Into the Trenches of Regression Analysis (Part 2)

Smita Skrivanek
MoreSteam.com
February 13, 2013

MoreSteam.com®

Agenda

- Welcome
- Introduction of MBB Webcast Series
  - Larry Goldman, MoreSteam.com
- Today's Session
  - Smita Skrivanek, MoreSteam.com
- Open Discussion and Questions

MoreSteam.com

- Founded in 2000
- Trained over 390,000 Lean Six Sigma professionals
- Served over 2,000 corporate customers (including 50+% of the F500)
- First firm to offer the complete Black Belt curriculum online and only firm to offer online DfLSS
- Courses reviewed and approved by ASQ and PMI
- Academic Partnership with Ohio State University

Select Customers:

Today's Presenter

Smita Skrivanek

Senior Statistician, MoreSteam.com

- *Develops content & software functions, reviews projects, and assists students with questions on advanced statistics*
- *Heads research & development for EngineRoom® software*
- *Masters in Applied Statistics from The Ohio State University and an MBA from Indiana University Kelley School of Business*

Discussion Points

- *Multiple regression output*
- *Criteria for variable selection*
- *Alternative approaches to model building/selection*
- *Recommendations*
- *Brief overview of Generalized Linear Models*

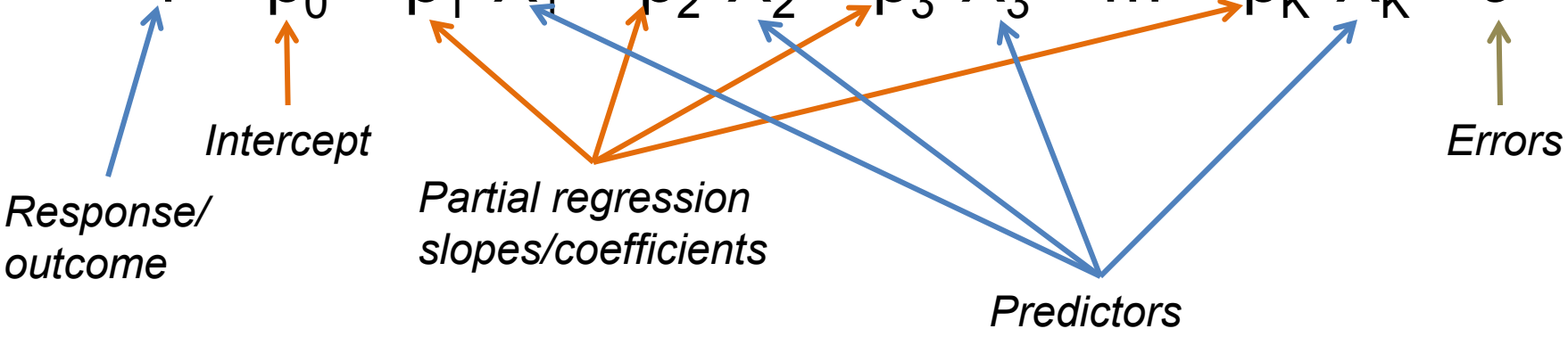
Dataset

- *Dataset from Journal of Statistics Education (amstat.org)*
- *Kelly Blue Book for several hundred 2005 used GM cars used to predict car value based on several characteristics*
- *Dependent (Y): Price*
- *Potential predictors (Xs): Price, Mileage, Make, Model, Trim, Type, Cylinder, Liter, Doors, Cruise, Sound, Leather*
- *JMP9 used to analyze data*

The multiple regression model

True:

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \ldots + \beta_K * X_K + \varepsilon$$



Estimated:

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \ldots + b_K * x_K$$

Anatomy of the regression output

Coefficient Table

Predictor	Coefficient	SE(Coeff)	t [H0: βi = 0]	P-value
Intercept	b0	se(b0)	t_{b0}	
X1	b1	se(b1)	t_{b1}	
X2	b2	se(b2)	t_{b2}	

$s = \sqrt{MS(Error)}$ R-Sq R-Sq(adj) PRESS

ANOVA Table

Source	DF	SumSq	MeanSumSq	F	P-value
Regression	k-1	RegSS	MS(Reg)	F_{Reg}	Model p
Error	n-k	ErrorSS	MS(Error)		
Total	n-1	TotalSS	MS(Total)		

Which predictors are important?

