ENV 710

linear models

Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT

roadmap

- questions?
- where we are going
- pod work!

where we are

linear models

↓

lm with continuous & categorical IV's

⇣

model selection/reduction

I – epa gas standards

Do the type of fuel, drive systems, number of cylinders or CO2 emissions explain the city gas mileage of 2020 vehicles?

1. download data from Sakai:
   `epa-gas.csv`
2. explore data
3. build the model
4. validate the model assumptions with residuals
5. interpret model fit and parameters
6. conclusion?

https://www.fueleconomy.gov/feg/download.shtml

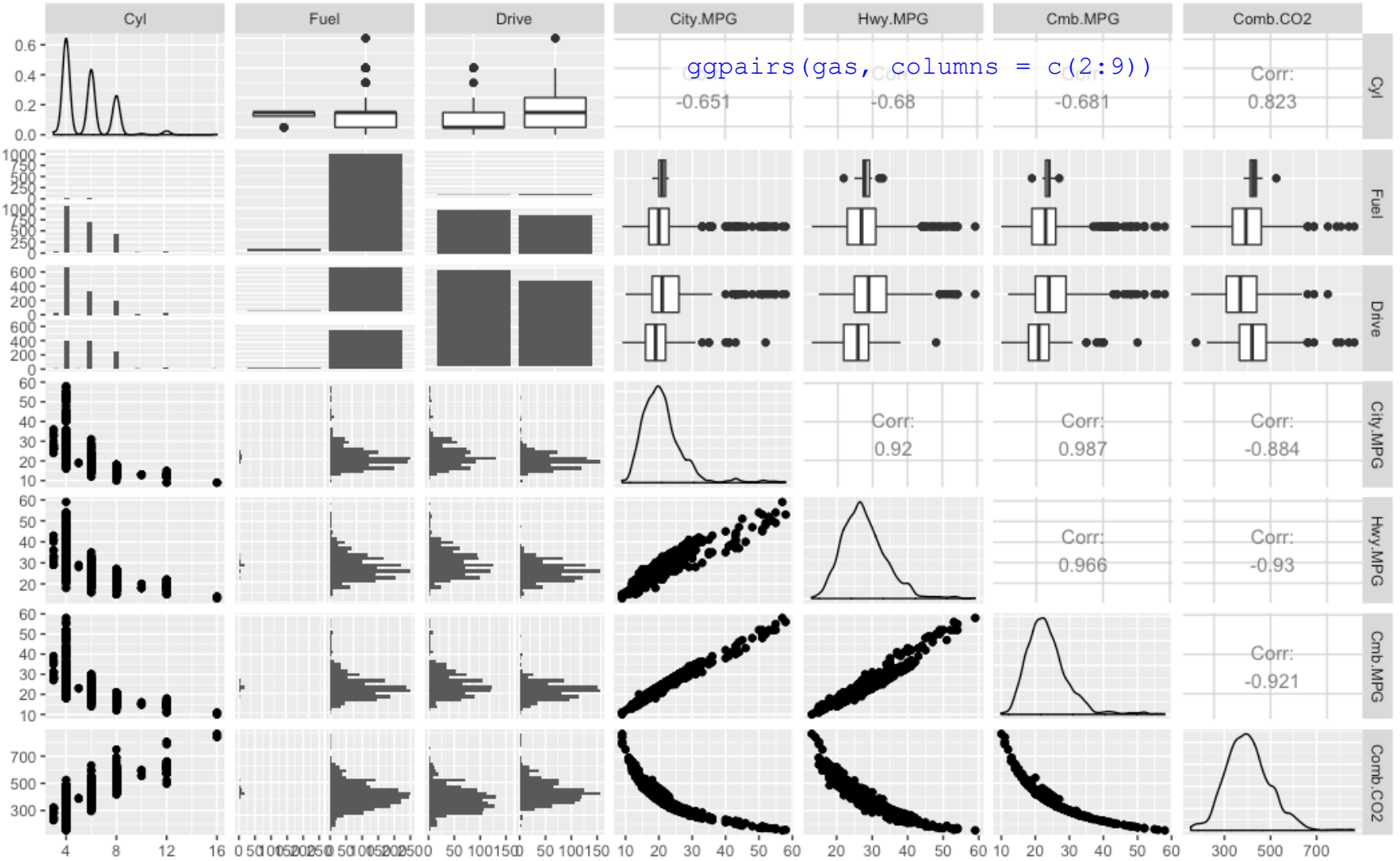
I – epa gas standards

Do the type of fuel, drive systems, number of cylinders or CO2 emissions explain the city gas mileage of 2020 vehicles?

