

432 Class 5 Slides

github.com/THOMASELOVE/432-2018

2018-01-30

Setup

```
library(skimr)
library(broom)
library(modelr)
library(leaps)
library(tidyverse)

oh_count <- read.csv("data/counties2017a.csv") %>% tbl_df
```

Today's Materials

- Review of Minute Papers after Class 04
- Discussion of Homework 1
- Ohio County Health Rankings Data
- Variable Selection via Best Subsets
 - Adjusted R^2
 - Mallows' C_p
 - AIC after Correction for Bias
 - BIC
- Cross-Validating to Compare Two Model-Building Approaches
- Assessing Residual Diagnostic Plots

Homework 1

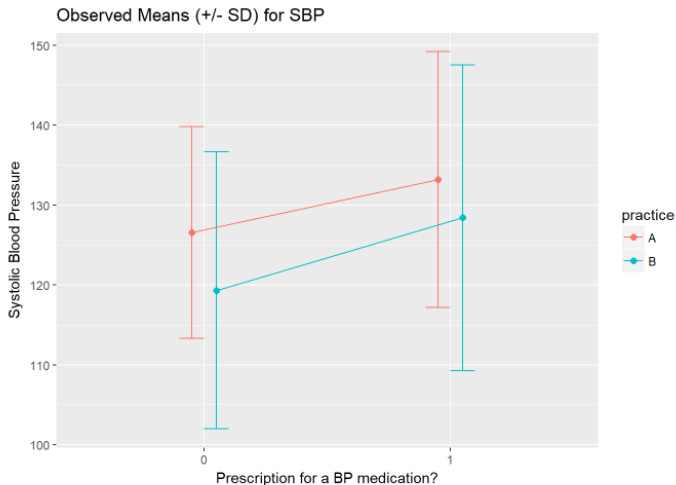
Table 1

n	Stratified by practice		p	test
	A	B		
	180	150		
age (mean (sd))	56.34 (11.17)	54.17 (11.89)	0.088	
race (%)			<0.001	
Asian/PI	0 (0.0)	10 (6.7)		
Black/AA	166 (92.7)	14 (9.4)		
Multi-Racial	4 (2.2)	3 (2.0)		
White	9 (5.0)	122 (81.9)		
eth_hisp = Yes (%)	2 (1.1)	62 (41.6)	<0.001	
sex = M (%)	61 (33.9)	66 (44.0)	0.077	
insurance (%)			0.016	
Commercial	35 (19.4)	18 (12.0)		
Medicaid	66 (36.7)	68 (45.3)		
Medicare	76 (42.2)	54 (36.0)		
Uninsured	3 (1.7)	10 (6.7)		
bmi (mean (sd))	35.20 (8.20)	34.39 (7.83)	0.365	
bmi_cat (%)			0.587	
Underweight	1 (0.6)	1 (0.7)		
Normal	11 (6.1)	14 (9.3)		
Overweight	32 (17.8)	31 (20.7)		
Obese	136 (75.6)	104 (69.3)		
sbp (mean (sd))	130.82 (15.38)	125.44 (19.00)	0.005	
dbp (mean (sd))	74.49 (11.40)	75.05 (8.58)	0.617	

Notes for Table 1:

1. There are 4 subjects missing Hispanic ethnicity status in practice A, and 1 in practice B.
2. There is 1 subject in each practice missing Race.
3. Results are shown in terms of means and standard deviations for quantitative variables, and t tests are used for comparisons, because a Normal approximation was a reasonable choice for each such variable.
4. For categorical variables, we display counts and percentages, and use Pearson chi-square tests of significance.

Question 2



I don't see much to suggest a meaningful interaction here. The lines joining the points are essentially parallel. It looks like the group with the lowest (healthiest) mean SBP are the subjects in practice B without a medication.

