Bayesian Linear Regression

STAT-S 626 FINAL PROJECT

Ashish Patidar
Saumya Mehta

EXECUTIVE SUMMARY

We aimed at predicting the apparent temperature for Szeged, Hungary based on different explanatory variables. To dive deep further into the aspects of depression the main research questions that we've tried to address are the following research questions:

- How do Humidity, Visibility, and Wind Speed affect the apparent temperature?
- Does interaction between explanatory variables generate a better regression model?

During our initial analysis, we found a few columns like temperature, humidity, visibility, wind_speed, and wind_degrees relevant for our analysis based on the correlation pair plot. The temperature was highly correlated with apparent temperature and including it in our analysis resulted in better prediction from our models. Without temperature, the model needs to use every available variable to reasonable prediction for apparent temperature.

In our analysis, we found that humidity was a strong predictor(with a larger CI) when using Zellner-g prior while it was non-significant when using a default prior. Humidity was also non-significant when adding interaction between explanatory variables in either case. Wind speed, on the other hand, remained a significant factor when adding interactions with the default prior. We observed interactions between different groups of explanatory variables for each case. For the case of Zellner-g priors, the interaction between temperature and wind_speed, humidity and wind_speed and an interaction between temperature, wind_speed and humidity was significant. While in default prior case, the interaction between temperature and wind_speed was non significant.

The samples in all cases were independent. We found that for Zellner-g prior, the samples become more efficient when adding interactions.

In both cases, adding an interaction caused an improvement in the prediction error.

DATASET DESCRIPTION

The weather in the Szeged dataset is available on Kaggle for 2006-2016. The dataset consists of 12 columns including the apparent temperature which is the prediction variable.

COLUMN NAME	DESCRIPTION
Date	Contains the date and time on which data was collected
Summary	Contains the summarized weather of the day, eg: Partly cloudy, overcast, etc.
Precipitation Type	Contains if the precipitation is rain, snow or other
Temperature	Contains the actual temperature of the day at the time when data is collected
Apparent Temperature	Contains the temperature affected by other factors, like wind speed, visibility, humidity, etc.

The rest of other columns are self-explanatory containing the values for humidity, wind speed, wind bearing, visibility, loud cover, and pressure.

The daily summary column is almost the same as the summary column described above.

Figure 1 shows the correlation plot between explanatory variables of interest. We can see that temperature and apparent temperature are highly correlated. Humidity has the next highest (albeit in the negative side) correlation with apparent temperature followed by visibility and wind_speed. Wind degrees has barely any correlation with our variable of interest. The figure also shows the distribution of variables and we can see that temperature and apparent temperature almost seem distributed like a mixture of two normals. Wind degrees and visibility seem to be distributed in a multimodal fashion and humidity and wind_speed appear to be skewed to the left and right respectively.

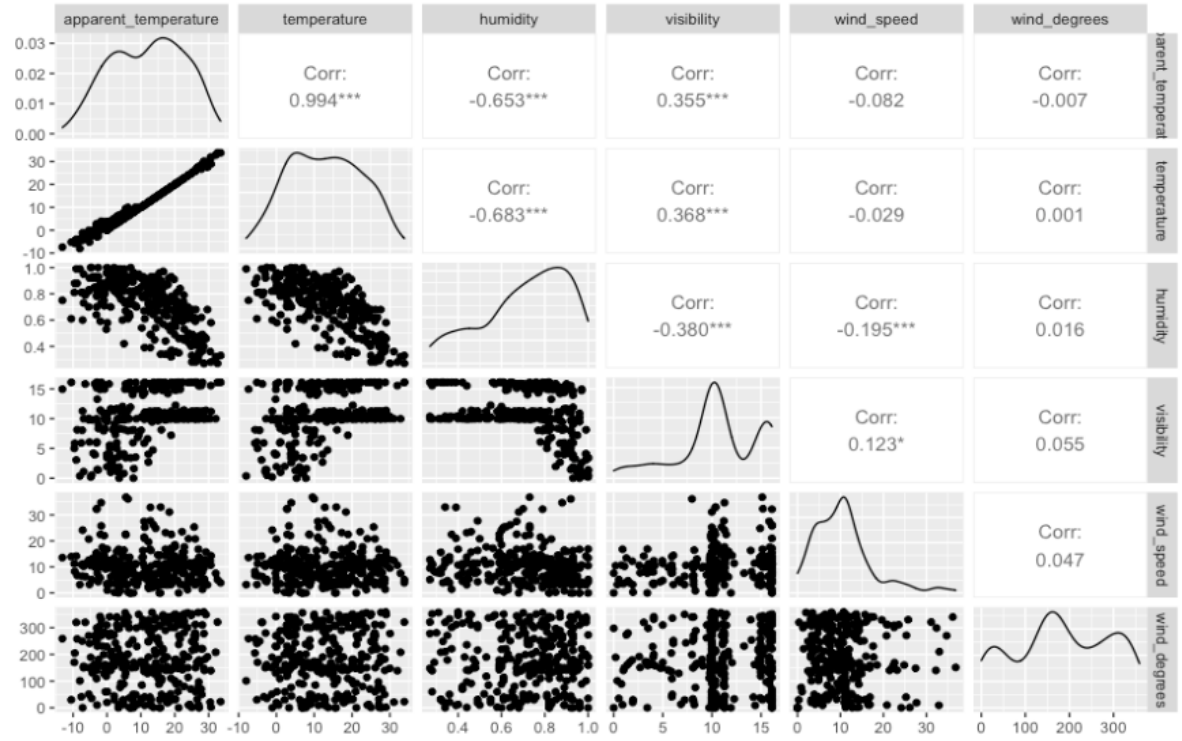

Figure 1: Pairs plots of all the predictors

BAYESIAN ANALYSIS

Bayesian Linear Regression is the conditional modeling of data to obtain a posterior probability of regression coefficients from the assumed prior probabilities of them. The standard syntax for Bayesian Linear Regression is given as:

$$y \sim N(\beta^T X, \sigma^2 I)$$

Here, the response variable is not a point estimate but a normal distribution with mean $\beta^T X$ and variance $\sigma^2 I$.

