Bayesian Analysis on Teen Smoking: Which groups of Teenagers Smoke?

Written by Liman Wei in Nov. 2020 and Posted on https://github.com/weiliman/Course-Assignments

Introduction

When people talk about the smoking rate of teenagers, it is believed rural-urban difference is one of the most important factors, and variation of smoking rate among states is larger than the variation between schools. Besides these variations, there are also some age differences, race differences and gender differences. In this report, we are investigating what are major factors influencing the smoking rate for students aged 10-19. The data we are using comes from 2014 American National Youth Tobacco Survey, which includes over 18,000 samples.

Abstract

We built a generalized linear mixed model to analyze the major factors influencing the smoking rate using a sample of 18,000 students aged 10-19. As a result, we find the largest effect on smoking rate is the school level differences, while rural-urban difference is the second largest and state difference is the least influential one. Hence tobacco control programs should focus on schools with high smoking rates, and put more effort on rural areas. For individual students, there is a gender difference on rates of smoking, but no race difference.

Methods

The data set contains 18,399 observations of students from different schools and different states. Each observation records a student's age, race, school, living area (rural/urban), state, and whether this student ever smokes a cigarette or not.

We built a generalized linear mixed model using Bayesian inference, and the model could be described as

$$Y_{ijk} \sim Bernoulli(p_{ijk}) \qquad log(\frac{p_{ijk}}{1-p_{ijk}}) = \mu + X_{ijk}\beta + U_i + V_{ij}$$

with prior distributions:

$$\begin{array}{ll} \mu \sim N(0, 0.2) & \beta \sim N(0, 10) \\ U_i \sim N(0, \sigma_1) & V_{ij} \sim N(0, \sigma_2) \\ P(\sigma_1 > log(2)) = 0.5 & P(\sigma_2 > log(1.2)/1.3) = 0.5 \end{array}$$

where Y_{ijk} represents the the predicted probability of smoking for student k at school j in state i, X_{ijk} is the covariates containing interaction term between age and race, interaction term between age and sex, and interaction term between age and rural-urban (notice that age is treated as a categorical variable). U_i represents the state level difference and V_{ij} represents the school level difference.

We used $log(2)$ as a threshold for the median of σ_1 because we want one standard deviation of U_i to double the probability odds. We used $log(1.2)/1.3$ as a threshold for median of σ_2 because we want to set the IQR of V_{ij} to 1.2, and hence the odds ratio of IQR is a 20% increase.

To compare the effects of rural-urban difference, state level difference and school difference, we listed the posterior quantiles of the interaction term between rural-urban and age in Table 1, and we listed the posterior quantiles for σ_1 and σ_2 in Table 2. We also plotted the prior distributions and posterior distributions for these variables in Figure 1 and Figure 2. Furthermore, to visualize how the fixed effects influence the response variable, we plotted the predicted probabilities of smoking for different age, race, gender groups living in rural or urban areas (with 95% confidence intervals) in Figure 3 - 6.

Results

From table 1 and table 2, we can see the most influential factor is the school level difference, the second largest effect is the rural-urban difference, and the lease important effect is state level difference.

In Figure 3 - 6, we can see the predicted probabilities of smoking are increasing as age increases. For the race difference,

there are lots of overlapping between the predicted 95% confidence intervals of the probabilities of smoking for Hispanic, black and white people , hence we conclude that there are no significant evidence showing the probabilities of smoking for Hispanic, black and white people are different when holding other variables constant.

From Figure 3-6, the predicted probabilities for rural males/females are about 0.1 higher than urban males/females. Hence students from rural areas are more likely to smoke. From the same graphs, we can see the predicted probabilities for rural/urban males are 0.05 higher than rural/urban females. Hence there is a sexual difference, males have a higher smoking rate than females.

