CIS 8695

Linear Regression

Ling Xue

Computer Information System

Georgia State University

Key Issues

- Problems that can be addressed using linear regression
  - Explanatory vs. predictive modeling
- Linear regression model specification
- Explanations on estimation results
- Model fit and assessing predictive accuracy
- Selecting a subset of predictors

Explanatory vs Predictive Use of Regression

- **Explanatory Modeling**
- Fit the data well and understand the contribution of explanatory variables.
- Use entire dataset for analysis
- Model performance is measured by "goodness-of-fit": R^2, residual analysis, p-values.
- Explaining role of predictors is a primary purpose.

- **Predictive Modeling**
- Optimize predictive accuracy.
- Train model on training data, assess performance on validation (hold-out) data.
- Model performance is measured by predictive accuracy: total sum of squared errors (SSE), mean squared error (MSE), root mean squared error (RMSE).
- Explaining role of predictors is not primary purpose (but useful). Predicting target values in other data where we have predictor values, but not target values.

Problem: Prices of Toyota Corolla

- **Data:** Prices of 1442 used Toyota Corollas, with their specification information (CIS4930_ToyotaCorolla_Lab.xlsx)

(Original Sample)

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

ToyotaCorolla.xls

Problem: Prices of Toyota Corolla

- **Potential Questions from the dealer:**

1. Does mileage has an impact on price or not?

2. What is the relationship between price and # of doors, fuel_type, weight, and other car features? E.g., if the # of doors goes up, does the price go up or down?

3. If the age of a car goes up by 1 year, how much does the price increase or decrease?

4. How should I price a used 3-door, automatic, Toyota Corollas with 27 months in age, a mileage of 19700, 192 horsepower, and using Petrol?

Variables Used

- **Price** in Euros
- **Age** in months as of 8/04
- **KM** (kilometers)
- **Fuel Type** (diesel, petrol, CNG) → Converted to three dummies
  - Fuel_Type_CNG, Fuel_Type_Diesel, Fuel_Type_Petrol
- **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)

Linear Regression: Basic Idea

- **Linear regression** assumes the following relationship between predictors and target variable

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \varepsilon$$

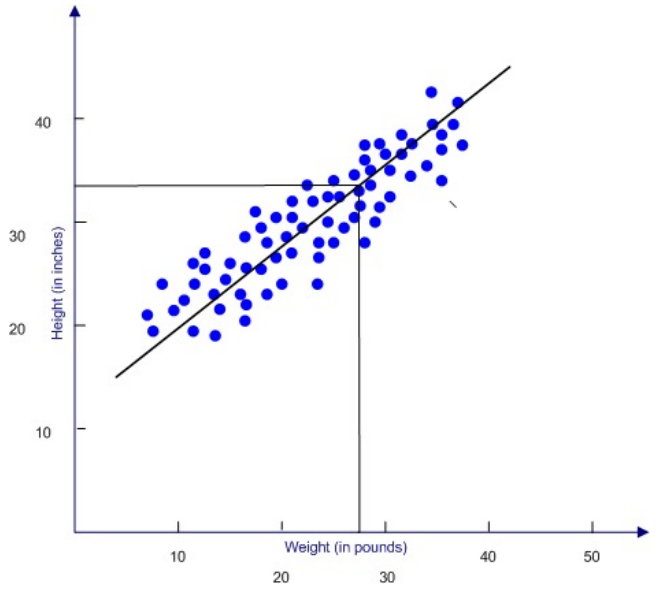
- It uses the following for **prediction**:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_q x_q$$

- Training: choose the **intercept** $\hat{\beta}_0$ and **coefficients** $\hat{\beta}_k, k = 1..q$ to minimize (least square):

$$\sum_{\substack{training \\ data}} (y_i - \hat{y}_i)^2$$

Fit a Linear Relationship

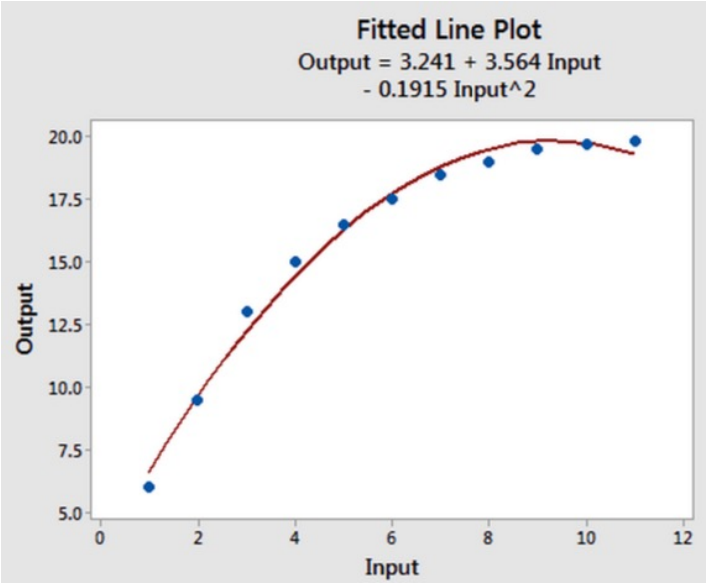


Issues to consider when using linear regression for prediction

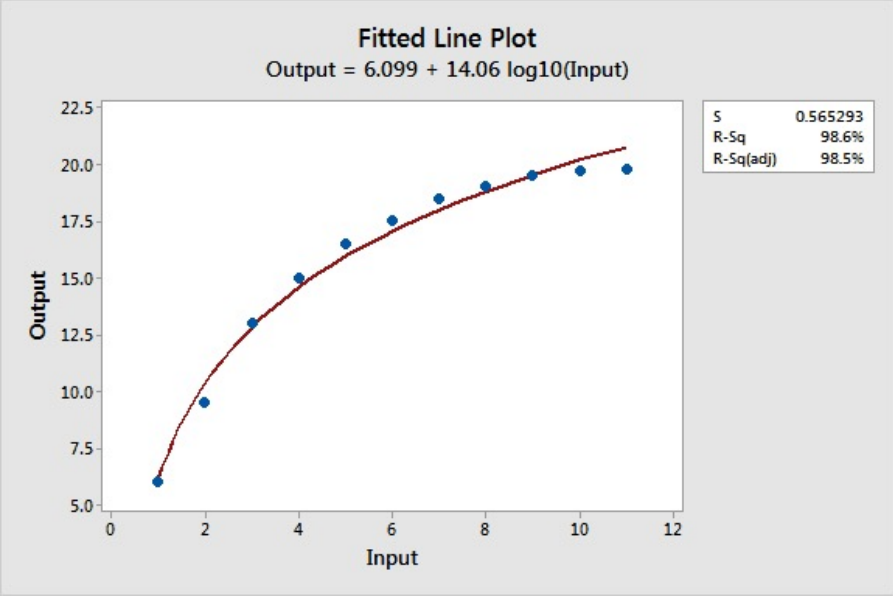
- Missing input values
  - At training time, what to do with missing values.
  - Skip? Mean replacement? Using estimated values?
- Categorical inputs
  - Need to be transformed into dummies
- Extreme values
  - May bias the model
  - Remove outlier cases

Issues to consider when using linear regression for prediction

- How to choose inputs
  - May consider variable selection (more later)
- What if the relationship is nonlinear?
  - May consider variable transformations



fit a non-linear relationship with a square term



fit a non-linear relationship with log-transformed variables.

Data Preprocessing

- Fuel type is categorical, must be transformed into binary variables
  - Add three dummies, but only use **two**: Fuel_Type_Diesel, Fuel_Type_CNG.
  - No dummy needed for “Petrol” (reference category)
- Data partition (training 60% validation 40%)
- Other considerations
  - Correlational analysis
  - Transformation (e.g., log)
  - Add square, interaction terms
  - Variable selection (more later)
  - Outlier removal

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

Regression Model

Input Variables	Coefficient	Std. Error	t-Statistic	P-Value	CI Lower	CI Upper	RSS Reduction
Intercept	-1833.3	1363.468725	-1.34458531	0.179	-4509.46	842.8602	99626055314
Age_08_04	-121.6074	3.320627169	-36.6218141	0.000	-128.125	-115.09	8611425804
KM	-0.019282	0.001699871	-11.3431708	0.000	-0.02262	-0.01595	264977161.1
HP	29.75675	4.224060494	7.044584823	0.000	21.46594	38.04756	207959275
Met_Color	101.1609	97.8079189	1.034280977	0.301	-90.8125	293.1342	567466.6314
Automatic	456.6505	199.4455759	2.289599302	0.022	65.18689	848.114	30074032.84
cc	-0.022305	0.091510026	-0.2437426	0.807	-0.20192	0.157307	26418651.34
Doors	-118.3781	50.63479867	-2.33788041	0.020	-217.762	-18.9942	11322463.67
Quarterly_Tax	11.75023	2.201777534	5.336700883	0.000	7.42867	16.07179	281862947.6
Weight	16.06755	1.414469773	11.35941787	0.000	13.29129	18.84382	256344071.7
Fuel_Type_CN	-2367.255	448.7481365	-5.27524242	0.000	-3248.04	-1486.47	34110235.19
Fuel_Type_Di	-1077.4	354.0357916	-3.04319608	0.002	-1772.29	-382.513	16040259.65

Error Reports

Training Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1472212319	1306.868	-2.17983E-12

Validation Data Scoring - Summary Report

Total sum of squared errors	RMS Error	Average Error
1033384804	1341.761	-58.72311001

Predicted Values

Predicted price computed using regression coefficients

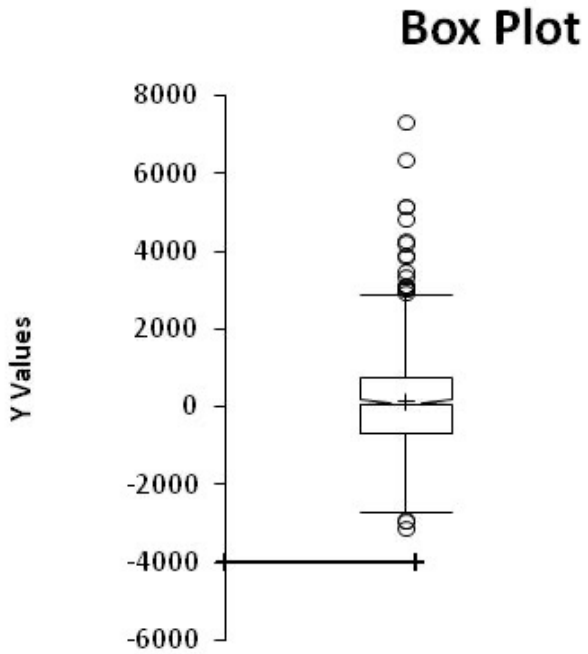
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_q x_q$$

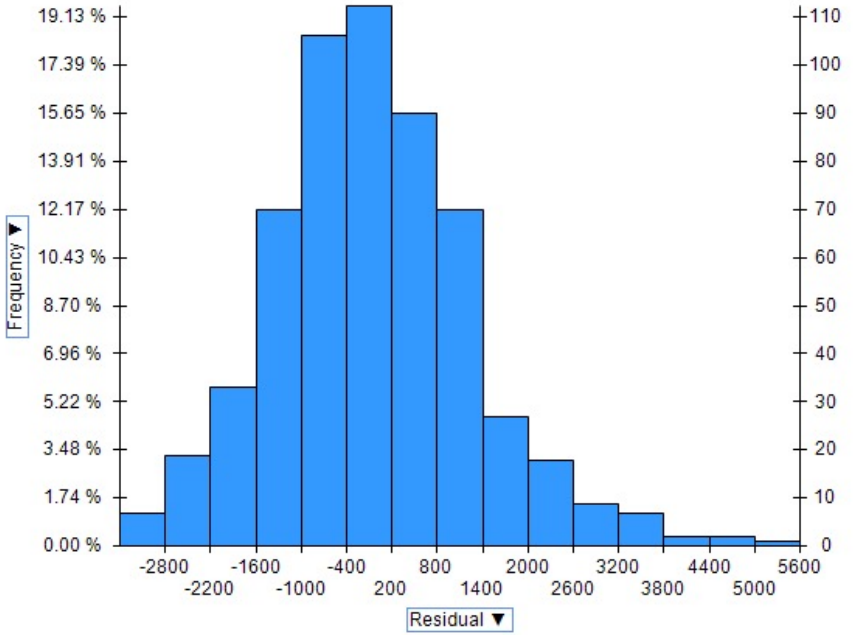
Residuals = difference between actual and predicted prices

