Week 6 Question Sheet

Lab Example

Part i)

```
# use the leaps function
leap.mod <- leaps(housing[,-ncol(housing)], housing$MEDV,
                  method="adjr2", nbest=5, names=names(housing)[-ncol(housing)])
```

Part ii)

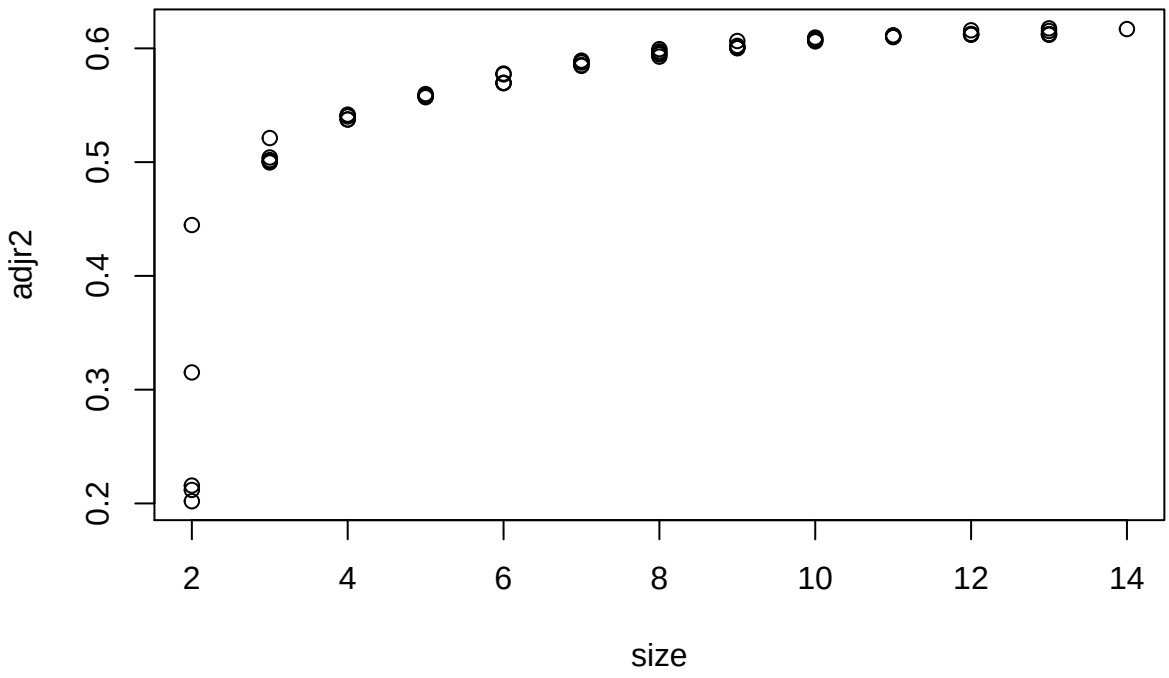
```
result.tab <- data.frame(adjr2=leap.mod$adjr2,
                         size=leap.mod$size,
                         leap.mod$which,
                         row.names=NULL)

# This table is quite large - the first few rows are:
head(result.tab)
```

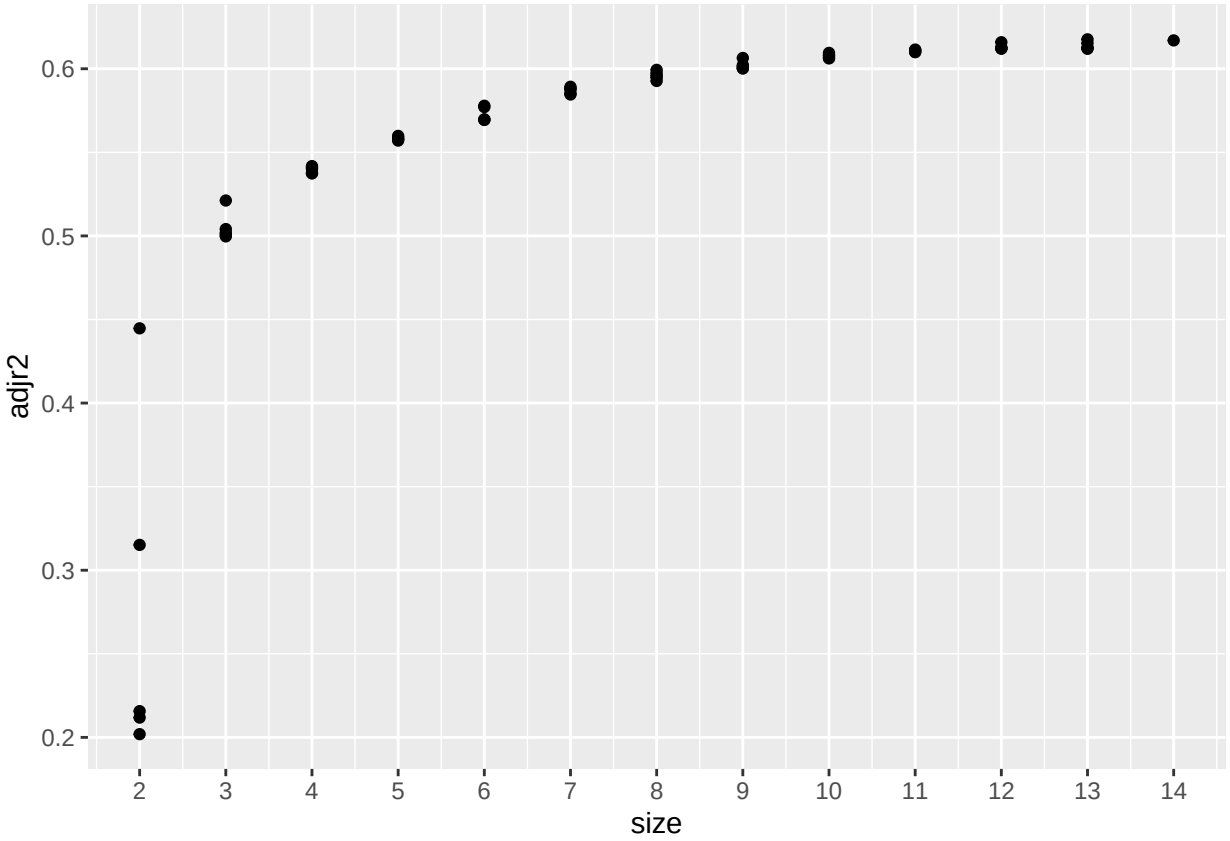
```
##       adjr2 size  CRIM    ZN INDUS  CHAS   NOX    RM   AGE   DIS   RAD   TAX
## 1 0.4447169    2 FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
## 2 0.3151611    2 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 3 0.2156906    2 FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 4 0.2118456    2 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
## 5 0.2019258    2 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 6 0.5211816    3 FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE  TRUE
##   PTRATIO     B LSTAT
## 1   FALSE FALSE FALSE
## 2   FALSE FALSE  TRUE
## 3   FALSE FALSE FALSE
## 4   FALSE FALSE FALSE
## 5    TRUE FALSE FALSE
## 6   FALSE FALSE FALSE
```

