Research Article

Quality and Reliability Engineering International

(wileyonlinelibrary.com) DOI: 10.1002/qre.1715

Published online 27 August 2014 in Wiley Online Library

Statistical Tests to Validate Predictive Models

M. S. Hamada*† and J. I. Abes

This article presents validation tests to check a predictive model over time. These validation tests are like individuals, $\bar{X}$ and S control charts. The individuals validation test is used when there are no replicates at the checked time points. The $\bar{X}$ and S validation tests can be used when there are replicates. Their power is evaluated under various scenarios via a simulation study. Based on these results, the $\bar{X}$ and S validation tests are recommended when there are replicate measurements.

Keywords: individuals control chart; regression control chart; power; $\bar{X}$ and S control charts

1. Introduction

One activity in surveilling a population is the monitoring of trends over time, where units are sampled periodically and their measurements are checked against a predictive model. The predictive model may be solely empirical based on a fit of previous measurements or a synthesis of available scientific knowledge (e.g., a science model), previous measurements, and other data such as from relevant scientific experiments. The purpose of ongoing sampling is to validate the predictive model, where discrepancies indicate inadequacies in the predictive model that need to be explored.

Bierbaum *et al.*[1] consider the power to detect specified discrepancies (2σ difference from the predicted mean, where σ is the population standard deviation, as well as a doubling of the population standard deviation) based on the amount of data that will be collected in the next 4 years. They assume that the true predicted mean and population standard deviation are known exactly so that the validation tests consist of chi-squared tests to detect such discrepancies. They used this simplification because their goal was to provide general guidance on the impact of the amount of data collected. However, in actually checking a predictive model with newly collected data, uncertainty in the predictive model needs to be accounted for. This article presents such validation tests when there are no replicates at a time point and also when there are replicates.

For illustrative purposes, we will use the simple linear regression model to describe the population as it changes over time. Consequently, the data are generated by a simple linear regression model, and a simple linear regression model is fit to the data. The parameters of the simple linear regression model are unknown and have to be estimated from the data. We also assume that all units in the population have the same age. See Figure 1, which displays stress measurements of polymer samples for a given amount of strain over time in years. A downward trend is displayed, which we model by a simple linear regression model. The validation tests use data collected in the future to use to check for discrepancies from this predictive model.

An outline of this article is as follows. We review the simple linear regression model in Section 2. Then, in Section 3, we introduce an alternative validation test when there are no replicates at a time point so that it tests the individual values, that is, an 'individuals' validation test. In Section 4, we illustrate the individuals validation test with the stress data displayed in Figure 1. In Section 5, we present a simulation study that shows the impact of various factors on the power of the individuals validation test; these factors include the amount of data and the spread in its times, the true slope, the true population standard deviation, and the number of validation samples and how they are distributed over a 4-year period. In Section 6, we present other alternative validation tests when there are replicates, that is, when more than one unit is sampled and measured at a time point. We also present a simulation study of the power of these alternative tests. We conclude with a discussion in Section 7.

2. Simple linear regression model

Now, we consider the situation where the population is aging as depicted in Figure 2. Suppose that the property Y is an impurity level whose mean increases linearly with age. At age t, the population of property Y values has a normal distribution $Normal(\beta_0 + \beta_1 t, \sigma^2)$ with mean $\beta_0 + \beta_1 t$ and standard deviation σ. That is,

$$Y_t \sim Normal(\beta_0 + \beta_1 t, \sigma^2). \quad (1)$$

Los Alamos National Laboratory, Los Alamos, NM, USA
**Correspondence to: M. S. Hamada, Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos, NM 87545, USA.*
†E-mail: hamada@lanl.gov

