**Logistic Regression Analysis Using ‘mtcars’ Data Set**

1. **Overview**

This analysis is very similar to the previous regression analysis using the ‘mtcars’ data set, except this time we’re going to see if we can apply a logistic regression model to the data in order to answer some common question about the dependent variable: transmission (0 = automatic, 1 = manual).

1. **Data**

|  |  |  |
| --- | --- | --- |
| Variable Name | Type | Description |
| mpg | Numerical | Miles per gallon (US) |
| cyl | Categorical | Number of cylinders |
| disp | Numerical | Engine displacement (in.3) |
| hp | Numerical | Gross horsepower |
| drat | Numerical | Rear axle ratio |
| wt | Numerical | Weight (lb/1000) |
| gsec | Numerical | ¼ mile time (s) |
| vs | Categorical | V/S |
| am† | Categorical | Transmission (0 = automatic, 1 = manual) |
| gear | Categorical | Number of forward gears |
| carb | Categorical | Number of carburators |
| † dependent variable | | |

With **am** as our dependent variable, the question we want answered begins to write itself. That is, can we use any combination of the other remaining variables to predict whether a car has an automatic or a manual transmission?

For the sake of simplicity, I will use three explanatory variables: **wt** (weight), **hp** (horsepower) and **disp** (engine displacement, in.3) and see if these variables can predict the type of transmission in the car.

1. **Logistic Regression Model(s)**

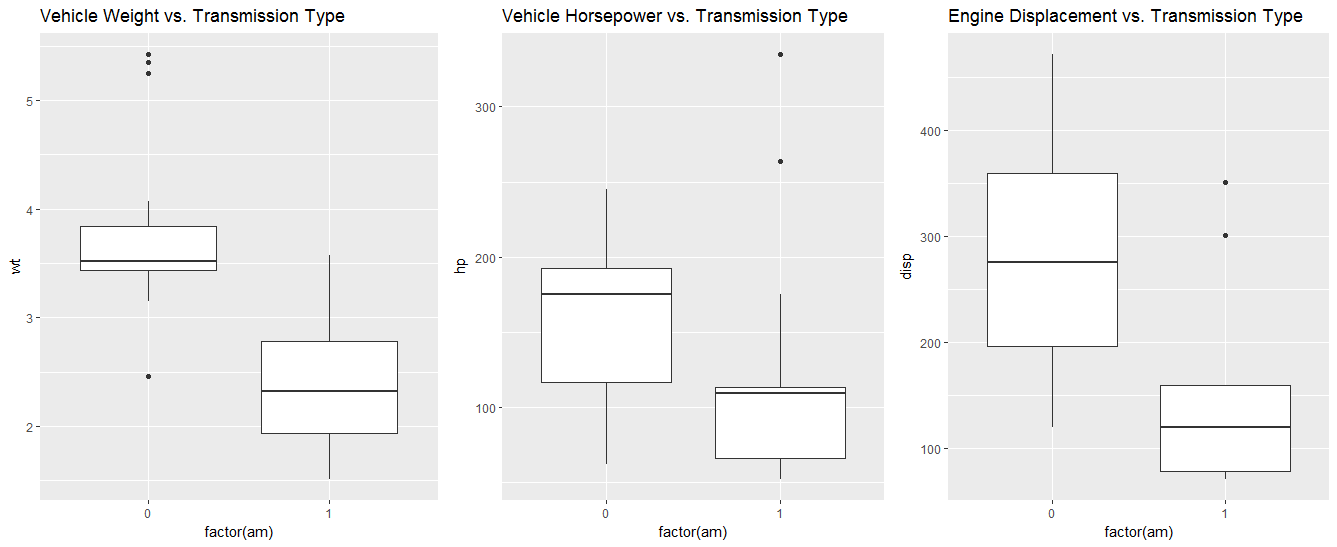
The general plan is to first run a basic regression model where all the independent variables are used to explain the dependent variable. Depending on how accurate the model is, either conclude with that model or attenuate it to try and get a better model.

