Stat 5100 Handout 2.6.1 - R: Inference with Multiple Predictors

Stat 5100: Dr. Bean

Example: (Table 7.1) Study seeks to relate (in females) amount of body fat (Y) to triceps skinfold thickness (X_1), thigh circumference (X_2), and midarm circumference (X_3). Amount of body fat is expensive to measure, requiring immersion of person in water. This expense motivates the desire for a predictive model based on these inexpensive predictors.

Q1: Do thigh and midarm both have no effect on body fat when triceps is in the model?

Q2: Do the relationships among the predictors cause any problems in the fitted model?

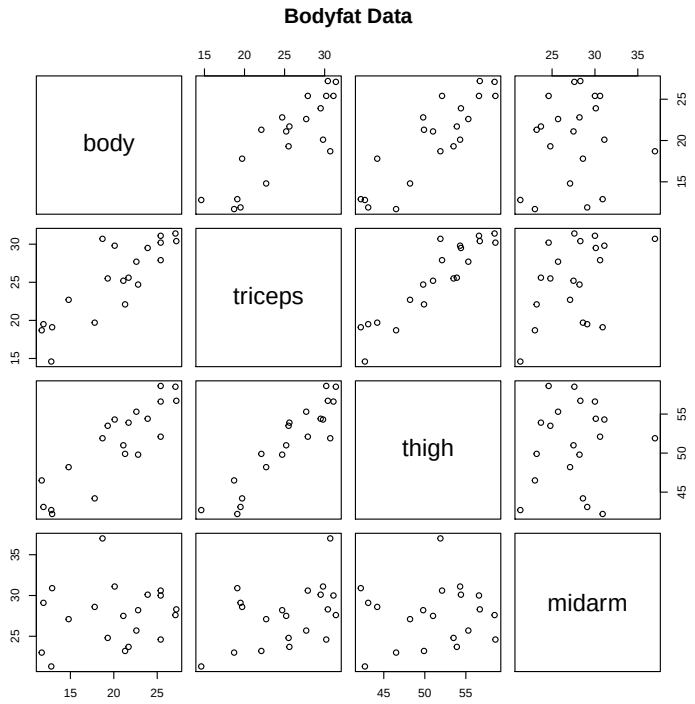
```
# Make output easier to read by disabling scientific notation
options(scipen = 999)

# Input data and take a look at the first few observations
library(stat5100)
data(bodyfat)
head(bodyfat)
```

```
##   triceps thigh midarm body
## 1    19.5  43.1   29.1 11.9
## 2    24.7  49.8   28.2 22.8
## 3    30.7  51.9   37.0 18.7
## 4    29.8  54.3   31.1 20.1
## 5    19.1  42.2   30.9 12.9
## 6    25.6  53.9   23.7 21.7
```

```
# Look at the correlation matrix
cor(bodyfat)
```

```
##           triceps     thigh    midarm      body
## triceps 1.0000000 0.9238425 0.4577772 0.8432654
## thigh   0.9238425 1.0000000 0.0846675 0.8780896
## midarm  0.4577772 0.0846675 1.0000000 0.1424440
## body    0.8432654 0.8780896 0.1424440 1.0000000
```

```
# Look at the scatterplot
pairs( ~ body + triceps + thigh + midarm, data = bodyfat,
       main = "Bodyfat Data")
```



Question 1: Test whether thigh and midarm BOTH have no effect on body when triceps is in the model

```
bodyfat_lm_full <- lm(body ~ triceps + thigh + midarm, data = bodyfat)
anova(bodyfat_lm_full)

## Analysis of Variance Table
##
## Response: body
##           Df Sum Sq Mean Sq F value      Pr(>F)
## triceps    1 352.27  352.27 57.2768 0.000001131 ***
## thigh      1  33.17   33.17  5.3931     0.03373 *
## midarm     1  11.55   11.55  1.8773     0.18956
## Residuals 16  98.40    6.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

bodyfat_lm_reduced <- lm(body ~ triceps, data = bodyfat)
anova(bodyfat_lm_reduced)

## Analysis of Variance Table
##
## Response: body
##           Df Sum Sq Mean Sq F value      Pr(>F)
## triceps    1 352.27  352.27  44.305 0.000003024 ***
## Residuals 18 143.12    7.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Perform the subset F-test by hand

