Solutions to BDA Assignment 1, 2020/2021 Semester 2

Theodoros Ladas, s2124289

March 2, 2021

1)

(a) Explanation: A state-space model is built in this assignment in order to model the population (log-population) of whales, with data from 1952 up to 1997. The model is divided into the state model $x_t = bx_{t-1} + u + w_t; \quad w_t \sim N(0, \sigma^2)$, and an observation model $y_t = x_t + v_t, \quad v_t \sim N(0, \eta^2)$. Here the meaning of the parameters of the model will be discussed, specifically, b, u, σ^2, η^2.

The b parameter is the coefficient that connects the previous log-population, to the current one. In other words, if there is no noise on our data, therefore the relationship being completely deterministic, the log-population of the whales would change from $t = i$ to $t = i + 1$ by b.

The u parameter is the intersect of the state model, and its interpretation is: u would be the log-population of whales at time $t = 0$.

The σ^2 parameter is the parameter of the variance of random noise of the underlying state model described above.

Finally, the η^2 parameter has a similar explanation to the σ^2 parameter as it also is the parameter of variance of random noise, but this time it represents the noise of the observation equation. That means that even if our underlying model was deterministic, there is also an extra uncertainty in our model, because of the observation error that might occur.

The point of modeling the log-population of whales with this state-space model is that now, we can have priors on all these parameters, that are going to be updated from our dataset in later stages.

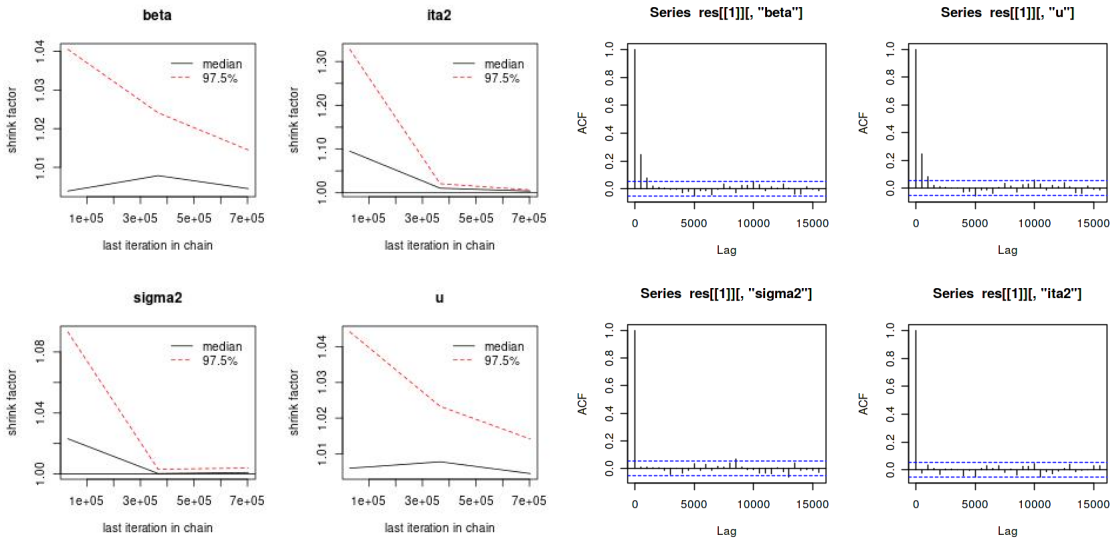
(b) Explanation: The dataset had some missing values that needed to be addressed before continuing with building the engine of the model. First of all the missing data came in two different ways, first of all, there were years in the dataset where the popuplation observed was reported as NA, and secondly, there were missing years from the data, indicating that the observed population for that year, is NA. On the first group, no additional imputation step was taken, because the model was going to be evaluated using JAGS (Just Another Gibbs Sampler). JAGS, treats these kinds of NAs as stochastic nodes, meaning that they are another part of our model. On the other hand, the second kind of NAs, was treated by injecting into the dataset the years that were completely missing along with an NA for the observed population value.

The priors for the model are stated here. It was assumed that $x_0 \sim N(\log(2500), 1)$, $b \sim U(0, 1)$, $u \sim \exp(1)$ and finaly, σ^2, η^2 inv-Gamma$(0.1, 0.1)$

The burn-in period for which the data are discarded for computing estimated of these parameters, was selected to be $2,000$ iterations, and the whole simulation run for $700,000$ iterations.

