Variable Selection for Health Care Demand in Germany

Zhu Wang
Connecticut Children's Medical Center
University of Connecticut School of Medicine
zwang@connecticutchildrens.org

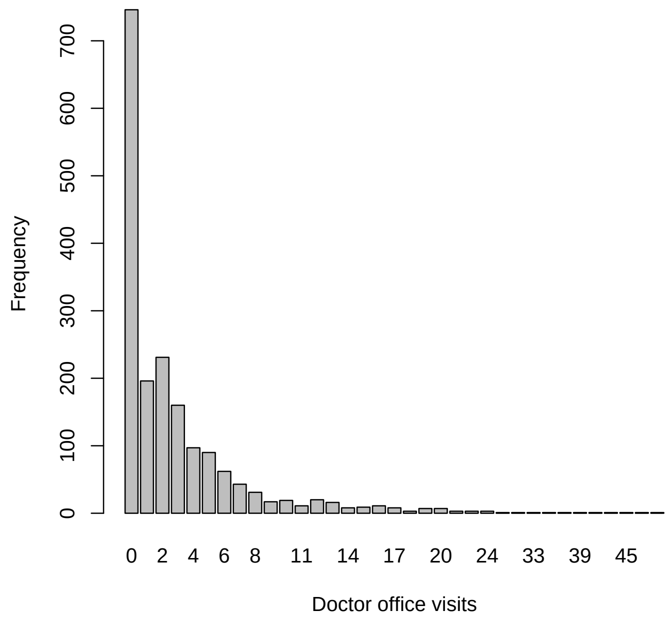
November 9, 2014

This document reproduces the data analysis presented in Wang et al. (2014). For a description of the theory behind application illustrated here we refer to the original manuscript.

Riphahn et al. (2003) utilized a part of the German Socioeconomic Panel (GSOEP) data set to analyze the number of doctor visits. The original data have twelve annual waves from 1984 to 1995 for a representative sample of German households, which provide broad information on the health care utilization, current employment status, and the insurance arrangements under which subjects are protected. The data set contains number of doctor office visits for 1,812 West German men aged 25 to 65 years in the last three months of 1994. As shown in the figure, many doctor office visits are zeros, which can be difficult to fit with a Poisson or negative binomial model. Therefore, zero-inflated negative binomial (ZINB) model is considered.

```
R> library("mpath")
R> library("zic")
R> library("pscl")
R> data(docvisits)
```

```
R> barplot(with(docvisits, table(docvisits)), ylab = "Frequency",
+      xlab = "Doctor office visits")
```



We include the linear spline variables *age30* to *age60* and their interaction terms with the health satisfaction *health*.

```
R> dt <- docvisits[, -(2:3)]
R> tmp <- model.matrix(~age30 * health + age35 * health +
+      age40 * health + age45 * health + age50 * health +
+      age55 * health + age60 * health, data = dt)[, -(1:9)]
R> dat <- cbind(dt, tmp)
```

Full ZINB model with all predictor variables.

```
R> m1 <- zeroinfl(docvisits ~ . | ., data = dat, dist = "negbin")
R> summary(m1)
R> cat("loglik of zero-inflated model", logLik(m1))
R> cat("BIC of zero-inflated model", AIC(m1, k = log(dim(dat)[1])))
R> cat("AIC of zero-inflated model", AIC(m1))
```

Backward stepwise variable selection with significance level alpha=0.01.

```
R> fitbe <- be.zeroinfl(m1, data = dat, dist = "negbin",
+      alpha = 0.01, trace = FALSE)
R> summary(fitbe)
R> cat("loglik of zero-inflated model with backward selection",
+      logLik(fitbe))
R> cat("BIC of zero-inflated model with backward selection",
+      AIC(fitbe, k = log(dim(dat)[1])))
```

Compute LASSO estimates.

```
R> fit.lasso <- zipath(docvisits ~ . | ., data = dat, family = "negbin",
+      nlambda = 100, lambda.zero.min.ratio = 0.001, maxit.em = 300,
+      maxit.theta = 25, theta.fixed = FALSE, trace = FALSE,
+      penalty = "enet", rescale = FALSE)
```

Estimated coefficient parameters with smallest BIC value.

```
R> minBic <- which.min(BIC(fit.lasso))
R> coef(fit.lasso, minBic)
R> cat("theta estimate", fit.lasso$theta[minBic])
```

Compute standard errors of coefficients and theta (the last one for theta).

```
R> se(fit.lasso, minBic, log = FALSE)
```

Compute AIC, BIC, log-likelihood values of the selected model.

```
R> AIC(fit.lasso)[minBic]
R> BIC(fit.lasso)[minBic]
R> logLik(fit.lasso)[minBic]
```

Compute log-likelihood value via 10-fold cross-validation.

```
R> n <- dim(dat)[1]
R> K <- 10
R> foldid <- split(sample(1:n), rep(1:K, length = n))
R> fitcv <- cv.zipath(docvisits ~ . | ., data = dat, family = "negbin",
+      nlambda = 100, lambda.count = fit.lasso$lambda.count[1:30],
+      lambda.zero = fit.lasso$lambda.zero[1:30], maxit.em = 300,
+      maxit.theta = 1, theta.fixed = FALSE, trace = FALSE,
+      penalty = "enet", rescale = FALSE, foldid = foldid)
R> cat("max of cv loglik", max(fitcv$cv))
```

Compute MCP estimates. We compute solution paths for the first 30 pairs of shrinkage parameters (the EM algorithm can be slow), and then evaluate results as for the LASSO estimates. For cross-validation, set maximum number of iterations in estimating scaling parameter 1 (maxit.theta=1) to reduce computation costs.

```
R> fit.mcp <- zipath(docvisits ~ . | ., data = dat, family = "negbin",
+      gamma.count = 2.7, gamma.zero = 2.7, lambda.count = fit.lasso$lambda.count[1:30],
+      lambda.zero = fit.lasso$lambda.zero[1:30], maxit.em = 300,
+      maxit.theta = 10, theta.fixed = FALSE, penalty = "mnet")
R> minBic <- which.min(BIC(fit.mcp))
R> coef(fit.mcp, minBic)
R> cat("theta estimate", fit.mcp$theta[minBic])
R> se(fit.mcp, minBic, log = FALSE)
R> AIC(fit.mcp)[minBic]
R> BIC(fit.mcp)[minBic]
R> logLik(fit.mcp)[minBic]
R> fitcv <- cv.zipath(docvisits ~ . | ., data = dat, family = "negbin",
+      gamma.count = 2.7, gamma.zero = 2.7, lambda.count = fit.lasso$lambda.count[1:30],
+      lambda.zero = fit.lasso$lambda.zero[1:30], maxit.em = 300,
+      maxit.theta = 1, theta.fixed = FALSE, trace = FALSE,
+      penalty = "mnet", rescale = FALSE, foldid = foldid)
R> cat("max of cv loglik", max(fitcv$cv))
```

Compute SCAD estimates.

```
R> fit.scad <- zipath(docvisits ~ . | ., data = dat, family = "negbin",
+      gamma.count = 2.5, gamma.zero = 2.5, lambda.count = fit.lasso$lambda.count[1:30],
+      lambda.zero = fit.lasso$lambda.zero[1:30], maxit.em = 300,
+      maxit.theta = 10, theta.fixed = FALSE, penalty = "snet")
R> minBic <- which.min(BIC(fit.scad))
R> coef(fit.scad, minBic)
R> cat("theta estimate", fit.scad$theta[minBic])
R> se(fit.scad, minBic, log = FALSE)
R> AIC(fit.scad)[minBic]
R> BIC(fit.scad)[minBic]
R> logLik(fit.scad)[minBic]
R> fitcv <- cv.zipath(docvisits ~ . | ., data = dat, family = "negbin",
+      gamma.count = 2.5, gamma.zero = 2.5, lambda.count = fit.lasso$lambda.count[1:30],
+      lambda.zero = fit.lasso$lambda.zero[1:30], maxit.em = 300,
+      maxit.theta = 1, theta.fixed = FALSE, trace = FALSE,
+      penalty = "snet", rescale = FALSE, foldid = foldid)
R> cat("max of cv loglik", max(fitcv$cv))
```

References

Regina T Riphahn, Achim Wambach, and Andreas Million. Incentive effects in the demand for health care: a bivariate panel count data estimation. *Journal of Applied Econometrics*, 18(4):387–405, 2003.

Zhu Wang, Shuangge Ma, and Ching-Yun Wang. Variable selection for zero-inflated and overdispersed data with application to health care demand in germany. *Biometric Journal*, 2014. under review.