STAT2170 Applied Statistics

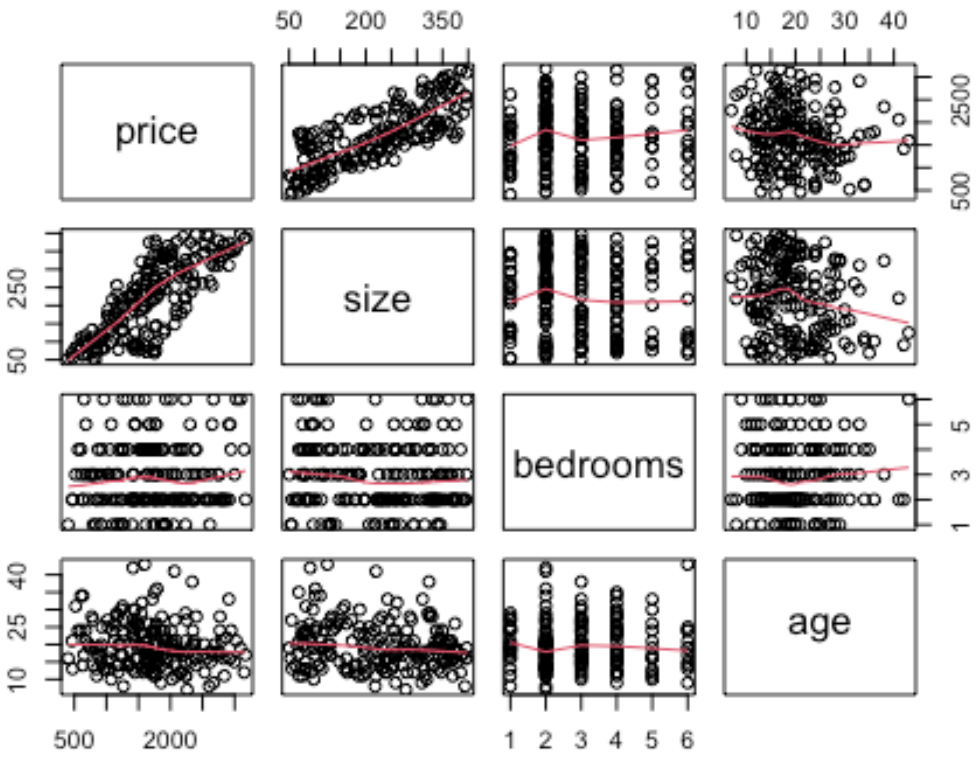
Assignment 2

2024-10-02

Question 1

a. Inputting the data and producing the scatterplot and correlation matrix:

```
realestate <- read.csv("realestate2024.csv", header = TRUE)
pairs(realestate, panel = panel.smooth)
```




```
cor(realestate)
```

```
##                price        size    bedrooms         age
## price     1.00000000  0.77994644  0.05560245 -0.12347514
## size      0.77994644  1.00000000 -0.07285563 -0.16695401
## bedrooms  0.05560245 -0.07285563  1.00000000  0.02850195
## age      -0.12347514 -0.16695401  0.02850195  1.00000000
```

- The response variable price has a moderate positive linear relationship with the predictor size. The response price has no obvious relationship with the predictor bedrooms and predictor age.
- There doesn't seem to be a relationship present between the predictors themselves.

b. Fit the full model:

```
r1 <- lm(price ~., data = realestate)
summary(r1)

##
## Call:
## lm(formula = price ~ ., data = realestate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -748.08 -318.57  -54.74  366.46  784.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 449.2955   133.7219   3.360  0.00094 ***
## size          4.9371     0.2819  17.514  < 2e-16 ***
## bedrooms     53.6872    21.1222   2.542  0.01182 *
## age           0.4821     4.3038   0.112  0.91092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 398.7 on 193 degrees of freedom
## Multiple R-squared:  0.621,  Adjusted R-squared:  0.6152
## F-statistic: 105.4 on 3 and 193 DF,  p-value: < 2.2e-16



summary.r1 <- summary (r1)
se <- sqrt(diag(summary.r1$cov.unscaled * summary.r1$sigma^2))
```

