Homework 4

Charles Hwang

11/13/2022

Charles Hwang

Dr. Xi

STAT 408-001

2022 November 13

Problem 1

```
rm(list=ls())
p<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 408 - Applied Regression Analysis/prostate.csv")
s<-lm(lpsa~.,data=p)
```

Problem 1a

```
summary(s) # Lecture 10, Slides 5-10
```

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

```
summary(update(s,.~.-gleason)) # p = 0.7750328
```

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

```
summary(update(update(s,.~.-gleason),.~.-lcp)) # p = 0.2512688
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + pgg45,
##     data = p)
##
## 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
```

```
summary(update(update(update(s,.~.-gleason),.~.-lcp),.~.-pgg45)) # p = 0.2533092
```

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

```
summary(update(update(update(update(s,.~.-gleason),.~.-lcp),.~.-pgg45),.~.-age)) # p = 0.1695282
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + lbph + svi, data = p)
##
## 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
```

```
summary(update(update(update(update(update(s,.~.-gleason),.~.-lcp),.~.-pgg45),.~.-age),.~.-lbph)) #p=.1
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi, data = p)
##
```

```
## 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
```

Problem 1b

```
step(s) # Lecture 10, Slides 15-16
```

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

```
summary(lm(lpsa~lcavol+lweight+age+lbph+svi,data=p)) # Summary of "best" model (AIC = -61.37)
```

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

Problem 1c

We can see the two model selection methods do not give us the same result. The model from the backward elimination method at the $\alpha = 0.05$ level includes the `lcavol`, `lweight`, and `svi` variables, while the model from the Akaike information criterion (AIC) method includes these and the `age` and `lbph` variables. We can also see the adjusted r^2 values for the two models are different (0.6143899 vs. 0.6245476).

I do not believe it is an issue that the two models are different because it simply illustrates the difference between the two methods. Changing the α level (to $\alpha \geq 0.1121295$) for the backward elimination method would also yield a different model.

Problem 2

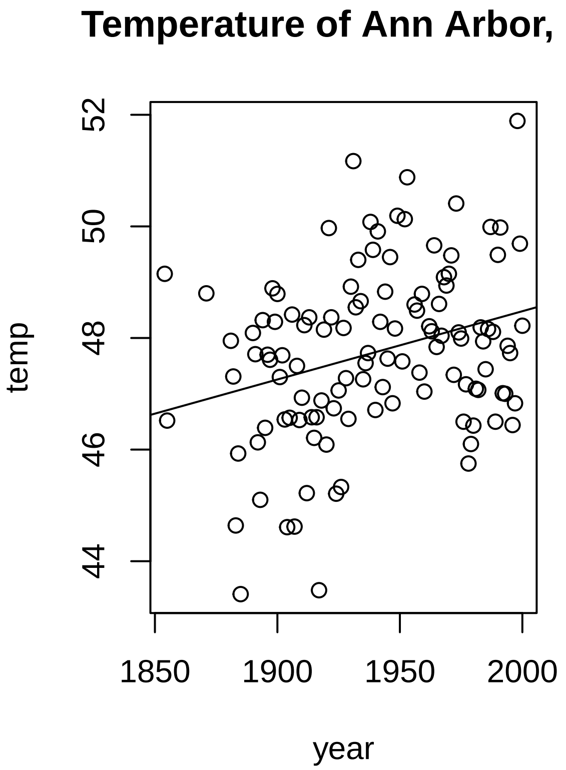
```
a<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 408 - Applied Regression Analysis/aatemp.csv")
```