For this experiment we are only considering uninformative priors:

- Default Prior
- Zellner-g prior

The default prior in bayesian statistics is nothing but the ordinary least square values of beta and sigma, which are calculated as

- $_{OLS} = (X^T X)^{-1} X^T y$
- $\sigma^2 = SSR/(n-p)$

Here, SSR = $(y - X_{OLS})^T (y - X_{OLS})$, and n = number of rows in X, p = number of columns or features we are considering for linear model.

Zellner-g prior is a special case of default(OLS) prior for which beta and sigma can be calculated as

- $= (g/g + 1)^* (X^T X)^{-1} X^T y$
- $\sigma^2 = SSR_g/(n-p)$(not sure)

For this $SSR_g = y^T (I - (g/g + 1) X(X^T X)^{-1} X^T) y$.

Default Priors(without interaction):

This weather dataset contains 7 numerical variables which could be used as a feature in linear regression but not all of them are relevant, therefore to identify the important variables a graph for 95% confidence intervals of coefficients of all the variables is constructed.

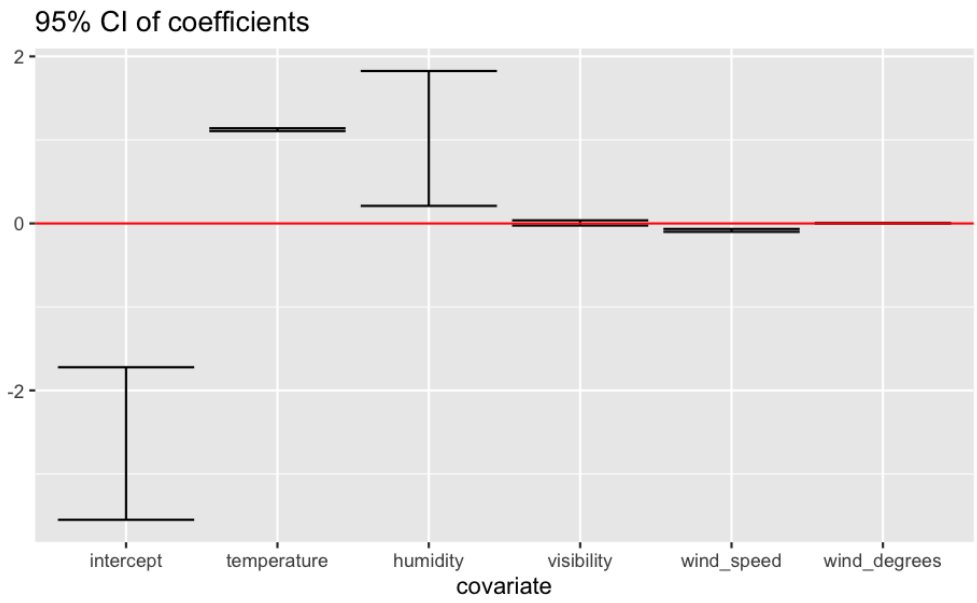

Figure 2: 95% CI for Coefficients

From this plot and model selection, intercept, temperature, and wind speed are the most important variables for creating our bayesian linear regression model to predict apparent temperature without any interaction between the predictors.

The below graph shows the comparison between predicted and observed points,

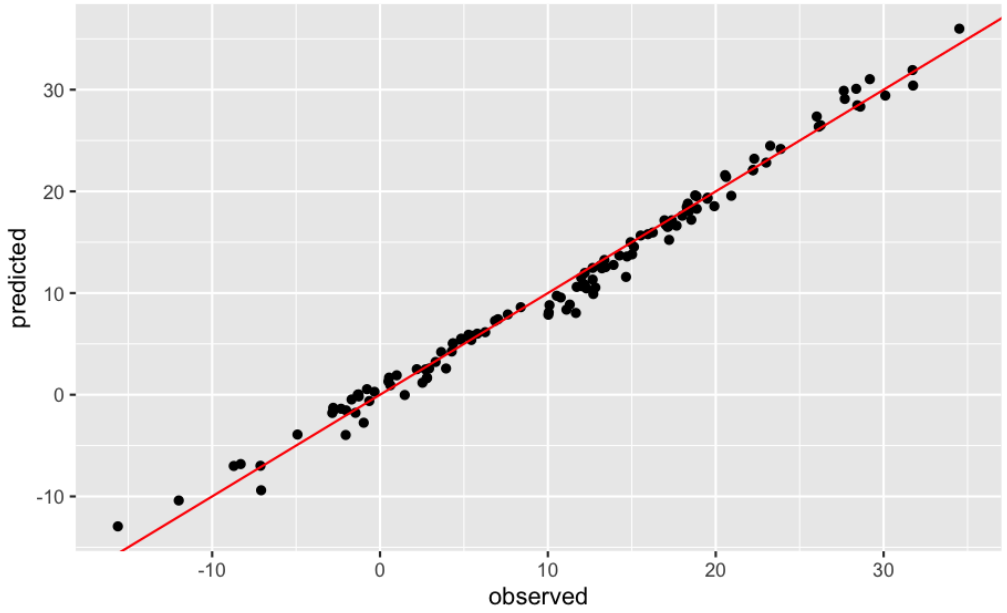

Figure 3: Predicted vs Observed outcomes

From figure 3, it looks clear that our model is performing very well for the data, and the mean square error for this model is 1.40.

For all the predictors values of beta are moving constantly around the mode value which shows that our sample size of data is good for modelling.