Discussion

From the tables and graphs, we find the most influential factor of probabilities of smoking is the school level difference. Hence if some people whats to lower the rates of student smoking, they should target at individual schools where smoking is a big concern, rather than target at individual states. Rural-urban difference is the second largest effect on probabilities of smoking, hence paying more attention to rural areas will be a good strategy as well.

For individual students, the probabilities to smoke increases as age increases, and male students are more likely to smoke, but there is no significant race difference. Hence tobacco control program should put more effort on older, male students.

Appendix: Graphs and Tables

Tables of Posterior quantitles

Table 1: Posterior quantiles of the fixed effects (variables relating to rural-urban only)

	0.5quant	0.025quant	0.975quant
RuralUrbanRural	0.207	-0.010	0.422
ageFac12:RuralUrbanRural	-0.068	-0.323	0.187
ageFac13:RuralUrbanRural	0.185	-0.051	0.420
ageFac14:RuralUrbanRural	0.091	-0.122	0.304
ageFac15:RuralUrbanRural	0.192	-0.019	0.403
ageFac16:RuralUrbanRural	0.323	0.115	0.532
ageFac17:RuralUrbanRural	0.243	0.035	0.452
ageFac18:RuralUrbanRural	0.350	0.123	0.576
ageFac19:RuralUrbanRural	0.224	-0.104	0.551

Table 2: Posterior quantiles of the SD of random effects

	0.5quant	0.025quant	0.975quant
SD for state	0.157	0.038	0.359
SD for school	0.606	0.540	0.708

