Non-nested Comparisons

David Gerard

2018-12-07

Learning Objective

- Sections 10.4.1 and 12.4
- Choosing Between Non-nested Models

Case Study and EDA

Case Study: Sex Descrimination

- Same study as in Case Study 0102
- Looked at beginning salary at a bank with respect to sex.
- Want to control for many different variables.

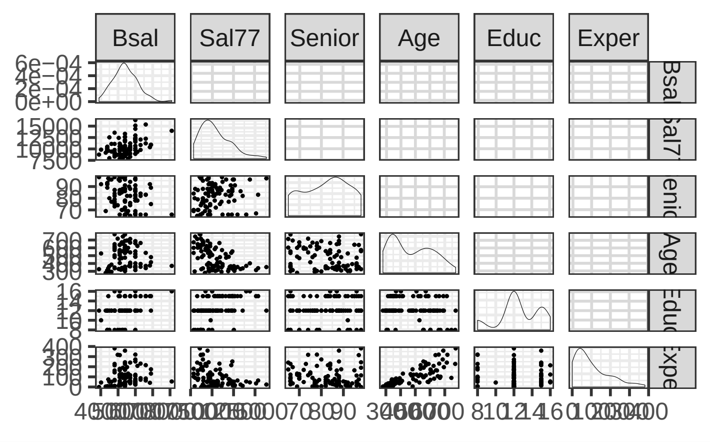
Case Study: Sex Descrimination

```r
library(Sleuth3)
data(case1202)
head(case1202)
```

```
##   Bsal Sal77  Sex Senior Age Educ Exper
## 1 5040 12420 Male     96 329   15  14.0
## 2 6300 12060 Male     82 357   15  72.0
## 3 6000 15120 Male     67 315   15  35.5
## 4 6000 16320 Male     97 354   12  24.0
## 5 6000 12300 Male     66 351   12  56.0
## 6 6840 10380 Male     92 374   15  41.5
```

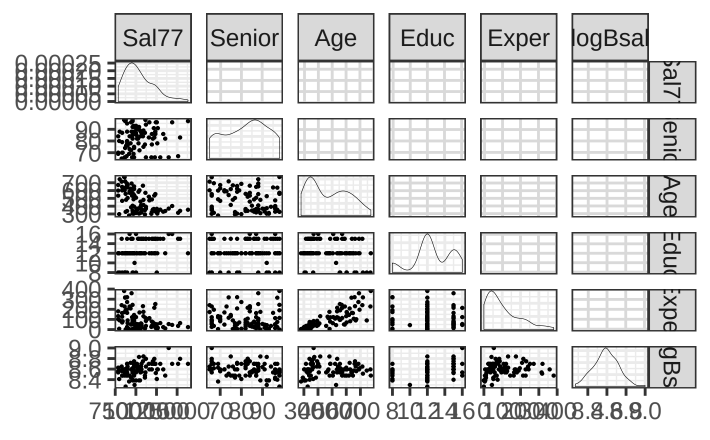
EDA

```r
library(GGally)
ggpairs(case1202, columns = c(1, 2, 4, 5, 6, 7),
        aes(size = I(0.1)))
```



EDA

```r
case1202$logBsal <- log(case1202$Bsal)
ggpairs(case1202, columns = c(2, 4, 5, 6, 7, 8),
        aes(size = I(0.1)))
```



EDA Summary

- Loging `Bsal` seems to help a lot.
- Age and Experience might need a quadratic transformation.

Step-wise Procedures (Section 12.3)

Step-wise Regression

- Start with a complicated model.
- Look at p-values (when testing that a coefficient is 0)
- Drop the one with the largest p-value.
- Continue until all p-values are less than some threshold (usually 0.05).
- Note, you cannot interpret p-values the way we define them anymore if you do this.

Step-wise Regression, the manual way

```r
case1202$Age2   <- case1202$Age ^ 2
case1202$Exper2 <- case1202$Exper ^ 2
lm1 <- lm(logBsal ~ Senior + Age + Age2 +
            Senior + Educ + Exper + Exper2,
          data = case1202)
```

Step-wise Regression, the manual way

```r
coef(summary(lm1))
```

```
##                  Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)  8.630e+00  2.139e-01 40.35158 1.155e-57
## Senior      -3.242e-03  1.150e-03 -2.82072 5.948e-03
## Age         -3.094e-04  9.167e-04 -0.33749 7.366e-01
## Age2        -2.788e-08  8.828e-07 -0.03159 9.749e-01
## Educ         2.063e-02  5.095e-03  4.04819 1.125e-04
## Exper        1.960e-03  6.091e-04  3.21735 1.825e-03
## Exper2      -4.098e-06  1.657e-06 -2.47275 1.538e-02
```

- Drop `Age2` (*p*-value of 0.97)

Step-wise Regression, the manual way

```
##                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  8.635e+00  1.324e-01  65.200 1.207e-75
## Senior      -3.234e-03  1.114e-03  -2.902 4.693e-03
## Age         -3.380e-04  1.449e-04  -2.332 2.200e-02
## Educ         2.065e-02  5.003e-03   4.128 8.372e-05
## Exper        1.970e-03  5.053e-04   3.900 1.892e-04
## Exper2      -4.130e-06  1.301e-06  -3.175 2.071e-03
```

Step-wise Regression

- Can also start at the simplest model,
  - add the variable that has the smallest p-value
  - continue until no new variables would have a p-value less than 0.05
- Can also both add and drop variables based on p-values.

Step-wise Regression in R

- Use the step() function to do this automatically

```
lm1 <- lm(logBsal ~ Senior + Age + Age2 +
            Senior + Educ + Exper + Exper2,
          data = case1202)
stepout <- step(object = lm1, trace = FALSE)
stepout
```

```
##
## Call:
## lm(formula = logBsal ~ Senior + Age + Educ + Exper + Exper2,
##     data = case1202)
##
## Coefficients:
## (Intercept)       Senior          Age         Educ        Exper
##    8.63e+00    -3.23e-03    -3.38e-04     2.07e-02     1.97e-03
##      Exper2
##   -4.13e-06
```

Step-wise Regression in R

- The output of `step()` is also an `lm` object, so you can get coefficients, *p*-values, confidence intervals, fits, predictions, residuals, etc directly from it.

```r
confint(stepout)
```

```
##                    2.5 %        97.5 %
## (Intercept)  8.372e+00  8.898e+00
## Senior      -5.450e-03 -1.019e-03
## Age         -6.260e-04 -4.994e-05
## Educ         1.071e-02  3.060e-02
## Exper        9.661e-04  2.975e-03
## Exper2      -6.716e-06 -1.545e-06
```

Comparing Non-nested Models (Section 12.4)

Motivation

- What if we want to decide between the following two models
- $\mu(logBsal|...) = Senior + Educ + Exper + Exper^2$
- $\mu(logBsal|...) = Senior + Educ + Age + Age^2$
- These models are non-nested, so we cannot apply F-test techniques to them.

BIC and AIC

- BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion) return the log of the sum of square residuals **plus** a penalty due to the number of parameters in the model.
- Best model has the smallest BIC or AIC.
- BIC: $n \log(SSR/n) + \log(n)(p+1)$
- AIC: $n \log(SSR/n) + 2(p+1)$
- BIC penalizes more when the sample size is larger.
- BIC is better for model selection (get interpretable model), AIC is better for prediction (goal is prediction).

BIC and AIC in R

- Fit both models, then use the `AIC()` and `BIC()` functions.

```
lm_mod1 <- lm(logBsal ~ Senior + Educ + Exper + Exper2,
              data = case1202)
lm_mod2 <- lm(logBsal ~ Senior + Educ + Age + Age2,
              data = case1202)
BIC(lm_mod1)
```

```
## [1] -131.2
```

```
BIC(lm_mod2)
```

```
## [1] -123.6
```

```
AIC(lm_mod1)
```

```
## [1] -146.4
```

```
AIC(lm_mod2)
```

Mallow's C_p statistic

- $Bias(\hat{Y}_i) = \mu(\hat{Y}_i) - \mu(Y_i)$
- $MSE(\hat{Y}_i) = Bias(\hat{Y}_i)^2 + Var(\hat{Y}_i)$
- $TMSE = \sum_{i=1}^{n} MSE(\hat{Y}_i)$
- We don't know the $TMSE$, but Mallow's C_p **estimates** it.

C_p plot

- You obtain Mallow's C_p for **every** possible model.
- Only feasible if you have less than $p = 10$ or so explanatory variables (2^p models are possible).
- Plot C_p on the y-axis and the number of parameters on the x-axis.
- Models below the $y = x$ line are candidate models
  - Models without bias should have a C_p of about p
  - So if C_p is below p, the model probably does not have any bias issues.

