STAT 425 Assignment 5

Due Tuesday, April 6, 11:59 pm. Submit through Moodle.

Name: (insert your name here)

Netid: (insert)

Submit your computational work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown.

Most relevant class notes: 5.2.Spline, 6.Ancova, 7.VarSelect

Problem 1

Consider the `prostate` cancer surgery data from the `faraway` library in **R**. The variable `lpsa` is a measurement of prostate specific antigen on the log scale. Treat `lpsa` as the response and all the other variables in the data frame as potential predictors.

a) Using backward elimination testing and an alpha cutoff of 0.10, find the best model according to this procedure. Be sure to include the steps and the final model.

Answer:

```
library(faraway)

fit_full = lm(lpsa ~ ., data = prostate)
summary(fit_full)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## 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
```

```r
#The gleason coefficient has the highest p-value above alpha = 0.10.

fit2 = lm(lpsa ~ . -gleason, data = prostate)
summary(fit2)
```

```
##
## Call:
## lm(formula = lpsa ~ . - gleason, data = prostate)
##
## 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
```

```r
#The lcp coefficient has the highest p-value above alpha = 0.10.

fit3 = lm(lpsa ~ . - gleason - lcp, data = prostate)
summary(fit3)
```

```
##
## Call:
## lm(formula = lpsa ~ . - gleason - lcp, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77711 -0.41708  0.00002  0.40676  1.59681
##
## 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
```

```r
#The pgg45 coefficient has the highest p-value above alpha = 0.10.

fit4 = lm(lpsa ~ . - gleason - lcp - pgg45, data = prostate)
summary(fit4)
```

```
##
## Call:
## lm(formula = lpsa ~ . - gleason - lcp - pgg45, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83505 -0.39396  0.00414  0.46336  1.57888
##
## 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
```

```
#The age coefficient has the highest p-value above alpha = 0.10.

fit5 = lm(lpsa ~ . - gleason - lcp - pgg45 - age, data = prostate)
summary(fit5)
```

```
##
## Call:
## lm(formula = lpsa ~ . - gleason - lcp - pgg45 - age, data = prostate)
##
## 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
```

```
#The lbph coefficient has the highest p-value above alpha = 0.10.

fit6 = lm(lpsa ~ . - gleason - lcp - pgg45 - age - lbph, data = prostate)
summary(fit6)
```

```
##
## Call:
```

```
## lm(formula = lpsa ~ . - gleason - lcp - pgg45 - age - lbph, data = prostate)
##
## 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 of these predictors are significant at the alpha = .10 level, so this is the fina
```

b) Use backward selection in a stepwise algorithm to find the best model according to the AIC criterion. Be sure to include the steps and the final model.

Answer:

```
step(fit_full, direction = "backward")
```

```
## Start:  AIC=-58.32
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##
##           Df Sum of Sq    RSS     AIC
## - gleason  1    0.0412 44.204 -60.231
## - pgg45    1    0.5258 44.689 -59.174
## - lcp      1    0.6740 44.837 -58.853
## <none>                 44.163 -58.322
## - age      1    1.5503 45.713 -56.975
## - lbph     1    1.6835 45.847 -56.693
## - lweight  1    3.5861 47.749 -52.749
## - svi      1    4.9355 49.099 -50.046
## - lcavol   1   22.3721 66.535 -20.567
##
## Step:  AIC=-60.23
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45
##
##           Df Sum of Sq    RSS     AIC
```

```
## - lcp      1    0.6623 44.867 -60.789
## <none>                 44.204 -60.231
## - pgg45    1    1.1920 45.396 -59.650
## - age      1    1.5166 45.721 -58.959
## - lbph     1    1.7053 45.910 -58.560
## - lweight  1    3.5462 47.750 -54.746
## - svi      1    4.8984 49.103 -52.037
## - lcavol   1   23.5039 67.708 -20.872
##
## Step:  AIC=-60.79
## lpsa ~ lcavol + lweight + age + lbph + svi + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - pgg45    1    0.6590 45.526 -61.374
## <none>                 44.867 -60.789
## - age      1    1.2649 46.131 -60.092
## - lbph     1    1.6465 46.513 -59.293
## - lweight  1    3.5647 48.431 -55.373
## - svi      1    4.2503 49.117 -54.009
## - lcavol   1   25.4189 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    0.9592 46.485 -61.352
## - lbph     1    1.8568 47.382 -59.497
## - lweight  1    3.2251 48.751 -56.735
## - svi      1    5.9517 51.477 -51.456
## - lcavol   1   28.7665 74.292 -15.871
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)
##
## Coefficients:
## (Intercept)       lcavol      lweight          age         lbph          svi
##     0.95100      0.56561      0.42369     -0.01489      0.11184      0.72095
```

Step 1: gleason is removed. Step 2: lcp is removed. Step 3: pgg45 is removed. Full model has lcavol, lweight, age, lbph, and svi as predictors.

c) Use stepwise selection with the "both" option to find the best model according to the BIC criterion. Include the steps and the final model.

