### **Problem 1: 10.12**

Refer to Commercial Properties Problem 6.18.

#### Obtain the studentized deleteted residuals and identify any outlying Y observations. Use the Bonferroni outlier test procedure with a = .01. State the decision rule and conclusion.

*Answer:*

Below, I load in the dataset and set the column names:

commercial\_df <- **read.table**(**url**(["http://people.stat.sc.edu/Hitchcock/commercialproperties.txt"](http://people.stat.sc.edu/Hitchcock/commercialproperties.txt)), header

**head**(commercial\_df)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## | Rental\_Rates Age Operating\_Costs | | | Vacancy\_Rates Total\_Sqft | |
| ## 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 |
| **summary**(commercial\_df) | | | | | |
| ## | Rental\_Rates | Age | | Operating\_Costs | Vacancy\_Rates |
| ## | 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 |
| ## | Total\_Sqft |  | |  |  |
| ## | Min. : 27000 |  | |  |  |
| ## | 1st Qu.: 70000 |  | |  |  |
| ## | Median :129614 |  | |  |  |
| ## | Mean :160633 |  | |  |  |
| ## | 3rd Qu.:236000 |  | |  |  |
| ## | Max. :484290 |  | |  |  |
| **nrow**(commercial\_df) | | | | | |

## [1] 81

Below, I fit the model:

linear\_fit <- **lm**(commercial\_df**$**Rental\_Rates**~**.,commercial\_df) **summary**(linear\_fit)

##

## Call:

## lm(formula = commercial\_df$Rental\_Rates ~ ., data = commercial\_df) ##

## 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\_Costs 2.820e-01 6.317e-02 4.464 2.75e-05 \*\*\* ## Vacancy\_Rates 6.193e-01 1.087e+00 0.570 0.57

## Total\_Sqft 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

Below, I am obtaining the studentized deleted residuals using the rstudent() function:

stud.del.resids=**rstudent**(linear\_fit) stud.del.resids

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## | 1 | 2 | 3 | 4 | 5 |
| ## | -0.939937884 | -1.392595843 | -0.577011367 | -0.119075516 | 0.278241102 |
| ## | 6 | 7 | 8 | 9 | 10 |
| ## | -3.072104540 | -0.481687004 | 0.231248688 | 1.875611605 | 0.093764786 |
| ## | 11 | 12 | 13 | 14 | 15 |
| ## | 0.020906531 | -0.301082297 | 0.639839452 | -0.353610361 | -0.180875431 |
| ## | 16 | 17 | 18 | 19 | 20 |
| ## | -0.744640979 | 0.090403735 | -1.615503751 | -1.107548586 | -0.562849841 |
| ## | 21 | 22 | 23 | 24 | 25 |
| ## | -0.327339329 | 0.258289216 | -0.083335150 | 0.208326555 | -0.772290974 |
| ## | 26 | 27 | 28 | 29 | 30 |
| ## | -1.978297968 | 0.416597491 | -0.511757560 | -0.956028829 | -0.176356568 |
| ## | 31 | 32 | 33 | 34 | 35 |
| ## | -1.005353208 | -0.154727268 | -0.924215724 | -0.081555334 | 0.192083043 |
| ## | 36 | 37 | 38 | 39 | 40 |
| ## | 0.697628356 | 0.984132623 | -1.993210409 | -0.164752214 | -1.018159727 |
| ## | 41 | 42 | 43 | 44 | 45 |
| ## | -0.011515894 | 2.323017752 | -1.486145080 | 0.837830828 | 0.354117109 |
| ## | 46 | 47 | 48 | 49 | 50 |
| ## | 0.104403896 | 0.742467390 | 1.473706856 | 0.500426349 | 0.444803618 |
| ## | 51 | 52 | 53 | 54 | 55 |
| ## | 0.187431458 | -0.028642460 | 1.124081029 | 0.217241661 | -0.958385664 |
| ## | 56 | 57 | 58 | 59 | 60 |
| ## | 0.950756394 | -0.234536027 | 0.921077316 | -0.307615280 | 0.181887056 |
| ## | 61 | 62 | 63 | 64 | 65 |
| ## | 0.967199714 | 2.784062894 | 2.279078543 | 1.740674980 | 1.376416945 |
| ## | 66 | 67 | 68 | 69 | 70 |
| ## | -0.435479665 | -0.677252624 | 1.841761349 | 0.071777857 | 0.008824756 |
| ## | 71 | 72 | 73 | 74 | 75 |
| ## | 1.602838759 | -0.415806676 | -0.456318996 | -0.094082551 | 1.101175784 |
| ## | 76 | 77 | 78 | 79 | 80 |
| ## | -0.232389250 | -0.562616429 | 0.830178801 | -0.492369611 | -1.923161181 |
| ## | 81 |  |  |  |  |
| ## | -0.809548097 |  |  |  |  |

We can identify any outlying Y observations using the Bonferroni outlier test procedure: The appropriate

Bonferroni critical value is *t*(*l* − *a/*2*n*; *n* − *p* − 1), where a is .01 - in this case it is t(1-a/162; 81-5-1). If |*ti*| ≤ *t*(*l* − *a/*2*n*; *n* − *p* − 1), the case is not an outlier; otherwise, the case is an outlier.