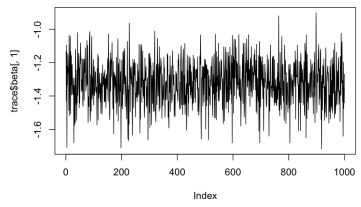

Figure 4.1: Trace Plots beta(intercept)

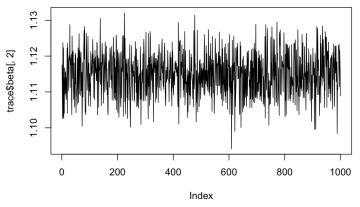

Figure 4.1: Trace Plots beta(temperature)

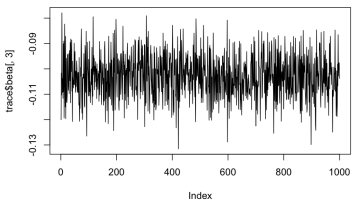

Figure 4.1: Trace Plots beta(wind speed)

The auto correlation is almost zero so samples are relatively independent which is the ideal condition of sampling.

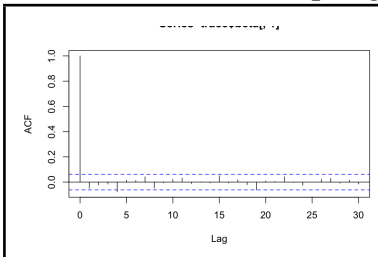

Figure 5.1: ACF of beta(intercept)

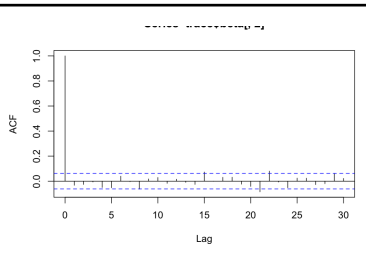

Figure 5.2: ACF of beta(temperature)

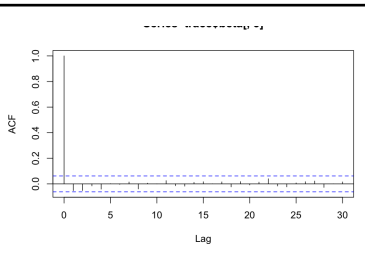

Figure 5.3: ACF of beta(wind speed)

The effective sample size for intecept, temperature, and wind speed.

```
    var1     var2     var3
1406.446 1000.000 1236.656
```

Figure 6: Effective sample size for each variable

Zellner-g Priors(without Interaction):

The above steps of model selection and 95% confidence intervals of coefficients are performed for all the variables using Zellner-g priors. The results are almost the same as Zellner-g is a special case of default OLS priors.

The below graph proves our assumption of considering intercept, temperature, humidity, and wind speed as important features for modeling our bayesian linear regression model.

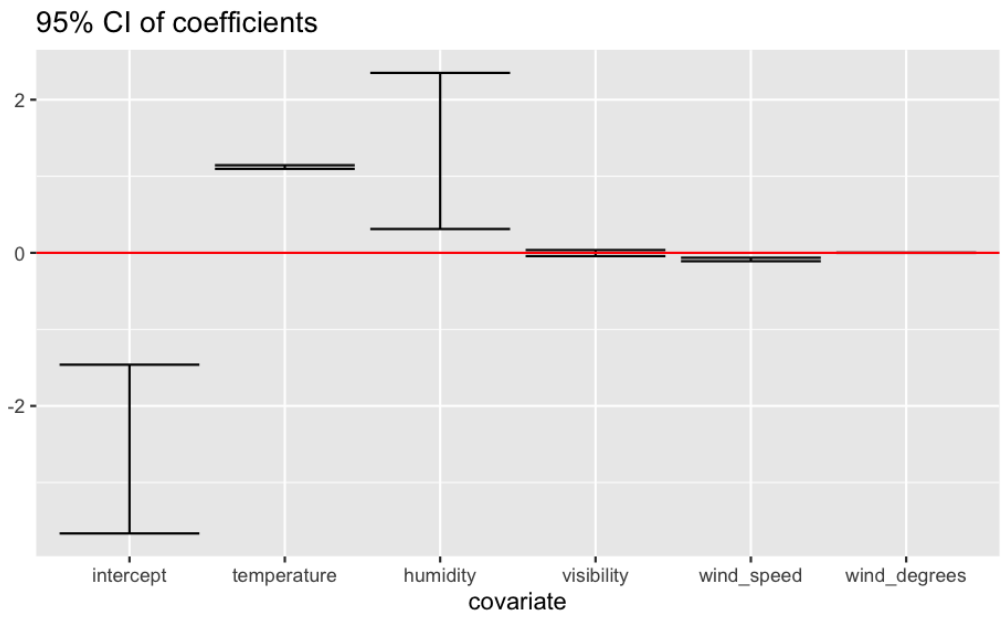

Figure 7: 95% CI of coefficients for Zellner-g Priors

Also, the results after considering default OLS prior and Zellner-g prior with no interaction of predictors are almost the same which can be seen from the predicted vs observed points graph below.

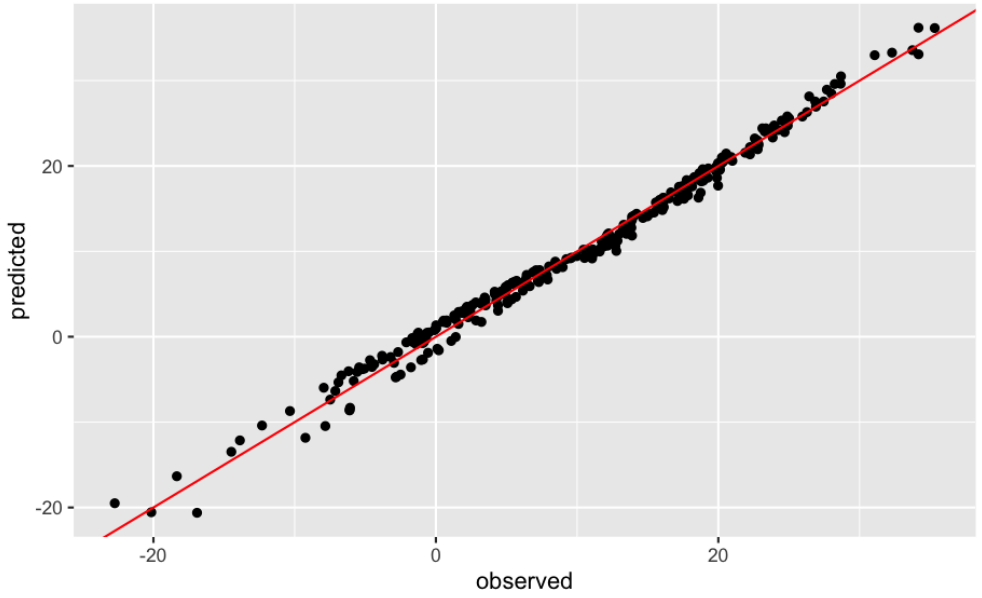

Figure 8: Predicted vs Observed outcomes with Zellner-g Priors

From figure 8 it is clear that both the prior models are giving good results but in the case of Zellner-g, there is a slight improvement, with mean square error = 1.13(approx).