Before getting to that, however, we should perform some diagnostics on the data that we’re going to use. In particular, we’ll plot the three independent variables at the two factor levels of **am** to see how they compare.

library(ggplot2)

library(gridExtra)

library(datasets)  
library(dplyr)

  
am\_wt <- ggplot(mtcars, aes(x=factor(am), y=wt)) + geom\_boxplot() + ggtitle("Vehicle Weight vs. Transmission Type")  
am\_hp <- ggplot(mtcars, aes(x=factor(am), y=hp)) + geom\_boxplot() + ggtitle("Vehicle Horsepower vs. Transmission Type")  
am\_disp <- ggplot(mtcars, aes(x=factor(am), y=disp)) + geom\_boxplot() + ggtitle("Engine Displacement vs. Transmission Type")  
grid.arrange(am\_wt, am\_hp, am\_disp, ncol=3)

In every case, it appears there is a notable difference in the explanatory variable depending on the type of transmission. Therefore, there is no need to exclude any of the variables from our initial model.

am\_prob1 <- glm(formula= am ~ wt + hp + disp, data=mtcars, family=binomial(link="logit"))  
summary(am\_prob1)  
## Call:  
## glm(formula = am ~ wt + hp + disp, family = binomial(link = "logit"),   
## data = mtcars)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2074 -0.1285 -0.0092 0.1346 1.3480   
##   
## Coefficients:  
## Estimate Std. Error z value **Pr(>|z|)**   
## (Intercept) 14.37948 7.65348 1.879 **0.0603 .**  
## **wt**  -5.95398 3.23118 -1.843 **0.0654 .**  
## **hp** 0.06105 0.05219 1.170 **0.2421**   
## **disp** -0.02731 0.03922 -0.696 **0.4863**   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 43.2297 on 31 degrees of freedom  
## Residual deviance: 9.1412 on 28 degrees of freedom  
## AIC: 17.141  
## Number of Fisher Scoring iterations: 8

Taking a look at the summary of our first model, we see that the p-value associated with **wt** is approaching statistical significance, given *α* = 0.05, and that **hp** and **dist** are nowhere near significant. Going forward, we’ll keep **wt** as is and test the interaction term of **hp** and **dist** to see if that changes their significance to the model.

am\_prob2 <- glm(formula= am ~ wt + hp:disp, data=mtcars, family=binomial(link="logit"))  
summary(am\_prob2)

## Call:  
## glm(formula = am ~ wt + hp:disp, family = binomial(link = "logit"),   
## data = mtcars)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.36010 -0.24476 -0.01513 0.10979 1.28637   
##   
## Coefficients:  
## Estimate Std. Error z value **Pr(>|z|)**   
## (Intercept) 2.248e+01 9.176e+00 2.450 **0.0143 \***  
## **wt** -8.384e+00 3.396e+00 -2.469 **0.0136 \***  
## **hp:disp** 7.787e-05 3.998e-05 1.948 **0.0514 .**  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 43.230 on 31 degrees of freedom  
## Residual deviance: 12.468 on 29 degrees of freedom  
## AIC: 18.468  
##   
## Number of Fisher Scoring iterations: 7

Using the interaction term of **hp:dist**, we see the p-value approaching significance. I chose to keep it in the model because of the nature of this analysis. We’re attempting to explain the type of transmission based on other vehicle metrics. Our output is one of two categories and that lack of specificity in the output means we don’t need to achieve a model where all terms are significant with a p-value under 0.05. Essentially, we can “get by” with a lower significance especially when it is only 0.0014 units away from the assumed *α*.

1. **Model Evaluation**

Now that I have a model that I’m satisfied with, I want to test the null hypothesis that the two explanatory variables taken together don’t explain the dependent variable, **am**. The statistical test to use is a chi-square test with two degrees of freedom, shown below.

with(am\_prob2, pchisq(null.deviance - deviance, df.null - df.residual,   
 lower.tail = FALSE))

## [1] 2.089904e-07

Since 2.09e-7 is less than *α* = 0.05, we reject the null hypothesis that the type of transmission is independent of our inputs—**wt** and **hp:dist**. The last thing to do is to analyze the odds ratios of the inputs and interpret what they mean in the context of this analysis.

exp(coef(am\_prob2))

## (Intercept) wt hp:disp   
## 5.791245e+09 **2.285936e-04** **1.000078e+00**

exp(cbind(OR = coef(am\_prob2), confint(am\_prob2)))

## OR 2.5 % 97.5 %  
## (Intercept) 5.791245e+09 1.528354e+04 2.127118e+21  
## wt 2.285936e-04 1.115814e-08 2.629151e-02  
## hp:disp 1.000078e+00 1.000016e+00 1.000186e+00

* **wt**: the odds of a vehicle having a manual transmission rather than automatic increases by **2.29** for each unit increase in the vehicle’s weight.
* **hp:dist**: the odds of a vehicle having a manual transmission rather than automatic increases by **1** for each unit increase in horsepower between two vehicles with an equal engine displacement.