

# Multiple Regression Analysis

Applied Regression and Other Multivariable Methods  
Sections 8-1 - 8-6

8

## Preview

- In many ways, extension of single variable regr
- Consider more than one indep variable (predictor)
- Considerably more difficult because
  - More difficult to choose “best” model
  - More difficult to interpret “best” model
  - Harder to visualize relationship (multi-dimensional)
  - More difficult to estimate “best” model
- Consider response  $Y$  and predictors  $X_1$  and  $X_2$
- Numerous more relationships to consider
  - Model 1:  $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + E$
  - Model 2:  $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + E$
  - Model 3:  $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 + \beta_4X_1^2 + \beta_5X_2^2 + E$
- Must also consider relationships among predictors

8-1

## Preview

- Consider response  $Y$  and predictors  $X_1$  and  $X_2$
- Three variables  $\rightarrow$  three dimensional problem
- In general,  $k$  predictors  $\rightarrow (k + 1)$  dimensions
- **Goal:** find  $k$  dimensional surface which minimizes

$$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

- Deviations still measured in  $Y$  direction only
- In two dimensions, most simple surface is a plane
  - Model:  $\mu_{Y|X_1, X_2} = \beta_0 + \beta_1X_1 + \beta_2X_2$
- Can add curve to surface using  $X_1^2$ ,  $X_2^2$  or  $X_1X_2$

8-2

## Assumptions

- Similar assumptions as single variable regression
- **Existence**
  - For any combo of  $X$ 's,  $Y$  is a rv with finite mean and var
- **Independence**
  - The  $Y$ 's are statistically independent of one another
  - Common violation - same indiv measured over time
  - Repeated measures, Mixed models, GEE techniques
- **Linearity**
  - Mean of  $Y$  is a linear function of  $X$ 's
$$\hat{Y}_i = \mu_{Y|X_1, X_2, \dots, X_k} = \beta_0 + \sum_{i=1}^k \beta_i X_i$$
  - Results in response surface
  - Can have  $X_5 = X_1^2$  and  $X_6 = X_2X_3$
  - $E$  represents the error component,  $E_i = Y_i - \hat{Y}_i$

8-3

## Assumptions

- **Homoscedasticity**

- The variance of  $Y$  is constant,  $\text{Var}(Y|X_1, X_2, \dots, X_k) \equiv \sigma^2$
- This implies  $\text{Var}(E) = \sigma^2$

- **Normality**

- For any fixed combination of the  $X$ 's,

$$Y \sim N(\mu_{Y|X_1, X_2, \dots, X_k}, \sigma)$$

- This in turn means

$$E \sim N(0, \sigma)$$

- If needed, can **transform**  $Y$  to achieve normality
- Distribution justifies the use of  $t$  and  $F$  tests

8-4

## Least Squares Approach

- Similar idea to single variable regression
- Estimates are determined by minimizing

$$\sum_{i=1}^n \left( Y_i - \beta_0 - \sum_{j=1}^k \beta_j X_{ji} \right)^2$$

- Solution more complicated due to more predictors
- Can be presented succinctly using matrices

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

- This approach discussed in Appendix B (we'll rely on SAS)
- **Properties**
  - Each of the  $\hat{\beta}$ 's is a linear function of the  $Y$ 's
  - Correlation between  $Y$  and  $\hat{Y}$  is maximized
  - Approach coincides with multivariate normal dist

8-5

## ANOVA Table

- Very similar to simple linear regression

$$\begin{aligned} \sum (Y_i - \bar{Y})^2 &= \sum (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 \\ \text{SSY} &= (\text{SSY} - \text{SSE}) + \text{SSE} \end{aligned}$$

- Differences are

Regression or model df equals  $k$  (# of predictors)

Error df equals  $n - 1 - k$

More parameter estimates and std errors

The  $\sqrt{R^2}$  is the multiple correlation coeff

8-6

## SAS Procedures

- SAS has interactive data analysis

Solutions → Analysis → Interactive Data Analysis

Select work library and appropriate SAS dataset

- Gives pairwise scatterplots
- Can click on point, highlighted in all plots

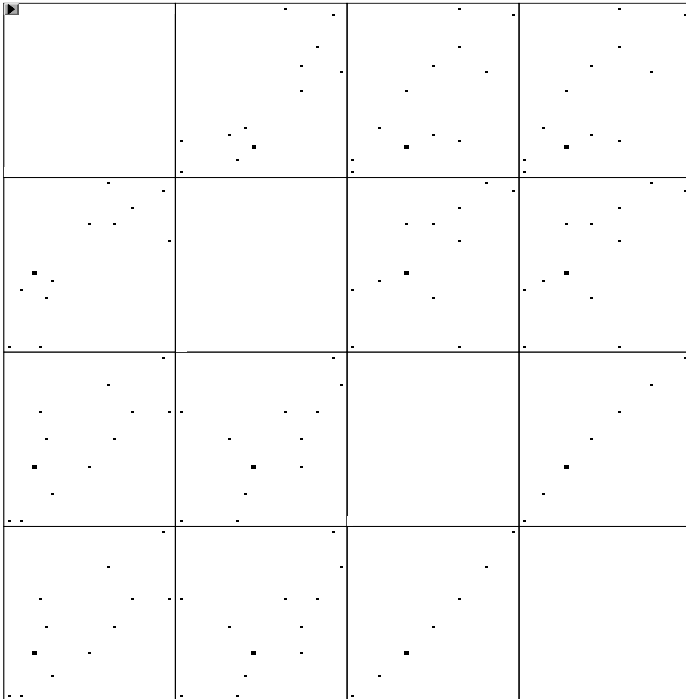
```
options nocenter ls=72;
```

```
data table8;
infile 'I:\www\datasets502\example8-1.dat';
input child wgt hgt age;
agesq = age*age;
```

```
proc print;
```

```
proc reg simple corr;
model wgt = hgt;
model wgt = age;
model wgt = hgt age;
model wgt = hgt agesq;
model wgt = hgt age agesq;
run;
```

8-7



8-8

Descriptive Statistics

Variable	Sum	Mean	Uncorrected	
			SS	Variance
Intercept	12.00000	1.00000	12.00000	0
hgt	633.00000	52.75000	33903	46.56818
age	106.00000	8.83333	976.00000	3.60606
agesq	976.00000	81.33333	91684	1118.42424
wgt	753.00000	62.75000	48139	80.75000

Variable	hgt	Correlation		
		age	agesq	wgt
hgt	1.0000	0.6138	0.6154	0.8143
age	0.6138	1.0000	0.9944	0.7698
agesq	0.6154	0.9944	1.0000	0.7665
wgt	0.8143	0.7698	0.7665	1.0000

The REG Procedure  
Model: MODEL1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	588.92252	588.92252	19.67	0.0013
Error	10	299.32748	29.93275		
Corrected Total	11	888.25000			

Root MSE	5.47108	R-Square	0.6630
Dependent Mean	62.75000	Adj R-Sq	0.6293
Coeff Var	8.71886		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	6.18985	12.84875	0.48	0.6404
hgt	1	1.07223	0.24173	4.44	0.0013

8-9

Model: MODEL2

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	526.39286	526.39286	14.55	0.0034
Error	10	361.85714	36.18571		
Corrected Total	11	888.25000			

Root MSE	6.01546	R-Square	0.5926
Dependent Mean	62.75000	Adj R-Sq	0.5519
Coeff Var	9.58638		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	30.57143	8.61371	3.55	0.0053
age	1	3.64286	0.95512	3.81	0.0034

Model: MODEL3

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	692.82261	346.41130	15.95	0.0011
Error	9	195.42739	21.71415		
Corrected Total	11	888.25000			

Root MSE	4.65984	R-Square	0.7800
Dependent Mean	62.75000	Adj R-Sq	0.7311
Coeff Var	7.42605		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	6.55305	10.94483	0.60	0.5641
hgt	1	0.72204	0.26081	2.77	0.0218
age	1	2.05013	0.93723	2.19	0.0565

8-10

Model: MODEL4

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	689.64995	344.82498	15.63	0.0012
Error	9	198.60005	22.06667		
Corrected Total	11	888.25000			

Root MSE	4.69752	R-Square	0.7764
Dependent Mean	62.75000	Adj R-Sq	0.7267
Coeff Var	7.48608		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	15.11754	11.79690	1.28	0.2321
hgt	1	0.72598	0.26333	2.76	0.0222
agesq	1	0.11480	0.05373	2.14	0.0614

Model: MODEL5

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	693.06046	231.02015	9.47	0.0052
Error	8	195.18954	24.39869		
Corrected Total	11	888.25000			

Root MSE	4.93950	R-Square	0.7803
Dependent Mean	62.75000	Adj R-Sq	0.6978
Coeff Var	7.87172		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	3.43843	33.61082	0.10	0.9210
hgt	1	0.72369	0.27696	2.61	0.0310
age	1	2.77687	7.42728	0.37	0.7182
agesq	1	-0.04171	0.42241	-0.10	0.9238

8-11