Model selection

We'll now begin a more systematic approach to model selection. The idea in model selection is to pick a reasonable subset of possible predictors from those available in the data set. Other issues include whether or not to include interaction or quadratic terms, and whether or not to transform variables.

In some ways, model selection is as much as an art as a science. There can be different goals in model selection which could lead to different models in particular cases, and there can be different opinions about which model is "best". There are some ways of doing automated model selection in the computer, but you should be aware that these approaches tend to treat each predictor as being equally important. Scientifically, some predictors might be more interesting than others, and there might be reasons for including them in the model whether or not they are statistically significant.

Model selection

The goals of prediction versus explanation can also lead to differences in choosing models. In prediction, you are interested in predicting future values of the response variable. This might lead to wanting to know which combination and levels of predictors can maximize a response, for example.

In explanation, you might be less interested in the response itself and more interested in which variables contribute to the response.

Model selection

To take a particular example where either prediction or explanation could be of interest, consider universities modeling student success (measured as years to graduation, probability of graduating within 5 years, cumulative GPA, future income after graduating, or some other measure) as predicted by high school GPA, high school class rank, ACT/SAT score, and some measure of socioeconomic status. A regression model treating success as the response and these other variables as predictors is easy enough to build, but what is the point of the model?

One possible point is to determine future criteria for enrollment. Here they might be able to predict how much changing the formula for admissions would affect graduation rates. In this case, the prediction might be more important than whether a variable passes a particular threshold for significance.

Model selection

For the same example, users of the regression might be interested whether socioeconomic status is an important variable. In this case, the absolute graduation rates aren't as significant as whether or not socioeconomic status helps explain differences graduate rates for different students. In this case also, if socioeconomic status isn't statistically significant, you might still be interested in keeping it in the model in order to compare the differences between socioeconomic groups adjusting for other variables in the model. In this case it would just be important to note that the differences are not statistically significant (although they could be practically significant).

Eliminating it from the model would essentially mean dropping the initial research question altogether, which might not make sense. In that case you might want to compare models with and without the variable of interest.

Model selection

If scientific questions aren't an issue, then usually we prefer models with fewer variables. This is often expressed as the principle of "Ockham's Razor" (or "Occam"). William of Ockham was a medieval theologian (died 1347) and philosopher, and the idea is named after him, although the idea appeared earlier, including in Artistotle, who is quoted (on Wikipedia) as saying

"We may assume the superiority ceteris paribus [other things being equal] of the demonstration which derives from fewer postulates or hypotheses."

Another saying that is similar is "do not multiple entities beyond necessity". In statistics, this tends to get interpreted as "use as few parameters as possible", although one could use the reasoning to prefer non-parametric methods.

Model selection

The use of Ockham's razor is also sometimes called the principle of parsimony. This has also been used extensively in evolutionary theory by a method literally called parsimony. The idea there is something like this: if a feature or trait is difficult to evolve, then it is (often) better to assume an evolutionary tree for which the trait only evolves once (or as few times as possible) rather than a tree that requires a trait to have evolved multiple times. Here Ockham's razor is interpreted to mean something like "do not multiply mutations beyond necessity".

The method doesn't always work well. For example, winged flight (bats, birds, insects), echolocation (bats and whales), bioluminescence, and fingerprints (koalas and humans). Although it would be simpler in some ways to have an evolutionary tree where these things arose once, there is a lot of genetic evidence to suggest otherwise in some cases. The moral is that simpler models are not always better.

Model selection

The idea that simpler models are better seems to me more a philosophical idea than statistical. Philosophers of science try to think about what makes good scientific theories and hypotheses, and usually the list includes things like

- simplicity (i.e., parsimony)
- predictive ability
- conservativism (or coherence with existing theory)
- verifiability/falsifiability
- fruitfulness (i.e., leads to more theories and hypothesis)
- accuracy–(being true, and able to account for existing evidence)
- precision–(making predictions that are as exact as possible).
- not being ad hoc

Model selection

Often, philosophers are interested in big scientific theories, such as Copernicus's sun-centered solar system versus earth-centered solar system models, Darwin's theory of natural selection, Freud's theories about the subconscious, Relativity, etc.

In statistics, our goals are usually more modest, and often we are not looking for models that are literally true. We are usually quite happy with models that find relationships between variables that are approximately correct and that find trends in the data rather than exact relationships. A famous saying from the statistician George Box is

"All models are wrong, but some are useful"

Here usefulness might mean that we can make predictions that help us plan for the future, or that we can be convinced that certain variables are more important than others for understanding things like graduation rates.

Model selection

For model selection in statistics, we're mostly interested in finding variables that are most predictive of the response variable, and leaving those in the model, while eliminating variables that are less useful for predicting the repsonse variable.

Rather than big, philosophical motivations, this is often motivated by some practical reasons. Here are some:

- models with lots of predictors (especially interactions) are harder to interpret
- models with lots of predictors will tend to have larger confidence intervals for their estimates
- models with too many predictors can be "overfitted"–they account for the current data but are unlikley to generalize well to future data sets
- often we have more predictors than observations!
- often predictors are very closely related, and so have redundant information (collinearity, more on this later)

Model selection

There are different strategies for dealing with model selection. A nice one to use if you don't have too many predictors is called **backward elimination**. The idea is to start with all variables that could potentially be used as predictors (all variables available).

Once you fit the model with all variables, you decide whether to accept the model or delete one of the variables from the model. Criteria for choosing the variable to delete include using the variable with the highest p-value (if it is above a minimum threshold), or choosing the variable that would have the minimum impact on adjusted R^2 if deleted. Once you delete a variable, you fit the model again with the reduced set of variables, and repeat the procedure (either accept the model or find another variable to delete). You repeat the process over and over until you have a model where all variables meet the threshold where they should be retained.

Model selection

We'll use an example which has more predictors than previous data sets we've used. The example is for salaries at a small college in the 1970s and compares salaries of male (0) versus female (1) professors, and includes variables for their rank (assistant, associate, full), number of years in current rank, degree (1=doctorate or 0=other), and yd for years since highest degree completed.

```
> x <- read.table("salary.dat",header=T)
> head(x)
  id sex rank year degree yd salary
1  1   0    3   25      1 35  36350
2  2   0    3   13      1 22  35350
3  3   0    3   10      1 23  28200
4  4   1    3    7      1 27  26775
5  5   0    3   19      0 30  33696
6  6   0    3   16      1 21  28516
```

Model selection

Although this data set is old, comparing pay for men versus women is still quite a timely topic. A recent paper discusses gender pay differences for Uber drivers:

Discussion on podcast: http://one.npr.org/i/583678276:583678278

Paper: http://www.math.unm.edu/~james/STAT428/Uber.pdf

This data set had over 1.8 million drivers, and who knows how many variables. The goal of the researchers was not so much to determine the incomes of the drivers (dollars per hour, which is the response), but rather to see if there were differences in pay (was gender a significant predictor of pay?), and whether gender could be made insignificant by accounting for other variables (time of service, experience of drivers, speed of drivers, etc.)

