Ruoheng Du

Regression and Multivariate Data Analysis

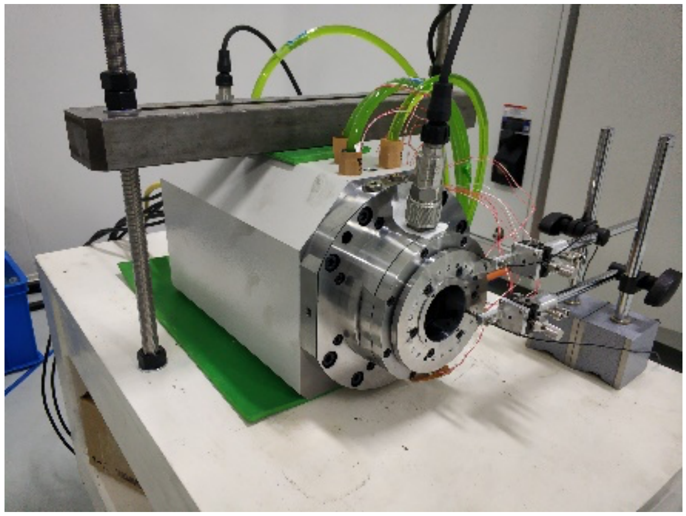
Prof. Jeff Simonoff

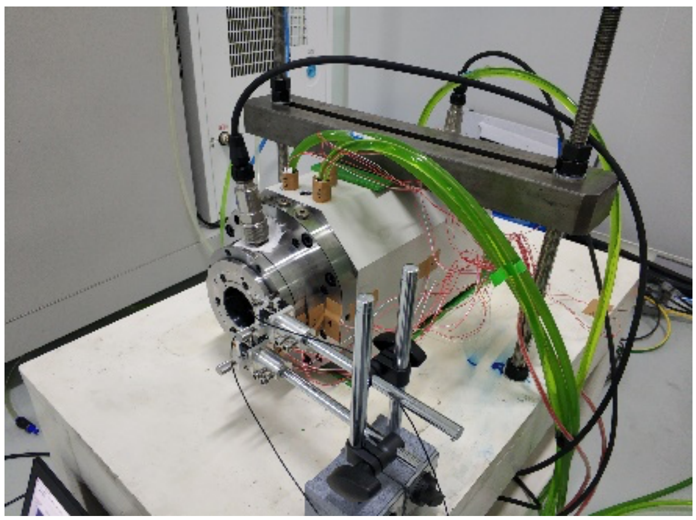
Mar 28, 2023

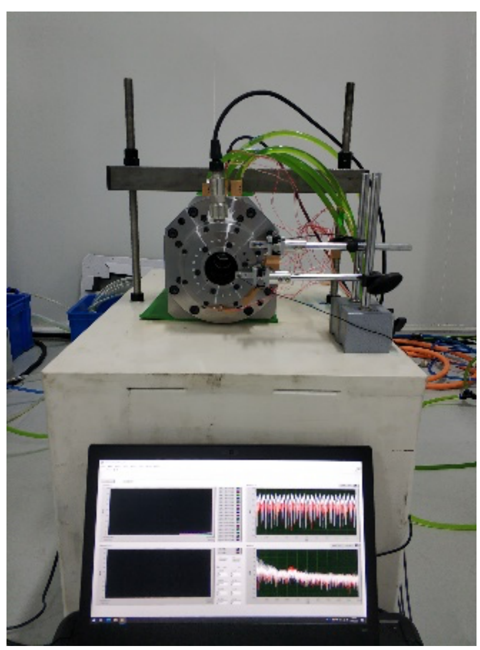
Thermal Expansion Error Modeling of High-speed Electric Spindle

High-speed electric spindles are frequently utilized in the machining industry to achieve great precision and efficiency. However, these spindles are subject to thermal expansion, which can cause errors in the machining process and reduce accuracy (Bryan, 1990). Monitoring and predicting the thermal error of the electric spindles is essential for maintaining high precision in machining operations. Developments in this area could have a significant impact on the accuracy and effectiveness of contemporary manufacturing processes. For instance, manufacturers can boost productivity, reduce waste, and improve product quality by detecting and compensating for thermal expansion errors.

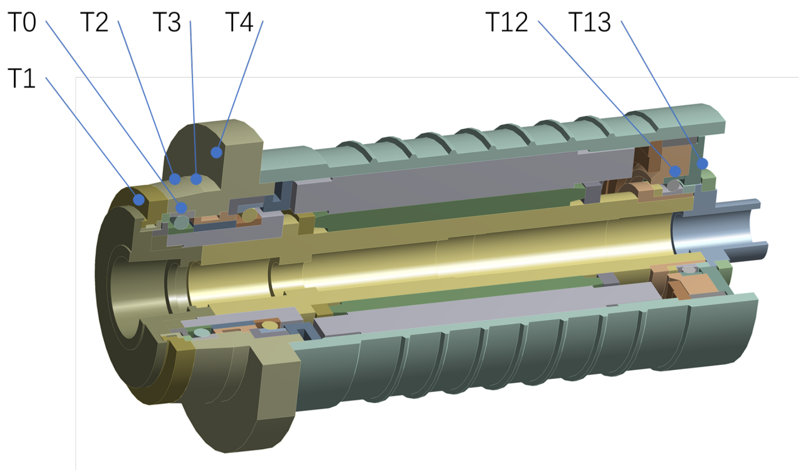
Therefore, it is crucial to track and predict the thermal expansion error of high-speed electric spindles and ensure high precision in machining operations. Such thermal expansion error is affected by many factors, including the spindle design, material properties, environmental conditions, and operating parameters (Li et al., 2019). And this report aims to investigate the relationship between thermal expansion error of the central shaft of the electric spindle (i.e., delta L) and the temperatures of the environment and different parts of the electric spindle. This is an experiment conducted at the School of Mechanical Engineering, Shanghai Jiao Tong University, and the following are photos of the experiment.







Below is a diagram of the temperature sensors showing the locations of different sensors on the central shaft of the electric spindle with their detected values as the predicting variables for this report.



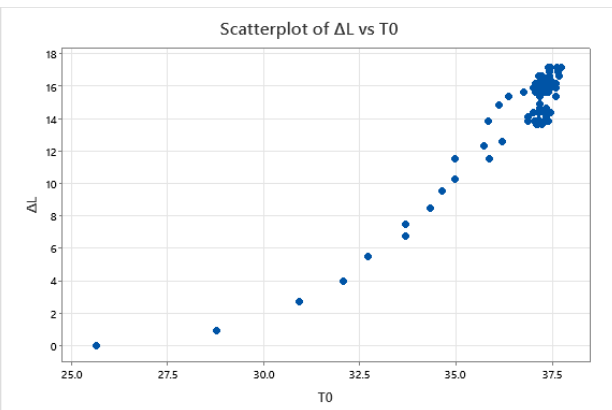
Here, T0 is the temperature measured inside the front bearing, T1 is the temperature of the front bearing housing end face, T2 and T3 represent the temperature of two front bearing housing side, T4 is the temperature of the front bearing housing flange surface, T12 is the temperature inside rear bearing, and T13 is the temperature outside rear bearing. Besides, there's another predicting variable named as T15, which is the environment temperature. The relationship between delta L and temperatures of the environment and different parts of the electric spindle can be expressed by a multiple regression model:

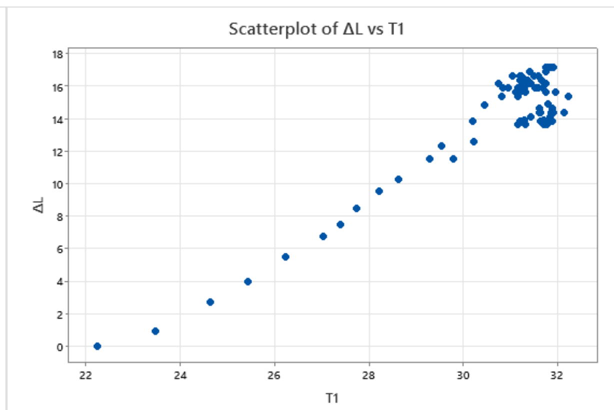
$$delta\ L = \beta_o + \beta_1 * T0 + \beta_2 * T1 + \beta_3 * T2 + \beta_4 * T3 + \beta_5 * T4 + \beta_6 * T12 + \beta_7 * T13 + \beta_8 * T15 + random\ error.$$

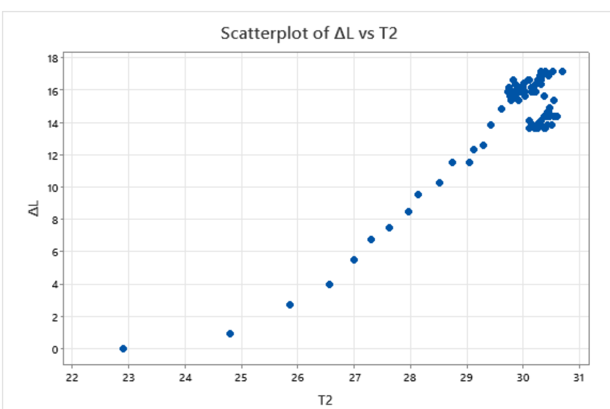
The following analysis is based on data from a sample of 79 set of thermal expansion error of the central shaft of the electric spindle and the temperatures, which are collected from the experiment. The fact that the data are collected at one-minute intervals, representing the relative change in thermal expansion error per minute under various temperature settings, is an important feature of the data in this report. Besides, the unit of the target variable is micrometer (μm) and all temperatures are measured in Celsius (°C).

