Mixed GWR Simulation Write-up

Aaron Swoboda

October 17, 2013

What are our research questions? Basically:

1. Can we find the "true" model among the eight different possibilities with three model parameters?
2. Are there differences in the results based on the metric used?
3. What happens as we change the sample size and amount of error in the model?
4. How much does it really matter if we are concerned with coefficient estimates?
5. What happens when we use other decision tools to help with model selection? (Monte Carlo simulations and test statistics)

```
## Loading required package:  animation
```

1 Methodology

Imagine a simple linear model with two explanatory variables,

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \epsilon \tag{1}$$

Now imagine that each of our N individual observations in this model occur at geographical locations on a Cartesian plane. Thus our data consists of an $Nx5$ matrix, where Y may be house prices, X_1 and X_2 could be the living space and lot size associated with each house, and the final two columns determine the location of the observations (for instance, latitude and longitude, or distances north and east from a prescribed point serving as the origin).

The above simple model exemplifies spatial stationarity in the parameters. The β coefficients are constant over space. Instead, the coefficients could exhibit spatial non-stationarity, in which case one, two, or all three of the β coefficients are a function of location. This has a natural interpretation in the current real estate example. Location matters. Location can matter in different ways.

For instance, if the value of land varies over space, then we would expect the coefficient on lot size to vary over space, while it is also possible that the intercept varies over space to reflect variation in prices of similar houses in different locations.

While it is possible to parameterize the variation in coefficients, for instance researchers often (CITATION?) include a variable measuring the distance from an observation to an important amenity such as the Central Business District and then this distance variable could be interacted with variables whose value are predicted to vary over space. However, it is not implausible to believe that the variation in coefficients might not be easily parameterized (for instance, if land values are a non-monotonic function of distance). Researchers may instead interact variables with fixed effects for cities or census tracts. However, such strategies require the analyst to make assumptions that severely limit the type and degree of variation in the parameters. For instance, interaction terms with geographic boundaries assume discrete differences in the value of parameters across the boundaries, while instead the parameters may instead be a continuous function of location.

1.1 Local Regression to the Rescue?

1.2 Experimental Design

We generate data in the following format:

$$Y = \beta_0(location) + \beta_1(location) * X_1 + \beta_2(location) * X_2 + \epsilon, \qquad (2)$$

where sometimes the coefficient is in fact stationary, $\beta_i(location) = \beta$, and other times it is non-stationary. With three coefficients each having the possibility of being stationary or not, there are eight different possible combinations, ranging from (stationary, stationary, stationary) to (non-stationary, non-stationary, non-stationary).

We generate data using all eight different combinations and then estimate all eight possible LWR models (assuming non-stationarity or not for each variable). We then calculate different Cross-Validation metrics and compare their values across models and bandwidths.

We have three different values for each coefficient in our DGP, no variation, some variation, and more variation.

We also change the sample size of our data as well as the variance of the model error term.

2 Simulation Results

We have seven different ways to pick the "best" model (the AIC, GCV, SCV, LOOCV, and RMSEs for the three different coefficients). Here are tables showing the relative frequency (in percentage) that each model number was selected

