Multiple Regression, Variable Selection

Case Study 12-02 from Sleuth 3: Sex Discrimination in Employment

Here's the description from the book: "Data on employees from one job category (skilled, entry–level clerical) of a bank that was sued for sex discrimination. The data are on 32 male and 61 female employees, hired between 1965 and 1975."

We have the following variables:

- Bsal: Annual salary at time of hire
- Sex: Sex of employee
- Senior: Seniority (months since first hired)
- Age: Age of employee (in months)
- Educ: Education (in years)
- Exper: Work experience prior to employment with the bank (months)

One of the claims in the court case was that women were paid a lower starting salary than men of comparable experience and education when they were first hired. Our response variable in this analysis will be Bsal.

The code below loads the data:

```
##    Sex Senior Age Educ Exper Bsal
## 1 Male     96 329   15  14.0 5040
## 2 Male     82 357   15  72.0 6300
## 3 Male     67 315   15  35.5 6000
## 4 Male     97 354   12  24.0 6000
## 5 Male     66 351   12  56.0 6000
## 6 Male     92 374   15  41.5 6840
```

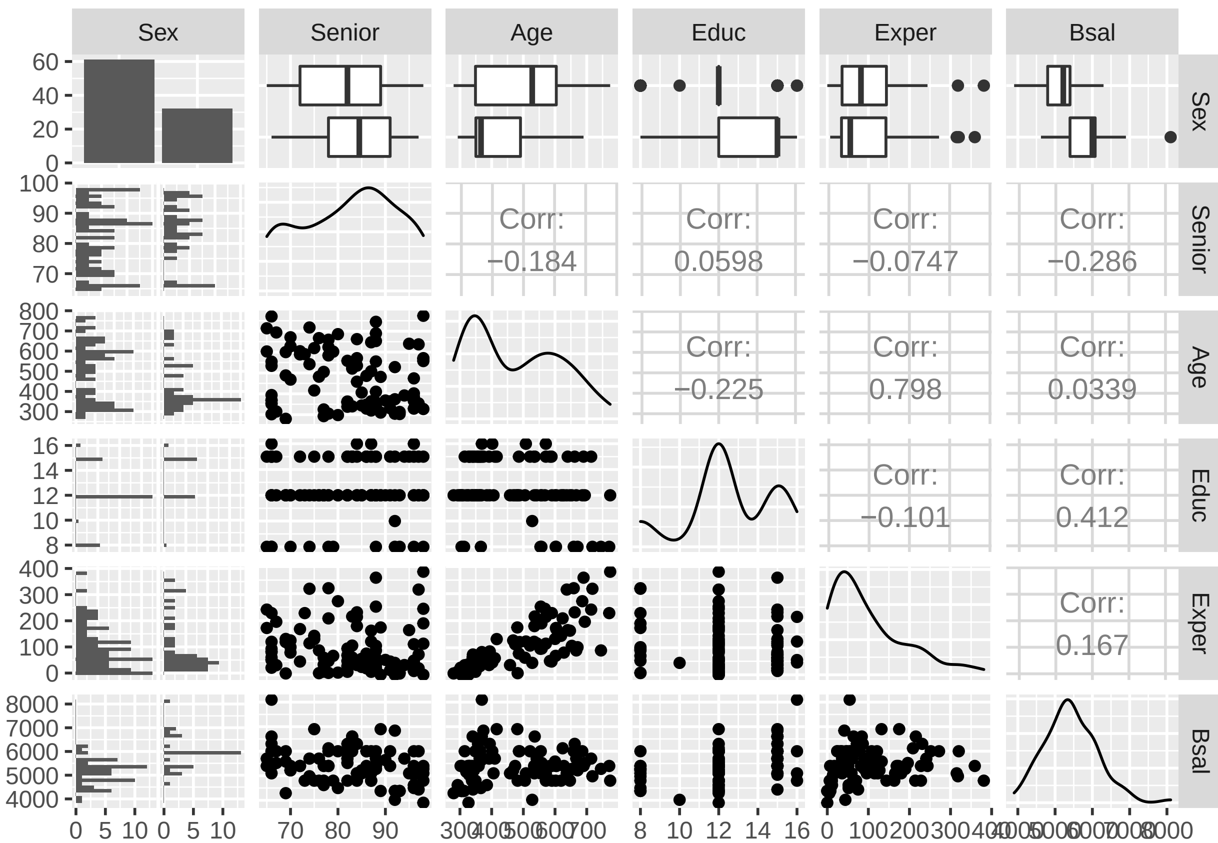
We will follow the following outline for our analysis:

1. Make initial plots
2. Do our best to identify necessary data transformations from the plots
3. Fit a model including all variables
4. Look at residuals plots from that model; tweak data transformations or add non-linear terms to the model if necessary
5. Consider outliers. Do outliers seem to be affecting inferences?
6. Select variables to include in a final model. These should definitely include `Sex` since that variable is related to the primary purpose of our analysis.
7. Fit final model(s) and double check residuals one more time.
8. Summarize our findings across all combination of models with and without outliers (if necessary) and with various sets of explanatory variables (if necessary).