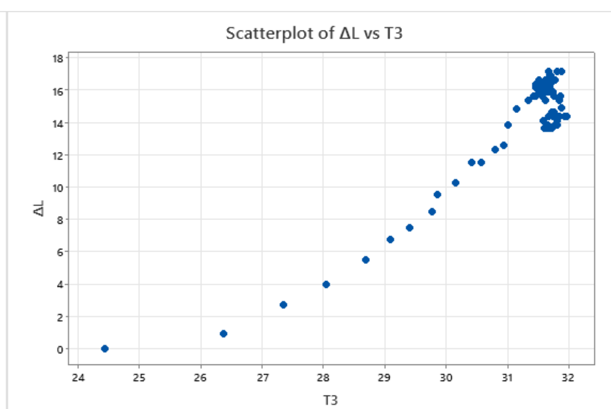
Here is the descriptive statistics and the scatterplot of the variables.

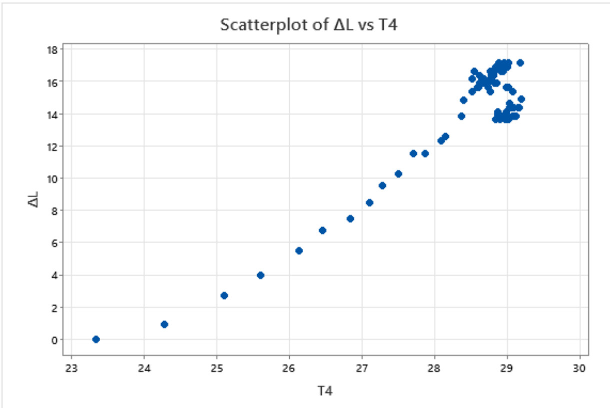
Variable	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
T0	79	36.517	0.226	2.011	25.634	36.883	37.205	37.418	37.752
T1	79	30.701	0.222	1.976	22.239	30.956	31.343	31.737	32.230
T2	79	29.673	0.151	1.340	22.898	29.776	30.144	30.356	30.692
T3	79	31.185	0.145	1.288	24.433	31.464	31.638	31.744	31.959
T4	79	28.445	0.123	1.094	23.339	28.596	28.803	28.996	29.192
T12	79	49.537	0.616	5.477	28.130	49.529	52.210	52.415	52.625
T13	79	31.378	0.255	2.263	24.301	31.313	32.395	32.727	32.945
T15	79	20.619	0.0814	0.724	18.686	20.258	20.735	21.147	21.630
ΔL	79	13.949	0.413	3.668	0.000	13.818	15.331	16.105	17.117

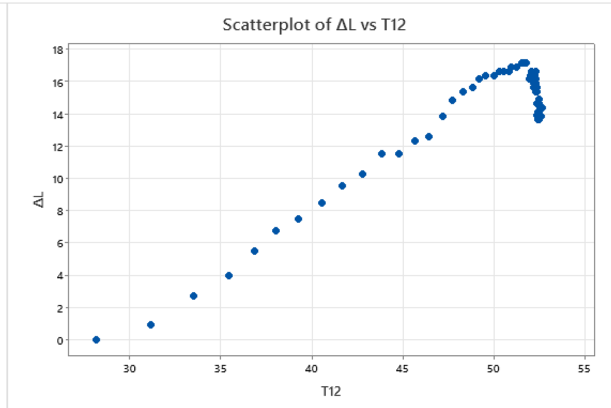


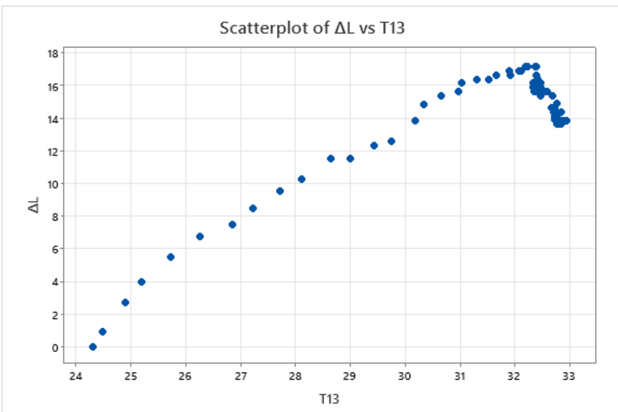


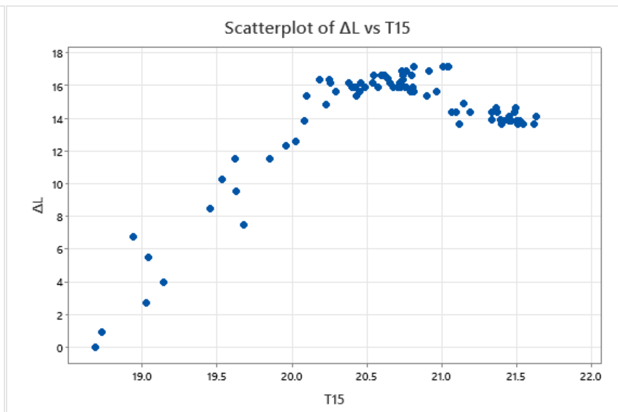












The target variable, delta L, has a mean of 13.949µm and a standard deviation of 3.668µm, which means that on average, the thermal expansion error of the central shaft of the electric spindle is 13.949µm. For predicting variables, the mean of T0, T1, T2, T3, T4, T12, T13, and T15 are 36.517µm, 30.701µm, 29.673µm, 31.185µm, 28.445µm, 49.537µm, 31.378µm, and 20.619µm respectively and the standard deviation of these predicting variables are 2.011µm, 1.976µm, 1.340µm, 1.288µm, 1.094µm, 5.477µm, 2.263µm, and 0.0814µm respectively. In addition, as can be seen from these scatterplots, there does appear to be a positive linear relationship between each of all the eight predicting variables and the target variable delta L, although these relationships do not appear to be absolutely linear and the possible quadratic relationship will be discussed later. The below is the least squares regression with all the eight predicting variables.

Regression Equation

$\Delta L = -44.9 + 1.188\ T0 + 0.727\ T1 - 0.71\ T2 + 0.21\ T3 + 1.03\ T4 - 0.936\ T12 + 2.84\ T13 - 3.122\ T15$

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-44.9	25.3	-1.77	0.081	
T0	1.188	0.567	2.10	0.040	129.93
T1	0.727	0.592	1.23	0.223	136.80
T2	-0.71	1.20	-0.60	0.554	256.78
T3	0.21	1.81	0.11	0.909	544.66
T4	1.03	1.83	0.56	0.575	399.88
T12	-0.936	0.618	-1.52	0.134	1144.92
T13	2.84	1.22	2.34	0.022	758.41
T15	-3.122	0.441	-7.08	0.000	10.20

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.882909	94.80%	94.21%	87.14%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	8	995.14	124.393	159.57	0.000
T0	1	3.43	3.426	4.40	0.040
T1	1	1.18	1.178	1.51	0.223
T2	1	0.28	0.276	0.35	0.554
T3	1	0.01	0.010	0.01	0.909
T4	1	0.25	0.248	0.32	0.575
T12	1	1.79	1.792	2.30	0.134
T13	1	4.26	4.255	5.46	0.022
T15	1	39.03	39.030	50.07	0.000
Error	70	54.57	0.780		
Total	78	1049.71			

Overall, the F-statistic of this regression has a p-value of 0.000, which is statistically significant, and this regression is quite strong as the R-squared value is 94.80% and adjusted R-squared value is 94.21%. In other words, T0, T1, T2, T3, T4, T12, T13, and T15 are able to account well for the observed variability in delta L. The constant coefficient indicates that when all the temperature variables equal zero Celsius (i.e., T0, T1, T2, T3, T4, T12, T13, and T15 are all zero Celsius), the estimated expected delta L is -44.9µm (that is, the thermal expansion error of the central shaft of the electric spindle is expected to be -44.9µm. But this point does not have any practical interpretation, since it is meaningless to discuss a negative thermal expansion error.

As for T0, T1, T2, T3, T4, T12, T13, and T15 coefficients, in general, for one specific coefficient, holding all other coefficients constant, 1 ℃ change in the temperature of this coefficient is associated with a certain micrometer estimated expected increase or decrease in the variation of the thermal expansion error. Take T15 coefficient as an example. It indicates that holding other predicting variables constant (i.e., holding T0, T1, T2, T3, T4, T12, and T13 constant), 1 ℃ change in the temperature of the environment is associated with an estimated expected -3.122μm change in the delta L (i.e., thermal expansion error of the central shaft of the electric spindle). The standard error of the estimate of 0.882909μm says that this model could be used to predict delta L within ±1.765818μm, roughly 95% of the time.