Model selection

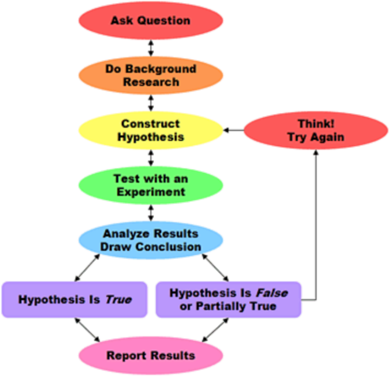
Some interesting features of their data are that the raw data would consist of one row per drive, rather than one row per driver, so that the size of the data set would be enormous. If the average Uber driver gave 100 rides over the data collection period (something like 2 years), the data set would have 180 million rows. Note that Excel has a maximum of $2^{20} = 1,048,576$ rows for a single spreadsheet.

Explanatory variables could have included driver GPS coordinates (latitude and longitude) for place of pickup, GPS coordinates for drop off, number of passengers, number of miles driven, time of pickup, time of drop off, date, day of week, CC information of the passenger (and whatever variables they can get from that), fare paid, plus variables associated with the driver such as age, sex, time that they started working for Uber, year, make, and model of the car. Researchers would have also wanted to determine things like type of locations: airport, business, residential.

Model selection

The topic of big data deals with very large data sets like these. What if there is too much to load into R (I think R would struggle with this one). Uber data will be small compared to say, Amazon.com, or Medicare. What if the data doesn't even fit on one computer?

Many of the variables might not have been relevant for the study questions, but many data sets are like this. Lots of data is collected, then questions about the data are asked later. This reverses the usual high school science fair presentation of the scientific method
http://astro1.panet.utoledo.edu/~ljc/ScientificMethod.htm



Model selection

To go back to our smaller data set, we have only 52 observations. The predictors are sex, rank, year since attaining rank, an indicator for doctorate degree, and year since highest degree, so there are five predictors. It's not necessary to convert binary variables to factors, but this can be done anyway to make sure that one category is the baseline. Thus, full professor here is made the baseline.

```
> x$sex <- factor(x$sex,labels=c("Male","Female"))
> x$degree <- factor(x$degree,
labels=c("Other", "Doctorate"))
> faculty$rank <- factor(faculty$rank , levels=c(3,2,1),
label=c("Full","Assoc","Assist"))
> head(x)
  id    sex rank year    degree yd salary
1  1   Male    3   25 Doctorate 35  36350
2  2   Male    3   13 Doctorate 22  35350
3  3   Male    3   10 Doctorate 23  28200
4  4 Female    3    7 Doctorate 27  26775
5  5   Male    3   19     Other 30  33696
6  6   Male    3   16 Doctorate 21  28516
```

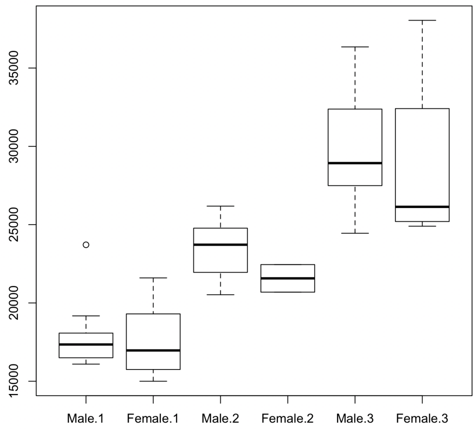
Note that the distribution of ranks appears to be different for male versus female professors

```
> attach(x)
> table(sex,rank)
        rank
sex       1  2  3
  Male   10 12 16
  Female  8  2  4

> chisq.test(table(sex,rank))
Pearson's Chi-squared test
X-squared = 4.4323, df = 2, p-value = 0.109

Warning message:
In chisq.test(table(sex, rank)) :
  Chi-squared approximation may be incorrect
```

Also note that male professors had a sligtly lower proportion of doctorates.

```
> table(sex,degree)
        degree
sex      Other Doctorate
  Male      14        24
  Female     4        10
> 24/38
[1] 0.6315789
> 10/14
[1] 0.7142857
```

Here is a boxplot of salary against combinations of sex and faculty rank. It doesn't adjust for years of experience or years since highest degree.



Model selection

The full model, allowing for two-way interactions only, is:

```
> m1 <- lm(salary ~ sex + degree + rank + year +
yd + sex*degree + sex*rank + sex*year + sex*yd +
degree*rank + degree*year + degree*yd + rank*year +
rank*yd + year*yd)
```

For a model with p predictors, the number of possible two-way interactions is $\binom{p}{2} = p(p-1)/2$. For 5 predictors, there are $(5)(4)/2 = 10$ possible interactions. For 10 predictors, there would be 45 possible interactions.

```
> Anova(m1,type=3)
Anova Table (Type III tests)

Response: salary
               Sum Sq Df F value  Pr(>F)
(Intercept)  22605087  1  3.6916 0.06392 .
sex           4092995  1  0.6684 0.41984
degree        4137628  1  0.6757 0.41735
rank          5731837  2  0.4680 0.63059
year          2022246  1  0.3302 0.56966
yd            3190911  1  0.5211 0.47578
sex:degree    7164815  1  1.1701 0.28773
sex:rank       932237  2  0.0761 0.92688
sex:year      7194388  1  1.1749 0.28676
sex:yd        2024210  1  0.3306 0.56947
degree:rank  13021265  2  1.0632 0.35759
degree:year   4510249  1  0.7366 0.39735
degree:yd     6407880  1  1.0465 0.31424
rank:year     1571933  2  0.1284 0.88001
rank:yd       9822382  2  0.8020 0.45750
```

Model selection

Here we'll use backward elimination using p-value as the criterion. The idea is to first consider removing interactions. Remove the interaction with the highest p-value greater than $\alpha = .05$. You can also remove main effects if they have higher p-values than any interactions **and** are not involved in any interactions.

