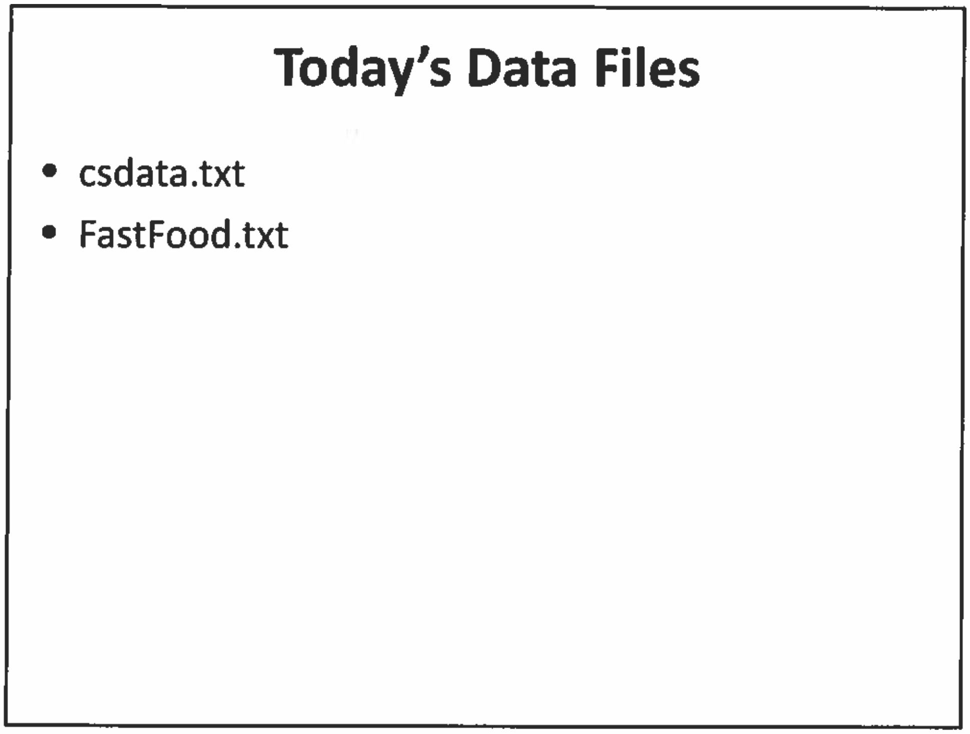
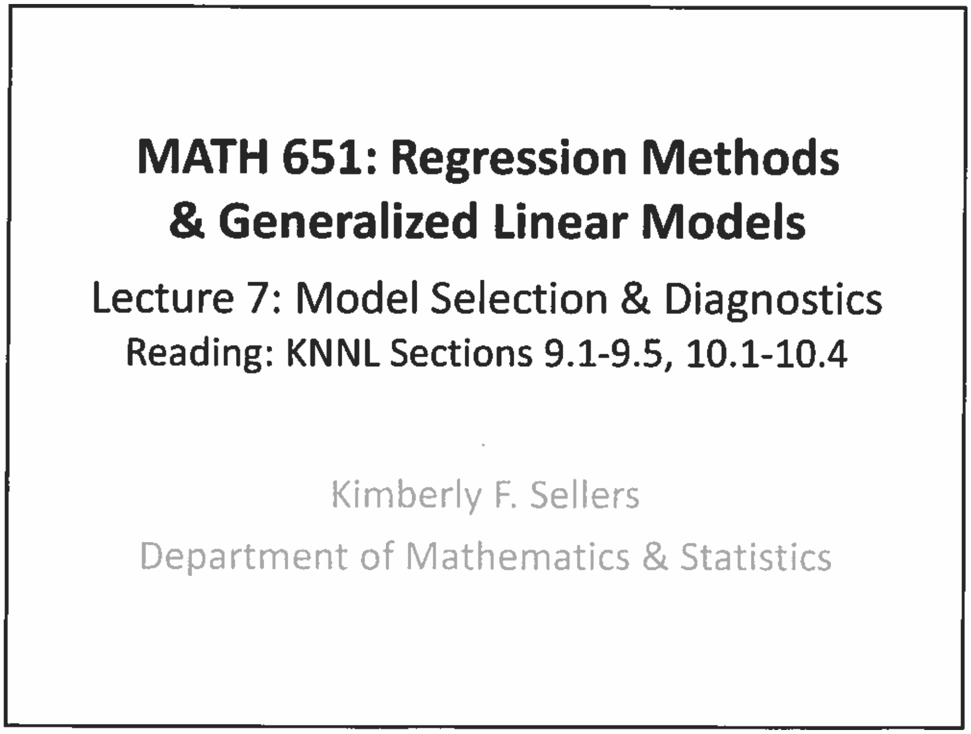


