MSCI152: Introduction to Business Intelligence and Analytics

Lecture 11: Multiple Linear Regression

Dr Anna Sroginis

Lancaster University Management School

Agenda

"All models are wrong, but some are useful" George Box

1. Recap
2. Categorical independent variables
3. Linear versus non-linear relationships
4. Forecasting with regression
5. Powerful but complex regression

More details can be found in Camm et al., Section 7.6, 7.7 & 7.10

Recap of the previous lecture

- Multiple linear regression gives an **average estimate of linear relation**:
  - $\hat{y}_j = b_0 + b_1 x_{1,j} + b_2 x_{2,j} + \ldots + b_{k-1} x_{k-1,j}$, where
    - y_j is a dependent (response) variable (the one that we want to model/predict);
    - $x_{1,j}$, $x_{2,j}$, ... , $x_{k-1,j}$ are independent variables (explanatory);
- As soon as you fit any model, you need to **validate this model**
- Confidence intervals help you to **test and interpret the coefficients**
- We can **measure the quality of a fit of a model**: Adjusted R^2 is better for multiple regression
- But we should be careful with **over/under fitting**

General Approach

1. Plot charts for each variable
   - As before, look for the shape of relationship and outliers
   - But, shape may be obscured by effect of other variables
2. Think what variables to include and how
3. Use Excel or stats package to fit regression equation
4. Validate your model
5. Use Excel output to assess the strength of relationship overall and for each variable (parameter estimation)
   - Any statistically insignificant or missing variables? Wrong specification?
6. Consider alternative models
   - We have to decide which variables to include, so there are lots of choices

Agenda

Qualitative predictors

So far, we have assumed that all variables are *quantitative*. But in practice, this is not always the case - often, some predictors are *qualitative*.

Let's consider this **credit card debt** dataset:

	A	B	C	D	E	F	G	H	I	J	K
1	Income	Limit	Rating	Cards	Age	Education	Own	Student	Married	Region	Balance
2	14.891	3606	283	2	34	11	No	No	Yes	South	333
3	106.025	6645	483	3	82	15	Yes	Yes	Yes	West	903
4	104.593	7075	514	4	71	11	No	No	No	West	580
5	148.924	9504	681	3	36	11	Yes	No	No	West	964
6	55.882	4897	357	2	68	16	No	No	Yes	South	331
7	80.18	8047	569	4	77	10	No	No	No	South	1151
8	20.996	3388	259	2	37	12	Yes	No	No	East	203
9	71.408	7114	512	2	87	9	No	No	No	West	872
10	15.125	3300	266	5	66	13	Yes	No	No	South	279
11	71.061	6819	491	3	41	19	Yes	Yes	Yes	East	1350

- Which variables are quantitative? Qualitative?
- What could be our response variable?

Qualitative predictors with only two levels

Suppose that we wish to investigate differences in **credit card balance** between those who **own a house** and those who **don't** (ignoring the other variables).

Definition

A qualitative predictor (also known as a factor) with two levels is called a **dummy** variable.

Based on the **OWN** variable, we can create a new variable that takes the form

$$x_i = \begin{cases} 1 \text{ if } i\text{th person owns a house} \\ 0 \text{ if } i\text{th person does not own a house,} \end{cases} \tag{1}$$

Qualitative predictors with only two levels

If we use this dummy variable as a predictor in a simple regression of this form:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{2}$$

We will get this:

$$y_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does not own a house,} \end{cases} \tag{3}$$

- β_0 is the average credit card balance among those who **do not own**,
- $\beta_0 + \beta_1$ is the average credit card balance among those who **do own their house**
- β_1 is the average **difference in balance** between owners and non-owners.

Credit card balance vs owning a house

Regression Statistics	
Multiple R	0.021474
R Square	0.0004611
Adjusted R Square	-0.0020503
Standard Error	460.22995
Observations	400

ANOVA

	df	SS	MS	F	Significance F
Regression	1	38891.914	38891.914	0.1836156	0.6685161
Residual	398	84301020	211811.61		
Total	399	84339912			

	Coefficients	tandard Erro	t Stat	P-value	Lower 95%	Upper 95%
Intercept	509.80311	33.128077	15.388853	2.909E-42	444.67522	574.931
Own	19.733123	46.05121	0.4285039	0.6685161	-70.8009	110.26715

- What is the regression equation?
- What do you think about this model?

Qualitative predictors with more than two levels

When a qualitative predictor has **more than two levels**, a single **dummy** variable cannot represent all possible values. For example, **REGION** variable has three values: West, East and South.

$$x_{i1} = \begin{cases} 1 \text{ if } i\text{th person from the South} \\ 0 \text{ if } i\text{th person is not from the South,} \end{cases} \tag{4}$$

$$x_{i2} = \begin{cases} 1 \text{ if } i\text{th person from the West} \\ 0 \text{ if } i\text{th person is not from the West,} \end{cases} \tag{5}$$

- We always need to create
  - 2 dummies for a 3-level variable;
  - 3 dummies for a 4-level variable etc.
- There will always be one fewer dummy variable than the number of levels.

Qualitative predictors with more than two levels

So we add these dummy variables:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \tag{6}$$

We will get this:

$$y_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i \text{ if } i\text{th person is from the South} \\ \beta_0 + \beta_2 + \epsilon_i \text{ if } i\text{th person is from the West} \\ \beta_0 + \epsilon_i \text{ if } i\text{th person is from the East} \end{cases} \tag{7}$$

- β_0 is the average credit card balance for people from the **East**,
- β_1 is the difference in the average balance people from the **South versus the East**
- β_2 is the difference in the average balance people from the **West versus the East**.

Credit card balance vs region

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.0147921
R Square	0.0002188
Adjusted R Sc	-0.0048179
Standard Err(	460.86508
Observations	400

ANOVA

	df	SS	MS	F	Significance F
Regression	2	18454.2	9227.1002	0.0434428	0.9574919
Residual	397	84321458	212396.62		
Total	399	84339912			

	Coefficients	tandard Erro	t Stat	P-value	Lower 95%	Upper 95%
Intercept	531	46.318683	11.464057	1.774E-26	439.93944	622.06056
South	-12.502513	56.681038	-0.2205766	0.8255355	-123.93502	98.929995
West	-18.686275	65.021075	-0.287388	0.7739652	-146.51494	109.14239

- What is the regression equation?
- What do you think about this model?

Exercise

Remember this **credit card debt** dataset?

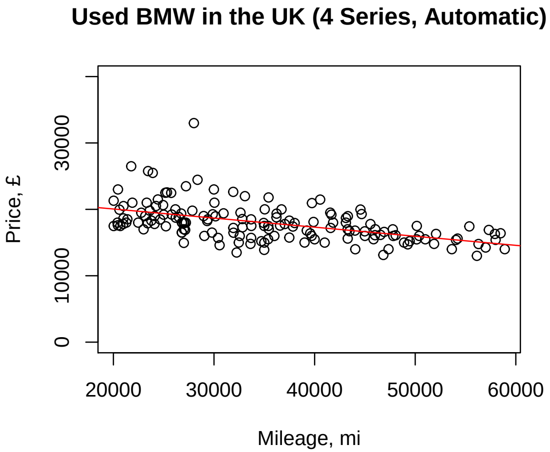
	A	B	C	D	E	F	G	H	I	J	K
1	Income	Limit	Rating	Cards	Age	Education	Own	Student	Married	Region	Balance
2	14.891	3606	283	2	34	11	No	No	Yes	South	333
3	106.025	6645	483	3	82	15	Yes	Yes	Yes	West	903
4	104.593	7075	514	4	71	11	No	No	No	West	580
5	148.924	9504	681	3	36	11	Yes	No	No	West	964
6	55.882	4897	357	2	68	16	No	No	Yes	South	331
7	80.18	8047	569	4	77	10	No	No	No	South	1151
8	20.996	3388	259	2	37	12	Yes	No	No	East	203
9	71.408	7114	512	2	87	9	No	No	No	West	872
10	15.125	3300	266	5	66	13	Yes	No	No	South	279
11	71.061	6819	491	3	41	19	Yes	Yes	Yes	East	1350

- To include all qualitative variables in the regression, how many dummy variables do you need to create in total?