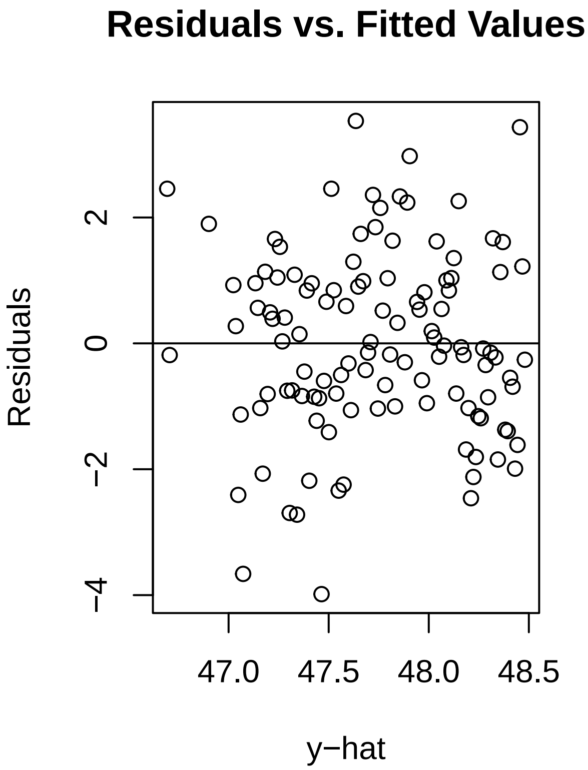
Problem 2a

```
ty<-lm(temp~year,data=a)
summary(ty)
```

```
##
## Call:
## lm(formula = temp ~ year, data = a)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9843 -0.9113 -0.0820  0.9946  3.5343
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.005510   7.310781   3.284  0.00136 **
## year         0.012237   0.003768   3.247  0.00153 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.466 on 113 degrees of freedom
## Multiple R-squared:  0.08536,    Adjusted R-squared:  0.07727
## F-statistic: 10.55 on 1 and 113 DF,  p-value: 0.001533
```

```
par(mfrow=c(1,2))
plot(temp~year,main="Temperature of Ann Arbor, MI",data=a)
abline(ty$coefficients["(Intercept)"],ty$coefficients["year"])
plot(ty$residuals~ty$fitted.values,xlab="y-hat",ylab="Residuals",main="Residuals vs. Fitted Values")
abline(h=0)
```





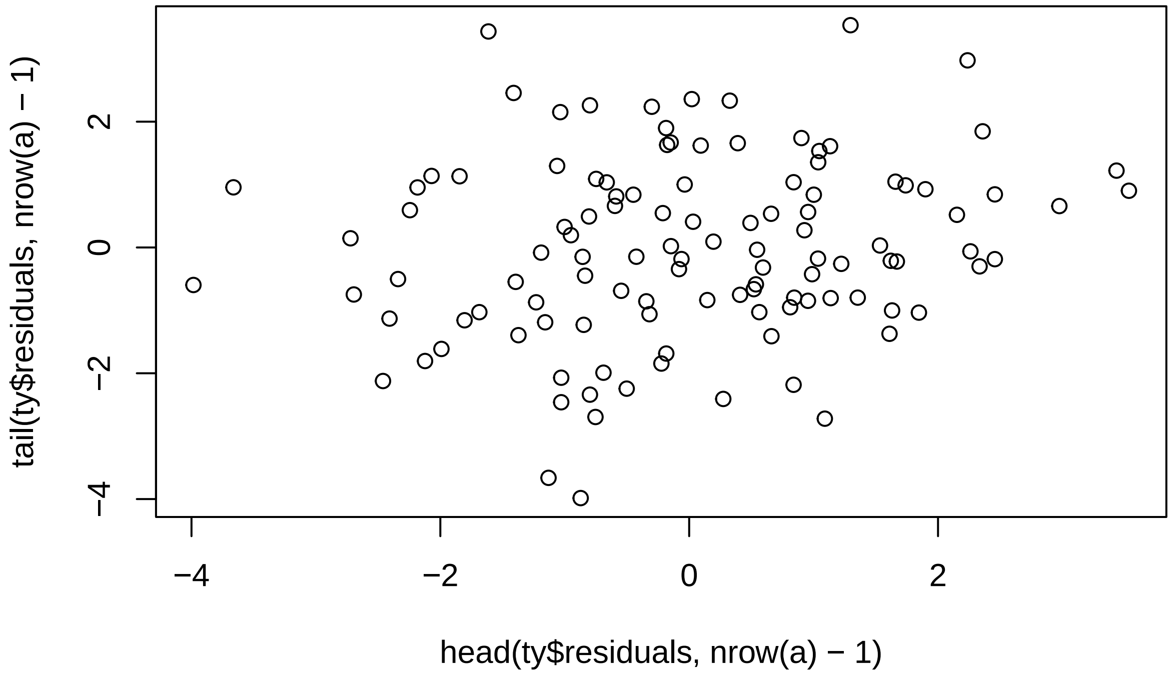
From the output produced, it is possible there is a linear trend in the data, but it would be rather weak. For example, we can see the adjusted r^2 for the model is only 0.0772653.

