Homework 5

Greg DePaul

2023-03-28

Problem 6 - A multiple linear regression case study by R.

You need to submit your codes alongside 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')
```

(a) Read data into R. What is the type of each variable? Draw plots to depict the distribution of each variable and obtain summary statistics for each variable. Comment on the distributions of these variables. We see that

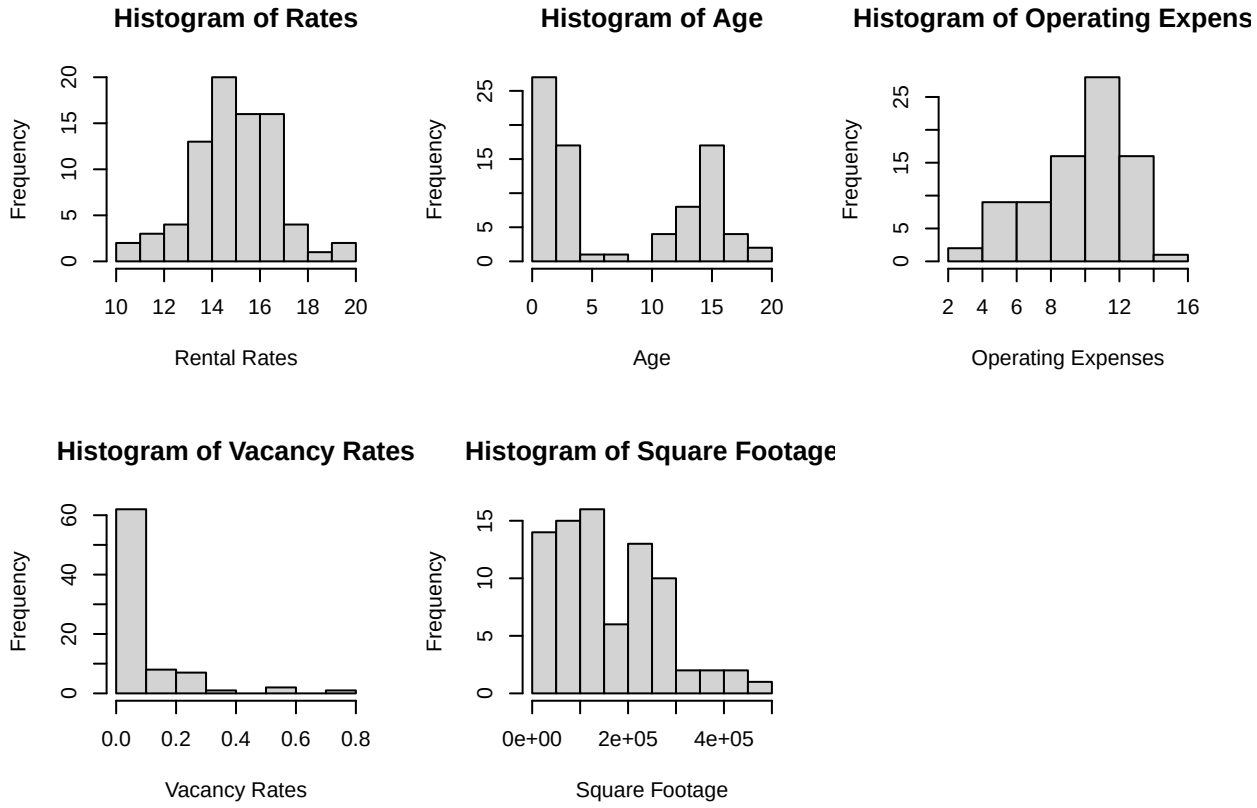
- rental_rates = FLOAT
- age = INT
- operating_expense = FLOAT
- vacancy_rates = FLOAT
- square_footage = INT

