Bayesian Regression

Priors and Pooling

Introduction

The objective is to examine the impact of pooling on a Bayesian regression models. We will develop 2 models: first with no pooling and next with partial pooling, followed by a discussion of results.

The data comes from the *Automobile Price Prediction.csv* dataset *(which you should be familiar with)*. The data are nested by model, and because our purpose is to examine the effect of pooling, we will only use 2 variables:

- model *(the grouping variable)*
- horsepower *(the indepedent variable)*

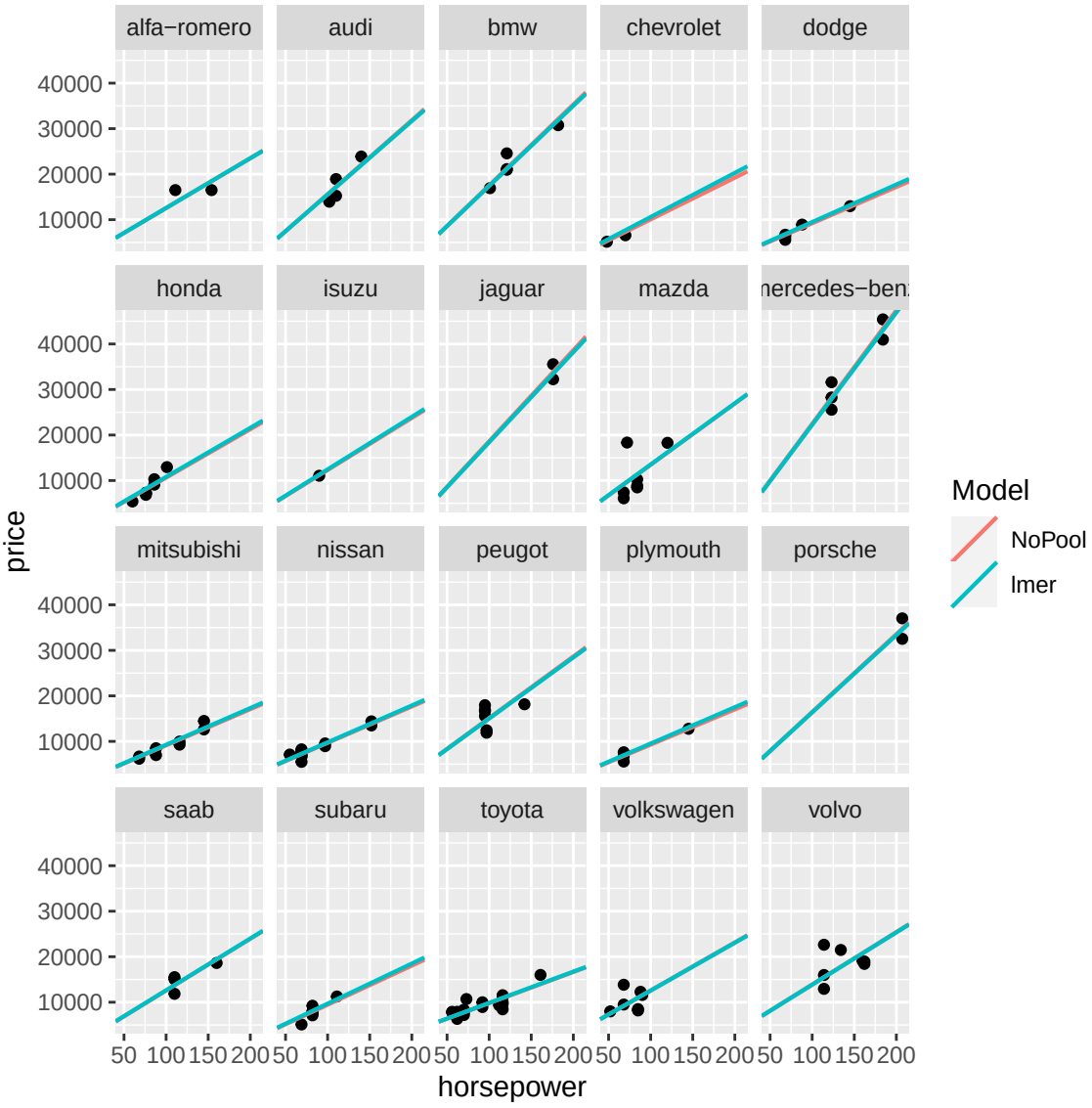
Model 1 - No Pooling

The no-pooled model with comments is shown below:

Note how the equation for y_hat is nested with a and b defined by primary key with foreign keys in the data (x).