Answer:

```r
n = length(prostate[,1])
step(fit_full, direction = "both", k = log(n))
```

```
## Start:  AIC=-35.15
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##
##           Df Sum of Sq    RSS     AIC
## - gleason  1    0.0412 44.204 -39.634
## - pgg45    1    0.5258 44.689 -38.576
## - lcp      1    0.6740 44.837 -38.255
## - age      1    1.5503 45.713 -36.377
## - lbph     1    1.6835 45.847 -36.095
## <none>                 44.163 -35.149
## - lweight  1    3.5861 47.749 -32.151
## - svi      1    4.9355 49.099 -29.448
## - lcavol   1   22.3721 66.535   0.030
##
## Step:  AIC=-39.63
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - lcp      1    0.6623 44.867 -42.766
## - pgg45    1    1.1920 45.396 -41.627
## - age      1    1.5166 45.721 -40.936
## - lbph     1    1.7053 45.910 -40.537
## <none>                 44.204 -39.634
## - lweight  1    3.5462 47.750 -36.723
## + gleason  1    0.0412 44.163 -35.149
## - svi      1    4.8984 49.103 -34.014
## - lcavol   1   23.5039 67.708  -2.849
##
## Step:  AIC=-42.77
## lpsa ~ lcavol + lweight + age + lbph + svi + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - pgg45    1    0.6590 45.526 -45.926
## - age      1    1.2649 46.131 -44.644
## - lbph     1    1.6465 46.513 -43.844
## <none>                 44.867 -42.766
## - lweight  1    3.5647 48.431 -39.925
## + lcp      1    0.6623 44.204 -39.634
## - svi      1    4.2503 49.117 -38.561
## + gleason  1    0.0296 44.837 -38.255
```

```
## - lcavol   1   25.4189 70.285  -3.800
##
## Step:  AIC=-45.93
## lpsa ~ lcavol + lweight + age + lbph + svi
##
##           Df Sum of Sq    RSS     AIC
## - age      1    0.9592 46.485 -48.478
## - lbph     1    1.8568 47.382 -46.623
## <none>                  45.526 -45.926
## - lweight  1    3.2251 48.751 -43.862
## + pgg45    1    0.6590 44.867 -42.766
## + gleason  1    0.4560 45.070 -42.328
## + lcp      1    0.1293 45.396 -41.627
## - svi      1    5.9517 51.477 -38.583
## - lcavol   1   28.7665 74.292  -2.997
##
## Step:  AIC=-48.48
## lpsa ~ lcavol + lweight + lbph + svi
##
##           Df Sum of Sq    RSS     AIC
## - lbph     1    1.3001 47.785 -50.377
## <none>                  46.485 -48.478
## - lweight  1    2.8014 49.286 -47.377
## + age      1    0.9592 45.526 -45.926
## + pgg45    1    0.3533 46.131 -44.644
## + gleason  1    0.2126 46.272 -44.348
## + lcp      1    0.1023 46.383 -44.117
## - svi      1    5.8063 52.291 -41.636
## - lcavol   1   27.8298 74.315  -7.542
##
## Step:  AIC=-50.38
## lpsa ~ lcavol + lweight + svi
##
##           Df Sum of Sq    RSS     AIC
## <none>                  47.785 -50.377
## + lbph     1    1.3001 46.485 -48.478
## + pgg45    1    0.5735 47.211 -46.974
## + age      1    0.4025 47.382 -46.623
## + gleason  1    0.3890 47.396 -46.596
## + lcp      1    0.0641 47.721 -45.933
## - svi      1    5.1814 52.966 -44.966
## - lweight  1    5.8924 53.677 -43.673
## - lcavol   1   28.0445 75.829 -10.160
##
```

```
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)
##
## Coefficients:
## (Intercept)       lcavol      lweight          svi
##     -0.2681       0.5516       0.5085       0.6662
```

Step 1: gleason is removed. Step 2: lcp is removed. Step 3: pgg45 is removed. Step 4: age is removed. Step 5: lbph is removed. Full model includes lcavol, lweight, and svi as predictors.

d) Use the leaps and bounds algorithm to determine the model with smallest residual sums of squares for each model size from 2 to the maximum possible based on the number of columns in the data frame. Display the results by showing "which" variables were selected for model sizes 2, 3, etc.