I implement this below, and we see *t*(1 − *a/*162; 81 − 5 − 1) = 4*.*05.

bonferonni\_result <- **qt**(1-0.01**/**(2**\*nrow**(commercial\_df)), **nrow**(commercial\_df) **-** 5 **-** 1) bonferonni\_result

## [1] 4.050335

We can use the any() function to check if *ti* 4*.*05; the result below is False, which indicates that *ti* 4*.*05

| | ≤ | | ≤

and we conclude there are no outliers.

**any**(stud.del.resids**>=**bonferonni\_result)

## [1] FALSE

#### Obtain the diagonal elements of the hat matrix. Identify any outlying X observations.

*Answer:*

The course Lecture 10 slides illustrate the use of the hatvalues function for obtaining the hat matrix. We note in this case 2p/n will be equal to 2\*5/81 - since p = 5 and n = 81.

hii<-**hatvalues**(linear\_fit)

index<-hii**>**2**\***5**/nrow**(commercial\_df) index

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ## | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| ## | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE |
| ## | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| ## | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE |
| ## | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 |
| ## | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE |
| ## | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 |
| ## | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE |
| ## | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 |
| ## | FALSE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE |
| ## | 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 | 71 | 72 |
| ## | TRUE | FALSE | FALSE | FALSE | TRUE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE |
| ## | 73 | 74 | 75 | 76 | 77 | 78 | 79 | 80 | 81 |  |  |  |
| ## | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE | FALSE |  |  |  |

We see *h*3, *h*8, *h*53, *h*61, and *h*65 are outlying X observations.

#### The researcher wishes to estimate the rental rates of a property whose age is 10 years, whose operating expenses and taxes are 12.00, whose ocupancy rate is 0.05, and whose square footage is 350,000. Use (10.29) to determine whether this estimate will involve a hidden extrapolation.

*Answer:*

10.29 states:

First, we create X:

*hnew.new* = *Xn*r *ew* (*X*r*X*)−1*Xnew*

X <- **rep**(**c**(1))

X <- **cbind**(X, **data.matrix**(commercial\_df[2**:**5], rownames.force = NA))

**head**(X)

## X Age Operating\_Costs Vacancy\_Rates Total\_Sqft

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## [1,] 1 | 1 | 5.02 | 0.14 | 123000 |
| ## [2,] 1 | 14 | 8.19 | 0.27 | 104079 |
| ## [3,] 1 | 16 | 3.00 | 0.00 | 39998 |
| ## [4,] 1 | 4 | 10.70 | 0.05 | 57112 |
| ## [5,] 1 | 11 | 8.97 | 0.07 | 60000 |
| ## [6,] 1 | 15 | 9.45 | 0.24 | 101385 |

Below, I am including the results for (*X*r*X*)−1:



X\_X <- **solve**(**t**(X) **\*** X) X\_X

## X Age Operating\_Costs Vacancy\_Rates ## X 2.584382e-01 -3.048114e-04 -2.510574e-02 -2.508322e-01 ## Age -3.048114e-04 3.524211e-04 -2.094253e-04 3.144336e-03 ## Operating\_Costs -2.510574e-02 -2.094253e-04 3.087599e-03 2.191073e-02 ## Vacancy\_Rates -2.508322e-01 3.144336e-03 2.191073e-02 9.138530e-01 ## Total\_Sqft 1.235553e-07 -4.310481e-09 -3.072159e-08 -3.746468e-07

## Total\_Sqft

## X 1.235553e-07

## Age -4.310481e-09

## Operating\_Costs -3.072159e-08 ## Vacancy\_Rates -3.746468e-07 ## Total\_Sqft 1.483630e-12

In this case *Xn*r *ew* = 1*,* 10*,* 12*.*00*,* 0*.*05*,* 350000:



X\_new <- **c**(1, 10, 12, 0.05, 350000)

x\_X\_X <- **t**(X\_new) **\*** X\_X H\_extrapolate <- x\_X\_X **\*** X\_new H\_extrapolate

## [,1]

## [1,] 0.05292296

We see *hnew.new* = 0*.*05, approximately.

