STAT 420 Fall 2012

Homework #7

(due Friday, October 26, by 3:00 p.m.)

1. For the `prostate` data, fit a model with `lpsa` as the response and the other variables as predictors.

a) Implement the Backward Elimination variable selection method to determine the "best" model. Use $\alpha_{crit} = 0.10$.

```
> library(faraway)
> data(prostate)
> attach(prostate)
> fit = lm(lpsa~lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45)
> summary(fit)

Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
    gleason + pgg45)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7331 -0.3713 -0.0170  0.4141  1.6381

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.669337   1.296387   0.516  0.60693
lcavol       0.587022   0.087920   6.677 2.11e-09 ***
lweight      0.454467   0.170012   2.673  0.00896 **
age         -0.019637   0.011173  -1.758  0.08229 .
lbph         0.107054   0.058449   1.832  0.07040 .
svi          0.766157   0.244309   3.136  0.00233 **
lcp         -0.105474   0.091013  -1.159  0.24964
gleason      0.045142   0.157465   0.287  0.77503
pgg45        0.004525   0.004421   1.024  0.30886
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7084 on 88 degrees of freedom
Multiple R-squared: 0.6548,     Adjusted R-squared: 0.6234
F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

`gleason` is the least significant variable, p-value = 0.77503.

```
> fit1 = update(fit, .~. - gleason)
> summary(fit1)

Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
    pgg45)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.73117 -0.38137 -0.01728  0.43364  1.63513 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.953926   0.829439   1.150  0.25319    
lcavol       0.591615   0.086001   6.879 8.07e-10 ***
lweight      0.448292   0.167771   2.672  0.00897 ** 
age         -0.019336   0.011066  -1.747  0.08402 .  
lbph         0.107671   0.058108   1.853  0.06720 .  
svi          0.757734   0.241282   3.140  0.00229 ** 
lcp         -0.104482   0.090478  -1.155  0.25127    
pgg45        0.005318   0.003433   1.549  0.12488    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.7048 on 89 degrees of freedom
Multiple R-squared: 0.6544,     Adjusted R-squared: 0.6273 
F-statistic: 24.08 on 7 and 89 DF,  p-value: < 2.2e-16
```

`lcp` is the least significant variable, p-value = 0.25127.

```
> fit1 = update(fit1, .~. - lcp)
> summary(fit1)

Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + pgg45)

Residuals:
       Min         1Q     Median         3Q        Max 
-1.777e+00 -4.171e-01  1.733e-05  4.068e-01  1.597e+00 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.980085   0.830665   1.180  0.24116    
lcavol       0.545770   0.076431   7.141 2.31e-10 ***
lweight      0.449450   0.168078   2.674  0.00890 ** 
age         -0.017470   0.010967  -1.593  0.11469    
lbph         0.105755   0.058191   1.817  0.07249 .  
svi          0.641666   0.219757   2.920  0.00442 ** 
pgg45        0.003528   0.003068   1.150  0.25331    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.7061 on 90 degrees of freedom
Multiple R-squared: 0.6493,     Adjusted R-squared: 0.6259 
F-statistic: 27.77 on 6 and 90 DF,  p-value: < 2.2e-16
```

`pgg45` is the least significant variable, p-value = 0.25331.

```
> fit1 = update(fit1, .~. - pgg45)
> summary(fit1)

Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi)

Residuals:
      Min        1Q    Median        3Q       Max 
-1.835049 -0.393961  0.004139  0.463365  1.578879 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.95100    0.83175   1.143 0.255882    
lcavol       0.56561    0.07459   7.583 2.77e-11 ***
lweight      0.42369    0.16687   2.539 0.012814 *  
age         -0.01489    0.01075  -1.385 0.169528    
lbph         0.11184    0.05805   1.927 0.057160 .  
svi          0.72095    0.20902   3.449 0.000854 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.7073 on 91 degrees of freedom
Multiple R-squared: 0.6441,     Adjusted R-squared: 0.6245 
F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

