STAT-S632

Assignment 1

John Koo

```r
# packages, etc.
import::from(magrittr, `%>%`, `%<>%`)
dp <- loadNamespace('dplyr')
library(ggplot2)
import::from(GGally, ggpairs)
```

Problem 1

[From ALR 10.2]

```r
# load the data
highway.df <- alr4::Highway %>%
  dp$mutate(sigs1 = (sigs * len + 1) / len)
```

Part 1

Forward selection

```r
# formula for full model
full.formula <- ~ log(len) + shld + log(adt) +
  log(trks) + lane + slim +
  lwid + itg + log(sigs1) +
  acpt + htype

# forward selection using AIC
forward.mod <- lm(log(rate) ~ log(len), data = highway.df) %>%
  step(scope = full.formula, direction = 'forward')
```

```
Start:  AIC=-72.51
log(rate) ~ log(len)

             Df Sum of Sq    RSS     AIC
+ slim        1   2.54718 2.9366 -94.866
+ acpt        1   2.10148 3.3823 -89.355
+ shld        1   1.70693 3.7769 -85.052
+ log(sigs1)  1   0.96128 4.5225 -78.025
+ htype       3   1.33997 4.1438 -77.436
+ log(trks)   1   0.72812 4.7557 -76.065
+ log(adt)    1   0.42857 5.0552 -73.682
<none>                    5.4838 -72.509
+ lane        1   0.26267 5.2211 -72.423
+ itg         1   0.21704 5.2667 -72.084
+ lwid        1   0.18502 5.2988 -71.847
```

```
Step:  AIC=-94.87
log(rate) ~ log(len) + slim

             Df Sum of Sq    RSS     AIC
+ acpt        1   0.28844 2.6482 -96.898
+ log(trks)   1   0.26317 2.6734 -96.528
<none>                    2.9366 -94.866
+ log(sigs1)  1   0.14671 2.7899 -94.865
+ htype       3   0.33646 2.6002 -93.612
+ shld        1   0.03265 2.9040 -93.302
+ log(adt)    1   0.02563 2.9110 -93.208
+ lwid        1   0.01664 2.9200 -93.088
+ lane        1   0.00343 2.9332 -92.912
+ itg         1   0.00265 2.9340 -92.901

Step:  AIC=-96.9
log(rate) ~ log(len) + slim + acpt

             Df Sum of Sq    RSS     AIC
+ log(trks)   1  0.172940 2.4752 -97.532
<none>                    2.6482 -96.898
+ log(sigs1)  1  0.120061 2.5281 -96.708
+ shld        1  0.034595 2.6136 -95.411
+ log(adt)    1  0.015190 2.6330 -95.122
+ lane        1  0.014872 2.6333 -95.118
+ itg         1  0.013501 2.6347 -95.097
+ lwid        1  0.012646 2.6355 -95.085
+ htype       3  0.217478 2.4307 -94.240

Step:  AIC=-97.53
log(rate) ~ log(len) + slim + acpt + log(trks)

             Df Sum of Sq    RSS     AIC
<none>                    2.4752 -97.532
+ shld        1  0.065299 2.4099 -96.575
+ log(sigs1)  1  0.050568 2.4247 -96.337
+ log(adt)    1  0.031220 2.4440 -96.027
+ htype       3  0.259505 2.2157 -95.851
+ lwid        1  0.019009 2.4562 -95.833
+ itg         1  0.010964 2.4643 -95.705
+ lane        1  0.003299 2.4719 -95.584
```

```
summary(forward.mod)
```

```

Call:
lm(formula = log(rate) ~ log(len) + slim + acpt + log(trks),
    data = highway.df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.43125 -0.17980  0.03907  0.16660  0.55657

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  4.166541   0.741065   5.622 2.67e-06 ***
log(len)    -0.235735   0.084897  -2.777  0.00887 **
slim        -0.031852   0.010262  -3.104  0.00383 **
acpt         0.011004   0.006669   1.650  0.10815
log(trks)   -0.329037   0.213484  -1.541  0.13251
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2698 on 34 degrees of freedom
Multiple R-squared:  0.6961,    Adjusted R-squared:  0.6603
F-statistic: 19.47 on 4 and 34 DF,  p-value: 2.067e-08
```

Backward elimination

```
# backward elimination
backward.mod <- step(forward.mod, scope = c(lower = ~ log(len)),
                     direction = 'backward')
```

```
Start:  AIC=-97.53
log(rate) ~ log(len) + slim + acpt + log(trks)

            Df Sum of Sq    RSS     AIC
<none>                   2.4752 -97.532
- log(trks)  1   0.17294 2.6482 -96.898
- acpt       1   0.19821 2.6734 -96.528
- slim       1   0.70140 3.1766 -89.802
```