Question 2 ANOVA (no interaction)

```
Call:
lm(formula = sbp ~ practice + bpmed, data = hbp330)

Residuals:
    Min       1Q   Median       3Q      Max
-41.844 -11.961  -0.702   9.369  63.039

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  125.844      1.768  71.164 < 2e-16 ***
practiceB     -5.600      1.852  -3.023  0.0027 **
bpmed         7.716      1.944   3.970 8.85e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.75 on 327 degrees of freedom
Multiple R-squared:  0.06889,    Adjusted R-squared:  0.06319
F-statistic: 12.1 on 2 and 327 DF,  p-value: 8.548e-06
```

Since each of the two factors is binary, we can simply read off that both `practice` and `bpmed` appear to have a significant impact on SBP, with practice B having lower SBP levels, on average, and subjects without BP medications having lower SBP levels, on average.

Question 3 (ANOVA test to compare models)

```
anova(hw1_q3, hw1_q2_no_int)
```

Analysis of Variance Table

Model 1: sbp ~ practice + bpmed + age

Model 2: sbp ~ practice + bpmed

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	326	90178				
2	327	91712	-1	-1534.3	5.5467	0.01911 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

It does appear that `age` adds significant predictive value to the no-interaction model.

Question 3 (Fit Summaries)

```
glance(hw1_q3)
```

	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik
1	0.08446494	0.07603977	16.63185	10.02531	2.452493e-06	4	-1393.973
	AIC	BIC	deviance	df.residual			
1	2797.946	2816.941	90177.6	326			

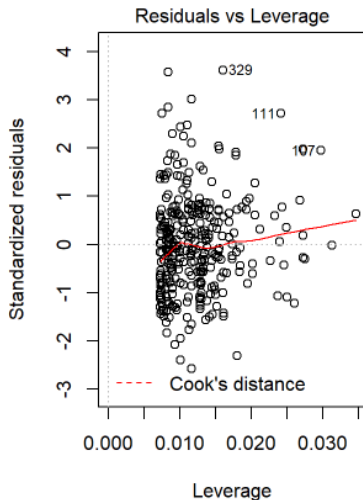
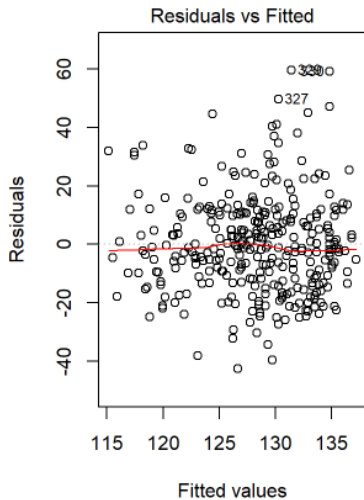
Hide

```
glance(hw1_q2_no_int)
```

	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik
1	0.06888764	0.06319276	16.74708	12.09642	8.547555e-06	3	-1396.757
	AIC	BIC	deviance	df.residual			
1	2801.513	2816.71	91711.92	327			

The model with `age` included performs a bit better in terms of adjusted (and raw) R^2 and AIC and performs comparably in terms of BIC.

Question 3 (Residual plots)



Ohio County Health Rankings Data
**[http://www.countyhealthrankings.org/
rankings/data/oh](http://www.countyhealthrankings.org/rankings/data/oh)**

Codebook (2017 County Health Rankings), I

Variable	Description
fips	FIPS code for county (an ID)
state	Ohio in all cases
county	County Name (88 counties in Ohio)
years_lost	Years of potential life lost before age 75 per 100,000 population (age-adjusted, 2012-14)
population	County population, Census Population Estimates, 2015
female	% female (Census Population Estimates, 2015)
rural	3 categories from % rural (0-20: Urban, 20.1-50: Suburban, 50.1+: Rural; Census 2015)
non_white	4 categories from 100 - % white non-hispanic: (> 20: High, 10.1-20: Medium, 5.1-10: Low, <=5: Very Low, Census 2015)

