Linear Regression

Thomas Lonon

Division of Financial Engineering
Stevens Institute of Technology

February 1, 2017

Advertising Data

1. Is there a relationship between advertising budget and sales?
2. How strong is the relationship between the advertising budget and sales?
3. Which media contribute to sales?
4. How accurately can we estimate the effect of each medium on sales?
5. How accurately can we predict future sales?
6. Is the relationship linear?
7. Is there synergy among the advertising media?

[1]

Simple Linear Regression: A very straightforward approach for predicting a quantitative response Y on the basis of a single predictor variable X that assumes there is an approximately linear relationship between X and Y. This is expressed as:

$$Y \approx \beta_0 + \beta_1 X$$

Once we have determined our estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we can predict future sales

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Let

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

represent n observation pairs. We are looking for coeffecient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that represent the data well.

In other words, we are looking for these parameters such that:

$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$$

This is done through **least squares**

Let

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

be the predictor for Y based on the i^{th} value of X. Then the i^{th} **residual**, the difference between the observed and the predicted response, is represented as

$$e_i = y_i - \hat{y}_i$$

The **Residual Sum of Squares (RSS)** is given as:

$$RSS = \sum_{i=1}^{n} e_i^2$$

which can also be represented as

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

For the second representation of RSS given, we can determine the parameters for $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize this value. We do this by taking the derivatives with respect to these parameters and setting them equal to 0 (standard approach). We get:

$$\frac{\partial RSS}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0$$

This system of equations is solved as:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

and

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

We have the assumption that the *true* relationship between X and Y takes the form

$$Y = f(X) + \epsilon$$

for some function f. If this function is a linear function, then we have:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

where β_0 is the intercept term and β_1 is the slope. This expression is the **population regression line**, the best linear approximation to the true relationship between X and Y.

For a set of i.i.d. random variable $\{x_i\}, i \in 1, \ldots, n$, what can we say about the average($\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$)?

- The expected value is:

$$\mathbb{E}[\bar{x}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_i] = \mathbb{E}[x_i]$$

- The variance is:

$$\begin{aligned}
\mathbb{V}(\bar{x}) &= \mathbb{E}[(\bar{x} - \mathbb{E}[\bar{x}])^2] \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_i]\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n}(x_i - \mathbb{E}[x_i])\right)^2\right] \\
&= \frac{1}{n}\mathbb{V}(x_i)
\end{aligned}$$

This estimator for $\bar{x}$ is an example of an **unbiased** estimator.

Based on this definition of the variance of a sample mean, we can also have the **standard error of the estimate (SE)** given by:

$$SE(\hat{\mu}) = \frac{1}{\sqrt{n}}\sigma$$

where σ is the standard deviation of each of the realizations y_i of Y.

Using this approach, we can get the standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$
$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where $\sigma^2 = \mathbb{V}(\epsilon)$. If σ isn't known, we use an estimate for σ known as the **residual standard error (RSE)** given by the formula:

$$RSE = \sqrt{\frac{RSS}{n-2}}$$

These standard errors can be used to calculate our **confidence intervals.** For linear regression, the 95% confidence intervals are given by:

$$\hat{\beta}_0 \pm 2SE(\hat{\beta}_0)$$
$$\hat{\beta}_1 \pm 2SE(\hat{\beta}_1)$$

That is, there is approximately a 95% chance that the true value of β_0 is contained within

$$[\hat{\beta}_0 - 2SE(\hat{\beta}_0), \hat{\beta}_0 + 2SE(\hat{\beta}_0)]$$

Hypothesis Testing

The most common approach is to test the null hypothesis (H_0) versus the alternate hypothesis (H_A). For linear regression an example of this test would be to check whether there is a relationship between X and Y.

- : H_0: There is no relationship between X and Y

$$H_0 : \beta_1 = 0$$

- : H_A: There is some relationship between X and Y

$$H_A : \beta_1 \neq 0$$

Note that if $\beta_1 = 0$, then $Y = \beta_0 + \epsilon$

To test these hypotheses, we compute a **t-statistic** given by

$$t = \frac{\hat{\beta}_1 - 0}{\widehat{SE}(\hat{\beta}_1)}$$

which measures the number of standard deviations that $\hat{\beta}_1$ is from 0. If there is no relationship between X and Y, then the value of t will have a t-distribution with $n-2$ degrees of freedom. The **p-value** is the probability of observing $|t|$ or larger with this distribution. If this p-value is small enough, we **reject the null hypothesis**

If we reject the null hypothesis, we will want to know the extent in which the model fits the data. This is assessed using the RSE and the **R^2 statistic**.

This RSE is considered a measure of the **lack of fit** of the model to the data.

To calculate R^2 we use:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

where

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

is the **total sum of squares (TSS)**

R^2

This R^2 statistic is a relative measure of fit, and so is easier to interpret than the RSE. It measures the proportion of variability in Y that can be expressed using X.

- a R^2 close to 1 indicates a large proportion of variability in the response has been explained by the regression
- R^2 close to 0 indicates the the regression did not explain much of the variability in the response.

There is still some leeway as to what constitutes a "good" R^2 value.

For multiple predictors, instead of simply running a linear regression on each predictor (which isn't efficient and leads to more questions) we extend the linear regression model so that it can accommodate multiple predictors. If we have a model with p predictors, then the linear regression model takes the form:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

where X_j references the j^{th} predictor and β_j quantifies the association between that variable and the response.

Given estimates for the β's, we can make predictions using the formula:

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_j$$

These β's are chosen to minimize the sum of squared residuals:

$$\begin{aligned} RSS &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij} \right)^2 \end{aligned}$$

[2]

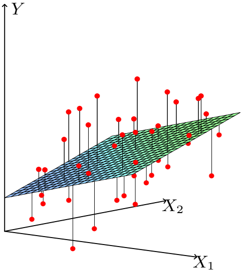

FIGURE 3.1. *Linear least squares fitting with $X \in \mathbb{R}^2$. We seek the linear function of X that minimizes the sum of squared residuals from Y.*

1. Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of particular values, what response value should we predict, and how accurate is the response?

[1]

Is there a relationship?

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_A : \text{ at least one } \beta_j \text{ is non-zero}$$

This test is performed using the **F-statistic**

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} = \frac{(TSS - RSS)(n-p-1)}{pRSS}$$

If the linear model assumption are correct:

$$\mathbb{E}\left[\frac{RSS}{n-p-1}\right] = \sigma^2$$

and provided that H_0 is true

$$\mathbb{E}\left[\frac{TSS - RSS}{p}\right] = \sigma^2$$

χ^2 Distribution

If $x_i, i \in 1, \ldots, n$ are iid standard normal variables, then

$$s = \sum_{i=1}^{n} x_i^2 \sim \chi_n^2$$

where χ_n^2 denotes a chi-square distribution with n degrees of freedom.

The F-Distribution

If u is a χ^2_n random variable, and v is a χ^2_m random variable, and u and v are independent, then

$$f = \frac{(u/n)}{(v/m)} = \frac{um}{vn} \sim F_{n,m}$$

Deciding on Important Variables

There are three classical approaches:

- Forward Selection: start with the null model, fit p simple linear regressions, add to the null model the one that has the lowest RSS, continue until stopping rule is satisfied
- Backward Selection: Start with all the variables and remove the largest p-value, continue until stopping rule is satisfied
- Mixed Selection: start with the null model, adding in variables one at a time, if the p-value for a variable rises above a threshold, remove the variable, repeat

Model Fit

The two measures for fit of the multiple linear regression models are still the RSE and R^2, calculated in a similar fashion.

The biggest change is in the RSE:

$$RSE = \sqrt{\frac{1}{n-p-1}RSS}$$

which simplifies to the single linear regression formula for $p = 1$.

Predictions

Once we have the model, we can make predictions using our predictors x_j by:

$$\hat{y} = \beta_0 + \sum_{j=1}^{p} x_j \beta_j$$

This contains three types of uncertainty:

- Reducible Error: least-squares plane is only approximation to the true population regression plane
- Model Bias: $f(x)$ is almost certainly nonlinear, so this is just an approximation
- Irreducible Error: Even if $f(x)$ was perfect, we would still have our ϵ

[1] Trevor Hastie Gareth James, Daniela Witten and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Number v. 6. Springer, 2013.

[2] Robert Tibshirani Trevor Hastie and Jerome Friedman. *The Elements of Stastical Learning: Data Mining, Inference, and Prediction*. Number v.2 in Springer Series in Statistics. Springer, 2009.