`age` is the least significant variable, p-value = 0.169528.

```
> fit1 = update(fit1, .~. - age)
> summary(fit1)

Call:
lm(formula = lpsa ~ lcavol + lweight + lbph + svi)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.82653 -0.42270  0.04362  0.47041  1.48530 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.14554    0.59747   0.244  0.80809    
lcavol       0.54960    0.07406   7.422 5.64e-11 ***
lweight      0.39088    0.16600   2.355  0.02067 *  
lbph         0.09009    0.05617   1.604  0.11213    
svi          0.71174    0.20996   3.390  0.00103 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.7108 on 92 degrees of freedom
Multiple R-squared: 0.6366,     Adjusted R-squared: 0.6208 
F-statistic: 40.29 on 4 and 92 DF,  p-value: < 2.2e-16
```

`lbph` is the least significant variable, p-value = 0.11213.

```
> fit1 = update(fit1, .~. - lbph)
> summary(fit1)

Call:
lm(formula = lpsa ~ lcavol + lweight + svi)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.72964 -0.45764  0.02812  0.46403  1.57013 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.26809    0.54350  -0.493  0.62298    
lcavol       0.55164    0.07467   7.388  6.3e-11 ***
lweight      0.50854    0.15017   3.386  0.00104 ** 
svi          0.66616    0.20978   3.176  0.00203 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.7168 on 93 degrees of freedom
Multiple R-squared: 0.6264,     Adjusted R-squared: 0.6144 
F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

All are significant at α_{crit}.

"Best" model:

lpsa ~ lcavol + lweight + svi

```
> anova(fit1,fit)
Analysis of Variance Table

Model 1: lpsa ~ lcavol + lweight + svi
Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + 
    pgg45
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     93 47.785                           
2     88 44.163  5     3.622 1.4434 0.2167
```

b) Implement the AIC variable selection method (any) to determine the "best" model.

```
> step(fit, direction = "backward")
Start:  AIC=-58.32
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45

          Df Sum of Sq    RSS     AIC
- gleason  1     0.041 44.204 -60.231
- pgg45    1     0.526 44.689 -59.174
- lcp      1     0.674 44.837 -58.853
<none>                 44.163 -58.322
- age      1     1.550 45.713 -56.975
- lbph     1     1.684 45.847 -56.693
- lweight  1     3.586 47.749 -52.749
- svi      1     4.936 49.099 -50.046
- lcavol   1    22.372 66.535 -20.567

Step:  AIC=-60.23
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45

          Df Sum of Sq    RSS     AIC
- lcp      1     0.662 44.867 -60.789
<none>                 44.204 -60.231
- pgg45    1     1.192 45.396 -59.650
- age      1     1.517 45.721 -58.959
- lbph     1     1.705 45.910 -58.560
- lweight  1     3.546 47.750 -54.746
- svi      1     4.898 49.103 -52.037
- lcavol   1    23.504 67.708 -20.872

Step:  AIC=-60.79
lpsa ~ lcavol + lweight + age + lbph + svi + pgg45

          Df Sum of Sq    RSS     AIC
- pgg45    1     0.659 45.526 -61.374
<none>                 44.867 -60.789
- age      1     1.265 46.131 -60.092
- lbph     1     1.647 46.513 -59.293
- lweight  1     3.565 48.431 -55.373
- svi      1     4.250 49.117 -54.009
- lcavol   1    25.419 70.285 -19.248

Step:  AIC=-61.37
lpsa ~ lcavol + lweight + age + lbph + svi

          Df Sum of Sq    RSS     AIC
<none>                 45.526 -61.374
- age      1     0.959 46.485 -61.352
- lbph     1     1.857 47.382 -59.497
- lweight  1     3.225 48.751 -56.735
- svi      1     5.952 51.477 -51.456
- lcavol   1    28.767 74.292 -15.871
```

```
Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi)

