

Motor Trend magazine - Data analysis of influence on MPG for Automatic vs. Manual Transmission.

by jmvilaverde

Thursday, June 18, 2015

Context

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

“Is an automatic or manual transmission better for MPG” “Quantify the MPG difference between automatic and manual transmissions”

Question

Take the mtcars data set and write up an analysis to answer their question using regression models and exploratory data analyses.

Your report must be:

Written as a PDF printout of a compiled (using knitr) R markdown document. Brief. Roughly the equivalent of 2 pages or less for the main text. Supporting figures in an appendix can be included up to 5 total pages including the 2 for the main report. The appendix can only include figures. Include a first paragraph executive summary. Upload your PDF by clicking the Upload button below the text box.

Peer Grading

The criteria that your classmates will use to evaluate and grade your work are shown below. Each criteria is binary: (1 point = criteria met acceptably; 0 points = criteria not met acceptably) Your Course Project score will be the sum of the points and will count as 40% of your final grade in the course.

Evaluation/feedback on the above work

- Did the student interpret the coefficients correctly?
- Did the student do some exploratory data analyses?
- Did the student fit multiple models and detail their strategy for model selection?
- Did the student answer the questions of interest or detail why the question(s) is (are) not answerable?
- Did the student do a residual plot and some diagnostics?
- Did the student quantify the uncertainty in their conclusions and/or perform an inference correctly?
- Was the report brief (about 2 pages long) for the main body of the report and no longer than 5 with supporting appendix of figures?
- Did the report include an executive summary?
- Was the report done in Rmd (knitr)?

Report brief (about 2 pages)

Executive summary (First paragraph)

Automatic/Manual transmission is better...

Data adquisition

1. Data adquisition and Initial structure analysis

```
data(mtcars)

#?mtcars

#Show colnames of mtcars
names(mtcars)
```

1.1. Variables

```
## [1] "mpg" "cyl" "disp" "hp" "drat" "wt" "qsec" "vs" "am" "gear"
## [11] "carb"
```

mtcars

Format: A data frame with 32 observations on 11 variables.

Variable	Units
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (lb/1000)
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

```
#Show structure of mtcars
str(mtcars)
```

```
## 'data.frame': 32 obs. of 11 variables:
## $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
## $ cyl : num 6 6 4 6 8 6 8 4 4 6 ...
## $ disp: num 160 160 108 258 360 ...
## $ hp : num 110 110 93 110 175 105 245 62 95 123 ...
## $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
```

```
## $ wt : num 2.62 2.88 2.32 3.21 3.44 ...
## $ qsec: num 16.5 17 18.6 19.4 17 ...
## $ vs : num 0 0 1 1 0 1 0 1 1 1 ...
## $ am : num 1 1 1 0 0 0 0 0 0 0 ...
## $ gear: num 4 4 4 3 3 3 3 4 4 4 ...
## $ carb: num 4 4 1 1 2 1 4 2 2 4 ...
```

2.Initial analysis Selected variables:

- Predictor: $X = \text{am}$, Transmission with values 0 for automatic, 1 for manual.
- Outcome: $Y = \text{mpg}$, Miles/(US) gallon

Linear regression model formula:

- $Y = \beta_0 + \beta_1 X \rightarrow \text{mpg} = \beta_0 + \beta_1 \text{am}$

Create initial model:

Calculate β_0 (Intercept): $\hat{\beta}_0 = \bar{Y} + \hat{\beta}_1 \bar{X}$

Calculate β_1 (Slope): $\hat{\beta}_1 = \text{Cor}(Y, X) \frac{Sd(Y)}{Sd(X)}$

Fitting the best line:

- $\sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_i))^2$ to minimize the distance.

```
#Theoretical formula
y <- mtcars$mpg
x <- mtcars$am
beta1 <- cor(y,x)*(sd(y)/sd(x))
beta0 <- mean(y) - beta1 * mean(x)
#R function
model.Initial <- lm(y ~ x)
coeffs <- coef(model.Initial)

#Comparison
rbind(c(beta0, beta1), coeffs)
```

```
##      (Intercept)      x
##      17.14737 7.244939
## coeffs 17.14737 7.244939
```