Part 2

```
# model for log(rate * len) that includes lwid
# using all three methods

part.2.forward.mod <- lm(log(rate * len) ~ lwid, data = highway.df) %>%
  step(scope = full.formula, direction = 'forward')
```

```
Start:  AIC=-54.06
log(rate * len) ~ lwid

             Df Sum of Sq    RSS     AIC
+ log(len)    1    3.5027 5.2988 -71.847
+ shld        1    2.3680 6.4335 -64.280
+ log(adt)    1    1.5735 7.2280 -59.738
+ htype       3    1.6682 7.1333 -56.253
+ lane        1    0.8697 7.9318 -56.115
+ slim        1    0.7962 8.0053 -55.755
+ itg         1    0.7002 8.1013 -55.290
+ acpt        1    0.4564 8.3451 -54.134
<none>                   8.8015 -54.057
+ log(sigs1)  1    0.0832 8.7183 -52.427
+ log(trks)   1    0.0238 8.7776 -52.163

Step:  AIC=-71.85
```

```
log(rate * len) ~ lwid + log(len)

             Df Sum of Sq    RSS     AIC
+ slim        1   2.37880 2.9200 -93.088
+ acpt        1   1.96443 3.3343 -87.912
+ shld        1   1.79989 3.4989 -86.034
+ log(sigs1)  1   0.86898 4.4298 -76.833
+ htype       3   1.18982 4.1089 -75.765
+ log(trks)   1   0.73475 4.5640 -75.669
+ log(adt)    1   0.36312 4.9356 -72.616
<none>                    5.2988 -71.847
+ lane        1   0.25251 5.0463 -71.752
+ itg         1   0.20235 5.0964 -71.366

Step:  AIC=-93.09
log(rate * len) ~ lwid + log(len) + slim

             Df Sum of Sq    RSS     AIC
+ acpt        1   0.28444 2.6355 -95.085
+ log(trks)   1   0.27104 2.6489 -94.887
<none>                    2.9200 -93.088
+ log(sigs1)  1   0.14118 2.7788 -93.020
+ shld        1   0.05556 2.8644 -91.837
+ htype       3   0.32443 2.5955 -91.681
+ log(adt)    1   0.02261 2.8973 -91.391
+ lane        1   0.00275 2.9172 -91.124
+ itg         1   0.00233 2.9176 -91.119

Step:  AIC=-95.08
log(rate * len) ~ lwid + log(len) + slim + acpt

             Df Sum of Sq    RSS     AIC
+ log(trks)   1  0.179304 2.4562 -95.833
<none>                    2.6355 -95.085
+ log(sigs1)  1  0.115847 2.5197 -94.838
+ shld        1  0.055489 2.5800 -93.915
+ lane        1  0.013518 2.6220 -93.285
+ log(adt)    1  0.013206 2.6223 -93.281
+ itg         1  0.012767 2.6227 -93.274
+ htype       3  0.212804 2.4227 -92.368

Step:  AIC=-95.83
log(rate * len) ~ lwid + log(len) + slim + acpt + log(trks)

             Df Sum of Sq    RSS     AIC
<none>                    2.4562 -95.833
+ shld        1  0.103548 2.3527 -95.512
+ log(sigs1)  1  0.045814 2.4104 -94.567
+ log(adt)    1  0.028029 2.4282 -94.280
+ htype       3  0.254456 2.2018 -94.098
+ itg         1  0.010106 2.4461 -93.993
+ lane        1  0.002395 2.4538 -93.871
```

```
part.2.backward.mod <- lm(as.formula(paste('log(rate * len)',
                                           paste(as.character(full.formula),
                                                 collapse = ' '))),
                          data = highway.df) %>%
  step(scope = c(lower = ~ lwid), direction = 'backward')
```

```
Start:  AIC=-94.2
log(rate * len) ~ log(len) + shld + log(adt) + log(trks) + lane + 
    slim + lwid + itg + log(sigs1) + acpt + htype

             Df Sum of Sq    RSS     AIC
- shld        1    0.0005 1.6999 -96.188
- itg         1    0.0015 1.7008 -96.166
- lane        1    0.0026 1.7019 -96.140
- acpt        1    0.0379 1.7372 -95.339
- log(trks)   1    0.0461 1.7455 -95.155
<none>                    1.6993 -94.199
- htype       3    0.3004 1.9998 -93.850
- log(adt)    1    0.1298 1.8291 -93.329
- slim        1    0.1790 1.8783 -92.294
- log(sigs1)  1    0.4426 2.1420 -87.172
- log(len)    1    4.1956 5.8949 -47.689

Step:  AIC=-96.19
log(rate * len) ~ log(len) + log(adt) + log(trks) + lane + slim + 
    lwid + itg + log(sigs1) + acpt + htype

             Df Sum of Sq    RSS     AIC
- itg         1    0.0013 1.7012 -98.157
- lane        1    0.0027 1.7026 -98.125
- acpt        1    0.0468 1.7466 -97.129
- log(trks)   1    0.0556 1.7555 -96.932
<none>                    1.6999 -96.188
- htype       3    0.3284 2.0283 -95.298
- log(adt)    1    0.1365 1.8364 -95.175
- slim        1    0.3405 2.0404 -91.067
- log(sigs1)  1    0.4814 2.1813 -88.463
- log(len)    1    4.5463 6.2462 -47.432

Step:  AIC=-98.16
log(rate * len) ~ log(len) + log(adt) + log(trks) + lane + slim + 
    lwid + log(sigs1) + acpt + htype

             Df Sum of Sq    RSS      AIC
- lane        1    0.0025 1.7037 -100.100
- acpt        1    0.0455 1.7467  -99.127
- log(trks)   1    0.0568 1.7580  -98.877
<none>                    1.7012  -98.157
- log(adt)    1    0.1552 1.8564  -96.752
- htype       3    0.5597 2.2609  -93.065
- slim        1    0.3795 2.0807  -92.304
- log(sigs1)  1    0.4812 2.1823  -90.443
- log(len)    1    4.5777 6.2788  -49.229
```

```
Step:  AIC=-100.1
log(rate * len) ~ log(len) + log(adt) + log(trks) + slim + lwid +
    log(sigs1) + acpt + htype

             Df Sum of Sq    RSS      AIC
- acpt        1    0.0469 1.7506 -101.040
- log(trks)   1    0.0548 1.7585 -100.865
<none>                    1.7037 -100.100
- log(adt)    1    0.1834 1.8871  -98.113
- slim        1    0.3845 2.0882  -94.164
- htype       3    0.6129 2.3166  -94.115
- log(sigs1)  1    0.4894 2.1930  -92.253
- log(len)    1    4.5822 6.2859  -51.185

Step:  AIC=-101.04
log(rate * len) ~ log(len) + log(adt) + log(trks) + slim + lwid +
    log(sigs1) + htype

             Df Sum of Sq    RSS      AIC
- log(trks)   1    0.0597 1.8103 -101.733
<none>                    1.7506 -101.040
- log(adt)    1    0.1593 1.9099  -99.645
- htype       3    0.7296 2.4802  -93.454
- log(sigs1)  1    0.5573 2.3079  -92.262
- slim        1    0.5777 2.3283  -91.919
- log(len)    1    4.5357 6.2863  -53.182

Step:  AIC=-101.73
log(rate * len) ~ log(len) + log(adt) + slim + lwid + log(sigs1) +
    htype

             Df Sum of Sq    RSS      AIC
<none>                    1.8103 -101.733
- log(adt)    1    0.1538 1.9640 -100.554
- slim        1    0.5562 2.3664  -93.285
- htype       3    0.8266 2.6369  -93.065
- log(sigs1)  1    0.7656 2.5759  -89.977
- log(len)    1    4.5348 6.3451  -54.820
```

```r
part.2.both.mod <- lm(log(rate * len) ~ lwid, data = highway.df) %>%
  step(scope = list(lower = ~ lwid, upper = full.formula), direction = 'both')
```

```
Start:  AIC=-54.06
log(rate * len) ~ lwid

             Df Sum of Sq    RSS     AIC
+ log(len)    1    3.5027 5.2988 -71.847
+ shld        1    2.3680 6.4335 -64.280
+ log(adt)    1    1.5735 7.2280 -59.738
+ htype       3    1.6682 7.1333 -56.253
+ lane        1    0.8697 7.9318 -56.115
+ slim        1    0.7962 8.0053 -55.755
+ itg         1    0.7002 8.1013 -55.290
+ acpt        1    0.4564 8.3451 -54.134
<none>                    8.8015 -54.057
```

```
+ log(sigs1)  1    0.0832 8.7183 -52.427
+ log(trks)   1    0.0238 8.7776 -52.163

Step:  AIC=-71.85
log(rate * len) ~ lwid + log(len)

             Df Sum of Sq    RSS     AIC
+ slim        1    2.3788 2.9200 -93.088
+ acpt        1    1.9644 3.3343 -87.912
+ shld        1    1.7999 3.4989 -86.034
+ log(sigs1)  1    0.8690 4.4298 -76.833
+ htype       3    1.1898 4.1089 -75.765
+ log(trks)   1    0.7347 4.5640 -75.669
+ log(adt)    1    0.3631 4.9356 -72.616
<none>                    5.2988 -71.847
+ lane        1    0.2525 5.0463 -71.752
+ itg         1    0.2023 5.0964 -71.366
- log(len)    1    3.5027 8.8015 -54.057

Step:  AIC=-93.09
log(rate * len) ~ lwid + log(len) + slim

             Df Sum of Sq    RSS     AIC
+ acpt        1    0.2844 2.6355 -95.085
+ log(trks)   1    0.2710 2.6489 -94.887
<none>                    2.9200 -93.088
+ log(sigs1)  1    0.1412 2.7788 -93.020
+ shld        1    0.0556 2.8644 -91.837
+ htype       3    0.3244 2.5955 -91.681
+ log(adt)    1    0.0226 2.8974 -91.391
+ lane        1    0.0028 2.9172 -91.124
+ itg         1    0.0023 2.9176 -91.119
- slim        1    2.3788 5.2988 -71.847
- log(len)    1    5.0853 8.0053 -55.755

Step:  AIC=-95.08
log(rate * len) ~ lwid + log(len) + slim + acpt

             Df Sum of Sq    RSS     AIC
+ log(trks)   1    0.1793 2.4562 -95.833
<none>                    2.6355 -95.085
+ log(sigs1)  1    0.1158 2.5197 -94.838
+ shld        1    0.0555 2.5800 -93.915
+ lane        1    0.0135 2.6220 -93.285
+ log(adt)    1    0.0132 2.6223 -93.281
+ itg         1    0.0128 2.6228 -93.274
- acpt        1    0.2844 2.9200 -93.088
+ htype       3    0.2128 2.4227 -92.368
- slim        1    0.6988 3.3343 -87.912
- log(len)    1    5.3612 7.9967 -53.797

Step:  AIC=-95.83
log(rate * len) ~ lwid + log(len) + slim + acpt + log(trks)
```

```
             Df Sum of Sq    RSS     AIC
<none>                    2.4562 -95.833
+ shld        1    0.1035 2.3527 -95.512
- log(trks)   1    0.1793 2.6355 -95.085
- acpt        1    0.1927 2.6489 -94.887
+ log(sigs1)  1    0.0458 2.4104 -94.567
+ log(adt)    1    0.0280 2.4282 -94.280
+ htype       3    0.2545 2.2018 -94.098
+ itg         1    0.0101 2.4461 -93.993
+ lane        1    0.0024 2.4538 -93.871
- slim        1    0.6608 3.1171 -88.540
- log(len)    1    5.2642 7.7205 -53.168
```

```r
summary(part.2.forward.mod)
```

```
Call:
lm(formula = log(rate * len) ~ lwid + log(len) + slim + acpt + 
    log(trks), data = highway.df)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.51518 -0.16169  0.03966  0.17333  0.55629 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.80121    1.46241   3.283  0.00243 ** 
lwid        -0.05235    0.10359  -0.505  0.61666    
log(len)     0.75168    0.08938   8.410 1.02e-09 ***
slim        -0.03117    0.01046  -2.980  0.00538 ** 
acpt         0.01086    0.00675   1.609  0.11713    
log(trks)   -0.33565    0.21626  -1.552  0.13018    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2728 on 33 degrees of freedom
Multiple R-squared:  0.7504,    Adjusted R-squared:  0.7125 
F-statistic: 19.84 on 5 and 33 DF,  p-value: 4.275e-09
```

```r
summary(part.2.backward.mod)
```

```
Call:
lm(formula = log(rate * len) ~ log(len) + log(adt) + slim + lwid + 
    log(sigs1) + htype, data = highway.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.4707 -0.1212 -0.0209  0.1019  0.4535 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.655848   1.327821   2.753  0.00992 ** 
log(len)     0.753327   0.086899   8.669 1.14e-09 ***
log(adt)    -0.136162   0.085296  -1.596  0.12089    
```

```
slim        -0.029711   0.009787  -3.036  0.00492 **
lwid         0.056762   0.098322   0.577  0.56804
log(sigs1)   0.217015   0.060925   3.562  0.00125 **
htypefai     0.157704   0.345084   0.457  0.65096
htypepa     -0.362487   0.232778  -1.557  0.12991
htypema     -0.123393   0.206890  -0.596  0.55537
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2456 on 30 degrees of freedom
Multiple R-squared:  0.816, Adjusted R-squared:  0.7669
F-statistic: 16.63 on 8 and 30 DF,  p-value: 4.336e-09
```

```
summary(part.2.both.mod)
```

```

Call:
lm(formula = log(rate * len) ~ lwid + log(len) + slim + acpt +
    log(trks), data = highway.df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51518 -0.16169  0.03966  0.17333  0.55629

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.80121    1.46241   3.283  0.00243 **
lwid        -0.05235    0.10359  -0.505  0.61666
log(len)     0.75168    0.08938   8.410 1.02e-09 ***
slim        -0.03117    0.01046  -2.980  0.00538 **
acpt         0.01086    0.00675   1.609  0.11713
log(trks)   -0.33565    0.21626  -1.552  0.13018
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2728 on 33 degrees of freedom
Multiple R-squared:  0.7504,    Adjusted R-squared:  0.7125
F-statistic: 19.84 on 5 and 33 DF,  p-value: 4.275e-09
```

The model found by backward elimination resulted in the lowest AIC value. It also is the largest model.

All three models have `log(len)` with the same coefficient estimate, which is expected.

The model found by backward elimination resulted in a positive coefficient estimate for `lwid` while the other two found a model with a negative coefficient for `lwid`. In either case, the result is not significant ($p > 0.5$).

