Linear Regression

Simple Linear Regression

It assumes dependency of Y on predictors X_1,X_2,X_3...X_n.Predicting a quantitative response Y on the basis of a singular predictor variable X. It assumes that there is approximately a linear relationship between X and Y . Mathematically, we can write this linear relationship as

Y ≈ β0 + β1X

You might read "≈" as "is approximately modeled as". We will sometimes describe by saying that we are regressing Y on X (or Y onto X).For example, X may represent TV advertising and Y may represent sales.Then we can regress sales onto TV by fitting the model

SALES ≈ β0 + β1 × TV+ε

β0 and β1 are two unknown constants that represent the intercept and slope terms in the linear model. ε-is the error term.

Together, β0 and β1 are known as the model *coefficients or parameters*.Once we have used our training data to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we can predict future sales on the basis of a particular value of TV advertising by computing

Y=$\hat{\beta}_0$ +$\hat{\beta}_1$X

where ˆy indicates a prediction of Y on the basis of X = x. Here we use a hat symbol ˆ , to denote the estimated value for an *unknown parameter or coefficient*, or to denote the predicted value of the response.

Estimating the Coefficients

Y ≈ β0 + β1X

In practice, β0 and β1 are unknown. So before we can use the above equation to make predictions, we must use data to estimate the coefficients. Let(x1, y1), (x2, y2), . . . , (xn, yn) represent n observation pairs, each of which consists of a measurement of X and a measurement of Y . In the Advertising example, this dataset consists of the TV advertising budget and product sales in n = 200 different markets.

Our goal is to obtain coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the above linear model fits the available data well—that is, so that $\hat{Y} \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ for i =1, . . . , n.

Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of x .Then $e_i = y_i - \hat{y}_i$ represents the ith residual—this is the difference between the ith observed response value and the ith response value that is predicted by our linear model .

We define the residual sum of squares (RSS) as RSS = $\sum_i^n e_i{}^2$

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. Using some calculus, one can show that the minimizers are

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

$\bar{y} = \frac{1}{n}\sum_i^n y_i$ $\quad and \quad$ $\bar{x} = \frac{1}{n}\sum_i^n x_i$ are the sample means

Linear Regression-Derivation

Derivation of intercept and co-efficient of Linear Regression

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. Using some calculus, one can show that the minimizers are

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x},$$

The below is derivation of intercept and co-efficient using calculus
Y = α + βx
$Y_i = \alpha + \beta x_i + \varepsilon_i$

$$\min_{\hat{\alpha},\hat{\beta}} \text{SSE}\left(\hat{\alpha},\hat{\beta}\right) \equiv \min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)^2$$

To find a minimum, we take partial derivatives with respect to $\hat{\alpha}$ and$\hat{\beta}$.

$$\frac{\partial}{\partial\hat{\alpha}}\left(\text{SSE}\left(\hat{\alpha},\hat{\beta}\right)\right) = -2\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}x_i\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n}\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} x_i$$

$$\Rightarrow \sum_{i=1}^{n} y_i = n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} x_i$$

$$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} y_i = \hat{\alpha} + \frac{1}{n}\hat{\beta}\sum_{i=1}^{n} x_i$$

$$\Rightarrow \bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$$

Before taking partial derivative with respect to $\hat{\beta}$, substitute the previous result for $\hat{\alpha}$.

$$\min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{n}\left[y_i - \left(\bar{y} - \hat{\beta}\bar{x}\right) - \hat{\beta}x_i\right]^2 = \min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{n}\left[\left(y_i - \bar{y}\right) - \hat{\beta}\left(x_i - \bar{x}\right)\right]^2$$

$$\frac{\partial}{\partial\hat{\beta}}\left(\text{SSE}\left(\hat{\alpha},\hat{\beta}\right)\right) = -2\sum_{i=1}^{n}\left[\left(y_i - \bar{y}\right) - \hat{\beta}\left(x_i - \bar{x}\right)\right]\left(x_i - \bar{x}\right) = 0$$

$$\Rightarrow \sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) - \hat{\beta}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = 0$$

$$\Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{\text{Cov}(x,y)}{\text{Var}(x)}$$

Finally substitute $\hat{\beta}$ to determine $\hat{\alpha}$.

