Stat 412/512

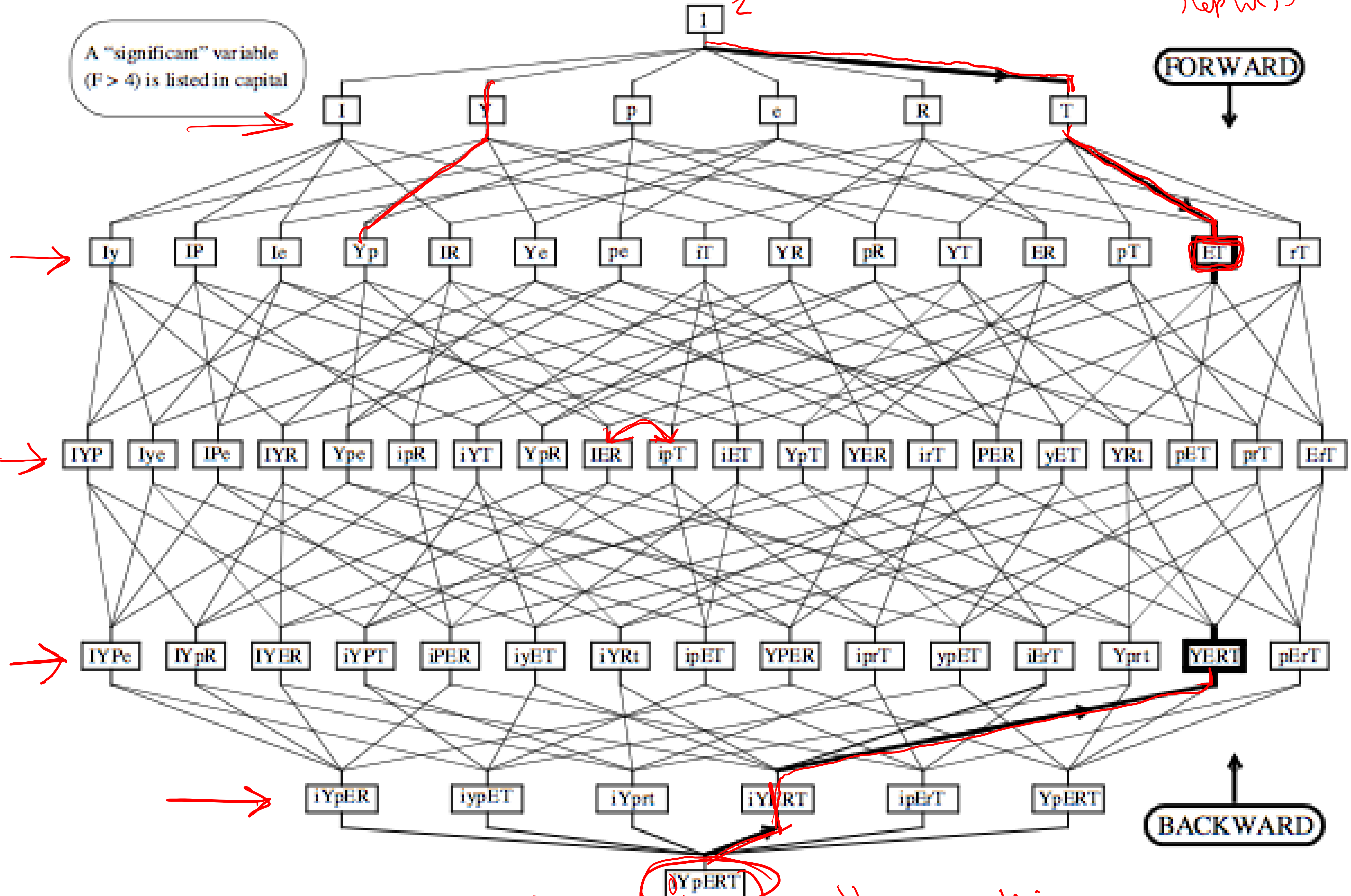
Variable Selection Example

Feb 20 2015

Charlotte Wickham

constant

stepwise



with everything

All subsets

Look at all possible models.