- patterns in the data?
- data types?

```
pacman::p_load(ggplot2, car, GGally)

gas <- read.csv("epa-gas.csv", header = T, stringsAsFactors = T)
 ggpairs(gas, columns = c(2:9))
```

```
ggpairs(gas, columns = c(2:9))
```



1 – epa gas standards

Run a model on city gas mileage with 4 explanatory variables: fuel (diesel or gasoline), drive system (2WD or 4WD), cylinder, and CO2 emissions. Treat fuel and drive as categorical variables and cylinder and CO2 emissions as continuous variables.

1. what does `Intercept` represent?
2. what does `factor(Fuel)Gasoline` mean?
3. what does `factor(Drive)4WD` mean?
4. what do `Cyl` and `Comb.CO2` represent?
5. does the data meet model assumptions?
6. conclusions

I – epa gas standards

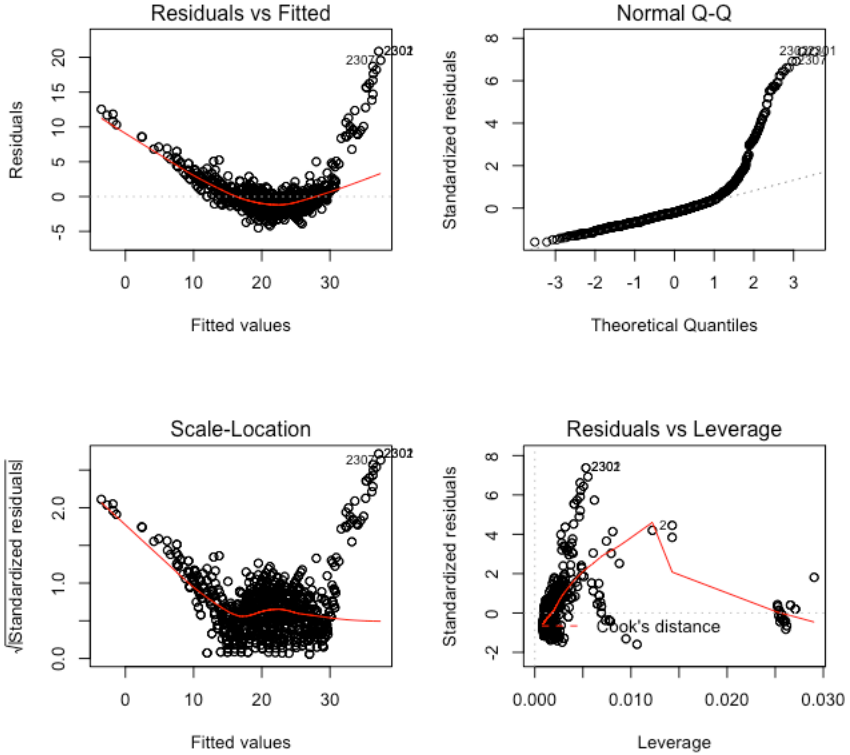
```
lm1 <- lm(City.MPG ~ factor(Fuel) +
         factor(Drive) + Cyl + Comb.CO2,
          data = gas)
 summary(lm1)
```

```
Call:
lm(formula = City.MPG ~ factor(Fuel) + factor(Drive) + Cyl +
    Comb.CO2, data = gas)

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          47.272611   0.524828  90.073  < 2e-16 ***
factor(Fuel)Gasoline -2.118673   0.453543  -4.671 3.16e-06 ***
factor(Drive)4WD     -0.070894   0.123834  -0.572    0.567
Cyl                   0.825006   0.056689  14.553  < 2e-16 ***
Comb.CO2             -0.071545   0.001119 -63.923  < 2e-16 ***
---

Residual standard error: 2.835 on 2303 degrees of freedom
Multiple R-squared:  0.8019,        Adjusted R-squared:  0.8015
F-statistic:  2330 on 4 and 2303 DF,  p-value: < 2.2e-16
```

