Regression Models Course Project

Teo Lo Piparo

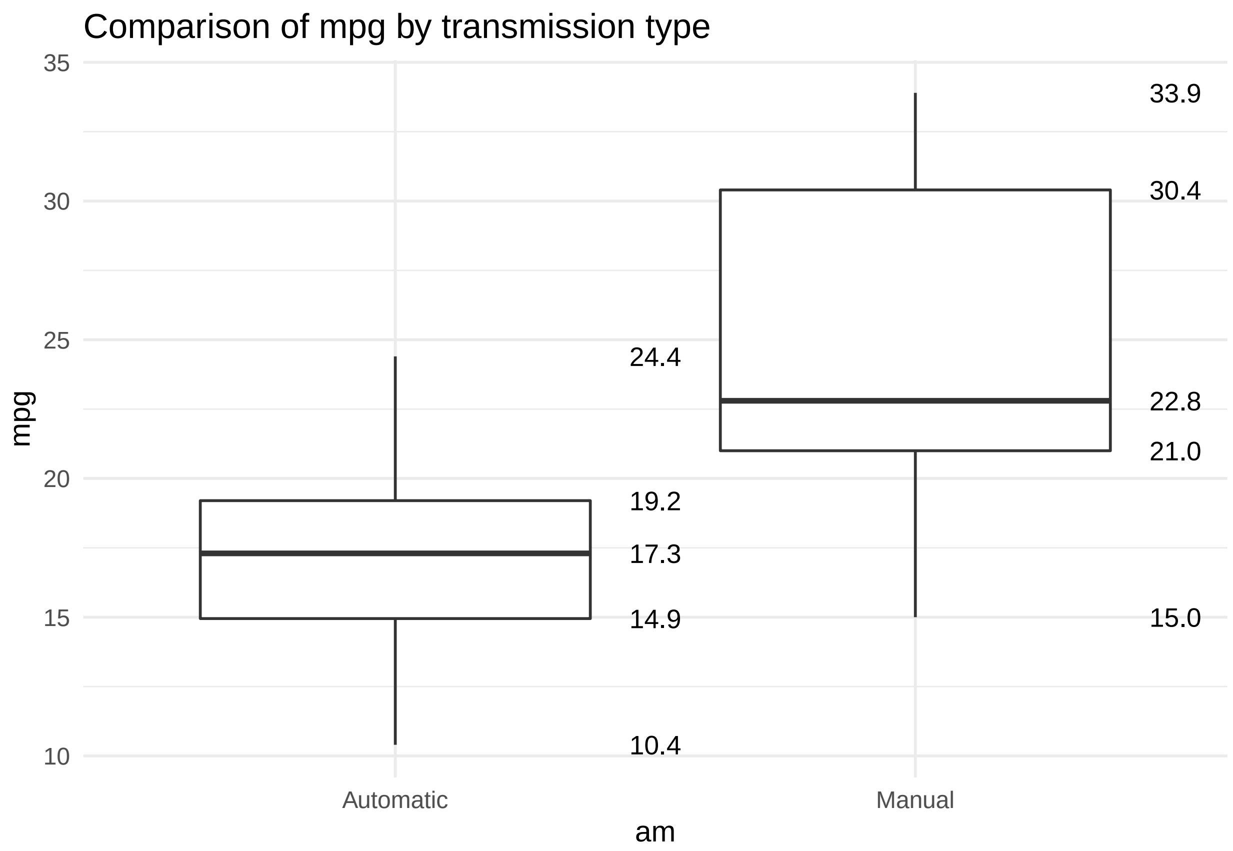
20/11/2020

```r
library(tidyverse)
library(ggplot2)
library(broom)
```

Exploratory Analysis

Boxplot

```r
mtcars %>%
    select(mpg, am) %>%
    group_by(am) %>%
    mutate(am = factor(am, levels = c(0,1))) %>%
    ggplot(aes(x = am, y = mpg, group = am)) +
    geom_boxplot() +
    scale_x_discrete(labels = c("Automatic", "Manual"),
                       breaks = c("0","1")) +
    theme_minimal() +
    ggtitle("Comparison of mpg by transmission type") +
    stat_summary(geom="text", fun=quantile,
                  aes(label=sprintf("%1.1f", ..y..)),
                  position=position_nudge(x=0.5), size=3.5)
```



- Manual transmissions, generally perform better in terms of miles/gallon, compared to Manual transmissions.

