Statistical Analysis Of The CO2 Emission For The Cars Sold In France In 2015

Francois Laforgia

24 Dec 2015

Summary

This report will explain how to predict the CO2 emissions based on a dataset which include all the cars sold in France during 2015 (up to October 2015). The analysis has been done on two different kind of gas, Diesel and Petrol. It has been decided to remove from the analysis the electric cars (hybrid or full electric) and also the gas type like the LPG for example. Also the study does not take into account the different methods that can be used t odecrease this level of emission like the particle filters or the catalytic converters.
The dataset has been used to generate two predictives models, one for each type of fuel. The models then were used in a shiny application to predict the level of CO2 for any of gas mileage.

Important Information About The Data

The dataset used is local to France and it include only the models sold in France. There could be some difference in countries were other models are sold.
The units used in this study are european units. So the gas mileage is not given in Miles Per Galon, like it is usually the case in the US but in Liter per 100 Kms, abreviated L/kms in the lectures. You need to convert you mileage in this unit before using the shiny app.
The data include the information for the VolksWagen car even if now we know that the data provided by the car maker are biaised because the company cheated on the test results.
This study focused on the CO2 (Carbon Dioxyde). This pollutant is one of the gas that are part of the greenhouse effect.

Information about this report

All the web links regarding the inforamtion provided in this report are given in the appendix.
The code will be shown and executed when possible but the full R code used for the study is also available on github and so is this report.

Data Exploratory

The dataset has been downloaded from the open data French government. This dataset is about the poluants emission for the cars sold in France during the year 2015 up to October 2015. The file comes as zipped a csv file. The csv used a ";" as separator and the encoding is not UTF-8 but latin1.

```
setwd("/Users/flaforgia/Documents/Data products/Polution CO2/Data sets/")
cars.2015 <- read.csv("fic_etiq_edition_40-mars-2015.csv", sep=";", encoding="latin1")
```

The dataset contains 20880 observations of 26 variables. Each observation is a model of cars. A detailed view of the variables shows that not all will be accurrate for this study. I selected only the most relevant and I gave up the adminstrative variable which are not part of this study.
Also the names are in French so for internaltionalization purpose, it has been decided to rename them with a more english explicit name.

```r
colnames(cars.2015)
```

```
##  [1] "lib_mrq_doss"      "lib_mod_doss"      "mrq_utac"
##  [4] "mod_utac"          "dscom"             "cnit"
##  [7] "tvv"               "energ"             "hybride"
## [10] "puiss_admin"       "puiss_max"         "puiss_heure"
## [13] "typ_boite_nb_rapp" "conso_urb_93"      "conso_exurb"
## [16] "conso_mixte"       "co2_mixte"         "co_typ_1"
## [19] "hc"                "nox"               "hcnox"
## [22] "ptcl"              "masse_ordma_min"   "masse_ordma_max"
## [25] "champ_v9"          "date_maj"
```

```r
select.2015 <- c("lib_mrq_doss", "lib_mod_doss", "energ", "hybride", "puiss_max", "typ_boite_nb_rapp",
                 "conso_urb_93", "conso_exurb" , "conso_mixte", "co2_mixte", "co_typ_1",
                 "hc", "nox", "hcnox", "ptcl", "masse_ordma_min", "masse_ordma_max")
cars.2015 <- cars.2015[,select.2015]
names.2015 <- c("maker", "model", "gas", "hybrid", "max_hp", "gear", "city_gas_mileage", "hghy_gas_mile
                "co2", "co_typ_1", "hc", "nox", "hcnox", "ptcl", "min_wght", "max_wght")

names(cars.2015) <- names.2015
```

