Examples of focused model comparison: linear regression

Christopher Jackson
chris.jackson@mrc-bsu.cam.ac.uk

Abstract

This vignette illustrates focused model comparison with the **fic** package for linear regression models. Examples are given of covariate selection and polynomial order selection, with focuses defined by the mean, median or other quantiles of the outcome.

Keywords: models.

The linear regression model considered here has the general form

$$y_i \sim N(\mu_i, \sigma^2), \quad \mu_i = \alpha + \sum \beta_s x_{is}.$$

for observations $i = 1, \ldots, n$. The regressors x_{is} might represent different covariates, contrasts between levels of a factor, functions of covariates such as polynomials, or interactions between different covariates.

1. Covariate selection in linear regression

Firstly we present a simple covariate selection problem in the well-known `mtcars` dataset from the **datasets** package distributed with standard R installations. The outcome y_i is the fuel efficiency of car model i measured in MPG. The wide model is taken to be the model suggested in Henderson and Velleman (1981) which includes the following predictors

- `am`: transmission type (0=automatic, 1=manual)
- `wt`: weight in 1000 lbs
- `qsec`: quarter mile time in seconds
- `disp`: displacement (cubic inches)
- `hp`: gross horsepower

Paired scatterplots of these variables suggest that `mpg` is correlated with all of these predictors, but many of the predictors themselves are correlated with each other.

```
library(GGally)
mtcars$am <- factor(mtcars$am)
ggpairs(mtcars[,c("mpg","am","wt","qsec","disp","hp")], aes(colour=am))
```


```
wide.lm <- lm(mpg ~ am + wt + qsec + disp + hp, data=mtcars)
```

We compare all submodels of this wide model, with the minimal model including only an intercept. The `all_inds` function constructs a matrix of indicators `inds` for whether each coefficient (column) is included in each submodel (row).

```
library(fic)
ncovs_wide <- length(coef(wide.lm)) - 1
inds0 <- c(1, rep(0, ncovs_wide))
inds <- all_inds(wide.lm, inds0)
```

The focus is taken as the mean outcome (`focus=mean_normal`) for a car with covariate values supplied in `X`: automatic transmission `am=0` and values of the other four continuous covariates defined by their means in the data.

```
cmeans <- colMeans(model.frame(wide.lm)[,c("wt","qsec","disp","hp")])
X <- rbind(
  "auto"   = c(intercept=1, am=0, cmeans),
  "manual" = c(intercept=1, am=1, cmeans)
)
ficres <- fic(wide.lm, inds=inds, focus=mean_normal, X=X)
summary(ficres)

## Model with lowest RMSE by focus
##        index focus                    pars
## auto      26  18.5 (Intercept),am1,disp,hp
## manual    26  22.3 (Intercept),am1,disp,hp
## ave       26  20.4 (Intercept),am1,disp,hp
##
## Range of focus estimates and RMSE over models
##        min(focus) max(focus) min(RMSE) max(RMSE)
## auto         16.5       20.1     0.572      2.22
## manual       20.1       25.4     0.699      3.21
## ave          20.1       20.9     0.638      2.76
```

There is a cluster of submodels whose focus estimates are judged to have relatively low bias and mean square error. The model with minimal mean square error, for either focus, omits `wt` and `qsec`. Given the strong correlation of `wt` with `disp` and `qsec` with `hp`, these two variables do not improve the precision of the focus estimate.

2. Polynomial order selection

A common model selection problem is to choose an appropriate level of flexibility for a nonlinear relationship of an outcome with a predictor. This is often implemented through polynomial regression.