For linear model $Y = \beta_0 + \beta_1 X \rightarrow \text{mpg} = \beta_0 + \beta_1 \text{am}$

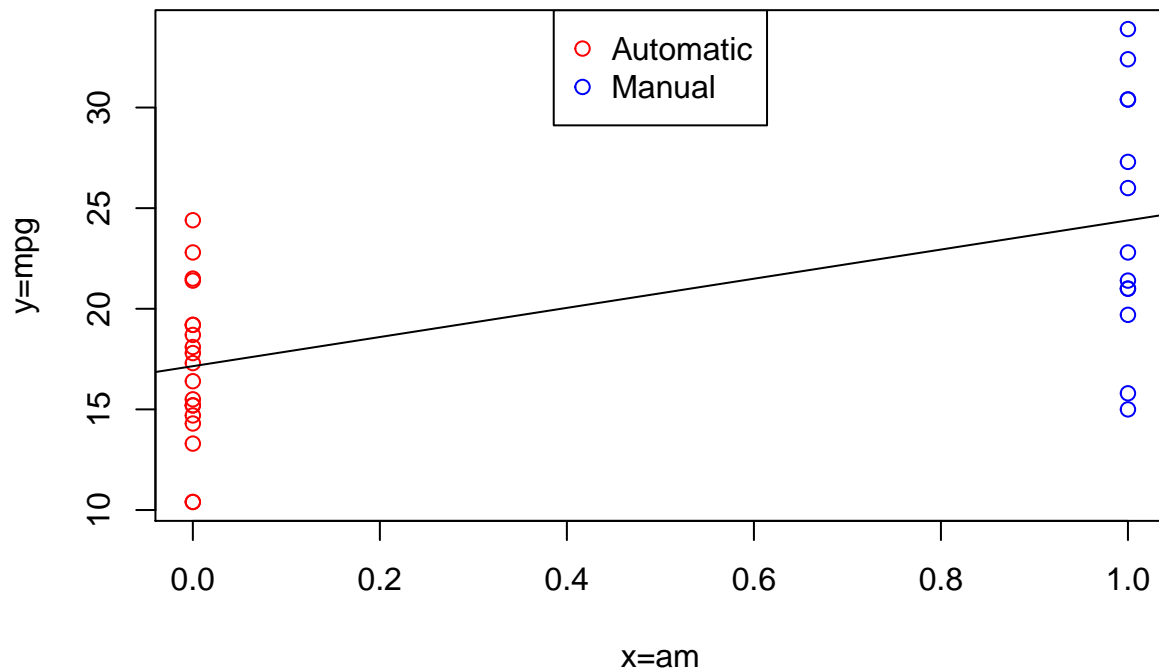
- Intercept $\beta_0 = 17.1473684$
- Slope $\beta_1 = 7.2449393$

For this linear regression model, the expected difference in MPG is 7.2449393 when the car have manual transmission in comparison to the same car with automatic transmission. The Intercept 17.1473684 is the expected MPG of a automatic transmission car.

```

color <- ifelse(x==0,"red","blue")
plot(x,y,col=color,xlab="x=am",ylab="y=mpg")
legend("top",c("Automatic","Manual"),col=c("red","blue"),pch=1)
abline(model.Initial)

```



View Figure 1 for plot with regression line.

Normalizing data????

3.Basic regression model with additive Gaussian errors. Into point 2 is obtained a linear regression model that fits the data, but doesn't take in consideration the impact of the others variables. We are going to analyze gaussian errors for Initial Model.

Selected variables:

- **Predictor:** X = am, Transmission with values 0 for automatic, 1 for manual.
- **Outcome:** Y = mpg, Miles/(US) gallon

Probabilistic model for linear regression:

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \rightarrow mpg_i = \beta_0 + \beta_1 am_i + \epsilon_i$
- ϵ_i are assumed iid $N(\mu_i, \delta^2)$.
- Note, $E[Y_i|X_i = x_i] = \mu_i = \beta_0 + \beta_1 x_i$ and $Var(Y_i|X_i = x_i) = \delta^2$.
- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ and $\hat{\beta}_1 = Cor(Y, X) \frac{Sd(Y)}{Sd(X)}$.

