Selecting the Best Model

As we discussed earlier, we may *not* want to include all the predictors in the model. Some of the predictors may **not** be statistically significant (i.e. they have a large p-value), in which case, they are not needed. In addition, a model with a smaller number of predictors is **easier to interpret**, so researchers often prefer the simplest possible model.

- For example, interaction terms may not be necessary. If they are **not** statistically significant, we should take them out of the model. (However, if they *are* significant, then we should keep them in the model).

If interaction X_1X_2 is SIG, then we keep X_1 and X_2 (lower-order terms) regardless of their p-values.

- If we are adding or taking out predictors in our new model, we should compare the new **adjusted** R^2 to the old model's adjusted R^2.

R^2 never goes down when we add a predictor variable no matter how bad that predictor is.
Hence, our goal is NOT maximizing R^2, but R^2_{adj}. We aim to find the simplest model that does a decent job of predicting y.

- We should also examine the p-values of the **individual** predictors. Should we throw out **all** the predictors that have a large p-value? **NO.** Why?

If higher-order term SIG → keep lower-order terms
Due to the possibility of multicolinearity, we do NOT throw out all predictors that have a large p-value.
We eliminate/add ONE predictor at a time.

- How do we select the best model? There are many ways to perform model selection, but in this class, we will learn three options.

  – Backwards elimination

• starts with ALL available predictors in model → means we need to refit the model after throwing out one variable, repeatedly
• throws out variables with high p-vals ONE at a time (usually >0.05) and each time we remove the variable with the highest p-val)

  – Forward selection

• starts with all models with one pred – pick the best one (can use R^2, R^2_{adj}, [redacted] and even p-val)
• adds all others as second pred – pick the best two
• continues until no more sig pred.

  – Best subsets regression

• Computer creats EVERY possible model with one pred, two pred, ..., all pred
• Prints a summary of the best 2 models for each p

y

EXAMPLE: Predicting $\boxed{\text{College GPA}}$ - data from book

p=16 predictor variables

n=59 ppl

Regression Analysis: CGPA versus Height, Gender, etc

* sample size should be 5-20 times bigger than # predictors.

→ We need to collect more data, or if possible, reduce ~~# pred~~ # predictors.

```
The regression equation is
CGPA = 0.53 + 0.0194 Height + 0.047 Gender - 0.00163 Haircut - 0.042 Job
       + 0.0004 Studytime - 0.375 Smokecig + 0.0488 Dated + 0.546 HSGPA
       + 0.00315 HomeDist + 0.00069 BrowseInternet - 0.00128 WatchTV
- 0.0117 Exercise + 0.0140 ReadNewsP + 0.039 Vegan
- 0.0139 PoliticalDegree - 0.0801 PoliticalAff
```

Predictor	Coef	SE Coef	T	P (P(>|t|))
Constant	0.532	1.496	0.36	0.724
Height	0.01942	0.01637	1.19	0.242
Gender	0.0468	0.1429	0.33	0.745
Haircut	-0.001633	0.001697	-0.96	0.341
Job	-0.0418	0.1024	-0.41	0.685
Studytime	0.00043	0.01921	0.02	0.982
Smokecig	-0.3746	0.2249	-1.67	0.103
Dated	0.04881	0.07111	0.69	0.496
HSGPA	0.5457	0.1776	3.07	0.004 ←
HomeDist	0.003147	0.003400	0.93	0.360
BrowseInternet	0.000689	0.001163	0.59	0.557
WatchTV	-0.0012840	0.0009710	-1.32	0.193
Exercise	-0.011657	0.005934	-1.96	0.056 ←
ReadNewsP	0.01395	0.02272	0.61	0.543
Vegan	0.0392	0.1578	0.25	0.805
PoliticalDegree	-0.01390	0.03185	-0.44	0.665
PoliticalAff	-0.08006	0.07741	-1.03	0.307

Only 2 predictor variables have small p-vals → We need to simplify the model slowly

Don't throw everything "bad" out of the model at once.

S = 0.322198 $\underbrace{\text{R-Sq = 43.2\%}}_{\text{Small, best we can do so far}}$ $\overbrace{\text{R-Sq(adj) = 21.5\%}}^{\text{Small, something we want to improve.}}$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	16	3.3135	0.2071	1.99	0.037
Residual Error	42	4.3601	0.1038		
Total	58	7.6736			

→ ANOVA p-val = 0.037
Pretty strong evidence that at least one pred good.

