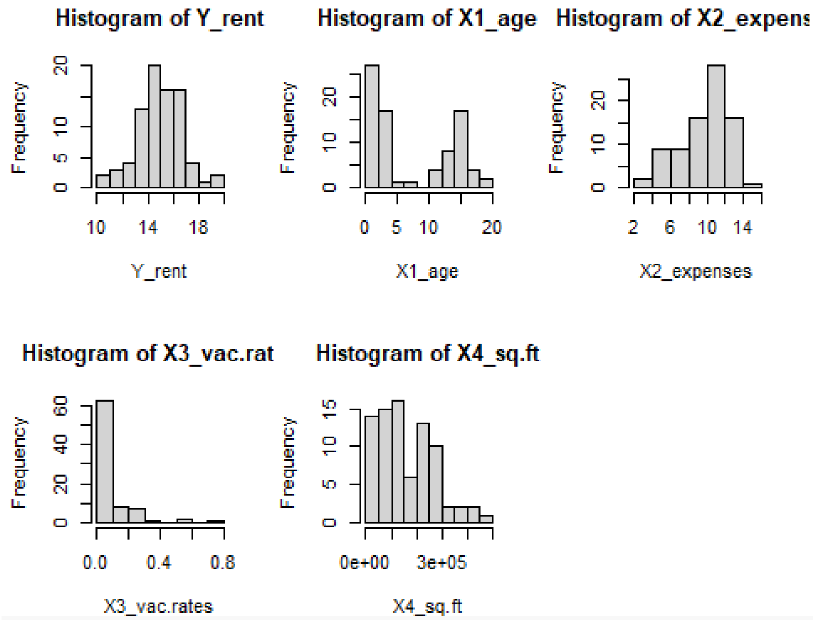
R codes

```
#
rm(list=ls()) # Clean up the workspace for the new analysis

# Set the following to your own folder
setwd("C:/Users/jyang/OneDrive - Arizona State University/10 Classes_OneDrive/2023_STP530_Regression/HW9")


library(rgl)
library(Hmisc)
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:base':
##
##     format.pval, units
library(Rmisc)
## Loading required package: lattice
## Loading required package: plyr
##
## Attaching package: 'plyr'
## The following objects are masked from 'package:Hmisc':
##
##     is.discrete, summarize
library(ggplot2)

# Import the dataset, and inspect the data.
mydata <- read.table("CH06PR18.txt")
head(mydata)
##     V1 V2    V3   V4     V5
## 1 13.5  1  5.02 0.14 123000
## 2 12.0 14  8.19 0.27 104079
## 3 10.5 16  3.00 0.00  39998
## 4 15.0  4 10.70 0.05  57112
## 5 14.0 11  8.97 0.07  60000
## 6 10.5 15  9.45 0.24 101385
# Re-name columns
# V1 = Y  = rental rates
# V2 = X1 = age
# V3 = X2 = operating expenses and taxes
# V4 = X3 = vacancy rates
# V5 = X4 = total square footage
```

```
colnames(mydata) <- c("Y_rent", "X1_age", "X2_expenses", "X3_vac.rates", "X4_sq.ft")
str(mydata)
