

### Overview

## Types of Studies

- Controlled Experiments (F = ma, PV = nRT)
- Controlled Experiments with Supplemental variables
- Confirmatory <u>observational</u> studies
- Exploratory <u>observational</u> studies
- Data Collection and preparation
  - How the data are collected? (Design, Nature of the study)
  - Data consistency. (Plots and numerical summaries, logical relations)
  - Is this data analysis-ready? (Format checking, file conversion, etc.)

Home Page

Tite Page

Conens

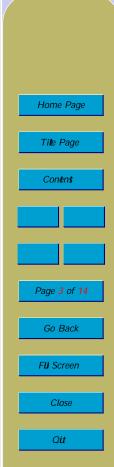
Page 2 of 14

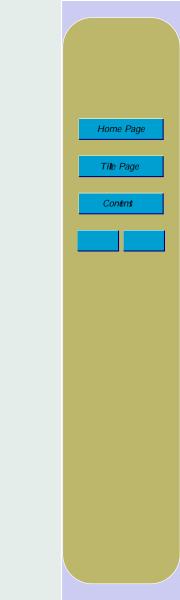
Go Back

F**U** Screen

Close

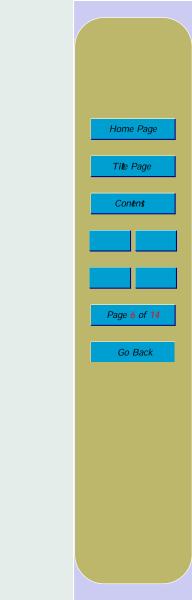
- GIGO (Garbage In; Garbage Out.)
- Reduction of explanatory or predictor variables
   Find parsimonious model with good explanatory/prediction power. Trade-o .
- Model refinement and selection
   Choosing from many "good" models, checking the adequacy of the models, sensitivity of the models, fixing the weak spots.
- Model Validation
   OK to explain what's going on? OK to predict what the future will be?





Home Page
Ti**t**e Page

Conens



## 3. Diagnosis

Checking the adequacy of a regression model

- Improper functional form of a predictor
- Outliers
- Influential observation
- Multicollinearity

## 3.1. Improper functional form of a predictor

- Goal: Detect the suitable form of Y vs  $X_q$  while  $X_1, \dots, X_{q-1}$  in the model.
- Partial Regression Plots:  $e(Y|X_1, \dots, X_{q-1})$  vs.  $e(X_q|X_1, \dots, X_{q-1})$ .
  - $-e(Y/X_1,\cdots,X_{q-1})$ : residual of Y regresses on  $X_1,\cdots,X_{q-1}$

Home Page

Tite Page

Conens

Page 7 of 14

Go Back

F**U** Screen

Close

- 
$$e(X_q/X_1, \dots, X_{q-1})$$
: residual of  $X_q$  regresses on  $X_1, \dots, X_{q-1}$ 

Why bother?

#### 3.2. Outliers

- Rationale: The model (fitted) shouldn't be a ect by just few points.
- LSE is EXTREMELY sensitive to outliers. Example.
- Detection: Residual-based tests and plots towards outlying Y

Tite Page

Conens

Page 8 of 14

Go Back

Fit Screen

Qit

Home Page

$$r_i = \frac{e_i}{MSE(1-h_{ii})}$$
 since 
$${}^2(e) = {}^2(I-H), \quad H = X(XX)^{-1}X.$$

Deleted Residuals: With or Without You. Outlying Y.

$$d_i = Y_i - \hat{Y_{i(i)}}$$

- Studentized Deleted Residual:

$$t_i = \frac{d_i}{s(d_i)}$$
 where  $s(d_i) = \overline{MSE_{(i)}(1 - h_{ii})}$ 

- Hat matrix Leverage values Outlying X
  - -0  $h_{ii}$  1,  $\sum_{i=1}^{n} h_{ii} = p$ .
  - $-\bar{h}_{ii} = \frac{p}{n}$ . 2p/n, extreme  $h_{ii}$ , outside (0.2, 0.5)
  - $-h_{new} =$

Home Page

Tite Page

Conens

Page 9 of 14

Go Back

F**U** Screen

Close

#### 3.3. Influential obs

 $(DFFITS)_i = \frac{Y_i - Y_{i(i)}}{\overline{MSE_{(i)bij}}}$ 

Flag: If |DFFITS| > 1 for small/medium data set or > 2  $\overline{p/n}$ , large data set.

Cook's Distance

$$D_{i} = \frac{\prod_{j=1}^{n} (\hat{Y}_{j} - \hat{Y}_{j(i)})}{pMSE} = \frac{e_{i}^{2}}{pMSE} \frac{h_{ii}}{(1 - h_{ii})^{2}} \qquad F_{p,n-p}$$

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

where  $c_{kk}$  is the diagonal entries of  $(X|X)^{-1}$  Flag: DFBETAS > 1 for small/medium data; > 2/ $\overline{n}$ . Change of signs.

Home Page

Tite Page

Conens

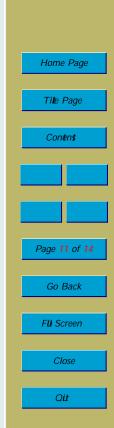
Page 10 of 14

Go Back

F**U** Screen

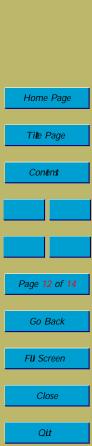
Close

# • DF<u>INF</u>



 $R_k^2$ :  $R^2$  of  $X_k$  regressing on the other X s.

Flag: Larger than 10 or VIF





Home Page

Tite Page

Conens