S631 HW10

Shibi He

2. ALR 10.6

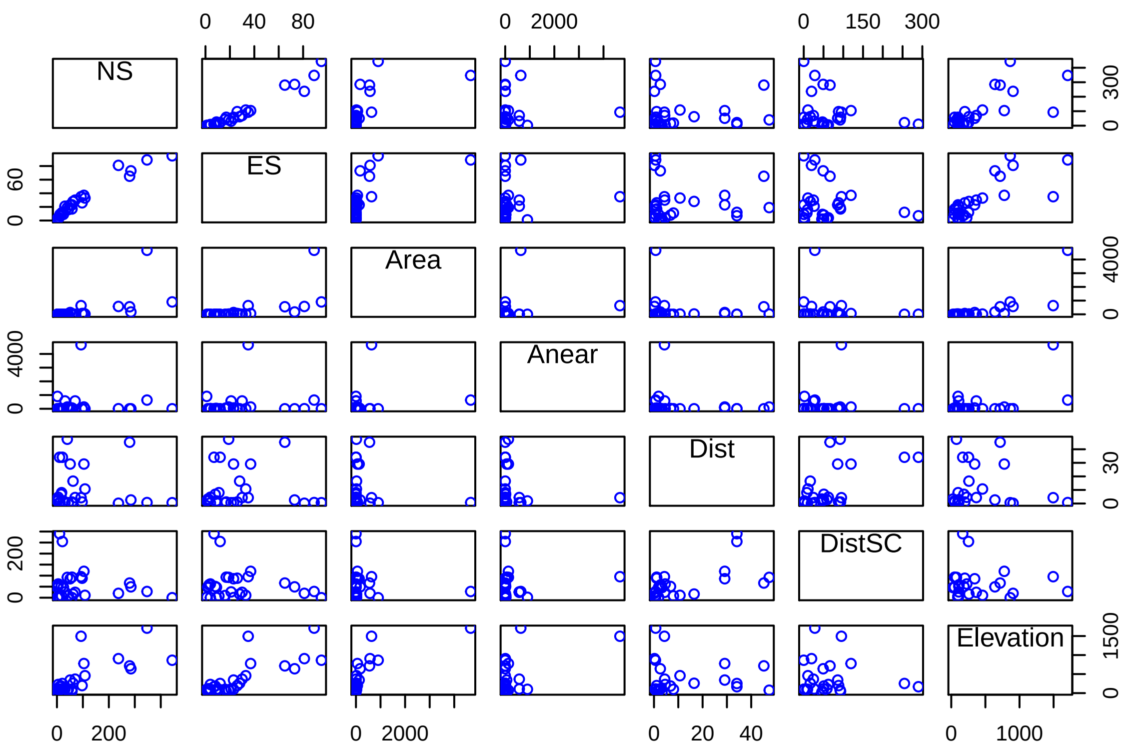
Handling the missing value in elevation: since none of the six islands exceeds 200m, I substitue the missing elevation with 100m.

I consider the ratio of endemic species to the total number of species as a measure for diversity, i.e. response variable = ES/NS.

```
#summary(galapagos)
#View(galapagos)

galapagos$Elevation[galapagos$EM == 0] = 100

scatterplotMatrix(~NS+ES+Area+Anear+Dist+DistSC+Elevation,
                  diagonal=F,
                  smooth=F,
                  regLine=F,
                  data = galapagos)
```



The scatterplot matrix show that many data points concentrate in the lower left area of the graphs and the relationships do not seem to be linear, suggesting transformation of the variables may be needed.

```
# transform predictors
# modify DistSC to be strictly positive
galapagos$DistSC = galapagos$DistSC + 0.5
```

```r
bc1 = powerTransform(cbind(Area, Anear, Dist, DistSC, Elevation) ~ 1, galapagos)
summary(bc1)
```

```
## bcPower Transformations to Multinormality
##           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Area         0.0285        0.00      -0.0725       0.1296
## Anear       -0.0441        0.00      -0.1699       0.0816
## Dist        -0.0756        0.00      -0.3141       0.1629
## DistSC       0.2841        0.33       0.0801       0.4880
## Elevation    0.0289        0.00      -0.2566       0.3143
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                    LRT df    pval
## LR test, lambda = (0 0 0 0 0) 9.003439  5 0.10893
##
## Likelihood ratio test that no transformations are needed
##                                    LRT df       pval
## LR test, lambda = (1 1 1 1 1) 570.0652  5 < 2.22e-16
```

```r
# transform response
model = lm(ES/NS ~ log(Area) + log(Dist) + log(DistSC) + log(Elevation), galapagos)
summary(powerTransform(model))
```