Just like the default prior samples our trace plots are showing that the beta values for all the predictors are moving really well along the mode value.

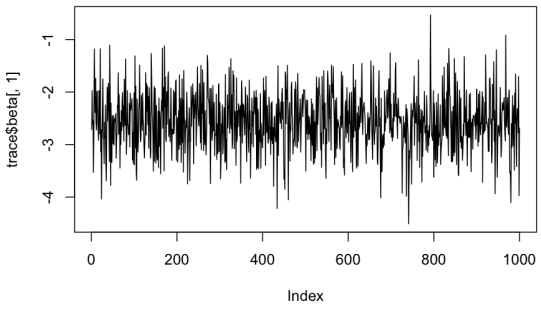

Figure 9.1: Trace plot for beta(intercept)

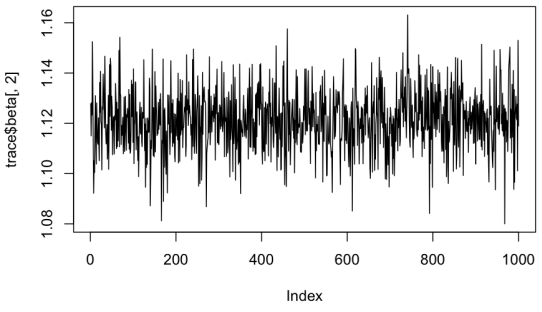

Figure 9.2: Trace plot for beta(temperature)

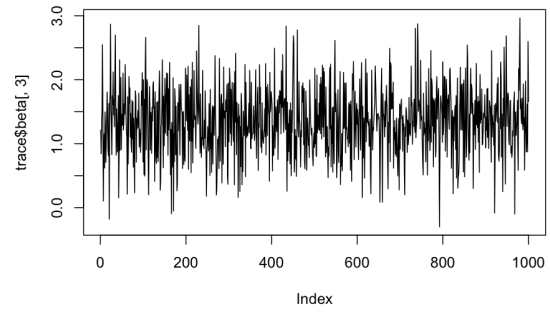

Figure 9.3: Trace plot for beta(humidity)

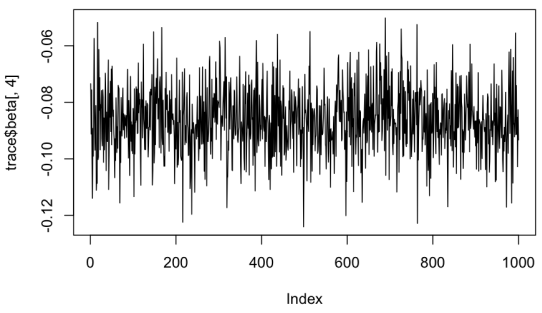

Figure 9.4: Trace plot for beta(wind speed)

In this case also, the auto correlation is approx. zero suggesting our samples are relatively independent.

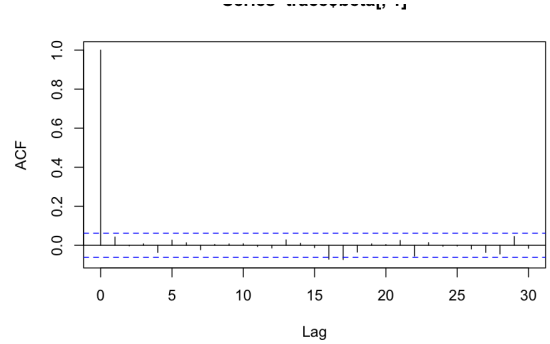

Figure 10.1: ACF of beta(intercept)

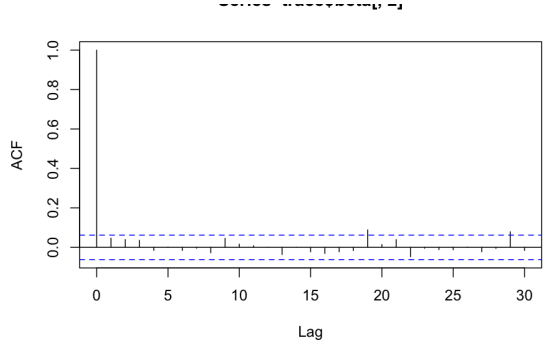

Figure 10.2: ACF for beta(temperature)

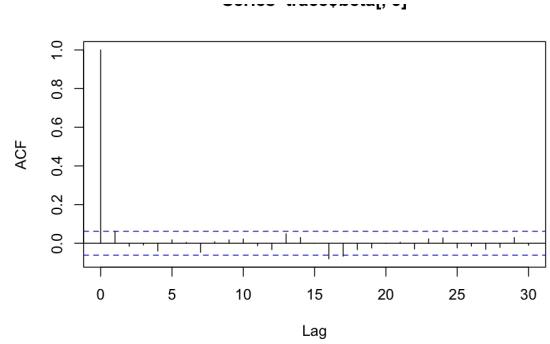

Figure 10.3: ACF of beta(humidity)

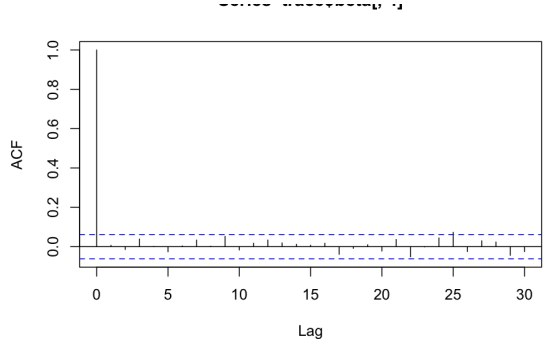

Figure 10.4: ACF of beta(wind speed)

The effective sample size of intercept, temperature, humidity, and wind speed respectively for this model.

```
     var1      var2      var3      var4
1000.0000  909.7779  885.0165 1000.0000
```

Figure 11: Effective sample size for each variable

These models are without interaction of the predictive variables, the interaction between all the relevant variables is also checked if it helps to improve the accuracy of our regression model.