Quantified difference

```
mtcars %>%
    select(mpg, am) %>%
    group_by(am) %>%
    summarise("mu.mpg" = mean(mpg))
```

```
## # A tibble: 2 x 2
##      am mu.mpg
##   <dbl>  <dbl>
## 1     0   17.1
## 2     1   24.4
```

- For Automatic transmissions, the car has a mean of ~17 miles/gallon
- For Manual transmissions, the car has a mean of ~24 miles/gallon

Regression analysis

Linear model for all variables

```
mtcars %>%
    mutate(am = factor(am, levels = c(0,1))) %>%
    group_by(am) %>%
    do(broom::tidy(lm(mpg ~ . -1, data = mtcars))) ## wt and qsec appear stat significant
```

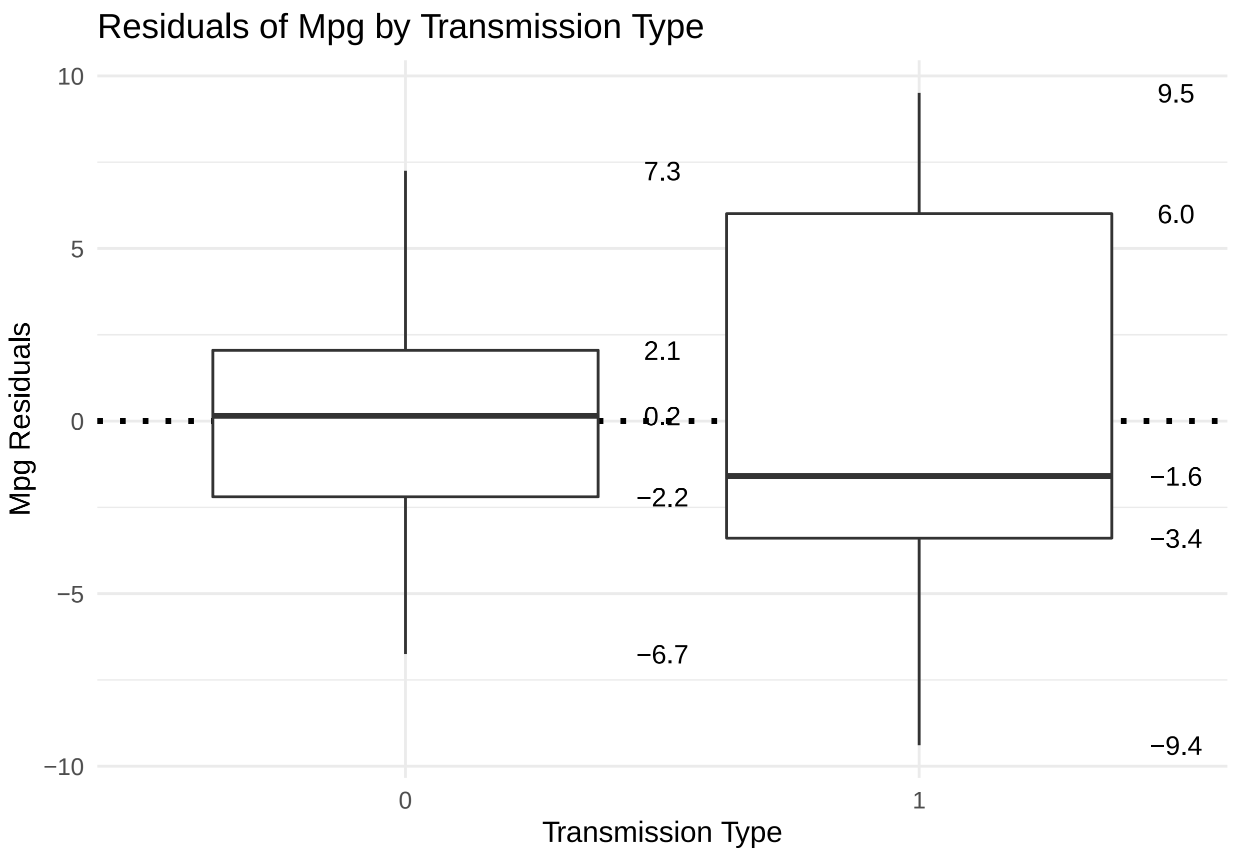
```
## # A tibble: 20 x 6
## # Groups:   am [2]
##    am    term  estimate std.error statistic p.value
##    <fct> <chr>    <dbl>     <dbl>     <dbl>   <dbl>
##  1 0     cyl      0.351     0.763     0.460   0.650
##  2 0     disp     0.0135    0.0176    0.768   0.450
##  3 0     hp      -0.0205    0.0214   -0.958   0.348
##  4 0     drat     1.24      1.46      0.849   0.405
##  5 0     wt      -3.83      1.86     -2.05    0.0520
##  6 0     qsec     1.19      0.459     2.59    0.0166
##  7 0     vs       0.190     2.07      0.0917  0.928
##  8 0     am       2.83      1.98      1.43    0.166
##  9 0     gear     1.05      1.35      0.783   0.442
## 10 0     carb    -0.263     0.812    -0.324   0.749
## 11 1     cyl      0.351     0.763     0.460   0.650
## 12 1     disp     0.0135    0.0176    0.768   0.450
## 13 1     hp      -0.0205    0.0214   -0.958   0.348
## 14 1     drat     1.24      1.46      0.849   0.405
## 15 1     wt      -3.83      1.86     -2.05    0.0520
## 16 1     qsec     1.19      0.459     2.59    0.0166
## 17 1     vs       0.190     2.07      0.0917  0.928
## 18 1     am       2.83      1.98      1.43    0.166
## 19 1     gear     1.05      1.35      0.783   0.442
## 20 1     carb    -0.263     0.812    -0.324   0.749
```

```
## Both variables are equally changing regardless the transmission type.
```

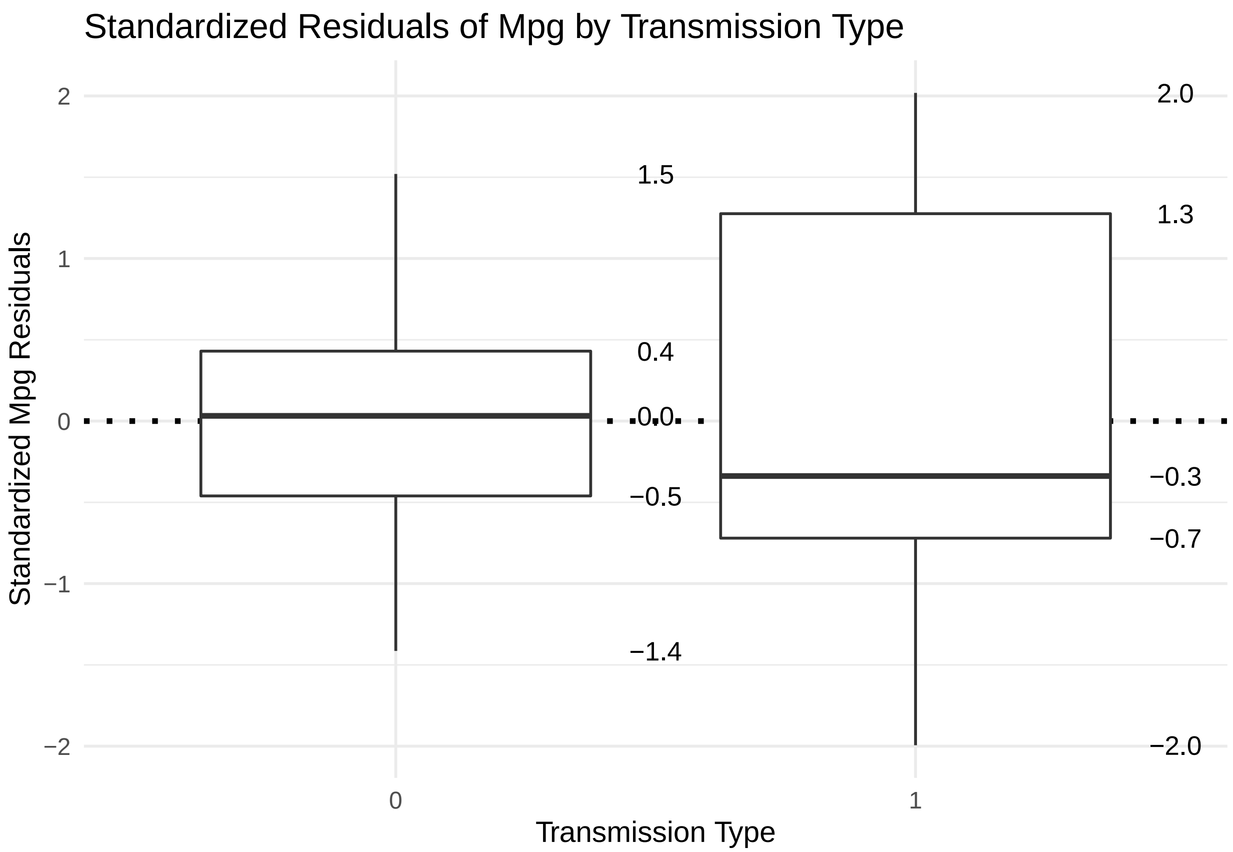
- For each increase of 1000 lbs the car decrease the miles/gallon by 3.83
- For each increase of sec necessary to travel 1/4 mile the miles/gallon increase by 1.19.

Residual analysis

```
mpg.lm = lm(mpg ~ factor(am), data=mtcars)
mpg.res = resid(mpg.lm)
ggplot(data = mtcars, aes(x = factor(am), y = mpg.res)) +
    ylab(label = "Mpg Residuals") + xlab(label = "Transmission Type") +
    ggtitle(label = "Residuals of Mpg by Transmission Type") +
    geom_hline(yintercept = 0, linetype = "dotted", lwd = 1) +
    geom_boxplot() +
    stat_summary(geom="text", fun=quantile,
                 aes(label=sprintf("%1.1f", ..y..)),
                 position=position_nudge(x=0.5), size=3.5) +
    theme_minimal()
```



```r
mpg.stdres = rstandard(mpg.lm)
ggplot(data = mtcars, aes(x = factor(am), y = mpg.stdres)) +
    ylab(label = "Standardized Mpg Residuals") + xlab(label = "Transmission Type") +
    ggtitle(label = "Standardized Residuals of Mpg by Transmission Type") +
    geom_hline(yintercept = 0, linetype = "dotted", lwd = 1) +
    geom_boxplot() +
    stat_summary(geom="text", fun=quantile,
                 aes(label=sprintf("%1.1f", ..y..)),
                 position=position_nudge(x=0.5), size=3.5) +
    theme_minimal()
```



- The standardized residual plot shows that the manual transmission presents a broader range of outliers and its mean value is not aligned with the mean value of the model, which indicate that the model accountability is way less reliable compared to automatic transmissions.

