

Lecture 13: Model Building in multiple regression

Stat 102

- We will illustrate the practice of model building and discuss the theory through examination of a model building exercise.
- Data set pollution. JMP provides information about the relationship between pollution and mortality for 60 cities between 1959-1961.
- Goal: Build a multiple regression model that can be used to examine the effect on mortality of several pollution related variables.
 - Build the most useful model based on the available data.

The Data

- The Dependent Variable (Y) is **MORT**ality
 - = total age adjusted mortality in deaths per 100,000 population
- The x -variables that could be used are
 - PRECIP=mean annual precipitation (in inches);
 - EDUC = median number of school years completed for persons 25 and older;
 - NONWHITE = percentage of 1960 population that is nonwhite;
 - NOX = relative pollution potential of N_2O (related to amount of tons of = “Nitrous Oxide” (N_2O) emitted per day per square kilometer);
 - SO2 = relative pollution potential of SO_2 .
- Among these, the pollution-related descriptors are NOX, SO2 and PRECIP (indirectly)
- The remaining 2 variables are included as “controls”.

Controls help answer whether pollution is important *after controlling for* other relevant (non-pollution) factors.

Choosing variables to include

1. Which form of the variable?

- For each variable, decide whether it is appropriate to transform it, e.g., use the log or square root of the variable.
- For the Y- variable the main reason for a transformation is to attain homoscedasticity of the final model.
- There are two main reasons to make a transformation of the x -variables:
 - (1) the relationship between the explanatory variable and the response variable is nonlinear;
 - (2) the explanatory variable is crunched together with a few outliers and/or some influential points.

2. Which of the explanatory variables to include in the model?

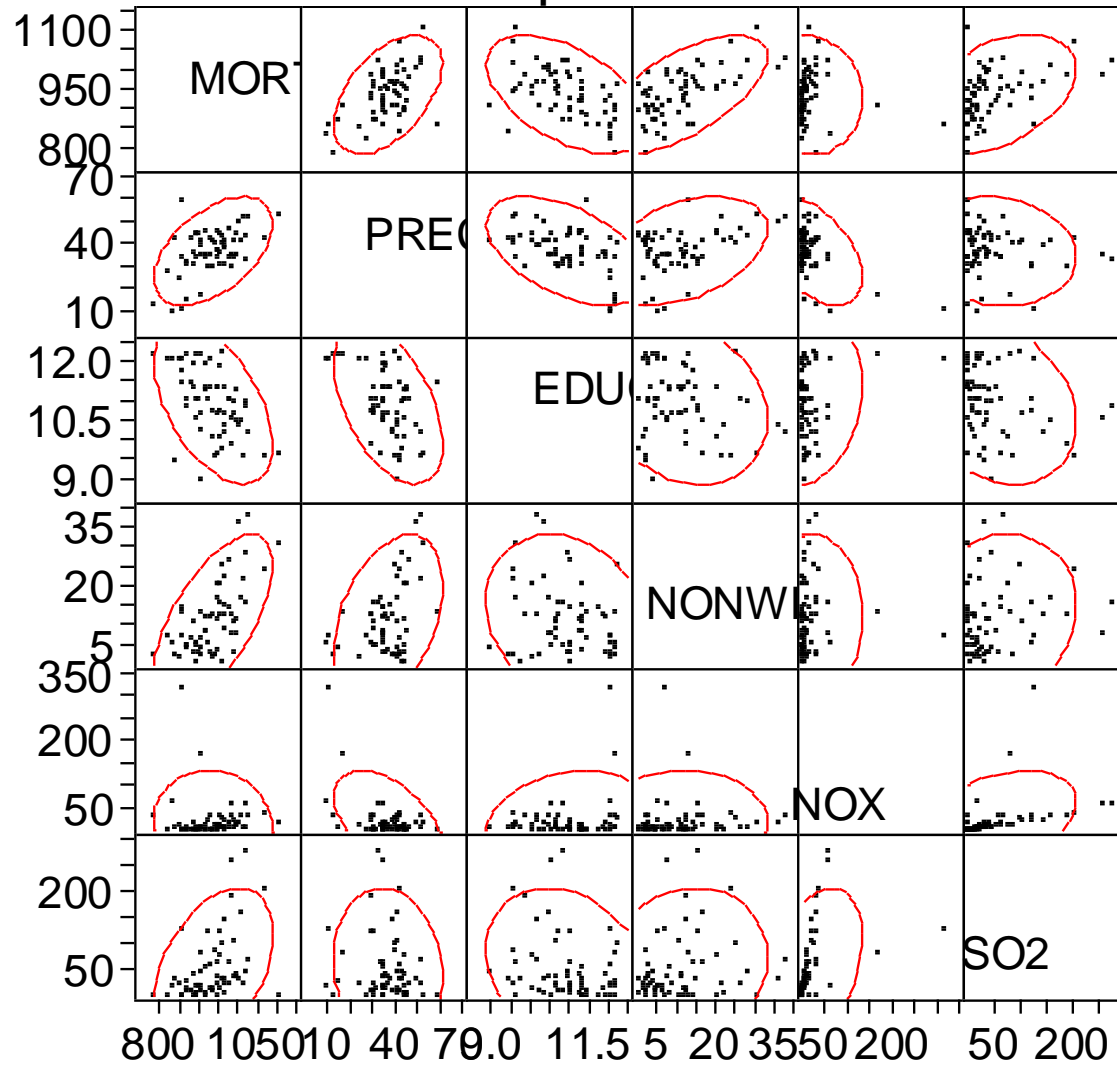
The Model Building Process is Interactive

