Coursera - Regression Models - Project 1

Charles Bradbury

Executive Summary

In this report, we will analyze `mtcars` data set and explore the relationship between a set of variables and miles per gallon (MPG). These data were obtained from the 1974 *Motor Trend* US magazine. They comprise the fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–1974). Regression models and exploratory data analysis will be used to explore how **automatic** (am = 0) and **manual** (am = 1) transmissions features affect the **MPG**. T-tests will show the performance difference between cars with automatic and manual transmission, which is about 7 MPG more for cars with manual transmission than those with automatic transmission. We will also fit some linear regression models and select the one with highest Adjusted R-squared value. Given that weight and 1/4 mile time are held constant, manual transmitted cars are 14.079 + (-4.141)*weight more MPG (miles per gallon) on average better than automatic transmitted cars. The analysis shows that cars that are lighter in weight with a manual transmission and cars that are heavier in weight with an automatic transmission will have higher MPG values.

Load needed libraries/packages

```
library(ggplot2)
```

Exploratory Data Analysis

Let's load the data set `mtcars` and transform some variables from `numeric` class to `factor` class.

```
data(mtcars)
mtcars[1:3, ] # Get a small sample of data.
```

```
##                mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
```

```
dim(mtcars)
```

```
## [1] 32 11
```

```
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
attach(mtcars)
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

Let's perform some basic exploratory data analyses. Refer to the **Appendix: Figures** section for the plots and charts. According to the box plot, it is shown that manual transmission yields higher values of MPG in general. And as for the pair graph, we can see some higher correlations between variables like "wt" (weight), "disp" (displacement), "cyl" (number of cylinders) and "hp" (horsepower).

Inference

At this step, we assume the null hypothesis as the MPG of the automatic and manual transmissions are from the same population (assuming the MPG has a normal distribution). Let's use the two sample T-test to show it.

```
result <- t.test(mpg ~ am)
result$p.value
```

```
## [1] 0.001373638
```

```
result$estimate
```

```
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

Since the p-value is 0.00137, we reject the null hypothesis. Therefore, the automatic and manual transmissions are from different populations. The average/mean for MPG of manual transmitted cars is about 7 more MPG than that of automatic transmitted cars.

Regression Analysis

Let's fit the full model as the following.

```
fullModel <- lm(mpg ~ ., data=mtcars)
summary(fullModel) # These resusults are hidden from the report.
```

This model has the Residual standard error as 2.833 on 15 degrees of freedom. The Adjusted R-squared value is 0.779, which means that the model can explain about 78% of the variance of the MPG variable. However, none of the coefficients are significant at 0.05 significant level.

Now, let's use backward selection to select some statistically significant variables.

```
stepModel <- step(fullModel, k=log(nrow(mtcars)))
summary(stepModel) # These results are hidden from the report.
```

This model is "mpg ~ wt + qsec + am". It has the Residual standard error as 2.459 on 28 degrees of freedom. The Adjusted R-squared value is 0.8336, which means that the model can explain about 83% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level.

Refer to the **Appendix: Figures** section for the plots/charts. According to the scatter plot, there appears to be an interaction term between "wt" variable and "am" variable, since automatic cars tend to weigh more than manual cars. Therefore, we have the following model including the interaction term:

```
amIntWtModel<-lm(mpg ~ wt + qsec + am + wt:am, data=mtcars)
summary(amIntWtModel) # These results are hidden from the report.
```

This model has the Residual standard error as 2.084 on 27 degrees of freedom. The Adjusted R-squared value is 0.8804, which indicates the model can explain about 88% of the variance of the MPG variable. All coefficients are significant at 0.05 significant level.

Let's fit the simple model with MPG as the outcome variable and Transmission as the predictor variable.

```
amModel<-lm(mpg ~ am, data=mtcars)
summary(amModel) # These results are hiiden from the report.
```

On average, a car with an automatic transmission achieves 17.147 mpg; if it has manual transmission, it acheives 7.245 mpg more. This model has the Residual standard error as 4.902 on 30 degrees of freedom. The Adjusted R-squared value is 0.3385, which means that the model can explain about 34% of the variance of the MPG variable. The low Adjusted R-squared value also indicates that we need to add other variables to the model.

