

Emergency calls to the New York Auto Club

Topics covered: Prediction. Residual plots. Simple regression. Tests of hypotheses.

Key words: F -statistic. Fitted value. Normal plot. p -value. Prediction interval. R^2 . Residual. Scatter plot. Type I error.

Data File: ersRev.dat

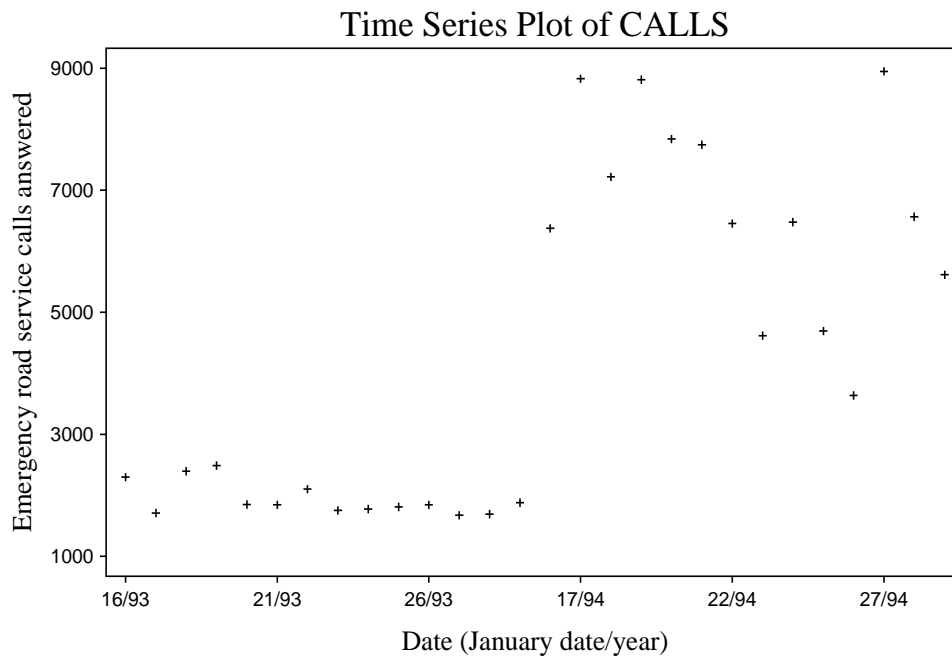
The Automobile Club of New York provides many services to its members, including travel planning, traffic safety classes and discounts on insurance. The service with the highest profile is its Emergency Road Service (ERS). If a club member's car breaks down, the member can call the Club to send out a tow truck for assistance. This service is especially useful in the winter months, when Club members can be stranded with frozen locks, dead batteries, weather induced accidents and spinning tires.

If the weather is very bad, the Club can be overwhelmed with calls. By tracking the weather conditions the Club can divert resources from other Club activities to the ERS for projected peak days. This will lead to better service for Club members and also greatly reduce stress on the Club staff.

Are the number of calls that the Club will receive in a day predictable from the weather forecast given on the previous day?

We will investigate this with data from the second-half of January in 1993 and 1994. The Club reports the number of ERS calls answered each day (CALLS). We have also recorded the lowest temperature for the previous night (LOW), the highest temperature for the day (HIGH), and forecasts of both of these from the previous day (FLOW and FHIGH, respectively).

Here is a **time series** plot of CALLS for these two periods:

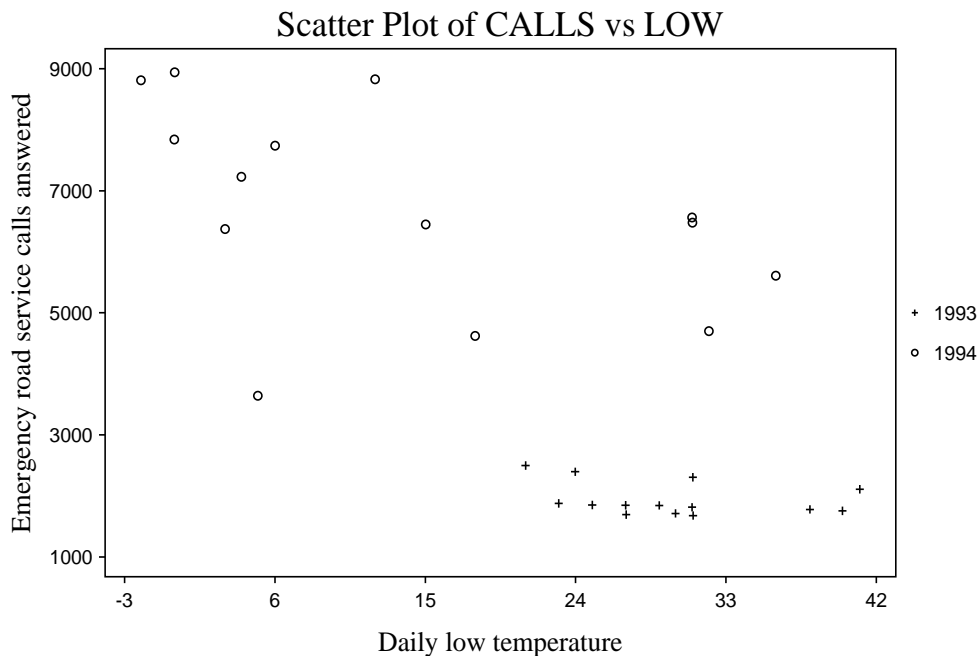


The first half of the plot (16/93 – 29/93) refers to values for 1993 and the latter half (16/94 – 29/94) refers to the corresponding period in 1994.

Several things are apparent from the plot. First, the number of calls in 1994 is much larger than the number of calls in 1993. The general level of calls is greater; the smallest number of calls in 1994 is far greater than the largest number of calls in 1993. Second, the variation in the number of calls from day to day in 1994 is far greater than the variation in the number of calls in 1993.

What could cause these differences? An obvious possibility is the low temperature for the day. In January, cold arctic air can sweep over the New York area bringing snow and freezing rain. Below freezing temperatures also reduce the effectiveness of car batteries.

Let's look at a scatter plot of the relationship between the number of calls and the low temperature:



The values for 1993 have been represented by a + and those in 1994 have been represented by a o. It is clear that the temperature differences between the two years explain much of the difference in the number of calls. As the temperature decreases, the number of calls tends to increase. However, there is still a lot of variation.

Another factor is the cumulative effects of previous bad weather. Starting in mid-December 1993 a succession of heavy snow storms hit the New York area. The accumulated snow and ice did not melt away because the temperatures remained below normal. These cumulative effects, in addition to the daily low temperatures, could lead to a dramatic increase in the number of calls. In fact, this period in 1994 saw the largest number of calls in Club history: 167,492 in January 1994, compared with 64,132 in January 1993.

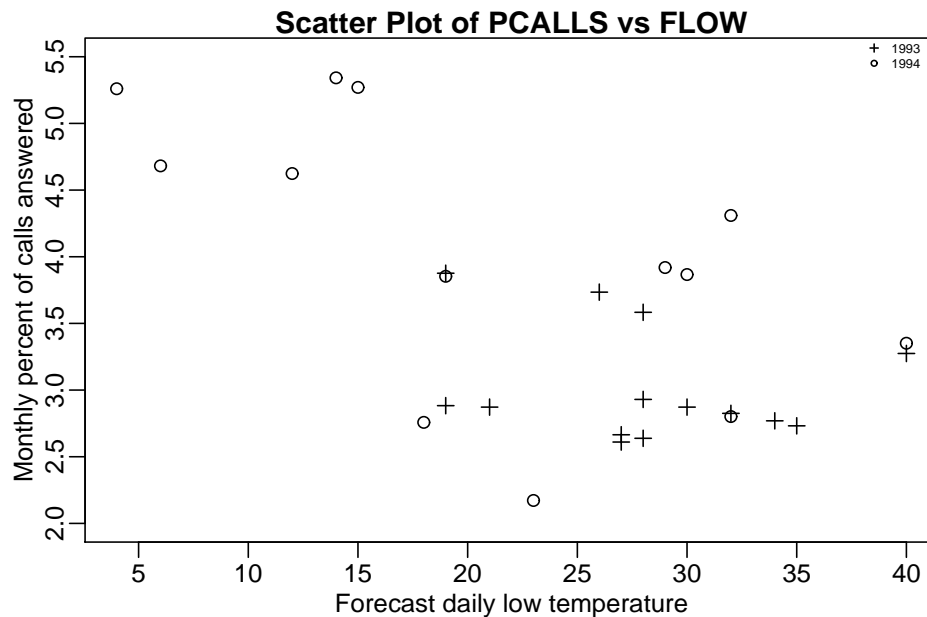
Our objective is to predict the levels of calls for the next day. It is not necessary to predict the total number (although that would be nice), but only to predict the level compared with other days in that month. The reason predicting the total number of

calls is difficult is that it depends on the cumulative effects of the weather over time. This is difficult to measure and to take into account. An alternative is to model the percentage of the monthly ERS calls made on each day:

$$PCALLS = 100 \times CALLS / \text{TOTAL MONTHLY CALLS}.$$

For example, the percentages for 1/16/93 and 1/17/93 are 3.6% and 2.7%, respectively. This suggests that the resources necessary on the 17th would be about $75\% = 2.7/3.6$ of those necessary on the 16th. Thus 25% of those people working for the ERS on the 16th could be re-assigned or given a rest day. The advantage of considering PCALLS is that it adjusts for the total levels of calls for that month due to the cumulative effects of weather.

