Time Series Analytics Notes
Model Building

General Comments on Statistical Model Building

A **theory** [1] has three objectives:

1) to explain existing data,
2) to predict new data, and
3) to do so simply.

It is the same with a **statistical theory**. A statistical theory is a **statistical model** that explains the data we have, and predicts data that we do not yet have, all the while trying to be as simple as possible. You may think about this in terms of the General Statistical Model, **Actual = Fit + Residual**. The data that we must explain and future data that we must predict are the Actuals. The explanations and predictions are embodied in the Fits, which are the model-estimated or model-predicted values. There are many different types of statistical models: Random Sample, Simple Linear Regression, Multiple Linear Regression, Random Walk, Autoregression, Nonlinear Regression, ARIMA, Vector Autoregression, Neural Nets, Machine Learning varieties, etc. Each different type produces a different Fit. This applies both to time series and cross-sectional types of models. The complexity increases as more predictors are added and as nonlinear prediction methodologies are incorporated. We need methods for selecting good models that will meet the above three objectives of explanation, prediction, and simplicity. We also need both general principles and quantitative measures to assess and to compare how well different models meet the above three criteria.

> Ex: (Estimated) RENT = 143.6693 + .3875 * AREA + 89.9290 * BATHROOMS is a statistical theory (or model) [2] for explaining/predicting the data, which are a cross-section of monthly rents of Austin apartments. For this to be a good theory, it must do a good job of explaining rents of apartments that you know about and predicting rents for apartments that you do not yet know about. That is, the model should work well for both the sample and the population. How well this theory explains can be evaluated by comparing its FITs with the actual rents in the sample data (by R-square and the residual standard deviation, for example). The equation also stands ready to predict rents of apartments not in the sample, given knowledge of their areas and numbers of bathrooms. Using the OLS fitting criterion, there is no better model (FIT) using these two predictor variables and these data. As data on additional apartments become available, we can compare the rent predictions made by the model with the reality of the actual rents of the additional apartments.[3] There may be better-fitting models that use more than two predictors or that use nonlinear functions of the variables. Such models are more complicated.

A statistical model has built-in measures for evaluating how well it explains the sample data: The closer the Residuals are to zero, and the closer the Fits are to the Actuals, the better. The residual

---

[1] The term *theory* is used here in the scientific sense. The scientific meaning is neutral and has none of the pejorative popular sense, in which "theory" connotes a "wild or half-baked guess" or an "impractical ivory-tower absurdity." A good theory explains well and predicts well the data of the real world; a bad theory does not.

[2] The theory includes the model assumptions LHIN, in addition to the calculation formula.

[3] Ideally, we would like to know *in advance* how well the model will predict new data. If this regression model satisfies its LHIN assumptions and sound sampling protocols have been followed (so that the sample is representative of the population from which it came) and if sound model-building protocols have been followed (so that the model is not overfit, etc.), then additional population data should be well-predicted by the model. Moreover, given these conditions, the existing data provide the means to estimate how well additional data will be predicted (via R-square, *s*, confidence intervals, etc.).

standard deviation (RMSE) and R-square quantify these measures. However, sometimes a better fit must be purchased at the cost of increased complexity. In such a case, is the improvement in fit worth the loss of simplicity, given that all regression specifications (LHIN) are satisfied? Simplicity in a model is called **parsimony**. Statistics offers some tools to help build models that explain and predict well and are parsimonious. For regression models, the major tool is **stepwise regression**, which is a strategy for statistical model building one predictor at a time. But I strongly advise you that model-building software like stepwise regression should be viewed as a tool to *inform* your judgment. I caution you against letting the computer *replace* your judgment. At times you may need to over-rule the software to get sensible results.

Statistical Model Building with Multiple Regression

Statistical model building with multiple regression begins with understanding your problem: Do you need to explain or to forecast? Which variable do you want to explain or to forecast? That is your response variable that provides the supervision for your project. You will need access to empirical data for this variable.

Next, make a list of possible relevant predictors, including nonlinear functions of predictors for which you have data. How to choose predictors?

- Past experience, if any, with predictors that have worked well in similar models
- Any underlying theory that applies to your situation. The theory may suggest factors that should be important.
- Your inspection of plots and graphs of your data. This is especially important with time series data, as many time series have common features like trends, seasonality, increasing variability that are easy to see in plots.

