Correlation and Ordinary Least Squares

ITEC 621 W3

Diamonds data

```
library(ggplot2)
head(diamonds)
```

```
# A tibble: 6 x 10
  carat       cut color clarity depth table price     x     y     z
  <dbl>     <ord> <ord>   <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>
1  0.23     Ideal     E     SI2  61.5    55   326  3.95  3.98  2.43
2  0.21   Premium     E     SI1  59.8    61   326  3.89  3.84  2.31
3  0.23      Good     E     VS1  56.9    65   327  4.05  4.07  2.31
4  0.29   Premium     I     VS2  62.4    58   334  4.20  4.23  2.63
5  0.31      Good     J     SI2  63.3    58   335  4.34  4.35  2.75
6  0.24 Very Good     J    VVS2  62.8    57   336  3.94  3.96  2.48
```

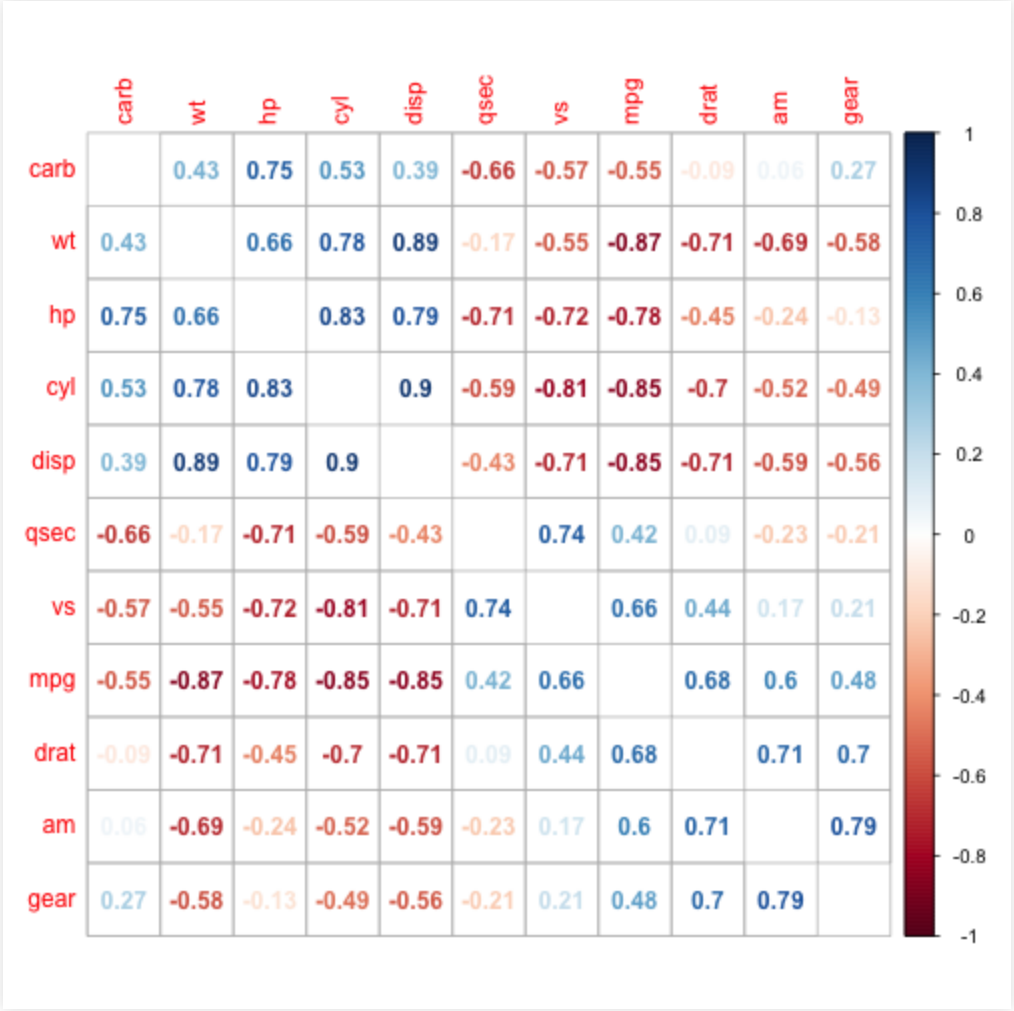
Correlations

```
diamonds.matrix = as.matrix(diamonds[, c('price', 'carat', 'depth')])
cor(diamonds.matrix)
```

```
            price      carat       depth
price  1.0000000 0.92159130 -0.01064740
carat  0.9215913 1.00000000  0.02822431
depth -0.0106474 0.02822431  1.00000000
```