## 'data.frame':    81 obs. of  5 variables:
##  $ Y_rent      : num  13.5 12 10.5 15 14 10.5 14 16.5 17.5 16.5 ...
##  $ X1_age      : int  1 14 16 4 11 15 2 1 1 8 ...
##  $ X2_expenses : num  5.02 8.19 3 10.7 8.97 ...
##  $ X3_vac.rates: num  0.14 0.27 0 0.05 0.07 0.24 0.19 0.6 0 0.03 ...
##  $ X4_sq.ft    : int  123000 104079 39998 57112 60000 101385 31300 248172 215000 251015 ...
Hmisc::describe(mydata)
## mydata
##
##  5  Variables      81  Observations
## --------------------------------------------------------------------------------
## Y_rent
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       81        0       28    0.995    15.14     1.91    12.00    13.00
##      .25      .50      .75      .90      .95
##    14.00    15.00    16.50    17.00    17.75
##
## lowest : 10.5  11.5  12    12.5  13   , highest: 17.5  17.75 18    18.75 19.25
## --------------------------------------------------------------------------------
## X1_age
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       81        0       16    0.985    7.864    7.293        1        1
##      .25      .50      .75      .90      .95
##        2        4       15       16       18
##
## Value          0     1     2     3     4     6     8    11    12    13    14
## Frequency      1    15    11    12     5     1     1     2     2     3     5
## Proportion 0.012 0.185 0.136 0.148 0.062 0.012 0.012 0.025 0.025 0.037 0.062
##
## Value         15    16    17    18    20
## Frequency      8     9     1     3     2
## Proportion 0.099 0.111 0.012 0.037 0.025
##
## For the frequency table, variable is rounded to the nearest 0
## --------------------------------------------------------------------------------
## X2_expenses
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       81        0       75        1    9.688    2.909     5.00     5.50
##      .25      .50      .75      .90      .95
##     8.13    10.36    11.62    12.68    12.86
##
## lowest : 3     4     4.82  4.99  5    , highest: 12.86 12.97 12.99 13.23 14.62
## --------------------------------------------------------------------------------
## X3_vac.rates
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       81        0       24    0.946  0.08099   0.1143     0.00     0.00
##      .25      .50      .75      .90      .95
##     0.00     0.03     0.09     0.22     0.27
##
## lowest : 0    0.02 0.03 0.04 0.05, highest: 0.27 0.33 0.57 0.6  0.73
## --------------------------------------------------------------------------------
## X4_sq.ft
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##       81        0       76        1   160633   121736    32000    40500
##      .25      .50      .75      .90      .95
##    70000   129614   236000   296966   359665
##
## lowest :  27000  30005  31300  31750  32000, highest: 359665 366013 421000 434746 484290
## --------------------------------------------------------------------------------
pairs(mydata)
```




```
# Dependent variable Y: rental rates
attach(mydata)
summary(Y_rent)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.50   14.00   15.00   15.14   16.50   19.25
sd(Y_rent, na.rm=T)
## [1] 1.719584
par(mfrow=c(2, 3))
hist(Y_rent)

