Rexample

Lingxiao Zhou

Bodyfat example

Load dataset

```
dat = read.table("http://www.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData/Cha
    col.names = c("X1", "X2", "X3", "Y"))
```

Get sequential sum of sqaures

- In R, we obtain SSR decomposed by sequential sums of squares which differ depending on the order the variables are entered.

X1, X2|X1, X3|X1,X2

```
### Choosing model using sequential sums of squares
reg123 = lm(Y ~ X1 + X2 + X3, data = dat)
summary(reg123)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7263 -1.6111  0.3923  1.4656  4.1277
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  117.085     99.782   1.173    0.258
## X1             4.334      3.016   1.437    0.170
## X2            -2.857      2.582  -1.106    0.285
## X3            -2.186      1.595  -1.370    0.190
##
## Residual standard error: 2.48 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: 7.343e-06
```

```r
anova(reg123)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 352.27  352.27 57.2768 1.131e-06 ***
## X2         1  33.17   33.17  5.3931   0.03373 *
## X3         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
```

- At least one predictor is significant (F-test)
- No predictor is significant given other two (t-test)
- The anova table shows $\text{SSR}(X_1) = 352.27$, $\text{SSR}(X_2|X_1) = 33.17$ and , $\text{SSR}(X_3|X_1, X_2) = 11.55$
- predictor X_1 is significant if the other two predictors are not included in the model (p-value = 1.131e-06)
- predictor X_2 is significant given X_1 (p-value = 0.03373)

X2, X1|X2, X3|X1,X2

```r
reg213 = lm(Y ~ X2 + X1 + X3, data = dat)
anova(reg213)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X2         1 381.97  381.97 62.1052 6.735e-07 ***
## X1         1   3.47    3.47  0.5647    0.4633
## X3         1  11.55   11.55  1.8773    0.1896
## Residuals 16  98.40    6.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

X3, X2|X3, X1|X2,X3

```r
reg321 = lm(Y ~ X3 + X2 + X1, data = dat)
anova(reg321)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X3         1  10.05   10.05  1.6343    0.2193
## X2         1 374.23  374.23 60.8471 7.684e-07 ***
## X1         1  12.70   12.70  2.0657    0.1699
## Residuals 16  98.40    6.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

General linear test

H_0: $\beta_1 = \beta_3 = 0$ **(given** X_2**)**

- test whether we can remove X1 and X3 simultaneously
- full: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$
- reduced: $Y = \beta_0 + \beta_2 X_2 + \epsilon$

```
reg2 = update(reg123, . ~ . - X1 - X3)  # model with X2 only
anova(reg2, reg123)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X2
## Model 2: Y ~ X1 + X2 + X3
##   Res.Df     RSS Df Sum of Sq     F Pr(>F)
## 1     18 113.424
## 2     16  98.405  2    15.019 1.221  0.321
```

H_0: β_2**=0 (given** X_1**,** X_3**)**

- test whether we can remove X2 (given other two predictors)
- full: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$
- reduced: $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \epsilon$
- Note the equivalence of F-test and t-test for testing x_2 given X_1, X_3

```
reg13 = update(reg123, . ~ . - X2)
summary(reg13)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X3, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8794 -1.9627  0.3811  1.2688  3.8942
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.7916     4.4883   1.513   0.1486
## X1            1.0006     0.1282   7.803 5.12e-07 ***
## X3           -0.4314     0.1766  -2.443   0.0258 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.496 on 17 degrees of freedom
## Multiple R-squared:  0.7862, Adjusted R-squared:  0.761
## F-statistic: 31.25 on 2 and 17 DF,  p-value: 2.022e-06
```

```
anova(reg13, reg123)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X3
## Model 2: Y ~ X1 + X2 + X3
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     17 105.934
## 2     16  98.405  1    7.5293 1.2242 0.2849
```

```
summary(reg123)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7263 -1.6111  0.3923  1.4656  4.1277
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  117.085     99.782   1.173    0.258
## X1             4.334      3.016   1.437    0.170
## X2            -2.857      2.582  -1.106    0.285
## X3            -2.186      1.595  -1.370    0.190
##
## Residual standard error: 2.48 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: 7.343e-06
```