1. Make a pairs plot of the data

```r
ggpairs(discrim)
```

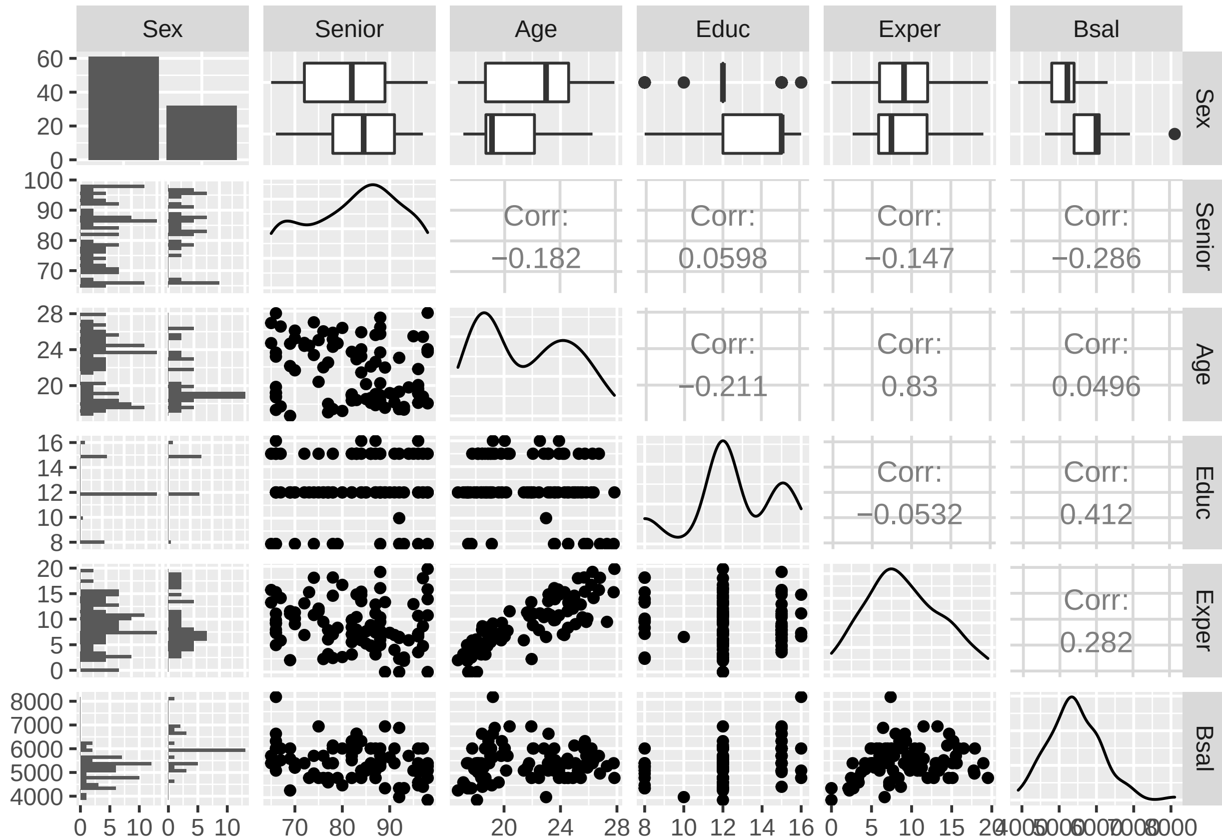
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



2. See if you can identify transformations to address any problems you can see in the pairs plots. Note: the model is much more interpretable if you can justify not transforming the response (i.e., transforming the response variable is only worth it if you don't trust the model othewise, not to fix minor problems).

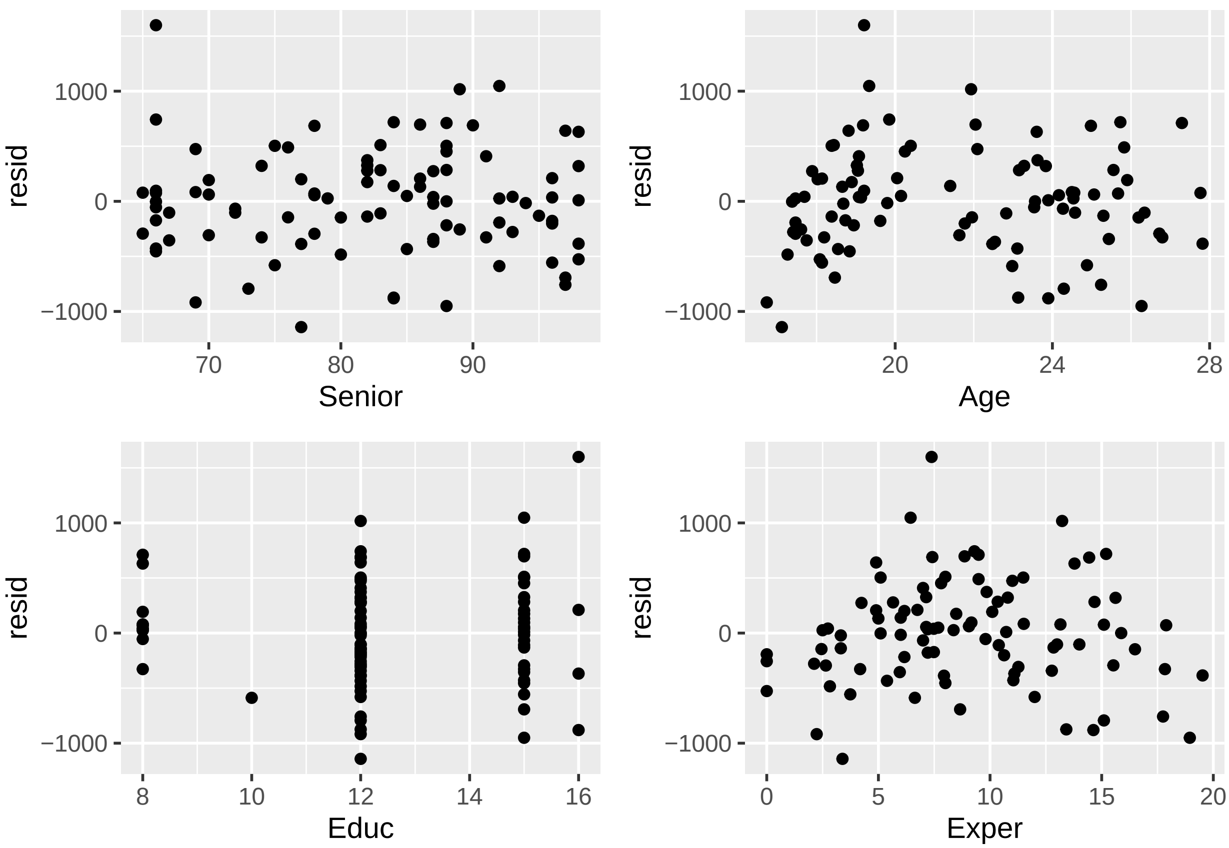
```r
discrim_transformed <- discrim %>% mutate(Age = sqrt(Age), Exper = sqrt(Exper))
ggpairs(discrim_transformed)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



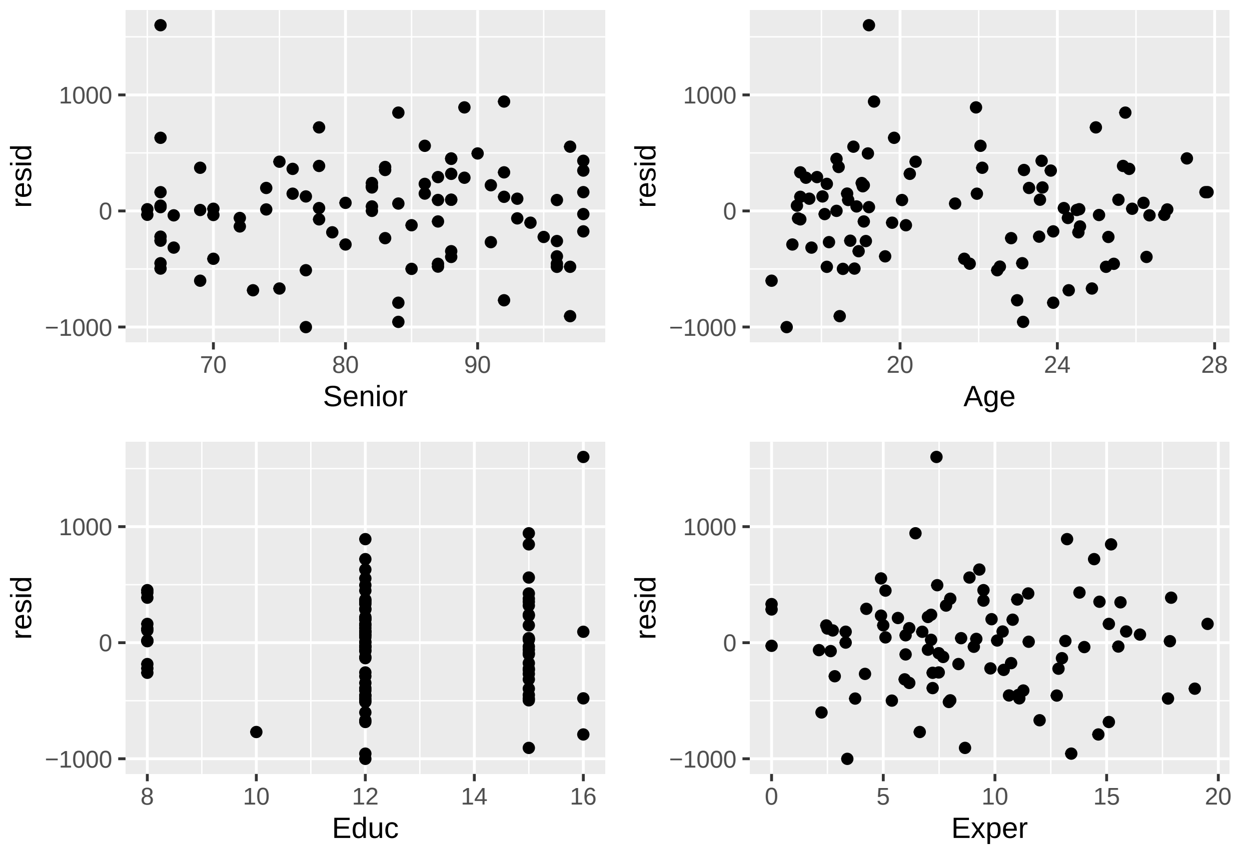
3. Fit a model including all explanatory variables and create plots of the residuals vs explanatory variables

