Regression

Day 2

Transformations

What needs to be done when regression assumptions are not met?

Transformations

- If an examination of the residuals indicate that the assumptions underlying linear regression are not met, one must transform the variables.

- Such as taking logs, square roots, or reciprocals, can stabilize the variance, make the distribution normal, or make the relationship linear.

Transformations

Transforming Variables to Achieve Linearity

Method	Transform	Regression equation	Predicted value ($\hat{y}$)
Standard linear regression	None	$y = b_0 + b_1x$	$\hat{y} = b_0 + b_1x$
Exponential model	DV = log(y)	$\log(y) = b_0 + b_1x$	$\hat{y} = 10^{b_0 + b_1x}$
Quadratic model	DV = sqrt(y)	$\text{sqrt}(y) = b_0 + b_1x$	$\hat{y} = (b_0 + b_1x)^2$
Reciprocal model	DV = 1/y	$1/y = b_0 + b_1x$	$\hat{y} = 1 / (b_0 + b_1x)$
Logarithmic model	IV = log(x)	$y= b_0 + b_1\log(x)$	$\hat{y} = b_0 + b_1\log(x)$
Power model	DV = log(y) IV = log(x)	$\log(y)= b_0 + b_1\log(x)$	$\hat{y} = 10^{b_0 + b_1\log(x)}$

- Each Row shows a different transformation method.
- Transform column shows the method of transformation to be applied on DV or IV.
- Regression equation is the equation used in analysis.
- Last Column shows the equation of Prediction.

Transformations

greatlearning

Steps involved in conversion:

- Create Linear Regression Model.
- Construct a residual plot
- If the plot is random, don't transform the data.
- Compute the Coefficient of Determination (R2)
- Choose a Transformation method as mentioned in table in previous slide.
- Transform IV or DV or both.
- Apply Regression
- If the Transformed R2 is greater than the previous score, the transformation is a success.

The best transformation depends on the data and the best model will give the highest coefficient of determination.

Model Tuning

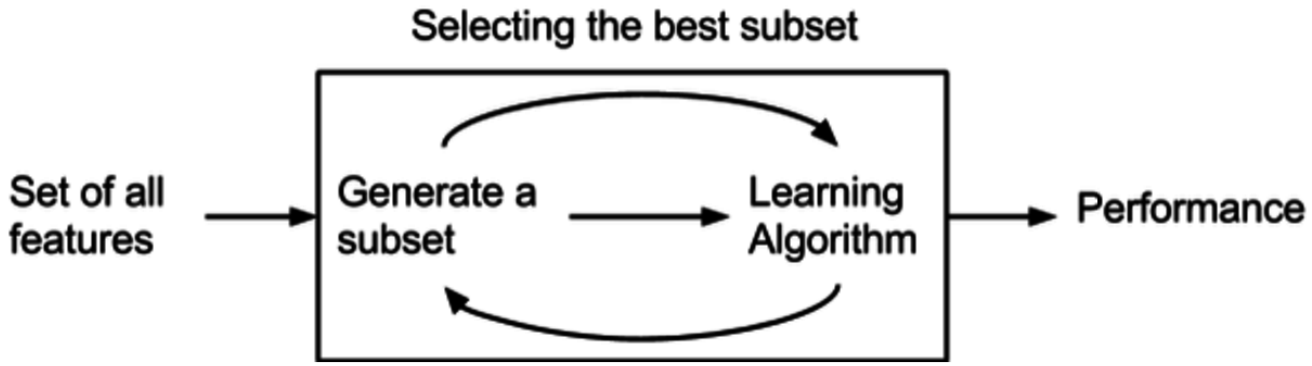
Variable/Feature Selection

Variable selection is intended to select the “best” subset of predictors

Prior to variable selection:

- Identify outliers and influential points - maybe exclude them at least temporarily or treat them.
- Add in any transformations of the variables that seem appropriate.

Variable/Feature Selection



- Source: https://en.wikipedia.org/wiki/Feature_selection#/media/File:Feature_selection_Wrapper_Method.png

Variable/Feature Selection

- Stepwise Regression

- Few independent variables considered in a study often turn out not to be significant.
- When there are large number of independent variables and the researcher suspects that not all of them are significant, stepwise regression should be used

- It is a regression procedure in which the predictor variables enter or leave the regression equation one at a time.