Problem 2b

```
cor(tail(ty$residuals,nrow(a)-1),head(ty$residuals,nrow(a)-1)) # Lecture 9, Slide 14
```

```
## [1] 0.1809103
```

```
plot(tail(ty$residuals,nrow(a)-1)~head(ty$residuals,nrow(a)-1))
```



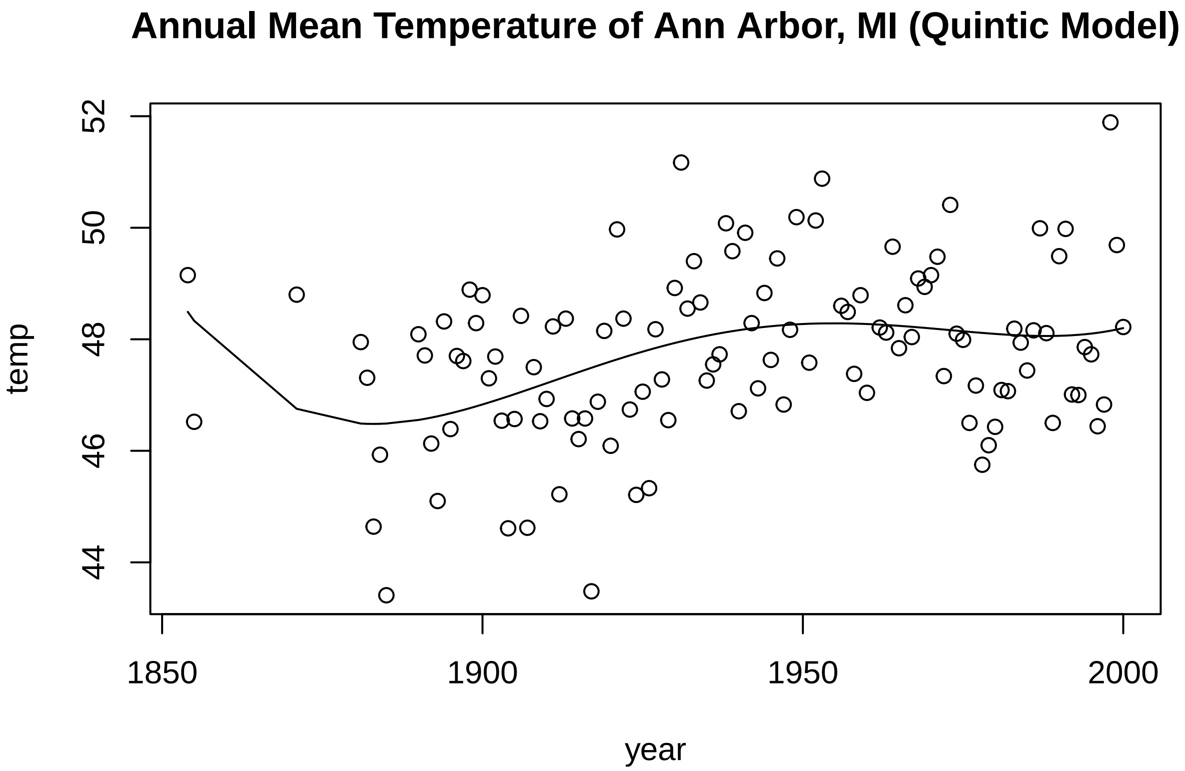
This output does not change my opinion about the trend. There appears to be weak correlation ($r^2 = 0.1809103$) between years.