(c) Explanation: In order to check the mixing of the chains, the Gelman-Rubin statistic was calculated. The Diagrams below, show this statistic per iteration. One indicates that there is no evidence that shows a problem of the mixing of the chains. We can see that on iteration $700,000$ the statistic is either at one, of very close to it (1.01) for all the parameters of the model, so this is a good sign. Also, a second way that the mixing of the chains was evaluated, was by the autocorrelation plots. These are shown below as well per parameter. What we can observe is that for the first iteration the autocorrelation is 1, which is expected since we are using a Markov-Chain model. However, either on the second lag or very quickly after that the autocorrelation of the point explored by the algorithm, versus the previous one, goes to 0 and stays there. That is also a very good indication that the algorithm is not 'stuck' on a small space of the parameter space and instead explores the whole space.

Results:



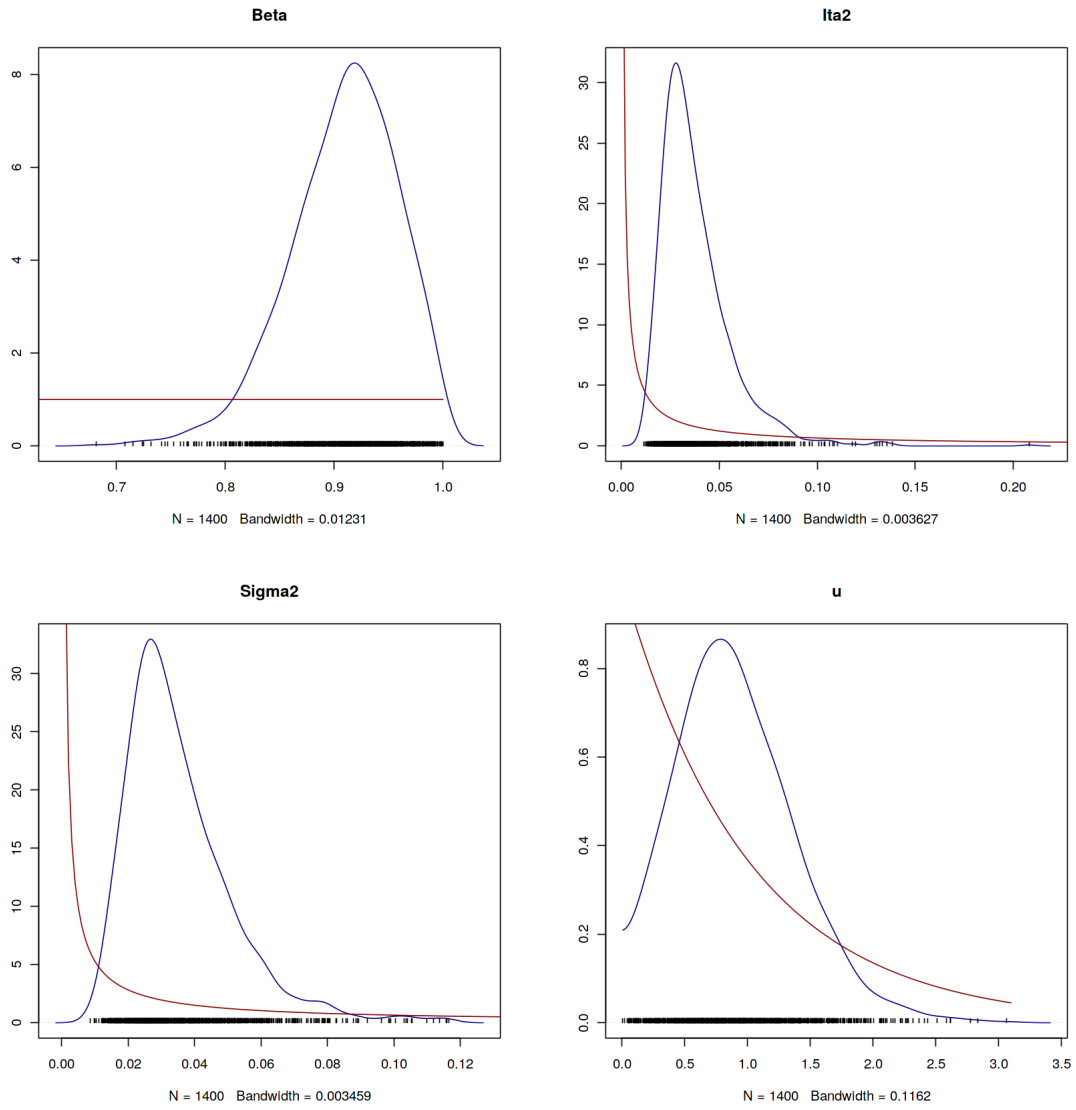
(d) Explanation: Here the results from the above simulation are presented.

```
Iterations = 3500:703000
Thinning interval = 500
Number of chains = 3
Sample size per chain = 1400

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean      SD  Naive SE Time-series SE
beta    0.90627 0.04964 0.0007659      0.0011033
ita2    0.03803 0.01769 0.0002730      0.0002822
sigma2 0.03609 0.01768 0.0002729      0.0002871
u       0.93320 0.47055 0.0072608      0.0104600

2. Quantiles for each variable:

          2.5%     25%     50%     75%   97.5%
beta    0.79885 0.87510 0.91058 0.94284 0.98938
ita2    0.01540 0.02571 0.03410 0.04613 0.08304
sigma2 0.01452 0.02433 0.03231 0.04326 0.07766
u       0.14655 0.58468 0.89420 1.22722 1.95804
```