Today's Data Files

- csdata.txt
- FastFood.txt



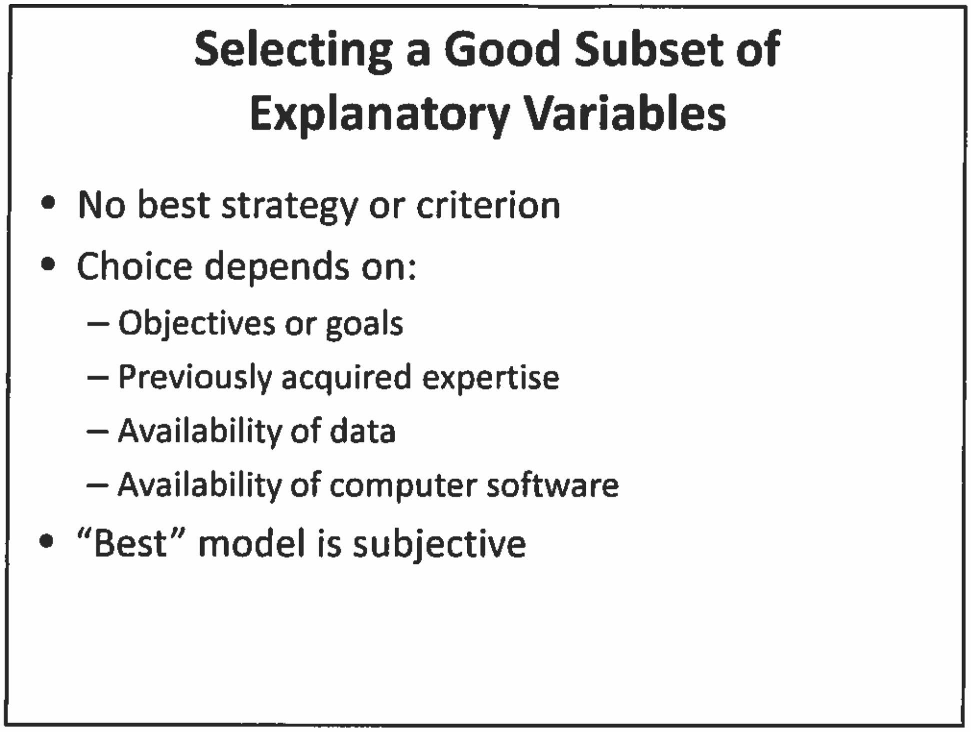
MATH 651: Regression Methods & Generalized Linear Models

Lecture 7: Model Selection & Diagnostics

Reading: KNNL Sections 9.1-9.5, 10.1-10.4

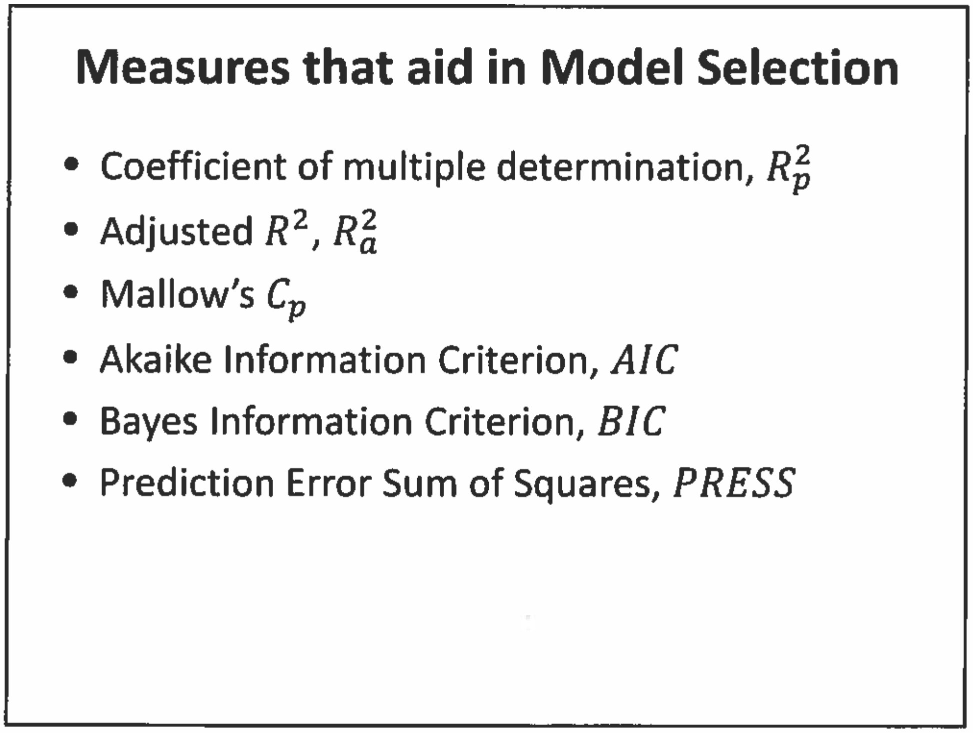
Kimberly F. Sellers

Department of Mathematics & Statistics



Selecting a Good Subset of Explanatory Variables

- No best strategy or criterion
- Choice depends on:
  - Objectives or goals
  - Previously acquired expertise
  - Availability of data
  - Availability of computer software
- "Best" model is subjective



Measures that aid in Model Selection

- Coefficient of multiple determination, R_p^2
- Adjusted R^2, R_a^2
- Mallow's C_p
- Akaike Information Criterion, AIC
- Bayes Information Criterion, BIC
- Prediction Error Sum of Squares, $PRESS$

R_p^2: Coeff of Multiple Determination

- $R_p^2 =$ coefficient of multiple determination for $p-1$ variables
$$R_p^2 = 1 - \frac{SSE_p}{SSTO}$$
- Always increases with additional X variables $\Rightarrow$ look for rate of increase
- R_p^2 most useful for comparing models containing the same number of explanatory variables

This is the usual R^2; the subscript allows you to keep track of how many coeffs in the model

R_a^2: Adjusted R^2

- Recall definition:
$$R_a^2 = 1 - \frac{(n-1)SSE_p}{(n-p)SSTO} = 1 - \frac{MSE_p}{SSTO/n-1}$$
- Choosing model with maximum R_a^2 equates to choosing model with minimum MSE
- R_a^2 will not necessarily increase with the inclusion of an additional variable

$V(X) = E(X^2) - E^2(X)$

Mallow's C_p Statistic

(Mallows (1973), *Technometrics*, 15, 661-675)

- Measures predictive ability of a fitted model
- Let $\hat{Y}_{ip}$ = predicted Y_i value associated with p-parameter model
- Total standardized MSE of prediction is

$$\Gamma_p = \frac{\sum_{i=1}^{n} E\left(\hat{Y}_{ip} - E(Y_i)\right)^2}{\sigma^2}$$

$$= \frac{\sum_{i=1}^{n}\left[E(\hat{Y}_{ip}) - E(Y_i)\right]^2 + \sum_{i=1}^{n} Var(\hat{Y}_{ip})}{\sigma^2}$$

  - First numerator term is the square of the bias; second term is the prediction variance

Mallow's C_p Statistic (cont.)

$$\Gamma_p = \frac{\sum_{i=1}^{n}\left[E(\hat{Y}_{ip}) - E(Y_i)\right]^2 + \sum_{i=1}^{n} Var(\hat{Y}_{ip})}{\sigma^2}$$

- Bias decreases as more variables enter model
  - Assume full model ($p = P$) is true, then $E\left(\hat{Y}_{ip}\right) - E(Y_i) = 0$, i.e. bias equals zero
- Prediction variance increases as more variables enter model ($\sum Var\left(\hat{Y}_{ip}\right) = p\sigma^2$)
- Tradeoff between bias and variance terms achieved by minimizing Γ_p

Mallow's C_p Statistic (cont.)