Default (with Interaction):

For default prior (with interaction) between predictors, the graph below suggests that temperature, wind speed and interactions between temperature, humidity and windspeed are strongly predictive

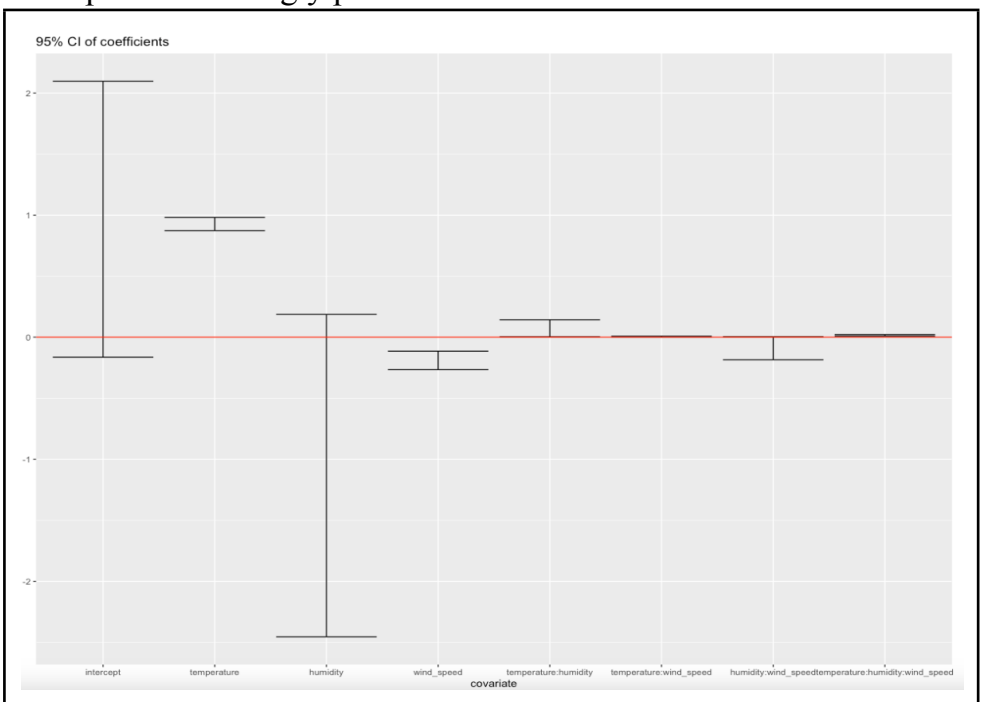

Figure 12: 95% CI of coefficients

The predicted vs observed graph below suggests an improvement compared to the model with default priors where the interaction of predictors was not considered.

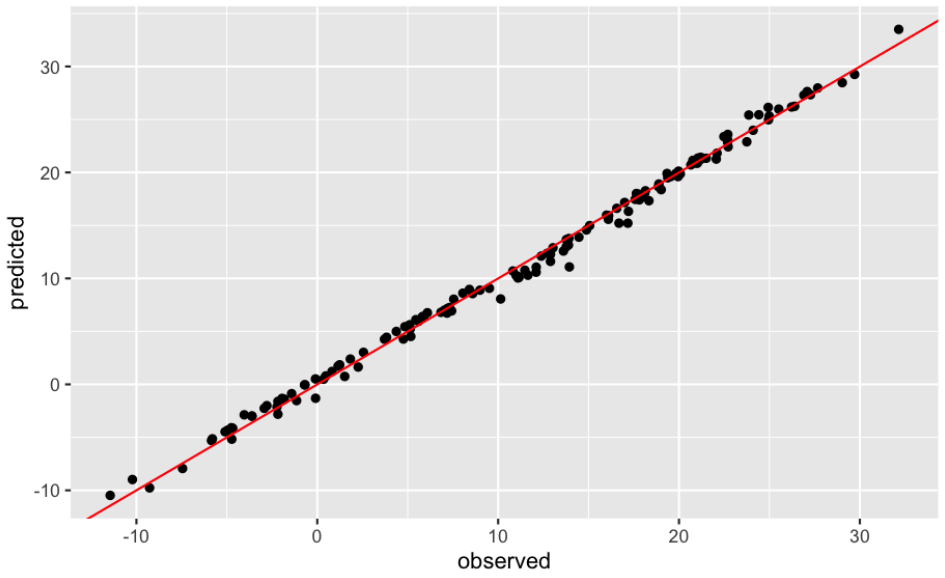

Figure 13: Predicted vs Observed

The error rate, in this case, is 0.45 which is very much lesser than the error rate without interaction.

