STOR 455
STATISTICAL METHODS I

Jan Hannig

Why REGRESSION?

- Remember in SLR
  - b_1=SXY/SSX=r (SSY/SSX)$^{1/2}$= r S_Y/S_X
  - b_0= $\overline{Y}$ - b_1 $\overline{X}$
  - The equation

$$\hat{Y} = b_0 + b_1 X = \bar{Y} + r\frac{S_y}{S_X}(X - \bar{X})$$

- Remark
  - If predictor is at the mean of X, response is predicted at the mean of Y
  - If the predictor is 1 sd above the mean of X then the response is predicted r sd above the mean of Y
  - Regression toward the mean

Multiple Regression (Chapter 4)

- More than one predictor.
- Relationship still linear
- Questions
  - How to fit
  - How to spot departures from assumptions
  - How to select important predictors

Observables in MLR (Section 4.2)

- For each population item we observe p variables $Y, X_1, \ldots, X_{p-1}$
  - Y is the response (only one response value per item)
  - $X_1, \ldots, X_{p-1}$ are the predictors (multiple predictors per item)
- Two possible assumptions
  - $(Y, X_1, \ldots, X_{p-1})$ are jointly normal
  - $X_1, \ldots, X_{p-1}$ are fixed and every subpopulation $Y|X_1, \ldots, X_{p-1}$ is normal

Studios Example

- Dwaine Studios currently operates in 21 medium size cities, specialized in portraits of children.
- They want to expand to other cities.
- Want to know: relationship between sales, young population, and disposable income

Graphical and numerical summaries for individual and pairs of observables

- Histogram (proc univariate)
- Scatter plot (SAS Macro scatter.sas)
- Mean, s.d., min, max (proc means)
- Correlation (proc corr)

Do It in SAS

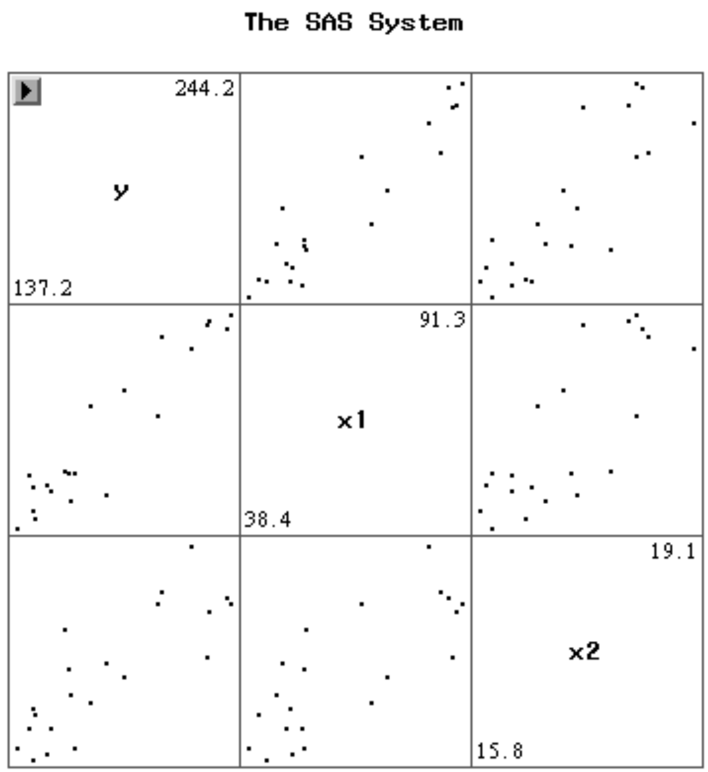
```
*Data shown on page 237 of the OPTIONAL
  textbook - file CH06FI05.txt;
data studios;
  input x1 x2 y;
  x1x2=x1*x2;
  label x1='targtpop'
        x2='dispoinc';
cards;
  68.5  16.7  174.4
  45.2  16.8  164.4
  91.3  18.2  244.2
  ...

  52.3  16.0  166.5
  ;

run;
```

Do it in SAS

```
* Descriptive statistics;
proc means data=studios;
run;
* Check correlation;
proc corr data = studios;
run;
proc univariate data = studios
  noprint;
var y x1 x2;
histogram y x1 x2;
run;
```

Do it in SAS

```
* Making scatter
  plot using
  macro;
%include "T:\...
  \Macro
  \scatter.sas";
%scatter(data =
  studios, var = y
  x1 x2);
```



Multiple Regression Model (Section 4.3)

- $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{ip-1} + \xi_i$
- Y_i is the value of the response variable for the i^{th} case
- β_0 is the intercept
- $\beta_1, \beta_2, \ldots, \beta_{p-1}$ are the regression coefficients for the explanatory variables
- X_{ik} is the value of the k^{th} explanatory variable for the i^{th} case
- ξ_i are independent normally distributed random errors with mean 0 and variance σ^2

Do it in SAS

```
proc reg data = studios;
   model y = x1;
run;
proc reg data = studios;
   model y = x2;
run;
*MLR;
proc reg data=studios;
*  model y=x1 x2;
run;
```

Parameters of MLR

- β_0 the intercept
- $\beta_1, \beta_2, \ldots, \beta_{p-1}$ the regression coefficients for the explanatory variables
- Interpretation of β_i
- σ^2 the variance of the error term

Predictors

- Two types of predictors
  - Basic observable variables X
  - Derived variables X^2,log(X),…
- Both are used!
- Example
  - In our car.dat we have maintenance cost as response (Y) and miles driven as predictor (X_1)
  - The straight line does not fit, quadratic regression is needed. Create $X_2=X_1^2$ and use $Y=\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon$

MLR in Matrix Form (Section 4.4)

$$\underset{n \times 1}{Y} = \underset{n \times p}{X} \underset{p \times 1}{\beta} + \underset{n \times 1}{\varepsilon}$$

$$\varepsilon \sim N(0, \sigma^2 I)$$

$$Y \sim N(X\beta, \sigma^2 I)$$

LS Estimation

$$Y = X\beta + \varepsilon$$

$$\min(Y - Xb)'(Y - Xb)$$

$$X'Xb = X'Y$$

LS Estimator

$$b = (X'X)^{-1}X'Y$$

$$\begin{aligned}\hat{Y} &= Xb \\ &= X(X'X)^{-1}X'Y \\ &= HY\end{aligned}$$

Residuals

$$\begin{aligned} e &= Y - \hat{Y} \\ &= Y - HY \\ &= (I - H)Y \end{aligned}$$

$$s^2 = \frac{e'e}{n-p}$$

Distribution of b

$$b = (X'X)^{-1}X'Y$$

$$Y \sim N(X\beta, \sigma^2 I)$$

$$\begin{aligned} E(b) &= ((X'X)^{-1}X')X\beta \\ &= \beta \end{aligned}$$

$$\begin{aligned} cov(b) &= \sigma^2((X'X)^{-1}X')((X'X)^{-1}X') \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

Estimation of variance of b

$$b \sim N(\beta, \sigma^2 (X'X)^{-1})$$

$$\sigma^2 (X'X)^{-1}$$

is estimated by

$$s^2 (X'X)^{-1}$$

Estimation of σ^2

$s^2 = (\Sigma e_i^2)/(n\text{-}p)$

$= SSE/dfE$

$= MSE$

$s = sqrt(s^2)$

$= \text{Root } MSE$

Residual Analysis (Section 4.5)

- Recall $H=X(X'X)^{-1}X'$ - matrix $H=(h_{ij})$
- Standardized residuals

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$$

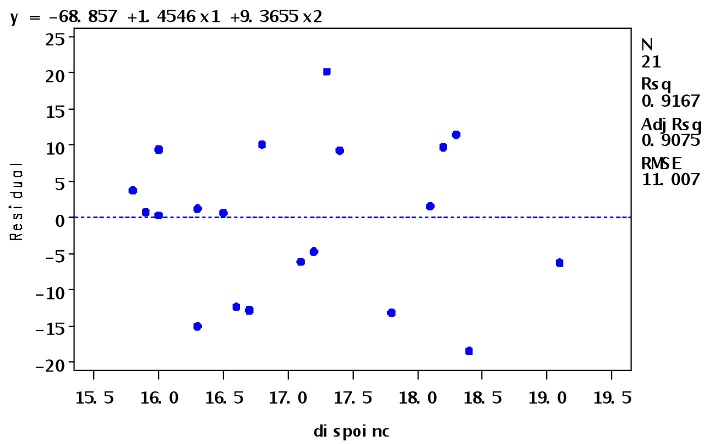
- Similarly as with SLR we should look at
  - Plot of r vs predictors X_i (p –plots)
  - Plot of r vs predicted values $\hat{Y}$
  - Gaussian QQ- Plot of r

