Module 2

Supervised learning

The specialized statistical learning techniques are required for large-scale data analytics. Two basic data analytics are supervised or unsupervised. The term “supervised learning” is rooted in statistical learning or machine learning, where it describes the analysis of data via a focused structure – sometimes called “learning with a teacher”. The observations may belong to a set of training data, used to identify a model that relates the inputs, usually a vector or matrix of predictor or feature variables X, to an output response variable Y.

Two basic types of supervised statistical learning are regression analysis and classification analysis. Both are predictive in nature: they combine input variables with the eventual model to forecast or classify future realizations of the outcome variable.

Regression Versus Classification Problems

Variables can be characterized as either quantitative or qualitative (also known as categorical). Quantitative variables take on numerical values. Examples include a person’s age, height, or income, the value of a house, and the price of a stock. In contrast, qualitative variables take on values in one of K different classes, or categories. Examples of qualitative variables include a person’s gender (male or female), the brand of product purchased (brand A, B, or C), whether a person defaults on a debt (yes or no), or a cancer diagnosis (Acute Myelogenous Leukemia, Acute Lymphoblastic Leukemia, or No Leukemia).

We tend to refer to problems with a quantitative response as regression problems, while those involving a qualitative response are often referred to as classification problems. However, the distinction is not always that crisp.

2.1 Simple linear regression

2.1.1 Simple Linear Model

The simplest form of predictive model, simple linear regression (SLR), where a single quantitative variable, x, is used to describe a single outcome variable, Y. The SLR paradigm involves a single predictor variable, x, also called an input variable or feature variable or independent variable, used to describe an output or response variable or dependent variable, Y, via a simple linear model.

Formally, assume data are collected in matched pairs (xi, yi), i = 1, … , n, where the response variable is modeled as,

$$Y \approx \beta_0 + \beta_1 X \quad \text{--------(2.1)}$$

This equation means that regressing Y on X. β_0 and β_1 are two unknown constants that represent the intercept and slope terms in the linear model. Together, β0 and β1 are known as the regression coefficients or parameters. Once we have used the training data to produce estimates $\beta\hat{}_0$ and $\beta\hat{}_1$ for the model coefficients, we can predict future values on the basis of a particular value of an independent attribute x_i by computing

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

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

In practice, β_0 and β_1 are unknown. That is regression coefficients are assumed unknown and must be estimated from the data. 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 . Our goal is to obtain coefficient estimates βˆ0 and βˆ1 such that the linear model (2.1) fits the available data well—that is, so that yi ≈ βˆ0 + βˆ1xi for i = 1,...,n. In other words, we want to find an intercept βˆ0 and a slope βˆ1 such that the resulting line is as close as possible to the n data points. There are a number of ways of measuring closeness. However, by far the most common approach involves minimizing the Least Squares (LS) criterion. βˆ0 and βˆ1 are calculating using the following equations,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad (2.2)$$

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

In other words, $\hat{\beta}0$ and $\hat{\beta}1$ defines the least squares coefficient estimates for simple linear regression. It can be shown that the maximum likelihood estimators (MLEs) for $\beta 0$ and $\beta 1$ are identical to these LS estimators.

To estimate the mean response in (2.1), simply apply the LS estimators for the regression coefficients to the model equation. This gives $\hat{Y}$ i = $\hat{\beta}_0 + \hat{\beta}_1$xi , which are called the fitted values from the SLR.

The LS estimators in (2.2) possess a number of important qualities. One can show that $E[\hat{\beta}_j] = \beta_j$ for j = 0, 1; hence, the estimators are unbiased. Further, their sampling variances are

$$\mathrm{Var}[\hat{\beta}_0] = \sigma^2 \left\{ \frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right\}, \text{ and}$$

$$\mathrm{Var}[\hat{\beta}_1] = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

The linear model relates that the variances in the above equation are the minimum possible among all unbiased estimators for $\beta 0$ and $\beta 1$. Thus the $\hat{\beta}_j$s are known as “best linear unbiased estimators”. Where σ is the standard deviation

Under the SLR model, the estimated responses are the fitted values $\hat{Y}_i$, thus the pertinent deviations are

$$e_i = Y_i - \hat{Y}_i \quad , i = 1, \dots , n$$

This equation is known as the residuals from the model fit. Summing the squared residuals produces the residual sum of squares (RSS), also called the error sum of squares (SSE).

$$\mathrm{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

The SSE here has n− 2 d.f. (called error degrees of freedom). In effect, this is the number of observations minus the number of parameters estimated. Dividing the SSE by its d.f. produces a mean squared error (MSE):

$$\text{MSE} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2} \; .$$

The residuals, e_i, are useful for more than just constructing the SSE. In particular, they help in assessing quality of the model fit and in other regression diagnostics.

Assessing the Accuracy of the Coefficient Estimates:

We assume that the true relationship between X and Y takes the form

Y = β0 + β1X + ϵ, where ϵ is a mean-zero random error term.

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 model given by the above equation defines the population regression line, which is the best linear approximation to the true relationship between X and Y. The least squares regression coefficient estimates characterize the least squares line.

