Chapter 9 Model Selection and Validation

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

Salary example in proc glm

Model salary ($1000) as function of age in years, years post-high school education (educ), & political affiliation (pol), pol = D for Democrat, pol = R for Republican, and pol = O for other.

```
data salary;
input salary age educ pol$ @@;
datalines;
38 25 4 D 45 27 4 R 28 26 4 O 55 39 4 D 74 42 4 R 43 41 4 O
47 25 6 D 55 26 6 R 40 29 6 O 65 40 6 D 89 41 6 R 56 42 6 O
56 32 8 D 65 33 8 R 45 35 9 O 75 39 8 D 95 65 9 R 67 69 10 O
;
options nocenter;
proc glm; class pol; model salary=age educ pol / solution; run;
```

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		26.19002631 B	7.89909191	3.32	0.0056
age		0.89834968	0.19677236	4.57	0.0005
educ		1.50394642	1.18414843	1.27	0.2263
pol	D	-9.15869409 B	4.84816554	-1.89	0.0814
pol	O	-25.69911504 B	4.75120999	-5.41	0.0001
pol	R	0.00000000 B	.	.	.

The model is

$$Y_i = \beta_0 + \underbrace{\beta_1 \text{age}_i + \beta_2 \text{educ}_i}_{\text{2 continuous}} + \underbrace{\beta_{31} I\{\text{pol}_i = \text{D}\} + \beta_{32} I\{\text{pol}_i = \text{O}\} + \beta_{33} I\{\text{pol}_i = \text{R}\}}_{\text{1 categorical}} + \epsilon_i$$

and the coefficient vector is $\beta' = (\beta_0, \beta_1, \beta_2, \beta_{31}, \beta_{32}, \underbrace{\beta_{33}}_{=0})$.

General linear test in SAS

- The `contrast` statement in SAS PROC GLM lets you test whether one or more linear combinations of regression effects are (simultaneously) zero.
- To test no difference between Democrats and Republicans, $H_0 : \beta_{31} = \beta_{33}$ equivalent to $H_0 : \beta_{31} - \beta_{33} = 0$, use `contrast "Dem=Rep" pol 1 0 -1;`. Need to include the "-1" *even though SAS sets* $\beta_{33} = 0$!
- To test no difference among all political affiliations, use $H_0 : \beta_{31} - \beta_{32} = 0$ *and* $H_0 : \beta_{32} - \beta_{33} = 0$, given by `contrast "Dem=Rep=Other" pol 1 -1 0, pol 0 1 -1;`.

```
proc glm; class pol; model salary=age educ pol / solution;
contrast "Dem=Rep" pol 1 0 -1;
contrast "Dem=Rep=Other" pol 1 -1 0, pol 0 1 -1;
-------------------------------------------------------------
Contrast                    DF     Contrast SS     Mean Square    F Value    Pr > F
Dem=Rep                      1      240.483581      240.483581       3.57    0.0814
Dem=Rep=Other                2     2017.608871     1008.804436      14.97    0.0004
```

General linear test in SAS

- We can also test quadratic effects and interactions.
- From the initial fit, `educ` is not needed with `age` and `pol` in the model. Let's refit:

```
proc glm; class pol; model salary=age pol / solution; run;
-----------------------------------------------------------
Source                     DF     Type III SS     Mean Square    F Value    Pr > F
age                         1     2648.275862     2648.275862      37.65    <.0001
pol                         2     1982.208197      991.104098      14.09    0.0004

                                                  Standard
Parameter              Estimate                      Error    t Value    Pr > |t|
Intercept         30.15517241 B      7.41311553       4.07      0.0012
age                1.03448276       0.16859121       6.14      <.0001
pol       D       -8.63793103 B      4.93543380      -1.75      0.1020
pol       O      -25.37931034 B      4.84730261      -5.24      0.0001
pol       R        0.00000000 B       .                .          .
```

The Type III SS test $H_0 : \beta_1 = 0$ and $H_0 : \beta_{21} = \beta_{22} = \beta_{23} = 0$ in

$$Y_i = \beta_0 + \beta_1 \text{age}_i + \beta_{21} I\{\text{pol}_i = \text{D}\} + \beta_{22} I\{\text{pol}_i = \text{O}\} + \beta_{23} I\{\text{pol}_i = \text{R}\} + \epsilon_i$$

Drop quadratic effects?

A test of the main effects model versus the quadratic model

```
proc glm; class pol;
 model salary=age pol age*pol age*age / solution;
 contrast "drop quadratic effects?" age*age 1, age*pol 1 -1 0, age*pol 1 0 -1;
```

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
drop quadratic effects?	3	376.8443881	125.6147960	2.27	0.1369

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		-22.13053948	B	25.12432158	-0.88	0.3972
age		3.46694474	B	1.16442934	2.98	0.0126
pol	D	1.18699006	B	21.44129001	0.06	0.9568
pol	O	-15.72146564	B	13.51918833	-1.16	0.2695
pol	R	0.00000000	B	.	.	.
age*pol	D	-0.28943698	B	0.61955938	-0.47	0.6495
age*pol	O	-0.23843048	B	0.32387727	-0.74	0.4770
age*pol	R	0.00000000	B	.	.	.
age*age		-0.02513595		0.01254539	-2.00	0.0704

