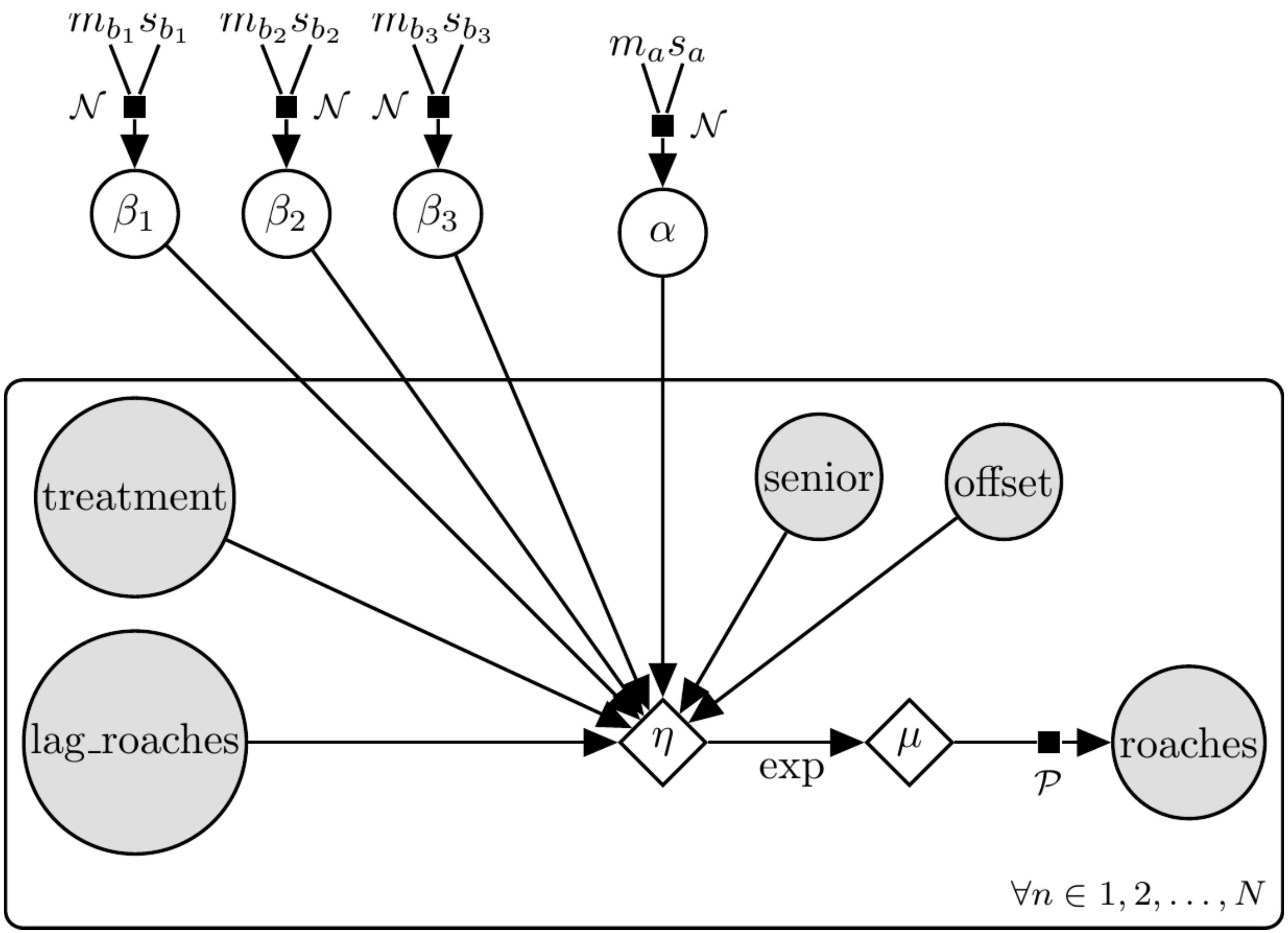
Generalized Linear Models with the rstanarm R Package

