Model Validation

Johns Hopkins Engineering

625.461 Statistical Models and Regression

Module 10 – Lecture 10C

Purpose of Model Validation

Model validation is directed toward determining if the model will function successfully in its intended operating environment.

In particular, there is no assurance that the equation that provides the best fit to existing data will be a successful predictor.

Validation Techniques

1. Analysis of model coefficients and predicted values

The coefficients in the final regression model should be studied to determine if they are stable and if their signs and magnitudes are reasonable.

Variance inflation factors (VIF) and other multicollinearity diagnostics are an important guide to model validity.

Validation Techniques

If VIF exceeds 5 or 10, that particular coefficient is poorly estimated or unstable because of near-linear dependencies among the regressors.

Predicted values inside and on the boundary of the regressors space provide a measure of the model's interpolation performance.

Predicted values outside this region are a measure of the model's extrapolation performance.

Validation Techniques

2. Collecting fresh data – confirmation runs

Collect new data (called confirmation runs) and directly compare the model predictions against them. In general, 15 – 20 new data are desirable to give a reliable assessment of the model's prediction performance.

The Delivery Time Data (Ex. 11.2, page 376)

The LS fit based on 25 observations that came from four cities: Austin, San Diego, Boston, Minneapolis gives

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

The Delivery Time Data (Ex. 11.2, page 376)

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

The Delivery Time Data (Ex. 11.2, page 376)

$$SS_{\mathrm{T}} = \sum_{i=26}^{40} (y_i - \bar{y})^2 = 3206.2338$$

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

$$R^2_{\text{prediction}} = 1 - \frac{332.2809}{3206.2338} = 0.8964$$

Validation Techniques

3. Data Splitting (Cross Validation)

Split the available data into two parts – estimation data and prediction data, or training sample and test sample; for example,

$$\text{PRESS} = \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

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

Ex. 11.2. $R^2_{\text{prediction}} = 1 - \frac{457.4000}{5784.5426} = 0.9206$

JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING