

BIOS 6612 Lecture 1

Model Selection and Outliers

KKMN Chapter 14 and 16

Review (BIOS 6611)/ Current (Lecture 1)/ Preview (Lecture 2)**BIOS 6611:**

- Distributions
 - Normal, Binomial
- t-tests
 - Paired vs 2 independent samples
 - Wilcoxon signed rank vs Wilcoxon rank sum (Mann-Whitney U)
- Linear Regression (Lecture 16-26):
 - Lecture 20: Matrix Approach to Linear Regression
 - Lecture 23-24: Categorical Predictors and General Linear Hypotheses
 - Lecture 25: Polynomial Regression
 - Lecture 26: Regression Diagnostics

Lecture 1:

- Model Selection Strategies
 - Adjusted R squared
 - Partial F-test
 - AIC
 - Mallows' Cp
 - Forward, Backwards and Stepwise selection
- Outliers
 - high leverage => potential influence

Lecture 2:

- Logistic Regression
 - Binary outcome vs continuous outcome

Variable Selection Strategies in Linear Regression

Reasons for including predictors in a regression model:

- 1) The variable is required to address the scientific question
i.e. primary explanatory variable, exposure, treatment
- Depends on the question of interest?
- 2) Assess potential confounding and adjust for it
To potentially provide a *valid* estimate for one or more regression coefficients of interest
- 3) Assess effect modification (interaction).
-Biological reasoning or question of interest
- 4) Assess mediation.
-Need to assume causal relationship between variables
- 5) Increase precision/efficiency or improve prediction.
-i.e. machine, study, batch effect

Model Selection Procedures

1. Epidemiologic – observational study
 - One or a few exposure variables of interest; can include interactions
 - Adjust for confounding
 - Collect a large number of potentially predictive variables without adequate scientific or historical information to tell us which variables are important predictors and/or potential confounders (Model selection needed)
2. Clinical – randomized study
 - Intervention effect is the variable of primary interest
 - Little or no confounding expected ; possible stratification in the design
 - Adjust for predictive variables (including stratification/matching variables) to increase precision, but in general, these variables should be identified during the design of the trial
 - Model selection is usually not necessary or desired
3. Prediction – randomized or observational study
 - Wish to identify a model that will best predict the outcome for future observations
 - Model selection possible, but may not be necessary
 - Hold-out sample when interested in prediction

Major Considerations for Model Selection Procedures

- Goal of modeling: estimation, hypothesis testing, prediction, etc.
- Type I error inflation
- Adequate power for hypothesized effects, extent of confounding
- Form of the model: linear, generalized linear, nonlinear, etc.
- Functional form of variables to suit the model assumptions of linearity

Other considerations

- We want the model to be as parsimonious as possible because:
 - Interpretation is difficult with a large number of predictors (particularly when a large number of interaction terms or complex interaction terms are included)
 - Excess variables may increase the variability of our estimates of β
 - Estimation problems may occur with too many variables
 - We may want the smallest subset for prediction for practical reasons (e.g., cost, time, availability, etc.)
- Promising models can be selected based on comparing the MS_{error} , Adjusted R^2 , Partial F test, AIC, Mallow's C_p .
- How should models be compared?
 - Numerical differences?
 - Statistically significant differences?
 - Scientifically important differences?



R^2 and Adjusted R^2

- No matter how strong or weak an additional explanatory variable is, SS_{Model} **never** decreases (it will increase or stay the same):
 - If the extra explanatory variable is a strong predictor, SS_{Model} significantly increase.
 - If the extra explanatory variable is a poor predictor, SS_{Model} may change very little.

- What will happen to the R^2 ? $R^2 = \frac{SS_{\text{model}}}{SS_{\text{total}}}$

Either increases or stays the same

- **Adjusted R^2** : The Adjusted R^2 “adjusts” for the number of parameters in the model (i.e., it penalizes for each additional variables), and thus the adjusted R^2 can increase or decrease when a variable is added to the model.

$$\text{Adjusted } R^2 = 1 - \left(\frac{n-1}{n-p-1} \right) (1 - R^2),$$

where p is the number of variables (predictors) in the model, excluding the intercept.

MSE vs SS_{error}

- If X_2 is a significant explanatory variable given that X_1 is already in the model, the Adjusted R^2 should be higher in *Model 2* compared to *Model 1*.
- What will happen to the $MSE = \frac{SS_{\text{error}}}{n - p - 1}$?
 - If X_2 is a significant explanatory variable given that X_1 is already in the model, you will see a “large” drop from $SS_{\text{Error-1}}$ to $SS_{\text{Error-2}}$.
 - This will result (most likely) in a decrease in $MS_{\text{Error-2}}$ from $MS_{\text{Error-1}}$, regardless of the fact that you are dividing by fewer degrees of freedom.
 - If X_2 is a *not* a significant explanatory variable given that X_1 is already in the model, there will be very little change from $SS_{\text{Error-1}}$ to $SS_{\text{Error-2}}$.
 - This will result (most likely) in an increase in $MS_{\text{Error-2}}$ from $MS_{\text{Error-1}}$ since you are dividing by fewer degrees of freedom.
- Thus, the SS_{error} will always decrease or stay the same as the number of regressors in the model increases
 - In general, the MSE initially decreases, then stabilizes, then may increase as the number of regressors in the model increases

Partial F tests

Partial F test (F_p): Compare *nested* models by performing a partial F test.

- Nested Models=two models in which one model contains a subset of the variables included in the larger model

$$\left\{ \begin{array}{l} E[Y] = \beta_0 + \beta_1 X \\ E[Y] = \beta_0 + \beta_1 X + \beta_2 X^2 \end{array} \right\} \text{Nested Models}$$

$$\left\{ \begin{array}{l} E[Y] = \beta_0 + \beta_1 X \\ E[Y] = \beta_0 + \beta_1 Z \end{array} \right\} \text{Non-Nested Models}$$

- This cannot be used for *non-nested* models (for example to compare a model with age and height as predictors to a model with age and weight as predictors).

$$F = \frac{[SS_{\text{model}}(\text{full}) - SS_{\text{model}}(\text{reduced})] / k}{MS_{\text{error}}(\text{full})} \sim F_{k, n-p-k-1}$$

where p is the number of independent variables in the reduced model and k is the number of regression coefficients specified to be zero under the null hypothesis

AIC and BIC

AIC (Akaike Information Criterion): Based on goodness of fit and includes a penalty for increasing the number of parameters in the model.

- Models do not need to be **nested**.

$$AIC = 2k - 2\ln(L)$$

where k is the number of parameters in the model and L is the maximized likelihood function for the model.

- The best model has the lowest AIC
- Arbitrary rule of thumb, difference of 2 is significant

BIC (Bayesian Information Criterion or Schwarz Criterion): Based on goodness of fit and includes a penalty for increasing the number of parameters in the model.

- Models do not need to be **nested**.

$$BIC = k * \ln(n) - 2\ln(L)$$

where k is the number of parameters in the model and L is the maximized likelihood function for the model and n is the number of subjects.