Part iii)

```
# base R
plot(adjr2~size, data=result.tab)
```




```
# or ggplot
ggplot(result.tab) +
  geom_point(aes(size, adjr2)) +
  scale_x_continuous(breaks=1:14)
```



The curve here isn't particularly dramatic, so I would probably pick the best model of sizes 8, 9, and 10 which is roughly where the plot looks like its reaching an upper limit.

```
# best model contains:
# CRIM, NOX, RM, AGE, DIS, PTRATIO & B
result.tab %>% filter(size==8)
```

```
##       adjr2 size  CRIM    ZN INDUS  CHAS  NOX   RM   AGE  DIS   RAD   TAX
## 1 0.5993404    8  TRUE FALSE FALSE FALSE TRUE TRUE  TRUE TRUE FALSE FALSE
## 2 0.5974042    8 FALSE FALSE FALSE  TRUE TRUE TRUE  TRUE TRUE FALSE FALSE
## 3 0.5958389    8  TRUE FALSE FALSE  TRUE TRUE TRUE  TRUE TRUE FALSE FALSE
## 4 0.5946570    8  TRUE FALSE FALSE  TRUE TRUE TRUE FALSE TRUE FALSE FALSE
## 5 0.5927144    8  TRUE FALSE FALSE FALSE TRUE TRUE FALSE TRUE FALSE FALSE
##   PTRATIO     B LSTAT
## 1    TRUE  TRUE FALSE
## 2    TRUE  TRUE FALSE
## 3    TRUE FALSE FALSE
## 4    TRUE  TRUE FALSE
## 5    TRUE  TRUE  TRUE
```

```
# best model contains:
# CRIM, CHAS, NOX, RM, AGE, DIS, PTRATIO & B
result.tab %>% filter(size==9)
```

```
##       adjr2 size  CRIM    ZN INDUS  CHAS  NOX   RM  AGE  DIS   RAD   TAX
## 1 0.6064624    9  TRUE FALSE FALSE  TRUE TRUE TRUE TRUE TRUE FALSE FALSE
## 2 0.6021899    9  TRUE  TRUE FALSE FALSE TRUE TRUE TRUE TRUE FALSE FALSE
## 3 0.6012865    9  TRUE FALSE FALSE FALSE TRUE TRUE TRUE TRUE FALSE FALSE
```

```
## 4 0.6002879    9  TRUE FALSE  TRUE FALSE TRUE TRUE TRUE TRUE FALSE FALSE
## 5 0.6002154    9 FALSE FALSE FALSE  TRUE TRUE TRUE TRUE TRUE FALSE FALSE
##   PTRATIO    B LSTAT
## 1    TRUE TRUE FALSE
## 2    TRUE TRUE FALSE
## 3    TRUE TRUE  TRUE
## 4    TRUE TRUE FALSE
## 5    TRUE TRUE  TRUE
```

```
# best model contains:
# CRIM, ZN, CHAS, NOX, RM, AGE, DIS, PTRATIO & B
result.tab %>% filter(size==10)
```

```
##       adjr2 size CRIM    ZN INDUS CHAS  NOX   RM  AGE  DIS   RAD   TAX PTRATIO
## 1 0.6094973   10 TRUE  TRUE FALSE TRUE TRUE TRUE TRUE TRUE FALSE FALSE    TRUE
## 2 0.6080315   10 TRUE FALSE FALSE TRUE TRUE TRUE TRUE TRUE FALSE FALSE    TRUE
## 3 0.6078712   10 TRUE FALSE  TRUE TRUE TRUE TRUE TRUE TRUE FALSE FALSE    TRUE
## 4 0.6065012   10 TRUE FALSE FALSE TRUE TRUE TRUE TRUE TRUE  TRUE FALSE    TRUE
## 5 0.6061945   10 TRUE FALSE FALSE TRUE TRUE TRUE TRUE TRUE FALSE  TRUE    TRUE
##       B LSTAT
## 1 TRUE FALSE
## 2 TRUE  TRUE
## 3 TRUE FALSE
## 4 TRUE FALSE
## 5 TRUE FALSE
```

```
## fit the models

size8 <- lm(MEDV ~ CRIM + NOX + RM +
              AGE + DIS + PTRATIO + B, data=housing)

size9 <- lm(MEDV ~ CRIM + CHAS + NOX + RM +
              AGE + DIS + PTRATIO + B, data=housing)

size10 <- lm(MEDV ~ CRIM + ZN + CHAS + NOX + RM +
              AGE + DIS + PTRATIO + B, data=housing)
```