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Linear Regression-Accuracy

Assessing the Accuracy of the Coefficient Estimates

We assume that the true relationship between X and Y takes the form Y = f(X) + ε for some unknown function f, where ε is a mean-zero random error term. If f is to be approximated by a linear function, then we can write this relationship as Y = β0 + β1X + ε.

Here β0 is the intercept term—that is, the expected value of Y when X = 0, and β1 is the slope—the average increase in Y associated with a one-unit increase in X.

The above equation defines the *population regression line*, which is the best linear approximation to the true relationship between X and Y. In real applications, we have access to a set of observations from which we can compute the least squares line; however, the population regression line is unobserved

In general the standard error $\frac{\sigma}{\sqrt{n}}$ explains the average amount that this estimate $\widehat{\mu}$ differs from the actual value of μ.

We can wonder how close $\hat{\beta}_0$ and $\hat{\beta}_1$ are to the true values β0 and β1.To compute the standard errors associated with c and $\hat{\beta}_1$, we use the following formulas:

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

Where σ^2 = Var(ε)For these formulas to be strictly valid, we need to assume that the errors ε for each observation are uncorrelated with common variance σ^2 .

The estimate of σ is known as the residual standard error, and is given by the formula

$$\mathrm{RSE} = \sqrt{\mathrm{RSS}/(n-2)}$$

Standard errors can be used to compute confidence intervals. That is, there is approximately a 95% chance that the interval will contain the true value of β1

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

Similarly, a confidence interval for β0 approximately takes the form

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

In the case of the advertising data, the 95% confidence interval for β0 is [6.130, 7.935] and the 95% confidence interval for β1 is [0.042, 0.053].Therefore, we can conclude that in the absence of any advertising, sales will on average fall somewhere between 6,130 and 7,940 units. Furthermore, for each $1,000 increase in television advertising, there will be an average increase in sales of between 42 and 53 units.

Linear Regression-Hypothesis testing

Hypothesis testing

Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of

H0 : There is no relationship between X and Y

versus the alternative hypothesis

Ha : There is some relationship between X and Y

Mathematically, this corresponds to testing H0 : β1 = 0 versus Ha : β1 ≠ 0,
To test the null hypothesis, we need to determine whether $\hat{\beta}$, our estimate for β1, is sufficiently far from zero that we can be confident that β1 is non-zero.

If SE($\hat{\beta}1$), is small, then even relatively small values of $\hat{\beta}1$ may provide strong evidence that β1 ≠ 0, and hence that there is a relationship between X and Y . Incontrast, if SE($\hat{\beta}1$) is large, then $\hat{\beta}1$ must be large in absolute value in order for us to reject the null hypothesis. In practice, we compute a t-statistic

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

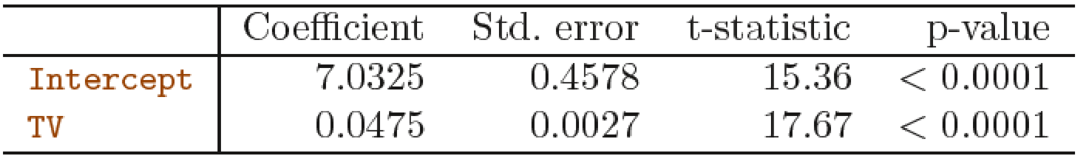
which measures the number of standard deviations that $\hat{\beta}1$ is away from 0. If there really is no relationship between X and Y , then we expect that it will have a t-distribution with n−2 degrees of freedom.

The t-distribution has a bell shape and for values of n greater than approximately 30 it is quite similar to the normal distribution.

Consequently, it is a simple matter to compute the probability of observing any number equal to |t| or larger in absolute value, assuming β1= 0. We call this probability the p-value.

We interpret the p-value as follows: a small p-value indicates that it is unlikely to observe such a substantial association been predictor and the response due to chance, in the absence of any real association between predictor and response.
Hence, if we see a small p-value, then we can infer that there is an association between the predictor and the response.

We reject the null hypothesis—that is, we declare a relationship to exist between X and Y —if the p-value is small enough. Typical p-value cut offs for rejecting the null hypothesis are 5 or 1%. When n = 30, these correspond to t-statistics (3.14) of around 2 and 2.75, respectively.



