Mathematics Learning Support Centre

Statistics: Regression

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 changes in the 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 a person contracting heart disease vary with their blood pressure?

Regression modeling can often 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.

In Section 8, we explain how simple linear regression can be generalized to deal with situations involving multiple predictors and categorical variables.

For details of how to fit a simple linear regression model in SPSS, see separate handout.

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 simple linear model - the gas consumption

[1]Source of data: Hand (1994)

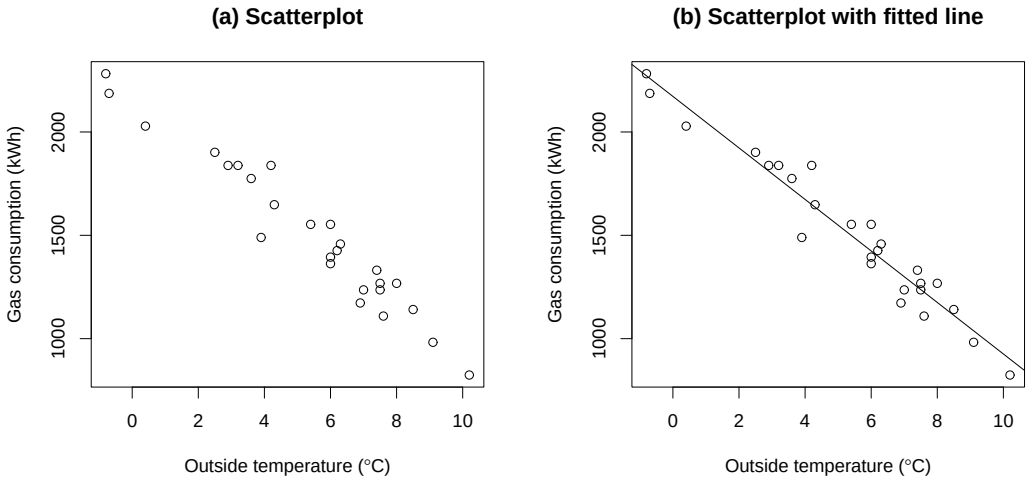

Figure 1: Gas consumption vs Temperature

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.

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 regresion,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.

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.

Coefficients

Model		Unstandardized Coefficients		t	Sig.
		B	Std.Error		
1	(Constant)	2172.174	37.532	57.876	.000
	T	-124.629	6.207	-.971	.000

Table 1: SPSS output for Gas data

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.

By making some assumptions about the unexplained variation, we can quantify the uncertainty and calculate a confidence interval, or range of plausible values, for the gas consumption.

5 Assumptions of simple linear regression

We make the following assumptions...

- Mean response varies linearly with predictor
- Unexplained variation is Normally and independently distributed with constant variance

To check these assumptions, we look at plots of the *residuals* and *fitted values*. The fitted values are the values of the response predicted by the model. The residuals are obtained by taking the observed values of the response and subtracting the fitted values. The two most useful plots are...

- Plot of Residuals vs Fitted values
  - We can use this plot to check the assumptions of linearity and constant variance. For example, Figure 2 shows some plots for a regression model relating stopping distance to speed[2]. The plot on the left shows the data, with a fitted linear model. The plot on the right shows the residuals plotted against the fitted values - a smooth curve has been added to highlight the pattern of the plot.