```
summary(my_data)
```

```
##   rental_rates        age         operating_expenses 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
##  square_footage
##  Min.   : 27000
##  1st Qu.: 70000
##  Median :129614
##  Mean   :160633
##  3rd Qu.:236000
##  Max.   :484290
```

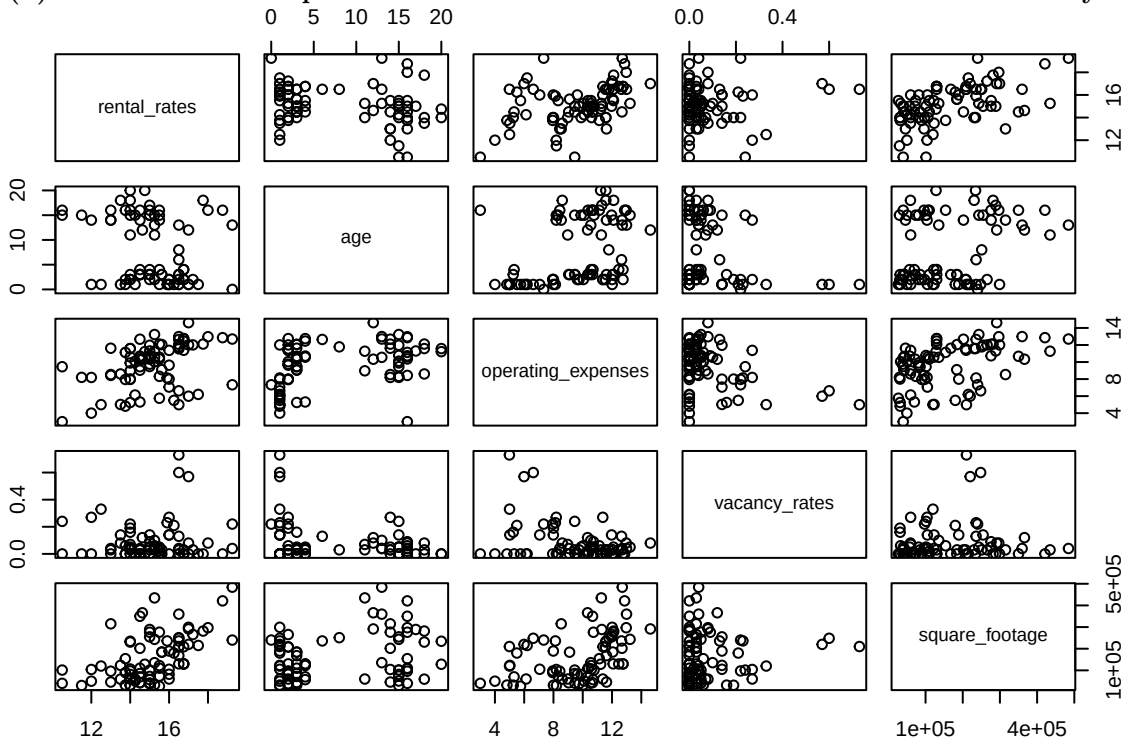
```r
par(mfrow = c(2, 3))
hist(my_data$rental_rates, xlab='Rental Rates', main='Histogram of Rates')
hist(my_data$age, xlab='Age', main='Histogram of Age')

hist(my_data$operating_expenses, xlab='Operating Expenses', main='Histogram of Operating Expenses')
hist(my_data$vacancy_rates, xlab='Vacancy Rates', main='Histogram of Vacancy Rates')
hist(my_data$square_footage, xlab='Square Footage', main='Histogram of Square Footage')
```




```r
pairs(my_data)
```

(b) Draw the scatter plot matrix and obtain the correlation matrix. What do you observe?




```
cor(my_data)
```

```
##                    rental_rates        age operating_expenses vacancy_rates
## rental_rates         1.00000000 -0.2502846          0.4137872    0.06652647
## age                 -0.25028456  1.0000000          0.3888264   -0.25266347
## operating_expenses   0.41378716  0.3888264          1.0000000   -0.37976174
## vacancy_rates        0.06652647 -0.2526635         -0.3797617    1.00000000
## square_footage       0.53526237  0.2885835          0.4406971    0.08061073
##                    square_footage
## rental_rates           0.53526237
## age                    0.28858350
## operating_expenses     0.44069713
## vacancy_rates          0.08061073
## square_footage         1.00000000
```

```
fit_1 = lm(rental_rates ~ age + operating_expenses + vacancy_rates + square_footage, data=my_data)
summary(fit_1)
```

(c) Perform regression of the rental rates Y on the four predictors X1, X2, X3, X4 (Model 1). What are the Least-squares estimators? Write down the fitted regression function. What are MSE, R2 and Ra2?