Ben Goodrich

March 28, 2022

Prior Predictive Distribution for Roach Study



Roach Model

Prior Predictive Distribution in Symbols

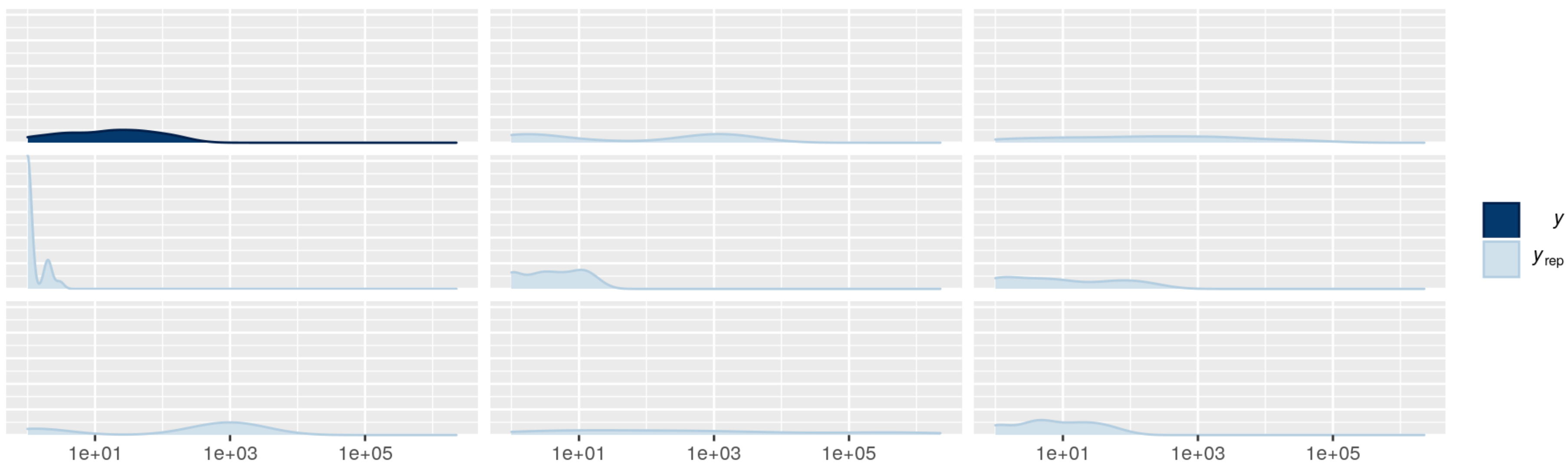
$$\alpha \sim \mathcal{N}\left(m_\alpha, s_\alpha\right)$$
$$\beta_1 \sim \mathcal{N}\left(m_{\beta_1}, s_{\beta_1}\right)$$
$$\beta_2 \sim \mathcal{N}\left(m_{\beta_2}, s_{\beta_2}\right)$$
$$\beta_3 \sim \mathcal{N}\left(m_{\beta_3}, s_{\beta_3}\right)$$
$$\forall n : \eta_n \equiv \alpha + OFFSET_n + \beta_1 \times \log LAG_n + \beta_2 \times SENIOR_n + \beta_3 \times T_n$$
$$\forall n : \mu_n \equiv e^{\eta_n}$$
$$\forall n : Y_n \sim \mathcal{P}\left(\mu_n\right)$$

- In this case, the inverse link function mapping the linear predictor η_n on $\mathbb{R}$ to the outcome's conditional expectation μ_n on $\mathbb{R}_+$ is the antilog function.

Prior Predictive Distribution with **`stan_glm`**

```
roaches <- roaches[roaches$roach1 > 0, ]
priors  <- stan_glm(y ~ senior + log(roach1) + treatment, data = roaches,
                    family = poisson, offset = log(exposure2), QR = TRUE, prior_PD = TRUE)
```

```
pp_check(priors, plotfun = "dens") + scale_x_continuous(trans = "log10")
```