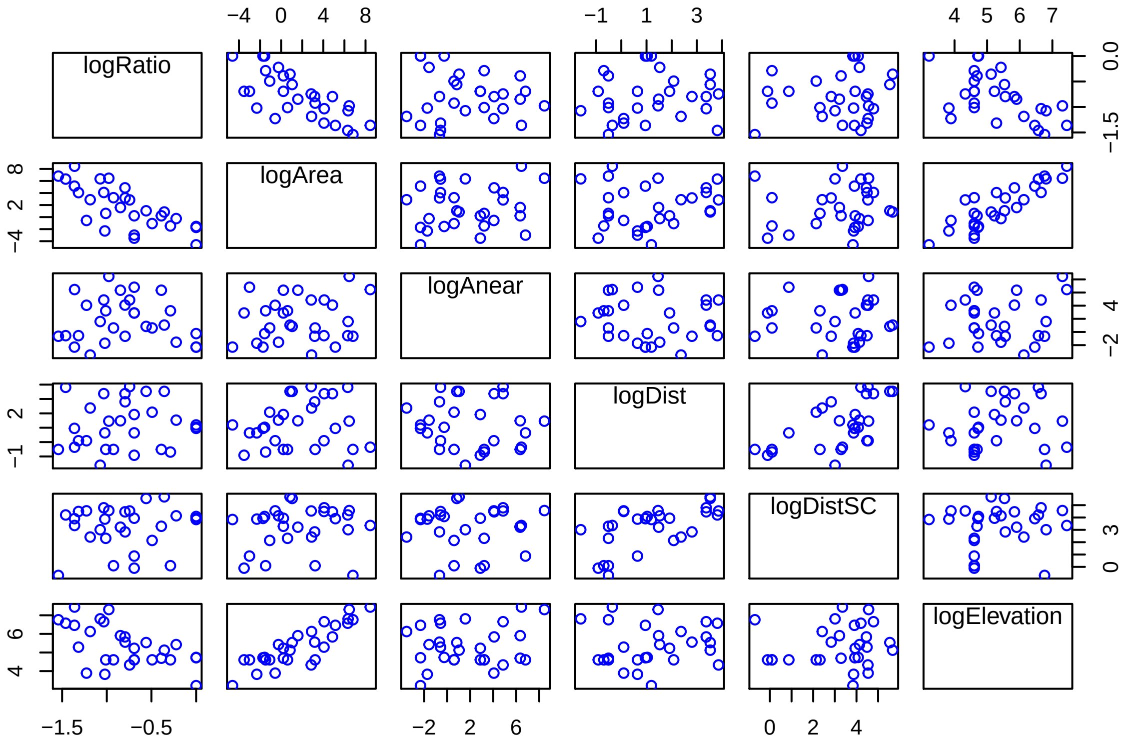
```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts
## argument ignored
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1   -0.3978           0      -1.1299       0.3343
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                            LRT df    pval
## LR test, lambda = (0) 1.131175  1 0.28752
##
## Likelihood ratio test that no transformation is needed
##                           LRT df       pval
## LR test, lambda = (1) 13.1862  1 0.00028202
```

The LR test results suggest to take log transformation for all variables. Check the scatter plots after transformation:

```r
galapagos$logRatio = log(galapagos$ES/galapagos$NS)
galapagos$logArea = log(galapagos$Area)
galapagos$logAnear = log(galapagos$Anear)
galapagos$logDist = log(galapagos$Dist)
galapagos$logDistSC = log(galapagos$DistSC)
galapagos$logElevation = log(galapagos$Elevation)

scatterplotMatrix(~logRatio+logArea+logAnear+logDist+logDistSC+logElevation,
                  diagonal=F,
                  smooth=F,
                  regLine=F,
                  data = galapagos)
```



The scatter plots look better after the log transformation.

Model selection:

```r
m3.small = lm(logRatio ~ 1, data = galapagos)
m3.full = lm(logRatio ~ logArea + logAnear +
                 logDist + logDistSC + logElevation,
                 data = galapagos)
```

```r
## Forward Selection
m3.fwd = step(m3.small, scope= ~ logArea + logAnear +
                  logDist + logDistSC + logElevation,
                  direction="forward", trace = FALSE)
m3.fwd$anova
```

```
##          Step Df  Deviance Resid. Df Resid. Dev       AIC
## 1             NA        NA        28   5.556980 -45.91499
## 2    + logArea -1 2.8914718        27   2.665508 -65.22014
## 3 + logDistSC -1 0.2197162        26   2.445791 -65.71488
```

```r
## Backward Elimination
m3.bck = step(m3.full, scope = ~ 1, direction = "backward", trace = FALSE)
m3.bck$anova
```

```
##             Step Df    Deviance Resid. Df Resid. Dev       AIC
## 1                NA          NA        23   2.332745 -61.08725
## 2      - logAnear  1 0.002205313        24   2.334950 -63.05985
## 3       - logDist  1 0.023095822        25   2.358046 -64.77441
## 4 - logElevation  1 0.087745148        26   2.445791 -65.71488
```

```r
## Bidirectional Stepwise method
m3.bi = step(m3.small,
            scope=list(lower = m3.small, upper = m3.full),
```

```r
            direction = "both", trace = FALSE)
m3.bi$anova
```

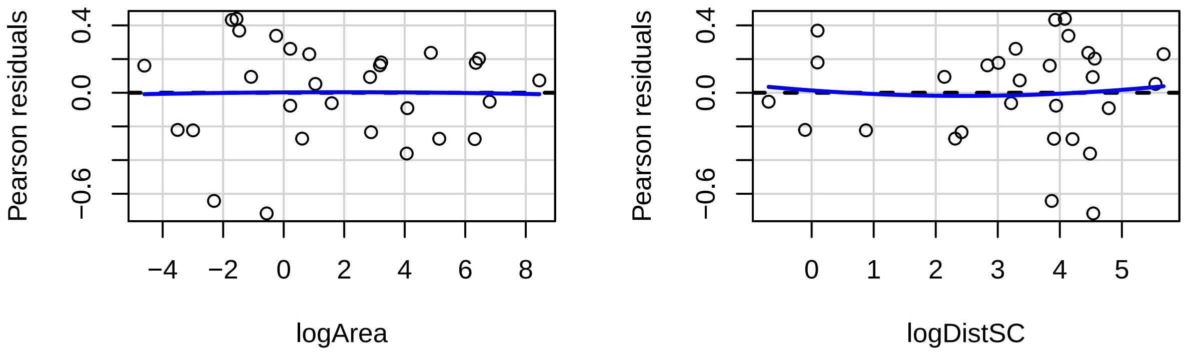
```
##          Step Df  Deviance Resid. Df Resid. Dev       AIC
## 1              NA        NA        28   5.556980 -45.91499
## 2   + logArea -1 2.8914718        27   2.665508 -65.22014
## 3 + logDistSC -1 0.2197162        26   2.445791 -65.71488
```