However, not all the coefficients are statistically significant. As for the constant coefficient, with the null hypothesis of $\beta_0 = 0$ and the alternative hypothesis of $\beta_0 \neq 0$, the t–statistic for the slope is:

$$t = \frac{\widehat{\beta_0} - Hypothesized\ Value}{\delta_{\widehat{\beta_0}}} = \frac{-44.9-0}{25.3} = -1.77.$$

This is marginally significant compared with the critical value of -1.645. And the p-value of this coefficient is 0.081, which is smaller than 0.10. So, we are 90% confident that the null hypothesis could be rejected and the alternative hypothesis of $\beta_0 \neq 0$ could be accepted. T0 coefficient is also statistically significant, as its t-statistics with the null hypothesis of $\beta_1 = 0$ can be calculated as below:

$$t = \frac{\widehat{\beta_1} - Hypothesized\ Value}{\delta_{\widehat{\beta_1}}} = \frac{1.188-0}{0.567} = 2.10.$$

So, it's obvious that this value is greater than the critical value of 1.96, and we are 95% confident that the alternative hypothesis of $\beta_1 \neq 0$ could be accepted. However, T1 coefficient is not statistically significant because its t-statistics is 1.23, which is less than 1.645, and p-value is

0.223, which is greater than 0.1. So, we are 90% confident that the null hypothesis of $\beta_2 = 0$ could be accepted. Similarly, T2, T3, and T4 coefficient are not statistically significant with their t-statistics of -0.60, 0.11, and 0.56 respectively, and p-values of 0.554, 0.909, and 0.575 respectively. As for T12 coefficient, with the null hypothesis of $\beta_6 = 0$, its t-statistics is:

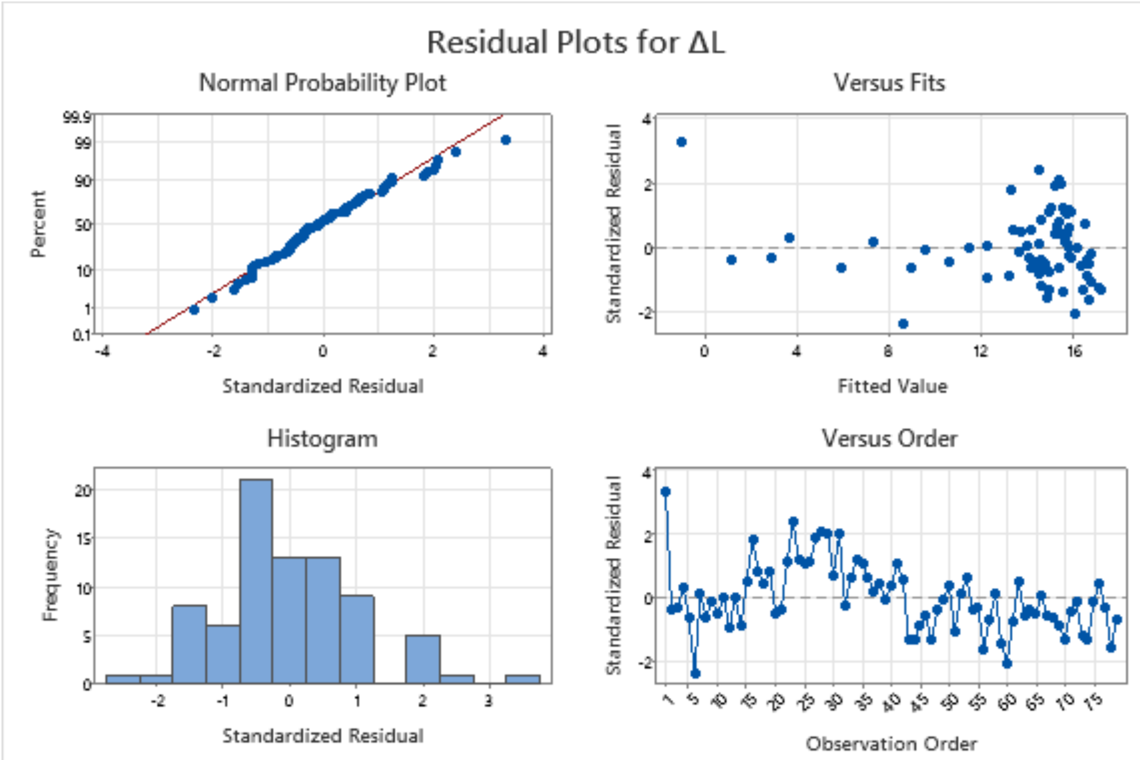
$$t = \frac{\widehat{\beta_6} - Hypothesized\ Value}{\delta_{\widehat{\beta_6}}} = \frac{-0.936-0}{0.618} = -1.52.$$

This value is slightly greater than the -1.645, so, we are 90% confident that the null hypothesis of $\beta_6 = 0$ could be accepted and the alternative hypothesis could be rejected. But T13 and T15 coefficient are statistically significant with their t-statistics of 2.34 and -7.08, and p-value of 0.022 and 0.000, which is definitely smaller than any reasonable significance level. In conclusion, among all the nine coefficients in this regression model, constant coefficient, T0 coefficient, T13 coefficient, and T15 coefficient are statistically significant, and we are 90% confident that $\beta_0 \neq 0$, $\beta_1 \neq 0$, $\beta_7 \neq 0$, and $\beta_8 \neq 0$. Besides, multicollinearity does seem to be a problem in this regression, as shown by the VIF value of all the variables above. The general guideline for VIF value is that

$$\max\left(10, \frac{1}{1-R^2_{model}}\right) = max\left(10, \frac{1}{1-94.80\%}\right) = \max(10, 19.23) = 19.23.$$

So, all the VIF values above, except for T15 coefficient, are problematic because they are greater than 19.23.

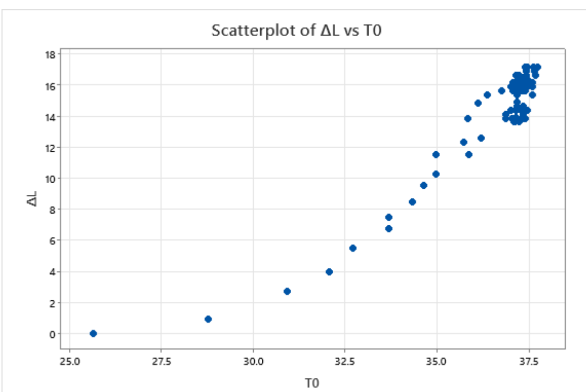
Now, let's take a look at the "four in one" plot to see if our assumptions of least squares regression hold true.

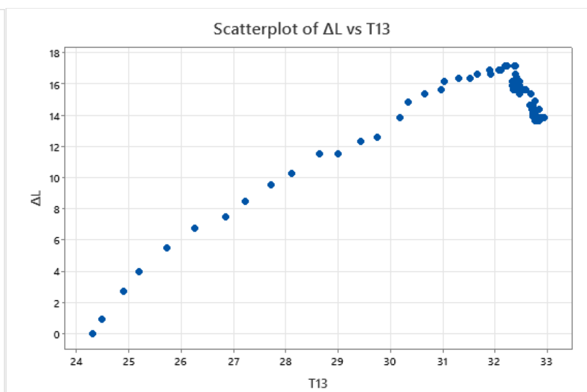


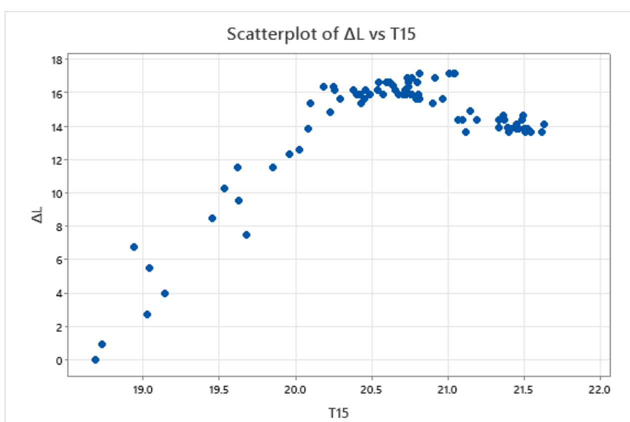
There's insufficient evidence to say that the first assumption is violated and the error has a non-zero expected value because there does not exist specific subgroups with residuals that systematically appear to be deviated from the zero line. As for the second assumption of homoscedasticity, the points are more concentrated in the right half of the plot of standardized residuals versus fitted values, which does not exhibit a lack of pattern, and therefore, indicates nonconstant variance and violation of this assumption. The third assumption suggests that the errors should be uncorrelated with each other, and this is not violated because delta L is only very much influenced by the temperature variables in this model, so, it is impossible to know the error of one specific delta L given the error of another delta L. However, the fourth assumption, that errors are normally distributed, is slightly violated. As we can see from the normal probability plot, one point on the top-right deviates from the line. Similarly, as shown in the histogram, the distribution is skewed to the right, indicating non-normality.

Here, the top-right point in the normal probability plot, the top-left point in the plot of standardized residuals versus fitted values, and the top-left point in the plot of standardized residuals versus order are the same point, which is the first point recorded in the experiment. This is a leverage point that has unusual predicting variable values as well as an outlier point with an unusual response variable value of 0. But, as it is the first point in the experiment with particular significance, this is not a major issue. For the time being, let's instead focus on improving our model rather than considering eliminating this point.

As discussed above, T0, T13, and T15 coefficients are statistically significant, which indicates that these predicting variables may be associated with the change in delta L. Let's take a look at the scatterplot of delta L versus T0, T13, and T15 respectively.







As we can see from the plots, each of the three predictive variables and the response variable delta L do appear to be correlated positively, and the new multiple regression model can be expressed as:

$$delta\ L = \beta_o + \beta_1 * T0 + \beta_2 * T13 + \beta_3 * T15 + random\ error.$$

The below is the least squares regression with T0, T13, and T15 as predicting variables.

