STOR 455

STATISTICAL METHODS I

Jan Hannig

Multivariate Regression

- $Y = \mathbf{X}\,\beta + \varepsilon$
  - $\mathbf{X}$ is a regression matrix, β is a vector of parameters and ε are independent $N(0,\sigma)$
- Estimated parameters $b = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$
- Predicted responses $\hat{Y} = \mathbf{H}\,Y$, $\mathbf{H} = \mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
- Residuals $e = (\mathbf{I} - \mathbf{H})Y$
- Estimated regression variance $s^2 = e'e/(n-p)$

Inference for b (Section 4.6+4.7)

- $b \sim N(\beta, \sigma^2(X'X)^{-1})$
  - Estimate $Var(b_i)$ by $s^2(b_i) = MSE * (X'X)^{-1}_{i,i}$
- CI: $b_i \pm t^* s(b_i)$
  - proc reg, option /clb
- Significance test for $H_{0i}: \beta_i, = 0$ uses the test statistic $t = b_i/s(b_i)$, df=dfE=n-p, and the P-value computed from the t(n-p) distribution
  - Automatically part of SAS output

Do it in SAS

```
proc reg data=studios;
  model y=x1 x2/clb clm cli  alpha=0.01
                 xpx i covb corrb;
/*clb CI for b;
clm CI for mean;
cli CI for prediction;
xpx gives the X'X matrix
i gives the inverse of the X'X matrix
covb gives the variance covariance matrix of
  b
corrb gives the correlation matrix of b*/
run;
```

ANOVA (Section 4.8)

- Sources of variation are
  - Model or Regression (SSM or SSR)
  - Error or Residual (SSE)
  - Total (SSTO)
- SS and df add
  - SSM + SSE =SSTO
  - dfM + dfE = dfT

Sum of Squares

$$\mathrm{SSM} = \Sigma_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

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

$$\mathrm{SST} = \Sigma_{i=1}^{n}(Y_i - \bar{Y})^2$$

Degree of Freedom and Mean Squares

$$df_M = p-1$$
$$df_E = n-p$$
$$df_T = n-1$$

$$MSM = SSM/df_M$$
$$MSE = SSE/df_E$$
$$MST = SST/df_T$$

ANOVA Table

Source	SS	df	MS	F
Model	SSM	dfM	MSM	MSM/MSE
Error	SSE	dfE	MSE	
Total	SST	dfT	(MST)	

ANOVA F test

- H_0: $\beta_1 = \beta_2 = \ldots \beta_{p-1} = 0$
- H_a: β_k neq 0, for at least one *k=1, … , p-1*
- Under H_0, F ~ F(p-1,n-p)
- Reject H_0 if F is large, use P value
- Studios example: see SAS output

Interpret F-test

- The p-value for the F significance test tells us one of the following:
  – p-value large: there is no evidence to conclude that *any* of our explanatory variables can help us to model the response variable using this kind of model
  – P-value small: one or more of the explanatory variables in our model *is* potentially useful for predicting the response variable in a linear model

Difference between F and T tests

- In SLR the T test for $\beta_1=0$ is the same as the F test (same p-value, the statistics $T^2=F$)
- In MLR this is not true
  - The T tests test whether each $\beta_i=0$ individually
  - The F test tests whether all $\beta_1=\ldots=\beta_{p-1}=0$

Do it in SAS

```
                              The REG Procedure
•                               Model: MODEL1
•                        Dependent Variable: revenue

•                   Number of Observations Read          21
•                   Number of Observations Used          21


•                            Analysis of Variance

•                                     Sum of           Mean
•          Source                DF        Squares         Square    F Value    Pr > F

•          Model                  2          24015          12008      99.10    <.0001
•          Error                 18     2180.92741      121.16263
•          Corrected Total       20          26196


•                    Root MSE             11.00739    R-Square     0.9167
•                    Dependent Mean      181.90476    Adj R-Sq     0.9075
•                    Coeff Var             6.05118

•                             Parameter Estimates

•                        Parameter       Standard
•     Variable       DF       Estimate          Error    t Value    Pr > |t|       95% Confidence Limits

•     Intercept       1      -68.85707       60.01695      -1.15      0.2663     -194.94801       57.23387
•     disp_income     1        9.36550        4.06396       2.30      0.0333        0.82744       17.90356
•     proportion      1        1.45456        0.21178       6.87      <.0001        1.00962        1.89950
```

R^2

- The squared multiple regression correlation (R^2) gives the proportion of variation in the response variable explained by the explanatory variables included in the model
- It is usually expressed as a percent
- It is sometimes called the coefficient of multiple determination

R^2 (continue)