Then judge them on some measure of fit.

Generally learn the most by looking at a few good models.

Measures of fit

p = # β's in model

If the number of parameters are the same, we prefer the model with smaller residual sum of squares (RSS).

If the number of parameters are different, we want to balance smaller RSS with fewer parameters.

remember RSS always gets smaller if you add another parameter

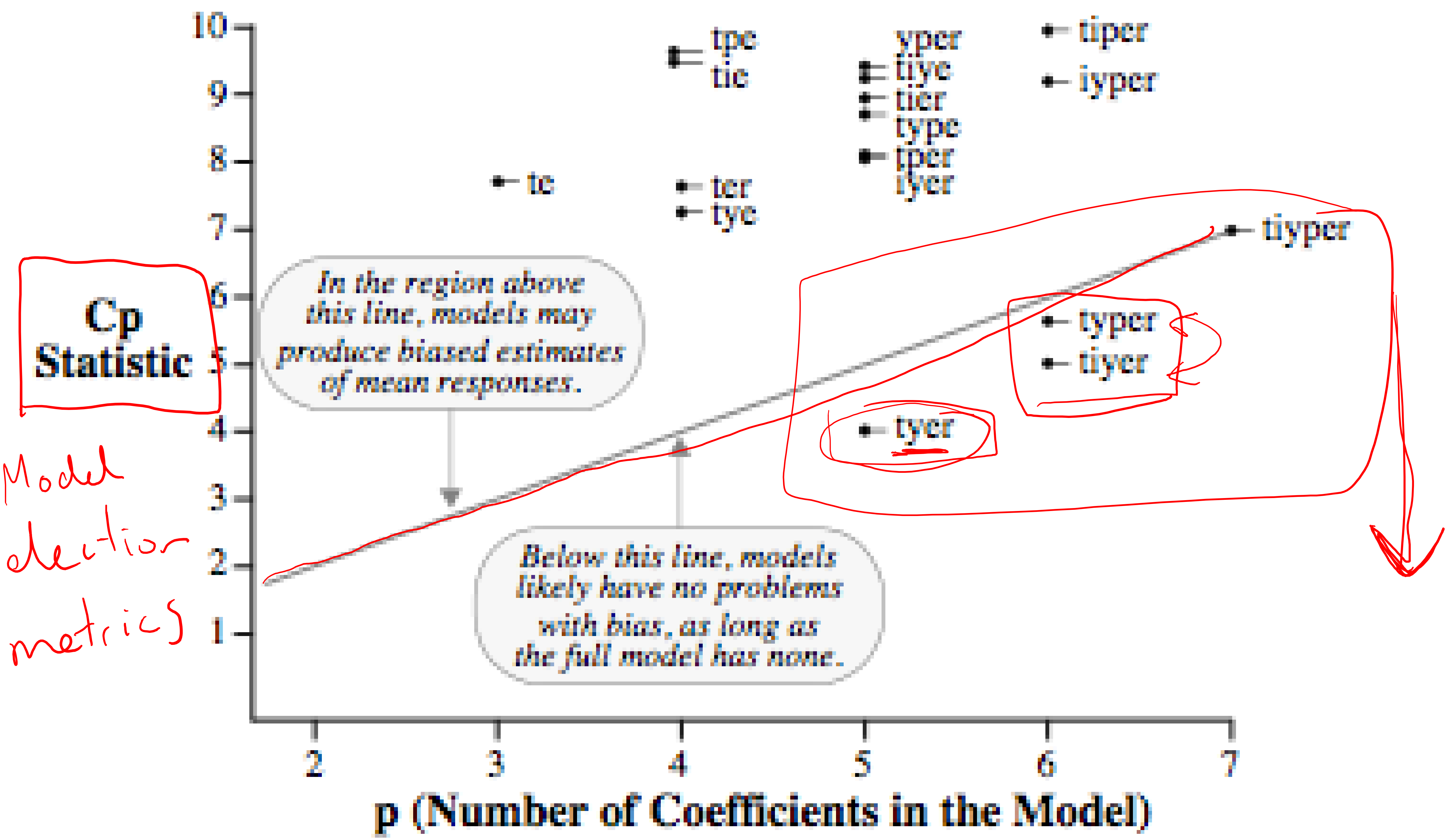
big is bad!