We can inspect the range of the leverage values *hii*:

**range**(hii)

## [1] 0.02419885 0.30367144

From the above result, we see the value of *hnew.new* is well within the range of the leverage values *hii* for the cases in the data set, so no hidden extrapolation is involved for this estimate.

#### Cases 61, 8, 3, and 53 appear to be outlying X observations, and cases 6 and 62 appear to be outlying Xobservations. Obtain the DFFITS, DFBETAS, and Cook’s distance values for each case to assess its influence. What do you conclude?

*Answer:*

The course Chapter 10 lecture slides illustrate the use of the influence.measures() functions to obtain the values of the metrics.

influence\_results <- **influence.measures**(linear\_fit)

For Case 61:

influence\_results**$**infmat[61,]

## dfb.1\_ dfb.Age dfb.Op\_C dfb.Vc\_R dfb.Tt\_S ## -0.055415285 0.024248530 -0.007608429 0.545712701 0.003819789

## dffit cov.r cook.d hat ## 0.638720760 1.442215560 0.081662174 0.303671442

DFBETAS *b*0 = -0.0554153. *b*1 = 0.0242485. *b*2 = -0.0076084. *b*3 = 0.5457127. *b*4 = 0.0038198.

Since the dataset is small.medium, if the absolute value of DFBETAS exceeds 1, we can consider the case as influential. In this case, the DEFBETAS metric suggests the case is not influential for all betas.

DFFITS = 0.6387208.

Since the dataset is small.medium, if the absolute value of DFFITS exceeds 1, we can consider the case as influential. In this case, the DFFITS metric suggests the case is not influential.

Cook’s Distance = 0.0816622.

The Cook’s Distance value is small, suggesting this is not an influential case. Conclusion: Not Influential.

For Case 8:

influence\_results**$**infmat[8,]

## dfb.1\_ dfb.Age dfb.Op\_C dfb.Vc\_R dfb.Tt\_S ## -0.014210232 -0.007197892 0.003014250 0.095519272 0.012599064

## dffit cov.r cook.d hat ## 0.116413642 1.334480430 0.002744609 0.202185915

DFBETAS *b*0 = -0.0142102. *b*1 = -0.0071979. *b*2 = 0.0030143. *b*3 = 0.0955193. *b*4 = 0.0125991.

DFFITS = 0.1164136.

Cook’s Distance = 0.0027446.

In this case, the DEFBETAS metric suggests the case is not influential for all betas. The DFFITS metric suggests the case is not influential.

The Cook’s Distance value is small, suggesting this is not an influential case. Conclusion: Not Influential.

For Case 3:

influence\_results**$**infmat[3,]

## dfb.1\_ dfb.Age dfb.Op\_C dfb.Vc\_R dfb.Tt\_S dffit ## -0.23178572 -0.15532832 0.23641364 0.10078041 -0.01149395 -0.28428045

## cov.r cook.d hat ## 1.29873472 0.01630620 0.19532064

DFBETAS *b*0 = -0.2317857. *b*1 = -0.1553283. *b*2 = 0.2364136. *b*3 = 0.1007804. *b*4 = -0.0114939.

DFFITS = -0.2842805.

Cook’s Distance = 0.0163062.

In this case, the DEFBETAS metric suggests the case is not influential for all betas.

The DFFITS metric suggests the case is not influential.

The Cook’s Distance value is small, suggesting this is not an influential case. Conclusion: Not Influential.

For Case 53:

influence\_results**$**infmat[53,]

## dfb.1\_ dfb.Age dfb.Op\_C dfb.Vc\_R dfb.Tt\_S dffit ## -0.01962803 -0.02398353 -0.02434044 0.41796384 0.04896786 0.52522645

## cov.r cook.d hat ## 1.19741531 0.05498189 0.17919908

DFBETAS *b*0 = -0.019628. *b*1 = -0.0239835. *b*2 = -0.0243404. *b*3 = 0.4179638. *b*4 = 0.0489679.

DFFITS = 0.5252264.

Cook’s Distance = 0.0549819.

In this case, the DEFBETAS metric suggests the case is not influential for all betas. The DFFITS metric suggests the case is not influential.

The Cook’s Distance value is small, suggesting this is not an influential case. Conclusion: Not Influential.

For Case 6:

influence\_results**$**infmat[6,]

## dfb.1\_ dfb.Age dfb.Op\_C dfb.Vc\_R dfb.Tt\_S dffit ## 0.1951155 -0.5648515 -0.1767223 -0.6171914 0.4481729 -0.8735488

## cov.r cook.d hat ## 0.6384837 0.1373665 0.0748058

DFBETAS *b*0 = 0.1951155. *b*1 = -0.5648515. *b*2 = -0.1767223. *b*3 = -0.6171914. *b*4 = 0.4481729.