```
lm_fit <- lm(Bsal ~ Sex + Senior + Age + Educ + Exper, data = discrim_transformed)
discrim_transformed <- discrim_transformed %>%
  mutate(
    resid = residuals(lm_fit)
  )
p1 <- ggplot(data = discrim_transformed, mapping = aes(x = Senior, y = resid)) +
  geom_point()
p2 <- ggplot(data = discrim_transformed, mapping = aes(x = Age, y = resid)) +
  geom_point()
p3 <- ggplot(data = discrim_transformed, mapping = aes(x = Educ, y = resid)) +
  geom_point()
p4 <- ggplot(data = discrim_transformed, mapping = aes(x = Exper, y = resid)) +
  geom_point()
grid.arrange(p1, p2, p3, p4)
```



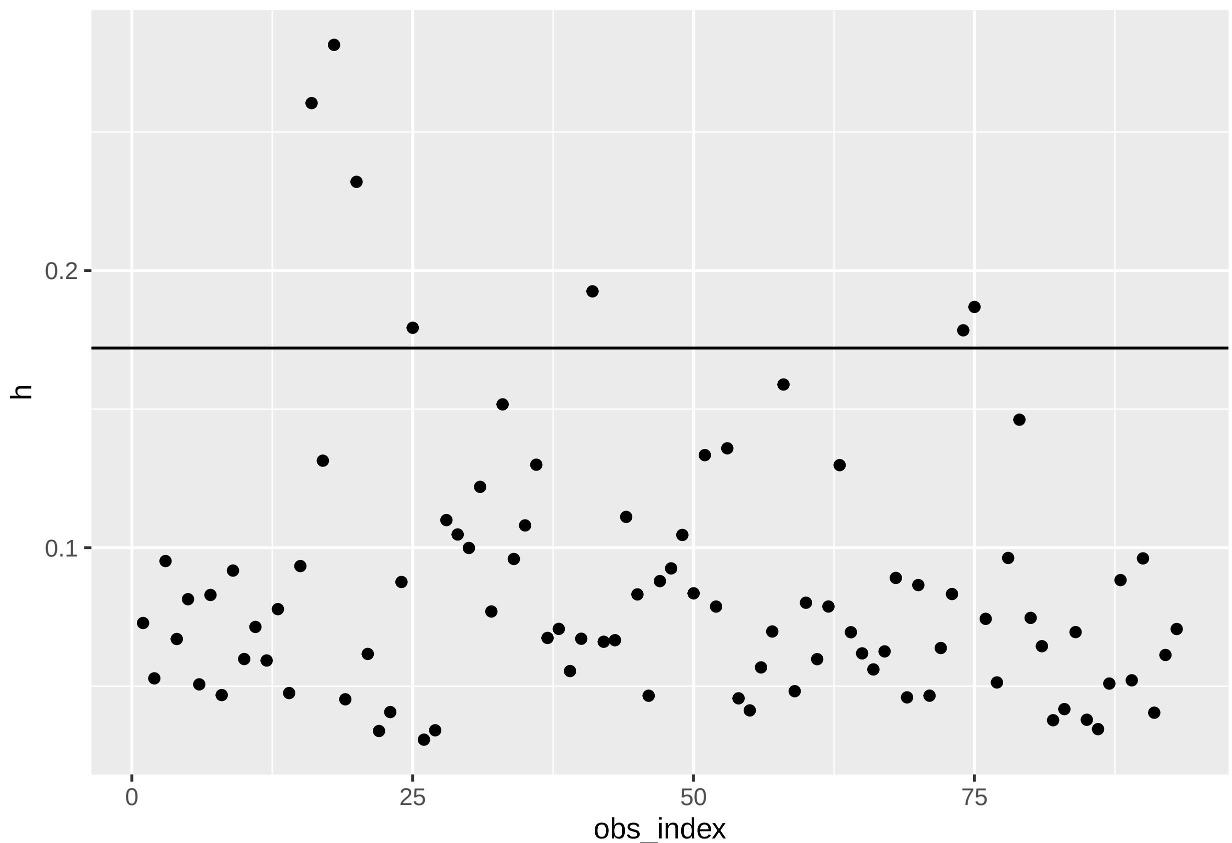
4. Tweak data transformations or add non-linear terms to the model if necessary

```
lm_fit <- lm(Bsal ~ Sex + Senior + Age + I(Age^2) + Educ + Exper + I(Exper^2), data = discrim_transforme
discrim_transformed <- discrim_transformed %>%
  mutate(
    resid = residuals(lm_fit)
  )
p1 <- ggplot(data = discrim_transformed, mapping = aes(x = Senior, y = resid)) +
  geom_point()
p2 <- ggplot(data = discrim_transformed, mapping = aes(x = Age, y = resid)) +
  geom_point()
p3 <- ggplot(data = discrim_transformed, mapping = aes(x = Educ, y = resid)) +
  geom_point()
p4 <- ggplot(data = discrim_transformed, mapping = aes(x = Exper, y = resid)) +
  geom_point()
grid.arrange(p1, p2, p3, p4)
```

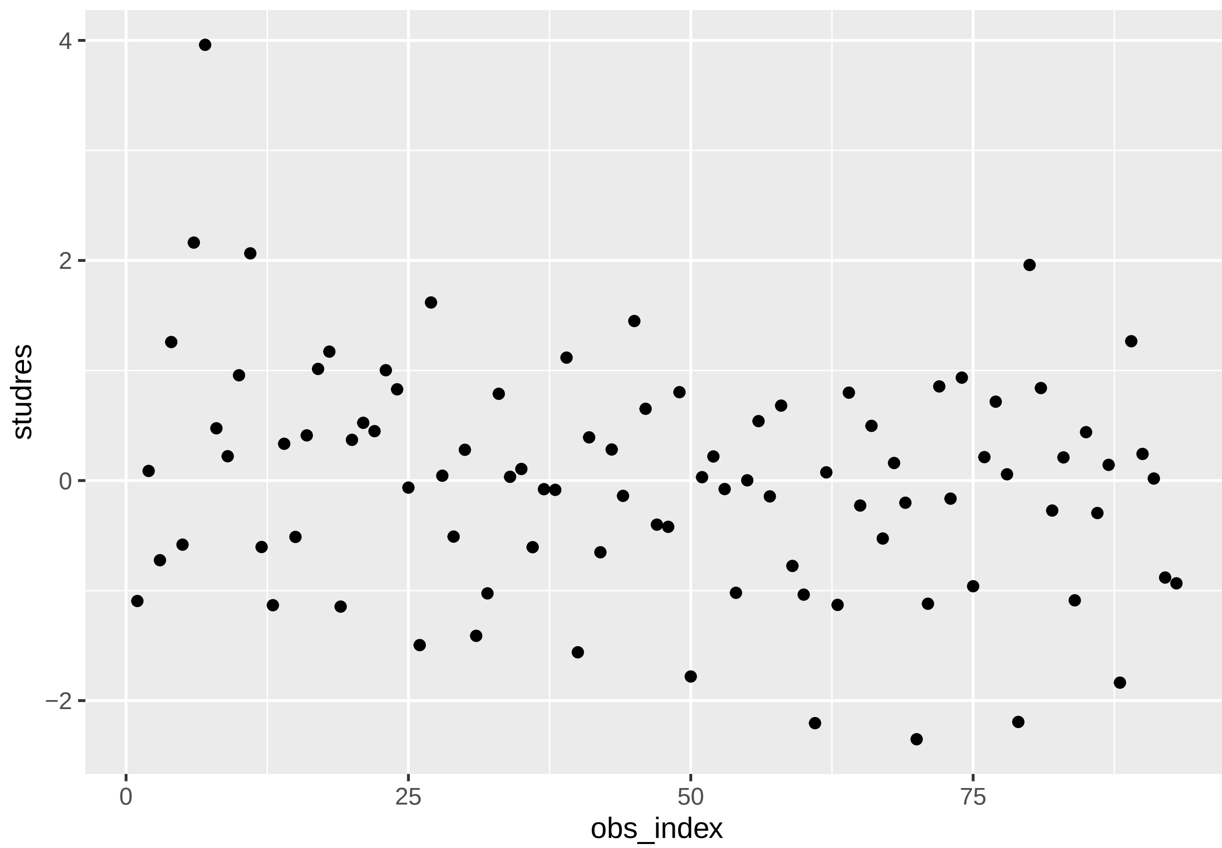


5. Consider outliers. Do outliers seem to be affecting inferences?

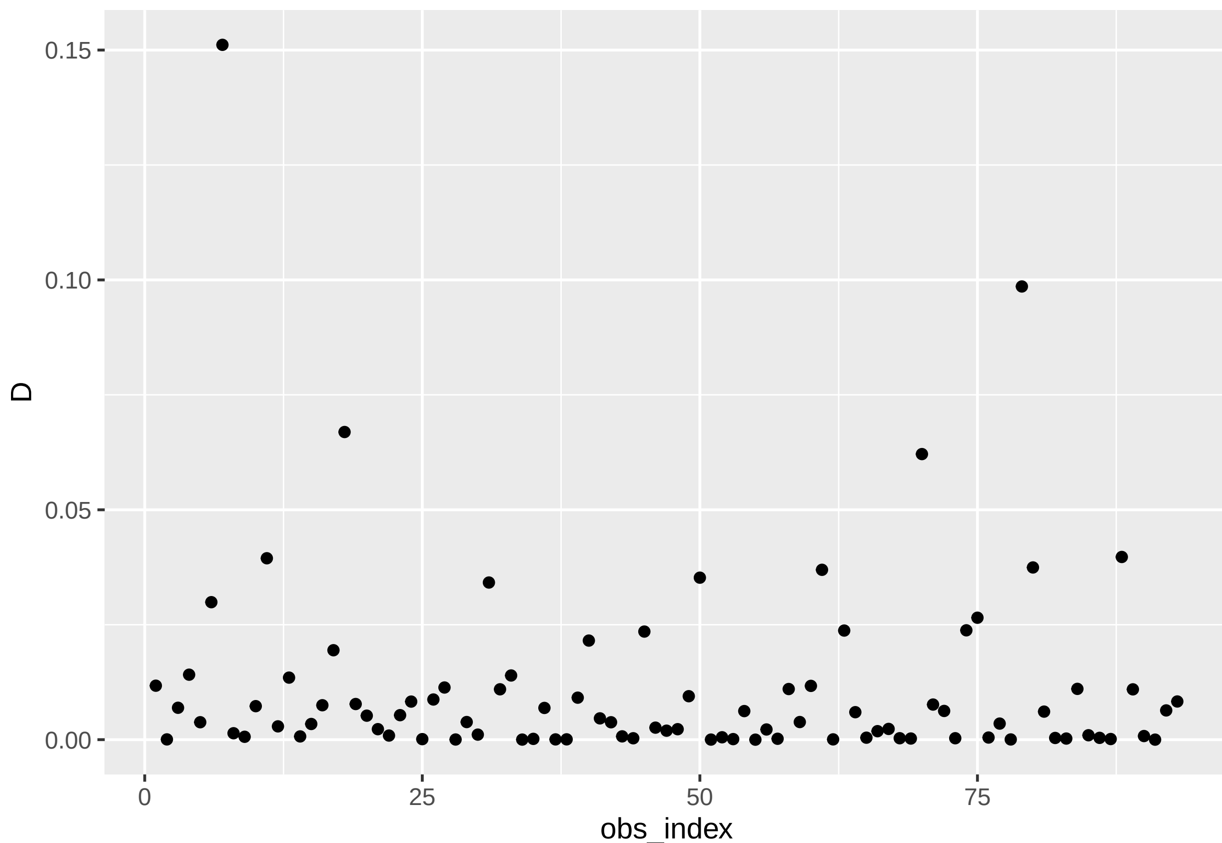
```
discrim_transformed <- discrim_transformed %>%
  mutate(
    obs_index = row_number(),
    h = hatvalues(lm_fit),
    studres = rstudent(lm_fit),
    D = cooks.distance(lm_fit)
  )
ggplot(data = discrim_transformed, mapping = aes(x = obs_index, y = h)) +
  geom_hline(yintercept = 2 * 8 / nrow(discrim_transformed))+
  geom_point()
```