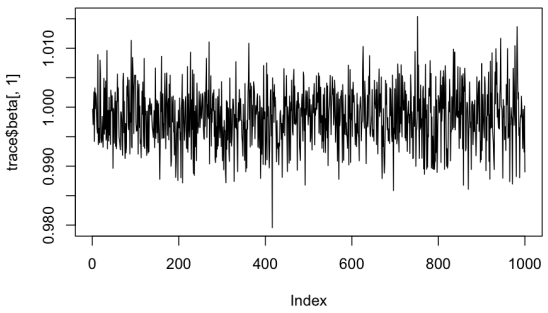

Figure 14.1: Trace Plot of beta(temperature)

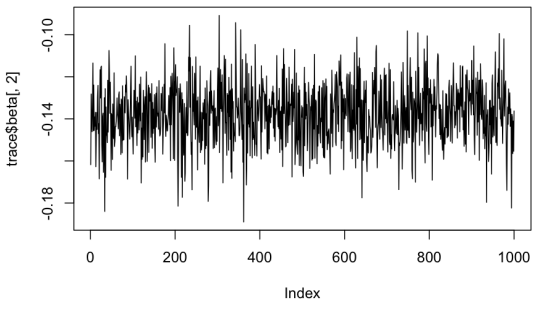

Figure 14.2: Trace Plot of beta(wind speed)

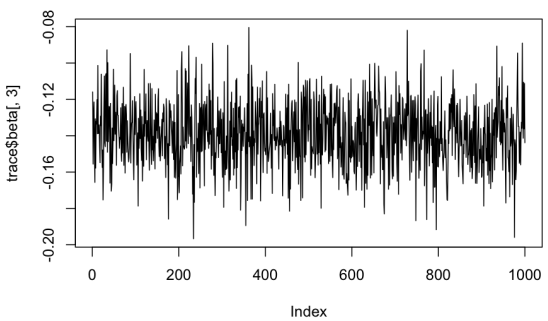

Figure 14.3: Trace Plot of beta(humidity x wind speed)

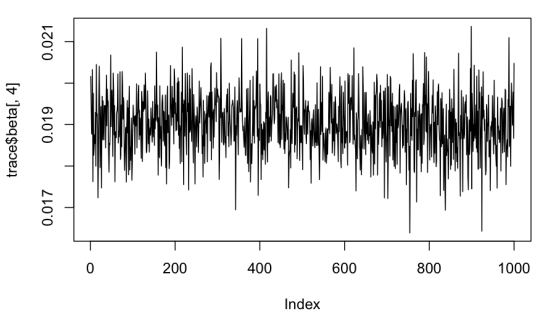

Figure 14.4: Trace Plot of beta(temperature x humidity x wind speed)

Just like the cases with no interaction both trace plots and ACF suggests that the data is well sampled and good for conducting our analysis.

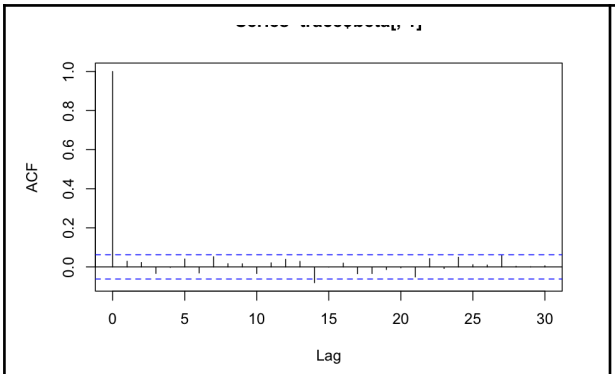

Figure 15.1: ACF of beta(temperature)

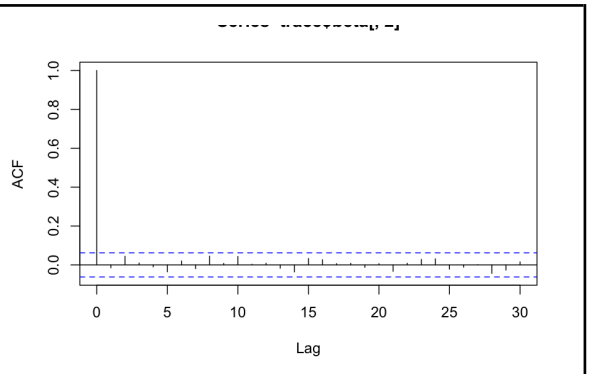

Figure 15.2: ACF for beta(wind speed)

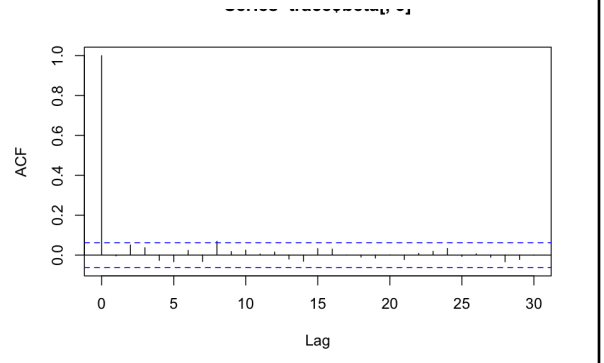

Figure 15.3: ACF of beta(humidity x wind speed)

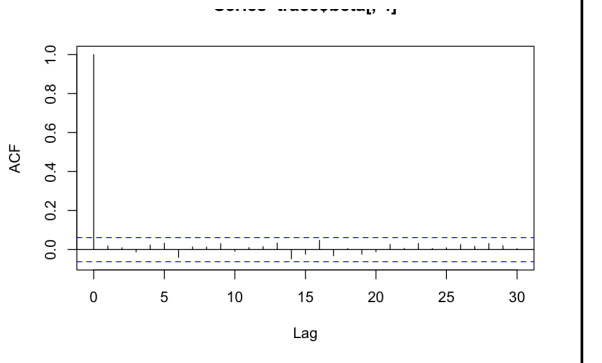

Figure 15.4: ACF of beta(temperature x humidity x wind speed)

The effective sample size of intercept, temperature, humidity, and wind speed respectively, for the model with default prior and interaction between predictors.

```
var1 var2 var3 var4
1000 1000 1000 1000
```

Figure 16: Effective sample size for each variable

Zellner-g Prior(with Interaction):

The graph below suggests that temperature and the interactions between temperature, humidity and windspeed are strongly predictive.