- Γ_p unknown because it involves β_is
- Consider estimate for Γ_p:

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - (n - 2p)$$

  - $\hat{\sigma}^2 = MSE(X_1, \ldots, X_{P-1})$, i.e. the MSE with all possible X variables in the model
  - (Almost) unbiased estimate of Γ_p
- As more variables are added to the model, SSE_p decreases but $2p$ increases
- No bias $\Rightarrow E(C_p) \approx p$, i.e. good models have C_p close to p

Mallow's C_p: Useful Guidelines

Hocking (1976), *Biometrics*, 32, 1-49.

- For prediction:
  - Consider models with $C_p \leq p$
- For parameter estimation:
  - Consider models with $C_p \leq 2p - (P - 1)$
    - Fewer variables should be eliminated from the model to avoid excess bias in the estimates

AIC: Akaike Information Criterion

Akaike (1969), *Annals of Inst. Stat. Math.*

$$AIC = n \ln\left(\frac{SSE_p}{n}\right) + 2p$$

- First RHS term decreases with increasing p
- Second RHS term = penalty paid for increasing the number of parameters in the model

- AIC represents tradeoff between precision of fit against number of parameters used
  - Look for models with small AIC
  - If AIC increases when a parameter is added to the model, the parameter is not needed

BIC: Bayes Information Criterion

Schwarz (1978)

- AIC tends to select models with larger numbers of parameters than the true model
- Alternatively, consider Bayes (or Schwarz) Information Criterion:

$$BIC = n \ln\left(\frac{SSE_p}{n}\right) + (\ln n)p$$

- Coefficient $\ln n$ (in front of p) tends to penalize more heavily models with a larger number of parameters as compared to AIC

PRESS: Prediction Error Sum of Squares

$$PRESS_p = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{i(i)}\right)^2$$

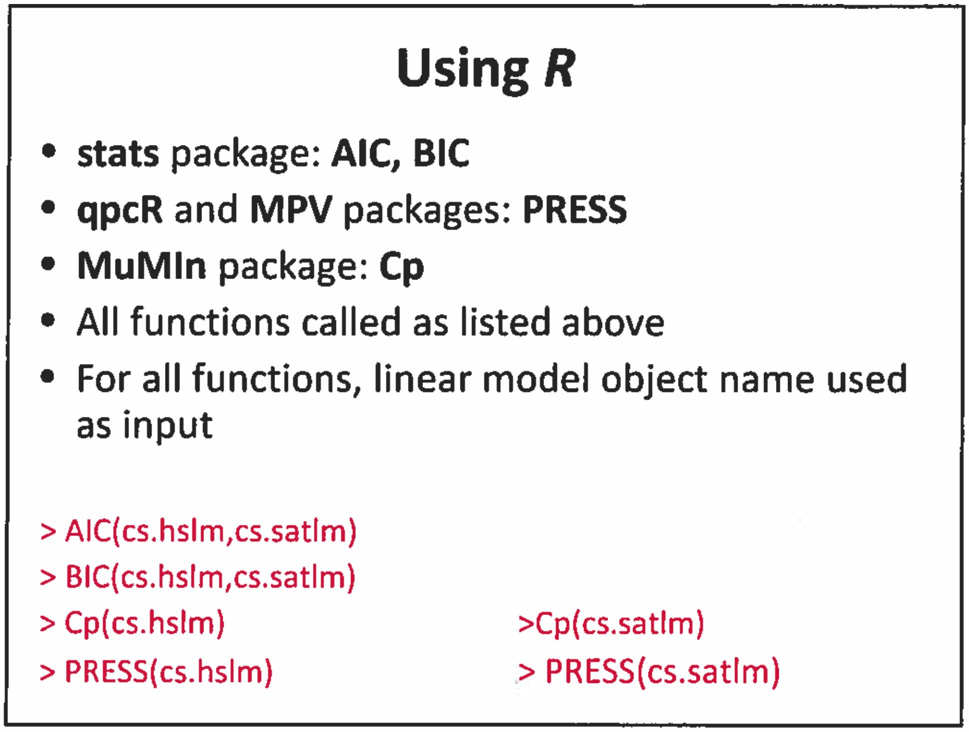
where $\hat{Y}_{i(i)}$ = prediction of ith response when ith observation is not used; obtained for the model with p parameters

- *PRESS* evaluates predictive ability of postulated model by omitting one observation at a time
- Want small $PRESS_p$ values for good models

Example

Data collected from a large university looks at first-year computer science majors in a particular year (**csdata.txt**, available on Blackboard). The purpose of the study was to attempt to predict success in the early university years. One measure of success was the cumulative grade point average (GPA) after three semesters. Among the explanatory variables recorded at the time the students enrolled in the university were average high school grades in mathematics (HSM), science (HSS), and English (HSE). We also have the students' SAT verbal (SATV) and mathematics scores (SATM). High school grades are coded on scale from 1-10: 10 = A, 9 = A-, 8 = B+, etc.