$$e = Y - \hat{Y}$$

Predicted Value	Actual Value	Residual
12912.82	12450	-462.824
9536.681	10950	1413.319
18255.13	18500	244.8665
7148.686	7750	601.3145
9075.332	10500	1424.668
8260.328	9950	1689.672
11315.24	9980	-1335.24
6813.801	8950	2136.199
18585.64	19600	1014.357
10466.63	7900	-2566.63
7682.82	9900	2217.18
10770.61	10990	219.3929
10608.4	11950	1341.599
13114.49	12950	-164.488
8239.33	7500	-739.33

Distribution of Residuals





Symmetric distribution and some outliers on the right

Multicollinearity

- **Problem:** If one predictor is a linear combination of other predictor(s), model estimation will fail
  - Note that in such a case, we have at least one redundant predictor
- **Solution:** Remove extreme redundancies
  - Drop predictors via variable selection
  - Use data reduction methods such as PCA (Principal Component Analysis) (more details later)

Variable selection

- Reasons for variable selection: selecting a subset of predictors (rather than including them all):
  - *Parsimony*: simpler model is more robust; including inputs uncorrelated to response reduces predictive accuracy.
  - *Multicollinearity*: redundancy in inputs (two or more predictors share the same relationship with the response) can cause unstable results
- Exhaustive search
- Popular variable selection algorithms
  - Forward, backward, stepwise

Exhaustive Search

- All possible subsets of predictors assessed (single, pairs, triplets, etc.)
- Computationally intensive
- Judge by "adjusted R^2"

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

Penalty for number of predictors

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
- Alternative criterions may include:
  - Adjusted R-square (penalizes complex models)
  - F statistics
  - AIC (Akaike's information criterion), etc.

Forward Selection

	X1	X2	X3	X4	X5	X6	X7	X8	X9
1			√						
2	√		√						
3	√		√	√					
4	√		√	√		√			
5	√		√	√		√			√
6	√		√	√		√		√	√
7	√	√	√	√		√		√	√

Backward Elimination

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

Backward Elimination

	X1	X2	X3	X4	X5	X6	X7	X8	X9
1	√	√	√	√	√	√	√	√	√
2	√	√		√	√	√	√	√	√
3	√	√		√	√	√		√	√
4	√	√		√	√			√	√
5	√	√		√	√			√	
6	√	√			√			√	
7	√				√			√	

Stepwise

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

	X1	X2	X3	X4	X5	X6	X7	X8	X9
1	√								
2	√	√							
3	√	√	√						
4	√	√		√					
5	√	√		√	√				
6	√			√	√	√			
7	√			√	√		√		
8	√			√	√		√	√	
9	√			√	√		√		√

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, Cp = # predictors +1

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

Model Fit →

Training Data scoring - Summary Report

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

Predictive performance →

(compare to 12-predictor model!)

Validation Data scoring - Summary Report

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

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.

Additional In-class Exercise: Chapter Q2

- Explore SPENDING vs. FREQ, and SPENDING vs. LAST_UPDATE;
- To fit a predictive model for SPENDING:
  - Partition the 1000 records into training (60%) and validation (40%) sets;
  - Run a MLR for SPENDING versus all six predictors. Give the estimated predictive equation;
  - Based on this model, what type of purchaser is most likely to spend a large amount of money?
  - If we used backward elimination, which predictor would be dropped first?
  - Show how the prediction and the prediction error are computed for the first purchase in the validation set?
  - Evaluate the predictive accuracy on the validation set;
  - Create a histogram of the model residuals. Normal distribution?