Codebook (2017 County Health Rankings), II

Variable	Description
sroh_fairpoor	% of adults reporting fair or poor health (age-adjusted via 2015 BRFSS)
smoker_pct	% of adults who currently smoke (2015 BRFSS)
food_envir	Food environment index (0 = worst, 10 = best) (via USDA Map the Meal 2014)
exer_access	% of population with adequate access to locations for physical activity (several sources)
income_ratio	Ratio of household income at the 80th percentile to income at the 20th percentile (ACS 2011-15)
air_pollution	Mean daily density of fine particulate matter in micrograms per cubic meter (PM2.5)
health_costs	Health Care Costs (from Dartmouth Atlas, 2014)

Basic Data Summaries

```
oh_count %>% select(-fips, -state, -county) %>% skim()
```

Skim summary statistics

n obs: 88

n variables: 12

Variable type: factor

variable	missing	complete	n	n_unique	top_counts	ordered
non_white	0	88	88	4	Low: 30, Ver: 27, Med: 23, Hig: 8	FALSE
rural	0	88	88	3	Rur: 43, Sub: 31, Urb: 14, NA: 0	FALSE

Variable type: integer

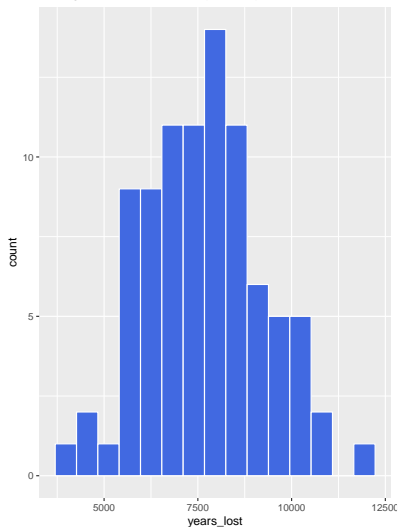
variable	missing	complete	n	mean	sd	p0	p25	median	p75	p100	hist
population	0	88	88	131970.72	216261.12	13048	36982.25	57733.5	123712.75	1255921	
years_lost	0	88	88	7659.12	1563.34	4129	6538.75	7700	8597.5	12091	

Variable type: numeric

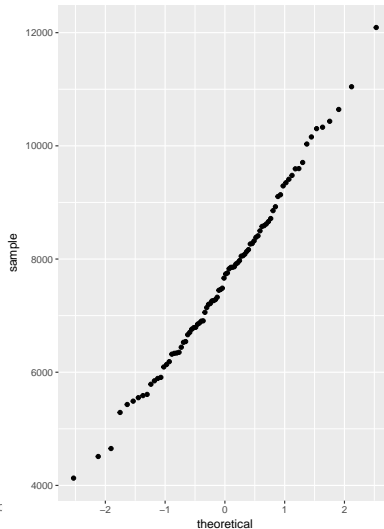
variable	missing	complete	n	mean	sd	p0	p25	median	p75	p100	hist
air_pollution	0	88	88	11.38	0.47	10.5	11.1	11.3	11.7	13	
exer_access	0	88	88	68.19	17.43	26.2	58.18	69.73	80.09	96.23	
female	0	88	88	50.34	1.38	41.78	50.05	50.58	50.96	52.41	
food_envir	0	88	88	7.4	0.67	5.3	7	7.45	7.8	8.9	
health_costs	0	88	88	10158.06	859.43	8274.48	9650.2	10093.36	10577.49	13702.91	
income_ratio	0	88	88	4.33	0.6	3.45	3.94	4.21	4.57	7.24	
smoker_pct	0	88	88	19.33	2.05	13.82	18.23	19.28	20.61	24.53	
sroh_fairpoor	0	88	88	15.99	2.14	10.31	14.58	15.86	17.21	21.86	