Nested lineam models with ANOVA

```
fit1 <- lm(mpg ~ factor(am) + factor(gear) -1, mtcars)
fit2 <- lm(mpg ~ factor(am) + factor(gear) + factor(carb) + factor(vs) -1, mtcars)
fit3 <- lm(mpg ~ factor(am) + factor(gear) + factor(carb) + factor(vs) + qsec + wt -1, mtcars)
fit4 <- lm(mpg ~ factor(am) + factor(gear) + factor(carb) + factor(vs) + qsec + wt + drat + hp -1, mtca
fit5 <- lm(mpg ~ factor(am) + factor(gear) + factor(carb) + factor(vs) + qsec + wt + drat + hp + disp +

anova(fit1, fit2, fit3, fit4, fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am) + factor(gear) - 1
## Model 2: mpg ~ factor(am) + factor(gear) + factor(carb) + factor(vs) -
##     1
## Model 3: mpg ~ factor(am) + factor(gear) + factor(carb) + factor(vs) +
##     qsec + wt - 1
## Model 4: mpg ~ factor(am) + factor(gear) + factor(carb) + factor(vs) +
##     qsec + wt + drat + hp - 1
## Model 5: mpg ~ factor(am) + factor(gear) + factor(carb) + factor(vs) +
##     qsec + wt + drat + hp + disp + factor(cyl) - 1
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     28 570.00
## 2     22 196.30  6    373.70 7.7594 0.0006273 ***
## 3     20 155.44  2     40.87 2.5456 0.1117185
## 4     18 144.18  2     11.26 0.7014 0.5114585
## 5     15 120.40  3     23.77 0.9873 0.4252633
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## The linear models are not responding very well with inflated factorial variables. However, the conti

fit0 <- lm(mpg ~ factor(am) + wt + qsec -1, mtcars)

as.data.frame(summary(fit0)$coef) %>%
    select(Estimate, `Pr(>|t|)`) %>%
    filter(`Pr(>|t|)` <= 0.05)
```

```
##                Estimate     Pr(>|t|)
## factor(am)1 12.553618 4.754335e-02
## wt          -3.916504 6.952711e-06
## qsec         1.225886 2.161737e-04
```

- Transmission type is impacting the miles/gallon but not more than the weight variable and the 1/4 mile time variable.

Predicted Probability

```
fit1.2 <- glm(am ~ mpg + wt + qsec - 1, mtcars, family = "binomial")
am.predict = data.frame(mpg = mean(mtcars$mpg), wt=mean(mtcars$wt), qsec=mean(mtcars$qsec))
round(predict(fit1.2, am.predict, type="response"),2)
```

```
##   1
## 0.4
```

- Given a car with an average Miles/Gallon, an average Weight and an average 1/4 Miles Time the percentage that a car will have a manual transmission is estimated to be ~ 40%

Generalized linear model

```
fit1.1 <- glm(am ~ mpg, mtcars, family = "binomial")
summary(fit1.1)
```

```
##
## Call:
## glm(formula = am ~ mpg, family = "binomial", data = mtcars)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5701  -0.7531  -0.4245   0.5866   2.0617
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.6035     2.3514  -2.808  0.00498 **
## mpg           0.3070     0.1148   2.673  0.00751 **
## ---
## 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: 29.675  on 30  degrees of freedom
## AIC: 33.675
##
## Number of Fisher Scoring iterations: 5
```