Agenda

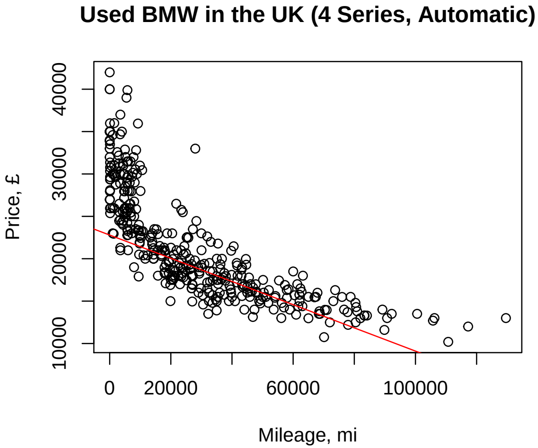
Linear relationships

- Do you remember my favourite example?



Non-linear relationships

- If we take the whole available sample, we see a non-linear pattern in this data:

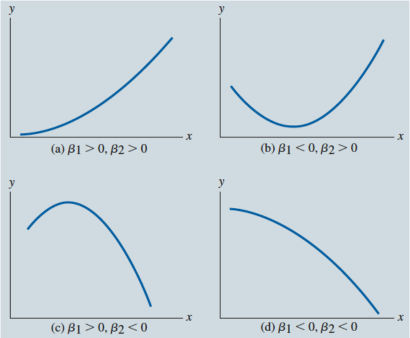


Polynomial linear regression

A simple approach for incorporating non-linear associations in a linear model is to include **transformed** versions of the predictors.

- One of the ways to model non-linear relationships is **polynomial** regression. The simplest form is to add a **quadratic** shape:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \tag{8}$$



But it is still a linear model!

Quadratic linear regression

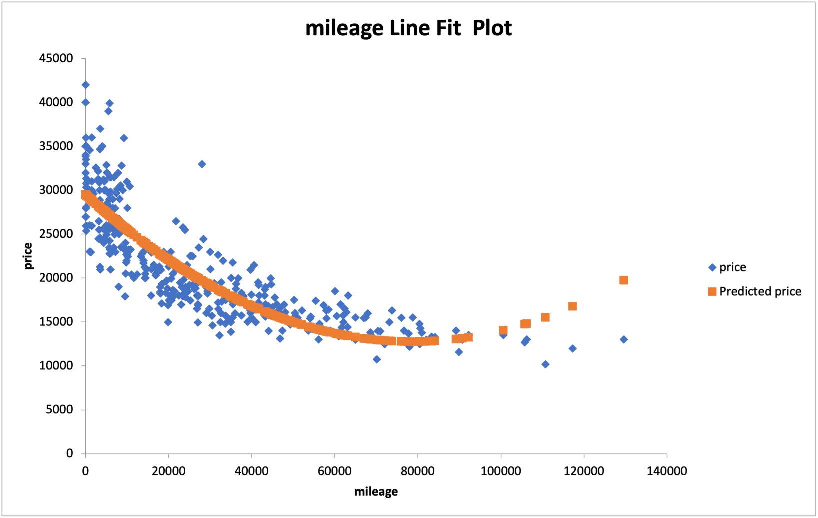
- Remember to include initial term!
- In business analytics applications, polynomial regression models of higher than second (x_i^2) or third-order (x_i^3) are rarely used. Be careful and always justify these transformations!
- There are alternatives for certain cases.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.85347226
R Square	0.72841489
Adjusted R Sc	0.72695867
Standard Errc	3361.9783
Observations	376

	Coefficients	*Standard Erroi*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	29497.7662	322.581577	91.4428109	1.605E-257	28863.4597	30132.0726
mileage	-0.4253346	0.01872869	-22.71032	2.5624E-72	-0.4621617	-0.3885076
mileage^2	2.7021E-06	2.0798E-07	12.9921274	4.265E-32	2.2931E-06	3.1111E-06

Used BMW example: quadratic variable



- What do you think about this fit?