Our Outcome: Age-Adjusted Years Lost

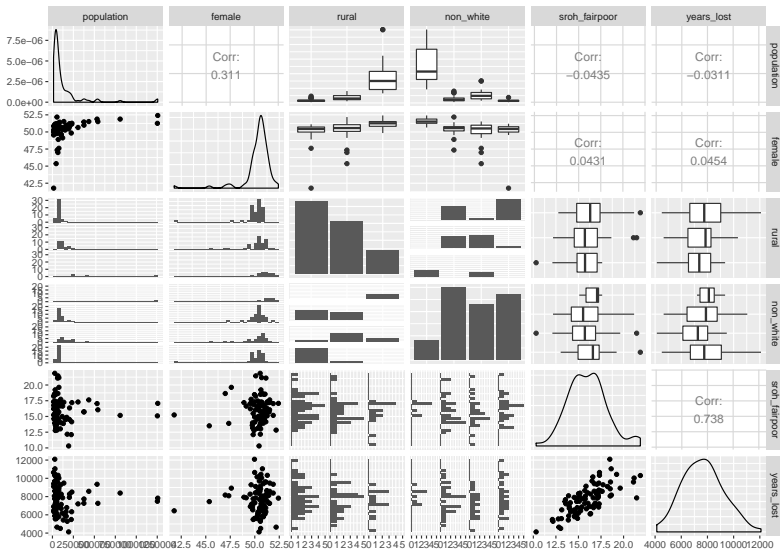
Histogram of Years Lost, by County



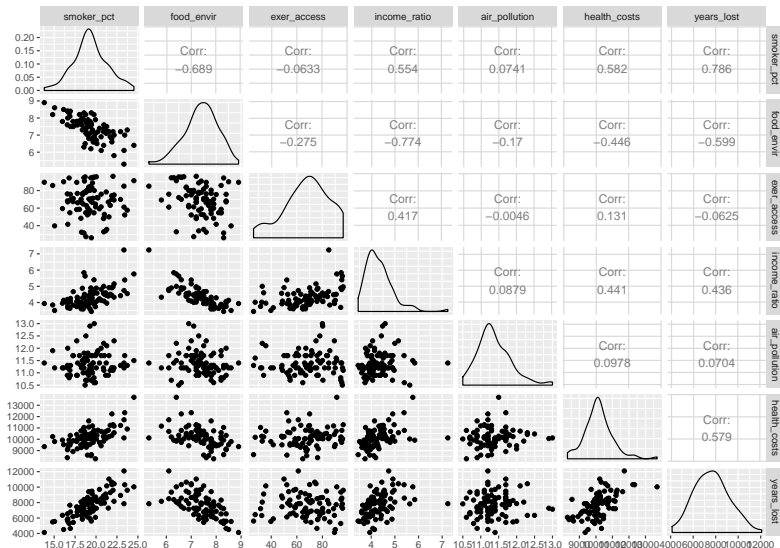
Normal Q-Q of Years Lost



Scatterplot Matrix with GGally, Part I



Scatterplot Matrix with GGally, Part II



Using “Best Subsets” to Select Variables

Using “Best Subsets” to Select Variables

We'll consider models using some combination of the 11 available meaningful predictors.

```
bs_preds <- with(oh_count, cbind(population, female, rural,  
                                non_white, sroh_fairpoor,  
                                smoker_pct, food_envir,  
                                exer_access, income_ratio,  
                                air_pollution, health_costs))
```

We'll look for models using up to 8 of those predictors.

```
bs_subs <- regsubsets(bs_preds,  
                     y = oh_count$years_lost,  
                     nvmax = 8)  
bs_mods <- summary(bs_subs)
```

Looking at bs_mods

bs_mods

[illegible]

Look at the models that “win”

```
bs_mods$which
```

```
> bs_mods$which
(Intercept) population female rural non_white sroh_fairpoor smoker_pct food_envir exer_access income_ratio air_pollution health_costs
1      TRUE      FALSE  FALSE  FALSE      FALSE      FALSE      TRUE      FALSE      FALSE      FALSE      FALSE      FALSE
2      TRUE      FALSE  FALSE  FALSE      FALSE      FALSE      TRUE      FALSE      FALSE      FALSE      FALSE      TRUE
3      TRUE      FALSE  FALSE  FALSE      FALSE      TRUE       TRUE      FALSE      FALSE      FALSE      FALSE      TRUE
4      TRUE      FALSE  FALSE  FALSE      FALSE      FALSE      TRUE      TRUE      FALSE      TRUE      FALSE      TRUE
5      TRUE      FALSE  TRUE   FALSE      FALSE      FALSE      TRUE      TRUE      FALSE      TRUE      FALSE      TRUE
6      TRUE      FALSE  TRUE   FALSE      FALSE      FALSE      TRUE      TRUE      TRUE      TRUE      FALSE      TRUE
7      TRUE      FALSE  TRUE   FALSE      FALSE      TRUE       TRUE      TRUE      TRUE      TRUE      FALSE      TRUE
8      TRUE      FALSE  TRUE   FALSE      TRUE       TRUE       TRUE      TRUE      TRUE      TRUE      FALSE      TRUE
```

Sometimes easier to transpose this...

```
t(bs_mods$which)
```

```
> t(bs_mods$which)
```

	1	2	3	4	5	6	7	8
(Intercept)	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
population	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
female	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
rural	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
non_white	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
sroh_fairpoor	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE
smoker_pct	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
food_envir	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
exer_access	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
income_ratio	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
air_pollution	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
health_costs	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Look at the R-square values for each “winning” model

```
bs_mods$rsq
```

```
[1] 0.6172471 0.6397030 0.6460605 0.6530869 0.6649312  
[6] 0.6730306 0.6783975 0.6802613
```

```
bs_mods$adjr2
```

```
[1] 0.6127964 0.6312255 0.6334198 0.6363682 0.6445001  
[6] 0.6488107 0.6502573 0.6478827
```

Place winning results in bs_winners

```
bs_winners <- tbl_df(bs_mods$which)
bs_winners$k <- 2:9 ## in general, this is 2:(nvmax + 1)
bs_winners$r2 <- bs_mods$rsq
bs_winners$adjr2 <- bs_mods$adjr2
bs_winners$cp <- bs_mods$cp
bs_winners$bic <- bs_mods$bic
```


Calculate Bias-Corrected AIC from Residual Sum of Squares

This requires specifying the sample size (`temp.n`) and the number of inputs that you'll look at in your largest subset (here, we limited the number of variables to 8 with `nvmax` and so that's 9 inputs, including the intercept term.)

```
temp.n <- nrow(oh_count)
temp.inputs <- 9 ## nvmax + 1

bs_mods$aic.corr <- temp.n*log(bs_mods$rss / temp.n) +
  2*(2:temp.inputs) +
  (2 * (2:temp.inputs) * ((2:temp.inputs)+1) /
    (temp.n - (2:temp.inputs) - 1))

bs_winners$aic.corr <- bs_mods$aic.corr
```

Detailed Breakdown: bs_winners

Inputs	Predictors	Raw r^2	Adj. r^2	C_p	BIC	AIC_c
2	smoker_pct	.617	.613	8.0	-75.6	1213.0
3	+ health_costs	.640	.631	4.6	-76.4	1209.9
4	+ sroh_fairpoor	.646	.633	5.1	-73.5	1210.5
5	(see below)	.653	.636	5.4	-70.8	1211.0
6	+ female	.665	.645	4.5	-69.4	1210.2
7	+ exer_access	.673	.649	4.6	-67.0	1210.4
8	+ sroh_fairpoor	.678	.650	5.3	-64.0	1211.4
9	+ non_white	.680	.648	6.9	-60.0	1213.4

- The “best” model with 5 inputs includes smoker_pct, health_costs, food_envir and income_ratio.
- That model forms the basis for the “best” models with 6-9 inputs.