Answer:

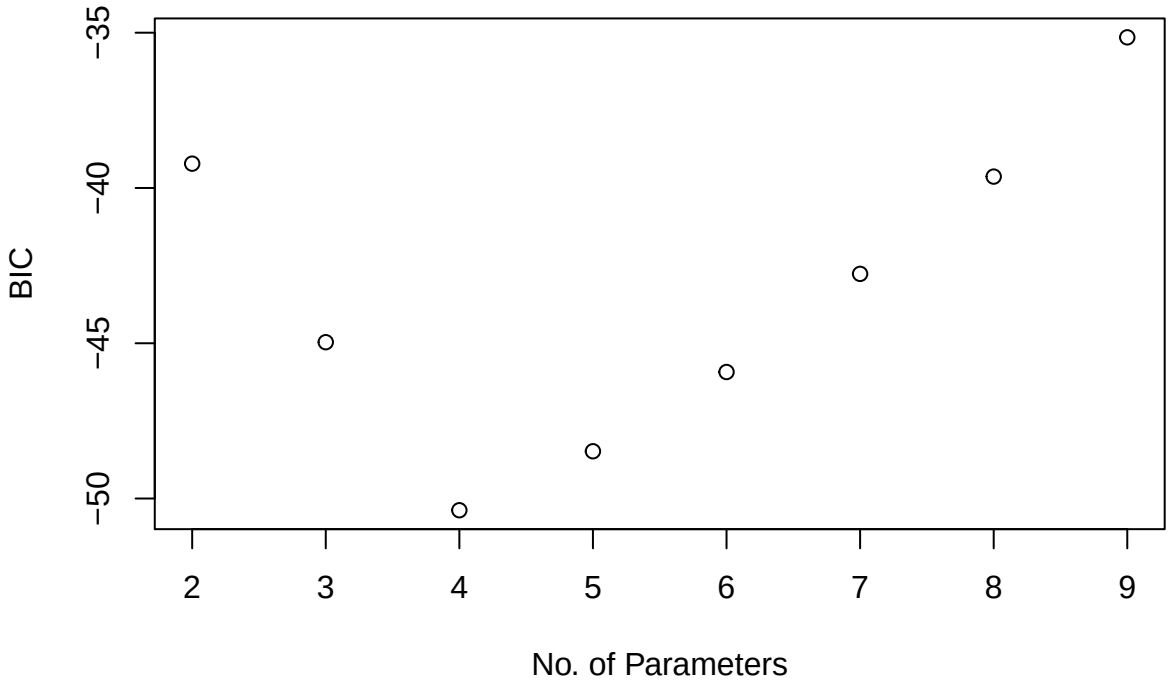
```
library(leaps)
b=regsubsets(lpsa ~ ., data = prostate)
rs = summary(b)
rs$which
```

```
##   (Intercept) lcavol lweight   age  lbph   svi   lcp gleason pgg45
## 1        TRUE   TRUE   FALSE FALSE FALSE FALSE FALSE   FALSE FALSE
## 2        TRUE   TRUE    TRUE FALSE FALSE FALSE FALSE   FALSE FALSE
## 3        TRUE   TRUE    TRUE FALSE FALSE  TRUE FALSE   FALSE FALSE
## 4        TRUE   TRUE    TRUE FALSE  TRUE  TRUE FALSE   FALSE FALSE
## 5        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE FALSE   FALSE FALSE
## 6        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE FALSE   FALSE  TRUE
## 7        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE  TRUE   FALSE  TRUE
## 8        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE  TRUE    TRUE  TRUE
```

e) Using the results from d) and further calculations, graph BIC versus model size for the models selected in part d). Which model is the overall best model according to BIC?

Answer:

```
msize = 2:9
Bic = n*log(rs$rss/n) + msize*log(n)
plot(msize, Bic, xlab="No. of Parameters", ylab = "BIC")
```



As shown by the graph, model size 4, with an intercept and three predictors has the lowest BIC. In the table shown in part d, a model with lcavol, lweight, and svi as predictors is the best overall model according to BIC.

