4.6 Statistical inference

otexts.org/fpp/4/6

(This section is an optional digression.)

As well as being used for forecasting, simple linear regression models are also valuable in studying the historical effects of predictors. The analysis of the historical effect of a predictor uses statistical inference methods.

Hypothesis testing

If you are an analyst then you may also be interested in testing whether the predictor variable xx has had an identifiable effect on yy. That is, you may wish to explore whether there is enough evidence to show that xx and yy are related.

We use statistical hypothesis testing to formally examine this issue. Ifxx and yy are unrelated, then the slope parameter β1=0β1=0. So we can construct a test to see if it is plausible that β1=0β1=0 given the observed data.

The logic of hypothesis tests is to assume the thing you want to disprove, and then to look for evidence that the assumption is wrong. In this case, we assume that there is no relationship between xx and yy. This is called the “null hypothesis” and is stated as

H0:β1=0.H0:β1=0.

Evidence against this hypothesis is provided by the value of β^1β^1, the slope estimated from the data. If β^1β^1 is very different from zero, we conclude that the null hypothesis is incorrect and that the evidence suggests there really is a relationship between xx and yy.

To determine how big the difference between β^1β^1 and β1β1 must be before we would reject the null hypothesis, we calculate the probability of obtaining a value of β^1β^1 as large as we have calculated if the null hypothesis were true. This probability is known as the “P-value”.

The details of the calculation need not concern us here, except to note that it involves assuming that the errors are normally distributed. R will provide the P-values if we need them.

In the car fuel example, R provides the following output.

R output

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.525647   0.199232   62.87   <2e-16 ***
City        -0.220970   0.008878  -24.89   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Residual standard error: 0.4703 on 132 degrees of freedom
Multiple R-squared: 0.8244, Adjusted R-squared: 0.823
F-statistic: 619.5 on 1 and 132 DF, p-value: < 2.2e-16

The column headed `Pr>|t|` provides the P-values. The P-value corresponding to the slope is in the row beginning `City` and takes value `<2e-16` . That is, it is so small, it is less than 2×10−16=0.00000000000000022×10−16=0.0000000000000002. In other words, the observed data are extremely unlikely to have arisen if the null hypothesis were true. It is much more likely that there is a relationship between city fuel consumption and the carbon footprint of a car. (Although, as we have seen, that relationship is more likely to be non-linear than linear.)

The asterisks to the right of the P-values give a visual indication of how small the P-values are. Three asterisks correspond to values less than 0.001, two asterisks indicate values less than 0.01, one asterisk means the value is less than 0.05, and so on. The legend is given in the line beginning `Signif. codes` .

There are two other P-values provided in the above output. The P-value corresponding to the intercept is also <2×10−16<2×10−16; this P-value is usually not of great interest --- it is a test on whether the intercept parameter is zero or not. The other P-value is on the last line and, in the case of simple linear regression, is identical to the P-value for the slope.

Confidence intervals

It is also sometimes useful to provide an interval estimate forβ1β1, usually referred to as a confidence interval (and not to be confused with a forecast or prediction interval). The interval estimate is a range of values that probably contain β1β1. In the car fuel example, R provides the following output.

R output

```
> confint(fit,level=0.95)
                2.5 %     97.5 %
(Intercept) 12.1315464 12.9197478
City        -0.2385315 -0.2034092
```

So if the linear model is a correct specification of the relationship between xx and yy, then the interval [−0.239,−0.203][−0.239,−0.203] contains the slope parameter, β1β1, with probability 95%. Intervals for the intercept and for other probability values are obtained in the same way.

There is a direct relationship between P-values and confidence intervals. If the 95% confidence interval for β1β1 does not contain 0, then the associated P-value must be less than 0.05. More generally, if the 100(1−α)%100(1−α)% confidence interval for a parameter does not contain 0, then the associated P-value must be less than ααα.