H_0: $\beta_1 = \beta_2 = \beta_3$

- Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$
- Reduced model:

$$\begin{aligned} Y =& \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \beta_1 X_3 + \epsilon \\ =& \beta_0 + \beta_1 (X_1 + X_2 + X_3) + \epsilon \\ =& \beta_0 + \beta_1 Z + \epsilon \end{aligned}$$

where $Z = X_1 + X_2 + X_3$

- resulting F-test will be $F_{2,n-4}$

```
dat$Z <- dat$X1 + dat$X2 + dat$X3
regz <- lm(Y ~ Z, data = dat)
anova(regz, reg123)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ Z
## Model 2: Y ~ X1 + X2 + X3
##   Res.Df     RSS Df Sum of Sq     F  Pr(>F)
## 1     18 173.328
```

```
## 2     16  98.405  2    74.923 6.091 0.01079 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Since p-value is small, we will reject the null hypothesis.

Unstandardized regression

```
library(car)  # for vif()

reg13 = lm(Y ~ X1 + X3, data = dat)
summary(reg13)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X3, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8794 -1.9627  0.3811  1.2688  3.8942
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.7916     4.4883   1.513   0.1486
## X1            1.0006     0.1282   7.803 5.12e-07 ***
## X3           -0.4314     0.1766  -2.443   0.0258 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.496 on 17 degrees of freedom
## Multiple R-squared:  0.7862, Adjusted R-squared:  0.761
## F-statistic: 31.25 on 2 and 17 DF,  p-value: 2.022e-06
```

```
sqrt(vif(reg13))  # compute the square root of VIF
```

```
##       X1       X3
## 1.124775 1.124775
```

```
round(cor(dat[, 1:4]), 2)  # get the pairwise correlation between Y,X1,X2,X3
```

```
##      X1   X2   X3    Y
## X1 1.00 0.92 0.46 0.84
## X2 0.92 1.00 0.08 0.88
## X3 0.46 0.08 1.00 0.14
## Y  0.84 0.88 0.14 1.00
```

- For a model with only two predictors, the VIF for these predictors are always the same.
- Standard error for the coefficient of X_1 variable is 1.125 times as large as it would be if X_1 were uncorrelated with X_3.
- X_1 and X_2 are highly correlated ($\text{cor}(X_1,X_2) = 0.92$)

Standardized regression

```
# Define a function to compute the correlation
# transformation
cor.trans = function(y) {
    n = length(y)
    1/sqrt(n - 1) * (y - mean(y))/sd(y)
}

dat_trans = as.data.frame(apply(dat[, 1:4], 2, cor.trans))  # obtain the transformed data
reg13_trans = lm(Y ~ 0 + X1 + X3, data = dat_trans)
summary(reg13_trans)
```

```
##
## Call:
## lm(formula = Y ~ 0 + X1 + X3, data = dat_trans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17430 -0.08818  0.01712  0.05701  0.17496
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## X1   0.9843     0.1226   8.029 2.33e-07 ***
## X3  -0.3082     0.1226  -2.514   0.0217 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.109 on 18 degrees of freedom
## Multiple R-squared:  0.7862, Adjusted R-squared:  0.7624
## F-statistic: 33.09 on 2 and 18 DF,  p-value: 9.35e-07
```

- Note that the standard errors decrease in the standardized regression.

Multicolinearity

```
reg123 = lm(Y ~ X1 + X2 + X3, data = dat)
vif(reg123)
```

```
##       X1       X2       X3
## 708.8429 564.3434 104.6060
```

```
summary(reg123)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = dat)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.7263 -1.6111  0.3923  1.4656  4.1277
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  117.085     99.782   1.173    0.258
## X1             4.334      3.016   1.437    0.170
## X2            -2.857      2.582  -1.106    0.285
## X3            -2.186      1.595  -1.370    0.190
##
## Residual standard error: 2.48 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: 7.343e-06
```

- Standard errors are greatly inflated for the model with all three