You will need to obtain or calculate the values to use for these predictors. For example, a linear trend may be represented by the sequence 1, 2, 3, … . A quadratic trend may be represented by the sequence 1, 4, 9, 16, … . Long-term interest rates may be represented by the yield on the 25-year U.S. Treasury bond.

The next step is to identify possible *model forms*. That is, will the variables be analyzed in original scale, or in some transformed scale, such as squares or logarithms, as changes between time periods, as percentage changes, etc.? [4]

Finally, **variable selection** culls the list of predictors with three principles in mind:

- **Explanatory power**. A good model should explain as much of the Y-variability as possible. Explanatory power can be assessed by R-square and the root mean square error (RMSE). The higher the R-square and the smaller the RMSE, the better. One achieves higher R-square and (up to a point) lower RMSE by adding additional predictors.[5] This principle conflicts with parsimony.

[4] Often, you will pick a form and go on to the next step (variable selection) to see if the form works. If it does not work, try another form. How to pick a form? Be guided by your experience of what has worked in similar cases. If you have no idea, start with original scale and simple forms before trying more complicated ones.

[5] R-square never decreases when an additional predictor is added. RMSE will decrease when an additional predictor is added to the model, as long as the |T-value| of the new predictor is above 1 in magnitude. If the magnitude of the new predictor's T-value is below 1, RMSE will increase.

- **Parsimony**. A good model should be as simple as possible.[6] Parsimony can be assessed by the number of predictors in the model. The fewer the predictors, the better. One way to quantify parsimony is to calculate the data-to-predictor ratio, n / p, where n is the number of rows of data and p is the number of predictors. The higher this ratio, the better. A common rule-of-thumb is that $n / p > 15$ is good; $n / p < 5$ is bad; and $5 < n / p < 15$ is marginal or suspect.[7] One achieves parsimony by removing predictors from the model. This principle conflicts with explanatory power, so a trade-off is necessary.
- **Validity**. A model is valid if it satisfies its model assumptions. For example, if our final model is a regression model, it should satisfy the LHIN specifications of regression in order for us to rely on it.

The principles of explanatory power and parsimony must be balanced against each other. Ultimately, it is a judgment call as to whether to prefer more explanatory power or more parsimony, based on the requirements of the problem and one's personal preferences. But statistical considerations can help inform that judgment.

Comments on Overfitting Models:

- The model fitting process (called least squares – or ordinary least squares (OLS)) for regression tries to get the model as close to the sample data as the form of the model will permit. But the sample data include not only features that the sample data share with other population data, but also idiosyncratic features peculiar to the particular sample and that are absent in other population data. We want the model to capture the features of the sample data that are shared with other population data, so that we can apply the model to the rest of the population. We want the model to capture the *signal* in the sample and ignore the *noise*. However, if the model overfits (gets too close to) the sample data, then the model will capture the idiosyncratic features of the sample data as well as the features shared with the population. Thus an overfit model will not predict other population data very well because the predictions of an overfit model will include the idiosyncrasies of the sample data.
- The danger of overfitting increases as the model consumes more degrees of freedom -- one degree of freedom per predictor. In fact, in a model with the number of predictors greater than or equal to n - 1, the model consumes all degrees of freedom, and the model fits "perfectly"! In such a model, R-square = 1 and RMSE = 0. Moreover, it does not matter which variables are used as predictors! If you were to fill n-1 predictors with completely random numbers, then R-square = 1 and RMSE = 0. In spite of perfect R-square and RMSE, this would not be a good model. Although that model would "explain" those data "perfectly", the model would fail utterly to explain any new data.
- To avoid the problems of overfitting, it is desirable to keep the ratio of data n to predictors p high. Studies indicate that a data-to-predictors ratio of $n/p > 15$ or more is generally safe.[8] A ratio of $n/p < 5$ usually causes problems. Between 5 and 15 is a judgment call. I advise beginning modelers to play it safe and use higher ratios.

[6] Simplicity is desirable not only for esthetic reasons and for ease of use, but also because of the dangers of overfitting in complex models.
[7] Some authorities tolerate a lower upper limit ratio for this rule-of-thumb, like 10 instead of 15.
[8] Some authorities tolerate a lower upper limit ratio for this rule-of-thumb, like 10 instead of 15.

- **Cross-validation.** To diagnose potential overfitting, the cross-validation method is available. Cross-validation is a clever idea to assess how well the model will apply to new data without actually getting new data. In cross-validation, the sample data are split into two subsets, called the training sample and holdout sample. The idea is to develop the model on the training sample, then see how well the model works on the holdout sample. If the model works well on the training sample, but not on the holdout sample, the reason is probably that the model has been overfit to the training sample. This is indicated by a much lower R-square and/or larger RMSE in the holdout sample than in the training sample.