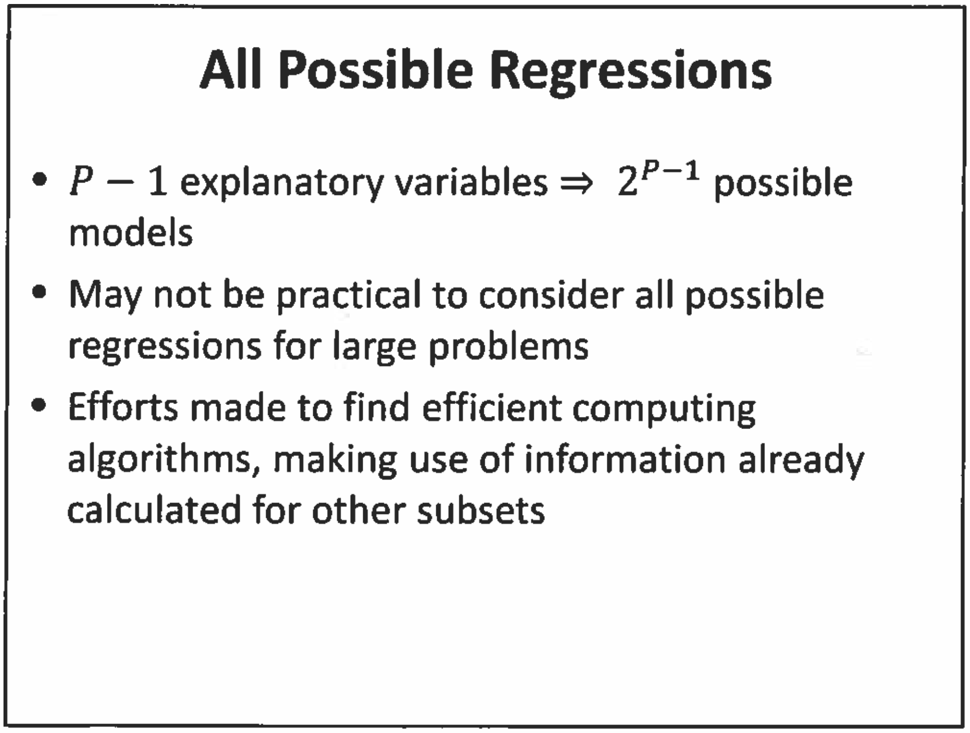
```
> cs <- read.table("C:/MATH651data/csdata.txt",header=TRUE)
> cs.hslm <- lm(gpa ~ hsm + hss + hse, data=cs)
> cs.satlm <- lm(gpa ~ satm + satv, data=cs)
```



Using *R*

- **stats** package: **AIC, BIC**
- **qpcR** and **MPV** packages: **PRESS**
- **MuMIn** package: **Cp**
- All functions called as listed above
- For all functions, linear model object name used as input

```
> AIC(cs.hslm,cs.satlm)
> BIC(cs.hslm,cs.satlm)
> Cp(cs.hslm)                      >Cp(cs.satlm)
> PRESS(cs.hslm)                   > PRESS(cs.satlm)
```



All Possible Regressions

- $P - 1$ explanatory variables $\Rightarrow$ 2^{P-1} possible models
- May not be practical to consider all possible regressions for large problems
- Efforts made to find efficient computing algorithms, making use of information already calculated for other subsets

Best Subsets Algorithm

- "Leaps and bounds" algorithm (Furnival & Wilson, 1974) combines SSE comparison across different subset models with control over sequence in which subset regressions are computed
- Guarantees finding best m subset regressions within each subset size
- Using *R*: **leaps** (in **leaps** package)

```
> cs.leaps <- leaps(y=cs$gpa, x=cs[,3:8])
> cs.leapsfull <- cbind(cs.leaps$which,cs.leaps$Cp)
```

- Best to use this approach as screening tool to suggest several models for further investigation

Stepwise Selection Procedures

- For cases where very large numbers (100s or 1000s) of variables to consider, "best" subsets procedure may not be feasible computationally
- Instead, possible to consider automatic search procedure that develops "best" subset of variables sequentially.
- Many tend to favor these approaches because they give "best" model automatically.
  - Not true!

Forward Stepwise Regression

- Most widely used sequential procedure
  - Finds plausible subset sequentially
  - At each step, variable added or deleted
  - Criterion for adding or deleting based on SSE, R^2, T-or F-statistic
  - Chosen model not guaranteed to be "best" model

Forward Stepwise Regression Procedure

1. Fit all single variable regressions and calculate $F_j = \frac{MSR(X_j)}{MSE(X_j)}$
   - If largest marginal F_j exceeds predetermined limit, add X_j to model; otherwise procedure stops
2. Given X_j in the model, fit two-variable regressions $X_k, k \neq j$, and calculate $F_k = \frac{MSR(X_k|X_j)}{MSE(X_j,X_k)}$
   - If largest marginal F_k exceeds predetermined limit, add X_k to model; otherwise procedure stops
3. Given X_k in the model, check if any $X_l, l \neq k$, already in model should be dropped
   - If smallest marginal F_l below predetermined limit, remove associated X_l from model; keep otherwise
4. Continue Steps 2 and 3 until STOP

Forward Stepwise Regression: Notes

- Computer packages usually specify "significance" level instead of F-statistic threshold to stay in model
  - SLE = "significance" level to enter
  - SLS = "significance" level to stay
- No clear recommendation for SLE, SLS selection
- Choice of SLE and SLS represents balance of opposing tendencies
  - Large SLE values ⇒ too many predictor variables
  - Small SLE ⇒ underspecified models, σ^2 overestimated
- People choose SLE values $\in [0.05, 0.5]$
  - Bendel and Afifi (1977) suggest $SLE \approx 0.25$ based on simulation studies

Forward Stepwise Regression: Notes (cont.)