Problem 2c

```
summary(lm(temp~year+I(year^2)+I(year^3)+I(year^4)+I(year^5),data=a)) # year^5 = NA
```

```
##
## Call:
## lm(formula = temp ~ year + I(year^2) + I(year^3) + I(year^4) +
##     I(year^5), data = a)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0085 -0.9618 -0.0913  0.9926  3.7370
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.497e+06  8.553e+05   1.750   0.0829 .
## year        -3.086e+03  1.775e+03  -1.739   0.0849 .
## I(year^2)    2.385e+00  1.381e+00   1.727   0.0869 .
## I(year^3)   -8.189e-04  4.773e-04  -1.716   0.0890 .
## I(year^4)    1.054e-07  6.186e-08   1.704   0.0912 .
## I(year^5)           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.431 on 110 degrees of freedom
## Multiple R-squared:  0.1522, Adjusted R-squared:  0.1213
## F-statistic: 4.936 on 4 and 110 DF,  p-value: 0.001068
```

```
plot(temp~year,main="Annual Mean Temperature of Ann Arbor, MI (Quintic Model)",data=a)
lines(a$year,predict(lm(temp~year+I(year^2)+I(year^3)+I(year^4)+I(year^5),data=a)))
```

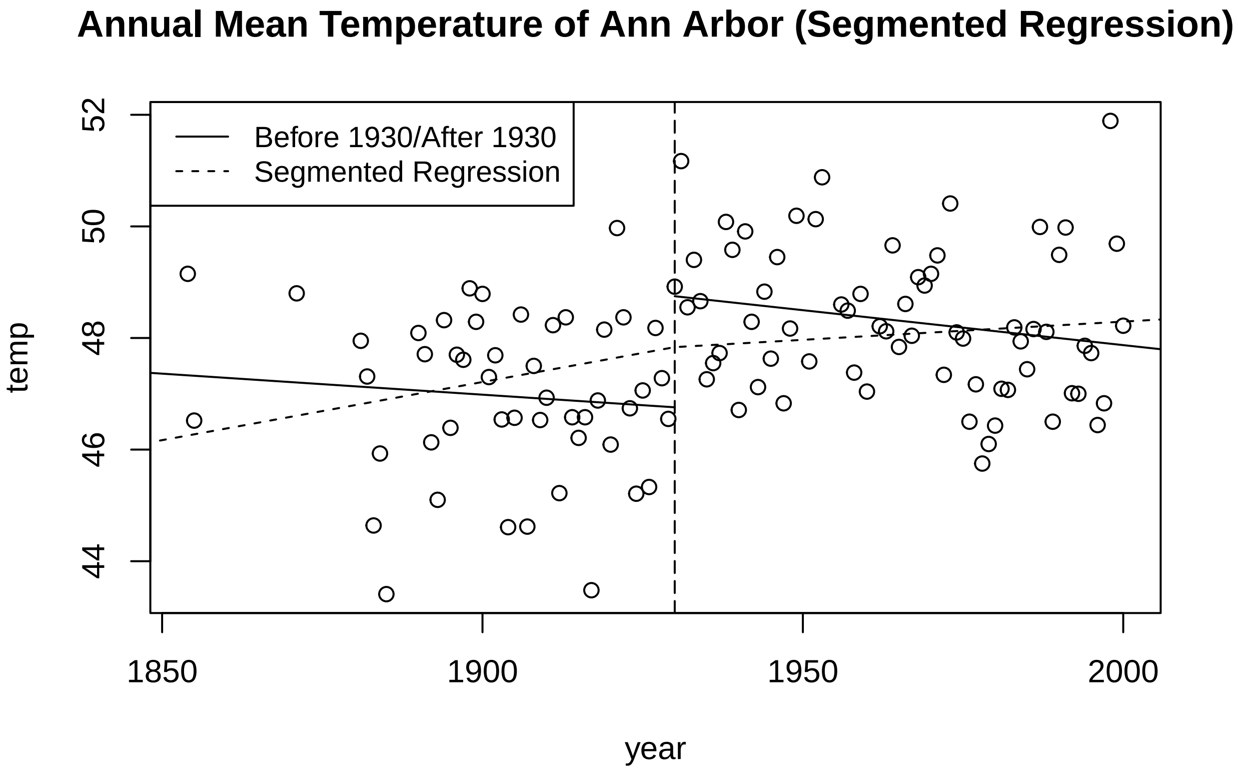


```
#lines(a$year,predict(loess(temp~year,data=a)),col="blue") # Local regression (LOESS)
#legend("topleft",c("Quintic","LOESS"),col=c("black","blue"),lty=1)
```