Answer:

Problem 2

The `aatemp` data in the `faraway` library comes from the U.S. Historical Climatological Network. The data report annual mean temperatures in Ann Arbor Michigan for roughly 150 years.

a) With `temp` as the response, fit a regression spline with intercept using B-spline basis functions of `year` and 8 degrees of freedom. Show the fitted curve on the scatter plot of `temp` versus `year`.

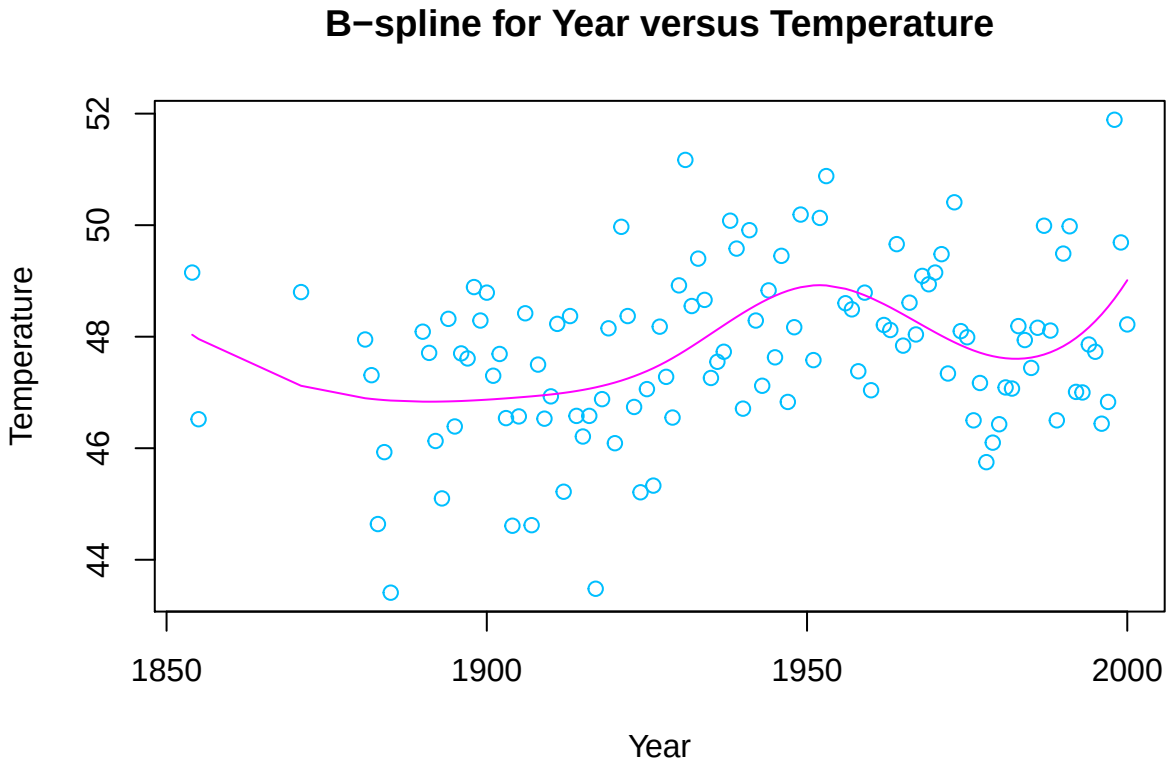
Answer: We first create a basis function for `year` using `bs` and plot the fitted curve as below.

```
# load package and data
require('faraway')
require('splines')
```

```
## Loading required package: splines
```

```
data('aatemp')
# B-spline
bs.basis <- bs(aatemp$year, df = 8, intercept = TRUE)
# regression
```

```r
lmod <- lm(aatemp$temp ~ bs.basis - 1)
# plot
plot(temp ~ year, data = aatemp, col = 'deepskyblue',
     xlab = 'Year', ylab = 'Temperature',
     main = 'B-spline for Year versus Temperature')
lines(x = aatemp$year, y = fitted.values(lmod), col = 'magenta')
```



b) How many knots does the model in a) have?

Answer: Based on the formula df $= m + 4$ for B-spline with an intercept, where m is the number of knots. We know that we have $m = 4$ knots for model in a).

c) Compute AIC, BIC and adjusted R-square for the model in a).