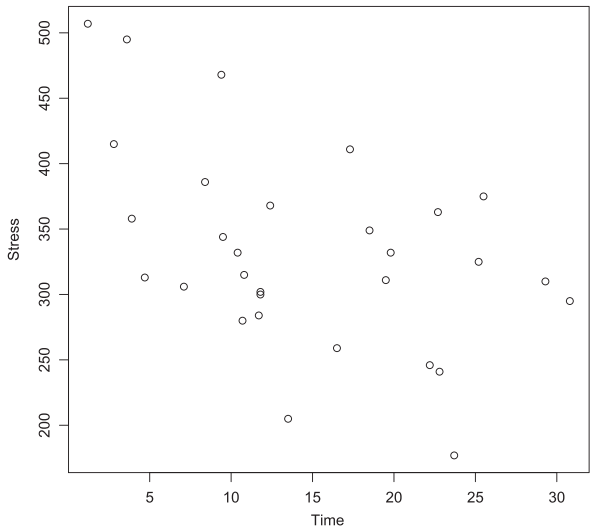

Figure 1. Stress measurements sampled over time in years

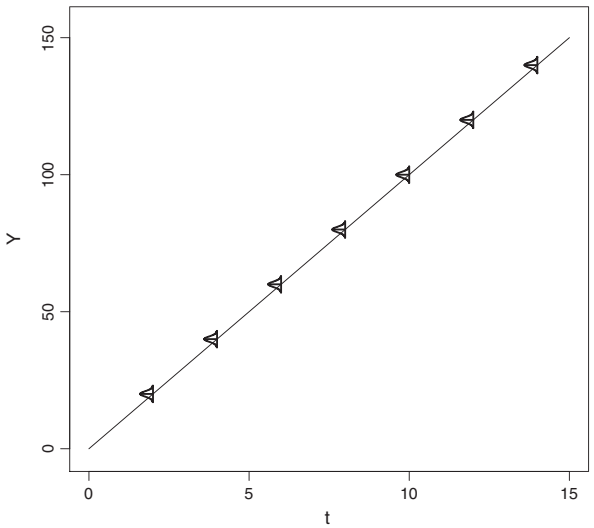

Figure 2. A trend of the population of properties Y such as impurities with standard deviation $\sigma = 1$. The solid trend line or mean is given by $\beta_0 + \beta_1 t = 0 + 10t$

Figure 2 shows a population with $\beta_0 = 0, \beta_1 = 10$, and $\sigma = 1$.

Using the data consisting of n pairs $(Y_1, t_1), \ldots, (Y_n, t_n)$, the simple linear regression model is fit by least squares producing estimates for (β_0, β_1) denoted by $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ and an estimate for σ^2; $\hat{\sigma}^2$ is the residual sum of squares divided by $n-2$.[2] Consequently, the prediction for Y_t is its estimated mean $\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 t$. $\hat{Y}_t$ is normally distributed with mean $\beta_0 + \beta_1 t$ and standard deviation $\sigma\sqrt{H_t}$, where

$$H_t = \mathbf{x}_t(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_t',$$

$\mathbf{x}_t = (1\ t)$ and $\mathbf{X} = \begin{pmatrix} 1 & t_1 \\ \ldots & \ldots \\ 1 & t_n \end{pmatrix}$. That is,

$$\hat{Y}_t \sim Normal(\beta_0 + \beta_1 t, \sigma^2 H(t)). \tag{2}$$

3. An individuals validation test that accounts for prediction error

From Equations (1) and (2), we have

$$Y_t - \hat{Y}_t \sim Normal(0, \sigma^2(1 + H(t)). \tag{3}$$

A $(1-\alpha)\times 100\%$ prediction interval for Y_t based on Equation (3) is

$$\hat{Y}_t \pm t_{1-\alpha/2,n-2}\hat{\sigma}\sqrt{1+H(t)}, \tag{4}$$

where $t_{1-\alpha/2,n-2}$ is the $1-\alpha/2$ quantile of a t distribution with $n-2$ degrees of freedom.

For m validation data points, choose $\alpha = 0.05/m$, the Bonferroni correction.[3] Then, the validation test rejects the null hypothesis (that the current model is valid) if any of the validation data points fall outside of their respective prediction intervals given in Equation (4). If the tests for the individual validation data points were independent, the Bonferroni correction would ensure an overall test type I error of 0.05. However, the tests are not independent because all the prediction intervals depend on $\hat{\beta}_0$, $\hat{\beta}_1$, and s. The overall test type I error is conservative being less than 0.05.

