Linear Regression Models

We consider the following simple linear model for the data

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \text{with } \epsilon_i \overset{iid}{\sim} N(0, \sigma^2),\ i = 1, ..., n$$

or equivalently

$$y_i \mid x_i \sim N\Big(\beta_0 + \beta_1 x_i, \sigma^2\Big)$$

using vector notation, we have

$$\mathbf{y} = \mathbf{x}\beta + \epsilon\ , \qquad \mathbf{y} \mid \mathbf{x}\beta \sim N(\mathbf{x}\beta, \sigma^2)$$

Note: The parameters are β_0, β_0 and the variance σ^2 assumed to be constant accross all observations. They are assumed to be fixed and unknown. The x_i are assumed to be fixed and the y_i are assumed to be random, because of the random error terms e_i.

Classical approach to Bayesian Linear Regression Models

As usual in Bayesian statistics, the posterior distribution is proportional to the likelihood times the prior distributions of the parameters, so that we have $\pi(\beta, \sigma^2 \mid \mathbf{y}) \propto L(\beta, \sigma^2)\ \pi(\beta, \sigma^2)$.

Many choices for priors are available. For example, if the following noninformative reference priors are choosen $\pi(\beta \mid \sigma^2) \propto 1$ and $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$, then, the posterior, up to proportionality, becomes

$$\begin{aligned}
\pi(\beta, \sigma^2 \mid \mathbf{y}) &\propto \prod_{i=1}^{n} L(\beta, \sigma^2)\ \pi(\beta, \sigma^2) \\
&\propto (\sigma^2)^{-n/2}\ \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2 \right] \left(\frac{1}{\sigma^2}\right) \\
&\propto (\sigma^2)^{-(n+2)/2}\ \exp\left[-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{x}\beta)^T (\mathbf{y} - \mathbf{x}\beta)) \right]
\end{aligned}$$

Inference for the Bayesian Linear Regression Models

As usual in Bayesian statistics, the posterior distribution is proportional to the likelihood times the prior distributions of the parameters, so that we have $\pi(\beta, \sigma^2 \mid \mathbf{y}) \propto L(\beta, \sigma^2)\ \pi(\beta, \sigma^2)$.

Many choices for priors are available. For example, if the following noninformative reference priors are choosen $\pi(\beta \mid \sigma^2) \propto 1$ and $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$, then, the posterior, up to proportionality, becomes

$$\begin{aligned}
\pi(\beta, \sigma^2 \mid \mathbf{y}) &\propto \prod_{i=1}^{n} L(\beta, \sigma^2)\ \pi(\beta, \sigma^2) \\
&\propto (\sigma^2)^{-n/2}\ \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2 \right] \left(\frac{1}{\sigma^2}\right) \\
&\propto (\sigma^2)^{-(n+2)/2}\ \exp\left[-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{x}\beta)^T (\mathbf{y} - \mathbf{x}\beta)) \right]
\end{aligned}$$