Some type of fuel are too marginal to influence greatly the result and it has been taken the decision to give them up. Moreover the Diesel (GO) and the Petrol (ES) cars represent 93.91 % of the whole dataset.
For the gas mileage, we will focus on the variable names mix_gas_mileage. This variable represent the gas mileage for a car that is used both to drive in a city (city_gas_mileage) and also on an highway (hghy___gas_mileage) and in general it is the main usage of the cars. The dataset was split in two parts, one for each type of fuel and the N/A values were removed from the cols co2 for each sub-datasets.

```r
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
```

```r
cars.ES <- filter(cars.2015, gas=="ES ")
cars.GO <- filter(cars.2015, gas=="GO ")

cars.GO <- cars.GO[complete.cases(cars.GO$co2),]
cars.ES <- cars.ES[complete.cases(cars.ES$co2),]
```

Data Analysis

The pollutant we want to check is the co2. This pollutant is strongly connected to the gas mileage. To confirm that a correlation analysis is done of the co2 against all the other variables. To do that it has been decided to remove all the non-numeric variables which have no influence on the study.

```
cars.GO <- select(cars.GO,
                  mix_gas_mileage,
                  max_hp, co2,
                  co_typ_1, nox,
                  hcnox, ptcl,
                  min_wght,
                  max_wght,
                  hghy_gas_mileage,
                  city_gas_mileage)

cars.ES <- select(cars.ES,
                  mix_gas_mileage,
                  max_hp, co2,
                  co_typ_1, nox,
                  hcnox, ptcl,
                  min_wght,
                  max_wght,
                  hghy_gas_mileage,
                  city_gas_mileage)

cor(cars.GO$co2, cars.GO[,])
```

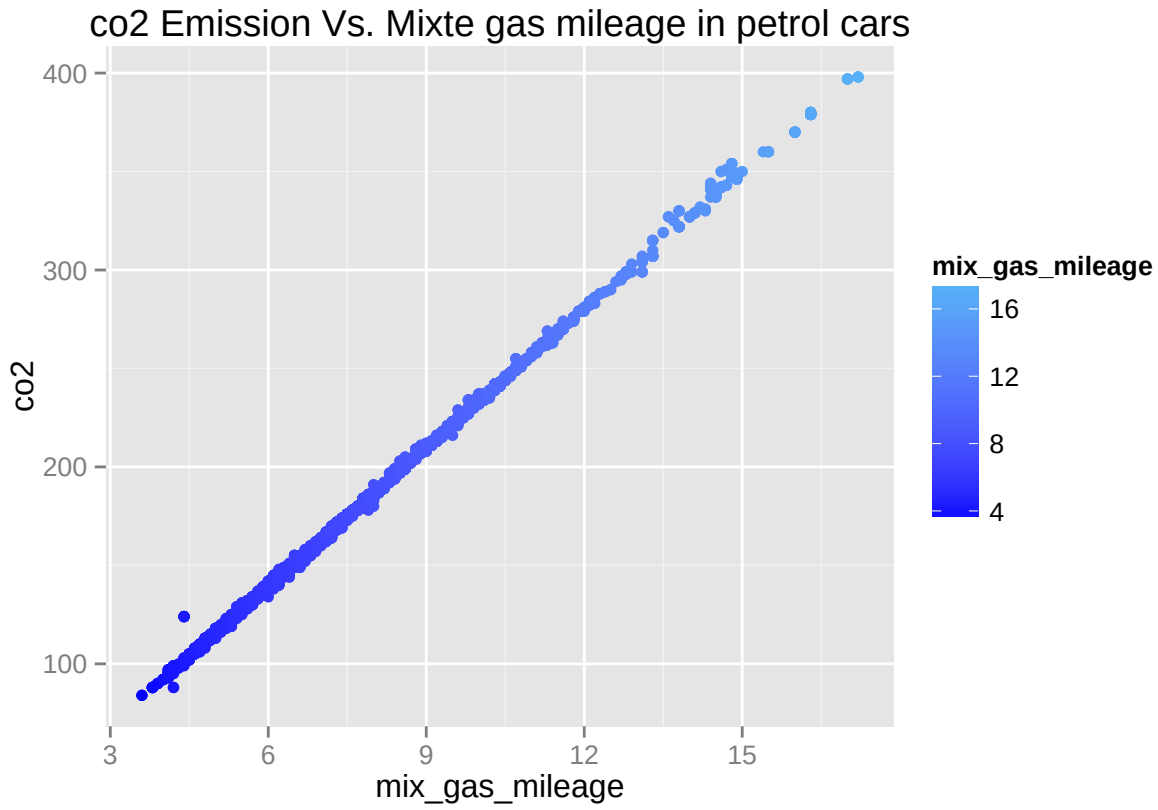
```
##      mix_gas_mileage    max_hp co2 co_typ_1 nox hcnox ptcl  min_wght
## [1,]       0.9993103 0.1867888   1       NA  NA    NA   NA 0.8974891
##       max_wght hghy_gas_mileage city_gas_mileage
## [1,] 0.8986818        0.9859308        0.9755979
```

```
cor(cars.ES$co2, cars.ES[,])
```

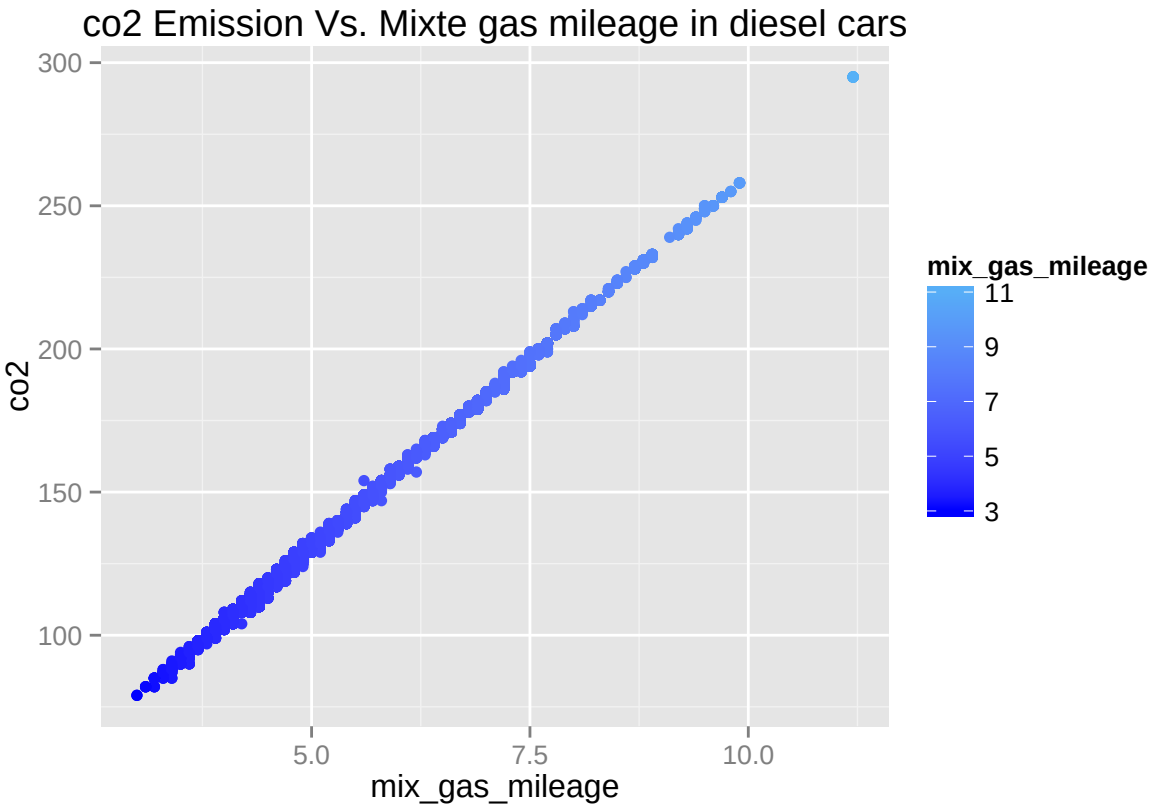
```
##      mix_gas_mileage    max_hp co2 co_typ_1 nox hcnox ptcl  min_wght
## [1,]       0.9996904 0.8651351   1       NA  NA    NA   NA 0.7880659
##       max_wght hghy_gas_mileage city_gas_mileage
## [1,] 0.7915338          0.98844        0.9932782
```