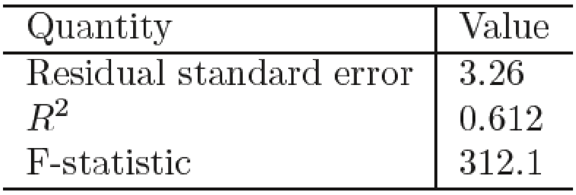
	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Table provides details of the least squares model for the regression of number of units sold on TV advertising budget for the Advertising data. Notice that the coefficients for $\hat{\beta}0$ and $\hat{\beta}1$ are very large relative to their standard errors, so the t-statistics are also large; the probabilities of seeing such values if H0 is true are virtually zero. Hence we can conclude thatβ0≠ 0 and β1≠0.

Linear Regression-Accuracy of the model

Assessing the Accuracy of the Model

It is natural to want to quantify the extent to which the model fits the data. The quality of a linear regression fit is typically assessed using two related quantities: the residual standard error (RSE) and the R^2 statistic



Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

Residual Standard Error

Due to the presence of these error terms, even if we knew the true regression line (i.e. even if β0 and β1 were known), we would not be able to perfectly predict Y from X.The RSE is an estimate of the standard deviation of ε.Roughly speaking, it is the average amount that the response will deviate from the true regression line. It is computed using the formula

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

In the case of the advertising data, we see from the linear regression output in Table 3.2 that the RSE is 3.26. In other words, actual sales in each market deviate from the true regression line by approximately 3,260

units, on average. Another way to think about this is that even if the model were correct and the true values of the unknown coefficients β0 and β1 were known exactly, any prediction of sales on the basis of TV

advertising would still be off by about 3,260 units on average. Of course, whether or not 3,260 units is an acceptable prediction error depends on the problem context.

R^2 Statistic

The RSE provides an absolute measure of lack of fit of the model to the data. But since it is measured in the units of Y , it is not always clear what constitutes a good RSE. The R^2 statistic provides an alternative measure of fit. It takes the form of a proportion—the proportion of variance explained and so it always takes on a value between 0 and 1, and is independent of the scale of Y .

To calculate R2, we use the formula

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

TSS = $\sum(y_i - \bar{y})^2$

TSS measures the total variance in the response Y , and can be thought of as the amount of variability inherent in the response before the regression is performed.

RSS measures the amount of variability that is left unexplained after performing the regression

Hence, TSS—RSS measures the amount of variability in the response that is explained (or removed) by performing the regression, and R2 measures the proportion of variability in Y that can be explained using X.

Linear Regression-Explain R^2

Explain R^2

The R^2 statistic has an interpretational advantage over the RSE since unlike the RSE, it always lies between 0 and 1. However, it can still be challenging to determine what is a good R^2 value, and in general, this will depend on the application.

For instance, in certain problems in physics, we may know that the data truly comes from a linear model with a small residual error. In this case, we would expect to see an R^2 value that is extremely close to 1, and a substantially smaller R^2 value might indicate a serious problem with the experiment in which the data were generated.

On the other hand, in typical applications in biology, psychology, marketing, and other domains, the linear model is at best an extremely rough approximation to the data, and residual errors due to other unmeasured factors are often very large. In this setting, we would expect only a very small proportion of the variance in the response to be explained by the predictor, and an R^2 value well below 0.1 might be more realistic

Correlation is another measure of the linear relationship between XX and Y.Y. Correlation of can be calculated as

$$\text{Cor(X,Y)}=\sum_{i=1}^{n}(\text{xi}-\bar{x})(y\text{i}-\bar{y})\ /\sqrt{\sum_{i=1}^{n}(\text{xi}-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(\text{yi}-\bar{y})^2}$$

This suggests that r=Cor(X,Y) could be used instead of R^2 to assess the fit of the linear model, however for simple linear regression it can be shown that $R^2=r^2$. More concisely, for simple linear regression, the squared correlation and the R^2 statistic are equivalent. Though this is the case for simple linear regression, correlation does not extend to multiple linear regression since correlation quantifies the association between a single pair of variables. R^2 can, however, be applied to multiple regression.

Logistic Regression

Logistic regression is technique borrowed by machine learning from the field of statistics. It's best