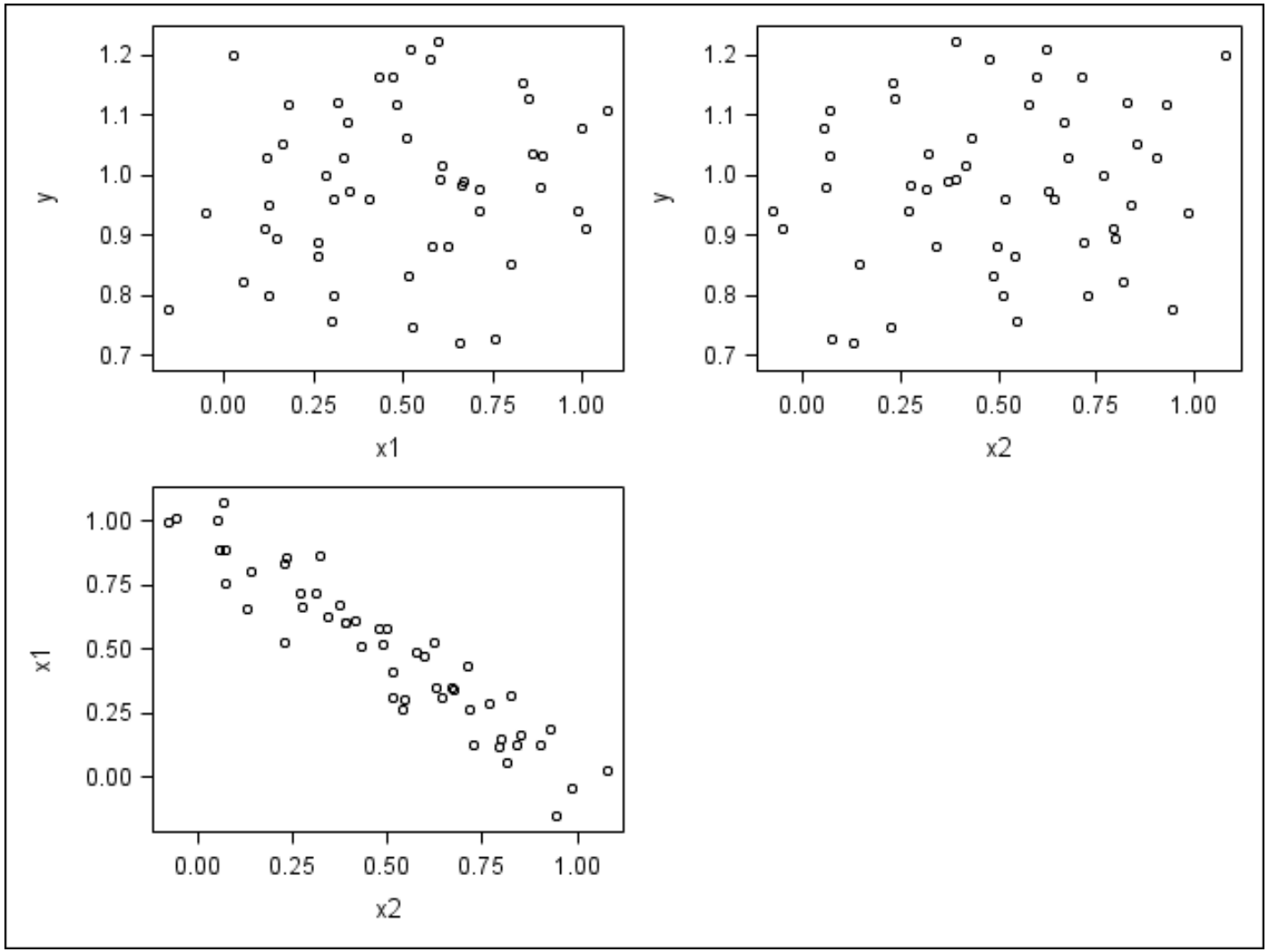
We'll work this out on the board. We can drop the quadratic effects (p=0.137), although there's some indication in the table of regression effects that age_i^2 might be needed.

Scatterplots

- Scatterplots show the *marginal* relationship between Y and each of the $x_1, \ldots, x_k$. They *cannot* show you anything about the *joint* relationship among the $Y, x_1, \ldots, x_k$.
- If a nonlinear relationship between Y and x_j $(j = 1, \ldots, k)$ *marginally* may or may not be present in the *joint* relationship.
- Actually, any strong relationship between Y and x_j marginally doesn't mean that x_j will be needed in the presence of other variables.
- Seeing no marginal relationship between Y and x_j *does not* mean that x_j is not needed in a model including other predictors.

No relationship?

Here Y vs. x_1 and Y vs. x_2 shows nothing. There seems to be some multicollinearity though.



proc reg output

x_1 important marginally? $Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.94576	0.03602	26.26	<.0001
x1	1	0.06974	0.06311	1.11	0.2745

x_2 important marginally? $Y_i = \beta_0 + \beta_2 x_{i2} + \epsilon_i$

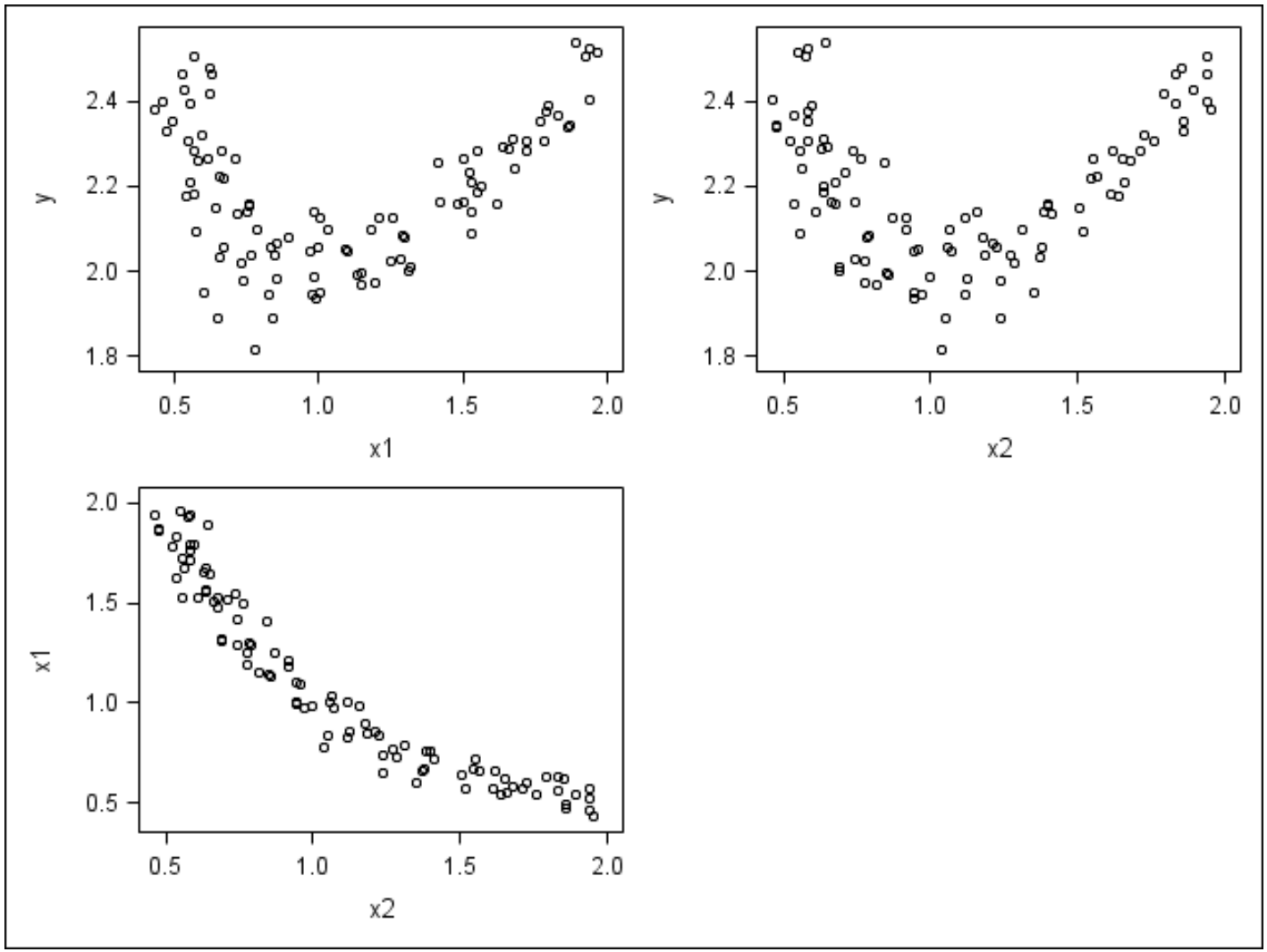
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.95180	0.03730	25.52	<.0001
x2	1	0.05603	0.06458	0.87	0.3898

x_1, x_2 important jointly? $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.08151	0.10876	-0.75	0.4572
x1	1	1.07327	0.11065	9.70	<.0001
x2	1	1.08548	0.11271	9.63	<.0001

Nonlinear relationship?

Marginally, x_1 and x_2 have highly nonlinear relationships with Y. Should we transform?



proc reg output

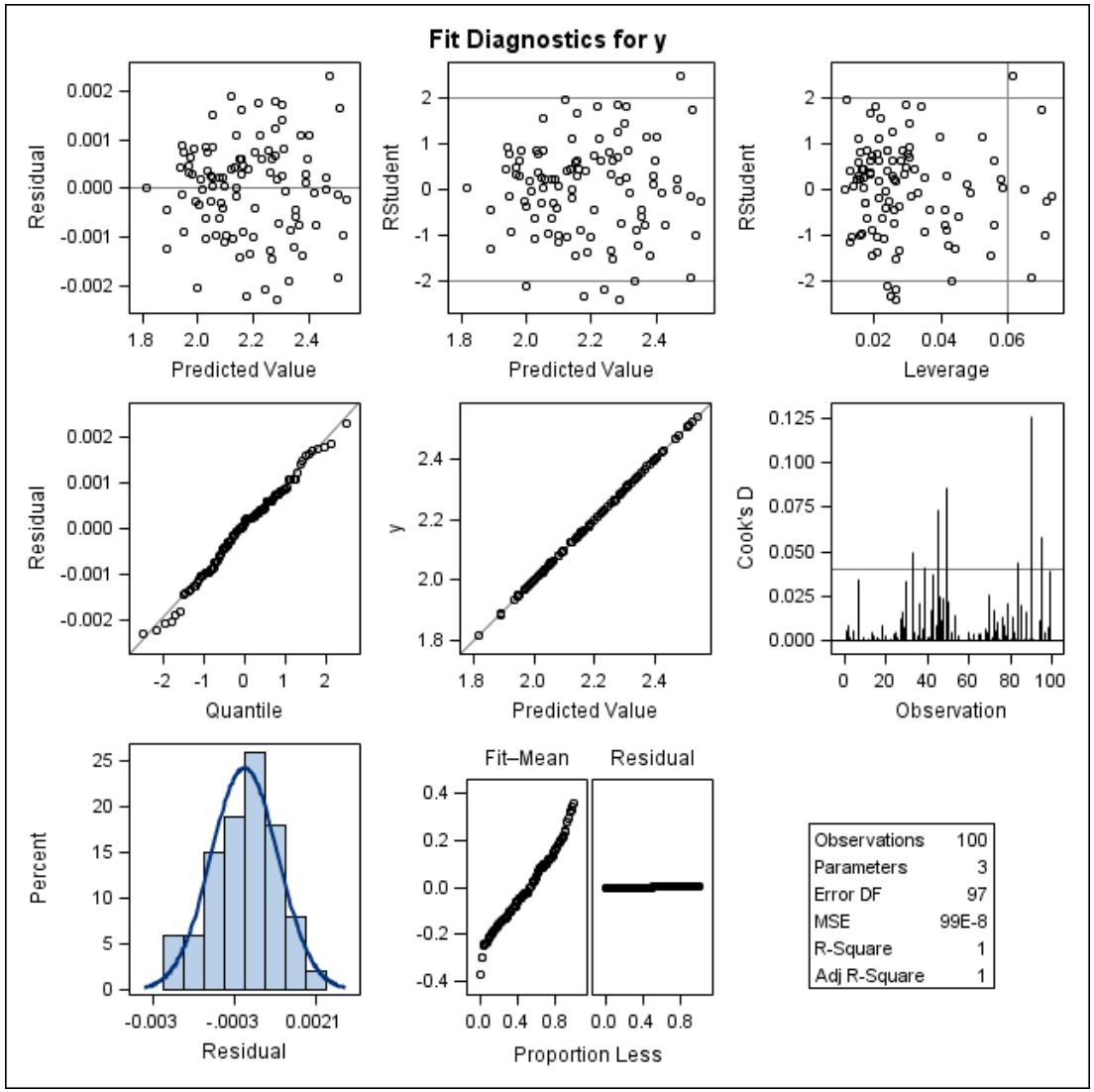
Let's try fitting a simple main effects model without any transformation.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.00036626	0.00130	-0.28	0.7791
x1	1	1.00022	0.00059936	1668.80	<.0001
x2	1	1.00009	0.00060998	1639.54	<.0001

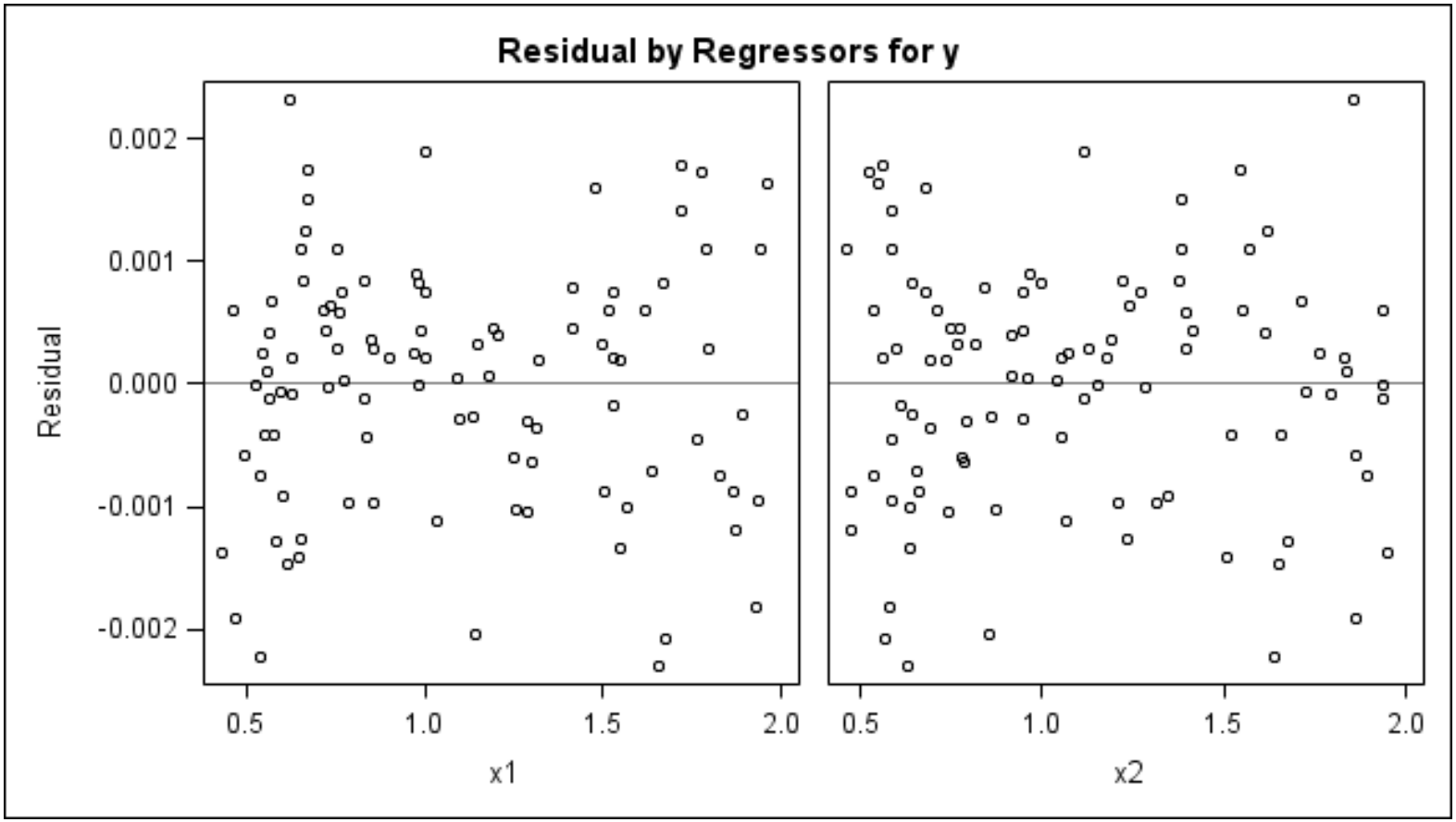
both x_1 and x_2 are important, but does the model fit okay?