Coefficients:
(Intercept)       lcavol      lweight          age         lbph          svi
    0.95100      0.56561      0.42369     -0.01489      0.11184      0.72095
```

OR

```
> step(fit, direction = "both")
Start:  AIC=-58.32
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45

          Df Sum of Sq    RSS     AIC
- gleason  1     0.041 44.204 -60.231
- pgg45    1     0.526 44.689 -59.174
- lcp      1     0.674 44.837 -58.853
<none>                 44.163 -58.322
- age      1     1.550 45.713 -56.975
- lbph     1     1.684 45.847 -56.693
- lweight  1     3.586 47.749 -52.749
- svi      1     4.936 49.099 -50.046
- lcavol   1    22.372 66.535 -20.567

Step:  AIC=-60.23
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45

          Df Sum of Sq    RSS     AIC
- lcp      1     0.662 44.867 -60.789
<none>                 44.204 -60.231
- pgg45    1     1.192 45.396 -59.650
- age      1     1.517 45.721 -58.959
- lbph     1     1.705 45.910 -58.560
+ gleason  1     0.041 44.163 -58.322
- lweight  1     3.546 47.750 -54.746
- svi      1     4.898 49.103 -52.037
- lcavol   1    23.504 67.708 -20.872

Step:  AIC=-60.79
lpsa ~ lcavol + lweight + age + lbph + svi + pgg45

          Df Sum of Sq    RSS     AIC
- pgg45    1     0.659 45.526 -61.374
<none>                 44.867 -60.789
+ lcp      1     0.662 44.204 -60.231
- age      1     1.265 46.131 -60.092
- lbph     1     1.647 46.513 -59.293
+ gleason  1     0.030 44.837 -58.853
- lweight  1     3.565 48.431 -55.373
- svi      1     4.250 49.117 -54.009
- lcavol   1    25.419 70.285 -19.248
```

```
Step:  AIC=-61.37
lpsa ~ lcavol + lweight + age + lbph + svi

          Df Sum of Sq    RSS     AIC
<none>                 45.526 -61.374
- age      1     0.959 46.485 -61.352
+ pgg45    1     0.659 44.867 -60.789
+ gleason  1     0.456 45.070 -60.351
+ lcp      1     0.129 45.396 -59.650
- lbph     1     1.857 47.382 -59.497
- lweight  1     3.225 48.751 -56.735
- svi      1     5.952 51.477 -51.456
- lcavol   1    28.767 74.292 -15.871

Call:
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi)

Coefficients:
(Intercept)       lcavol      lweight          age         lbph          svi
    0.95100      0.56561      0.42369     -0.01489      0.11184      0.72095
```

OR

```
> step(lm(lpsa ~ 1), lpsa~lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45,
direction = "forward")
Start:  AIC=28.84
lpsa ~ 1

          Df Sum of Sq     RSS     AIC
+ lcavol   1    69.003  58.915 -44.366
+ svi      1    41.011  86.907  -6.658
+ lcp      1    38.528  89.389  -3.926
+ pgg45    1    22.814 105.103  11.783
+ gleason  1    17.416 110.501  16.641
+ lweight  1    16.041 111.876  17.840
+ lbph     1     4.136 123.782  27.650
+ age      1     3.679 124.238  28.007
<none>                 127.918  28.837

Step:  AIC=-44.37
lpsa ~ lcavol

          Df Sum of Sq    RSS     AIC
+ lweight  1     5.949 52.966 -52.690
+ svi      1     5.237 53.677 -51.397
+ lbph     1     3.266 55.649 -47.898
+ pgg45    1     1.698 57.217 -45.203
<none>                 58.915 -44.366
+ lcp      1     0.656 58.259 -43.453
+ gleason  1     0.416 58.499 -43.053
+ age      1     0.003 58.912 -42.370
```