```
ggplot(data = discrim_transformed, mapping = aes(x = obs_index, y = studres)) +
  geom_point()
```




```
ggplot(data = discrim_transformed, mapping = aes(x = obs_index, y = D)) +
  geom_point()
```

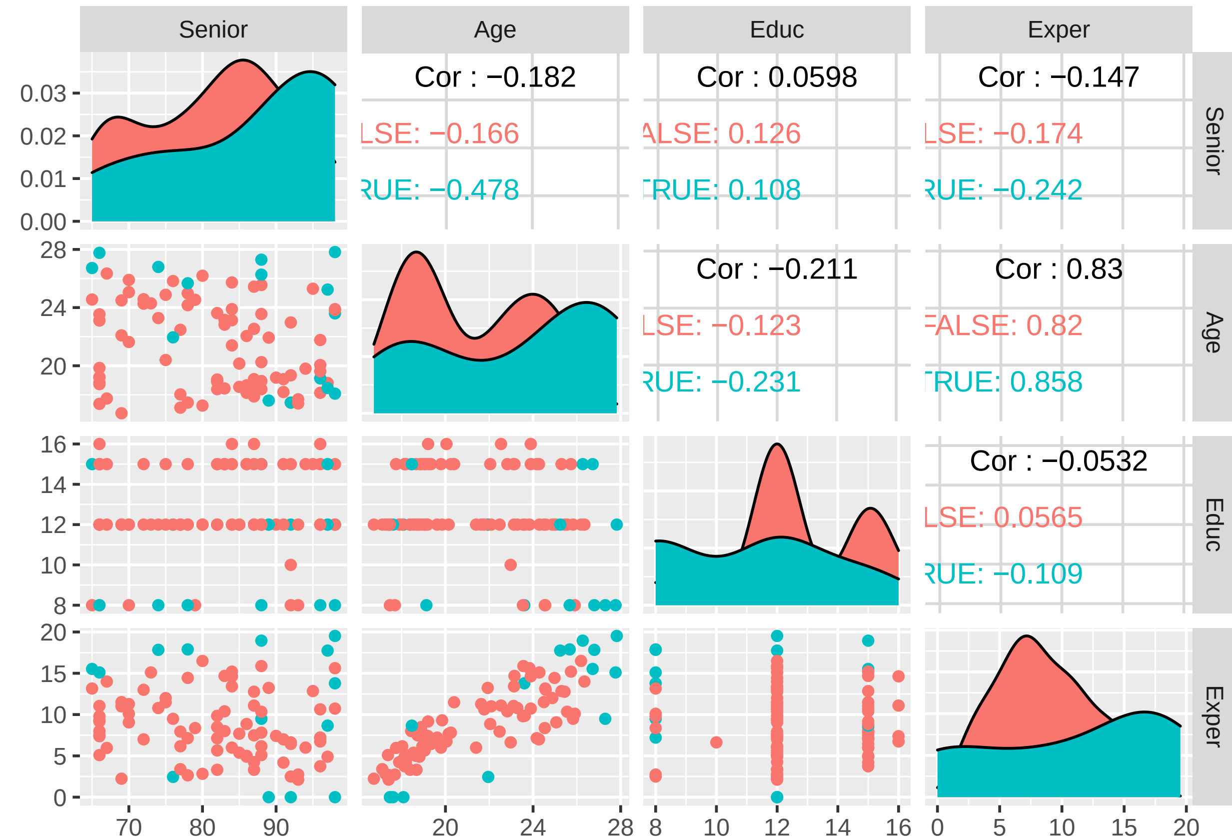



