MechaCarChallenge

Ray Whelan

MPG Regession

```
#Load the Mecha Car data
mecha_mpg <- read.csv(file='MechaCar_mpg.csv',check.names=F,stringsAsFactors = F)

#Print out a Covariance matrix
mpg_matrix <-as.matrix(mecha_mpg[,c("vehicle_length", "vehicle_weight", "spoiler_angle", "ground_clearar
cor(mpg_matrix)
```

```
                  vehicle_length vehicle_weight spoiler_angle ground_clearance
vehicle_length       1.00000000    -0.12271790    0.02577114      -0.31663112
vehicle_weight      -0.12271790     1.00000000   -0.11307851       0.08511338
spoiler_angle        0.02577114    -0.11307851    1.00000000      -0.21112057
ground_clearance    -0.31663112     0.08511338   -0.21112057       1.00000000
AWD                  0.08565668    -0.03698098   -0.09120266      -0.15214456
mpg                  0.60947984     0.09068314   -0.02083999       0.32874886
                         AWD         mpg
vehicle_length    0.08565668  0.60947984
vehicle_weight   -0.03698098  0.09068314
spoiler_angle    -0.09120266 -0.02083999
ground_clearance -0.15214456  0.32874886
AWD               1.00000000 -0.14166977
mpg              -0.14166977  1.00000000
```

The above correlation matrix shows us that two of our five independent variables have an outsized impact on our dependent variables, the prototypes fuel economy ('mpg'): vehicle length and ground clearance. The former alone has 60% correlation while the latter has 32% correlation. Thus these two variables should be our strongest predictors in the multiple linear regression we will run in the next step.

```
#Create our multiple linear regression model and print out the summary table
mecha_multi <- lm(mpg~ vehicle_length + vehicle_weight + spoiler_angle + ground_clearance + AWD, data=me
summary(mecha_multi)
```

```
Call:
lm(formula = mpg ~ vehicle_length + vehicle_weight + spoiler_angle +
    ground_clearance + AWD, data = mecha_mpg)

Residuals:
     Min       1Q   Median       3Q      Max
-19.4701  -4.4994  -0.0692   5.4433  18.5849
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.040e+02  1.585e+01  -6.559 5.08e-08 ***
vehicle_length    6.267e+00  6.553e-01   9.563 2.60e-12 ***
vehicle_weight    1.245e-03  6.890e-04   1.807   0.0776 .
spoiler_angle     6.877e-02  6.653e-02   1.034   0.3069
ground_clearance  3.546e+00  5.412e-01   6.551 5.21e-08 ***
AWD              -3.411e+00  2.535e+00  -1.346   0.1852
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.774 on 44 degrees of freedom
Multiple R-squared:  0.7149,    Adjusted R-squared:  0.6825
F-statistic: 22.07 on 5 and 44 DF,  p-value: 5.35e-11
```

From the above code we can see two variables aside from the intercept have a statistically significant impact on the car's MPG: vehicle length and ground clearance (as denoted by the asterisks). The model estimates that each additional unit of vehicle length increases the car's fuel economy by 6.2 mpg and each additional unit of ground clearance increases fuel economy by 3.5 mpg.

With an R-squared value of ~.7, we can confirm that this model is an effective predictor of prototype mpg as the five dependent variables in our model can combined account for 70% of the variance in our dependent variable. Thus, this model will prove useful to Autos'R'Us's engineers in maximizing fuel efficiency.

```
mecha_length<-lm(mpg~vehicle_length, data=mecha_mpg)
summary(mecha_length)
```

```
Call:
lm(formula = mpg ~ vehicle_length, data = mecha_mpg)

Residuals:
    Min      1Q  Median      3Q     Max
-26.303  -7.160  -1.231   9.374  26.670

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -25.0622    13.2960  -1.885   0.0655 .
vehicle_length   4.6733     0.8774   5.326 2.63e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.47 on 48 degrees of freedom
Multiple R-squared:  0.3715,    Adjusted R-squared:  0.3584
F-statistic: 28.37 on 1 and 48 DF,  p-value: 2.632e-06
```