```
Step:  AIC=-52.69
lpsa ~ lcavol + lweight

          Df Sum of Sq    RSS     AIC
+ svi      1     5.181 47.785 -60.676
+ pgg45    1     1.949 51.017 -54.327
<none>                 52.966 -52.690
+ lcp      1     0.837 52.129 -52.236
+ gleason  1     0.781 52.185 -52.131
+ lbph     1     0.675 52.291 -51.935
+ age      1     0.420 52.546 -51.463

Step:  AIC=-60.68
lpsa ~ lcavol + lweight + svi

          Df Sum of Sq    RSS     AIC
+ lbph     1     1.300 46.485 -61.352
<none>                 47.785 -60.676
+ pgg45    1     0.573 47.211 -59.847
+ age      1     0.403 47.382 -59.497
+ gleason  1     0.389 47.396 -59.469
+ lcp      1     0.064 47.721 -58.806

Step:  AIC=-61.35
lpsa ~ lcavol + lweight + svi + lbph

          Df Sum of Sq    RSS     AIC
+ age      1     0.959 45.526 -61.374
<none>                 46.485 -61.352
+ pgg45    1     0.353 46.131 -60.092
+ gleason  1     0.213 46.272 -59.796
+ lcp      1     0.102 46.383 -59.565

Step:  AIC=-61.37
lpsa ~ lcavol + lweight + svi + lbph + age

          Df Sum of Sq    RSS     AIC
<none>                 45.526 -61.374
+ pgg45    1     0.659 44.867 -60.789
+ gleason  1     0.456 45.070 -60.351
+ lcp      1     0.129 45.396 -59.650

Call:
lm(formula = lpsa ~ lcavol + lweight + svi + lbph + age)

Coefficients:
(Intercept)       lcavol      lweight          svi         lbph          age
    0.95100      0.56561      0.42369      0.72095      0.11184     -0.01489
```

"Best" model:

lpsa ~ lcavol + lweight + age + lbph + svi

```
> fit2 = lm(lpsa ~ lcavol + lweight + age + lbph + svi)
> anova(fit2,fit)
Analysis of Variance Table

Model 1: lpsa ~ lcavol + lweight + age + lbph + svi
Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     91 45.526
2     88 44.163  3     1.363 0.905 0.4421
```

c) Compare the values of Adjusted R^2 for the full model, the "best" model from part (a), and the "best" model from part (b). Which model is the "best" model out of the three? ***Justify your answer***.