DFFITS = -0.8735488.

Cook’s Distance = 0.1373665.

In this case, the DEFBETAS metric suggests the case is not influential for all betas. The DFFITS metric suggests the case is not influential.

The Cook’s Distance value is small, suggesting this is not an influential case.

Conclusion: Not Influential. However, the values of DFBETAS, DFFITS, and Cook’s Distance are larger compared to the other cases, sugessting it might be slightly influential.

For Case 62:

influence\_results**$**infmat[62,]

## dfb.1\_ dfb.Age dfb.Op\_C dfb.Vc\_R dfb.Tt\_S dffit ## 0.27581469 -0.33349618 -0.25947037 0.06272880 0.40507814 0.69033187

## cov.r cook.d hat ## 0.69360904 0.08753589 0.05792213

DFBETAS *b*0 = 0.2758147. *b*1 = -0.3334962. *b*2 = -0.2594704. *b*3 = 0.0627288. *b*4 = 0.4050781.

DFFITS = 0.6903319.

Cook’s Distance = 0.0875359.

In this case, the DEFBETAS metric suggests the case is not influential for all betas. The DFFITS metric suggests the case is not influential.

The Cook’s Distance value is small, suggesting this is not an influential case. Conclusion: Not Influential.

My final conclusion is that it seems that the metric sugges the given cases are not influential (we may consider, however, case 6 to be influential).

#### Calculate the average absolute percent difference in the fttted values with and without each of the cases. What does this measure indicate about the influence of each case’?

*Answer:*

The measure for average absolute percent difference is:

( (((*Y*ˆ*i*(*k*) − *Y*ˆ*i*)*/Y*ˆ*i*))100)*/n*

Σ

where *Y*ˆ*i*(*k*) are fitted value obtained when case k is omitted.

**for** (i **in c**(61, 8, 3, 53, 6, 62)) {

new\_df <- commercial\_df[**-c**(i), ] new\_linear\_fit <- **lm**(Rental\_Rates**~**.,new\_df)

newer\_result <- **mean**(**abs**((new\_linear\_fit**$**fitted.values **-** linear\_fit**$**fitted.values)**/**linear\_fit**$**fitted newer\_new\_result <- (100**\***newer\_result)**/nrow**(commercial\_df)

**print**(newer\_new\_result)

}

## Warning in new\_linear\_fit$fitted.values - linear\_fit$fitted.values: longer ## object length is not a multiple of shorter object length

## [1] 0.03393956

## Warning in new\_linear\_fit$fitted.values - linear\_fit$fitted.values: longer ## object length is not a multiple of shorter object length

## [1] 0.07969013

## Warning in new\_linear\_fit$fitted.values - linear\_fit$fitted.values: longer ## object length is not a multiple of shorter object length

## [1] 0.0890982

## Warning in new\_linear\_fit$fitted.values - linear\_fit$fitted.values: longer ## object length is not a multiple of shorter object length

## [1] 0.03845253

## Warning in new\_linear\_fit$fitted.values - linear\_fit$fitted.values: longer ## object length is not a multiple of shorter object length

## [1] 0.08266518

## Warning in new\_linear\_fit$fitted.values - linear\_fit$fitted.values: longer ## object length is not a multiple of shorter object length

## [1] 0.03420592

From the results above, we see that the measure indicates that cases 6 and 62 have a higher/larger influence on the fitted regression function in the range of X observations in a direct fashion.

#### Calculate Cook’s distance D; for each case and prepare an index plot. Are any cases influential according to this measure?

*Answer:*

From one of the previous-subproblems, we have Cook’s Distance = 0.0816622.

Cook’s Distance = 0.0027446. Cook’s Distance = 0.0163062. Cook’s Distance = 0.0549819. Cook’s Distance = 0.1373665. Cook’s Distance = 0.0875359.

influence\_results**$**infmat[61,][8]

## cook.d ## 0.08166217

influence\_results**$**infmat[8,][8]

## cook.d ## 0.002744609

influence\_results**$**infmat[3,][8]

## cook.d ## 0.0163062

influence\_results**$**infmat[53][8]

## [1] NA

influence\_results**$**infmat[6,][8]

## cook.d ## 0.1373665

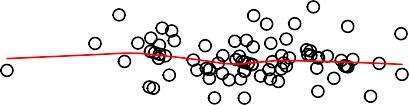
influence\_results**$**infmat[62,][8]

## cook.d ## 0.08753589

Below, I plot the diagnostics plot and the index plot:

