Lab 4

Adam Shen

October 7, 2020

Packages

```r
library(car)
library(broom)
```

Homes Data

```r
homes <- read.table("./house.txt", header=TRUE)
```

Fit multiple linear regression model

```r
m123 <- lm(Price ~ Improve + Area + Land, x=TRUE, data=homes)
summary(m123)
```

```
##
## Call:
## lm(formula = Price ~ Improve + Area + Land, data = homes, x = TRUE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.856  -2.897   1.797   2.783  16.246
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 229.5069    97.9279   2.344  0.03234 *
## Improve       0.7932     0.2232   3.553  0.00265 **
## Area         13.3934     6.6878   2.003  0.06246 .
## Land          1.0104     0.6735   1.500  0.15299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.979 on 16 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8763
## F-statistic: 45.88 on 3 and 16 DF,  p-value: 4.397e-08
```

Correlations and variance inflation factors

```r
X <- m123$x
cor(X)
```

```
## Warning in cor(X): the standard deviation is zero
```

```
##             (Intercept)   Improve      Area      Land
## (Intercept)           1        NA        NA        NA
## Improve              NA 1.0000000 0.7881460 0.7866654
## Area                 NA 0.7881460 1.0000000 0.7328524
## Land                 NA 0.7866654 0.7328524 1.0000000
```

We get a bunch of `NA` values for the rows and columns involving `(Intercept)` because our design matrix has an entire column of ones in the `(Intercept)` column.

```r
head(X)
```

```
##   (Intercept) Improve  Area  Land
## 1           1  94.967 2.573 162.0
## 2           1  77.860 1.628 165.0
## 3           1  81.439 1.826 165.5
## 4           1  89.592 1.965 166.0
## 5           1 122.827 2.914 174.0
## 6           1  77.317 1.612 164.5
```

As such, the resulting sample standard deviation is zero. Therefore it may be of interest to use `cor()` on the data frame rather than the design matrix for "prettier" output.

```r
cor(homes)
```

```
##             Price   Improve      Area      Land
## Price   1.0000000 0.9156607 0.8489815 0.8295975
## Improve 0.9156607 1.0000000 0.7881460 0.7866654
## Area    0.8489815 0.7881460 1.0000000 0.7328524
## Land    0.8295975 0.7866654 0.7328524 1.0000000
```

```r
vif(m123)
```

```
##  Improve     Area     Land
## 3.516128 2.895052 2.877341
```

From *Module 4.8*, variance inflation factors measure how the correlations among the predictor variables affect the variances of the least squares estimators. If the value of the VIF is greater than 10, then there is evidence of severe multicollinearity.

Since none of the VIF values are greater than 10 here, there is insufficient evidence of severe multicollinearity.

Partial F-tests

```r
anova(m123)
```

```
## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## Improve    1 8202.5  8202.5 128.8261 4.589e-09 ***
## Area       1  418.5   418.5   6.5736   0.02081 *
## Land       1  143.3   143.3   2.2511   0.15299
## Residuals 16 1018.7    63.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The `Sum Sq` column represents the sequential increase of SSR (or equivalently, the sequential decrease in SSE) as the predictors `Improve`, `Area`, and `Land` are included into the model that came before it.

Build all the models we will use

```
m1 <- lm(Price ~ Improve, data=homes)
m2 <- lm(Price ~ Area, data=homes)
m13 <- lm(Price ~ Improve + Land, data=homes)
summary(m1)
```

```
##
## Call:
## lm(formula = Price ~ Improve, data = homes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7199  -3.3395  -0.6258   5.0580  18.8687
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 376.3830    12.1101  31.080  < 2e-16 ***
## Improve       1.3513     0.1398   9.665 1.51e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.371 on 18 degrees of freedom
## Multiple R-squared:  0.8384, Adjusted R-squared:  0.8295
## F-statistic: 93.41 on 1 and 18 DF,  p-value: 1.505e-08
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = Price ~ Area, data = homes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.2919  -4.7552   0.2277   9.1076  25.4814
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  405.485     12.939  31.338  < 2e-16 ***
## Area          41.364      6.068   6.816 2.21e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.32 on 18 degrees of freedom
## Multiple R-squared:  0.7208, Adjusted R-squared:  0.7053
## F-statistic: 46.46 on 1 and 18 DF,  p-value: 2.213e-06
```

```
summary(m13)
```

```
##
## Call:
## lm(formula = Price ~ Improve + Land, data = homes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -18.8960  -1.3223  -0.0745   2.5029  20.9228
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 172.1617   101.6027   1.694 0.108417
## Improve       1.0184     0.2092   4.868 0.000145 ***
## Land          1.4109     0.6977   2.022 0.059171 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.657 on 17 degrees of freedom
## Multiple R-squared:  0.8698, Adjusted R-squared:  0.8544
## F-statistic: 56.77 on 2 and 17 DF,  p-value: 2.987e-08
```

Partial F-tests: round 1

```r
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Improve    1 8202.5  8202.5   93.41 1.505e-08 ***
## Residuals 18 1580.6    87.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(m1, m123)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Improve
## Model 2: Price ~ Improve + Area + Land
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     18 1580.6
## 2     16 1018.7  2    561.88 4.4123 0.02978 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The associated hypotheses for the partial F-test are:

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{vs} \quad H_A : \text{At least one of } \beta_2, \beta_3 \text{ non-zero}$$

The p-value (0.02978) is less than 0.05 so we reject the null hypothesis. We conclude that at least one of β_2, β_3 is non-zero.

Partial F-tests: round 2

```r
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Area       1 7051.4  7051.4  46.463 2.213e-06 ***
## Residuals 18 2731.8   151.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(m2, m123)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Area
## Model 2: Price ~ Improve + Area + Land
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     18 2731.8
## 2     16 1018.7  2      1713 13.452 0.0003741 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The associated hypotheses for the partial F-test are:

$$H_0 : \beta_1 = \beta_3 = 0 \quad \text{vs} \quad H_A : \text{At least one of } \beta_1,\ \beta_3 \text{ non-zero}$$

The p-value (0.0003741) is less than 0.05 so we reject the null hypothesis. We conclude that at least one of β_1, β_3 is non-zero.

Partial F-tests: round 3

```r
anova(m13)
```

```
## Analysis of Variance Table
##
## Response: Price
##           Df Sum Sq Mean Sq  F value   Pr(>F)
## Improve    1 8202.5  8202.5 109.4437 7.96e-09 ***
## Land       1  306.5   306.5   4.0897  0.05917 .
## Residuals 17 1274.1    74.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(m13, m123)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Improve + Land
## Model 2: Price ~ Improve + Area + Land
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     17 1274.1
## 2     16 1018.7  1    255.37 4.0107 0.06246 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The associated hypotheses for the partial F-test are:

$$H_0 : \beta_2 = 0 \quad \text{vs} \quad H_A : \beta_2 \neq 0$$

The p-value (0.06246) is greater than 0.05 so we fail to reject the null hypothesis. There is insufficient evidence to support the claim that β_2 is non-zero. In other words, we can safely drop `Area` from the model.

Brand preference data

```r
muffin <- read.table("./brandpref.txt", header=TRUE)
```

Fit multiple linear regression model

```r
m12 <- lm(Liking ~ Moisture + Sweetness, x=TRUE, data=muffin)
summary(m12)
```

```
##
## Call:
## lm(formula = Liking ~ Moisture + Sweetness, data = muffin, x = TRUE)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.400 -1.762  0.025  1.587  4.200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.6500     2.9961  12.566 1.20e-08 ***
## Moisture      4.4250     0.3011  14.695 1.78e-09 ***
## Sweetness     4.3750     0.6733   6.498 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.693 on 13 degrees of freedom
## Multiple R-squared:  0.9521, Adjusted R-squared:  0.9447
## F-statistic: 129.1 on 2 and 13 DF,  p-value: 2.658e-09
```