# X1: number of copiers served
summary(X1_age)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   4.000   7.864  15.000  20.000
hist(X1_age)

# X2: operating expenses and taxes
summary(X2_expenses)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   8.130  10.360   9.688  11.620  14.620
hist(X2_expenses)

# X3: vacancy rates
summary(X3_vac.rates)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.03000 0.08099 0.09000 0.73000
hist(X3_vac.rates)

# X4: total square footage
summary(X4_sq.ft)
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   27000   70000  129614  160633  236000  484290
hist(X4_sq.ft)
```

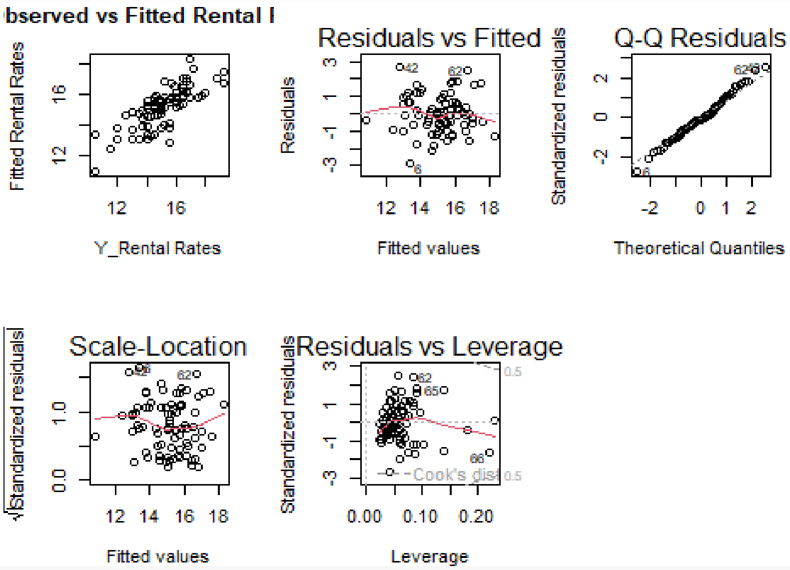



```
# ------------------------------------------------------------------------------
# Part I. regression models, rename column
colnames(mydata) <- c("Y", "X1", "X2", "X3", "X4")

# m0: Additive
m0 <- lm(Y ~ X1 + X2 + X3 + X4, data=mydata)
library(car)
## Loading required package: carData
vif(m0)
##       X1       X2       X3       X4
## 1.240348 1.648225 1.323552 1.412722
summary(m0)
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = mydata)
##
## 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
summary(m0)$r.squared
## [1] 0.5847496
# m.X4:
m.X4 <- lm(Y ~ X4, data=mydata)
summary(m.X4)
##
## Call:
## lm(formula = Y ~ X4, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```r
## -4.1390 -0.7930  0.2890  0.9653  3.4415
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.378e+01  2.903e-01  47.482  < 2e-16 ***
## X4          8.437e-06  1.498e-06   5.632 2.63e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.462 on 79 degrees of freedom
## Multiple R-squared:  0.2865, Adjusted R-squared:  0.2775
## F-statistic: 31.72 on 1 and 79 DF,  p-value: 2.628e-07
summary(m.X4)$r.squared # 0.2865058
## [1] 0.2865058
# m.X1X4: X1|X4
m.X1X4 <- lm(Y ~ X4 + X1, data=mydata)
vif(m.X1X4)
##       X4       X1
## 1.090846 1.090846
summary(m.X1X4)
##
## Call:
## lm(formula = Y ~ X4 + X1, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2032 -0.4593  0.0641  0.7730  2.5083
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.436e+01  2.771e-01  51.831  < 2e-16 ***
## X4           1.045e-05  1.363e-06   7.663 4.23e-11 ***
## X1          -1.145e-01  2.242e-02  -5.105 2.27e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.274 on 78 degrees of freedom
## Multiple R-squared:  0.4652, Adjusted R-squared:  0.4515
## F-statistic: 33.93 on 2 and 78 DF,  p-value: 2.506e-11
summary(m.X1X4)$r.squared # 0.4652132
## [1] 0.4652132
# m.X2X14: X2|X1,X4
m.X2X14 <- lm(Y ~ X4 + X1 + X2, data=mydata)
vif(m.X2X14)
##       X4       X1       X2
## 1.266471 1.202271 1.367789
summary(m.X2X14)
##
## Call:
## lm(formula = Y ~ X4 + X1 + X2, data = mydata)
##
## 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 ***
## X4           8.178e-06  1.305e-06   6.265 1.97e-08 ***
## X1          -1.442e-01  2.092e-02  -6.891 1.33e-09 ***
## X2           2.672e-01  5.729e-02   4.663 1.29e-05 ***
## ---
## 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
summary(m.X2X14)$r.squared # 0.5829752
## [1] 0.5829752
```

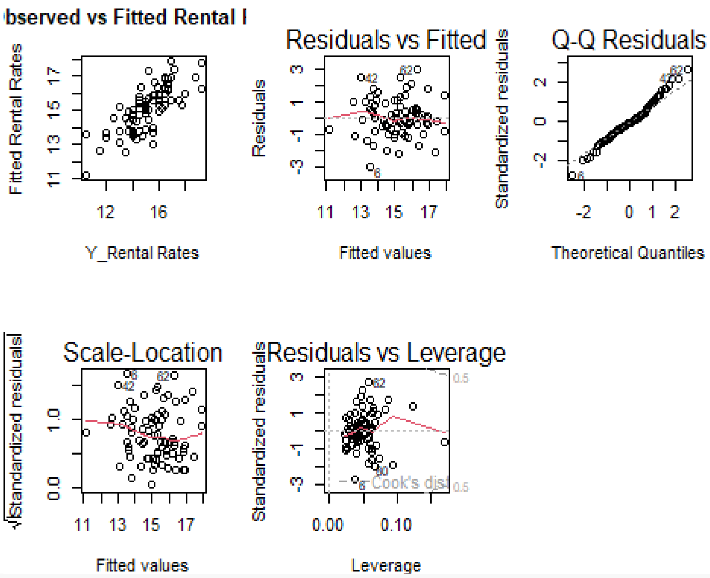
```r
# m.X3X124: X3|X1,X2,X4
m.X3X124 <- lm(Y ~ X4 + X1 + X2 + X3, data=mydata)