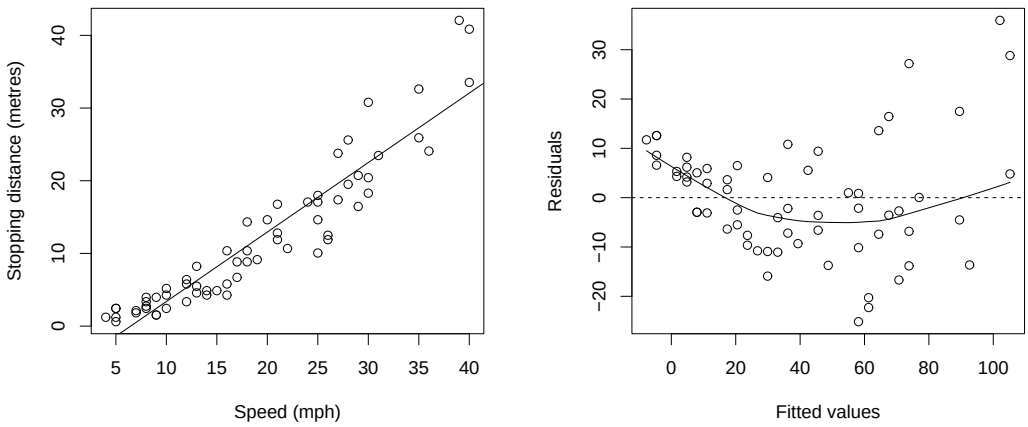

Figure 2: Stopping distance vs Speed

Ideally, the residual plot should show a horizontal band of roughly equal width. In this case, we have a strong 'U' shape, suggesting that the residuals go from positive to negative to positive. This suggests that we're fitting a line to a non-linear relationship - see plot of original data. In addition, the width of the band of data increases from the left to the right, suggesting that the variance is increasing. There are various courses of action that we can take to deal with these problems - for details, consult a Statistician.

- Normal probability plot of residuals
  - This plot is used to check the assumption that the unexplained variation follows a Normal distribution. If the ϵ's are roughly Normal, this plot should be roughly linear. Any strong systematic curvature suggests a non-Normal distribution. Figure 3 shows a Normal plot for the Gas data. The plot seems to be roughly linear, suggesting that there is no evidence of non-Normality.

There are several ways of checking the assumption that the random errors, or ϵ's are statistically independent. For details, see Koop (2008). The assumption of independence is not usually a problem except for data that has been collected at successive points in time - e.g. monthly unemployment figures.

[2]Source of data: Hand (1994)

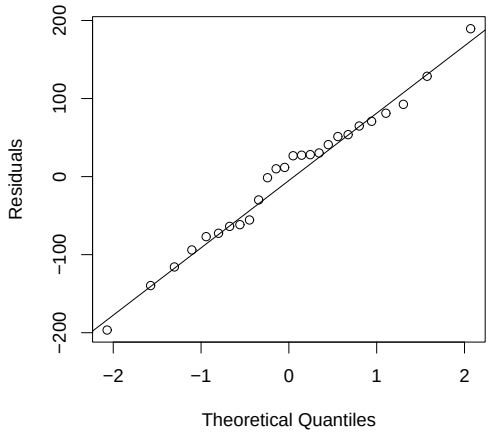

Figure 3: Normal probability plot for gas data

6 Confidence intervals for predictions

Provided the assumptions in Section 5 are satisfied, we can obtain confidence intervals for any predictions that we make. There are two types of interval...

- Confidence interval for individual case
  - Range of plausible values for a single case - e.g. for the gas consumption in a single 1-week period
- Confidence interval for mean
  - Range of plausible values for the mean - e.g. for the mean gas consumption over a large number of 1-week periods, all with the same average outside temperature

In Section 4, we made a prediction of gas consumption for a week in which the outside temperature is 6 deg C. To put a confidence interval on this prediction, we use the SPSS output in Table 2.

Temp	PRE_1	LMCI_1	UMCI_1	LICI_1	UICI_1
6.00	1424	1387	1461	1237	1612

Table 2: Prediction and confidence intervals

If we're predicting the gas consumption for a single week, we use the columns headed LICI_1 (Lower Individual Confidence Interval) and UICI_1 (Upper Individual Confidence Interval).

95% confidence interval 1237 to 1612 kWh

We can be fairly sure that the gas consumption will lie within this range.

If we wished to predict the *mean* gas consumption over a large number of weeks in which the temperature was 6 deg C, we would use the columns LMCI_1 and UMCI_1.

95% confidence interval for mean 1387 to 1461 kWh

This interval is much narrower. We're much less sure about the gas consumption in a single week than we are about the mean consumption over a large number of weeks.

7 Confidence interval for slope of regression model

We're sometimes interested in the change in the response corresponding to a given change in the predictor. For example, how much will our stopping distance increase if we travel 10mph faster? We can answer this kind of question by looking at the *slope* of the regression line.