In this example, a linear model with orthogonal polynomials is used to represent the relationship of life expectancy to GDP per capita for 1704 countries (worldwide) and years from 1952 to 2007, using data from `http://www.gapminder.org`, packaged by Bryan (2017). The dataset used for analysis excludes Kuwait, whose data follow a distinct pattern. The scatterplot shows a diminishing increase in life expectancy as GDP increases above a certain level.

```
library(gapminder)
gap2 <- gapminder[gapminder$country !="Kuwait",]
pal <- heat.colors(5)
p <- ggplot(gap2, aes(x=gdpPercap, y=lifeExp)) +
  geom_point() +
  xlab("GDP per capita (US$, inflation-adjusted)") +
  ylab("Life expectancy (years)") +
  geom_point(data=gapminder[gapminder$country =="Kuwait",], col="gray") +
  annotate("text", x=80000, y=70, label="Kuwait", col="gray")
```

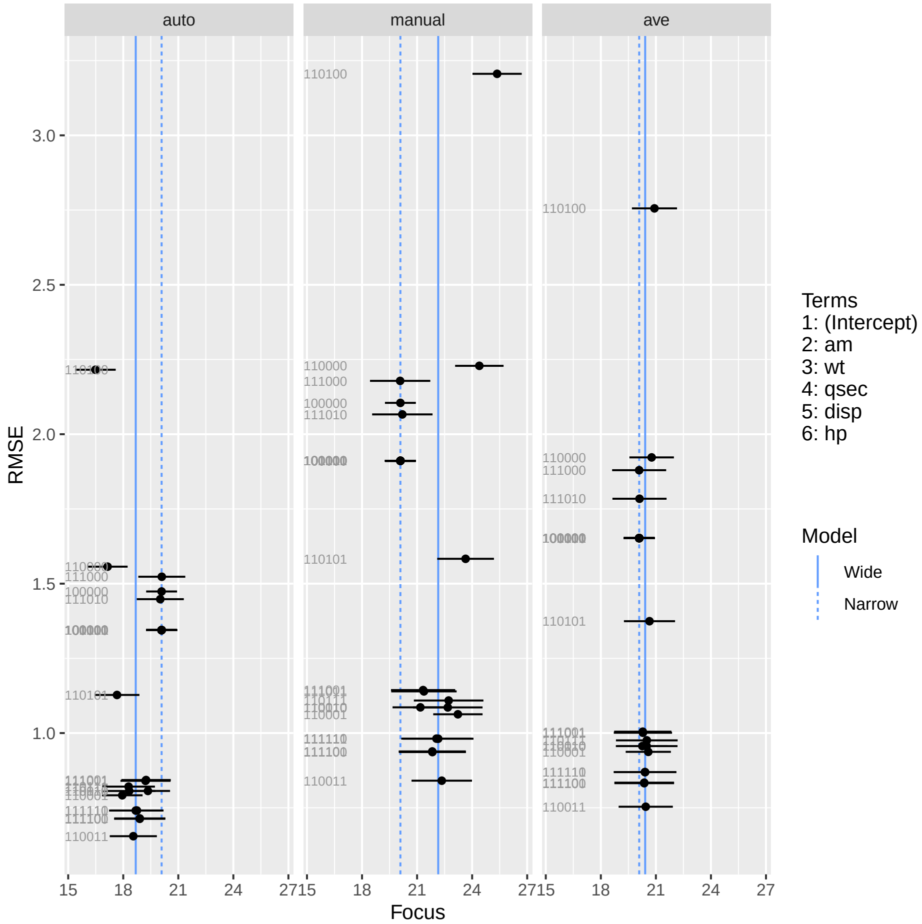
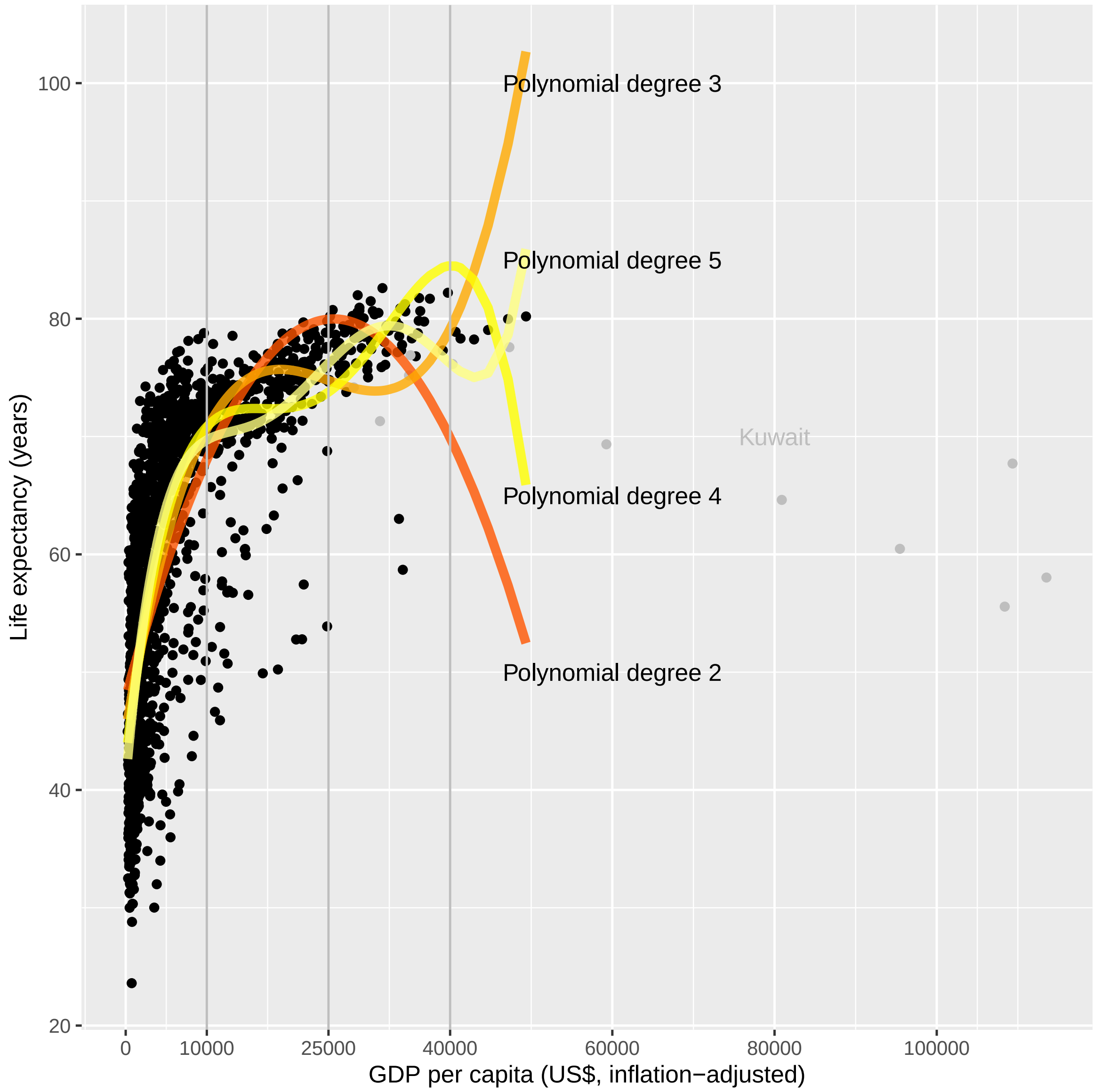


Figure 1: Focused comparison of linear regression models for the `mtcars` data.

A wide model is fitted with a polynomial relationship of degree 5. Fitted values from each model are added to the scatterplot.

```
wide.lm <- lm(lifeExp ~ poly(gdpPercap,5), data=gap2)
yilab <- c(0, 50, 100, 65, 85)
for (i in 2:5) {
    poly.lm <- lm(lifeExp ~ poly(gdpPercap,i), data=gap2)
    ft <- data.frame(x=gap2$gdpPercap, y=fitted(poly.lm))
    ft <- ft[order(ft$x),]
    p <- p +
      geom_line(data=ft, aes(x=x,y=y), col=pal[i], lwd=2, alpha=0.8) +
      annotate("text", x=60000, y=yilab[i], col="black",
               label=sprintf("Polynomial degree %s", i))
}
gdp_focus <- c(10000, 25000, 40000)
p <- p +
```

```
  geom_vline(xintercept=gdp_focus, col="gray") +
  scale_x_continuous(breaks=c(0, gdp_focus, 60000, 80000, 100000))
p
```



Submodels of degrees 2, 3 and 4 are compared in terms of how well they estimate three focuses: the average life expectancy at GDP per capita of $10,000, $25,000 and $40,000. Note that the parameters include the intercept, so, for example, the simplest model, the quadratic polynomial model, has three parameters indicated by entries of 1 in the first row of `inds`.

```
inds <- rbind("quadratic"= c(1,1,1,0,0,0),
              "cubic"     =c(1,1,1,1,0,0),
              "quartic"   =c(1,1,1,1,1,0),
              "degree 5"  =c(1,1,1,1,1,1))
```

```
X <- newdata_to_X(list(gdpPercap=gdp_focus), wide.lm, intercept=TRUE)
rownames(X) <- gdp_focus
(ficres <- fic(wide.lm, inds=inds, focus=mean_normal, X=X))

##     vals      mods  rmse rmse.adj      bias    se    FIC focus
## 1  10000 quadratic 1.701    1.701 -1.68e+00 0.286   4757  68.0
## 5  10000     cubic 0.989    0.989  9.32e-01 0.333   1615  70.7
## 9  10000   quartic 1.153    1.153  1.10e+00 0.334   2209  70.8
## 13 10000  degree 5 0.373    0.373  0.00e+00 0.373    194  69.7
## 2  25000 quadratic 3.768    3.768  3.74e+00 0.444  23689  80.0
## 6  25000     cubic 1.518    1.518 -1.41e+00 0.559   3999  74.8
## 10 25000   quartic 2.483    2.483 -2.42e+00 0.566  10535  73.8
## 14 25000  degree 5 0.673    0.673  0.00e+00 0.673    867  76.2
## 3  40000 quadratic 6.459    6.459 -6.33e+00 1.299  67731  69.9
## 7  40000     cubic 2.970    2.970  2.60e+00 1.433  15660  79.1
## 11 40000   quartic 8.331    8.331  8.19e+00 1.513 118171  84.5
## 15 40000  degree 5 1.952    1.952  0.00e+00 1.952   7180  76.2
## 4    ave quadratic 4.335    4.335 -4.26e+00 0.809  30955  72.6
## 8    ave     cubic 2.009    2.009  1.79e+00 0.909   5987  74.9
## 12   ave   quartic 5.063    5.063  4.97e+00 0.952  42534  76.4
## 16   ave  degree 5 1.211    1.211 -1.35e-31 1.211   1643  74.0

summary(ficres)

## Model with lowest RMSE by focus
##       index focus
## 10000     4  69.7
## 25000     4  76.2
## 40000     4  76.2
## ave       4  74.0
##
## 10000 (Intercept),poly(gdpPercap, 5)1,poly(gdpPercap, 5)2,poly(gdpPercap, 5)3,poly(gdpP
## 25000 (Intercept),poly(gdpPercap, 5)1,poly(gdpPercap, 5)2,poly(gdpPercap, 5)3,poly(gdpP
## 40000 (Intercept),poly(gdpPercap, 5)1,poly(gdpPercap, 5)2,poly(gdpPercap, 5)3,poly(gdpP
## ave   (Intercept),poly(gdpPercap, 5)1,poly(gdpPercap, 5)2,poly(gdpPercap, 5)3,poly(gdpP
##
## Range of focus estimates and RMSE over models
##       min(focus) max(focus) min(RMSE) max(RMSE)
## 10000       68.0       70.8     0.373      1.70
## 25000       73.8       80.0     0.673      3.77
## 40000       69.9       84.5     1.952      8.33
## ave         72.6       76.4     1.211      5.06
```

While the most complex model gives the most precise estimates of mean life expectancy at all focuses, the preference for the complex model is less strong for GDP=10000 — at this point there are more data, the models give more consistent focus estimates, and the bias incurred by using a simpler model is less.

This is a simplified example — alternative approaches to nonlinear regression might involve, e.g. splines or fractional polynomials. In theory, these can be implemented as linear additive models of the form shown here. Though exact details of implementing focused model comparison have not been investigated for these classes of models — note that this would require all submodels to be nested within a single wide model. Note also the importance of considering knowledge of the underlying mechanism when building a regression model, for example, we might be sure that the relationship is monotonic.