**par**(mfrow=**c**(2,2)) **plot**(linear\_fit)

### Residuals vs Fitted Normal Q−Q



42

62

6



42 62

6

Residuals

0

3

Standardized residuals

0 2

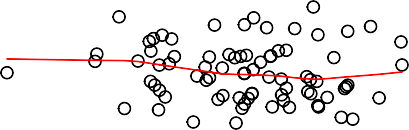
11 12 13 14 15 16 17 18 −2 −1 0 1 2

−3

−3

Fitted values Theoretical Quantiles

### Scale−Location



42

6

62

Standardized residuals

0 2

Residuals vs Leverage

1

0.5

1.0

11 12 13 14 15 16 17 18

0.0

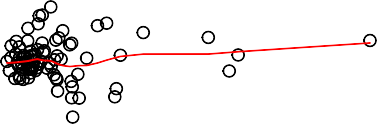
0.00 0.10 0.20

−3

0.30

0.5

1



62

Co8o0k's distance

6

Fitted values Leverage

Standardized residuals

**influenceIndexPlot**(linear\_fit, vars=**c**("Cook"))

## Diagnostic Plots

0.08

0.12

### 0 20 40 60 80



2

6

6

Cook's distance

0.00

0.04

Index

From the index plot, we see cases 6 (Cook’s distance value: 0.1373665) and 62 (Cook’s distance value: 0.08753589) seem to have relatively higher Cook’s distance points compared to the other cases, suggesting cases 6 and 62 are influential according to this measure.

### **Problem 2: 10.13**

Cosmetics sales. An assistant in the district sales office of a national cosmetics firm obtained data, shown below, on advertising expenditures and sales last year in the district’s 44 territories. X1 denotes expenditures for point-of-sale displays in beauty salons anti department stores (in thousand dollars), and X2 and X3 represent the corresponding expenditures for local media advertising and prorated ~hare of national media adveL1ising, respectively. Y denote~ sales (in thousand cases). The assistant was instructed to estimate the increase in expected sales when X1 is increased by 1 thousand dollars and X2 and Xl are held constant, anti was told to use an ordinary multiple regression model with linear terms for the predictor variables and with independent normal error terms.

#### State the regression model to be employed and ftt it to the data.

*Answer:*

First, I read in the dataset:

cosmetic\_data <- **read.table**(**url**(["http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatas](http://users.stat.ufl.edu/%7Errandles/sta4210/Rclassnotes/data/textdatas) col.names=**c**("Yi", "Xi1", "Xi2", "Xi3"))

**head**(cosmetic\_data)

## Yi Xi1 Xi2 Xi3 ## 1 12.85 5.6 5.6 3.8

## 2 11.55 4.1 4.8 4.8

## 3 12.78 3.7 3.5 3.6

## 4 11.19 4.8 4.5 5.2

## 5 9.00 3.4 3.7 2.9

## 6 9.34 6.1 5.8 3.4

The regression model to be employed is the following (since there is no need for interaction terms):

*Yi*r = *β*0 + *β*1*Xi*1 + *β*2*Xi*2 + *β*3*Xi*3 + *ei*

Below, I fit this regression model:

cosmetic\_fit <- **lm**(Yi**~**Xi1**+**Xi2**+**Xi3, data=cosmetic\_data) cosmetic\_fit

##

## Call:

## lm(formula = Yi ~ Xi1 + Xi2 + Xi3, data = cosmetic\_data) ##

## Coefficients:

|  |  |  |  |
| --- | --- | --- | --- |
| ## (Intercept) | Xi1 | Xi2 | Xi3 |
| ## 1.0233 | 0.9657 | 0.6292 | 0.6760 |

We see the answer is:

*Y*ˆ = 1*.*0233 + 0*.*9657*X*1 + 0*.*6292*X*2 + 0*.*6760*X*3

#### Test whether there is a regression relation between sales and the three predictor variables; use a = .05. State the alternatives, decision rule, and conclusion.

*Answer* :

To test whether whether there is a regression relation between sales and the three predictor variables, we need to use an overall F test; here, we are testing the alternatives:

*H*0 : *β*1 = *β*2 = *β*3 = 0

*Ha* : not all *βk*(*k* = 1*,* 2*,* 3) equal zero. In this case, the test statistic here is:

*F* ∗ = *MSR/MSE*

If *H*0 holds, *F* ∗ ~ *F* (*p* − 1*, n* − *p*). Large values of F\* lead to conclusion *Ha*. From the below results from anova(), we see the F\* is 38.28.

**summary**(cosmetic\_fit)

##

## Call:

## lm(formula = Yi ~ Xi1 + Xi2 + Xi3, data = cosmetic\_data) ##

## Residuals:

## Min 1Q Median 3Q Max ## -5.4217 -0.9115 0.0703 1.1420 3.5479 ##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|) ## (Intercept) 1.0233 1.2029 0.851 0.4000

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## | Xi1 | 0.9657 | 0.7092 | 1.362 | 0.1809 |
| ## | Xi2 | 0.6292 | 0.7783 | 0.808 | 0.4237 |
| ## | Xi3 | 0.6760 | 0.3557 | 1.900 | 0.0646 . |
| ## | --- |  |  |  |  |

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 ##

## Residual standard error: 1.825 on 40 degrees of freedom ## Multiple R-squared: 0.7417, Adjusted R-squared: 0.7223 ## F-statistic: 38.28 on 3 and 40 DF, p-value: 7.821e-12

For a = 0.05, we require F(0.95;3,40).

F\_statistic <- **qf**(0.95, 3, 40) *#Since the significance is 0.05.*

F\_statistic

## [1] 2.838745

We can see F(0.95;3,40) = 2.84, approximately.

If *F* ∗ ≤ *F statistic*, conclude *H*0, else conclude *Ha*.

Since *F* ∗ *>* 2*.*838745, we conclude *Ha*: Not all *βk* equal zero.

1. Test for each of the regression coetficients *βk*(*k* = 1*,* 2*,* 3) individually whether or not *βk* = 0; use a =

.05 each time. Do the conclusions ofthese tests correspond to that obtained in part (b)?

*Answer:*

To test whether *Xk* can be dropped from the multiple regression model:

*H*0 : *βk* = 0

*Ha* : *βk* ƒ= 0

In this case, the test statistics are:

*F* ∗ = *SSR*(*X*1|*X*2*, X*3)*/*1*/SSE*(*X*1*, X*2*, X*3)*/n* − 4

*F* ∗ = *SSR*(*X*2|*X*1*, X*3)*/*1*/SSE*(*X*1*, X*2*, X*3)*/n* − 4

*F* ∗ = *SSR*(*X*3|*X*1*, X*2)*/*1*/SSE*(*X*1*, X*2*, X*3)*/n* − 4

IF *H*0 holds, *F* ∗ ~ *F* (1*, n* − *p*).

We have *F* (0*.*95; 1*, n* − *p*), where n is 44 and p is 4.

**nrow**(cosmetic\_data)

## [1] 44

F\_statistic <- **qf**(0.95, 1, 40) *#Since the significance is 0.01.*

F\_statistic

## [1] 4.084746

We can see the F\_statistic is 4.0847457.

If *F* ∗ ≤ *F statistic*, conclude *H*0, else conclude *Ha*.

As taught in the course lectures, we can fit the models and use the anova() function for comparing between the reduced and the original model.

For X1:

X1\_cosmetic\_fit <- **lm**(Yi**~**Xi2**+**Xi3, data=cosmetic\_data)

**anova**(cosmetic\_fit, X1\_cosmetic\_fit)

## Analysis of Variance Table ##

## Model 1: Yi ~ Xi1 + Xi2 + Xi3 ## Model 2: Yi ~ Xi2 + Xi3

## Res.Df RSS Df Sum of Sq F Pr(>F)

|  |  |  |  |
| --- | --- | --- | --- |
| ## | 1 | 40 | 133.29 |
| ## | 2 | 41 | 139.46 -1 |

-6.1778 1.854 0.1809

We see F\* = 1.854.

Since *F* ∗ *<* 4*.*084746, we conclude *H*0: *B*1 = 0. For X2:

X2\_cosmetic\_fit <- **lm**(Yi**~**Xi1**+**Xi3, data=cosmetic\_data)

**anova**(cosmetic\_fit, X2\_cosmetic\_fit)

## Analysis of Variance Table ##

## Model 1: Yi ~ Xi1 + Xi2 + Xi3 ## Model 2: Yi ~ Xi1 + Xi3

## Res.Df RSS Df Sum of Sq F Pr(>F)

|  |  |  |  |
| --- | --- | --- | --- |
| ## | 1 | 40 | 133.29 |
| ## | 2 | 41 | 135.46 -1 |

-2.1775 0.6535 0.4237

We see F\* = 0.6535.

Since *F* ∗ *<* 4*.*084746, we conclude *H*0: *B*2 = 0. For X3:

X3\_cosmetic\_fit <- **lm**(Yi**~**Xi1**+**Xi2, data=cosmetic\_data)

**anova**(cosmetic\_fit, X3\_cosmetic\_fit)

## Analysis of Variance Table ##

## Model 1: Yi ~ Xi1 + Xi2 + Xi3 ## Model 2: Yi ~ Xi1 + Xi2

## Res.Df RSS Df Sum of Sq F Pr(>F) ## 1 40 133.29

## 2 41 145.32 -1 -12.033 3.6113 0.06461 . ## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We see F\* = 3.6113.

Since *F* ∗ *<* 4*.*084746, we conclude *H*0: *B*3 = 0.

