SOCIETY OF ACTUARIES

Article from:

Forecasting & Futurism

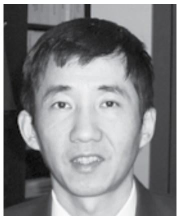
July 2012 – Issue 5

How to Win an iPad2

By Richard Xu

In the January 2012 issue of *Forecasting & Futurism Newsletter*, there was an announcement about the 2nd annual iPad2 forecasting competition. The competition is to develop a model to predict the monthly unemployment rate (UNRATE) from March 2012 to September 2012. The winner's model should have the smallest sum of squared deviations between model predicted values and the actual data over the forecast period of six months. Data is limited to the Federal Reserve Economic Data (FRED) database, which has about 35,000 historical economic data and is available free of charge, with the exclusion of variables that have direct unemployment information.

Many actuaries are tempted to get into the race and win a nice iPad2, but find themselves with a lack of available model to start with. But in fact, almost all actuaries have the educational background to build a regression or time series model from either their college courses or required actuarial exams. The problem that many actuaries are facing is that they infrequently, if ever, use these in their actuarial works. Without much work experience, many actuaries may forget linear regression models or time series analysis, and so winning an iPad may look like a daunting task.



Richard Xu

Richard Xu, FSA, Ph.D., is a modeling actuary with RGA Reinsurance Company in Chesterfield, Mo. He can be reached at *rxu@rgare.com.*

The purpose of this article is to provide a refresher on regression and time series models so that actuaries will feel more comfortable and confident to build a forecasting model based on these fundamental tools and apply them in their actuarial works if such models are appropriate.

Simply put, a linear regression model can be described by an equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} + \varepsilon = \Sigma_j \beta_j x_{ij} + \varepsilon_i$$

where y_i is called *response variable*, or *dependent variable.* This is the variable that has been observed in experience and is to be predicted by model. x_{ij} are called the *explanatory variables*, *covariates*, *input variables*, or *independent variables.* β_j are coefficients to be estimated in model building process, and ε_i is error term.

To make a valid linear regression in this basic form, several assumptions are needed. A linear relationship between response and explanatory variables is obviously one. In most applications in finance, this usually is not a problem. Either the relationship is inherently linear, or it can be well-approximated by a linear equation over short ranges. In addition, the error term ε_i must follow normal distribution with mean value at zero and a constant variance, i.e., $\varepsilon_i \sim N(0, \sigma^2)$. Other requirements include that y_i is representative of population, observations are independent from each other, and x_{ij} is error-free.

The most common method of estimating β_j is least squares, in which β_j is chosen such that $RSS = \sum_i (\hat{y}_i - y_i)^2 = \sum_i (\Sigma_j \beta_j x_{ij} - y_i)^2$ is at its minimum, where RSS stands for Residual Sum Square, and $\hat{y}_i$ is the fitted value. There are close form solutions for β_j in matrix form. The other estimation is maximum likelihood to find β_j so that product of probability at all data points is at its maximum. Under the normal distribution, it can be proven that both estimations will give the same result.

Unless it is a very small data set, it is not possible to build a real model just with pen and paper. You have to rely on computing software to find β_j. The choice of statistical software is quite abundant, such as R, SAS, SPSS, MatLab, MiniTab, etc. Actually, for a very small simple application, you can use Excel built-in function by selection "Data" -> "Data Analysis," but it has the limit of only 16 explanatory variables. For a large or complicated model, computing software is the only viable choice. Among the actuarial community, the two most commonly used are R and SAS. The R is free software under GUN license, while the later one is a commercial product. The examples in this article will be illustrated by using R. R is unique, not only because it is free, but also because there is a large online community and a core statistics team to support it. You have a wide choice of education and academic materials about R, and there will never be a shortage of statistic tools in R to build any particular model. As of now, there are 3,738 packages available on top of already abundant basic tools that come with the R system, and the number is still growing.

Let's look at an example on how you can work out a linear regression model. Here is a 10-year revenue data of a public insurance company. We would like to know how the revenue grew in the past 10 years and predict what revenue will be for 2012. With the data, you can save to a text file or CSV file called "iData.txt" with common as separator and header included.

Year	Revenue
2002	2.382
2003	3.175
2004	4.021
2005	4.585
2006	5.194
2007	5.718
2008	5.681
2009	7.067
2010	8.262
2011	8.830

Inside R, you can use the following command to first load data into R system, build a linear regression model, and show summary.

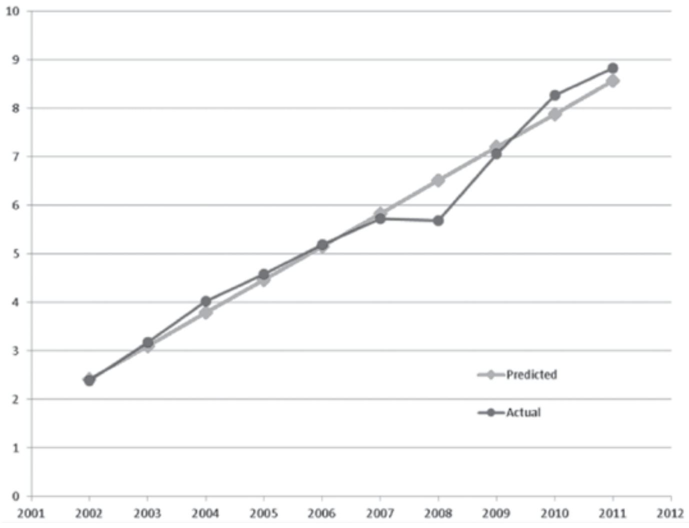
```
>iData<-read.table("iData1.txt", header = TRUE, sep=",")
>iModel<- lm(Revenue~Year, data=iData)
>summary(iModel)
Call:
lm(formula = Revenue ~ Year)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83489 -0.09530  0.05885  0.20709  0.38025

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.365e+03  7.893e+01  -17.29 1.27e-07 ***
year         6.829e-01  3.934e-02   17.36 1.24e-07 ***
---
Signif.codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3573 on 8 degrees of freedom
Multiple R-squared: 0.9741,    Adjusted R-squared: 0.9709
F-statistic: 301.4 on 1 and 8 DF,  p-value: 1.235e-07
```

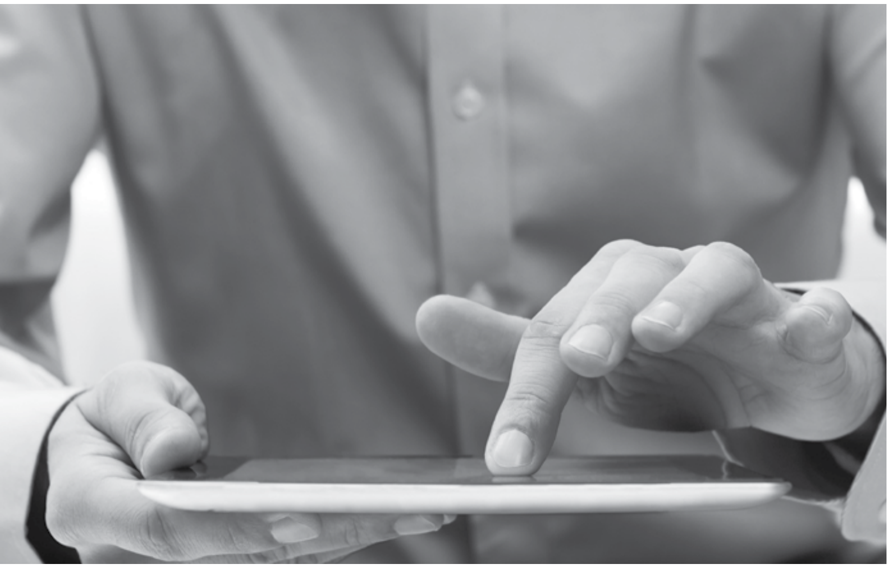
Revenue (2002-2011) - In Billions



In summary, we can see that the slope is 0.68, which is the annual revenue increase rate. For the year 2012, the predicted revenue is 9.25. Actually, you will have more statistical information about the model, such as the confidence level of the coefficient, goodness of fittings, etc.

CONTINUED ON **PAGE 10**

HOW TO WIN AN IPAD2 | FROM PAGE 9



For linear regression models, the most often used criteria to assess the goodness of fitting is R^2, which is defined as a ratio of variance that has been explained by the model to the total variance in the data. However, the R^2 could be misleading as more explanatory variables will always increase R^2 even though the additional variables may totally be irrelevant, such as purenoise. This is called an over-fitting problem in modeling, and can be a serious issue as a model may have a perfect fit to data that are used for modeling, but very poor in application of real life. The adjusted $\bar{R}^2$ is better, as a penalty is added to it such that the increase of R^2 has to be statistically large enough to overcome the penalty of additional variable. A more universal approach is the maximum likelihood, where Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) can be used to assess the model and to avoid the overfitting issue.

LINEAR REGRESSION AND TIME SERIES ARE VERY BASIC STATISTIC TOOLS. YOU ARE NEVER SHORT OF APPLICATIONS IN ALMOST ALL INDUSTRY FIELDS.

Once you are comfortable enough to build a linear regression model, you can naturally extend your skills to time series, where input data is a sequence of data points at successive time instants usually with uniform time intervals. There are two basic models that are conceptual extensions of linear regression model. One is the autoregressive model (*AR*), in which explanatory variables include the response variable itself, but at an earlier time. For example, the unemployment rate at a certain month is highly correlated to levels of several previous months, and can be explained in large part by its immediately previous monthly rate. A mathematical equation for a simple *AR* model can be stated as

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_n y_{t-p} + \varepsilon_t$$

This is an autoregressive model with p terms, usually denoted as *AR(p)*. The other model is called moving average (*MA*) model, where response variable is a function of previous error terms. An *MA* model with q terms can be represented by

$$y_t = \mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_n \epsilon_{t-q} + \varepsilon_t$$

When you combine these two models, you have the autoregressive–moving-average (ARMA) models, sometimes called Box–Jenkins models. Usually the notation ARMA (p,q) is used to refer to the model with p autoregressive terms and q moving-average terms.

There are basic functions in *R* that you can use to model the time series, such as “arima.” Also, there are several packages you can install within *R* to do some special analysis. Just like linear regression models, you need the right commands to load data, build a model, and assess the model. Usually, the procedure is iterative in nature. You will try different variables and can even include an interaction term, until you find an optimal model that best explains the data.

Linear regression and time series are very basic statistic tools. You are never short of applications in almost all industry fields. Extensions of these two techniques to overcome various limits have led to numerous other modeling tools, such as generalized linear model (GLM), mixed effect model, GARCH, etc. Although the direct application of linear regression and time series in insurance is very limited, the GLM eventually finds its way into actuarial science and now we are witnessing the explosive applications of GLM in insurance, known as predictive modeling.

Data is always a concern in modeling, but actuaries are considered as number experts and never underestimate the importance of data and difficulty of understanding and cleaning data. In reality you have all different kinds of issues to consider, such as sources of data, quality and quantity, missing variable, etc, and actuaries are usually clever in finding their way out. Luckily, in this competition, data is less an issue. The main question is to find the right explanatory variables from the list of 35,000 series.

A few words about *R*. Many actuaries find it very intimidating to start to learn *R* after they are used to graphic user interface (GUI) by clicking on buttons or menu for so many years. It truly is, at the beginning, especially when you have never had the exposure to the command line environment. A good start will be a few simple examples that you already know what the model is all about so that uncertainty about the model itself is removed leaving only questions about *R*. As you progress in both scope and depth of modeling skills, you will find that *R* is a very powerful and versatile tool for data analysis and visualization.

To win an iPad2 would be a very nice achievement, but you will gain more even by just participation. With the coming of the big data era and wide acceptance of predictive modeling in insurance, actuaries are faced with more demands on their modeling skills and their tool choices. This competition is a very good starting point for actuaries to try these, which is perhaps the main reason why you should participate in it.

Have you used the Competency Framework Self-Assessment Tool?

We want to know what you think.

How has the tool and personal planning workbook helped you design your future?

Contact **Jacquenette Moody**, professional development manager, with feedback at **jmoody@soa.org.**

CompetencyFRAMEWORK
DESIGN your future.