CHAPTER 11

Validation of Regression Models

Introduction

- Model Adequacy Checking
  - Residual analysis, lack of fit testing, determining influential observations
  - Checks the fit of the model to the available data
- Model Validation
  - Determining if the model will behave or function as it was intended in the operating environment

Validation Techniques

- **Analysis of model coefficients and predicted values**
  - Check for “inappropriate” signs on the coefficients
  - Check for unusual magnitudes on the coefficients
  - Check for stability in the coefficient estimates
  - Check the predicted values

Example 11.1 The Hald Cement Data

Consider the Hald cement data introduced in Example 9.1. We used all possible regressions to develop two possible models for these data, model 1,

$$\hat{y} = 52.58 + 1.468x_1 + 0.662x_2$$

and model 2,

$$\hat{y} = 71.65 + 1.452x_1 + 0.416x_2 - 0.237x_4$$

- Coefficients of x_1 very similar, coefficients of x_2 and the intercept moderately different
- Examine differences in predicted values

Example 11.1 The Hald Cement Data

TABLE 11.1 Prediction Values for Two Models for Hald Cement Data

y	x_1	x_2	x_3	x_4	Model 1	Model 2
78.5	7	26	6	60	80.074	78.438
74.3	1	29	15	52	73.251	72.867
104.3	11	56	8	20	105.815	106.191
87.6	11	31	8	47	89.258	89.402
95.9	7	52	6	33	97.293	95.644
109.2	11	55	9	22	105.152	105.302
102.7	3	71	17	6	104.002	104.129
72.5	1	31	22	44	74.575	75.592
93.1	2	54	18	22	91.275	91.818
115.9	21	47	4	26	114.538	115.546
83.8	1	40	23	34	80.536	81.702
113.3	11	66	9	12	112.437	112.244
109.4	10	68	8	12	112.293	111.625

Validation Techniques

- **Collection of new data**
  - Usually 15-20 new observations are adequate
- Example
  - *The Delivery Time Data*

$$\hat{y} = 2.3412 + 1.6159x_1 + 0.0144x_2$$

Example 11.2 The Delivery Time Data

New data:

TABLE 11.2 Prediction Data Set for the Delivery Time Example

	(1)	(2)	(3)	(4)	(5)	(6)
					Least-Squares Fit	
Observation	City	Cases, x_1	Distance, x_2	Observed Time, y	$\hat{y}$	$y - \hat{y}$
26	San Diego	22	905	51.00	50.9230	0.0770
27	San Diego	7	520	16.80	21.1405	–4.3405
28	Boston	15	290	26.16	30.7557	–4.5957
29	Boston	5	500	19.90	17.6207	2.2793
30	Boston	6	1000	24.00	26.4366	–2.4366
31	Boston	6	225	18.55	15.2766	3.2734
32	Boston	10	775	31.93	29.6602	2.2698
33	Boston	4	212	16.95	11.8576	5.0924
34	Austin	1	144	7.00	6.0307	0.9693
35	Austin	3	126	14.00	9.0033	4.9967
36	Austin	12	655	37.03	31.1640	5.8660
37	Louisville	10	420	18.62	24.5482	–5.9282
38	Louisville	7	150	16.10	15.8125	0.2875
39	Louisville	8	360	24.38	20.4524	3.9276
40	Louisville	32	1530	64.75	76.0820	–11.3320

Average squared prediction error →

$$\frac{\sum_{i=26}^{40} (y_i - \hat{y}_i)^2}{15} = \frac{332.2809}{15} = 22.1521$$

Example 11.2 The Delivery Time Data

- Compare the residual mean square

$$MS_{\text{Res}} = 10.6239$$

to the average squared prediction error

$$\frac{\sum_{i=26}^{40} (y_i - \hat{y}_i)^2}{15} = \frac{332.2809}{15} = 22.1521$$

Example 11.2 The Delivery Time Data

It is also instructive to compare R^2 from the least-squares fit (0.9596) to the percentage of variability in the new data explained by the model, say

$$R^2_{\text{Prediction}} = 1 - \frac{\sum_{i=26}^{40} (y_i - \hat{y}_i)^2}{\sum_{i=26}^{40} (y_i - \bar{y})^2} = 1 - \frac{332.2809}{3206.2338} = 0.8964$$

Once again, we see that the least-squares model does not predict new observations as well as it fits the original data. However, the "loss" in R^2 for prediction is slight.