```
##
## Call:
## lm(formula = rental_rates ~ age + operating_expenses + vacancy_rates +
##     square_footage, 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 ***
## vacancy_rates        6.193e-01  1.087e+00   0.570     0.57
## square_footage       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
```

This gives us the regression:

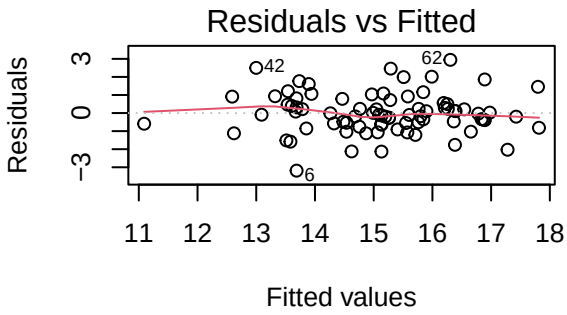
$$\hat{Y}_i = 12.20059 - 0.1420336X_1 + 0.2820165X_2 + 0.6193435X_3 + (7.924302e^{-06})X_4$$

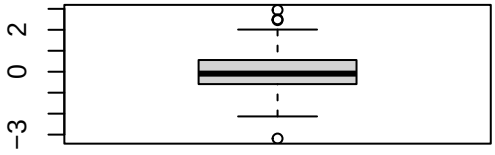
$$R^2 = 0.5847, \quad R_a^2 = 0.5629$$

$$\sqrt{MSE} = 1.137 \implies MSE = 1.293$$

```
par(mfrow = c(2, 2))
plot(fit_1,which=1) ##residuals vs. fitted values
plot(fit_1,which=2) ##residuals Q-Q plot
boxplot(fit_1$residuals) ## residuals boxplot
```

(d) Draw residuals vs. fitted values plot, residuals Normal Q-Q plot and residuals box- plot. Comment on the model assumptions based on these plots. (Hint: for a compact report, use par(mfrow) to create one multiple paneled plot).

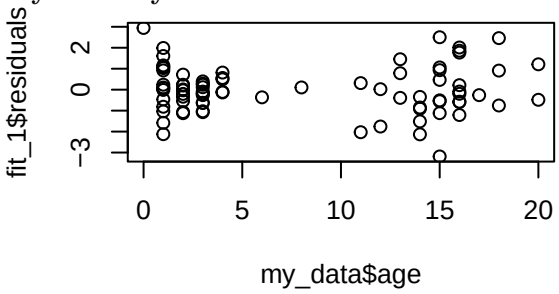


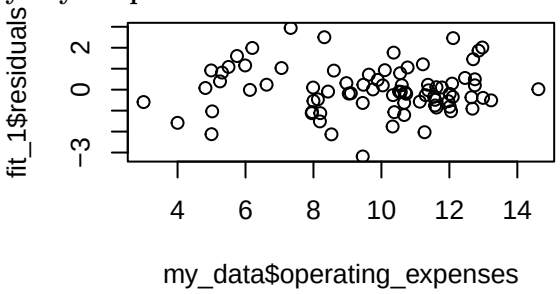


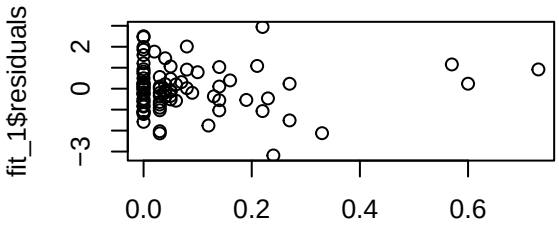
No obvious nonlinearity is scene.

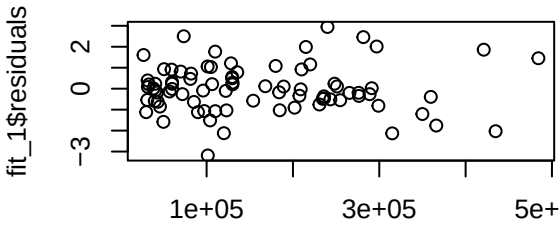
```
par(mfrow = c(2, 2))
plot(my_data$age,fit_1$residuals)
plot(my_data$operating_expenses,fit_1$residuals)
plot(my_data$vacancy_rates,fit_1$residuals)
plot(my_data$square_footage,fit_1$residuals)
```

(e) Draw residuals vs. each predictor variable plots, and residuals vs. each two-way interaction term plots. How many two-way interaction terms are there? Analyze your plots and summarize your findings.



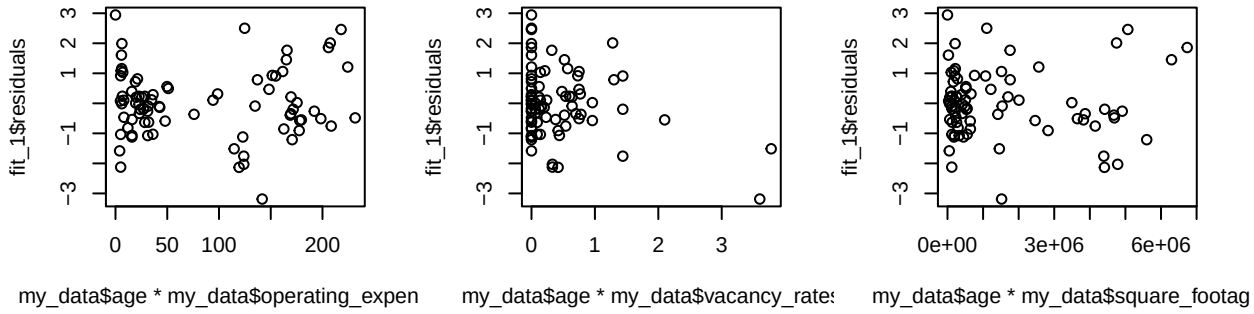


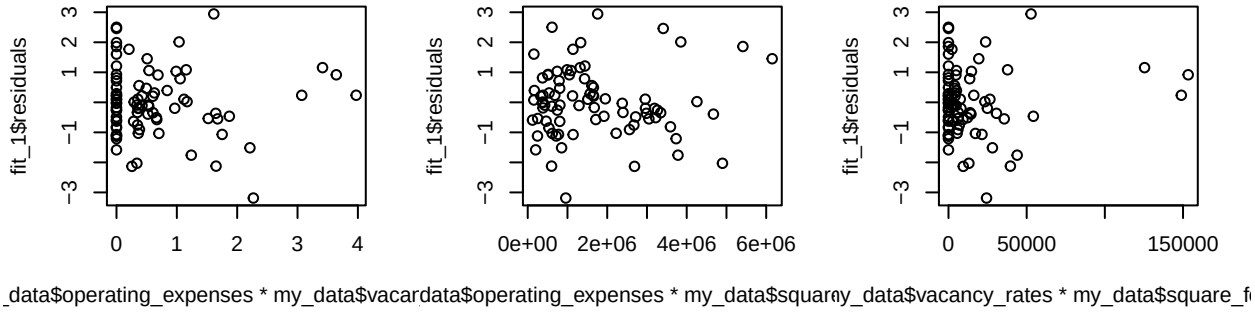




There should be $\binom{4}{2} = 6$ first order interation terms