I – epa gas standards

Do the type of fuel, drive system, number of cylinders or CO2 emissions explain the city gas mileage of 2020 vehicles?

```
lm1 <- lm(City.MPG ~ factor(Fuel) +
        factor(Drive) + Cyl + Comb.CO2,
         data = my.gas)
 summary(lm1)
  par(mfrow = c(2,2)); plot(lm1)
```



I – epa gas standards

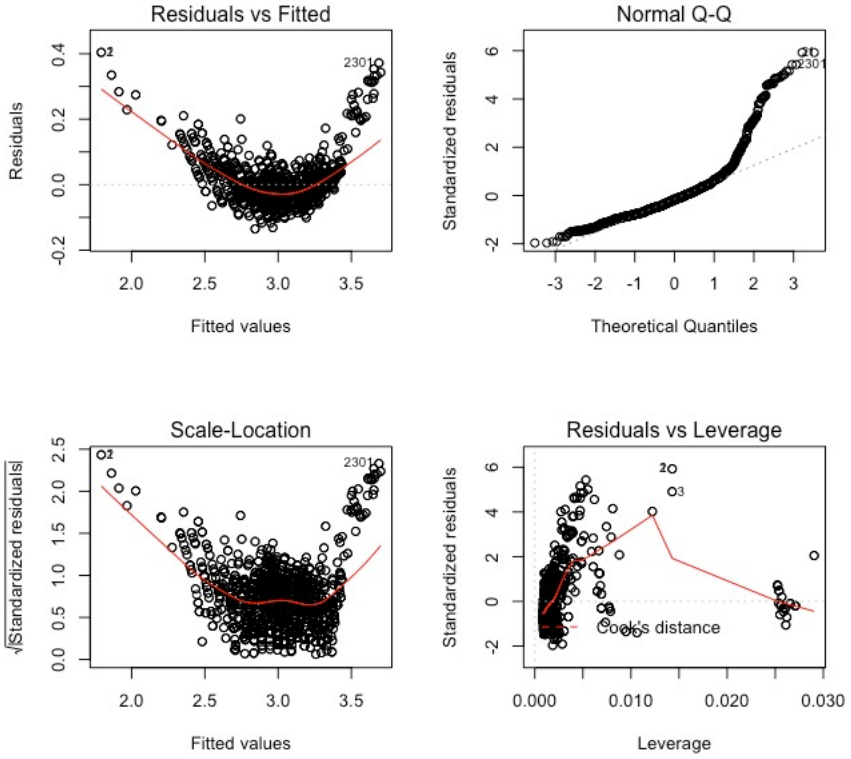
Run a model on city gas mileage with 4 explanatory variables: fuel (diesel or gasoline), drive system (2WD or 4WD), cylinder, and CO2 emissions. Treat fuel and drive as categorical variables and cylinder and CO2 emissions as continuous variables.

what next?

I – epa gas standards

log-transform `City.MPG`

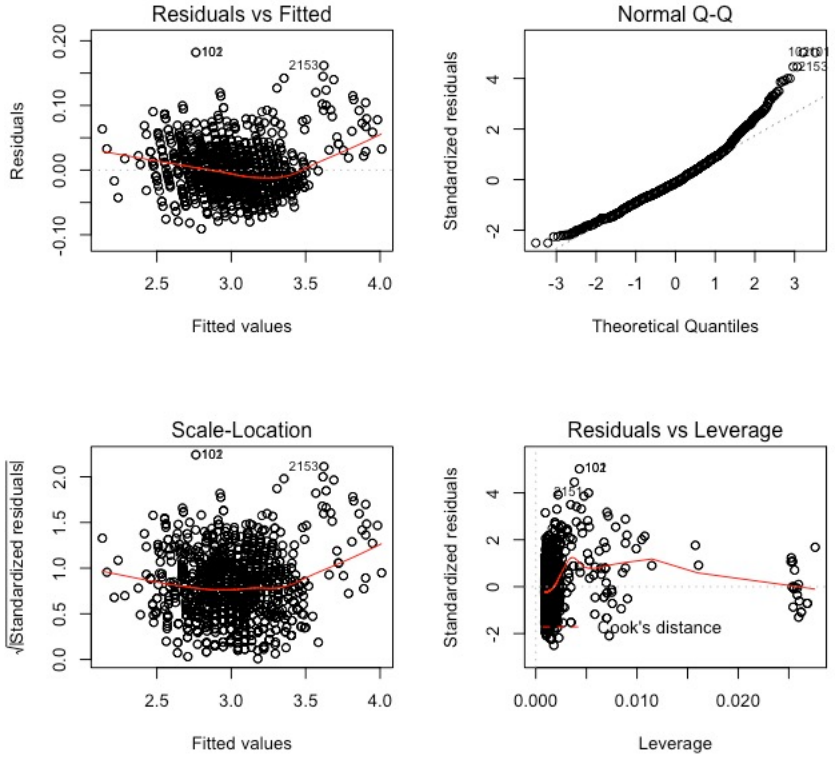
```
lm2 <- lm(log(City.MPG) ~ factor(Fuel) +
         factor(Drive) + Cyl + Comb.CO2,
          data = gas)
summary(lm2)
 par(mfrow = c(2,2)); plot(lm2)
```