Regression Equation

ΔL = -12.13 + 1.039 T0 + 1.384 T13 - 2.682 T15

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-12.13	5.10	-2.38	0.020	
T0	1.039	0.135	7.71	0.000	6.56
T13	1.384	0.180	7.67	0.000	14.89
T15	-2.682	0.364	-7.36	0.000	6.20

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.934586	93.76%	93.51%	90.24%

Analysis of Variance

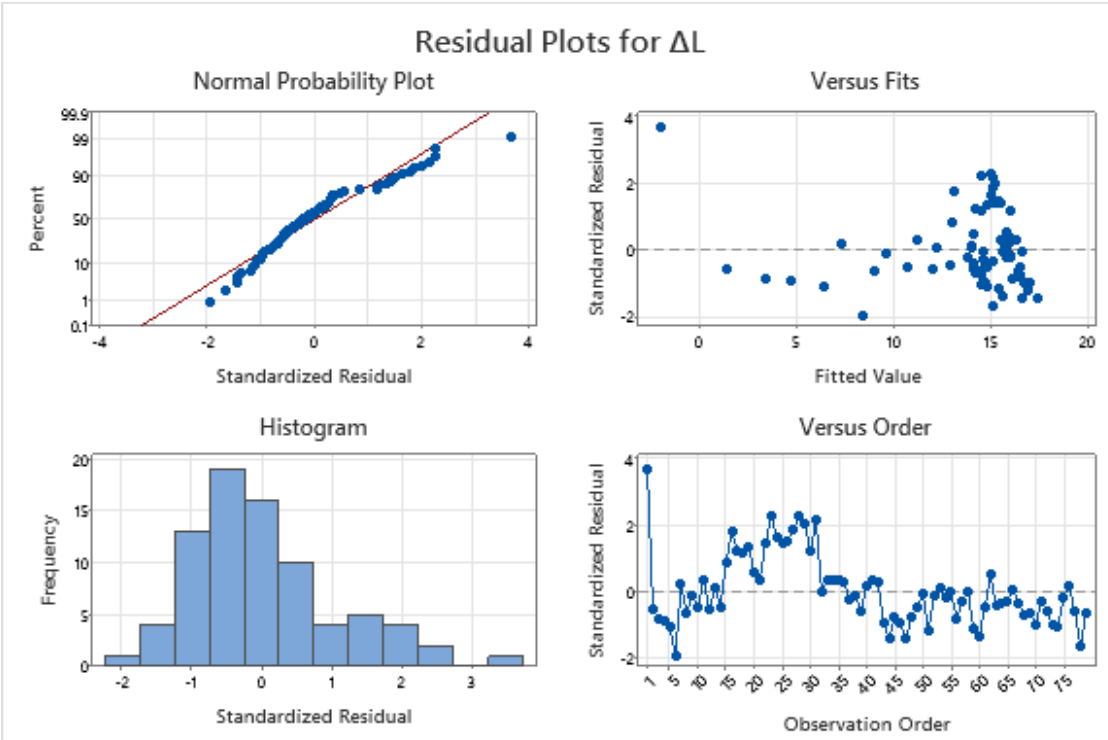
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	984.20	328.067	375.60	0.000
T0	1	51.95	51.951	59.48	0.000
T13	1	51.42	51.415	58.86	0.000
T15	1	47.35	47.354	54.21	0.000
Error	75	65.51	0.873		
Total	78	1049.71			

This regression is still quite strong as the R-squared value is 93.76% and adjusted R-squared value is 93.51%. In other words, T0, T13, and T15 are able to account for the observed variability in delta L well. The new constant coefficient indicates that when the temperature of T0, T13, and T15 are all zero Celsius, the thermal expansion error of the central shaft of the electric spindle is expected to be -12.13µm. But this point is meaningless because it has a negative value. As for T0, T13, and T15 coefficients, they indicate that 1) holding T13 and T15 constant, 1 °C change in the temperature inside the front bearing is associated with an estimated expected 1.039μm change in the thermal expansion error of the central shaft of the electric spindle; 2) holding T0 and T15 constant, 1 ℃ change in the temperature inside the front bearing is associated with an estimated expected 1.384μm change in the thermal expansion error of the central shaft of the electric spindle; 3) holding T0 and T13 constant, 1 ℃ change in the temperature inside the front bearing is associated with an estimated expected -2.682μm change in the thermal expansion error of the central shaft of the electric spindle. This time, all the four coefficients are statistically significant as all the absolute values of the t-statistics are greater than 1.96, which is the critical value of the 5% significance level, and all the p-values are less than 0.05. Besides, multicollinearity is not a problem in this regression. The new general guideline for VIF value is that

$$\max\left(10, \frac{1}{1-R^2_{model}}\right) = max\left(10, \frac{1}{1-93.76\%}\right) = \max(10, 16.03) = 16.03.$$

So, as shown above, the VIF values for all predicting variables are less than 16.03.

The “four in one” plot of the new regression is shown below.



This “four in one” plot is almost the same as the one with eight predicting variables before. The second assumption of homoscedasticity is somewhat violated as the points are still more concentrated in the right half of the plot of standardized residuals versus fitted values, indicating nonconstant variance. And the fourth assumption of the normal distribution of errors is still slightly violated by top-right point in the normal probability plot. The unusual point in this regression model is still the first point recorded in the experiment, and given its uniqueness, let’s keep this point for now.

Let’s take a further look at the current two models. This model, $delta\ L = \ \beta_o + \beta_1 * T0 + \beta_2 * T13 + \beta_3 * T15 + random\ error$, is a simpler model (i.e., restricted model), which is a special case of our previous full model (i.e., unrestricted model), $delta\ L = \ \beta_o + \beta_1 * T0 + \beta_2 * T1 + \beta_3 * T2 + \beta_4 * T3 + \beta_5 * T4 + \beta_6 * T12 + \beta_7 * T13 + \beta_8 * T15 + random\ error$, with eight predicting variables. Therefore, we can use a F-test to examine whether restricting T1, T2, T3, T4, and T12 coefficients is meaningful. This can be formulated as a hypothesis testing problem of

$$H_0: \beta_{2_{full}} = \beta_{3_{full}} = \beta_{4_{full}} = \beta_{5_{full}} = \beta_{6_{full}} = 0$$

versus

$$H_a: at\ least\ one\ of\ \beta_{2_{full}}, \beta_{3_{full}}, \beta_{4_{full}}, \beta_{5_{full}}, \beta_{6_{full}} \neq 0\ .$$

The F-statistics can be calculated as

$$F = \frac{\frac{R^2_{full} - R^2_{subset}}{5}}{\frac{1 - R^2_{full}}{79 - 8 - 1}} = \frac{\frac{94.80\% - 93.76\%}{5}}{\frac{1 - 94.80\%}{79 - 8 - 1}} = 2.8$$

on (5, 70) degree of freedom, which is greater than the critical value of 2.346 for significance level of 5% and smaller than the critical value of 3.291 for significance level of 1%.

Therefore, to select relevant predicting variables more systematically, let's perform model selection and see if leaving out some variables in the full model will be a good idea.

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	AICc	BIC	Cond No	T0	T1	T2	T3	T4	T12	T13	T15
1	88.4	88.3	163.9	84.4	80.5	1.2549	264.356	271.144	1.000	X							
1	88.1	88.0	129.5	87.7	84.9	1.2721	266.518	273.306	1.000		X						
1	87.1	86.9	164.5	84.3	99.1	1.3275	273.244	280.032	1.000				X				
2	92.4	92.2	88.9	91.5	29.8	1.0267	233.835	242.772	11.910		X						X
2	92.0	91.8	89.2	91.5	34.5	1.0501	237.392	246.329	15.078						X		X
2	89.5	89.2	177.9	83.1	68.8	1.2058	259.248	268.185	74.944	X	X						
3	94.4	94.2	64.0	93.9	4.7	0.88718	211.995	223.021	89.524		X					X	X
3	94.2	93.9	69.3	93.4	7.5	0.90338	214.855	225.880	132.337		X				X		X
3	94.1	93.9	78.8	92.5	8.1	0.90651	215.402	226.427	54.991			X				X	X
4	94.5	94.3	83.6	92.0	4.4	0.87958	211.920	224.970	254.499	X	X					X	X
4	94.4	94.1	77.6	92.6	5.9	0.88810	213.444	226.494	359.073		X		X			X	X
4	94.4	94.1	74.3	92.9	5.9	0.88821	213.464	226.514	698.354		X	X				X	X
5	94.7	94.4	80.1	92.4	3.8	0.86951	211.438	226.447	5143.772	X	X				X	X	X
5	94.7	94.3	82.0	92.2	4.6	0.87414	212.277	227.286	4188.574	X				X	X	X	X
5	94.6	94.2	88.5	91.6	5.5	0.87968	213.274	228.283	4407.600	X			X		X	X	X
6	94.8	94.3	80.9	92.3	5.4	0.87276	213.417	230.315	7195.615	X	X			X	X	X	X
6	94.8	94.3	86.9	91.7	5.5	0.87357	213.564	230.462	9615.160	X	X		X		X	X	X
6	94.7	94.3	91.8	91.3	5.8	0.87549	213.910	230.808	6386.390	X	X	X			X	X	X
7	94.8	94.3	100.5	90.4	7.0	0.87675	215.584	234.301	8767.080	X	X	X		X	X	X	X
7	94.8	94.3	92.3	91.2	7.3	0.87866	215.927	234.643	13218.193	X	X	X	X		X	X	X
7	94.8	94.3	97.9	90.7	7.4	0.87889	215.968	234.685	13948.525	X	X		X	X	X	X	X
8	94.8	94.2	135.0	87.1	9.0	0.88291	218.196	238.655	16712.756	X	X	X	X	X	X	X	X

Based on the above model selection outputs by Minitab and the selection criteria of high R_a^2, small Mallows C_p, high R_{pred}^2, and small AIC_c, the highlighted rows are the regression models with different number of predicting variables that have the best performance among all. In the simplest model in the output, it takes T1, T13, and T15 as the predicting variables, which is consistent with the result of the F-test. And the below is the regression model.