Plots of posterior distributions comparing to prior distributions

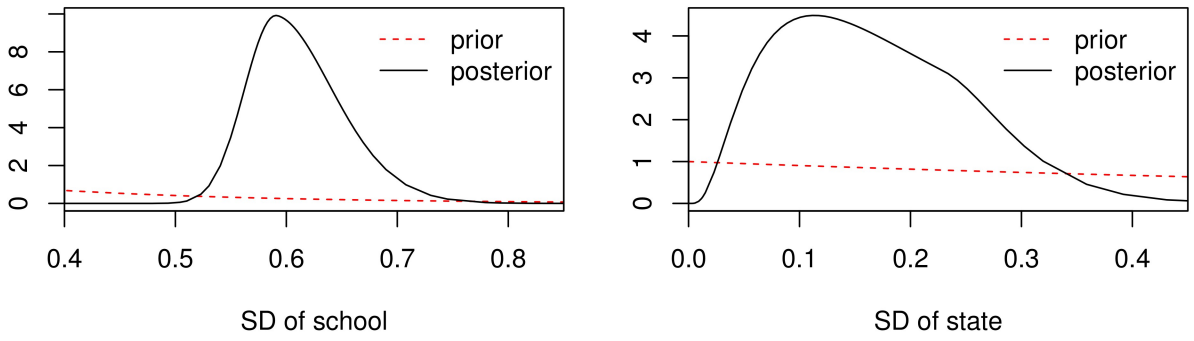

Figure 1: Priors and Posteriors for SD of random effects

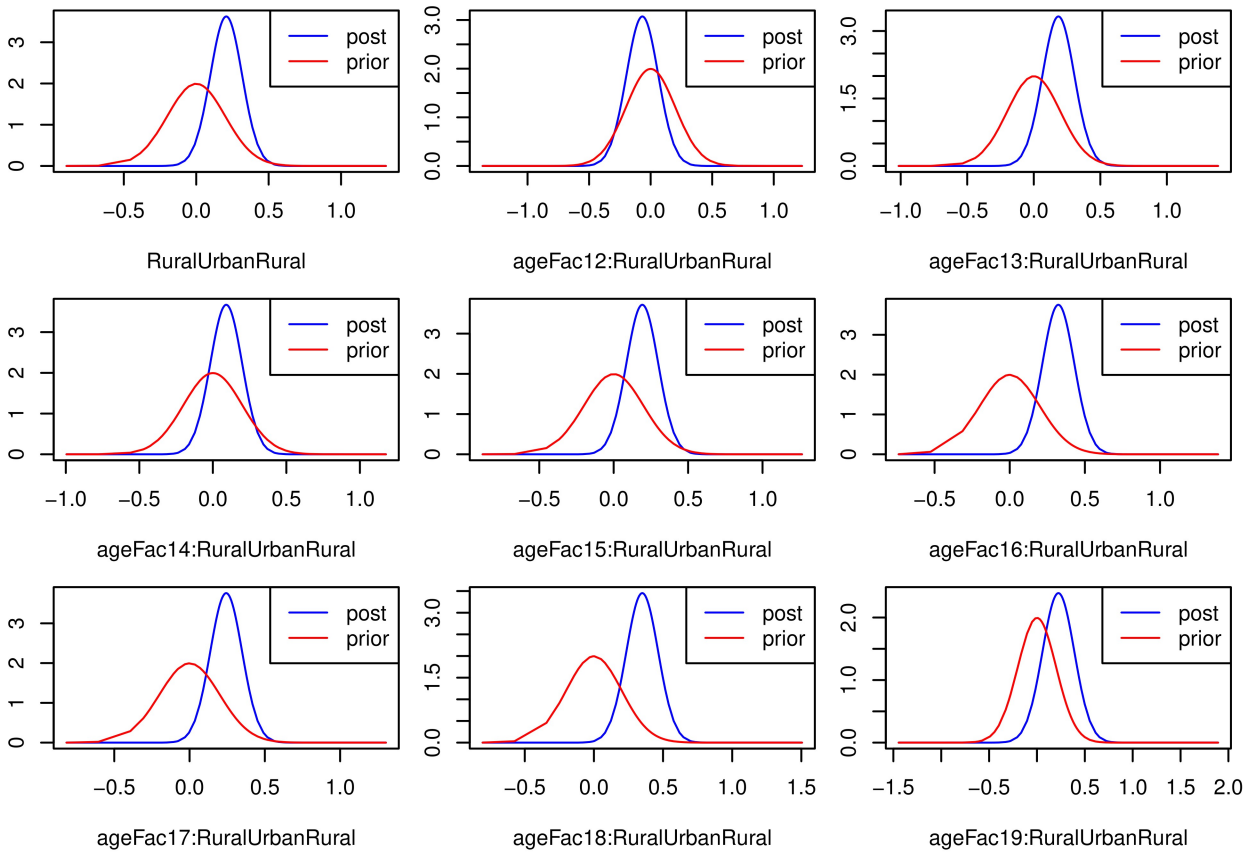

Figure 2: Priors and Posteriors for fixed effects (variables relating to rural-urban only)