This validation test is like an individuals control chart, where individual values are plotted against their respective control limits. Consequently, we refer to this test as an individuals validation test. Note that the confidence intervals in Equation (4) are equivalent to the control limits for the regression control chart,[4] except that the regression control chart plots the residuals $Y_t - \hat{Y}_t$. However, here, their use is intended for the next 4-year period. After 4 years, assuming that no discrepancies had been detected, the data collected up to then would be used to refit the simple linear model and obtain revised prediction intervals for the next 4-year period.

4. Illustration of individuals validation test

Consider the stress data in Figure 1 up to 25 years with four data after 25 years. The simultaneous 95% prediction intervals in Equation (4) for the latter data (at four time points) based on fitting the earlier data are displayed in Figure 3 as dashed lines. Note that the four data measured after 25 years are within their prediction intervals so that none of these are discrepancies from the predicted model.

5. A simulation study to evaluate individuals validation test power

There are a number of factors to consider in a simulation study of power. For the data to fit the model, these factors include the times to collect the data, how much data, and parameter values, especially for β_1 and σ. For the validation data, these factors include the times to collect the data and how much to collect. Without loss of generality, we set $\beta_0 = 0$ and consider values (1, 10) for β_1 and values (0.1, 1) for σ. We found that the differences in the results obtained by varying β_1 and σ are not significant, that is, within the simulation error based on 10,000 simulated fitting and validation data sets. Consequently, we present the $\beta_1 = 10$ and $\sigma = 1$ results in the succeeding text. For the fitting data, we let the t_i be $1, 2, \ldots, 10$ and consider 1, 4, 7, and 11 replicates at each t_i. For example, four replicates mean that we measure four units at time 1, four units at time 2, and so on.

For the validation data, we consider collecting them within 4 years of the last fitting datum, so years 11–14. We consider five cases; cases 1–4 are all at year 11 to all at year 14. Case 5 is an equal number of units at each of the years 11–14. We consider 1–3 validation replicates of four so that case 1 with three replicates mean that 12 units are measured at year 11.

Table I displays the individuals validation test type I errors of which a number is smaller than 0.05. Table II displays power for a 2σ shift in the mean; in the validation period, the observations are simulated from $Normal(\beta_0 + \beta_1 t + 2\sigma, \sigma^2)$. Note that after four

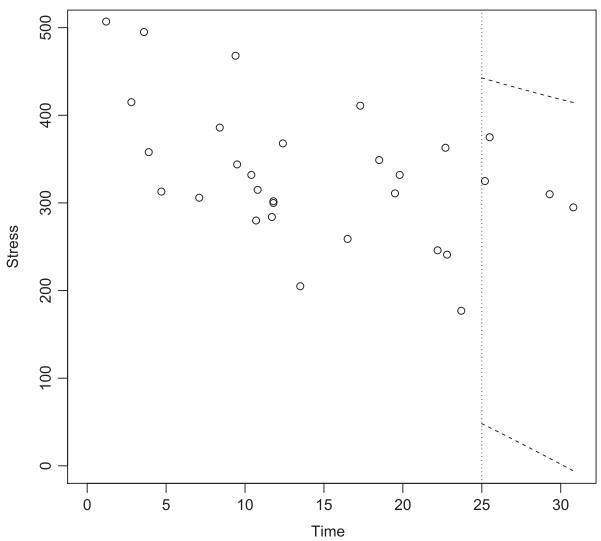

Figure 3. Stress measurements sampled over time in years with simultaneous 95% prediction intervals for the last four time points

Table I. Individuals validation test type I error

Validation replicates	Fitting replicates	Validation case 1	2	3	4	5
1	1	0.04	0.04	0.04	0.04	0.04
	4	0.04	0.04	0.05	0.05	0.05
	7	0.05	0.05	0.05	0.05	0.05
	11	0.05	0.05	0.05	0.05	0.05
2	1	0.04	0.04	0.03	0.03	0.04
	4	0.05	0.04	0.05	0.05	0.05
	7	0.05	0.05	0.05	0.05	0.05
	11	0.05	0.05	0.05	0.05	0.05
3	1	0.03	0.03	0.03	0.03	0.03
	4	0.05	0.05	0.04	0.05	0.04
	7	0.05	0.05	0.04	0.05	0.05
	11	0.05	0.05	0.05	0.05	0.05