Regression Equation

ΔL = -3.42 + 1.554 T1 + 1.029 T13 - 3.037 T15

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-3.42	4.31	-0.79	0.430	
T1	1.554	0.180	8.62	0.000	12.59
T13	1.029	0.199	5.17	0.000	20.06
T15	-3.037	0.334	-9.10	0.000	5.78

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.887178	94.38%	94.15%	93.91%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	990.68	330.226	419.56	0.000
T1	1	58.43	58.428	74.23	0.000
T13	1	21.08	21.078	26.78	0.000
T15	1	65.20	65.199	82.84	0.000
Error	75	59.03	0.787		
Total	78	1049.71			

With the R-squared value is 94.38% and adjusted R-squared value is 94.15%, this regression model of delta L versus T1, T13, and T15 seems to work slightly better than the regression model of delta L versus T0, T13, and T15. However, different from the regression model of delta L versus T0, T13, and T15, the constant coefficient in this model is not statistically significant with a large p-value of 0.430, which is greater than any reasonable significance level. Besides, multicollinearity is again a problem. As guideline suggests:

$$\max\left(10, \frac{1}{1-R_{model}^2}\right) = max\left(10, \frac{1}{1-94.38\%}\right) = \max(10, 17.79) = 17.79.$$

So, the VIF value of 20.06 of T13 coefficient in this model is above the guideline.

It's interesting that despite some minor differences, the regression model of delta L versus T1, T13, and T15 and the regression model of delta L versus T0, T13, and T15 are similar in general. Let's take a look at the correlation matrix between some variables.

Correlations

	T0	T1	T2	T3
T1	0.974			
T2	0.982	0.990		
T3	0.991	0.984	0.993	
T4	0.977	0.992	0.996	0.994

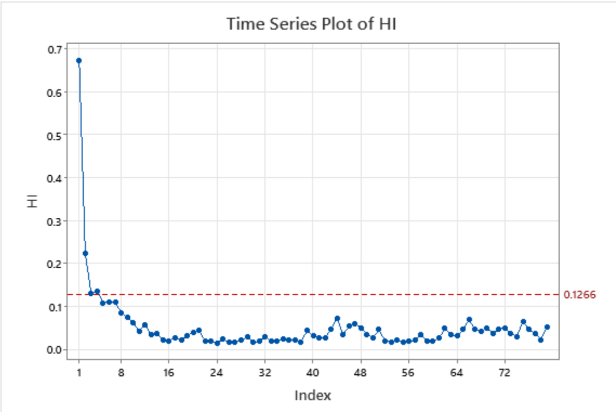
Correlations

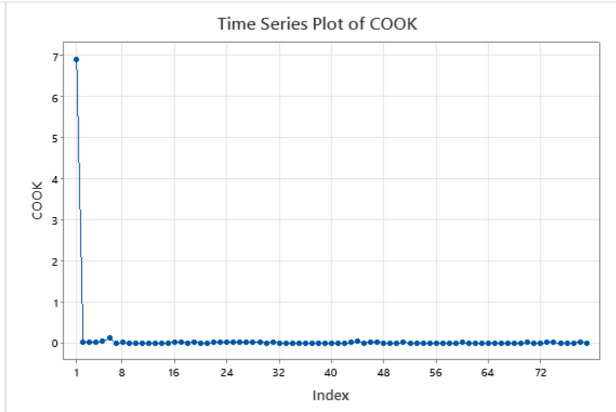
	T12
T13	0.992

T0, T1, T2, T3 and T4 are all located around the front bearing of the center shaft of the electric spindle. As a result, the temperatures measured by the sensors are quite close to each other, resulting in a significant correlation between each pair of the five variables. Similarly, T12 and T13 are both situated around the rear bearing of the center shaft of the electric spindle. Because of the close proximity of the temperatures detected by their sensors, there is also a strong correlation between them. In view of this, considering the complexity and validity of the model, we can choose one out of five from T0, T1, T2, T3 and T4, and one out of two from T12 and T13 to represent the temperature of the front and rear bearings. Given that the regression model of delta L versusT1, T13, and T15 and the regression model of delta L versus T0, T13, and T15 have very close R-squared values, but the coefficient of the latter is more statistically significant, we will continue to study the regression models of delta L versus T0, T13, and T15 in the following analysis.

Before examining new variables in the model, let's perform the diagnostics by calculating leverage values and Cook's distances. The reference line in the HI plot is

$$2.5 * \frac{p+1}{n} = 2.5 * \frac{3+1}{79} = 0.1266.$$

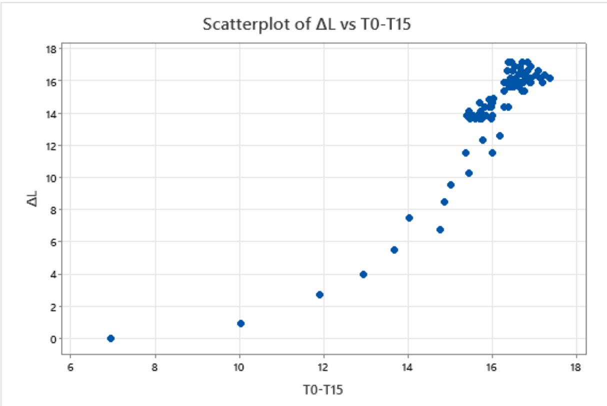


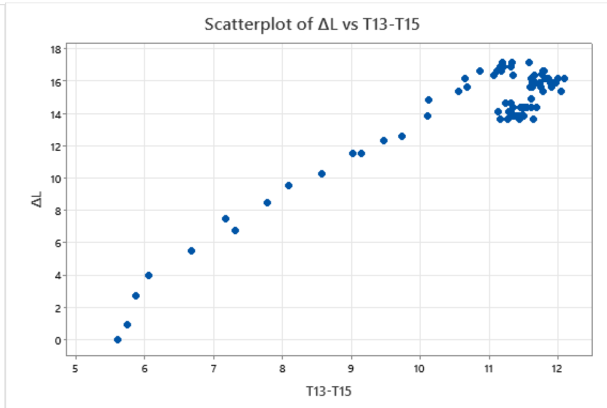


As shown in the Time Series Plot of HI, based on the guideline for a large leverage value, the first two points recorded in the experiment are clearly above 0.1266. In the Time Series Plot of Cook's Distance, most of the points have a Cook's distance close to 0, and only the first recorded point appears to be particularly influential. So, obviously, the first point of this experiment is really worth discussing. Let's put this point aside for the time being and look at the new model.

It might be beneficial to alter the model by introducing new predicting variables that accounts for the difference between the environment temperature and temperatures measured on the different parts of the electric spindle. This is due to the fact that temperature is a relative measurement and that the thermal expansion error may not be accurately represented by simply measuring the temperature at a particular location. As a result, taking into account the difference between the environment temperature and temperatures measured on the different parts of the electric spindle may offer a more accurate way to assess how temperature variables and the thermal expansion error of the electric spindle are associated. In this case, we can modify the three predicting variables T0, T13, and T15 to two predicting variables T0-T15 and T13-T15. And here's the descriptive statistics.

Variable	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
T0-T15	79	15.898	0.172	1.525	6.948	15.722	16.283	16.683	17.353
T13-T15	79	10.759	0.184	1.637	5.615	10.858	11.342	11.664	12.098





Here, the mean of T0-T15 and T13-T15 is 15.898μm and 10.759μm respectively, and the standard deviation is 1.525μm and 1.637μm respectively. The plots also show that 1) in the relationship between delta L and T0-T15, delta L increases with T0-T15, but the rate of increase varies; and 2) in the relationship between delta L and T13-T15, delta L increases with T13-T15, but the rate of growth is relatively stable. Therefore, we can express the relationship in a multiple regression model:

$$delta\ L = \beta_o + \beta_1 * (T0 - T15) + \beta_2 * (T13 - T15) + random\ error.$$

And the below is the result of the least squares regression.