As it is shown by the correlations, the co2 varaiable is strongly linked to the gas mileage, almost 1 which is the prefect correlation. This can be proved by plotting the co2 values against the mix_gas_mileage values.

```
library(ggplot2)
co2.ES <- ggplot(cars.ES, aes(x=mix_gas_mileage, y=co2))
co2.ES <- co2.ES + geom_point(aes(, colour=mix_gas_mileage)) + scale_colour_gradient(low="blue")
co2.ES <- co2.ES + ggtitle("co2 Emission Vs. Mixte gas mileage in petrol cars")
co2.ES
```




```
co2.GO <- ggplot(cars.GO, aes(x=mix_gas_mileage, y=co2))
co2.GO <- co2.GO + geom_point(aes(, colour=mix_gas_mileage)) + scale_colour_gradient(low="blue")
co2.GO <- co2.GO + ggtitle("co2 Emission Vs. Mixte gas mileage in diesel cars")
co2.GO
```



Based on the shape of the plot, this almost perfect plot can also be the sign of heteroskedasticity. To confirm or not that, we need to perform a residual analysis. To do that we will plot the residual given by the model against the X axis.
As starting poimnt we fit a linear model for both fuel type (co2~mix_gas_mileage).

```
modelFit.co2.ES <- lm(co2~mix_gas_mileage, data=cars.ES)
summary(modelFit.co2.ES)
```

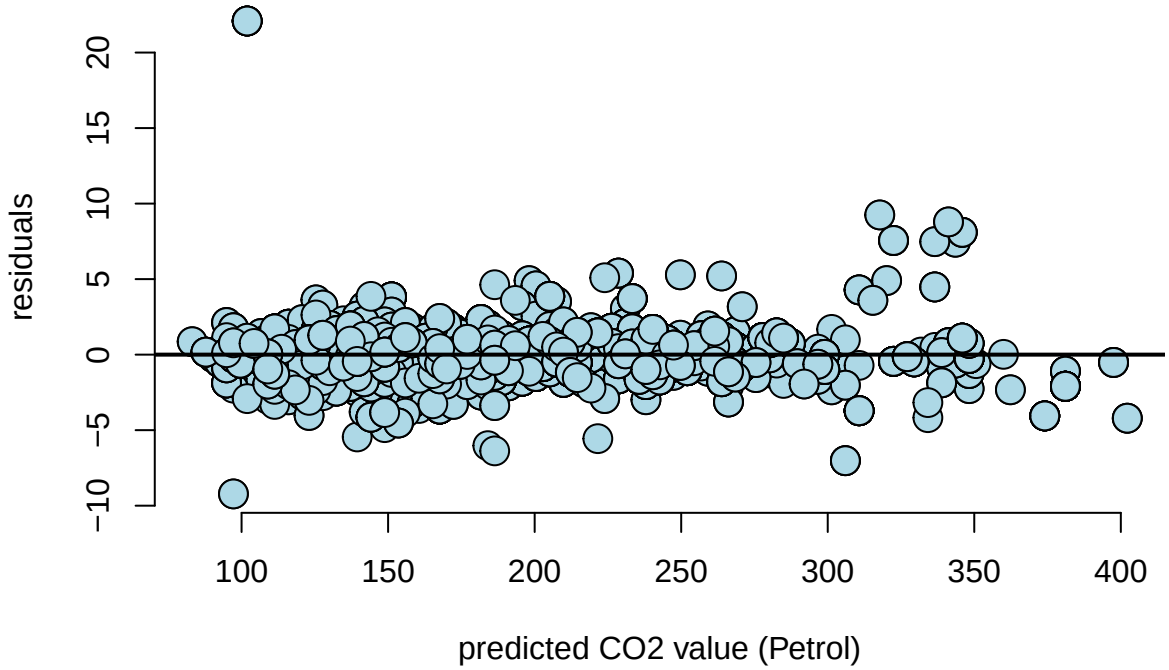
```
##
## Call:
## lm(formula = co2 ~ mix_gas_mileage, data = cars.ES)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2203 -0.7149 -0.0270  0.6651 22.0876
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.314721   0.053992  -24.35   <2e-16 ***
## mix_gas_mileage 23.460710   0.007014 3344.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.349 on 6932 degrees of freedom
## Multiple R-squared:  0.9994, Adjusted R-squared:  0.9994
## F-statistic: 1.119e+07 on 1 and 6932 DF,  p-value: < 2.2e-16
```

```r
modelFit.co2.GO <- lm(co2~mix_gas_mileage, data=cars.GO)
summary(modelFit.co2.GO)
```