Figure 3 shows some SPSS output giving the coefficients of the Gas model, togethre with confidence intervals for both the slope and intercept.

Coefficients

Model		Unstandardized Coefficients		95% Conf Int for B	
		B	Std.Error	Lower Bound	Upper Bound
1	(Constant)	2172.174	37.532	2094.712	2249.636
	T	-124.629	6.207	-137.440	-111.817

Table 3: Confidence intervals for coefficients

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. The columns on the right of the table give a confidence interval for this figure.

95% confidence interval 111.817 to 137.440 kWh

This gives us a range of plausible values for the reduction in gas consumption corresponding to an increase of 1 deg C in the temperature.

If we're interested in the change in gas usage corresponding to say a 5 deg C increase in temperature, we can obtain a confidence interval by simply multiplying the lower and upper ends of our confidence interval by 5 to give...

95% confidence interval 559.085 to 687.200 kWh

8 Extending simple linear regression

This section indicates some of the ways in which simple linear regression can be extended to model more complex behaviour. For more details, see Freund and Wilson (1998).

- Multiple regression
  - In multiple regression, we can introduce several predictors, rather than just one. For example, in trying to explain variation in the crime rate in different cities, we might use a model of the form...

    $$R = \beta_0 + \beta_1 Ed + \beta_2 Ex + \beta_3 U + \epsilon$$

    ..., where R is the crime rate, Ed is a measure of educational level, Ex is police expenditure and U is unemployment.

    Fitting a multiple regression model is quite simple, but interpreting the fitted model can be quite challenging. There are often strong near dependencies among the predictors and this can make it difficult to separate out the effect of each individual predictor.
- Allowing for curvature
  - We can sometimes allow for curvature by introducing a squared term into our model. The following model allows for some curvature in the relationship between Expired ventilation (E) and Oxygen uptake (O).

    $$E = \beta_0 + \beta_1 O + \beta_2 O^2 + \epsilon$$
- Categorical predictors
  - We can represent a categorical predictor by the use of *dummy* variables. For example, suppose we have some data on gas consumption and temperature for several weeks before and after the installation of roof insulation. The period of the data - before or after insulation - is a categorical variable with two possible values. We introduce a dummy variable I which takes the value 0 for the *before* and 1 for *after*. We can now extend our earlier model to...

    $$G = \beta_0 + \beta_1 T + \beta_2 I + \epsilon$$

    Our estimate of the coefficient β_2 will give us an estimate of the change in gas consumption resulting from the insulation.
- Allowing for interactions
  - Sometimes the effect of one predictor will vary according to the setting of the other one. In the example on roof insulation, the effect of temperature on gas consumption may well become smaller after the insulation has been installed - i.e. the effect of the predictor T will be less when $I = 1$ than when $I = 0$.

We can allow for an interactive effect by introducing a cross-product term in T and I to give...

$$G = \beta_0 + \beta_1 T + \beta 2I + \beta_3 TI + \epsilon$$

The cross-product term allows the coefficient of T to vary according to the setting of I.

- Categorical response
  - In a study of the relationship between heart disease and blood pressure, the response is a binary categorical variable with the two values 'Patient has heart disease' and 'Patient does not have heart disease'. This kind of problem can be handled using a technique known as *logistic* regression. For a brief introduction, see Freund and Wilson (1998).

    Sometimes, we have an ordinal response, with more than two categories. For example, the quality of a patient's life following treatment may take on the values 'Excellent', 'Good', 'Fair' or 'Poor'. This kind of problem can be handled by a related technique, known as ordinal logistic regression. For details, see Agresti (2002).

9 References

For a simple *introduction* to regression, see Moore and McCabe (2004). For a more comprehensive treatment, see Freund and Wilson (1998).

Agresti, A. (2002). Categorical data analysis, Wiley.

Freund, R.J. and Wilson, W.J. (1998). Regression Analysis: statistical Modeling of a Response Variable, Academic Press.

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

Koop, G. (2008). Introduction to Econometrics, Wiley.

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