Resulting bs_winners tibble

```
head(bs_winners, 2)
```

```
# A tibble: 2 x 18
#   `(Intercept)` population female rural non_white
#   <lgl>          <lgl>          <lgl>  <lgl> <lgl>
1 T              F              F      F      F
2 T              F              F      F      F
# ... with 13 more variables: sroh_fairpoor <lgl>,
#   smoker_pct <lgl>, food_envir <lgl>, exer_access <lgl>,
#   income_ratio <lgl>, air_pollution <lgl>,
#   health_costs <lgl>, k <int>, r2 <dbl>, adjr2 <dbl>,
#   cp <dbl>, bic <dbl>, aic.corr <dbl>
```

If You're Curious: A Stepwise Fit

```
step(lm(years_lost ~ population + female + rural +  
        non_white + sroh_fairpoor + smoker_pct +  
        food_envir + exer_access + income_ratio +  
        air_pollution + health_costs, data = oh_count))
```

using backwards elimination produces the model containing:

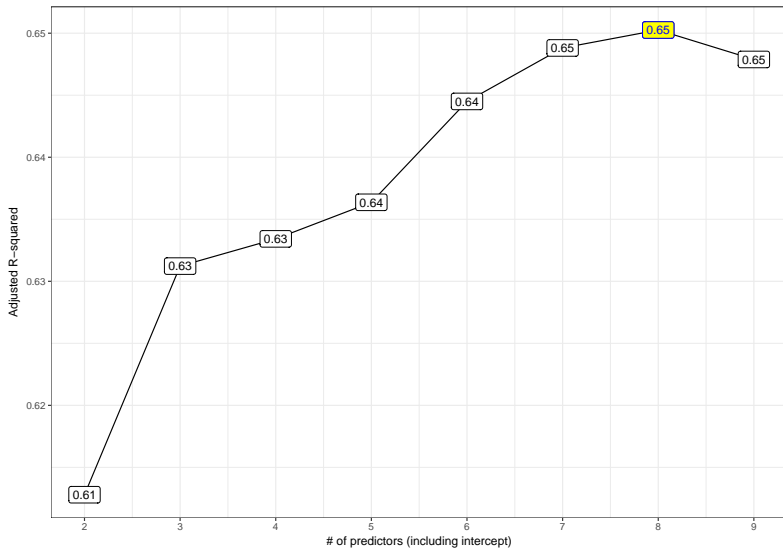
- smoker_pct, health_costs, food_envir, income_ratio, female, and exer_access
- also known as what “best subsets” chose for its model 7.

Building the “Best Subsets” Plots

Adjusted R-square plot using ggplot2

```
p1 <- ggplot(bs_winners, aes(x = k, y = adjr2,  
                             label = round(adjr2,2))) +  
  geom_line() +  
  geom_label() +  
  geom_label(data = subset(bs_winners,  
                           adjr2 == max(adjr2)),  
             aes(x = k, y = adjr2, label = round(adjr2,2)),  
             fill = "yellow", col = "blue") +  
  theme_bw() +  
  scale_x_continuous(breaks = 2:9) +  
  labs(x = "# of predictors (including intercept)",  
       y = "Adjusted R-squared")
```