Collecting new data has indicated that the least-squares fit for the delivery time data results in a reasonably good prediction equation. The interpolation performance of the model is likely to be better than when the model is used for extrapolation.

- From original model:

$$R^2_{\text{Prediction}} = 1 - \frac{\text{PRESS}}{SS_T} = 1 - \frac{457.4000}{5784.5426} = 0.9209$$

Validation Techniques

- **Data Splitting (Cross Validation)**
  - Divide the data into two parts: estimation data and prediction data
  - The PRESS statistic is an estimate of performance based on data splitting:

$$PRESS = \sum [y_i - y_{(i)}]^2$$

  - PRESS can be used to compute an R^2 type statistic for prediction:

$$R^2_{\text{Prediction}} = 1 - \frac{PRESS}{SS_T}$$

Validation Techniques

- **Data Splitting**
  - If the time sequence is known, data splitting can be done by time order
  - Given other characteristics of the data, data splitting based on group (operator, machine, location, etc.)
  - Random selection

Example 11.3 The Delivery Time Data

Observation, i	Cases, x_1	Distance, x_2	Delivery Time, y	Estimation (E) or Prediction (P) Data Set
1	7	560	16.68	P
2	3	220	11.50	P
3	3	340	12.03	P
4	4	80	14.88	E
5	6	150	13.75	E
6	7	330	18.11	E
7	2	110	8.00	E
8	7	210	17.83	E
9	30	1460	79.24	E
10	5	605	21.50	E

A portion of Table 11.3 showing prediction and estimation data

Example 11.3 The Delivery Time Data

TABLE 11.5 Summary of Least-Squares Fit to the Delivery Time Data

A. Analysis Using Estimation Data				B. Analysis Using All Data			
Variable	Coefficient Estimate	Standard Error	t_0	Variable	Coefficient Estimate	Standard Error	t_0
Intercept	2.4123	1.4165	1.70	Intercept	3.9840	0.9861	4.04
x_1	1.6392	0.1769	9.27	x_1	1.4877	0.1376	10.81
x_2	0.0136	0.0036	3.78	x_2	0.0134	0.0028	4.72
$MS_{Res} = 13.9145$, $R^2 = 0.952$				$MS_{Res} = 13.6841$, $R^2 = 0.944$			

Example 11.3 The Delivery Time Data

TABLE 11.6 Prediction Performance for the Model Developed from the Estimation Data

	(1)	(2)	(3)
		Least-Squares Fit	
Observation, i	Observed, y_i	Predicted, $\hat{y}_i$	Prediction Error, $e_i = y_i - \hat{y}_i$
1	16.68	21.4976	−4.8176
2	11.50	10.3199	1.1801
3	12.03	11.9508	0.0792
11	40.33	37.9901	2.3399
12	21.00	21.7264	−0.7264
14	19.75	18.5265	1.2235
16	29.00	29.3509	−0.3509
17	15.35	14.9657	0.3843
19	9.50	7.8192	1.6808
26	51.00	50.7746	0.2254
28	26.16	30.9417	−4.7817
32	31.93	29.3373	2.5927
33	16.95	11.8504	5.0996
34	7.00	6.0086	0.9914
35	14.00	9.0424	4.9576
36	37.03	30.9848	6.0452
37	18.62	24.5125	−5.8925
38	16.10	15.9254	0.1746
39	24.38	20.4187	3.9613
40	64.75	75.6609	−10.9109

Example 11.3 The Delivery Time Data

- Estimated Data: $R^2 = 0.952$
- Prediction Data: $R^2_{\text{Prediction}} = 0.922$

Using SAS and R for Validation

- R

```
tableb1 <- read.delim("e:\\data-table-B1.txt",h=T)
temp <- tableb1[sample(1:nrow(tableb1), 10,
replace=FALSE),]
temp
```