Lecture 20 notes
Model Selection

We'll again work with the big mac data, but this time (almost) all predictors will be logged.

```
> bm <- read.table("bigmac.dat", header=T)
> names(bm)
 [1] "bigmac"   "bread"    "busfare"  "engsal"   "engtax"
"service"
 [7] "teachsal" "teachtax" "vacdays"  "workhrs"  "city"
> attach(bm)
> lbread <- log(bread)
> lbusfare <- log(busfare)
> lengsal <- log(engsal)
> lengtax <- log(engtax)
> lservice <- log(service)
> lteachsal <- log(teachsal)
> lteachtax <- log(teachtax)

> bmnew <- data.frame(bigmac, lbread, lbusfare, lengsal,
lengtax, lservice, lteachsal, lteachtax,vacdays, workhrs)
```

First, we fit the full model.

```
> full <- lm(bigmac~., data=bmnew)
> summary(full)

Call:
lm(formula = bigmac ~ ., data = bmnew)

Residuals:
    Min      1Q  Median      3Q     Max
-40.019 -12.672  -3.009  11.370  76.543

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  70.91348  114.55044   0.619  0.53989
lbread        7.83756    7.71698   1.016  0.31678
lbusfare    -24.00557    9.17079  -2.618  0.01299 *
lengsal     -46.06286   15.95945  -2.886  0.00664 **
lengtax      23.03688   16.38954   1.406  0.16866
```

```
lservice     22.01032   15.07442   1.460  0.15318
lteachsal     4.74398   13.45146   0.353  0.72645
lteachtax     3.36451   15.06750   0.223  0.82460
vacdays      -1.22750    0.71497  -1.717  0.09484 .
workhrs      -0.04842    0.03400  -1.424  0.16324
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 23.4 on 35 degrees of freedom
Multiple R-Squared: 0.7856,   Adjusted R-squared: 0.7305
F-statistic: 14.25 on 9 and 35 DF,  p-value: 2.618e-09
```

Only busfare and engineer's salries appear to be useful predictors, although the number of vacation days might also play a role.

We'll do a stepwise regression, backwards, using AIC as our comparison criteria.

```
> step(full)
Start:  AIC= 292.45
 bigmac ~ lbread + lbusfare + lengsal + lengtax + lservice
+ lteachsal +
    lteachtax + vacdays + workhrs

            Df Sum of Sq     RSS     AIC
- lteachtax  1      27.3 19196.4   290.5
- lteachsal  1      68.1 19237.2   290.6
- lbread     1     564.9 19734.0   291.8
<none>                   19169.1   292.4
- lengtax    1    1082.0 20251.1   292.9
- workhrs    1    1110.9 20280.0   293.0
- lservice   1    1167.6 20336.7   293.1
- vacdays    1    1614.4 20783.5   294.1
- lbusfare   1    3752.7 22921.8   298.5
- lengsal    1    4562.5 23731.5   300.1

Step:  AIC= 290.51
 bigmac ~ lbread + lbusfare + lengsal + lengtax + lservice
+ lteachsal +
    vacdays + workhrs
```

```
            Df Sum of Sq     RSS     AIC
- lteachsal  1     123.0 19319.3   288.8
- lbread     1     543.0 19739.4   289.8
<none>                   19196.4   290.5
- workhrs    1    1085.0 20281.4   291.0
- lservice   1    1394.8 20591.2   291.7
- vacdays    1    1592.6 20789.0   292.1
- lengtax    1    3046.9 22243.3   295.1
- lbusfare   1    3751.1 22947.5   296.5
- lengsal    1    5654.3 24850.7   300.1

Step:  AIC= 288.8
 bigmac ~ lbread + lbusfare + lengsal + lengtax + lservice
+ vacdays +
    workhrs

           Df Sum of Sq     RSS     AIC
- lbread    1     467.7 19787.0   287.9
<none>                  19319.3   288.8
- workhrs   1     991.2 20310.6   289.1
- lservice  1    1541.7 20861.1   290.3
- vacdays   1    1619.7 20939.0   290.4
- lengtax   1    2962.5 22281.9   293.2
- lbusfare  1    5049.1 24368.4   297.2
- lengsal   1    9920.1 29239.4   305.4

Step:  AIC= 287.88
 bigmac ~ lbusfare + lengsal + lengtax + lservice + vacdays
+
    workhrs

           Df Sum of Sq   RSS   AIC
- workhrs   1       769 20556   288
<none>                  19787   288
- vacdays   1      1570 21357   289
- lservice  1      2105 21892   290
- lengtax   1      3667 23454   294
- lbusfare  1      4927 24714   296
- lengsal   1     22565 42352   320
```

```
Step:  AIC= 287.59
 bigmac ~ lbusfare + lengsal + lengtax + lservice + vacdays

           Df Sum of Sq   RSS   AIC
- vacdays   1       841 21397   287
<none>                 20556   288
- lservice  1      2389 22945   291
- lbusfare  1      4179 24736   294
- lengtax   1      4448 25005   294
- lengsal   1     23146 43702   320

Step:  AIC= 287.4
 bigmac ~ lbusfare + lengsal + lengtax + lservice

           Df Sum of Sq   RSS   AIC
<none>                 21397   287
- lservice  1      1914 23311   289
- lengtax   1      3629 25026   292
- lbusfare  1      4246 25643   294
- lengsal   1     22579 43976   318

Call:
lm(formula = bigmac ~ lbusfare + lengsal + lengtax +
lservice,     data = bmnew)

Coefficients:
(Intercept)     lbusfare      lengsal      lengtax
lservice
     -21.10       -18.04       -46.03        23.86
25.17
```

We started with AIC at 292 and ended with it at 287. This suggests a good model is one including busfare, engineer's salary, engineers' tax, and the service.

You won't necessarily get the same formula if you do forward regression.

You won't necessarily get the same formula if you use BICs or Mallows.

In fact, "automated" procedures are dangerous. The pvalues that are viewed at the end are not quite right, and the result is that you might end up with less predictive ability than if you had kept he full model. STill, they are a useful exploratory technique. In the end,

they should be examined in full, to step by step, and should be compared with expert knowledge.

Note that also for this particular model, the fit is not too good and more care needs to be taken to find suitable transformations and to consider maybe higher orders of the fit.