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

H_0 : There is no relationship between X and Y

versus the alternative hypothesis

H_a : There is some relationship between X and Y

Mathematically, this corresponds to testing

H_0 : β1 = 0

versus

H_a : β1 ≠ 0,

Since if $\beta_1 = 0$ then the model reduces to Y = β0 + ϵ, and X is not associated with Y . To test the null hypothesis, we need to determine whether βˆ1, our estimate for β1, is sufficiently far from zero that we can be confident that β1 is non-zero.

If SE($\beta\hat{}_1$) is small, then even relatively small values of $\beta\hat{}_1$ may provide strong evidence that $\beta_1 \neq 0$, and hence that there is a relationship between X and Y . In contrast, if SE(βˆ1) is large, then βˆ1 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 βˆ1 is away from 0. If there really is no relationship between X and Y , then we expect that above equation 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 value equal to |t| or larger, assuming β1 = 0. We call this probability the p-value.

Roughly speaking, we interpret the p-value as follows: a small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance, in the absence of any real association between the predictor and the 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.

Accessing the accuracy of the model

Once we have rejected the null hypothesis in favor of the alternative hypothesis, 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 R2 statistic.

- **Residual standard error (RSE)**

The linear regression model that associated with each observation is an error term ϵ. Due to the presence of this error term, even if we know the true regression line, we would not be able to perfectly predict Y from X. The RSE is an estimate of the standard deviation of ϵ. It is the average amount that the response will deviate from the true regression line. It is computed using the formula,

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

where RSS is

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

The RSE is considered a measure of the lack of fit of the regression model to the data. If the predictions obtained using the model are very close to the true outcome values—that is, if ˆyi ≈ yi for i = 1,...,n—then RSE will be small, and we can conclude that the model fits the data very well. On the other hand, if ˆyi is very far from yi for one or more observations, then the RSE may be quite large, indicating that the model doesn't fit the data well.

R^2 Statistic

The RSE provides an absolute measure of lack of fit of the regression 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 R^2, we use the formula

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

Where TSS is total sum of squares $\mathrm{TSS} = \sum (y_i - \bar{y})^2$

TSS measures the total variance in the response Y , and can be squares thought of as the amount of variability inherent in the response before the regression is performed. In contrast, 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 R^2 measures the proportion of variability in Y that can be explained using X. An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression. A number near 0 indicates that the regression did not explain much of the variability in the response; this might occur because the linear model is wrong.

Notice that $0 \leq R2 \leq 1$. This quantity is known as the coefficient of determination. Its intuitive interpretation makes it an often-used (and sometimes, over-used) summary for the value of the SLR.

2.1.2 Multiple inferences and simultaneous confidence bands

An important consideration in data analytics is that of multiple inferences or multiple comparisons, that is, when two or more inferences are performed on a single set of data. The multiplicity problem is common when there are many population parameters under study and confidence intervals or hypothesis tests are desired on each. For instance, suppose a hypothesis test is conducted on each of p parameters from the same set of data. Without correction for the multiplicity of tests being undertaken, there will be p opportunities to make a false positive error. Thus the experiment wise or family wise false positive error rate (FWER) will be much larger than the point wise significance level, α. Analogous concerns regarding the confidence coefficient, $1-\alpha$, arise when constructing multiple point wise confidence intervals. The simplest way to adjust for multiple inferences is to build the multiplicity into the estimation or testing scenario; for example, construct joint hypothesis tests or joint confidence regions that simultaneously contain all p parameters of interest.

The adjustments for multiplicity are prescribed whenever more than one inference is performed on a single set of data. This issue is a pertinent one in regression analysis, because there are often many different parameters or model features under study. For instance, analysts may wish to perform joint inferences on both of the unknown regression parameters, $\beta 0$ and $\beta 1$.

2.1.3 Regression diagnostics

The quality of an SLR fit can vary, and it is good analytic practice to examine the fit for potential violations of the model assumption, outlying observations, or other unusual features. A variety of diagnostic tools are available for this task. One of the most basic, and also most useful, is analysis of the residuals, $\mathbf{e}_\mathbf{i}$. In effect, the residuals estimate variation in the $\mathbf{Y}_\mathbf{i}$ s that remains – hence their name – after the SLR effect has been accounted for via the model fit. A simple tool to study residual variation graphs $\mathbf{e}_\mathbf{i}$ against the fitted values $\hat{Y}_i$. For the SLR model, one could equivalently plot **e$_i$** against the predictor variable xi. Known as a residual plot, this graphic very effectively visualizes the quality of an SLR fit. When the model is correct, E[**e$_i$**] = 0. Thus we expect the residual plot to display essentially random scatter about e = 0. Following figure shows a prototypical example.

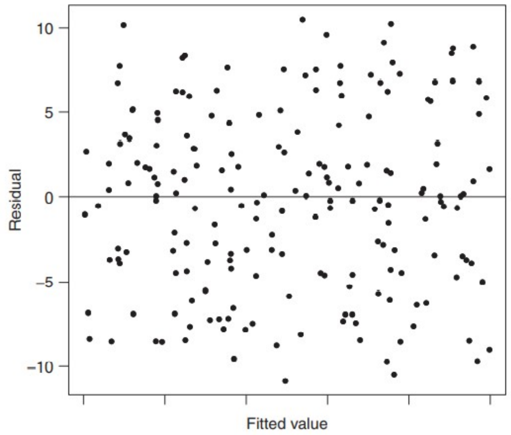
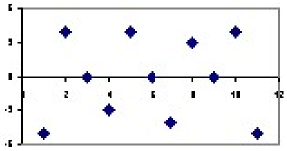


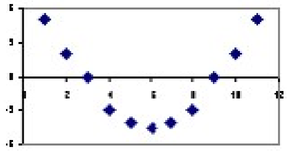
Figure Prototypical residual plot with desired random scatter about $e = 0$. Horizontal line at $e = 0$ included for reference.

A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a nonlinear model is more appropriate. The residual plot shows a fairly random pattern - the first residual is positive, the next two are negative, the fourth is positive, and the last residual is negative. This random pattern indicates that a linear model provides a decent fit to the data.

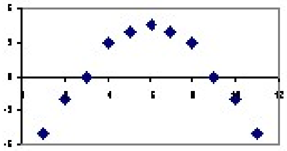
Below, the residual plots show three typical patterns. The first plot shows a random pattern, indicating a good fit for a linear model.



Random pattern



Non-random: U-shaped



Non-random: Inverted U

The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a nonlinear model.

By contrast, if a residual plot displays a clear pattern – of which there are many different types – then some model violation or discrepancy may be indicated. A common violation is heterogeneous variance, i.e. a departure from the assumption that Var[Yi] = σ2 is constant. This can be quickly recognized in a residual plot by variation(s) in the width of the residual pattern. For instance, if Var[Yi] increases with increasing xi, a plot of ei vs. xi will show a broadening of the residuals around e = 0 as xi increases. If the variation drops with increasing xi, then the pattern will be reversed. Plots against the fitted values, Ŷi, will show similar effects.

Another form of model violation observable in a residual plot is departure from linearity, where the mean response takes some form of nonlinear function. For instance, if E[Y] is quadratic in x but an SLR model is fit, the fitted values will under- and overestimate the true response in a recognizable.

The residuals are also valuable in assessing the quality of the normality. If normality is valid, the raw residuals should display normal variation, at least to a good approximation. Histograms, density estimates, or stemplots of the $\mathbf{e_i}$ may help in visualizing the effect. Or, the normal quantile plot. Under normality, a normal quantile plot of the residuals will appear roughly linear. If the plot deviates strongly from normal variation, however, a transformation of the original Yi. Residual plots can also aid in the identification of potential outliers.

Multiple linear regression

Multicollinearity

Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model. This means that an independent variable can be predicted from another independent variable in a regression model. For example, height and weight, household income and water consumption, mileage and price of a car, study time and leisure time, etc.

Multicollinearity can be a problem in a regression model because we would not be able to distinguish between the individual effects of the independent variables on the dependent variable. For example, let’s assume that in the following linear equation:

$$Y = \beta0+\beta1*X1+\beta2*X2$$

Coefficient β1 is the increase in Y for a unit increase in X1 while keeping X2 constant. But since X1 and X2 are highly correlated, changes in X1 would also cause changes in X2 and we would not be able to see their individual effect on Y. “ This makes the effects of X1 on Y difficult to distinguish from the effects of X2 on Y. ”

Multicollinearity may not affect the accuracy of the model as much. But we might lose reliability in determining the effects of individual features in your model – and that can be a problem when it comes to interpretability.

Multicollinearity could occur due to the following problems:

- Multicollinearity could exist because of the problems in the dataset at the time of creation. These problems could be because of poorly designed experiments, highly observational data, or the inability to manipulate the data:

  For example, determining the electricity consumption of a household from the household income and the number of electrical appliances. Here, we know that the number of electrical appliances in a household will increase with household income. However, this cannot be removed from the dataset
- Multicollinearity could also occur when new variables are created which are dependent on other variables:

  For example, creating a variable for BMI from the height and weight variables would include redundant information in the model
- Including identical variables in the dataset:

  For example, including variables for temperature in Fahrenheit and temperature in Celsius

- Inaccurate use of dummy variables can also cause a multicollinearity problem. This is called the Dummy variable trap:

  For example, in a dataset containing the status of marriage variable with two unique values: 'married', 'single'. Creating dummy variables for both of them would include redundant information. We can make do with only one variable containing 0/1 for 'married'/'single' status.

- Insufficient data in some cases can also cause multicollinearity problems

Detecting Multicollinearity using VIF

Multicollinearity can be detected via various methods. In this article, we will focus on the most common one – **VIF (Variable Inflation Factors)**.

" VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable. "

or

VIF score of an independent variable represents how well the variable is explained by other independent variables.

R^2 value is determined to find out how well an independent variable is described by the other independent variables. A high value of **R^2** means that the variable is highly correlated with the other variables. This is captured by the **VIF** which is denoted below:

$$\text{VIF} = \frac{1}{1-R^2}$$

So, the closer the **R^2** value to 1, the higher the value of VIF and the higher the multicollinearity with the particular independent variable.