At this first step, the interaction with the highest p-value is year with yd. An interaction here would have meant that the effect of year in rank would depend on the number of years since graduating.

```
> m2 <- lm(salary ~ sex + degree + rank + year +
yd + sex*degree + sex*rank + sex*year + sex*yd +
degree*rank + degree*year + degree*yd + rank*year +
rank*yd)
```

```
> Anova(m2,type=3)
Anova Table (Type III tests)

Response: salary
                Sum Sq Df F value  Pr(>F)
(Intercept)   26986124  1  4.5480 0.04073 *
sex            4442691  1  0.7487 0.39332
degree         4089226  1  0.6892 0.41260
rank           6079684  2  0.5123 0.60394
year           7029024  1  1.1846 0.28455
yd             3912094  1  0.6593 0.42280
sex:degree     7341235  1  1.2372 0.27429
sex:rank        907205  2  0.0764 0.92657
sex:year       7178186  1  1.2097 0.27959
sex:yd         2152917  1  0.3628 0.55118
degree:rank   13240859  2  1.1157 0.34008
degree:year    4601976  1  0.7756 0.38506
degree:yd      6443383  1  1.0859 0.30519
rank:year      1930802  2  0.1627 0.85054
rank:yd        9944911  2  0.8380 0.44184
Residuals    189876375 32
```

Here the sex by rank interaction had the highest p-value, so we remove it.

```
> m2 <- lm(salary ~ sex + degree + rank + year +
yd + sex*degree + sex*year + sex*yd +
degree*rank + degree*year + degree*yd + rank*year +
rank*yd)
```

```
> Anova(m3,type=3)
Anova Table (Type III tests)

Response: salary
               Sum Sq Df F value Pr(>F)
(Intercept)  37666808  1  6.7127 0.0140 *
sex          12952041  1  2.3082 0.1379
degree        3814698  1  0.6798 0.4154
rank          8196244  2  0.7303 0.4892
year         14777996  1  2.6336 0.1139
yd            4812803  1  0.8577 0.3609
sex:degree   10640012  1  1.8962 0.1775
sex:year     10690026  1  1.9051 0.1765
sex:yd        3614221  1  0.6441 0.4278
degree:rank  16341405  2  1.4561 0.2473
degree:year   4894265  1  0.8722 0.3569
degree:yd     6719487  1  1.1975 0.2815
rank:year     5037089  2  0.4488 0.6421
rank:yd      15110673  2  1.3465 0.2737
Residuals   190783580 34
```

Here the rank by year interaction had the highest p-value, so we remove it.

```
> m3 <- lm(salary ~ sex + degree + rank + year +
yd + sex*degree + sex*year + sex*yd +
degree*rank + degree*year + degree*yd + rank*yd)
```

```
> Anova(m3,type=3)
Anova Table (Type III tests)

Response: salary
               Sum Sq Df F value   Pr(>F)
(Intercept)  56455344  1 10.3788 0.002705 **
sex          13042634  1  2.3978 0.130255
degree        5336283  1  0.9810 0.328555
rank          8406030  2  0.7727 0.469276
year         12790031  1  2.3513 0.133918
yd            9295736  1  1.7089 0.199411
sex:degree   12831931  1  2.3590 0.133302
sex:year     13646799  1  2.5089 0.121955
sex:yd        2456466  1  0.4516 0.505866
degree:rank  21836322  2  2.0072 0.149124
degree:year   7414066  1  1.3630 0.250690
degree:yd     9232872  1  1.6974 0.200903
rank:yd      41051000  2  3.7734 0.032525 *
Residuals   195820669 36
```

Here the sex by yd interaction had the highest p-value, so we remove it.

```
> m4 <- lm(salary ~ sex + degree + rank + year +
yd + sex*degree + sex*year +
degree*rank + degree*year + degree*yd + rank*yd)
```

```
> Anova(m4,type=3)
Anova Table (Type III tests)

Response: salary
               Sum Sq Df F value   Pr(>F)
(Intercept)  54336558  1 10.1396 0.002941 **
sex          10838535  1  2.0226 0.163354
degree        5696946  1  1.0631 0.309204
rank         10610665  2  0.9900 0.381199
year         10334602  1  1.9285 0.173225
yd           13494052  1  2.5181 0.121057
sex:degree   10394382  1  1.9397 0.172017
sex:year     22789419  1  4.2527 0.046263 *
degree:rank  21157939  2  1.9741 0.153243
degree:year   8497324  1  1.5857 0.215833
degree:yd     9463400  1  1.7659 0.192023
rank:yd      42516602  2  3.9670 0.027486 *
Residuals   198277134 37
```

Here rank had the highest p-value, but it is involved in some interactions, so we don't consider removing it. Degree has the second highest, but again is involved in interactions. The third highest is degree by year, with a p-value of 0.21, so we remove it.

```
> m5<- lm(salary ~ sex + degree + rank + year +
yd + sex*degree + sex*year +
degree*rank  + degree*yd + rank*yd)
```

Next we'll remove degree by yd.

```
> Anova(m5,type=3)
Anova Table (Type III tests)
Response: salary
               Sum Sq Df F value    Pr(>F)
(Intercept)  77962216  1 14.3275 0.0005312 ***
sex           3548444  1  0.6521 0.4243835
degree        1652083  1  0.3036 0.5848523
rank          5984927  2  0.5499 0.5815072
year         81988541  1 15.0675 0.0004005 ***
yd            6103883  1  1.1217 0.2962298
sex:degree    3189136  1  0.5861 0.4486666
sex:year     14489584  1  2.6628 0.1109792
degree:rank  13515717  2  1.2419 0.3002849
degree:yd     1695058  1  0.3115 0.5800292
rank:yd      34725539  2  3.1908 0.0523619 .
Residuals   206774458 38
```

Next we'll remove sex by degree

```
> Anova(m6,type=3)
Anova Table (Type III tests)

Response: salary
               Sum Sq Df F value    Pr(>F)
(Intercept) 252985654  1 47.3280 3.138e-08 ***
sex           2656144  1  0.4969 0.4850519
degree          26167  1  0.0049 0.9445786
rank          4326190  2  0.4047 0.6699681
year         80806360  1 15.1171 0.0003821 ***
yd            5098991  1  0.9539 0.3347463
sex:degree    2505272  1  0.4687 0.4976433
sex:year     12832093  1  2.4006 0.1293665
degree:rank  15741805  2  1.4725 0.2418287
rank:yd      38135455  2  3.5671 0.0377828 *
Residuals   208469515 39
```

Next we'll remove degree by rank

```
> Anova(m7,type=3)
Anova Table (Type III tests)

Response: salary
               Sum Sq Df F value    Pr(>F)
(Intercept) 252298089  1 47.8347 2.453e-08 ***
sex            486921  1  0.0923 0.7628253
degree         179478  1  0.0340 0.8545786
rank          6294899  2  0.5967 0.5554288
year         92097669  1 17.4614 0.0001546 ***
yd            4252203  1  0.8062 0.3746187
sex:year     11377954  1  2.1572 0.1497226
degree:rank  14519997  2  1.3765 0.2641686
rank:yd      38113373  2  3.6131 0.0361035 *
Residuals   210974787 40
```

At this point, degree is not involved in any interactions, so we can consider removing it. It has a higher p-value than the two remaining interaction terms, so we remove degree as a main effect.

```
> Anova(m8,type=3)
Anova Table (Type III tests)

Response: salary
               Sum Sq Df F value    Pr(>F)
(Intercept) 482851531  1 89.9345 5.335e-12 ***
sex            936435  1  0.1744 0.6783426
degree        8902098  1  1.6581 0.2049131
rank         91805630  2  8.5497 0.0007673 ***
year        101743686  1 18.9505 8.422e-05 ***
yd             640363  1  0.1193 0.7315491
sex:year     14134386  1  2.6326 0.1121718
rank:yd      24905278  2  2.3194 0.1108009
Residuals   225494784 42
```

Now we remove sex by year.

```
> Anova(m9,type=3)
Anova Table (Type III tests)

Response: salary
               Sum Sq Df  F value    Pr(>F)
(Intercept) 657912109  1 120.6937 4.606e-14 ***
sex           1311737  1   0.2406 0.6262400
rank         91215249  2   8.3667 0.0008529 ***
year         92989960  1  17.0590 0.0001638 ***
yd            6925991  1   1.2706 0.2659107
sex:year     11545391  1   2.1180 0.1528391
rank:yd      27221003  2   2.4968 0.0942138 .
Residuals   234396882 43
```

If doing automated selection, we would next remove sex. However, that was the research question. What to do depends on your research goals. Are you looking for the parsimonious model? or are you looking for the most parsimonious model that includes sex as a predictor? You can also fit several models (both most parsimonious and most parsimonious with sex). You think about removing the rank by yd interaction and then seeing whether sex should still be obtained.

```
> Anova(m10,type=3)
Anova Table (Type III tests)

Response: salary
               Sum Sq Df  F value    Pr(>F)
(Intercept) 682341395  1 122.0734 2.822e-14 ***
sex           5552916  1   0.9934 0.3243537
rank        122529231  2  10.9605 0.0001372 ***
year        106510254  1  19.0551 7.587e-05 ***
yd            4472402  1   0.8001 0.3759217
rank:yd      23603682  2   2.1114 0.1331705
Residuals   245942272 44
```

```
> Anova(m11,type=3)
Anova Table (Type III tests)

Response: salary
                Sum Sq Df  F value    Pr(>F)
(Intercept) 2518702345  1 429.8351 < 2.2e-16 ***
sex            5132365  1   0.8759    0.3542
rank         479020588  2  40.8742 6.270e-11 ***
year         134188974  1  22.9003 1.799e-05 ***
yd             5142131  1   0.8775    0.3538
Residuals    269545954 46
```

```
> Anova(m11,type=3)
Anova Table (Type III tests)

Response: salary
                Sum Sq Df  F value    Pr(>F)
(Intercept) 2518702345  1 429.8351 < 2.2e-16 ***
sex            5132365  1   0.8759    0.3542
rank         479020588  2  40.8742 6.270e-11 ***
year         134188974  1  22.9003 1.799e-05 ***
yd             5142131  1   0.8775    0.3538
Residuals    269545954 46
```

```
> Anova(m12,type=3)
Anova Table (Type III tests)

Response: salary
                Sum Sq Df  F value    Pr(>F)
(Intercept) 3585257969  1 613.4490 < 2.2e-16 ***
sex            2304648  1   0.3943    0.5331
rank         634005385  2  54.2402 6.165e-13 ***
year         157183229  1  26.8945 4.473e-06 ***
Residuals    274688086 47
```

We might also look at what happens if we only use sex as a predictor.

```
> Anova(m13,type=3)
                 Sum Sq Df F value Pr(>F)
(Intercept) 2.3177e+10  1 693.260 <2e-16 ***
sex         1.1411e+08  1   3.413 0.0706 .
Residuals   1.6716e+09 50

> t.test(salary ~ sex,var.equal=TRUE)

Two Sample t-test

data:  salary by sex
t = 1.8474, df = 50, p-value = 0.0706
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -291.257 6970.550
sample estimates:
  mean in group Male mean in group Female
            24696.79             21357.14
```

If you don't assume equal variances in the t.test:

```
> t.test(salary ~ sex)

Welch Two Sample t-test

data:  salary by sex
t = 1.7744, df = 21.591, p-value = 0.09009
```

Although the t-test isn't significant at the .05 level, by having a p-value less than .10, you can say that there is some evidence (although not strong) of a difference in salaries. The evidence is much weaker when rank and year are taken into account. To see the effect of sex, use `summary(m13)`

```
> summary(m12)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25390.65    1025.14  24.768  < 2e-16 ***
sexFemale     524.15     834.69   0.628    0.533
rankAssoc   -5109.93     887.12  -5.760 6.20e-07 ***
rankAssist  -9483.84     912.79 -10.390 9.19e-14 ***
year          390.94      75.38   5.186 4.47e-06 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 2418 on 47 degrees of freedom
Multiple R-squared:  0.8462,Adjusted R-squared:  0.8331
F-statistic: 64.64 on 4 and 47 DF,  p-value: < 2.2e-16
```

Note that in the model, the baseline salary is for male full professors. The model therefore predicts that being an assistant professor reduces salary by an average of \$9483.84, that being an associate professor reduces the salary by \$5109.93 (compare to a full professor), and that being female increases the salary by \$524.15. On average, female professors made \$3339.68 dollars less (you can see this from the t-test output). However, based on the model, this is accounted for female professors tending to be younger (in academic age–years since highest degree) and having lower rank. For example, 42% of male professors were full professors, while 28.5% of full professors were female, and full professors tend to get paid more than other ranks.

Based on the results, can you conclude that there is no discrimination against female professors in terms of salary?

No. There could be a number of explanations for the patterns in the data. It could be that male professors at this university tend to be older and therefore have had to more time to be promoted in terms of academic rank. On the other hand, it could be that male professors are promoted more easily, and this leads to them having higher ranks. Adjusting for academic rank might therefore might sweep some things under the rug that are due to a form of discrimination.

Another variable not accounted for in the data is the department that the professors are from. STEM fields and business, for example, tend to pay better than humanities subjects at US universities. Where I worked in New Zealand, every professor at the same academic rank and grade within rank got the same pay, regardless of department, so this would not have been an issue there. However, it was still probably easier to get promoted more quickly in STEM fields than non-STEM fields.

What is tricky in statistics, particularly in observational studies, is knowing whether you have accounted for the relevant variables. To give another example outside of the regression/ANOVA setting, consider the voting records for the Civil Rights Act of 1964. Sometimes republicans claim to have had a better voting record (i.e., higher proportion voting in favor of the act) for the Civil Rights Act than did democrats. Is this true? Here are the raw numbers for the House of Representatives (data from Wikipedia):

Party	Yes	No
Democrat	152 (61%)	96 (39%)
Republican	138 (80.2%)	34 (19.8%)

Overall, a higher proportion of republicans voted for the Civil Rights Act than did democrats. However, there were different voting patterns in Southern versus other states.

Party	Yes	No
Democrat, Southern	7 (7%)	94 (93%)
Democrat, Other	145 (94%)	9 (6%)
Republican, Southern	0 (0%)	10 (100%)
Republican, Other	138 (85%)	24 (15%)

This might seem paradoxical: republicans were more likely to favor the Act than democrats overall, but Southern republicans were less likely to than Southern democrats, and non-Southern republicans were less likely to than non-Southern democrats. This is an example of something called **Simpson's paradox**, where the relationships between two variables seem to be reversed when a third variable is taken into account.