Model fit is okay



Look at Y_i vs. $\hat{Y}_i$ and R^2!

No pattern here, either



9.1 Model building overview (pp. 343–349)

Chapter 9 Model variable selection and validation

Book outlines four steps in data analysis

1. Data collection and preparation (acquiring and "cleaning").
2. Reduction of explanatory variables (for exploratory observational studies). Mass screening for "decent" predictors.
3. Model refinement and selection.
4. Model validation.

We usually get data after step 1.

9.1 Model building overview

- Book has flowchart for model building process on p. 344.
- *Designed experiments* are typically easy; experimenter *manipulates* treatment variables during experiment (and expects them to be significant); other variables are collected to adjust for.
- With *confirmatory* observational studies, the goal is to determine whether (or how) the response is related to one or more pre-specified explanatory variables. No need to weed them out.
- *Exploratory* observational studies are done when we have little previous knowledge of exactly which predictors are related to the response. Need to "weed out" good from useless predictors.
- We may have a list of *potentially* useful predictors; *variable selection* can help us "screen out" useless ones and build a good, predictive model.

Controlled experiments

- These include clinical trials, laboratory experiments on monkeys and pigs, etc., community-based intervention trials, etc.
- The experimentors control one or more variables that are related to the reponse. Often these variables are "treatment" and "control." Can ascribe causality if populations are the same except for the control variables.
- Sometimes other variables (not experimentally assigned) that may also affect the response are collected too, e.g. gender, weight, blood chemistry levels, viral load, whether other family members smoke, etc.
- When building the model the treatment is always included. Other variables are included as needed to reduce variability and zoom in on the treatment factors. Some of these variables may be useful and some not, so part of the model building process is weeding out "noise" variables.

Confirmatory observational studies

- Used to test a hypothesis built from other studies or a "hunch."
- Variables involved in the hypothesis (amount of fiber in diet) that affect the response (cholesterol) are measured along with other variables that can affect the outcome (age, exercise, gender, race, etc.) – nothing is controlled. Variables involved in the hypothesis are called *primary* variables; the others are called risk factors; epidemiologists like to "adjust" for "risk factors."
- Note that your book discusses Vitamin E and cancer on p. 345. Recall what Stan Young discussed in his seminar a few weeks back?
- Usually all variables are retained in the analysis; they were chosen ahead of time.

Observational studies

- When people are involved, often not possible to conduct controlled experiments.
- Example: maternal smoking affects infant birthweight. One would have to randomly allocate the treatments "smoking" and "non-smoking" to pregnant moms – ethical problems.
- Investigators consider *anything* that is easy to measure that might be related to the response. Many variables are considered and models painstakenly built. Often called "data dredging."

Observational studies

- There's a problem here – one is sure to find *something* if they look hard enough. Often "signals" are there spuriously, and sometimes *in the wrong direction*.
- The number of variables to consider can be large; there can be high multicollinearity. Keeping too many predictors can make prediction *worse*.
- Your textbook says "*The identification of "good"...variables to be included in the...regression model and the determination of appropriate functional and interaction relations...constitute some of the most difficult problems in regression analysis.*"

Section 9.2: Surgical unit example

- First steps often involve plots:
  - Plots to indicate correct functional form of predictors and/or response.
  - Plots to indicate possible interaction.
  - Exploration of correlation among predictors (maybe).
  - Often a first-order model is a good starting point.
- Once a reasonable set of potential predictors is identified, formal model selection begins.
- If the number of predictors is large, say $k \geq 10$, we can use (automated) stepwise procedures to reduce the number of variables (and models) under consideration.

9.3 Model selection (pp. 353–361)

Once we reduce the set of potential predictors to a reasonable number, we can examine all possible models and choose the "best" according to some criterion.

Say we have k predictors $x_1, \ldots, x_k$ and we want to find a good subset of predictors that predict the data well. There are several useful criteria to help choose a subset of predictors.

Adjusted-R^2, R^2_a

"Regular" R^2 measures how well the model predicts the data that built it. It is possible to have a model with $R^2 = 1$ (predicts the data that built it perfectly), but has *lousy out-of-sample prediction*. The adjusted R^2, denoted R^2_a provides a "fix" to R^2 to provide a measure of how good the model will predict data not used to build the model. For a candidate model with $p-1$ predictors

$$R^2_a = 1 - \frac{n-1}{n-p}\frac{SSE_p}{SSTO}\left(= 1 - \frac{MSE_p}{s^2_y}\right).$$

- Equivalent to choosing the model with the *smallest* MSE_p.
- If irrelevant variables are added, R^2_a may decrease unlike "regular" R^2 (R^2_a can be negative!).
- R^2_a penalizes model for being too complex.
- Problem: R^2_a is greater for a "bigger" model whenever the F-statistic for comparing bigger to smaller is greater than 1. We usually want F-statistics to be a lot bigger than 1 before adding in new predictors $\Rightarrow$ *too liberal*.

AIC

Choose model with smallest Akaike Information Criterion (AIC). For normal error model,

$$AIC = n \log(SSE_p) - n \log(n) + 2p.$$

- $n \log(SSE_p) - n \log(n) = C - 2 \log\{\mathcal{L}(\hat{\beta}, \hat{\sigma}^2)\}$ from the normal model where C is a constant; we'll show this on the board.
- $2p$ is "penalty" term for adding predictors.
- Like R_a^2, AIC favors models with small SSE, but penalizes models with too many variables p.

SBC (or BIC)

Models with smaller Schwarz Bayesian Criterion (SBC) are estimated to predict better. SBC is also known as *Bayesian Information Criterion*:

$$BIC = n\log(SSE_p) - n\log(n) + p\log(n).$$

- BIC is similar to AIC, but for $n \geq 8$, the BIC "penalty term" is more severe.
- Chooses model that "best predicts" the observed data according to asymptotic criteria.

Mallow's C_p

Let F be the full model with all k predictors and R be a reduced model with $p-1$ predictors to be compared to the full model. Mallows C_p is

$$C_p = \frac{SSE(R)}{MSE(F)} - n + 2p.$$

- Measures the bias in the reduced regression model relative full model having all k candidate predictors.
- The full model is chosen to provide an unbiased estimate $\hat{\sigma}^2 = MSE(x_1, \ldots, x_k)$. Predictors must be in "correct form" and important interactions included.
- If a reduced model is unbiased, $E(\hat{Y}_i) = \mu_i$, then $E(C_p) = p$ (pp. 357–359).
- The full model always has $C_p = k + 1$.
- If $C_p \approx p$ then the reduced model predicts as well as the full model. If $C_p < p$ then the reduced model is estimated to predict *better* than the full model.
- In practice, just choose model with smallest C_p.

Which criteria to use?

R_a^2, AIC, BIC, and C_p may given different "best" models, or they may agree. Ultimate goal is to find model that balances:

- A good fit to the data.
- Low bias.
- Parsimony.

All else being equal, the simpler model is often easier to interpret and work with. Christensen (1996) recommends C_p and notes the similarity between C_p and AIC.

Two methods for "automatically" picking variables

- Any regression textbook will caution against not thinking about the data at all and simply using automated procedures.
- Automated procedures cannot assess a good functional form for a predictor, cannot think about which interactions might be important, etc.
- Anyway, automated procedures are widely used and *can* produce good models. They can also produce models that are *substantially inferior* to other models built from the same predictors using scientific input and common sense.

Two methods for "automatically" picking variables

- Two methods are **best subsets** and **stepwise** procedures.
- Best subsets simply finds the models that are best according to some statistic, e.g. smallest C_p of a given size. `proc reg` does this automatically, but does not enforce hierarchical model building.
- Stepwise procedures add and/or subtract variables one at a time according to prespecified inclusion/exclusion criteria. Useful when you have a very large number of variables (e.g. $k > 30$). Both `proc reg` and `proc glmselect` incorporate stepwise methods.

Best subsets for blood pressure data Problem 9.13

- Increased arterial blood pressure in lungs can lead to heart failure in patients with chronic obstructive pulmonary disease (COPD).
- Determining arterial lung pressure is invasive, difficult, and can hurt patient.
- Radionuclide imaging is noninvasive, less risky way to estimate arterial pressure in lungs.
- A cardiologist measured three potential proxies and the invasive measure on $n = 19$ COPD patients.
  1. x_1 = emptying rate of blood into the pumping chamber of the heart
  2. x_2 = ejection rate of blood pumped out of the heart into the lungs
  3. x_3 = a blood gas.
  4. Y = invasive measure of systolic pulmonary arterial pressure

Best subsets using C_p

```
data lung;
input y x1 x2 x3 @@; x12=x1*x2; x13=x1*x3; x23=x2*x3; x1sq=x1*x1; x2sq=x2*x2; x3sq=x3*x3;
label y="pulmonary arterial pressure" x1="emptying rate" x2="ejection rate" x3="blood gas";
datalines;
   49.0   45.0   36.0   45.0   55.0   30.0   28.0   40.0   85.0   11.0   16.0   42.0
   32.0   30.0   46.0   40.0   26.0   39.0   76.0   43.0   28.0   42.0   78.0   27.0
   95.0   17.0   24.0   36.0   26.0   63.0   80.0   42.0   74.0   25.0   12.0   52.0
   37.0   32.0   27.0   35.0   31.0   37.0   37.0   55.0   49.0   29.0   34.0   47.0
   38.0   26.0   32.0   28.0   41.0   38.0   45.0   30.0   12.0   38.0   99.0   26.0
   44.0   25.0   38.0   47.0   29.0   27.0   51.0   44.0   40.0   37.0   32.0   54.0
   31.0   34.0   40.0   36.0
;
* best subset in proc reg, show 5 models with smallest Cp out of all possible models;
proc reg; model y=x1 x2 x3 x12 x13 x23 x1sq x2sq x3sq / selection=cp best=5;
```

Number in Model	C(p)	R-Square	Variables in Model
3	-0.0561	0.7922	x1 x2 x12
3	0.6717	0.7784	x1 x2 x1sq
4	1.2140	0.8061	x1 x2 x1sq x2sq
4	1.3025	0.8044	x1 x3 x23 x1sq
4	1.4108	0.8023	x1 x13 x23 x1sq

Only models x_1, x_2, x_1x_2; x_1, x_2, x_1^2; and x_1, x_2, x_1^2, x_2^2 are hierarchical.

9.4 automated variable search (pp. 361–368)

Forward stepwise regression (pp. 364–365)
We start with k potential predictors $x_1, \ldots, x_k$. We add and delete predictors one at a time until all predictors are significant at some preset level. Let α_e be the significance level for adding variables, and α_r be significance level for removing them.

Note: We should choose $\alpha_e < \alpha_r$; in book example $\alpha_e = 0.1$ & $\alpha_r = 0.15$.

Forward stepwise regression

1. Regress Y on x_1 only, Y on x_2 only, up to Y on x_k only. In each case, look at the p-value for testing the slope is zero. Pick the x variable with the smallest p-value to include in the the model.
2. Fit all possible 2-predictor models (in general j-predictor models) than include the initially chosen x, along with each remaining x variable in turn. Pick new x variable with smallest p-value for testing slope equal to zero in model that already has first one chosen, as long as p-value $< \alpha_e$. Maybe nothing is added.
3. Remove the x variable with the *largest* p-value as long as p-value $> \alpha_r$. Maybe nothing is removed.
4. Repeat steps (2)-(3) until no x variables can be added or removed.

proc glmselect

- *Forward selection* and *backwards elimination* are similar procedures; see p. 368. I suggest stepwise of the three.
- proc glmselect implements automated variable selection methods for regression models.
- Does stepwise, backwards, and forwards procedures as well as least angle regression (LAR) and lasso. Flom and Casell (2007) recommend either of these last two overall traditional stepwise approaches & note they both do about the same.
- The syntax is the same as `proc glm`, and you can include `class` variables, interactions, etc.

proc glmselect

- The `hier=single` option builds hierarchical models. To do stepwise as in your textbook cutoffs suggested in your textbook include `select=sl`, also `sle=0.1` is entry cutoff and `sls=0.15` is cutoff for staying in the model (used by your book). You can also do model selection using any of AIC, BIC, C_p, R_a^2 rather than p-value cutoffs.
- proc glmselect will stop when you cannot add or remove any predictors, but the "best" model may have been found in an earlier iteration. Using `choose=cp`, for example, gives the model with the lowest C_p as the final model, regardless where the procedure stops.
- `include=`p includes the first p variables listed in the model statement in every model. Why might this be necessary?
- Salary data: stepwise selection, choosing hierarchical model with smallest C_p during stepwise procedure (which happens to be at the end!)

```
proc glmselect;
model y=x1 x2 x3 x1*x1 x2*x2 x3*x3 x1*x2 x1*x3 x2*x3 /
      selection=stepwise(select=sl choose=cp sle=0.1 sls=0.15) hier=single;
```

proc glmselect output

The GLMSELECT Procedure

Selection Method	Stepwise
Select Criterion	Significance Level
Stop Criterion	Significance Level
Choose Criterion	C(p)
Entry Significance Level (SLE)	0.1
Stay Significance Level (SLS)	0.15
Effect Hierarchy Enforced	Single

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	CP	F Value	Pr > F
0	Intercept		1	1	211.6955	0.00	1.0000
1	age		2	2	111.4284	13.05	0.0023
2	pol		3	4	31.6296	14.09	0.0004
3	age*age		4	5	19.2714	6.84	0.0213
4	educ		5	6	12.1660	6.01	0.0305
5	educ*educ		6	7	6.2608*	8.47	0.0142

* Optimal Value Of Criterion

Selection stopped because the candidate for entry has SLE > 0.1 and the candidate for removal has SLS < 0.15.

Stop Details

Candidate For	Effect	Candidate Significance		Compare Significance	
Entry	age*educ	0.3056	>	0.1000	(SLE)
Removal	age*age	0.0266	<	0.1500	(SLS)

proc glmselect output

The selected model, based on C(p), is the model at Step 5.