We can see that the mean and standard deviation values for all the parameters, as well as confidence intervals for each parameter.

Below the plot of the posterior densities and their prior densities are also plotted. From these diagrams we can clearly see that the prior is not dominating the posterior in any case, so our choice of prior densities and hyperparameters is reasonable. In order to further robustify the method, later a sensitivity analysis, will be performed.

Results:

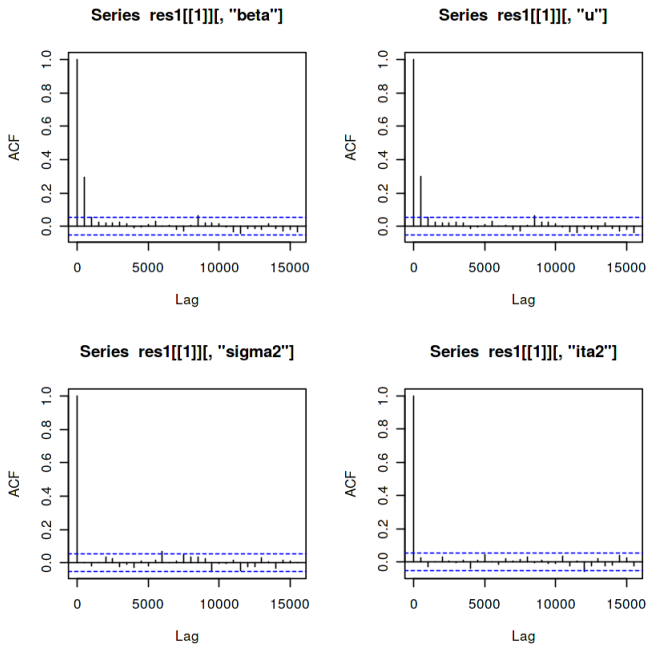


(e) Explanation: Here the sensitivity analysis is discussed. The priors were chosen as follows. The distribution for the prior of each parameter stayed the same and the choice of starting hyperparameters was vastly different, in order to see whether the results will change or not. More specifically, $b \sim U(-1, 2)$, $u \sim \exp(3)$ and finaly, σ^2, η^2 inv-Gamma$(1/0.1, 1/0.1)$. The prior for x_0 stayed the same.

```
Iterations = 3500:703000
Thinning interval = 500
Number of chains = 3
Sample size per chain = 1400

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

          Mean      SD  Naive SE Time-series SE
beta   0.90816 0.04842 0.0007472      0.0010115
ita2   0.03886 0.01831 0.0002826      0.0002826
sigma2 0.03619 0.01761 0.0002717      0.0002717
u      0.91439 0.45794 0.0070662      0.0096503

2. Quantiles for each variable:

          2.5%     25%     50%     75%   97.5%
beta   0.79962 0.88022 0.91213 0.94281 0.98807
ita2   0.01567 0.02571 0.03407 0.04579 0.08501
sigma2 0.01456 0.02450 0.03249 0.04322 0.08215
u      0.15977 0.58298 0.87057 1.17453 1.94474
```

As we can see the results changed to the fourth significant digit. This clearly indicates that the prior doesn't dominate the posterior and it makes it even more likely that the posterior density is converging to the true value of the parameters. The autocorrelation plot figure for this model is also present in the results graph, used as in the first model to check the mixing of the new chains.