Posterior Distribution

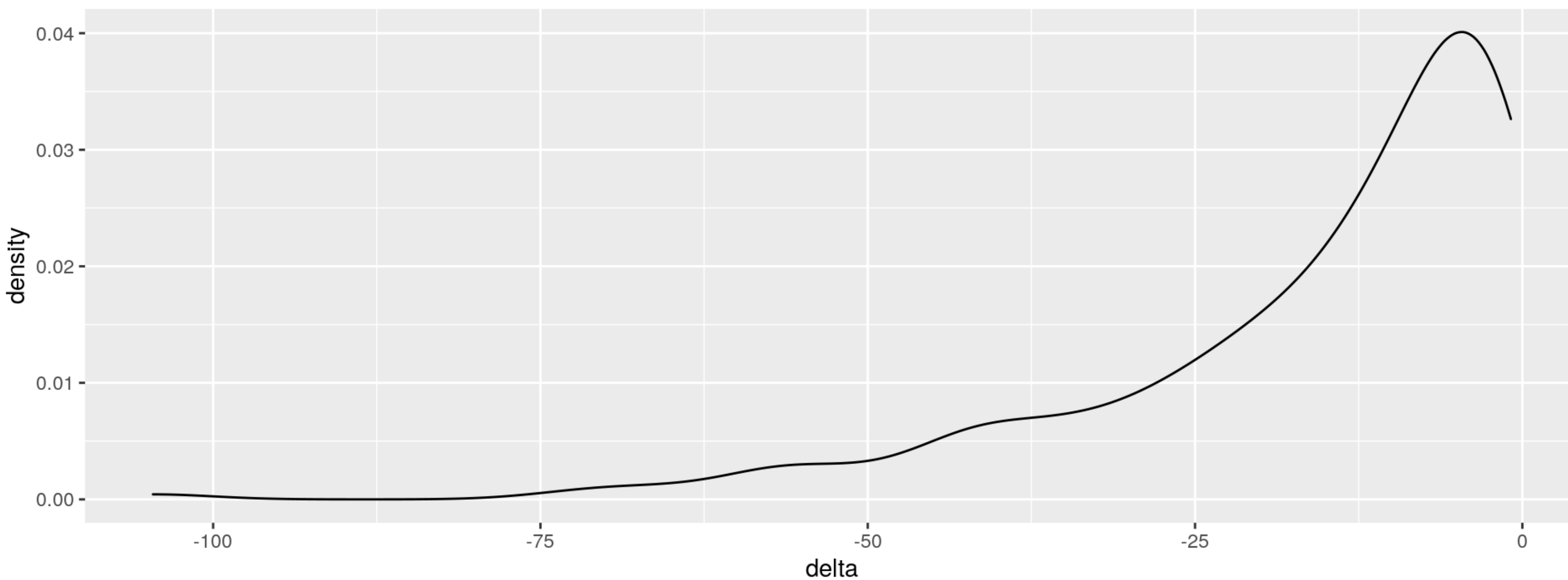
```
post <- update(priors, prior_PD = FALSE)
```

```
print(post, digits = 2)
```

```
...
##             Median MAD_SD
## (Intercept)  1.58   0.04
## senior      -0.46   0.04
## log(roach1)  0.62   0.01
## treatment   -0.49   0.03
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
...
```

Estimating Treatment Effects

```
df <- roaches; df$treatment <- 0
Y_0 <- posterior_epred(post, newdata = df, offset = log(df$exposure2))
df$treatment <- 1
Y_1 <- posterior_epred(post, newdata = df, offset = log(df$exposure2))
ggplot(data.frame(delta = colMeans(Y_1 - Y_0))) + geom_density(aes(x = delta))
```



Numerical Assessment of Calibration

```
PPD <- posterior_predict(post); dim(PPD)
```

```
## [1] 4000  202
```

```
lower <- apply(PPD, MARGIN = 2, FUN = quantile, probs = 0.25)
upper <- apply(PPD, MARGIN = 2, FUN = quantile, probs = 0.75)
mean(roaches$y > lower & roaches$y < upper) # bad fit
```

```
## [1] 0.04950495
```

- Overall, the model is fitting the data poorly
- You will often overfit when you lazily use all predictors that are available in the dataset