```
> summary(fit)$adj.r.squared           – full
[1] 0.6233681
> summary(fit1)$adj.r.squared          – part (a)
[1] 0.6143899
> summary(fit2)$adj.r.squared          – part (b)
[1] 0.6245476                                        – largest of the three
```

"Best" model:

`lpsa ~ lcavol + lweight + age + lbph + svi`,

the "best" model from part (b).

2. A survey was conducted to study teenage gambling in Britain. (Ide-Smith & Lea, 1988, Journal of Gambling Behavior, 4, 110-118) The data is stored in the data frame `teengamb` (library `faraway`). This data frame contains the following columns:

`sex`	0 = male, 1 = female,
`status`	Socioeconomic status score based on parents' occupation,
`income`	in pounds per week,
`verbal`	verbal score in words out of 12 correctly defined,
`gamble`	expenditure on gambling in pounds per year.

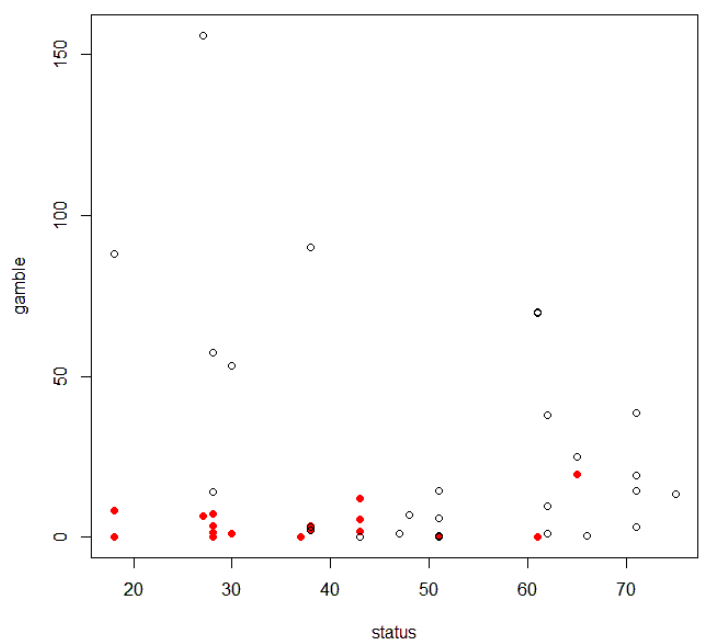
```
> library(faraway)
> data(teengamb)
```

The data are also stored in `teengamb.csv`.

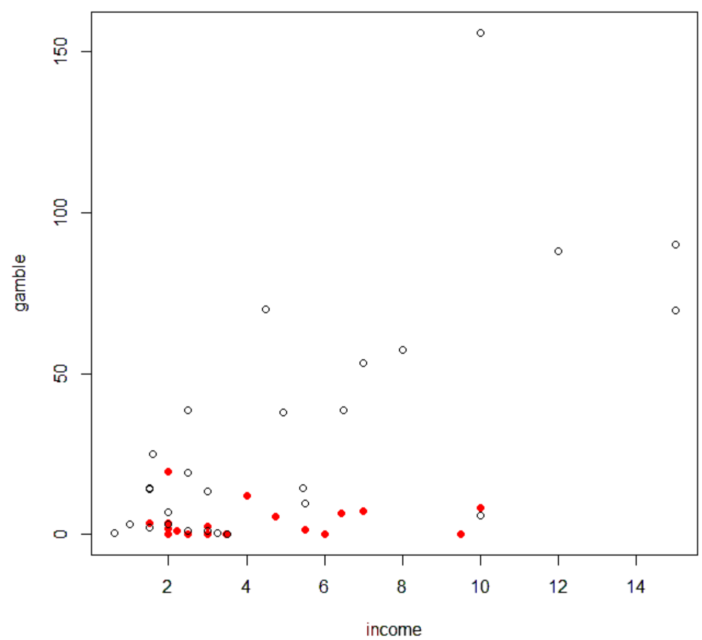
We will try to model `gamble` as the response and the other variables as predictors.

a) Plot `gamble` vs `status`, `gamble` vs `income`, and `gamble` vs `verbal`, using different symbols for males and females. Do these plots suggest the possible need for the interaction terms between `sex` and the other predictors?

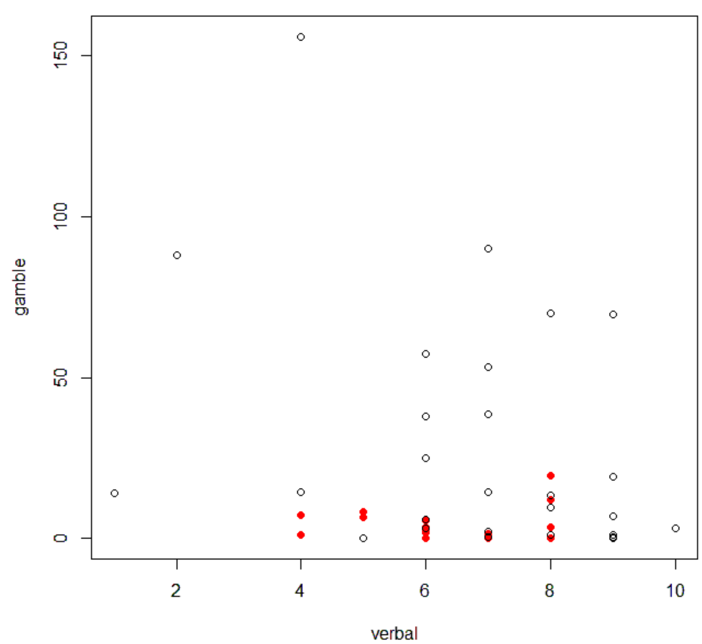
```
> library(faraway)
> data(teengamb)
> attach(teengamb)
> plot(status,gamble,pch=1+15*sex,col=sex+1)
```