Analysis - Model 1

After mapping the model parameters by make *(see R file)*, the following plot summarizes the no-pool approach, compared to a frequentist mixed-effects model in lme4:



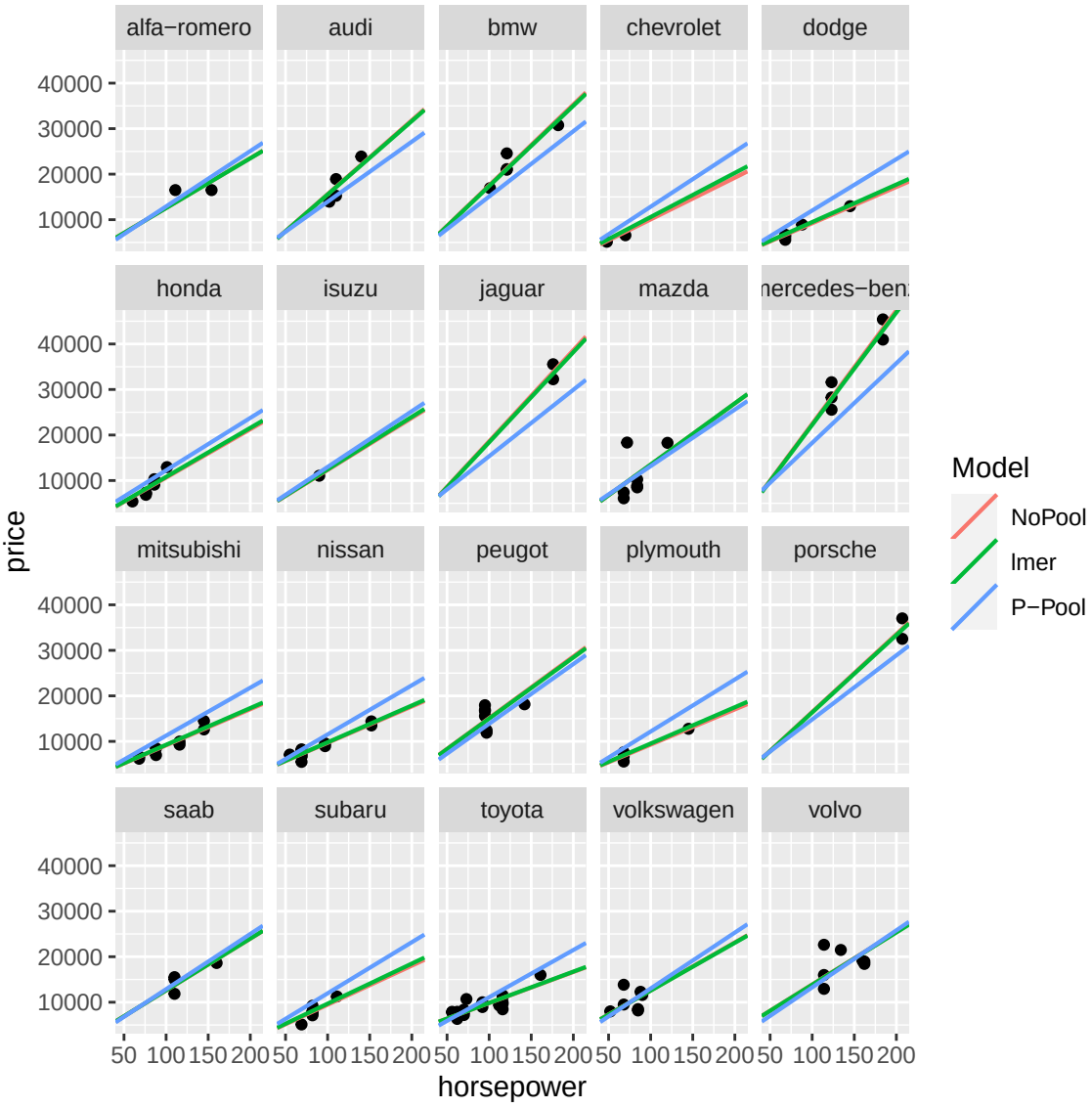
Comparing the RMSEs between models, we would expect the results to be very close:

```
## [1] 2839.092
```

```
## [1] 2869.748
```

Model 2 - Partial Pooling

The partial-pooling model is uses the same Stan model, but this time we use priors to pull parameters towards the mean *(which is the partial pooling effect)*. We can do this as follows:



Notice how the p-pool regression lines varies less across models *(notice how the slopes are more consistent)*. Also notice how, in test data, the groups with less data has less impact on the partailly pooled model.

Analysis

There are many possible combinations of pooling and with most data, even a simple dataset like this one. The usage of Bayesian priors gives us great flexiblity in controlling the effect of pools *(note that we can set a prior mean AND variance for EACH grouping)*. To restate a few of the advantages:

- Crossed effects let us differentiate pricing between models *(a shopper expecting to buy a Mercedes based on an average of all models is going to be very disappointed)*. So we have the ability to target expected values.
- Partial pooling lets us tune effects for each group - data tends to normalize inter-group, as well in intra-group and inter-group. In many cases, neither no-pooling nor complete pooling will be a good approach.
- Partial pooling lets us create predictions for groups that have little data *(a no pooled model will fail if there are few data points)*

- Generalization. Using nested models with priors gives us the ability to generalize models in a very targeted way - by group, by paramter. This level of control is just not possible with any other approach to modeling.

Comparing RMSE's for all models *(Bayes No-Pool, LME4, Bayes Partial-Pool)*:

```
## [1] 2839.092
```

```
## [1] 2869.748
```

```
## [1] 3352.525
```

Note that partial pooling does not always yield a lower RMSE. But this is RMSE with the **TEST** data. As you've learned from generalization, the point is to prepare models for real data - to avoid overfitting. The idea of pooling **BY PARAMETER** gives us total control over our models.