Unusual Observations

Obs	Height	CGPA	Fit	SE Fit	Residual	St Resid
28	67.0	2.9800	3.5898	0.2442	-0.6098	-2.90R
40	65.0	3.9300	3.3458	0.2176	0.5842	2.46R
59	62.0	2.5000	3.4718	0.1352	-0.9718	$\boxed{\text{-3.32R}}$ Large than 3 in absolute value

R denotes an observation with a large standardized residual.

Best Subsets Regression: CGPA versus Height, Gender, ...

Best subsets regression.

Response is CGPA

Good for interpretation
Bad for model selection. (R-Sq)

larger better (R-Sq(adj))

smaller better (Mallows C-p)

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	Height	Gender	Haircut	Job	Studytime	Smokecig	Dated	HSGPA	HomeDist	BrowseInternet	WatchTV	Exercise	ReadNewsP	Vegan	PoliticalDegree	PoliticalAff
1	25.5	24.2	0.1	0.31667								X								
1	13.0	11.5	9.3	0.34217												X				
2	31.6	29.2	-2.4	0.30613								X				X				
2	29.4	26.9	-0.8	0.31109								X			X					
3	33.8	30.2	-2.1	0.30389			X					X				X				
3	33.7	30.0	-2.0	0.30423								X			X	X				
4	35.7	31.0	-1.5	0.30223						X		X			X	X				
4	35.3	30.5	-1.2	0.30320			X					X			X	X				
5	37.3	31.4	-0.6	0.30132	X					X		X			X	X				
5	37.0	31.1	-0.4	0.30198			X			X		X			X	X				
6	38.3	31.2	0.6	0.30163	X		X			X		X			X	X				
6	38.3	31.2	0.6	0.30164	X					X		X			X	X				X
7	39.6	31.3	1.7	0.30150	X		X			X		X			X	X				X
7	39.3	30.9	1.9	0.30231	X					X		X	X		X	X				X
8	40.4	30.8	3.1	0.30249	X		X			X		X			X	X			X	X
8	40.4	30.8	3.1	0.30256	X		X			X		X	X		X	X				X
9	41.5	30.8	4.2	0.30266	X		X			X		X	X		X	X			X	X
9	41.0	30.2	4.6	0.30395	X		X	X		X		X	X		X	X				X
10	41.9	29.8	6.0	0.30478	X		X			X	X	X	X		X	X			X	X
10	41.8	29.7	6.0	0.30492	X		X	X		X		X	X		X	X			X	X
11	42.2	28.7	7.7	0.30712	X		X	X		X	X	X	X		X	X			X	X
11	42.2	28.7	7.7	0.30715	X		X			X	X	X	X		X	X	X		X	X
12	42.6	27.6	9.4	0.30945	X		X			X	X	X	X	X	X	X	X		X	X
12	42.6	27.6	9.5	0.30954	X		X	X		X	X	X	X	X	X	X			X	X
13	42.9	26.4	11.2	0.31205	X		X	X		X	X	X	X	X	X	X	X		X	X
13	42.8	26.3	11.3	0.31229	X	X	X			X	X	X	X	X	X	X	X		X	X
14	43.1	25.0	13.1	0.31502	X	X	X	X		X	X	X	X	X	X	X	X		X	X
14	43.0	24.9	13.1	0.31526	X		X	X		X	X	X	X	X	X	X	X	X	X	X
15	43.2	23.4	15.0	0.31843	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X
15	43.1	23.2	15.1	0.31866	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X
16	43.2	21.5	17.0	0.32220	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

(2-variable model, Mallows C-p = -2.4) has the smallest Cp.

In total, prints $16 \times 2 - 1 = 31$ models
(although the computer fits ALL models)

Sidenote: one can ~~use~~ employ the AIC criterion / BIC criterion
the smaller the AIC/BIC the better the model.

Regression Analysis: CGPA versus HSGPA, Exercise

The regression equation is
CGPA = 1.55 + 0.560 HSGPA - 0.0111 Exercise

Predictor	Coef	SE Coef	T	P
Constant	1.5489	0.5551	~~2.79~~	~~0.007~~
HSGPA	0.5599	0.1436	3.90	0.000
Exercise	-0.011138	0.004985	-2.23	0.029

} both p small — good

S = 0.306126   R-Sq = 31.6%   R-Sq(adj) = 29.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	2.4256	1.2128	12.94	0.000
Residual Error	56	5.2479	0.0937		
Total	58	7.6736			