The required CI is

$$\hat{\beta}_{\text{size}} \pm t_{n-p,1-\alpha/2} s.e.\left(\hat{\beta}_{\text{size}}\right)$$

$$= \hat{\beta}_{\text{size}} \pm t_{193,0.975} s.e.\left(\hat{\beta}_{\text{size}}\right)$$

$$= 4.9371 \pm 1.972332 \times 0.2819$$

$$= (4.3811, 5.4931)$$

That is, we are 95% confident that for every extra square meter of property size, the price quantifies will increase between \$4.3811 and \$5.4931 on average.

c.

- Theoretical Model is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1,2,\dots,n$$

- Y_i is the response variable price
- X_{ij} are the predictors variables for the i-th observation
  - X_{i1} = annual mean size of the test locations
  - X_{i2} = annual mean bedrooms of the test locations
  - X_{i3} = annual mean age of the test locations
- $\epsilon_i \sim N(0, \sigma^2)$ denotes the random variation with constant variance;

Conducting the F-test we have,

- Hypotheses: H_0: $\beta_1 = \cdots = \beta_4 = 0$ vs H_1:not all $\beta_i = 0$; $i = 1,2,3$.
- Standard R output ANOVA table

```
anova(r1)

## Analysis of Variance Table
##
## Response: price
##            Df   Sum Sq  Mean Sq  F value  Pr(>F)
## size        1 49256631 49256631 309.8153 < 2e-16 ***
## bedrooms    1  1028915  1028915   6.4717 0.01174 *
## age         1     1995     1995   0.0125 0.91092
## Residuals 193 30684511   158987
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Can show the reduced Overall ANOVA table as

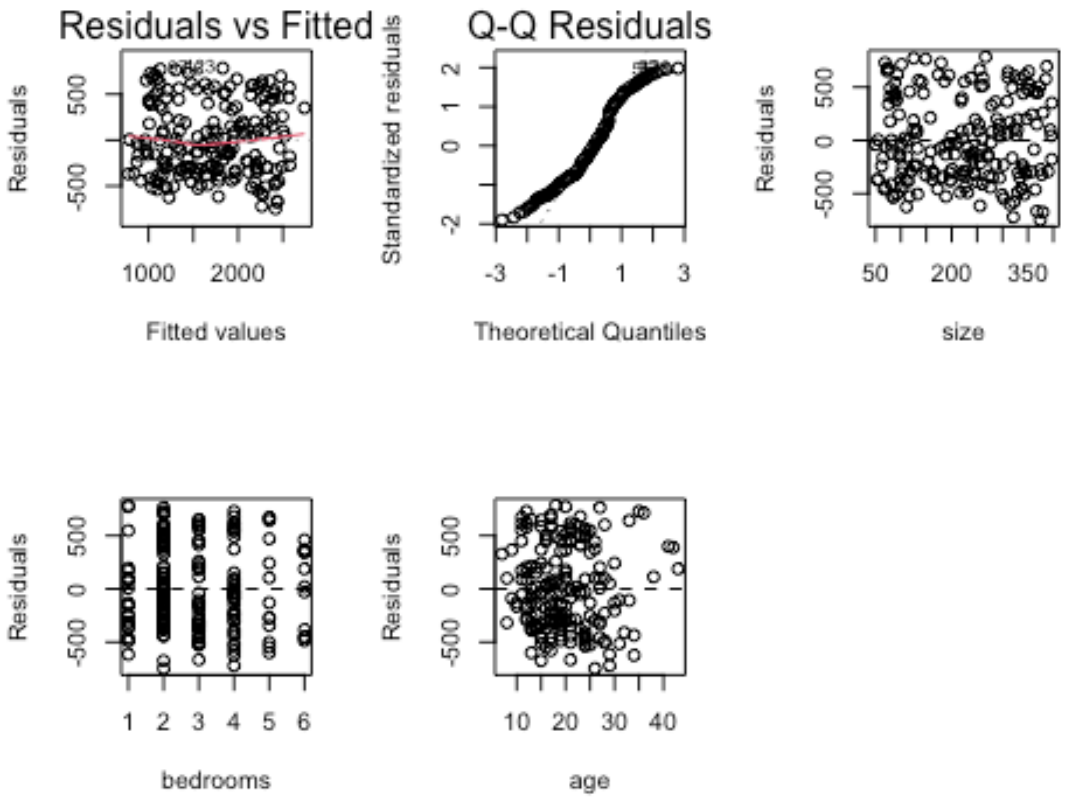
Df	Sum.Sq	Mean.Sq	F.value	Pr..F.
3	50287541	16762514	105.4332	0
193	30684511	158987	NA	NA

- Note the Regression SS = 49256631 + 1028915 + 1995 = 50287541
- Therefore the Mean Square Reg = Reg SS/Reg df = 50287541/3 = 16762514
- Test statistic: $F_{obs} = MS_{Reg}/MS_{Res}$ = 16762514/158987 = 105.4332;
- The null distribution for the test stattistics is $F_{3,193}$.
- P-value: P($F_{3,193} \geq 105.4332$) = 0 < 0.05
- As the P-value is small,
  - (Statistical) There is enough evidence to reject H_0.

- (Contextual) That is, there is a significant linear relationship between price and at least one of the 3 predictor variables.

d. For the diagnostic:

```
par(mfrow = c(2,3))
plot(r1, which = 1:2)
plot(resid(r1) ~ size, data = realestate, xlab = "size", ylab = "Residuals")
abline(h = 0, lty = 2)
plot(resid(r1) ~ bedrooms, data = realestate, xlab = "bedrooms", ylab =
"Residuals")
abline(h = 0, lty = 2)
plot(resid(r1) ~ age, data = realestate, xlab = "age", ylab = "Residuals")
abline(h = 0, lty = 2)
```



- The quantile plot of residuals look approximately linear, suggesting the normality assumption for residuals is appropriate.
- There is no obvious pattern in any of the residual plots so it appears the linearity and constant variance assumptions of the multiple linear model are justified.

e. Here R^2 = 0.621 = 62.1%, which is a goodness of fit metric. It means 61.2% of the variation in price is explained by the full linear regession model.

f. Starting with all the predictors

```
summary(r1)

##
## Call:
## lm(formula = price ~ ., data = realestate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -748.08 -318.57  -54.74  366.46  784.33
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 449.2955   133.7219   3.360  0.00094 ***
## size          4.9371     0.2819  17.514  < 2e-16 ***
## bedrooms     53.6872    21.1222   2.542  0.01182 *
## age           0.4821     4.3038   0.112  0.91092
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 398.7 on 193 degrees of freedom
## Multiple R-squared:  0.621,  Adjusted R-squared:  0.6152
## F-statistic: 105.4 on 3 and 193 DF,  p-value: < 2.2e-16
```

age has the highest P-value so we shall remove it first.

```
r2 <- update(r1, . ~. - age)
summary(r2)