- The goal is the best end product.
- There is no uniquely correct order in which to proceed.
- Decisions as to which variables to use, and how to transform them may need to be revisited as the analysis progresses.

Scatterplot Matrix

- This is often a useful first step.
- Examine the Simple Regression Plots of Y on the potential x -variables
 - See whether they seem mostly linear.
 - And see whether the Y -variable looks as if it will be homoscedastic in future analyses.
 - If not, try a transformation such as $\text{Log}(Y)$ before going further.
- Our data looks good in terms of heteroscedasticity
- It also looks pretty good in terms of nonlinearity but we will still find some useful corrective actions.
- *To get the Scatterplot Matrix in JMP click Multivariate and then put the Y -variable and all x -variables into the “ Y , Columns” box.*

Scatterplot Matrix



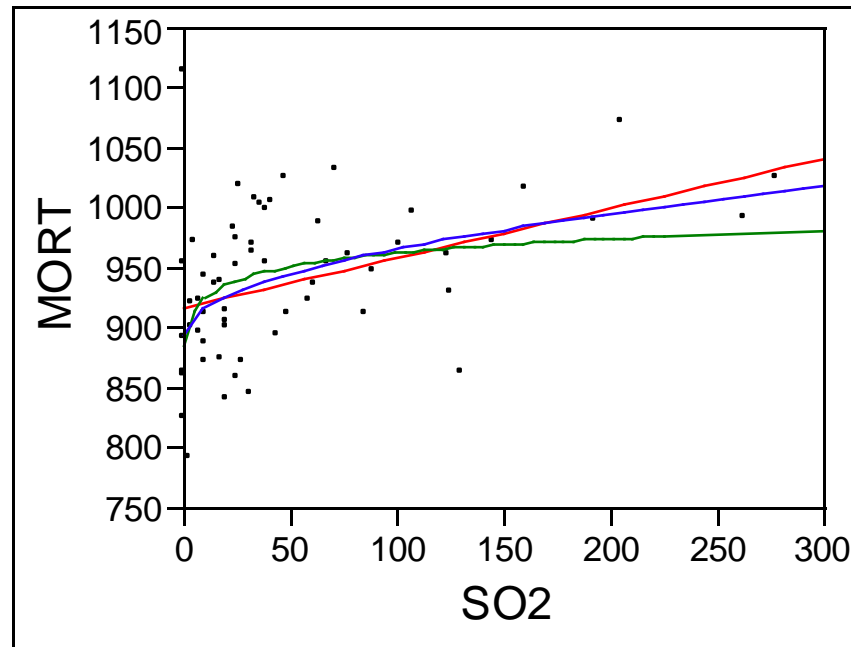
Transformation to Correct Non-Linearities

- If non-linearity is present in the relationship between Y and some x it is usually a good idea to try to correct it now, before going further.

We'll want to use only transformations of x – We probably don't want to transform Y since we seem to already have desirable homoscedasticity.

- It's not very evident from the small size plot in the Scatterplot Matrix, but there is some nonlinearity in the plot of MORT on SO2. We'll work on that first --

Plot of MORT on SO2



— Linear Fit: $R^2 = .181$

— Transformed Fit to Log: $R^2 = .163$

— Transformed Fit to SqRt: $R^2 = .199$

SqRt(SO2) is the best, if only by a modest amount.

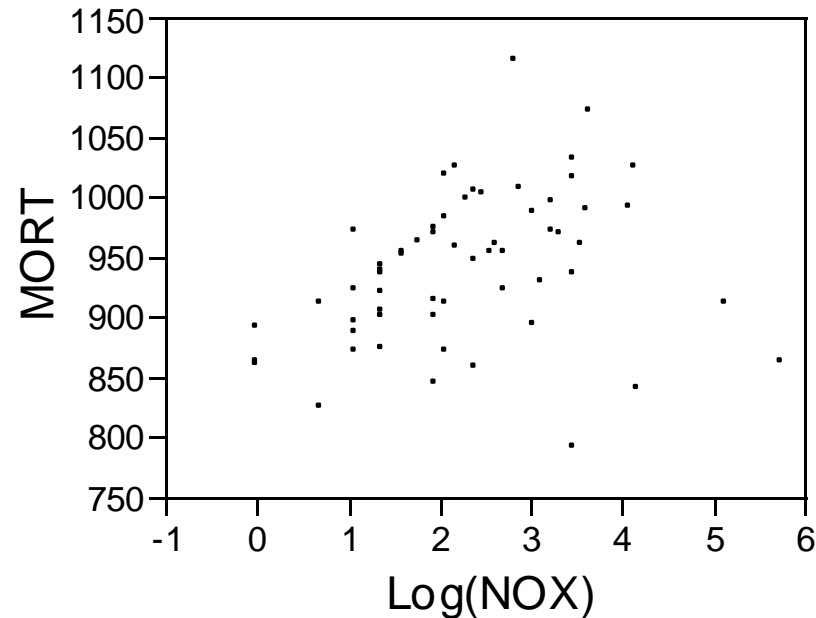
And it has another advantage we'll look at soon.

{& The residual plot also looks good.}

Transformations to Correct Crunching and bad influence situations

- All the scatterplots involving NOX as independent variable look odd. Including that for MORT on NOX.
- The problem is that the values of NOX are “crunched”.
- They mostly lie near one end (the left) with a few high influence points hanging out.
- High influence points are (usually) undesirable, as we’ve seen before.
- This type of situation can sometimes be improved by transforming the offending x -variable.
 - Use $\text{Log}(x)$ for right skewness; e^x for left skewness.
- $\text{Log}(x)$ works well for our data.

Scatterplot of MORT by Log(NOX)