```r
discrim_transformed %>%
  filter(h >  2 * 6 / nrow(discrim_transformed))
```

```
##       Sex Senior      Age Educ     Exper Bsal      resid obs_index         h     studres
## 1  Female     98 27.82086   12 19.519221 4800  162.13445        16 0.2604446  0.41044658
## 2  Female     98 23.60085    8 13.784049 5280  432.06862        17 0.1314032  1.01442344
## 3  Female     88 27.29469    8  9.486833 5280  452.93313        18 0.2814736  1.17155813
## 4  Female     76 21.95450   12  2.449490 4800  148.89207        20 0.2320333  0.36981562
## 5  Female     98 18.08314   12  0.000000 3900  -26.75971        25 0.1793659 -0.06424641
## 6  Female     92 17.46425   12  0.000000 4380  332.62779        33 0.1517253  0.78835650
## 7  Female     96 19.13113    8  7.211103 4500 -259.23873        36 0.1299445 -0.60576558
## 8  Female     66 27.76689    8 15.099669 5400  161.96947        41 0.1925257  0.39237206
## 9  Female     74 26.79552    8 17.832555 4980   13.05876        51 0.1334047  0.03050898
## 10 Female     65 26.72078   15 15.524175 5700  -33.06952        53 0.1358698 -0.07737228
## 11 Female     89 17.60682   12  0.000000 4380  286.30088        58 0.1588934  0.68080649
## 12   Male     97 25.23886   12 17.748239 5100 -481.06306        63 0.1297929 -1.13005270
## 13   Male     78 25.67100    8 17.888544 6000  387.77194        74 0.1784541  0.93528017
## 14   Male     88 26.26785   15 18.947295 5400 -396.03240        75 0.1868844 -0.96041225
## 15 Female     97 18.46619   15  8.660254 4440 -906.44963        79 0.1461865 -2.19377230
##               D
## 1  7.489234e-03
## 2  1.945305e-02
## 3  6.691651e-02
## 4  5.218210e-03
## 5  1.141081e-04
```

```
## 6  1.395774e-02
## 7  6.902030e-03
## 8  4.634581e-03
## 9  1.812403e-05
## 10 1.190511e-04
## 11 1.101443e-02
## 12 2.373135e-02
## 13 2.378640e-02
## 14 2.652424e-02
## 15 9.857835e-02
```

```r
discrim_transformed <- discrim_transformed %>%
  mutate(suspicious = (h >  2 * 6 / nrow(discrim_transformed)))
ggpairs(discrim_transformed, mapping = aes(color = suspicious), columns = 2:5)
```




```r
discrim_no_suspicious <- discrim_transformed %>%
  filter(!suspicious)
lm_fit2 <- lm(Bsal ~ Sex + Senior + Age + I(Age^2) + Educ + Exper + I(Exper^2), data = discrim_no_suspi
summary(lm_fit)
```

```
##
## Call:
## lm(formula = Bsal ~ Sex + Senior + Age + I(Age^2) + Educ + Exper +
##     I(Exper^2), data = discrim_transformed)
##
```

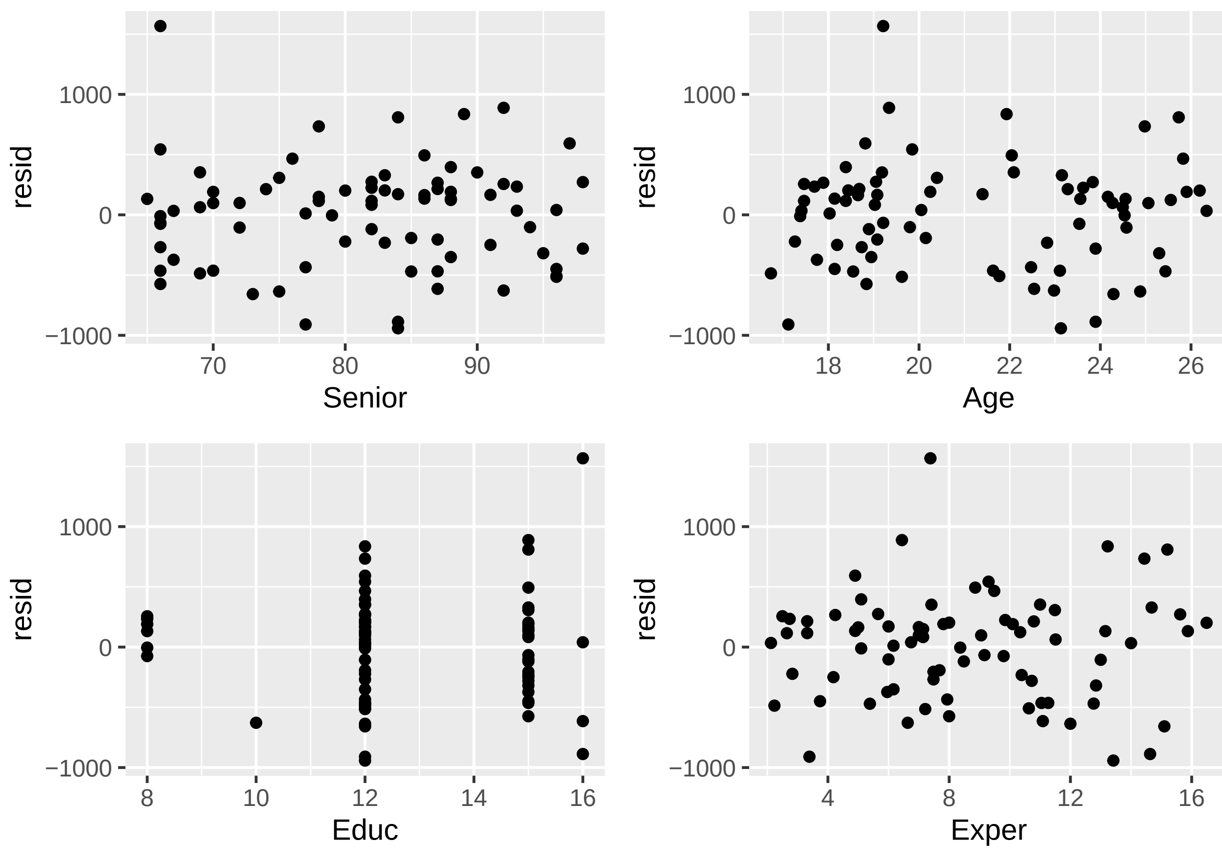
```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1000.74  -268.76    19.08   240.72  1600.57
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5400.5424  3504.9611   1.541  0.12707
## SexMale      686.2562   117.4113   5.845 9.12e-08 ***
## Senior       -16.9148     5.0367  -3.358  0.00118 **
## Age          -37.3359   341.9422  -0.109  0.91331
## I(Age^2)       0.1811     7.7085   0.023  0.98132
## Educ          66.6511    23.2124   2.871  0.00516 **
## Exper        211.4974    54.7914   3.860  0.00022 ***
## I(Exper^2)    -8.2271     2.5238  -3.260  0.00160 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 457.1 on 85 degrees of freedom
## Multiple R-squared:  0.6166, Adjusted R-squared:  0.5851
## F-statistic: 19.53 on 7 and 85 DF,  p-value: 2.44e-15
```