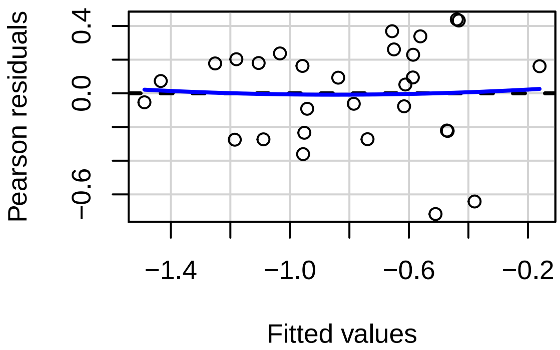
Forward selection, backward elimination, and bidirectional stepwise method all suggest the model with "logArea" and "logDistSC" has the lowest AIC of -65.71488. Therefore, "Area" and "DistSC" are two factor that influences the ratio of number of endemic species to the total number of species on an island, i.e. the diversity. The final model I consider is as follows:

```r
m.diversity = lm(logRatio ~ logArea + logDistSC, galapagos)
summary(m.diversity)
```

```
##
## Call:
## lm(formula = logRatio ~ logArea + logDistSC, data = galapagos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71663 -0.22333  0.07318  0.20257  0.43915
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.80196    0.12428  -6.453 7.74e-07 ***
## logArea     -0.09557    0.01680  -5.688 5.52e-06 ***
## logDistSC    0.05235    0.03426   1.528    0.139
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3067 on 26 degrees of freedom
## Multiple R-squared:  0.5599, Adjusted R-squared:  0.526
## F-statistic: 16.54 on 2 and 26 DF,  p-value: 2.326e-05
```

```r
residualPlots(m.diversity)
```






```
##             Test stat Pr(>|Test stat|)
## logArea       -0.0573           0.9548
## logDistSC      0.2727           0.7873
## Tukey test     0.1497           0.8810
```

The residual plots look like null plots and do not have any visual evidence for curvature. So I believe this model does not violate the assumptions for linear model.

3.a Select model to predict baseball pitchers' salaries.

```
baseball = read.table("BaseballPitchers.txt", header = TRUE)
#str(baseball)
```

```
m.small = lm(salary ~ 1, data = baseball)
m.full = lm(salary ~ team86+league86+W86+L86+ERA86+
                     G86+IP86+SV86+years+careerW+careerL+
                     careerERA+careerG+careerIP+careerSV+
                     league87+team87, data = baseball)
```

```
## Forward Selection
m.fwd = step(m.small, scope = ~ team86+league86+W86+L86+ERA86+
                     G86+IP86+SV86+years+careerW+careerL+
                     careerERA+careerG+careerIP+careerSV+
                     league87+team87,
                     direction="forward", trace = FALSE)
m.fwd$anova
```

```
##          Step  Df  Deviance Resid. Df Resid. Dev      AIC
## 1              NA        NA       175   24213875 2084.424
## 2     + years  -1 6922565.0       174   17291310 2027.161
```

```
## 3 + careerERA  -1 1204207.1       173   16087103 2016.456
## 4       + IP86  -1 1357011.2       172   14730092 2002.946
## 5    + team87 -23 3645451.0       149   11084641 1998.903
## 6  + careerSV  -1  395108.0       148   10689533 1994.515
## 7  + league87  -1  274475.7       147   10415057 1991.937
```

```
## Backward Elimination
m.bck = step(m.full, scope = ~ 1, direction = "backward", trace = FALSE)
m.bck$anova
```

```
##          Step Df    Deviance Resid. Df Resid. Dev      AIC
## 1               NA          NA       123    8628912 2006.826
## 2    - team86 14 1272685.960       137    9901598 2003.039
## 3     - ERA86  1    2410.761       138    9904009 2001.082
## 4  - careerSV  1    4935.047       139    9908944 1999.170
## 5   - careerL  1   33623.939       140    9942568 1997.766
## 6       - W86  1   38549.328       141    9981117 1996.447
## 7       - L86  1   45321.324       142   10026438 1995.244
## 8   - careerG  1   57221.880       143   10083660 1994.246
## 9  - league87  1   79611.700       144   10163272 1993.630
## 10      - G86  1  109981.619       145   10273254 1993.524
## 11     - SV86  1   67869.646       146   10341123 1992.683
```

```
## Bidirectional Stepwise method
m.bi = step(m.small,
            scope=list(lower = m.small, upper = m.full),
            direction = "both", trace = FALSE)
m.bi$anova
```

```
##          Step  Df  Deviance Resid. Df Resid. Dev      AIC
## 1                NA        NA       175   24213875 2084.424
## 2     + years  -1 6922565.0       174   17291310 2027.161
## 3 + careerERA  -1 1204207.1       173   16087103 2016.456
## 4      + IP86  -1 1357011.2       172   14730092 2002.946
## 5    + team87 -23 3645451.0       149   11084641 1998.903
## 6  + careerSV  -1  395108.0       148   10689533 1994.515
## 7  + league87  -1  274475.7       147   10415057 1991.937
```

Both forward selection and bidirectional stepwise methods suggest that the model with variables “years”, “careerERA”, “IP86”, “team87”, “careerSV”, “league87” has the lowest AIC of 1991.4. Therefore, these variables should be considered as predictors for baseball pitchers’ salaries in 1987.

