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

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).

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

4. 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.

```
library(leaps)
```

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). Focus on the estimated coefficient for sex. It's always nice to get confidence intervals for effects you want to describe.