Let's select the final model.

```
anova(amModel, stepModel, fullModel, amIntWtModel)
confint(amIntWtModel) # These results are hidden from the report.
```

We choose the model with the highest Adjusted R-squared value, "mpg ~ wt + qsec + am + wt:am".

```
summary(amIntWtModel)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.723053  5.8990407  1.648243 0.1108925394
## wt          -2.936531  0.6660253 -4.409038 0.0001488947
## qsec         1.016974  0.2520152  4.035366 0.0004030165
## am1         14.079428  3.4352512  4.098515 0.0003408693
## wt:am1      -4.141376  1.1968119 -3.460340 0.0018085763
```

The result shows that when "wt" (weight lb/1000) and "qsec" (1/4 mile time) remain constant, cars with manual transmission has 14.079 + (-4.141)*wt more MPG (miles per gallon) on average than cars with automatic transmission. Therefore, a car with a manual transmission that weighs 2000 lbs acheives 5.797 more MPG than a car with an automatic transmission that has both the same weight and 1/4 mile time.
Residual Analysis and Diagnostics
Refer to the **Appendix: Figures** section for the plots/charts. According to the residual plots, we can verify the following: 1. The Residuals vs. Fitted plot shows no consistent pattern, supporting the accuracy of the independence assumption.
2. The Normal Q-Q plot indicates that the residuals are normally distributed because the points lie closely to the line.
3. The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed.
4. The Residuals vs. Leverage argues that no outliers are present, as all values fall well within the 0.5 bands.

As for the Dfbetas, the measure of how much an observation has effected the estimate of a regression coefficient, we get the following result:

```
sum((abs(dfbetas(amIntWtModel)))>1)
```

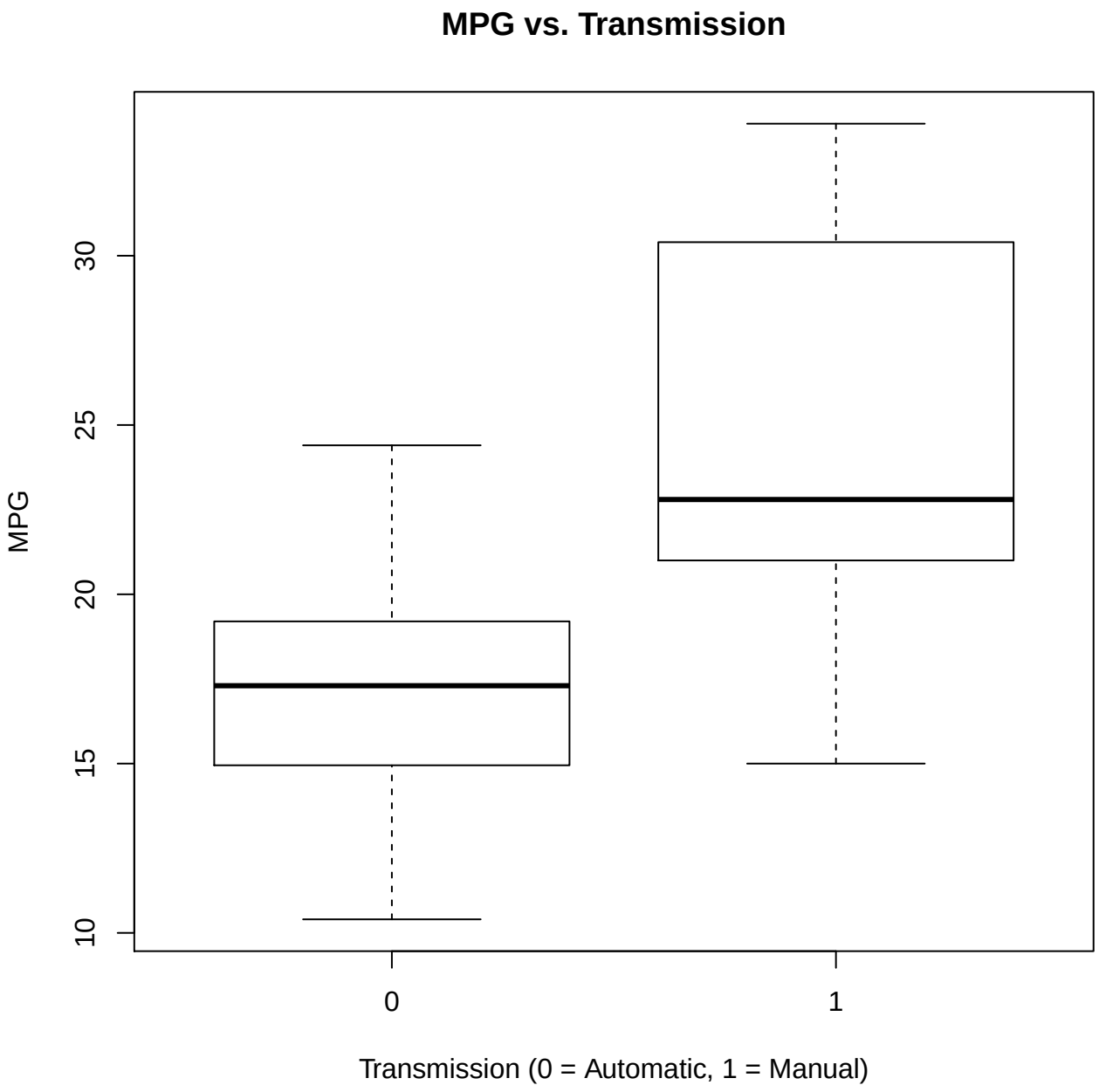
```
## [1] 0
```

Therefore, the above analyses meet all basic assumptions of linear regression and well answer the questions.