Adding Overdispersion

$$\alpha \sim \mathcal{N}\left(m_\alpha, s_\alpha\right)$$
$$\beta_1 \sim \mathcal{N}\left(m_{\beta_1}, s_{\beta_1}\right)$$
$$\beta_2 \sim \mathcal{N}\left(m_{\beta_2}, s_{\beta_2}\right)$$
$$\beta_3 \sim \mathcal{N}\left(m_{\beta_3}, s_{\beta_3}\right)$$
$$\forall n : \eta_n \equiv \alpha + OFFSET_n + \beta_1 \times \log LAG_n + \beta_2 \times SENIOR_n + \beta_3 \times T_n$$
$$\forall n : \mu_n \equiv e^{\eta_n}$$
$$\phi \sim \mathcal{E}\left(r\right)$$
$$\forall n : \epsilon_n \sim \mathcal{G}\left(\phi, \phi\right)$$
$$\forall n : Y_n \sim \mathcal{Poisson}\left(\epsilon_n \mu_n\right)$$

- The conditional distribution of Y_n given μ_n and a Gamma-distributed ϵ_n and is Poisson, but the conditional distribution of Y_n given μ_n irrespective of ϵ_n is negative binomial with expectation μ_n and variance $\mu_n + \mu_n^2/\phi$

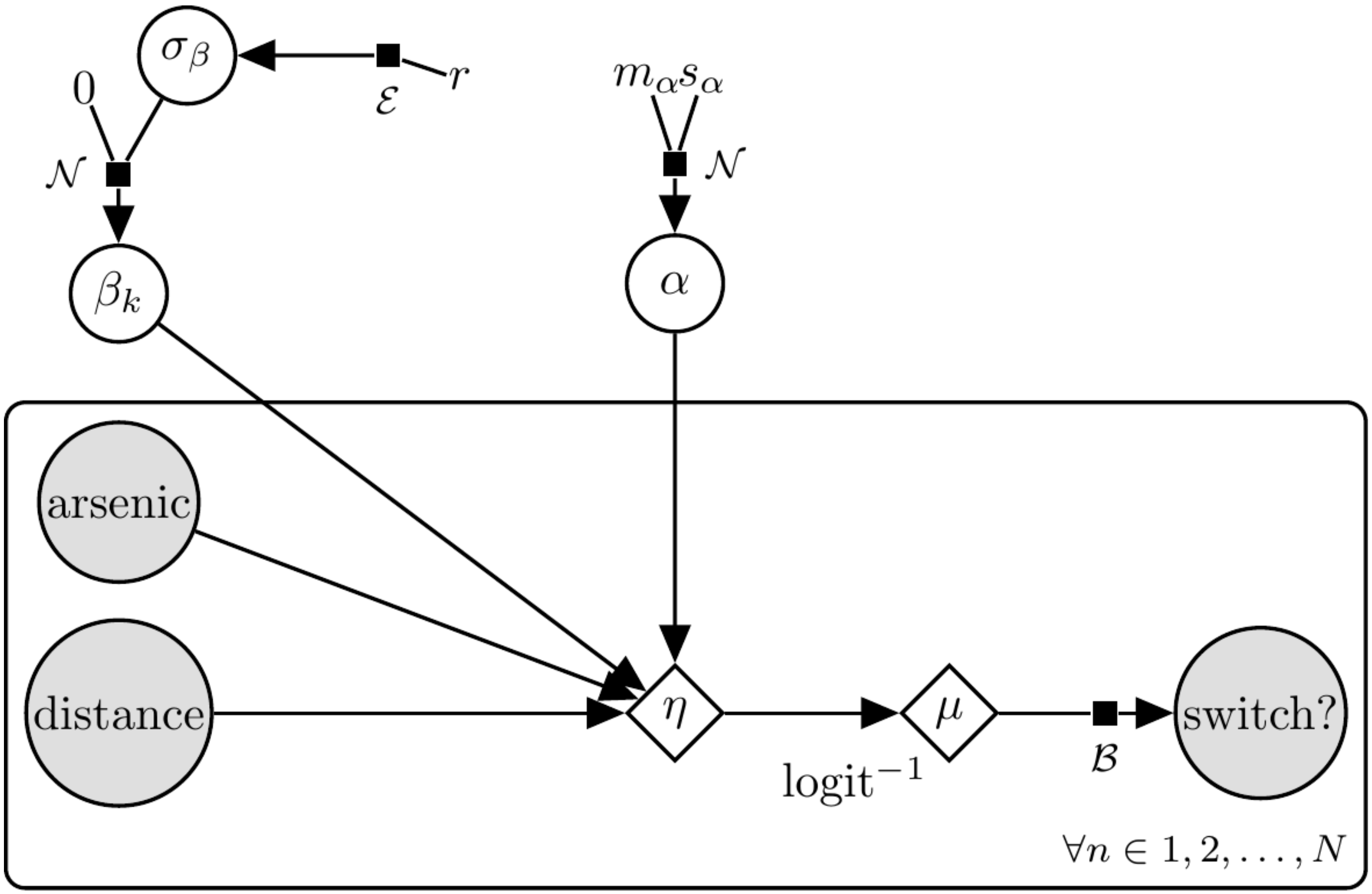
Posterior if Likelihood Is Negative Binomial

```
post <- update(post, family = neg_binomial_2)
```

```
print(post, digits = 2)
```

```
...
##             Median MAD_SD
## (Intercept)  1.33   0.26
## senior      -0.20   0.24
## log(roach1)  0.70   0.07
## treatment   -0.62   0.22
##
## Auxiliary parameter(s):
##                       Median MAD_SD
## reciprocal_dispersion 0.47   0.05
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
...
```

