Chapter 6: Multiple Linear Regression

Machine Learning for Business Analytics (4th ed.)

Shmueli, Bruce, K. Deokar & Patel

We assume a linear relationship between predictors and outcome:

outcome

coefficients

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon,$$

constant

error (noise)

predictors

Topics

- Explanatory vs. predictive modeling with regression
- Example: prices of Toyota Corollas
- Fitting a predictive model
- Assessing predictive accuracy
- Selecting a subset of predictors

Explanatory Modeling

Goal: Explain relationship between predictors (explanatory variables) and target

- Familiar use of regression in data analysis
- Model Goal: Fit the data well and understand the contribution of explanatory variables to the model
- Metrics: "goodness-of-fit" - R^2, residual analysis, p-values

Predictive Modeling

Goal: predict target values in other data where we have predictor values, but not target values

- Classic data mining context
- Model Goal: Optimize predictive accuracy
- Train model on training data
- Assess performance on validation (hold-out) data
- Explaining role of predictors is not primary purpose (but useful)

Example: Prices of Toyota Corolla
ToyotaCorolla.xls

Goal: predict prices of used Toyota Corollas based on their specification

Data: Prices of 1442 used Toyota Corollas, with their specification information

Data Sample

(showing only the variables to be used in analysis)

Price	Age	KM	Fuel_Type	HP	Metallic	Automatic	cc	Doors	Quarterly_Tax	Weight
13500	23	46986	Diesel	90	1	0	2000	3	210	1165
13750	23	72937	Diesel	90	1	0	2000	3	210	1165
13950	24	41711	Diesel	90	1	0	2000	3	210	1165
14950	26	48000	Diesel	90	0	0	2000	3	210	1165
13750	30	38500	Diesel	90	0	0	2000	3	210	1170
12950	32	61000	Diesel	90	0	0	2000	3	210	1170
16900	27	94612	Diesel	90	1	0	2000	3	210	1245
18600	30	75889	Diesel	90	1	0	2000	3	210	1245
21500	27	19700	Petrol	192	0	0	1800	3	100	1185
12950	23	71138	Diesel	69	0	0	1900	3	185	1105
20950	25	31461	Petrol	192	0	0	1800	3	100	1185

Variables Used

Price in Euros

Age in months as of 8/04

KM (kilometers)

Fuel Type (diesel, petrol, CNG)

HP (horsepower)

Metallic color (1=yes, 0=no)

Automatic transmission (1=yes, 0=no)

CC (cylinder volume)

Doors

Quarterly_Tax (road tax)

Weight (in kg)

Preprocessing

Fuel type is categorical, must be transformed into binary variables

Diesel (1=yes, 0=no)

CNG (1=yes, 0=no)

None needed for “Petrol” (reference category)

The Fitted Regression Model

Coefficients

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Standard Error	T-Statistic	P-Value
Intercept	-3092.3662	-6251.7994	67.0670	1608.6673	-1.9223	0.0550
Age_08_04	-132.6615	-141.8995	-123.4236	4.7036	-28.2041	0.0000
KM	-0.0212	-0.0256	-0.0168	0.0022	-9.4588	0.0000
HP	42.0362	32.7017	51.3708	4.7528	8.8445	0.0000
Met_Color	165.2588	-75.0337	405.5514	122.3481	1.3507	0.1773
Automatic	454.3259	-44.8779	953.5298	254.1763	1.7874	0.0744
CC	0.0037	-0.1815	0.1888	0.0943	0.0389	0.9690
Doors	-129.4727	-249.6833	-9.2621	61.2068	-2.1153	0.0348
Quarterly_Tax	15.3352	10.1466	20.5237	2.6418	5.8048	0.0000
Weight	13.9023	10.8882	16.9163	1.5347	9.0589	0.0000
Fuel_Type_Diesel	1389.9271	466.3193	2313.5350	470.2672	2.9556	0.0032
Fuel_Type_Petrol	2515.8331	1568.7759	3462.8902	482.2067	5.2173	0.0000

Predicted Values (Validation set)