Answer: We can compute AIC, BIC, and adjusted R-square as below.

```r
# AIC
AIC(lmod)
```

```
## [1] 414.0364
```

```r
# BIC
BIC(lmod)
```

```
## [1] 438.7408
```

```r
# Adjusted R square
n <- nrow(aatemp)
```

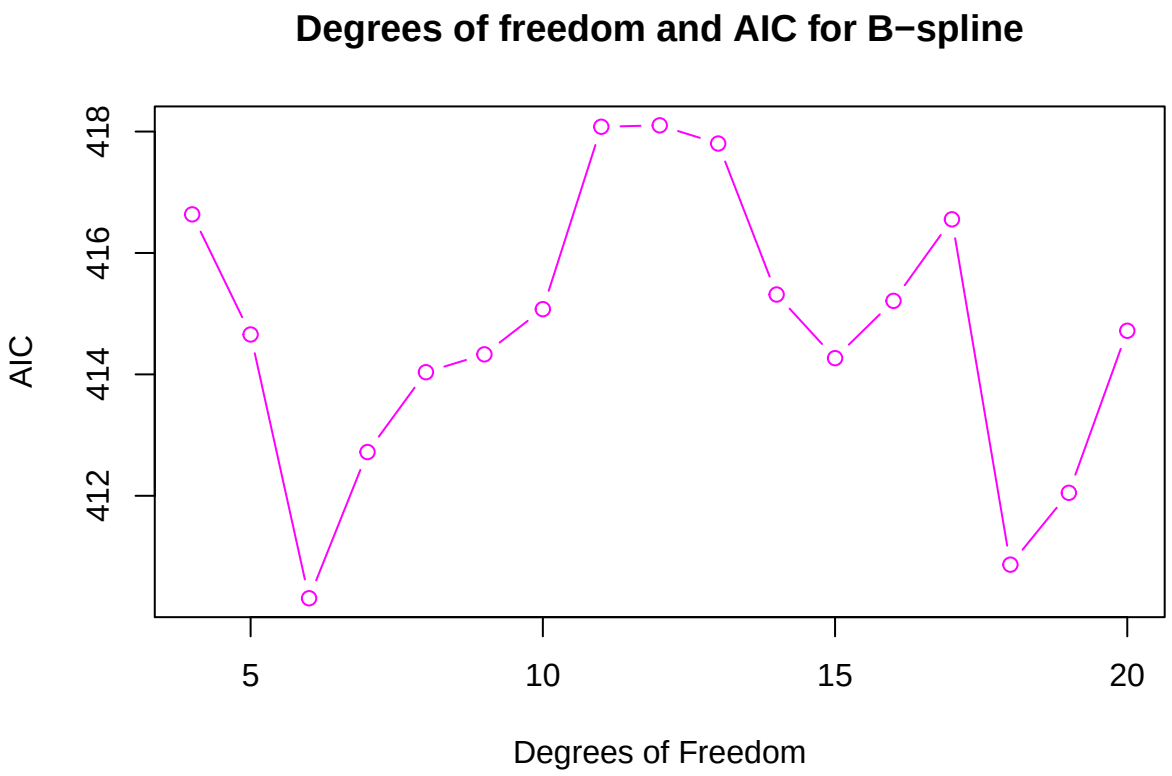
```r
sse <- as.numeric(crossprod(residuals(lmod)))
sst <- var(aatemp$temp) * (nrow(aatemp) - 1)
r2 <- 1 - sse / sst
adj.r2 <- 1 - (1 - r2) * (n - 1) / (n - 8)
```

Note that if you calculate the AIC, BIC, and adjusted R^2 by yourself, your answer should be correct if the formula are applied correctly. Also, if you use p or $p+1$ in the computation of AIC, where p is the dimension of the design matrix, you should also be fine.

d) Compute AIC for the b-spline models with degrees of freedom 4, 5, 6, ... 20. Plot AIC versus degrees of freedom. Which of these models is the best, according to AIC?