Approaches to stepwise regression

- Forward Inclusion
- Initially, there are no variables in the regression equation.
- Predictor variables are added one at a time, only if they meet certain criteria specified in terms of the F-ratio

- Backward Elimination
- Initially, all the predictor variables are included in the regression equation.
- Predictors are then removed once at a time based on the F-ratio.

- Stepwise Solution
- Forward inclusion is combined with the removal of predictors that no longer meet the specified criterion at each step

Forward Inclusion

Forward Inclusion

- Start with no variables in the model.
- For all predictors not in the model, check their p-value if they are added to the model. Choose the one with lowest p-value less than α_{crit} .
- Continue until no new predictors can be added.

- In this procedure, the first independent variable entered into the model is the one with the highest correlation with the dependent variable.

CARDIO Sample data

OBS	AGE	BMI	FFNUM	EXERCISE	BEER
1	26	23.2	0	621	3
2	30	30.2	9	201	6
3	32	28.1	17	240	10
4	27	22.7	1	669	5
5	33	28.9	7	1,140	12
6	29	22.4	3	445	9
7	32	23.2	1	710	15
8	33	20.3	0	783	11
9	31	25.6	1	454	0
10	33	21.2	3	432	2
11	26	22.3	5	1,562	13
12	34	23.0	2	697	1
13	33	26.3	4	280	2
14	31	22.2	1	449	5
15	31	19.0	0	689	4
16	27	20.8	2	785	3
17	36	20.9	2	350	7
18	35	36.4	14	48	11
19	31	28.6	11	285	12
20	36	27.5	8	85	5
Total	626	492.8	91	10,925	136
Mean	31.3	24.6	4.6	546.3	6.8

CARDIO Example

Dependent Variable

y = BMI

Independent Variables

x_1 = Age in years

x_2 = FFNUM, a measure of fast food usage,

x_3 = Exercise, an exercise intensity score

x_4 = Beers per day

Forward Inclusion

The REG Procedure

Model: MODEL1
Dependent Variable: bmi

Stepwise Selection: Step 1

Variable ffnum Entered: R-Square = 0.6613 and C(p) = 8.5625

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	228.24473	228.24473	35.15	<.0001
Error	18	116.88327	6.49351		
Corrected Total	19	345.12800			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	21.43827	0.78506	4842.33895	745.72	<.0001
ffnum	0.70368	0.11869	228.24473	35.15	<.0001

Stepwise Selection: Step 2

Variable beer Entered: R-Square = 0.7788 and C(p) = 2.0402

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	268.79888	134.39944	29.93	<.0001
Error	17	76.32912	4.48995		
Corrected Total	19	345.12800			

Forward Inclusion

Model: MODEL1
Dependent Variable: bmi

Stepwise Selection: Step 2

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	20.29360	0.75579	3237.09859	720.97	<.0001
ffnum	0.46380	0.12693	59.94878	13.35	0.0020
beer	0.33375	0.11105	40.55414	9.03	0.0080

Bounds on condition number: 1.654, 6.6161

All variables left in the model are significant at the 0.0500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number VarsIn	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	ffnum		1	0.6613	0.6613	8.5625	35.15	<.0001
2	beer		2	0.1175	0.7788	2.0402	9.03	0.0080

Backward Elimination

- Start with all the predictors in the model
- Remove the predictor with highest p-value greater than αcrit
- Refit the model and goto 2
- Stop when all p-values are less than αcrit .