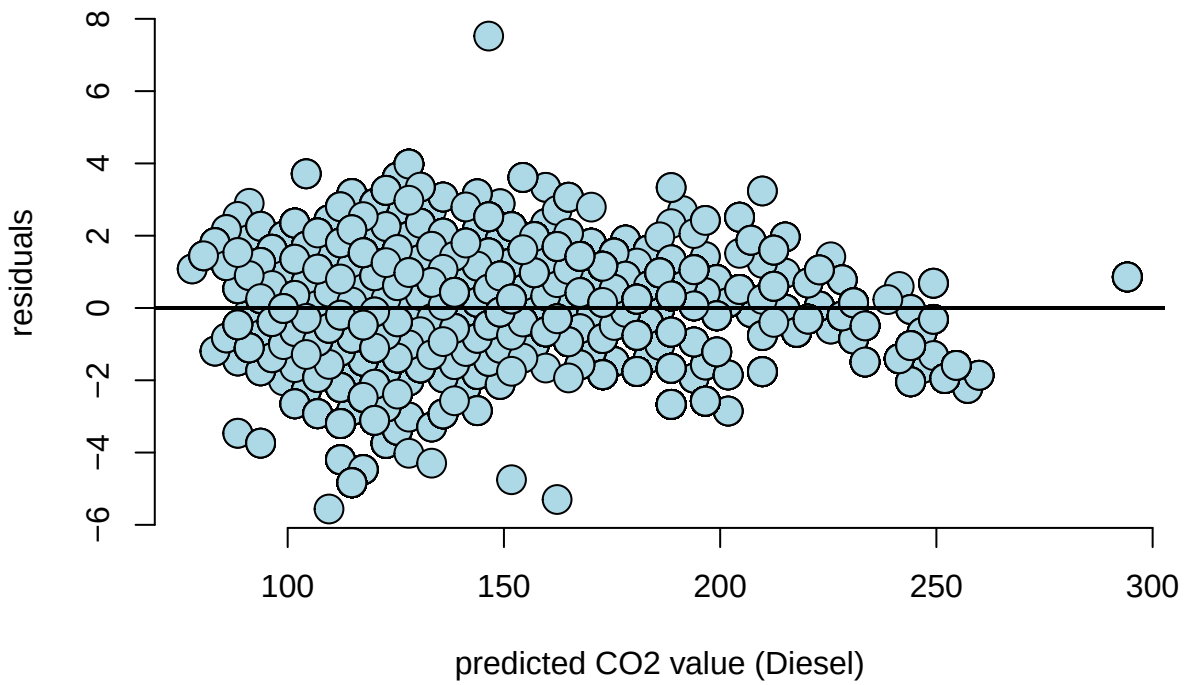
```
##
## Call:
## lm(formula = co2 ~ mix_gas_mileage, data = cars.GO)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5613 -0.7457  0.1539  0.8580  7.5225
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.187056   0.050666  -23.43   <2e-16 ***
## mix_gas_mileage 26.368667   0.008704 3029.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.276 on 12673 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9986
## F-statistic: 9.178e+06 on 1 and 12673 DF,  p-value: < 2.2e-16
```

