Chapter 6 – Multiple Linear Regression

Instructor: Zach Zhizhong ZHOU, Shanghai Jiao Tong University
主讲教师：周志中，上海交通大学

Data Mining for Business Intelligence

Shmueli, Patel & Bruce

Topics

□ Explanatory vs. predictive modeling with regression 解释性 vs. 预测性模型的回归分析

□ Example: prices of Toyota Corollas 例子：丰田卡罗拉的价格预测

□ Fitting a predictive model 拟合一个预测模型

□ Assessing predictive accuracy 评价预测的准确度

□ Selecting a subset of predictors 选择预测因子的子集

Explanatory Modeling 解释性模型

Goal: Explain relationship between predictors (explanatory variables) and target 目标：解释预测因子（解释变量）和目标之间的关系。

□Familiar use of regression in data analysis 数据分析中常见的回归分析。

□Model Goal: Fit the data well and understand the contribution of explanatory variables to the model 模型的目标：数据拟合很好并理解解释变量对模型的贡献。

□ “goodness-of-fit” : R^2, residual analysis, p-values 拟合优度：R^2，残差分析，p值

Predictive Modeling 预测性模型

Goal: predict target values in other data where we have predictor values, but not target values 目标：在只有预测因子而没有目标值的情况下预测目标值。

□Classic data mining context 经典数据挖掘内容

□Model Goal: Optimize predictive accuracy 模型目标：优化预测准确度。

□Train model on training data 使用训练数据来训练模型

□Assess performance on validation (hold-out) data 在验证数据集上评价预测模型的表现。

□Explaining role of predictors is not primary purpose (but useful) 解释预测因子所扮演的角色不是主要的目的（但对预测是有用的）。

Example: Prices of Toyota Corolla

ToyotaCorolla.xls

Goal: predict prices of used Toyota Corollas based on their specification 预测二手丰田卡罗拉轿车的价格

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 数据预处理

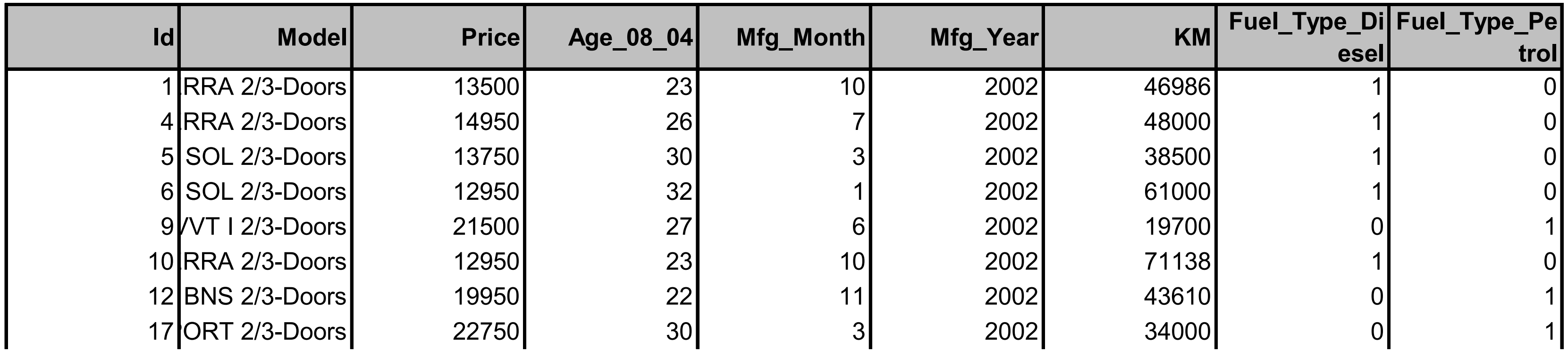
Fuel type is categorical, must be transformed into binary variables 燃料类型是分类型数据，需要构建0-1型虚拟变量表达。

Diesel (1=yes, 0=no)

CNG (1=yes, 0=no)

None needed for “Petrol” (reference category)