1. *Tests on partial regression slopes*
2. *Standardized partial regression slopes*
3. *Incremental variance explained (ΔR-squared)*

Example

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	1372.4266	1434.5	0.96	0.3390
Cylinder	2976.3618	719.8049	4.13	<.0001*
Liter	1412.1981	903.3883	1.56	0.1184

Example

***Unstandardized**:*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3145.7503	1325.934	2.37	0.0179*
Mileage	-0.152433	0.034638	-4.40	<.0001*
Cylinder	4027.6746	204.6118	19.68	<.0001*

Price = 3146 – 0.152*Mileage + 4028*Cylinder

$3146 = Price with zero mileage and zero cylinders (intercept)

$0.152 = Price reduces by $0.152 for each 1 mile increase in mileage while holding number of cylinders fixed

$4028 = Price increases by $4028 for 1 extra cylinder while holding mileage fixed

Example

***Standardized**:*

Parameter Estimates

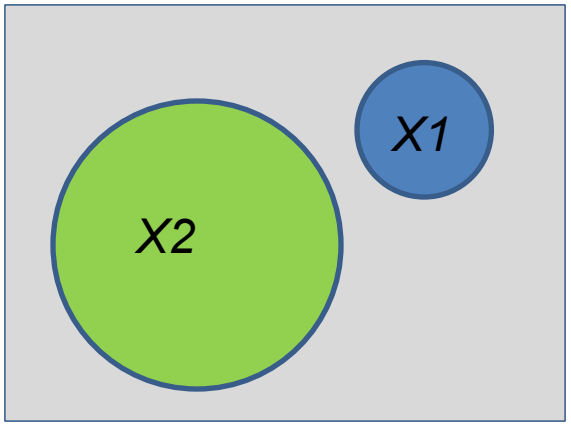
Term	Estimate	Std Error	t Ratio	Prob>|t|	Std Beta
Intercept	3145.7503	1325.934	2.37	0.0179*	0
Mileage	-0.152433	0.034638	-4.40	<.0001*	-0.12639
Cylinder	4027.6746	204.6118	19.68	<.0001*	0.565362

$$Z_{Price} = -0.126 * Z_{Mileage} + 0.565 * Z_{Cylinder}$$

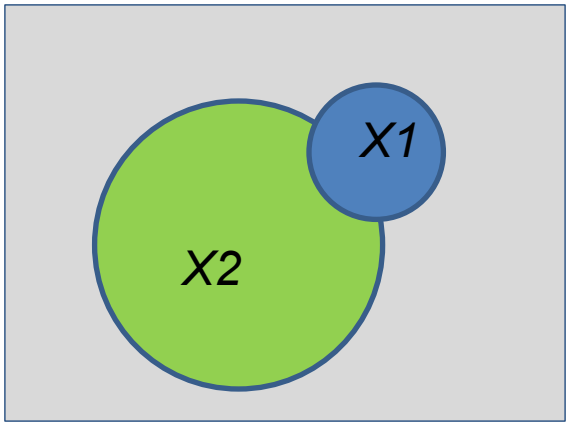
0.126 = Price reduces by 0.126 SD for each 1 SD increase in mileage holding number of cylinders fixed

0.565 = Price increases by 0.565 SD for 1 extra cylinder while holding mileage fixed