Regression Equation

$$\Delta L = -16.85 + 1.090\,(T0\text{-}T15) + 1.252\,(T13\text{-}T15)$$

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-16.85	1.17	-14.36	0.000	
T0-T15	1.090	0.124	8.82	0.000	3.18
T13-T15	1.252	0.115	10.87	0.000	3.18

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.933992	93.68%	93.52%	91.16%

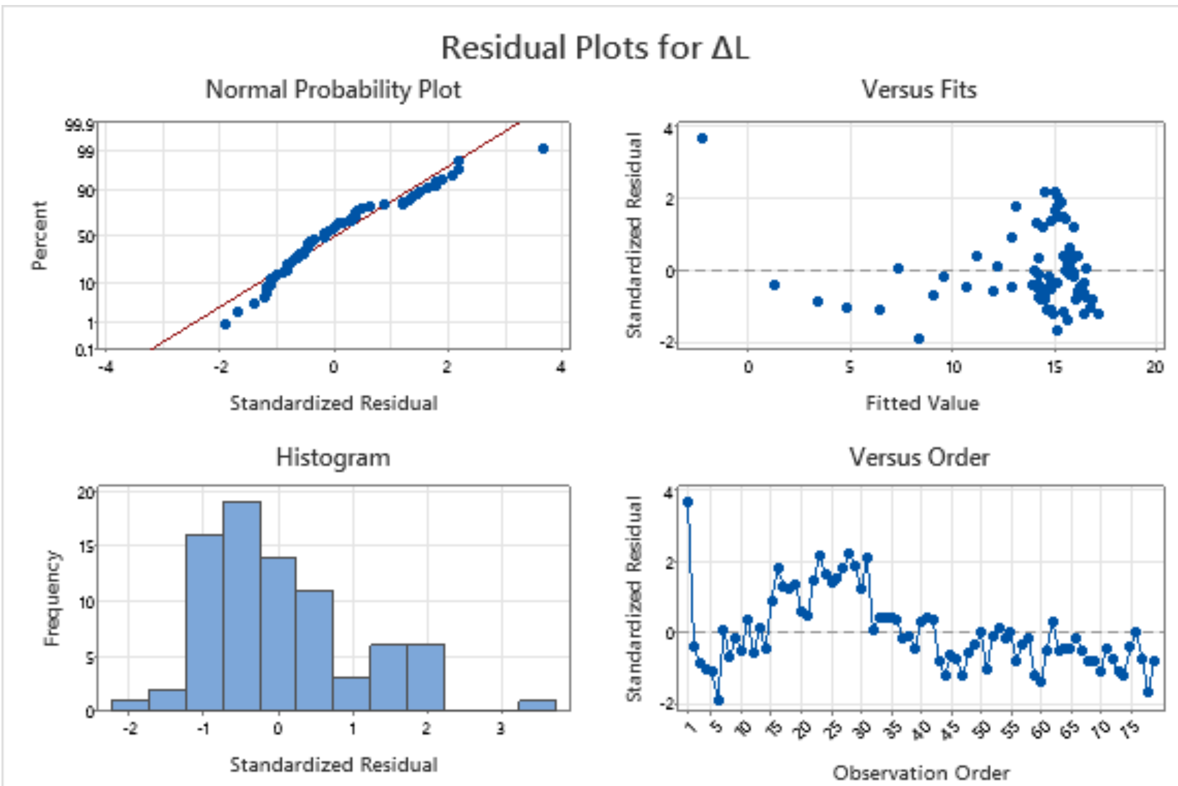
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	983.41	491.706	563.66	0.000
T0-T15	1	67.84	67.838	77.77	0.000
T13-T15	1	103.06	103.064	118.15	0.000
Error	76	66.30	0.872		
Total	78	1049.71			

This regression is strong because the R-squared value is 93.68%, which is slightly lower than that of the regression model of delta L versus T0, T13, and T15, and adjusted R-squared value is 93.52%, which is slightly higher than that of the regression model of delta L versus T0, T13, and T15. This means that T0-T15 and T13-T15 are able to account for the observed variability in delta L quite well. The constant coefficient indicates that when the temperature of T0-T15 and T13-T15 are zero Celsius, the thermal expansion error of the central shaft of the electric spindle is expected to be -16.85µm, which is negative and meaningless. As for T0-T15 coefficient, it means that holding T13-T15 constant, 1 °C change in the difference between the temperature measured by the sensor inside the front bearing and the environment temperature is associated with an estimated expected 1.090µm change in the thermal expansion error of the central shaft of the electric spindle. As for T13-T15 coefficient, it means that holding T0-T15 constant, 1 °C change in the difference between the temperature measured by the sensor outside the rear bearing and the environment temperature is associated with an estimated expected

1.252μm change in the thermal expansion error of the central shaft of the electric spindle. What's more, all the coefficients are statistically significant as all the p-values are around 0.000, which are equal or below any reasonable significance level. Besides, multicollinearity is no longer a problem in this regression since the VIF values of T0-T15 and T13-T15 are 3.18, which is obviously below the VIF value of 10 in the guideline.

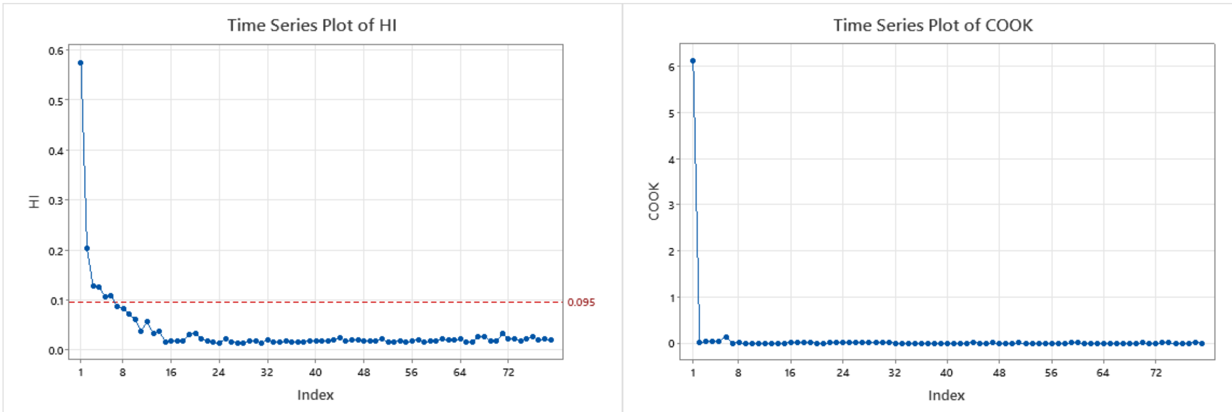
The following is the "four in one" plot for this regression.



As we have seen before, there is an obvious violation of the assumptions: nonconstant variance and non-normality of the residuals. The unusual point here is still an outlier and a leverage point, which is the first point recorded in the experiment and we will discuss it later.

Let's perform the diagnostics again and check the points. The reference line in this HI plot is

$$2.5 * \frac{p+1}{n} = 2.5 * \frac{2+1}{79} = 0.095.$$



The result is quite similar as the regression model of delta L versus T0, T13, and T15. In the Time Series Plot of HI, based on the guideline for a large leverage value, this time, the first six points recorded in the experiment are greater than 0.095. But in the Time Series Plot of Cook's Distance, most of the points still have a Cook's distance close to 0, and again, the first recorded point appears to be particularly influential. So, obviously, both the regression model of delta L versus T0, T13, and T15 and the regression model of delta L versus T0-T15 and T13-T15 indicate that the first point of this experiment may have a great impact on the model, and we will discuss this point once the preferred model is chosen.

In addition to these simpler regression models, to investigate the relationship between thermal expansion error of the central shaft of the electric spindle and the temperatures, we can also take into account the inclusion of quadratic variables. As shown in the scatterplots of delta L versus T0-T15 and T13-T15, their relationship is not a straight line, and a quadratic curvilinear relationship is possible. So, we would like to see whether the following four predicting variables, $(T0 - T15)^2$, $(T13 - T15)^2$, $(T0 - T15)$, and $(T13 - T15)$, would result in a better model and we use Minitab to perform the model selection.

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	AICc	BIC	Cond No	$(T0-T15)^2$	$(T13-T15)^2$	$T0-T15$	$T13-T15$
1	88.2	88.0	148.6	85.8	145.4	1.2689	266.113	272.901	1.000	X			
1	87.2	87.1	143.9	86.3	163.4	1.3199	272.335	279.124	1.000				X
1	83.9	83.7	255.5	75.7	226.1	1.4831	290.757	297.545	1.000			X	
2	94.6	94.5	64.1	93.9	27.3	0.86162	206.141	215.078	12.691	X			X
2	94.0	93.9	71.4	93.2	38.2	0.90709	214.266	223.204	10.645	X	X		
2	93.7	93.5	92.8	91.2	44.9	0.93399	218.885	227.822	10.622			X	X
3	95.2	95.0	59.6	94.3	18.4	0.81868	199.300	210.325	1193.766	X	X		X
3	94.9	94.7	63.2	94.0	24.0	0.84433	204.173	215.199	407.161	X		X	X
3	94.3	94.0	79.4	92.4	36.2	0.89688	213.714	224.740	414.946	X	X	X	
4	96.0	95.8	59.2	94.4	5.0	0.75003	186.747	199.797	1895.375	X	X	X	X

Based on the above outputs and the selection criteria of high R_a^2, small Mallows C_p, high R_{pred}^2, and small AIC_c, the highlighted rows are the regression models with different number of predicting variables that have the best performance among all. In the simplest model in the output, it takes $(T13 - T15)$, $(T0 - T15)^2$, and $(T13 - T15)^2$ as the predicting variables. And the below is the expression of the quadratic regression model and the regression result:

$$delta\ L = \beta_o + \beta_1 * (T13 - T15) + \beta_2 * (T0 - T15)^2 + \beta_3 * (T13 - T15)^2 + random\ error.$$

Regression Equation

ΔL = -20.03 + 3.673 (T13-T15) + 0.03930 (T0-T15)^2 - 0.1315 (T13-T15)^2

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-20.03	3.39	-5.90	0.000	
T13-T15	3.673	0.859	4.28	0.000	229.92
(T0-T15)^2	0.03930	0.00518	7.59	0.000	5.08
(T13-T15)^2	-0.1315	0.0434	-3.03	0.003	198.31