  How to do cross-validation:
  - First, split the data into a training set and a hold-out sample.[9]
  - Second, build the model with the training data.
  - Third, calculate the "FITs" for the holdout sample by plugging the holdout X's into the training equation.
  - Fourth, calculate an "R-square" and "RMSE" for these holdout FITS.[10]
  - Finally, compare the R-square and RMSE from the training sample to the R-square and RMSE for the holdout sample. If R-square does not decline significantly between the training and holdout samples, and if RMSE does not increase significantly between the training and holdout samples, then there has been little overfitting and the model should generalize well to other data.

  If cross-validation indicates little problem with overfitting, then we can combine the training sample with the holdout sample and build the final model on the combined and complete data.
- **The jackknife.** Cross-validation is clearly not feasible if a split of the sample into two parts would leave the training sample too small for meaningful modeling. If there would be too few data for a reasonable split between training and holdout samples, the jackknife method of cross-validation can often be employed. In the jackknife, the training sample consists of all but one of the data; the holdout sample is a single datum. After developing the model on the training sample, the X's for the holdout datum are plugged into the training equation to yield a "FIT", but no "R-square" or "RMSE" is calculated.[11] Then the procedure is repeated with a different datum as the holdout sample (but still only one datum!), and the remaining *n*-1 data for the training sample. The model is refit to the new training sample and applied to the new holdout datum. This procedure is repeated until each datum has served as the singleton

[9] It is often appropriate to split the data randomly into two equal parts, or to make the training sample larger than the holdout sample in perhaps a 2-to-1 training to holdout ratio.

[10] [This is a technical note.] How to calculate an R-square and RMSE for the holdout sample is a good question! There is a problem in that the regression equation for the training sample is not the same as the regression equation *would be* for the holdout sample. So the "FITs" you get from plugging the holdout X's into the training regression equation are not the same as the real FITs you would get from plugging the holdout X's into the holdout regression equation. Several proposals have been advanced to calculate an "R-square" and an "RMSE" for the holdout sample. Here is one simple way to do this for R-square: "R-square" = CORR(holdout ACTUAL, holdout FIT)^2, where you get the "holdout FIT" by plugging the holdout X-values into the training regression equation. For "RMSE", we can calculate "RESIDUAL" = holdout ACTUAL – holdout FIT. Then "RMSE" = StDev("RESIDUALS") (remember to divide by *n - number of predictors* - 1). However, these are not the only reasonable definitions of R-square and RMSE for cross-validation. With all definitions, some anticipated properties do not hold. For example, the mean of the "RESIDUALS" as defined in this footnote is not exactly zero, and the "R-square" defined here is not exactly the proportion of holdout Y-variability explained by the holdout X's, but this R-square does lie between 0 and 1, whereas other R-square definitions may not.

[11] R-square and RMSE cannot be computed with only one datum.

holdout. By rotating the singleton holdout sample among all sample data individually, a collection of *n* "FITs" and *n* "RESIDUALS" is ultimately obtained and then analyzed for deterioration in "R-square" and "RMSE" as in cross-validation.[12]

I now discuss some variable selection techniques.

A. All Possible Regressions

If there are k possible predictors, then there are 2^k possible regression models that can be formed.[13] One approach to variable selection is to run all 2^k possible regressions and select the optimal model from the 2^k outputs. This can be a very large number of models, even for relatively small k. Therefore all possible regressions is not a practical strategy, except for small k.[14]

B. Stepwise Regression

Stepwise regression is a strategy for model building that examines a fraction of all possible regressions to come up with a model that is pretty good, even if not optimal. Stepwise regression examines one predictor at a time, deciding to put it into the model or leave it out. There are several main varieties of stepwise regression. In PROC REG, SAS has separate options for backward, forward, and regular stepwise regression strategies, among others. Out-of-the-box Excel does not have stepwise regression. I will next discuss these three options in general terms.

