Linear Regression Chapter 3

Johann PIedras

11/11/2020

Contents

Simple Linear Regression

Simple linear regression lives up to its name: it is a very straightforward approach for predicting a quantitative response Y on the basis of a sin- gle predictor variable X. It assumes that there is approximately a linear relationship between X and Y . Mathematically, we can write this linear relationship as

(3.1)$Y \approx \beta_0 + \beta_1 X$

We are 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

$Sales \approx \beta_0 + \beta_1 \times TV$

In Equation 3.1,β_0 and β_1 are two unknown constants that represent the intercept and slope terms in the linear model. Together, β_0 and β_1 are known as the model coefficients or parameters. 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
### Prediction of y and the basis of x (3.2)$\hat{y} = \hat{\beta_0} + \hat{\beta_1} x$

where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$. Here we use a hat symbol,$\hat{}$, to denote the estimated value for an unknown parameter or coefficient, or to denote the predicted value of the response.

Linear regression represented as a function

The full linear regression function is also just like below
$\hat{Y} = \hat{f}(X) + \varepsilon$

Residual Error

This method include ε which is also considered the **residual error**.
One method of calculating the ith residual is through using the predicting and actual value.
$e_i = y_i - \hat{y}_i$
There are properties of a fitted regression line. One to keep in mind is that the sum of residuals is zero.
$\sum_{i=1}^{n} e_i = 0$

Resdiaul Sum Of Squares

If we add all the residual errors we get the **RSS** also known as the **Residual Sum Squares**
$RSS = \sum e_i^2 = e_1^2 + e_2^2 + e_3^2 + ... + e_n^2$, with n being the number of data points. The sum of squared residuals is a minimum. This was the requirement to be satisfied in deriving the east squares estimators of the regression parameters since the criterion is to minimize residuals for our estimation parameters.

Least Square Coefficients

To find the $\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

To find the $\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$

Sample Means

$\bar{y}$ is $= \frac{1}{n}\sum_{i=1}^{n} y_i$ which is also the sample mean
$\bar{x}$ is $= \frac{1}{n}\sum_{i=1}^{n} x_i$

EXAMPLE We use the advertising data set to quickly represent how to implement a simple linear regression model.
First we import the data into our R environment

```
data <- read.csv('advertising.csv', header = TRUE)
```