2.1. Quantiles as the focus

Claeskens and Hjort (2008) show that for a normal linear regression model, FIC and MSE are the same for a focus defined by the mean outcome as for a focus defined by any quantile of the outcome.

We can check this in this example, while demonstrating how to implement quantiles as focus functions in **fic**.

Firstly, the median of a normal distribution is equal to the mean, and is independent of the variance. Therefore we will get identical answers to the results for `focus=mean_normal` above by doing:

```
median_normal<- function(par,X){
    qnorm(0.5, mean = as.numeric(X %*% par))
}
ficres <- fic(wide.lm, inds=inds, focus=median_normal, X=X)
```

Other quantiles, however, depend on the variance. Therefore a `sigma` argument should be defined for the focus function. This allows, e.g. a 10% quantile focus to be implemented as

```
q10_normal <- function(par, X, sigma){
    qnorm(0.1, mean = as.numeric(X %*% par), sd=sigma)
}
ficres <- fic(wide.lm, inds=inds, focus=q10_normal, X=X)
```

However, we can define focus functions with arbitrary additional arguments. This allows any quantile to be defined using one common function, with an argument, say, `focus_p`, specifying the particular quantile to return. in a

```
quantile_normal <- function(par, X, sigma, focus_p=0.5){
    qnorm(focus_p, mean = as.numeric(X %*% par), sd=sigma)
}
```

This argument can be passed to `fic`, along with the focus function, to fully specify the focus of interest. If a vector of values is supplied in `focus_p`, then multiple focuses are evaluated at once.[1]

[1]Note that vectors for `X` are treated differently from vectors for other focus arguments. If a named vector is

```
ficres <- fic(wide.lm, inds=inds, focus=quantile_normal,
              X=X[1,], focus_p=c(0.1,0.5,0.9))
```

We can check that the results match between the alternative ways of setting up `fic` for the same focus.

3. Relation of focused model comparison with AIC

Using the `mtcars` example, we illustrate when focused model comparison agrees with model comparison using AIC. The following code performs focused model comparison for 32 distinct focus parameters, defined as the log likelihood contribution from each of the 32 observed covariate combinations in the `mtcars` data.

Firstly the focus function is defined as the log density for an individual outcome. Claeskens and Hjort (2003) show that differences between submodels in the expected mean square error of this focus are asymptotically equivalent to differences in AIC.

```
focus_loglik <- function(par,X,sigma,Y){
    mu <- as.numeric(X %*% par)
    dnorm(Y,mu,sigma,log=TRUE)
}
```