by optimizing a given metric. Note that the columns in the following tables may not sum exactly to 100 due to rounding.

```
for (i in 1:8) {
    temp2 = which(mcOutput[, "True Model", ] == i, arr.ind = TRUE)
    temp3 = mcOutput[8:14, "Model Number", unique(temp2[, 2])]
    temp4 = factor(temp3)
    newdata = data.frame(ModelNum = temp4, Metric = factor(rownames(temp3),
        levels = rownames(temp3)))
    cat(paste("\n true model =", i, "\n"))
    cat("spatial variation...\n")
    print(models[i, ])
    print(round(table(newdata) * 100 * 7/sum(table(newdata)), 0))
}
```

```
##
##  true model = 1
## spatial variation...
##   beta0 beta1 beta2
## 1    no    no    no
##         Metric
## ModelNum AIC GCV SCV LOOCV B0RMSE B1RMSE B2RMSE
##        1   0   0   0    75      6      7      6
##        2  26  26   0     6      5     22     20
##        3  34  33   0     6     21      4     22
##        4   1   1   8     2      5      3     37
##        5  38  38   0     7     20     21      4
##        6   1   1   8     2      5     35      3
##        7   1   1   5     1     32      4      5
##        8   0   0  78     0      5      3      2
##
##  true model = 2
## spatial variation...
##   beta0 beta1 beta2
## 2   yes    no    no
##         Metric
## ModelNum AIC GCV SCV LOOCV B0RMSE B1RMSE B2RMSE
##        1   0   0   0     4      0      4      4
##        2  91  91  37    89     94     29     29
##        3   4   4   0     2      1      6     25
##        4   0   1  16     1      1      1     32
##        5   4   4   0     3      1     24      6
##        6   0   1  16     1      1     33      1
##        7   0   0   3     0      1      2      2
##        8   0   0  28     0      0      1      1
##
```

```
##  true model = 3
## spatial variation...
##   beta0 beta1 beta2
## 3    no   yes    no
##         Metric
## ModelNum AIC GCV SCV LOOCV B0RMSE B1RMSE B2RMSE
##        1   0   0   0    12     13      0     13
##        2  11  11   1     7      1      2     24
##        3  77  76   0    65     23     66     22
##        4   2   2  12     4      4      6     32
##        5   8   8   0     5     24      2      1
##        6   0   0   9     1      2      3      2
##        7   2   2   6     5     29     15      3
##        8   0   0  72     1      4      6      2
##
##  true model = 4
## spatial variation...
##   beta0 beta1 beta2
## 4   yes   yes    no
##         Metric
## ModelNum AIC GCV SCV LOOCV B0RMSE B1RMSE B2RMSE
##        1   0   0   0     2      1      0      6
##        2  70  68  29    65     65      6     29
##        3  10   9   0     8      2     34     25
##        4  16  19  25    21     25     14     34
##        5   2   2   0     2      2      4      2
##        6   0   1  13     1      2      8      1
##        7   0   0   3     1      2     22      2
##        8   0   0  30     0      2     12      1
##
##  true model = 5
## spatial variation...
##   beta0 beta1 beta2
## 5    no    no   yes
##         Metric
## ModelNum AIC GCV SCV LOOCV B0RMSE B1RMSE B2RMSE
##        1   0   0   0    12     13     13      0
##        2  11  10   1     7      1     24      2
##        3   8   8   0     4     24      1      2
##        4   0   1   9     1      1      2      3
##        5  78  77   0    66     23     23     66
##        6   1   2  12     5      5     32      6
##        7   1   2   6     4     30      3     15
##        8   0   0  73     1      4      2      6
##
```

```
##  true model = 6
## spatial variation...
##   beta0 beta1 beta2
## 6   yes    no   yes
##         Metric
## ModelNum AIC GCV SCV LOOCV B0RMSE B1RMSE B2RMSE
##        1   0   0   0     2      1      6      0
##        2  71  68  29    66     65     30      7
##        3   3   2   0     2      2      2      4
##        4   0   0  13     1      2      1      7
##        5  10   9   0     8      2     24     34
##        6  16  19  24    20     25     33     14
##        7   0   1   3     1      2      2     21
##        8   0   0  30     0      2      1     13
##
##  true model = 7
## spatial variation...
##   beta0 beta1 beta2
## 7    no   yes   yes
##         Metric
## ModelNum AIC GCV SCV LOOCV B0RMSE B1RMSE B2RMSE
##        1   0   0   0     5     13      1      1
##        2   8   8   1     6      0      2      2
##        3  25  24   0    18     25     16      2
##        4   1   2   9     3      1     12      4
##        5  25  24   0    18     25      2     17
##        6   2   2   9     3      1      5     12
##        7  39  40   8    44     31     50     51
##        8   1   1  72     3      4     11     11
##
##  true model = 8
## spatial variation...
##   beta0 beta1 beta2
## 8   yes   yes   yes
##         Metric
## ModelNum AIC GCV SCV LOOCV B0RMSE B1RMSE B2RMSE
##        1   0   0   0     1      1      1      1
##        2  60  57  23    53     48      7      7
##        3   6   5   0     5      3     15      5
##        4   9  11  19    12     14     16      8
##        5   6   5   0     5      3      5     16
##        6   9  11  19    11     14      8     15
##        7   3   3   3     4      4     32     33
##        8   6   7  35     8     14     16     15
```