Plots of Predicted probabilities of smoking

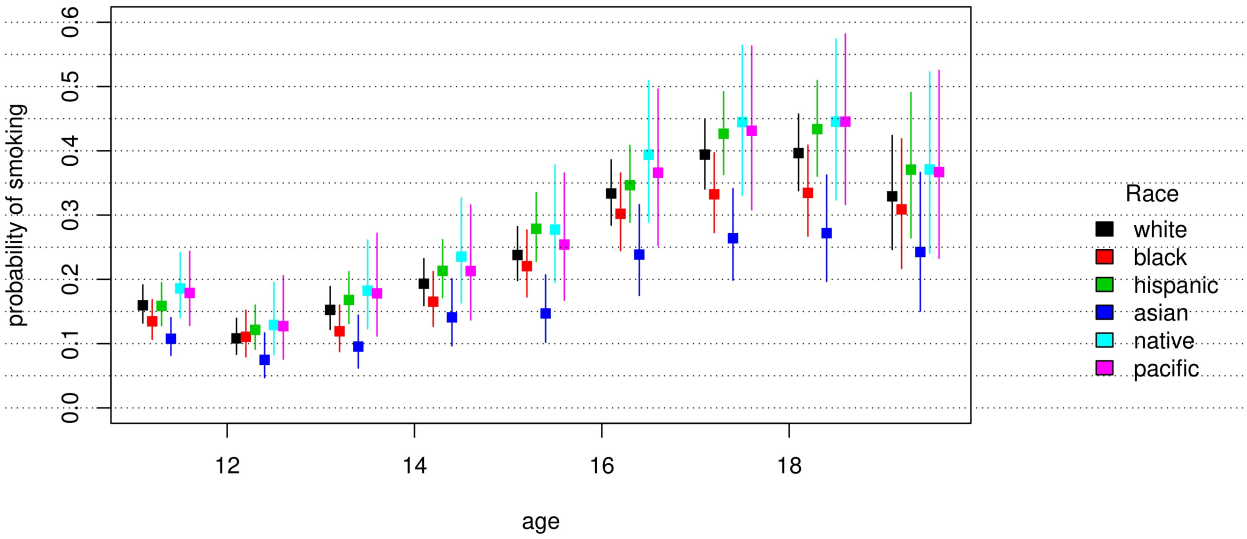

Figure 3: Predicted probabilities of smoking for RURAL MALES by ages with 95% CIs

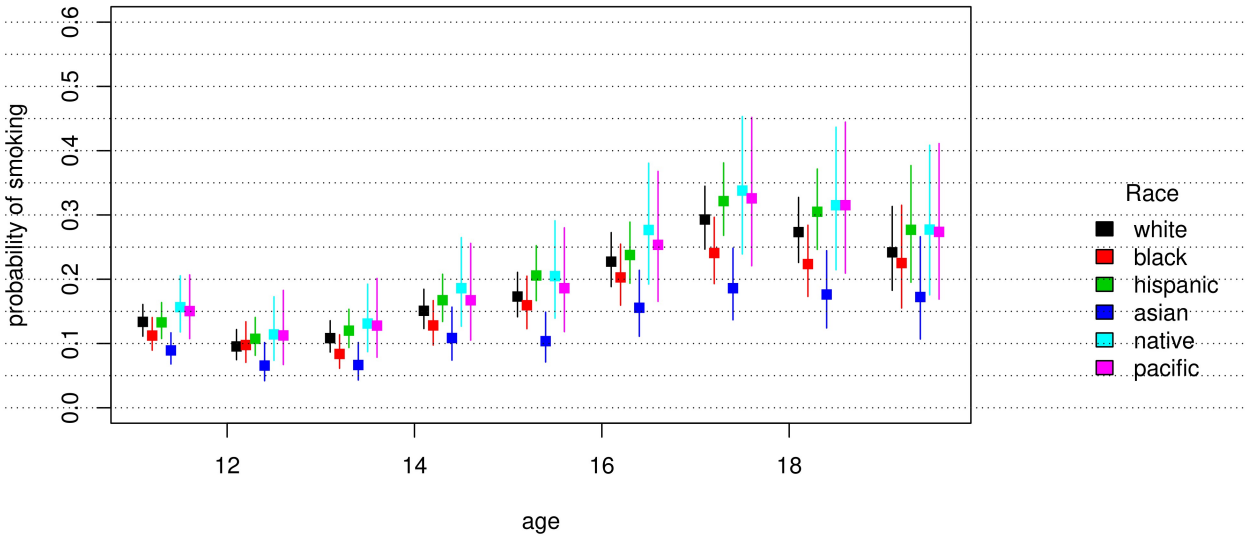

Figure 4: Predicted probabilities of smoking for URBAN MALES by ages with 95% CIs