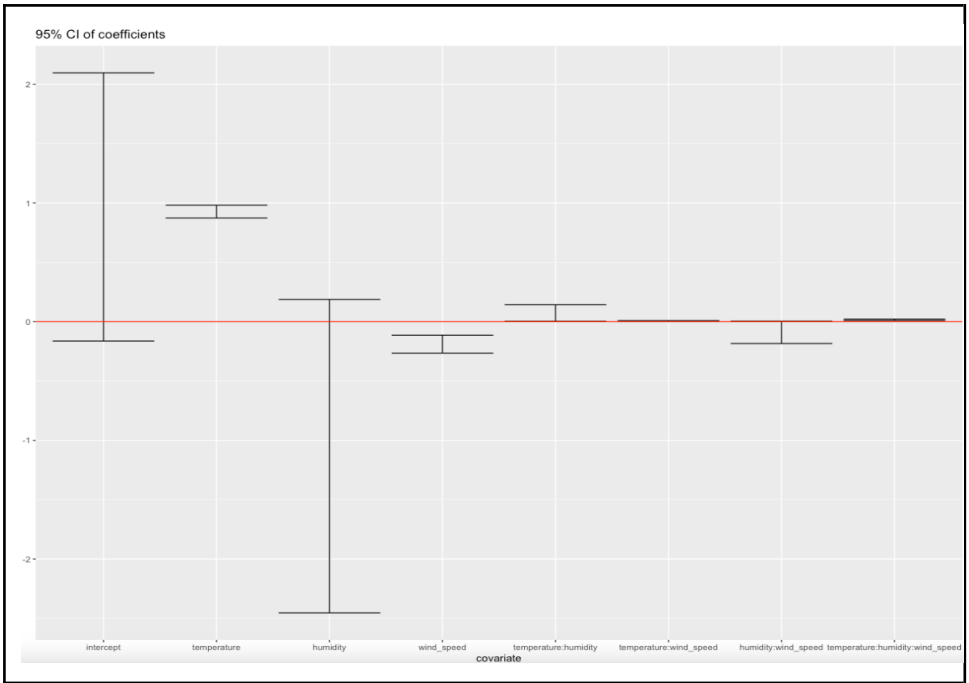

Figure 17: 95% CI of Coefficients

The predictive vs observed outcomes for this model lies almost on the straight line and our predictions from this model are the best among all others.

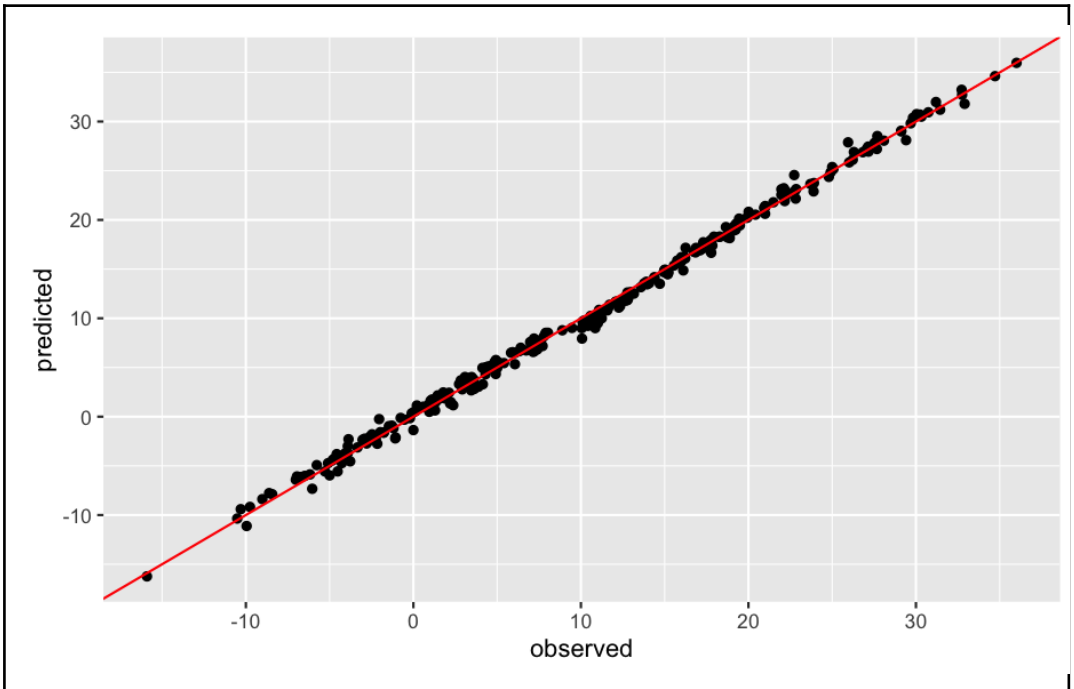

Figure 18: Predicted vs Observed points

The error rate, in this case, is 0.39, which is even lesser than the model of default prior with interaction between variables.

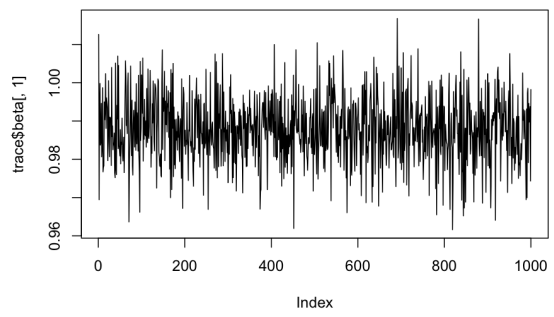

Figure 19.1: Trace Plot of beta(temperature)

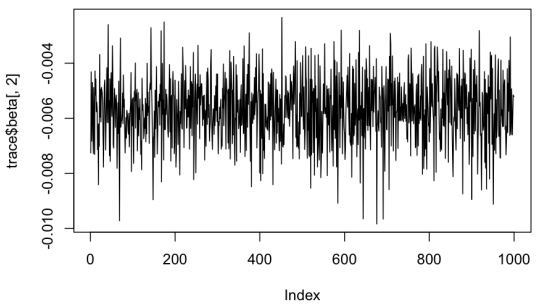

Figure 19.2: Trace Plot of beta(temperature x wind speed)