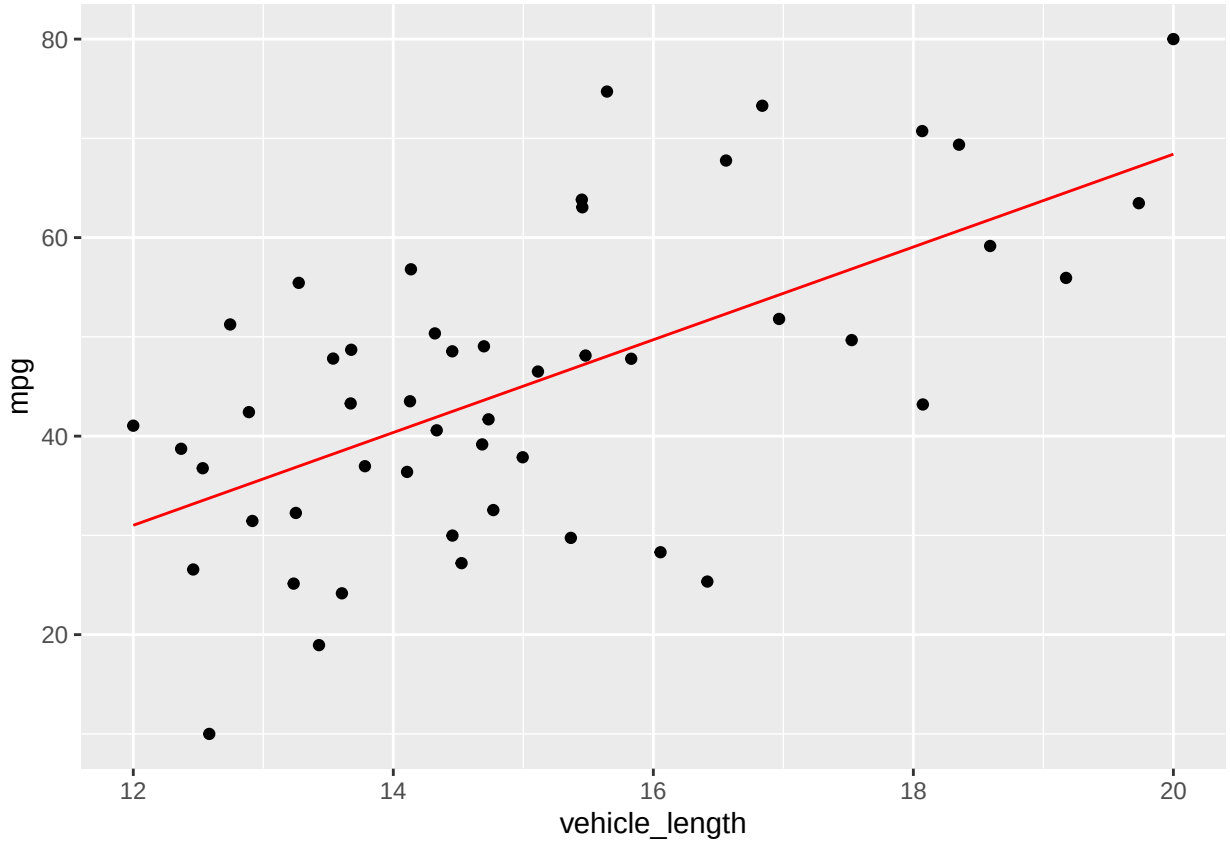
```
yvals <- mecha_length$coefficients['vehicle_length']*mecha_mpg$vehicle_length + mecha_length$coefficien

plt <- ggplot(mecha_mpg,aes(x=vehicle_length,y=mpg))
yvals
```

```
 [1] 43.62138 33.51360 68.40332 37.69282 47.13982 42.48332 40.85899 47.15849
 [9] 42.81741 47.27356 43.95884 41.85839 43.55905 40.96428 61.81690 35.17424
```

```
[17] 45.56345 35.29848 36.96579 32.73731 59.37544 45.01550 33.74836 42.49014
[25] 38.51692 51.65188 60.69871 48.91519 38.20126 33.17129 43.78710 54.23089
[33] 59.39765 34.50025 49.96907 31.01711 38.83198 36.78267 36.86110 53.62683
[41] 39.34478 56.84094 67.16165 52.32340 40.99761 48.04505 38.85108 64.54569
[49] 41.92478 46.74610
```

```
plt + geom_point()  + geom_line(aes(y=yvals), color = "red")
```



To further prove the point of our model's effectiveness, the above plot of mpg against vehicle length shows graphically what we have confirmed already that there is strong positive correlation between the two variables.

Suspension Coil Summary

```
#Read in the suspension coil data
susp <- read.csv(file='Suspension_Coil.csv',check.names=F,stringsAsFactors = F)

susp_sum <- susp %>%
  group_by(Manufacturing_Lot) %>%
  summarize(Mean_PSI= mean(PSI),Median_PSI= median(PSI),Variance_PSI= var(PSI), StdDev_PSI= sd(PSI))

susp_sum
```

```
# A tibble: 3 x 5
```

```
  Manufacturing_Lot Mean_PSI Median_PSI Variance_PSI StdDev_PSI
  <chr>                <dbl>      <dbl>        <dbl>      <dbl>
1 Lot1                 1500       1500         0.980      0.990
2 Lot2                 1500.      1500         7.47       2.73
3 Lot3                 1496.      1498.      170.        13.0
```

Our supension coil summary table shows us there is significant difference in the quality control of the three lots of suspension coils. Our first lost is highly controlled with almost negligible variance from our mean of 1500 PSI. The second lot shows more variance but is still well within our variance tolerance of 100 PSI. Meanwhile our third lot shows significantly higher variance, exceeding our control variance tolerance at 170 PSI. This lot should be discarded and the manufacturing process investigated to determine the causes of the drop in quality control.