- The best model has the lowest BIC.
- Penalizes the number of parameters more than AIC

Mallow's C_p

$$\text{Mallow's } C_p = \frac{SSE(p)}{MSE(k)} - [n - 2(p + 1)]$$

where $SSE(p)$ is the sums of squares error from the reduced model with p variables and $MSE(k)$ is the mean square error from the full model with all k predictors

- The C_p statistic can be used to compare **non-nested** models.
- For two models with the same number of predictors, p , the model with the lower C_p is considered the better model
- The C_p “penalizes” the addition of variables to the model
- The C_p criterion can be used to decide how many variables to put in the best model, since it achieves a value of approximately $p+1$ if $MSE(p)$ is roughly equal to $MSE(k)$
i.e., if the correct model is of size p
- If important predictors are omitted (from those considered), C_p should be larger than $p+1$.
- The full model, with all k predictors has a C_p value exactly equal to $k+1$.
- The C_p statistic is the following simple function of the F_p statistic (F_p partial F comparing full model to current): $C_p = (k-p)F_p + (2p-k+1)$

Common Model Selection Procedures

Regardless of which model selection strategy is used, you should first determine the maximum size of the model you will consider for number of covariates p

- As a general rule, $p < n/10$, $p < n/15$, or $p < n/20$.
- Another general rule is to **always have a minimum of 10df for your MSE**
 - This will not guarantee a desired level of power but results in a minimally adequate sample size-to-variable ratio
 - If the ratio is too small the model will tend to “overfit” the data, and will not be valid because of unstable and/or biased estimates

All Possible Subsets:

- This procedure involves fitting all possible subsets of variables.
- The most promising models can be selected based on comparing the MS_{error} , R^2 , Adjusted R^2 , Mallows' C_p , AIC across models.
 - With p independent variables, the total number of equations fit is 2^p ($2^p - 1$ excluding the intercept only model)
 - The number of equations can become prohibitive as p increases (e.g., $2^7 - 1 = 127$)

Reducing the Number of Considered Variables

To reduce the number of considered variables:

- 1) Use the literature and knowledge of the area to eliminate unimportant variables.
- 2) Eliminate variables with distributions that are too narrow.
 - a. i.e. everyone is 25 years old, then age is not a useful variable in this study
- 3) Eliminate variables with a large percentage of missing data
 - a. If future observations are also likely to be missing on these variables
- 4) Perform statistical data reduction – clustering, means, combine into new variable (e.g., BMI), etc.

$$\text{BMI} = \text{wt}(\text{kg}) / \text{ht}(\text{m})^2$$

Common method to determine the number of covariates included in model selection:

- Run univariate analysis
- Choose covariates for model selection with p-value < 0.05, 0.1 or 0.2
- Run this set of variable through step wise model selection
- **Problems of this approach:**
 - Not thinking about polynomial trends or collinearity
 - Covariate could not be significant in univariate analysis but significant in multiple linear regression
 - Not thinking about the question of interest

Forward Selection and Backwards Elimination: ***NOT PREFERRED***

Forward Selection:

- Step 1. Start with a model that contains only an intercept and enter the variable most highly correlated with the dependent variable, if $p < 0.05$ (or other pre-determined α).
- Step 2. Calculate the Partial F Test (or t Test) for each variable not in the model based on a regression equation containing that variable and all other variables already in the model.
- Step 3. Add the most significant variable if its p-value is less than some pre-selected value (an alpha level of .10 is often used).
- Step 4. Repeat steps 2 and 3 until no additional variables can be added to the model (until all remaining variables are not significant).

Backward Elimination:

- Step 1. Start with all potential variables in the model.
- Step 2. Calculate the Partial F Test (or t Test) for each variable in the model.
- Step 3. Remove the least significant variable if its p-value is greater than some pre-selected value (an alpha level of .10 is often used).
- Step 4. Re-compute the regression equation for the remaining variables and repeat steps 2 and 3 until all of the remaining variables are statistically significant.

Stepwise Selection

Stepwise:

- Step 1. Start with a model that contains only an intercept and enter the variable most highly correlated with the dependent variable.
- Step 2a. Calculate the Partial F Test (or t Test) for each variable in the model. Remove the least significant variable if its p-value is greater than some pre-selected value (an alpha level of .15 is often used). If a variable is removed, re-calculate the regression equation and repeat Step 2.
- Step 2b. Calculate the Partial F Test (or t Test) for each variable not in the model based on a regression equation containing that variable and all other variables already in the model.
- Step 3. Add the most significant variable if its p-value is less than some pre-selected value (an alpha level of .10 is often used).
- Step 4. Repeat steps 2 and 3 until no additional variables can be removed from or added to the model.

Alternatively, the stepwise procedure can begin with all variables in the model.

Done with AIC, BIC, Mallow's C_p , adjusted R^2 , etc

Generalizations of Variable Selection Procedures

- ***Chunkwise methods:*** Sets or “chunks” of predictors can be considered for addition or deletion together using the partial F -test.
- Interaction terms can (also) be included in variable selection procedures.
 - **As a general rule, main effect terms and lower-degree interaction terms need to be included in a model before a higher-order interaction term is added.**
 - Likewise, main effect terms and lower-degree interaction terms are *usually* not eligible for deletion from a model if they are components of higher-degree interaction terms that are still in the model.
- You may wish to include important variables (e.g., the primary explanatory variable(s) or treatment variable) in all models.
 - These variable(s) would enter the model on the first step and are not candidates for deletion.
 - This can be accomplished using the INCLUDE option in SAS.

Model Selection with Interactions



Preliminary Questions:

- Which set of predictor variables will be considered as possible candidates and which, if any, predictor variables should be retained in ALL models (e.g., exposure of interest, treatment group)?
- Which interaction terms will be considered as possible candidates?
 - Will you consider all interactions or only interactions involving the primary explanatory variable?
 - If you will consider interactions that don't include the primary explanatory variable, should they be limited on scientific grounds? (answer should usually be 'yes')
 - What is the highest order interaction you wish to consider (two-factor or three-factor interactions or no interactions)?
 - Interactions of degree higher than three are rarely, if ever, etiologically interpretable.
 - The power to detect them as significant is usually very low.
 - And even if they are significant, the validity of statistical tests about such higher-order terms can be seriously questioned.

Keep in mind that the total number of predictors you wish to consider (including interaction terms!) should be less than $n/10$ (or $n/15$ or $n/20$).

Example

Weight (WGT), height (HGT), and age (AGE) of a random sample of 12 nutritionally deficient children. (KKM, p. 385). HGT², and AGE² are height and age squared, the HGT*AGE interaction is also considered.

```
DATA wgt;  
input wgt hgt age;  
agehgt = age*hgt;  
age2 = age*age;  
hgt2 = hgt*hgt;  
datalines;  
64 57 8  
71 59 10  
53 49 6  
67 62 11  
55 51 8  
58 50 7  
77 55 10  
57 48 9  
56 42 10  
51 42 6  
76 61 12  
68 57 9  
;  
  
PROC REG;  
  MODEL wgt = age hgt agehgt age2 hgt2  
  / SELECTION=cp rsquare adjrsq mse B;  
RUN;
```

All variables we'd consider for "saturated" model.

Number in Model	C(p)	R-Square	Adjusted R-Square	MSE
1	-0.4453	0.7535	0.7289	21.89185
2	0.7440	0.7800	0.7311	21.71415
2	0.8535	0.7764	0.7267	22.06667
2	0.8809	0.7755	0.7256	22.15495
2	0.9539	0.7731	0.7227	22.38984
2	0.9765	0.7724	0.7218	22.46250
2	1.0152	0.7711	0.7203	22.58723
2	1.1206	0.7677	0.7161	22.92643
2	1.1459	0.7669	0.7151	23.00799
3	2.0454	0.8028	0.7288	21.89768
3	2.1249	0.8002	0.7253	22.18565
1	2.1916	0.6675	0.6343	29.53302
3	2.2574	0.7959	0.7193	22.66559
1	2.3295	0.6630	0.6293	29.93275
3	2.7358	0.7803	0.6978	24.39869
3	2.7367	0.7802	0.6978	24.40178
3	2.8376	0.7769	0.6933	24.76744
3	2.8534	0.7764	0.6926	24.82438
3	2.9311	0.7739	0.6891	25.10614
3	2.9464	0.7734	0.6884	25.16136
3	2.9742	0.7725	0.6872	25.26208
4	4.0001	0.8043	0.6924	24.83873
4	4.0344	0.8031	0.6906	24.98033
4	4.1127	0.8006	0.6866	25.30469
2	4.1729	0.6681	0.5944	32.75423
1	4.4874	0.5926	0.5519	36.18571
1	4.6413	0.5876	0.5464	36.63180
4	4.7356	0.7803	0.6547	27.88308
4	4.8338	0.7771	0.6497	28.28980
5	6.0000	0.8043	0.6411	28.97780
2	6.4844	0.5927	0.5022	40.19683

```
PROC REG;
```

```
MODEL wgt = age hgt agehgt age2 hgt2  
/ SELECTION=cp rsquare adjrsq mse B;
```

```
RUN;
```

Saturated/ Full Model

Number in	-----Parameter Estimates-----						
Model	C(p)	Intercept	age	hgt	agehgt	age2	hgt2
1	-0.4453	37.59975	.	.	0.05314	.	.
2	0.7440	6.55305	2.05013	0.72204	.	.	.
2	0.8535	15.11754	.	0.72598	.	0.11480	.
2	0.8809	25.06226	.	0.36759	0.03866	.	.
2	0.9539	35.43113	.	.	0.07647	-0.10909	.
2	0.9765	25.51240	1.98704	.	.	.	0.00697
2	1.0152	34.75979	.	.	0.03919	.	0.00334
2	1.1206	33.94445	.	.	.	0.11079	0.00701
2	1.1459	42.37088	-1.53252	.	0.07167	.	.
3	2.0454	-89.23617	.	4.86082	0.05313	.	-0.04586
3	2.1249	-93.09350	.	4.96869	.	0.15261	-0.04200
1	2.1916	33.49427	0.01036
3	2.2574	-79.88380	2.51363	4.03411	.	.	-0.03269
1	2.3295	6.18985	.	1.07223	.	.	.
3	2.7358	3.43843	2.77687	0.72369	.	-0.04171	.
3	2.7367	1.06471	2.63866	0.83032	-0.01146	.	.
3	2.8376	23.76128	2.76001	.	0.07622	-0.26389	.
3	2.8534	14.25089	.	0.75722	-0.00335	0.12467	.
3	2.9311	36.35244	.	.	0.11660	-0.22335	-0.00376
3	2.9464	19.60223	3.37735	.	.	-0.08007	0.00702
3	2.9742	27.26447	1.61138	.	0.00753	.	0.00626
4	4.0001	-92.43339	-1.84901	5.32331	0.09078	.	-0.05389
4	4.0344	-86.09507	.	4.76070	0.08138	-0.08233	-0.04746
4	4.1127	-97.07052	-1.03966	5.29692	.	0.21413	-0.04524
2	4.1729	48.02672	.	-0.56763	.	.	0.01581
1	4.4874	30.57143	3.64286
1	4.6413	45.99764	.	.	.	0.20597	.
4	4.7356	2.34281	2.77820	0.76299	-0.00421	-0.02937	.
4	4.8338	24.74161	2.62091	.	0.09332	-0.30473	-0.00160
5	6.0000	-91.93718	-1.79686	5.29733	0.09337	-0.01064	-0.05387
2	6.4844	32.40411	3.20536	.	.	0.02498	.

Model: age & height
Highest adjusted R squared
Lowest MSE
Low C(p)

NOT hierarchical

Caveats

- Model selection is an art
 - There is usually no single “best” method
 - But there are several wrong models
 - Collinearity
 - Not hierarchical
- Model selection should first and foremost be driven by your knowledge of the subject area and by your hypotheses.
- Different selection strategies can lead to different sets of variables being included in the final model.
- A good analysis should point out that there are different possible models when more than one “adequate” model is detected in an analysis.
 - This could help shed light on the structure of the data and can aid in the understanding of the underlying process.
- When modeling is performed to find the best prediction equation, it is often desirable to perform a split-sample analysis, where part of the sample is used to obtain the predication equation (the *training sample*) and the remaining sample is used to validate the equation (the *holdout* or *validation* sample).

Frank Harrell's concerns with variable selection

(*Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*, Harrell, 2001)

The automated techniques covered in this lecture have many scientific and statistical drawbacks.

They should NEVER replace careful scientific thought and consideration in model building.


More specifically, variable selection procedures can lead to the following statistical and scientific problems:

- It yields R^2 values that are biased high.
- The ordinary F and χ^2 statistics do not have the claimed distribution. Variable selection is based on methods that were intended for testing only pre-specified hypotheses.
- The method yields standard errors of regression coefficient estimates that are biased low and confidence intervals for effects and predicted values that are falsely narrow.
- It yields p -values that are too small (i.e., there are severe multiple testing problems), and the proper correction for them is a difficult problem.
- It provides regression coefficients that are biased high in absolute value and need shrinkage.
- Rather than solving problems caused by collinearity, variable selection is made arbitrary by collinearity.
- **It allows us to NOT think about the problem.**

Analyzing Data and Model Selection

- The following steps are **just one approach** to model selection and analyzing data
 - There are many other valid approaches

Step 1: Literature Review

- Don't reinvent the wheel
- a. Make sure you have a clear hypothesis or question of interest
 - i. Write out your scientific question of interest
- b. What other confounders have people considered in the past
 - i. Are these covariates available in your cohort?
 - ii. Is the cohort appropriate for the question of interest?
- c. Does the outcome need to be transformed
 - i. Have other studies transformed the outcome? 
 - 1. If so, what transformations, were used
 - 2. Do these transformations make sense for your study or in general?
- d. Quadratic trends to consider
 - i. Does BMI have a quadratic relationship with the outcome?
- e. Use this as a starting block
 - i. Current literature may or may not be required or useful in your scenario

Analyzing Data and Model Selection

Step 2: Consider Transforming the Outcome

1. First check the diagnostic plots to see if this necessary
2. Then apply one of the following transformations and check if the diagnostic plots improve
 - a. Some possible transformations:
 - ii. Log (e.g. $\log(y)$)
 1. If left or right skewed and greater than zero
 2. Exponential model
 - iii. Square Root (e.g. \sqrt{y})
 1. Quadratic model
 - iv. Reciprocal Model (e.g. $1/y$)
 - v. Inverse normal transformation
 1. Beasley TM, Erickson S. (2009) Rank-Based Inverse Normal Transformations are Increasingly Used, But are They Merited? Behav Genet. 39(5): 580–595.
 - b. Consider a different model (e.g. not linear regression)
 - i. If there is large point mass at zero for count data
 - a. Consider a zero inflated negative binomial
 - b. Consider a zero inflated Poisson

Analyzing Data and Model Selection

Step 3: Consider the Covariates


- a. Non-linear trends

Be careful with collinearity

- i. Fit the polynomial model- see if the higher order terms are needed
 - 1. Make sure this makes sense
 - a. Quadratic binary variables don't make sense or variables with 3 levels
 - b. Just fit the appropriate number of indicator variables
- ii. Check diagnostic plots
- iii. Use VIF to make sure lower order term is not collinear with higher order term
 - 1. If so, then consider making the variable categorical
 - 2. Give BMI categories (i.e. $BMI < 25$, $25 < BMI < 30$, $BMI > 30$)

Analyzing Data and Model Selection


Step 3: Consider the Covariates

- b. 2 variables are too similar (collinear) then drop one
 - i.e. height, weight, bmi
 - ii. Use correlation AND variance inflation factor to check
 - iii. Consider dropping the variable that is self reported or has a higher level of missing data
- c. Violation of model assumptions 
 - a. Normality or homoscedasticity
 - i. This can occur if a variable is very skewed
 - ii. Categorize variable (binary at median)
- d. Confounders
 - i. Make sure they are in the model even if they are not significant
- e. Interactions
 - i. Biological reasoning
 - ii. Question of interest
- f. Lower order terms
 - a. Polynomial models
 - b. Interactions

Analyzing Data and Model Selection

Step 4: Consider the Univariate Analysis of Each Covariate with the Outcome

Step 5: Consider the Full Model

- What is the full model?
 - All covariates
 - Too many covariates
 - Ridge regression 
 - LASSO
 - Bayesian approach
 - Only covariates with a p-value less than 0.1 or XXXXX from the univariate analysis
 - What about confounders and lower order terms?
 - Can this approach be justified?
- What is the reduced model?
 - Reduce using
 - partial F-tests
 - Stepwise (force in lower order terms)

Example

A hypothetical analysis was performed to examine if height and weight are associated with systolic blood pressure.

SAS Output:

```
PROC REG;
  MODEL sbp = weight height /VIF;
RUN;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2336.31739	1168.15870	4.81	0.0102
Error	97	23566	242.94665		
Corrected Total	99	25902			

Significant

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-15.78944	23.64042	-0.67	0.5058
weight	1	0.32599	0.22603	1.44	0.1525
height	1	0.14993	0.19400	0.77	0.4415

NOT Significant

As a set, are weight and height significantly associated with SBP?

Yes. $F=4.81$, $p=0.0102$

Univariate Models

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2191.22534	2191.22534	9.06	0.0033
Error	98	23711	241.94813		
Corrected Total	99	25902			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-1.69786	15.01476	-0.11	0.9102
weight	1	0.45550	0.15136	3.01	0.0033

```
PROC REG;
  MODEL sbp = weight;
RUN;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1830.98504	1830.98504	7.45	0.0075
Error	98	24071	245.62405		
Corrected Total	99	25902			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-20.80585	23.51164	-0.88	0.3784
height	1	0.35738	0.13089	2.73	0.0075

```
PROC REG;
  MODEL sbp = height;
RUN;
```

What happened? Multicollinearity

Multicollinearity

Multicollinearity occurs when there is a linear relationship among one or more of the independent variables

Recall: for $Y = X\beta + \epsilon$, the least squares estimator for β is

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (1)$$

- If X_2 is a linear combination of X_1 then $(X^T X)$ is not invertible

For example: consider $X_{2i} = 1 + 2X_{1i}$

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \\ Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 (1 + 2X_{1i}) + \epsilon_i \\ Y_i &= (\beta_0 + \beta_2) + (\beta_1 + 2\beta_2) X_{1i} + \epsilon_i \\ Y_i &= \beta_0^* + \beta_1^* X_{1i} + \epsilon_i \end{aligned} \quad (2)$$

Two parameters (β_0^*, β_1^*) even though the original model had three parameters $(\beta_0, \beta_1, \beta_2)$. There are multiple values of $\beta_0, \beta_1, \beta_2$ which solve the two equations:

$$\begin{aligned} \beta_1^* &= (\beta_1 + 2\beta_2) \\ \beta_0^* &= (\beta_0 + \beta_2) \end{aligned} \quad (3)$$

The model is not identified, meaning we cannot estimate the separate influence of X_1 and X_2 on Y . Also, the $Var(\beta_1)$ can be rewritten as follows:

$$Var(\beta_1) = \frac{\sigma^2}{\sum x_{1i}^2 (1 - r_{12})^2} \quad (4)$$

where r_{12} is the correlation between X_1 and X_2 . Since X_1 and X_2 are linearly related $r_{12} = 1$ and the denominator goes to zero and the variance goes to infinity.

Multicollinearity

(KKMN Section 12.5, pp.237-248)

- Often, when numerous explanatory variables are considered, some offer redundant information concerning the response.
- By including such explanatory variables in the model, some or all may be classified as nonsignificant while the overall F - test provides a significant result.
 - May have inflated SE, especially for the intercept
- A high degree of collinearity may result in inflated standard errors for the parameter estimates, and there will be low power for hypothesis testing.
- The *variance inflation factor* (VIF) is often used to measure collinearity in a multiple regression analysis. It is computed for the j th predictor variable as:

$$VIF_j = \frac{1}{1 - R_j^2}, j = 1, 2, \dots, k$$

where R_j^2 is the squared multiple correlation based on regressing X_j on the remaining $k-1$ predictors.

Multicollinearity

- Some people prefer to consider the tolerance of a variable, which is computed as:

$$Tolerance_j = \frac{1}{VIF_j} = 1 - R_j^2$$

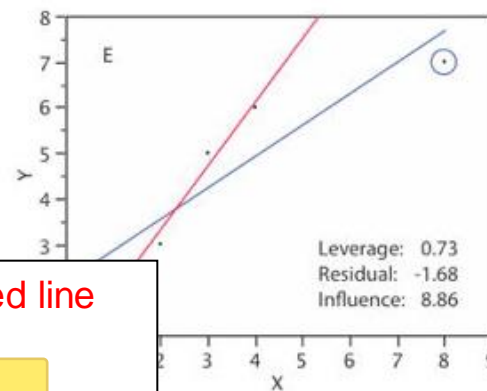
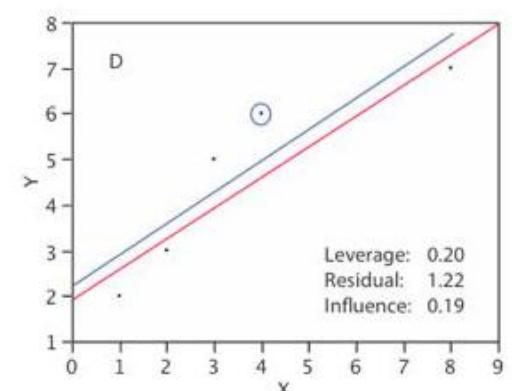
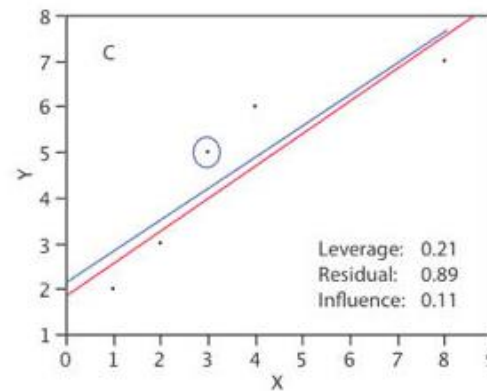
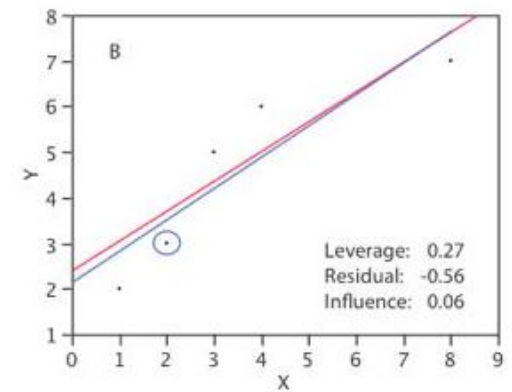
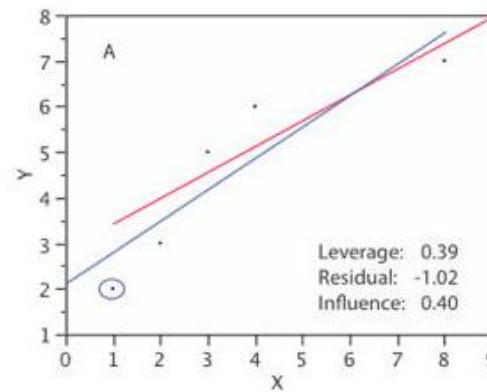
- The choice among R_j^2 , VIF_j , and $tolerance_j$ is a matter of personal preference since they all contain exactly the same information.
- A rule of thumb is to be concerned with a $VIF > 10.0$, which corresponds to an $R_j^2 > 0.90$ or a tolerance < 0.10 .
- High multiple correlations may be enough support to leave some explanatory variables out of the model, for using data reduction techniques to generate functions of the collinear variables (e.g., use BMI rather than height and weight), or for doing multiple- df rather than 1- df partial F -tests.
 - Better to use VIF than correlations

Note: The correlation between height and weight in the above example is 0.74.

- So, is height or weight a better predictor of SBP?
 - Choose variable with largest adjusted R squared
 - BMI
 - PCA of the variables (Larger # of covariates)

Outliers

- High leverage => potential influence
- In the plots, E represents an outlier we would want to investigate
- How to determine an outlier's influence
 - Cook's Distance (Cook's D)
 - DFFITS
 - DFBETAS
- How to handle Outliers
 - If an outliers is a false value, then delete it
 - Otherwise, **report both models with and without the deleted observation(s)**





The blue line includes the circled point and the red line excludes that point

- A. High leverage & low influence
- B. Lower Leverage => lower influence
- C. Low Leverage => low influence
- D. Low leverage => low influence
- E. High leverage & high influence

Outliers

- **Outlier** = an observation with a residual that is much larger than the rest.

Explanations for Outliers:

- 1) The value(s) for the observation was measured, recorded, or entered incorrectly.
In which case the value(s) for the observation should be corrected or the observation should be deleted from the analysis (but ONLY if it is KNOWN to be wrong).
- 2) Inadequacies in the model. 
The model may fail to fit the data well for certain values of the predictor due to non-linearity, non-homogeneity of the variance, or an important variable may have been omitted 
from the model or a strong interaction effect may have been overlooked.

In this case, it could be disastrous to delete the observation from the analysis.

- 3) Outliers can occur because of poor sampling of observations in the tail of the distribution.
- **Use extreme caution when deleting observations unless the values are not plausible or you are positive a coding error has occurred!**
 - **If you choose to delete an observation(s), it is often best to report both models (with and without the deleted observation(s)).**

Assessing Outliers

- Jackknife residuals outside the range of ± 3 are often considered potential outliers with residuals outside the range of ± 4 being of great concern
 - But recall that the expected number outside this range will depend on the sample size
- Just because an observation appears to be unusual when compared with the rest of the data does not automatically mean that it should be dropped!
- Deleting observations from the analysis can lead to an underestimation of the variability and result in p-values that are optimistically small.
- The observation should be evaluated for its effect on the analysis.
 - Depending on its location in the prediction space, an outlier can have severe effects on the regression model.
- **Use extreme caution when deleting observations unless the values are not plausible or you are positive a coding error has occurred!**
 - **If you choose to delete an observation(s), it is often best to report both models (with and without the deleted observation(s)).**
- KKNM briefly discuss statistical tests for detecting outliers using a Bonferroni correction.

High Leverage and Influential Points

- The **leverage**, h_i , of an observation is a measure of the geometric distance of the observation's predictor point $(X_{i1}, X_{i2}, \dots, X_{ik})$ from the center point $(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k)$ of the predictor space.
- The quantity h_i , known as the **leverage**, is the i th element on the diagonal of the hat matrix.
 - Recall that the hat matrix, H , is calculate as $X(X'X)^{-1}X'$
 - An element of the hat matrix can be calculated as:

$$h_{ij} = \mathbf{x}_i' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_j$$

- High leverage observations have the potential to be very influential observations, but are not necessarily influential.

High Leverage \rightarrow Potential Influence

- An ***influential observation*** is defined as an observation that has an important influence on the coefficient(s) of the fitted regression line
 - i.e., the observation has a noticeable effect on the regression coefficients

Measures of Influence:

- Cook's Distance (Cook's D).
- DFFITS
- DFBETAS
- An observation can have high leverage (potential influence), but little or no influence
 - But a low leverage point cannot have dramatic influence on the regression coefficients (particularly not on the slope coefficients).

Cook's distance (or Cook's D)

Cook's D is a measure of the influence of an observation. It measures how much the regression coefficients are changed by deleting the particular observation in question.

- Each d_i measures the influence of the i th observation on all n fitted values and is given by:

$$d_i = \frac{(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(-i)})' (\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_{(-i)})}{(p+1)MSE} \quad \& \quad D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})' (\mathbf{X}'\mathbf{X})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})}{(p+1)MSE}$$

where $\hat{\mathbf{Y}}$ is the vector of fitted values when all n observations are included and $\hat{\mathbf{Y}}_{(-i)}$ is the vector of fitted values when the i th observation is deleted. Cook's D can also be expressed as:

$$d_i = \left(\frac{r_i^2}{p+1} \right) \left(\frac{h_i}{1-h_i} \right) = \frac{e_i^2 h_i}{(p+1)MSE(1-h_i)^2}$$

- The d_i depends on both the size of its residual, e_i , and its leverage, h_i
 - The quantity h_i , known as the **leverage**, is the i th element on the diagonal of the hat matrix.
 - Recall that the hat matrix, H , is calculate as $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. An element of the hat matrix can be calculated as: $h_{ij} = \mathbf{x}_i' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_j$
- Cook's D values > 1.0 should be examined more closely, although recent work questions the useful of this measure.

DFFITS

Cook's D measures the influence of the i th observation on all n fitted values. In contrast, $(DFFITS)_i$ is a measure of the influence of the i th observation on the fitted value \hat{Y}_i . The measure is given by

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{(-i)}}{\sqrt{MSE_{(-i)} h_i}} \quad \text{🗨️}$$

where $\hat{Y}_{(-i)}$ is the fitted value of Y_i from the regression model fit with the i th observation deleted.

The denominator is the estimated standard deviation of \hat{Y}_i and is based on the MSE calculated from the regression model fit when the i th observation is deleted.

The resulting standardization represents the number of estimated standard deviations of Y_i that the fitted value increases or decreases with the inclusion of the i th observation in the model.

$(DFFITS)_i$ can also be calculated without refitting the model n times as

$$(DFFITS)_i = r_{(-i)} \sqrt{\frac{h_i}{1 - h_i}}$$

Recall jackknife residual:

$$r_{(-i)} = \frac{e_i}{\sqrt{MSE_{-i} (1 - h_i)}}$$

If h_i near 0 or $r_{(-i)}$ near 0, then little effect of observation.

Any observation with $(DFFITS)_i$ outside the range of $\pm 2\sqrt{(p+1)/n}$ warrants further investigation.

DFBETAS

Both DFFITS and Cook's D measure the influence of an observation on the fitted values (and therefore its effect on the overall model for Cook's D).

Alternatively, we may be interested in an observation's influence on the individual coefficient estimates.

We can get this by measuring the difference between the coefficient estimated with and without the i th observation and standardize this difference by dividing by an estimate of the standard error.

The measure, called $(DFBETAS)_{k,i}$ is given by

$$(DFBETAS)_{k,i} = \frac{\hat{\beta}_k - \hat{\beta}_{k(-i)}}{\sqrt{C_{kk} MSE_{(-i)}}}$$

The denominator is the SE of $\hat{\beta}_k$ without the i th observation.

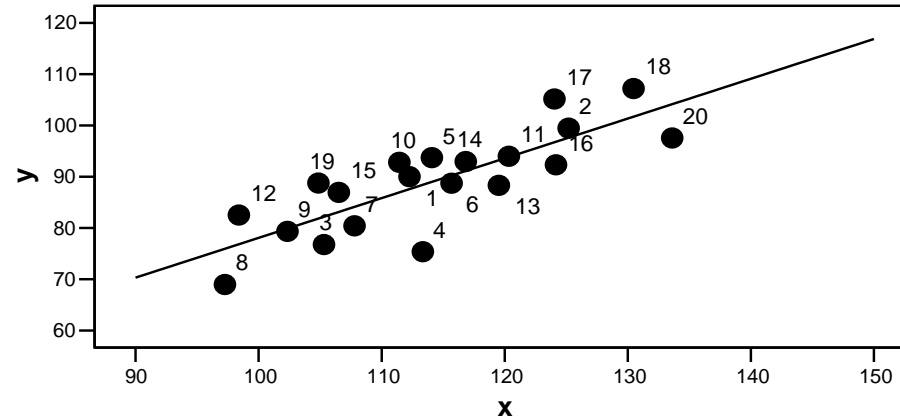
Where C_{kk} is the k th diagonal element of $(X'X)^{-1}$. A large value of DFBETAS indicates that the i th observation has a sizable impact on the k th regression coefficient. The sign of the DFBETAS is also meaningful.

Any observation with $(DFBETAS)_{k,i}$ outside the range of $\pm 2/\sqrt{n}$ warrants further investigation. In smaller data sets, a larger absolute value may be considered meaningful.

No outliers, influential points, or high leverage points

NOTES:

Sxx = 1988.6533
Xbar = 114.175



Largest Cook's D? **#20**

Largest $r(-i)$? **#4**

The REG Procedure

Model: MODEL1

Dependent Variable: y

```
PROC REG DATA=lecture11 a;
  MODEL y = x / R INFLUENCE;
RUN;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1196.83758	1196.83758	34.86	<.0001
Error	18	617.97068	34.33170		
Corrected Total	19	1814.80826			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.48192	15.05876	0.03	0.9748
x	1	0.77578	0.13139	5.90	<.0001

The REG Procedure
 Model: MODEL1
 Dependent Variable: y

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual
1	90.0100	87.5709	1.3341	2.4391	5.705	0.428
2	99.4700	97.6095	1.9532	1.8605	5.524	0.337
3	76.7600	82.1792	1.7531	-5.4192	5.591	-0.969
4	75.3600	88.4165	1.3147	-13.0565	5.710	-2.287
5	93.7000	88.9828	1.3102	4.7172	5.711	0.826
6	88.7200	90.2240	1.3250	-1.5040	5.708	-0.264
7	80.4100	84.1109	1.5551	-3.7009	5.649	-0.655
8	68.9600	75.9420	2.5788	-6.9820	5.261	-1.327
9	79.3300	79.8829	2.0324	-0.5529	5.496	-0.101
10	92.7900	86.9347	1.3586	5.8553	5.700	1.027
11	93.9800	93.8392	1.5404	0.1408	5.653	0.0249
12	82.5100	76.8186	2.4521	5.6914	5.322	1.070
13	88.3100	93.2030	1.4865	-4.8930	5.668	-0.863
14	92.9500	91.1240	1.3562	1.8260	5.700	0.320
15	86.9300	83.1179	1.6517	3.8121	5.622	0.678
16	92.3100	96.8027	1.8541	-4.4927	5.558	-0.808
17	105.1500	96.7096	1.8430	8.4404	5.562	1.518
18	107.1900	101.6978	2.5101	5.4922	5.294	1.037
19	88.7500	81.8301	1.7929	6.9199	5.578	1.241
20	97.5400	104.1338	2.8701	-6.5938	5.108	-1.291

Note:
 $X_1=112.26$
 $Y_1=90.010$

Obs	-2 -1 0 1 2	Cook ' s D	RStudent	Hat Diag H	Cov Ratio	DFFITS
1		0.005	0.4176	0.0518	1.1585	0.0976
2		0.007	0.3283	0.1111	1.2454	0.1161
3	*	0.046	-0.9676	0.0895	1.1061	-0.3034
4	****	0.139	-2.6382	0.0503	0.5943	-0.6074
5	*	0.018	0.8184	0.0500	1.0924	0.1878
6		0.002	-0.2566	0.0511	1.1724	-0.0596
7	*	0.016	-0.6444	0.0704	1.1492	-0.1774
8	**	0.212	-1.3578	0.1937	1.1317	-0.6655
9		0.001	-0.0978	0.1203	1.2730	-0.0362
10	**	0.030	1.0290	0.0538	1.0499	0.2453
11		0.000	0.0242	0.0691	1.2043	0.0066
12	**	0.121	1.0741	0.1751	1.1919	0.4949
13	*	0.026	-0.8569	0.0644	1.1011	-0.2248
14		0.003	0.3122	0.0536	1.1711	0.0743
15	*	0.020	0.6676	0.0795	1.1565	0.1961
16	*	0.036	-0.8002	0.1001	1.1571	-0.2669
17	***	0.126	1.5793	0.0989	0.9462	0.5233
18	**	0.121	1.0397	0.1835	1.2138	0.4929
19	**	0.079	1.2606	0.0936	1.0345	0.4052
20	**	0.263	-1.3169	0.2399	1.2146	-0.7399

Recommendations for Further Investigation:

Jackknife Residual: ± 3 ; ± 4

Cook's D > 1.0

Leverage: $2(p+1)/n = 0.2$

DFFITS $\pm 2 \sqrt{((p+1)/n)} = 2\sqrt{(2/20)} = 0.63$

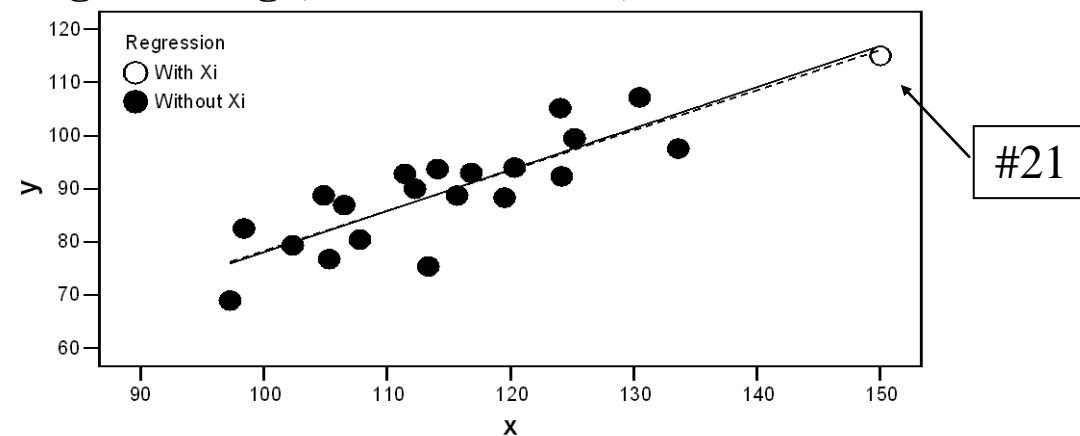
DFBETAS $\pm 2\sqrt{n} = 2\sqrt{20} = 0.45$

The SAS option R requests the statistics through Cook's D.

The SAS option INFLUENCE requests the remaining statistics.

```
-----DFBETAS-----  
Obs  Intercept          x  
  1      0.0267      -0.0184  
  2     -0.0790       0.0861  
  3     -0.2205       0.2016  
  4     -0.1026       0.0501  
  5      0.0181     -0.0018  
  6      0.0037     -0.0089  
  7     -0.1082       0.0955  
  8     -0.6005       0.5732  
  9     -0.0296       0.0277  
 10      0.0852     -0.0649  
 11     -0.0030       0.0035  
 12      0.4398     -0.4183  
 13      0.0885     -0.1062  
 14     -0.0129       0.0192  
 15      0.1325     -0.1194  
 16      0.1717     -0.1889  
 17     -0.3343       0.3680  
 18     -0.3965       0.4204  
 19      0.3013     -0.2766  
 20      0.6264     -0.6583
```

High leverage, little influence, not an outlier



```
PROC REG DATA=lecture11_b;
  MODEL y = x ;
  OUTPUT OUT=diags P=yhat R=e H=h STUDENT=r RSTUDENT=r_i COOKD=cookd DFFITS=dffit;
RUN;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1835.83597	1835.83597	56.26	<.0001
Error	19	619.98676	32.63088		
Corrected Total	20	2455.82272			

Parameter Estimates

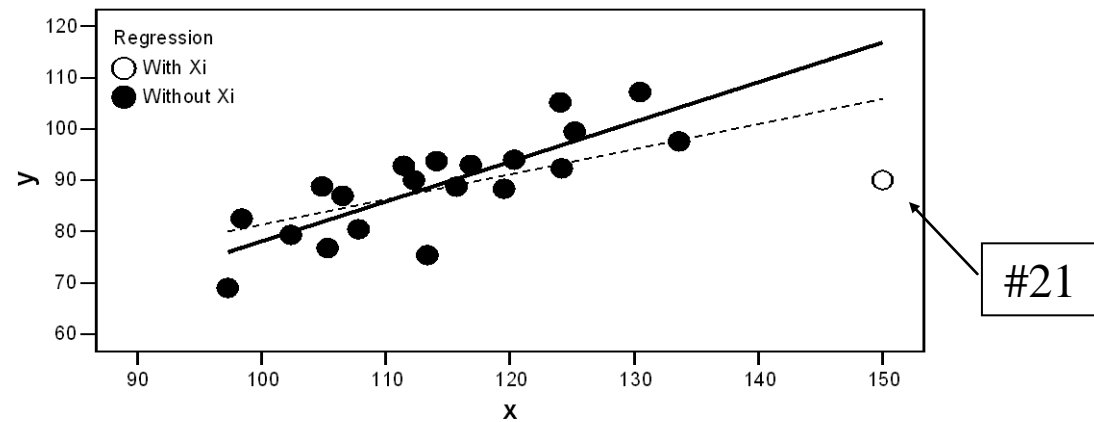
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	2.67034	11.74808	0.23	0.8226
x	1	0.75613	0.10081	7.50	<.0001

```
PROC PRINT DATA=diags;
VAR x y yhat e h r r_i cookd dffit;
RUN;
```

Obs	x	y	yhat	e	h	r	r_i	cookd	dffit
1	112.26	90.01	87.554	2.4560	0.05170	0.44151	0.43196	0.00531	0.10086
2	125.20	99.47	97.338	2.1316	0.07467	0.38793	0.37909	0.00607	0.10768
3	105.31	76.76	82.299	-5.5388	0.08242	-1.01224	-1.01293	0.04602	-0.30358
4	113.35	75.36	88.378	-13.0182	0.04961	-2.33768	-2.69581	0.14264	-0.61594
5	114.08	93.70	88.930	4.7699	0.04863	0.85608	0.84980	0.01873	0.19213
6	115.68	88.72	90.140	-1.4200	0.04763	-0.25472	-0.24835	0.00162	-0.05554
7	107.80	80.41	84.182	-3.7716	0.06796	-0.68390	-0.67401	0.01705	-0.18200
8	97.27	68.96	76.220	-7.2595	0.15549	-1.38290	-1.41934	0.17605	-0.60902
9	102.35	79.33	80.061	-0.7307	0.10464	-0.13518	-0.13164	0.00107	-0.04500
10	111.44	92.79	86.934	5.8561	0.05376	1.05388	1.05713	0.03155	0.25198
11	120.34	93.98	93.664	0.3165	0.05381	0.05695	0.05544	0.00009	0.01322
12	98.40	82.51	77.074	5.4360	0.14279	1.02784	1.02945	0.08799	0.42015
13	119.52	88.31	93.044	-4.7335	0.05174	-0.85095	-0.84451	0.01976	-0.19727
14	116.84	92.95	91.017	1.9329	0.04791	0.34679	0.33861	0.00303	0.07595
15	106.52	86.93	83.214	3.7162	0.07491	0.67639	0.66642	0.01852	0.18964
16	124.16	92.31	96.552	-4.2420	0.06897	-0.76961	-0.76104	0.02194	-0.20713
17	124.04	105.15	96.461	8.6888	0.06835	1.57586	1.64510	0.09110	0.44559
18	130.47	107.19	101.323	5.8668	0.11390	1.09106	1.09687	0.07651	0.39327
19	104.86	88.75	81.959	6.7914	0.08545	1.24320	1.26248	0.07220	0.38589
20	133.61	97.54	103.697	-6.1574	0.14551	-1.16609	-1.17792	0.11577	-0.48608
21	150.00	115.00	116.090	-1.0905	0.41016	-0.24856	-0.24233	0.02148	-0.20208

Obs	Residual	RStudent	Hat	Diag	Cov	-----DFBETAS-----	
			H	Ratio		Intercept	x
1	2.4560	0.4320	0.0517	1.1510	0.1009	0.0385	-0.0283
2	2.1316	0.3791	0.0747	1.1851	0.1077	-0.0553	0.0648
3	-5.5388	-1.0129	0.0824	1.0868	-0.3036	-0.2206	0.1973
4	-13.0182	-2.6958	0.0496	0.5950	-0.6159	-0.1868	0.1235
5	4.7699	0.8498	0.0486	1.0825	0.1921	0.0477	-0.0277
6	-1.4200	-0.2483	0.0476	1.1619	-0.0555	-0.0068	0.0009
7	-3.7716	-0.6740	0.0680	1.1373	-0.1820	-0.1152	0.0996
8	-7.2595	-1.4193	0.1555	1.0671	-0.6090	-0.5402	0.5073
9	-0.7307	-0.1316	0.1046	1.2420	-0.0450	-0.0363	0.0332
10	5.8561	1.0571	0.0538	1.0439	0.2520	0.1099	-0.0852
11	0.3165	0.0554	0.0538	1.1772	0.0132	-0.0031	0.0045
12	5.4360	1.0295	0.1428	1.1593	0.4202	0.3668	-0.3430
13	-4.7335	-0.8445	0.0517	1.0871	-0.1973	0.0353	-0.0557
14	1.9329	0.3386	0.0479	1.1555	0.0760	0.0022	0.0059
15	3.7162	0.6664	0.0749	1.1471	0.1896	0.1299	-0.1145
16	-4.2420	-0.7610	0.0690	1.1233	-0.2071	0.0963	-0.1152
17	8.6888	1.6451	0.0684	0.9037	0.4456	-0.2046	0.2454
18	5.8668	1.0969	0.1139	1.1048	0.3933	-0.2713	0.3000
19	6.7914	1.2625	0.0854	1.0282	0.3859	0.2859	-0.2568
20	-6.1574	-1.1779	0.1455	1.1240	-0.4861	0.3669	-0.3987
21	-1.0905	-0.2423	0.4102	1.8767	-0.2021	0.1816	-0.1900

High leverage, strong influence, outlier



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	772.49534	772.49534	14.07	0.0014
Error	19	1043.16072	54.90320		
Corrected Total	20	1815.65606			

Parameter Estimates

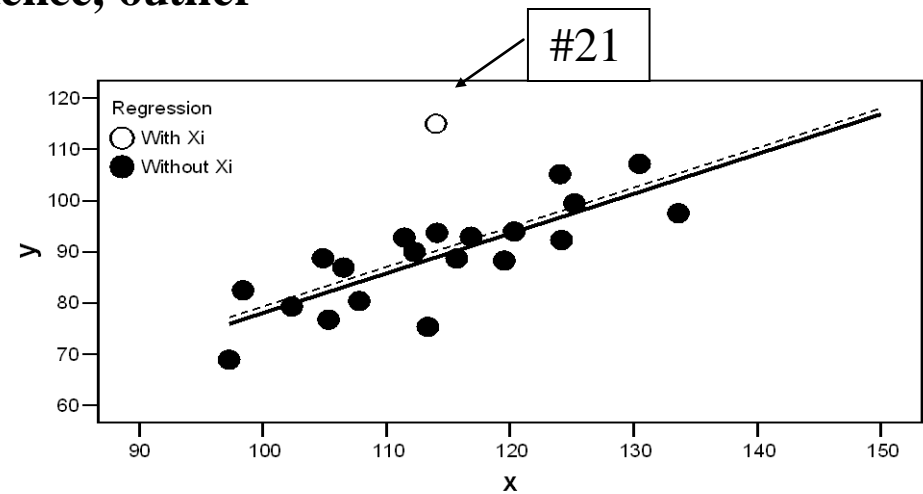
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	32.26301	15.23882	2.12	0.0477
x	1	0.49049	0.13076	3.75	0.0014

Obs	x	y	yhat	e	h	r	r_i	cookd	dffit
1	112.26	90.01	87.325	2.6846	0.05170	0.37206	0.36346	0.00377	0.08487
2	125.20	99.47	93.672	5.7977	0.07467	0.81340	0.80586	0.02669	0.22891
3	105.31	76.76	83.916	-7.1565	0.08242	-1.00827	-1.00874	0.04566	-0.30233
4	113.35	75.36	87.860	-12.5000	0.04961	-1.73046	-1.83512	0.07816	-0.41929
5	114.08	93.70	88.218	5.4819	0.04863	0.75851	0.74971	0.01470	0.16950
6	115.68	88.72	89.003	-0.2829	0.04763	-0.03912	-0.03808	0.00004	-0.00852
7	107.80	80.41	85.138	-4.7278	0.06796	-0.66091	-0.65081	0.01592	-0.17573
8	97.27	68.96	79.973	-11.0129	0.15549	-1.61734	-1.69522	0.24081	-0.72740
9	102.35	79.33	82.465	-3.1346	0.10464	-0.44708	-0.43747	0.01168	-0.14955
10	111.44	92.79	86.923	5.8668	0.05376	0.81396	0.80644	0.01882	0.19222
11	120.34	93.98	91.289	2.6915	0.05381	0.37342	0.36480	0.00397	0.08700
12	98.40	82.51	80.527	1.9828	0.14279	0.28903	0.28194	0.00696	0.11507
13	119.52	88.31	90.886	-2.5763	0.05174	-0.35706	-0.34871	0.00348	-0.08146
14	116.84	92.95	89.572	3.3782	0.04791	0.46724	0.45742	0.00549	0.10260
15	106.52	86.93	84.510	2.4200	0.07491	0.33957	0.33152	0.00467	0.09434
16	124.16	92.31	93.162	-0.8522	0.06897	-0.11920	-0.11606	0.00053	-0.03159
17	124.04	105.15	93.103	12.0466	0.06835	1.68438	1.77754	0.10407	0.48147
18	130.47	107.19	96.257	10.9328	0.11390	1.56744	1.63501	0.15791	0.58621
19	104.86	88.75	83.696	5.0542	0.08545	0.71327	0.70373	0.02377	0.21510
20	133.61	97.54	97.797	-0.2573	0.14551	-0.03757	-0.03657	0.00012	-0.01509
21	150.00	90.00	105.836	-15.8365	0.41016	-2.78287	-3.51920	2.69262	-2.93463

Obs	Residual	RStudent	Hat Diag	Cov	DFFITS	-----DFBETAS-----	
			H			Intercept	x
1	2.6846	0.3635	0.0517	1.1579	0.0849	0.0324	-0.0239
2	5.7977	0.8059	0.0747	1.1217	0.2289	-0.1176	0.1378
3	-7.1565	-1.0087	0.0824	1.0878	-0.3023	-0.2197	0.1965
4	-12.5000	-1.8351	0.0496	0.8319	-0.4193	-0.1272	0.0841
5	5.4819	0.7497	0.0486	1.1013	0.1695	0.0421	-0.0244
6	-0.2829	-0.0381	0.0476	1.1697	-0.0085	-0.0010	0.0001
7	-4.7278	-0.6508	0.0680	1.1411	-0.1757	-0.1112	0.0961
8	-11.0129	-1.6952	0.1555	0.9811	-0.7274	-0.6452	0.6059
9	-3.1346	-0.4375	0.1046	1.2184	-0.1496	-0.1205	0.1104
10	5.8668	0.8064	0.0538	1.0968	0.1922	0.0838	-0.0650
11	2.6915	0.3648	0.0538	1.1603	0.0870	-0.0207	0.0295
12	1.9828	0.2819	0.1428	1.2884	0.1151	0.1005	-0.0939
13	-2.5763	-0.3487	0.0517	1.1593	-0.0815	0.0146	-0.0230
14	3.3782	0.4574	0.0479	1.1435	0.1026	0.0030	0.0079
15	2.4200	0.3315	0.0749	1.1898	0.0943	0.0646	-0.0569
16	-0.8522	-0.1161	0.0690	1.1949	-0.0316	0.0147	-0.0176
17	12.0466	1.7775	0.0684	0.8654	0.4815	-0.2210	0.2652
18	10.9328	1.6350	0.1139	0.9533	0.5862	-0.4044	0.4472
19	5.0542	0.7037	0.0854	1.1539	0.2151	0.1594	-0.1431
20	-0.2573	-0.0366	0.1455	1.3037	-0.0151	0.0114	-0.0124
21	-15.8365	-3.5192	0.4102	0.6629	-2.9346	2.6374	-2.7590

#21 has same leverage as previous example, but much more influence

Low leverage, little influence, outlier



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1190.12071	1190.12071	17.87	0.0005
Error	19	1265.70201	66.61590		
Corrected Total	20	2455.82272			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	1.97332	20.97092	0.09	0.9260
x	1	0.77359	0.18302	4.23	0.0005

Obs	x	y	yhat	e	h	r	r_i	cookd	dffit
1	112.26	90.01	88.817	1.1931	0.04945	0.14993	0.14602	0.00058	0.03330
2	125.20	99.47	98.827	0.6428	0.10883	0.08342	0.08121	0.00042	0.02838
3	105.31	76.76	83.440	-6.6804	0.08706	-0.85664	-0.85037	0.03499	-0.26261
4	113.35	75.36	89.660	-14.3001	0.04795	-1.79565	-1.91807	0.08121	-0.43048
5	114.08	93.70	90.225	3.4751	0.04762	0.43629	0.42680	0.00476	0.09544
6	115.68	88.72	91.463	-2.7426	0.04877	-0.34453	-0.33640	0.00304	-0.07617
7	107.80	80.41	85.367	-4.9567	0.06800	-0.62906	-0.61876	0.01444	-0.16714
8	97.27	68.96	77.221	-8.2608	0.19118	-1.12539	-1.13382	0.14968	-0.55124
9	102.35	79.33	81.151	-1.8206	0.11783	-0.23749	-0.23150	0.00377	-0.08461
10	111.44	92.79	88.183	4.6074	0.05136	0.57959	0.56918	0.00909	0.13243
11	120.34	93.98	95.068	-1.0876	0.06678	-0.13793	-0.13432	0.00068	-0.03593
12	98.40	82.51	78.095	4.4151	0.17262	0.59470	0.58430	0.03689	0.26689
13	119.52	88.31	94.433	-6.1232	0.06203	-0.77463	-0.76617	0.01984	-0.19703
14	116.84	92.95	92.360	0.5900	0.05121	0.07422	0.07225	0.00015	0.01679
15	106.52	86.93	84.376	2.5535	0.07702	0.32565	0.31785	0.00442	0.09182
16	124.16	92.31	98.023	-5.7127	0.09784	-0.73690	-0.72772	0.02944	-0.23965
17	124.04	105.15	97.930	7.2201	0.09664	0.93074	0.92730	0.04633	0.30329
18	130.47	107.19	102.904	4.2859	0.18127	0.58035	0.56994	0.03729	0.26818
19	104.86	88.75	83.092	5.6577	0.09117	0.72712	0.71779	0.02652	0.22735
20	133.61	97.54	105.333	-7.7931	0.23772	-1.09362	-1.09962	0.18648	-0.61406
21	114.00	115.00	90.163	24.8370	0.04763	3.11823	4.34360	0.24316	0.97141

Obs	Residual	RStudent	Hat	Diag	Cov	-----DFBETAS-----	
			H	Ratio		Intercept	x
1	1.1931	0.1460	0.0494	1.1694	0.0333	0.0092	-0.0064
2	0.6428	0.0812	0.1088	1.2494	0.0284	-0.0196	0.0213
3	-6.6804	-0.8504	0.0871	1.1280	-0.2626	-0.1926	0.1768
4	-14.3001	-1.9181	0.0480	0.8068	-0.4305	-0.0723	0.0360
5	3.4751	0.4268	0.0476	1.1466	0.0954	0.0090	-0.0008
6	-2.7426	-0.3364	0.0488	1.1567	-0.0762	0.0053	-0.0117
7	-4.9567	-0.6188	0.0680	1.1462	-0.1671	-0.1031	0.0915
8	-8.2608	-1.1338	0.1912	1.2000	-0.5512	-0.4993	0.4777
9	-1.8206	-0.2315	0.1178	1.2555	-0.0846	-0.0696	0.0653
10	4.6074	0.5692	0.0514	1.1334	0.1324	0.0464	-0.0357
11	-1.0876	-0.1343	0.0668	1.1915	-0.0359	0.0166	-0.0192
12	4.4151	0.5843	0.1726	1.2970	0.2669	0.2382	-0.2271
13	-6.1232	-0.7662	0.0620	1.1140	-0.1970	0.0800	-0.0950
14	0.5900	0.0722	0.0512	1.1737	0.0168	-0.0031	0.0044
15	2.5535	0.3179	0.0770	1.1937	0.0918	0.0627	-0.0567
16	-5.7127	-0.7277	0.0978	1.1654	-0.2396	0.1569	-0.1717
17	7.2201	0.9273	0.0966	1.1235	0.3033	-0.1971	0.2160
18	4.2859	0.5699	0.1813	1.3131	0.2682	-0.2178	0.2303
19	5.6577	0.7178	0.0912	1.1587	0.2273	0.1705	-0.1571
20	-7.7931	-1.0996	0.2377	1.2834	-0.6141	0.5238	-0.5491
21	24.8370	4.3436	0.0476	0.2789	0.9714	0.0991	-0.0166