```
> plot(income,gamble,pch=1+15*sex,col=sex+1)
```




```
> plot(verbal,gamble,pch=1+15*sex,col=sex+1)
```



All three plots suggest the need for the interaction term between `sex` and the other three predictors, since the rates of the relationships between `gamble` and `status`, `income`, and `verbal` are different for `sex` = 0 and `sex` = 1.

b) Fit a model with `gamble` as the response and the other variables as predictors that includes the interaction terms between `sex` and the other predictors. Determine whether this model may be reasonably simplified.

```
> fit = lm(gamble ~ sex + status + (sex*status) + income +
(sex*income) + verbal + (sex*verbal), data=teengamb)
> summary(fit)

Call:
lm(formula = gamble ~ sex + status + (sex * status) + income +
    (sex * income) + verbal + (sex * verbal), data = teengamb)

Residuals:
    Min      1Q  Median      3Q     Max
-56.654  -7.589  -1.016   3.323  83.903

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.6354    17.6218   1.568   0.1249
sex         -33.0132    35.0530  -0.942   0.3521
status       -0.1456     0.3316  -0.439   0.6631
income        6.0291     1.0538   5.721 1.26e-06 ***
verbal       -2.9748     2.4265  -1.226   0.2276
sex:status    0.3529     0.5492   0.643   0.5243
sex:income   -5.3478     2.4244  -2.206   0.0334 *
sex:verbal    2.8355     4.5973   0.617   0.5410
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.98 on 39 degrees of freedom
Multiple R-squared: 0.6243,     Adjusted R-squared: 0.5569
F-statistic:  9.26 on 7 and 39 DF,  p-value: 1.06e-06

> step(fit,direction="backward")
Start:  AIC=293.32
gamble ~ sex + status + (sex * status) + income + (sex * income) +
    verbal + (sex * verbal)

             Df Sum of Sq     RSS    AIC
- sex:verbal  1     167.4 17331.0  291.8
- sex:status  1     181.7 17345.2  291.8
<none>                    17163.5  293.3
- sex:income  1    2141.4 19304.9  296.8

Step:  AIC=291.77
gamble ~ sex + status + income + verbal + sex:status + sex:income
```

```
               Df Sum of Sq     RSS    AIC
- sex:status   1     393.9 17724.8  290.8
- verbal       1     494.6 17825.5  291.1
<none>                     17331.0  291.8
- sex:income   1    2189.5 19520.4  295.4

Step:  AIC=290.83
gamble ~ sex + status + income + verbal + sex:income

               Df Sum of Sq     RSS    AIC
- status       1      15.2 17740.1  288.9
- verbal       1     740.0 18464.8  290.8
<none>                     17724.8  290.8
- sex:income   1    3898.9 21623.8  298.2

Step:  AIC=288.87
gamble ~ sex + income + verbal + sex:income

               Df Sum of Sq     RSS    AIC
<none>                     17740.1  288.9
- verbal       1    1189.8 18929.9  289.9
- sex:income   1    3901.5 21641.5  296.2

Call:
lm(formula = gamble ~ sex + income + verbal + sex:income, data =
teengamb)

Coefficients:
(Intercept)          sex       income       verbal   sex:income
     17.833        4.625        6.247       -2.807       -6.385
```

"Best" model:

`gamble ~ sex + income + verbal + sex*income`

Note that `step` does not consider removing a predictor from the model if an interaction term involving that predictor is present in the model.

3. Suppose a complete second-order model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon$$

was fit to $n = 24$ data points.