I – epa gas standards

log-transform `Comb.CO2`

```
gas$Comb.CO2l <- with(gas, log(Comb.CO2))

lm3 <- lm(log(City.MPG) ~ factor(Fuel) +
        factor(Drive) + Cyl + Comb.CO2l,
         data = gas)
 summary(lm3)
  par(mfrow = c(2,2)); plot(lm3)
```



I – epa gas standards

log-transform `Comb.CO2`

```
summary(lm3)

Call: lm(formula = log(City.MPG) ~ factor(Fuel) +
      factor(Drive) + Cyl + Comb.CO2l, data = gas)

Residuals:
     Min       1Q   Median       3Q      Max
-0.09086 -0.02350 -0.00470  0.02019  0.18205

Coefficients:
                       Estimate Std. Error  t value Pr(>|t|)
(Intercept)            9.636053   0.028231  341.324   <2e-16 ***
factor(Fuel)Gasoline -0.140490   0.005816  -24.156   <2e-16 ***
factor(Drive)4WD       0.013445   0.001593    8.441   <2e-16 ***
Cyl                   -0.001159   0.000647   -1.791   0.0735 .
Comb.CO2l             -1.088030   0.005044 -215.712   <2e-16 ***
---

Residual standard error: 0.03633 on 2303 degrees of freedom
Multiple R-squared:  0.9818,        Adjusted R-squared:  0.9817
F-statistic: 3.098e+04 on 4 and 2303 DF,  p-value: < 2.2e-16
```

I – epa gas standards

remove `Cyl`

```
lm4 <- lm(log(City.MPG) ~ factor(Fuel) +
       factor(Drive) + Comb.CO2l,
       data = my.gas)
```

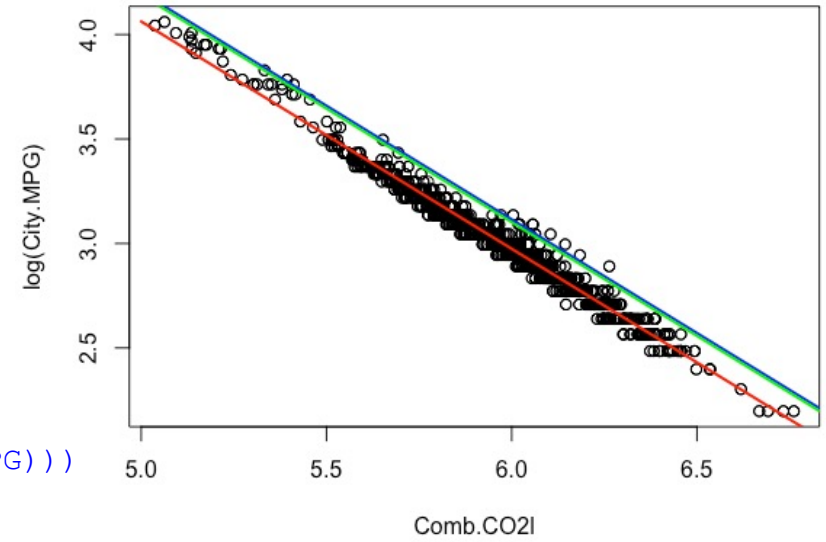
```
Call:
lm(formula = log(City.MPG) ~ factor(Fuel) + factor(Drive) + Comb.CO2l,
    data = gas)

Residuals:
      Min        1Q    Median        3Q       Max
-0.092280 -0.023685 -0.004398  0.020037  0.185520

Coefficients:
                     Estimate Std. Error  t value Pr(>|t|)
(Intercept)           9.671551   0.020111  480.913   <2e-16 ***
factor(Fuel)Gasoline -0.141140   0.005807  -24.303   <2e-16 ***
factor(Drive)4WD      0.013766   0.001583    8.693   <2e-16 ***
Comb.CO2l            -1.094976   0.003226 -339.446   <2e-16 ***
---

Residual standard error: 0.03634 on 2304 degrees of freedom
Multiple R-squared:  0.9817,        Adjusted R-squared:  0.9817
F-statistic: 4.127e+04 on 3 and 2304 DF,  p-value: < 2.2e-16
```

```
summary(lm3)$adj.r.squared
[1] 0.9817256
summary(lm4)$adj.r.squared
[1] 0.9817081
```

```
AIC(lm3, lm4)
    df       AIC
lm3  6 -8746.156
lm4  5 -8744.945
```