Cars dataset in R

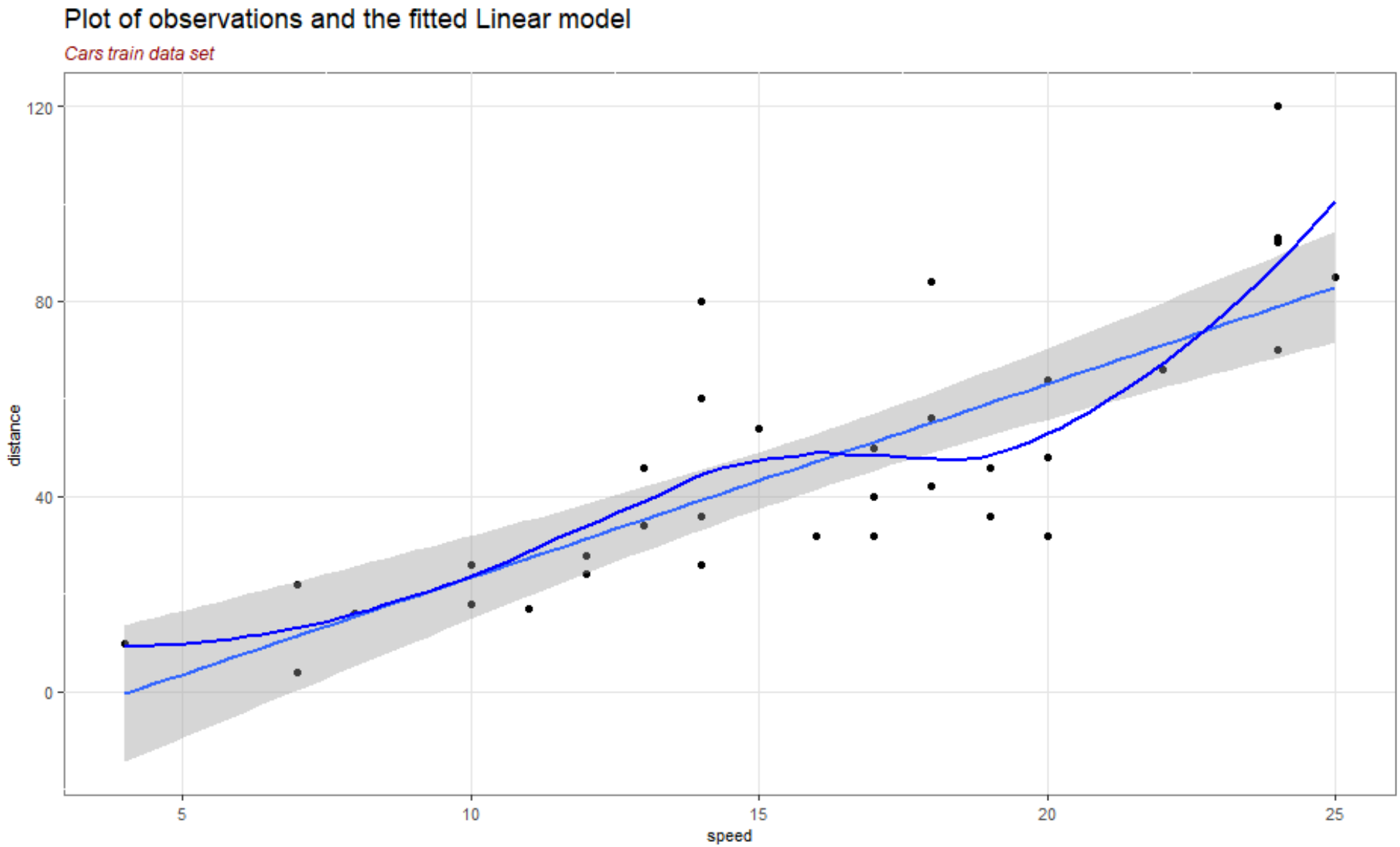
We work with the dataset 'cars' to illustrate linear and log-linear models, both in the traditional frequentist and bayesian framework. Two variables are considered: 'dist', which corresponds to speed (mph) and 'dist' which corresponds to the stopping distance (ft). The dataset, which consitst in 50 observations for each variable is then split into a train set (70%) and a test set. Bellow the R code to access the initial data set, partitioning it and creating linear models.

```r
library(tidyverse)
data(cars)

# 0.1 Split the dataset into train and test sets
set.seed(2024)
sample = sample(c(TRUE, FALSE), nrow(cars), replace=T, prob=c(0.7, 0.3))
#Split your data into training (70%) and test (30%) sets
train = cars[sample, ]
test = cars[!sample, ]

# 1.1 Create frequentist linear model, log linear model and plot
freq.lin.mod = lm(dist~speed, data=train)
freq.log.lin.mod = lm(log(dist)~speed, data = train)

summary(freq.lin.mod)
summary(freq.log.lin.mod)
```