Observation: The quintic (year)^5 term appears as `NA` in the summary output, functionally making this the same as a quartic model. This may be because the values of the year numbers raised to the fifth power are too big. Additionally, the fitted curve looks similar to one from a local regression (LOESS) model.

Problem 2d

```
plot(temp~year,main="Annual Mean Temperature of Ann Arbor (Segmented Regression)",data=a)
abline(v=1930,lty=5) # Lecture 11, Slides 7-13
ty29<-lm(temp~year,data=a,subset=(year<1930))$coefficients
ty31<-lm(temp~year,data=a,subset=(year>1930))$coefficients
segments(1830,ty29["(Intercept)"]+ty29["year"]*1830,1930,ty29["(Intercept)"]+ty29["year"]*1930)
segments(1930,ty31["(Intercept)"]+ty31["year"]*1930,2010,ty31["(Intercept)"]+ty31["year"]*2010)
bl<-function(x){ifelse(x<1930,1930-x,0)}
br<-function(x){ifelse(x>1930,x-1930,0)}
sm<-lm(temp~bl(year)+br(year),data=a)
x<-seq(1830,2010)
lines(x,sm$coefficients[1]+sm$coefficients[2]*bl(x)+sm$coefficients[3]*br(x),lty=2)
legend("topleft",c("Before 1930/After 1930","Segmented Regression"),lty=1:2,cex=0.9)
```



It does appear that the temperature trend is different before and after 1930. We can see the model for years before 1930 predicts a mean temperature of approximately 46.7580908 degrees Fahrenheit for Ann Arbor, Michigan in 1930, while the model for years after 1930 predicts a mean temperature of approximately 48.7487383 degrees Fahrenheit. The segmented regression model is between the two, predicting a mean temperature of approximately 47.8364118 degrees Fahrenheit.[1] We can see the actual mean temperature in Ann Arbor, Michigan in 1930 was approximately 48.92 degrees Fahrenheit.

Problem 3

```
l<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 408 - Applied Regression Analysis/longley.csv")
```

Problem 3a

```
round(cor(l[,-7]),6) # Lecture 11, Slides 21 and 24-25
```

```
##              GNP.deflator      GNP Unemployed Armed.Forces Population     Year
## GNP.deflator     1.000000 0.991589   0.620633     0.464744   0.979163 0.991149
## GNP              0.991589 1.000000   0.604261     0.446437   0.991090 0.995273
## Unemployed       0.620633 0.604261   1.000000    -0.177421   0.686552 0.668257
## Armed.Forces     0.464744 0.446437  -0.177421     1.000000   0.364416 0.417245
## Population       0.979163 0.991090   0.686552     0.364416   1.000000 0.993953
## Year             0.991149 0.995273   0.668257     0.417245   0.993953 1.000000
```

Looking at the correlation matrix, we can see the `GNP.deflator`, `GNP`, `Year`, and `Population` variables are all extremely highly correlated with one another ($|r| > 0.97$). No other pairs of variables are highly correlated with one another to this extent, with `Unemployed` and `Population` having the next highest correlation ($r = 0.6865515$).

[1] Observation: This appears to be an example of Simpson's paradox (https://en.wikipedia.org/wiki/Simpson's_paradox). We can see that the data before 1930 and the data after 1930 both individually have negative trends, but when combined, the data have a positive trend overall as indicated by the segmented regression line.

The `GNP.deflator` and `GNP` variables are extremely highly correlated with each other because the GNP implicit price deflator is calculated using gross national product (GNP). A potential reason that these and the `Population` and `Year` variables are extremely highly correlated with one another is because both GNP and population in the United States increased at a consistent rate from 1947 to 1962 as a result of post-World War II economic production.