R-squared - variance explained



Uncorrelated

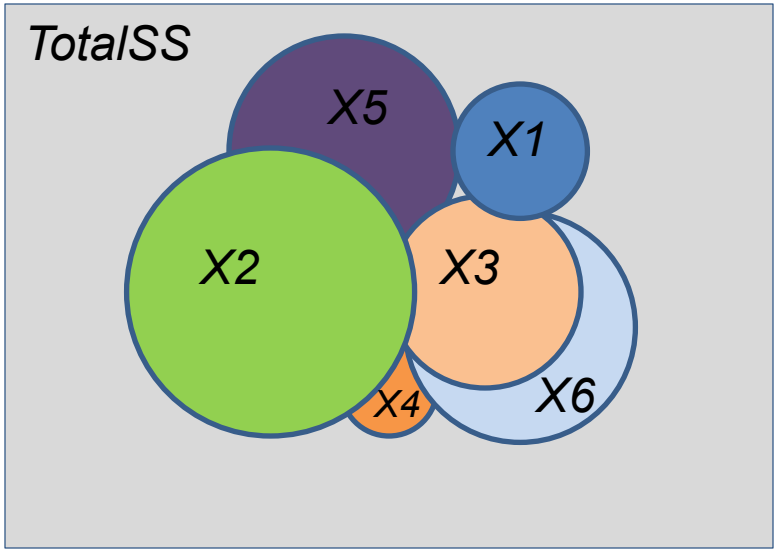


Correlated

$SS(X1,X2) = SS(X1) + SS(X2)$

$$R^2 = \frac{SS(X1,X2)}{TotalSS} = 1 - \frac{ErrorSS}{TotalSS}$$

What's wrong with incremental variance explained?



$$\Delta R^2 = \frac{SS(X1,X2,\ldots X6) - SS(X1,X2,\ldots X5)}{TotalSS}$$

$$F_{\Delta R^2} = \frac{\Delta R^2 / 1}{(1 - R^2_{with})/(n-k-1)}$$

$$F_{\Delta R^2} = \frac{SS(Extra\ \ due\ to\ added\ term)}{MS(Error)_{with}}$$

Problem:

Evaluating higher order terms

- Include polynomial terms: interactions (X1*X2), quadratic ($X1^2$)

$$y = b0 + b1*x1 + b2*x2 + b3*x1x2$$

$$y = b0 + b1*x1 + b2*x1^2$$

- Problem: Polynomial terms increase collinearity

  Solution: Run the model on centered predictors

Original X:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-17.05749	1126.944	-0.02	0.9879
Cylinder	4054.2025	206.8516	19.60	<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	63428.352	4516.492	14.04	<.0001*
Cylinder	-19725.35	1660.639	-11.88	<.0001*
Cylinder^2	2083.477	144.5988	14.41	<.0001*

Centered X:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	20081.396	411.9117	48.75	<.0001*
Cylinder-mean	5276.3696	203.0207	25.99	<.0001*
(Cylinder-mean)^2	2083.477	144.5988	14.41	<.0001*

The model selection problem

- k potential predictors = 2^k *potential subsets*

X1, X2 ⟹

Y = b0 (mean)
Y = b0 + b1X1
Y = b0 + b2X2
Y = b0 + b1X1 + b2X2

- 7 predictors => 128 possible models

- *One 'Best' model?*

- *Goal: Explanation vs. Prediction/Exploration*

Indices for selecting models

- *Model F, p-value*
  → *Large F, small p-value better*
- *Adjusted R-square*
  → *Larger values better*
- *MS(Error) or PRESS*
  → *Smaller values better*
- *Mallow's Cp*
  → *Cp =< p (number of model parameters including intercept)*
- *Information loss criteria – AIC, BIC*
  → *Smaller values better*

Model selection approaches

- *Simultaneous*
- *Sequential/Hierarchical*
- *Stepwise (automatic) procedures:*
    - *Forward selection*
    - *Backward elimination*
    - *Forward and Backward*
    - *Best subsets (All possible models)*

Simultaneous regression

- *All predictors enter model simultaneously*
- *Assess the amount of unique variance in the dependent variable explained by each independent variable.*
- *Strengths: Order of variables is unimportant. Useful for explanation, based on theory. Allows conclusions about relative effects. Estimates direct effects.*
- *Limitations: Regression slopes can change depending on the actual set of variables entered. Implies a theoretical model. Estimates only direct effects.*

Example: Response = Price/1000

- *Ordering of predictors unimportant:*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Std Beta
Intercept	7.3231643	1.770837	4.14	<.0001*	0
Mileage	-0.000171	3.186e-5	-5.35	<.0001*	-0.14139
Cylinder	3.2001246	0.202983	15.77	<.0001*	0.4492
Doors	-1.463399	0.308274	-4.75	<.0001*	-0.12586
Cruise	6.2055113	0.651463	9.53	<.0001*	0.271098
Sound	-2.024401	0.570718	-3.55	0.0004*	-0.09566
Leather	3.3271433	0.597114	5.57	<.0001*	0.150575

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Std Beta
Intercept	7.3231643	1.770837	4.14	<.0001*	0
Mileage	-0.000171	3.186e-5	-5.35	<.0001*	-0.14139
Leather	3.3271433	0.597114	5.57	<.0001*	0.150575
Doors	-1.463399	0.308274	-4.75	<.0001*	-0.12586
Cylinder	3.2001246	0.202983	15.77	<.0001*	0.4492
Cruise	6.2055113	0.651463	9.53	<.0001*	0.271098
Sound	-2.024401	0.570718	-3.55	0.0004*	-0.09566

- *Which predictors included is important:*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.6597701	1.661255	2.20	0.0279*
Cylinder	3.2471672	0.206276	15.74	<.0001*
Doors	-1.436854	0.313528	-4.58	<.0001*
Cruise	6.0796646	0.66222	9.18	<.0001*
Sound	-1.935754	0.580275	-3.34	0.0009*
Leather	3.2922223	0.607333	5.42	<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Std Beta
Intercept	13.898198	1.090933	12.74	<.0001*	0
Mileage	-0.000179	0.000016	-11.23	<.0001*	-0.14847
Make[Buick]	-3.667984	0.39584	-9.27	<.0001*	-0.15466
Make[Cadillac]	9.7795162	0.472062	20.72	<.0001*	0.412364
Make[Chevrolet]	-5.99373	0.250012	-23.97	<.0001*	-0.36809
Make[Pontiac]	-5.728626	0.291449	-19.66	<.0001*	-0.28917
Make[SAAB]	11.351309	0.39834	28.50	<.0001*	0.528955
Cylinder	3.7406511	0.139464	26.82	<.0001*	0.525073
Doors	-2.092318	0.160575	-13.03	<.0001*	-0.17995
Cruise	-0.095122	0.366817	-0.26	0.7955	-0.00416
Sound	0.0733854	0.295011	0.25	0.8036	0.003468
Leather	0.4903796	0.31561	1.55	0.1206	0.022193

Sequential/Hierarchical regression

- *Predictors are entered in steps, individually or in blocks, with each predictor or block being assessed in terms of what it adds to the prediction of Y after controlling for the previous entered predictors in the model.*

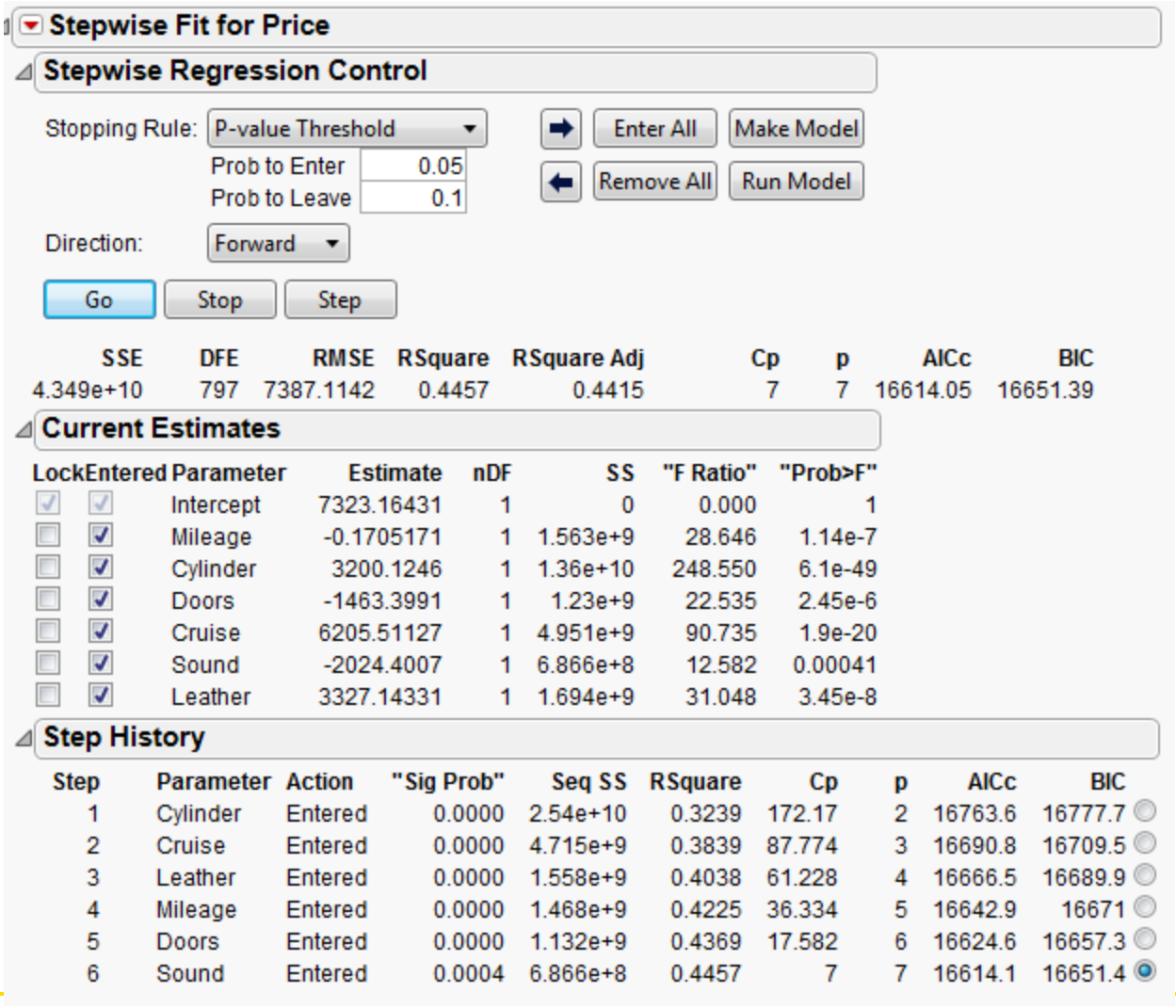
- *Strengths: useful for explanation, based on theory. Allows testing for curves/interactions. Estimates total effects.*

- *Limitations: incremental Rsq changes/can overestimate importance of variables depending on order of entry of variables. Order of entry implies a theoretical model. estimates only total effects.*

Stepwise regression

- *An automated program is used to select the variables and the order in which they are entered in the model based on pre-selected statistical criteria.*
- *Tells you how much unique variance in the dependent variable each of the independent variables explained.*
- *Strengths: can pick an efficient subset of variables for prediction based on purely statistical criteria. Doesn't need theoretical basis.*
- *Limitations: cannot use for explanation. Can produce nonsensical models.*

Forward Stepwise – p-value



Stepwise Fit for Price

Stepwise Regression Control

Stopping Rule: P-value Threshold

Prob to Enter 0.05

Prob to Leave 0.1

Direction: Forward

Enter All | Make Model | Remove All | Run Model

Go | Stop | Step

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
4.349e+10	797	7387.1142	0.4457	0.4415	7	7	16614.05	16651.39

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	7323.16431	1	0	0.000	1
☐	☑	Mileage	-0.1705171	1	1.563e+9	28.646	1.14e-7
☐	☑	Cylinder	3200.1246	1	1.36e+10	248.550	6.1e-49
☐	☑	Doors	-1463.3991	1	1.23e+9	22.535	2.45e-6
☐	☑	Cruise	6205.51127	1	4.951e+9	90.735	1.9e-20
☐	☑	Sound	-2024.4007	1	6.866e+8	12.582	0.00041
☐	☑	Leather	3327.14331	1	1.694e+9	31.048	3.45e-8