The summary of these model show a p-value very low for both model. This is the sign that the model is good and that we won't fail to predict correct values.
We than plotted the residuals.

```r
cars.ES.resid <- resid(modelFit.co2.ES)
cars.ES.yhat <- predict(modelFit.co2.ES)
cars.ES.y <- cars.ES$co2
plot(cars.ES.yhat, cars.ES.resid, xlab = "predicted CO2 value (Petrol)", ylab = "residuals", bg = "ligh
abline(h = 0, lwd = 2)
```



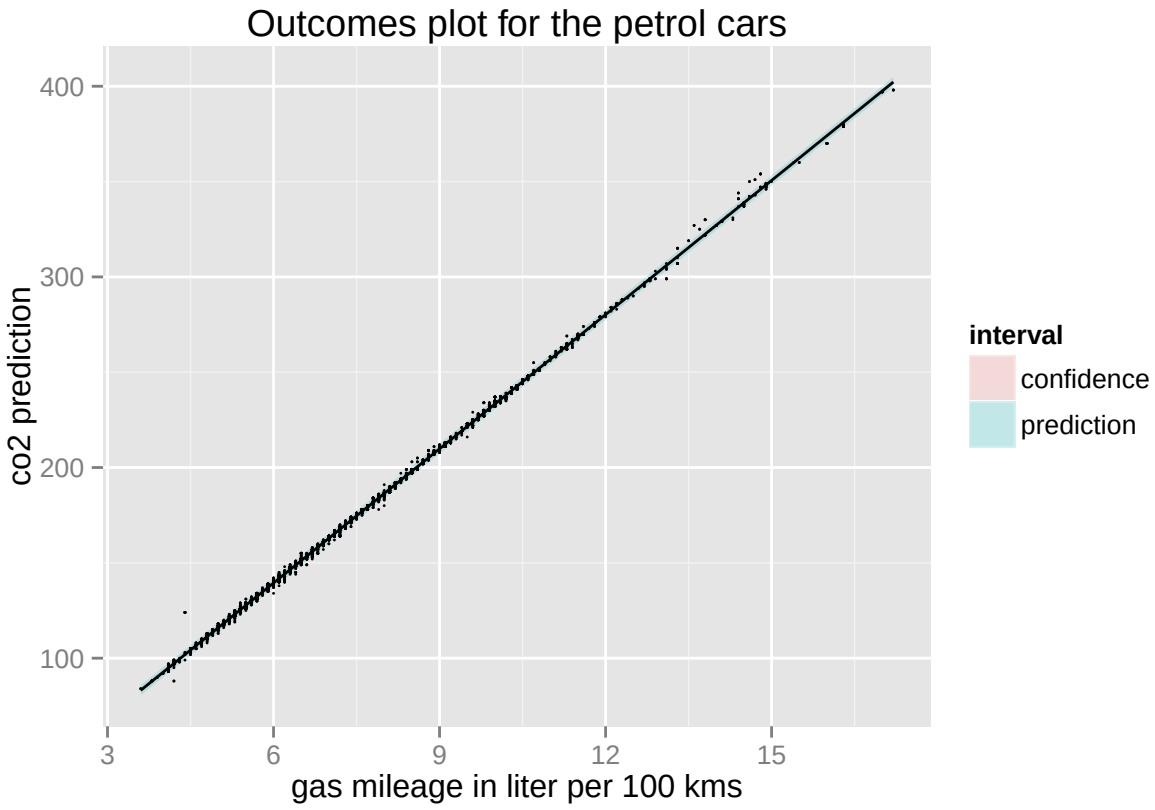
```
cars.GO.resid <- resid(modelFit.co2.GO)
cars.GO.yhat <- predict(modelFit.co2.GO)
cars.GO.y <- cars.GO$co2
plot(cars.GO.yhat, cars.GO.resid, xlab = "predicted CO2 value (Diesel)", ylab = "residuals", bg = "ligh
abline(h = 0, lwd = 2)
```