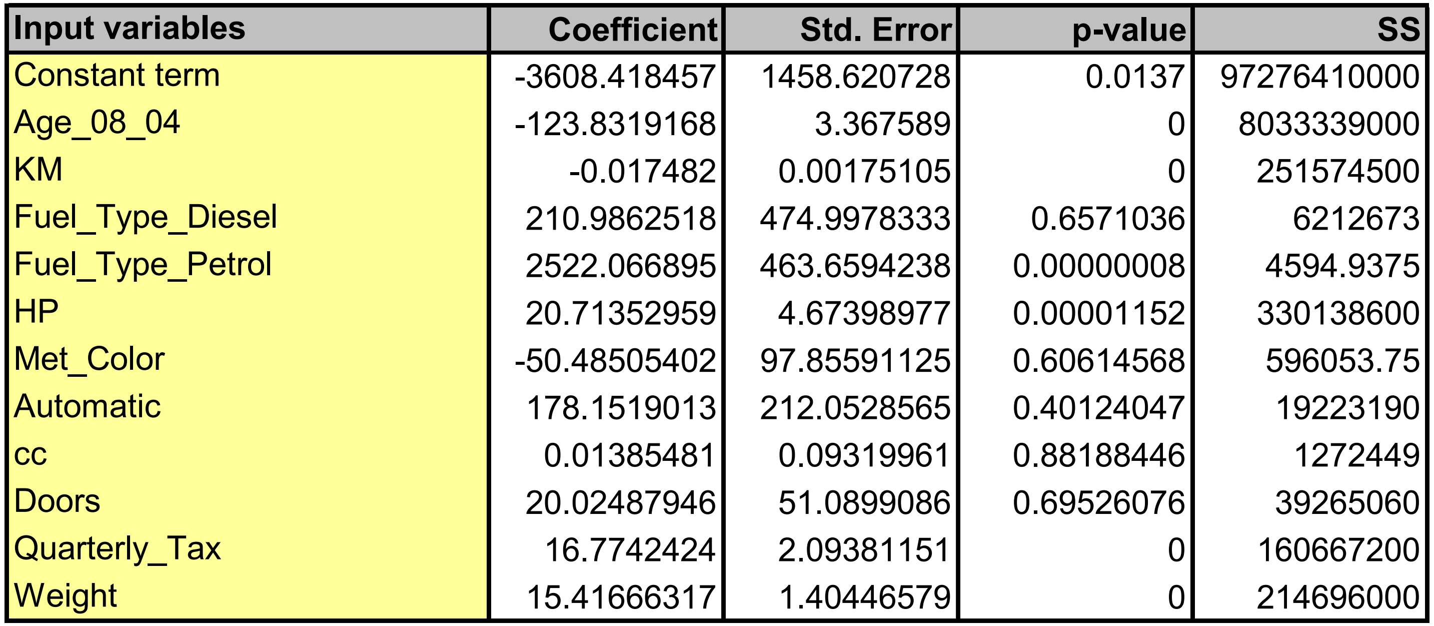
Subset of the records selected for training partition (limited # of variables shown)



Id	Model	Price	Age_08_04	Mfg_Month	Mfg_Year	KM	Fuel_Type_Diesel	Fuel_Type_Petrol
1	RRA 2/3-Doors	13500	23	10	2002	46986	1	0
4	RRA 2/3-Doors	14950	26	7	2002	48000	1	0
5	SOL 2/3-Doors	13750	30	3	2002	38500	1	0
6	SOL 2/3-Doors	12950	32	1	2002	61000	1	0
9	/VT I 2/3-Doors	21500	27	6	2002	19700	0	1
10	RRA 2/3-Doors	12950	23	10	2002	71138	1	0
12	BNS 2/3-Doors	19950	22	11	2002	43610	0	1
17	ORT 2/3-Doors	22750	30	3	2002	34000	0	1

60% training data / 40% validation data

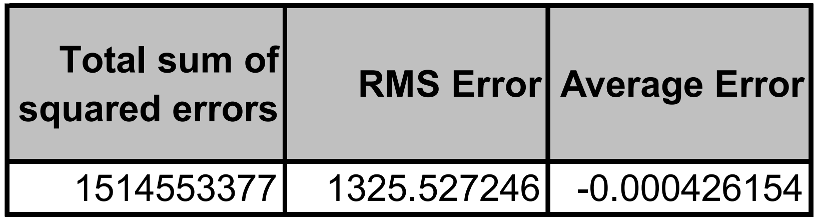
The Fitted Regression Model



Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3608.418457	1458.620728	0.0137	97276410000
Age_08_04	-123.8319168	3.367589	0	8033339000
KM	-0.017482	0.00175105	0	251574500
Fuel_Type_Diesel	210.9862518	474.9978333	0.6571036	6212673
Fuel_Type_Petrol	2522.066895	463.6594238	0.00000008	4594.9375
HP	20.71352959	4.67398977	0.00001152	330138600
Met_Color	-50.48505402	97.85591125	0.60614568	596053.75
Automatic	178.1519013	212.0528565	0.40124047	19223190
cc	0.01385481	0.09319961	0.88188446	1272449
Doors	20.02487946	51.0899086	0.69526076	39265060
Quarterly_Tax	16.7742424	2.09381151	0	160667200
Weight	15.41666317	1.40446579	0	214696000