Predicted price computed using regression coefficients

Record ID	Price	Prediction: Price	Residual
Record 774	10950	9282.1065	1667.8935
Record 104	18500	18694.6944	-194.6944
Record 903	9950	7633.4346	2316.5654
Record 660	10500	8810.3560	1689.6440
Record 575	9980	11185.8312	-1205.8312
Record 163	19600	19019.6580	580.3420
Record 411	7900	10398.2070	-2498.2070
Record 694	9900	7141.5331	2758.4669
Record 460	10990	10661.7933	328.2067
Record 795	11950	10415.4776	1534.5224
Record 290	12950	13190.5301	-240.5301
Record 207	12500	12703.5841	-203.5841
Record 328	12950	14729.9383	-1779.9383
Record 350	12750	14864.5311	-2114.5311
Record 971	9950	8457.1722	1492.8278
Record 470	11250	11464.4207	-214.4207
Record 869	9950	9166.8561	783.1439
Record 964	9950	10535.0076	-585.0076
Record 966	9900	9854.2586	45.7414
Record 667	9500	8320.2497	1179.7503

Residuals = errors = difference between actual and predicted prices

Model Evaluation (Validation Set)

Metric	Value
SSE	799341491.6681
MSE	1998353.7292
RMSE	1413.6314
MAD	1109.2115
R2	0.8582

Specialized Metrics Used in Regression (lower values are better)

Akaike Information Criterion (AIC)

AIC = $n \ln(SSE/n) + n(1 + \ln(2\pi)) + 2(p + 1)$

Bayesian Information Criterion (BIC)

BIC = $n \ln(SSE/n) + n(1 + \ln(2\pi)) + \ln(n)(p + 1)$

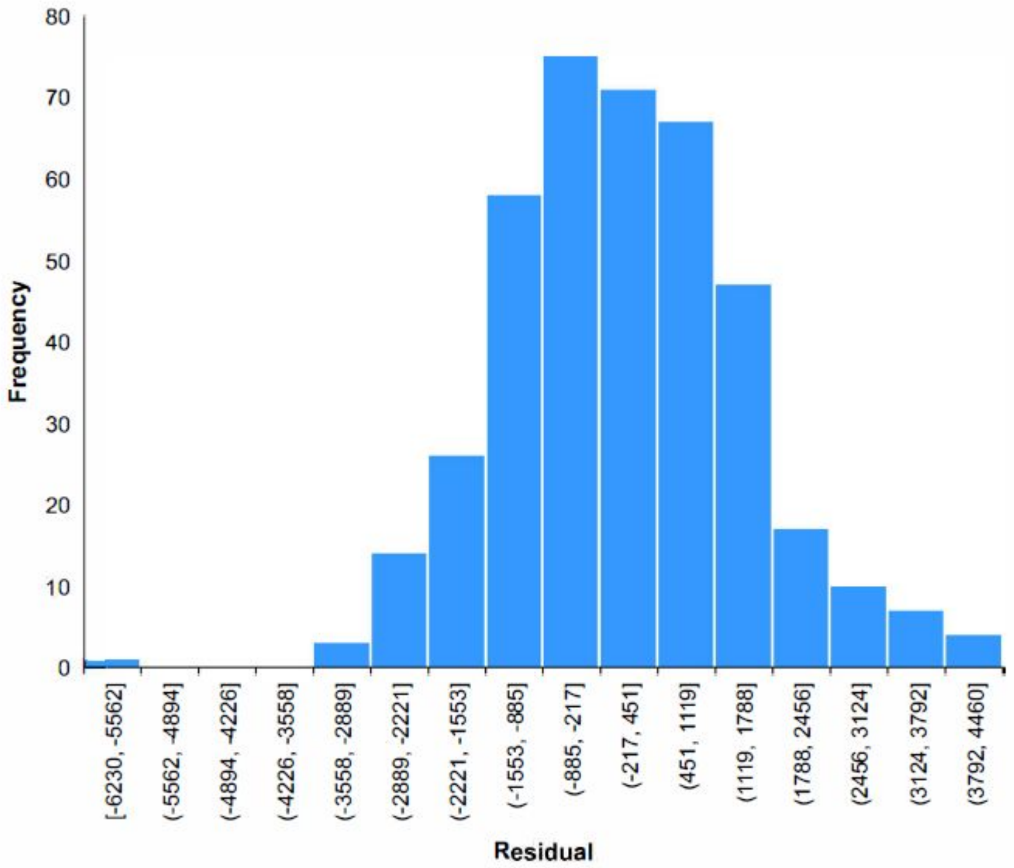
Mallow's Cp

Cp = $SSE/\sigma^2_{full} + 2(p+1) - n$

σ^2_{full} is the estimated MSE for the full model

Mallow's Cp is equivalent to AIC for large samples

Distribution of Residuals



Mostly symmetric distribution, a few large negative outliers

Feature (Variable, Predictor) Selection

- Why select a subset of attributes to predict the target?
- More predictors/attributes problems:
  - Expensive data collection
  - More missing data
  - Multicollinearity – some predictors behave the same way
  - Uncorrelation with target variable
- The goal
  - Find parsimonious model (simplest model that performs sufficiently well)
  - More robust & higher predictive accuracy
- Variable selection methods
  - Exhaustive search
  - Partial Subset selection: Forward
  - Partial Subset selection: Backward
  - Partial Subset selection: Stepwise