Getting back to model selection, we illustrated the idea of **backward elimination** as one technique for model selection. Other standard techniques are **forward selection** and **stepwise addition**.

In backward elimination, a full model is constructed, and then predictor variables (or interactions) are eliminated one by one until a final model is obtained.

In forward selection, we start with an intercept-only model, then add variables one at a time, adding more significant variables first, and only adding a new variable if it significantly improve the model.

The stepwise method tries to use advantages of both forward and backward methods. In the forward method, once a variable is included, it can never be removed, even though it might turn out to be redundant once other variables are in the model. Thus, at each step, you can either add a new variable or delete a variable, depending on what most improves the model. Eventually you reach a point where the model cannot be improved by either adding or removing variables.

A final method is called **best subsets regression**. You consider all possible subsets of predictors, and pick the model that is best according to some criterion. This method is feasible for small to moderate numbers of predictors. The number of subsets, only considering main effects (not considering interactions), is 2^p, where p is the number of predictors. For five predictors, you would therefore consider $2^5 = 32$ models. For 10 predictors, you would have to consider $2^{10} > 1000$ models, and for 20 predictors, there are over 1 million possible models.

In addition to the method (backward, forward, stepwise, best subsets), you have to pick a criterion by which to compare models and determine whether one model is significantly better than another. In the backward elimination example done earlier, we used p-values as a criterion. However, other criteria are possible, such as adjusted R^2, Mallow's C_p, AIC (Akaike Information criterion) and $BIC(Bayesianinformationcriterion)$. These choices are essentially independent of the method (backward, forward, stepwise, best subsets). With so many ways to do model selection, the "best" model chosen can depend on these choices, and there often isn't a clear answer to what model is best.

There are other more recent methods as well for doing model selection as well, including the lasso (least absolute shrinkage and selection operator), and cross-validation.

To discuss some of these alternative criteria for model selection, Mallow's C_p is

$$\frac{SSE_p}{\widehat{\sigma}^2_{\text{FULL}}} - N + 2p$$

where SSE_p is the sum of squared error on the model with p predictors, $\widehat{\sigma}^2_{\text{FULL}}$ is the mean square error for the full model, N is the sample size, and p is the number of predictors. A model is better if it has lower C_p, so you can think of the $2p$ term as penalizing having more parameters.

The AIC and BIC criteria are similar in that they penalize extra parameters, and smaller AIC/BIC values indicate preferred models. Here

$$AIC = -2 \log L + 2p$$

$$BIC = -2 \log L + p \log n$$

where $\log L$ is the log-likelihood, related to the probability of observing data similar to what is observed under the model. BIC tends to have a stronger penalty for more parameters (especially for larger sample sizes) than AIC, so tends to prefer fewer predictors.

Note that in forward and backward methods, two consecutively considered models are related by setting one of the parameters equal to 0 or nonzero. In this case, the models are **nested**, meaning that one model has predictors that are subsets of the other. Testing whether the coefficient is equal to 0 is therefore equivalent to testing whether the fuller model is significantly better than the reduced model.

For best subsets regression, we have to compare models that aren't necessarily nested within each other. Criteria such as AIC and BIC can be used to compare models that are based on different predictors and don't have to be nested.

We'll illustrate these approaches using the salary data. The function `step()` carry's out automated model selection using AIC by default. I'll include one possible interaction for the forward selection to possibly test, although interactions won't be significant.

```
> m.empty <- lm(salary ~ 1)
> m.forward <- step(m.empty,salary ~ sex + rank+ year +
degree + yd + rank*year, direction="forward")
```

```
> m.forward <- step(m.empty,salary ~ sex + rank + degree + year +
Start:  AIC=904.3
salary ~ 1
         Df  Sum of Sq        RSS    AIC
+ rank    2 1346783800  438946058 835.33
+ year    1  876680907  909048951 871.19
+ yd      1  813271618  972458240 874.69
+ sex     1  114106220 1671623638 902.86
<none>                 1785729858 904.30
+ degree  1    8681649 1777048209 906.04

Step:  AIC=835.33
salary ~ rank
         Df Sum of Sq       RSS    AIC
+ year    1 161953324 276992734 813.39
+ yd      1  23162091 415783967 834.51
<none>                438946058 835.33
+ degree  1  10970082 427975976 836.01
+ sex     1   7074743 431871315 836.48
```

```
Step:  AIC=813.39
salary ~ rank + year

              Df Sum of Sq       RSS    AIC
<none>                     276992734 813.39
+ rank:year    2  15215454 261777280 814.45
+ yd           1   2314414 274678320 814.95
+ sex          1   2304648 274688086 814.95
+ degree       1   1127718 275865016 815.18
```

The best model based on forward selection and AIC as the criterion is salary $=$ rank $+$ year. The model is

```
> m.forward

Call:
lm(formula = salary ~ rank + year)

Coefficients:
(Intercept)        rank2        rank3         year
    16203.3       4262.3       9454.5        375.7
```

Now we'll look at what happens with backward model selection. Here we'll start with the full model and all interaction terms.

```
m.backward <- step(m2,salary ~ sex + rank + degree + year
+ yd + sex*rank + sex*degree + sex*year + sex*yd +
rank*degree + rank*year + rank*yd + degree*year +
degree*yd + year*yd,direction="backward")
```

```
Start:  AIC=822
salary ~ sex + degree + rank + year + yd + sex * degree + sex *
    year + sex * yd + degree * rank + degree * year + degree *
    yd + rank * year + rank * yd

              Df Sum of Sq       RSS    AIC
- rank:year    2   5037089 195820669 819.36
- sex:yd       1   3614221 194397801 820.98
- degree:year  1   4894265 195677844 821.32
- degree:yd    1   6719487 197503067 821.80
- rank:yd      2  15110673 205894253 821.96
<none>                     190783580 822.00
- degree:rank  2  16341405 207124985 822.27
- sex:degree   1  10640012 201423592 822.82
- sex:year     1  10690026 201473606 822.84
```

```
Step:  AIC=819.36
salary ~ sex + degree + rank + year + yd + sex:degree + sex:year +
    sex:yd + degree:rank + degree:year + degree:yd + rank:yd

              Df Sum of Sq       RSS    AIC
- sex:yd       1   2456466 198277134 818.00
- degree:year  1   7414066 203234734 819.29
<none>                     195820669 819.36
- degree:yd    1   9232872 205053541 819.75
- sex:degree   1  12831931 208652600 820.66
- degree:rank  2  21836322 217656990 820.85
- sex:year     1  13646799 209467467 820.86
- rank:yd      2  41051000 236871669 825.25
```

