

Introduction to Building a Linear Regression Model

Leslie A. Christensen

The Goodyear Tire & Rubber Company, Akron Ohio

Abstract

This paper will explain the steps necessary to build a linear regression model using the SAS System®. The process will start with testing the assumptions required for linear modeling and end with testing the fit of a linear model. This paper is intended for analysts who have limited exposure to building linear models. This paper uses the REG, GLM, CORR, UNIVARIATE, and PLOT procedures.

Topics

The following topics will be covered in this paper:

1. assumptions regarding linear regression
2. examining data prior to modeling
3. creating the model
4. testing for assumption validation
5. writing the equation
6. testing for multicollinearity
7. testing for auto correlation
8. testing for effects of outliers
9. testing the fit
- 10 modeling without code.

Assumptions

A linear model has the form $Y = b_0 + b_1X + \varepsilon$. The constant b_0 is called the intercept and the coefficient b_1 is the parameter estimate for the variable X . The ε is the error term. ε is the residual that can not be explained by the variables in the model. Most of the assumptions and diagnostics of linear regression focus on the assumptions of ε . The following assumptions must hold when building a linear regression model.

1. The dependent variable must be continuous. If you are trying to predict a categorical variable, linear regression is not the correct method. You can investigate discrim, logistic, or some other categorical procedure.
2. The data you are modeling meets the "iid" criterion. That means the error terms, ε , are:
 - a. independent from one another and
 - b. identically distributed.

If assumption 2a does not hold, you need to investigate time series or some other type of method. If assumption 2b does not hold, you need to investigate methods that do not assume

normality such as non-parametric procedures.

3. The error term is normally distributed with a mean of zero and a standard deviation of σ^2 , $N(0, \sigma^2)$.

Although not an actual assumption of linear regression, it is good practice to ensure the data you are modeling came from a random sample or some other sampling frame that will be valid for the conclusions you wish to make based on your model.

Example Dataset

We will use the SASUSER.HOUSES dataset that is provided with the SAS System for PCs v6.10. The dataset has the following variables: PRICE, BATHS, BEDROOMS, SQFEET and STYLE. STYLE is a categorical variable with four levels. We will model PRICE.

Initial Examination Prior to Modeling

Before you begin modeling, it is recommended that you plot your data. By examining these initial plots, you can quickly assess whether the data have linear relationships or interactions are present.

The code below will produce three plots.

```
PROC PLOT DATA=HOUSES;
  PLOT PRICE*(BATHS
              BEDROOMS SQFEET);
RUN;
```

An X variable (e.g. SQFEET) that has a linear relationship with Y (PRICE) will produce a plot that resembles a straight line. (Note Figure 1.) Here are some exceptions you may come across in your own modeling.

If your data look like Figure 2, consider transforming the X variable in your modeling to $\log_{10}X$ or \sqrt{X} .

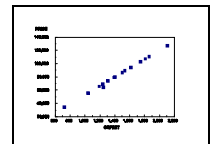


Figure 1

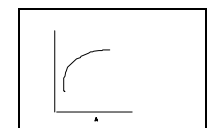


Figure 2

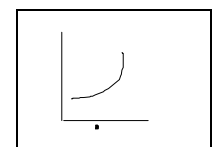


Figure 3



Figure 4

If your data look like Figure 3, consider transforming the X variable in your modeling to X^2 or $\exp(X)$.

If your data look like Figure 4, consider transforming the X variable in your modeling to $1/X$ or $\exp(-X)$

This SAS code can be used to visually inspect for interactions between two variables.

```
PROC PLOT DATA=HOUSES;
  PLOT BATHS*BEDROOMS;
RUN;
```

Additionally, running correlations among the independent variables is helpful. These correlations will help prevent multicollinearity problems later.

```
PROC CORR DATA=HOUSES;
  VAR BATHS BEDROOMS SQFEET;
RUN;
```

In our example, the output of the correlation analysis will contain the following.

Correlation Analysis			
3 'VAR' Variables: BATHS BEDROOMS SQFEET			
	BATHS	BEDROOMS	SQFEET
BATHS	1.00000	0.82492	0.71553
BEDROOMS	0.0002	1.00000	0.75714
SQFEET	0.0027	0.0011	1.00000

In the above example, the correlation coefficients are in bold. The correlation of 0.82492 between BATHS and BEDROOMS indicates that these variables are highly correlated. A decision should be made to include only one of them in the model. You might also argue that 0.71553 is high. For our example we will keep it in the model.

Creating the Model

As you read, learn and become experienced with linear regression you will find there is no *one* correct way to build a model. The method suggested here is to help you better understand the decisions required without having to learn a lot of SAS programming.

The REG procedure can be used to build and test the assumptions of the data we propose to model. However, PROC REG has some limitations as to how the variables in your model must be set up. REG can not handle interactions such as

BEDROOMS*SQFEET or categorical variables with more than two levels. As such you need to use a DATA step to manipulate your variables. Let's say you have two continuous variables (BEDROOMS and SQFEET) and a categorical variable with four levels (STYLE) and you want all of the variables plus an interaction term in the first pass of the model. You would have to have a DATA step to prepare your data as such:

```
DATA HOUSES;
  SET HOUSES;
  BEDSQFT = BEDROOMS*SQFEET;
  IF STYLE='CONDO' THEN DO;
    S1=0; S2=0; S3=0; END;
  ELSE IF STYLE='RANCH' THEN DO;
    S1=1; S2=0; S3=0; END;
  ELSE IF STYLE='SPLIT' THEN DO;
    S1=0; S2=1; S3=0; END;
  ELSE DO;
    S1=0; S2=0; S3=1; END;
RUN;
```

When creating a categorical term in your model, you will need to create dummy variables for "the number of levels minus 1". That is, if you have three levels, you will need to create two dummy variables and so on.

Once the variables are correctly prepared for REG we can run the procedure to get an initial look at our model.

```
PROC REG DATA=HOUSES;
  MODEL PRICE = BEDROOMS SQFEET S1 S2
              S3 BEDSQFT ;
RUN;
```

The GLM procedure can also be used to create a linear regression model. The GLM procedure is the safer procedure to use for your final modeling because it does not assume your data are balanced. That is with respect to categorical variables, it does not assume you have equal sample sizes for each level of each category. GLM also allows you to write interaction terms and categorical variables with more than two levels directly into the MODEL statement. (These categorical variables can even be character variables.) Thus using GLM eliminates some DATA step programming.

Unfortunately, the SAS system does not provide the same statistics in REG and GLM. Thus you may want to test some basic assumptions with REG and then move on to using GLM for final modeling. Using GLM we can run the model as:

```

PROC GLM DATA=HOUSES;
  CLASS STYLE;
  MODEL PRICE = BEDROOMS SQFEET STYLE
              BEDROOMS*SQFEET;
RUN;

```

The output from this initial modeling attempt will contain the following statistics:

```

General Linear Models Procedure
Dependent Variable: PRICE   Asking price
Sum of Mean F Pr>
Source DF Squares Square Value F
Model 6 7895558723 1315926453 410.66 0.0001

Error 8 25635276 3204409
Corrected
Total 14 7921194000
R-Square C.V. Root MSE PRICE Mean
0.996764 2.16 1790.08 82720.00

Source DF Type III Mean F Pr>
SS Square Value F
BEDROOMS 1 245191 245191 0.08 0.7891
SQFEET 1 823358866 823358866 256.95 0.0001
STYLE 3 5982918 1994306 0.62 0.6202
BEDROOMS*
SQFEET 1 712995 712995 0.22 0.6497

```

When building a model you may wonder which statistic tells whether the model is good. There is no *one* correct answer. Here are some approaches of statistics that are found in both REG and GLM.

R-square and Adj-Rsq

You want these numbers to be as high as possible. If your model has a lot of variables, use Adj-Rsq because a model with more variables will have a higher R-square than a similar model with fewer variables. Adj-Rsq takes the number of variables in your model into account. An R-square or 0.7 or higher is generally accepted as good.

Root MSE

You want this number to be small compared to other models. The value of Root MSE will be dependent on the values of the Y variable you are modeling. Thus, you can only compare Root MSE against other models that are modeling the same dependent variable.