##
## Call:
## lm(formula = price ~ size + bedrooms, data = realestate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -744.25 -321.86  -59.73  362.39  783.79
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 459.8715    94.4581   4.869 2.33e-06 ***
## size          4.9318     0.2773  17.785  < 2e-16 ***
## bedrooms     53.7265    21.0655   2.550   0.0115 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 397.7 on 194 degrees of freedom
## Multiple R-squared:  0.621,  Adjusted R-squared:  0.6171
## F-statistic:   159 on 2 and 194 DF,  p-value: < 2.2e-16
```

At this point, all remaining predictors are significant and should be kept in the model. The final (fitted) model equation is

$\widehat{Y}_i = 459.8715 + 4.9318X_{i1} + 53.7265X_{i2}$ or

$\widehat{price} = 459.8715 + 4.9318size + 53.7265bedrooms$

g. The R^2 goodness of fit metric always decreases/increases when a predictor is removed/added from/into the model. The adjusted R^2 has a penalty for the number of predictors in the model. So it will sometimes increase when a predictor is removed. In this case, from the full to final model, the R^2 stays at 62.1% but the adjusted R^2 increases from 61.52% to 61.71%. This indicates the final model is a better parsimonious model for the data.

Question 2

a. A study is balanced if there are equal number of replicates across all the levels factors in the study. Here we check the number of replicates with,

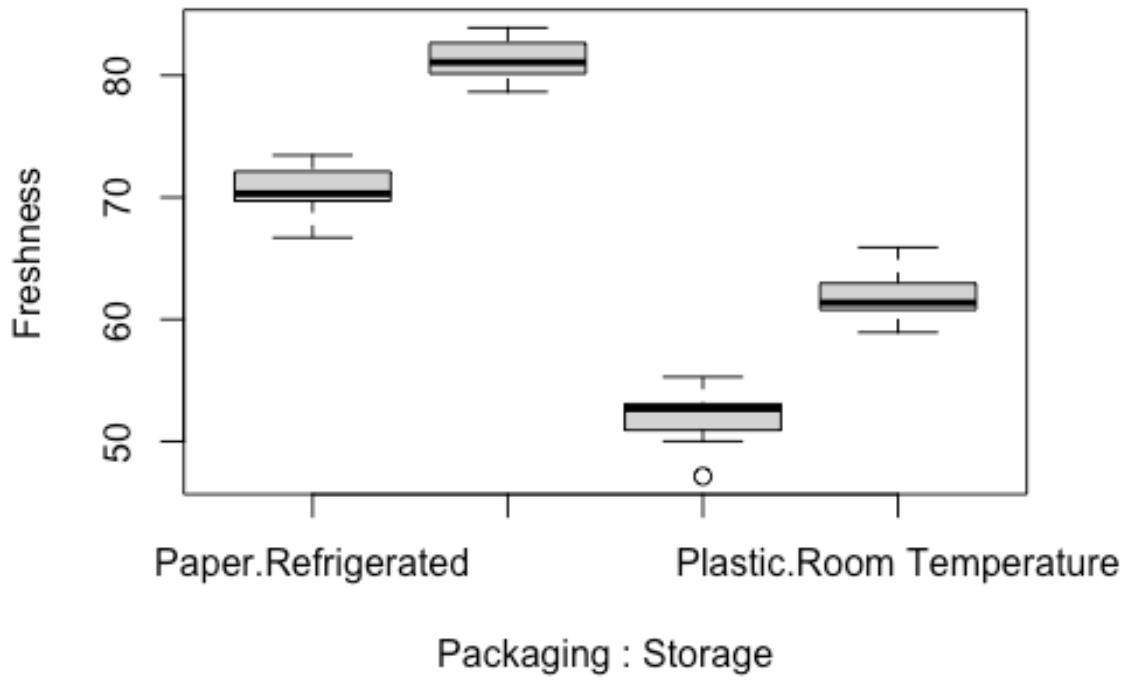
```
goods <- read.csv("goods.csv", header = TRUE, stringsAsFactors = TRUE)
table(goods[, c("Packaging", "Storage")])
```