→ at least one good pred

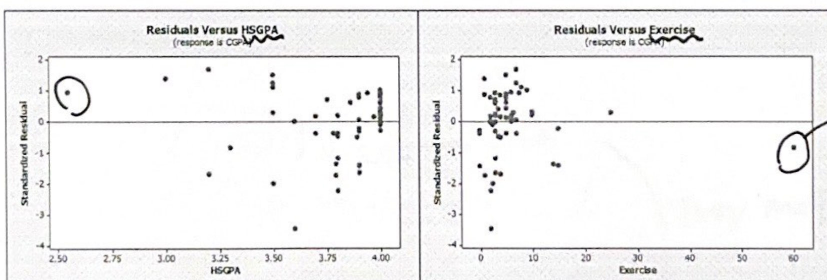
59-1

Unusual Observations

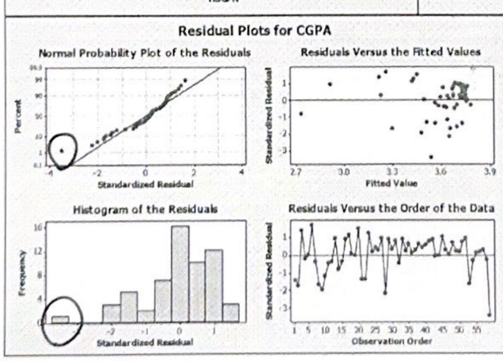
Obs	HSGPA	CGPA	Fit	SE Fit	Residual	St Resid
3	3.00	3.6000	3.2176	0.1297	0.3824	1.38 X
9	3.50	2.8800	3.4808	0.0642	-0.6008	-2.01R
14	3.30	2.6000	2.7284	0.2647	-0.1284	-0.83 X
27	2.55	3.1400	2.9099	0.1840	0.2301	0.94 X
28	3.80	2.9800	3.6544	0.0445	-0.6744	-2.23R
59	3.60	2.5000	3.5424	0.0556	-1.0424	$\boxed{-3.46R}$

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

Residual Plots for CGPA



Influential outlier
one person exercises
60 hrs/wk



Regression Analysis: CGPA versus HSGPA, Exercise

```
The regression equation is
CGPA = 1.54 + 0.554 HSGPA - 0.00432 Exercise
Predictor        Coef    SE Coef      T      P
Constant       1.5388     0.5568   2.76  0.008
HSGPA          0.5542     0.1441   3.85  0.000
Exercise    -0.004320   0.009596  -0.45  0.654
```

NOT SIG!

One person was removed (Ex=60hrs/wk) because that person was an influential outlier.

```
S = 0.306969   R-Sq = 21.9%   R-Sq(adj) = 19.0%

Analysis of Variance

Source          DF        SS       MS     F      P
Regression       2   1.45009  0.72504  7.69  0.001
Residual Error  55   5.18265  0.09423
Total           57   6.63274
```

58-1 because we removed one person

```
Unusual Observations
Obs  HSGPA    CGPA     Fit  SE Fit  Residual  St Resid
  3   3.00  3.6000  3.1970  0.1324    0.4030     1.45 X
 25   3.50  3.3100  3.3705  0.1974   -0.0605    -0.26 X
 26   2.55  3.1400  2.9261  0.1856    0.2139     0.87 X
 27   3.80  2.9800  3.6361  0.0497   -0.6561    -2.17R
 58   3.60  2.5000  3.5252  0.0594   -1.0252    -3.40R
```

Maybe need to remove this one as well

```
R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.
```

Throw out Exercise and fit a SLR model on HSGPA

Regression Analysis: CGPA versus HSGPA

```
The regression equation is
CGPA = 1.50 + 0.560 HSGPA

Predictor     Coef  SE Coef     T      P
Constant    1.4964   0.5448  2.75  0.008
HSGPA       0.5596   0.1426  3.92  0.000
```

→ good

They match

```
S = 0.304776   R-Sq = 21.6%   R-Sq(adj) = 20.2%

Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  1.4310  1.4310  15.41  0.000
Residual Error  56  5.2017  0.0929
Total           57  6.6327
```

→ good

```
Unusual Observations

Obs  HSGPA    CGPA     Fit  SE Fit  Residual  St Resid
  3   3.00  3.6000  3.1753  0.1223    0.4247     1.52 X
 26   2.55  3.1400  2.9234  0.1842    0.2166     0.89 X
 27   3.80  2.9800  3.6230  0.0400   -0.6430    -2.13R
 58   3.60  2.5000  3.5111  0.0500   -1.0111    -3.36R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.
```

The model is still NOT ideal. Possible next steps:

- Take out obs 58 and refit the model
- Gather more data
- Come up with more "useful" predictors
- Add higher-order terms
- Use other models (something "fancier" :))