Forward stepwise regression begins with the no-predictor model. For the first step, add the best predictor and run the regression with this predictor. For the second step, add the best predictor among those that remain and rerun the regression. At each step, add the best of the remaining predictors. The "best" predictor at any step is the predictor that increases the R-square by the most; it is not necessarily the predictor with the next best correlation with *Y*. The explanatory power of the model rises with each step. This is a "greedy" algorithm, trying to add as much as possible to R-square at every step. SAS has an option to specify a cut-off for stopping the forward selection: **SLENTRY = 0.05** (**S**ignificance **L**evel for **ENTRY**) says to stop adding predictors when each remaining candidate predictor would have p-value > 0.05 if added to the model. To be eligible for entry into the model, a predictor must have a p-value < 0.05, or whatever value is

---

[12] This has been a thumbnail sketch of cross-validation applicable to cross-sectional data. Proper cross-validation, in general, is trickier for time series than for cross-sectional data because of autocorrelation issues. The period-to-period linkages are broken whenever a period is removed for a hold-out sample.

[13] Each predictor can be included or excluded from the model. So there are 2x2x2x...x2 = 2^k possible models, based on including or excluding each predictor. This includes the model with no predictors, for which the FIT is the mean.

[14] *All possible regressions* also runs a danger of overfitting – not necessarily from a low data-to-predictors ratio, but from a low data-to-possible-models ratio. The more models we examine, the more likely it is that we will find among those models one that, by chance variation alone, happens to look like the sample data. We might call this the "Rorschach" model-building technique (after the psychology test that asks you to tell a story about a random inkblot): Given a random ink smear on paper, you may find a vaguely similar pattern by looking through a book of geometrical shapes. The bigger your book of shapes, the more likely you are to find one that resembles the ink smear. The shape may fit this ink smear well, but you do not expect it to generalize to the population of other ink smears. Similarly, if you look through enough statistical models, you may by chance find one that seems to have some explanatory power for your data, but the match will be misleading – the model will not fit new data.

specified in this option. This option is used in the model statement. For example, **MODEL Y = X1 X2 X3 X4 / SELECTION = FORWARD SLENTRY = 0.05;**

Backward stepwise regression begins with the model that includes all predictors. For the first step, remove the worst predictor and run the regression without this predictor. For the second step, remove the worst predictor among those that remain and rerun the regression. At each step, remove the worst of the remaining predictors. The "worst" predictor at any step is the predictor that decreases the R-square by the least. This will be the predictor with the smallest magnitude T-value (equivalently, the largest p-value) in the current model. The explanatory power of the model decreases with each step. This, too, is a "greedy" algorithm, trying to hang on to as much R-square as possible at every step. SAS has an option to specify a cut-off for stopping the backward elimination: **SLSTAY = 0.05** (**S**ignificance **L**evel to **STAY**) says to stop eliminating predictors when all remaining model predictors have p-values < 0.05. To be eligible for elimination from the model, a predictor must have a p-value > 0.05, or whatever value is specified in this option. This option is used in the model statement. For example, **MODEL Y = X1 X2 X3 X4 / SELECTION = BACKWARD SLSTAY = 0.05;**

Regular stepwise regression begins in forward mode, but at each step, it looks both forward and backward. That is, having added a predictor, regular stepwise regression then examines the current predictors already in the model to see if any have become superfluous by the addition. For example, a |T-value| of a previously added predictor may have fallen below an acceptable level by virtue of adding the new predictor.[15] If so, the earlier predictor is removed. At the next step, regular stepwise regression adds the best of the remaining possible predictors, then reviews all previously added predictors for possible removal. You may use the SAS options **SLENTRY** and **SLSTAY** to control how the forward and backward steps of regular stepwise regression work. These options work as described in the preceding sections of this topic note on forward and backward regression. For example, **MODEL Y = X1 X2 X3 X4 / SELECTION = STEPWISE SLENTRY=0.05 SLSTAY = 0.10;** says to stop adding predictors when all of the remaining candidates for entry have p-values > 0.05 and to stop removing predictors when all of the model predictors have p-values < 0.10. Notice that for consistency, you must specify **SLENTRY ≤ SLSTAY**.[16]

Miscellaneous General Remarks on Stepwise Regression

- No variable selection method is guaranteed to arrive at the optimal model, except for *all possible regressions*.
- Backward stepwise regression does not necessarily stop at the same final model as forward stepwise regression, nor as regular stepwise regression, even if they all are using the same criteria for explanatory power.
- In general, backward stepwise regression tends to favor retaining predictors in the model, hence favors the explanatory power criterion.

[15] This could happen if the added and previous predictors share significant overlapping explanatory power.