I – epa gas standards

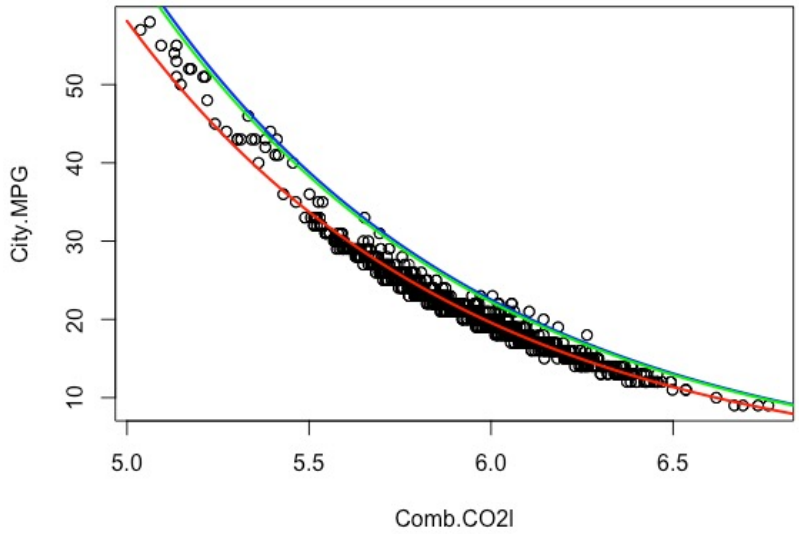
Do the type of fuel, drive, number of cylinders or CO2 emissions explain the city gas mileage of 2020 vehicles?




```
cf <- coef(lm3)
with(my.gas, plot(x = Comb.CO2l, y = log(City.MPG)))

 # curve for gasoline and 4wd
 curve(cf[1] + cf[2] + cf[3] + cf[4]*mean(my.gas$Cyl) + cf[5]*x, 5,7,
       add = T, lwd = 2, col = "red")

 # curve for electric and 4wd
  curve(cf[1] + cf[3] + cf[4]*mean(my.gas$Cyl) + cf[5]*x, 5,7,
        add = T, lwd = 2, col = "blue")

 # curve for electric and 2wd
  curve(cf[1] + cf[4]*mean(my.gas$Cyl) + cf[5]*x, 5,7,
        add = T, lwd = 2, col = "green")
```

I – epa gas standards

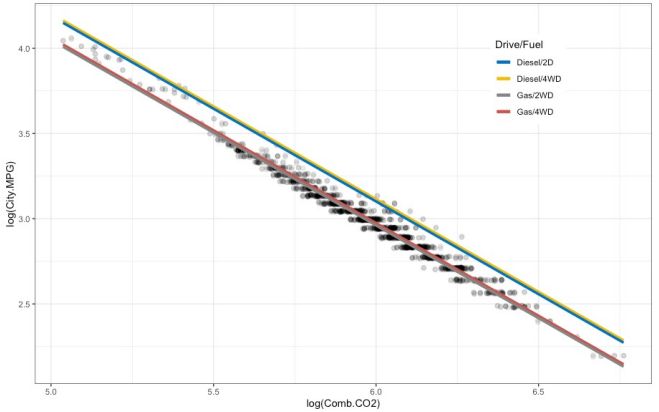
Do the type of fuel, drive, number of cylinders or CO2 emissions explain the city gas mileage of 2020 vehicles?



