Page Covered: PP 59 - 104

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

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

B0 and **B1** represent the intercept and slope terms in the linear model, which is commonly known as **model coefficients** and **parameters**.

Once we used training data to produce estimates B0 and B1, we can predict the following:

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

Where $\hat{y}$ indicates a prediction of Y on the basis of X = x. The hat symbol denotes the estimated value for the unknown coefficient /parameter.

Estimating the Coefficients
The goal is to obtain coefficient estimates B0 and B1 such that the linear model fits the available data well; therefore, we want to find an intercept B0 and a slope B1 such that the resulting line is as close as possible to the ground truth.

So.. to measure this closeness. The most common approach involves **minimizing the least-squares criterion**.

Let $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$ be the prediction for Y based on the ith value of X. Then *ei = yi - yi_hat* 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.

So.. we define **RSS** (residual sum of squares) as

$$RSS = e1^2 + e2^2 + \ldots + en^2$$

Or

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

The least squares approach chooses B0 and B1 to minimize the RSS. so we can obtain both as following:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad (3.4)$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

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, (3.4) defines the *least squares coefficient estimates* for simple linear regression.

Assessing the accuracy of the Coefficient Estimates

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

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

1. B0 is the intercept term, that is, the expected value of Y when X= 0, and
2. B1 is the slope --The average increase in Y associated with a one-unit increase in X.
3. e is the catch-all for what we miss with the model. (Typically assume to be independent)
    - Ie: the true function might not be linear
    - ie : may be there is other variables that cause variation in Y
    - Ie: measurement error

Suppose we are interested in finding out the population mean u of some random variable Y. But we dont know u.

One estimate for the population mean (u) :

- The sample mean (y^hat, $\widehat{u}$) is the same as the population mean (u)
- In the same way, coefficients B0 and B1 in linear regression define the population regression line and we want to estimate these two parameters with $\widehat{B0}$ and $\widehat{B1}$.

So if we use this estimate approach, this estimate is unbiased, in the sense that on average, we expect that u to equal to $\widehat{u}$.

Why? $\widehat{u}$ might overestimate u or underestimate u. So if we average it out from a large sample size of observations, the average value is u. **Hence, the unbiased estimator does not systematically over- or under-estimate the true parameter.**

HOW ACCURATE IS THE SAMPLE MEAN $\widehat{u}$ AS AN ESTIMATE OF u?

To understand how far off will that single estimate of $\widehat{u}$ be, we can compute the standard error of $\widehat{u}$, written as SE($\widehat{u}$).

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n},$$

Where σ is the standard deviation of each of the realizations yi of Y.

-The standard error tells us the average amount that this estimate $\widehat{u}$ differs from the actual value of u.

From the above equation, we know that the more observations we have, the smaller the standard error of $\widehat{u}$.

What about how close $\widehat{B0}$ and $\widehat{B1}$ are to the true values B0 **and** B1**?**

To compute the standard errors associated with $\widehat{B0}$ and $\widehat{B1}$, we can 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 σ2 = Var(ε). For these formulas to be strictly valid, we need to assume that the errors εi for each observation are uncorrelated with common variance σ2.

In general, σ2 is not known, but can be estimated from the data. The estimate of σ is known as the **residual standard error**, and is given by the formula RSE = Square Root of (RSS/(n-2)).

For linear regression, 95% confidence interval for B1 is approximately takes the form

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

For example, the 95% confidence interval for **B1** is [0.042, 0.053], so we can say that for each $1000 increase in B1 associated predictor, there will be an average increase in Y between 42 and 53 units, with the assumption that all other variables remained the

same. For the **B0**, when no predictor variable is considered, the result is fallen somewhere between 6130 and 7940.

To prove that a certain variable X truly has a relationship with response Y:

Standard errors can 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

H0: B1 = 0

Vs the alternative hypothesis

Ha: There is some relationship between X and Y

Ha: B1 != 0

Here is the logic:

- To overturn the null hypothesis and proof that there is some relationship between B1 and Y, it really depends on the accuracy of B1, that is , it depend on SE(B1).
- If SE(B1) is small, then small values of B1 still can provide evidence that B1 !=0 and therefore there is a relationship between X and Y.
- If SE(B1) is large, then B1 must be large in absolute value in order to prove that there is a relationship between X and Y.

So we can compute a t-statistic, given by

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

Which measure the number of standard deviation that B1 is away from 0. If there is no relationship between X and Y, then we expect t-test will have a t-distribution with n-2 degree of freedom. The t-distribution has a bell shape and for values of n greater than approximately 30, it is similar to normal distribution.

P-Value: the probability of observing any number equal to |t| or larger in absolute value, assuming B1 = 0.

- P-value is small: There is an association between predictor and response.

  → reject the H0.

- P-value is large: There is no association between predictor and response.

  → reject the Ha.

- P-value cutoff: 5 or 1%

Assessing the accuracy of the Model (quantify the extent to which the model fits the data):

1. Residual standard Error (RSE)
2. R^2 statistic

RSE: an estimate of the standard deviation of e (error term). It is the average amount that the response will deviate from the true regression line.

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \qquad (3.15)$$

Note that RSS was defined in Section 3.1.1, and is given by the formula

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \qquad (3.16)$$

The RSE is considered a measure of the *lack of fit* of the model to the data. But since we are measuring in the units of Y, we dont know what constitutes a good RSE.