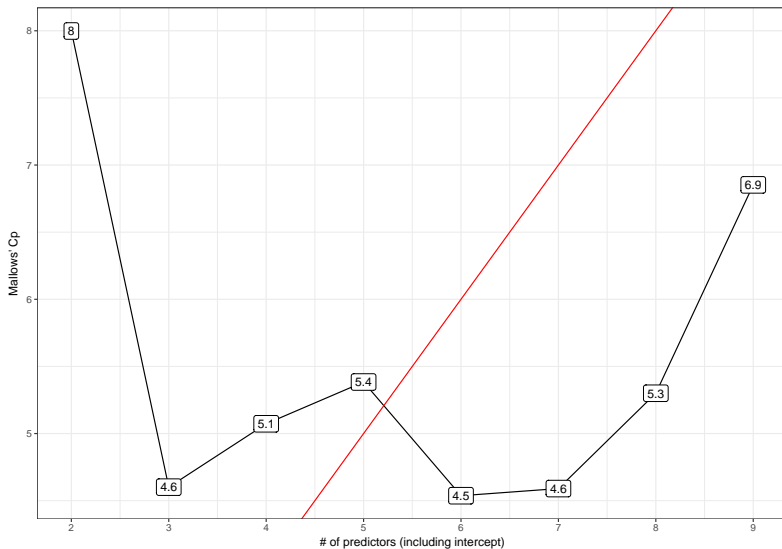
Adjusted R-square plot using ggplot2



Mallows' C_p plot using ggplot2

```
p2 <- ggplot(bs_winners, aes(x = k, y = cp,  
                             label = round(cp,1))) +  
  geom_line() +  
  geom_label() +  
  geom_abline(intercept = 0, slope = 1,  
              col = "red") +  
  theme_bw() +  
  scale_x_continuous(breaks = 2:9) +  
  labs(x = "# of predictors (including intercept)",  
       y = "Mallows' Cp")
```

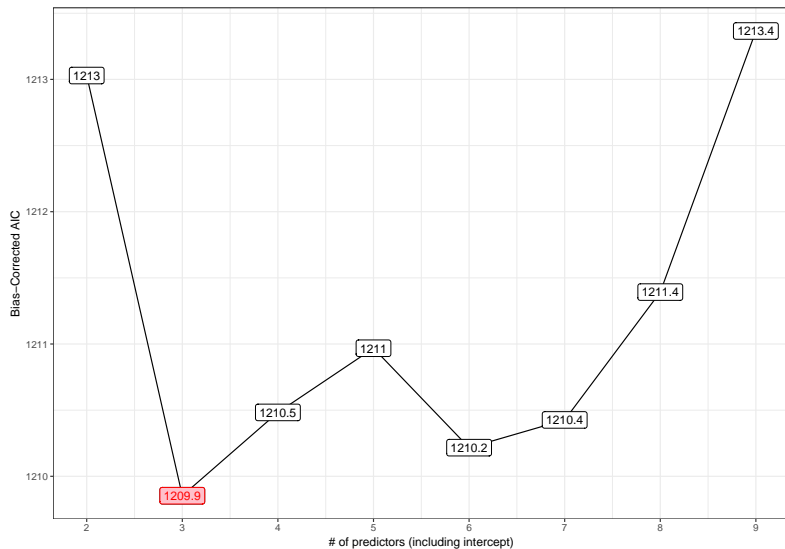

Mallows' C_p plot using ggplot2



Corrected AIC plot using ggplot2

```
p3 <- ggplot(bs_winners, aes(x = k, y = aic.corr,  
                             label = round(aic.corr,1))) +  
  geom_line() +  
  geom_label() +  
  geom_label(data = subset(bs_winners,  
                           aic.corr == min(aic.corr)),  
             aes(x = k, y = aic.corr),  
             fill = "pink", col = "red") +  
  theme_bw() +  
  scale_x_continuous(breaks = 2:9) +  
  labs(x = "# of predictors (including intercept)",  
       y = "Bias-Corrected AIC")
```