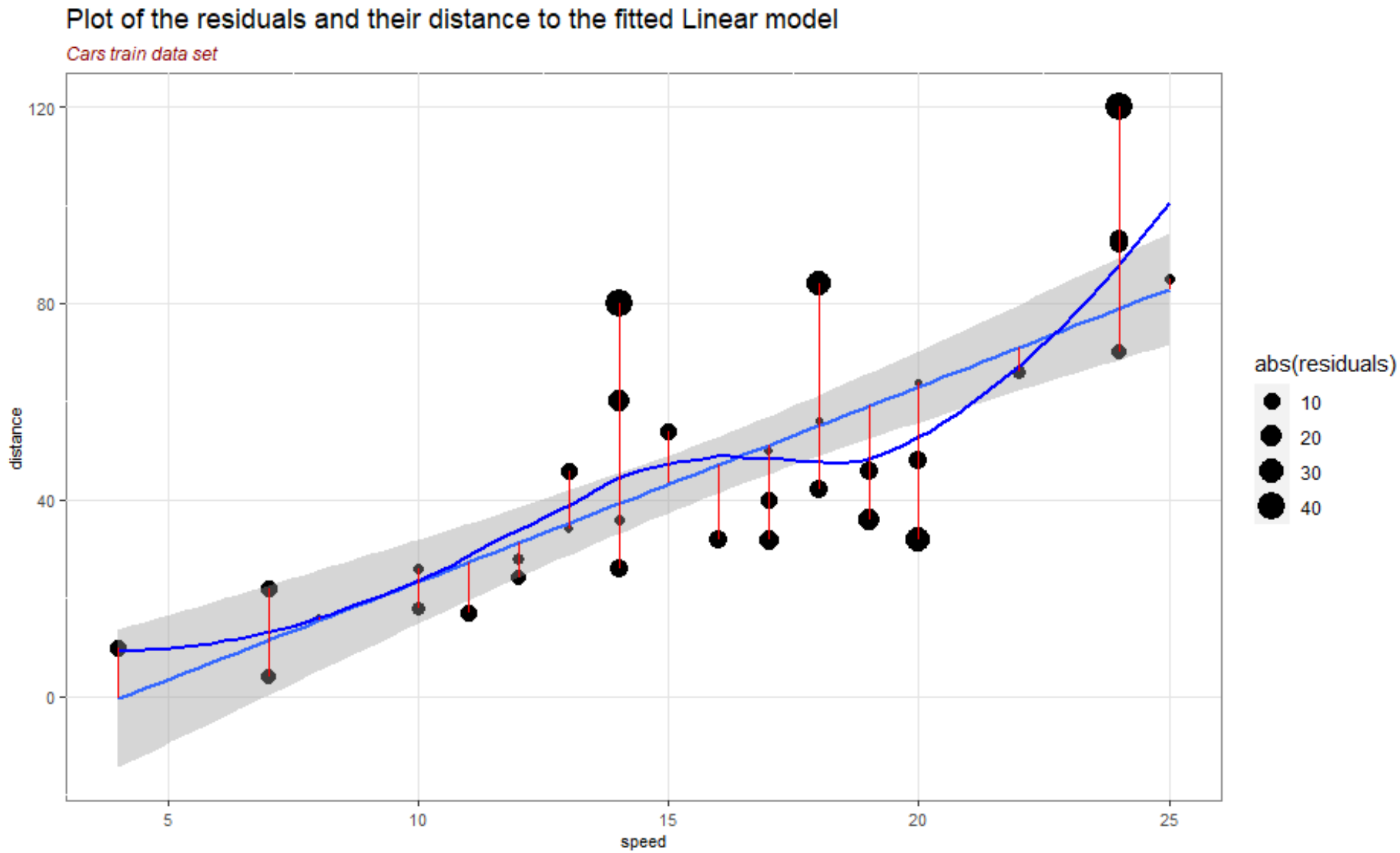
Frequentist LM outputs

	Dependent variable:
	dist
speed	3.968^{***} (0.527)
Constant	-16.293^{*} (8.843)
Observations	34
R^2	0.639
Adjusted R^2	0.628
Residual Std. Error	16.569 (df = 32)
F Statistic	56.652^{***} (df = 1; 32)
Note:	*p<0.1; **p<0.05; ***p<0.01