- $SLE > SLS \Rightarrow$ cycling pattern may occur (i.e., variable put in and immediately taken out, etc.)
  - Most computer packages detect this and stop when it happens
  - Scenario tends to allow removal of nonsignificant variables
  - Bendel & Afifi (1977) recommendation: $SLS = \frac{SLE}{2}$
- $SLE < SLS \Rightarrow$ conservative procedure, may cause variables whose contributions have weakened to be retained
  - Order in which variables enter the model does not reflect importance
  - Variables entered early may eventually be dropped as other variables are added

Other Automated Selection Procedures

- Forward selection
  - Same idea as forward stepwise procedure, but doesn't test if variables should be dropped once entered
  - Not as good as forward stepwise
- Backward elimination
  - Begins with all variables in model and identifies variable with smallest F-value as candidate to drop.
  - Procedure can be made stepwise if variables are allowed back in
- Using *R*: packages **forward, subselect** and **FWDselect**
- These procedures may give different final models

Partial Regression Plots: Background & Notes

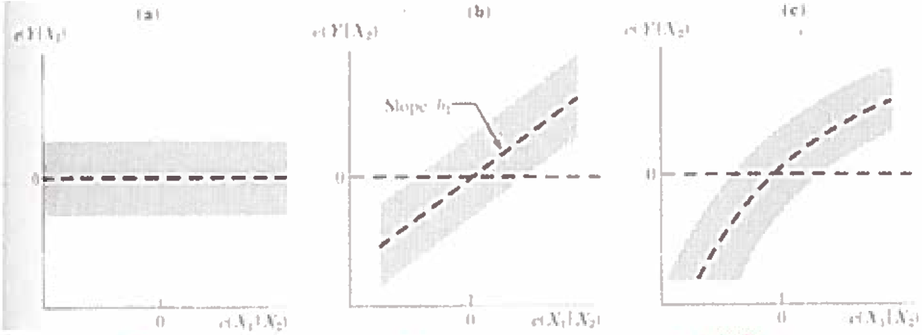
- Identify marginal relation for predictor variable X_i in regression model, given other predictor variables are in model
- Graphical indication of variable importance in model
- Identify outliers and important observations
- Be careful:
  - Sensitive to appropriateness of functional form for other variables in model
  - May not detect interaction
  - High multicollinearity may suggest improper functional relationships

Partial Regression Plots: Approach

1. Regress X_k against other X variables in model and find residuals,
$$e(X_k|X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{p-1})$$
2. Regress Y against the other X variables (i.e., not including X_k) and find the residuals,
$$e(Y|X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{p-1})$$
3. Plot two sets of residuals, $e(Y|\ldots)$ vs. $e(X_k|\ldots)$ and look for pattern and strength of linear association

Partial Regression Plots: Examples

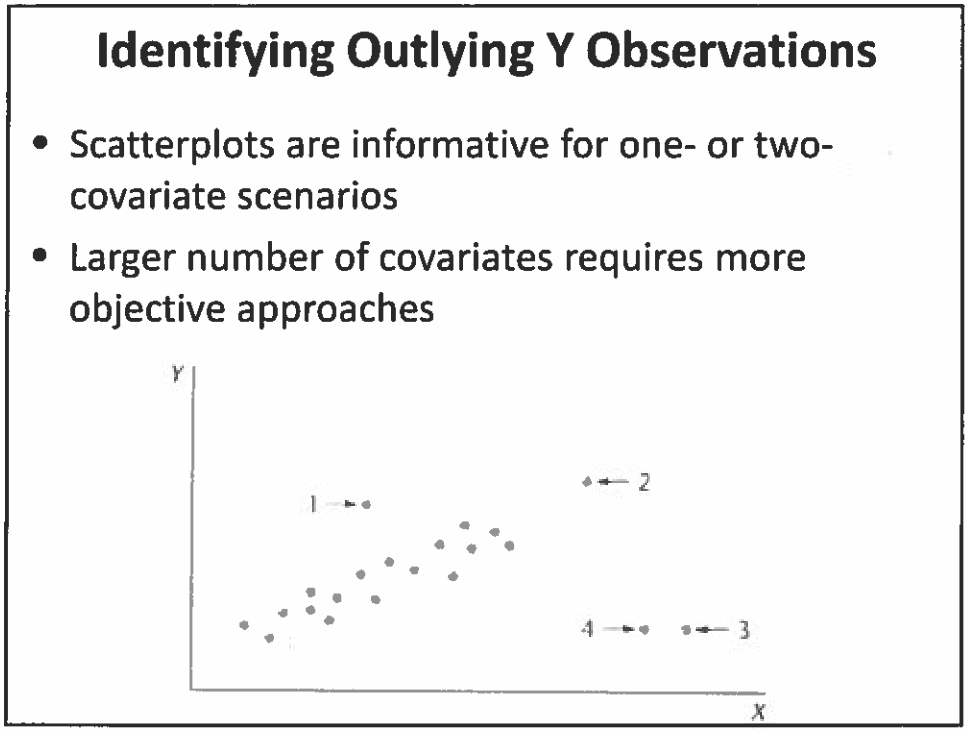
X_2 is already in the model. Should X_1 be added?