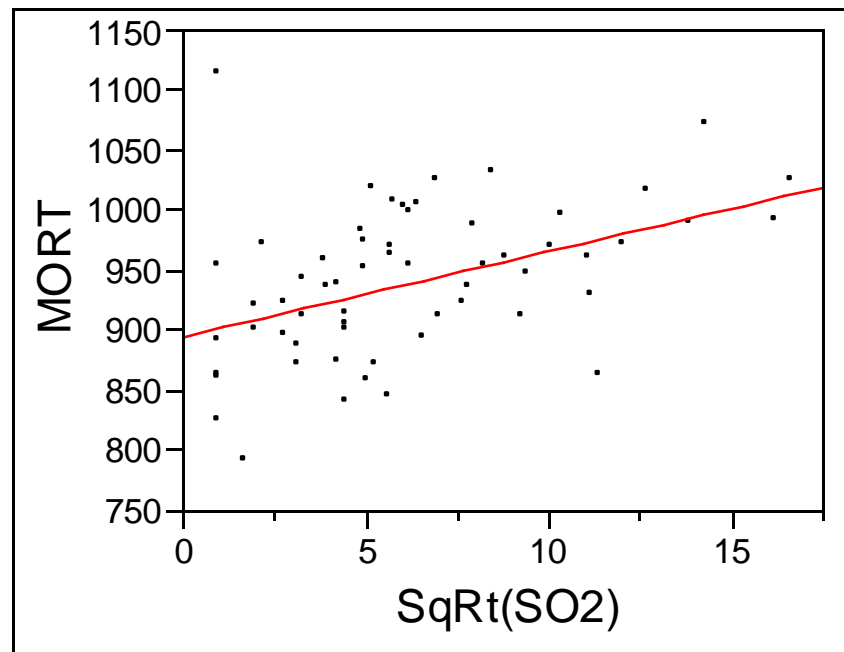


- The values of this x-variable are nicely spread out.
- The Y on x relationship looks not too non-linear.
 - It's not perfect, but it's not clear how we could get it to be better.
 - Mainly, there are 4 offending points (lower right); and we'll look more at these later.

The Bonus in Using SqRt(SO2)

- SO2 is also a little crunched.
- SqRt(SO2) “uncrunches” it.

Scatterplot of MORT by SqRt(SO2)



There's even a nicely linear pattern here!

Forward Selection

- It is important to use only the variables which are useful predictors.
 - Using other variables will result in worse predictions and higher standard errors of coefficient estimates (because degrees of freedom must be used to estimate the coefficients of the not-useful variables).
- So we usually don't keep variables which are insignificant. (P-value < 0.05)
- Proceed in steps:
 1. Choose the explanatory variable having the highest R^2 with Y . Include it **if** its P-value is < **0.05 (approx)**.
 2. Compute *residuals* from the simple linear regression.
 3. For the remaining x -variables calculate their values of R^2 with these *residuals*. Include the value with the highest such R^2 **if** it has **P < 0.05**.
 4. There are now two x -variables in the model being built. Compute the *residuals* from the multiple regression of Y on these two variables.
 5. For the remaining x -variables calculate their values of R^2 with these *residuals*. Include the value with the highest such R^2 **if** it has **P < 0.05 (approx)**. ETC.
 6. **Stop** when the relevant P-value is > **0.05**.

Forward Selection for Pollution Data

1. Multivariate Correlations –

	MORT	PRECIP	EDUC	NONWHITE	Log(NOX)	SqRt(SO2)
MORT	1.0000	0.5095	-0.5110	0.6437	0.2920	0.4458

- NONWHITE has the largest $R^2 = .6437^2 = .4143$.
- We can find the its P-value from the Parameter Estimates table:

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	887.1	10.37	85.53	<.0001
NONWHITE	4.488	0.7006	6.41	<.0001

- It has **P < .0001**. So we include this variable in our model.

2. In JMP save the residuals from this model.

3. Repeat Step #1, but using these residuals instead of the original Y-values.

Forward Selection for the Data (cont)

3(cont). Multivariate Correlations with the Residuals

	Residuals MORT	PRECIP	EDUC	NONWHITE	Log(NOX)	SqRt(SO2)
Residuals MORT	1.0000	0.3182	-0.4921	0.0000	0.2220	0.4616

- EDUC now has the largest $R^2 = (-.4921)^2 = .2422$.
- We can find the its P-value from the Parameter Estimates table:

Parameter Estimates (with Y = ResidualsMORT)

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	304.08	70.84	4.29	<.0001
EDUC	-27.71	6.44	-4.31	<.0001

- It has **P < .0001**. So we include this variable in our model.
- Additional Note: The correlation between ResidualsMORT and NONWHITE is **0.0000**. It is a property of linear regression (simple or multiple) that the correlation between the residuals and any explanatory variable in the model that produced them is **0**.