Frequentist LM, observations



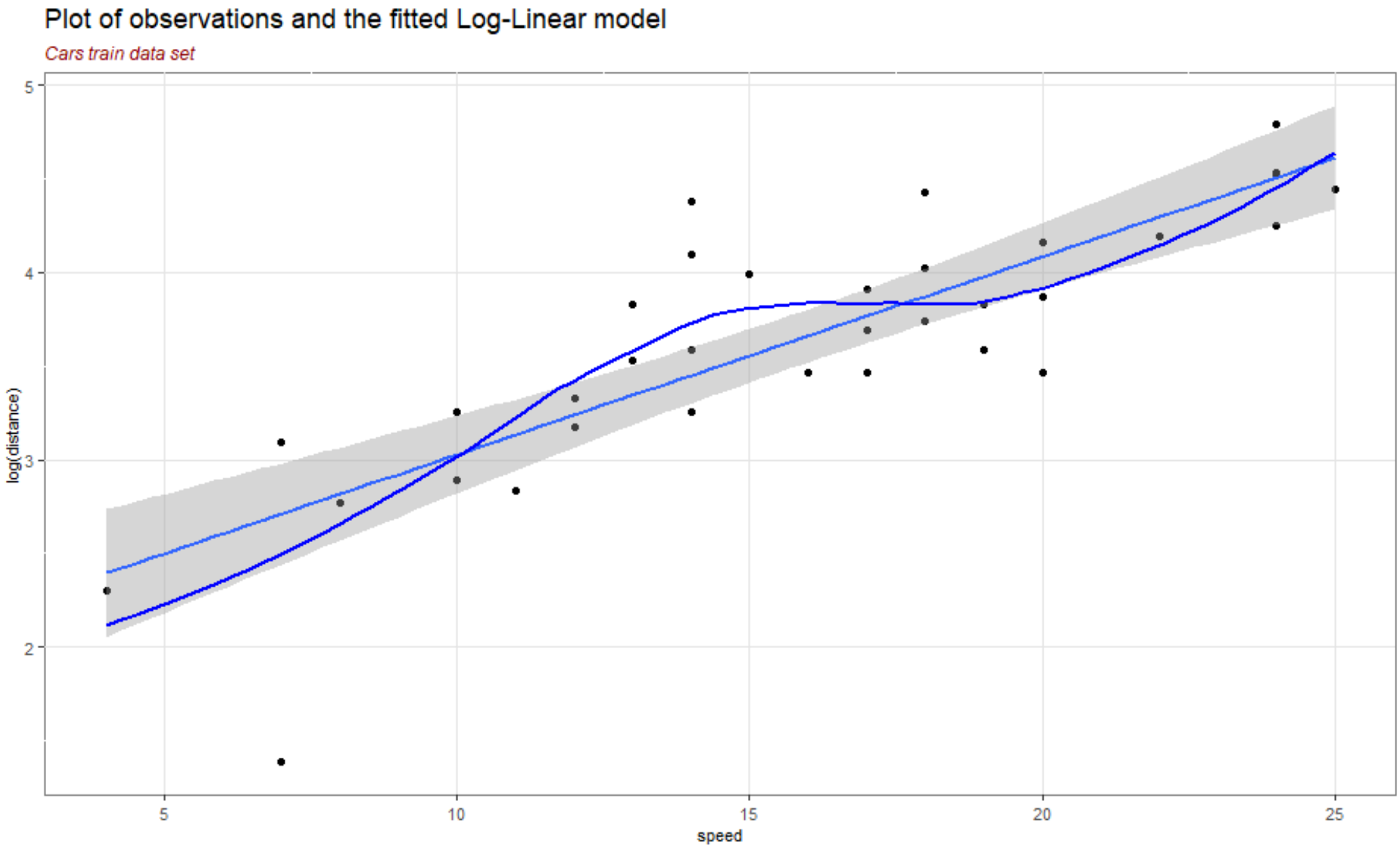
Frequentist LM, residuals



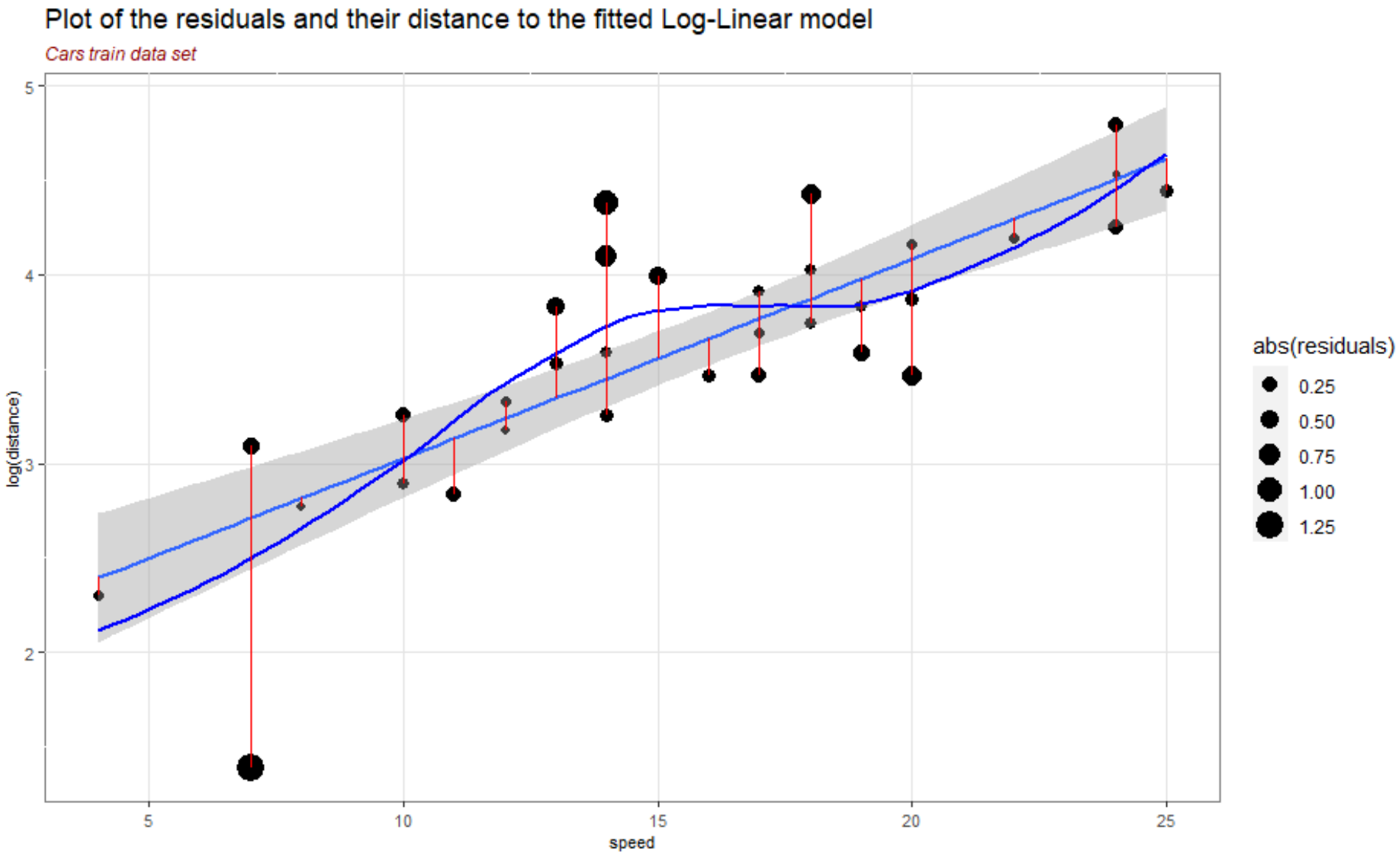
Frequentist LM outputs

	Dependent variable:
	dist
speed	3.968^{***} (0.527)
Constant	-16.293^{*} (8.843)
Observations	34
R^2	0.639
Adjusted R^2	0.628
Residual Std. Error	16.569 (df = 32)
F Statistic	56.652^{***} (df = 1; 32)
Note:	*p<0.1; **p<0.05; ***p<0.01