R^2 provides an alternative measure of fit, which takes the form of 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

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{3.17}$$

where $\text{TSS} = \sum(y_i - \bar{y})^2$ is the *total sum of squares*, and RSS is defined in (3.16). 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. 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, or the inherent error σ^2 is high,

TSS is the sum of square of the difference between each observation and the mean.

RSS is the sum of square of the difference between each observation and the predicted response.

R^2 will be in between 0 and 1 and the greater the R^2 value, the better the model is explaining the relationship, since it explains more variance of the data.

Multiple Linear Regression

In general, suppose that we have p distinct predictors, then the multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

Jan. 30th, 2021
bh563@cornell.edu

Estimating regression coefficients

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2. \end{aligned}$$

To assess the relationship between response and predictors, we are using hypothesis test and F-statistic

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

versus the alternative

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

This hypothesis test is performed by computing the *F-statistic*,

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

where, as with simple linear regression, $\text{TSS} = \sum(y_i - \bar{y})^2$ and $\text{RSS} = \sum(y_i - \hat{y}_i)^2$. If the linear model assumptions are correct, one can show that

$$E\{\text{RSS}/(n-p-1)\} = \sigma^2$$

and that, provided H_0 is true,

$$E\{(\text{TSS} - \text{RSS})/p\} = \sigma^2.$$

Hence, when there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1. On the other hand, if H_a is true, then $E\{(\text{TSS} - \text{RSS})/p\} > \sigma^2$, so we expect F to be greater than 1.

A large F-statistic suggests that at least one of the predictors is related to the response.

How large of F-statistic value for us to reject H0? (depends on n and p)

When n is large, an F-statistic that is just a little larger than 1 might still prove against H0.

When n is small, an F-statistic that is very large is needed to reject H0.

F-statistic ~<= 1, H0 is true

F-statistic >>1, Ha is true

F-statistic ~>= 1 && large n → can weakly reject H0

F-statistic >> 1 && small n → reject H0

Deciding on Variables:

1. **Forward Selection:** We begin with the null model -- a model that contains an intercept but no predictor. We then fit p simple linear regressions and add to the null model the variable that results in the lowest RSS and continue adding the model with lowest RSS variable for new two-variable model and keep going until triggering the stop rule.

2. **Backward Selection:** We start with all variables in the model and remove the variable with the largest p-value, that is the variable that is the least statistically significant. We continue to take away variables until a stopping rule is reached.

Model Fits (Multi-linear Regression)

When adding more variables, R^2 will increase. But the amount it increases speaks about the proportion of the variance it explains.

RSE is defined as

$$\text{RSE} = \sqrt{\frac{1}{n-p-1}\text{RSS}}, \tag{3.25}$$

which simplifies to (3.15) for a simple linear regression. Thus, models with more variables can have higher RSE if the decrease in RSS is small relative to the increase in p.

Potential Problems

1. Non-linearity of the response-predictor relationships.

   - What if the linear regression is not linear?

     Residual plots are a useful tool for identifying non-linearity. If it is a U-shape, then it means there is non-linearity in the true function.

2. Correlation of error terms.

   - We tend to assume that error term e1,e2,e3 are uncorrelated. However, there could be correlation between errors. In the context of time-seres data, we can see that residual plots shows a consistent of positive or negative errors (below/above line)

3. Non-constant variance of error terms.

   - We tend to assume that the variance of the error is constant, but it is not the case. We can identify non-constant variance in the errors, or heteroscedasticity, from the presence of a funnel shape in teh resident plot. This funnel shape is because the variances of the erro terms may increase wit the value of the response.
   - Solution: one possible solution is to transform the response Y using a concave function such as logY or Squart root of Y.

4. Outliers.

   - An outliner is a point for which yi is far from the value predicted by the model. Studentized residual plots can be used to identify outliners. We compute by dividing each residuals ei by its estimated standard error.

5. High-leverage points.

In order to quantify an observation's leverage, we compute the *leverage statistic*. A large value of this statistic indicates an observation with high leverage. For a simple linear regression,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}. \quad (3.37)$$

6. Collinearity.

Collinearity refers to the situation in which two or more predictor variables are closely related to one another. The presence of collinearity can pose problems in regression, since it can be difficult to separate out the individual affect of collinear variables on the response. Collinearity can result in a decline in the t-statistic because collinearity reduces the accuracy of the estimates of coefficients. A decline of t-statistic might lead to failing to reject H0.

- We can use a contour plot to detect this problem.
- We can look into correlation matrix of the predictors. (multicollinearity cannot be detected)
- We can look at VIF (variance inflation factor): (the ratio of the variance of B^hat when fitting the full model) divided by (the variance of B^hat if fit on its own).
    - The smallest possible value for VIF is 1, indicates no collinearity.
    - A VIF of 5 and above can be problematic

collinearity. The VIF for each variable can be computed using the formula

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors. If $R^2_{X_j|X_{-j}}$ is close to one, then collinearity is present, and so the VIF will be large.

When comparing KNN and linear regression, by increasing dimension, a given observation might not have nearby neighbors (Curse of Dimensionality) and so there will be a bad KNN fit.

So in general, **parametric methods will tend to outperform non-parametric approaches when there is a small number of observations per predictor.**