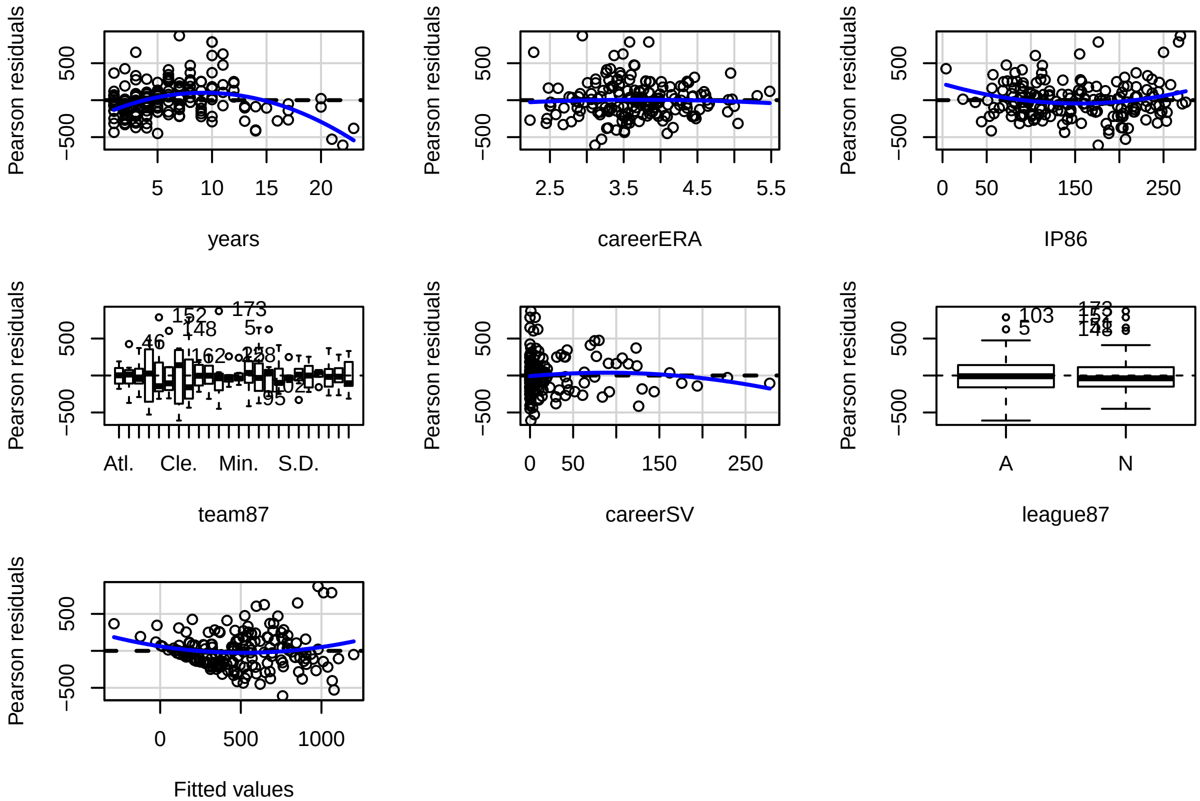
```
m.baseball = lm(salary ~ years+careerERA+IP86+
                    team87+careerSV+league87,
                    data = baseball)

summary(m.baseball)
```

```
##
## Call:
## lm(formula = salary ~ years + careerERA + IP86 + team87 + careerSV +
##     league87, data = baseball)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -609.2 -158.1  -26.6  125.1  871.1
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   57.6001   226.9173   0.254 0.799976
## years         34.2776     5.5267   6.202 5.36e-09 ***
## careerERA   -113.6827    41.9487  -2.710 0.007528 **
## IP86           1.9527     0.4111   4.750 4.78e-06 ***
## team87Bal.   443.1004   172.2881   2.572 0.011107 *
## team87Bos.   366.8161   172.5102   2.126 0.035145 *
## team87Cal.   253.6172   194.5625   1.304 0.194433
## team87Chi.   489.7911   129.4280   3.784 0.000224 ***
## team87Cin.   160.9367   144.1991   1.116 0.266212
## team87Cle.   -44.2065   178.0179  -0.248 0.804229
## team87Det.   548.6312   173.3258   3.165 0.001883 **
## team87Hou.    34.7395   134.1054   0.259 0.795962
## team87K.C.   385.5194   174.7426   2.206 0.028921 *
## team87L.A.   281.2451   133.5009   2.107 0.036843 *
## team87Mil.   217.7212   186.0869   1.170 0.243895
## team87Min.   380.3508   173.7509   2.189 0.030170 *
## team87Mon.  -148.4768   144.1560  -1.030 0.304714
## team87N.Y.   255.4010   133.5029   1.913 0.057683 .
## team87Oak.   292.4248   163.7470   1.786 0.076188 .
## team87Phi.    43.6651   138.4021   0.315 0.752834
## team87Pit.    57.3938   139.5902   0.411 0.681554
## team87S.D.   147.4488   138.6138   1.064 0.289192
## team87S.F.     6.6883   144.6113   0.046 0.963174
## team87Sea.   199.1914   187.0554   1.065 0.288677
## team87St.L.   82.6906   145.6265   0.568 0.571019
## team87Tex.   197.2861   169.4256   1.164 0.246132
## team87Tor.   310.1011   174.6704   1.775 0.077909 .
## careerSV       1.3671     0.6250   2.187 0.030311 *
## league87N    207.4445   105.3955   1.968 0.050921 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 266.2 on 147 degrees of freedom
##   (30 observations deleted due to missingness)
## Multiple R-squared:  0.5699, Adjusted R-squared:  0.4879
## F-statistic: 6.956 on 28 and 147 DF,  p-value: 8.51e-16
```

```r
residualPlots(m.baseball)
```




```
##              Test stat Pr(>|Test stat|)
## years          -6.7790        2.785e-10 ***
## careerERA      -0.4454          0.65671
## IP86            2.4940          0.01375 *
## team87
## careerSV       -1.1971          0.23322
## league87
## Tukey test      1.8417          0.06551 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The adjusted R-squared is 0.4879, suggesting that the model only explains 48.8% of the variation in the baseball pitchers' salaries. Moreover, the residual plots show that the residuals are somewhat large, suggesting the model is not very good at predicting salaries. The model makes substantive sense as it takes into account the players' performance over their entire career as well as which team they belong to in 1987 to predict the salaries.

3.b Cross-validation