```r
summary(lm_fit2)
```

```
##
## Call:
## lm(formula = Bsal ~ Sex + Senior + Age + I(Age^2) + Educ + Exper +
##     I(Exper^2), data = discrim_no_suspicious)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -941.76 -309.16   73.32  222.28 1567.25
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7644.529   5434.006   1.407  0.16391
## SexMale      609.433    130.628   4.665 1.44e-05 ***
## Senior       -13.029      6.202  -2.101  0.03925 *
## Age         -302.959    542.871  -0.558  0.57858
## I(Age^2)       5.130     12.200   0.420  0.67543
## Educ          82.575     28.767   2.870  0.00542 **
## Exper        328.899    107.881   3.049  0.00324 **
## I(Exper^2)   -12.545      4.954  -2.532  0.01358 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 462.8 on 70 degrees of freedom
## Multiple R-squared:  0.6043, Adjusted R-squared:  0.5648
## F-statistic: 15.27 on 7 and 70 DF,  p-value: 5.902e-12
```

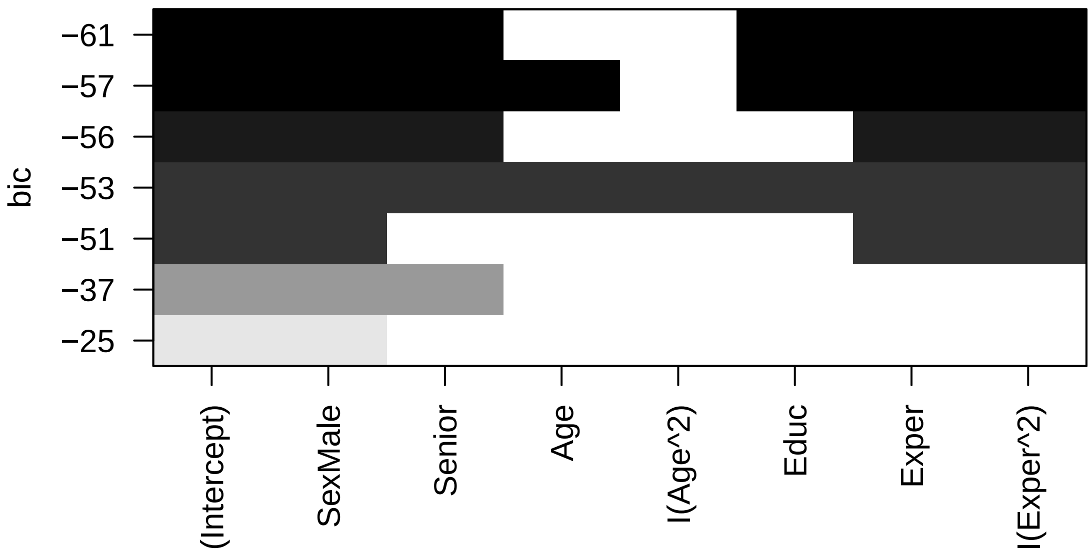
```r
discrim_no_suspicious <- discrim_no_suspicious %>%
  mutate(
    resid = residuals(lm_fit2)
  )
```