- R^2 = SSM/SSTO, the proportion of variation explained
- $R^2 = 1 - (SSE/SSTO)$, 1 – the proportion of variation not explained
- H_0: $\beta_1 = \beta_2 = \ldots \beta_{p-1} = 0$ is equivalent to H_0: the population R^2 is zero
- $F = [(R^2)/(p-1)] / [(1- R^2)/(n-p)]$

Estimation of $E(Y_h)$

- X_h is now a vector
- $(1, X_{h1}, X_{h2}, \dots, X_{h(p-1)})'$
- We want an point estimate and a confidence interval for the subpopulation mean corresponding to X_h
- SAS option /clm

Theory for E(Y_h)

$$E(Y_h) = X_h'\beta$$

$$E(\widehat{Y}_h) = X_h'b$$

$$\sigma^2(E(\widehat{Y}_h)) = \sigma^2 X_h'(X'X)^{-1}X_h$$

$$s^2(E(\widehat{Y}_h)) = \text{(MSE)}\ X_h'(X'X)^{-1}X_h$$

Prediction of Y_h

- X_h is now a vector
- $(1, X_{h1}, X_{h2}, \ldots , X_{h(p-1)})'$
- We want a prediction for Y_h with an interval that expresses the uncertainty in our prediction
- SAS option /cli

Theory for Y_h

$$Y_h = X_h'\beta + \varepsilon_h$$

$$\widehat{Y}_h = X_h' b$$

$$\sigma^2(\widehat{Y}_h) = \sigma^2 (1 + X_h'(X'X)^{-1}X_h)$$

$$s^2(\widehat{Y}_h) = \text{(MSE)} \ (1 + X_h'(X'X)^{-1}X_h)$$

Do it in SAS

```
* estimate mean response,
  predict new obs;
data stoPred;
  input x1 x2 y;
cards;
  68.5  16.7  174.4
  45.2  16.8  164.4
  91.3  18.2  244.2
  …
  82.7  19.1  224.1
  52.3  16.0  166.5
  65.4  17.6    .
  53.1  17.7    .
  ;
run;
```

```
proc reg data = stoPred;
  model y = x1 x2 / clm
  cli;
  output
  OutputStatistics=temp;
run;
quit;

proc print data = temp;
  where Observation >=
  22;
run;
```

Do it in SAS

Obs	Model	Dependent	Observation	DepVar	Predicted Value	StdErr Mean Predict
22	MODEL1	y	22	.	191.1039	2.7668
23	MODEL1	y	23	.	174.1494	4.5986

Obs	Lower CLMean	Upper CLMean	LowerCL	UpperCL	Residual
22	185.2911	196.9168	167.2589	214.9490	.
23	164.4881	183.8107	149.0867	199.2121	.

SAT Example

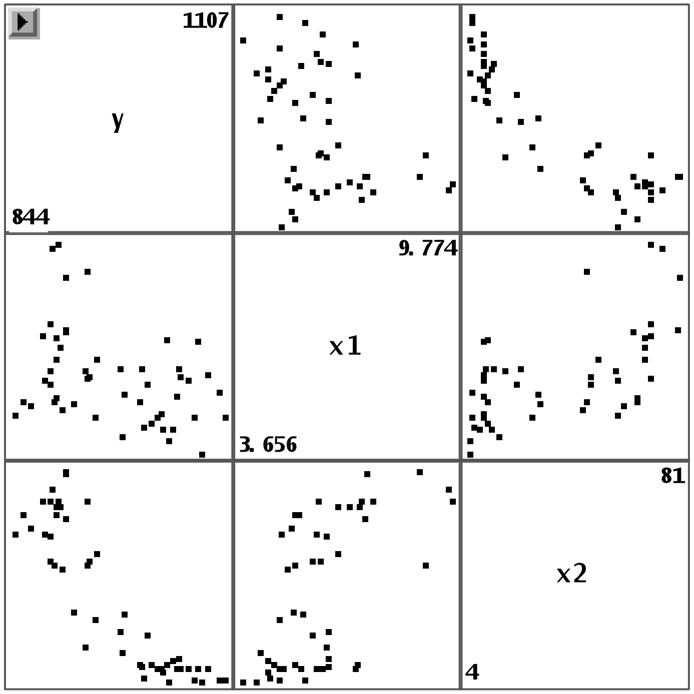
- School spending and academic performance are often thought to be positively correlated
- Some argue that they are statistically unrelated
- Data collected from 50 states in 1995 to address this question

Data

- Y: Average total score on the SAT for each state in US, 1994-95
- X_1: Current expenditure per pupil 1994-95 (in thousands of dollars)
- X_2: Percentage of all eligible students taking the SAT, 1994-95
- Ch6.sat.sas : SAS code for analyzing this data