Agenda

Predicting with regression

Once we have estimated a regression line, we can use it to forecast by simply using numerical inputs for the various variables in the model.

But we need to have x_i:

- X is **known ahead of time** (e.g., the size of a sales force, or demographic details concerning a consumer).
- X is **unknown** but can still be **forecast** (e.g., gross domestic product).
- X is **unknown**, but we wish to make **what-if** forecasts (e.g., the effects of different advertising or pricing policies).

To calculate a **point prediction**, substitute the given values of the X's into the estimated regression equation.

Source: https://otexts.com/fpp2/forecasting-regression.html

Prediction intervals

A **prediction interval** is an interval estimate of an individual y value given values of the independent variables.

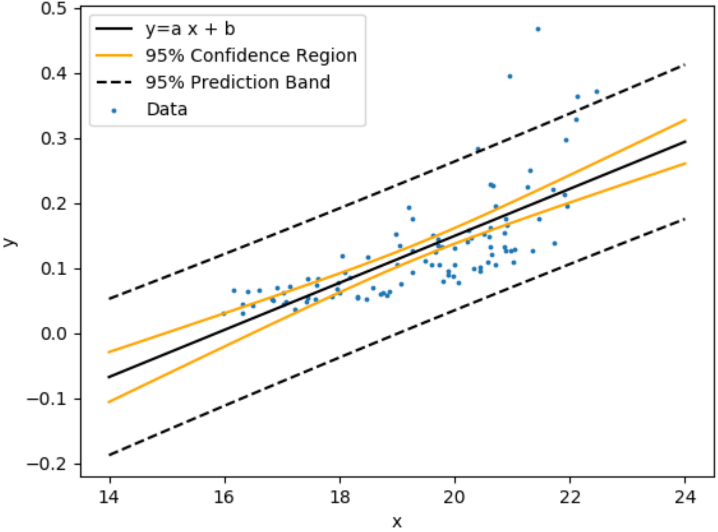
An approximate 95% prediction interval associated with this forecast is given by

$$\hat{y} \pm 1.96\hat{\sigma}_e\sqrt{1 + \frac{1}{T} + \frac{(x - \bar{x})^2}{(T - 1)s_x^2}},$$

where

- T is the total number of observations
- $\bar{x}$ is the mean of the observed x values
- s_x is the standard deviation of the observed x values
- σ_e is the standard error of the regression

Prediction intervals



Agenda

1. Recap
2. Categorical independent variables
3. Linear versus non-linear relationships
4. Forecasting with regression
5. Powerful but complex regression

Other regression features

- Other types of regression
  - Log-regression, logistic regression, multi-level regressions etc.
- Interaction between independent variables
  - $Sales = b_0 + b_1 Price + b_2 Advertising + b_3 Price \cdot Advertising$
- There are different variable selection methods
  - Backward elimination, forward selection, stepwise selection, best subsets

Don't worry about the details:

- You will not be asked about this in the exam or the coursework.

Just giving you an idea of the range of things you can do with regression!

A couple more words about regression

We briefly discussed the main assumptions to validate your model, but there are also other issues that you might have:

- Outliers
- Omitted variables
  - underfitting the data - we don't explain the structure well
- Redundant variables
  - overfitting the data - we explain the noise
- Multicolleniarity
  - two or more predictor variables are closely related to one another - avoid it!
- Heteroscedasticity
  - the variances of the error terms are non-constant - always plot your residuals!

Wrap up

Here we:

- Modelling relationships between two and more variables: **Multiple linear regression**

Next time:

- **Introduction to forecasting**