Effects: Intercept age educ age*age educ*educ pol

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	6	5143.28668	857.21445	38.85
Error	11	242.71332	22.06485	
Corrected Total	17	5386.00000		

Root MSE	4.69732	Dependent Mean	57.66667
R-Square	0.9549	Adj R-Sq	0.9304
AIC	80.82717	AICC	96.82717
BIC	72.24885	C(p)	6.26081
SBC	67.05977		

Parameter Estimates

Parameter		DF	Estimate	Standard Error	t Value
Intercept		1	-66.482220	18.342169	-3.62
age		1	2.787032	0.626151	4.45
educ		1	18.751324	5.739109	3.27
age*age		1	-0.018677	0.007298	-2.56
educ*educ		1	-1.342341	0.461108	-2.91
pol	D	1	-9.495127	2.790631	-3.40
pol	O	1	-23.472038	2.813063	-8.34
pol	R	0	0	.	.

```
proc glm; class pol;
 model salary=age age*age educ educ*educ pol / solution;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
age	1	437.146692	437.146692	19.81	0.0010
age*age	1	144.511058	144.511058	6.55	0.0266
educ	1	235.546067	235.546067	10.68	0.0075
educ*educ	1	186.991457	186.991457	8.47	0.0142
pol	2	1547.552133	773.776067	35.07	<.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		-66.48222024	B	18.34216851	-3.62	0.0040
age		2.78703226		0.62615068	4.45	0.0010
age*age		-0.01867654		0.00729787	-2.56	0.0266
educ		18.75132386		5.73910890	3.27	0.0075
educ*educ		-1.34234095		0.46110774	-2.91	0.0142
pol	D	-9.49512746	B	2.79063098	-3.40	0.0059
pol	O	-23.47203787	B	2.81306286	-8.34	<.0001
pol	R	0.00000000	B	.	.	.

Moral: when a predictor (e.g. education) is not included in proper functional form, it can be missed if one is looking for important main effects only.

Stepwise procedures vs. best subsets

- Forwards selection, backwards elimination, and stepwise procedures are designed for very large numbers of variables.
- Best subsets work well when the number of potential variables is smaller. Say have k predictors. The number of possible 9-variable models is $\begin{pmatrix} 10 \\ 9 \end{pmatrix} = 10$, the number of 8-variable models is $\begin{pmatrix} 10 \\ 8 \end{pmatrix} = 45$, 120 7-variable, 210 6-variable, 252 5-variable, 210 4-variable, etc.
- In `proc reg` you can find best subsets, but SAS will not weed out non-hierarchical models.

Stepwise with lung pressure data, proc glmselect

Implemented as described in your textbook.

```
* stepwise until all effect sig. at 0.1 and 0.15 levels, stop when cannot enter or remove
  variable & choose that model;
proc glmselect;
model y=x1 x2 x3 x1*x1 x2*x2 x3*x3 x1*x2 x1*x3 x2*x3 / selection=stepwise(select=sl stop=sl
                                                                          sle=0.1 sls=0.15) hier=single;
```

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	F Value	Pr > F
0	Intercept		1	0.00	1.0000
1	x2		2	21.54	0.0002
2	x2*x2		3	6.16	0.0246

Selection stopped because the candidate for entry has SLE > 0.1 and the candidate for removal has SLS < 0.15.

Stop Details

Candidate For	Effect	Candidate Significance		Compare Significance	
Entry	x1	0.3805	>	0.1000	(SLE)
Removal	x2*x2	0.0246	<	0.1500	(SLS)

This model, x_2, x_2^2 has $C_p = 3.8$, much larger than best model found using "best subsets."

Stepwise with lung pressure data, `proc reg`

Implemented as described in your textbook, *but non-hierarchical*.

```
proc glmselect;
proc reg;
 model y=x1 x2 x3 x12 x13 x23 x1sq x2sq x3sq / selection=stepwise sle=0.1 sls=0.15;
```

```
                                The REG Procedure

                            Stepwise Selection: Step 1

                 Variable x2 Entered: R-Square = 0.5589 and C(p) = 8.2349

                          Parameter     Standard
Variable                   Estimate        Error   Type II SS  F Value  Pr > F
Intercept                  72.87601      7.19467        21533   102.60  <.0001
x2                         -0.67707      0.14590   4519.89726    21.54  0.0002
```

```
                            Stepwise Selection: Step 2

                Variable x2sq Entered: R-Square = 0.6814 and C(p) = 3.7784

                          Parameter     Standard
Variable                   Estimate        Error   Type II SS  F Value  Pr > F
Intercept                 104.09098     14.07022   8812.77920    54.73  <.0001
x2                         -2.12876      0.59884   2034.77821    12.64  0.0026
x2sq                        0.01327      0.00535    991.41494     6.16  0.0246
```

Stepwise with lung pressure data, proc reg

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1000 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value
1	x2		ejection rate	1	0.5589	0.5589	8.2349	21.54
2	x2sq			2	0.1226	0.6814	3.7784	6.16

Summary of Stepwise Selection

Step	Pr > F
1	0.0002
2	0.0246

Both `proc glmselect` and `proc reg` do stepwise. Only `proc reg` does best subsets. Only `proc glmselect` does stepwise hierarchical model building, LASSO and LAR. Choose `proc glmselect` for "large p" problems and choose `proc reg` for smaller numbers of predictors, e.g. $k < 30$.