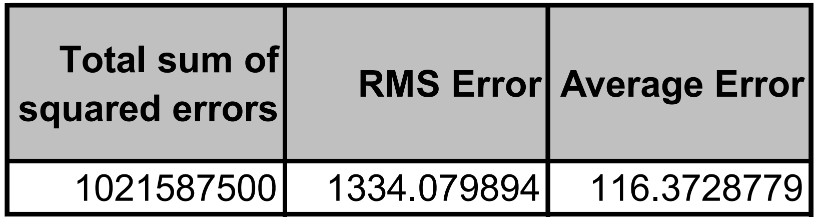
Error reports

Training Data scoring - Summary Report



Total sum of squared errors	RMS Error	Average Error
1514553377	1325.527246	-0.000426154

Validation Data scoring - Summary Report



Total sum of squared errors	RMS Error	Average Error
1021587500	1334.079894	116.3728779

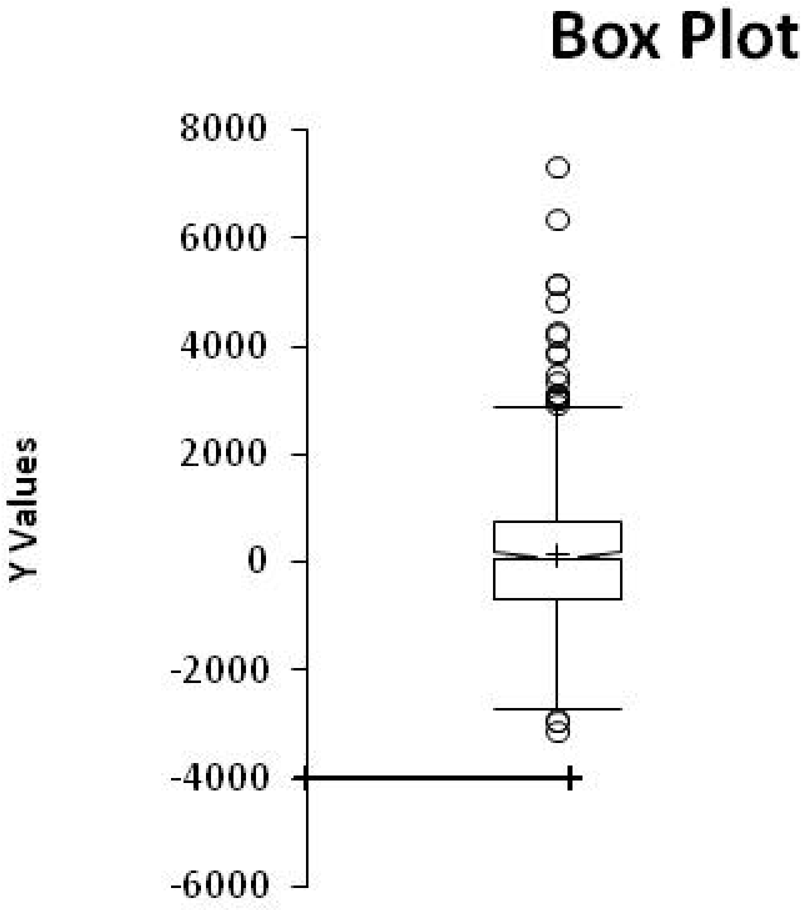
Predicted Values

Predicted price computed using regression coefficients

Predicted Value	Actual Value	Residual
15863.86944	13750	-2113.869439
16285.93045	13950	-2335.930454
16222.95248	16900	677.047525
16178.77221	18600	2421.227789
19276.03039	20950	1673.969611
19263.30349	19600	336.6965066
18630.46904	21500	2869.530964
18312.04498	22500	4187.955022
19126.94064	22000	2873.059357
16808.77828	16950	141.2217206
15885.80362	16950	1064.196384
15873.97887	16250	376.0211263
15601.22471	15750	148.7752903
15476.63164	15950	473.3683568
15544.83584	14950	-594.835836
15562.25552	14750	-812.2555172
15222.12869	16750	1527.871313
17782.33234	19000	1217.667664