Table II. Individuals validation test for mean power

Validation replicates	Fitting replicates	Validation case 1	2	3	4	5
1	1	0.30	0.28	0.24	0.21	0.26
	4	0.62	0.59	0.58	0.55	0.58
	6	0.69	0.67	0.65	0.63	0.66
	7	0.71	0.70	0.70	0.68	0.70
2	1	0.27	0.23	0.21	0.18	0.23
	4	0.68	0.66	0.63	0.60	0.63
	7	0.77	0.74	0.72	0.71	0.74
	11	0.81	0.79	0.79	0.77	0.79
3	1	0.24	0.21	0.19	0.16	0.21
	4	0.70	0.67	0.64	0.61	0.66
	7	0.80	0.78	0.76	0.74	0.77
	11	0.85	0.83	0.82	0.80	0.83

data fitting replicates, more replicates have diminishing returns. More validation replicates improve power very little after accounting for more individuals tests with more replicates through the Bonferroni correction. Cases 1–4 with all validation data at one time are better than case 5 (all data spread out) with case 1 with the least amount of extrapolation from the fitting data being the best. Similar results are seen in Table III that displays power for a doubling of the standard deviation; in the validation period, the observations are simulated from $Normal(\beta_0 + \beta_1 t, (2\sigma)^2)$. Note that from a surveillance standpoint, case 5 is preferred because units are sampled each year; Tables II and III show that there is a little loss in power over case 1.

6. '$\bar{X}/S$' validation tests and their power

When we saw that more validation replicates hardly impacted power, we realized that we were not taking advantage of the replicates. That is, we could calculate sample means and standard deviations and use them in validation tests. These tests are like $\bar{X}$ and S control charts so we will refer to them as '$\bar{X}/S$' validation tests.

If there is only one replicate at each validation time t, then the individuals validation test needs to be used. Suppose that there are m_t replicates at validation time t, and $Y_{t,1}, \ldots, Y_{t,m_t}$ are the corresponding measurements, then $\bar{Y}_t$ is the sample mean of these validation replicates. Like individual measurements, we obtain

$$\bar{Y}_t - \hat{Y}_t \sim Normal(0, \sigma^2(1/m_t + H(t)),$$

that yields a $(1-\alpha) \times 100\%$ prediction interval for $\bar{Y}_t$

$$\hat{Y}_t \pm t_{1-\alpha/2,n-2}\hat{\sigma}\sqrt{1/m_t + H(t)},$$

Table III. Individuals validation test for standard deviation power

Validation replicates	Fitting replicates	Validation case 1	2	3	4	5
1	1	0.31	0.30	0.26	0.24	0.28
	4	0.54	0.54	0.51	0.50	0.52
	7	0.57	0.56	0.56	0.55	0.55
	11	0.59	0.58	0.58	0.57	0.58
2	1	0.35	0.32	0.28	0.26	0.30
	4	0.67	0.66	0.65	0.63	0.65
	7	0.71	0.72	0.71	0.70	0.71
	11	0.74	0.75	0.73	0.73	0.73
3	1	0.36	0.33	0.30	0.26	0.31
	4	0.75	0.74	0.73	0.70	0.73
	7	0.80	0.80	0.78	0.78	0.79
	11	0.82	0.83	0.82	0.82	0.82

Table IV. $\bar{X}$ validation test for mean type I error

Validation replicates	Fitting replicates	Validation case 1	2	3	4	5
1	1	0.05	0.05	0.05	0.05	
	4	0.05	0.05	0.05	0.05	
	7	0.05	0.05	0.05	0.05	
	11	0.05	0.05	0.05	0.05	
2	1	0.05	0.05	0.05	0.05	0.04
	4	0.05	0.05	0.05	0.05	0.05
	7	0.05	0.05	0.05	0.05	0.05
	11	0.05	0.05	0.05	0.05	0.05
3	1	0.05	0.05	0.05	0.05	0.03
	4	0.05	0.05	0.05	0.05	0.05
	7	0.05	0.05	0.05	0.05	0.05
	11	0.05	0.05	0.05	0.05	0.05

Table V. S validation test for standard deviation type I error

Validation replicates	Fitting replicates	Validation case 1	2	3	4	5
1	1	0.05	0.05	0.05	0.05	
	4	0.05	0.05	0.05	0.05	
	7	0.05	0.05	0.05	0.05	
	11	0.05	0.05	0.05	0.05	
2	1	0.05	0.05	0.05	0.05	0.05
	4	0.05	0.05	0.05	0.05	0.05
	7	0.05	0.05	0.05	0.05	0.05
	11	0.05	0.05	0.05	0.05	0.05
3	1	0.05	0.05	0.05	0.05	0.05
	4	0.05	0.05	0.05	0.05	0.05
	7	0.05	0.05	0.05	0.05	0.05
	11	0.05	0.05	0.05	0.05	0.05

where $t_{1-\alpha/2,n-2}$ is the $1-\alpha/2$ quantile of a t distribution with $n-2$ degrees of freedom. As before, we use the Bonferroni correction so that α equals 0.05 divided by the number of unique validation times. Here, we consider the same cases 1–5 as in Section 5. For cases 1–4, $\alpha = 0.05$, but for case 5, $\alpha = 0.05/4$. The $\bar{X}$ validation test to detect a change in the mean rejects the null hypothesis if any $\bar{Y}_t$ lies outside its prediction interval.

To test for a change in the standard deviation, take the replicate measurements at time t and compute the sum of squares $SS_t = \sum_{j=1}^{m_t} \left(Y_{tj} - \bar{Y}_t\right)^2$ with associated degrees of freedom $m_t - 1$. If there are several unique validation times, then add the sums of squares and the degrees of freedom and denote them by $SS = \sum_t SS_t$ and $\nu_{SS} = \sum(m_t - 1)$. The S validation test uses the ratio $F = \frac{SS/\nu_{SS}}{s^2}$, where s is the estimate of σ from the simple linear regression model fitting. Under the null hypothesis, F has an F distribution with

Table VI. $\bar{X}$ validation test for mean power

Validation replicates	Fitting replicates	Validation case 1	2	3	4	5
1	1	0.55	0.47	0.40	0.36	
	4	0.90	0.86	0.84	0.79	
	7	0.94	0.93	0.91	0.88	
	11	0.96	0.95	0.94	0.93	
2	1	0.62	0.54	0.46	0.39	0.32
	4	0.98	0.96	0.93	0.90	0.81
	7	0.99	0.99	0.98	0.97	0.89
	11	1.00	1.00	0.99	0.99	0.93
3	1	0.66	0.56	0.47	0.41	0.35
	4	0.99	0.98	0.96	0.93	0.89
	7	1.00	1.00	0.99	0.99	0.96
	11	1.00	1.00	1.00	1.00	0.98

Table VII. S validation test for standard deviation power

Validation replicates	Fitting replicates	Validation case 1	2	3	4	5
1	1	0.44	0.43	0.43	0.44	
	4	0.55	0.55	0.57	0.55	
	7	0.57	0.57	0.56	0.57	
	11	0.57	0.58	0.57	0.58	
2	1	0.57	0.57	0.57	0.56	0.47
	4	0.77	0.78	0.77	0.78	0.63
	7	0.81	0.81	0.80	0.80	0.64
	11	0.81	0.81	0.82	0.82	0.66
3	1	0.63	0.61	0.62	0.63	0.57
	4	0.89	0.88	0.89	0.88	0.81
	7	0.91	0.90	0.91	0.91	0.83
	11	0.92	0.92	0.92	0.92	0.85

Table VIII. Difference between $\bar{X}$ and individuals validation test for mean power

Validation replicates	Fitting replicates	Validation case 1	2	3	4	5
1	1	0.25	0.19	0.16	0.15	
	4	0.28	0.27	0.26	0.24	
	7	0.25	0.26	0.26	0.25	
	11	0.25	0.25	0.24	0.25	
2	1	0.35	0.31	0.25	0.21	0.09
	4	0.30	0.30	0.30	0.30	0.18
	7	0.22	0.25	0.26	0.26	0.15
	11	0.19	0.21	0.20	0.22	0.14
3	1	0.42	0.35	0.28	0.25	0.14
	4	0.29	0.31	0.32	0.32	0.23
	7	0.20	0.22	0.23	0.25	0.19
	11	0.15	0.17	0.18	0.20	0.15