We want to be able to predict the usage for the next day. Hence, we should use the forecasted low temperature rather than the actual low temperature, because the latter number will not be available the previous day. The forecasted temperature and the actual temperature should be closely related. Let's look at a scatter plot of the relationship between the percentage of calls and the forecasted low temperature:



The values for 1993 have been represented by a + and those in 1994 have been represented by a o. As the forecasted temperature decreases, the percentage of calls tends to increase. Although there is still a lot of variation, the relationship appears to be more linear than the relationship for the number of calls itself.

Let's do a **least squares regression** with PCALLS as the dependent (target) variable, and FLOW as the independent (predicting) variable:

LINEAR REGRESSION OF PCALLS Monthly percentage of calls answered

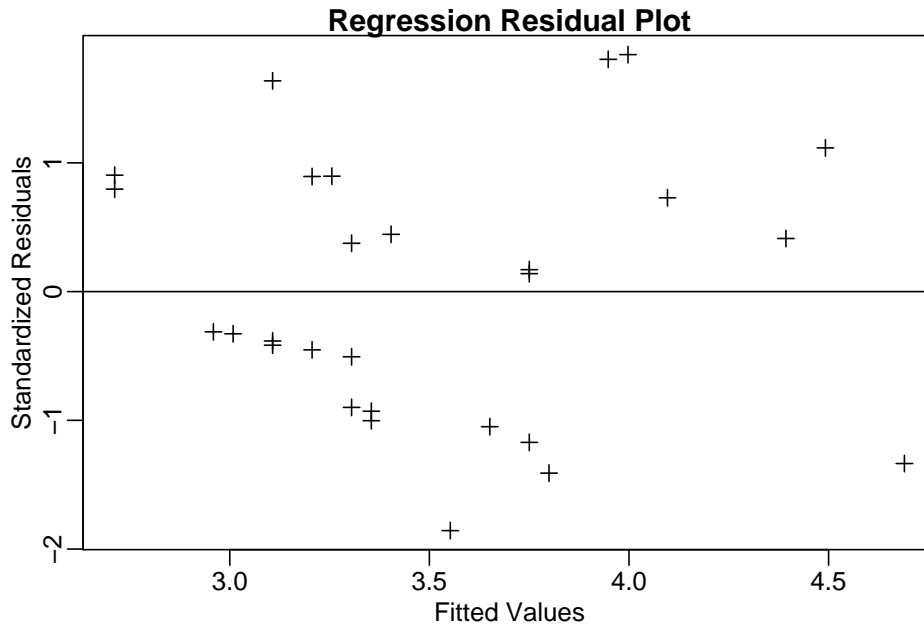
PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
CONSTANT	4.68996	0.36806	12.74	0.0000
FLOW	-0.04946	0.01421	-3.48	0.0018
R-SQUARED	0.3177	RESID. MEAN SQUARE (MSE)		0.5733
ADJUSTED R-SQUARED	0.2915	STANDARD DEVIATION		0.7572

SOURCE	DF	SS	MS	F	P
REGRESSION	1	6.9411	6.9411	12.11	0.0018
RESIDUAL	26	14.9056	0.5733		
TOTAL	27	21.8467			

CASES INCLUDED 28 MISSING CASES 0

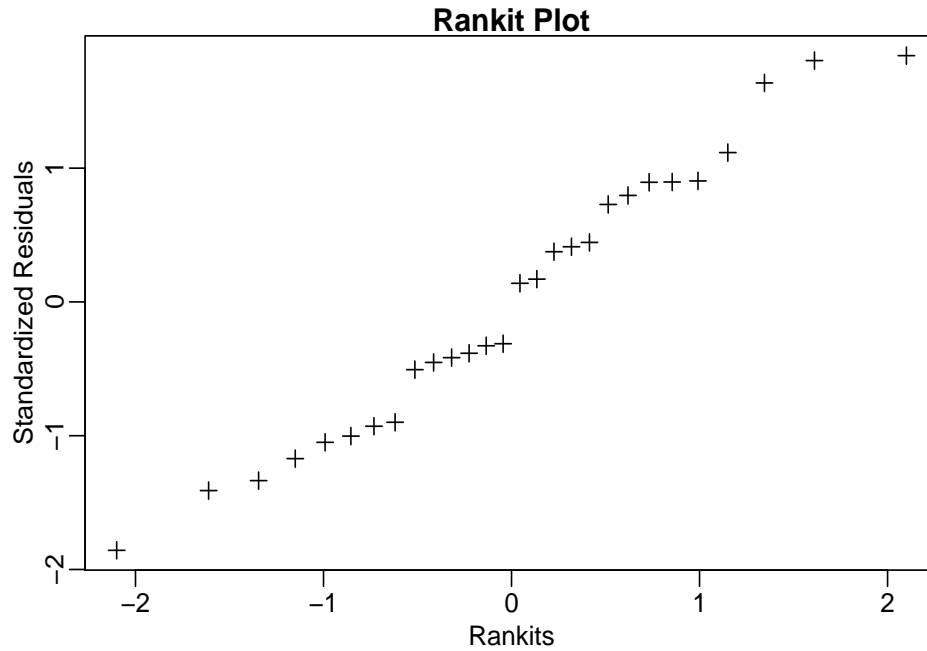
The strength of the regression is quite modest, with an R^2 of .32, but the F -statistic is significant. There is certainly a relationship between the forecasted low temperature and the percentage of monthly calls. The entry given under "STANDARD DEVIATION" is the *standard error of the estimate*, and is an estimate of the standard deviation of the error term. Thus, we would estimate that roughly 95% of all daily percentages would be within $\pm 2 \times .76 = 1.52\%$ of the population regression line. Given that the original range of PCALLS was roughly 2.0 – 5.5% , this reinforces the modest strength of the observed regression relationship.

To see if the assumptions underlying the regression model are correct, let's look at some diagnostic plots. Here is a plot of standardized residuals versus fitted values:

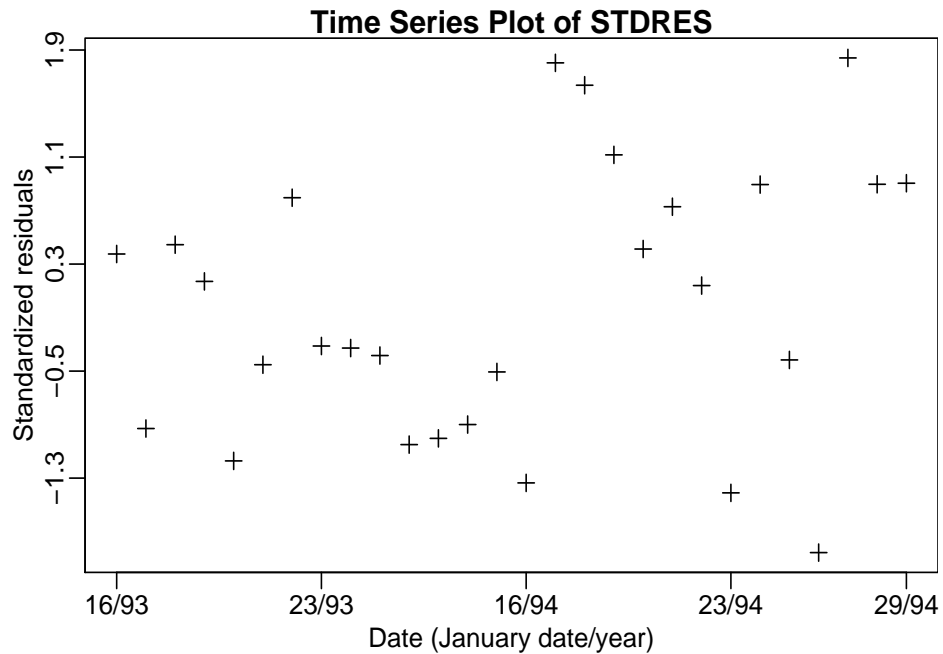


While there are no outlying values apparent in the plot, nor trends, it appears that the days with larger fitted values could have more variation than the days with smaller fitted values.

Here is a normal plot of the standardized residuals:



The standardized residuals appear to be roughly normally distributed. As a final diagnostic plot, let's look at a time series plot of the standardized residuals:



It appears that the variation in the residuals for the percentage usage in 1994 is greater than that for the values from 1993. However, the difference is not large, and we will not adjust for it here.

Thus, the regression assumptions, by and large, appear to hold. We can then interpret the coefficient of FLOW in the numerical output; each degree decrease in forecasted low temperature is associated with a 0.052% increase in the percentage of the monthly calls made on that day.