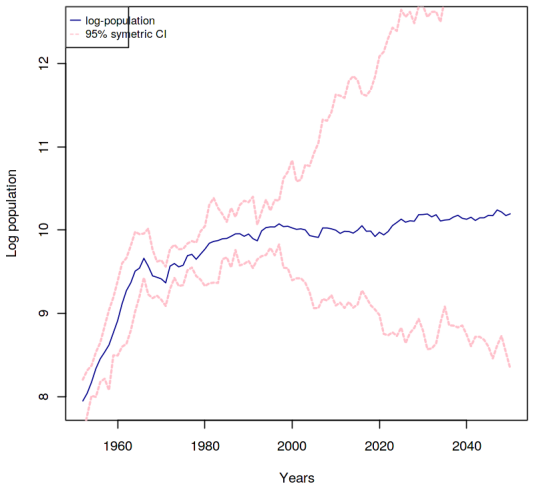
Results:



(f) Explanation: On the below graph, the evolution of the posterior mean of log-population of whales is presented. This line is projected to 2050. The dotted lines show the 95% Confidence intervals (the lower line is the 2.5% and the upper line the 97.5%). We see that the mean gets

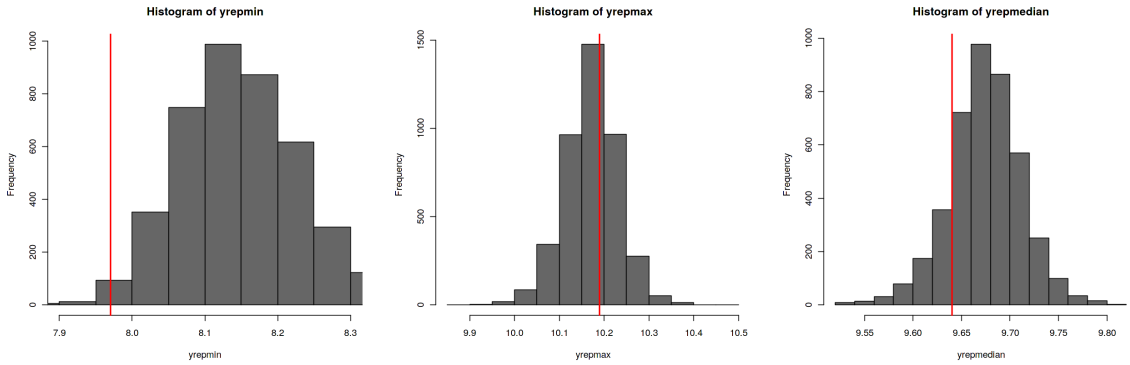
stable after a while, which is what we expected since the β coefficient was around 0.9. A β of 1 indicates a stationary process in the AR(1) model. We also see the 95% CI for the mean to diverge very heavily. This is because our last observed datapoint is in 1997 and we are trying to predict 2050. This divergence explained because the state-model has a random noise and the observed values model has another small random noise. All these 'mistakes' add up, making the projection for 2050 have a very big variance associated with it. Also the posterior probability that the population of gray whales becores smaller than 100 at any year from 1951 until the end of 2050 was calculated to be 0. Or mathematically, $p(\min_{t\in\{0,1...,99\}} x_t <= \log(100)|y) = 0$

Results:



(g) Explanation: Finally, posterior predictive checks for various statistics, specifically, the min, the max and the mean, to evaluate the fit of this model, where the priors are the original ones stated in section (b) was performed. The results are reported below in the form of histograms of the posterior predictive samples generated from the model. The red lines indicate the min, the max and the mean of the given sample that we observed respectively. We can clearly see that in all cases the line is well within the bounds of the histograms. The most extreme case is that of the min, but even this is still acceptable as it is within 2σ from the mean of its distribution.

Results:



2)

(a) Explanation: In this problem, the aim is to model house prices per unit area from recent transactions in Taipei. Our dataset consists of 414 rows of 6 features (house age, distance from nearest MRT (metro) station in meters, number of convenience stores within walking distance, latitude, longitude and the target of the price). First of all, the data matrix was centered. That means that for every feature, its mean has been subtracted in order to shift their distribution to have a zero mean, and then each value has been divided by their standard deviation in order to make every feature to have a unit variance. Then since we will again work on the log scale, the target variable of the price has been converted to log-price. Then a standard linear regression was fit, with the defaults lm function of R. This was done to have a benchmark. Later, more complicated models can be fit and their quality of fit can be compared to the standard linear regression.