Type III SS Pr>F

As a guideline, you want the value for each of the variables in your model to have a Type III SS p-value of 0.05 or less. This is a judgement call. If you have a p-value greater than .05 and are willing to accept a lesser confidence level, then you can use the model. Do not substitute Type I or Type II SS for Type III SS. They are different statistics and could lead to incorrect conclusions in some cases.

Other approaches to finding good models are having a small PRESS statistic (found in REG as Predicted Resid SS (Press)) or having a CP statistic of $p-1$ where p is the number of parameters in your model. CP can also be found using PROC REG.

When building a model only eliminate one term, variable or interaction, at a time. From examining the GLM printout, we will drop the interaction term of BEDROOMS*SQFEET as the Type III SS indicates it is not significant to the model. If an interaction term is significant to a model, its individual components are generally left in the model as well. It is also generally accepted to leave an intercept in your model unless you have a good reason for eliminating it. We will use BEDROOMS, S1, S2 and S3 in our final model.

Some of the approaches for choosing the best model listed above are available in SELECTION= options of REG and SELECTION= options of GLM.

For example:

```

PROC REG;
  MODEL PRICE = BEDROOMS SQFEET
              S1 S2 S3 / SELECTION = ADJRSQ;

```

will iteratively run models until the model with the highest adjusted R-square is found. Consult the SAS/STAT User's Guide for details.

Test of Assumptions

We will validate the "iid" assumption of linear regression by examining the residuals of our final model. Specifically, we will use diagnostic statistics from REG as well as create an output dataset of residual values for PROC UNIVARIATE to test. The following SAS code will do this for us.

```

PROC REG DATA=HOUSES;
  MODEL PRICE = BEDROOMS S1 S2 S3 /
              DW SPEC ;
  OUTPUT OUT=RESIDS R=RES;
RUN;
PROC UNIVARIATE DATA=RESIDS
  NORMAL PLOT;
  VAR RES;
RUN;

```

```

Dependent Variable: PRICE
Test of First and Second Moment Specification
DF: 8 Chisq Value: 7.0152 Prob>Chisq:0.5350

Durbin-Watson D          1.334
(For Number of Obs.)    15
1st Order Autocorrelation 0.197

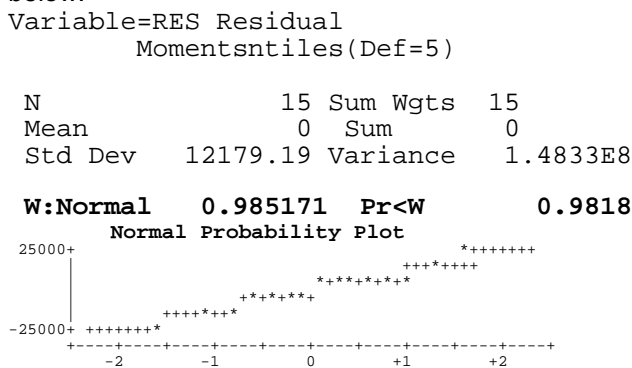
```

The bottom of the REG printout will have a statistic that jointly tests for heteroscedasticity (not identical distributions of error terms) and dependence of error terms. This test is run by using the SPEC option in the REG model statement. As the SPEC test is testing for the opposite of what you hope to conclude, a non-significant p-value indicates that the error variances are not identical and the error terms are not dependent. Thus the Prob>Chisq of 0.5350 > 0.05 lets us conclude our error terms are independent and identically distributed.

The Durbin-Watson statistic is calculated by using the DW option in REG. The Durbin-Watson statistic test for first order correlation of error terms. The Durbin Watson statistic ranges from 0 to 4.0. Generally a D-W statistic of 2.0 indicates the data are independent. A small (less than 1.60) D-W indicates positive first order correlation and a large D-W indicates negative first order correlation.

However, the D-W statistic is not valid with small sample sizes. If your data set has more than $p \times 10$ observations then you can consider it large. If your data set has less than $p \times 5$ observations than you can consider it small. Our data set has 15 observations, n , and 4 parameters, p . As n/p equals $15/4=3.75$, we will not rely on the D-W test.

The assumption that the error terms are normal can be checked in a two step process using REG and UNIVARIATE. Part of the UNIVARIATE output is below.