```
> sum( lm( y ~ 1 )$residuals^2 )
[1]  360
> sum( lm( y ~ x1 + x2 )$residuals^2 )
[1]  126
> sum( lm( y ~ x1 + x2 + I(x1*x2) )$residuals^2 )
[1]  100
> sum( lm( y ~ x1 + x2 + I(x1*x2) + I(x1^2) +I(x2^2) )$residuals^2 )
[1]  72
```

a) Perform the "significance of the regression" test at a 5% level of significance.

$\mathbf{H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0}$ vs H_1 : at least one of $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ is not zero.

Full model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon$ dim = 6

Null model: $Y = \beta_0 + \varepsilon$ dim = 1

$\text{SSResid}_{\text{Null}} = 360$ $\text{SSResid}_{\text{Full}} = 72$ $360 - 72 = 288$

ANOVA table:

Source	SS	DF	MS	F
Regression (Diff.)	288	6 – 1 = 5	57.6	**14.4**
Residuals (Full)	72	24 – 6 = 18	4	
Total (Null)	360	24 – 1 = 23		

$F_{0.05}(5, 18) = \mathbf{2.77}$ **Reject $\mathbf{H_0}$** at $\alpha = 0.05$

b) Test whether the second-order terms are significant at a 5% level of significance. What is the p-value of the test? (You may give a range.)

$\mathbf{H_0 : \beta_3 = \beta_4 = \beta_5 = 0}$ vs H_1 : at least one of $\beta_3, \beta_4, \beta_5$ is not zero.

Full model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon$ dim = 6

Null model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ dim = 3

$\text{SSResid}_{\text{Null}} = 126$ $\text{SSResid}_{\text{Full}} = 72$ $126 - 72 = 54$

ANOVA table:

Source	SS	DF	MS	F
Regression (Diff.)	54	6 – 3 = 3	18	**4.5**
Residuals (Full)	72	24 – 6 = 18	4	
Total (Null)	126	24 – 3 = 21		

$F_{0.05}(3, 18) = \mathbf{3.16}$ **Reject $\mathbf{H_0}$** at $\alpha = 0.05$

$3.16 = F_{0.05}(3, 18) < F < F_{0.01}(3, 18) = 5.09.$

0.01 < p-value < 0.05. (p-value ≈ 0.0159)

c) Find the values of Adjusted R^2 for the null and the full models from part (b). Which model is preferred? ***Justify your answer***.

$R^2_{\text{Null}} = 1 - \frac{126}{360} = 0.65.$ $R^2_{\text{Full}} = 1 - \frac{72}{360} = 0.80.$

Adjusted $R^2_{\text{Null}} = 1 - \frac{23}{21} \cdot (1 - 0.65)$ Adjusted $R^2_{\text{Full}} = 1 - \frac{23}{18} \cdot (1 - 0.80)$

$\approx \mathbf{0.61667}.$ **<** $\approx \mathbf{0.74444}.$

Full model is preferred.

d) Find the values of AIC for the null and the full models from part (b). Which model is preferred? ***Justify your answer***.

$$\text{AIC}_{\text{Null}} = n + n \ln(2\pi) + n \ln\left(\text{SSResid}_{\text{Null}}/n\right) + 2(3)$$

$$= 24 + 24 \cdot \ln(2\pi) + 24 \cdot \ln\left(\frac{126}{24}\right) + 2 \cdot 3 \approx \mathbf{113.9065}.$$

OR

R: $$\text{AIC}_{\text{Null}} = n \ln\left(\text{SSResid}_{\text{Null}}/n\right) + 2(3) = 24 \cdot \ln\left(\frac{126}{24}\right) + 2 \cdot 3 = \mathbf{45.7975}.$$