The results of the previous tables must be taken with a grain of salt, as there

are frequently times when a "true" model may include variation in a coefficient, but the degree of non-stationarity in the coefficient may be small. In such cases, choosing an incorrect model (such as one that keeps such a coefficient constant) may not be such a big problem.

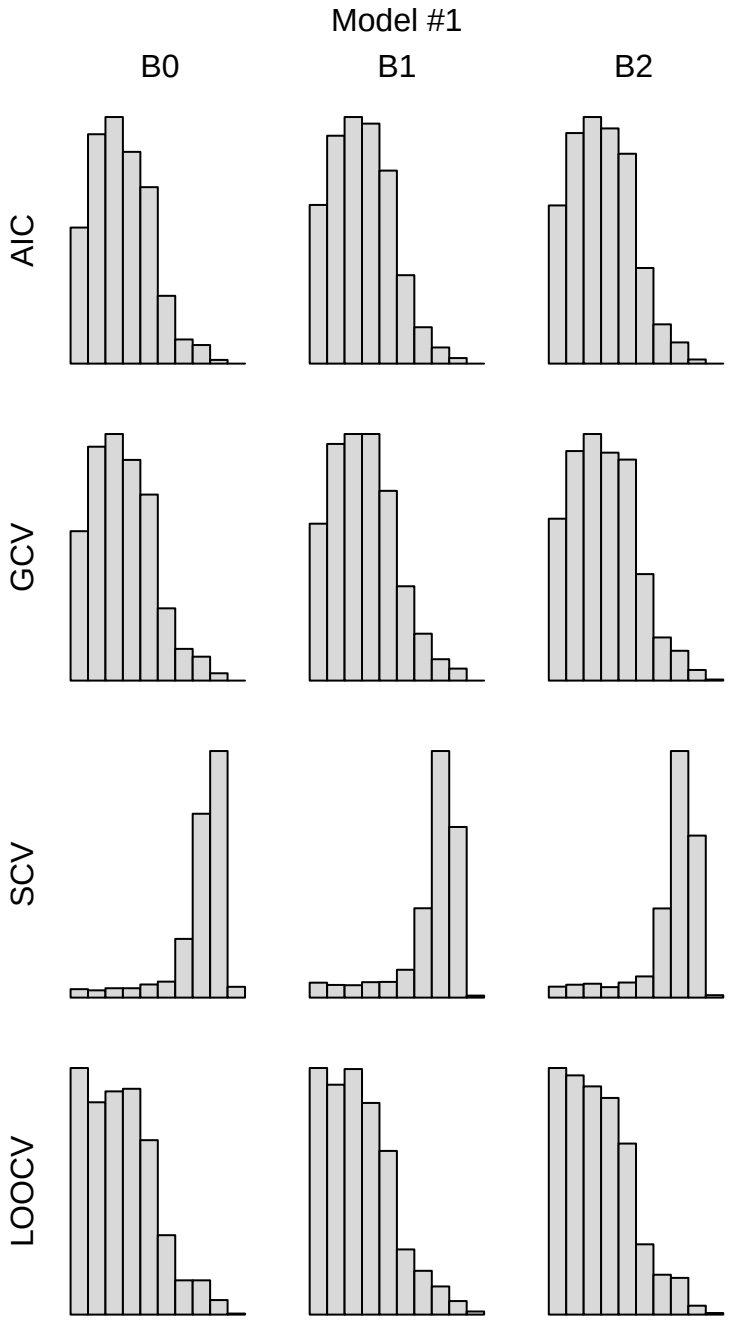
Some patterns clearly emerge.

- The AIC and GCV metrics *never* select Model 1, even when it is the actual model.
- There are several occasions where the model/bandwidth combination with the smallest RMSE is not the "correct" model.

2.1 Coefficient Formulation

Even if an incorrect model is chosen, the model may yield accurate estimates of the coefficients. In each run of the simulation we estimated 50 different model/bandwidth combinations (seven different bandwidths for each of the seven models with at least one coefficient varying over space, plus the standard OLS model). We calculate the Root Mean Squared Error for each model and can rank these values. For instance, it is possible, and often the case, that the "wrong" model yields the most accurate estimates of a coefficient.

```
colMat = matrix("black", 8, 3)
colMat[as.matrix(models) == "yes"] = "red"
par(oma = c(0, 2, 3, 0))
par(mar = rep(1, 4))
for (i in 1:8) {
    # i = 1
    layout(matrix(1:12, 4, 3, byrow = T))
    temp2 = which(mcOutput[1, "True Model", ] == i)
    inputMetrics = c("AIC", "GCV", "SCV", "LOOCV")
    inputStats = c("B0RMSE Rank", "B1RMSE Rank", "B2RMSE Rank")
    for (j in 1:4) {
        for (k in 1:3) {
            temp3 = mcOutput[inputMetrics[j], inputStats[k], temp2]
            # summary(temp3)
            hist(temp3, xlim = c(0, 50), breaks = 5 * (0:10), col = ifelse(colMat[i,
                k] == "red", "red", "grey85"), axes = F, ylab = "rel freq",
                xlab = "rank", main = "")
        }
    }
    mtext(text = paste0("Model #", i), side = 3, outer = TRUE, line = 1.8)
    mtext(inputMetrics, at = seq(0.875, 0.125, length.out = 4), side = 2, outer = T)
    mtext(paste0("B", 0:2), at = seq(0.17, 0.83, length.out = 3), side = 3,
        outer = T, line = 0, col = colMat[i, ])
}
```


Model #1
B0
B1
B2
AIC
GCV
SCV
LOOCV

3 Predicting Beta RMSE Rank

Let's convert ranks into a proportion and then use a logistic regression to look for patterns in rank vs. other variables like the metric and degree of spatial variation, etc.

```
temp2 = which(mcOutput[1, "True Model", ] == i)
inputMetrics = c("AIC", "GCV", "SCV", "LOOCV")
inputStats = c("B0RMSE Rank", "B1RMSE Rank", "B2RMSE Rank")
temp3 = mcOutput[inputMetrics[j], inputStats[k], temp2]
myVars = c("B0RMSE Rank", "B1RMSE Rank", "B2RMSE Rank", "Sample Size", "Error",
    "True Model", "B0 SpVar", "B1 SpVar", "B2 SpVar")
rankData1 = as.data.frame(t(mcOutput["AIC", myVars, ]))
rankData1$Metric = "AIC"
rankData2 = as.data.frame(t(mcOutput["GCV", myVars, ]))
rankData2$Metric = "GCV"
rankData3 = as.data.frame(t(mcOutput["SCV", myVars, ]))
rankData3$Metric = "SCV"
rankData4 = as.data.frame(t(mcOutput["LOOCV", myVars, ]))
rankData4$Metric = "LOOCV"
rankData = rbind(rankData1, rankData2, rankData3, rankData4)
rankData$B0rank = rankData[, "B0RMSE Rank"]/50
rankData$B1rank = rankData[, "B1RMSE Rank"]/50
rankData$B2rank = rankData[, "B2RMSE Rank"]/50
rankData$samplesize = rankData[, "Sample Size"]
rankData$error = as.factor(rankData$Error)
rankData$truemodel = as.factor(rankData[, "True Model"])
rankData$B0sv = as.factor(rankData[, "B0 SpVar"])
rankData$B1sv = as.factor(rankData[, "B1 SpVar"])
rankData$B2sv = as.factor(rankData[, "B2 SpVar"])
rankData$metric = as.factor(rankData$Metric)
summary(rankData)
```

```
##   B0RMSE Rank    B1RMSE Rank    B2RMSE Rank    Sample Size
##  Min.   : 1.0   Min.   : 1.0   Min.   : 1.0   Min.   : 100
##  1st Qu.: 1.0   1st Qu.: 6.0   1st Qu.: 6.0   1st Qu.: 250
##  Median : 6.0   Median :15.0   Median :15.0   Median : 450
##  Mean   :11.9   Mean   :17.4   Mean   :17.4   Mean   : 500
##  3rd Qu.:19.0   3rd Qu.:26.0   3rd Qu.:26.0   3rd Qu.: 700
##  Max.   :50.0   Max.   :50.0   Max.   :50.0   Max.   :1000
##
##      Error        True Model      B0 SpVar       B1 SpVar
##  Min.   :0.50   Min.   :1.00   Min.   :0.00   Min.   :0.00
##  1st Qu.:0.50   1st Qu.:4.00   1st Qu.:0.00   1st Qu.:0.00
##  Median :1.00   Median :6.00   Median :1.00   Median :1.00
##  Mean   :1.17   Mean   :5.67   Mean   :1.33   Mean   :1.33
```

```
##  3rd Qu.:2.00   3rd Qu.:8.00   3rd Qu.:3.00   3rd Qu.:3.00
##  Max.   :2.00   Max.   :8.00   Max.   :3.00   Max.   :3.00
##
##     B2 SpVar       Metric              B0rank          B1rank
##  Min.   :0.00   Length:259200      Min.   :0.020   Min.   :0.020
##  1st Qu.:0.00   Class :character   1st Qu.:0.020   1st Qu.:0.120
##  Median :1.00   Mode  :character   Median :0.120   Median :0.300
##  Mean   :1.33                      Mean   :0.239   Mean   :0.348
##  3rd Qu.:3.00                      3rd Qu.:0.380   3rd Qu.:0.520
##  Max.   :3.00                      Max.   :1.000   Max.   :1.000
##
##      B2rank        samplesize   error         truemodel     B0sv
##  Min.   :0.020   Min.   : 100   0.5:86400   8      :76800   0:86400
##  1st Qu.:0.120   1st Qu.: 250   1  :86400   4      :38400   1:86400
##  Median :0.300   Median : 450   2  :86400   6      :38400   3:86400
##  Mean   :0.348   Mean   : 500               7      :38400
##  3rd Qu.:0.520   3rd Qu.: 700               2      :19200
##  Max.   :1.000   Max.   :1000               3      :19200
##                                             (Other):28800
##  B1sv      B2sv        metric
##  0:86400   0:86400   AIC  :64800
##  1:86400   1:86400   GCV  :64800
##  3:86400   3:86400   LOOCV:64800
##                      SCV  :64800
##
##
##
```

How are the various parameters related to the results of the β_0 ranking?

```
mylogit <- glm(cbind(B0rank, 1 - B0rank) ~ samplesize + metric + B0sv + B1sv +
    B2sv + error, data = rankData, family = "binomial")

## Warning:  non-integer counts in a binomial glm!

summary(mylogit)

##
## Call:
## glm(formula = cbind(B0rank, 1 - B0rank) ~ samplesize + metric +
##     B0sv + B1sv + B2sv + error, family = "binomial", data = rankData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7108  -0.3114  -0.0971   0.1461   2.5635
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.14e+00   1.94e-02  -58.83  < 2e-16 ***
## samplesize  -7.64e-04   1.55e-05  -49.27  < 2e-16 ***
## metricGCV    5.47e-02   1.55e-02    3.53  0.00042 ***
## metricLOOCV  1.21e-01   1.54e-02    7.86  3.9e-15 ***
## metricSCV    1.47e+00   1.43e-02  102.15  < 2e-16 ***
## B0sv1       -6.37e-01   1.12e-02  -57.07  < 2e-16 ***
## B0sv3       -2.31e+00   1.58e-02 -146.08  < 2e-16 ***
## B1sv1        3.81e-02   1.26e-02    3.01  0.00258 **
## B1sv3        1.49e-01   1.25e-02   11.93  < 2e-16 ***
## B2sv1        2.80e-02   1.27e-02    2.21  0.02690 *
## B2sv3        1.57e-01   1.25e-02   12.56  < 2e-16 ***
## error1       4.54e-01   1.32e-02   34.44  < 2e-16 ***
## error2       9.30e-01   1.29e-02   72.37  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 101072  on 259199  degrees of freedom
## Residual deviance:  50829  on 259187  degrees of freedom
## AIC: 180384
##
## Number of Fisher Scoring iterations: 5
```