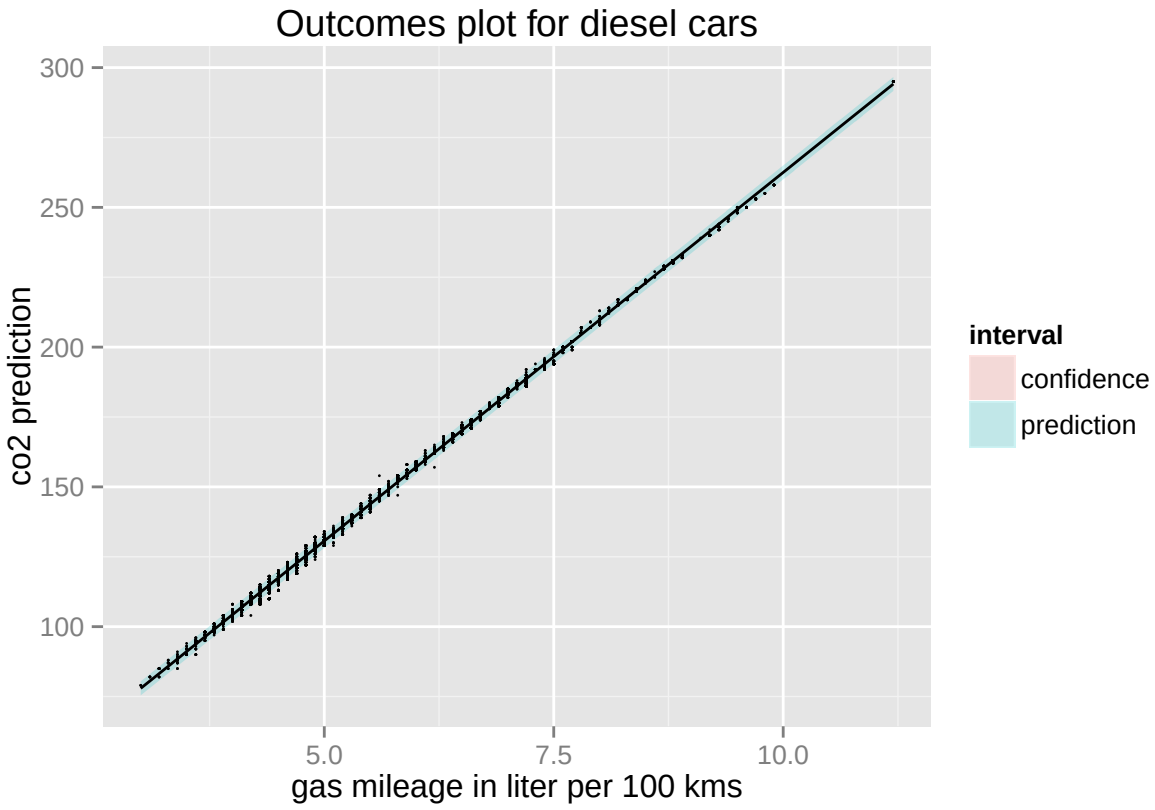
With these plots, we confirm there is not so much dispersion of the residuals and they stay grouped around the 0 line. This means we can use this model to predict the CO2 emission value for any gas mileage with the caveat that the gas mileage represent the mist gas mileage. We can now plot the prediction interval of the outcomes for each model to confirm that they are accurrate.

```
newset.ES <- data.frame(mix_gas_mileage= seq(min(cars.ES$mix_gas_mileage), max(cars.ES$mix_gas_mileage)
newset.GO <- data.frame(mix_gas_mileage= seq(min(cars.GO$mix_gas_mileage), max(cars.GO$mix_gas_mileage)

predict1.co2.ES <- data.frame(predict(modelFit.co2.ES, newset.ES, interval="confidence"))
predict2.co2.ES <- data.frame(predict(modelFit.co2.ES, newset.ES, interval="prediction"))
predict1.co2.ES$interval <- "confidence"
predict2.co2.ES$interval <- "prediction"
predict1.co2.ES$mix_gas_mileage <- newset.ES$mix_gas_mileage
predict2.co2.ES$mix_gas_mileage <- newset.ES$mix_gas_mileage
data <- rbind(predict1.co2.ES, predict2.co2.ES)
names(data)[1] <- "y"
plot.co2.ES <- ggplot(data, aes(x = mix_gas_mileage, y = y))
plot.co2.ES <- plot.co2.ES + xlab("gas mileage in liter per 100 kms")
plot.co2.ES <- plot.co2.ES + ylab("co2 prediction")
plot.co2.ES <- plot.co2.ES + geom_ribbon(aes(ymin=lwr, ymax=upr, fill=interval), alpha=0.2)
plot.co2.ES <- plot.co2.ES + geom_line()
plot.co2.ES <- plot.co2.ES + geom_point(data=data.frame(x=cars.ES$mix_gas_mileage, y=cars.ES$co2), aes(
plot.co2.ES <- plot.co2.ES + ggtitle("Outcomes plot for the petrol cars")
plot.co2.ES
```




```
predict1.co2.GO <- data.frame(predict(modelFit.co2.GO, newset.GO, interval="confidence"))
predict2.co2.GO <- data.frame(predict(modelFit.co2.GO, newset.GO, interval="prediction"))
predict1.co2.GO$interval <- "confidence"
predict2.co2.GO$interval <- "prediction"
predict1.co2.GO$mix_gas_mileage <- newset.GO$mix_gas_mileage
predict2.co2.GO$mix_gas_mileage <- newset.GO$mix_gas_mileage
data <- rbind(predict1.co2.GO, predict2.co2.GO)
names(data)[1] <- "y"
plot.co2.GO <- ggplot(data, aes(x = mix_gas_mileage, y = y))
plot.co2.GO <- plot.co2.GO + xlab("gas mileage in liter per 100 kms")
plot.co2.GO <- plot.co2.GO + ylab("co2 prediction")
plot.co2.GO <- plot.co2.GO + geom_ribbon(aes(ymin=lwr, ymax=upr, fill=interval), alpha=0.2)
plot.co2.GO <- plot.co2.GO + geom_line()
plot.co2.GO <- plot.co2.GO + geom_point(data=data.frame(x=cars.GO$mix_gas_mileage, y=cars.GO$co2), aes(
plot.co2.GO <- plot.co2.GO + ggtitle("Outcomes plot for diesel cars")
plot.co2.GO
```