Prior Predictive Distribution for Well Switching



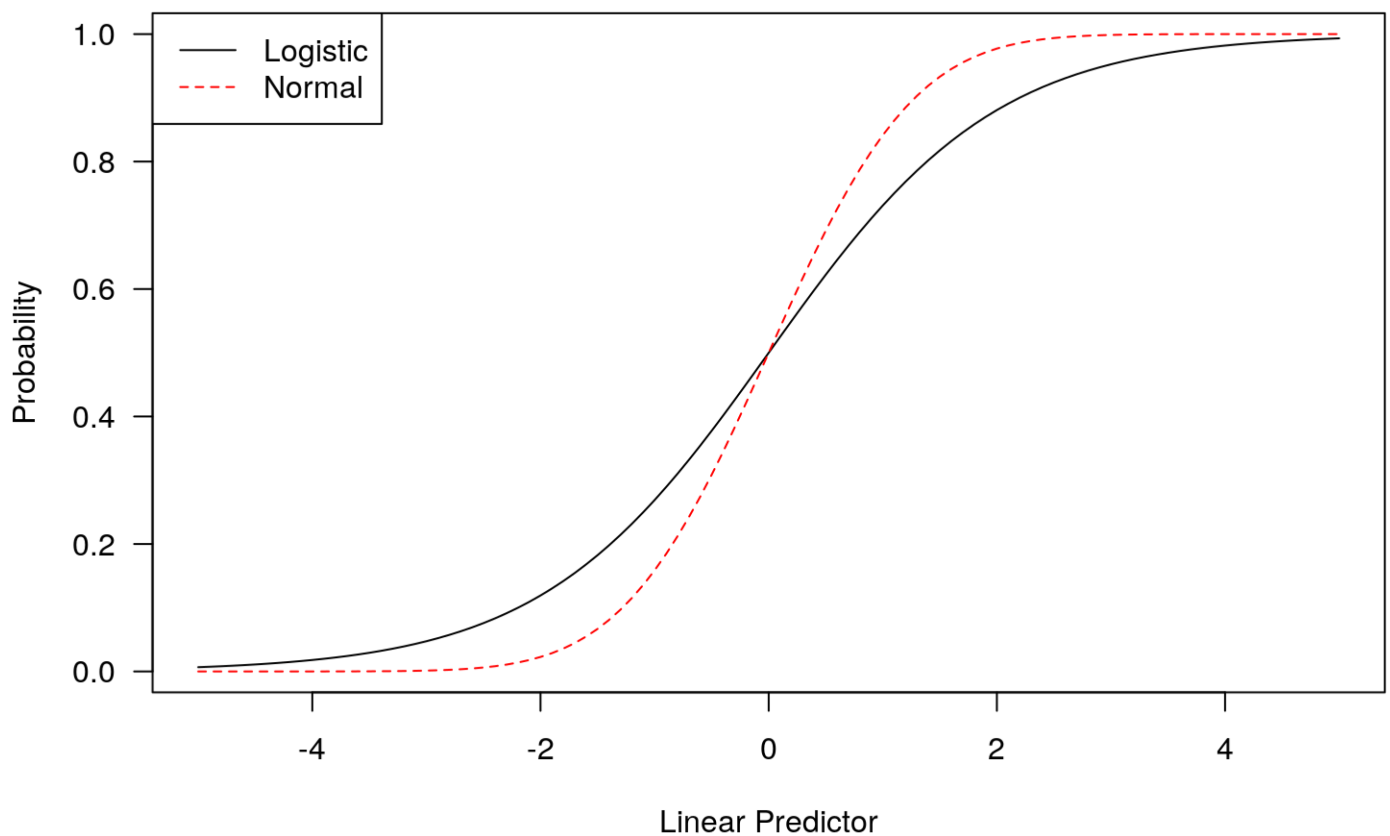
Well Switching Model

Prior Predictive Distribution in Symbols

$$\sigma_\beta :\sim \mathcal{E}\left(r\right)$$
$$\forall k : \beta_k \sim \mathcal{N}\left(0, \sigma_\beta\right)$$
$$\alpha \sim \mathcal{N}\left(m_\alpha, s_\alpha\right)$$
$$\forall n : \eta_n \equiv \alpha + s\left(ARSENIC_n, DISTANCE_n, \beta_1 \dots \beta_K\right)$$
$$\forall n : \epsilon_n \sim \mathcal{L}\left(0, 1\right)$$
$$\forall n : u_n \equiv \eta_n + \epsilon_n$$
$$\forall n : Y_n \equiv u_n > 0$$

- $s\left(\cdot\right)$ is a smooth but non-linear function of arsenic and well-distance that has many coefficients, each of which has a normal prior with expectation zero and standard deviation σ_β, which has an exponential prior with expectation r^{-1}
- $\Pr\left(y_n = 1 \mid \dots\right) = \Pr\left(\eta_n + \epsilon_n > 0\right) = \Pr\left(\epsilon_n > -\eta_n\right) = \Pr\left(\epsilon_n \leq \eta_n\right)$, which can evaluated using the standard logistic CDF, $F\left(\eta_n\right) = \frac{1}{1+e^{-\eta_n}}$