big if there are lots of parameters

3 common model selection criteria		
	Cp	$\frac{SSRes}{\hat{\sigma}^2_{full}} - n + p$
	BIC	$n \times \log\left(\frac{SSRes}{n}\right) + \log(n) \times (p+1)$
	AIC	$n \times \log\left(\frac{SSRes}{n}\right) + 2 \times (p+1)$

big if RSS is big

Cp plot for State SAT averages (showing only those models with Cp < 10); t = log takers, i = income, y = years, p = public, e = expend, and r = rank



Model selection metrics

complexity

You can’t trust inference after variable selection. **Why?**

We choose variables to be in the model if, in our data, they show some power to explain the response.

If a variable appears in our final model, it has by construction, shown some power to explain the response.

It doesn’t then make sense to ask if the variable is significant...we’ll get a small p-value because we only selected variables that gave low p-values!

Lab: with explanatory variables generated to have **absolutely no relationship** to the response, the best model selected by model selection has very small p-values!

Methods we won't talk about

but can be useful

Principal component based methods

Penalized methods (~~ridge~~, lasso, lars)

variable

case1202: Sex Discrimination

years

months

Sex Discrimination Data

Beginning Salary	1977 Salary	FSex(1=F)	Seniority	Age	Education	Experience
5040	12420	0	96	329	15	14
6300	12060	0	82	357	15	72
6000	15120	0	67	315	15	35.5
6000	16320	0	97	354	12	24
6000	12300	0	66	351	12	56
6840	10380	0	92	374	15	41.5
8100	13980	0	66	369	16	54.5

Sex

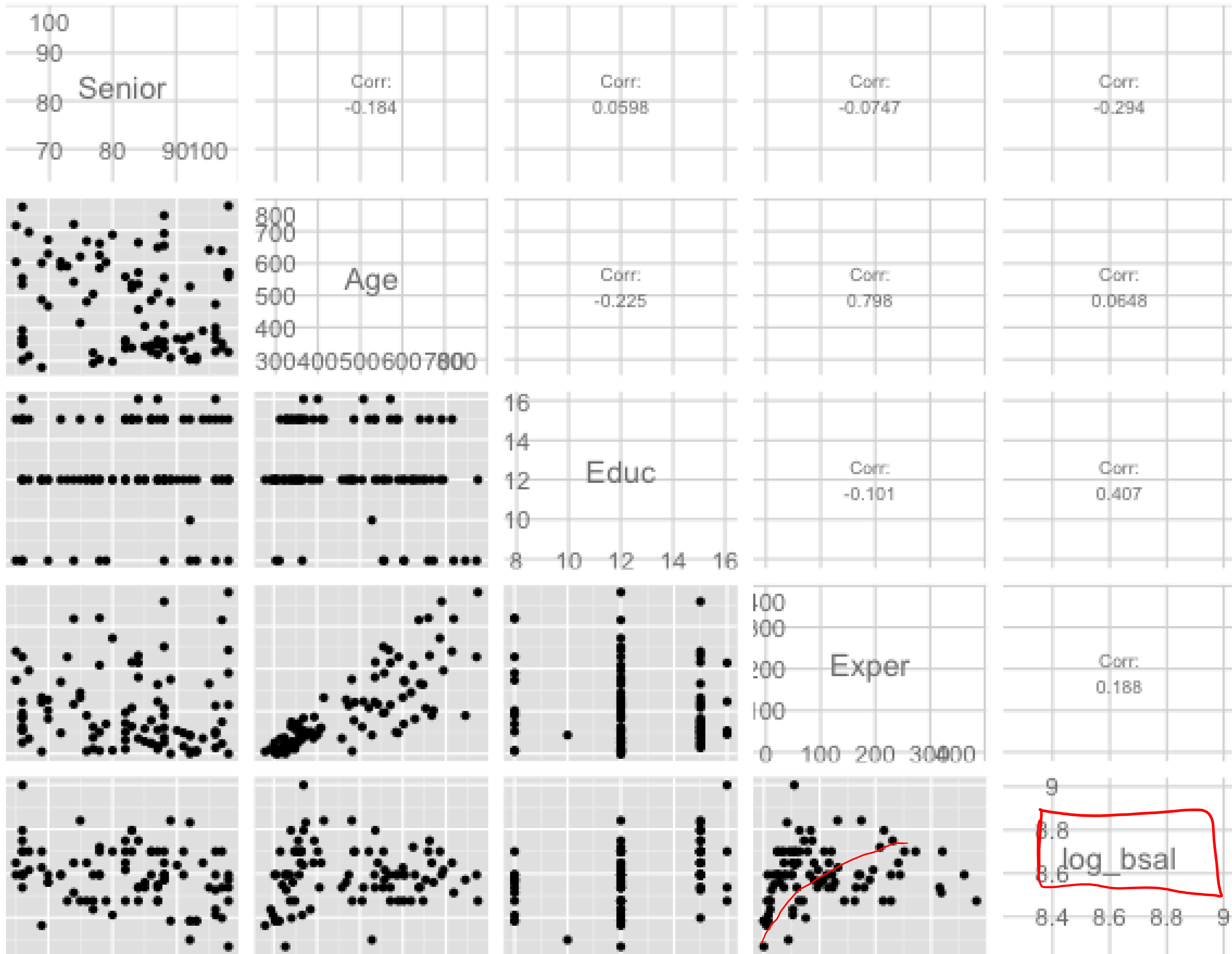
93 "skilled, entry-level clerical" employees at a bank.