The conclusions from these tests (*H*0) *do not* match with the conclusion from part\_b (*Ha*), suggesting that potentially the regression relation seems to exist only when all predictors are present.

1. Obtain the correlation matrix of the X variables.

*Answer:*

Below, I am obtaining the correlation matrix of the X variables (excluding the outcome).

**cor**(cosmetic\_data[2**:**4])

## Xi1 Xi2 Xi3 ## Xi1 1.0000000 0.9744313 0.3759509

## Xi2 0.9744313 1.0000000 0.4099208

## Xi3 0.3759509 0.4099208 1.0000000

As an interpretation, we see Xi1 and Xi2 have a significantly high correlation value, while Xi1 and Xi3 and Xi2 and Xi3 have moderately (but not significantly) high correlation values.

#### What do the results in parts (b), (c), and (d) suggest about the suitability of the data for the research objective?

*Answer:*

cosmetic\_fit

##

## Call:

## lm(formula = Yi ~ Xi1 + Xi2 + Xi3, data = cosmetic\_data) ##

## Coefficients:

|  |  |  |  |
| --- | --- | --- | --- |
| ## (Intercept) | Xi1 | Xi2 | Xi3 |
| ## 1.0233 | 0.9657 | 0.6292 | 0.6760 |

Depending on the use case, the data may/may not be suitable for the research objective. We note from part b that there is a regression relation between the outcome and the three predictor variables, but when testing individually for each of the regression coefficients, we saw *Bk* = 0, so this could perhaps further data engineering (e.g. interaction terms) might better help find a regression coefficient that is not *Bk* = 0.

As the textbook notes, holding *X*2 and *X*3 constant (assuming that these predictors have the same units as X1) might not be fully applicable in this case since multicollinearity exists (specifically, high correlation values between Xi and Xi2).

Therefore, additional modifications might need to be made in order to make the data more suitable for this research objective.

## **Problem 3: 10.18**

Refer to Commercial properties Problem 6.18b.

commercial\_df <- **read.table**(**url**(["http://people.stat.sc.edu/Hitchcock/commercialproperties.txt"](http://people.stat.sc.edu/Hitchcock/commercialproperties.txt)), header

**head**(commercial\_df)

## Rental\_Rates Age Operating\_Costs Vacancy\_Rates Total\_Sqft

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## 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 |

**summary**(commercial\_df) ## Rental\_Rates Age Operating\_Costs Vacancy\_Rates

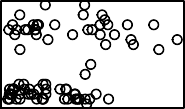
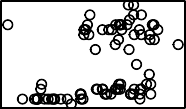
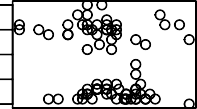
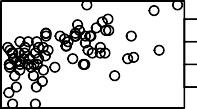
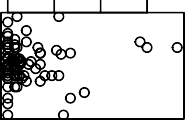
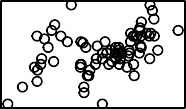
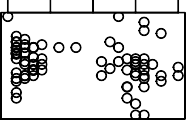
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## | Min. :10.50 | Min. : 0.000 | Min. : 3.000 | Min. :0.00000 |
| ## | 1st Qu.:14.00 | 1st Qu.: 2.000 | 1st Qu.: 8.130 | 1st Qu.:0.00000 |
| ## | Median :15.00 | Median : 4.000 | Median :10.360 | Median :0.03000 |
| ## | Mean :15.14 | Mean : 7.864 | 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 |
| ## | Total\_Sqft |  | | |
| ## | Min. : 27000 |
| ## | 1st Qu.: 70000 |
| ## | Median :129614 |
| ## | Mean :160633 |
| ## | 3rd Qu.:236000 |
| ## | Max. :484290 |

#### What do the scatter plot matrix and the correlation matrix show about pairwise linear associations among the predictor variables?

*Answer:*

Below, I plot the scatter plot matrix and print the correlation matrix:

**pairs**(commercial\_df)

0 5 10 15 20 0.0 0.4

Rental\_Rates

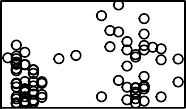
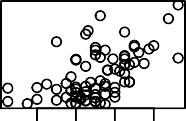
Age

0

10

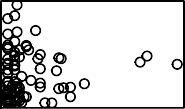
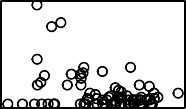
20