∴ no additional information beyond what's already addressed by X_2

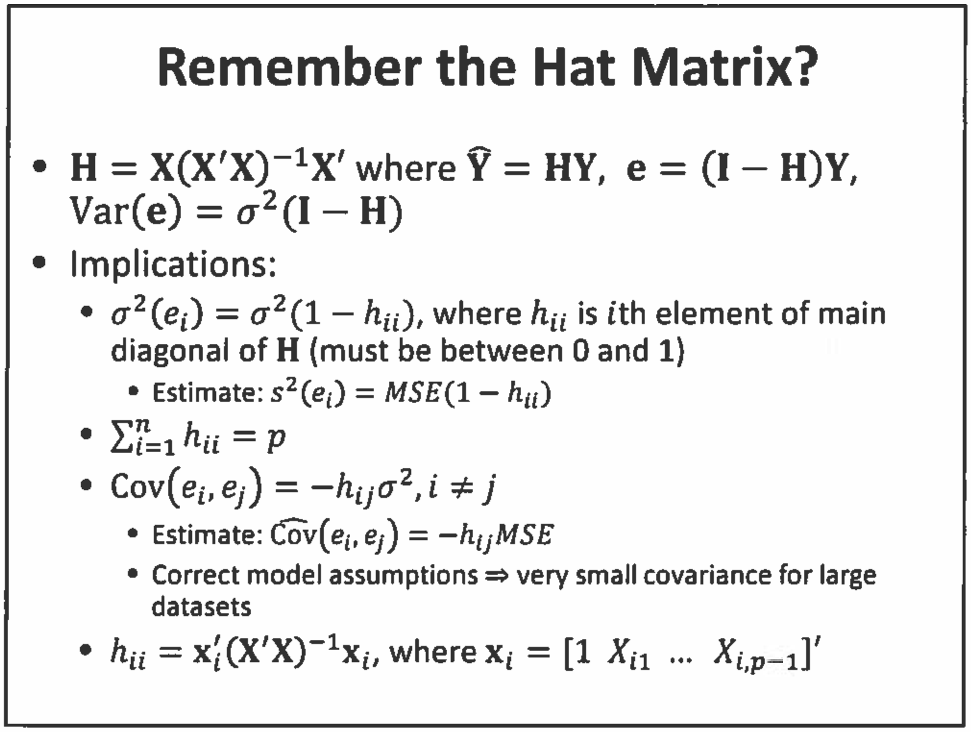
∴ linear X_1 may add to the model

∴ adding X_1 to the model may be helpful as a curvilinear model



Identifying Outlying Y Observations

- Scatterplots are informative for one- or two-covariate scenarios
- Larger number of covariates requires more objective approaches



Remember the Hat Matrix?

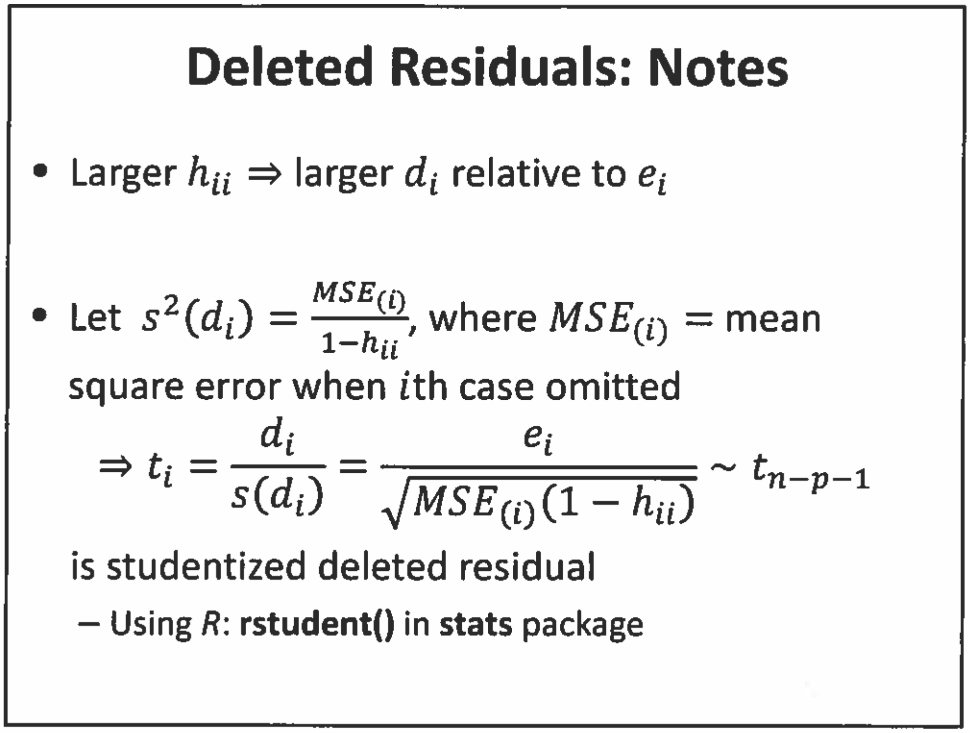
- $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ where $\hat{\mathbf{Y}} = \mathbf{HY}$, $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$, $\text{Var}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$
- Implications:
  - $\sigma^2(e_i) = \sigma^2(1 - h_{ii})$, where h_{ii} is ith element of main diagonal of $\mathbf{H}$ (must be between 0 and 1)
    - Estimate: $s^2(e_i) = MSE(1 - h_{ii})$
  - $\sum_{i=1}^{n} h_{ii} = p$
  - $\text{Cov}(e_i, e_j) = -h_{ij}\sigma^2, i \neq j$
    - Estimate: $\widehat{\text{Cov}}(e_i, e_j) = -h_{ij}MSE$
    - Correct model assumptions ⇒ very small covariance for large datasets
  - $h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$, where $\mathbf{x}_i = [1 \;\; X_{i1} \;\; \dots \;\; X_{i,p-1}]'$

Studentized Residuals

- Denoted $r_i = \frac{e_i}{s(e_i)}$; $s(e_i) = \sqrt{MSE(1 - h_{ii})}$
- Sometimes called "standardized residual"
- $r_i \sim N(0,1)$ approximately
- Note: we considered semi-studentized residual before: $e_i^* = \frac{e_i}{\sqrt{MSE}}$
  - Doesn't take into account different variances for each e_i
- Using R: **rstandard**

Deleted Residuals

- Reasonable to look at residual when model fit without the corresponding observation
- Procedure:
  1. Delete the ith case
  2. Fit regression on remaining $n - 1$ cases
  3. Get responses for ith case, $\hat{Y}_{i(i)}$, and find the difference, $d_i = Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1-h_{ii}}$
     - d_i = "deleted residual"
- Latter formulation shows that we don't need to recompute regression model for each case!