Selecting Subsets of Predictors

Goal: Find parsimonious model (the simplest model that performs sufficiently well)

- More robust
- Higher predictive accuracy

Exhaustive Search

Partial Search Algorithms

- Forward
- Backward
- Stepwise

Exhaustive Search = Best Subset

- All possible subsets of predictors assessed (single, pairs, triplets, etc.)
- Computationally intensive, not feasible for big data
- Judge by "adjusted R^2"

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

Penalty for number of predictors

Exhaustive output shows best model for each number of predictors

Best Subsets

Subset ID	Intercept	Age_08_04	KM	HP	Met_Color	Automatic	CC	Doors	Quarterly_Tax	Weight	Fuel_Type_Diesel	Fuel_Type_Petrol
Subset 1	1	0	0	0	0	0	0	0	0	0	0	0
Subset 2	1	1	0	0	0	0	0	0	0	0	0	0
Subset 3	1	1	0	1	0	0	0	0	0	0	0	0
Subset 4	1	1	1	0	0	0	0	0	0	1	0	0
Subset 5	1	1	1	1	0	0	0	0	0	1	0	0
Subset 6	1	1	1	1	0	0	0	0	1	1	0	0
Subset 7	1	1	1	1	0	0	0	0	1	1	0	1
Subset 8	1	1	1	1	0	0	0	0	1	1	1	1
Subset 9	1	1	1	1	0	0	0	1	1	1	1	1
Subset 10	1	1	1	1	0	1	0	1	1	1	1	1
Subset 11	1	1	1	1	1	1	0	1	1	1	1	1
Subset 12	1	1	1	1	1	1	1	1	1	1	1	1

Each row is the best model for a given # of predictors, “1” and “0” show whether the variable is included

Performance metrics for models with 1 predictor, 2 predictors, 3 predictors, etc. (exhaustive search method)

Best Subsets Details					
Subset ID	#Coefficients	RSS	Mallows's Cp	R2	Adjusted R2
Subset 1	1	8400665894	4058.5313	0.0000	0.0000
Subset 2	2	2052522022	541.7233	0.7557	0.7553
Subset 3	3	1694127987	345.0636	0.7983	0.7977
Subset 4	4	1348575230	155.5220	0.8395	0.8387
Subset 5	5	1153590564	49.4410	0.8627	0.8618
Subset 6	6	1125330324	35.7762	0.8660	0.8649
Subset 7	7	1092561716	19.6124	0.8699	0.8686
Subset 8	8	1077566856	13.3007	0.8717	0.8702
Subset 9	9	1069490729	10.8241	0.8727	0.8710
Subset 10	10	1064093126	9.8322	0.8733	0.8714
Subset 11	11	1060790523	10.0015	0.8737	0.8716
Subset 12	12	1060787798	12.0000	0.8737	0.8714

Metrics improve as you add predictors, then stabilize after you have about 9 predictors. ("Coefficients" is the number of predictors + 1, for the constant)

Exhaustive search may be computationally infeasible - some alternatives:

FORWARD SELECTION

- Start with no predictors
- Add them one by one (add the one with largest contribution)
- Stop when the addition is not statistically significant

BACKWARD ELIMINATION

- Start with all predictors
- Successively eliminate least useful predictors one by one
- Stop when all remaining predictors have statistically significant contribution

STEPWISE

- Like Forward Selection
- Except at each step, also consider dropping non-significant predictors

Next step

- Subset selection methods give candidate models that might be “good models”
- Do not guarantee that “best” model is indeed best
- Also, “best” model can still have insufficient predictive accuracy
- Must run the candidates and assess predictive accuracy (click “choose subset”)

Summary

- Linear regression models are very popular tools, not only for explanatory modeling, but also for prediction
- A good predictive model has high predictive accuracy (to a useful practical level)
- Predictive models are built using a training data set, and evaluated on a separate validation data set
- Removing redundant predictors is key to achieving predictive accuracy and robustness
- Subset selection methods help find “good” candidate models. These should then be run and assessed.