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC
1	Cylinder	Entered	0.0000	2.54e+10	0.3239	172.17	2	16763.6	16777.7
2	Cruise	Entered	0.0000	4.715e+9	0.3839	87.774	3	16690.8	16709.5
3	Leather	Entered	0.0000	1.558e+9	0.4038	61.228	4	16666.5	16689.9
4	Mileage	Entered	0.0000	1.468e+9	0.4225	36.334	5	16642.9	16671
5	Doors	Entered	0.0000	1.132e+9	0.4369	17.582	6	16624.6	16657.3
6	Sound	Entered	0.0004	6.866e+8	0.4457	7	7	16614.1	16651.4

Stepwise Fit for Price/1000

Stepwise Regression Control

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
9332.8443	796	3.4241318	0.8811	0.8800	6.1270858	8	4271.043	4313.023

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	19.5699611	1	0	0.000	1
☐	☑	Make{Saturn&Chevrolet&Pontiac&Buick-SAAB&Cadillac}	-8.4856365	4	42710.56	910.698	3e-295
☐	☑	Make{Saturn&Chevrolet-Pontiac&Buick}	-0.6708462	2	265.7372	11.332	1.4e-5
☐	☐	Make{Saturn-Chevrolet}	0	1	0.763986	0.065	0.7987
☐	☑	Make{Pontiac-Buick}	-0.8144438	1	134.0294	11.431	0.00076
☐	☑	Make{SAAB-Cadillac}	-1.1479221	1	177.7231	15.158	0.00011
☐	☐	Mileage/1000	0	1	0.942241	0.080	0.777
☐	☐	Cylinder-mean	0	1	14.60558	1.246	0.26464
☐	☑	M*C/1000	-0.0343035	1	1935.821	165.107	1.8e-34
☐	☑	Liter	5.18509832	1	14701.59	1253.901	1e-165
☐	☑	Doors	-1.6350969	1	1396.247	119.086	6.2e-26
☐	☐	Cruise	0	1	23.20105	1.981	0.15965
☐	☐	Sound	0	1	0.395574	0.034	0.8544
☐	☐	Leather	0	1	0.278081	0.024	0.87772

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	Make{SAAB-Cadillac}	Entered	0.0000	48505.66	0.6182	1750.9	3	5198.48	5217.18	○
2	Liter	Entered	0.0000	17182.09	0.8372	290.9	4	4515.22	4538.59	○
3	M*C/1000	Entered	0.0000	1996.214	0.8626	123.05	5	4380.63	4408.66	○
4	Doors	Entered	0.0000	1178.836	0.8777	24.739	6	4289.53	4322.22	○
5	Make{Pontiac-Buick}	Entered	0.0000	265.7372	0.8811	6.1271	8	4271.04	4313.02	○
6	Cruise	Entered	0.1596	23.20105	0.8813	6.1529	9	4271.09	4317.71	○
7	Cylinder-mean	Entered	0.2495	15.54356	0.8815	6.8303	10	4271.81	4323.06	○
8	Mileage/1000	Entered	0.3994	8.324862	0.8817	8.122	11	4273.15	4329.03	○
9	Make{Saturn-Chevrolet}	Entered	0.7875	0.852207	0.8817	10.049	12	4275.14	4335.64	○
10	Sound	Entered	0.8555	0.389774	0.8817	12.016	13	4277.18	4342.3	○
11	Leather	Entered	0.8985	0.191424	0.8817	14	14	4279.24	4348.97	○
12	Best	Specific	.	.	0.8811	6.1271	8	4271.04	4313.02	◉

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
9829.2011	795	3.516216	0.8747	0.8735	7.074855	9	4314.756	4361.374

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	32.3036167	1	0	0.000	1
☐	☐	Mileage/1000	0	1	10.42745	0.843	0.35876
☐	☑	Cylinder	-3.2727681	1	142.1567	11.498	0.00073
☐	☑	Cylinder^2	0.68212035	1	796.3568	64.410	3.6e-15
☐	☑	M*C/1000	-0.034969	1	1980.383	160.176	1.4e-33
☐	☑	Doors	-1.7934849	1	1635.379	132.272	2e-28
☐	☐	Cruise	0	1	8.244382	0.667	0.41451
☐	☐	Sound	0	1	16.45799	1.332	0.24885
☐	☐	Leather	0	1	3.680278	0.297	0.58567
☐	☑	Make{Saturn&Chevrolet&Pontiac&Buick-SAAB&Cadillac}	-7.2668766	4	30227.35	611.208	9e-241
☐	☑	Make{Saturn&Chevrolet-Pontiac&Buick}	-0.8742774	2	550.5265	22.264	3.9e-10
☐	☐	Make{Saturn-Chevrolet}	0	1	2.191488	0.177	0.67402
☐	☑	Make{Pontiac-Buick}	-1.4444095	1	399.1422	32.283	1.87e-8
☐	☑	Make{SAAB-Cadillac}	1.26373337	1	166.9088	13.500	0.00025

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	Make{SAAB-Cadillac}	Entered	0.0000	48505.66	0.6182	1619	3	5198.48	5217.18	○
2	Cylinder^2	Entered	0.0000	15919.17	0.8211	336.55	4	4591.02	4614.4	○
3	M*C/1000	Entered	0.0000	1997.672	0.8466	177.36	5	4469.62	4497.65	○
4	Doors	Entered	0.0000	1634.61	0.8674	47.474	6	4354.33	4387.02	○
5	Make{Pontiac-Buick}	Entered	0.0000	432.9064	0.8729	16.545	8	4324.25	4366.23	○
6	Cylinder	Entered	0.0007	142.1567	0.8747	7.0749	9	4314.76	4361.37	○
7	Sound	Entered	0.2489	16.45799	0.8749	7.7469	10	4315.46	4366.72	○
8	Mileage/1000	Entered	0.3531	10.6705	0.8751	8.886	11	4316.65	4372.53	○
9	Cruise	Entered	0.3956	8.933906	0.8752	10.165	12	4317.98	4378.49	○
10	Leather	Entered	0.6967	1.882116	0.8752	12.013	13	4319.9	4385.02	○
11	Make{Saturn-Chevrolet}	Entered	0.9083	0.164584	0.8752	14	14	4321.96	4391.7	○
12	Best	Specific	.	.	0.8747	7.0749	9	4314.76	4361.37	◉

Recommendations

- *Use prior knowledge and plot your data!*
- *Use manual methods for explanatory modeling.*
- *Use automated procedures only for preliminary exploratory/prediction modeling*
- *Obtain several plausible models and compare them or combine them.*
- *Don't forget multiple testing issues when 'fishing'*
- *Among automated procedures Best Subsets selection with AIC or BIC criteria are the best.*

Generalized Linear Models

$$g(\mu) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \ldots + \beta_K * X_K + \varepsilon$$

- *Random component*
- *Systematic component*
- *Link functions: Identity, Logit, Log*
- *Maximum Likelihood Estimation*