Part iv)

```
# fit the full model
full.mod <- lm(MEDV ~ ., data=housing)

# use drop1
drop1(full.mod, scope= ~ CRIM + ZN + INDUS + CHAS +
        NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT,
      test="F")
```

```
## Single term deletions
##
## Model:
## MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD +
##     TAX + PTRATIO + B + LSTAT
##         Df Sum of Sq   RSS    AIC  F value    Pr(>F)
## <none>               17082 1808.7
## CRIM     1     473.7 17555 1820.6  13.6442 0.0002457 ***
## ZN       1     259.5 17341 1814.3   7.4753 0.0064807 **
## INDUS    1       6.4 17088 1806.9   0.1830 0.6690157
```

```
## CHAS     1     290.6 17372 1815.3   8.3705 0.0039827 **
## NOX      1     430.3 17512 1819.3  12.3943 0.0004707 ***
## RM       1    5250.7 22332 1942.3 151.2357 < 2.2e-16 ***
## AGE      1     345.1 17427 1816.8   9.9405 0.0017156 **
## DIS      1    1608.5 18690 1852.3  46.3311 2.915e-11 ***
## RAD      1     235.4 17317 1813.7   6.7811 0.0094913 **
## TAX      1     230.9 17312 1813.5   6.6500 0.0102047 *
## PTRATIO  1     640.1 17722 1825.3  18.4373 2.116e-05 ***
## B        1     423.3 17505 1819.1  12.1920 0.0005232 ***
## LSTAT    1     109.3 17191 1810.0   3.1493 0.0765771 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

INDUS is least significant so refit the model without INDUS and repeat:

```
mod <- lm(MEDV ~ CRIM + ZN + CHAS +
        NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT, data=housing)

# use drop1
drop1(mod, scope= ~ CRIM + ZN + CHAS +
        NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT,
      test="F")
```

```
## Single term deletions
##
## Model:
## MEDV ~ CRIM + ZN + CHAS + NOX + RM + AGE + DIS + RAD + TAX +
##     PTRATIO + B + LSTAT
##         Df Sum of Sq   RSS    AIC  F value    Pr(>F)
## <none>               17088 1806.9
## CRIM     1     470.6 17558 1818.7  13.5778 0.0002542 ***
## ZN       1     271.7 17360 1812.9   7.8380 0.0053165 **
## CHAS     1     284.9 17373 1813.3   8.2193 0.0043221 **
## NOX      1     489.6 17577 1819.2  14.1264 0.0001914 ***
## RM       1    5388.9 22477 1943.6 155.4743 < 2.2e-16 ***
## AGE      1     346.8 17435 1815.1  10.0054 0.0016571 **
## DIS      1    1640.0 18728 1851.3  47.3147 1.840e-11 ***
## RAD      1     277.3 17365 1813.0   8.0003 0.0048675 **
## TAX      1     328.2 17416 1814.5   9.4682 0.0022069 **
## PTRATIO  1     664.8 17753 1824.2  19.1809 1.453e-05 ***
## B        1     428.0 17516 1817.4  12.3484 0.0004820 ***
## LSTAT    1     112.4 17200 1808.2   3.2427 0.0723533 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

At the 10% significance level everything should stay in the model, at 5% significance level you would remove LSTAT. Since the p-value for LSTAT is still low I am choosing to retain it in the model, however it is not wrong to exclude it - variable selection is subjective.

`mod` is my final model.

Part v)

```
# fit the full model
int.mod <- lm(MEDV ~ 1, data=housing)

# use drop1
```

```
add1(int.mod, scope= ~ CRIM + ZN + INDUS + CHAS +
        NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT,
      test="F")
```

```
## Single term additions
##
## Model:
## MEDV ~ 1
##         Df Sum of Sq   RSS    AIC F value    Pr(>F)
## <none>               45773 2281.5
## CRIM     1    6679.2 39094 2203.7  86.108 < 2.2e-16 ***
## ZN       1    5270.5 40503 2221.6  65.583 4.223e-15 ***
## INDUS    1    9944.0 35829 2159.6 139.879 < 2.2e-16 ***
## CHAS     1    1233.7 44540 2269.7  13.961  0.000208 ***
## NOX      1    7806.5 37967 2188.9 103.629 < 2.2e-16 ***
## RM       1   20406.5 25367 1984.8 405.446 < 2.2e-16 ***
## AGE      1    6155.7 39618 2210.4  78.310 < 2.2e-16 ***
## DIS      1    2532.3 43241 2254.7  29.515 8.650e-08 ***
## RAD      1    6681.8 39092 2203.7  86.147 < 2.2e-16 ***
## TAX      1    9768.3 36005 2162.0 136.737 < 2.2e-16 ***
## PTRATIO  1    9315.2 36458 2168.4 128.773 < 2.2e-16 ***
## B        1    4682.8 41091 2228.9  57.438 1.686e-13 ***
## LSTAT    1   14488.1 31285 2090.9 233.400 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A lot of p-values are very significant `< 2.2e-16` so look to the F-statistics and add the term with the largest F-statistic: `RM`.

```
mod2 <- lm(MEDV ~ RM, data=housing)

# use add1
add1(mod2, scope= ~ CRIM + ZN + INDUS + CHAS +
        NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT,
      test="F")
```

```
## Single term additions
##
## Model:
## MEDV ~ RM
##         Df Sum of Sq   RSS    AIC F value    Pr(>F)
## <none>               25367 1984.8
## CRIM     1    2653.7 22713 1930.9 58.7672 9.226e-14 ***
## ZN       1     905.3 24462 1968.4 18.6145 1.927e-05 ***
## INDUS    1    2256.2 23111 1939.7 49.1062 7.828e-12 ***
## CHAS     1     500.2 24867 1976.7 10.1184  0.001559 **
## NOX      1    2252.2 23115 1939.8 49.0111 8.181e-12 ***
## AGE      1    2103.6 23263 1943.0 45.4840 4.238e-11 ***
## DIS      1     493.8 24873 1976.9  9.9851  0.001673 **
## RAD      1    2757.3 22610 1928.6 61.3415 2.868e-14 ***
## TAX      1    3536.5 21830 1910.8 81.4860 < 2.2e-16 ***
## PTRATIO  1    2602.6 22764 1932.0 57.5062 1.639e-13 ***
## B        1    2558.1 22809 1933.0 56.4132 2.700e-13 ***
## LSTAT    1    2463.7 22903 1935.1 54.1083 7.765e-13 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