Problem 3b

```
c<-model.matrix(lm(Employed~.,data=l))[,-1]
r<-data.frame(rep(NA,6))
for(i in 1:(length(l)-1)){r[i,1]<-summary(lm(c[,i]~c[,-i]))$r.squared}
names(r)<-"Adjusted r^2"
rownames(r)<-colnames(c)
r
```

```
##              Adjusted r^2
## GNP.deflator    0.9926217
## GNP             0.9994409
## Unemployed      0.9702548
## Armed.Forces    0.7213654
## Population      0.9974947
## Year            0.9986824
```

Yes, regressing each predictor on the others results in the same conclusion. We can see there is clearly a lot of collinearity on the variables in the model.

Problem 3c

```
cor(l[,c(3:4,6)])
```

```
##              Unemployed Armed.Forces      Year
## Unemployed    1.0000000   -0.1774206 0.6682566
## Armed.Forces -0.1774206    1.0000000 0.4172451
## Year          0.6682566    0.4172451 1.0000000
```

```
summary(lm(Employed~Unemployed+Armed.Forces+Year,data=l))
```

```
##
## Call:
## lm(formula = Employed ~ Unemployed + Armed.Forces + Year, data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57285 -0.11989  0.04087  0.13979  0.75303
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.797e+03  6.864e+01 -26.183 5.89e-12 ***
## Unemployed   -1.470e-02  1.671e-03  -8.793 1.41e-06 ***
## Armed.Forces -7.723e-03  1.837e-03  -4.204  0.00122 **
## Year          9.564e-01  3.553e-02  26.921 4.24e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3321 on 12 degrees of freedom
## Multiple R-squared:  0.9928, Adjusted R-squared:  0.9911
## F-statistic: 555.2 on 3 and 12 DF,  p-value: 3.916e-13
```

```
anova(lm(Employed~Unemployed+Armed.Forces+Year,data=l),lm(Employed~.,data=l))
```

```
## Analysis of Variance Table
##
## Model 1: Employed ~ Unemployed + Armed.Forces + Year
## Model 2: Employed ~ GNP.deflator + GNP + Unemployed + Armed.Forces + Population +
##     Year
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     12 1.32336
## 2      9 0.83642  3   0.48694 1.7465  0.227
```

I removed the `GNP.deflator`, `GNP`, and `Population` variables from the full model Ω to create a smaller model ω with only `Unemployed`, `Armed.Forces`, and `Year` as predictors. We can see the strongest correlation in this smaller model is between `Unemployed` and `Year` ($r = 0.6682566$), and only one of the four variables highly correlated with one another as seen in Problem 3a is included. I believe this smaller model is better because we can see the comparison-of-models hypothesis test (Lecture 8, Slides 12-14) shows there is insufficient evidence at the $\alpha = 0.05$ level ($F = 1.7464946$, $p = 0.2270322$) that the full model Ω is better than the reduced model ω.

Problem 4

```
g<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 408 - Applied Regression Analysis/gala.csv")
gm<-read.csv("/Users/newuser/Desktop/Notes/Graduate/STAT 408 - Applied Regression Analysis/galamiss.csv
```

Problem 4a

```
summary(lm(Species~Area+Elevation+Nearest+Scruz+Adjacent,data=g))
```

```
##
## Call:
## lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
##     data = g)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -111.679  -34.898   -7.862   33.460  182.584
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.068221  19.154198   0.369 0.715351
## Area         -0.023938   0.022422  -1.068 0.296318
## Elevation     0.319465   0.053663   5.953 3.82e-06 ***
## Nearest       0.009144   1.054136   0.009 0.993151
## Scruz        -0.240524   0.215402  -1.117 0.275208
## Adjacent     -0.074805   0.017700  -4.226 0.000297 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.98 on 24 degrees of freedom
## Multiple R-squared:  0.7658, Adjusted R-squared:  0.7171
## F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

Problem 4b