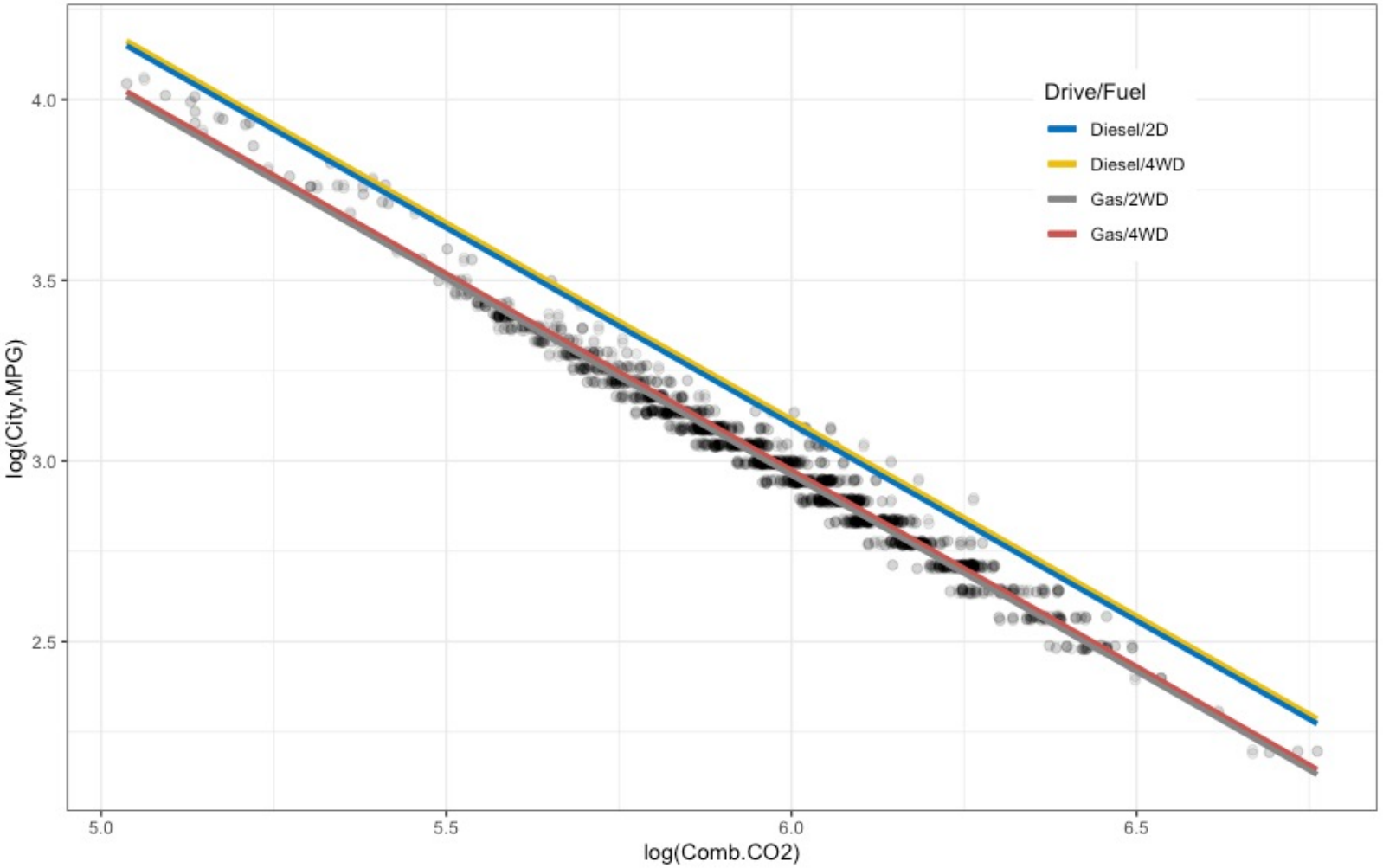

```
with(my.gas, plot(x = Comb.CO2l, y = City.MPG))

 # curve for gasoline and 4wd
 curve(exp(cf[1] + cf[2] + cf[3] + cf[4]*mean(my.gas$Cyl) + cf[5]*x), 5,7,
       add = T, lwd = 2, col = "red")

# curve for electric and 4wd
  curve(exp(cf[1] + cf[3] + cf[4]*mean(my.gas$Cyl) + cf[5]*x), 5,7,
       add = T, lwd = 2, col = "blue")

 # curve for electric and 2wd
  curve(exp(cf[1] + cf[4]*mean(my.gas$Cyl) + cf[5]*x), 5,7,
       add = T, lwd = 2, col = "green")
```

I – epa gas standards

Do the type of fuel, drive, number of cylinders or CO2 emissions explain the city gas mileage of 2020 vehicles?

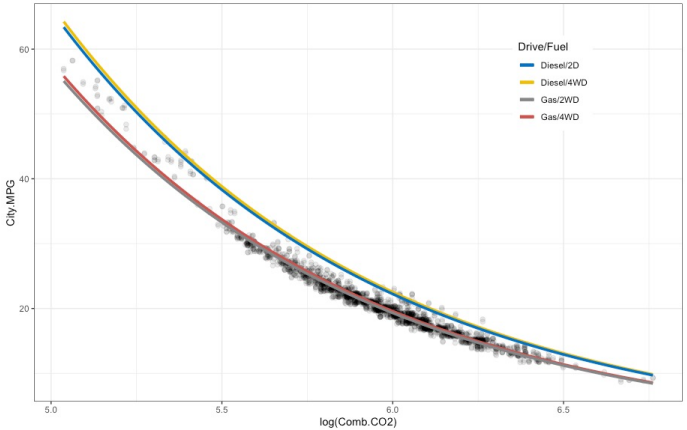



```
cfs <- coef(lm3)
eq1=function(x){cfs[1] + cfs[2] + cfs[3] + cfs[4]*mean(gas$Cyl) + cfs[5]*x}
eq2 =function(x){cfs[1] + cfs[3] + cfs[4]*mean(gas$Cyl) + cfs[5]*x}
eq3 =function(x){cfs[1] + cfs[2] + cfs[4]*mean(gas$Cyl) + cfs[5]*x}
eq4 =function(x){cfs[1]+ cfs[4]*mean(gas$Cyl) + cfs[5]*x}

my_jco <- c("#0073C2FF", "#EFC000FF", "#868686FF", "#CD534CFF", "#7AA6DCFF",
"#003C67FF", "#8F7700FF", "#3B3B3BFF", "#A73030FF", "#4A6990FF")

ggplot(data = gas, aes(x = log(Comb.CO2), y = log(City.MPG))) +
       geom_jitter(shape = 19, size = 2, fill = my_jco[3], alpha = 0.1) +
        stat_function(fun = eq1, geom="line", colour = my_jco[1]) +
        stat_function(fun = eq2, geom="line", colour = my_jco[2]) +
        stat_function(fun = eq3, geom="line", colour = my_jco[3]) +
        stat_function(fun = eq4, geom="line", colour = my_jco[4]) +
      theme_bw()
```