Part 3

```
# model for log(rate) with offset = len
# using all three methods

part.3.forward.mod <- lm(log(rate) ~ lwid, data = highway.df,
                         offset = log(len)) %>%
  step(scope = full.formula, direction = 'forward')
```

```
Start:  AIC=-2.77
log(rate) ~ lwid

             Df Sum of Sq    RSS     AIC
+ log(len)    1   27.4878  5.299 -71.847
+ log(sigs1)  1   13.9244 18.862 -22.330
+ log(trks)   1   10.7906 21.996 -16.335
+ acpt        1    9.6906 23.096 -14.432
+ slim        1    9.4825 23.304 -14.082
+ log(adt)    1    1.9238 30.863  -3.126
<none>                    32.787  -2.768
+ shld        1    0.5750 32.212  -1.458
+ lane        1    0.5730 32.214  -1.456
+ itg         1    0.4531 32.334  -1.311
+ htype       3    1.1578 31.629   1.830

Step:  AIC=-71.85
log(rate) ~ lwid + log(len)

             Df Sum of Sq    RSS     AIC
+ slim        1   2.37880 2.9200 -93.088
+ acpt        1   1.96443 3.3343 -87.912
+ shld        1   1.79989 3.4989 -86.034
+ log(sigs1)  1   0.86898 4.4298 -76.833
+ htype       3   1.18982 4.1089 -75.765
+ log(trks)   1   0.73475 4.5640 -75.669
+ log(adt)    1   0.36312 4.9356 -72.616
<none>                    5.2988 -71.847
+ lane        1   0.25251 5.0463 -71.752
+ itg         1   0.20235 5.0964 -71.366

Step:  AIC=-93.09
log(rate) ~ lwid + log(len) + slim

             Df Sum of Sq    RSS     AIC
+ acpt        1   0.28444 2.6355 -95.085
+ log(trks)   1   0.27104 2.6489 -94.887
<none>                    2.9200 -93.088
+ log(sigs1)  1   0.14118 2.7788 -93.020
+ shld        1   0.05556 2.8644 -91.837
+ htype       3   0.32443 2.5955 -91.681
+ log(adt)    1   0.02261 2.8973 -91.391
+ lane        1   0.00275 2.9172 -91.124
+ itg         1   0.00233 2.9176 -91.119

Step:  AIC=-95.08
log(rate) ~ lwid + log(len) + slim + acpt

             Df Sum of Sq    RSS     AIC
+ log(trks)   1  0.179304 2.4562 -95.833
<none>                    2.6355 -95.085
+ log(sigs1)  1  0.115847 2.5197 -94.838
+ shld        1  0.055489 2.5800 -93.915
+ lane        1  0.013518 2.6220 -93.285
+ log(adt)    1  0.013206 2.6223 -93.281
+ itg         1  0.012767 2.6227 -93.274
+ htype       3  0.212804 2.4227 -92.368

Step:  AIC=-95.83
log(rate) ~ lwid + log(len) + slim + acpt + log(trks)

             Df Sum of Sq    RSS     AIC
<none>                    2.4562 -95.833
+ shld        1  0.103548 2.3527 -95.512
+ log(sigs1)  1  0.045814 2.4104 -94.567
+ log(adt)    1  0.028029 2.4282 -94.280
+ htype       3  0.254456 2.2018 -94.098
+ itg         1  0.010106 2.4461 -93.993
+ lane        1  0.002395 2.4538 -93.871
```

```
part.3.backward.mod <- lm(as.formula(paste('log(rate)',
                                           paste(as.character(full.formula),
                                                 collapse = ' '))),
                          data = highway.df,
                          offset = log(len)) %>%
  step(scope = c(lower = ~ lwid), direction = 'backward')
```

```
Start:  AIC=-94.2
log(rate) ~ log(len) + shld + log(adt) + log(trks) + lane + slim +
    lwid + itg + log(sigs1) + acpt + htype

             Df Sum of Sq     RSS     AIC
- shld        1    0.0005  1.6999 -96.188
- itg         1    0.0015  1.7008 -96.166
- lane        1    0.0026  1.7019 -96.140
- acpt        1    0.0379  1.7372 -95.339
- log(trks)   1    0.0461  1.7455 -95.155
<none>                     1.6993 -94.199
- htype       3    0.3004  1.9998 -93.850
- log(adt)    1    0.1298  1.8291 -93.329
- slim        1    0.1790  1.8783 -92.294
- log(sigs1)  1    0.4426  2.1420 -87.172
- log(len)    1   10.0285 11.7279 -20.862

Step:  AIC=-96.19
log(rate) ~ log(len) + log(adt) + log(trks) + lane + slim + lwid +
    itg + log(sigs1) + acpt + htype

             Df Sum of Sq     RSS     AIC
- itg         1    0.0013  1.7012 -98.157
- lane        1    0.0027  1.7026 -98.125
- acpt        1    0.0468  1.7466 -97.129
- log(trks)   1    0.0556  1.7555 -96.932
<none>                     1.6999 -96.188
- htype       3    0.3284  2.0283 -95.298
- log(adt)    1    0.1365  1.8364 -95.175
- slim        1    0.3405  2.0404 -91.067
- log(sigs1)  1    0.4814  2.1813 -88.463
- log(len)    1   10.9819 12.6818 -19.812
```

```
Step:  AIC=-98.16
log(rate) ~ log(len) + log(adt) + log(trks) + lane + slim + lwid +
    log(sigs1) + acpt + htype

             Df Sum of Sq     RSS      AIC
- lane        1    0.0025  1.7037 -100.100
- acpt        1    0.0455  1.7467  -99.127
- log(trks)   1    0.0568  1.7580  -98.877
<none>                     1.7012  -98.157
- log(adt)    1    0.1552  1.8564  -96.752
- htype       3    0.5597  2.2609  -93.065
- slim        1    0.3795  2.0807  -92.304
- log(sigs1)  1    0.4812  2.1823  -90.443
- log(len)    1   11.1208 12.8220  -21.384

Step:  AIC=-100.1
log(rate) ~ log(len) + log(adt) + log(trks) + slim + lwid + log(sigs1) +
    acpt + htype

             Df Sum of Sq     RSS      AIC
- acpt        1    0.0469  1.7506 -101.040
- log(trks)   1    0.0548  1.7585 -100.865
<none>                     1.7037 -100.100
- log(adt)    1    0.1834  1.8871  -98.113
- slim        1    0.3845  2.0882  -94.164
- htype       3    0.6129  2.3166  -94.115
- log(sigs1)  1    0.4894  2.1930  -92.253
- log(len)    1   11.1850 12.8887  -23.181

Step:  AIC=-101.04
log(rate) ~ log(len) + log(adt) + log(trks) + slim + lwid + log(sigs1) +
    htype

             Df Sum of Sq     RSS      AIC
- log(trks)   1    0.0597  1.8103 -101.733
<none>                     1.7506 -101.040
- log(adt)    1    0.1593  1.9099  -99.645
- htype       3    0.7296  2.4802  -93.454
- log(sigs1)  1    0.5573  2.3079  -92.262
- slim        1    0.5777  2.3283  -91.919
- log(len)    1   11.4122 13.1628  -24.361

Step:  AIC=-101.73
log(rate) ~ log(len) + log(adt) + slim + lwid + log(sigs1) +
    htype

             Df Sum of Sq     RSS      AIC
<none>                     1.8103 -101.733
- log(adt)    1    0.1538  1.9640 -100.554
- slim        1    0.5562  2.3664  -93.285
- htype       3    0.8266  2.6369  -93.065
- log(sigs1)  1    0.7656  2.5759  -89.977
- log(len)    1   12.4192 14.2295  -23.322
```

```r
part.3.both.mod <- lm(log(rate) ~ lwid, data = highway.df, offset = log(len)) %>%
  step(scope = list(lower = ~ lwid, upper = full.formula), direction = 'both')
```

```
Start:  AIC=-2.77
log(rate) ~ lwid

             Df Sum of Sq    RSS     AIC
+ log(len)    1   27.4878  5.299 -71.847
+ log(sigs1)  1   13.9244 18.862 -22.330
+ log(trks)   1   10.7906 21.996 -16.335
+ acpt        1    9.6906 23.096 -14.432
+ slim        1    9.4825 23.304 -14.082
+ log(adt)    1    1.9238 30.863  -3.126
<none>                    32.787  -2.768
+ shld        1    0.5750 32.212  -1.458
+ lane        1    0.5730 32.214  -1.456
+ itg         1    0.4531 32.334  -1.311
+ htype       3    1.1578 31.629   1.830

Step:  AIC=-71.85
log(rate) ~ lwid + log(len)

             Df Sum of Sq    RSS     AIC
+ slim        1    2.3788  2.920 -93.088
+ acpt        1    1.9644  3.334 -87.912
+ shld        1    1.7999  3.499 -86.034
+ log(sigs1)  1    0.8690  4.430 -76.833
+ htype       3    1.1898  4.109 -75.765
+ log(trks)   1    0.7347  4.564 -75.669
+ log(adt)    1    0.3631  4.936 -72.616
<none>                     5.299 -71.847
+ lane        1    0.2525  5.046 -71.752
+ itg         1    0.2023  5.096 -71.366
- log(len)    1   27.4878 32.787  -2.768

Step:  AIC=-93.09
log(rate) ~ lwid + log(len) + slim

             Df Sum of Sq     RSS     AIC
+ acpt        1    0.2844  2.6355 -95.085
+ log(trks)   1    0.2710  2.6489 -94.887
<none>                     2.9200 -93.088
+ log(sigs1)  1    0.1412  2.7788 -93.020
+ shld        1    0.0556  2.8644 -91.837
+ htype       3    0.3244  2.5955 -91.681
+ log(adt)    1    0.0226  2.8974 -91.391
+ lane        1    0.0028  2.9172 -91.124
+ itg         1    0.0023  2.9176 -91.119
- slim        1    2.3788  5.2988 -71.847
- log(len)    1   20.3842 23.3041 -14.082

Step:  AIC=-95.08
log(rate) ~ lwid + log(len) + slim + acpt
```

```
              Df Sum of Sq     RSS     AIC
+ log(trks)    1    0.1793  2.4562 -95.833
<none>                      2.6355 -95.085
+ log(sigs1)   1    0.1158  2.5197 -94.838
+ shld         1    0.0555  2.5800 -93.915
+ lane         1    0.0135  2.6220 -93.285
+ log(adt)     1    0.0132  2.6223 -93.281
+ itg          1    0.0128  2.6228 -93.274
- acpt         1    0.2844  2.9200 -93.088
+ htype        3    0.2128  2.4227 -92.368
- slim         1    0.6988  3.3343 -87.912
- log(len)     1   18.7459 21.3814 -15.440

Step:  AIC=-95.83
log(rate) ~ lwid + log(len) + slim + acpt + log(trks)

              Df Sum of Sq     RSS     AIC
<none>                      2.4562 -95.833
+ shld         1    0.1035  2.3527 -95.512
- log(trks)    1    0.1793  2.6355 -95.085
- acpt         1    0.1927  2.6489 -94.887
+ log(sigs1)   1    0.0458  2.4104 -94.567
+ log(adt)     1    0.0280  2.4282 -94.280
+ htype        3    0.2545  2.2018 -94.098
+ itg          1    0.0101  2.4461 -93.993
+ lane         1    0.0024  2.4538 -93.871
- slim         1    0.6608  3.1171 -88.540
- log(len)     1   14.5184 16.9746 -22.442
```

```
summary(part.3.forward.mod)
```

```

Call:
lm(formula = log(rate) ~ lwid + log(len) + slim + acpt + log(trks),
    data = highway.df, offset = log(len))

Residuals:
     Min       1Q   Median       3Q      Max
-0.51518 -0.16169  0.03966  0.17333  0.55629

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.80121    1.46241   3.283  0.00243 **
lwid        -0.05235    0.10359  -0.505  0.61666
log(len)    -1.24832    0.08938 -13.966 2.08e-15 ***
slim        -0.03117    0.01046  -2.980  0.00538 **
acpt         0.01086    0.00675   1.609  0.11713
log(trks)   -0.33565    0.21626  -1.552  0.13018
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2728 on 33 degrees of freedom
Multiple R-squared:  0.6984,    Adjusted R-squared:  0.6527
F-statistic: 15.28 on 5 and 33 DF,  p-value: 8.762e-08
```

```
summary(part.3.backward.mod)
```

```
Call:
lm(formula = log(rate) ~ log(len) + log(adt) + slim + lwid +
    log(sigs1) + htype, data = highway.df, offset = log(len))

Residuals:
    Min      1Q  Median      3Q     Max
-0.4707 -0.1212 -0.0209  0.1019  0.4535

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.655848   1.327821   2.753  0.00992 **
log(len)    -1.246673   0.086899 -14.346  5.7e-15 ***
log(adt)    -0.136162   0.085296  -1.596  0.12089
slim        -0.029711   0.009787  -3.036  0.00492 **
lwid         0.056762   0.098322   0.577  0.56804
log(sigs1)   0.217015   0.060925   3.562  0.00125 **
htypefai     0.157704   0.345084   0.457  0.65096
htypepa     -0.362487   0.232778  -1.557  0.12991
htypema     -0.123393   0.206890  -0.596  0.55537
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2456 on 30 degrees of freedom
Multiple R-squared:  0.7777,    Adjusted R-squared:  0.7184
F-statistic: 13.12 on 8 and 30 DF,  p-value: 6.472e-08
```

```
summary(part.3.both.mod)
```

```
Call:
lm(formula = log(rate) ~ lwid + log(len) + slim + acpt + log(trks),
    data = highway.df, offset = log(len))

Residuals:
     Min       1Q   Median       3Q      Max
-0.51518 -0.16169  0.03966  0.17333  0.55629

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.80121    1.46241   3.283  0.00243 **
lwid        -0.05235    0.10359  -0.505  0.61666
log(len)    -1.24832    0.08938 -13.966 2.08e-15 ***
slim        -0.03117    0.01046  -2.980  0.00538 **
acpt         0.01086    0.00675   1.609  0.11713
log(trks)   -0.33565    0.21626  -1.552  0.13018
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2728 on 33 degrees of freedom
Multiple R-squared:  0.6984,    Adjusted R-squared:  0.6527
F-statistic: 15.28 on 5 and 33 DF,  p-value: 8.762e-08
```