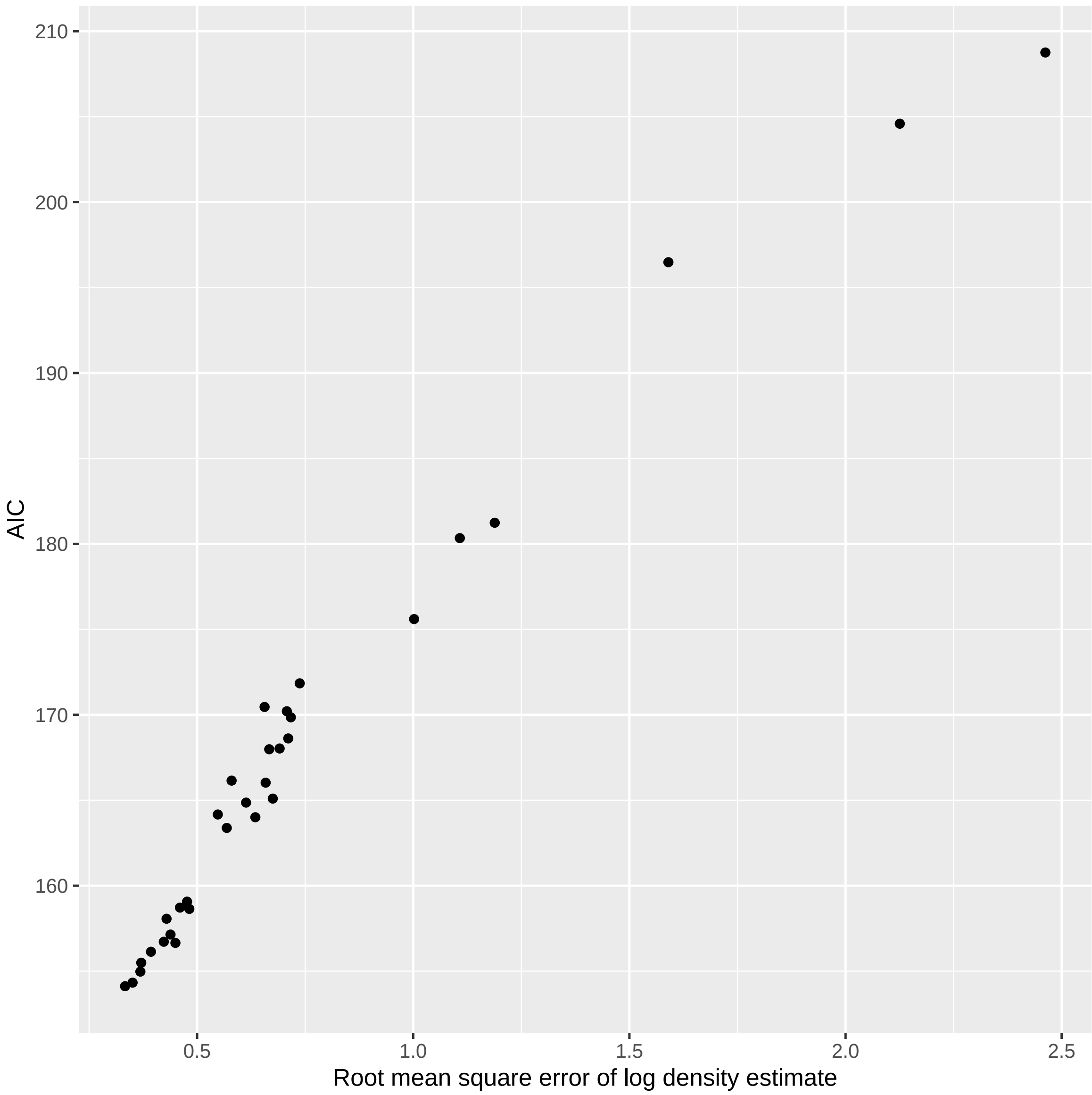
To illustrate this result, we run `fic` with $n = 32$ variants of this focus defined by the observed outcomes `Y` and covariates `X` in the `mtcars` data.

```
wide.lm <- lm(mpg ~ am + wt + qsec + disp + hp, data=mtcars)
ncovs_wide <- length(coef(wide.lm)) - 1
inds0 <- c(1, rep(0, ncovs_wide))
inds <- all_inds(wide.lm, inds0)
X <- model.matrix(wide.lm)
Y <- model.response(model.frame(wide.lm))
ficres <- fic(wide.lm, inds=inds, focus=focus_loglik, X=X, Y=Y)
```

We then extract the results averaged over these focuses, automatically computed by `fic` with each focus weighted equally, and extract the AICs of the submodels. The preference among models from the averaged FIC result agrees with AIC, up to sampling error.

```
ficres <- ficres[ficres$vals=="ave",]
aics <- sapply(attr(ficres,"sub"), AIC)
qplot(ficres$rmse, aics,
      xlab="Root mean square error of log density estimate",
      ylab="AIC")
```

supplied for `X` it is assumed to refer to multiple covariate values defining a single focus. If a vector is supplied for any other argument, it is assumed to identify multiple focuses. To completely avoid ambiguity for any argument, a matrix can be supplied, where the rows identify focuses and the columns identify, e.g. covariate values.



References

Bryan J (2017). ***gapminder**: Data from Gapminder.* R package version 0.3.0, URL https://CRAN.R-project.org/package=gapminder.

Claeskens G, Hjort N (2003). "The Focused Information Criterion (with discussion)." *Journal of the American Statistical Association*, **98**(464), 900–945.

Claeskens G, Hjort N (2008). *Model Selection and Model Averaging.* Cambridge University Press.

Henderson HV, Velleman PF (1981). "Building Multiple Regression Models Interactively." *Biometrics*, **37**(2), 391–411.