```
# (All these values are grabbed from the ANOVA tables above)
SSE_Reduced <- 143.12               # Sum of squared error for reduced model
SSE_Full <- 98.40                   # Sum of squared error for full model
MSE_Full <- 6.15                    # Mean square error for full model
MSR <- (SSE_Reduced - SSE_Full) / 2 # Mean square reduction
F_statistic <- MSR / MSE_Full
F_statistic

## [1] 3.635772

# The F-statistic above follows a F(2,16) distribution (16 denominator degrees
# of freedom because MSE_Full is calculated by SSE_Full / 16)
pvalue <- pf(F_statistic, 2, 16, lower.tail = FALSE)
pvalue

## [1] 0.04992961
```

Perform the subset F-test automatically

```
# This is a test for the null hypothesis that thigh = midarm= 0
bodyfat_lm_full <- lm(body ~ triceps + thigh + midarm, data = bodyfat)
bodyfat_lm_reduced <- lm(body ~ triceps, data = bodyfat)

anova(bodyfat_lm_full, bodyfat_lm_reduced)

## Analysis of Variance Table
##
## Model 1: body ~ triceps + thigh + midarm
## Model 2: body ~ triceps
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1     16  98.405
## 2     18 143.120 -2   -44.715 3.6352 0.04995 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Notice here that we get the same F-statistic and p-value as we did when
# we tested by hand.
```

Question 2: Investigate the effect of relationships among predictors

```
# Standardize the variables and create a standardized regression model
bodyfat_standardized <- data.frame(scale(bodyfat))
bodyfat_s_lm<- lm(body ~ triceps + thigh + midarm, data = bodyfat_standardized)
summary(bodyfat_s_lm)

##
## Call:
## lm(formula = body ~ triceps + thigh + midarm, data = bodyfat_standardized)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.72976 -0.31552  0.07683  0.28702  0.80837
##
## Coefficients:
##                            Estimate             Std. Error t value Pr(>|t|)
## (Intercept) -0.000000000000009096  0.1086017313054759154   0.000    1.000
## triceps      4.2637045669999400488  2.9665382148454386702   1.437    0.170
## thigh       -2.9287006520636991169  2.6469556563935556781  -1.106    0.285
## midarm      -1.5614167939150791486  1.1396021351584755266  -1.370    0.190
##
## Residual standard error: 0.4857 on 16 degrees of freedom
## Multiple R-squared:  0.8014,Adjusted R-squared:  0.7641
## F-statistic: 21.52 on 3 and 16 DF,  p-value: 0.000007343

# Test for multicollinearity
library(olsrr)

##
## Attaching package: 'olsrr'
## The following object is masked from 'package:datasets':
##
##     rivers

ols_coll_diag(bodyfat_lm_full)

## Tolerance and Variance Inflation Factor
## ---------------------------------------
##   Variables   Tolerance      VIF
## 1   triceps 0.001410750 708.8429
## 2     thigh 0.001771971 564.3434
## 3    midarm 0.009559681 104.6060
##
##
## Eigenvalue and Condition Index
## ------------------------------
##       Eigenvalue Condition Index      intercept        triceps          thigh
## 1 3.967956738483         1.00000 0.000001946679 0.000003196101 0.000001104251
## 2 0.020522792222        13.90482 0.000371520161 0.001319092929 0.000032620187
## 3 0.011511821361        18.56570 0.000599149515 0.000218746554 0.000325502370
## 4 0.000008647934       677.37207 0.999027383645 0.998458964415 0.999640773192
##           midarm
## 1 0.000009798156
## 2 0.001388740360
## 3 0.006933507626
## 4 0.991667953858
```