Appendix: Figures

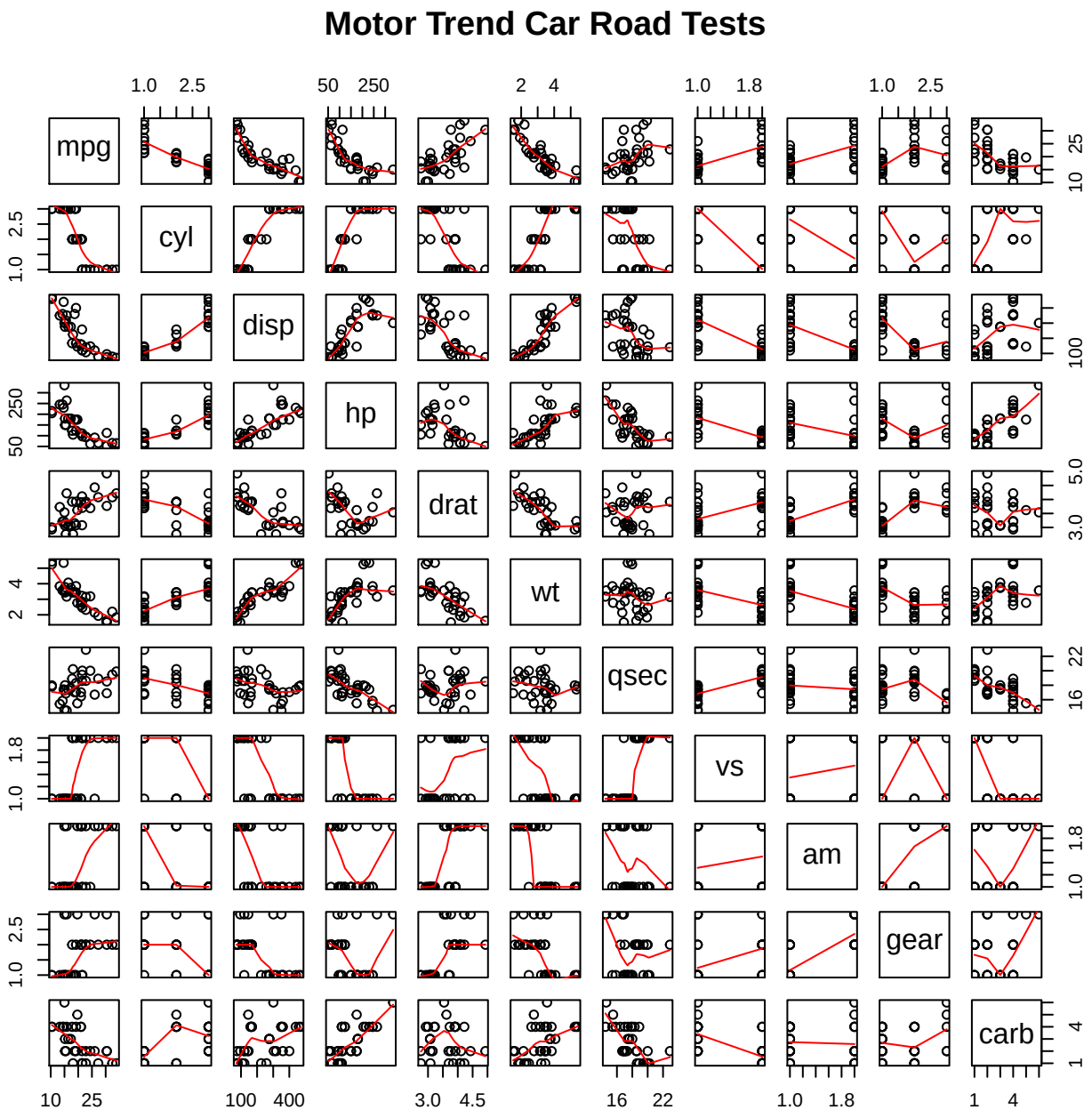
1. Boxplot: MPG vs. Transmission

```
boxplot(mpg ~ am, xlab="Transmission (0 = Automatic, 1 = Manual)", ylab="MPG",
        main="MPG vs. Transmission")
```