The coefficient estimates are the same as in part 2 except for the one for `log(len)`. The models created using `offset` =log(len)‘ also have the same AIC values.

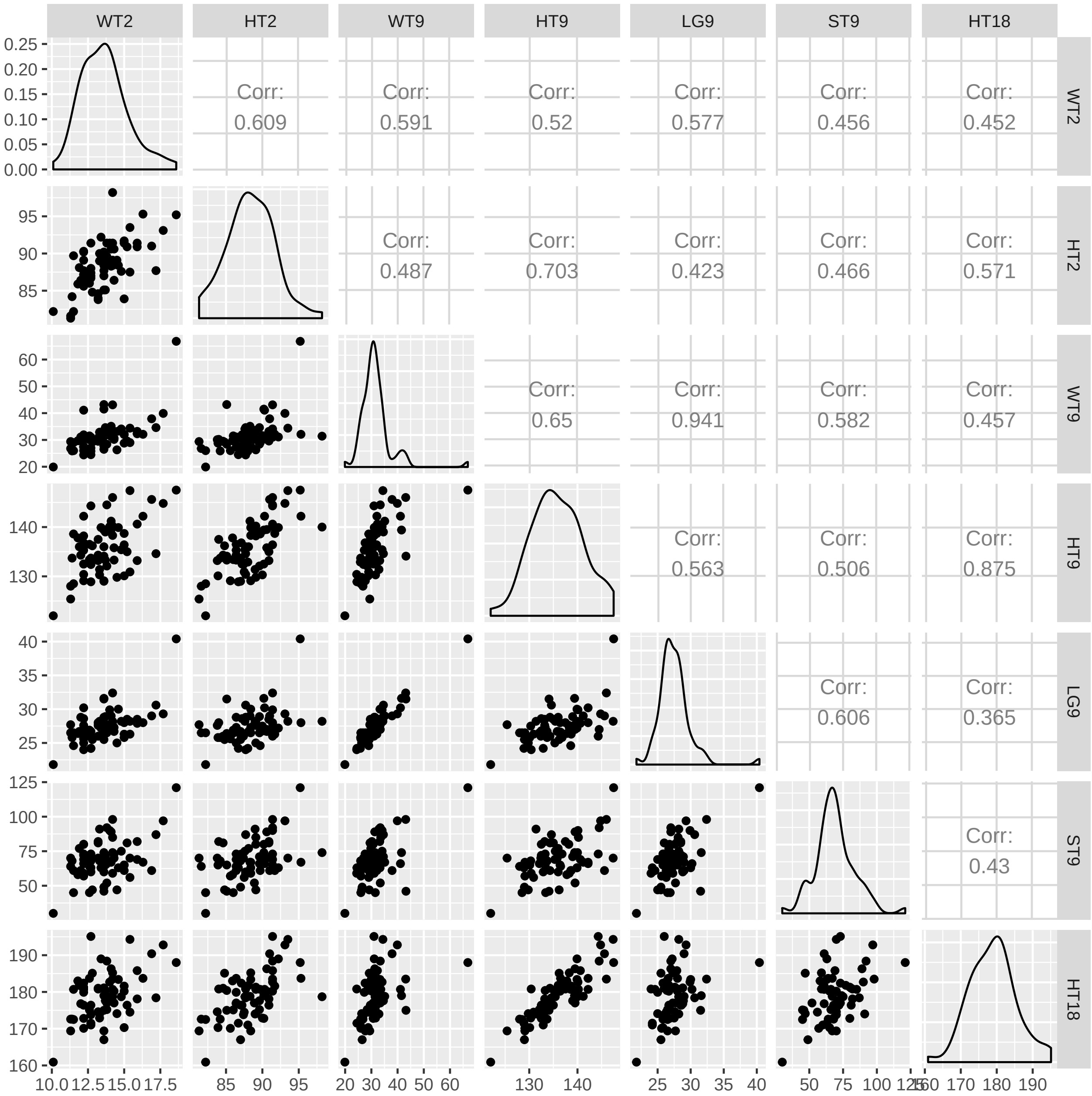
Problem 2

[From ALR 10.4]

```
bgsboys.df <- alr4::BGSboys %>%
  dp$select(WT2, HT2, WT9, HT9, LG9, ST9, HT18)
summary(bgsboys.df)
```

```
      WT2             HT2             WT9             HT9       
 Min.   :10.10   Min.   :81.30   Min.   :19.90   Min.   :122.0  
 1st Qu.:12.30   1st Qu.:86.40   1st Qu.:28.85   1st Qu.:132.5  
 Median :13.60   Median :88.35   Median :31.00   Median :135.6  
 Mean   :13.63   Mean   :88.37   Mean   :31.63   Mean   :135.9  
 3rd Qu.:14.30   3rd Qu.:90.60   3rd Qu.:33.27   3rd Qu.:139.5  
 Max.   :18.60   Max.   :98.20   Max.   :66.80   Max.   :147.5  
      LG9             ST9              HT18      
 Min.   :21.80   Min.   : 30.00   Min.   :160.9  
 1st Qu.:26.30   1st Qu.: 61.00   1st Qu.:174.5  
 Median :27.25   Median : 68.00   Median :178.9  
 Mean   :27.50   Mean   : 68.92   Mean   :179.0  
 3rd Qu.:28.43   3rd Qu.: 74.75   3rd Qu.:182.6  
 Max.   :40.40   Max.   :121.00   Max.   :195.1  
```

```
ggpairs(bgsboys.df)
```



Based on the plots of HT18 vs the predictors, there doesn't seem to be any reason to perform any transformations. It is worth noting that many of the predictors appear to be strongly correlated. This suggests that transformations are not necessary but there is strong reason to leave out some of the predictors.

We'll try both forward selection and backward elimination

```
bgsboys.forward.mod <- lm(HT18 ~ 1, data = bgsboys.df) %>%
  step(scope = ~ WT2 + HT2 + WT9 + HT9 + LG9 + ST9, direction = 'forward')
```

```
Start:  AIC=248.42
HT18 ~ 1

       Df Sum of Sq     RSS    AIC
+ HT9   1   2113.28  647.75 154.73
+ HT2   1    899.73 1861.30 224.40
+ WT9   1    576.22 2184.81 234.98
+ WT2   1    564.86 2196.17 235.32
+ ST9   1    510.51 2250.52 236.93
+ LG9   1    368.71 2392.32 240.96
<none>              2761.03 248.42

Step:  AIC=154.73
HT18 ~ HT9

       Df Sum of Sq    RSS    AIC
+ LG9   1    64.964 582.79 149.76
+ WT9   1    60.183 587.57 150.30
<none>              647.75 154.73
+ HT2   1    10.560 637.19 155.65
+ ST9   1     0.635 647.12 156.67
+ WT2   1     0.028 647.73 156.73

Step:  AIC=149.76
HT18 ~ HT9 + LG9

       Df Sum of Sq    RSS    AIC
<none>              582.79 149.76
+ WT2   1   11.2745 571.52 150.47
+ ST9   1   10.0847 572.71 150.61
+ HT2   1    8.2445 574.55 150.82
+ WT9   1    0.8820 581.91 151.66
```

```r
bgsboys.backward.mod <- lm(HT18 ~ ., data = bgsboys.df) %>%
  step(direction = 'backward')
```

```
Start:  AIC=152.2
HT18 ~ WT2 + HT2 + WT9 + HT9 + LG9 + ST9

       Df Sum of Sq     RSS    AIC
- WT9   1      0.65  536.35 150.28
- ST9   1     12.52  548.22 151.72
- LG9   1     15.24  550.94 152.05
<none>               535.70 152.20
- WT2   1     24.51  560.21 153.15
- HT2   1     26.80  562.50 153.42
- HT9   1   1083.06 1618.76 223.18

Step:  AIC=150.28
HT18 ~ WT2 + HT2 + HT9 + LG9 + ST9

       Df Sum of Sq     RSS    AIC
- ST9   1     13.09  549.44 149.87
<none>               536.35 150.28
- WT2   1     24.16  560.50 151.19
- HT2   1     26.98  563.32 151.52
- LG9   1     99.09  635.43 159.47
- HT9   1   1207.50 1743.85 226.10

Step:  AIC=149.87
HT18 ~ WT2 + HT2 + HT9 + LG9
```

```
       Df Sum of Sq     RSS    AIC
<none>              549.44 149.87
- HT2   1     22.08  571.52 150.47
- WT2   1     25.11  574.55 150.82
- LG9   1     86.23  635.67 157.49
- HT9   1   1242.74 1792.18 225.90
```

```
summary(bgsboys.forward.mod)
```

```

Call:
lm(formula = HT18 ~ HT9 + LG9, data = bgsboys.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.1632 -1.9599  0.4714  2.0057  6.6190 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 31.31920    9.63309   3.251  0.00185 ** 
HT9          1.18531    0.08475  13.986  < 2e-16 ***
LG9         -0.48762    0.18401  -2.650  0.01016 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.041 on 63 degrees of freedom
Multiple R-squared:  0.7889,	Adjusted R-squared:  0.7822 
F-statistic: 117.7 on 2 and 63 DF,  p-value: < 2.2e-16
```

```
summary(bgsboys.backward.mod)
```

```

Call:
lm(formula = HT18 ~ WT2 + HT2 + HT9 + LG9, data = bgsboys.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.3306 -1.5334  0.3825  1.7447  7.1090 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  42.4927    11.3368   3.748 0.000398 ***
WT2           0.5358     0.3209   1.670 0.100122    
HT2          -0.2717     0.1736  -1.566 0.122610    
HT9           1.2527     0.1067  11.746  < 2e-16 ***
LG9          -0.6194     0.2002  -3.094 0.002978 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.001 on 61 degrees of freedom
Multiple R-squared:  0.801,	Adjusted R-squared:  0.788 
F-statistic: 61.38 on 4 and 61 DF,  p-value: < 2.2e-16
```

```
anova(bgsboys.forward.mod, bgsboys.backward.mod)
```

```
Analysis of Variance Table

Model 1: HT18 ~ HT9 + LG9
Model 2: HT18 ~ WT2 + HT2 + HT9 + LG9
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     63 582.79
2     61 549.44  2    33.353 1.8515 0.1657
```