Residuals = difference between actual and predicted prices

Distribution of Residuals



Symmetric distribution

Some outliers

Selecting Subsets of Predictors 选择预测因子子集

Goal: Find parsimonious model (the simplest model that performs sufficiently well) 目标：找到精简模型（表现足够好的最简单的模型）

- More robust 更加健壮（高鲁棒性）
- Higher predictive accuracy 高预测准确度

Exhaustive Search 穷举搜索

Partial Search Algorithms 部分搜索算法

- Forward 前向搜索
- Backward 后向搜索
- Stepwise 逐步搜索

Exhaustive Search 穷举搜索

□All possible subsets of predictors assessed (single, pairs, triplets, etc.) 所有可能的子集都考虑到

□Computationally intensive 计算量大

□Judge by “adjusted R²” 使用adjusted R²评价模型

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

Penalty for number
of predictors

Forward Selection 前向选择

- □Start with no predictors 从没有预测因子开始

- □Add them one by one (add the one with largest contribution) 逐步增加预测因子，从对回归解释能力最大的因子开始。

- □Stop when the addition is not statistically significant 如果继续增加的因子不能显著性改善模型解释能力则停止。

Backward Elimination 后向排除

□Start with all predictors 从所有预测因子开始

□Successively eliminate least useful predictors one by one 逐步把最没用的预测因子排除

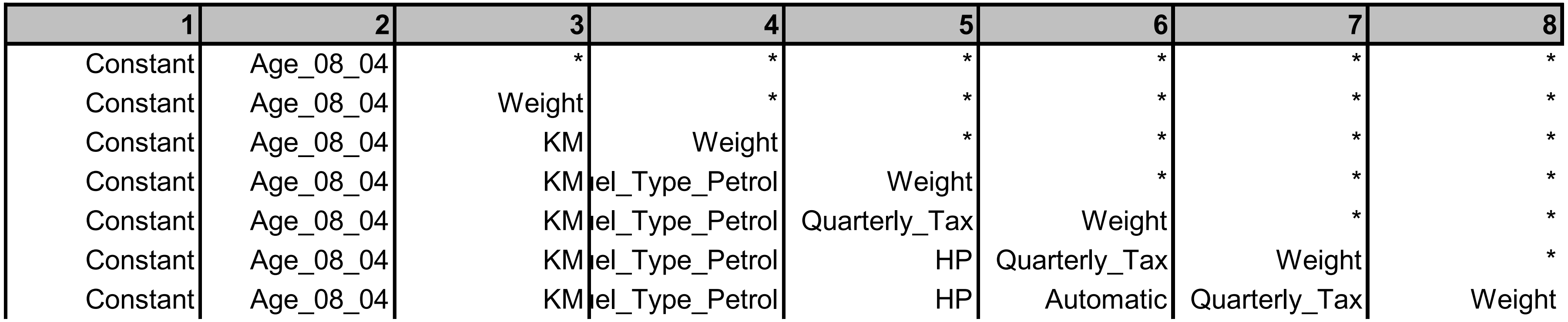
□Stop when all remaining predictors have statistically significant contribution 如果剩下的所有预测因子对模型解释能力都有显著贡献时停止。

Stepwise 逐步回归

□Like Forward Selection 类似于前向选择算法。

□Except at each step, also consider dropping non-significant predictors 但在每一步还会考虑到排除掉贡献度不大的预测因子的可能性。

Backward elimination (showing last 7 models)