Did women receive lower starting salaries than men, with similar qualifications and experience?

Strategy

Use model selection to find a suitable model to explain starting salary in terms of age, experience, seniority and education.

Once a good (or some good) models are found, add in the Sex indicator to estimate the Sex effect.


Senior
Age
Educ
Exper
log_bsal
Corr: -0.184
Corr: 0.0598
Corr: -0.0747
Corr: -0.294
Corr: -0.225
Corr: 0.798
Corr: 0.0648
Corr: -0.101
Corr: 0.407
Corr: 0.188

Possible model terms

Main Effect Variables	*Quadratic Variables*	*Interaction Variables*	
s = seniority	$t = s^2$	$m = s \times a$	$c = a \times e$
a = age	$b = a^2$	$n = s \times e$	$k = a \times x$
e = education	$f = e^2$	$v = s \times x$	$q = e \times x$
x = experience	$y = x^2$		

allow for curvature   allow for interaction

14 terms means $2^{14} = 16384$ possible models.

But:

models shouldn't include quadratic terms if they don't include the linear one

models shouldn't include interaction terms if they don't include the main effects

Strategy: find a subset of good models, then restrict attention to those that follow good practice.

Aside: model selection in R

The leaps package very quickly finds the best models for each size (number of parameters).

I.e. find the 6 best models of size 5.

It doesn't know about "good practice".

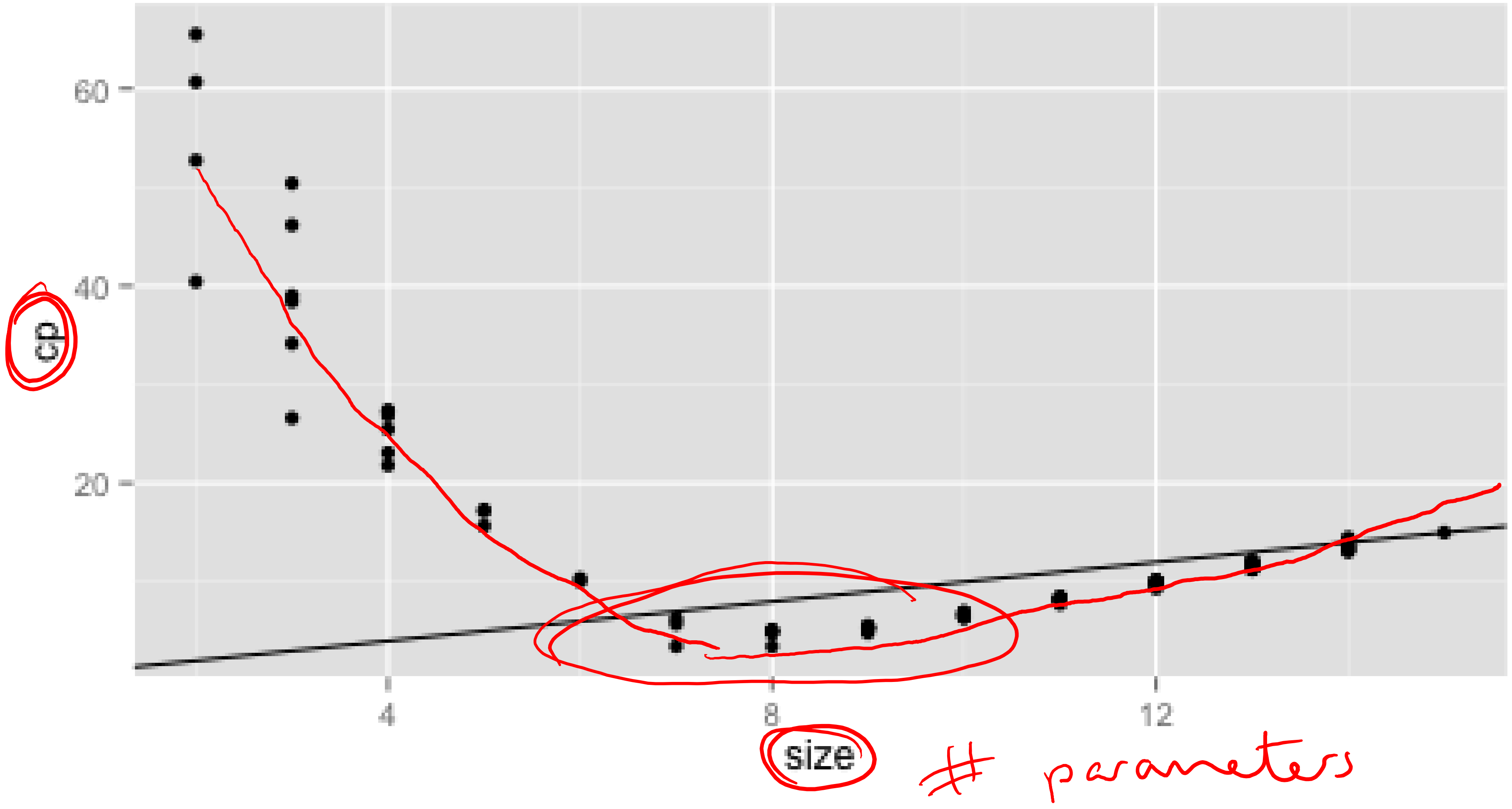
Find best 20 models of each size, then find the "good practice" models, and examine them.