References

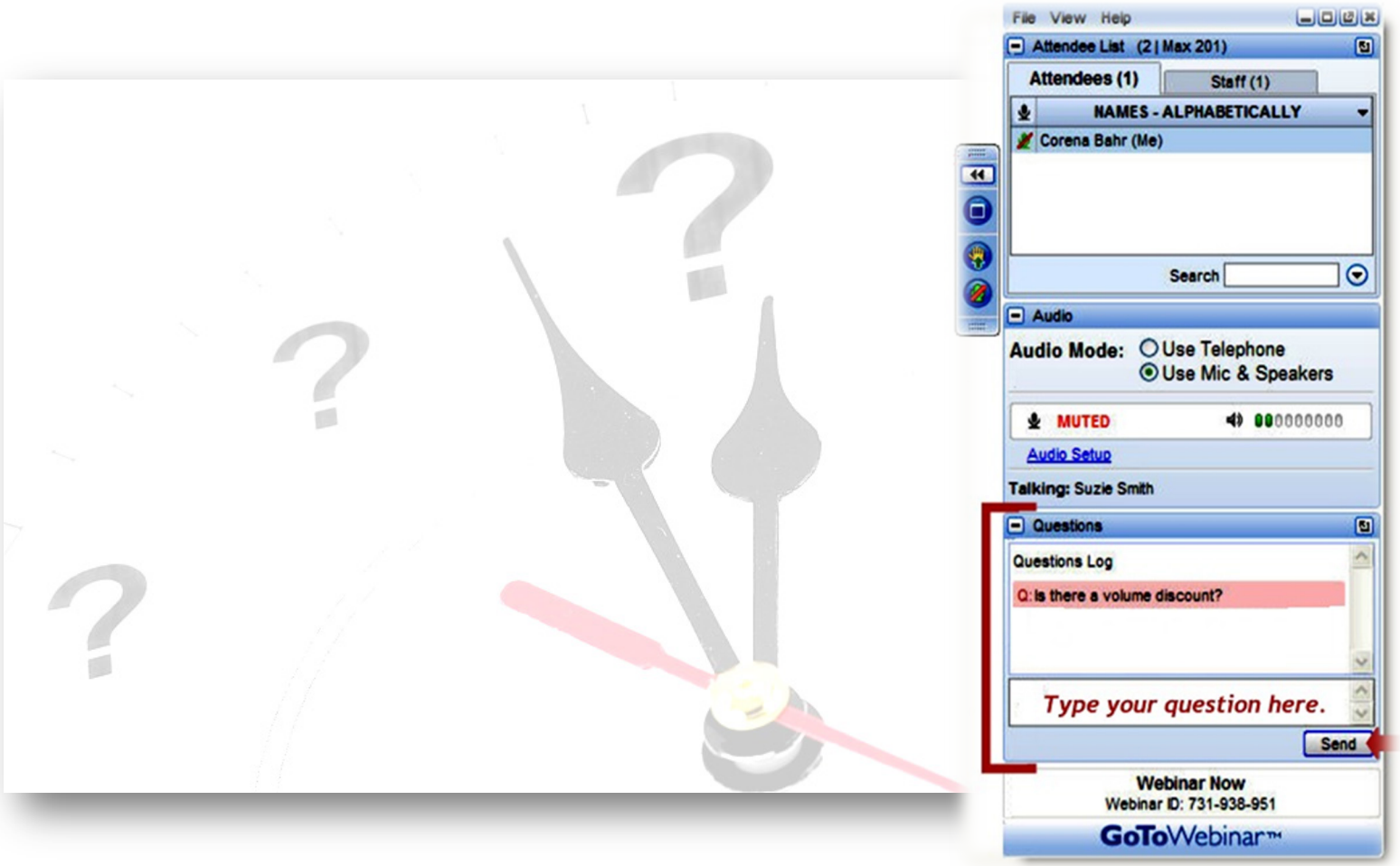
- *Data Analysis And Regression – a second course in statistics*
  - Frederick Mosteller, John W. Tukey

- *Introduction to Multiple Regression: How Much Is Your Car Worth? - Shonda Kuiper*
  http://www.amstat.org/publications/jse/v16n3/datasets.kuiper.html

- *Interpreting Multiple Linear Regression – A Guidebook of Variable Importance*
  - Laura Nathans, Frederick Oswald, Kim Nimon
  (Practical Assessment, Research & Evaluation, Volume 17, Number 9, April 2012)

- *Multiple Linear Regression in Data Mining – MIT OpenCourseWare*
  http://ocw.mit.edu/courses/sloan-school-of-management/15-062-data-mining-spring-2003/lecture-notes/lecture9.pdf

Thank You for Joining Us



Master Black Belt Program

- Offered in partnership with Fisher College of Business at The Ohio State University
- Employs a Blended Learning model with world-class instruction delivered in both the classroom and online
- Covers the MBB Body of Knowledge, topics ranging from advanced *DOE* to *Leading Change* to *Finance for MBBs*

Resource Links and Contacts

Questions? Comments? We'd love to hear from you.

Smita Skrivanek, Senior Statistician – MoreSteam.com
sskrivanek@moresteam.com

Larry Goldman, Vice President Marketing – MoreSteam.com
lgoldman@moresteam.com

Join us for our next Webcast on March 21st:

Eric Olsen, California Polytechnic State University, will discuss how to build 10,000 hours of experience with A3s.

Archived presentations and other materials:
http://www.moresteam.com/presentations/