Correlations and variance inflation factors

```r
X <- m12$x
head(X)
```

```
##   (Intercept) Moisture Sweetness
## 1           1        4         2
## 2           1        4         4
## 3           1        4         2
## 4           1        4         4
## 5           1        6         2
## 6           1        6         4
```

```r
cor(X)
```

```
## Warning in cor(X): the standard deviation is zero
```

```
##             (Intercept) Moisture Sweetness
## (Intercept)           1       NA        NA
## Moisture             NA        1         0
## Sweetness            NA        0         1
```

```r
cor(muffin)
```

```
##              Liking  Moisture Sweetness
## Liking    1.0000000 0.8923929 0.3945807
## Moisture  0.8923929 1.0000000 0.0000000
## Sweetness 0.3945807 0.0000000 1.0000000
```

```r
vif(m12)
```

```
##  Moisture Sweetness
##         1         1
```

Interestingly, the correlation between `Moisture` and `Sweetness` is zero. The resulting variance inflation factors are all one. This means that the correlation between `Moisture` and `Sweetness` is not affecting the variance of the least squares estimates (scaling by a factor of 1). Since none of the variance inflation factors are greater than 10, there is no evidence of severe multicollinearity (but we already knew that).

Brand preference data with centred predictors

```r
muffin <- transform(muffin,
                    Moistcen=Moisture - mean(Moisture),
                    Sweetcen=Sweetness - mean(Sweetness))
```

Fit multiple linear regression model with centred predictors

```r
m12c <- lm(Liking ~ Moistcen + Sweetcen, x=TRUE, data=muffin)
summary(m12c)
```

```
##
## Call:
## lm(formula = Liking ~ Moistcen + Sweetcen, data = muffin, x = TRUE)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.400 -1.762  0.025  1.587  4.200
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  81.7500     0.6733 121.413  < 2e-16 ***
## Moistcen      4.4250     0.3011  14.695 1.78e-09 ***
## Sweetcen      4.3750     0.6733   6.498 2.01e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.693 on 13 degrees of freedom
## Multiple R-squared:  0.9521, Adjusted R-squared:  0.9447
## F-statistic: 129.1 on 2 and 13 DF,  p-value: 2.658e-09
```

Correlations and variance inflation factors

```r
with(muffin, cor(Moistcen, Sweetcen))
```

```
## [1] 0
```

```r
Xc <- m12c$x
cor(Xc)
```

```
## Warning in cor(Xc): the standard deviation is zero
```

```
##             (Intercept) Moistcen Sweetcen
## (Intercept)           1       NA       NA
## Moistcen             NA        1        0
## Sweetcen             NA        0        1
```

```r
vif(m12c)
```

```
## Moistcen Sweetcen
##        1        1
```

The correlation of `Moistcen` and `Sweetcent` is zero. The variance inflation factors are 1, i.e. the variances of the least squares estimators are not being inflated.

Comparisons of centred and non-centred models

```r
options(pillar.sigfig=4)
```

For the code below, we will be creatings lists of tibbles (tibbles are special data frames). There is some oddness that will occur due to truncating/rounding when printing these lists of tibbles, making some numbers seem unequal when they are actually equal. We will change the number of significant digits (only for tibbles) to 4 (default is 3) as a workaround.

Fit the rest of the models

```r
m1 <- lm(Liking ~ Moisture, data=muffin)
m1c <- lm(Liking ~ Moistcen, data=muffin)
m2 <- lm(Liking ~ Sweetness, data=muffin)
m2c <- lm(Liking ~ Sweetcen, data=muffin)
```

Next page...