Table IX. Difference between S and individuals validation test for standard deviation power

Validation replicates	Fitting replicates	Validation case				
		1	2	3	4	5
1	1	0.13	0.13	0.17	0.20	
	4	0.01	0.01	0.06	0.05	
	7	0.00	0.01	0.00	0.02	
	11	−0.02	0.00	−0.01	0.01	
2	1	0.22	0.25	0.29	0.30	0.17
	4	0.10	0.12	0.12	0.15	−0.02
	7	0.10	0.09	0.09	0.10	−0.07
	11	0.07	0.06	0.09	0.09	−0.07
3	1	0.27	0.28	0.32	0.37	0.26
	4	0.14	0.14	0.16	0.18	0.08
	7	0.11	0.10	0.13	0.13	0.04
	11	0.10	0.09	0.10	0.10	0.03

ν_{SS} and ν degrees of freedom. Consequently, the S validation test rejects the null hypothesis if $F > F_{0.95,\nu_{SS},\nu}$, 0.95 quantile of an F distribution with ν_{SS} and ν degrees of freedom.

As before, we can simulate 10,000 fitting data and validation data sets to evaluate the $\bar{X}$ and S validation tests. Tables IV and V show the type I errors for the $\bar{X}$ and S validation tests; while in some situations they are conservative, in many situations they achieve the nominal 0.05 value.

Tables VI and VII display the power for the $\bar{X}$ and S validation tests. From Tables VIII and IX, which display the difference between the $\bar{X}$ and S validation tests and the individual tests, the $\bar{X}$ and S validation tests generally outperform the individuals validation test, where a positive difference means that the $\bar{X}/S$ validation test power is larger than the individuals validation test power. Based on these results, the $\bar{X}$ and S validation tests are recommended. From the Table VI results for case 5, the relevant validation scenario for surveillance, we see that for two validation replicates, seven fitting replicates have a power of 0.89, while for three validation replicates, only four fitting replicates have the same power for detecting a mean shift. From the Table VII results for case 5, we see that for two validation replicates, seven fitting replicates have a power of 0.64, while for three validation replicates, only four fitting replicates have an even larger power of 0.81 for detecting an increased standard deviation. Based on these results, three validation replicates and four fitting replicates are recommended.

7. Discussion

In this article, we have proposed validation tests that account for prediction error and have evaluated power under different scenarios. If there are no replicates at time points, then the individuals validation test should be used. If there are replicates at time points, then the $\bar{X}/S$ validation tests are preferred over the individuals validation test because they have higher power.

The validation scenario considered here is different than the usual control chart setting. Although the individuals test for a single replicate is like a regression control chart, we are focused on validating for the next 4 years and use a Bonferroni correction to account for testing at multiple years. Besides monitoring for shifts in the mean, we are interested in detecting increases in the standard deviation. Also, whereas average run length characterizes a control chart, here we compute power of the validation tests over the next 4 years. Consequently, we evaluate the power for both shifts in the mean and increases in the standard deviation. Finally, unlike the regression control chart, which typically has different covariate values, we can have replicates for each of the validation years so that the $\bar{X}$ and S validation tests are more powerful than the individuals validation test.

Acknowledgements

We thank two anonymous referees for comments on an earlier version of this article.

References

1. Bierbaum R, Diegert K, Hamada MS, Huzurbazar A, Robertson A. Using statistical methods to assess a surveillance program. Los Alamos National Lab Technical Report LA-UR-12-22724 and LA-UR-13-25760, Los Alamos National Lab, Los Alamos, NM, 87545 USA. Accepted by *Quality Engineering*, 2012.
2. Draper NR, Smith H. Applied Regression Analysis Third Edition. John Wiley and Sons, Inc.: New York, NY, 1998.

3. Dunn OJ. Journal of the American Statistical Association 1961; **56**:52–64.
4. Mandel BJ. The regression control chart. *Journal of Quality Technology* 1969; **1**:1–9.

Authors' biographies

M. S. Hamada is a Scientist and holds a Ph.D. in Statistics from the University of Wisconsin-Madison. He is a Fellow of the American Statistical Association.

J. I. Abes is an R&D Engineer and holds a Ph.D. in Chemical Engineering from MIT.