```
par(mfrow = c(2, 3))
plot(my_data$age * my_data$operating_expenses,fit_1$residuals)
plot(my_data$age * my_data$vacancy_rates,fit_1$residuals)
plot(my_data$age * my_data$square_footage,fit_1$residuals)
plot(my_data$operating_expenses * my_data$vacancy_rates,fit_1$residuals)
plot(my_data$operating_expenses * my_data$square_footage,fit_1$residuals)
plot(my_data$vacancy_rates * my_data$square_footage,fit_1$residuals)
```





(f) For each regression coefficient, test whether it is zero or not (under the Normal error model) at level 0.01. State the null and alternative hypotheses, the test statistic, its null distribution and the pvalue. Which regression coefficient(s) is (are) significant, which is/are not? What is the implication?

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 \neq 0$
- Test Statistic: $T^* = \frac{\hat{\beta}_1 - 0}{s\{\hat{\beta}_1\}}$
- Null distribution of T^* is $t_{n-2} = t_{82}$

We use the two-sided t-test. The rule becomes is we reject H_0 if the following comparison is true:

$$|T^*| = |\frac{\hat{\beta}_1}{SE\{\hat{\beta}_1\}}| > t_{n-p}(1 - \frac{\alpha}{2})$$

```
n <- 81
p <- 5
alpha <- 0.01
crit_val  <- qt(1 - alpha / 2, df = n - p)

betas <- fit_1$coefficients
beta_1 <- betas[2]
s_beta_1 <- summary(fit_1)$coefficients["age","Std. Error"]
T_star <- beta_1 / s_beta_1
print( abs(T_star) > crit_val)
```

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

```r
beta_2 <- betas[3]
s_beta_2 <- summary(fit_1)$coefficients["operating_expenses","Std. Error"]
T_star <- beta_2 / s_beta_2
print( abs(T_star) > crit_val)
```

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