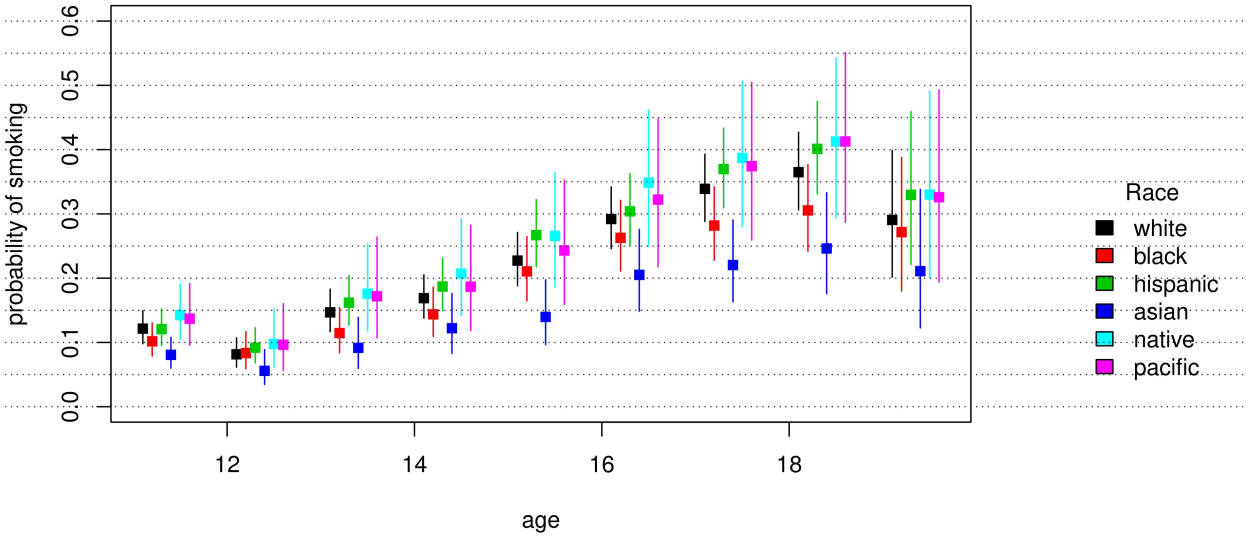

Figure 5: Predicted probabilities of smoking for RURAL FEMALES by ages with 95% CIs

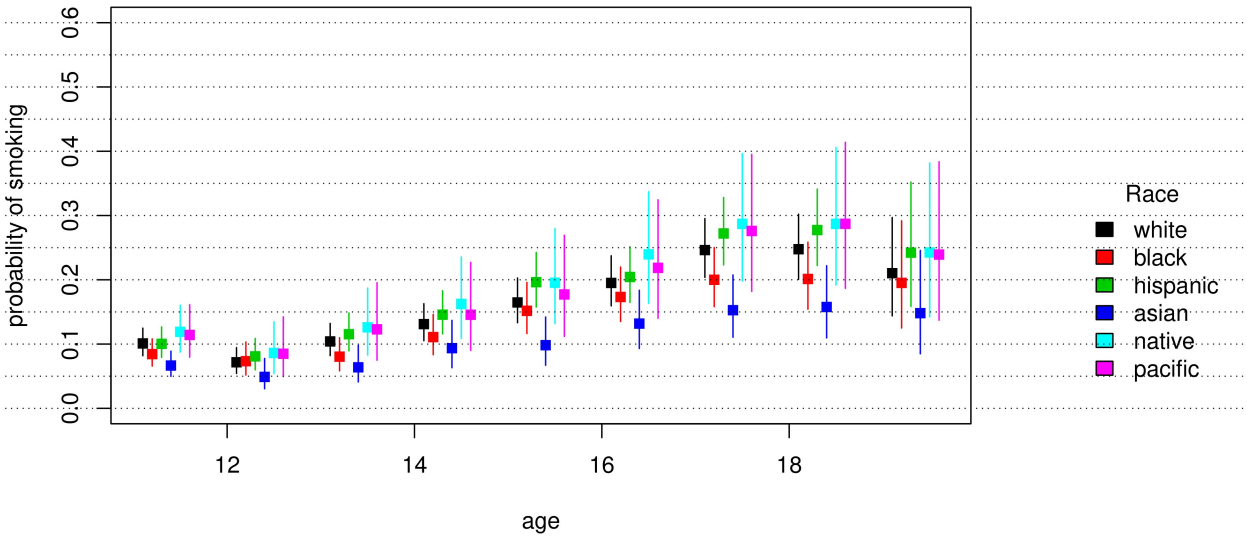

Figure 6: Predicted probabilities of smoking for URBAN FEMALES by ages with 95% CIs

Appendix: R codes

Data Cleaning

```
dataDir = "../data"
smokeFile = file.path(dataDir, "smoke2014.RData")
if (!file.exists(smokeFile)) {
download.file("http://pbrown.ca/teaching/appliedstats/data/smoke2014.RData",
smokeFile)
}
load(smokeFile)
forInla = smoke[smoke$Age > 10, c("Age", "ever_cigarettes",
"Sex", "Race", "state", "school", "RuralUrban",
"Harm_belief_of_chewing_to")]
forInla = na.omit(forInla)
forInla$y = as.numeric(forInla$ever_cigarettes)
forInla$ageFac = factor(as.numeric(as.character(forInla$Age)))
forInla$chewingHarm = factor(forInla$Harm_belief_of_chewing_to,
levels = 1:4, labels = c("less", "equal", "more",
"dunno"))
library("INLA")
```

Model

```
toPredict = expand.grid(ageFac = levels(forInla$ageFac),
                        RuralUrban = levels(forInla$RuralUrban),
                        Sex = levels(forInla$Sex), Race = levels(forInla$Race))
forLincombs = do.call(inla.make.lincombs,
                      as.data.frame(model.matrix(
                        ~ageFac * Race + ageFac * Sex + ageFac * RuralUrban, data = toPredict)))
model = inla(y ~ ageFac * Race + ageFac * Sex + ageFac * RuralUrban +
               f(state, model = "iid", prior = 'pc.prec', param = c(log(2), 0.5)) +
               f(school, model = "iid", prior = 'pc.prec', param = c(log(1.2)/1.3, 0.5)),
             data = forInla, family = "binomial",
             control.fixed = list(mean = 0, mean.intercept = 0,
                                  prec = (0.2)^(-2), prec.intercept = (10)^(-2)),
             lincomb = forLincombs)
```

Tables of Posterior quantitles

```
knitr::kable(model$summary.fixed[c(16, 65:72), c(4, 3, 5)], digits = 3,
             caption = "Posterior quantiles of the fixed effects
             (variables relating to rural-urban only)",
             label = 'tab:atable')
sdRes = Pmisc::priorPostSd(model)
knitr::kable(sdRes$summary[, c(4, 3, 5)], digits = 3,
             caption = "Posterior quantiles of the SD of random effects",
             label = 'tab:atable')
```

Plots of posterior distributions comparing to prior distributions

```
par(mfrow = c(1, 2), mar=c(4, 2, 2, 2))
sdRes$school$matplot$xlab = "SD of school"
do.call(matplot, sdRes$school$matplot)
do.call(legend, sdRes$legend)
sdRes$state$matplot$xlab = "SD of state"
do.call(matplot, sdRes$state$matplot)
do.call(legend, sdRes$legend)
```

```
par(mfrow = c(3, 3), mar=c(4, 1.8, 1, 1))
for (D in rownames(model$summary.fixed)[c(16, 65:72)]){
  plot(model$marginals.fixed[[D]], type = "l", col="blue", xlab = D)
  xseq= model$marginals.fixed[[D]][,'x']
  lines(xseq, dnorm(xseq, sd=0.2), type = "l", col="red")
  legend("topright", lty = 1, col = c("blue","red"), legend = c("post","prior"))
}
```