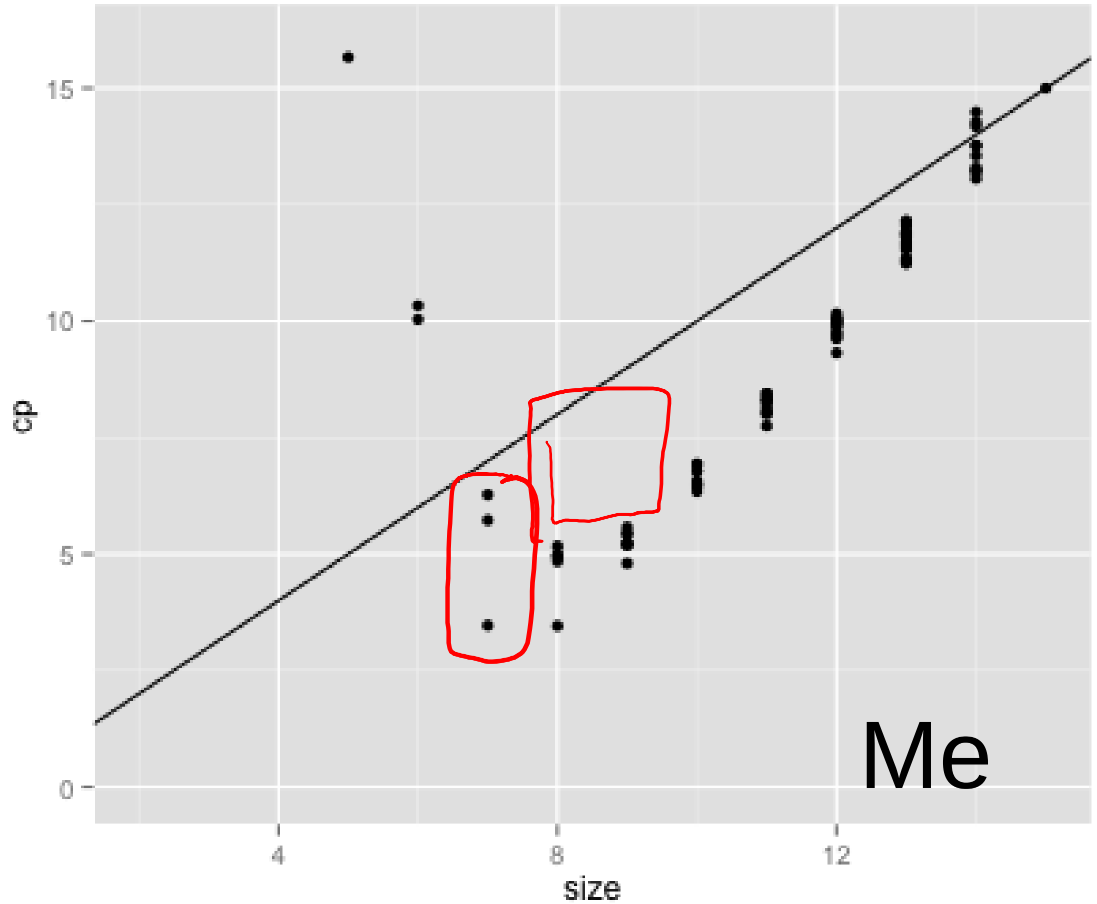
A numerical trick: center quadratic terms to remove correlation with linear terms.

```
case1202 <- mutate(case1202,
  s = Senior, a = Age, e = Educ, x = Exper,
  t = (s - mean(s))^2, b = (a - mean(a))^2,
  f = (e - mean(e))^2, y = (x - mean(x))^2,
  m = s*a, n = s*e, v = s*x, c = a*e,
  k = a*x, q = e*x)

# all subsets
all <- regsubsets(log_bsal ~ s + a + e + x + t + b + f + y + m + n + v + c + k + q,
  data = case1202,
  nbest = 30, method = "exhaustive", nvmax = 14)
```