Final Model

- When this process is continued the **FINAL MODEL** has the 4 explanatory variables PRECIP, EDUC, NONWHITE, and SqRt(SO2).
- Log(NOX) is not part of that model.
- Note: If we consider a variable substantively important, we might want to include it in the model even if it is not put into the model by the model building process. Since we are interested in studying the effect of Log(NOX) on mortality, we might want to include it in the model even though it is not put in the model by the model building process.

Automatic Model Building in JMP:

1. Click Analyze, Fit Model and add all variables under consideration to the Construct Model Effects box. Change the personality to Stepwise and click Run Model.
2. If there are variables which you would like to include in the model for substantive reasons, regardless of their significance, check Lock next to the variable.
3. Enter into the model the variable with the largest F ratio if Prob>F is less than .05 for this variable (do this by clicking the Enter box).
4. Enter into the model the variable that has not already been entered into the model with the largest F ratio if Prob>F is less than .05. The F ratio for a variable X_j that has not been included in the model is the F statistic for testing the reduced model that includes only the variables already included in the model versus the full model that includes variable X_j in addition to the variables that have already been included in the model.
5. Repeat Step 4 until no more variables can be entered into the model.

Here are the results of the model building process for the Pollution data:

Stepwise Fit: Response: MORT

Stepwise Regression Control

Prob to Enter 0.050
 Prob to Leave 0.100

Current Estimates

		SSE	DFE	MSE	RSquare	Cp	AIC		
		72504	55	1318	0.6824	4.79	435.8		
Lock	Entered	Parameter	Estimate	nDF	SS	F Ratio	Prob>F		
X	X	Intercept	956.7	1	0	0.000	1.0000		
	X	PRECIP	1.725	1	10167	7.713	0.0075		
	X	EDUC	-14.00	1	5404	4.100	0.0478		
	X	NONWHITE	3.048	1	34372	26.074	0.0000		
		Log(NOX)	0	1	1051	0.795	0.3767		
	X	SqRt(SO2)	5.83	1	24858	18.857	0.0001		

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p
1	NONWHITE	Entered	0.0000	94595	0.4144	45.03	2
2	EDUC	Entered	0.0000	33848	0.5627	21.45	3
3	SqRt(SO2)	Entered	0.0012	17157	0.6378	10.48	4
4	PRECIP	Entered	0.0075	10167	0.6824	4.79	5

Click Make Model to fit the model with the chosen variables:

Parameter Estimates				
Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	956.7	92.8	10.31	<.0001
PRECIP	1.725	0.621	2.78	0.0075
EDUC	-13.97	6.900	-2.02	0.0478
NONWHITE	3.047	0.597	5.11	<.0001
SqRt(SO2)	5.829	1.342	4.34	<.0001

Notes: 1. The final R^2 is $R^2 = .6824$.

2. The Final P-value for EDUC is $P = .0478$, even though its P-value when first entered at Step 2 of this analysis was $P = .0000$. This is the same as the value $P = .0001$ that occurred in step 3 on p. 14.

3. The Stepwise Fit process produces exactly the same sequence of choices and of P-values as did the earlier process described on p. 12 – 14.

You should be able to explain WHY.

Tables from the Full (5-factor) Model

Summary of Fit

RSquare	0.687
Observations	60

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	156820	31364	23.70
Error	54	71452	1323	Prob > F
C. Total	59	228273		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t
Intercept	950.35	93.23	10.19	<.0001
PRECIP	2.01	0.698	2.88	0.0058
EDUC	-14.71	6.963	-2.11	0.0392
NONWHITE	2.825	0.6485	4.36	<.0001
Log(NOX)	6.708	7.526	0.89	0.3767
SqRt(SO2)	4.389	2.102	2.09	0.0415

Effect Tests

Source	DF	Sum of Squares	F Ratio	Prob > F
PRECIP	1	10940	8.27	0.0058
EDUC	1	5907	4.46	0.0392
NONWHITE	1	25120	18.99	<.0001
Log(NOX)	1	1051	0.7955	0.3767
SqRt(SO2)	1	5767	4.36	0.0415

Log(NOX) is not statistically significant, as was also claimed on p. 15 and shown in the table on p. 17.