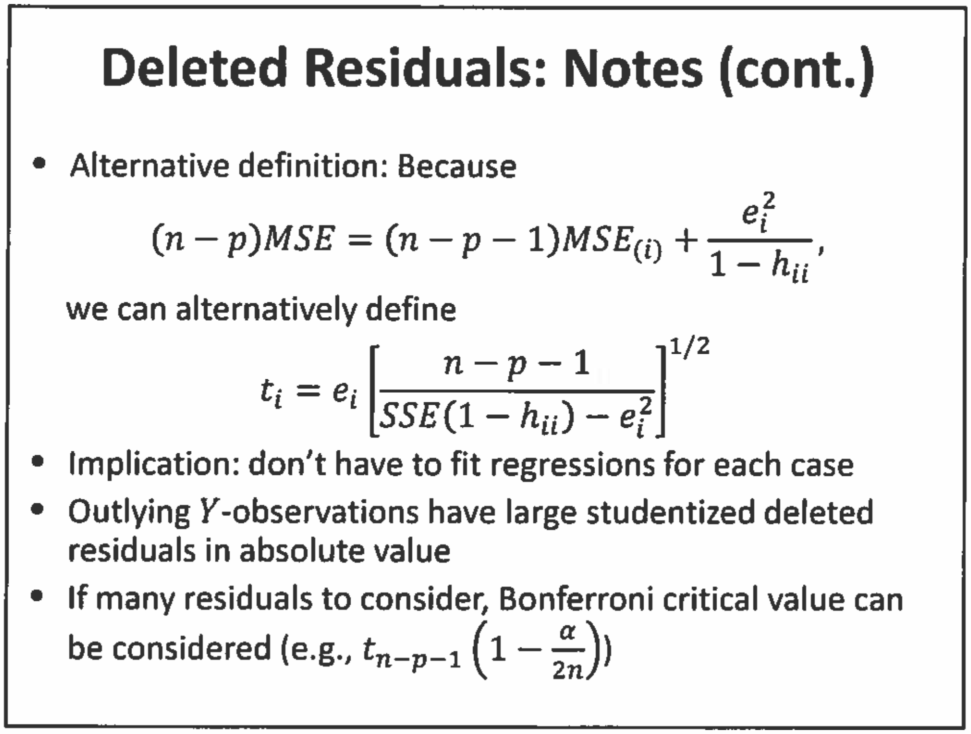
Deleted Residuals: Notes

- Larger $h_{ii} \Rightarrow$ larger d_i relative to e_i
- Let $s^2(d_i) = \frac{MSE_{(i)}}{1-h_{ii}}$, where $MSE_{(i)} =$ mean square error when ith case omitted

$$\Rightarrow t_i = \frac{d_i}{s(d_i)} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} \sim t_{n-p-1}$$

is studentized deleted residual

- Using *R*: **rstudent()** in **stats** package



Deleted Residuals: Notes (cont.)

- Alternative definition: Because

$$(n-p)MSE = (n-p-1)MSE_{(i)} + \frac{e_i^2}{1-h_{ii}},$$

we can alternatively define

$$t_i = e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}\right]^{1/2}$$

- Implication: don't have to fit regressions for each case
- Outlying Y-observations have large studentized deleted residuals in absolute value
- If many residuals to consider, Bonferroni critical value can be considered (e.g., $t_{n-p-1}\left(1-\frac{\alpha}{2n}\right)$)

Identifying Outlying X Observations

- Recall $0 \leq h_{ii} \leq 1$ and $\sum_{i=1}^{n} h_{ii} = p$ (i.e., total # of parameters)
- Large $h_{ii} \Rightarrow i$th case distant from center of all Xs (leverage of ith case); i.e. large value suggests observation exercises substantial leverage in determining $\hat{Y}_i$
- Question: why is this the case?
  - Answer: $\hat{\mathbf{Y}} = \mathbf{HY}$, linear combination of Y-value; h_{ii} is weight of Y_i, so h_{ii} measures role of X values in determining how important Y_i is in affecting $\hat{Y}_i$

What is "Large"?

- Large $h_{ii} \Rightarrow \text{Var}(e_i)$ small, so larger h_{ii} implies $\hat{Y}_i$ close to Y_i
- What is "large"?
  - Small datasets: $h_{ii} > 0.5$; large datasets: $h_{ii} > \frac{2p}{n}$
- Hat matrix for extrapolation identification: Consider
$$h_{new,new} = \mathbf{x'}_{new}(\mathbf{X'X})^{-1}\mathbf{x}_{new}$$
where $\mathbf{x}_{new}$ = new $\mathbf{X}$ values for inference about mean response or new observation; $\mathbf{X}$ = design matrix
  - If $h_{new,new}$ within range of h_{ii}s for cases in data set, no extrapolation involved; otherwise, extrapolation indicated

Identifying Influential Cases

- Influence on Single Fitted Values: DFFITS

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}h_{ii}}} = t_i\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$$

  - Standardized difference between ith fitted value with all observations and with ith case removed
  - Studentized deleted residual multiplied by factor that is function of ith leverage value
- Influence if

$$|DFFITS| > \begin{cases} 1, & \text{for small to medium datasets} \\ 2\sqrt{\frac{p}{n}}, & \text{for large datasets} \end{cases}$$

Influence on All Fitted Values

Cook's Distance:

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{p(MSE)} = \frac{e_i^2}{p(MSE)}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

- Gives influence of ith case on all fitted values
- If e_i increases or h_{ii} increases, then D_i increases
- D_i is a percentile of $F_{p,n-p}$ distribution
  - If percentile greater than 0.50, then ith case has major influence
- In practice, $D_i > \frac{4}{n} \Rightarrow i$th case has major influence