The algorithm stops here because no way of eliminating an interaction reduces the AIC. The algorithm is "greedy" in the sense that it only looks one step ahead. It will only continue if eliminating one term will reduce AIC. If you need to eliminate two terms to reduce AIC, the algorithm will not see this and will get stuck. We know from forward selection that there are smaller models with lower AIC.

```
Step:  AIC=818
salary ~ sex + degree + rank + year + yd + sex:degree +
 sex:year + degree:rank + degree:year + degree:yd + rank:yd

              Df Sum of Sq       RSS    AIC
<none>                     198277134 818.00
- degree:year  1   8497324 206774458 818.19
- degree:yd    1   9463400 207740534 818.43
- sex:degree   1  10394382 208671516 818.66
- degree:rank  2  21157939 219435073 819.28
- sex:year     1  22789419 221066553 821.66
- rank:yd      2  42516602 240793736 824.11
```

The BIC criterion penalizes larger models more, so we can check what happens in this case. here you need a parameter in the `step()` function that gives the log of the sample size. Here $\log(52) = 3.951244$. Thus $AIC = -2 \log L + 2p$, $BIC \approx -2 \log L + 3.95p$ for this sample size.

```
> m.backward2 <- step(m2,salary ~ sex + rank + degree
+ year + yd + sex*rank + sex*degree + sex*year + sex*yd
 + rank*degree + rank*year + rank*yd + degree*year +
 degree*yd + year*yd,direction="backward",k=log(52))
```

```
Start:  AIC=857.12
salary ~ sex + degree + rank + year + yd + sex * degree + sex *
    year + sex * yd + degree * rank + degree * year + degree *
    yd + rank * year + rank * yd

              Df Sum of Sq       RSS    AIC
- rank:year    2   5037089 195820669 850.58
- rank:yd      2  15110673 205894253 853.18
- degree:rank  2  16341405 207124985 853.49
- sex:yd       1   3614221 194397801 854.15
- degree:year  1   4894265 195677844 854.49
- degree:yd    1   6719487 197503067 854.97
- sex:degree   1  10640012 201423592 855.99
- sex:year     1  10690026 201473606 856.01
<none>                     190783580 857.12
```

```
Step:  AIC=850.58
salary ~ sex + degree + rank + year + yd + sex:degree + sex:year +
    sex:yd + degree:rank + degree:year + degree:yd + rank:yd

              Df Sum of Sq       RSS    AIC
- sex:yd       1   2456466 198277134 847.27
- degree:rank  2  21836322 217656990 848.17
- degree:year  1   7414066 203234734 848.56
- degree:yd    1   9232872 205053541 849.02
- sex:degree   1  12831931 208652600 849.93
- sex:year     1  13646799 209467467 850.13
<none>                     195820669 850.58
- rank:yd      2  41051000 236871669 852.57
```

```
Step:  AIC=847.27
salary ~ sex + degree + rank + year + yd + sex:degree + sex:year +
    degree:rank + degree:year + degree:yd + rank:yd

              Df Sum of Sq       RSS    AIC
- degree:rank  2  21157939 219435073 844.64
- degree:year  1   8497324 206774458 845.50
- degree:yd    1   9463400 207740534 845.75
- sex:degree   1  10394382 208671516 845.98
<none>                     198277134 847.27
- sex:year     1  22789419 221066553 848.98
- rank:yd      2  42516602 240793736 849.47
```

```
Step:  AIC=844.64
salary ~ sex + degree + rank + year + yd + sex:degree + sex:year +
    degree:year + degree:yd + rank:yd

              Df Sum of Sq       RSS    AIC
- degree:yd    1    361929 219797002 840.78
- degree:year  1    855102 220290175 840.89
- sex:degree   1   1616150 221051223 841.07
- rank:yd      2  24391011 243826084 842.22
- sex:year     1  10569795 230004869 843.14
<none>                     219435073 844.64
```

```
Step:  AIC=840.78
salary ~ sex + degree + rank + year + yd + sex:degree + sex:year +
    degree:year + rank:yd

              Df Sum of Sq       RSS    AIC
- sex:degree   1   3112507 222909509 837.56
- degree:year  1   4414318 224211320 837.86
- rank:yd      2  24695126 244492128 838.41
- sex:year     1  16645026 236442028 840.62
<none>                     219797002 840.78

Step:  AIC=837.56
salary ~ sex + degree + rank + year + yd + sex:year + degree:year +
    rank:yd

              Df Sum of Sq       RSS    AIC
- degree:year  1   2585275 225494784 834.21
- rank:yd      2  25367664 248277174 835.26
- sex:year     1  14770974 237680484 836.94
<none>                     222909509 837.56
```

```
Step:  AIC=834.21
salary ~ sex + degree + rank + year + yd + sex:year + rank:yd

           Df Sum of Sq       RSS    AIC
- rank:yd   2  24905278 250400062 831.75
- degree    1   8902098 234396882 832.27
- sex:year  1  14134386 239629170 833.42
<none>                  225494784 834.21

Step:  AIC=831.75
salary ~ sex + degree + rank + year + yd + sex:year

           Df Sum of Sq       RSS    AIC
- sex:year  1   8458303 258858365 829.53
- degree    1  11217823 261617885 830.08
- yd        1  16309342 266709404 831.08
<none>                  250400062 831.75
- rank      2 406263292 656663354 873.98
```

```
Step:  AIC=829.53
salary ~ sex + degree + rank + year + yd

         Df Sum of Sq       RSS    AIC
- sex     1   9134971 267993336 827.38
- degree  1  10687589 269545954 827.68
- yd      1  14868158 273726523 828.48
<none>                258858365 829.53
- year    1 144867403 403725768 848.69
- rank    2 399790682 658649047 870.19

Step:  AIC=827.38
salary ~ degree + rank + year + yd

         Df Sum of Sq       RSS    AIC
- degree  1   6684984 274678320 824.71
- yd      1   7871680 275865016 824.93
<none>                267993336 827.38
- year    1 147642871 415636208 846.25
- rank    2 404108665 672102002 867.29
```

```
Step:  AIC=824.71
salary ~ rank + year + yd

       Df Sum of Sq       RSS    AIC
- yd    1   2314414 276992734 821.19
<none>              274678320 824.71
- year  1 141105647 415783967 842.32
- rank  2 478539101 753217421 869.26

Step:  AIC=821.19
salary ~ rank + year

       Df Sum of Sq       RSS    AIC
<none>              276992734 821.19
- year  1 161953324 438946058 841.18
- rank  2 632056217 909048951 875.09
```

We see that backward seelction with BIC lead to the same model as forward selection with AIC. Using forward selection with BIC also leads to the same model (no interactions and only rank and year as predictors). The same is true with forward selection where all interactions are allowed.

```
> m.both <- step(m.empty,salary ~ sex + rank + degree
+ year + yd + sex*rank + sex*degree + sex*year +
sex*yd + rank*degree + rank*year + rank*yd +
degree*year + degree*yd +
year*yd,direction="both",k=log(52))
```

```
Start:  AIC=906.25
salary ~ 1

         Df  Sum of Sq        RSS    AIC
+ rank    2 1346783800  438946058 841.18
+ year    1  876680907  909048951 875.09
+ yd      1  813271618  972458240 878.60
<none>                 1785729858 906.25
+ sex     1  114106220 1671623638 906.76
+ degree  1    8681649 1777048209 909.95
```

```
Step:  AIC=841.18
salary ~ rank

         Df  Sum of Sq        RSS    AIC
+ year    1  161953324  276992734 821.19
<none>                  438946058 841.18
+ yd      1   23162091  415783967 842.32
+ degree  1   10970082  427975976 843.82
+ sex     1    7074743  431871315 844.29
- rank    2 1346783800 1785729858 906.25
```

```
Step:  AIC=821.19
salary ~ rank + year

             Df Sum of Sq       RSS    AIC
<none>                    276992734 821.19
+ yd          1   2314414 274678320 824.71
+ sex         1   2304648 274688086 824.71
+ degree      1   1127718 275865016 824.93
+ rank:year   2  15215454 261777280 826.16
- year        1 161953324 438946058 841.18
- rank        2 632056217 909048951 875.09
```

You can also summarize the sequence of model selection models by using extracting the sequence:

```
> m.backward2$anova
             Step Df   Deviance Resid. Df Resid. Dev      AIC
1                 NA         NA        34  190783580 857.1235
2     - rank:year  2  5037089.1        36  195820669 850.5761
3        - sex:yd  1  2456465.5        37  198277134 847.2731
4   - degree:rank  2 21157939.1        39  219435073 844.6430
5     - degree:yd  1   361928.9        40  219797002 840.7774
6    - sex:degree  1  3112507.1        41  222909509 837.5574
7   - degree:year  1  2585274.6        42  225494784 834.2058
8       - rank:yd  2 24905278.5        44  250400062 831.7509
9      - sex:year  1  8458302.6        45  258858365 829.5272
10          - sex  1  9134971.4        46  267993336 827.3794
11       - degree  1  6684983.5        47  274678320 824.7093
12           - yd  1  2314414.0        48  276992734 821.1944
```

To do best subset regression, we can use the `leaps` library. The following will give the two best models with up to 4 predictor variables.

```
> install.packages("leaps")
> library(leaps)
> m.subset <- regsubsets(salary ~ sex + rank
+ year + degree + yd,data=x,nvmax=6,nbest=3)
```

```
> summary(m.subset)
Subset selection object
Call: regsubsets.formula(salary ~ sex + rank + year + degree,
    +yd, data = x, nvmax = 6, nbest = 2)
6 Variables  (and intercept)
       Forced in Forced out
sex        FALSE      FALSE
rank2      FALSE      FALSE
...
2 subsets of each size up to 6
Selection Algorithm: exhaustive
         sex rank2 rank3 year degree yd
1  ( 1 ) " " " "   "*"   " "  " "    " "
1  ( 2 ) " " " "   " "   "*"  " "    " "
2  ( 1 ) " " " "   "*"   "*"  " "    " "
2  ( 2 ) " " "*"   "*"   " "  " "    " "
3  ( 1 ) " " "*"   "*"   "*"  " "    " "
3  ( 2 ) " " " "   "*"   "*"  " "    "*"
4  ( 1 ) " " "*"   "*"   "*"  " "    "*"
4  ( 2 ) "*" "*"   "*"   "*"  " "    " "
```

Because the function reports the best models of each number of predictors, the penalty term for the number of predictors doesn't matter–the best two models with three variables will be the same three models whether using AIC or BIC, for example. However, you can also get statistics for the models out of the function as follows. The minimum BIC is -81.1 which corresponds to the fifth model, which has rank2, rank3, and year. BIC here seems to be computed differently from the step() function.

```
> m.subset.summary <- summary(m.subset)
> names(m.subset.summary)
[1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"
 "outmat" "obj"
> m.subset.summary$bic
[1] -43.13556 -27.20706 -64.47238 -61.11298 -81.10177 -63.72225
-77.58684 -77.58499
```

Note that the different criteria will rank these 8 models (the two best with 1–4 predictors) nearly the same. BIC versus C_p and adjusted R^2 only differ by swapping the 4th and 5th best models. R^2, as opposed to adjusted R^2, will tend to favor larger models.

```
> rank(m.subset.summary$bic)
[1] 7 8 4 6 1 5 2 3
> rank(m.subset.summary$cp)
[1] 7 8 5 6 1 4 2 3
> rank(1-m.subset.summary$adjr2)
[1] 7 8 5 6 1 4 2 3
> rank(1-m.subset.summary$rsq)
[1] 7 8 5 6 3 4 1 2
```

You can also easily get regression coefficients for the different models. Here I get the coefficients for the first 5 listed models. Model 5 has the best BIC.

```
> coef(m.subset,1:5)
(Intercept)       rank3
  20134.344    9524.606

(Intercept)        year
 18166.1475    752.7978

(Intercept)       rank3        year
 17607.0603   7158.1499    459.5061

(Intercept)       rank2       rank3
  17768.667    5407.262   11890.283

(Intercept)       rank2       rank3        year
 16203.2682   4262.2847   9454.5232    375.6956
```

To understand the output, the package is turning the factor variable rank into 0/1 variables called dummy variables. When a factor variable has only two levels, you can treat this as an indicator variable. For example, the `degree` variable can only take two values. In the regression setting, this is equivalent to letting degree be a numeric value of either 0 or 1. The coefficient associated with degree gets multiplied by 0 for those without a doctorate and multiplied by 1 for those with a doctorate.

For a categorical variable with three levels (such as rank), regression creates dummy variables (with only 0/1) values as well. For a factor with k levels, the idea is to create $k-1$ 0/1 variables. Each of these variables is an indicator (i.e. 0/1) variable indicating whether or not that observation belongs to the particular category. In particular, we can represent rank being category 1, 2, or 3, by having a 0/1 variable for rank 2, and a 0/1 variable for rank 3.

We'll give an example of how to represent the data with dummy variables:

```
id sex rank year degree yd salary
19   0    2   10      0 15  22906
20   0    3    6      0 21  24450
21   0    1   16      0 23  19175
22   0    2    8      0 31  20525
----
id sex rank2 rank3 year degree yd salary
19  0      1     0    10  0   15 22906
20  0      0     1     6  0   21 24450
21  0      0     0    16  0   23 19175
22  0      1     0     8  0   31 20525
```

Now we'll show using the regression coefficients to predict salary for the model with rank, year, and rank*year interaction

```
id sex rank2 rank3 year degree yd salary
19  0     1     0    10   0   15 22906
20  0     0     1     6   0   21 24450
21  0     0     0    16   0   23 19175
22  0     1     0     8   0   31 20525

> mm2
Call:
lm(formula = salary ~ rank + year + rank * year)
Coefficients:
Coefficients:
(Intercept)     rank2     rank3      year rank2:year rank3:year
    16416.6    5354.2    8176.4     324.5     -129.7      151.2
```

Now we'll show using the regression coefficients to predict salary for the model with rank, year, and rank*year interaction. We can think of the model as

$$y = \beta_0 + \beta_1 \text{rank2} + \beta_2 \text{rank3} + \beta_3 \text{year} + \beta_4 \text{rank2*year} + \beta_5 \text{rank3*year}$$

$$\begin{aligned}
\widehat{y}_{19} &= \widehat{\beta}_0 + \widehat{\beta}_1 + \widehat{\beta}_3 \text{year} + \widehat{\beta}_4 \text{year} \\
\widehat{y}_{20} &= \widehat{\beta}_0 + \widehat{\beta}_2 + \widehat{\beta}_3 \text{year} + \widehat{\beta}_5 \text{year} \\
\widehat{y}_{21} &= \widehat{\beta}_0 + \widehat{\beta}_3 \text{year} \\
\widehat{y}_{22} &= \widehat{\beta}_0 + \widehat{\beta}_1 + \widehat{\beta}_3 \text{year} + \widehat{\beta}_4 \text{year}
\end{aligned}$$

Note that comparing say, associate and full professors, they have both different intercepts and different effects for year. The intercept for associate professors is $(\beta_0 + \beta_1)$, and the slope for the year is $(\beta_3 + \beta_4)$, whereas the intercept and slope for full professors are $(\beta_0 + \beta_2)$ and $(\beta_3 + \beta_5)$. For assistant professors, the intercept and slope are β_0 and β_3, respectively.

```
Coefficients:
(Intercept)      rank2      rank3       year rank2:year rank3:year
    16416.6     5354.2     8176.4      324.5     -129.7      151.2
```

To interpret the coefficients, salary is predicted to increase by \$324.5 for each year in the rank for assistant professors. For associate professors (rank 2), their salary is predicted to increase by $\$324.5 - \$129.7 = \$194.8$ for each year in their current rank. While for full professors, their salary is predicted to increase by $\$324 + \$151.2 = \$475.7$ for each year in their current rank.

In the model, because there is an interaction, the effect of being an associate versus a full professor by itself doesn't tell you directly the change in salary because it depends on the number of years in the rank.

Another package that does model selection is "MuMIn". To install, note that I is capitalized. This package uses AICc instead of AIC. AICc is used for small samples and involves a "correction" to AIC. Theoretically, these are based on approximations to information loss (from information theory), where AIC is a first-order approximation and AICc is a second-order approximation.

In general, the correction factor can depend on the model. For multiple linear regression, the correction is

$$AICc = AIC + \frac{2p^2 + 2p}{n - p - 1}$$

when comparing two models with the same number of parameters, AIC and AICc will rank the models the same. They can possibly rank models differently when comparing models with different numbers of parameters. For example, with $n = 52$ and $p = 5$ versus $p = 4$, we get

```
> n <- 52
> p <- 5
> (2*p^2+2*p)/(n-p-1)
[1] 1.304348
> p <- 4
> (2*p^2+2*p)/(n-p-1)
[1] 0.8510638
```

Consequently, the penalty for 5 versus 4 parameters is larger for AICc than for AIC. This makes it less likely that you would select a model with a smaller number of parameters using AICc than AIC.

Here m2 has the full model with all two-way interactions for the salary data. There are 480 models fitted!

```
> install.packages("MuMIn")
> library(MuMin)
> options(na.action=na.fail)
> models<-dredge(m2)
> models
Global model call: lm(formula = salary ~ sex + degree + rank + yea
    degree + sex * year + sex * yd + degree * rank + degree *
    year + degree * yd + rank * year + rank * yd)
---
Model selection table
19   16200   +            375.700          5 -476.480  964.3
27   16320   +  -34.3200  400.500          6 -476.261  966.4
23   15910   +            390.900          6 -476.262  966.4
1043 16420   +            324.500          7 -475.011  966.6
...
> dim(models)
[1] 480  19
```

The information saved in `models` above is hard to interpret. The left column is simply the id for the model that was fitted. The following helps

```
> summary(model.avg(get.models(models, subset=TRUE)))
                              df  logLik    AICc delta weight
2/5                            5 -476.48  964.26  0.00   0.17
2/4/5                          6 -476.26  966.39  2.13   0.06
2/3/5                          6 -476.26  966.39  2.13   0.06
2/5/11                         7 -475.01  966.57  2.30   0.05
...
> mm1 <- lm(salary ~ rank + year)
> AICc(mm1)
[1] 964.2634
> mm2 <- lm(salary ~ rank + year + rank*year)
> AICc(mm2)
[1] 966.5666
```

To list the variables, the function numbers them, in what appears to be alphabetical order. Effects 2 and 5 correspond to rank and year, while 11 corresponds to rank*year.

Notice that the difference in AICc values is 2.13 between the best model and second best. How big is this? Let AICc1 be the AICc for the best model and L1 be the likelihood for the best model. Here, I'll use the formula for AIC rather than AICc to get the idea of how big a difference of 2.0 is for AIC.

$$
\begin{aligned}
& AIC1 - AIC2 = 2 \\
& \Rightarrow (-2\log L1 + 2p1) - (-2\log L2 + 2p2) = 2 \\
& \Rightarrow (2\log L2 - 2\log L1) + 2(p1 - p2) = 2 \\
& \Rightarrow (2\log L2 - 2\log L1) + 2 = 2 \\
& \Rightarrow (2\log L2 - 2\log L1) = 4 \\
& \Rightarrow 2\log\left(\frac{L2}{L1}\right) = 4 \\
& \Rightarrow \frac{L2}{L1} = e^2 \approx 7.39 \\
& \Rightarrow L2 \approx 7.39L1
\end{aligned}
$$

This means that model 1 is preferred even though model L2 has higher likelihood by a factor of about 7.

Theoretically, for nested models like these, the statistic $2\log\left(\frac{L2}{L1}\right)$ has an approximate (large-sample) χ^2 distribution where the degrees of freedom is the difference in the number of parameters. This is called likelihood-ratio testing. The critical value for 1 degree of freedom is 3.84, so roughly a difference in AIC of 2 units (where the smaller model has an AIC of 2 units better than the larger model with one extra parameter is "statistically signficant". Usually statisticians don't mix significance testing with AIC-based model selection, however. Still, it is useful to think that a difference of 2 for AIC is somewhat substantial. Some authors use differences of 10 in AIC instead.

The difference in AIC is actually less than 2.0 for these models. From the likelihood ratio point of view, the difference being small suggests that the two models are not significantly different, in which case the simpler model should be preferred.

```
> AIC(mm1)
[1] 962.959
> AIC(mm2)
[1] 964.0212
```