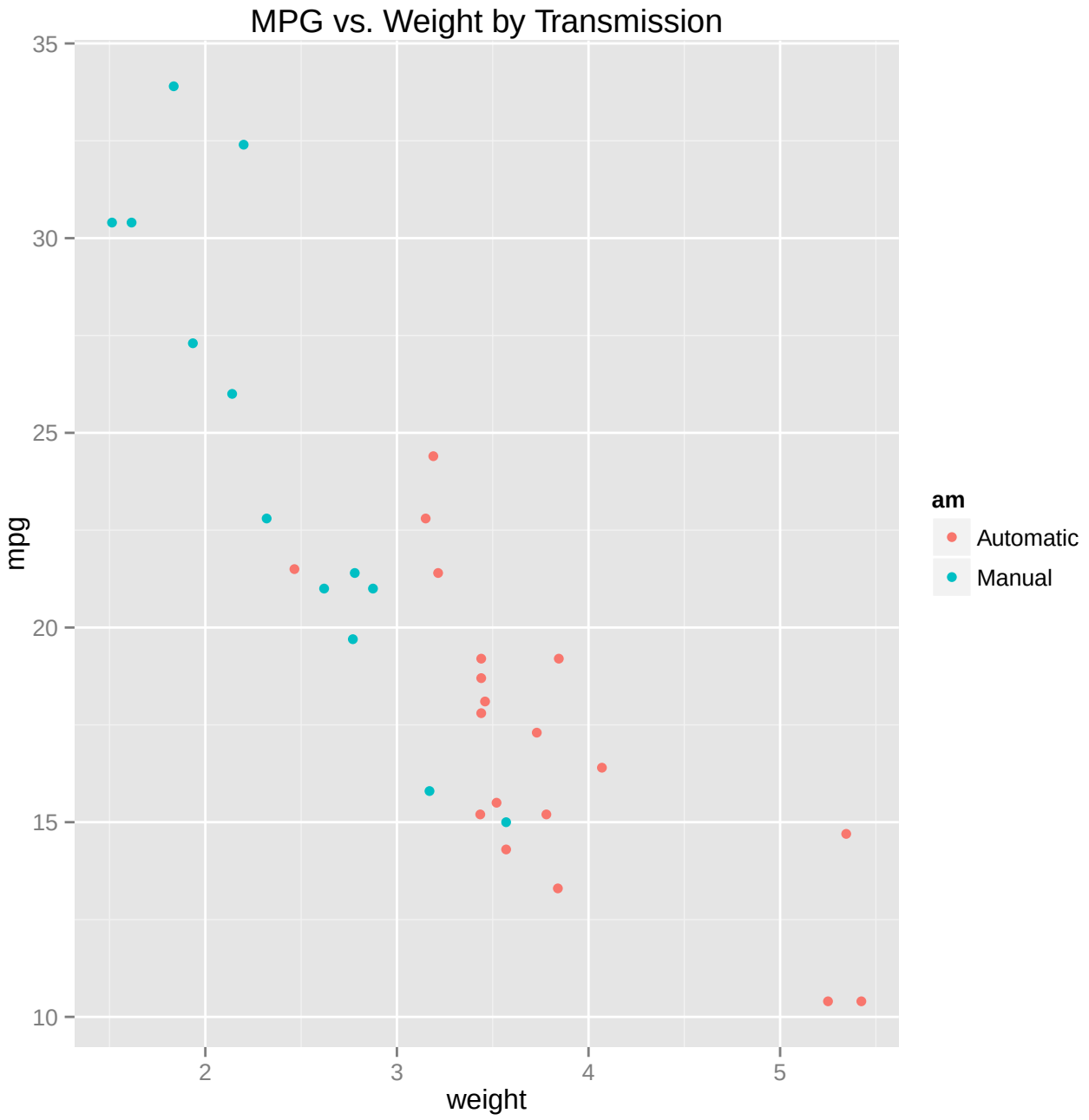
2. Pair Graph: Motor Trend Car Road Tests

```
pairs(mtcars, panel=panel.smooth, main="Motor Trend Car Road Tests")
```



3. Scatter Plot: MPG vs. Weight by Transmission

```
ggplot(mtcars, aes(x=wt, y=mpg, group=am, color=am, height=3, width=3)) + geom_point() +
scale_colour_discrete(labels=c("Automatic", "Manual")) +
xlab("weight") + ggtitle("MPG vs. Weight by Transmission")
```



4. Residual Plots

```
par(mfrow = c(2, 2))
plot(amIntWtModel)
```