Influence on Regression Coefficients: DFBETAS

- Definition: for $k = 0, \dots, p-1$, and $c_{kk} = k$th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$,

$$DFBETAS_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} c_{kk}}}$$

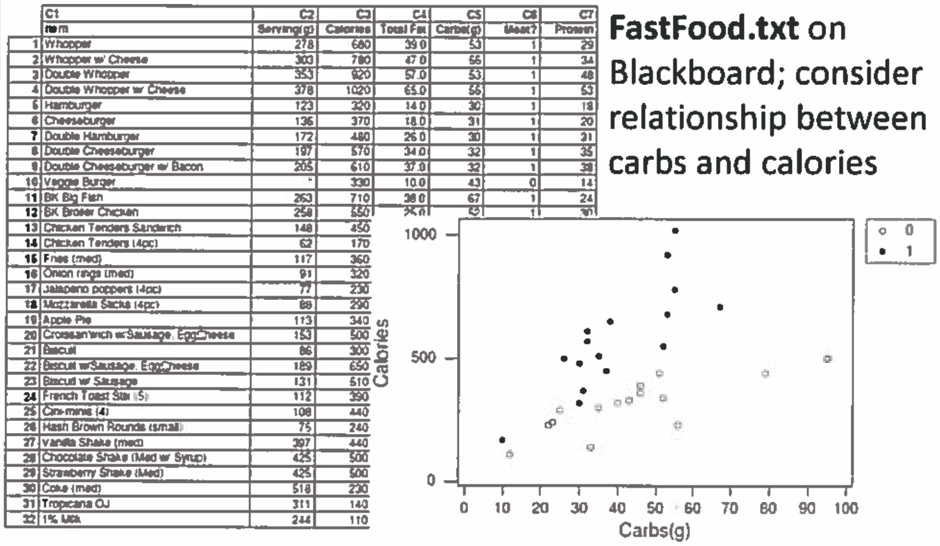
- Influence of ith case on each regression coefficient b_k = difference between estimated regression coefficients based on all n cases and regression coefficients obtained when ith case omitted

Influence on Regression Coefficients: DFBETAS (cont.)

- Sign of DFBETA indicates whether inclusion of a case leads to increase or decrease in estimates of regression coefficient

$$|DFBETA| > \begin{cases} 1, & \text{for small datasets} \\ \dfrac{2}{\sqrt{n}}, & \text{for large datasets} \end{cases}$$

Example: Fast Food



	C1 Item	C2 Serving(g)	C3 Calories	C4 Total Fat	C5 Carbs(g)	C6 Meat?	C7 Protein
1	Whopper	278	680	39.0	53	1	29
2	Whopper w/ Cheese	303	780	47.0	55	1	34
3	Double Whopper	353	920	57.0	53	1	48
4	Double Whopper w/ Cheese	378	1020	65.0	55	1	53
5	Hamburger	123	320	14.0	30	1	18
6	Cheeseburger	136	370	18.0	31	1	20
7	Double Hamburger	172	480	26.0	30	1	31
8	Double Cheeseburger	197	570	34.0	32	1	35
9	Double Cheeseburger w/ Bacon	205	610	37.0	32	1	38
10	Veggie Burger	*	330	10.0	43	0	14
11	BK Big Fish	263	710	38.0	67	1	24
12	BK Broiler Chicken	258	550	25.0	52	1	30
13	Chicken Tenders Sandwich	148	450				
14	Chicken Tenders (4pc)	62	170				
15	Fries (med)	117	360				
16	Onion rings (med)	91	320				
17	Jalapeno poppers (4pc)	77	230				
18	Mozzarella Sticks (4pc)	88	290				
19	Apple Pie	113	340				
20	Croissan'wich w/Sausage, Egg,Cheese	153	500				
21	Biscuit	86	300				
22	Biscuit w/Sausage, Egg,Cheese	189	650				
23	Biscuit w/ Sausage	131	510				
24	French Toast Stix (5)	112	390				
25	Cini-minis (4)	108	440				
26	Hash Brown Rounds (small)	75	240				
27	Vanilla Shake (med)	397	440				
28	Chocolate Shake (Med w/ Syrup)	425	500				
29	Strawberry Shake (Med)	425	500				
30	Coke (med)	518	230				
31	Tropicana OJ	311	140				
32	1% Milk	244	110				

FastFood.txt on Blackboard; consider relationship between carbs and calories

Using R: Model Diagnostics

- Regress **Calories** on **Carbs**, **Meat?**, and interaction

```
> food <- read.table("C:/MATH651data/FastFood.txt",header=TRUE,sep="\t")
> food.lm <- lm(Calories ~ Carbs..g. + Meat. + Carbs..g. * Meat. , data=food)
> summary(food.lm)
```

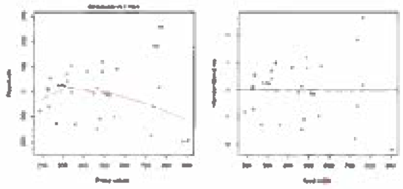
- Perform model diagnostics:

```
> influence.measures(food.lm)
```

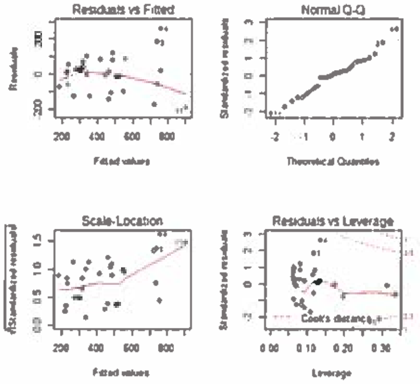
  - Outputs (among other things) DFBETAS for each model variable, DFFITS, Cook's distances and hat matrix diagonal elements
  - Influential cases wrt any measure marked with (*)

Using R: Model Diagnostics (cont.)

```
> par(mfrow=c(1,2))
> plot(food.lm,which=1)
> plot(food.lm$fit,rstandard(food.lm))
> abline(h=0)
```



```
> par(mfrow=c(2,2))
> plot(food.lm)
> dev.off()
```



For Completeness: COVRATIO

- Measures change in determinant of covariance matrix of the estimates by deleting ith observation:

$$COVRATIO = \frac{\det\left(s^2(i)\left(\mathbf{X}_{(i)}'\mathbf{X}_{(i)}\right)^{-1}\right)}{\det(s^2(\mathbf{X}'\mathbf{X})^{-1})}$$

- Belsley, Kuh, and Welsch (1980) suggest investigating observations with

$$|COVRATIO - 1| \geq \frac{3p}{n}$$

where p = # of model parameters, n = # of observations used to fit model