The W:Normal statistic is called the Shapiro-Wilks statistic. It tests that the error terms come from a normal distribution. If the test's p-value is less than significant (eg. 0.05) then the errors are not from a

normal distribution. Thus, $Pr < W$ or 0.9818 indicates the error terms are normally distributed.

The normal probability plot plots the errors as asterisks, *, against a normal distribution,+. If the *'s look linear and fall in line with the +'s then the errors are normally distributed. The outcome of the W:Normal statistic will coincide with the outcome of the normal probability plot. These tests will work with small sample sizes.

Writing the Equation

The final output of a model is an equation. To write the equation we need to know the parameter estimates for each term. A term can be a variable or interaction. In REG the parameter estimates print out and are called Parameter Estimate under a section of the printout with the same name.

Parameter Estimates					
Variable	DF	Parameter Estimate	Std. Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	53856	12758.2	4.221	0.0018
BEDROOMS	1	16530	3828.6	4.317	0.0015
S1	1	-22473	10368.1	-2.167	0.0554
S2	1	-19952	11010.9	-1.812	0.1001
S3	1	-19620	10234.7	-1.917	0.0842

We could argue that we should remove the STYLE dummy variables. For the sake of this example, however, we will keep them in the model. The parameter estimates from REG are used to write the following equations. There is an equation for each combination of the dummy variables.

$$\begin{aligned} \text{PRICE}_{\text{CONDO}} &= 53856 + 16530(\text{BEDROOMS}) \\ \text{PRICE}_{\text{RANCH}} &= 53856 + 16530(\text{BEDROOMS}) - 22473 \\ &= 31383 + 16530(\text{BEDROOMS}) \\ \text{PRICE}_{\text{SPLIT}} &= 33904 + 16530(\text{BEDROOMS}) \\ \text{PRICE}_{\text{2STORY}} &= 34236 + 16530(\text{BEDROOMS}) \end{aligned}$$

If using GLM, you need to add the SOLUTION option to the model statement to print out the parameter estimates.

```
PROC GLM DATA=HOUSES;
  CLASS STYLE;
  MODEL PRICE=BEDROOMS
  STYLE/SOLUTION;
RUN;
```

Parameter	Estimate	T for H0: Parameter=0	Pr > T
INTERCEPT	34235.8	2.52	0.0301
BEDROOMS	16529.7	4.32	0.0015
STYLE			
CONDO	19619.9	1.92	0.0842
RANCH	-2852.7	-0.27	0.7931
SPLIT	-331.7	-0.03	0.9767
TWOESTORY	0.0	.	.

NOTE: The X'X matrix has been found to be

singular and a generalized inverse was used to solve the normal equations. Estimates followed by the letter 'B' are biased, and are not unique estimators of the parameters.

When using GLM with categorical variables, you will always get the NOTE: that the solution is not unique. This is because GLM creates a separate dummy variable for each level of each categorical variable. (Remember in the REG example, we created n-1 dummy variables.) It is due to this difference in approaches that the GLM categorical variable parameter estimates will always be biased with the intercept estimate. The GLM parameter estimates can be used to write equations but it is understood that other parameter estimates exist that would also fit the equation.

Writing the equations using GLM output is done the same way as when using REG output. The equation for the price of a condo is shown below.

$$\text{PRICE}_{\text{CONDO}} = 34236 + 16530(\text{BEDROOMS}) + 19620 \\ = 53856 + 16530(\text{BEDROOMS})$$

Testing for Multicollinearity

Multicollinearity is when your independent, X, variables are correlated. A statistic called the Variance Inflation Factor, VIF, can be used to test for multicollinearity. A cut off of 10 can be used to test if a regression function is unstable. If VIF > 10 then you should search for causes of multicollinearity.

If multicollinearity exists, you can try: changing the model, eg. drop a term; transform a variable; or use Ridge Regression (consult a text, e.g., [SAS System for Regression](#)).