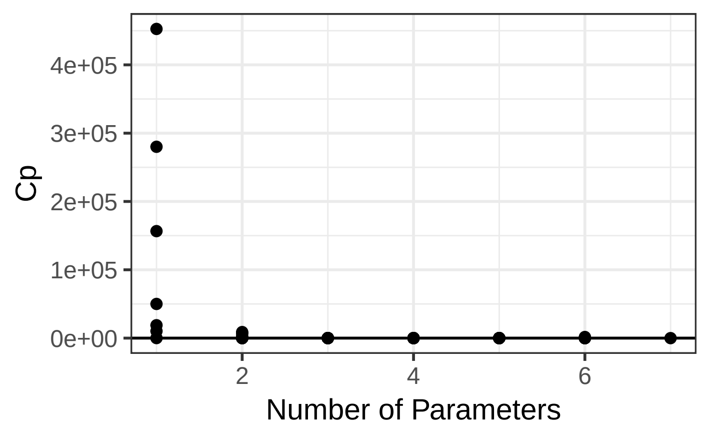
C_p in R

- We will use the `leaps()` function in the `leaps` library.

```
library(leaps)
lm1 <- lm(logBsal ~ Senior + Age + Age2 +
            Senior + Educ + Exper + Exper2,
          data = case1202)
X <- model.matrix(lm1)
leapsout <- leaps(x =  X,
                  y = case1202$logBsal,
                  int = FALSE)
```

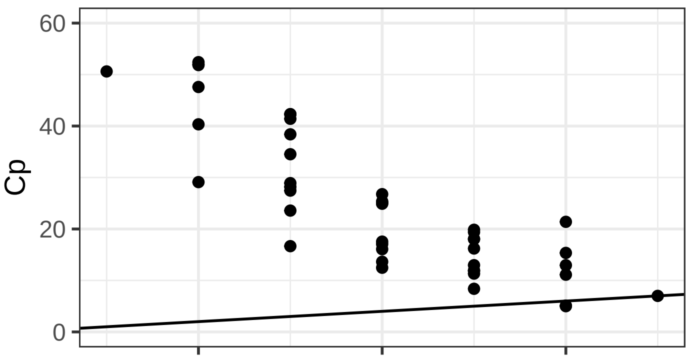
C_p in R

```r
qplot(leapsout$size, leapsout$Cp,
      xlab = "Number of Parameters",
      ylab = "Cp") +
  geom_abline(slope = 1, intercept = 0)
```



C_p in R

```
goodmodel <- leapsout$Cp < 1000
qplot(leapsout$size[goodmodel], leapsout$Cp[goodmodel],
      xlab = "Number of Parameters",
      ylab = "Cp") +
  geom_abline(slope = 1, intercept = 0) +
  ylim(0, 60)
```



Back to Case Study

Back to Case Study

- We chose a model with

  $$\mu(logBsal|...) = Senior + Age + Senior + Educ + Exper + Exper^2$$

- Now let's answer the question if `Sex` is still associated with base salary after adjusting for these variables.

Results

```r
lmfinal <- lm(logBsal ~ Sex + Senior + Age +
                Senior + Educ + Exper + Exper2,
              data = case1202)
coef(summary(lmfinal))
```

```
##                  Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  8.567e+00  1.097e-01 78.1245 1.199e-81
## SexMale      1.405e-01  2.167e-02  6.4842 5.401e-09
## Senior      -3.261e-03  9.186e-04 -3.5497 6.279e-04
## Age         -2.079e-05  1.291e-04 -0.1611 8.724e-01
## Educ         1.373e-02  4.260e-03  3.2232 1.792e-03
## Exper        1.549e-03  4.215e-04  3.6755 4.123e-04
## Exper2      -4.128e-06  1.072e-06 -3.8502 2.264e-04
```

Results

```r
cbind(coef(lmfinal), confint(lmfinal))
```

```
##                                2.5 %     97.5 %
## (Intercept)  8.567e+00  8.349e+00  8.785e+00
## SexMale      1.405e-01  9.742e-02  1.836e-01
## Senior      -3.261e-03 -5.087e-03 -1.435e-03
## Age         -2.079e-05 -2.774e-04  2.358e-04
## Educ         1.373e-02  5.262e-03  2.220e-02
## Exper        1.549e-03  7.113e-04  2.387e-03
## Exper2      -4.128e-06 -6.260e-06 -1.997e-06
```

Results

```r
exp(coef(lmfinal)[2])
```

```
## SexMale
##   1.151
```

```r
exp(confint(lmfinal)[2, ])
```

```
##  2.5 % 97.5 %
##  1.102  1.201
```