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.818680	95.21%	95.02%	94.32%

Analysis of Variance

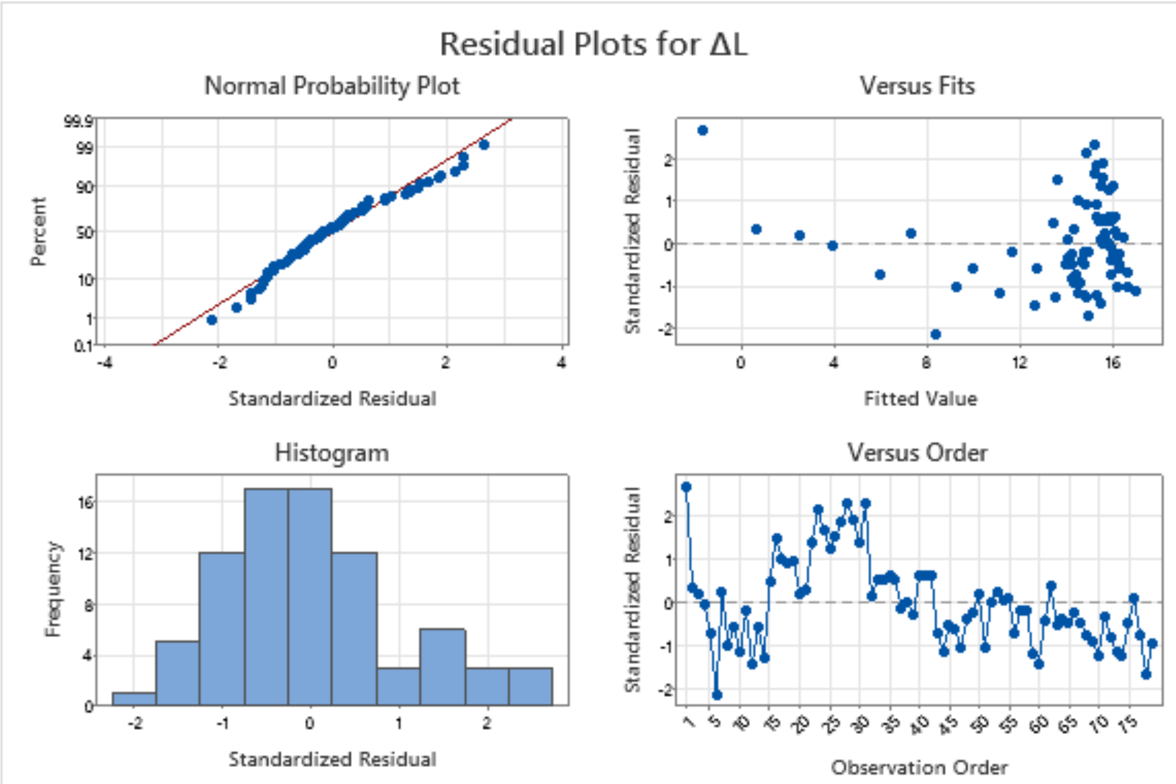
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	999.44	333.147	497.06	0.000
T13-T15	1	12.27	12.265	18.30	0.000
(T0-T15)^2	1	38.62	38.620	57.62	0.000
(T13-T15)^2	1	6.15	6.153	9.18	0.003
Error	75	50.27	0.670		
Total	78	1049.71			

Overall, this quadratic regression is relatively strong as the R-squared value is 95.02%, which is the highest among all the models we have performed and adjusted R-squared value is 94.32%. In other words, $(T13 - T15)$, $(T0 - T15)^2$, and $(T13 - T15)^2$ are able to account for the observed variability in delta L relatively well. All the four coefficients are statistically significant because their p-values are all equal or below any reasonable significance level. However, multicollinearity appears to be a great issue in this regression. The new general guideline for VIF value is that

$$\max\left(10, \frac{1}{1-R^2_{model}}\right) = max\left(10, \frac{1}{1-95.02\%}\right) = \max(10, 20.08) = 20.08,$$

and the VIF value of $(T13 - T15)$ and $(T13 - T15)^2$ are significantly greater than 20.08. This is because that in this model, it is not possible to avoid the interplay between the $(T13 - T15)$ coefficient and the $(T13 - T15)^2$ coefficient by holding one of them constant.

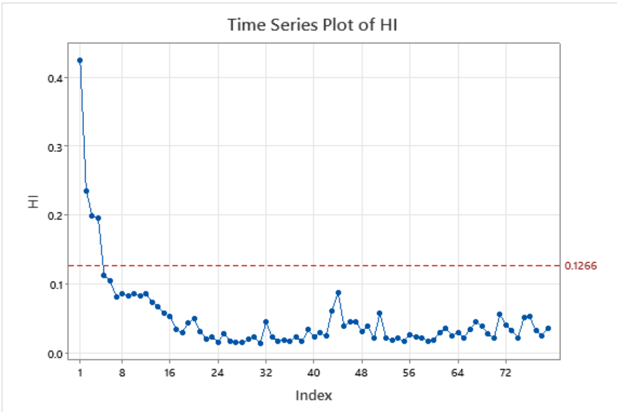
Now, let's look at its "four in one" plot to find out whether this model violates the four assumptions.

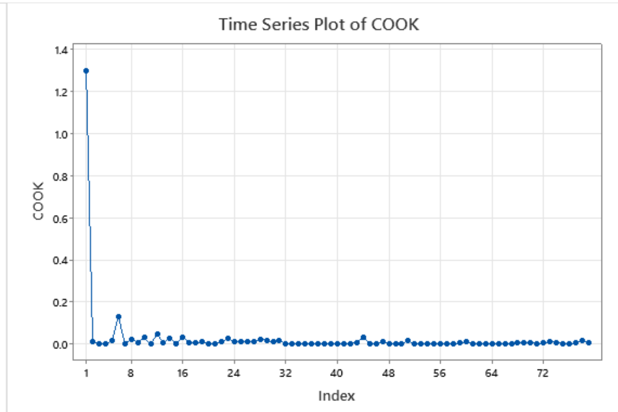


It is clear that there exists nonconstant variance in the plot of standardized residual versus fitted value, but in the normal probability plot, almost all points are located near the straight line with slight deviation. As for the histogram, it looks more like a normal distribution than the previous ones, but it still has a long right tail, implying that non-normality still exists. The unusual point, which is the first point recorded in the experiment, is still obvious on the top-left of plot of standardized residual versus fitted value as well as on the top-left of plot of standardized residual versus order, while it is not that obvious in the normal probability plot this time.

As for the leverage values and Cook's distances, the reference line of this HI plot is

$$2.5 * \frac{p+1}{n} = 2.5 * \frac{3+1}{79} = 0.1266.$$





This time, in the Time Series Plot of HI, the first four points in the experiment are above the guideline for a large leverage value of 0.1266, while in the Time Series Plot of Cook's Distance, in addition to the first point, there is another relatively noticeable point, but that point is still less than the suggested guideline value of Cook's D of 1. As a result, although it makes no sense to delete all the unusual points that may potentially influence the model, we can try to eliminate the first point in the following analysis because it has a rather unique target variable (i.e., delta L) of zero.

In short, so far, among all the regression models we have performed, two regression models stand out. The first one is a simpler multiple regression:

$$delta\ L = \beta_o + \beta_1 * (T0 - T15) + \beta_2 * (T13 - T15) + random\ error$$

or

$$delta\ L = -16.85 + 1.090 * (T0 - T15) + 1.252 * (T13 - T15).$$

The second one is a quadratic regression:

$$delta\ L = \beta_o + \beta_1 * (T13 - T15) + \beta_2 * (T0 - T15)^2 + \beta_3 * (T13 - T15)^2 + random\ error$$

or

$$delta\ L = -20.03 + 3.673 * (T13 - T15) + 0.03930 * (T0 - T15)^2 - 0.1315 * (T13 - T15)^2.$$

Both models have advantages, with the former requiring only a single term and the latter having a higher R-squared value and adjusted R-squared value.

Now, let's take a look at the unusual point that has appeared in all previous models. It is the first point in the experiment with T0 value of 25.634399, T13 value of 24.300758, T15 value of 18.686065 and the delta L value of 0. So, we will remove this point and check how eliminating it affects the two best models we mentioned previously.

For the simpler model, the regression statistics and plots are as follows.