TAX is the most significant so add this in and repeat:

```
mod2 <- lm(MEDV ~ RM + TAX, data=housing)

# use add1
add1(mod2, scope= ~ CRIM + ZN + INDUS + CHAS +
        NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT,
      test="F")
```

```
## Single term additions
##
## Model:
## MEDV ~ RM + TAX
##         Df Sum of Sq   RSS    AIC F value    Pr(>F)
## <none>               21830 1910.8
## CRIM     1    493.35 21337 1901.3 11.6072 0.0007097 ***
## ZN       1    256.96 21573 1906.9  5.9793 0.0148175 *
## INDUS    1     84.41 21746 1910.9  1.9486 0.1633531
## CHAS     1    474.54 21356 1901.7 11.1548 0.0009005 ***
## NOX      1    160.04 21670 1909.1  3.7075 0.0547336 .
## AGE      1    414.30 21416 1903.2  9.7113 0.0019363 **
## DIS      1     80.40 21750 1911.0  1.8558 0.1737237
## RAD      1     12.99 21817 1912.5  0.2989 0.5847854
## PTRATIO  1    940.27 20890 1890.6 22.5952 2.616e-06 ***
## B        1    783.10 21047 1894.4 18.6779 1.867e-05 ***
## LSTAT    1    787.73 21043 1894.2 18.7923 1.762e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Then add PTRATIO... continue to repeat this process until none of the variables can be added anymore. (I've included the code for the rest, but not the tables.)

```
mod2 <- lm(MEDV ~ RM + TAX + PTRATIO, data=housing)

# use add1
add1(mod2, scope= ~ CRIM + ZN + INDUS + CHAS +
        NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT,
      test="F")
```

```
## Single term additions
##
## Model:
## MEDV ~ RM + TAX + PTRATIO
##         Df Sum of Sq   RSS    AIC F value    Pr(>F)
## <none>               20890 1890.6
## CRIM     1    476.04 20414 1880.9 11.6829 0.0006821 ***
## ZN       1     71.14 20819 1890.8  1.7120 0.1913259
## INDUS    1     66.95 20823 1890.9  1.6108 0.2049698
## CHAS     1    348.73 20541 1884.0  8.5055 0.0036996 **
## NOX      1    377.86 20512 1883.3  9.2290 0.0025064 **
## AGE      1    386.17 20504 1883.1  9.4358 0.0022436 **
## DIS      1     74.73 20815 1890.8  1.7986 0.1804906
## RAD      1     65.16 20825 1891.0  1.5676 0.2111468
## B        1    823.13 20067 1872.2 20.5507 7.272e-06 ***
## LSTAT    1    582.98 20307 1878.2 14.3829 0.0001673 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod2 <- lm(MEDV ~ RM + TAX + PTRATIO + B, data=housing)

# use add1
add1(mod2, scope= ~ CRIM + ZN + INDUS + CHAS +
        NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT,
      test="F")
```

```
## Single term additions
##
## Model:
## MEDV ~ RM + TAX + PTRATIO + B
##         Df Sum of Sq   RSS    AIC F value   Pr(>F)
## <none>               20067 1872.2
## CRIM     1    291.27 19776 1866.8  7.3644 0.006882 **
## ZN       1     47.09 20020 1873.0  1.1760 0.278688
## INDUS    1     38.13 20029 1873.3  0.9519 0.329700
## CHAS     1    306.46 19761 1866.4  7.7545 0.005561 **
## NOX      1    251.83 19815 1867.8  6.3544 0.012019 *
## AGE      1    315.16 19752 1866.2  7.9781 0.004924 **
## DIS      1    117.86 19949 1871.2  2.9541 0.086281 .
## RAD      1    130.99 19936 1870.9  3.2852 0.070506 .
## LSTAT    1    359.22 19708 1865.1  9.1136 0.002667 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod2 <- lm(MEDV ~ RM + TAX + PTRATIO + B + LSTAT, data=housing)

# use add1
add1(mod2, scope= ~ CRIM + ZN + INDUS + CHAS +
        NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT,
      test="F")
```

```
## Single term additions
##
## Model:
## MEDV ~ RM + TAX + PTRATIO + B + LSTAT
##         Df Sum of Sq   RSS    AIC F value   Pr(>F)
## <none>               19708 1865.1
## CRIM     1    213.67 19494 1861.6  5.4694 0.019746 *
## ZN       1     13.93 19694 1866.7  0.3531 0.552652
## INDUS    1      2.61 19705 1867.0  0.0661 0.797196
## CHAS     1    318.31 19389 1858.8  8.1919 0.004384 **
## NOX      1    104.90 19603 1864.4  2.6703 0.102865
## AGE      1    134.02 19574 1863.6  3.4167 0.065132 .
## DIS      1    298.43 19409 1859.4  7.6724 0.005816 **
## RAD      1    125.45 19582 1863.9  3.1968 0.074391 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod2 <- lm(MEDV ~ RM + TAX + PTRATIO + B + LSTAT + CHAS, data=housing)

# use add1
add1(mod2, scope= ~ CRIM + ZN + INDUS + CHAS +
```

```
        NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT,
      test="F")
```

```
## Single term additions
##
## Model:
## MEDV ~ RM + TAX + PTRATIO + B + LSTAT + CHAS
##         Df Sum of Sq   RSS    AIC F value  Pr(>F)
## <none>               19389 1858.8
## CRIM     1   196.954 19192 1855.7  5.1105 0.02421 *
## ZN       1    32.937 19357 1860.0  0.8474 0.35774
## INDUS    1    21.094 19368 1860.3  0.5424 0.46180
## NOX      1   175.974 19213 1856.2  4.5611 0.03319 *
## AGE      1   202.948 19186 1855.5  5.2677 0.02214 *
## DIS      1   219.430 19170 1855.1  5.7004 0.01733 *
## RAD      1    98.417 19291 1858.3  2.5407 0.11158
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod2 <- lm(MEDV ~ RM + TAX + PTRATIO  + B + LSTAT + CHAS + DIS, data=housing)

# use add1
add1(mod2, scope= ~ CRIM + ZN + INDUS + CHAS +
        NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT,
      test="F")
```

```
## Single term additions
##
## Model:
## MEDV ~ RM + TAX + PTRATIO + B + LSTAT + CHAS + DIS
##         Df Sum of Sq   RSS    AIC F value    Pr(>F)
## <none>               19170 1855.1
## CRIM     1    219.92 18950 1851.2  5.7679  0.016688 *
## ZN       1    381.20 18789 1846.9 10.0834  0.001589 **
## INDUS    1    205.56 18964 1851.6  5.3871  0.020689 *
## NOX      1    834.26 18336 1834.6 22.6131 2.599e-06 ***
## AGE      1    852.09 18318 1834.1 23.1190 2.021e-06 ***
## RAD      1     92.66 19077 1854.6  2.4141  0.120886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod2 <- lm(MEDV ~ RM + TAX + PTRATIO  + B + LSTAT + CHAS + DIS + AGE, data=housing)

# use add1
add1(mod2, scope= ~ CRIM + ZN + INDUS + CHAS +
        NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT,
      test="F")
```

```
## Single term additions
##
## Model:
## MEDV ~ RM + TAX + PTRATIO + B + LSTAT + CHAS + DIS + AGE
##         Df Sum of Sq   RSS    AIC F value    Pr(>F)
## <none>               18318 1834.1
## CRIM     1    230.63 18087 1829.7  6.3246 0.0122225 *
## ZN       1    241.86 18076 1829.4  6.6366 0.0102784 *
```

```
## INDUS   1    121.90 18196 1832.7  3.3228 0.0689283 .
## NOX     1    450.03 17868 1823.5 12.4925 0.0004468 ***
## RAD     1     75.71 18242 1834.0  2.0586 0.1519788
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod2 <- lm(MEDV ~ RM + TAX + PTRATIO  + B + LSTAT + CHAS + DIS + AGE + NOX, data=housing)

# use add1
add1(mod2, scope= ~ CRIM + ZN + INDUS + CHAS +
        NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT,
      test="F")
```

```
## Single term additions
##
## Model:
## MEDV ~ RM + TAX + PTRATIO + B + LSTAT + CHAS + DIS + AGE + NOX
##        Df Sum of Sq   RSS    AIC F value   Pr(>F)
## <none>              17868 1823.5
## CRIM    1   272.122 17596 1817.7  7.6553 0.005872 **
## ZN      1   195.379 17673 1819.9  5.4725 0.019715 *
## INDUS   1    37.254 17831 1824.4  1.0342 0.309665
## RAD     1    97.485 17770 1822.7  2.7155 0.100013
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod2 <- lm(MEDV ~ RM + TAX + PTRATIO  + B + LSTAT + CHAS + DIS + AGE + NOX +
             CRIM, data=housing)

# use add1
add1(mod2, scope= ~ CRIM + ZN + INDUS + CHAS +
        NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT,
      test="F")
```

```
## Single term additions
##
## Model:
## MEDV ~ RM + TAX + PTRATIO + B + LSTAT + CHAS + DIS + AGE + NOX +
##     CRIM
##        Df Sum of Sq   RSS    AIC F value   Pr(>F)
## <none>              17596 1817.7
## ZN      1   230.610 17365 1813.0  6.5604 0.010723 *
## INDUS   1    67.041 17529 1817.8  1.8894 0.169896
## RAD     1   236.233 17360 1812.9  6.7225 0.009803 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod2 <- lm(MEDV ~ RM + TAX + PTRATIO  + B + LSTAT + CHAS + DIS + AGE + NOX +
             CRIM + RAD, data=housing)

# use add1
add1(mod2, scope= ~ CRIM + ZN + INDUS + CHAS +
        NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT,
      test="F")
```

```
## Single term additions
```

```
##
## Model:
## MEDV ~ RM + TAX + PTRATIO + B + LSTAT + CHAS + DIS + AGE + NOX +
##     CRIM + RAD
##        Df Sum of Sq   RSS    AIC F value   Pr(>F)
## <none>              17360 1812.9
## ZN      1   271.673 17088 1806.9  7.8380 0.005317 **
## INDUS   1    18.496 17341 1814.3  0.5258 0.468712
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod2 <- lm(MEDV ~ RM + TAX + PTRATIO  + B + LSTAT + CHAS + DIS + AGE + NOX +
             CRIM + RAD + ZN, data=housing)

# use add1
add1(mod2, scope= ~ CRIM + ZN + INDUS + CHAS +
        NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT,
      test="F")
```

```
## Single term additions
##
## Model:
## MEDV ~ RM + TAX + PTRATIO + B + LSTAT + CHAS + DIS + AGE + NOX +
##     CRIM + RAD + ZN
##        Df Sum of Sq   RSS    AIC F value Pr(>F)
## <none>              17088 1806.9
## INDUS   1    6.3526 17082 1808.7   0.183  0.669
```

So our final model based on this is the same as for backwards selection.

One thing that I want to draw your attention to that we'll come on to next week:

Notice that as you work through the forward selection process that some values switch between being non-significant and highly significant. This can be a sign of correlation between pairs of independent variables - or multicollinearity. Putting highly correlated covariates into a model together can cause serious problems with coefficient estimates because they try to balance each other out, and therefore don't accurately reflect the true association - these large changes in significance, or large changes in coefficient values are red flags for this issue.

Part vi)

```
# use int.mod from earlier

step(int.mod, scope= ~ CRIM + ZN + INDUS + CHAS +
        NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT)
```

```
## Start:  AIC=2281.49
## MEDV ~ 1
##
##           Df Sum of Sq   RSS    AIC
## + RM       1   20406.5 25367 1984.8
## + LSTAT    1   14488.1 31285 2090.9
## + INDUS    1    9944.0 35829 2159.6
## + TAX      1    9768.3 36005 2162.0
## + PTRATIO  1    9315.2 36458 2168.4
## + NOX      1    7806.5 37967 2188.9
## + RAD      1    6681.8 39092 2203.7
## + CRIM     1    6679.2 39094 2203.7
```

```
## + AGE      1    6155.7 39618 2210.4
## + ZN       1    5270.5 40503 2221.6
## + B        1    4682.8 41091 2228.9
## + DIS      1    2532.3 43241 2254.7
## + CHAS     1    1233.7 44540 2269.7
## <none>                 45773 2281.5
##
## Step:  AIC=1984.82
## MEDV ~ RM
##
##           Df Sum of Sq   RSS    AIC
## + TAX      1    3536.5 21830 1910.8
## + RAD      1    2757.3 22610 1928.6
## + CRIM     1    2653.7 22713 1930.9
## + PTRATIO  1    2602.6 22764 1932.0
## + B        1    2558.1 22809 1933.0
## + LSTAT    1    2463.7 22903 1935.1
## + INDUS    1    2256.2 23111 1939.7
## + NOX      1    2252.2 23115 1939.8
## + AGE      1    2103.6 23263 1943.0
## + ZN       1     905.3 24462 1968.4
## + CHAS     1     500.2 24867 1976.7
## + DIS      1     493.8 24873 1976.9
## <none>                 25367 1984.8
## - RM       1   20406.5 45773 2281.5
##
## Step:  AIC=1910.85
## MEDV ~ RM + TAX
##
##           Df Sum of Sq   RSS    AIC
## + PTRATIO  1     940.3 20890 1890.6
## + LSTAT    1     787.7 21043 1894.2
## + B        1     783.1 21047 1894.4
## + CRIM     1     493.4 21337 1901.3
## + CHAS     1     474.5 21356 1901.7
## + AGE      1     414.3 21416 1903.2
## + ZN       1     257.0 21573 1906.9
## + NOX      1     160.0 21670 1909.1
## <none>                 21830 1910.8
## + INDUS    1      84.4 21746 1910.9
## + DIS      1      80.4 21750 1911.0
## + RAD      1      13.0 21817 1912.5
## - TAX      1    3536.5 25367 1984.8
## - RM       1   14174.7 36005 2162.0
##
## Step:  AIC=1890.57
## MEDV ~ RM + TAX + PTRATIO
##
##           Df Sum of Sq   RSS    AIC
## + B        1     823.1 20067 1872.2
## + LSTAT    1     583.0 20307 1878.2
## + CRIM     1     476.0 20414 1880.9
## + AGE      1     386.2 20504 1883.1
## + NOX      1     377.9 20512 1883.3
```

```
## + CHAS     1     348.7 20541 1884.0
## <none>                 20890 1890.6
## + DIS      1      74.7 20815 1890.8
## + ZN       1      71.1 20819 1890.8
## + INDUS    1      66.9 20823 1890.9
## + RAD      1      65.2 20825 1891.0
## - PTRATIO  1     940.3 21830 1910.8
## - TAX      1    1874.2 22764 1932.0
## - RM       1   11651.0 32541 2112.8
##
## Step:  AIC=1872.23
## MEDV ~ RM + TAX + PTRATIO + B
##
##           Df Sum of Sq   RSS    AIC
## + LSTAT    1     359.2 19708 1865.1
## + AGE      1     315.2 19752 1866.2
## + CHAS     1     306.5 19760 1866.4
## + CRIM     1     291.3 19776 1866.8
## + NOX      1     251.8 19815 1867.8
## + RAD      1     131.0 19936 1870.9
## + DIS      1     117.9 19949 1871.2
## <none>                 20067 1872.2
## + ZN       1      47.1 20020 1873.0
## + INDUS    1      38.1 20029 1873.3
## - TAX      1     783.6 20851 1889.6
## - B        1     823.1 20890 1890.6
## - PTRATIO  1     980.3 21047 1894.4
## - RM       1   11626.3 31693 2101.5
##
## Step:  AIC=1865.09
## MEDV ~ RM + TAX + PTRATIO + B + LSTAT
##
##           Df Sum of Sq   RSS    AIC
## + CHAS     1     318.3 19389 1858.8
## + DIS      1     298.4 19409 1859.4
## + CRIM     1     213.7 19494 1861.6
## + AGE      1     134.0 19574 1863.6
## + RAD      1     125.5 19582 1863.9
## + NOX      1     104.9 19603 1864.4
## <none>                 19708 1865.1
## + ZN       1      13.9 19694 1866.7
## + INDUS    1       2.6 19705 1867.0
## - LSTAT    1     359.2 20067 1872.2
## - TAX      1     504.2 20212 1875.9
## - B        1     599.4 20307 1878.2
## - PTRATIO  1     801.0 20509 1883.2
## - RM       1    7484.7 27192 2026.0
##
## Step:  AIC=1858.85
## MEDV ~ RM + TAX + PTRATIO + B + LSTAT + CHAS
##
##           Df Sum of Sq   RSS    AIC
## + DIS      1     219.4 19170 1855.1
## + AGE      1     202.9 19186 1855.5
```

```
## + CRIM     1     197.0 19192 1855.7
## + NOX      1     176.0 19213 1856.2
## + RAD      1      98.4 19291 1858.3
## <none>                 19389 1858.8
## + ZN       1      32.9 19357 1860.0
## + INDUS    1      21.1 19368 1860.3
## - CHAS     1     318.3 19708 1865.1
## - LSTAT    1     371.1 19761 1866.4
## - TAX      1     535.5 19925 1870.6
## - B        1     560.6 19950 1871.3
## - PTRATIO  1     687.8 20077 1874.5
## - RM       1    7285.1 26675 2018.2
##
## Step:  AIC=1855.09
## MEDV ~ RM + TAX + PTRATIO + B + LSTAT + CHAS + DIS
##
##           Df Sum of Sq   RSS    AIC
## + AGE      1     852.1 18318 1834.1
## + NOX      1     834.3 18336 1834.6
## + ZN       1     381.2 18789 1846.9
## + CRIM     1     219.9 18950 1851.2
## + INDUS    1     205.6 18964 1851.6
## + RAD      1      92.7 19077 1854.6
## <none>                 19170 1855.1
## - DIS      1     219.4 19389 1858.8
## - CHAS     1     239.3 19409 1859.4
## - LSTAT    1     520.0 19690 1866.6
## - B        1     577.7 19748 1868.1
## - PTRATIO  1     657.9 19828 1870.2
## - TAX      1     727.3 19897 1871.9
## - RM       1    7014.3 26184 2010.9
##
## Step:  AIC=1834.08
## MEDV ~ RM + TAX + PTRATIO + B + LSTAT + CHAS + DIS + AGE
##
##           Df Sum of Sq   RSS    AIC
## + NOX      1     450.0 17868 1823.5
## + ZN       1     241.9 18076 1829.4
## + CRIM     1     230.6 18087 1829.7
## + INDUS    1     121.9 18196 1832.7
## + RAD      1      75.7 18242 1834.0
## <none>                 18318 1834.1
## - LSTAT    1     225.3 18543 1838.3
## - CHAS     1     301.2 18619 1840.3
## - TAX      1     561.4 18879 1847.4
## - B        1     602.6 18921 1848.5
## - PTRATIO  1     642.9 18961 1849.5
## - AGE      1     852.1 19170 1855.1
## - DIS      1     868.6 19186 1855.5
## - RM       1    7193.0 25511 1999.7
##
## Step:  AIC=1823.5
## MEDV ~ RM + TAX + PTRATIO + B + LSTAT + CHAS + DIS + AGE + NOX
##
```