```
##          Storage
## Packaging Refrigerated Room Temperature
##   Paper             14               17
##   Plastic           16               18
```

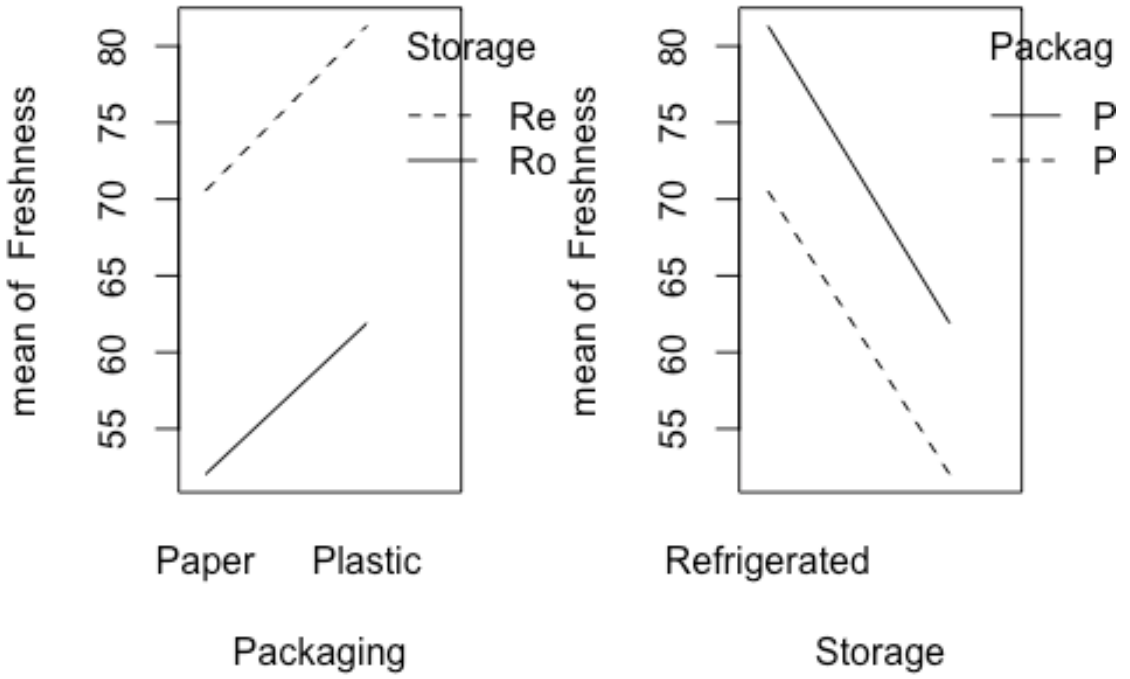
From the above table, we can see that the design is unbalanced with an unequal number of replicates for each combination of levels of the two factors.

b. Constructing the preliminary plots

```
boxplot(Freshness ~ Packaging + Storage, data = goods)
```




```
par(mfrow = c(1,2))
with(goods, interaction.plot(Packaging, Storage, Freshness))
with(goods, interaction.plot(Storage, Packaging, Freshness))
```



- From both interaction plots we can see the parallel lines for the means of each group at different levels of independent variables, this indicates an insignificant interaction effect between the two independent variables
- From the boxplot, we can see that the assumption of equal variance among levels seems approximately valid due to the similar box sizes. Optionally we can also compute the standard deviation for each group:

```
with(goods, aggregate(Freshness ~ Packaging + Storage, FUN = sd))
```

```
##   Packaging          Storage Freshness
## 1     Paper     Refrigerated  1.738885
## 2   Plastic     Refrigerated  1.582789
## 3     Paper Room Temperature  1.911079
## 4   Plastic Room Temperature  1.672555
```

The standard deviations are quite similar.

c. The full Two-Way ANOVA model with interation is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

where the paramaters are:

- Y_{ijk}: the brand recall score response;
- α_i: the Packaging effect, there are four levels - Paper and Plastic
- β_j: the Storage effect, there are two levels - Refrigerated and Room Temperature
- Y_{ij}: interaction effect between Packaging and Storage
- $\epsilon_{ijk} \sim N(0, \sigma^2)$ is the unexplained variation

d. We wish to first test

$$H_0: \gamma_{ij} = 0 \text{ for all } i, j \text{ against } H_1: \text{at least one } \gamma_{ij} \neq 0;$$

Fitting this interaction model

```
goods.int <- lm(Freshness ~ Packaging * Storage, data = goods)
anova(goods.int)

## Analysis of Variance Table
##
## Response: Freshness
##                   Df Sum Sq Mean Sq   F value Pr(>F)
## Packaging          1 1839.8  1839.8  613.6776 <2e-16 ***
## Storage            1 5824.5  5824.5 1942.7752 <2e-16 ***
## Packaging:Storage  1    3.4     3.4    1.1295 0.2921
## Residuals         61  182.9     3.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We can see that the interaction terms are insignificant since the F-test of the interaction term has a P-value 0.2921 > 0.05, they should be removed from the model.