Using the lm() function in R

```
model.1 = lm(sales ~ TV, data = data)
```

Now we want to see the statistics we cam derive by implementing the linear model.

```
summary(model.1)
```

```
##
## Call:
## lm(formula = sales ~ TV, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

From the output, one can see that there are the residuals split into quartiles (min, 1Q, median, 3Q, Max).

Additionally we can observe the coefficients. $\hat{\beta}_0$ being 7.0325935 and $\hat{\beta}_1$ being 0.0475366 . In other words, we can translate this analysis as "ab additional $1000 spent on TV advertising would be associated with selling approximately 47.5 more units".
Example) $\hat{\beta}_1 \times 1000 == .047537 * 1000 = 47.5$ Units

Assessing the Accuracy of the Coefficient Estimates

If f is to be approximated by a linear function, then we can write this relationship as
(3.5) $Y = \beta_0 + \beta_1 X + \varepsilon$

Where the intercept is β_0 , which is the expected value of Y when X = 0
And the slope is β_1 which would be the average increase in Y associated with one-unit increase in X.
ε is a catch all because the relationship is probably not linear, and there may be other variables that cause variation in Y, so we typically assum ethat the error term is independent of X.

Population Regression Line

Earlier I shared $Y = \beta_0 + \beta_1 X + \varepsilon$. This is considered the **population regression line** which is the best linear approximation to the true relationship between X and Y.

Least Square Line

The least squares regression coefficient estimates characterize the **least squares line** $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Bias and Unbiased

If we use the sample mean $\hat{\mu}$ to estimate μ, the population mean, this estimate would be considered unbiased because on average we expect the sample mean to be close to the population mean. If we used other variables to estimate the population mean, then we could be overstimating or underestimating the population mean which would then make our statistic bias. This idea translate over to how we estimate our $\beta_0 and \beta_1$. We wont always get the exact value of our coefficient estimators, but we can get close if we have a HUGE amount of data.

Standard Error of Population Example

Continuing with population mean, to assess the accuracy of our estimates we can use the standard error of $\hat{\mu}$
(3.7) $Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n}$ Where σ is the standard deviation of each of the realizations of y_i of Y. Roughly speaking the standard error tells us the average amount that this estimate $\hat{\mu}$ differs from the actual value of μ. TAKE NOTE that the larger n is smaller the standard error for $\hat{\mu}$ is

Standard Error of Point Estimators

$SE(\hat{\beta_0})^2 = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}]$ $SE(\hat{\beta_1})^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
Where $\sigma^2 = Var(\varepsilon)$. These formulas are strictly valid however if we assume that the errors for each observation are uncorrelated with common variance σ^2

Residual Stanard Error

Generally we do not know what σ^2 is not known, but it can be estimated with data. The estimate of σ is known as the **Residual Standard Error**
$RSE = \sqrt{\frac{RSS}{n-2}}$

Confidence Intervals

Standard errors can be used to compute confidence intervals. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. The range is defined in terms of lower and upper limits computed from the sample of data. For linear regression, the 95% confidence interval for β_1 approximately takes the form:
(3.9) $\hat{\beta}_1 \pm 2 \times SE(\hat{\beta}_2)$
That is, there is approximately a 95% chance that the interval
$[\hat{\beta}_1 - 2 \times SE(\hat{\beta}_2), \hat{\beta}_1 + 2 \times SE(\hat{\beta}_2)]$
will contain the true value of β_1

Hypothesis Test

Standard errors can be used to perform a **hypothesis test** on coefficients.
(3.12) $H_0 : \beta_1 = 0$ or in otherwords, “There is no relationship between X and Y. This is the **null hypothesis**
(3.13) $H_a or H_1 : \beta_1 \neq 0$ which in other words is”There is some relationship between X and Y"

Note that since $\beta_1 = 0$ The model is reduced to $Y = \beta_0 + \varepsilon$, and X is not associated with Y.

To test the null hypothesis we need to determine whether $\hat{\beta_1}$ is sufficiently far from zero that we can be confident that β_1 is non-zero. BUT how far is enough? Well that depends on the estimated coefficient which is used to find SE($\hat{\beta_1}$), the standard error.

T-Statistic

In contrast if the 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. At this point we would computer the **t-statistic**. The test statistic is used to measure the number of standard deviations the $\hat{\beta_1}$ is away from 0. $t = \frac{\hat{\beta_1} - 0}{SE(\hat{\beta_1})}$

P-Value

Consequently we can also create a p-value rather than the T-Statistic. P-values are hotly debated and I’m not knowledgeable enough to defend, so roughly speaking a small p-value invalidates 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 a SMALL p-value can be used to infer that there is an association between the predictor and the response. Therefore we **we reject the null hypothesis**, that is that we declare that there exist a relationship between X and Y - given that the p-value is small.

Aseessing the Accuracy of the Model

If the null hypothesis is reject, which in our example means that there is a relationship between X and Y, the next move would be to quantify the extent to which the model fits the data.
For linear regression that quantity usually is the **RSE and R Square**

Residual Standard Error

$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}$
Recall that RSS is $\sum_{i=1}^{n}(y_i - \hat{y_i})^2$
The RSE provides an absolute measure of lack of fit the model to the data, but its measured in the units of Y so its not always clear. An alternative measure would be the R^2

R Squared

R^2 takes the form of a proportion. The proportion of variance explained and so it always take the value between 0 and 1, and is independent of the scale of Y.
$R^2 = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS}$
where 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.
In contrast the 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 preforming the regression, and R^2 measures the proportion of variability in Y that can be explained using X.
if R^2 is near 0, that 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 σ^2 is high, or both.