The busiest day of all in 1994 was January 27, when the daily low temperature was zero degrees and the ground was under six inches of snow. The Club answered 8947 calls, which was 5.3% of the monthly total (compared with 3.2% on the average day).

Could this have been predicted from the low temperature forecast only?

The forecast low for January 27 was 14 degrees. Let's predict the percentage of calls we would expect to be answered, given that the forecasted temperature was 14 degrees. To make the comparison fair, we refitted the above model with January 27 removed and then predicted the value of PCALLS based on the forecasted temperature:

PREDICTED VALUES OF PCALLS

LOWER PREDICTED BOUND	2.3586
PREDICTED VALUE	3.8964
UPPER PREDICTED BOUND	5.4342
SE (PREDICTED VALUE)	0.7481

PERCENT COVERAGE 95.0

PREDICTOR VALUES: FLOW = 14.000

The 95% prediction interval of PCALLS for a day with that forecasted low temperature is (2.36%, 5.43%). Thus, although the percentage of calls on January 27, 1994 was above the predicted level for such a day (PCALLS = 3.95%), the percent of calls is not inconsistent with the data as a whole. The Club can choose the confidence level to adjust how confident it wants to be to meet the demand and then staff accordingly. For example, if it wanted to be 95% confident of meeting the demand on that day, while also not assigning too many people to ERS, this interval suggests it should provide enough staff to meet between 2.36% and 5.43% of the monthly total. As the previous day's percentage was 2.17%, this suggests that the necessary resources would be about $250\% = 5.43/2.17$ of those necessary on the 26th.

Let's calculate a 93% prediction interval for January 27th:

PREDICTED VALUES OF PCALLS

LOWER PREDICTED BOUND	2.4830
PREDICTED VALUE	3.8964
UPPER PREDICTED BOUND	5.3098
SE (PREDICTED VALUE)	0.7479

PERCENT COVERAGE 93.0

PREDICTOR VALUES: FLOW = 14.000

The 93% prediction interval of PCALLS for a day with that forecasted low temperature is (2.48%, 5.31%). The observed value falls just inside the upper bound of this interval. Thus, we can be more precise; we would expect about 3.5% of days with a forecasted low temperature of 14 degrees to have a percentage monthly calls at or above the actual observed level.

Summary

The Automobile Club of New York would find it very useful to be able to forecast the demand for its Emergency Road Service (ERS) based on readily available information.

We can build a linear regression model for the daily percentage of the monthly number of calls to the ERS in terms of the daily low temperature forecasted the previous day.

We find that each degree decrease in forecasted low temperature is associated with a 0.049% increase in percentage of the monthly calls made on the next day.

Based on the model, we show that the busiest day of the year is under-predicted by the model. However, the model forecasts that about 3.5% of such days will have a percentage monthly calls at or above the observed level for that day. The model also predicts that the resources necessary to meet the demand on such a day would be at most 2.5 times the resources that were needed on the previous day, with 97.5% confidence.

Technical terms

***F* –test:** a method used to assess the statistical significance of the linear relationship between the target variable and a set of predicting variables. The test is appropriate when the errors in the regression model can be considered to be a random sample from a (roughly) normal distribution.

Prediction interval: a confidence interval constructed for the predicted value of the target variable for a particular member of the population with given value(s) of the predicting variables. The interval has the usual *t*–based construction, using the standard error of the predicted value.

***R*²:** an estimate of the proportion of the variability in the target variable accounted for by the regression. It equals the ratio of the regression sum of squares to the (corrected) total sum of squares.

Residual plot: a graphical test of the adequacy of regression assumptions. The residual for an observation is the difference between the observed target value and the value predicted by the fitted regression model (the **fitted value**). A **standardized residual** is a residual divided by its standard error,

$$\text{standardized residual} = \frac{\text{residual}}{\text{standard error of the residual}}.$$

The advantage of the standardized residuals is that, unlike the residuals, each one has the same variance. A successful fitting of a regression model should produce standardized residuals lying within $(-2.5, 2.5)$, roughly speaking, with no discernible pattern in their distribution. Thus, a scatter plot of standardized residuals versus fitted values should exhibit no apparent pattern. No regression analysis is complete without an examination of residuals.

Standard error of the estimate: an estimate of the standard deviation of the errors in a regression model. It equals the square root of the residual mean square, and can be used to assess the ability of the regression model to predict the target variable accurately.

Target variable: the goal of a regression model is to predict the value of a single **target** variable based on the value of one or more **predicting variables**. The ubiquity of the model has led to many synonym pairs, including dependent/independent, response/explanatory, and even simply Y/X .

Time series plot: a scatter plot of a variable versus time. The plot indicates (among other things) whether seasonal effects might be present in the variable.