Correlation heat map

```
library(corrplot)
options(scipen="4")
mtCorr = cor(mtcars)
corrplot(mtCorr, method = "number", order = "hclust", diag = FALSE)
```


carb wt hp cyl disp qsec vs mpg drat am gear
carb 0.43 0.75 0.53 0.39 -0.66 -0.57 -0.55 -0.09 0.06 0.27
wt 0.43 0.66 0.78 0.89 -0.17 -0.55 -0.87 -0.71 -0.69 -0.58
hp 0.75 0.66 0.83 0.79 -0.71 -0.72 -0.78 -0.45 -0.24 -0.13
cyl 0.53 0.78 0.83 0.9 -0.59 -0.81 -0.85 -0.7 -0.52 -0.49
disp 0.39 0.89 0.79 0.9 -0.43 -0.71 -0.85 -0.71 -0.59 -0.56
qsec -0.66 -0.17 -0.71 -0.59 -0.43 0.74 0.42 0.09 -0.23 -0.21
vs -0.57 -0.55 -0.72 -0.81 -0.71 0.74 0.66 0.44 0.17 0.21
mpg -0.55 -0.87 -0.78 -0.85 -0.85 0.42 0.66 0.68 0.6 0.48
drat -0.09 -0.71 -0.45 -0.7 -0.71 0.09 0.44 0.68 0.71 0.7
am 0.06 -0.69 -0.24 -0.52 -0.59 -0.23 0.17 0.6 0.71 0.79
gear 0.27 -0.58 -0.13 -0.49 -0.56 -0.21 0.21 0.48 0.7 0.79
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Correlation p-values

Note we never prove/disprove a hypothesis. At best, we show one is extremely unlikely.

Null hypothesis: Correlation = 0

Alternative hypothesis: Correlation ≠ 0

p-value: Estimated probability that null hypothesis is true, given the data

```
library(Hmisc)
rcorr(diamonds.matrix, type="pearson")
```

```
      price carat depth
price  1.00  0.92 -0.01
carat  0.92  1.00  0.03
depth -0.01  0.03  1.00

n= 53940


P
      price  carat  depth
price        0.0000 0.0134
```

```
carat 0.0000        0.0000
depth 0.0134 0.0000
```

Comparing models

```r
# Null model
fit0 = lm(price ~ 1, data=diamonds)
summary(fit0)
```

```
Call:
lm(formula = price ~ 1, data = diamonds)

Residuals:
   Min     1Q Median     3Q    Max
 -3607  -2983  -1532   1392  14890

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3932.80      17.18     229   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3989 on 53939 degrees of freedom
```

```r
# Carat model (2 parameters)
fit1 = lm(price ~ 1 + carat, data=diamonds)
summary(fit1)
```

```
Call:
lm(formula = price ~ 1 + carat, data = diamonds)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-18585.3   -804.8    -18.9    537.4  12731.7

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2256.36      13.06  -172.8   <2e-16 ***
carat        7756.43      14.07   551.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1549 on 53938 degrees of freedom
Multiple R-squared:  0.8493,    Adjusted R-squared:  0.8493
F-statistic: 3.041e+05 on 1 and 53938 DF,  p-value: < 2.2e-16
```

Comparing models

```r
# Null model (1 parameter)
fit0 = lm(price ~ 1, data=diamonds)
# Carat model (2 parameters)
fit1 = lm(price ~ 1 + carat, data=diamonds)
# Is carat model an improvement over null model?
anova(fit0, fit1)
```

```
Analysis of Variance Table

Model 1: price ~ 1
Model 2: price ~ 1 + carat
  Res.Df          RSS Df    Sum of Sq      F    Pr(>F)
1  53939 858473135517
2  53938 129345695398  1 729127440120 304051 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing models

```r
# Null model (1 parameter)
fit0 = lm(price ~ 1, data=diamonds)
# Carat  model (2 parameters)
fit1 = lm(price ~ carat, data=diamonds)
# Carat + clarity  model (8 parameters)
fit2 = lm(price ~ carat + clarity, data=diamonds)
summary(fit2)
```

```
Call:
lm(formula = price ~ carat + clarity, data = diamonds)

Residuals:
     Min       1Q   Median       3Q      Max 
-17355.1   -639.9   -110.6    480.4  11162.8 

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)    
(Intercept) -2977.27      13.11 -227.080   <2e-16 ***
carat        8440.06      12.65  667.132   <2e-16 ***
clarity.L    4216.78      33.65  125.297   <2e-16 ***
clarity.Q   -1931.41      31.87  -60.601   <2e-16 ***
clarity.C    1005.85      27.39   36.719   <2e-16 ***
clarity^4    -480.18      21.94  -21.883   <2e-16 ***
clarity^5     283.94      17.98   15.796   <2e-16 ***
clarity^6      12.66      15.68    0.808    0.419    
clarity^7     198.05      13.82   14.336   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1294 on 53931 degrees of freedom
Multiple R-squared:  0.8949,    Adjusted R-squared:  0.8948
F-statistic: 5.737e+04 on 8 and 53931 DF,  p-value: < 2.2e-16
```

Comparing models

```r
# Carat  model (2 parameters)
fit1 = lm(price ~ 1 + carat, data=diamonds)
# Carat + clarity  model (8 parameters)
fit2 = lm(price ~ 1 + carat + clarity, data=diamonds)

# Is carat + clarity model an improvement over carat model?
anova(fit1, fit2)
```

```
Analysis of Variance Table

Model 1: price ~ 1 + carat
Model 2: price ~ 1 + carat + clarity
  Res.Df          RSS Df   Sum of Sq      F    Pr(>F)
1  53938 129345695398
2  53931  90263944583  7 39081750815 3335.8 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```