Mathematics Learning Support Centre

Simple Linear Regression: Introduction

Richard Buxton. 2008.

1 Introduction

We often want to predict, or explain, one variable in terms of others.

- How does a household's gas consumption vary with outside temperature?
- How does the crime rate in an area vary with differences in police expenditure, unemployment, or income inequality?
- How does the risk of heart disease vary with blood pressure?

Regression modeling can help with this kind of problem.

The aim of this handout is to introduce the simplest type of regression modeling, in which we have a single predictor, and in which both the response variable - e.g. gas consumption - and the predictor - e.g. outside temperature - are measured on numerical scales.

2 Model for simple linear regression

Figure 1 (a) shows a scatterplot of gas consumption and average outside temperature for 26 one-week periods[1].

As we'd expect, higher outside temperatures tend to be associated with lower gas consumption. The relationship between the two variables can be approximated roughly with a straight line - see Figure 1 (b) - and we could use this fitted line to predict the expected gas consumption for any given outside temperature.

But even with a very strong relationship, as here, there's still some variation in gas consumption that can't be accounted for by our linear model - the gas consumption sometimes lies above the line and sometimes below. In simple linear regression, we take account of this unexplained variation by using a model of the form...

$$G = \beta_0 + \beta_1 T + \epsilon$$

... where G is the gas consumption, T is the temperature and ϵ represents the unexplained variation.

[1] Source of data: Hand (1994)

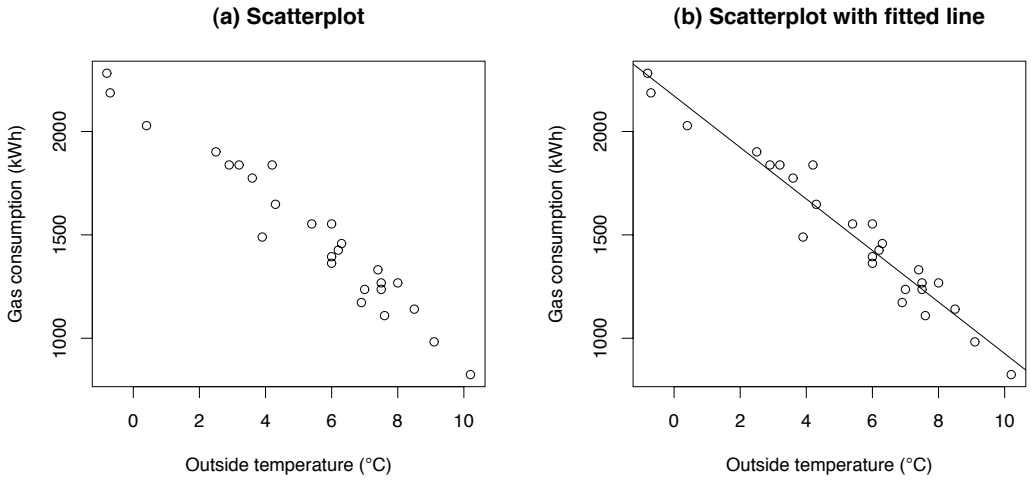

Figure 1: Gas consumption vs Temperature

We can't predict the size or direction of the ϵ's, but we can say something about how large they're likely to be. Looking at Figure 1 (b), a discrepancy from the line of say 50kWh would seem to be quite normal, but a discrepancy as large as 500kWh would be very surprising. In simple linear regression,we assume that the ϵ's vary according to a Normal distribution.

3 Fitting the model

Before we can use our model to make predictions, we need to *estimate* the coefficients β_0 and β_1. We do this by fitting a line to our data, using the criterion of *least squares*. The idea is to choose the line that minimizes the sum of the squares of the distances between the observed values of the response (gas consumption) and the values predicted by the model. Any statistical software will carry out the required calculations. Table 1 shows an extract from the SPSS output for the Gas data.

Coefficients

Model		Unstandardized Coefficients		t	Sig.
		B	Std.Error		
1	(Constant)	2172.174	37.532	57.876	.000
	T	-124.629	6.207	-20.078	.000

Table 1: SPSS output for Gas data

The coefficients are contained in the column headed 'B'. Rounding the figures to the nearest whole number, the fitted model is...

$$G = 2172 - 125\ T$$

Notice that the coefficient of T is negative, reflecting the fact that higher temperatures are associated with lower gas consumption.

4 Using the model

Once we've fitted a model, we can use it to make predictions - e.g. to predict the gas consumption corresponding to an outside temperature of 6 deg C, or the reduction in gas consumption corresponding to a 5 deg C *increase* in temperature.

For a temperature of 6 deg C, we predict a gas consumption of ...

$$\begin{aligned} G &= 2172.174 - (124.629 * T) \\ &= 2172.174 - (124.629 * 6) \\ &\simeq 1424 \text{ kWh} \end{aligned}$$

This figure gives us a rough idea of the gas consumption, but it is subject to some uncertainty - the actual consumption may be a bit higher, or lower, than our estimate suggests.

To predict the *reduction* in gas comsumption corresponding to a given *increase* in temperature, we need to look at the slope of our regression line.

Looking at Table 1, wee see that the coefficient of T is -124.629. This tells us that an increase of 1 deg C in the temperature is associated with a reduction in gas consumption of around 124.629kWh.

To predict the reduction in gas consumption corresponding to say a 5 deg C increase in temperature, we multiply by 5 to give a reduction of about $(5 \times 124.629) = 623.145kWh$.

5 How reliable are our predictions?

As we mentioned in Section 4, our predictions are subject to some uncertainty. This uncertainty arises because not all of the variation in gas consumption can be explained by outside temperature.

By making some assumptions about the unexplained variation, we can derive a *confidence interval*, or range of plausible values, for a prediction. For details, see the leaflet 'Linear Regression: Reliability of predictions'.

6 Simple Linear Regression in SPSS

The first step in Simple Linear Regression is to draw a scatter plot of the data.

—**Graphs**
—**Chart Builder**
—choose **Scatter/Dot**
—drag the **Simple Scatter** plot into the plotting region
—drag the response variable into the **Y-Axis** box
—drag the predictor variable into the **X-Axis** box
—Click **OK**

Does the relationship between response and predictor look roughly linear? If not, you'll need to think about using a more flexible model, such as a quadratic.

To fit a linear model, we use the *Regression* procedure.

—**Analyse**
—**Regression**
—**Linear**
—drag the response variable into the **Dependent** box
—drag the predictor variable into the **Independent(s)** box
—Click **OK**

7 References

For a clear introduction to regression analysis, see Moore and McCabe (2004).

Hand, D.J. (1994). A Handbook of Small Data Sets, Chapman and Hall.

Moore, D.S. and McCabe, G.P. (2004). Introduction to the practice of statistics, 5th edition, W.H.Freeman.