How are the various parameters related to the results of the β_1 ranking?

```
mylogit <- glm(cbind(B1rank, 1 - B1rank) ~ samplesize + metric + B0sv + B1sv +
    B2sv + error, data = rankData, family = "binomial")

## Warning:  non-integer counts in a binomial glm!

summary(mylogit)

##
## Call:
## glm(formula = cbind(B1rank, 1 - B1rank) ~ samplesize + metric +
##     B0sv + B1sv + B2sv + error, family = "binomial", data = rankData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3801  -0.4638  -0.0544   0.3592   1.7640
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -1.36e+00   1.64e-02  -82.87  < 2e-16 ***
## samplesize  -1.99e-04   1.25e-05  -15.96  < 2e-16 ***
## metricGCV    5.47e-02   1.22e-02    4.47  7.8e-06 ***
## metricLOOCV  8.33e-02   1.22e-02    6.82  8.9e-12 ***
## metricSCV    7.75e-01   1.18e-02   65.63  < 2e-16 ***
## B0sv1        4.95e-01   1.03e-02   48.05  < 2e-16 ***
## B0sv3        5.79e-02   1.06e-02    5.48  4.2e-08 ***
## B1sv1        2.12e-01   1.04e-02   20.43  < 2e-16 ***
## B1sv3        1.55e-01   1.04e-02   14.89  < 2e-16 ***
## B2sv1        1.91e-02   1.03e-02    1.85    0.064 .
## B2sv3       -5.25e-02   1.04e-02   -5.06  4.1e-07 ***
## error1       3.15e-01   1.05e-02   29.88  < 2e-16 ***
## error2       5.18e-01   1.04e-02   49.63  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 80156  on 259199  degrees of freedom
## Residual deviance: 68248  on 259187  degrees of freedom
## AIC: 279685
##
## Number of Fisher Scoring iterations: 4
```

How are the various parameters related to the results of the β_2 ranking?

```
mylogit <- glm(cbind(B2rank, 1 - B2rank) ~ samplesize + metric + B0sv + B1sv +
    B2sv + error, data = rankData, family = "binomial")

## Warning:  non-integer counts in a binomial glm!

summary(mylogit)

##
## Call:
## glm(formula = cbind(B2rank, 1 - B2rank) ~ samplesize + metric +
##     B0sv + B1sv + B2sv + error, family = "binomial", data = rankData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3731  -0.4656  -0.0568   0.3582   1.7645
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.36e+00   1.64e-02  -83.12  < 2e-16 ***
## samplesize  -2.05e-04   1.25e-05  -16.36  < 2e-16 ***
```

```
## metricGCV    5.68e-02   1.22e-02    4.64  3.5e-06 ***
## metricLOOCV  8.61e-02   1.22e-02    7.05  1.8e-12 ***
## metricSCV    7.74e-01   1.18e-02   65.49  < 2e-16 ***
## B0sv1        4.92e-01   1.03e-02   47.82  < 2e-16 ***
## B0sv3        5.44e-02   1.06e-02    5.16  2.5e-07 ***
## B1sv1        1.69e-02   1.03e-02    1.64      0.1
## B1sv3       -4.72e-02   1.04e-02   -4.55  5.3e-06 ***
## B2sv1        1.96e-01   1.04e-02   18.91  < 2e-16 ***
## B2sv3        1.60e-01   1.04e-02   15.41  < 2e-16 ***
## error1       3.34e-01   1.05e-02   31.68  < 2e-16 ***
## error2       5.29e-01   1.04e-02   50.70  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 80411  on 259199  degrees of freedom
## Residual deviance: 68477  on 259187  degrees of freedom
## AIC: 279952
##
## Number of Fisher Scoring iterations: 4
```