Regression Equation

ΔL = -21.82 + 1.551 (T0-T15) + 1.026 (T13-T15)

Coefficients

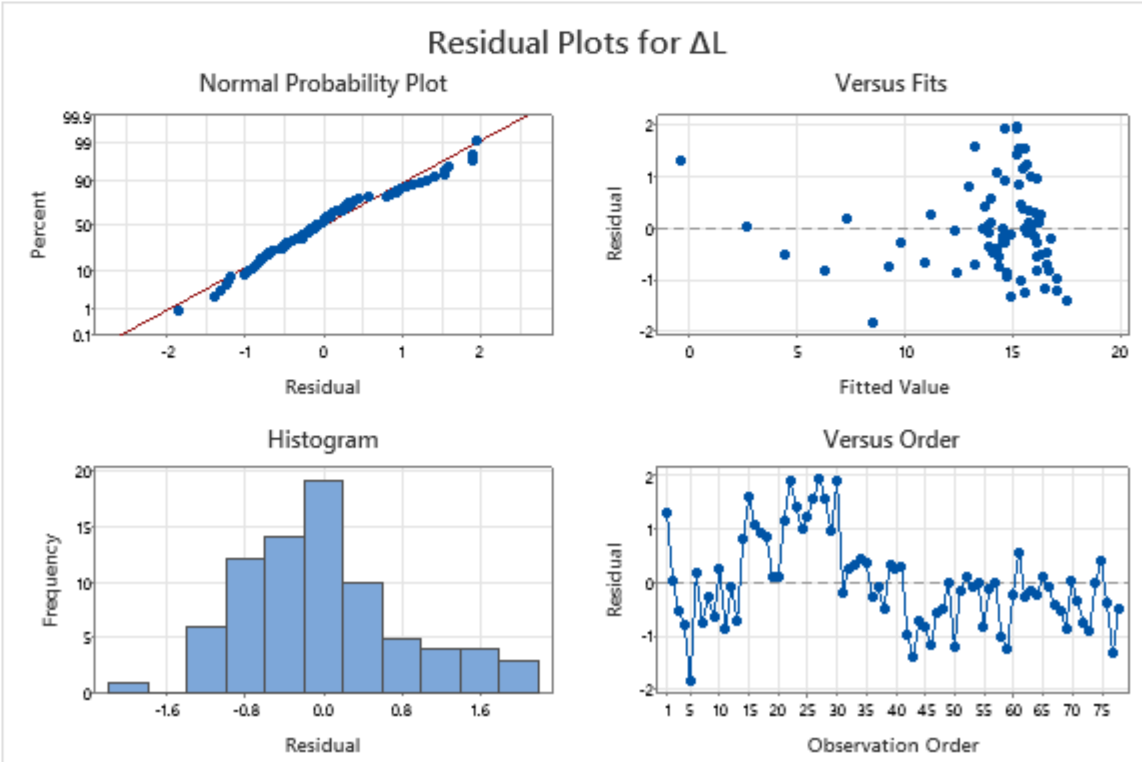
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-21.82	1.63	-13.40	0.000	
T0-T15	1.551	0.160	9.68	0.000	3.55
T13-T15	1.026	0.119	8.63	0.000	3.55

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.851815	93.62%	93.45%	92.79%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	798.23	399.116	550.06	0.000
T0-T15	1	67.95	67.954	93.65	0.000
T13-T15	1	54.05	54.047	74.49	0.000
Error	75	54.42	0.726		
Total	77	852.65			



As we can see from the regression equation, omitting the first point in the experiment will result in a new R-squared value of 93.62% and adjusted R-squared value of 93.45%, which slightly decreases compared with the previous R-squared value of 93.68% and adjusted R-squared value of 93.52%. This indicates that this new regression is a little bit weaker than before. The constant coefficient is smaller than before, and this means that in the current regression model, the estimated expected delta L with zero difference between the temperature measured by the sensor inside the front bearing (i.e., T0) and the environment temperature (i.e., T15) and zero difference between the temperature measured by the sensor outside the rear bearing (i.e., T13) and the environment temperature (i.e., T15) will be -21.82μm. But this is not practical since it's meaningless to discuss a negative delta L. All coefficients are statistically significant. The t-statistics of -13.40, 9,68, and 8.63 reject the null hypothesis of $\beta_0 = 0$, $\beta_1 = 0$, and $\beta_2 = 0$. For the normal probability plot, all the points are situated near the straight line, and the histogram does appear to be concentrated around 0. However, this regression still displays nonconstant variance.

For the quadratic model, the regression statistics and plots are as follows.

Regression Equation

ΔL = -21.79 + 3.759 (T13-T15) + 0.04689 (T0-T15)^2 - 0.1411 (T13-T15)^2

Coefficients

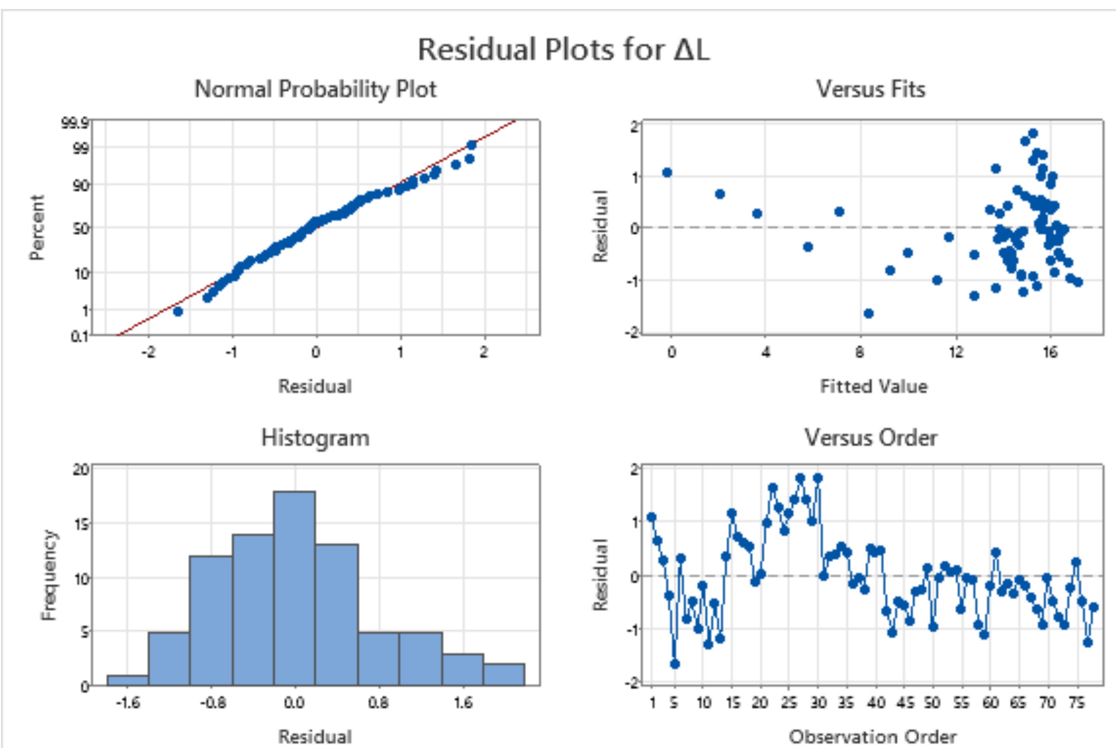
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-21.79	3.31	-6.58	0.000	
T13-T15	3.759	0.823	4.57	0.000	200.72
(T0-T15)^2	0.04689	0.00567	8.27	0.000	4.38
(T13-T15)^2	-0.1411	0.0417	-3.38	0.001	178.06

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.784493	94.66%	94.44%	93.93%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	807.109	269.036	437.15	0.000
T13-T15	1	12.826	12.826	20.84	0.000
(T0-T15)^2	1	42.133	42.133	68.46	0.000
(T13-T15)^2	1	7.033	7.033	11.43	0.001
Error	74	45.542	0.615		
Total	77	852.651			



As we can see, it results in a new R-squared value of 94.66% and adjusted R-squared value of 94.44%, which decreases compared with the previous R-squared value of 95.21% and adjusted R-squared value of 95.02%. This indicates that this new regression is somewhat weaker than before. The constant coefficient is smaller than before, and this means that in the current regression model, the estimated expected delta L with zero difference between the temperature measured by the sensor inside the front bearing (i.e., T0) and the environment temperature (i.e., T15) and zero difference between the temperature measured by the sensor outside the rear bearing (i.e., T13) and the environment temperature (i.e., T15) will be -21.79µm. But this is not practical as well. All coefficients are statistically significant. The t-statistics of -6.58, 4.57, 8.27, and -3.38 reject the null hypothesis of $\beta_0 = 0$, $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_3 = 0$. For the normal probability plot, all the points are situated around the straight line, and the histogram does appear to be concentrated around 0. However, this regression still displays nonconstant variance as shown in the plot.

In conclusion, from the analysis above, taking the thermal expansion error of the central shaft of the electric spindle (i.e., delta L) as the target variable and the temperature inside the front bearing (i.e., T0), the temperature outside rear bearing (i.e., T13), and the environment temperature (i.e., T15) as the predicting variables would result in the below two types of regression equation:

$$delta\ L = -21.82 + 1.551 * (T0 - T15) + 1.026 * (T13 - T15)$$

for the simpler multiple regression, and

$$delta\ L = -21.79 + 3.759 * (T13 - T15) + 0.04689 * (T0 - T15)^2 - 0.1411 * (T13 - T15)^2$$

for the quadratic regression with higher R-squared value and adjusted R-squared value.

To summarize, investigating the relationship between thermal expansion error of the central shaft of the electric spindle (i.e., delta L) and the temperatures of the environment and different parts of the electric spindle is crucial for maintaining high precision in machining processes. By studying this, researchers can develop effective compensation methods, optimize spindle design and operating parameters, thus improving machining accuracy.

Works Cited

Bryan, J. (1990). International status of Thermal Error Research (1990). *CIRP Annals, 39*(2), 645–656. https://doi.org/10.1016/s0007-8506(07)63001-7

Li, B., Tian, X., & Zhang, M. (2019). Thermal error modeling of machine tool spindle based on the improved algorithm optimized BP Neural Network. *The International Journal of Advanced Manufacturing Technology, 105*(1-4), 1497–1505. https://doi.org/10.1007/s00170-019-04375-w