Forward selection results in a simpler model `HT18 ~ HT9 + LG9` which is a sub-model of the result of backward elimination (`HT18 ~ HT9 + LG9 + WT2 + HT2`). An ANOVA test reveals that there's no significant difference between the two, and the smaller model has a lower AIC. This suggests that the smaller model is a better fit (also suggests that the larger model is overfitting rather than the smaller model underfitting).

Problem 3

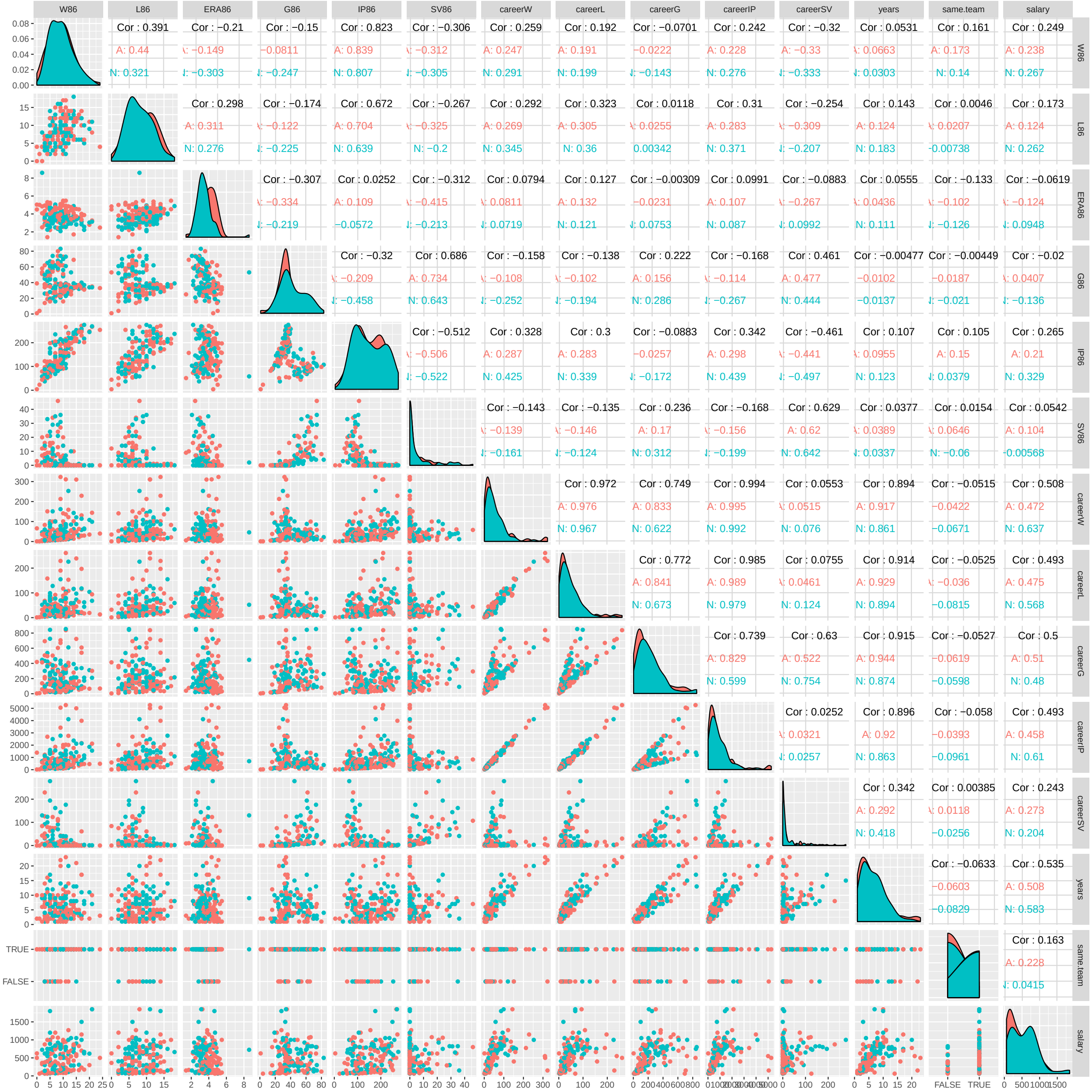
```
pitchers.df <- readr::read_tsv('~/dev/stats-hw/stat-s632/BaseballPitchers.txt') %>%
  # might be interesting
  dp$mutate(same.team = (team86 == team87)) %>%
  # also remove NA values
  na.omit()
summary(pitchers.df)
```

```
  firstName           lastName            team86
 Length:176         Length:176         Length:176
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character



   league86              W86              L86             ERA86
 Length:176         Min.   : 0.000   Min.   : 0.000   Min.   :1.410
 Class :character   1st Qu.: 6.000   1st Qu.: 5.000   1st Qu.:3.140
 Mode  :character   Median : 9.000   Median : 8.000   Median :3.735
                    Mean   : 9.205   Mean   : 8.312   Mean   :3.760
                    3rd Qu.:12.000   3rd Qu.:11.000   3rd Qu.:4.340
                    Max.   :24.000   Max.   :18.000   Max.   :8.590
      G86             IP86              SV86             years
 Min.   : 1.00   Min.   :  4.00   Min.   : 0.000   Min.   : 1.000
 1st Qu.:32.00   1st Qu.: 97.08   1st Qu.: 0.000   1st Qu.: 3.000
 Median :35.00   Median :144.00   Median : 0.000   Median : 5.000
 Mean   :40.27   Mean   :149.64   Mean   : 4.699   Mean   : 6.278
 3rd Qu.:51.00   3rd Qu.:203.28   3rd Qu.: 6.000   3rd Qu.: 9.000
 Max.   :83.00   Max.   :275.10   Max.   :46.000   Max.   :23.000
    careerW          careerL          careerERA        careerG
 Min.   :  1.00   Min.   :  1.00   Min.   :2.230   Min.   :  4.0
 1st Qu.: 16.00   1st Qu.: 14.00   1st Qu.:3.290   1st Qu.: 80.0
 Median : 36.00   Median : 33.00   Median :3.650   Median :155.5
 Mean   : 52.39   Mean   : 46.88   Mean   :3.669   Mean   :212.5
 3rd Qu.: 62.25   3rd Qu.: 62.00   3rd Qu.:4.030   3rd Qu.:295.0
 Max.   :323.00   Max.   :261.00   Max.   :5.480   Max.   :853.0
    careerIP         careerSV          salary          league87
 Min.   :  19.0   Min.   :  0.00   Min.   :  62.5   Length:176
 1st Qu.: 264.1   1st Qu.:  0.00   1st Qu.: 158.8   Class :character
 Median : 563.7   Median :  3.00   Median : 417.5   Mode  :character
 Mean   : 865.1   Mean   : 22.18   Mean   : 497.5                    
 3rd Qu.:1101.8   3rd Qu.: 17.00   3rd Qu.: 756.2                    
 Max.   :5264.2   Max.   :278.00   Max.   :1850.0                    
    team87          same.team      
 Length:176         Mode :logical  
 Class :character   FALSE:19       
 Mode  :character   TRUE :157      
```

```r
ggpairs(pitchers.df,
        columns = c('W86', 'L86', 'ERA86', 'G86', 'IP86', 'SV86',
                    'careerW', 'careerL', 'careerG', 'careerIP', 'careerSV',
                    'years', 'same.team', 'salary'),
        mapping = aes(colour = league86))
```