Do it in SAS

```
*Inputting the SAT data;
data sat;
  infile 'satexp.dat';
  input x1 x2 y;
  x1x2=x1*x2;
  label x1='Expenditure'
        x2='Percent'
    y='SATscore';
run;
* Making scatter plot
  using macro;
%include "C:\Macro
  \scatter.sas";
%scatter(data = sat, var
  = y x1 x2);
```



Do it in SAS

```
* Check correlation;
proc corr data = sat;
run;

*SLR with each variable,
  note sign of b1;
symbol1 v=dot h=.8
  c=blue;
proc reg data = sat;
  model y=x1;
run;
proc reg data = sat;
  model y=x2;
run;
```

```
* multivariate reg.;
proc reg data = sat;
* ask for covariance
   matrix of b and CI;
  model y = x1 x2/covb
   clb;
  plot rstudent. * (x1 x2
   p.) ;
  plot r. * nqq.;
run;
```

Do it in SAS

The CORR Procedure

4 Variables: x1 x2 y x1x2

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
x1	50	5.90526	1.36281	295.26300	3.65600	9.77400	Expenditure
x2	50	35.24000	26.76242	1762	4.00000	81.00000	Percent
y	50	965.92000	74.82056	48296	844.00000	1107	SATscore
x1x2	50	229.28338	206.17969	11464	14.62400	714.17700	

Pearson Correlation Coefficients, N = 50
Prob > |r| under H0: Rho=0

	x1	x2	y	x1x2
x1	1.00000	0.59263	-0.38054	0.77503
Expenditure		<.0001	0.0064	<.0001
x2	0.59263	1.00000	-0.88712	0.95115
Percent	<.0001		<.0001	<.0001
y	-0.38054	-0.88712	1.00000	-0.77923
SATscore	0.0064	<.0001		<.0001
x1x2	0.77503	0.95115	-0.77923	1.00000
	<.0001	<.0001	<.0001	

Do it in SAS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	39722	39722	8.13	0.0064
Error	48	234586	4887.20043		
Corrected Total	49	274308			

Root MSE	69.90851	R-Square	0.1448
Dependent Mean	965.92000	Adj R-Sq	0.1270
Coeff Var	7.23751		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1089.29372	44.38995	24.54	<.0001
x1	Expenditure	1	-20.89217	7.32821	-2.85	0.0064

Do it in SAS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	215875	215875	177.33	<.0001
Error	48	58433	1217.35723		
Corrected Total	49	274308			

Root MSE	34.89065	R-Square	0.7870
Dependent Mean	965.92000	Adj R-Sq	0.7825
Coeff Var	3.61217		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1053.32036	8.21121	128.28	<.0001
x2	Percent	1	-2.48015	0.18625	-13.32	<.0001

Do it in SAS

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	224788	112394	106.67	<.0001
Error	47	49520	1053.61828		
Corrected Total	49	274308			

Root MSE	32.45949	R-Square	0.8195
Dependent Mean	965.92000	Adj R-Sq	0.8118
Coeff Var	3.36047		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	993.83166	21.83323	45.52	<.0001
x1	Expenditure	1	12.28652	4.22432	2.91	0.0055
x2	Percent	1	-2.85093	0.21511	-13.25	<.0001

Do it in SAS

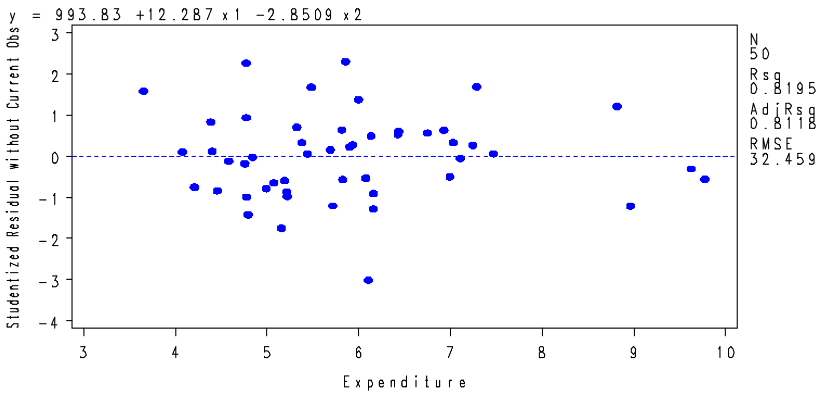
Parameter Estimates

Variable	Label	DF	95% Confidence Limits	
Intercept	Intercept	1	949.90886	1037.75446
x1	Expenditure	1	3.78829	20.78475
x2	Percent	1	-3.28368	-2.41818