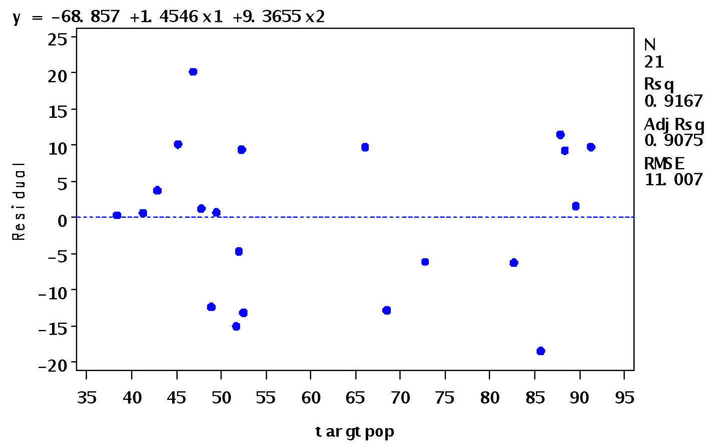
Do it in SAS

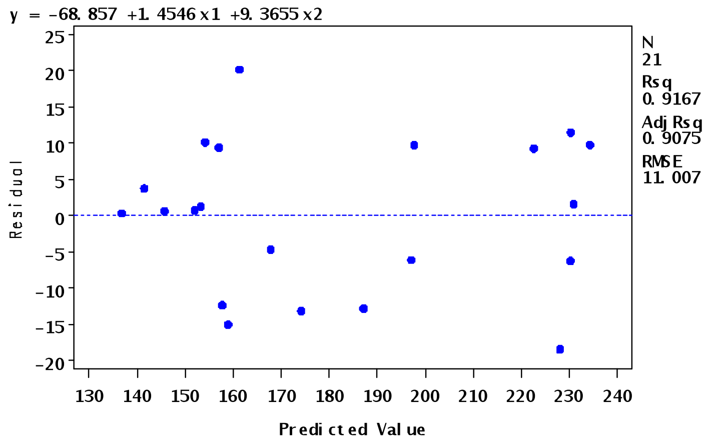
```
* plot residuals, QQ plot;
proc reg data = studios
   noprint;
  model y = x1 x2;
  plot student. * (x1 x2
   p.) ;
  plot student. * nqq.;
run;


*Alternative way of
   plotting;
proc reg data = studios;
  model y = x1 x2;
  output out=output p =
   fitted student =
   residual;
run;
```

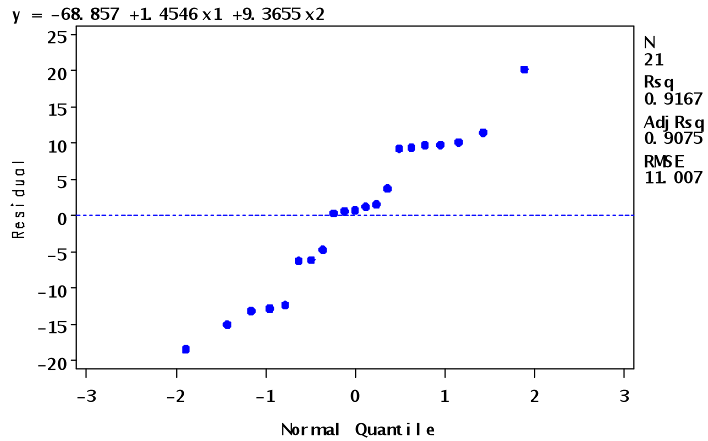
```
proc gplot data = output;
  plot residual*fitted;
  plot residual*x1;
  plot residual*x2;
  plot residual*x1x2;
run;

proc univariate data =
   output noprint ;
  qqplot residual / normal;
run;
*End of an alternative way
   of plotting;
```

Do it in SAS


y = -68. 857 +1. 4546 x1 +9. 3655 x2
Residual
di spoi nc
N 21
Rsq 0. 9167
Adj Rsq 0. 9075
RMSE 11. 007


y = -68. 857 +1. 4546 x1 +9. 3655 x2
Residual
t argt pop
N 21
Rsq 0. 9167
Adj Rsq 0. 9075
RMSE 11. 007


y = -68. 857 +1. 4546 x1 +9. 3655 x2
Residual
Predicted Value
N 21
Rsq 0. 9167
Adj Rsq 0. 9075
RMSE 11. 007


y = -68. 857 +1. 4546 x1 +9. 3655 x2
Residual
Normal Quantile
N 21
Rsq 0. 9167
Adj Rsq 0. 9075
RMSE 11. 007