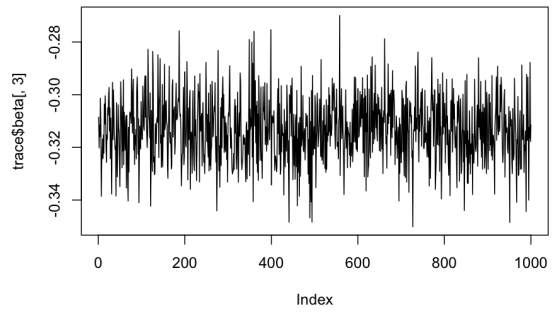

Figure 19.3: Trace Plot of beta(humidity x wind speed)

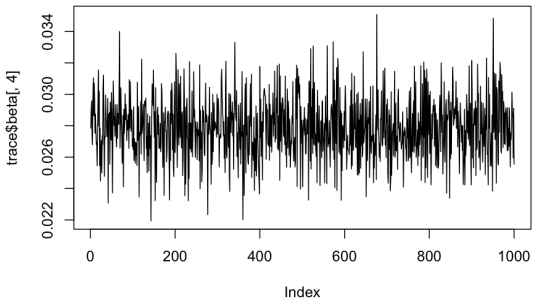

Figure 19.4: Trace Plot of beta(temperature x humidity x wind speed)

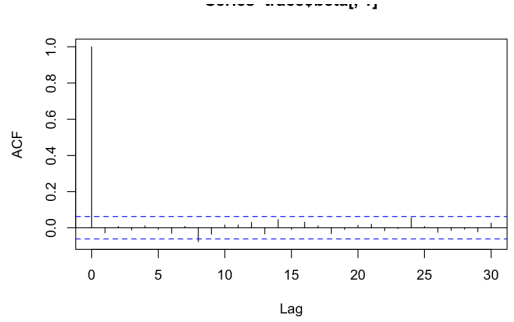

Figure 20.1: Trace Plot of beta(temperature)

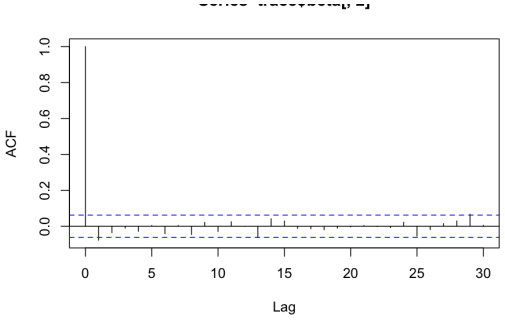

Figure 20.2: Trace Plot of beta(temperature x wind speed)

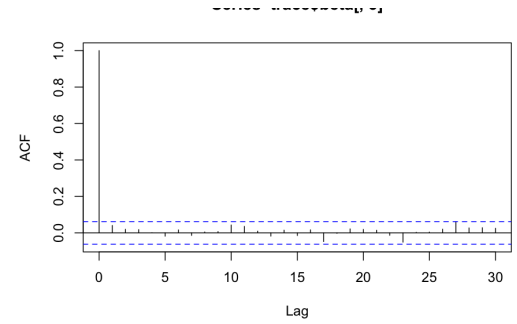

Figure 20.3: Trace Plot of beta(humidity x wind speed)

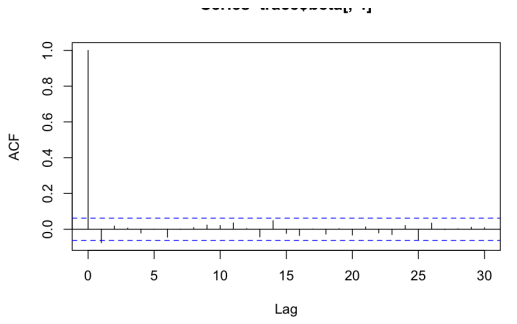

Figure 20.4: Trace Plot of beta(temperature x humidity x wind speed)

The effective sample size of intercept, temperature, humidity, and wind speed respectively, for the model with zellner-g prior and interactions amongst predictors.

```
    var1     var2     var3     var4
1000.000 1167.297 1000.000 1161.966
```

Figure 21: Effective sample size for each variable

CONCLUSION

In this paper we discussed the how useful uninformative priors can be and we also compared the results of both type of priors and which is one is best suited for this case. With model selection we also discussed the important predictors for different priors and how much there interaction could affect predicting the apparent temperature.

We also discussed performance of both the priors, i.e. default and zellner-g and different in results is not that huge to conclude which one is better.
From our analysis, we were able to arrive at the following conclusions:

- Results are much better when the interaction between predictors is considered.
- For the model with default priors and without interaction between predictors, only temperature and wind speed are used to create the best model but in the case of the model with default priors and with the interaction between predictors wind speed, temperature, and humidity gave the best model.
- Samples for default priors were more efficient compare to zellner-g priors without interaction
- Sample efficiency improved for zellner-g prior when we considered interactions between explanatory variables

FUTURE SCOPE

Since, our dataset contains a column for date and time, one obvious idea that strikes is to apply time series analysis. A model can be created to predict the temperature and apparent temperature based on the temperature and an apparent temperature of the previous 10 years of data, for this size of data a very accurate time series model can be developed.

Even though our model is performing well given that we are using uninformative prior, some other kinds of models like the Poisson-gamma model can be built instead of the regression model to get the prediction values for apparent temperature.

FURTHER QUESTIONS

Our model suggests that apparent temperature is mainly dependent on wind speed, and humidity apart from temperature. This study does raise some questions that we would like to investigate further:

- Can there be any other variables apart from wind speed and humidity which is not present in the dataset if we want to build a generic model for apparent temperature?
- Would it be helpful to create a time series-based bayesian model to predict apparent temperatures over time?
- Is this weather pattern localized to Szeged or can it be generalized to other regions in Hungary?
- Can we use a more informative prior to predicting weather patterns when we have fewer data?

STATEMENT OF SUBJECT MATTER IMPLICATION

Our analysis shows that we can successfully apply bayesian analysis to predict weather patterns even from an uninformative prior.