```
##           Df Sum of Sq   RSS    AIC
## + CRIM     1     272.1 17596 1817.7
## + ZN       1     195.4 17673 1819.9
## + RAD      1      97.5 17770 1822.7
## <none>                 17868 1823.5
## + INDUS    1      37.3 17831 1824.4
## - TAX      1     141.7 18010 1825.5
## - LSTAT    1     157.0 18025 1825.9
## - CHAS     1     354.0 18222 1831.4
## - NOX      1     450.0 18318 1834.1
## - AGE      1     467.9 18336 1834.6
## - B        1     511.7 18380 1835.8
## - PTRATIO  1     932.1 18800 1847.2
## - DIS      1    1278.2 19146 1856.5
## - RM       1    6598.1 24466 1980.5
##
## Step:  AIC=1817.73
## MEDV ~ RM + TAX + PTRATIO + B + LSTAT + CHAS + DIS + AGE + NOX +
##     CRIM
##
##           Df Sum of Sq   RSS    AIC
## + RAD      1     236.2 17360 1812.9
## + ZN       1     230.6 17365 1813.0
## - TAX      1      26.2 17622 1816.5
## <none>                 17596 1817.7
## + INDUS    1      67.0 17529 1817.8
## - LSTAT    1     108.8 17705 1818.8
## - CRIM     1     272.1 17868 1823.5
## - CHAS     1     332.0 17928 1825.2
## - B        1     396.0 17992 1827.0
## - AGE      1     463.5 18059 1828.9
## - NOX      1     491.5 18087 1829.7
## - PTRATIO  1     953.7 18550 1842.4
## - DIS      1    1356.0 18952 1853.3
## - RM       1    6602.9 24199 1977.0
##
## Step:  AIC=1812.89
## MEDV ~ RM + TAX + PTRATIO + B + LSTAT + CHAS + DIS + AGE + NOX +
##     CRIM + RAD
##
##           Df Sum of Sq   RSS    AIC
## + ZN       1     271.7 17088 1806.9
## <none>                 17360 1812.9
## - LSTAT    1      91.6 17451 1813.5
## + INDUS    1      18.5 17341 1814.3
## - TAX      1     231.5 17591 1817.6
## - RAD      1     236.2 17596 1817.7
## - CHAS     1     285.9 17645 1819.2
## - CRIM     1     410.9 17770 1822.7
## - AGE      1     429.5 17789 1823.3
## - B        1     438.7 17798 1823.5
## - NOX      1     541.2 17901 1826.4
## - PTRATIO  1    1097.0 18457 1841.9
## - DIS      1    1370.5 18730 1849.3
```

```
## - RM       1    6073.9 23433 1962.7
##
## Step:  AIC=1806.91
## MEDV ~ RM + TAX + PTRATIO + B + LSTAT + CHAS + DIS + AGE + NOX +
##     CRIM + RAD + ZN
##
##           Df Sum of Sq   RSS    AIC
## <none>                 17088 1806.9
## - LSTAT    1     112.4 17200 1808.2
## + INDUS    1       6.4 17082 1808.7
## - ZN       1     271.7 17360 1812.9
## - RAD      1     277.3 17365 1813.0
## - CHAS     1     284.9 17373 1813.3
## - TAX      1     328.2 17416 1814.5
## - AGE      1     346.8 17435 1815.1
## - B        1     428.0 17516 1817.4
## - CRIM     1     470.6 17558 1818.7
## - NOX      1     489.6 17577 1819.2
## - PTRATIO  1     664.8 17753 1824.2
## - DIS      1    1640.0 18728 1851.3
## - RM       1    5388.9 22477 1943.6
##
## Call:
## lm(formula = MEDV ~ RM + TAX + PTRATIO + B + LSTAT + CHAS + DIS +
##     AGE + NOX + CRIM + RAD + ZN, data = housing)
##
## Coefficients:
## (Intercept)           RM          TAX      PTRATIO            B        LSTAT
##    17.43724      5.92495     -0.01294     -0.68797      0.01167     -0.08941
##        CHAS          DIS          AGE          NOX         CRIM          RAD
##     3.04772     -1.67992     -0.04958    -17.23013     -0.14961      0.22310
##          ZN
##     0.04745
```

The resulting model is identical to that produced by forwards and backwards selection - note that this will not always be the case!

part vii)

$$CP = \frac{RSS_p}{\hat{\sigma}_q^2} + 2(p+1) - n$$

$$PRESS = \sum_{i=1}^{n} \frac{\hat{\varepsilon}_i^2}{(1-h_{ii})^2}$$

```
CP_PRESS <- function(model, sigma_full){
  res <- resid(model)

  hat_mod <- hatvalues(model)

  CP <- sum(res^2)/sigma_full + 2*length(coef(model)) - length(res)

  PRESS <- sum(res^2/(1-hat_mod)^2)
```

```
  list(Cp=CP, PRESS=PRESS)