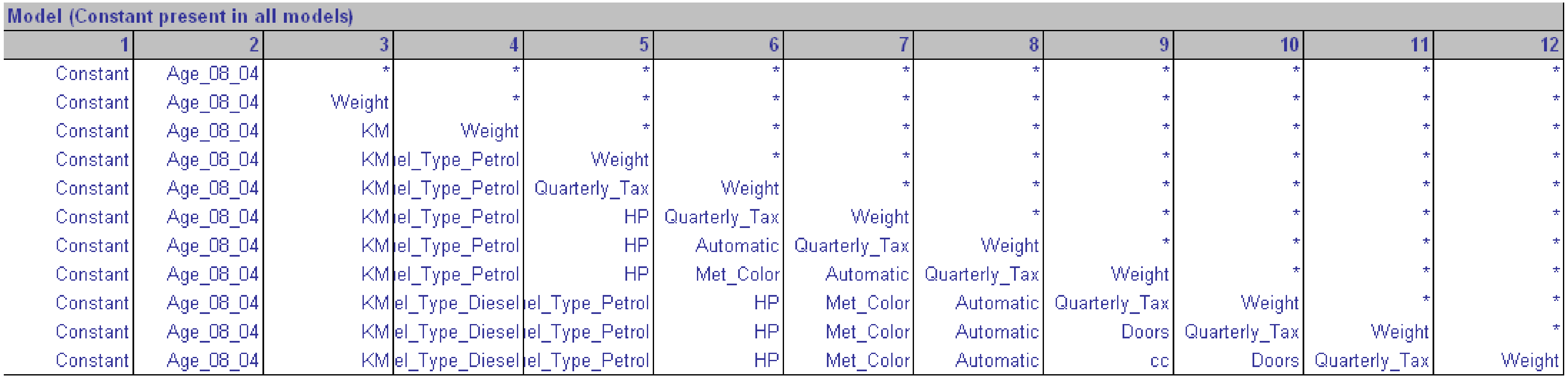


1	2	3	4	5	6	7	8
Constant	Age_08_04	*	*	*	*	*	*
Constant	Age_08_04	Weight	*	*	*	*	*
Constant	Age_08_04	KM	Weight	*	*	*	*
Constant	Age_08_04	KM	el_Type_Petrol	Weight	*	*	*
Constant	Age_08_04	KM	el_Type_Petrol	Quarterly_Tax	Weight	*	*
Constant	Age_08_04	KM	el_Type_Petrol	HP	Quarterly_Tax	Weight	*
Constant	Age_08_04	KM	el_Type_Petrol	HP	Automatic	Quarterly_Tax	Weight

Top model has a single predictor (Age_08_04)

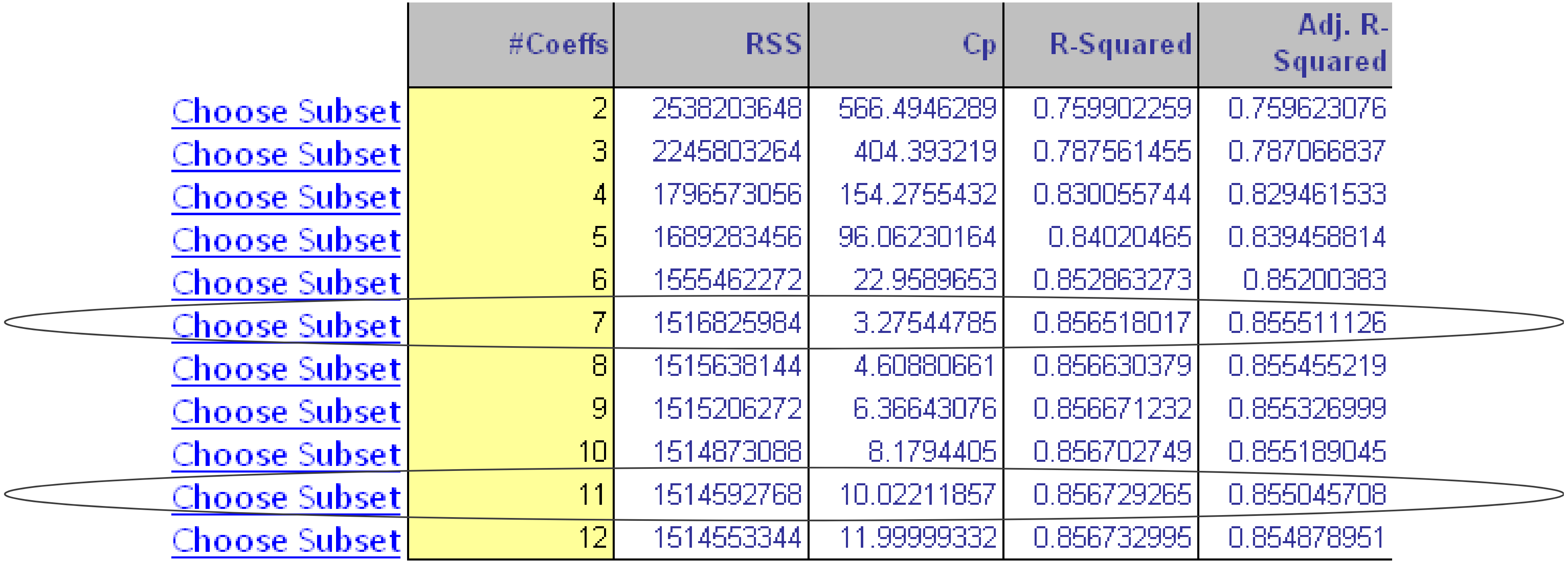
Second model has two predictors, etc.