And also the confidence interval for each model.

```
sumCoef.ES <- summary(modelFit.co2.ES)$coefficients
sumCoef.GO <- summary(modelFit.co2.GO)$coefficients
paste("The slope for petrol cars: ", (sumCoef.ES[2,1] + c(-1, 1) * qt(0.975, df = modelFit.co2.ES$df) *
```

```
## [1] "The slope for petrol cars:  1.1315418898399"
## [2] "The slope for petrol cars:  1.13404173083361"
```

```
#(sumCoef.ES[2,1] + c(-1, 1) * qt(0.975, df = modelFit.co2.ES$df) * sumCoef.ES[2,2])/length(cars.ES)-1
paste("The slope for petrol cars: ", (sumCoef.GO[2,1] + c(-1, 1) * qt(0.975, df = modelFit.co2.GO$df) *
```

```
## [1] "The slope for petrol cars:  1.39560051114346"
## [2] "The slope for petrol cars:  1.39870256361612"
```

Conclusion

As define by the chemistry, the CO2 emission is higly dependent of the gas mileage. We have used this strong and almost perfect correlation to build a predictive model and withis model we can predict the CO2 emission for any gas mileage provided w/ the caveat this mileage must be provided in the european unit. The CO2 correlation to the gas mileage can be explained by the way of the CO2 is generated in the engine. The CO2 is the result of a combustion between the oxygen and the carbone contained in the fuel. So the volume of CO2 depends partly on the volume of fuel injected in the combustion chamber. This volume is reported as the comsuption of fuel by the car makers, so it is the gas mileage. The second part of the equation is the volume of oxygen, which is represented by the air injected in the combustion chamber. This volume is more or less constant between the different model of cars and so it does not impact the CO2 emission prediction as the fuel volume.