$\boxed{\text{Chapter 9}}$ Model Building

1. Data collection and preparation
2. Reduction of predictor variables
3. Model refinement and selection
4. Model validation

$\boxed{\text{4 Types of studies}}$

Controlled: · Experimenter - controls levels of explanatory variables
- assigns treatments

Controlled w/ covariates: Same as controlled experiment except here we consider attributes of participants, ie age, education, ad hoc - after the experiment

Confirmatory-Observational: · Based on observational data (not experimental) used to test hunches
· Not controlled as the two above
· control variables: known risk factors
· explanatory variables: hunch factors

Exploratory Observational: · Looking for predictor variables that are related to a desired response variable

Data prep: · Check for outliers

$\boxed{\text{Reduction of explanatory variables}}$

Controlled: usually not important

controlled w/ covariates: covariates may need reduction

confirmatory: usually not important

exploratory: IMPORTANT - there may be a lot of explanatory variables so we can try to find a 'best subset'

Important variables left out are sometimes called latent vars

9.2 Example

1.) Screen for outliers

2.) Start with a first order model

(a) Does the residual plot indicate

i) non-constant variance?

ii) curvature?

(b) Does the Normality QQ Plot indicate

i) departures from normality?

(c) Fix issues

i) Try box-cox transformation to fix constant-variance

(d) Check that predictors are

i) linearly associated w/ response

ii) not too strongly associated

(e) Fix issues

i) Add higher order terms

ii) Collapse variables

(f) Refinement

i) Can we drop predictors?

ii) Is there an adequete subset?

- From any set of $p-1$ predictors there are 2^{p-1} models to choose from

- R_p^2 or SSE_p Criterion $\quad R_p^2 = 1 - \frac{SSE_p}{SSTO}$
  - 'Good' when SSE is small and R_p^2 is close to 1
  - Used to see when adding variables is no longer useful
- $R_{a,p}^2$ or MSE_p Criterion $\quad R_{a,p}^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE_p}{SSTO} = 1 - \frac{MSE_p}{SSTO/n-1)}$
  - 'Good' when $R_{a,p}^2$ is close to 1

Mallow' C_p Criterion Concerned w/ total mean squared error $\forall$ subset regression model

$$C_p = \frac{SSE_p}{MSE(X_1, X_2, \ldots X_{p-1})} - (n-2p)$$

$E(C_p) \approx p$ when $E(\hat{Y}_i) = \mu_i$

* Models with little bias will fall near $C_p = p$ line
* Models with bias will fall considerably above $C_p = p$

AIC_p and SBC_p Criteria

$AIC_p = n\ln(SSE_p) - n\ln(n) + 2p$

$SBC_p = n\ln(SSE_p) - n\ln(n) + \ln(n)p$

* Smaller = better for goodness of fit
* Both penalizes models w/ a lot of predictions

$Press_p$ Criterion

$Press_p = \sum_{i}^{n}(Y_i - \hat{Y}_{i(i)})^2$ - Sum of Squared prediction error

SAS

```
proc reg;
  model y= x1 x2 x3 / selection= cp aic sbc press;
run;
```

# of params in model	C(p)	R-sq	AIC	SBC	Variables in model
#					
#					
#					
⋮					

"Best" Subsets algorithm · Checks different models for C_p AIC_p SBC_p and $PRESS_p$ to get a small subset of "good" models

SAS

```
proc reg;
  model y = x1 x2 x3 / selection = cp best = 3;
run;
```

* Note: we prefer hierarchical models

"Stepwise Regression Models" · When we have, say 30-40 variables and "Best" would be too cumbersome to run we use this

2 types
- Forward Selection
- Backward Elimination

Forward: 1) · Simple regression model for each predictor
- $t_k^* = \frac{b_k}{s\{b_k\}}$ is calculated
· largest t values yield candidate for addition

2) · Regression model with first candidate and each other pred.
* $t_k^* = \frac{b_k}{s\{b_k\}}$
· largest t values yield candidate

⋮ continue in this fashion

SAS

```
proc reg;
  model y = x1 x2 x3 / selection = stepwise slentry = .15 slstay = .20;
run;
```

backward or forward } can be used instead