- The αcrit is sometimes called the “p-to-remove” and does not have to be 5%. If prediction performance is the goal, then a 15-20% cut-off may work best, although methods designed more directly for optimal prediction should be preferred

Backward Elimination

Global hypothesis

The REG Procedure

Model: MODEL1
Dependent Variable: bmi

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.7932 and C(p) = 5.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	273.74877	68.43719	14.38	<.0001
Error	15	71.37923	4.75862		
Corrected Total	19	345.12800			

b_0
b_1
b_2
b_3
b_4

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	18.47774	6.45406	39.00436	8.20	0.0119
age	0.08424	0.18931	0.94239	0.20	0.6627
ffnum	0.42292	0.13671	45.53958	9.57	0.0074
exercise	-0.00107	0.00170	1.87604	0.39	0.5395
beer	0.32601	0.11518	38.12111	8.01	0.0127

Backward Elimination Step 1

Variable age Removed: R-Square = 0.7904 and C(p) = 3.1980

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	272.80638	90.93546	20.12	<.0001
Error	16	72.32162	4.52010		
Corrected Total	19	345.12800			

The REG Procedure
Model: MODEL1
Dependent Variable: bmi

Backward Elimination: Step 1

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	21.28788	1.30004	1211.98539	268.13	<.0001
ffnum	0.42963	0.13243	47.57610	10.53	0.0051
exercise	-0.00140	0.00149	4.00750	0.89	0.3604
beer	0.32275	0.11203	37.51501	8.30	0.0109

Bounds on condition number: 1.7883, 14.025

Backward Elimination: Step 2

Variable exercise Removed: R-Square = 0.7788 and C(p) = 2.0402

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	268.79888	134.39944	29.93	<.0001
Error	17	76.32912	4.48995		
Corrected Total	19	345.12800			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	20.29360	0.75579	3237.09859	720.97	<.0001
ffnum	0.46380	0.12693	59.94878	13.35	0.0020
beer	0.33375	0.11105	40.55414	9.03	0.0080

Bounds on condition number: 1.654, 6.6161

All variables left in the model are significant at the 0.0500 level.

The SAS System
The REG Procedure
Model: MODEL1
Dependent Variable: bmi

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	age	3	0.0027	0.7904	3.1980	0.20	0.6627
2	exercise	2	0.0116	0.7788	2.0402	0.89	0.3604

Stepwise Regression

- It is a combination of backward elimination and forward selection.
- This addresses the situation where variables are added or removed early in the process and we want to change our mind about them later
- At each stage a variable may be added or removed and there are several variations on exactly how this is done

Drawbacks of Stepwise Regression

- Due to the “one-at-a-time” nature of adding/dropping variables, it’s possible to miss the “optimal” model.
- The p-values used should not be treated too literally. The removal of less significant predictors tends to increase the significance of the remaining predictors.
- Variables that are dropped can still be correlated with the response. It would be wrong to say these variables are unrelated to the response.
- Stepwise variable selection tends to pick models that are smaller than desirable for prediction purposes

Cross Validation

Cross Validation

- Before assessing the relative importance of the predictors or drawing any other inferences, it is necessary to cross-validate the regression model.

- Cross-validation examines whether the regression model continues to hold on comparable data not used in the estimation.

Cross-Validation Procedure

1. The regression model is estimated using the entire data set.

2. The available data are split into two parts, the estimation sample and the validation sample. The estimation sample generally contains 50 – 90% of the total sample.

3. The regression model is estimated using the data from the estimation sample only. This model is compared with the model estimated on the entire sample to determine the agreement in terms of the signs and magnitudes of the partial regression coefficients.

Cross-Validation Procedure

4. The estimated model is applied to the data in the validation sample to predict the values of the dependent variable,Y ^ i,for the observations in the validation sample.

5. The observed values, Yi, and the predicted values, Y ^ i, in the validation sample are correlated to determine the simple r2.This measure, r2, is compared with R2 for the total sample and with R2 for the estimation sample to assess the degree of shrinkage.

References

- Books:
- Marketing Research An Applied Orientation – Naresh K.Malhotra and Satyabhushan Dash
- Business Statistics A First Course – David M Levine, Kathryn A Szabat, David F Stephan and P.K. Viswanathan
- Applied Predictive Analytics – Dean Abbott

- Others:
- https://cran.r-project.org/web/packages/olsrr/vignettes/residual_diagnostics.html
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6387894/
- http://www.biostat.jhsph.edu/~iruczins/teaching/jf/ch10.pdf
- http://scott.fortmann-roe.com/docs/BiasVariance.html