

15.7.3 Perform Multiple Linear Regression

After reminding himself of how helpful linear regression is, Jeremy decides to keep going—he's really on a roll, and multiple linear regression is ready for him!

Multiple linear regression is a statistical model that extends the scope and flexibility of a simple linear regression model. Instead of using a single independent variable to account for all variability observed in the dependent variable, a multiple linear regression uses multiple independent variables to account for parts of the total variance observed in the dependent variable.

As a result, the linear regression equation is no longer $y = mx + b$. Instead, the multiple linear regression equation becomes $y = m_1x_1 + m_2x_2 + \dots + m_nx_n + b$, for all independent x variables and their m coefficients.

In actuality, a multiple linear regression is a simple linear regression in disguise—all of the assumptions, hypotheses, and outputs are the same. The only difference between multiple linear regression and simple linear regression is how we will evaluate the outputs.

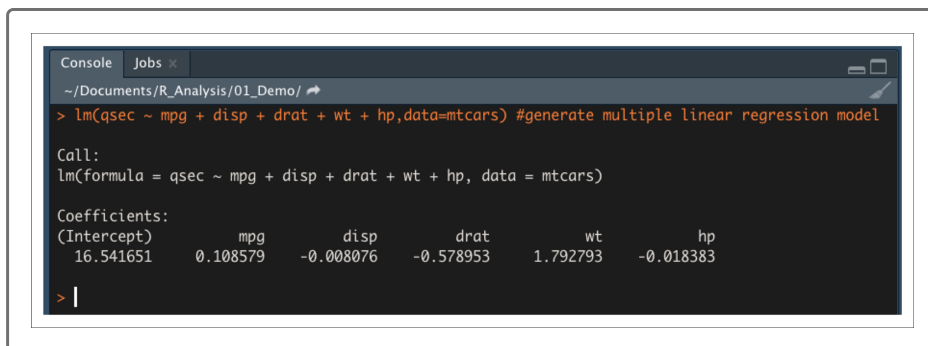
When it comes to multiple linear regression, we'll look at each independent variable to determine if there is a significant relationship with the dependent variable. Once we have evaluated each independent variable, we'll evaluate the r-squared value of the model to determine if the model sufficiently predicts our dependent variable.

To practice multiple linear regression, let's revisit our mtcars dataset. From our last example, we determined that quarter-mile time was not

adequately predicted from just horsepower. To better predict the quarter-mile time (`qsec`) dependent variable, we can add other variables of interest such as fuel efficiency (`mpg`), engine size (`disp`), rear axle ratio (`drat`), vehicle weight (`wt`), and horsepower (`hp`) as independent variables to our multiple linear regression model.

In R, our multiple linear regression statement is as follows:

```
> lm(qsec ~ mpg + disp + drat + wt + hp, data=mtcars) #generate multiple line
```



The screenshot shows an R console window with the following content:

```
~/Documents/R_Analysis/01_Demo/ ➔  
> lm(qsec ~ mpg + disp + drat + wt + hp, data=mtcars) #generate multiple linear regression model  
  
Call:  
lm(formula = qsec ~ mpg + disp + drat + wt + hp, data = mtcars)  
  
Coefficients:  
(Intercept)      mpg      disp      drat      wt      hp  
16.541651    0.108579   -0.008076   -0.578953    1.792793   -0.018383  
> |
```

Similar to the simple linear regression, the output of multiple linear regression using the `lm()` function produces the coefficients for each variable in the linear equation.

NOTE

Because multiple linear regression models use multiple variables and dimensions, they are almost impossible to plot and visualize.

Now that we have our multiple linear regression model, we need to obtain our statistical metrics using the `summary()` function. In your R console, use the following statement:

```
>summary(lm(qsec ~ mpg + disp + drat + wt + hp,data=mtcars)) #generate summa
```

```

Console Jobs
~/Documents/R_Analysis/01_Demo/ ➔
> summary(lm(qsec ~ mpg + disp + drat + wt + hp,data=mtcars))

Call:
lm(formula = qsec ~ mpg + disp + drat + wt + hp, data = mtcars)

Residuals:
    Min       1Q   Median       3Q      Max
-1.6628 -0.6138  0.0706  0.4087  3.3885

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.541651   3.413109   4.847 5.04e-05 ***
mpg          0.108579   0.077911   1.394  0.17523
disp        -0.008076   0.004384  -1.842  0.07689 .
drat         -0.578953   0.551771  -1.049  0.30371
wt           1.792793   0.513897   3.489  0.00175 **
hp           -0.018383   0.005421  -3.391  0.00223 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.053 on 26 degrees of freedom
Multiple R-squared:  0.7085,    Adjusted R-squared:  0.6524
F-statistic: 12.64 on 5 and 26 Df,    p-value: 2.767e-06
> |

```

In addition to overall model fit and the statistical test for slope, most data scientists would be curious about the contribution of each variable to the multiple linear regression model. To determine which variables provide a significant contribution to the linear model, we must look at the individual variable p-values.

In the summary output, each `Pr(>|t|)` value represents the probability that each coefficient contributes a random amount of variance to the linear model. According to our results, vehicle weight and horsepower (as well as intercept) are statistically unlikely to provide random amounts of variance to the linear model. In other words the vehicle weight and horsepower have a significant impact on quarter-mile race time. When an intercept is statistically significant, it means that the intercept term explains a significant amount of variability in the dependent variable when all independent variables are equal to zero. Depending on our dataset, a significant intercept could mean that the significant features (such as

weight and horsepower) may need scaling or transforming to help improve the predictive power of the model. Alternatively, it may mean that there are other variables that can help explain the variability of our dependent variable that have not been included in our model. Depending on the dataset and desired performance of the model, you may want to change your independent variables and/or transform them and then re-evaluate your coefficients and significance.

Despite the number of significant variables, the multiple linear regression model outperformed the simple linear regression. According to the summary output, the r-squared value has increased from 0.50 in the simple linear regression model to 0.71 in our multiple linear regression model while the p-value remained significant.

CAUTION

Although the multiple linear regression model is far better at predicting our current dataset, the lack of significant variables is evidence of overfitting. Overfitting means that the performance of a model performs well with a current dataset, but fails to generalize and predict future data correctly. Later in this course we'll learn more about overfitting and ways to avoid it.

Depending on the dataset, the questions being asked, and the audience, a simple linear regression model may be more appropriate than a multiple linear regression model. However, the amount of information that can be obtained and analyzed will be far greater using a multiple linear regression.

As with any data model, it takes practice to learn how to identify variables of interest, select an appropriate model, and refine a model to increase performance. Before moving to the next section, take some time to perform correlation analysis on our previous datasets. Then use the correlation analysis to identify potential variables of interest. Once you have variables of interest, practice generating simple and multiple linear regression models to try and create accurate predictive models.