Suspension Coil T-Test

```
#Run T-Test for the whole sample group against the expected mean of 1500 PSI
t.test(susp$PSI, mu = 1500)
```

```
    One Sample t-test

data:  susp$PSI
t = -1.8931, df = 149, p-value = 0.06028
alternative hypothesis: true mean is not equal to 1500
95 percent confidence interval:
 1497.507 1500.053
sample estimates:
mean of x
  1498.78
```

The above t-test shows that we can not with certainty reject the null hypothesis there is no statistical difference between our sample mean and our population mean. The p-value is both too high at .06 and our 95% confidence interval just barely extends past our expected mean. However, given the variance we saw between lots, it would be wise to break our data set down by lot and run a t-test on each one.

```
#Subset our dataset by lot number
Lot1 <- susp %>%
  subset(Manufacturing_Lot == "Lot1", select = PSI)

Lot2 <- susp %>%
  subset(Manufacturing_Lot == "Lot2", select = PSI)

Lot3 <- susp %>%
  subset(Manufacturing_Lot == "Lot3", select = PSI)

#Run T-Test for each lot against the expected mean of 1500 PSI
t.test(Lot1$PSI, mu = 1500)
```

```
    One Sample t-test
```

```
data:  Lot1$PSI
t = 0, df = 49, p-value = 1
alternative hypothesis: true mean is not equal to 1500
95 percent confidence interval:
 1499.719 1500.281
sample estimates:
mean of x
     1500
```

```
t.test(Lot2$PSI, mu = 1500)
```

```

    One Sample t-test

data:  Lot2$PSI
t = 0.51745, df = 49, p-value = 0.6072
alternative hypothesis: true mean is not equal to 1500
95 percent confidence interval:
 1499.423 1500.977
sample estimates:
mean of x
   1500.2
```

```
t.test(Lot3$PSI, mu = 1500)
```

```

    One Sample t-test

data:  Lot3$PSI
t = -2.0916, df = 49, p-value = 0.04168
alternative hypothesis: true mean is not equal to 1500
95 percent confidence interval:
 1492.431 1499.849
sample estimates:
mean of x
  1496.14
```

The above t-tests confirm that our first and second lot means are statistically identical to our expected mean of 1500 PSI. However, we can confirm with greater than 95% certainty that our third lot's variance from our expected mean is statistically significant with a greater than 95% confidence.

Self Designed Study

One potential area of inquiry Autos'R'Us should explore to test the MechaCar prototype against the current market selection is to test how well new vehicle model sales perform against how similar they are to other offerings already on the market. In other words, will customers buy a new model of car just because it's new or does it have to be substantially different in size, capactiy, form factor, performance, etc. from other market offerings in order to generate positive sales? By testing this, Autos'R'Us can determine whether the market's needs are being met by current market offerings or if the market demands innovation, providing incentive for Autos'R'Us to take greater risks with the MechaCar prototype.

In order to test this, Autos'R'Us should sort recent vehicle model launches into two classes, "similar to current market offerings" and "unique new offerings". Because this is a dichotomous class, the study analysts will

have to, perhaps somewhat arbitrarily, determine how different a car needs to be to fall into the latter category, but for example, a four door economy sedan would fall into the former while a high-performance SUV might fall into the ladder. Once again, there will need to be man judgment calls made here and the model may need to be tweaked with different categorizing criteria.

Once our categories have been set, a series of multiple linear regressions should be run with several control criteria (e.g. cost, mpg, number of seats, etc.), or independent dichotomous variable, and several different sales performance statistics such as: first year sales as a percentage of total vehicle sales, 5-year average sales percentage, and the difference between the first year sales and 5 year average. The first will tell us if new car sales perform better or worse when the car is unique to the market, and the second two will tell us how much holding power more unique offerings have relative to tried and true car models once their novelty has worn off and customers have decided whether or not the car fills a legitimate market niche.

Autos'R'Us will need to determine from the p-values of the slopes of our dichotomous variable on our outcome variables whether or nt there is any statistically significant effect and what to infer from the information they provide. The null hypothesis would be that the uniqueness of the car has no effect on sales performance, while the alternative hypothesis is that the uniqueness effects sales performance one way or the other.