vif(m.X3X124)
##       X4       X1       X2       X3 
## 1.412722 1.240348 1.648225 1.323552
summary(m.X3X124)
## 
## Call:
## lm(formula = Y ~ X4 + X1 + X2 + X3, data = mydata)
## 
## Residuals:
##     Min      1Q  Median      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 ***
## X4           7.924e-06  1.385e-06   5.722 1.98e-07 ***
## 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    
## ---
## 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
summary(m.X3X124)$r.squared # 0.5847496
## [1] 0.5847496
# Compare models
anova(m.X2X14, m.X3X124) # p=0.5704, X3 can be dropped
## Analysis of Variance Table
## 
## Model 1: Y ~ X4 + X1 + X2
## Model 2: Y ~ X4 + X1 + X2 + X3
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     77 98.650                           
## 2     76 98.231  1   0.41975 0.3248 0.5704
rsq::rsq.partial(objF = m.X3X124, objR = m.X2X14) # 0.004254889
## $adjustment
## [1] FALSE
## 
## $variables.full
## [1] "X4" "X1" "X2" "X3"
## 
## $variables.reduced
## [1] "X4" "X1" "X2"
## 
## $partial.rsq
## [1] 0.004254889
#-------------------------------------------------------------------------------
# Part 2: polynomial models
attach(mydata)
# 1) Center X1, to resolve multicollinearity
mydata$X1.centered <- mydata$X1 - mean(mydata$X1)

# 2) squared term of the centered X1 (property age)
mydata$X1.centered.sq <- mydata$X1.centered ^ 2
head(mydata)
##      Y X1    X2   X3     X4 X1.centered X1.centered.sq
## 1 13.5  1  5.02 0.14 123000   -6.864198      47.117208
## 2 12.0 14  8.19 0.27 104079    6.135802      37.648072
## 3 10.5 16  3.00 0.00  39998    8.135802      66.191282
## 4 15.0  4 10.70 0.05  57112   -3.864198      14.932023
## 5 14.0 11  8.97 0.07  60000    3.135802       9.833257
## 6 10.5 15  9.45 0.24 101385    7.135802      50.919677
# 3) Fit the polynominal regression model
m1 <- lm(Y ~ X1.centered + X1.centered.sq + X2 + X4, data = mydata) # X3 dropped
summary(m1)
```

```
##
## Call:
## lm(formula = Y ~ X1.centered + X1.centered.sq + X2 + X4, data = mydata)
##
## 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.centered    -1.818e-01  2.551e-02  -7.125 5.10e-10 ***
## X1.centered.sq  1.415e-02  5.821e-03   2.431   0.0174 *
## X2              3.140e-01  5.880e-02   5.340 9.33e-07 ***
## X4              8.046e-06  1.267e-06   6.351 1.42e-08 ***
## ---
## 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
# E{Y} = 10.19 - 0.1818(X1.centered) + 0.01415(X1.centered^2) + 0.314(X2) + 0.000008046(X4)
# E{Y} = 12.495128 – 0.404364 (X1) + 0.01415(X1^2) + 0.314(X2) + 0.000008046(X4)

vif(m1) # acceptable
##    X1.centered X1.centered.sq             X2             X4
##       1.901945       1.608797       1.532560       1.268814
# Plot the observed Y values against the fitted values from the model
par(mfrow=c(2, 3))
plot(mydata$Y, m1$fitted.values,
     xlab = "Y_Rental Rates",
     ylab = "Fitted Rental Rates",
     main = "Observed vs Fitted Rental Rates")
# residual plots for model diagnostics
# Checking for linearity, constant variance, and normality of residuals
plot(m1)
```




```
#---------------------------------------------------------------------------------