Part a

We'll try both forward selection and backward elimination. Name and team won't be considered, although we'll see if the fact that they switched teams had any effect.

```
full.form <- ~ W86 + L86 + ERA86 + G86 + IP86 + SV86 +
  careerW + careerL + careerG + careerIP + careerSV + same.team + league86

pitchers.forward.mod <- lm(salary ~ 1, data = pitchers.df) %>%
  step(scope = full.form, direction = 'both')
```

```
Start:  AIC=2084.42
salary ~ 1
```

```
            Df Sum of Sq      RSS    AIC
+ careerW    1   6258880 17954995 2033.8
+ careerG    1   6044133 18169743 2035.9
+ careerL    1   5887923 18325953 2037.4
+ careerIP   1   5883709 18330166 2037.4
+ IP86       1   1705607 22508268 2073.6
+ W86        1   1495645 22718230 2075.2
+ careerSV   1   1434419 22779456 2075.7
+ L86        1    726273 23487602 2081.1
+ same.team  1    642263 23571613 2081.7
+ league86   1    583827 23630048 2082.1
<none>                   24213875 2084.4
+ ERA86      1     92661 24121214 2085.8
+ SV86       1     71015 24142860 2085.9
+ G86        1      9729 24204146 2086.3

Step:  AIC=2033.79
salary ~ careerW

            Df Sum of Sq      RSS    AIC
+ careerSV   1   1125696 16829299 2024.4
+ same.team  1    867631 17087365 2027.1
+ careerG    1    779601 17175395 2028.0
+ league86   1    678731 17276264 2029.0
+ SV86       1    397196 17557800 2031.8
+ W86        1    355233 17599762 2032.3
+ careerIP   1    307085 17647911 2032.8
+ IP86       1    263079 17691917 2033.2
+ ERA86      1    254768 17700227 2033.3
<none>                   17954995 2033.8
+ G86        1     90974 17864022 2034.9
+ L86        1     16363 17938632 2035.6
+ careerL    1       685 17954310 2035.8
- careerW    1   6258880 24213875 2084.4

Step:  AIC=2024.39
salary ~ careerW + careerSV

            Df Sum of Sq      RSS    AIC
+ IP86       1   1493467 15335832 2010.0
+ W86        1   1057146 15772153 2015.0
+ same.team  1    854440 15974859 2017.2
+ league86   1    519807 16309492 2020.9
+ L86        1    198869 16630430 2024.3
<none>                   16829299 2024.4
+ ERA86      1    166191 16663108 2024.7
+ careerIP   1     76674 16752625 2025.6
+ G86        1     53943 16775356 2025.8
+ careerG    1     17940 16811359 2026.2
+ careerL    1     15788 16813511 2026.2
+ SV86       1      4852 16824447 2026.3
- careerSV   1   1125696 17954995 2033.8
- careerW    1   5950157 22779456 2075.7
```

```
Step:  AIC=2010.04
salary ~ careerW + careerSV + IP86

            Df Sum of Sq      RSS    AIC
+ same.team  1    554391 14781442 2005.6
+ league86   1    379844 14955988 2007.6
<none>                   15335832 2010.0
+ L86        1    127387 15208446 2010.6
+ ERA86      1    115128 15220704 2010.7
+ careerIP   1     87634 15248199 2011.0
+ SV86       1     64804 15271028 2011.3
+ G86        1     25354 15310478 2011.8
+ careerG    1     14819 15321014 2011.9
+ W86        1     12913 15322919 2011.9
+ careerL    1      5035 15330797 2012.0
- IP86       1   1493467 16829299 2024.4
- careerSV   1   2356084 17691917 2033.2
- careerW    1   3055602 18391435 2040.0

Step:  AIC=2005.56
salary ~ careerW + careerSV + IP86 + same.team

            Df Sum of Sq      RSS    AIC
+ league86   1    309249 14472192 2003.8
<none>                   14781442 2005.6
+ L86        1     85868 14695574 2006.5
+ ERA86      1     62503 14718939 2006.8
+ careerIP   1     61179 14720263 2006.8
+ SV86       1     48332 14733109 2007.0
+ G86        1     23244 14758198 2007.3
+ careerG    1     20548 14760894 2007.3
+ careerL    1      4598 14776843 2007.5
+ W86        1       585 14780857 2007.5
- same.team  1    554391 15335832 2010.0
- IP86       1   1193418 15974859 2017.2
- careerSV   1   2149448 16930890 2027.5
- careerW    1   3309698 18091140 2039.1

Step:  AIC=2003.84
salary ~ careerW + careerSV + IP86 + same.team + league86

            Df Sum of Sq      RSS    AIC
<none>                   14472192 2003.8
+ careerIP   1     79596 14392596 2004.9
+ L86        1     58974 14413219 2005.1
+ SV86       1     54021 14418172 2005.2
+ G86        1     49335 14422857 2005.2
- league86   1    309249 14781442 2005.6
+ ERA86      1     11138 14461054 2005.7
+ careerL    1      4270 14467923 2005.8
+ careerG    1      3499 14468693 2005.8
+ W86        1      1626 14470567 2005.8
- same.team  1    483796 14955988 2007.6
- IP86       1   1098873 15571065 2014.7
- careerSV    1   1920877 16393069 2023.8
- careerW     1   3408017 17880210 2039.0
```

```
pitchers.backward.mod <- lm(salary ~ . - firstName - lastName - team86 - team87,
                            data = pitchers.df) %>%
  step(direction = 'backward')
```

```
Start:  AIC=2005.1
salary ~ (firstName + lastName + team86 + league86 + W86 + L86 + 
    ERA86 + G86 + IP86 + SV86 + years + careerW + careerL + careerERA + 
    careerG + careerIP + careerSV + league87 + team87 + same.team) - 
    firstName - lastName - team86 - team87

            Df Sum of Sq      RSS    AIC
- careerG    1      9161 12872978 2003.2
- careerSV   1     21902 12885718 2003.4
- league87   1     25917 12889733 2003.5
- W86        1     42063 12905879 2003.7
- L86        1     80165 12943981 2004.2
- league86   1     81375 12945191 2004.2
- SV86       1    138479 13002295 2005.0
- ERA86      1    145769 13009585 2005.1
<none>                   12863816 2005.1
- G86        1    147562 13011378 2005.1
- careerL    1    175717 13039533 2005.5
- careerW    1    282318 13146134 2006.9
- same.team  1    334045 13197861 2007.6
- years      1    349559 13213375 2007.8
- careerIP   1    365030 13228846 2008.0
- IP86       1    641915 13505731 2011.7
- careerERA  1    755473 13619289 2013.2

Step:  AIC=2003.23
salary ~ league86 + W86 + L86 + ERA86 + G86 + IP86 + SV86 + years + 
    careerW + careerL + careerERA + careerIP + careerSV + league87 + 
    same.team

            Df Sum of Sq      RSS    AIC
- careerSV   1     13844 12886821 2001.4
- league87   1     27249 12900227 2001.6
- W86        1     40323 12913301 2001.8
- L86        1     76570 12949547 2002.3
- league86   1     84087 12957065 2002.4
<none>                   12872978 2003.2
- ERA86      1    152633 13025610 2003.3
- careerL    1    166697 13039675 2003.5
- SV86       1    214472 13087449 2004.1
- G86        1    253759 13126736 2004.7
- careerW    1    273624 13146602 2004.9
- same.team  1    334400 13207377 2005.7
- careerIP   1    357094 13230072 2006.0
- years      1    483179 13356157 2007.7
- IP86       1    653677 13526654 2010.0
- careerERA  1    750041 13623018 2011.2
```

```
Step:  AIC=2001.42
salary ~ league86 + W86 + L86 + ERA86 + G86 + IP86 + SV86 + years +
    careerW + careerL + careerERA + careerIP + league87 + same.team

            Df Sum of Sq      RSS    AIC
- league87   1     28426 12915248 1999.8
- W86        1     42408 12929230 2000.0
- L86        1     78012 12964833 2000.5
- league86   1     85115 12971936 2000.6
<none>                   12886821 2001.4
- ERA86      1    185548 13072370 2001.9
- careerL    1    187098 13073920 2002.0
- G86        1    249306 13136128 2002.8
- SV86       1    282228 13169050 2003.2
- careerW    1    330103 13216924 2003.9
- same.team  1    337713 13224534 2004.0
- careerIP   1    474417 13361239 2005.8
- IP86       1    658029 13544851 2008.2
- careerERA  1    874033 13760855 2011.0
- years      1    882683 13769505 2011.1

Step:  AIC=1999.8
salary ~ league86 + W86 + L86 + ERA86 + G86 + IP86 + SV86 + years +
    careerW + careerL + careerERA + careerIP + same.team

            Df Sum of Sq      RSS    AIC
- W86        1     45598 12960846 1998.4
- league86   1     70566 12985814 1998.8
- L86        1     87198 13002445 1999.0
<none>                   12915248 1999.8
- careerL    1    178435 13093683 2000.2
- ERA86      1    194713 13109961 2000.4
- G86        1    253912 13169160 2001.2
- careerW    1    311207 13226455 2002.0
- SV86       1    320391 13235639 2002.1
- same.team  1    364739 13279987 2002.7
- careerIP   1    451413 13366660 2003.8
- IP86       1    682231 13597478 2006.9
- careerERA  1    860305 13775552 2009.2
- years      1    872534 13787782 2009.3

Step:  AIC=1998.42
salary ~ league86 + L86 + ERA86 + G86 + IP86 + SV86 + years +
    careerW + careerL + careerERA + careerIP + same.team

            Df Sum of Sq      RSS    AIC
- L86        1     55272 13016118 1997.2
- league86   1     83024 13043869 1997.5
<none>                   12960846 1998.4
- careerL    1    152847 13113693 1998.5
- ERA86      1    241824 13202669 1999.7
- careerW    1    266852 13227697 2000.0
- G86        1    286392 13247238 2000.3
- SV86       1    308708 13269553 2000.6
- same.team  1    353884 13314730 2001.2
- careerIP   1    408786 13369632 2001.9
- careerERA  1    862900 13823746 2007.8
- years      1    930360 13891205 2008.6
- IP86       1   1222636 14183482 2012.3

Step:  AIC=1997.17
salary ~ league86 + ERA86 + G86 + IP86 + SV86 + years + careerW + 
    careerL + careerERA + careerIP + same.team

            Df Sum of Sq      RSS    AIC
- league86   1     75004 13091122 1996.2
- careerL    1    114028 13130146 1996.7
<none>                   13016118 1997.2
- ERA86      1    190601 13206719 1997.7
- careerW    1    256649 13272767 1998.6
- SV86       1    266512 13282630 1998.7
- G86        1    293908 13310026 1999.1
- careerIP   1    375046 13391164 2000.2
- same.team  1    378127 13394245 2000.2
- careerERA  1    875688 13891806 2006.6
- years      1   1049785 14065903 2008.8
- IP86       1   1642995 14659113 2016.1

Step:  AIC=1996.19
salary ~ ERA86 + G86 + IP86 + SV86 + years + careerW + careerL + 
    careerERA + careerIP + same.team

            Df Sum of Sq      RSS    AIC
- careerL    1    115572 13206694 1995.7
<none>                   13091122 1996.2
- ERA86      1    168073 13259194 1996.4
- careerW    1    225964 13317086 1997.2
- SV86       1    228456 13319578 1997.2
- G86        1    276806 13367928 1997.9
- careerIP   1    359104 13450226 1999.0
- same.team  1    400903 13492025 1999.5
- careerERA  1   1136836 14227958 2008.8
- years      1   1173336 14264458 2009.3
- IP86       1   1685330 14776452 2015.5

Step:  AIC=1995.73
salary ~ ERA86 + G86 + IP86 + SV86 + years + careerW + careerERA + 
    careerIP + same.team

            Df Sum of Sq      RSS    AIC
- careerW    1    142116 13348809 1995.6
<none>                   13206694 1995.7
- ERA86      1    197470 13404163 1996.3
- G86        1    253355 13460048 1997.1
- careerIP   1    258594 13465288 1997.1
- SV86       1    268856 13475549 1997.3
- same.team  1    456950 13663643 1999.7
- careerERA  1   1037312 14244005 2007.0
- IP86       1   1646309 14853003 2014.4
- years      1   1663914 14870607 2014.6

Step:  AIC=1995.62
salary ~ ERA86 + G86 + IP86 + SV86 + years + careerERA + careerIP +
    same.team

            Df Sum of Sq      RSS    AIC
<none>                   13348809 1995.6
- ERA86      1    204432 13553241 1996.3
- careerIP   1    271908 13620717 1997.2
- G86        1    325940 13674750 1997.9
- SV86       1    345036 13693845 1998.1
- same.team  1    474136 13822946 1999.8
- careerERA  1   1347366 14696176 2010.5
- IP86       1   1627833 14976643 2013.9
- years      1   1634678 14983487 2014.0
```

```r
summary(pitchers.forward.mod)
```

```

Call:
lm(formula = salary ~ careerW + careerSV + IP86 + same.team +
    league86, data = pitchers.df)

Residuals:
   Min     1Q Median     3Q    Max
-877.8 -194.5  -35.9  145.0 1148.6

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -133.9275    89.4704  -1.497 0.136276
careerW          2.7321     0.4318   6.327 2.14e-09 ***
careerSV         2.7672     0.5825   4.750 4.30e-06 ***
IP86             1.5701     0.4370   3.593 0.000428 ***
same.teamTRUE  171.8330    72.0805   2.384 0.018232 *
league86N       85.2596    44.7334   1.906 0.058344 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 291.8 on 170 degrees of freedom
Multiple R-squared:  0.4023,    Adjusted R-squared:  0.3847
F-statistic: 22.89 on 5 and 170 DF,  p-value: < 2.2e-16
```

```r
summary(pitchers.backward.mod)
```

```

Call:
lm(formula = salary ~ ERA86 + G86 + IP86 + SV86 + years + careerERA +
    careerIP + same.team, data = pitchers.df)

Residuals:
    Min      1Q  Median      3Q     Max
-790.32 -180.46   -1.19  154.36 1065.27
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    412.14444  214.65421   1.920   0.0566 .  
ERA86           47.73317   29.84755   1.599   0.1117    
G86             -3.92809    1.94525  -2.019   0.0451 *  
IP86             2.05644    0.45569   4.513 1.20e-05 ***
SV86             8.06380    3.88124   2.078   0.0393 *  
years           56.75477   12.55016   4.522 1.16e-05 ***
careerERA     -186.45547   45.41458  -4.106 6.30e-05 ***
careerIP        -0.12274    0.06655  -1.844   0.0669 .  
same.teamTRUE  171.00970   70.21536   2.436   0.0159 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 282.7 on 167 degrees of freedom
Multiple R-squared:  0.4487,    Adjusted R-squared:  0.4223 
F-statistic: 16.99 on 8 and 167 DF,  p-value: < 2.2e-16
```

```r
anova(pitchers.forward.mod, pitchers.backward.mod)
```

```
Analysis of Variance Table

Model 1: salary ~ careerW + careerSV + IP86 + same.team + league86
Model 2: salary ~ ERA86 + G86 + IP86 + SV86 + years + careerERA + careerIP + 
    same.team
  Res.Df      RSS Df Sum of Sq      F   Pr(>F)   
1    170 14472192                                
2    167 13348809  3   1123383 4.6847 0.003613 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Although the first method could use both, in this case only forward selection was used.

As usual, forward selection results in a smaller model. This time, backward elimination results in a smaller AIC, and an ANOVA test suggests that there is a significant difference between the two models. This suggests that the model found by backward elimination should be used.

Part b

Fit the models on the training set:

```r
# split the data
set.seed(632)
prop.train <- 2 / 3
train.df <- dp$sample_frac(pitchers.df, prop.train)
test.df <- dp$setdiff(pitchers.df, train.df)

# fit models using training data
pitchers.forward.mod <- lm(salary ~ 1, data = train.df) %>% 
  step(scope = full.form, direction = 'both')
```

```
Start:  AIC=1388.26
salary ~ 1

            Df Sum of Sq      RSS    AIC
+ careerW    1   3583581 12779642 1361.3
+ careerIP   1   3379524 12983698 1363.2
+ careerL    1   3198402 13164821 1364.8
+ careerG    1   2962154 13401069 1366.9
+ W86        1   2162511 14200711 1373.7
+ IP86       1   2104329 14258893 1374.2
+ same.team  1    994752 15368470 1382.9
+ league86   1    959439 15403784 1383.2
+ ERA86      1    665779 15697443 1385.4
+ careerSV   1    587285 15775937 1386.0
+ L86        1    466091 15897131 1386.9
<none>                  16363223 1388.3
+ SV86       1      8671 16354552 1390.2
+ G86        1      7933 16355290 1390.2

Step:  AIC=1361.34
salary ~ careerW

            Df Sum of Sq      RSS    AIC
+ same.team  1   1202176 11577466 1351.8
+ league86   1   1156367 11623275 1352.2
+ W86        1    978521 11801121 1354.0
+ ERA86      1    941284 11838357 1354.4
+ IP86       1    715582 12064060 1356.6
+ careerSV   1    429021 12350620 1359.3
<none>                  12779642 1361.3
+ careerG    1    187657 12591984 1361.6
+ careerIP   1    185102 12594539 1361.6
+ SV86       1    146498 12633144 1362.0
+ careerL    1     59814 12719828 1362.8
+ G86        1     43135 12736507 1362.9
+ L86        1       708 12778933 1363.3
- careerW    1   3583581 16363223 1388.3

Step:  AIC=1351.78
salary ~ careerW + same.team

            Df Sum of Sq      RSS    AIC
+ league86   1    816998 10760468 1345.2
+ W86        1    616006 10961460 1347.4
+ ERA86      1    546554 11030913 1348.1
+ IP86       1    440956 11136510 1349.2
+ careerSV   1    373181 11204286 1350.0
<none>                  11577466 1351.8
+ careerG    1    188055 11389411 1351.9
+ careerIP   1    152948 11424518 1352.2
+ SV86       1    109791 11467675 1352.7
+ G86        1     60240 11517226 1353.2
+ careerL    1     34612 11542854 1353.4
+ L86        1       389 11577077 1353.8
- same.team  1   1202176 12779642 1361.3
- careerW    1   3791004 15368470 1382.9

Step:  AIC=1345.22
salary ~ careerW + same.team + league86
```

```
            Df Sum of Sq      RSS    AIC
+ W86        1    616646 10143822 1340.3
+ IP86       1    410034 10350434 1342.7
+ careerSV   1    265241 10495228 1344.3
+ careerIP   1    200411 10560057 1345.0
+ ERA86      1    189845 10570624 1345.1
<none>                   10760468 1345.2
+ SV86       1     98390 10662078 1346.1
+ careerG    1     71829 10688639 1346.4
+ careerL    1     70878 10689590 1346.5
+ G86        1      7664 10752805 1347.1
+ L86        1         5 10760463 1347.2
- league86   1    816998 11577466 1351.8
- same.team  1    862807 11623275 1352.2
- careerW    1   3932819 14693287 1379.7

Step:  AIC=1340.31
salary ~ careerW + same.team + league86 + W86

            Df Sum of Sq      RSS    AIC
+ careerSV   1    667274  9476547 1334.3
+ careerG    1    406370  9737452 1337.5
+ SV86       1    323642  9820179 1338.5
<none>                   10143822 1340.3
+ careerIP   1     77031 10066790 1341.4
+ L86        1     72359 10071463 1341.5
+ ERA86      1     54136 10089686 1341.7
+ G86        1     22177 10121644 1342.1
+ careerL    1        95 10143727 1342.3
+ IP86       1        30 10143792 1342.3
- same.team  1    569589 10713410 1344.7
- W86        1    616646 10760468 1345.2
- league86   1    817638 10961460 1347.4
- careerW    1   2797274 12941096 1366.8

Step:  AIC=1334.35
salary ~ careerW + same.team + league86 + W86 + careerSV

            Df Sum of Sq      RSS    AIC
<none>                    9476547 1334.3
+ IP86       1    138832  9337716 1334.6
+ G86        1     56281  9420266 1335.7
+ careerG    1     18560  9457988 1336.1
+ L86        1     15424  9461124 1336.2
+ SV86       1     13765  9462783 1336.2
+ careerL    1      1208  9475340 1336.3
+ careerIP   1       374  9476173 1336.3
+ ERA86      1       131  9476416 1336.3
- same.team  1    458709  9935257 1337.9
- league86   1    645955 10122502 1340.1
- careerSV   1    667274 10143822 1340.3
- W86        1   1018680 10495228 1344.3
- careerW    1   2294804 11771352 1357.7
```

```r
pitchers.backward.mod <- lm(salary ~ . - firstName - lastName - team86 - team87,
                            data = train.df) %>%
  step(direction = 'backward')
```

```
Start:  AIC=1341.22
salary ~ (firstName + lastName + team86 + league86 + W86 + L86 +
    ERA86 + G86 + IP86 + SV86 + years + careerW + careerL + careerERA +
    careerG + careerIP + careerSV + league87 + team87 + same.team) -
    firstName - lastName - team86 - team87

            Df Sum of Sq     RSS    AIC
- W86        1       508 8327668 1339.2
- league87   1      1300 8328461 1339.2
- G86        1     30102 8357263 1339.7
- careerL    1     53180 8380341 1340.0
- ERA86      1     58588 8385749 1340.0
- SV86       1     77586 8404746 1340.3
- careerG    1     85254 8412414 1340.4
- careerW    1     92414 8419575 1340.5
- careerSV   1     98030 8425191 1340.6
- careerIP   1    101033 8428194 1340.6
- league86   1    108940 8436101 1340.8
<none>                   8327161 1341.2
- L86        1    166918 8494079 1341.5
- same.team  1    222876 8550036 1342.3
- years      1    274892 8602052 1343.0
- careerERA  1    301783 8628944 1343.4
- IP86       1    410420 8737580 1344.8

Step:  AIC=1339.23
salary ~ league86 + L86 + ERA86 + G86 + IP86 + SV86 + years +
    careerW + careerL + careerERA + careerG + careerIP + careerSV +
    league87 + same.team

            Df Sum of Sq     RSS    AIC
- league87   1      1271 8328940 1337.2
- G86        1     29605 8357274 1337.7
- careerL    1     57954 8385623 1338.0
- ERA86      1     58163 8385831 1338.0
- SV86       1     77617 8405286 1338.3
- careerG    1     86657 8414325 1338.4
- careerSV   1     99055 8426723 1338.6
- league86   1    108446 8436114 1338.8
- careerW    1    119114 8446782 1338.9
- careerIP   1    123685 8451353 1339.0
<none>                   8327668 1339.2
- L86        1    200709 8528377 1340.0
- same.team  1    224753 8552421 1340.3
- years      1    274393 8602061 1341.0
- careerERA  1    301721 8629389 1341.4
- IP86       1   1196439 9524107 1352.9

Step:  AIC=1337.25
salary ~ league86 + L86 + ERA86 + G86 + IP86 + SV86 + years +
    careerW + careerL + careerERA + careerG + careerIP + careerSV +
    same.team

            Df Sum of Sq     RSS    AIC
- G86        1     29418 8358358 1335.7
- careerL    1     57062 8386001 1336.0
- ERA86      1     57902 8386842 1336.1
- SV86       1     77796 8406735 1336.3
- careerG    1     88449 8417388 1336.5
- careerSV   1    100611 8429550 1336.7
- careerW    1    117952 8446891 1336.9
- careerIP   1    122433 8451373 1337.0
<none>                   8328940 1337.2
- L86        1    200961 8529900 1338.0
- same.team  1    262550 8591490 1338.9
- years      1    274177 8603116 1339.0
- careerERA  1    301638 8630577 1339.4
- league86   1    345544 8674484 1340.0
- IP86       1   1196625 9525565 1351.0

Step:  AIC=1335.66
salary ~ league86 + L86 + ERA86 + IP86 + SV86 + years + careerW +
    careerL + careerERA + careerG + careerIP + careerSV + same.team

            Df Sum of Sq     RSS    AIC
- SV86       1     48420 8406778 1334.3
- careerL    1     69093 8427451 1334.6
- ERA86      1     85043 8443401 1334.8
<none>                   8358358 1335.7
- careerIP   1    153247 8511605 1335.8
- careerW    1    171342 8529699 1336.0
- careerSV   1    179508 8537866 1336.2
- careerG    1    202440 8560798 1336.5
- L86        1    253017 8611375 1337.2
- careerERA  1    282277 8640635 1337.5
- same.team  1    317951 8676308 1338.0
- years      1    351866 8710224 1338.5
- league86   1    359761 8718119 1338.6
- IP86       1   1180662 9539020 1349.1

Step:  AIC=1334.34
salary ~ league86 + L86 + ERA86 + IP86 + years + careerW + careerL +
    careerERA + careerG + careerIP + careerSV + same.team

            Df Sum of Sq     RSS    AIC
- ERA86      1     52500 8459278 1333.1
- careerL    1     81304 8488082 1333.5
<none>                   8406778 1334.3
- careerIP   1    165212 8571990 1334.6
- careerW    1    183127 8589905 1334.9
- L86        1    216689 8623467 1335.3
- careerG    1    285912 8692690 1336.2
- careerERA  1    308065 8714843 1336.5
- same.team  1    325885 8732663 1336.8
- league86   1    326522 8733300 1336.8
- careerSV   1    331544 8738322 1336.9
- years      1    391498 8798276 1337.7
- IP86       1   1211401 9618179 1348.1

Step:  AIC=1333.07
salary ~ league86 + L86 + IP86 + years + careerW + careerL + 
    careerERA + careerG + careerIP + careerSV + same.team

            Df Sum of Sq     RSS    AIC
- careerL    1     65675 8524953 1332.0
- careerIP   1    139901 8599179 1333.0
<none>                   8459278 1333.1
- careerW    1    166051 8625329 1333.3
- L86        1    176646 8635924 1333.5
- careerERA  1    262901 8722179 1334.7
- league86   1    289051 8748328 1335.0
- same.team  1    309463 8768741 1335.3
- careerG    1    316017 8775294 1335.4
- careerSV   1    352739 8812017 1335.8
- years      1    413291 8872569 1336.7
- IP86       1   1173510 9632787 1346.3

Step:  AIC=1331.97
salary ~ league86 + L86 + IP86 + years + careerW + careerERA + 
    careerG + careerIP + careerSV + same.team

            Df Sum of Sq     RSS    AIC
- careerIP   1     75057 8600010 1331.0
- careerW    1    108041 8632994 1331.4
- L86        1    129891 8654844 1331.7
<none>                   8524953 1332.0
- careerERA  1    216195 8741148 1332.9
- careerG    1    265342 8790294 1333.6
- league86   1    302831 8827784 1334.1
- same.team  1    323348 8848301 1334.3
- careerSV   1    323996 8848949 1334.3
- years      1    463918 8988871 1336.2
- IP86       1   1108444 9633397 1344.3

Step:  AIC=1331
salary ~ league86 + L86 + IP86 + years + careerW + careerERA + 
    careerG + careerSV + same.team

            Df Sum of Sq     RSS    AIC
- careerW    1     48038 8648047 1329.7
<none>                   8600010 1331.0
- L86        1    152828 8752838 1331.1
- careerERA  1    256681 8856690 1332.4
- careerG    1    271199 8871208 1332.6
- league86   1    280152 8880161 1332.8
- same.team  1    314471 8914481 1333.2
- years      1    392664 8992674 1334.2
- careerSV   1    461455 9061464 1335.1
- IP86       1   1141748 9741758 1343.6

Step:  AIC=1329.65
salary ~ league86 + L86 + IP86 + years + careerERA + careerG + 
    careerSV + same.team

            Df Sum of Sq     RSS    AIC
- L86        1    127883 8775930 1329.4
<none>                   8648047 1329.7
- careerG    1    232433 8880480 1330.8
- league86   1    236792 8884839 1330.8
- same.team  1    306338 8954385 1331.7
- careerERA  1    384667 9032714 1332.7
- careerSV   1    435662 9083709 1333.4
- years      1    676911 9324958 1336.5
- IP86       1   1299845 9947892 1344.0

Step:  AIC=1329.37
salary ~ league86 + IP86 + years + careerERA + careerG + careerSV + 
    same.team

            Df Sum of Sq      RSS    AIC
<none>                    8775930 1329.4
- careerG    1    225933  9001864 1330.3
- league86   1    231349  9007280 1330.4
- same.team  1    324277  9100207 1331.6
- careerSV   1    403722  9179652 1332.6
- careerERA  1    539437  9315367 1334.3
- years      1    641991  9417921 1335.6
- IP86       1   1356476 10132407 1344.2
```

```r
summary(pitchers.forward.mod)
```

```

Call:
lm(formula = salary ~ careerW + same.team + league86 + W86 + 
    careerSV, data = train.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-681.43 -199.02  -12.99  152.91  974.47 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -146.5210    99.4896  -1.473 0.143655    
careerW          2.4261     0.4679   5.185 9.82e-07 ***
same.teamTRUE  223.2705    96.3222   2.318 0.022284 *  
league86N      152.9066    55.5890   2.751 0.006947 ** 
W86             21.4011     6.1956   3.454 0.000782 ***
careerSV         1.7144     0.6132   2.796 0.006104 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 292.2 on 111 degrees of freedom
Multiple R-squared:  0.4209,    Adjusted R-squared:  0.3948 
```

```
F-statistic: 16.13 on 5 and 111 DF,  p-value: 6.218e-12
```

```
summary(pitchers.backward.mod)
```

```

Call:
lm(formula = salary ~ league86 + IP86 + years + careerERA + careerG + 
    careerSV + same.team, data = train.df)

Residuals:
   Min     1Q Median     3Q    Max 
-624.7 -200.0  -17.9  133.3 1062.7 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    211.7107   230.5756   0.918  0.36055    
league86N       96.0982    56.6911   1.695  0.09291 .  
IP86             2.0203     0.4922   4.105 7.84e-05 ***
years           60.7033    21.4972   2.824  0.00564 ** 
careerERA     -122.8204    47.4497  -2.588  0.01096 *  
careerG         -1.0621     0.6340  -1.675  0.09677 .  
careerSV         2.4769     1.1061   2.239  0.02717 *  
same.teamTRUE  189.5825    94.4656   2.007  0.04724 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 283.7 on 109 degrees of freedom
Multiple R-squared:  0.4637,    Adjusted R-squared:  0.4292 
F-statistic: 13.46 on 7 and 109 DF,  p-value: 1.975e-12
```

```
AIC(pitchers.forward.mod)
```

```
[1] 1668.384
```

```
AIC(pitchers.backward.mod)
```

```
[1] 1663.397
```

Compute errors on test set and find AICs

```
forward.k <- length(pitchers.forward.mod$coefficients)
backward.k <- length(pitchers.backward.mod$coefficients)
n <- nrow(test.df)

sse.forward <- (predict(pitchers.forward.mod,
                        newdata = test.df) - test.df$salary) ** 2 %>%
  sum()
sse.backward <- (predict(pitchers.backward.mod,
                         newdata = test.df) - test.df$salary) ** 2 %>%
  sum()

aic.forward <- n * log(sse.forward / n) + 2 * forward.k
aic.backward <- n * log(sse.backward / n) + 2 * backward.k

aic.forward
```

```
[1] 692.9141
```

```
aic.backward
```

```
[1] 691.0268
```

On the test set, the AIC using the model found via backward elimination is slightly smaller than the one found using forward selection. This is consistent with the findings from the training sets, suggesting we're not overfitting by using the larger backward elimination model. Although I am not an expert on baseball, the regressors included seem to make sense, as most of them are performance metrics. The two factor variables show that the national league pays better than the American league, and players that change teams tend to make less (suggesting that underperforming players are shuffled around).