```r
p1 <- ggplot(data = discrim_no_suspicious, mapping = aes(x = Senior, y = resid)) +
  geom_point()
p2 <- ggplot(data = discrim_no_suspicious, mapping = aes(x = Age, y = resid)) +
  geom_point()
p3 <- ggplot(data = discrim_no_suspicious, mapping = aes(x = Educ, y = resid)) +
  geom_point()
p4 <- ggplot(data = discrim_no_suspicious, mapping = aes(x = Exper, y = resid)) +
  geom_point()
grid.arrange(p1, p2, p3, p4)
```



6. Select variables to include in a final model. These should definitely include `Sex` since that variable is related to the primary purpose of our analysis.

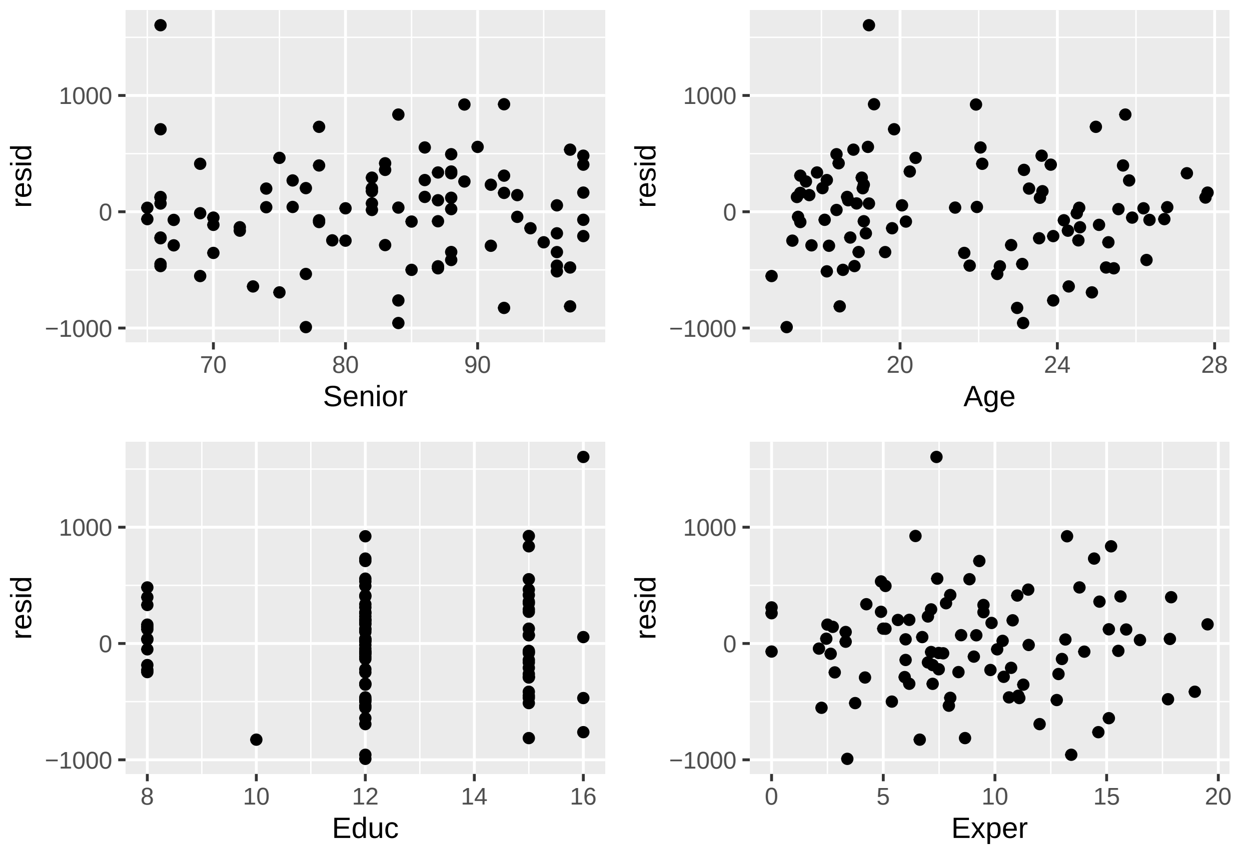
```r
library(leaps)
candidate_models <- regsubsets(Bsal ~ Sex + Senior + Age + I(Age^2) + Educ + Exper + I(Exper^2), data =
plot(candidate_models)
```



I will include all the variables above other than Age and Age squared.

7. Fit final model(s) and double check residuals one more time.

```
lm_fit <- lm(Bsal ~ Sex + Senior + Educ + Exper + I(Exper^2), data = discrim_transformed)
discrim_transformed <- discrim_transformed %>%
  mutate(
    resid = residuals(lm_fit)
  )
p1 <- ggplot(data = discrim_transformed, mapping = aes(x = Senior, y = resid)) +
  geom_point()
p2 <- ggplot(data = discrim_transformed, mapping = aes(x = Age, y = resid)) +
  geom_point()
p3 <- ggplot(data = discrim_transformed, mapping = aes(x = Educ, y = resid)) +
  geom_point()
p4 <- ggplot(data = discrim_transformed, mapping = aes(x = Exper, y = resid)) +
  geom_point()
grid.arrange(p1, p2, p3, p4)
```




```
lm_fit2 <- lm(Bsal ~ Sex + Senior + Educ + Exper + I(Exper^2), data = discrim_no_suspicious)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = Bsal ~ Sex + Senior + Educ + Exper + I(Exper^2),
##     data = discrim_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -991.93 -286.49   22.71  269.24 1604.24
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4763.043    522.722   9.112 2.65e-14 ***
## SexMale      733.453    105.482   6.953 6.26e-10 ***
## Senior       -16.713      4.862  -3.438 0.000902 ***
## Educ          70.337     22.481   3.129 0.002389 **
## Exper        192.071     40.178   4.780 7.06e-06 ***
## I(Exper^2)    -8.092      2.027  -3.992 0.000137 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 454.2 on 87 degrees of freedom
## Multiple R-squared:  0.6126, Adjusted R-squared:  0.5904
## F-statistic: 27.52 on 5 and 87 DF,  p-value: < 2.2e-16
```

```r
summary(lm_fit2)
```

```
##
## Call:
## lm(formula = Bsal ~ Sex + Senior + Educ + Exper + I(Exper^2),
##     data = discrim_no_suspicious)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -945.96 -329.22    8.84  260.94 1547.38
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4223.188    632.372   6.678 4.31e-09 ***
## SexMale      725.465    118.661   6.114 4.54e-08 ***
## Senior       -15.185      5.778  -2.628  0.01049 *
## Educ          94.500     28.512   3.314  0.00144 **
## Exper        216.581     69.368   3.122  0.00258 **
## I(Exper^2)    -9.260      3.724  -2.487  0.01522 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 469.3 on 72 degrees of freedom
## Multiple R-squared:  0.5816, Adjusted R-squared:  0.5525
## F-statistic: 20.02 on 5 and 72 DF,  p-value: 1.864e-12
```

Overall, things look pretty good. There is increasing standard deviation of residuals for higher education levels. It seems unlikely we could fix that, but also unlikely that that is going to affect our inferences substantially enough to change our conclusions.

8. Summarize our findings across all combination of models with and without outliers (if necessary) and with various sets of explanatory variables (if necessary). Focus on the estimated coefficient for sex. It's always nice to get confidence intervals for effects you want to describe.

```r
confint(lm_fit)
```

```
##                  2.5 %      97.5 %
## (Intercept) 3724.07675 5802.009754
## SexMale      523.79607  943.108964
## Senior       -26.37570   -7.050095
## Educ          25.65276  115.020802
## Exper        112.21277  271.929732
## I(Exper^2)   -12.12158   -4.062766
```

```r
confint(lm_fit2)
```

```
##                  2.5 %      97.5 %
## (Intercept) 2962.57672 5483.799699
## SexMale      488.91871  962.011343
## Senior       -26.70338   -3.667328
```

```
## Educ           37.66254  151.338057
## Exper          78.29859  354.862595
## I(Exper^2)   -16.68445   -1.836197
```

There is extremely strong evidence that men were paid higher base salaries than women, after accounting for seniority, education level, and experience. We estimate that the difference in population mean starting salaries between men and women starting at this bank between 1965 and 1975 is approximately $730, with a 95% confidence interval ranging from about $500 to about $950. These estimates were fairly stable whether or not several outlying or high leverage observations were included.