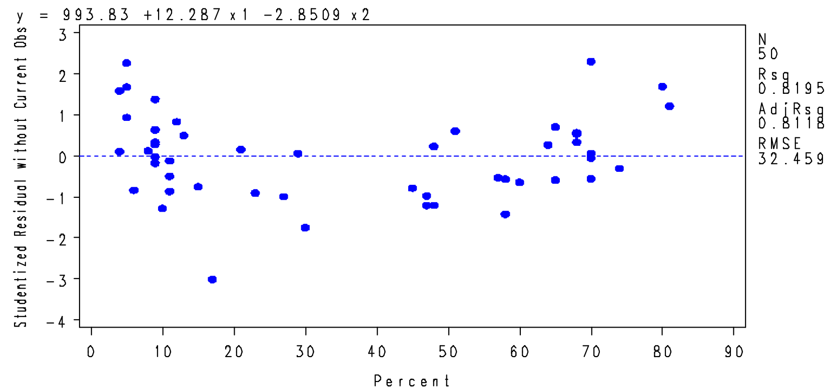
Covariance of Estimates

Variable	Label	Intercept	x1	x2
Intercept	Intercept	476.69008596	-86.40093833	1.5494405413
x1	Expenditure	-86.40093833	17.844845216	-0.538521916
x2	Percent	1.5494405413	-0.538521916	0.0462733084

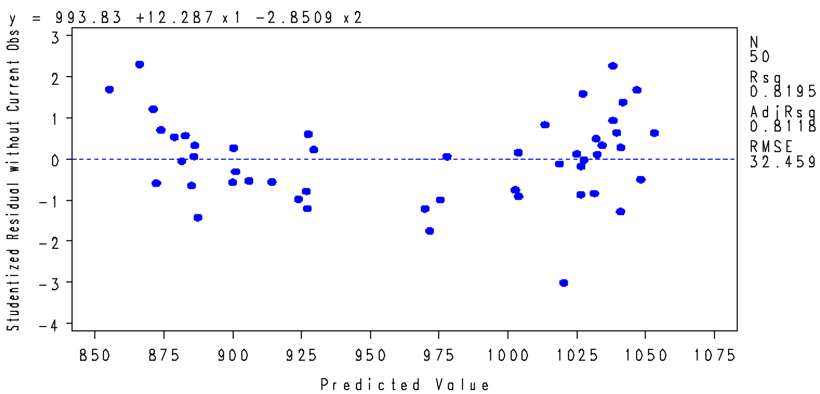
Do it in SAS



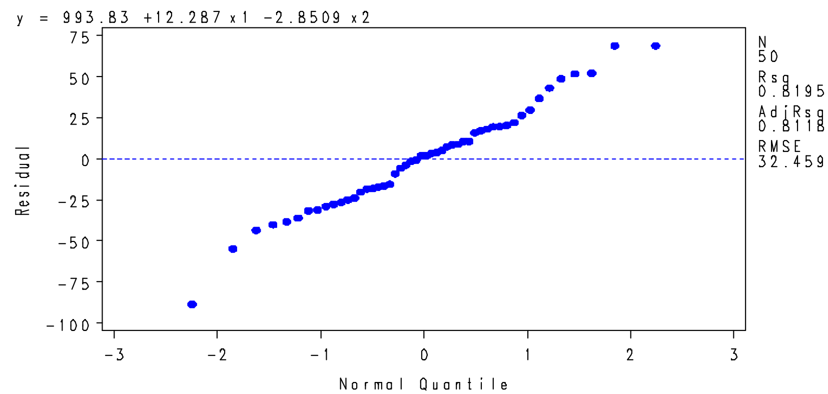
y = 993.83 +12.287 x1 -2.8509 x2
Studentized Residual without Current Obs
Expenditure
N 50
Rsq 0.8195
AdjRsq 0.8118
RMSE 32.459



y = 993.83 +12.287 x1 -2.8509 x2
Studentized Residual without Current Obs
Percent
N 50
Rsq 0.8195
AdjRsq 0.8118
RMSE 32.459



y = 993.83 +12.287 x1 -2.8509 x2
Studentized Residual without Current Obs
Predicted Value
N 50
Rsq 0.8195
AdjRsq 0.8118
RMSE 32.459



y = 993.83 +12.287 x1 -2.8509 x2
Residual
Normal Quantile
N 50
Rsq 0.8195
AdjRsq 0.8118
RMSE 32.459

Notes:

- The relationship between a response variable and an explanatory variable depends on what other explanatory variables are in the model
- Regression coefficients, standard errors and the results of significance tests can change dramatically when different explanatory variables are included in the model

SENIC Data

- Info about 113 hospitals between 1975 and 1976
- Variables: id, length of stay, age, infection risk, routine culturing ratio, routine chest X-ray ratio, number of beds, medical school affiliation, region (1=NE, 2=NC, 3=S, 4=W), average daily census, number of nurses, available facilities
- File AppendixC01.txt in the extra data sets

SENIC Data

- Compare different models using R^2
- Same model for different subset