Residuals analysis:

- Observed outcome i is Y_i at a predictor value X_i .
- Predicted outcome i is \hat{Y}_i at a predictor value X_i is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.
- Residual is the difference between observed and predicted: $e_i = Y_i - \hat{Y}_i$, the vertical distance between the observed data point and the regression line.
- Least squares minimizes $\sum_{i=1}^n e_i^2$.
- e_i can be thought of as estimates of the ϵ_i .

```
#Calculate residuals
```

```
e.ModelInitial <- y-(beta0+beta1*x)
e.ModelInitial
```

```
## [1] -3.3923077 -3.3923077 -1.5923077  4.2526316  1.5526316  0.9526316
## [7] -2.8473684  7.2526316  5.6526316  2.0526316  0.6526316 -0.7473684
## [13]  0.1526316 -1.9473684 -6.7473684 -6.7473684 -2.4473684  8.0076923
## [19]  6.0076923  9.5076923  4.3526316 -1.6473684 -1.9473684 -3.8473684
## [25]  2.0526316  2.9076923  1.6076923  6.0076923 -8.5923077 -4.6923077
## [31] -9.3923077 -2.9923077
```

```
#R function for residuals
```

```
resid(model.Initial)
```

```
##          1          2          3          4          5          6
## -3.3923077 -3.3923077 -1.5923077  4.2526316  1.5526316  0.9526316
##          7          8          9         10         11         12
## -2.8473684  7.2526316  5.6526316  2.0526316  0.6526316 -0.7473684
##         13         14         15         16         17         18
##  0.1526316 -1.9473684 -6.7473684 -6.7473684 -2.4473684  8.0076923
##         19         20         21         22         23         24
##  6.0076923  9.5076923  4.3526316 -1.6473684 -1.9473684 -3.8473684
##         25         26         27         28         29         30
##  2.0526316  2.9076923  1.6076923  6.0076923 -8.5923077 -4.6923077
##         31         32
## -9.3923077 -2.9923077
```

```
#Calculate predicted y (mpg)
```

```
yhat <- predict(model.Initial)
```

```
#Calculate max difference between residual and observed Y - predicted Y (y hat)
```

```
max(abs(e.ModelInitial -(y - yhat)))
```

```
## [1] 3.552714e-15
```

```
library(dplyr)
```

```
##
```

```
## Attaching package: 'dplyr'
```

```
##
```

```
## The following object is masked from 'package:stats':
```

```
##
```

```
## filter
```

```
##
## The following objects are masked from 'package:base':
##
## intersect, setdiff, setequal, union
```

```
with(mtcars, aggregate(mpg, list(c(am)), mean))
```

```
##   Group.1      x
## 1      0 17.14737
## 2      1 24.39231
```

```
#Create initial model mpg vs. am
model <- with(mtcars, lm(mpg ~ am))
with(mtcars, aggregate(mpg, list(c(am)), mean))
```

```
##   Group.1      x
## 1      0 17.14737
## 2      1 24.39231
```

```
summary(model)$coef[1,1]
```

```
## [1] 17.14737
```

```
summary(model)$coef[1,1]+summary(model)$coef[2,1]
```

```
## [1] 24.39231
```

```
print("The mean of mpg is higher in vehicles with manual transmission")
```

```
## [1] "The mean of mpg is higher in vehicles with manual transmission"
```

```
summary(model)$coef
```

```
##           Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 17.147368   1.124603  15.247492 1.133983e-15
## am          7.244939   1.764422   4.106127 2.850207e-04
```

FORMULAS

- Empirical mean: $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
- $\tilde{X} = X_i - \bar{X}$, the mean of \tilde{X} is 0. This is to centering the randoms.

“Is an automatic or manual transmission better for MPG”

Main Body + Appendix only figures (not more than 5)