Coefficient comparisons

```
(all_coefs <- list(m12=tidy(m12), m12c=tidy(m12c),
                   m1=tidy(m1), m1c=tidy(m1c),
                   m2=tidy(m2), m2c=tidy(m2c)))
```

```
## $m12
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   37.65     2.996      12.57  1.200e-8
## 2 Moisture       4.425    0.3011     14.70  1.778e-9
## 3 Sweetness      4.375    0.6733      6.498 2.011e-5
##
## $m12c
## # A tibble: 3 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   81.75     0.6733   121.4    3.017e-21
## 2 Moistcen       4.425    0.3011     14.70  1.778e- 9
## 3 Sweetcen       4.375    0.6733      6.498 2.011e- 5
##
## $m1
## # A tibble: 2 x 5
##   term        estimate std.error statistic       p.value
##   <chr>          <dbl>     <dbl>     <dbl>         <dbl>
## 1 (Intercept)   50.78     4.395      11.55  0.00000001519
## 2 Moisture       4.425    0.5980      7.399 0.000003356
##
## $m1c
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   81.75     1.337      61.13  2.115e-18
## 2 Moistcen       4.425    0.5980      7.399 3.356e- 6
##
## $m2
## # A tibble: 2 x 5
##   term        estimate std.error statistic     p.value
##   <chr>          <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept)   68.62      8.610      7.970 0.000001431
## 2 Sweetness      4.375     2.723      1.607 0.1304
##
## $m2c
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)   81.75      2.723     30.02  4.127e-14
## 2 Sweetcen       4.375     2.723      1.607 1.304e- 1
```

ANOVA comparisons

```
(all_anova <- list(m12=tidy(anova(m12)), m12c=tidy(anova(m12c)),
                   m1=tidy(anova(m1)), m1c=tidy(anova(m1c)),
                   m2=tidy(anova(m2)), m2c=tidy(anova(m2c))))
```

```
## $m12
## # A tibble: 3 x 6
##   term          df  sumsq   meansq statistic    p.value
##   <chr>      <int>  <dbl>    <dbl>     <dbl>      <dbl>
## 1 Moisture       1 1566.  1566.        215.9   1.778e-9
## 2 Sweetness      1  306.2  306.2        42.22  2.011e-5
## 3 Residuals     13   94.3    7.254      NA    NA
##
## $m12c
## # A tibble: 3 x 6
##   term          df  sumsq   meansq statistic    p.value
##   <chr>      <int>  <dbl>    <dbl>     <dbl>      <dbl>
## 1 Moistcen       1 1566.  1566.        215.9   1.778e-9
## 2 Sweetcen       1  306.2  306.2        42.22  2.011e-5
## 3 Residuals     13   94.3    7.254      NA    NA
##
## $m1
## # A tibble: 2 x 6
##   term          df  sumsq  meansq statistic       p.value
##   <chr>      <int>  <dbl>   <dbl>     <dbl>         <dbl>
## 1 Moisture       1 1566.  1566.       54.75   0.000003356
## 2 Residuals     14  400.5   28.61     NA    NA
##
## $m1c
## # A tibble: 2 x 6
##   term          df  sumsq  meansq statistic       p.value
##   <chr>      <int>  <dbl>   <dbl>     <dbl>         <dbl>
## 1 Moistcen       1 1566.  1566.       54.75   0.000003356
## 2 Residuals     14  400.5   28.61     NA    NA
##
## $m2
## # A tibble: 2 x 6
##   term          df  sumsq meansq statistic p.value
##   <chr>      <int>  <dbl>  <dbl>     <dbl>   <dbl>
## 1 Sweetness      1  306.2  306.2     2.582  0.1304
## 2 Residuals     14 1661.   118.6    NA     NA
##
## $m2c
## # A tibble: 2 x 6
##   term          df  sumsq meansq statistic p.value
##   <chr>      <int>  <dbl>  <dbl>     <dbl>   <dbl>
## 1 Sweetcen       1  306.2  306.2     2.582  0.1304
## 2 Residuals     14 1661.   118.6    NA     NA
```