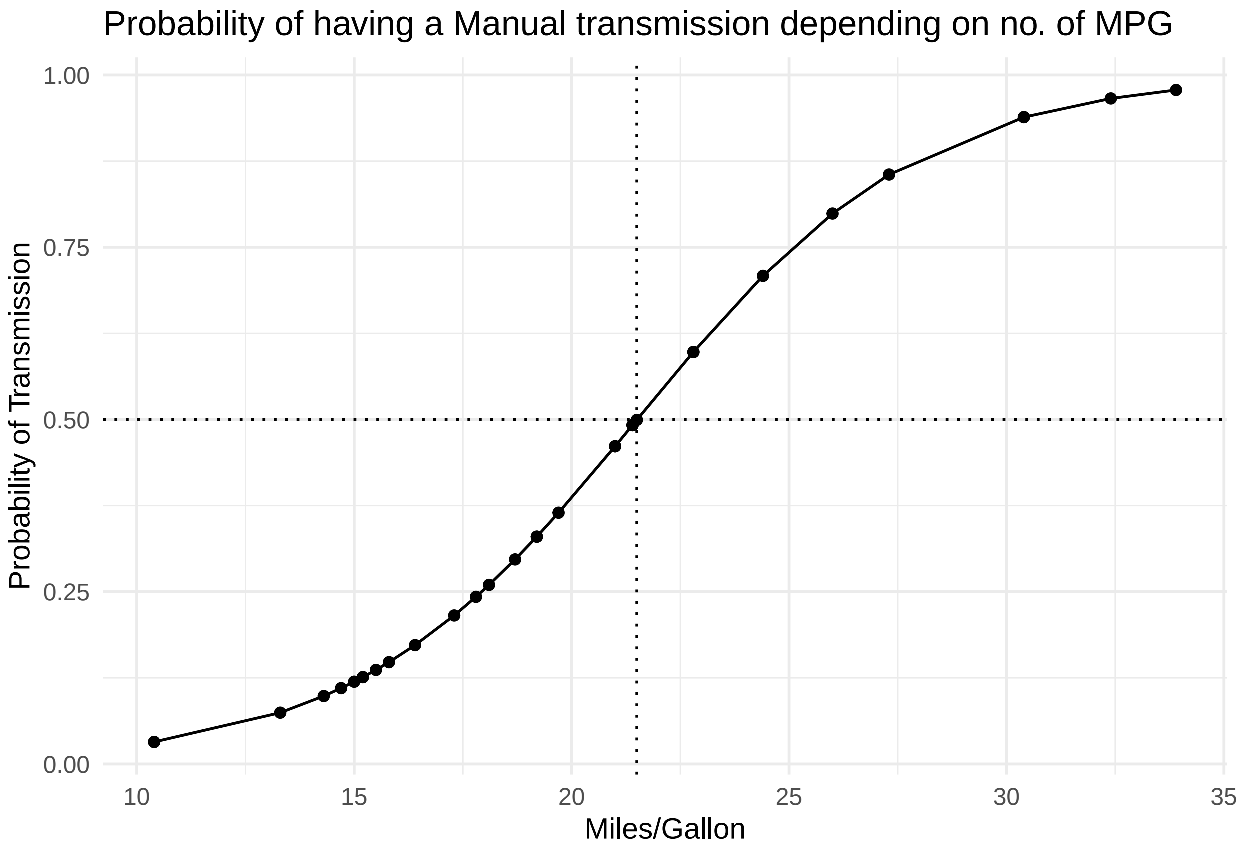
```
exp(fit1.1$coef)
```

```
## (Intercept)         mpg
## 0.001355579 1.359379288
```

```
exp(confint(fit1.1))
```

```
##                    2.5 %     97.5 %
## (Intercept) 4.425443e-06 0.06255158
## mpg         1.129764e+00 1.79946863
```

```
ggplot(data = mtcars, aes(x=mpg, y=fit1.1$fitted.values)) +
    geom_line() +
    geom_point() +
    theme_minimal() +
    xlab(label = "Miles/Gallon") + ylab(label = "Probability of Transmission") +
    ggtitle(label = "Probability of having a Manual transmission depending on no. of MPG") +
    geom_hline(yintercept = 0.5, linetype = "dotted") +
    geom_vline(xintercept = 21.5, linetype = "dotted")
```



- Above 21.5 miles/gallon there is more probability that the car has a manual transmission and below there is more probability that the car is automatic