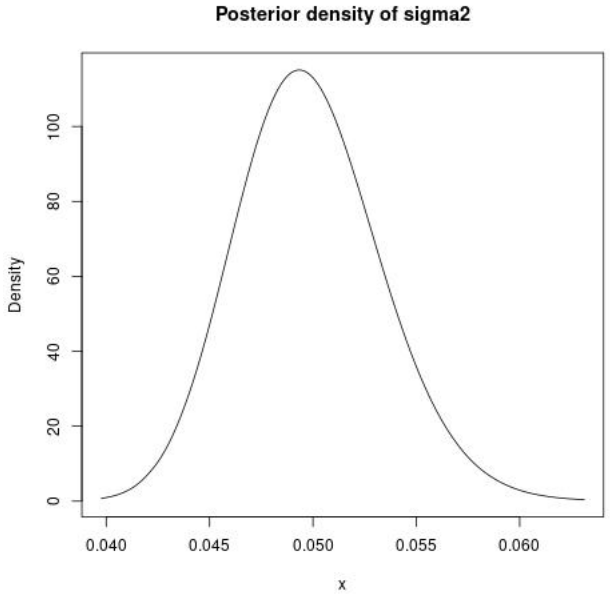
Results:

```
Call:
lm(formula = log(house.c$Y.house.price.of.unit.area) ~ house.c$X1.transaction.date +
    house.c$X2.house.age + house.c$X3.distance.to.the.nearest.MRT.station +
    house.c$X4.number.of.convenience.stores + house.c$X5.latitude +
    house.c$X6.longitude)

Residuals:
     Min       1Q   Median       3Q      Max
-1.68095 -0.11498 -0.00267  0.11540  1.04849

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               3.566695   0.010892 327.447  < 2e-16 ***
house.c$X1.transaction.date               0.038198   0.010985   3.477 0.000561 ***
house.c$X2.house.age                     -0.079373   0.010983  -7.227 2.46e-12 ***
house.c$X3.distance.to.the.nearest.MRT.station -0.183625   0.022675  -8.098 6.54e-15 ***
house.c$X4.number.of.convenience.stores   0.081731   0.013868   5.894 7.94e-09 ***
house.c$X5.latitude                       0.098348   0.013839   7.107 5.36e-12 ***
house.c$X6.longitude                      0.005659   0.018656   0.303 0.761766
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2216 on 407 degrees of freedom
Multiple R-squared:  0.6857,    Adjusted R-squared:  0.6811
F-statistic:   148 on 6 and 407 DF,  p-value: < 2.2e-16
```

(b) Explanation: For the purposes of this problem, an Integrated Nested Laplace Approximation (INLA) model was chosen. The posterior of the model, was a Gamma$(0.1, 0.1)$ and the Gaussian prior with mean zero and variance $1e6$ was chosen in order to represent prior ignorance of the value of the parameters of the model. The mean standard error of the fit was 4.47. More summary statistics are presented in the Results section. This model, immediately had a very positive effect on the standard deviation of the mean residuals of the model, since it is 0.22. The negative-sum log of the conditional predictive ordinate (NSLCPO) was also produced and it is -28.78.

Finally, a plot of the posterior density for the variance parameter σ^2 is also shown below. The figure shows that the variance parameter σ^2 has a mean value of around 0.05 with very high confidence, since we can see that most of the density on the diagram is on the interval $\{0.04, 0.06\}$. So we are fairly sure that a small, non-zero variance exists in the noise parameter of the model.

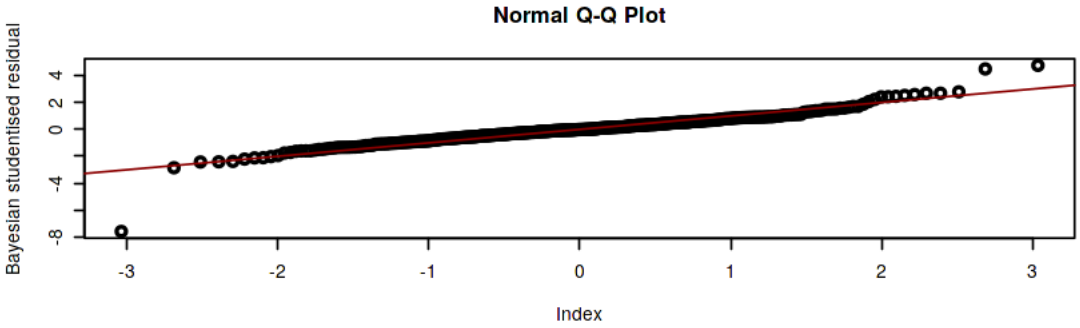
Results:

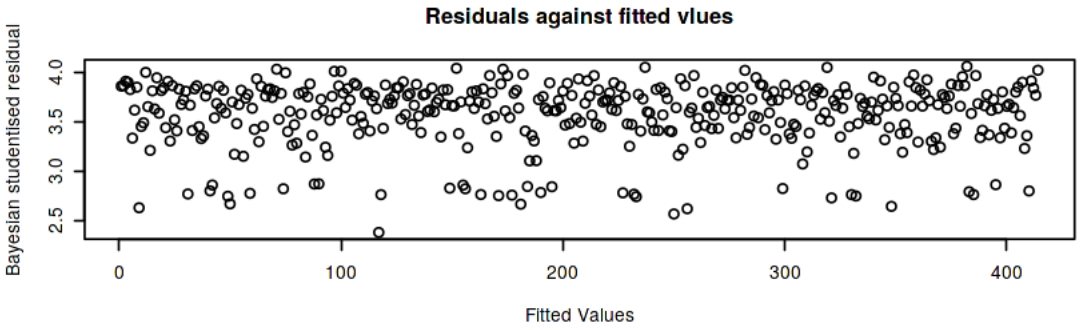


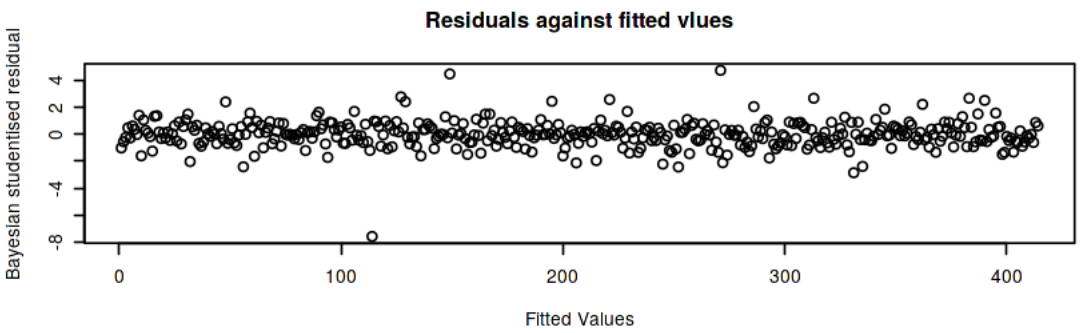
(c) Explanation: Next we evaluate the fit of the model, by producing multiple plots for the residuals of the fit. On the first diagram we see the normal quantile-quantile plot for the Bayesian studentized residuals of the fit. In red we see the line of 45 degrees. We can observe that almost all the residuals fall onto this line, and that means that

the distribution of the residuals is not far from a Normal distribution. So our error term has a Normal distribution and our assumption of homoskedasticity holds. On the next diagrams we see the Bayesian studentized residuals against the fitted values. This plot was done in order to check if any patterns are apparent. That would indicate that there is some correlation between the residuals and the fitted values which is another important assumption of our linear model. It is clear that no apparent pattern emerges from this diagram so we can continue with our analysis.

Results:

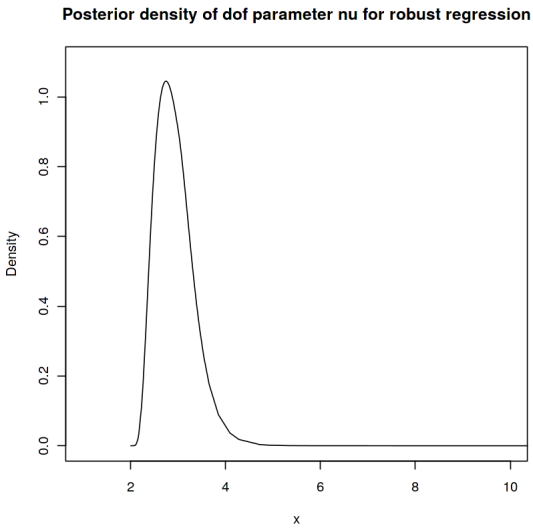


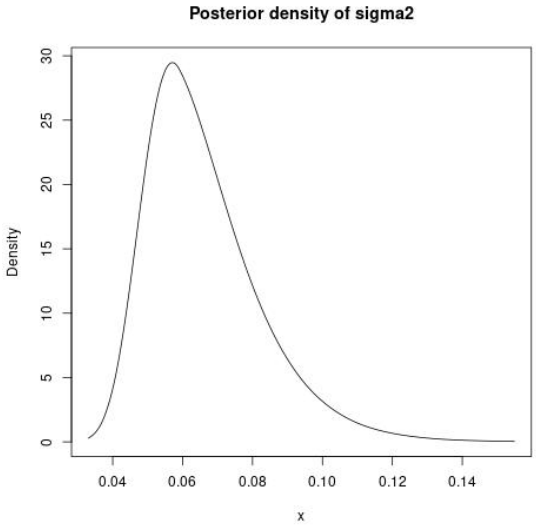




(d) Explanation: Next, a robust to outliers version of the model was considered. This has exactly the same priors as the original model described in (b). The NSLCPO score of the model was -77.9, which is lower than the first model, indicating a better fit. The corresponding DIC value was -157.09 and the standard deviation of the mean residuals was 0.222 which also indicated a better fit. In the results section, there are two posterior density graphs, one for the degrees of freedom and another of σ^2 as before. We can clearly see that the mean value for the degrees of freedom is around 3, which indicated that there are some outliers that skew the Student-t distribution. Lastly, we see a small increase in the mean variance. This is expected, as we are not using the information of some data points due to the robust regression. The summary statistic is also included below.

Results:





```
Time used:
    Pre = 0.322, Running = 3.23, Post = 0.084, Total = 3.63
Fixed effects:
             mean    sd 0.025quant 0.5quant 0.975quant   mode kld
(Intercept)  3.564 0.009      3.547    3.564      3.580  3.564   0
x1           0.023 0.008      0.007    0.023      0.040  0.023   0
x2          -0.089 0.009     -0.106   -0.089     -0.071 -0.089   0
x3          -0.226 0.020     -0.264   -0.226     -0.186 -0.226   0
x4           0.073 0.012      0.050    0.073      0.097  0.073   0
x5           0.089 0.012      0.067    0.089      0.112  0.089   0
x6          -0.011 0.014     -0.038   -0.011      0.017 -0.011   0

Model hyperparameters:

Expected number of effective parameters(stdev): 7.08(0.006)
Number of equivalent replicates : 58.43

Deviance Information Criterion (DIC) ...............: -157.09
Deviance Information Criterion (DIC, saturated) ....: -17132.73
Effective number of parameters .....................: 9.05

Marginal log-Likelihood:  0.684
CPO and PIT are computed

Posterior marginals for the linear predictor and
 the fitted values are computed
```

d

(e) Explanation: The next step was to consider our variable x_4 (number of convenience stores within walking distance) as a categorical variable with multiple factors. We then fitted the two models (simple and robust) with the new modified variable in order to check whether this change will affect the model in a positive way.

NSLCPO simple: -28.02

NSLCPO robust: -84.45

MSE simple: 0.2155

MSE robust: 0.2172

It is clear from the NSLCPO score, that the best model so far, is the one with x_4 as a categorical variable and with robust regression as this produces the lowest overall NSLCPO score. Lastly, the Mean standard deviation of the mean residuals of both models is lower than the previous one. The non-robust model seems to have a bit lower MSE, but this is obvious only on the third significant digit.

Below you can find summary statistics for both models.

