SMM637 Answers - Multiple Linear Regression

1. (a)
```
library(faraway)
data(savings)

g <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data=savings)
summary(g)

Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2422 -2.6857 -0.2488  2.4280  9.7509

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared:  0.3385,    Adjusted R-squared:  0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```
The fitted model is:

$$\hat{y} = 28.5660865 - 0.4611931 pop15 - 1.6914977 pop75 - 0.0003369 dpi + 0.4096949 ddpi$$

(b) The F-statistics is 5.756 and P-value 0.0007904. We reject the null hypothesis that all coefficients associated with the covariates are zeros. This means that at least one predictor is linearly associated with the response variable.

(c) The P-value associated with pop15 is 0.002603.This means that we reject the null hypothesis its coefficient is zero. That is, pop15 explains part of the variability of sr.

(d) This can be obtained by using the `anova` function in R. Specifically:
```
g2 <- lm(sr ~  pop75 + dpi + ddpi, data=savings)
anova(g2, g)

Analysis of Variance Table

Model 1: sr ~ pop75 + dpi + ddpi
Model 2: sr ~ pop15 + pop75 + dpi + ddpi
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     46 797.72
2     45 650.71  1    147.01 10.167 0.002603 **
---
```

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

We can notice that we reject again the null hypothesis.

(e) The value of the F-test is twice that of the t-test.

(f)
```
library(MASS)
g0 <- lm(sr ~ 1, data=savings)

stepAIC(g, ~ pop15 + pop75 + dpi + ddpi, data=savings)
Start:  AIC=138.3
sr ~ pop15 + pop75 + dpi + ddpi

        Df Sum of Sq    RSS    AIC
- dpi    1     1.893 652.61 136.45
<none>               650.71 138.30
- pop75  1    35.236 685.95 138.94
- ddpi   1    63.054 713.77 140.93
- pop15  1   147.012 797.72 146.49

Step:  AIC=136.45
sr ~ pop15 + pop75 + ddpi

        Df Sum of Sq    RSS    AIC
<none>               652.61 136.45
- pop75  1    47.946 700.55 137.99
+ dpi    1     1.893 650.71 138.30
- ddpi   1    73.562 726.17 139.79
- pop15  1   145.789 798.40 144.53

Call:
lm(formula = sr ~ pop15 + pop75 + ddpi, data = savings)

Coefficients:
(Intercept)        pop15        pop75         ddpi
    28.1247      -0.4518      -1.8354       0.4278


stepAIC(g0, ~ pop15 + pop75 + dpi + ddpi, data=savings)
Start:  AIC=150.96
sr ~ 1

        Df Sum of Sq    RSS    AIC
+ pop15  1   204.118 779.51 141.33
+ pop75  1    98.545 885.08 147.68
+ ddpi   1    91.374 892.25 148.09
+ dpi    1    47.763 935.87 150.47
<none>               983.63 150.96

Step:  AIC=141.33
sr ~ pop15
```

```
        Df Sum of Sq    RSS    AIC
+ ddpi   1     78.959 700.55 137.99
+ pop75  1     53.343 726.17 139.79
+ dpi    1     35.387 744.12 141.01
<none>                779.51 141.33
- pop15  1    204.118 983.63 150.96

Step:  AIC=137.99
sr ~ pop15 + ddpi

        Df Sum of Sq    RSS    AIC
+ pop75  1     47.946 652.61 136.45
<none>                700.55 137.99
+ dpi    1     14.603 685.95 138.94
- ddpi   1     78.959 779.51 141.33
- pop15  1    191.702 892.25 148.09

Step:  AIC=136.45
sr ~ pop15 + ddpi + pop75

        Df Sum of Sq    RSS    AIC
<none>                652.61 136.45
- pop75  1     47.946 700.55 137.99
+ dpi    1      1.893 650.71 138.30
- ddpi   1     73.562 726.17 139.79
- pop15  1    145.789 798.40 144.53

Call:
lm(formula = sr ~ pop15 + ddpi + pop75, data = savings)

Coefficients:
(Intercept)        pop15         ddpi        pop75
    28.1247      -0.4518       0.4278      -1.8354
```

Both procedures reach the same model:

$$Y = \alpha_0 + \beta_1 pop15 + \beta_2 pop75 + \beta_3 ddpi + \epsilon$$

(g)
```
g3 <- lm(sr ~ pop15 + pop75 + ddpi, data=savings)
summary(g3)

Call:
lm(formula = sr ~ pop15 + pop75 + ddpi, data = savings)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2539 -2.6159 -0.3913  2.3344  9.7070
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  28.1247     7.1838   3.915 0.000297 ***
pop15        -0.4518     0.1409  -3.206 0.002452 **
pop75        -1.8354     0.9984  -1.838 0.072473 .
ddpi          0.4278     0.1879   2.277 0.027478 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.767 on 46 degrees of freedom
Multiple R-squared:  0.3365,    Adjusted R-squared:  0.2933
F-statistic: 7.778 on 3 and 46 DF,  p-value: 0.0002646
```

In terms of estimates coefficients and of significance, the results are very similar. The only difference is that the coefficient associated with pop75 is more significant.

(h) By looking at the scatterplots pop15 and pop75 appear highly correlated which means that their standard errors will be inflated. There is some minor correlation between pop15 and dpi and pop75 and dpi. Based on this we could fit another model where pop75 is excluded from the model:

```
g4 <- lm(sr ~ pop15 + dpi + ddpi, data=savings)
summary(g4)

Call:
lm(formula = sr ~ pop15 + dpi + ddpi, data = savings)

Residuals:
    Min      1Q  Median      3Q     Max
-7.6889 -2.8813  0.0296  1.7989 10.4330

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.2771687  4.3888974   4.392 6.53e-05 ***
pop15       -0.2883861  0.0945354  -3.051  0.00378 **
dpi         -0.0008704  0.0008795  -0.990  0.32755
ddpi         0.3929355  0.1989390   1.975  0.05427 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.862 on 46 degrees of freedom
Multiple R-squared:  0.3026,    Adjusted R-squared:  0.2572
F-statistic: 6.654 on 3 and 46 DF,  p-value: 0.0007941
```

Because dpi is not significant we could fit a more parsimonius model:

```
g5 <- lm(sr ~ pop15 + ddpi, data=savings)
> summary(g5)

Call:
lm(formula = sr ~ pop15 + ddpi, data = savings)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-7.5831 -2.8632  0.0453  2.2273 10.4753

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.59958    2.33439   6.682 2.48e-08 ***
pop15       -0.21638    0.06033  -3.586 0.000796 ***
ddpi         0.44283    0.19240   2.302 0.025837 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.861 on 47 degrees of freedom
Multiple R-squared:  0.2878,    Adjusted R-squared:  0.2575
F-statistic: 9.496 on 2 and 47 DF,  p-value: 0.0003438
```

Compring this last model with the first in (a), we can see that the estimated coefficient of pop15 are dissimilar. The effects goes from -0.4611931 to -0.21638. This difference is due to the fact the before there was pop75 which was strongly correlated with pop15.