Answer: We can write a loop as below to compute the AIC and find which model is best according to AIC.

```r
# loop
DF <- 4:20
AICs <- c()
for (df in DF) {
  # B-spline
  bs.basis <- bs(aatemp$year, df = df, intercept = TRUE)
  # regression
  lmod <- lm(aatemp$temp ~ bs.basis)
  # AIC
  AICs <- c(AICs, AIC(lmod))
}
# plot
plot(AICs ~ DF, xlab = 'Degrees of Freedom',
     ylab = 'AIC',
     main = 'Degrees of freedom and AIC for B-spline', col = 'magenta',
     type = 'b')
```




```
# which minimizes
DF[which.min(AICs)]
```

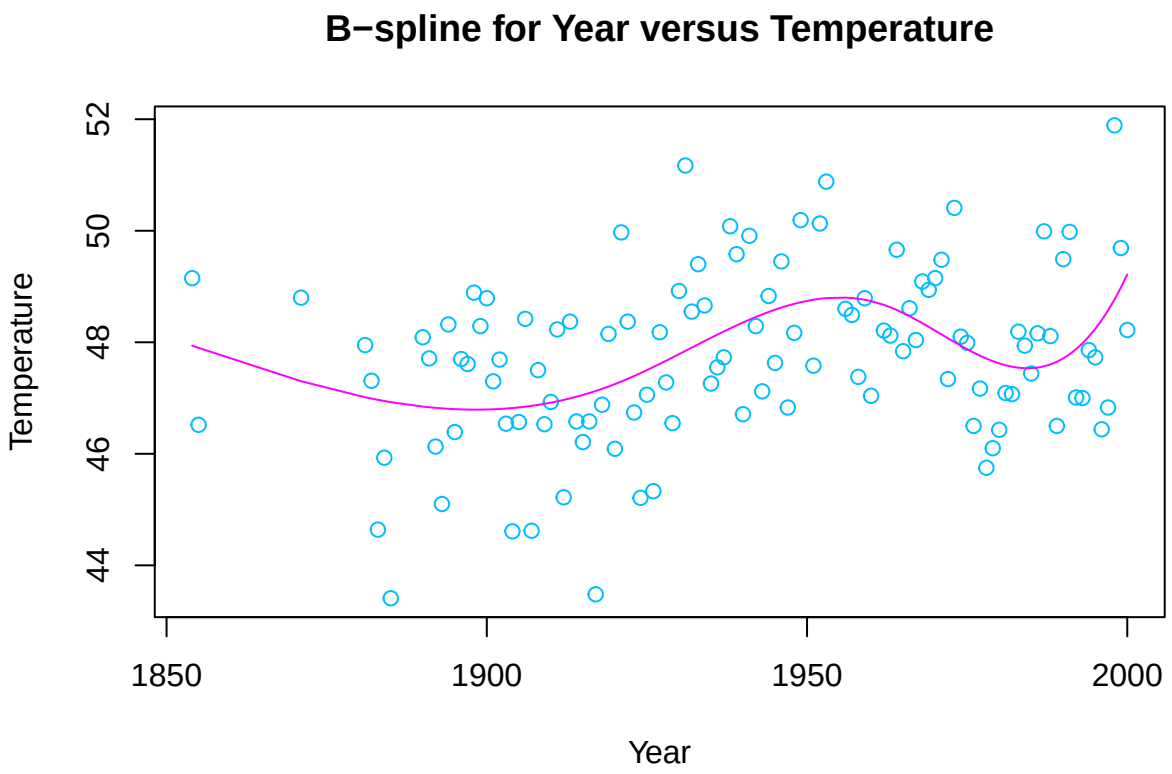
```
## [1] 6
```

As we can see, the B-spline model is degrees of freedom of 6 is considered the best according to AIC.

e) Show the fitted curve from the best model selected in d) on the scatter plot of `temp` versus `year`. How does it compare with the curve in a)?

Answer: We can plot the fitted curve of the best model in d) as below.

```
# B-spline
bs.basis <- bs(aatemp$year, df = 6, intercept = TRUE)
# regression
lmod <- lm.fit(x = bs.basis, y = aatemp$temp)
# plot
plot(temp ~ year, data = aatemp, col = 'deepskyblue',
     xlab = 'Year', ylab = 'Temperature',
     main = 'B-spline for Year versus Temperature')
lines(x = aatemp$year, y = fitted.values(lmod), col = 'magenta')
```



As we can see, compared to a), the plot is almost identical.

Problem 3:

In this problem we model data from a study of the prevalence of obesity, diabetes and cardiovascular disease among 403 African Americans in central Virginia. The data are in the `diabetes` dataset in the `faraway` library. One of the blood measurements is glycosolated hemoglobin (`glyhb`). A value higher than 7 is often considered to be a positive diagnosis for diabetes. Here we treat the numerical value og `glyhb` as the response, and consider possible predictor variables `gender` (a factor variable), `waist`, `age`, and stabilized glucose (`stab.glu`).

a) Fit the ancova model that includes the main effects for `gender`, `waist`, `age`, `stab.glu` as well as the two-way interactions between gender and each of the other three predictor variables. Show the model summary and indicate which coefficients are significant at level 0.05.