```
apply(apply(gm,1,is.na),1,sum)
```

```
##   Species  Endemics      Area Elevation   Nearest     Scruz  Adjacent
##         0         0         0         6         0         0         0
```

We can see the `Elevation` variable contains 6 missing values. This is the only variable with any missing values.

Problem 4c

```
summary(lm(Species~Area+Elevation+Nearest+Scruz+Adjacent,data=gm))
```

```
##
## Call:
## lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
##     data = gm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -115.17  -37.60  -10.08   35.17  172.54
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.32286   27.47417   0.558  0.58391
## Area        -0.02765    0.02557  -1.081  0.29388
## Elevation    0.32550    0.06476   5.026 8.78e-05 ***
## Nearest     -0.11042    1.17784  -0.094  0.92635
## Scruz       -0.28427    0.25422  -1.118  0.27818
## Adjacent    -0.07880    0.02092  -3.766  0.00141 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67.36 on 18 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.7668, Adjusted R-squared:  0.702
## F-statistic: 11.83 on 5 and 18 DF,  p-value: 3.54e-05
```

There are slight differences between the original model in Problem 4a and the model where rows with missing values are deleted. We can see the adjusted r^2 values are very similar (0.7170651 vs. 0.7019736) and notably, the *sign* of the coefficient for the `Nearest` variable changed (0.009144 vs. -0.1104193).

Problem 4d

```
gmm<-gm
for(j in 1:length(gm)){gmm[is.na(gmm[,j]),j]<-colMeans(gm,na.rm=TRUE)[j]}  # Lecture 12, Slide 8
summary(lm(Species~Area+Elevation+Nearest+Scruz+Adjacent,data=gmm))        # Lecture 12, Slide 9
```

```
##
## Call:
## lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
##     data = gmm)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
## -94.710 -42.598  -9.742  26.146 220.893 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -12.48266   28.62644  -0.436 0.666695    
## Area         -0.00137    0.02683  -0.051 0.959697    
## Elevation     0.27388    0.06891   3.975 0.000562 ***
## Nearest       0.37776    1.28270   0.295 0.770905    
## Scruz        -0.08544    0.27140  -0.315 0.755629    
## Adjacent     -0.06553    0.02215  -2.958 0.006856 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 74.52 on 24 degrees of freedom
## Multiple R-squared:  0.6503, Adjusted R-squared:  0.5774 
## F-statistic: 8.925 on 5 and 24 DF,  p-value: 6.77e-05
```

This model with mean value imputation may be weaker than the previous two models. We can see the adjusted r^2 values is lower (0.5774087 vs. 0.7019736 and 0.7170651). This may be due to some of the disadvantages of mean value imputation which are well-known through previous research (Lecture 12, Slide 10).

Problem 4e

```
gmr<-gm
ri<-lm(Elevation~Area+Nearest+Scruz+Adjacent,data=gm)
gmr[is.na(gm$Elevation),"Elevation"]<-predict(ri,gm[is.na(gm$Elevation),]) # Lecture 12, Slide 11
summary(lm(Species~Area+Elevation+Nearest+Scruz+Adjacent,data=gmr))
```

```
## 
## Call:
## lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent, 
##     data = gmr)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -112.11  -31.39  -15.30   24.36  194.34 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -13.03649   23.96067  -0.544 0.591407    
## Area         -0.02048    0.02508  -0.817 0.422167    
## Elevation     0.32550    0.06401   5.085 3.35e-05 ***
## Nearest       0.02966    1.15467   0.026 0.979720    
## Scruz        -0.11001    0.23946  -0.459 0.650066    
## Adjacent     -0.07916    0.02050  -3.862 0.000746 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 66.58 on 24 degrees of freedom
## Multiple R-squared:  0.7208, Adjusted R-squared:  0.6627 
## F-statistic: 12.39 on 5 and 24 DF,  p-value: 5.197e-06
```

This model with regression-based imputation appears to be stronger than the model with mean value imputation. We can see the adjusted r^2 value is higher (0.6626815 vs. 0.5774087). While it is lower than the models without imputation from Problems 4a and 4c (0.7170651 and 0.7019736 respectively), this model may still be better because it imputes missing values rather than simply deleting their entire rows.