To check the VIF statistic for each variable you can use REG with the VIF option in the model statement. We'll use our original model as an example.

```
PROC REG DATA=HOUSES;
  MODEL PRICE = BEDROOMS SQFEET S1 S2
    S3 BEDSQFT / VIF;
```

```
RUN;
```

Variable	Parameter Estimates
	Variance Inflation
INTERCEP	0.00000000
BEDROOMS	21.54569800
SQFEET	8.74473195
S1	2.14712669
S2	1.85694384
S3	2.01817464

BEDSQFT 39.07446066

The VIF of 39 for the interaction term BEDSQFT would suggest that BEDSQFT is correlated to another variable in the model. If you rerun the model with this term removed, we'll see that the VIF statistics change and all are now in acceptable limits.

Variable	Parameter Estimates
	Variance Inflation
INTERCEP	0.00000000
BEDROOMS	3.04238083
SQFEET	3.67088471
S1	2.13496991
S2	1.85420689
S3	2.00789593

Testing for Autocorrelation

Autocorrelation is when an error term is related to a previous error term. This situation can happen with time series data such as monthly sales. The Durbin-Watson statistic can be used to check if autocorrelation exist. The Durbin-Watson statistic is demonstrated in the Testing the Assumptions section.

Testing for Outliers

Outliers are observations that exert a large influence on the overall outcome of a model or a parameter's estimate. When examining outlier diagnostics, the size of the dataset is important in determining cutoffs. A data set where $(2(p/n)^{1/2} > 1)$ is considered large. p is the number of terms in the model excluding the intercept. n is the sample size.

The HOUSES data set we are using has n=15 and p=4 (BEDROOMS, S1, S2, S3). Since $2 \times (4/15)^{1/2} = 1.0$ we will consider the HOUSES dataset small.

The REG procedure can be used to view various outlier diagnostics. The Influence option requests a host of outlier diagnostic tests. The R option is used to print out Cook's D. The output prints statistics for each observation in the dataset.

```
PROC REG DATA=HOUSES;
  MODEL PRICE = BEDROOMS S1 S2 S3 /
    INFLUENCE R;
```

```
RUN;
```

Cook's D is a statistic that detects outlying observations by evaluating all the variables simultaneously. SAS prints a graph that makes it easy to spot outliers using Cook's D. A Cook's D greater than the absolute value of 2 should be

investigated.

		Dep	Var	Predict							
Cook's		Obs	PRICE	Value	Residual	-2	-1	-0	1	2	D
1	64000.0	64442.6	-442.6								0.000
2	65850.0	50433.8	15416.2				***				0.547
3	80050.0	86915.2	-6865.2		*						0.026
4	107250	100355	6895.3			*					0.032
5	86650.0	80972.3	5677.7								0.018

RSTUDENT is the studentized deleted residual. The studentized deleted residual checks if the model is significantly different if an observation is removed. An RSTUDENT whose absolute value is larger than 2 should be investigated.

The DFFITS statistic also test if an observation is strongly influencing the model. Interpretation of DFFITS depends on the size of the dataset. If your dataset is small to medium size, you can use 1.0 as a cut off point. DFFITS greater than 1.0 warrant investigating. For large datasets, investigate observations where $DFFITS > 2(p/n)^{1/2}$. DFFITS can also be evaluated by comparing the observations among themselves. Observations whose DFFITS values are extreme in relation to the others should be investigated. An abbreviated version of the printout is listed.

Obs	Rstudent	Dffits	INTERCEP Dfbetas	BEDROOMS Dfbetas	s1 Dfbetas
1	-0.0337	-0.0197	-0.0021	0.0026	-0.0131
2	1.7006	1.8038	0.9059	-1.0977	-0.2027
3	-0.5450	-0.3481	-0.2890	0.1289	0.2485
4	0.5602	0.3848	0.1490	0.1806	0.0333
5	0.4484	0.2864	-0.0875	0.1060	0.2045

In addition to tests for outliers that affect the overall model, the Dfbetas statistics can be used to find outliers that influence an particular parameter's coefficient. There will be a Dfbetas statistic for each term in the model. For small to medium sized datasets, a Dfbetas over 1.0 should be investigated. Suspected outliers of large datasets have Dfbetas greater than $2/\sqrt{n}$.

Outliers can be addressed by: assigning weights, modifying the model (eg. transform variables) or deleting the observations (eg. if data entry error suspected). Whatever approach is chosen to deal with outliers should be done for a reason and not just to get a better fit!