$$\text{AIC}_{\text{Full}} = n + n \ln(2\pi) + n \ln\left(\text{SSResid}_{\text{Full}}/n\right) + 2(6)$$

$$= 24 + 24 \cdot \ln(2\pi) + 24 \cdot \ln\left(\frac{72}{24}\right) + 2 \cdot 6 \approx \mathbf{106.4757}.$$

OR

R: $$\text{AIC}_{\text{Full}} = n \ln\left(\text{SSResid}_{\text{Full}}/n\right) + 2(6) = 24 \cdot \ln\left(\frac{72}{24}\right) + 2 \cdot 6 = \mathbf{38.3667}.$$

$\text{AIC}_{\text{Full}} < \text{AIC}_{\text{Null}}$ **Full** model is preferred.

4. The grade point averages of students participating in college sports programs at Anytown State University are compared.*

Football	2.3	2.9	3.1	3.1	3.6	$\overline{y}_1 = 3.0$	$s_1^2 = 0.220$
Basketball	2.8	3.3	3.8	3.1	3.5	$\overline{y}_2 = 3.3$	$s_2^2 = 0.145$
Hockey	1.9	2.6	3.1	2.0	2.4	$\overline{y}_3 = 2.4$	$s_3^2 = 0.235$

Consider the model $\mathrm{Y}_{ij} = \mu_j + \varepsilon_{ij}$, where ε_{ij}'s are i.i.d. $\mathrm{N}(0, \sigma^2)$.

At $\alpha = 0.05$, can one conclude that there is a difference in the mean GPA of the three groups? State the null and alternative hypotheses, construct the ANOVA table and state your conclusion at $\alpha = 0.05$. Do NOT use a computer for this problem.

$\mathrm{H}_0: \mu_1 = \mu_2 = \mu_3$

$\mathrm{H}_1:$ not all μ_j's are the same

$\mathrm{H}_1:$ at least two of the μ_j's are different

$J = 3.$ $N = n_1 + n_2 + \ldots + n_J = 5 + 5 + 5 = 15.$

$$\overline{y} = \frac{n_1 \cdot \overline{y}_1 + n_2 \cdot \overline{y}_2 + \ldots + n_J \cdot \overline{y}_J}{N} = \frac{5 \cdot 3.0 + 5 \cdot 3.3 + 5 \cdot 2.4}{15} = 2.9.$$

$$\mathrm{SSB} = n_1 \cdot \left(\overline{y}_1 - \overline{y}\right)^2 + n_2 \cdot \left(\overline{y}_2 - \overline{y}\right)^2 + \ldots + n_J \cdot \left(\overline{y}_J - \overline{y}\right)^2$$

$$= 5 \cdot (3.0 - 2.9)^2 + 5 \cdot (3.3 - 2.9)^2 + 5 \cdot (2.4 - 2.9)^2 = 2.1.$$

$$\mathrm{MSB} = \frac{\mathrm{SSB}}{J - 1} = \frac{2.1}{2} = 1.05.$$

$$\mathrm{SSW} = \left(n_1 - 1\right) \cdot s_1^2 + \left(n_2 - 1\right) \cdot s_2^2 + \ldots + \left(n_J - 1\right) \cdot s_J^2$$

$$= 4 \cdot 0.220 + 4 \cdot 0.145 + 4 \cdot 0.235 = 2.4.$$

$$\mathrm{MSW} = \frac{\mathrm{SSW}}{N - J} = \frac{2.4}{12} = 0.2.$$

$$\mathrm{SSTot} = \mathrm{SSB} + \mathrm{SSW} = 2.1 + 2.4 = 4.5.$$

* The data does NOT represent the instructor's opinion of hockey and the brave men who participate in this sport.

Test Statistic: $F = \frac{MSB}{MSW} = \frac{1.05}{0.2} = \mathbf{5.25}$.

ANOVA table:

Source	SS	DF	MS	F
Between	2.1	2	1.05	5.25
Within	2.4	12	0.2	
Total	4.5	14		

Critical Value(s): $F_{0.05}(2, 12) = \mathbf{3.89}$.

Decision: **Reject H_0 at $\alpha = 0.05$**.