```r
beta_3 <- betas[4]
s_beta_3 <- summary(fit_1)$coefficients["vacancy_rates","Std. Error"]
T_star <- beta_3 / s_beta_3
print( abs(T_star) > crit_val)
```

```
## vacancy_rates
##         FALSE
```

```r
beta_4 <- betas[5]
s_beta_4 <- summary(fit_1)$coefficients["square_footage","Std. Error"]
T_star <- beta_4 / s_beta_4
print( abs(T_star) > crit_val)
```

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

```r
SSE <- sum((fitted(fit_1) - my_data$rental_rates)^2)
print(SSE)
```

(g) Obtain SSTO, SSR, SSE and their degrees of freedom. Test whether there is a regression relation at = 0.01. State the null and alternative hypotheses, the test statistic, its null distribution, the decision rule and your conclusion.

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

```r
SSR <- sum((fitted(fit_1) - mean(my_data$rental_rates))^2)
print(SSR)
```

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

```r
SSTO <- SSE + SSR
print(SSTO)
```

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

Suppose we wanted to conduct the F test to determine if there is a regression relation. Here is how we could set it up for $\alpha = 0.01$:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad \text{vs} \quad H_1 : \text{ not all of the } \beta\text{'s are } 0$$

We reject H_0 if $F^* > F(1-\alpha; p-1, n-p)$. Now, we can use R to find the critical value.

In this case, $n = 81, p = 5$, so we calculate:

```r
alpha <- 0.01
n <- 81
p <- 5
F_crit <- qf(1 - alpha, p - 1, n - p)
```

```
fit_2 = lm(rental_rates ~ age + operating_expenses + square_footage, data=my_data)
summary(fit_2)
```

(h) You now decide to fit a different model by regressing the rental rates Y on three predictors X1,X2,X4 (Model 2). Why would you make such a decision? Get the Least-squares estimators and write down the fitted regression function. What are MSE, R2 and Ra2? How do these numbers compare with those from Model 1?

```
##
## Call:
## lm(formula = rental_rates ~ age + operating_expenses + square_footage,
##     data = my_data)
##
## 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 ***
## age                -1.442e-01  2.092e-02  -6.891 1.33e-09 ***
## operating_expenses  2.672e-01  5.729e-02   4.663 1.29e-05 ***
## square_footage      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
```

This gives us the regression:

$$\hat{Y}_i = 12.37 - 0.1442X_1 + 0.2672X_2 + (8.178e^{-06})X_4$$

$$R^2 = 0.583, \quad R_a^2 = 0.5667$$

$$\sqrt{MSE} = 1.132 \implies MSE = 1.281$$

(i) Compare the standard errors of the regression coefficient estimates for X1,X2,X4 under Model 2 with those under Model 1. What do you find? Construct 95% confidence intervals for regression coefficients for X1,X2,X4 under Model 2. Had these intervals been constructed under Model 1, how would their widths compare with the widths of the intervals you just constructed, i.e., being wider or narrower? Justify your answer. The standard errors are smaller than those listed under model 1.

```
confint(fit_2,parm=c("age","operating_expenses", "square_footage"),level=.95)
```

```
##                              2.5 %        97.5 %
## age                -1.858219e-01 -1.025074e-01
## operating_expenses  1.530784e-01  3.812557e-01
## square_footage      5.578873e-06  1.077755e-05
```

We would expect these intervals to be narrower than model 1 simply because of the reduced standard errors.

```
newX = data.frame(age = 4, operating_expenses = 10, vacancy_rates = 0.1, square_footage = 80000)
predict(fit_1, newX, interval='confidence', level=0.99)
```

(j) Consider a property with the following characteristics: X1 = 4,X2 = 10,X3 = 0.1, X4 = 80, 000. Construct 99% prediction intervals under Model 1 and Model 2, respectively. Compare these two sets of intervals, what do you find?

```
##       fit      lwr      upr
## 1 15.1485 14.64413 15.65286
```

```
predict(fit_2, newX, interval='confidence', level=0.99)
```

```
##        fit      lwr      upr
## 1 15.11985 14.63558 15.60412
```

(k) Which of the two Models you would prefer and why? If we're attempting to use either of these models for prediction, it would be better to use a model where the confidence interval on new predictions would be smaller. Model 2 would allow us to bound our inferences better, so model 2.