12 16

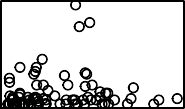
12 16

0.0

0.4

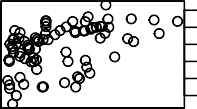
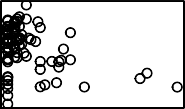
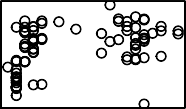
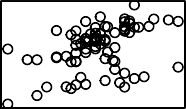
4 8 12

Vacancy\_Rates

1e+05 4e+05

Total\_Sqft

1e+05 5e+05

 **cor**(commercial\_df) ## Rental\_Rates Age Operating\_Costs Vacancy\_Rates

Operating\_Costs

4 8

14

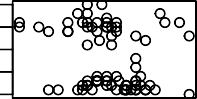
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ## | Rental\_Rates | 1.00000000 | -0.2502846 | 0.4137872 | 0.06652647 |
| ## | Age | -0.25028456 | 1.0000000 | 0.3888264 | -0.25266347 |
| ## | Operating\_Costs | 0.41378716 | 0.3888264 | 1.0000000 | -0.37976174 |
| ## | Vacancy\_Rates | 0.06652647 | -0.2526635 | -0.3797617 | 1.00000000 |
| ## | Total\_Sqft | 0.53526237 | 0.2885835 | 0.4406971 | 0.08061073 |
| ## | Rental\_Rates | 0.53526237 | | | |
| ## | Age | 0.28858350 | | | |
| ## | Operating\_Costs | 0.44069713 | | | |
| ## | Vacancy\_Rates | 0.08061073 | | | |
| ## | Total\_Sqft | 1.00000000 | | | |

## Total\_Sqft

We can use the pairs() function to plot the scatter plots and the confidence values in one plot:

**pairs**(commercial\_df,upper.panel=**function**(x,y)**legend**("topleft",**paste**("",**signif**(**cor**(x,y),2)),bty="n"), ma

# Scatter Plot Matrix

0 5 10 15 20 0.0 0.4

−0.25

0.067

0.54

0.29

−0.25

0.39

Age

0.41

Rental\_Rates

0

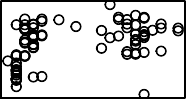
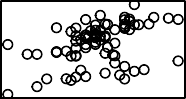
10

20

4 8 14

12

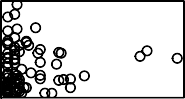
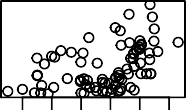
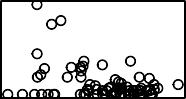
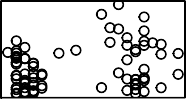
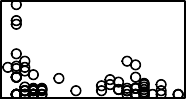
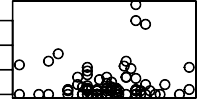
18



Operating\_Costs

−0.38

0.44

12 16

Total\_Sqft

0.081

Vacancy\_Rates

0.0

0.4

1e+05 5e+05

4 8 12

1e+05 4e+05

In terms of correlation with the outcome - we see that the Total\_Sqft and the Operating\_Costs variables have high positive correlation values with the outcome Rental\_Rates. In addition, we see Age has a negative correlation with Rental\_Rates, and Vacancy\_Rates has a weak positive correlation with Rental\_Rates.

In terms of pairwise associations, we see moderately high correlations (both positive and negative) between Age and the other variables (e.g. Age and Operating Costs), and Operating\_Costs and the other variables (e.g. the relatively high correlation value between Vacancy\_Rates and Operating\_Costs).

The pairwise correlation between Vacancy\_Rates and Total\_Sqft is comparatively weaker.

1. *Obtain the four variance inflation factors. Do they indicate that a serious multicollinearity problem exists here?*

*Answer:*

In lecture, the use of the vif() function from the car library was illustrated for the variance inflation factor, which is what I use below:

commercial\_fit <- **lm**(Rental\_Rates**~**Age**+**Operating\_Costs**+**Vacancy\_Rates**+**Total\_Sqft, data=commercial\_df)

**vif**(commercial\_fit)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ## | Age | Operating\_Costs | Vacancy\_Rates | Total\_Sqft |
| ##  We see: | 1.240348 | 1.648225 | 1.323552 | 1.412722 |

$(VIF)\_1 = $ 1.240348

$(VIF)\_2 = $ 1.648225

$(VIF)\_3 = $ 1.323552

$(VIF)\_4 = $ 1.412722

We note that the (*V IF* )*k* is equal to 1 when *Xk* is not linearly related to the other X variables, and that (*V IF* )*k* is greater than 1 due to inflated variance as a result of the intercorrelations among the X variables.

For the results obtained above, we see they *do not indicate a serious multicollinearity problem*. We note (*V IF* )2 or the Operating\_Costs variable seems to have the highest variance inflation factor, suggesting it is more intercorrelated with the other variables; however, the value of the VIF is not significantly high.