Answer:

```
library("faraway")
model1 = lm(glyhb ~ gender + waist + age + stab.glu + gender:waist + gender:age + gender
summary(model1)
```

```
##
## Call:
## lm(formula = glyhb ~ gender + waist + age + stab.glu + gender:waist +
##     gender:age + gender:stab.glu, data = diabetes)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2562 -0.7162 -0.1203  0.4528  9.6957
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.393920   0.838982   2.853  0.00456 **
## genderfemale           -2.492367   1.058500  -2.355  0.01905 *
## waist                  -0.011856   0.022109  -0.536  0.59210
## age                     0.013983   0.007507   1.863  0.06328 .
## stab.glu                0.027386   0.001891  14.484  < 2e-16 ***
## genderfemale:waist      0.048418   0.027563   1.757  0.07979 .
## genderfemale:age        0.003624   0.009664   0.375  0.70789
## genderfemale:stab.glu   0.005616   0.003001   1.872  0.06203 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.45 on 380 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.5898, Adjusted R-squared:  0.5822
## F-statistic: 78.06 on 7 and 380 DF,  p-value: < 2.2e-16
```

Based on the p values of summary result, we can see coefficients of "gender" and "stab.glu" are significant.

b) By comparing with a simplified model, test the null hypothesis that the three interaction coefficients all equal zero versus the alternative that at least one of them is nonzero. Use overall significance level of 0.05 and state your conclusion.

Answer:

```
model2 = lm(glyhb ~ gender + waist + age + stab.glu, data = diabetes)
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: glyhb ~ gender + waist + age + stab.glu + gender:waist + gender:age +
##     gender:stab.glu
## Model 2: glyhb ~ gender + waist + age + stab.glu
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1    380 798.52
## 2    383 818.95 -3   -20.432 3.241 0.02216 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the p value we should reject the null hypothesis, which means at least one of interaction

coefficients is nonzero.

c) The additive model has only the main effects; no interactions are included. (Main effects are terms involving only the original variables, not their products). Fit this model and determine which of the main effects appear to have statistically significant coefficients at the 0.05 level.

Answer:

```
modeladd = lm(glyhb ~ gender + waist + age + stab.glu, data = diabetes)
summary(modeladd)
```

```
##
## Call:
## lm(formula = glyhb ~ gender + waist + age + stab.glu, data = diabetes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7806 -0.7293 -0.1601  0.4072  9.6192
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.835159   0.520004   1.606 0.109084
## genderfemale 0.109011   0.151993   0.717 0.473683
## waist        0.020016   0.013315   1.503 0.133584
## age          0.017042   0.004755   3.584 0.000382 ***
## stab.glu     0.029261   0.001474  19.851  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.462 on 383 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.5793, Adjusted R-squared:  0.5749
## F-statistic: 131.9 on 4 and 383 DF,  p-value: < 2.2e-16
```

From the result we can see the variables "age" and "stab.glu" are two significant coefficients.

d) Compute AIC for the model in a) and the model in c). Which model is preferred according to this criterion?

Answer:

```
AIC(model1)
```

```
## [1] 1399.137
```

```
AIC(modeladd)
```

```
## [1] 1402.94
```

As we can see, the model in a) is prefered here since the AIC value is smaller.

e) Use the sequential analysis of variance table for the full model in a) to determine which main effects and interactions are significant. What do you conclude?

Answer:

```
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: glyhb
##                   Df Sum Sq Mean Sq  F value    Pr(>F)
## gender             1   4.81    4.81   2.2875   0.13125
## waist              1 101.82  101.82  48.4538 1.492e-11 ***
## age                1 178.50  178.50  84.9424 < 2.2e-16 ***
## stab.glu           1 842.61  842.61 400.9834 < 2.2e-16 ***
## gender:waist       1  11.37   11.37   5.4109   0.02054 *
## gender:age         1   1.70    1.70   0.8094   0.36887
## gender:stab.glu    1   7.36    7.36   3.5028   0.06203 .
## Residuals        380 798.52    2.10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the sequential anova test, we can see that all of three main effects and interaction between "gender" and "waist" are significant.