Testing the Fit of the Model

The overall fit of the model can be checked by looking at the F-Value and its corresponding p-value (Prob >F) for the total model under the Analysis of Variance portion of the REG or GLM print out. Generally, you want a Prob>F value less than 0.05.

If your dataset has "replicates", you can perform a formal Lack of Fit test. This test can be run using PROC RSEG with option LACKFIT in the model statement.

```
PROC RSREG DATA=HOUSES
  MODEL PRICE = BEDROOMS
  STYPE/LACKFIT ;

RUN;
```

If the p-value for the Lack of Fit test is greater than 0.05 then your model is a good fit and no additional terms are needed.

Another check you can perform on your model is to plot the error term against the dependent variable Y. The "shape" of the plot will indicate whether the function you've modeled is actually linear.

A shape that centers around 0 with 0 slope indicates the function is linear. Figure 5 is an example of a linear function.

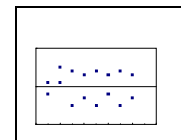


Figure 5

Two other shapes of Y plotted against residuals are a curvilinear shape (Figure 6) and an increasing area shape (Figure 7). If the plot of residuals is curvilinear, you should try changing the X variable in your model to X^2 .

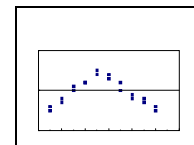


Figure 6

If the plot of residuals looks like Figure 7, you should try transforming the predicted Y variable.

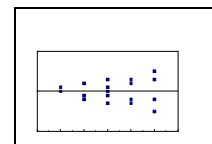
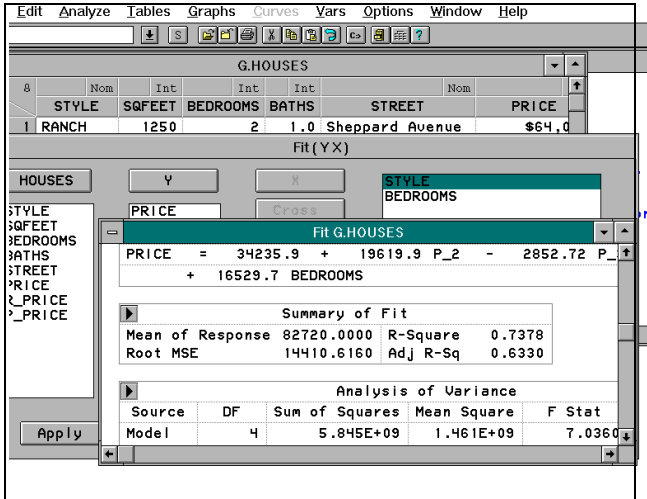


Figure 7

Modeling Without Code

Interactive SAS will let you run linear regression without writing SAS code. To do this invoke SAS/INSIGHT® (in PC SAS click Globals / Analyze / Interactive Data Analysis and select the Fit (Y X)

option). Use your mouse to click on the data set variables to build the model and then click on Apply to run the model.



References

Neter, John, Wasserman, William, Kutner, Micheal H, [Applied Linear Statistical Models](#), Richard D. Irwin, Inc., 1990, pp.113-133

Brocklebank PhD, John C, Dickey PhD, David A, [SAS System for Forecasting Time Series](#), Cary, NC: SAS Institute Inc., 1986, p.9

Fruend PhD, Rudolf J, Littell PhD, Ramon C, [SAS System for Linear Regression, Second Edition](#), Cary, NC: SAS Institute Inc., 1991, pp.59-99.

SAS Institute Inc., [SAS/STAT User's Guide, Version 6, Fourth Edition](#), Vol 2, Cary, NC: SAS Institute Inc., 1990, pp.1416-1431.

SAS Institute Inc., [SAS Procedures Guide, Version 6, Third Edition](#), Cary, NC: SAS Institute Inc., 1990, p.627.

SAS and SAS/Insight are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.

About The Author

Leslie A. Christensen, Market Research Analyst Sr.
 The Goodyear Tire & Rubber Company
 1144 E. Market St, Akron, OH 44316-0001
 (330) 796-4955 USGTRDRB@IBMMAIL.COM