All 12 Models



Model (Constant present in all models)											
1	2	3	4	5	6	7	8	9	10	11	12
Constant	Age_08_04	*	*	*	*	*	*	*	*	*	*
Constant	Age_08_04	Weight	*	*	*	*	*	*	*	*	*
Constant	Age_08_04	KM	Weight	*	*	*	*	*	*	*	*
Constant	Age_08_04	KM	el_Type_Petrol	Weight	*	*	*	*	*	*	*
Constant	Age_08_04	KM	el_Type_Petrol	Quarterly_Tax	Weight	*	*	*	*	*	*
Constant	Age_08_04	KM	el_Type_Petrol	HP	Quarterly_Tax	Weight	*	*	*	*	*
Constant	Age_08_04	KM	el_Type_Petrol	HP	Automatic	Quarterly_Tax	Weight	*	*	*	*
Constant	Age_08_04	KM	el_Type_Petrol	HP	Met_Color	Automatic	Quarterly_Tax	Weight	*	*	*
Constant	Age_08_04	KM	el_Type_Diesel	el_Type_Petrol	HP	Met_Color	Automatic	Quarterly_Tax	Weight	*	*
Constant	Age_08_04	KM	el_Type_Diesel	el_Type_Petrol	HP	Met_Color	Automatic	Doors	Quarterly_Tax	Weight	*
Constant	Age_08_04	KM	el_Type_Diesel	el_Type_Petrol	HP	Met_Color	Automatic	cc	Doors	Quarterly_Tax	Weight

Diagnostics for the 12 models



	#Coeffs	RSS	Cp	R-Squared	Adj. R-Squared
Choose Subset	2	2538203648	566.4946289	0.759902259	0.759623076
Choose Subset	3	2245803264	404.393219	0.787561455	0.787066837
Choose Subset	4	1796573056	154.2755432	0.830055744	0.829461533
Choose Subset	5	1689283456	96.06230164	0.84020465	0.839458814
Choose Subset	6	1555462272	22.9589653	0.852863273	0.85200383
Choose Subset	7	1516825984	3.27544785	0.856518017	0.855511126
Choose Subset	8	1515638144	4.60880661	0.856630379	0.855455219
Choose Subset	9	1515206272	6.36643076	0.856671232	0.855326999
Choose Subset	10	1514873088	8.1794405	0.856702749	0.855189045
Choose Subset	11	1514592768	10.02211857	0.856729265	0.855045708
Choose Subset	12	1514553344	11.99999332	0.856732995	0.854878951

Good model has:

High adj-R^2, low Cp, low # 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”)

Model with only 6 predictors

The Regression Model

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-3874.492188	1415.003052	0.00640071	97276411904
Age_08_04	-123.4366303	3.33806777	0	8033339392
KM	-0.01749926	0.00173714	0	251574528
Fuel_Type_Petrol	2409.154297	319.5795288	0	5049567
HP	19.70204735	4.22180223	0.00000394	291336576
Quarterly_Tax	16.88731384	2.08484554	0	192390864
Weight	15.91809368	1.26474357	0	281026176

Training Data scoring - Summary Report

Model Fit →

Total sum of squared errors	RMS Error	Average Error
1516825972	1326.521353	-0.000143957

Validation Data scoring - Summary Report

Predictive performance →

(compare to 12-predictor model!)

Total sum of squared errors	RMS Error	Average Error
1021510219	1334.029433	118.4483556

Summary 总结

□Linear regression models are very popular tools, not only for explanatory modeling, but also for prediction 线性回归对解释性模型和预测性模型都有用。

□A good predictive model has high predictive accuracy (to a useful practical level) 好的预测性模型有高准确度的预测结果。

□Predictive models are built using a training data set, and evaluated on a separate validation data set 预测性模型建立在训练数据集上并用验证数据集验证模型。

□Removing redundant predictors is key to achieving predictive accuracy and robustness 删掉多余预测因子对提高模型预测准确度和鲁棒性有关键性作用。

□Subset selection methods help find "good" candidate models. These should then be run and assessed. 预测因子子集的选择方法有助于找到好的模型。这些模型需要用在验证数据集上进行运行和评估。