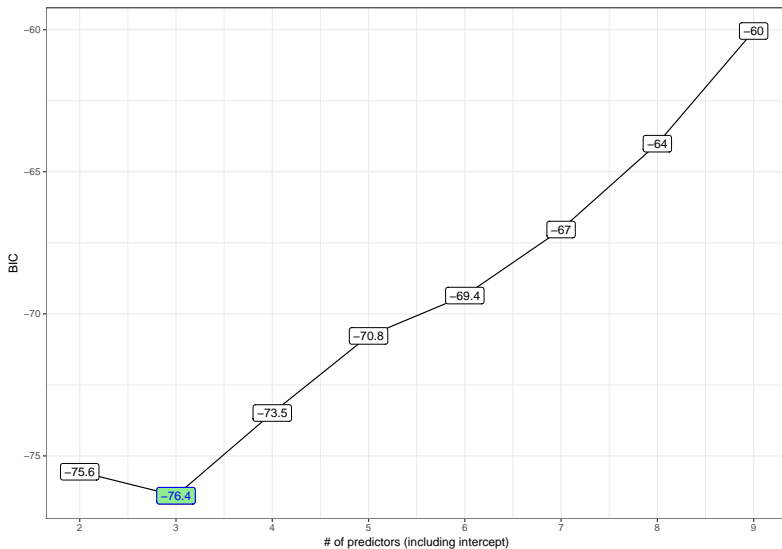
Corrected AIC plot using ggplot2



BIC plot using ggplot2

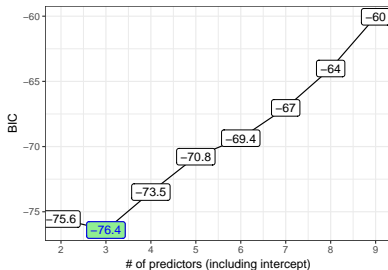
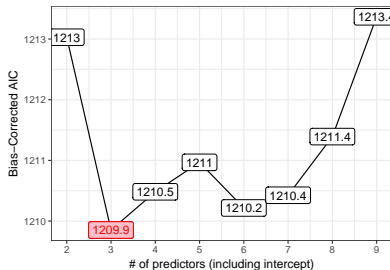
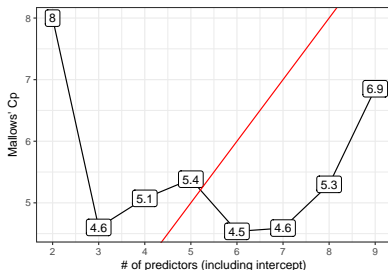
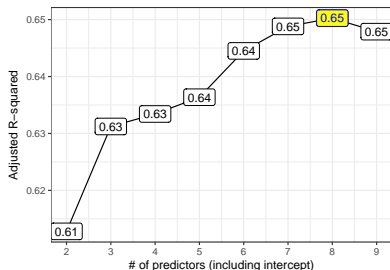
```
p4 <- ggplot(bs_winners, aes(x = k, y = bic,  
                             label = round(bic,1))) +  
  geom_line() +  
  geom_label() +  
  geom_label(data = subset(bs_winners, bic == min(bic)),  
             aes(x = k, y = bic),  
             fill = "lightgreen", col = "blue") +  
  theme_bw() +  
  scale_x_continuous(breaks = 2:9) +  
  labs(x = "# of predictors (including intercept)",  
       y = "BIC")
```

BIC plot using ggplot2



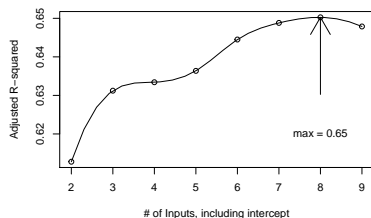
All Four Plots Together

```
gridExtra::grid.arrange(p1, p2, p3, p4, nrow = 2)
```

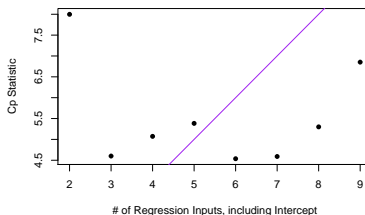


The Four Plots (using Base R plotting)

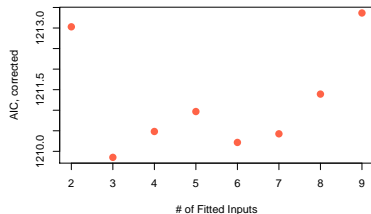
Adjusted R-squared



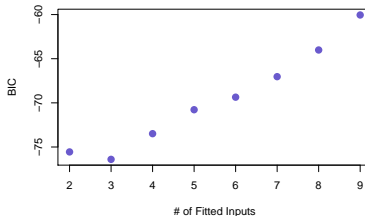
Cp Plot



AIC (corrected) Plot



BIC Plot



Next Time

- Best Subsets and Cross-Validation
- Stepwise Regression and the Allen-Cady Procedure
- (soon) Making Decisions about Non-Linearity in Y or the X s