Results:

```
house.fac.x42  -0.032 0.056     -0.142   -0.032      0.077 -0.032   0
house.fac.x43  -0.039 0.046     -0.130   -0.039      0.052 -0.039   0
house.fac.x44   0.046 0.053     -0.058    0.046      0.150  0.046   0
house.fac.x45   0.151 0.045      0.061    0.151      0.240  0.151   0
house.fac.x46   0.181 0.052      0.079    0.181      0.282  0.181   0
house.fac.x47   0.181 0.054      0.074    0.181      0.287  0.181   0
house.fac.x48   0.224 0.056      0.115    0.224      0.334  0.224   0
house.fac.x49   0.243 0.059      0.128    0.243      0.357  0.243   0
house.fac.x410  0.185 0.081      0.026    0.185      0.345  0.185   0
x5              0.110 0.015      0.081    0.110      0.139  0.110   0
x6             -0.020 0.020     -0.059   -0.020      0.019 -0.020   0

Model hyperparameters:

Expected number of effective parameters(stdev): 16.05(0.004)
Number of equivalent replicates : 25.79

Deviance Information Criterion (DIC) ...............: -62.03
Deviance Information Criterion (DIC, saturated) ....: 428.80
Effective number of parameters .....................: 17.19

Marginal log-Likelihood:  -125.68
CPO and PIT are computed

Posterior marginals for the linear predictor and
 the fitted values are computed
```

```
house.fac.x42  -0.048 0.046     -0.137   -0.048      0.042 -0.048   0
house.fac.x43  -0.055 0.039     -0.130   -0.055      0.022 -0.055   0
house.fac.x44   0.002 0.045     -0.085    0.001      0.091  0.000   0
house.fac.x45   0.133 0.041      0.052    0.133      0.214  0.133   0
house.fac.x46   0.175 0.047      0.083    0.175      0.266  0.175   0
house.fac.x47   0.159 0.048      0.064    0.159      0.254  0.159   0
house.fac.x48   0.195 0.048      0.101    0.195      0.290  0.194   0
house.fac.x49   0.184 0.050      0.084    0.184      0.283  0.184   0
house.fac.x410  0.163 0.059      0.047    0.162      0.279  0.162   0
x5              0.104 0.013      0.079    0.103      0.131  0.103   0
x6             -0.036 0.015     -0.066   -0.036     -0.007 -0.036   0

Model hyperparameters:

Expected number of effective parameters(stdev): 16.09(0.006)
Number of equivalent replicates : 25.73

Deviance Information Criterion (DIC) ...............: -171.49
Deviance Information Criterion (DIC, saturated) ....: -20395.90
Effective number of parameters .....................: 17.92

Marginal log-Likelihood:  -75.04
CPO and PIT are computed

Posterior marginals for the linear predictor and
 the fitted values are computed
```

(f) Explanation: Lastly a non linear in feature model was considered to explore the possibility of non-linear relationship in the data. The model that we considered is the following:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_6 x_6 + \hat{\beta}_7 x_1^2 + \hat{\beta}_8 x_2^2 + \hat{\beta}_8 x_3^2 + \hat{\beta}_9 x_5^2 + \hat{\beta_{10}} x_6^2$$

Essentially taking into consideration a relationship of the second degree.

NSLCPO transformed: -112.82

DIC transformed: -231.32

MSE transformed: 0.203

The NSLCPO score shows a significant improvement over all other models, which is also apparent on the MSE score of the model. Summary statistics are presented below. At this point, a very important caveat needs to be mentioned. Up until now, we are only considering in sample *goodness of fit* measures. That means that there is a chance that we have overfitted our model to our observed data. In order to check that we could compute all the above quantities on a held-out set, or even perform a k-fold cross validation. This is left for further development of the model.

Results:

```
house.fac.x47   0.151 0.048      0.057    0.151      0.246  0.150   0
house.fac.x48   0.198 0.046      0.108    0.197      0.290  0.196   0
house.fac.x49   0.155 0.050      0.059    0.155      0.253  0.154   0
house.fac.x410  0.199 0.057      0.087    0.199      0.312  0.199   0
x5              0.119 0.014      0.091    0.119      0.148  0.119   0
x6              0.064 0.025      0.015    0.064      0.114  0.064   0
I(x1^2)         0.010 0.009     -0.008    0.010      0.027  0.010   0
I(x2^2)         0.039 0.009      0.022    0.039      0.057  0.039   0
I(x3^2)         0.124 0.024      0.077    0.124      0.170  0.124   0
I(x5^2)        -0.063 0.018     -0.098   -0.063     -0.026 -0.063   0
I(x6^2)        -0.143 0.041     -0.223   -0.143     -0.063 -0.143   0

Model hyperparameters:

Expected number of effective parameters(stdev): 21.10(0.006)
Number of equivalent replicates : 19.62

Deviance Information Criterion (DIC) ...............: -231.32
Deviance Information Criterion (DIC, saturated) ....: -17968.70
Effective number of parameters .....................: 22.91

Marginal log-Likelihood:  -98.67
CPO and PIT are computed

Posterior marginals for the linear predictor and
 the fitted values are computed
```

(g) Explanation: Finally, using our original model from part (b), we will predict the prices for the three next years. Firstly 3 new rows were created and bound to the dataset with NA on the column of the price for prediction. The values for the features were the average of the

columns, except for the transaction date column, which was 2014, 2015 and 2016 for the 3 rows (the years we want to predict) and the house age, which was calculated as 2020-(transaction date). The mean values the three years are: 2938.587, 2943.112, 2947.637. These prices are not correct however, and further root cause analysis needs to be conducted in order to understand why these values have been produced. Below is the final Posterior predictive Density produces by the final model.

Results: