Multiple Regression

Definition

Multiple Regression Equation

A linear relationship between a dependent variable y and two or more independent variables ($x_1, x_2, x_3 \ldots, x_k$)

Model:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + e$$

Multiple Regression

Definition

Multiple Regression Equation

A linear relationship between a dependent variable y and two or more independent variables ($x_1, x_2, x_3 \ldots, x_k$)

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$$

$$\hat{y} = b_0 + b_1\,x_1 + b_2\,x_2 + b_3\,x_3 + \ldots + b_k\,x_k$$

(General form of the estimated multiple regression equation)

n = sample size

k = number of independent variables

$\hat{y}$ = predicted value of the dependent variable y

$x_1, x_2, x_3 \ldots, x_k$ are the independent variables

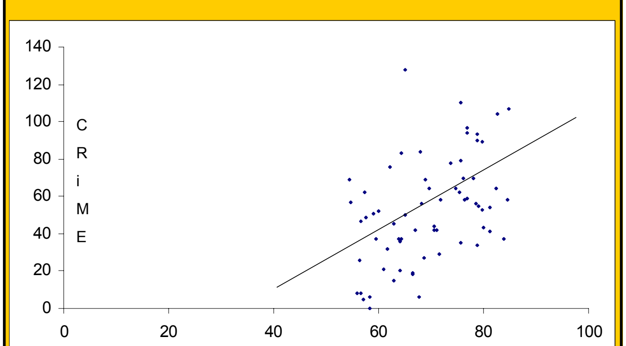
Notation

β_0 = the y-intercept, or the value of y when all of the predictor variables are 0

b_0 = estimate of β_0 based on the sample data

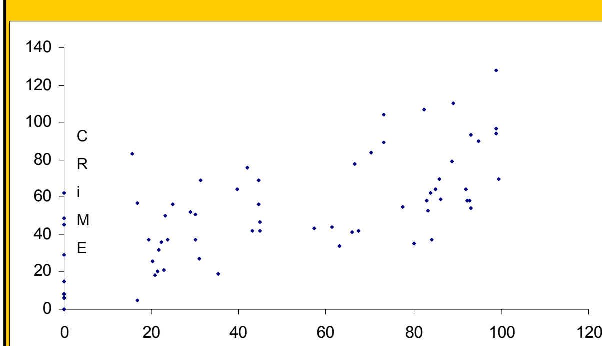
$\beta_1, \beta_2, \beta_3 \ldots, \beta_k$ are the population coefficients of the independent variables $x_1, x_2, x_3 \ldots, x_k$

$b_1, b_2, b_3 \ldots, b_k$ are the sample estimates of the coefficients $\beta_1, \beta_2, \beta_3 \ldots, \beta_k$



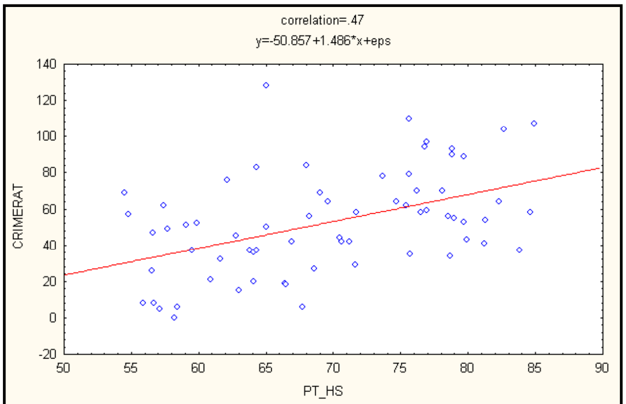
Percent High School

R=.47



Urbanization

R=.68



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	24732	12366	28.54	<.0001
Error	64	27730	433.28847		
Corrected Total	66	52462			

Overall Regression Analysis: A Test of the Multiple R

Root MSE	20.81558	R-Square	0.4714 (R=.686)
Dependent Mean	52.40299	Adj R-Sq	0.4549
Coeff Var	39.72213		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	59.11807	28.36531	2.08	0.0411
hs	1	-0.58338	0.47246	-1.23	0.2214
Urb	1	0.68250	0.12321	5.54	<.0001

Notice that the slope of hs is negative

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-50.85690	24.45065	-2.08	0.0415
hs	1	1.48598	0.34908	4.26	<.0001

When hs is by itself the slope is positive

Pearson Correlation Coefficients, N = 67
Prob > |r| under H0: Rho=0

	crate	incom	hs	Urb
crate	1.00000	0.43375 0.0002	0.46691 <.0001	0.67737 <.0001
incom	0.43375 0.0002	1.00000	0.79262 <.0001	0.73070 <.0001
hs	0.46691 <.0001	0.79262 <.0001	1.00000	0.79072 <.0001
Urb	0.67737 <.0001	0.73070 <.0001	0.79072 <.0001	1.00000

Multiple Regression SAS Setup

```
• proc corr; run;
• proc reg; model crate= hs Urb;
• plot crate*hs;run;
• proc reg; model crate= hs;run;
```

Adjusted R^2

Definitions

- **Multiple coefficient of determination**

  a measure of how well the multiple regression equation fits the sample data

- **Adjusted coefficient of determination**

  the multiple coefficient of determination R^2 modified to account for the number of variables and the sample size

Adjusted R^2

Adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{(n - 1)}{[n - (k + 1)]}(1 - R^2)$$

Adjusted R^2

$$\text{Adjusted } R^2 = 1 - \frac{(n - 1)}{[n - (k + 1)]}(1 - R^2)$$

where n = sample size

k = number of independent (x) variables

Including the Three Variables

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	24804	8268.16424	18.83	<.0001
Error	63	27658	439.00995		
Corrected Total	66	52462			

Overall Test

Root MSE	20.95256	**R-Square**	0.4728
Dependent Mean	52.40299	**Adj R-Sq**	0.4477
Coeff Var	39.98353		

Tests of Hypotheses

- Overall test:
  Null: All of the population regression weights are zero.
  Alternative: Not all are zero

The overall F Test

$$F_{k,n-k-1} = \frac{R^2 / k}{(1-R^2)/(n-k-1)}$$

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	59.71473	28.58953	2.09	0.0408
incom	1	-0.38309	0.94053	-0.41	0.6852
hs	1	-0.46729	0.55443	-0.84	0.4025
Urb	1	0.69715	0.12913	5.40	<.0001

Individual Tests

- Test for β_1:
  $Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$
  $Y = \beta_o + \beta_2 x_2 + \beta_3 x_3 + e$
- Test for β_2:
  $Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$
  $Y = \beta_o + \beta_1 x_1 + \beta_3 x_3 + e$
- Test for β_3:
  $Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$
  $Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + e$

F Test for Restricted Models

$$F_{k-g,n-k-1} = \frac{(R_f^2 - R_r^2)/(k-g)}{(1-R_f^2)/(n-k-1)}$$

More on SAS

- Model options:

  Model y = x1 x2 / R partial p stb;

  R– residual analysis

  Partial– partial regression scatter plot

  P– predicted values

  Stb– standardized regression weights

Standardized Regression Weights

$$b_i^* = b_i\left(\frac{S_{x_i}}{S_y}\right)$$

Generally, the standardized regression weights fall between 1 and –1. However, they can larger than one (or less than –1).

Obtaining the Standardized Regression Weights

- Model Statement
  - Model crate = incom hs urb /stb;

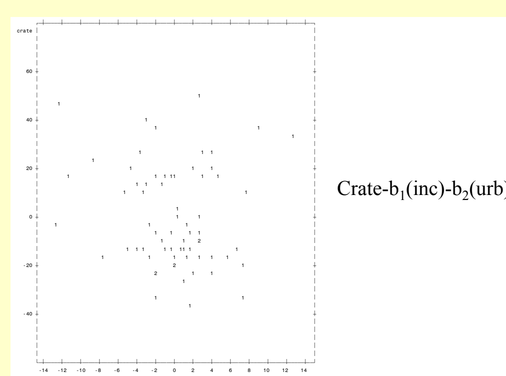
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	59.71473	28.58953	2.09	0.0408	0
incom	1	-0.38309	0.94053	-0.41	0.6852	-0.06363
hs	1	-0.46729	0.55443	-0.84	0.4025	-0.14683
Urb	1	0.69715	0.12913	5.40	<.0001	0.83996

Partial Regression Plots

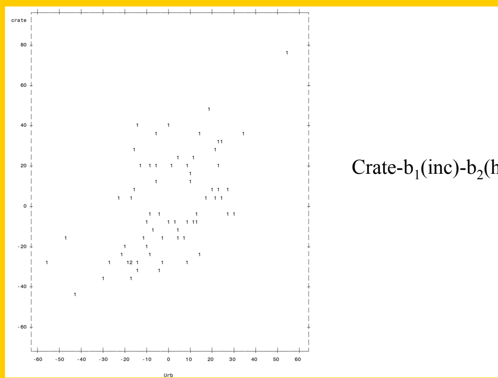
- Plot of two residuals
- The regression line in this plot corresponds to the regression weight in the overall model.
- Model Statement

  / partial



Crate-b_1(inc)-b_2(urb)

Hs-b_1(inc)-b_2(urb)



Crate-b_1(inc)-b_2(hs)

Urb-b_1(inc)-b_2(hs)

Interaction of two Variables

- Just as in ANOVA we can have interaction effects in a multiple regression analysis.
- For quantitative variables, interaction is present when the relationship between the explanatory variable and the response changes as the levels of another variable changes.
- Consider crime rate as a function of hs and urb. If the relationship (slope) between crime rate and urb changes as hs changes, we have an interaction between urb and hs.

Testing for an Interaction Effect

- We test for interaction effect by comparing a model with interaction to a model without interaction:

$$Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2) + e$$

$$Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + e$$

SAS Setup for Interaction

- Create a the product variable in the data statement:
  - Data new; input y x z; xz=x*z; cards;
  - Model y = x z xz;

Root MSE	20.82583	**R-Square**	0.4792
Dependent Mean	52.40299	**Adj R-Sq**	0.4544
Coeff Var	39.74168		

Model crate= hs urb hs*urb (Looking for an interaction)

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	19.31754	49.95871	0.39	0.7003
hs	1	0.03396	0.79381	0.04	0.9660
Urb	1	1.51431	0.86809	1.74	0.0860
hsurb	1	-0.01205	0.01245	-0.97	0.3367

Testing the Interaction

SAS Setup for Interaction

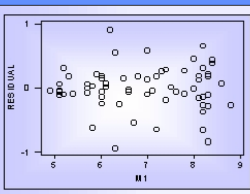
```
data crime; input crate incom
  hs Urb; hsurb= hs*urb;cards;
104 22.1    82.7    73.2

proc reg; model crate= hs Urb
  hsurb;
run;
```

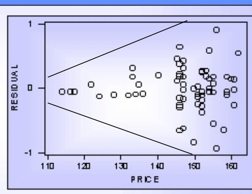
Assessing The Fit of the Model

Looking further at residuals

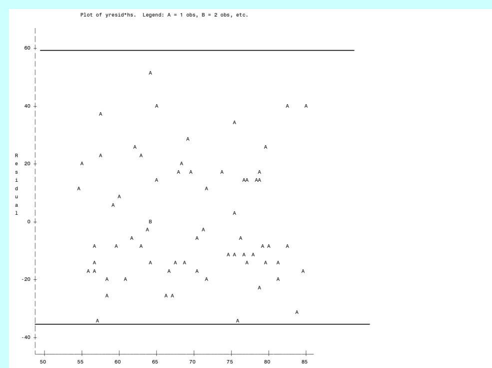
Residual Plots in Regression



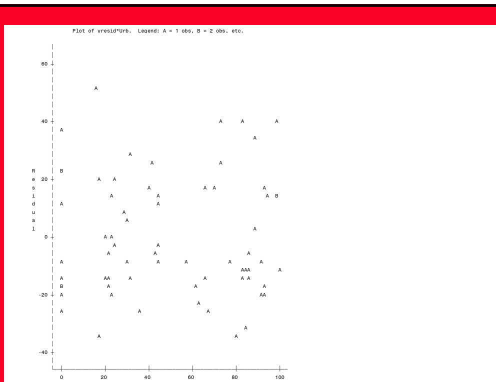
Residuals Plotted Against M1
(Apparently Random)



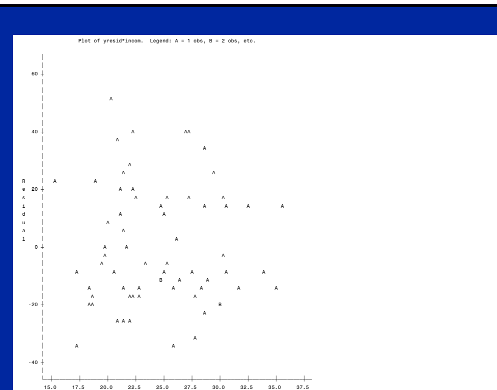
Residuals Plotted Against Price
(Apparent Heteroscedasticity)



Residual by hs



Residual by Urb



Residual by income

SAS Setup

```
• proc reg;
• model crate= incom hs Urb;
• output out=new p=yhat r=yresid;
• proc plot data=new;
• plot yresid*hs;
• plot yresid*urb;
• plot yresid*incom;run;
```

Assumptions

- Linearity– the relationship between the dependent variable and independent variables is linear.
- Normality– y is independently normally distributed
  Independently distributed random errors with a mean of zero.
- Homoskedasticity– the conditional variances of y given x are all equal.

Investigating Multicollinearty

- SAS Model Statement:
- `model crate= incom hs Urb / vif collin;`
- `When the explanatory variables are highly correlated (multicollinearity) the standard errors of the regression weights tend to get very large.`

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation			
			Intercept	incom	hs	Urb
1	3.78327	1.00000	0.00053670	0.00082225	0.00029978	0.00648
2	0.20397	4.30678	0.00735	0.00127	0.00107	0.39725
3	0.00983	19.61619	0.22868	0.81261	0.00944	0.29914
4	0.00293	35.92811	0.76343	0.18530	0.98919	0.29712

Using the Multicollinearity Indices

- Look for conditioned indices larger than 30
- If an index is large than 30, identify variables with proportion indices larger than .90
  - Proportion of variance in each coefficient attributable to the condition index.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	59.71473	28.58953	2.09	0.0408	0
incom	1	-0.38309	0.94053	-0.41	0.6852	2.91618
hs	1	-0.46729	0.55443	-0.84	0.4025	3.62675
Urb	1	0.69715	0.12913	5.40	<.0001	2.89274

Variance Inflation Factor (VIF)

$$R^2_{i.rest} = \frac{VIF - 1}{VIF}$$

VIF– the variance of the weight is inflated by this quantity.

Root MSE	4.72386	**R-Square**	0.7243
Dependent Mean	69.48955	**Adj R-Sq**	0.7157
Coeff Var	6.79794		

Model: hs = income urbanization

Collinearity Diagnostics

			Proportion of Variation		
Number	**Eigenvalue**	**Condition Index**	**Intercept**	**incom**	**Urb**
1	2.80034	1.00000	0.00293	0.00203	0.01645
2	0.19004	3.83868	0.03655	0.00499	0.50952
3	0.00962	17.06408	0.96052	0.99299	0.47402

Test that a subset of regression weights are equal to zero

- SAS Test Statement:

Or,

```
test incom, hs;
```

Results from the Joint test $\beta_{hs}=\beta_{incom}=0$

Test 1 Results for Dependent Variable crate

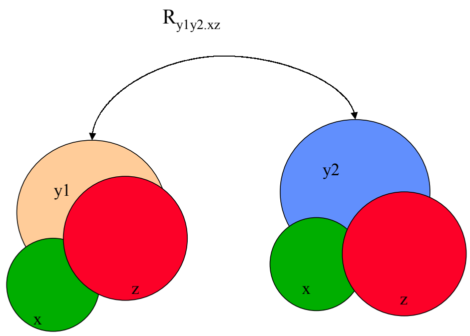
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	2	366.72473	0.84	0.4385
Denominator	63	439.00995		

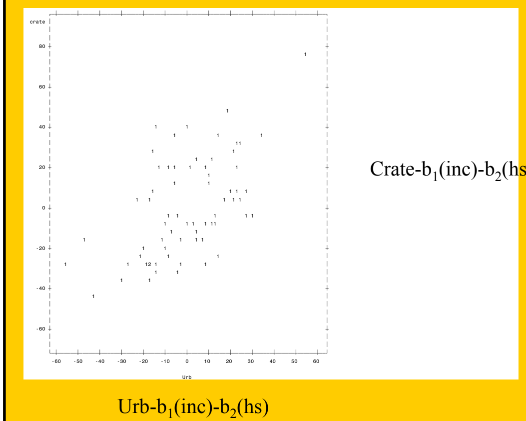
Partial Correlations

- In a partial correlation a variable is partial out of both variables

$r_{x1x2\,.\,x3}$

$$r^2_{12.3} = \frac{R^2_{1.23} - R^2_{1.3}}{1 - R^2_{1.23}}$$





Semipartial Correlations

- The variable is partial out from only one of the variables. The squared semipartial correlation is given by

$$r^2_{1(2.3)} = R^2_{1.23} - R^2_{1.3}$$

Relationship between Semipartials and the Multiple R^2

$$R^2_{y.1234} = r^2_{y.1} + r^2_{y(2.1)} + r^2_{y(3.12)} + r^2_{y(4.123)}$$

Finding the Best Multiple Regression Equation

1. Use common sense and practical considerations to include or exclude variables and always plot the data.

2. Instead of including almost every available variable, include relatively few independent (x) variables, weeding out independent variables that don't have an effect on the dependent variable, remember collinearity.

3. Select an equation having a value of adjusted R^2 with this property: If an additional independent variable is included, the value of adjusted R^2 does not increase by a substantial amount.

4. For a given number of independent (x) variables, select the equation with the largest value of adjusted R^2.

5. You want overall significance with all of the regression weights being significant also.