Frequentist LLM, observations



Frequentist LLM, residuals

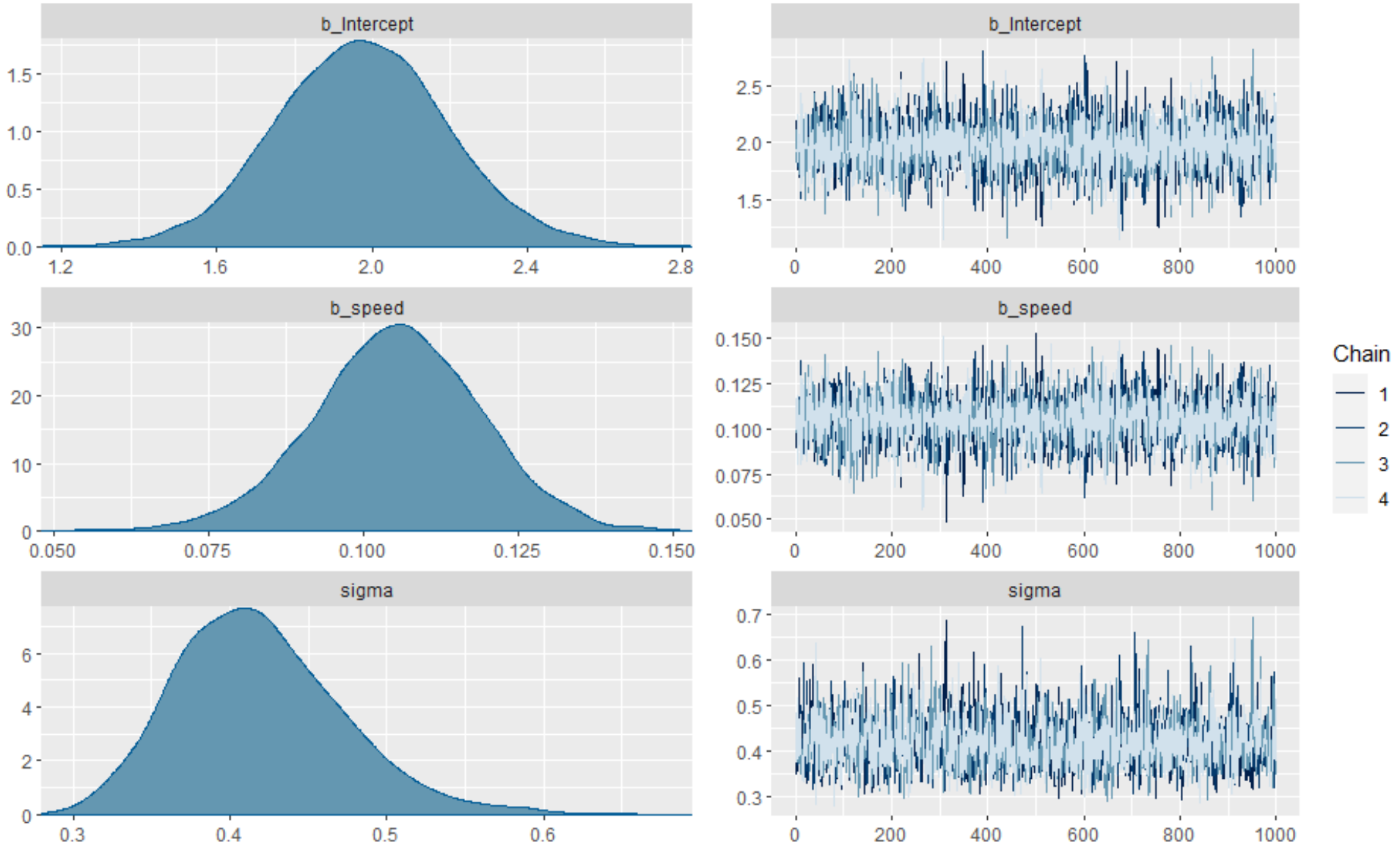


Bayesian LM outputs

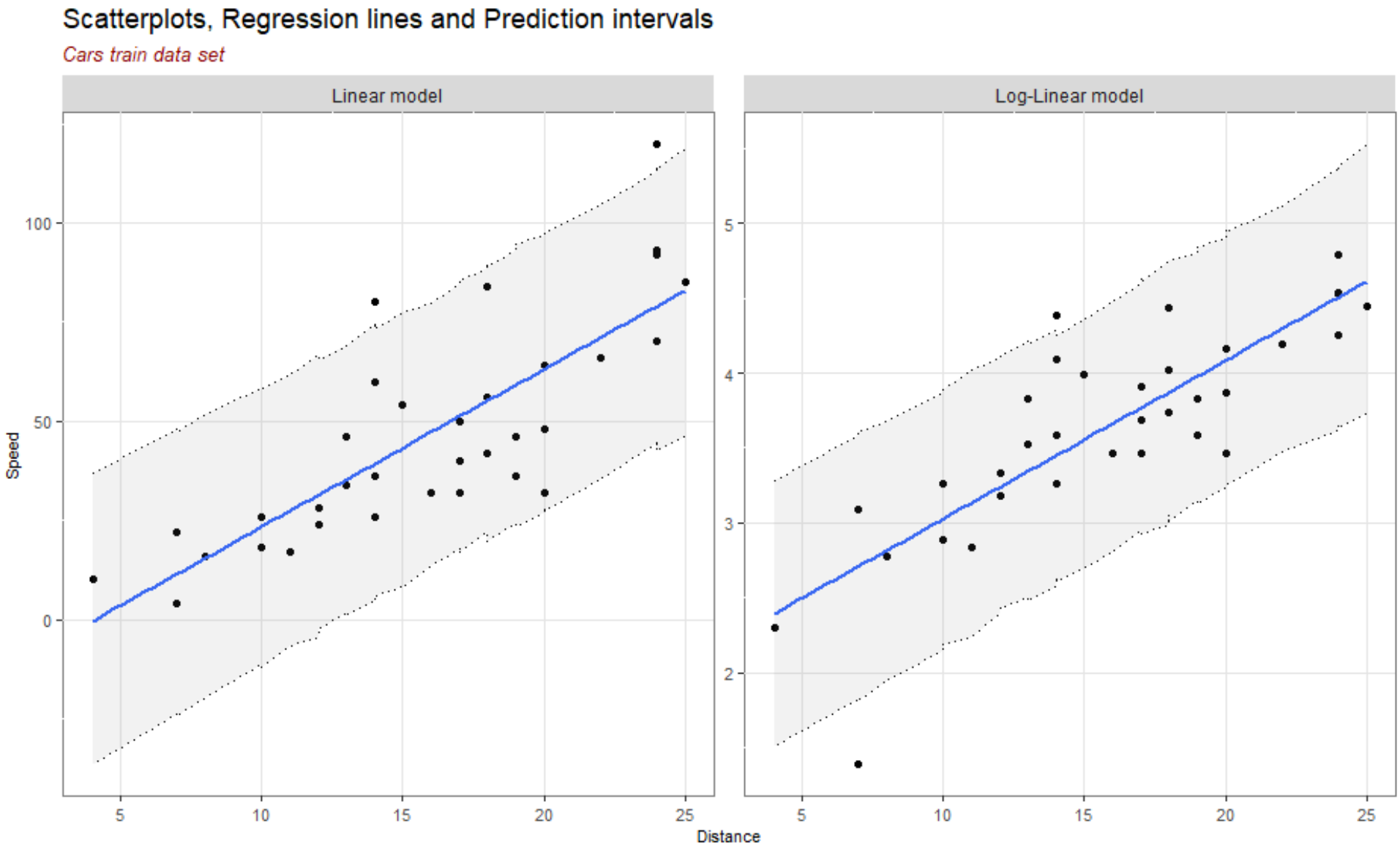
	Dependent variable:
	dist
speed	3.96 (0.54)
Constant	-16.27 (9.15)
Observations	34

Bayesian LLM outputs

Bayesian LLM, posterior densities



Bayesian LLM, fitted models



References

The R Project for Statistical Computing:
https://www.r-project.org/

Python:
https://www.python.org/

course notes