# Part 4: Test if the square of centered property age (x1^2) can be dropped
# from the model; use a = .05.
# State the alternatives, decision rule, and conclusion.
# What is the p-value of the test?

# m2: additive, reduced model
m2 <- lm(Y ~ X1.centered + +X2 + X4, data = mydata) # x1^2 dropped
summary(m2)
```

```r
##
## Call:
## lm(formula = Y ~ X1.centered + +X2 + X4, data = mydata)
##
## 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)  1.124e+01  5.303e-01  21.190  < 2e-16 ***
## X1.centered -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
vif(m2) # good
## X1.centered          X2          X4
##    1.202271    1.367789    1.266471
summary(m2)$r.squared
## [1] 0.5829752
summary(m1)$r.squared
## [1] 0.6130541
# Compare models
anova(m2, m1) # p=0.01743 < 0.05 = alpha, suggested to retain x1^2
## Analysis of Variance Table
##
## Model 1: Y ~ X1.centered + +X2 + X4
## Model 2: Y ~ X1.centered + X1.centered.sq + X2 + X4
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     77 98.650
## 2     76 91.535  1    7.1154 5.9078 0.01743 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
rsq::rsq.partial(objF = m1, objR = m2) # 0.07212732, 7% additional.
## $adjustment
## [1] FALSE
##
## $variables.full
## [1] "X1.centered"    "X1.centered.sq" "X2"             "X4"
##
## $variables.reduced
## [1] "X1.centered" "X2"          "X4"
##
## $partial.rsq
## [1] 0.07212732
# Take m1


# disgnostic
par(mfrow=c(2, 3))
plot(mydata$Y, m2$fitted.values,
     xlab = "Y_Rental Rates",
     ylab = "Fitted Rental Rates",
     main = "Observed vs Fitted Rental Rates")
# residual plots for model diagnostics
# Checking for linearity, constant variance, and normality of residuals
plot(m2)
```




```
#-------------------------------------------------------------------------------
# Part 5: Convert the final model (m1) to the original variable scales
m1 <- lm(Y ~ X1.centered + X1.centered.sq + X2 + X4, data = mydata)
summary(m1)
##
## Call:
## lm(formula = Y ~ X1.centered + X1.centered.sq + X2 + X4, data = mydata)
##
## 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.centered    -1.818e-01  2.551e-02  -7.125 5.10e-10 ***
## X1.centered.sq  1.415e-02  5.821e-03   2.431   0.0174 *
## X2              3.140e-01  5.880e-02   5.340 9.33e-07 ***
## X4              8.046e-06  1.267e-06   6.351 1.42e-08 ***
## ---
## 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
# E{Y} = 10.19 - 0.1818(X1-7.864) + 0.01415(X1-7.864)^2 + 0.314(X2) + 0.000008046(X4)
# E{Y} = 12.495128 – 0.404364 (X1) + 0.01415(X1^2) + 0.314(X2) + 0.000008046(X4)

b0.star <- 10.19
b1.star <- -0.1818
b11.star <- 0.01415

X1.bar <- 7.864198 #mean(X1)

b0 <- b0.star - b1.star*(X1.bar) + b11.star*(X1.bar^2) # 12.49483
b1 <- b1.star -2* (b11.star)*(X1.bar) # -0.4043568
b11 <- b11.star # 0.01415
b0; b1; b11
## [1] 12.49483
## [1] -0.4043568
## [1] 0.01415
# E{Y} = 12.495 – 0.40436 (X1) + 0.01415(X1^2) + 0.314(X2) + 0.000008046(X4)
#-------------------------------------------------------------------------------
# Estimate the mean rental rate (CI) when X1 = 8, X2 = 16, and X4 = 250,000;
# use a 95%  confidence interval. Interpret your interval.
```

```r
8 - mean(X1)
## [1] 0.1358025
# First transform the new X1 values
new_data <- data.frame(X1.centered= 8 - mean(X1),
                       X1.centered.sq =(8 - mean(X1))^2,
                       X2 = 16,
                       X4 = 250,000)

# 95% confidence interval of E{Y}
predict(m1, newdata=new_data,interval="confidence", level=.95)
##        fit      lwr      upr
## 1 15.19143 14.20138 16.18148
# Interpretation: We are 95% sure that the mean Y (Rental Rates) for
# all observations with X1 (property age) = 8, thus X1.centered = 0.1358025,
# X2 (operating expenses and taxes) = 16, and
# X4 (total square footage) = 250,000 falls between 14.20138 and 16.18148.

# 95% prediction interval of Y-hat
predict(m1, newdata=new_data, interval="prediction", level=.95)
##        fit      lwr      upr
## 1 15.19143 12.79189 17.59097
```