Bayesian Statistics: Techniques and Models Capstone by P.L. Resubmission (09.09.2023)

Executive Summary:

Data on the CO_2 emissions of vehicles was used to fit two different linear models in order to predict the CO_2 emissions of cars. While a simple linear regression model using fuel consumption as a regressor did not provide good results, a hierarchical model, with regression coefficients based on the fuel type, showed good results. CO_2 emissions of 2.34 kg/l for gasoline and 2.68 kg/l for Diesel were found (for both numbers rjags reported naive standard errors of approximately 10^{-4} kg/l).

1. The problem:

During the Bayesian Statistics: Techniques and Models course, hierarchical modeling was one of the most interesting topics for me. I wanted to look for a dataset that was related to environmental issues and at the same time allowed hierarchical modeling.

While searching for data, I saw various datasets on Kaggle (e.g. this one here). This dataset was based on data from the original source by “Canada Open Government” [1]. The dataset contains data on various cars: their CO_2 emissions, fuel consumption, fuel type, engine size, number of cylinders, vehicle type e.t.c. Task is predicting the CO_2 emissions of a vehicle.

2. Data

The Canada Open Government homepage has data for car models from different years and information on properties like vehicle type, brand, fuel consumption, fuel type, engine size, number of cylinders and the response variable CO_2 emissions.

Data is available in the csv format, only slight modifications to the header and removal of some auxiliary information at the end of the files are necessary to allow an easy import of data. After downloading data for the years 2019 - 2023, data for these 5 years was combined into one csv file for model fitting in R. Most data processing, exploratory analysis and plotting of results was done in Python, only the modeling itself was done in R. There were no major challenges, but the categorical values for fuel type had to be changed to integer in order to use them as factors in modeling.

The pair plot for some of the variables (Engine Size, Cylinders, Fuel consumption combined and CO2 emissions) shows that there is correlation between all of these variables. The best linear relationship can be observed between fuel consumption and CO_2 emissions, however, it appears as if there are multiple distinct lines.

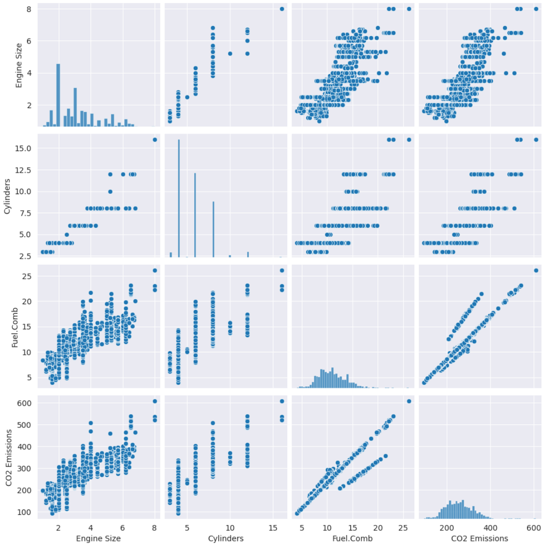


Figure 1: Pairs plot of the variables engine size, number cylinder, fuel consumption and CO2 emissions.

If we look at the scatter plot of fuel consumption vs. CO_2 emissions, where different colors represent different fuel types, we can observe that the different types of fuel lead to different relationships between CO_2 emissions and fuel consumption. If you remember your high-school chemistry class (or read up on Wikipedia [2]), you'll know that different fuel types lead to different amounts of CO_2 emitted per liter of fuel burned.

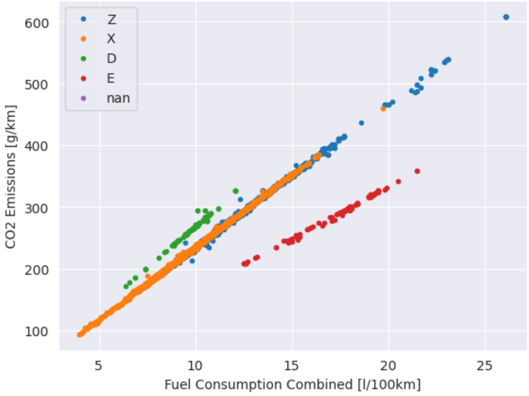
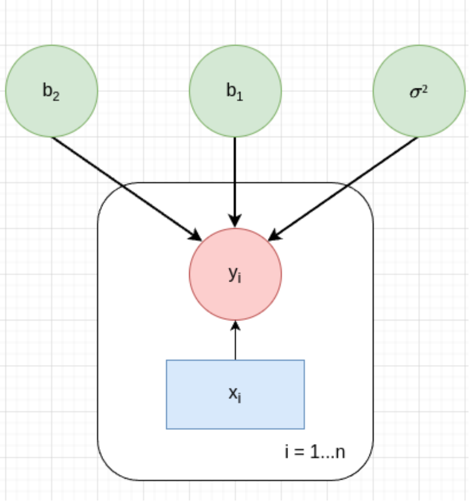
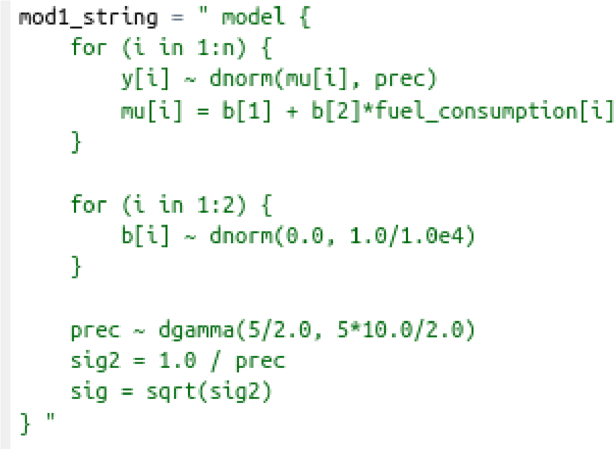


Figure 2: Fuel consumption vs. CO_2 emissions for different fuel types. (X=gasoline, Z=premium gasoline, D=Diesel, E=Ethanol)

3. Modeling

As a first baseline model a simple linear regression model with fuel consumption as the regressor was fitted using rjags. A model similar to Lesson 7 of the course was fitted, however with a smaller standard deviation, since we don't expect very large values for the bias or the regression coefficient. A graphical representation of the model and the rjags model definition can be found in Figure 3. After an initial burn-in phase of 1000 samples, 3 chains with 10.000 samples were calculated. The Gelman scale reduction factor was close to 1, the autocorrelation vanished at lags of 50 and the smallest effective sample size was around 1000. The MCMC chain seems to have converged. However, as can be seen in the CO_2 emission vs. residual plot in Figure 5 (left side), the residuals are rather large, as has to be expected since we earlier observed multiple linear curves. The posterior mean for the regression coefficient yields b1 = 2.12 kg/l.

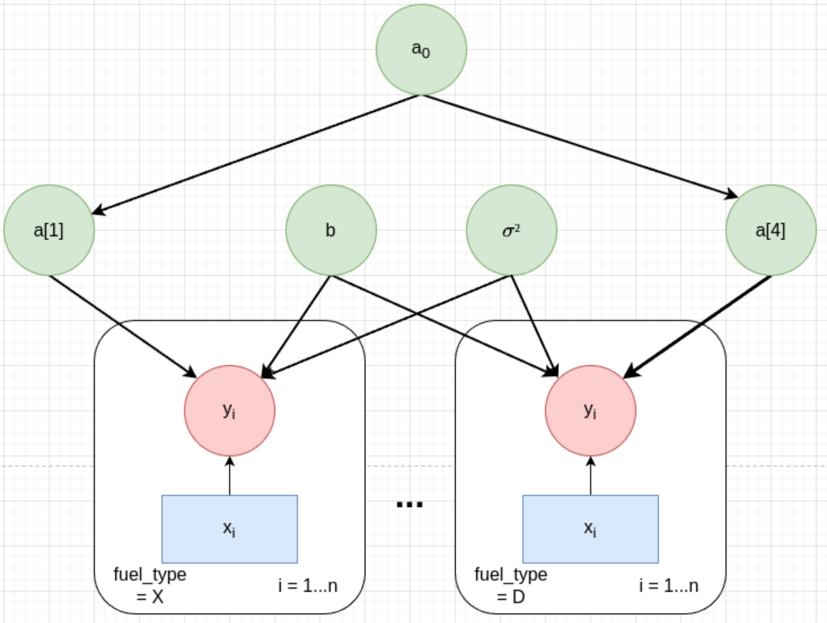


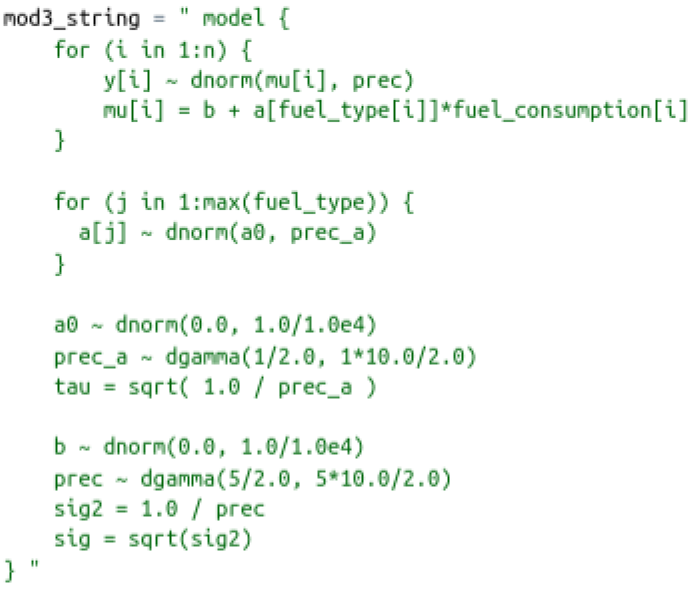


```
mod1_string = " model {
    for (i in 1:n) {
        y[i] ~ dnorm(mu[i], prec)
        mu[i] = b[1] + b[2]*fuel_consumption[i]
    }

    for (i in 1:2) {
        b[i] ~ dnorm(0.0, 1.0/1.0e4)
    }

    prec ~ dgamma(5/2.0, 5*10.0/2.0)
    sig2 = 1.0 / prec
    sig = sqrt(sig2)
} "
```

Figure 3: Graphical (left side) representation and rjags model definition for the simple linear regression model. Note that b_0 on the left side corresponds to b[1] in the code.

As we expect a different linear relationship for different fuel types, a second model was fitted in rjags. The second model was a hierarchical model with a fixed bias but regression coefficients depending on the fuel type. The regression coefficients are modeled to be normally and drawn from a normal distribution with mean 0 with a rather large standard deviation of 10000. The different convergence criteria also looked good after an initial 1000 steps, the MCMC chain had converged. The model works well and the residuals of the fit are better than for the first model (see Figure 5).





```
mod3_string = " model {
    for (i in 1:n) {
        y[i] ~ dnorm(mu[i], prec)
        mu[i] = b + a[fuel_type[i]]*fuel_consumption[i]
    }

    for (j in 1:max(fuel_type)) {
      a[j] ~ dnorm(a0, prec_a)
    }

    a0 ~ dnorm(0.0, 1.0/1.0e4)
    prec_a ~ dgamma(1/2.0, 1*10.0/2.0)
    tau = sqrt( 1.0 / prec_a )

    b ~ dnorm(0.0, 1.0/1.0e4)
    prec ~ dgamma(5/2.0, 5*10.0/2.0)
    sig2 = 1.0 / prec
    sig = sqrt(sig2)
} "
```

Figure 4: Graphical (left side) representation and rjags model definition for the hierarchical model, with a fuel type dependent regression coefficient b1. Note that for simplicity, the model is only shown for two different fuel types on the left side, while the model that was fitted, contained all four types.

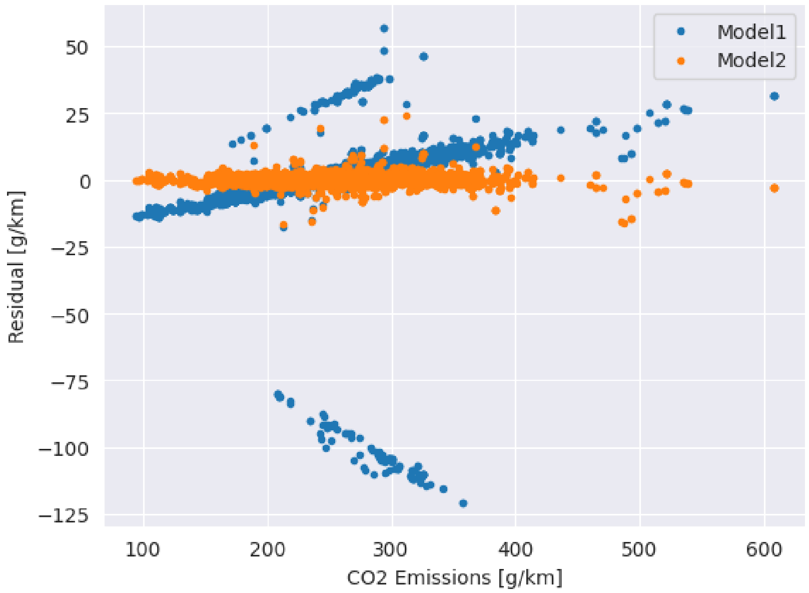


Figure 5: Residuals vs. CO_2 emissions. It can be clearly seen that the second model has much smaller and less structured residual structures.

The fit for the second model yielded posterior means of the regression coefficients of about 2.34 kg/l for both gasoline types, 2.68 kg/l for Diesel and 1.66 kg/l for Ethanol.

Conclusion:

The comparison of the two models clearly shows that we need to account for different fuel types when building a model to predict CO_2 emissions. A not very scientific, but quick comparison with EPA data [3] shows that the results for Gasoline and Diesel CO_2 emissions per liter fuel are close to the literature values.

References

[1] Open Government Canada, “**Fuel consumption ratings”,** URL: https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64, last accessed: 2023-09-06.
[2] Diesel fuel on Wikipedia, URL: https://en.wikipedia.org/wiki/Diesel_fuel#Carbon_dioxide_formation, last accessed: 2023-09-07.
[3] EPA: Greenhouse Gas Emissions from a Typical Passenger Vehicle, URL: https://www.epa.gov/greenvehicles/greenhouse-gas-emissions-typical-passenger-vehicle, last accessed: 2023--09-08.