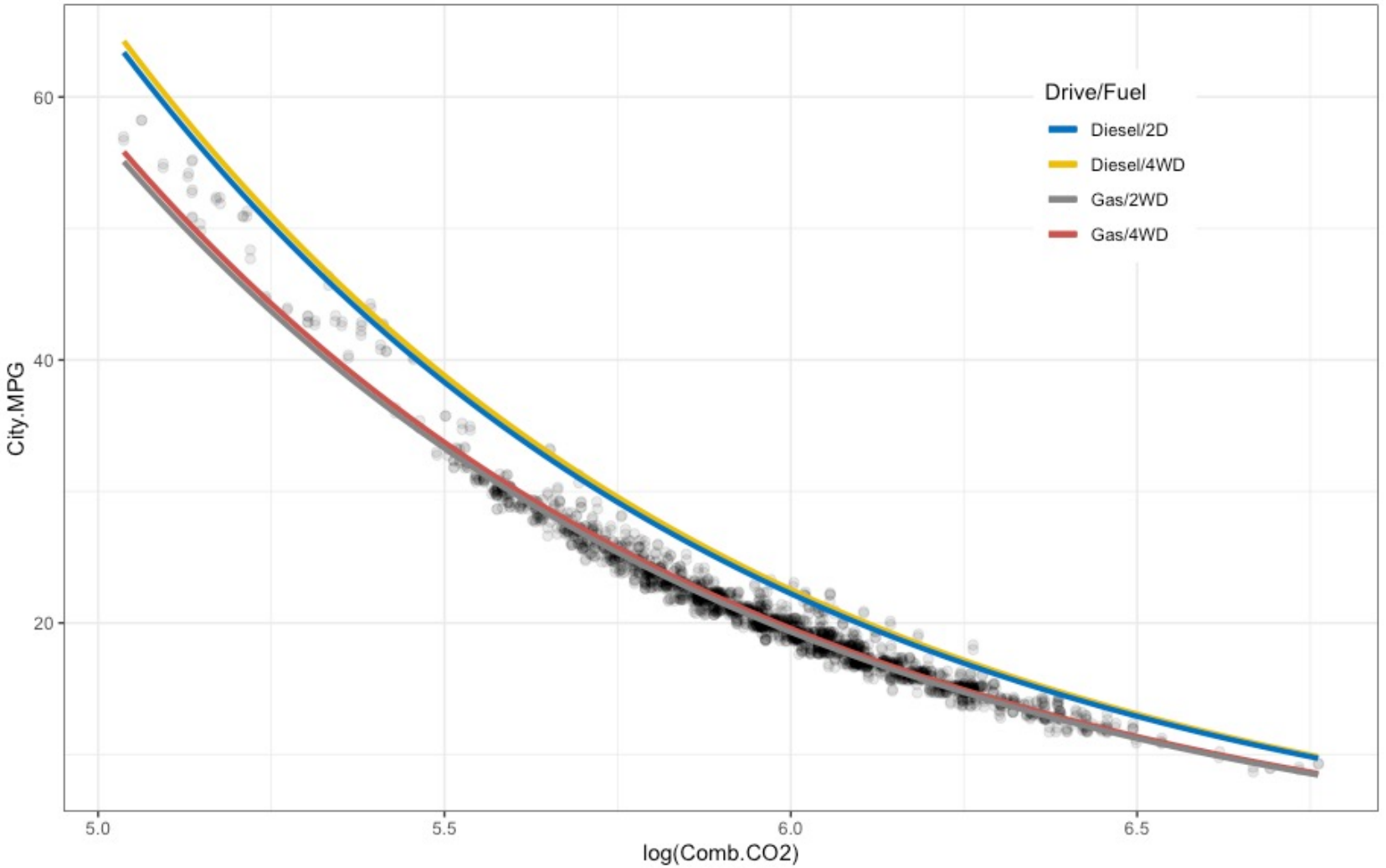

Drive/Fuel
Diesel/2D
Diesel/4WD
Gas/2WD
Gas/4WD
log(City.MPG)
4.0
3.5
3.0
2.5
5.0
5.5
6.0
6.5
log(Comb.CO2)

I – epa gas standards

Do the type of fuel, drive, number of cylinders or CO2 emissions explain the city gas mileage of 2020 vehicles?




```
eeq1=function(x){exp(cfs[1] + cfs[2] + cfs[3] + cfs[4]*mean(gas$Cyl) + cfs[5]*x)}
eeq2 =function(x){exp(cfs[1] + cfs[3] + cfs[4]*mean(gas$Cyl) + cfs[5]*x)}
eeq3 =function(x){exp(cfs[1] + cfs[2] + cfs[4]*mean(gas$Cyl) + cfs[5]*x)}
eeq4 =function(x){exp(cfs[1]+ cfs[4]*mean(gas$Cyl) + cfs[5]*x)}

ggplot(data = gas, aes(x = log(Comb.CO2), y = City.MPG)) +
       geom_jitter(shape = 19, size = 2, fill = my_jco[3], alpha = 0.1) +
        stat_function(fun = eeq1, geom="line", colour = my_jco[1]) +
        stat_function(fun = eeq2, geom="line", colour = my_jco[2]) +
        stat_function(fun = eeq3, geom="line", colour = my_jco[3]) +
        stat_function(fun = eeq4, geom="line", colour = my_jco[4]) +
      theme_bw()
```


Drive/Fuel
Diesel/2D
Diesel/4WD
Gas/2WD
Gas/4WD
City.MPG
60
40
20
5.0
5.5
6.0
6.5
log(Comb.CO2)

Questions?