Plots of Predicted probabilities of smoking

```
# create matrix of predicted probabilities
theCoef = exp(model$summary.lincomb.derived[, c("0.5quant",
"0.025quant", "0.975quant")])
theCoef = theCoef/(1 + theCoef)
# create an x axis, shift age by race group
toPredict$Age = as.numeric(as.character(toPredict$ageFac))
toPredict$shiftX = as.numeric(toPredict$Race)/10
toPredict$x = toPredict$Age + toPredict$shiftX

toPlot = toPredict$Sex == "M" & toPredict$RuralUrban == "Rural"
par(mar = c(4,4,1,10), xpd=TRUE)
plot(toPredict[toPlot, "x"], theCoef[toPlot, "0.5quant"],
     xlab = "age", ylab = "probability of smoking",
     ylim = c(0, 0.6), pch = 15, col = toPredict[toPlot, "Race"])
segments(toPredict[toPlot, "x"], theCoef[toPlot, "0.025quant"],
         y1 = theCoef[toPlot, "0.975quant"], col = toPredict[toPlot, "Race"])
abline(h=seq(0, 0.6, 0.05), col="gray8", lty = 3)
legend("topright", inset = c(-0.3, 0.4), fill = 1:nlevels(toPredict$Race),
       legend = levels(toPredict$Race), bty = "n", title = "Race")
```

```
par(mar = c(4,4,1,10), xpd=TRUE)
toPlot = toPredict$Sex == "M" & toPredict$RuralUrban == "Urban"
plot(toPredict[toPlot, "x"], theCoef[toPlot, "0.5quant"],
     xlab = "age", ylab = "probability of smoking",
     ylim = c(0, 0.6), pch = 15, col = toPredict[toPlot, "Race"])
segments(toPredict[toPlot, "x"], theCoef[toPlot, "0.025quant"],
         y1 = theCoef[toPlot, "0.975quant"], col = toPredict[toPlot, "Race"])
legend("topright", inset = c(-0.3, 0.4), fill = 1:nlevels(toPredict$Race),
       legend = levels(toPredict$Race), bty = "n", title = "Race")
abline(h=seq(0, 0.6, 0.05), col="gray8", lty = 3)
```

```r
par(mar = c(4,4,1,10), xpd=TRUE)
toPlot = toPredict$Sex == "F" & toPredict$RuralUrban == "Rural"
plot(toPredict[toPlot, "x"], theCoef[toPlot, "0.5quant"],
     xlab = "age", ylab = "probability of smoking",
     ylim = c(0, 0.6), pch = 15, col = toPredict[toPlot, "Race"])
segments(toPredict[toPlot, "x"], theCoef[toPlot, "0.025quant"],
         y1 = theCoef[toPlot, "0.975quant"], col = toPredict[toPlot, "Race"])
legend("topright", inset = c(-0.3, 0.4), fill = 1:nlevels(toPredict$Race),
       legend = levels(toPredict$Race), bty = "n", title = "Race")
abline(h=seq(0, 0.6, 0.05), col="gray8", lty = 3)
```

```r
par(mar = c(4,4,1,10), xpd=TRUE)
toPlot = toPredict$Sex == "F" & toPredict$RuralUrban == "Urban"
plot(toPredict[toPlot, "x"], theCoef[toPlot, "0.5quant"],
     xlab = "age", ylab = "probability of smoking",
     ylim = c(0, 0.6), pch = 15, col = toPredict[toPlot, "Race"])
segments(toPredict[toPlot, "x"], theCoef[toPlot, "0.025quant"],
         y1 = theCoef[toPlot, "0.975quant"], col = toPredict[toPlot, "Race"])
legend("topright", inset = c(-0.3, 0.4), fill = 1:nlevels(toPredict$Race),
       legend = levels(toPredict$Race), bty = "n", title = "Race")
abline(h=seq(0, 0.6, 0.05), col="gray8", lty = 3)
```