100 "good" models



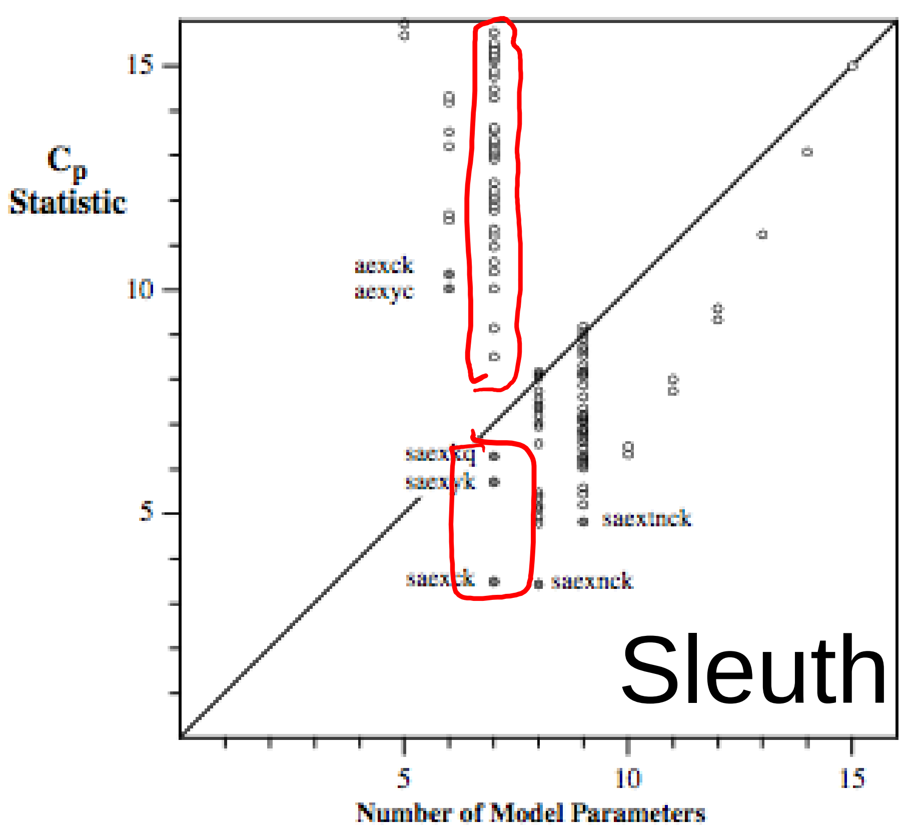


I ended up with fewer low Cp models.

Looks like I have at least the best 5 models.

After looking at this plot, I might try to get more models with 8 & 9 terms.

Cp Plot for the sex discrimination study



(ln)

"(ln)saexck" "(ln)saexnck" "(ln)saexyc" "(ln)saexkq" "(ln)saexbck"

5 best models according to BIC

"(ln)saexnck" "(ln)saexck" "(ln)saextnck" "(ln)saexbck" "(ln)saextck"

5 best models according to Cp

Picking a single model

"(In)saexck" + Sex

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.158e+00  2.205e-01  36.993  < 2e-16 ***
SexMale      1.196e-01  2.291e-02   5.219 1.25e-06 ***
s           -3.482e-03  9.090e-04  -3.830 0.000244 ***
a            9.147e-04  3.571e-04   2.562 0.012184 *
e            4.235e-02  1.568e-02   2.700 0.008356 **
x            2.181e-03  5.976e-04   3.650 0.000452 ***
c           -5.458e-05  2.909e-05  -1.876 0.064022 .
k           -3.231e-06  8.956e-07  -3.608 0.000520 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08528 on 85 degrees of freedom
Multiple R-squared: 0.5975,    Adjusted R-squared: 0.5644
F-statistic: 18.03 on 7 and 85 DF,  p-value: 1.786e-14
```

After adjusting for seniority, experience, age and education, the median salary for men is estimated to be 1.13 times the median salary for women (95% confidence interval 1.08 to 1.18).

Informally accounting for model selection

Bayesian posterior analysis of the difference between male and female log-beginning salaries

Model	p	BIC	Addition of sex indicator: coeff	SE	1-sided p-value
saexck	7	-401.40	-.1196	.0229	6.27E-7
saexyc	7	-398.89	-.1287	.0226	8.42E-8
saexkq	7	-398.28	-.1244	.0221	1.18E-7
saexnck	8	-398.08	-.1173	.0229	9.48E-7
aexyc	6	-397.81	-.1247	.0238	5.59E-7
aexck	6	-397.51	-.1135	.0246	6.94E-6
saexckb	8	-396.49	-.1195	.0229	6.70E-7
saexckt	8	-396.37	-.1189	.0232	9.10E-7
saexkqb	8	-396.36	-.1206	.0221	2.41E-7
saexycn	8	-396.33	-.1258	.0225	1.37E-7
saexk	6	-396.26	-.1331	.0221	1.96E-8
sexyq	6	-396.15	-.1345	.0201	1.02E-9
saexckf	8	-396.12	-.1196	.0230	6.93E-7
saexckq	8	-396.05	-.1208	.0230	5.54E-7
exyq	5	-395.93	-.1302	.0211	1.11E-8
saexcky	8	-395.91	-.1257	.0232	2.81E-7
saexyq	7	-398.89	-.1328	.0218	1.51E-8
saexckm	8	-395.84	-.1195	.0231	7.46E-7
saexckv	8	-395.80	-.1196	.0231	7.31E-7
saexbc	7	-395.20	-.1230	.0237	6.95E-7

Top models (Sleuth's) and the coefficient of Sex.

They are all very close, which gives us some relief that the actual model chosen doesn't matter too much.

Another look at BIC

A Bayesian approach to model selection places probability on models,

$$\Pr\{\underbrace{M_i \mid D}_{\text{}}\} = \underbrace{\Pr\{M_i\}}_{\text{prior belief}} \; \underbrace{\exp\{-BIC_i\}}_{} / \text{SUM}$$

posterior probability of model i

prior probability of model i

probability of seeing the data if model i is true

where $\text{SUM} = \sum_j \{\Pr\{M_j\} \exp\{-BIC_j\}\}$

It's convenient to say
"all models are equally probable before we see any data".

Formally accounting for model selection

Bayesian posterior analysis of the difference between male and female log-beginning salaries

				Addition of sex indicator		
Model	p	BIC	posterior probability	coeff	*SE*	1-sided p-value
saexck	7	-401.40	.7709	-.1196	.0229	6.27E-7
saexyc	7	-398.89	.0625	-.1287	.0226	8.42E-8
saexkq	7	-398.28	.0340	-.1244	.0221	1.18E-7
saexnck	8	-398.08	.0279	-.1173	.0229	9.48E-7
aexyc	6	-397.81	.0213	-.1247	.0238	5.59E-7
aexck	6	-397.51	.0157	-.1135	.0246	6.94E-6
saexckb	8	-396.49	.0057	-.1195	.0229	6.70E-7
saexckt	8	-396.37	.0051	-.1189	.0232	9.10E-7
saexkqb	8	-396.36	.0050	-.1206	.0221	2.41E-7
saexycn	8	-396.33	.0048	-.1258	.0225	1.37E-7
saexk	6	-396.26	.0045	-.1331	.0221	1.96E-8
sexyq	6	-396.15	.0040	-.1345	.0201	1.02E-9
saexckf	8	-396.12	.0039	-.1196	.0230	6.93E-7
saexckq	8	-396.05	.0037	-.1208	.0230	5.54E-7
exyq	5	-395.93	.0032	-.1302	.0211	1.11E-8
saexcky	8	-395.91	.0032	-.1257	.0232	2.81E-7
saexyq	7	-398.89	.0031	-.1328	.0218	1.51E-8
saexckm	8	-395.84	.0030	-.1195	.0231	7.46E-7
saexckv	8	-395.80	.0028	-.1196	.0231	7.31E-7
saexbc	7	-395.20	.0016	-.1230	.0237	6.95E-7

Formally accounting for model selection

Model averaging

Bayesian posterior estimate of the Sex effect:

$\sum_i$ Pr{ M_i | D } x estimate of Sex effect in Model i

= -0.1206

Bayesian posterior estimate of the p-value

$\sum_i$ Pr{ M_i | D } x p-value Sex effect in Model i

= 6.7×10^{-7}

We have allowed the data to dictate the model. All our traditional inferences act as though the model was pre-specified.

Estimates, confidence intervals and p-values should be used with caution.

There are approaches to try to fix this. A simple one, if you have enough data, is to split your data into one set for choosing the model and an independent one for estimation.