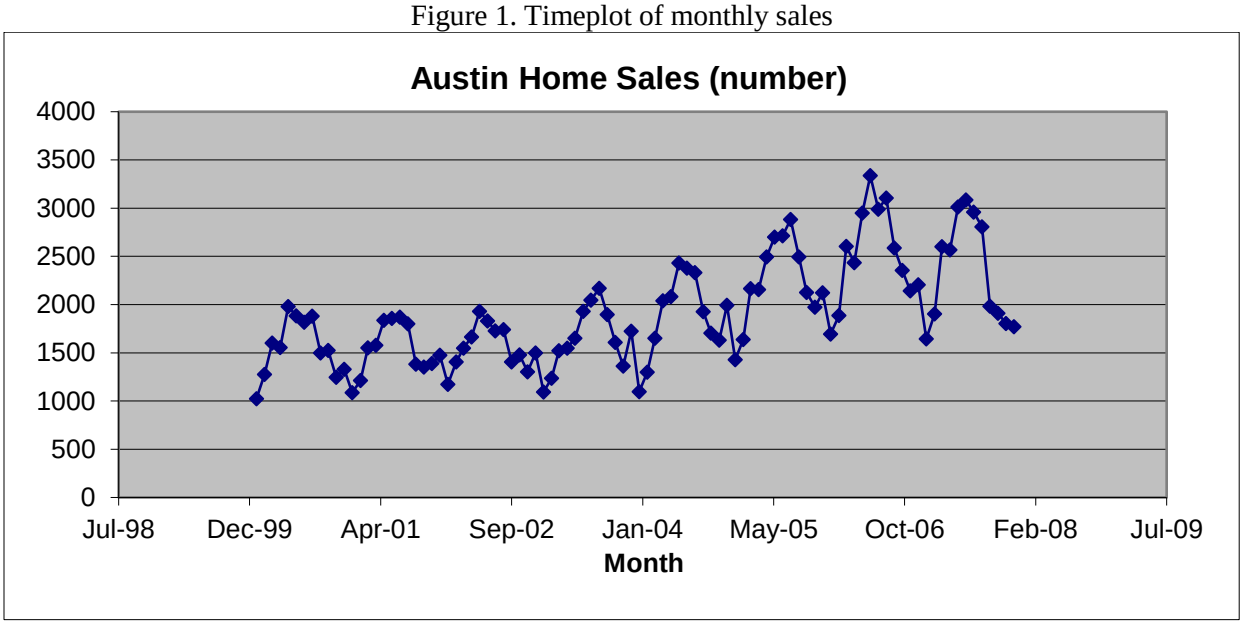
[16] Otherwise an infinite loop could be generated. E.g., if SLENTRY=0.10 SLSTAY = 0.05 and X is in the model with p-value = 0.08, then X will be removed. But then X is eligible to be entered. If it is, then it is eligible to be removed. And so on. Most software will not permit you to set SLENTRY > SLSTAY.

- In general, forward stepwise regression tends to favor a more parsimonious model than backward mode.
- Cut-off thresholds for adding and removing predictors are often relaxed in stepwise regression. For example, the T-threshold may be lowered from ±1.96 to ±1.645 or even ±1.282. Or equivalently, p-value criteria for adding or removing predictors may be relaxed from 0.05 to 0.10 or even 0.20.
- If you want high R-square and you think that most of the available predictors are important and should be in the final model, a good choice is backward stepwise regression with a loose standard for retaining predictors (low T or high p-value required to remove).
- If you want parsimony and you don't know which of the many available predictors are important, a good choice is forward stepwise regression with a strict standard for adding predictors (high T or low p-value required to enter).
- If you want to balance parsimony and explanatory power, a good choice is regular stepwise regression with the p-values set for entry or removal depending on how you want to balance entry and removal.
- Backward, forward, and regular stepwise regression are only a few among the many flavors of stepwise regression that the fertile brains of statisticians have invented.

1. An Example: Austin Home Sales

I will illustrate regular stepwise regression for one approach to forecasting in the case of monthly Austin home sales. The objective is to develop a statistical model to forecast the number of homes sold per month. Figure 1 shows the timeplot of monthly sales. Several features that are common in time series are evident from even cursory inspection of the timeplot: increasing trend, regular seasonal ups and downs, increasing variability. A successful forecasting model will have to deal with these features.

Figure 1. Timeplot of monthly sales



In this illustration, I will take a direct approach to forecasting sales with **deterministic** predictors. Plausible alternative approaches include the use of **stochastic** predictors like the lags of sales (plausibly lags 1 through 12), as well as switching the response variable to changes in sales or percent changes in sales, or the log of sales.

A **deterministic predictor** is a predictor variable whose future values are all known now, like a linear trend 1, 2, 3, … . A **stochastic predictor** is a predictor variable whose future values are not all known now, like the lag of sales – the next value of lag sales is known now (it is just the current value of sales), but no values after that are known now.

Predictors should represent the main features visible in the timeplot. These include the increasing long term trend, the regular up-and-down seasonal variation, and (arguably) the increasing variability of sales. Plausible variables to represent the long term trend are the linear trend (T1 = 1, 2, 3, 4, …), the quadratic trend (T2 = 1, 4, 9, 16, …), and perhaps cubic trend (T3 = 1, 8, 27, 64, …). These are all deterministic. Plausible predictors to represent seasonal (monthly) variation are indicators for each month (M1 = 1 if the month is Jan, M1 =0 if the month is not Jan, …, similarly for M2 – M12). These are also deterministic.

I will illustrate regular stepwise regression to select the important predictors of Sales from among these 15 predictors: T1, T2, T3, M1 – M12.

I ran the following SAS code:

```
proc reg data=HOMES;
  model sales = T1 T2 T3 M1-M12 / selection=stepwise slentry=0.05
slstay=0.10;
RUN;
```

At the end of the output is a summary of the step-by-step selection process, which I have extracted and displayed immediately below.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	T2		1	0.4443	0.4443	394.410	75.14	<.0001
2	M1		2	0.1022	0.5465	306.920	20.97	<.0001
3	M6		3	0.0595	0.6060	256.847	13.89	0.0003
4	M8		4	0.0535	0.6595	212.044	14.29	0.0003
5	M7		5	0.0627	0.7222	159.133	20.33	<.0001
6	M5		6	0.0661	0.7883	103.318	27.77	<.0001
7	T3		7	0.0231	0.8113	85.1311	10.76	0.0015
8	T1		8	0.0435	0.8549	49.0151	26.11	<.0001
9	M2		9	0.0227	0.8776	31.1173	15.98	0.0001
10	M11		10	0.0171	0.8947	18.1475	13.81	0.0004

Regular stepwise regression begins in forward mode. The summary shows that the selection procedure remained in forward mode for all steps. No predictors were removed. At the last step, there were 10 predictors that had been added to the model. All three of the trend variables had been added and 7 of the monthly indicators, representing seasonal effects. At step one, the quadratic trend indicator T2 was the single most important predictor and the R-square jumped to 0.4443. At step 2, the January indicator entered and added a further 0.1022 in R-square, bringing the total R-square after two steps to 0.5465. Finally, at step 10, the November indicator entered, adding only 0.0171 in R-square, and bringing the final total to 0.8947. The p-value of each predictor at the time it entered is shown in the extreme right-hand column of the summary table. All are less than the 0.05 level established for entry by the `slentry=0.05` option. However, none of the remaining five predictors, if added to the model at this point, would have a p-value < 0.05.

The SAS output includes the model at each step of the selection procedure. Each predictor in every model has a p-value that is less than the 0.10 level established in the `slstay=0.10` option. This means that no predictor was eligible for removal from the developing model. Hence, none was removed at any step. The model continued in forward selection mode.

The final model is:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	24338026	2433803	72.24	<.0001
Error	85	2863864	33693		
Corrected Total	95	27201890			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1770.64289	82.82863	15396942	456.98	<.0001
T1	-42.72241	6.99555	1256613	37.30	<.0001
T2	1.28867	0.16710	2003885	59.48	<.0001
T3	-0.00829	0.00113	1802282	53.49	<.0001
M1	-543.57927	71.56780	1943676	57.69	<.0001
M2	-344.51456	71.42948	783779	23.26	<.0001
M5	402.53930	71.16350	1078042	32.00	<.0001
M6	537.46789	71.11918	1924266	57.11	<.0001
M7	463.33048	71.09563	1430966	42.47	<.0001
M8	484.55180	71.09278	1565175	46.45	<.0001
M11	-264.63473	71.21487	465249	13.81	0.0004

The R-square is high (0.8947) and the RMSE = $\sqrt{33693}$ = 183.56. I judge that the explanatory power is good. Parsimony has $n / p = 96 / 10 = 9.6 < 10$. I might sacrifice the last predictor or two in order to boost the n / p ratio. Then the first two general model-building criteria would be comfortably satisfied. However, stepwise regression does not guarantee production of a valid model. This model still needs to be checked for compliance with regression specifications (LHIN) – omitted here.