```r
# Randomly split the data into two subsamples
set.seed(631)
n = nrow(baseball)
train_ind <- sample(seq_len(n), size =n/2 )

train <- baseball[train_ind, ]
test <- baseball[-train_ind, ]
```

```r
### Model selection on the train data
model.small = lm(salary ~ 1, data = train)
model.full = lm(salary ~ team86+league86+W86+L86+ERA86+
                     G86+IP86+SV86+years+careerW+careerL+
                     careerERA+careerG+careerIP+careerSV+
                     league87+team87, data = train)

## Forward Selection
model.fwd = step(model.small, scope = ~ team86+league86+W86+L86+ERA86+
                     G86+IP86+SV86+years+careerW+careerL+
                     careerERA+careerG+careerIP+careerSV+
                     league87+team87,
                     direction="forward", trace = FALSE)
model.fwd$anova
```

```
##          Step  Df   Deviance Resid. Df Resid. Dev      AIC
## 1              NA         NA        88   11995540 1053.215
## 2   + careerW  -1 2157924.91        87    9837615 1037.565
## 3 + careerERA  -1 1136124.19        86    8701491 1028.643
## 4      + IP86  -1  799778.26        85    7901712 1022.062
## 5  + careerSV  -1  717757.16        84    7183955 1015.586
## 6       + L86  -1  263861.70        83    6920093 1014.256
## 7    + team87 -23 2831969.68        60    4088124 1013.411
## 8  + league87  -1  171530.84        59    3916593 1011.597
## 9  + careerIP  -1   89290.25        58    3827303 1011.544
```

```r
## Backward Elimination
model.bck = step(model.full, scope = ~ 1, direction = "backward", trace = FALSE)
model.bck$anova
```

```
##            Step Df    Deviance Resid. Df Resid. Dev      AIC
## 1               NA          NA        41    3048027 1025.282
## 2      - team86  9 505231.3317        50    3553258 1020.932
## 3         - G86  1    311.4104        51    3553570 1018.940
## 4    - careerSV  1   1035.7083        52    3554605 1016.966
## 5    - league86  1   6113.1561        53    3560719 1015.118
## 6       - ERA86  1   6086.3718        54    3566805 1013.270
## 7     - careerL  1  21796.9200        55    3588602 1011.813
## 8   - careerERA  1  28770.3303        56    3617372 1010.523
## 9       - years  1  21814.6094        57    3639187 1009.058
## 10        - W86  1  68814.3018        58    3708001 1008.726
```

```r
## Bidirectional Stepwise method
model.bi = step(model.small,
            scope=list(lower = model.small, upper = model.full),
            direction = "both", trace = FALSE)
model.bi$anova
```

```
##            Step  Df   Deviance Resid. Df Resid. Dev      AIC
## 1                NA         NA        88   11995540 1053.215
## 2     + careerW  -1 2157924.91        87    9837615 1037.565
## 3   + careerERA  -1 1136124.19        86    8701491 1028.643
## 4        + IP86  -1  799778.26        85    7901712 1022.062
## 5    + careerSV  -1  717757.16        84    7183955 1015.586
## 6         + L86  -1  263861.70        83    6920093 1014.256
## 7      + team87 -23 2831969.68        60    4088124 1013.411
```

```
## 8   + league87   -1  171530.84         59    3916593 1011.597
## 9   - careerERA   1   45828.89         60    3962422 1010.632
## 10  + careerIP   -1  111500.88         59    3850921 1010.092
```

The forward selection method gives a model with careerW, careerERA, IP86, careerSV, L86, team87, league87, careerIP (AIC = 1011.544).

The Backward Elimination method gives a model with L86, IP86, SV86, careerW, careerG, careerID, league87, and team87 (AIC = 1008.73).

The bidirectional stepwise method gives a model with careerW, IP86, careerSV, L86, team87, league87, careerIP (AIC = 1010.092).

```r
# Three slected models
m.forward = lm(salary ~ careerW + careerERA + IP86 +
                   careerSV + L86 + team87 + league87 + careerIP,
                 data = train)

m.backward = lm(salary ~ L86 + IP86 + SV86 + careerW +
                  careerG + careerIP + league87 + team87,
                data = train)

m.bidirect = lm(salary ~ careerW + IP86 + careerSV +
                    L86 + team87 + league87 + careerIP,
                data = train)
```

```r
### Evaluate the selected models on the test data

pred.forward = predict(m.forward, newdata = test, type = "response")
pred.backward = predict(m.backward, newdata = test, type = "response")
pred.bidirect = predict(m.bidirect, newdata = test, type = "response")
name = paste(train$firstName, train$lastName, sep=" ")

compare.df = data.frame(Name = name,
                        trueSalary = test$salary,
                        forward = pred.forward,
                        backward = pred.backward,
                        bidirect = pred.bidirect)

# compute prediction errors
compare.df$forward.error = compare.df$forward - compare.df$trueSalary
compare.df$backward.error = compare.df$backward - compare.df$trueSalary
compare.df$bidirect.error = compare.df$bidirect - compare.df$trueSalary

# compute SD of the prediction errors
SD.forward = sd(na.omit(compare.df$forward.error))
SD.backward = sd(na.omit(compare.df$backward.error))
SD.bidirect = sd(na.omit(compare.df$bidirect.error))

SD.forward
```

```
## [1] 353.7076
```

```r
SD.backward
```

```
## [1] 350.769
```

```
SD.bidirect
```

```
## [1] 357.8788
```

I found the predicted values are not very close to the ture salary, suggesting the models are not very good at predicting. To compare models using different selection methods, I compare the standard deviations of their prediction errors. While all three models have large standard deviation, the model using forward selection has the relatively small standard deviation, suggesting this model is slightly better.