}

sigma_q <- summary(full.mod)$sigma^2

size8_stat <- CP_PRESS(size8, sigma_q)
size9_stat <- CP_PRESS(size9, sigma_q)
size10_stat <- CP_PRESS(size10, sigma_q)
sel_stat <- CP_PRESS(mod, sigma_q)
```

Model	p	adjr2	Cp	PRESS
Leaps - Size 8	7	59.93	30.91	18810.70
Leaps - Size 9	8	60.65	22.63	18610.79
Leaps - Size 10	9	60.95	19.67	18480.12
Selection	12	61.76	12.18	21017.75

I've also included adjusted R-squared, because it has a nice straightforward interpretation and sometimes it's easy to get caught up in Cp $<$ p. The points that I would want to make are:

- Not a great difference in adjusted r-squared across all of the models
- Cp is smallest in the 12 variable model from forwards selection, though in all of these CP $>$ p
- PRESS is smallest in the size 10 model

(in a report these should be a paragraph!)

Taking all of the above into consideration, I would choose the size 10 model, because I have a preference for adjusted R-squared over Cp (It's easier to determine what a "big" difference is), and the PRESS is smallest for the size 10 model, which indicates that it is a good predictor - the size 12 model chosen through model selection may be overfitting the data.

I do want to emphsise once more that none of the models above are "wrong" - you need to make a judgement base don the facts and justify it.

> *All models are wrong, but some are useful*
>
> - George E.P. Box

Written Example

(i) For forwards selection we add the variable at each step for which

$$F\text{-statistic} = \frac{RSS_{p-1} - RSS_p}{RSS_p/(n-p-1)}$$

is largest, and such that F-statistic $> F_{1,n-p-1}(\alpha)$.

- **Step 1:** Compare the model $Y = \beta_0 + \epsilon$ to each of $Y = \beta_0 + \beta_1 X_1 + \epsilon$, $Y = \beta_0 + \beta_2 X_2 + \epsilon$, $Y = \beta_0 + \beta_3 X_3 + \epsilon$ or $Y = \beta_0 + \beta_4 X_4 + \epsilon$. The associated F-statistics for these are already provided, and are 1.2, 4.6, 5.0 and 3.8. Only need to consider the largest of these which is 5. Critical value is $F_{1,84-1-1}(0.2) =$ `qf(0.80, 1, 82) = 1.662` < 5.0. Reject the null hypothesis $H_0 : \beta_3 = 0$, and add X_3 to the model.

- **Step 2:** Compare the single variable model $Y = \beta_0 + \beta_3 X_3 + \epsilon$ to models that include an additional X variable: ie. (X_1, X_3), (X_2, X_3), or (X_4, X_3). Need to obtain the largest F-statistic using the provided RSS values. The largest F-statistic (as defined above) occurs when the difference between $RSS_{p-1} - RSS_p$ is largest. Here $p = 2$ and this maximum happens when we compare $Y = \beta_0 + \beta_3 X_3 + \epsilon$ with $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \epsilon$:

$$F\text{-statistic} = \frac{RSS_{p-1} - RSS_p}{RSS_p/(n-p-1)} = \frac{40-9}{9/(84-2-1)} = 279$$

Critical value is $F_{1,84-2-1}(0.2) =$ `qf(0.80, 1, 81) = 1.6695` < 279. Reject the null hypothesis $H_0 : \beta_1 = 0$, and add X_1 to the model.

- **Step 3:** Compare the two variable model $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \epsilon$ to models that include an additional X variable: ie. (X_1, X_2, X_3) or (X_1, X_3, X_4). Need to obtain the largest F-statistic using the provided RSS values. The largest F-statistic (as defined above) occurs when the difference between $RSS_{p-1} - RSS_p$ is largest. Here $p = 3$ and this maximum happens when we compare $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \epsilon$ with $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$:

$$F\text{-statistic} = \frac{RSS_{p-1} - RSS_p}{RSS_p/(n-p-1)} = \frac{9-8.9}{8.9/(84-3-1} = 0.8989$$

Critical value is $F_{1,84-3-1}(0.2) =$ `qf(0.80, 1, 80) = 1.6698` > 0.8989. **Do not reject** the null hypothesis $H_0 : \beta_4 = 0$ and do not add X_4 to the model. Stop forward selection, concluding that the final model should be

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \epsilon$$

(ii) Using backward selection, we try to remove a single variable at each step in an attempt to simplify the model. We can either use the find the F-statistic that is largest when reducing the model with p variables to $p-1$ where

$$F\text{-statistic} = \frac{RSS_{p-1} - RSS_p}{RSS_p/(n-p-1)}$$

If F-statistic $< F_{1,n-p-1}(\alpha)$, **do not reject** the null hypothesis $H_0 : \beta_j = 0$, and consequently decide to removing X_j from the model. Equivalently, we can choose the variable with the t-statistic with smallest absolute value, where

$$t\text{-statistic} = \frac{\hat{\beta}_j}{\text{e.s.e}(\hat{\beta}_j)}$$

If $|t| < t_{n-p-1,\alpha/2}$ then do not reject $H_0 : \beta_j = 0$, and consequently remove X_j from the model. In this question it will be quicker to use the p value associated with the t-statistic as these are supplied in the above output. **Remember that the p value is the probability of finding a t-statistic with a more extreme absolute value under the null hypothesis - so if $p < \alpha_{stay}$, then reject** $H_0 : \beta_j = 0$

- **Step 1:** Begin with the largest model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$. Smallest t-statistic is associated with X_4. $p = 0.4507 > \alpha_{stay} = 0.1$. Therefore, **do not reject** $H_0 : \beta_4 = 0$, and remove X_4 from the model.

- **Step 2:** Next explore model excluding X_4: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$. Smallest t-statistic is associated with X_2. $p = 0.2095 > \alpha_{stay} = 0.1$. Therefore, **do not reject** $H_0 : \beta_2 = 0$, and remove X_2 from the model.

- **Step 3:** Next explore model excluding X_2, X_4: $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \epsilon$. Smallest t-statistic is associated with X_1. $p = 0.0265 < \alpha_{stay} = 0.1$. Therefore, **reject** $H_0 : \beta_1 = 0$, and conclude that X_1 cannot be removed from the model. **Stop** backwards selection and conclude that the final model should be

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \epsilon$$