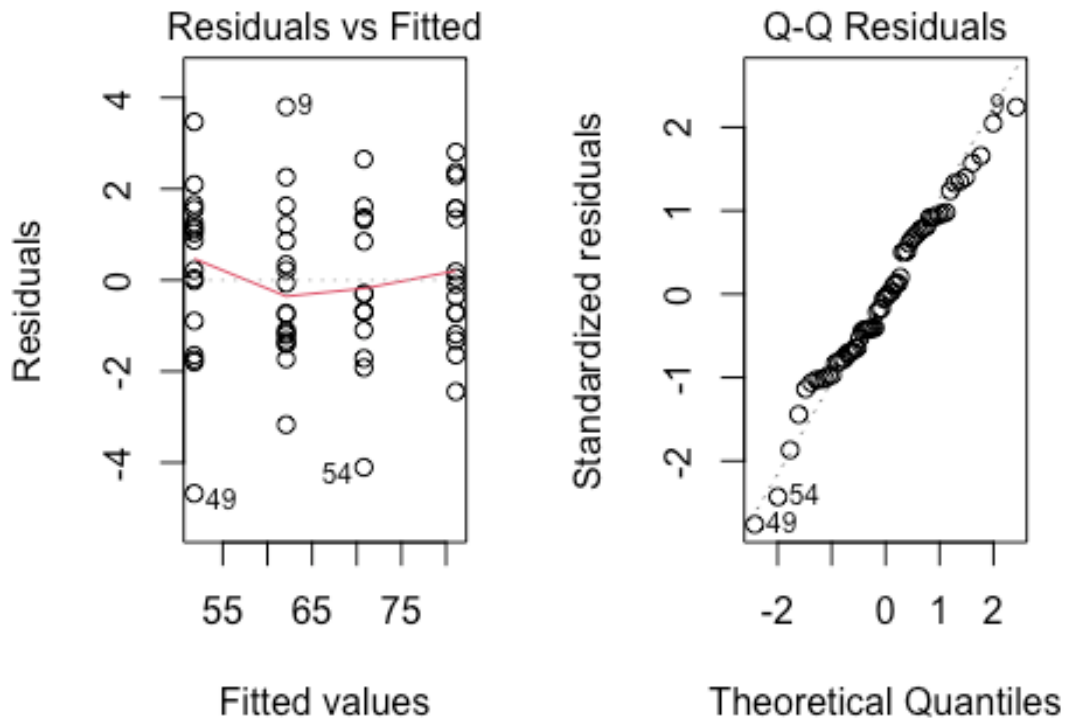
```
goods.int2 <- lm(Freshness ~ Packaging + Storage, data = goods)
anova(goods.int2)

## Analysis of Variance Table
##
## Response: Freshness
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Packaging  1 1839.8  1839.8   612.4 < 2.2e-16 ***
## Storage    1 5824.5  5824.5  1938.7 < 2.2e-16 ***
## Residuals 62  186.3     3.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Now the P-value for Packaging and Storage are both 2.2e-16 < 0.05, this also means we reached our final model.

We should validate the interaction model with diagnostic plots.

```
par(mfrow = c(1,2))
plot(goods.int2, which = 1:2)
```



The residuals are close to linear in the QQ-plot, and so the normal assumption should be valid. The residual plot seems to show equal spread around the fitted values and the constant variance assumption is also appropriate.