Inverse Link Functions



Posterior Distribution

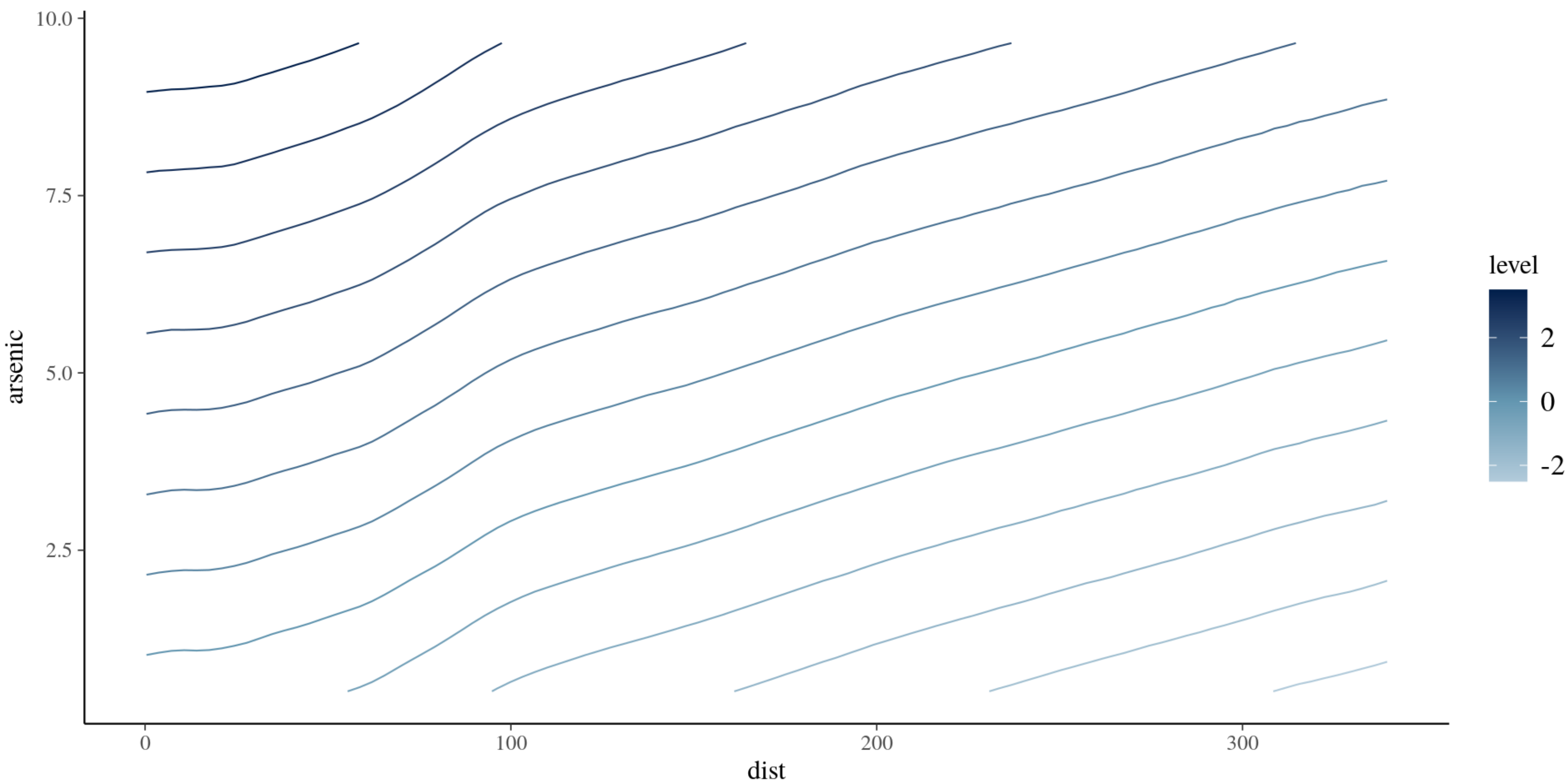
```
post <- stan_gamm4(switch ~ s(dist, arsenic), data = wells, family = binomial, adapt_delta = 0.98)
```

```
print(post, digits = 2)
```

```
...
##                       Median MAD_SD
## (Intercept)            0.33   0.04
## s(dist,arsenic).1     -0.04   0.53
## s(dist,arsenic).2      0.00   0.54
## s(dist,arsenic).3      0.00   0.56
## s(dist,arsenic).4      0.00   0.56
## s(dist,arsenic).5     -0.06   0.52
## s(dist,arsenic).6     -0.01   0.52
## s(dist,arsenic).7     -0.01   0.51
## s(dist,arsenic).8     -0.03   0.56
## s(dist,arsenic).9     -0.07   0.54
## s(dist,arsenic).10    -0.04   0.52
## s(dist,arsenic).11     0.04   0.55
## s(dist,arsenic).12     0.08   0.56
## s(dist,arsenic).13    -0.31   0.62
## s(dist,arsenic).14    -0.23   0.57
## s(dist,arsenic).15     0.03   0.54
## s(dist,arsenic).16     0.04   0.51
## s(dist,arsenic).17    -0.02   0.54
## s(dist,arsenic).18 -0.11   0.55
## s(dist,arsenic).19  0.08   0.51
## s(dist,arsenic).20  0.02   0.42
## s(dist,arsenic).21 -0.04   0.47
## s(dist,arsenic).22 -0.01   0.54
## s(dist,arsenic).23 -0.64   0.50
## s(dist,arsenic).24 -0.20   0.42
## s(dist,arsenic).25 -0.16   0.54
## s(dist,arsenic).26  0.09   0.54
## s(dist,arsenic).27 -0.01   0.43
## s(dist,arsenic).28  7.96   1.06
## s(dist,arsenic).29  6.90   2.11
##
## Smoothing terms:
##                              Median MAD_SD
## smooth_sd[s(dist,arsenic)1] 0.64   0.44
## smooth_sd[s(dist,arsenic)2] 4.58   1.21
##
## ------
## * For help interpreting the printed output see ?print.sta
## * For info on the priors used see ?prior_summary.stanreg
...
```

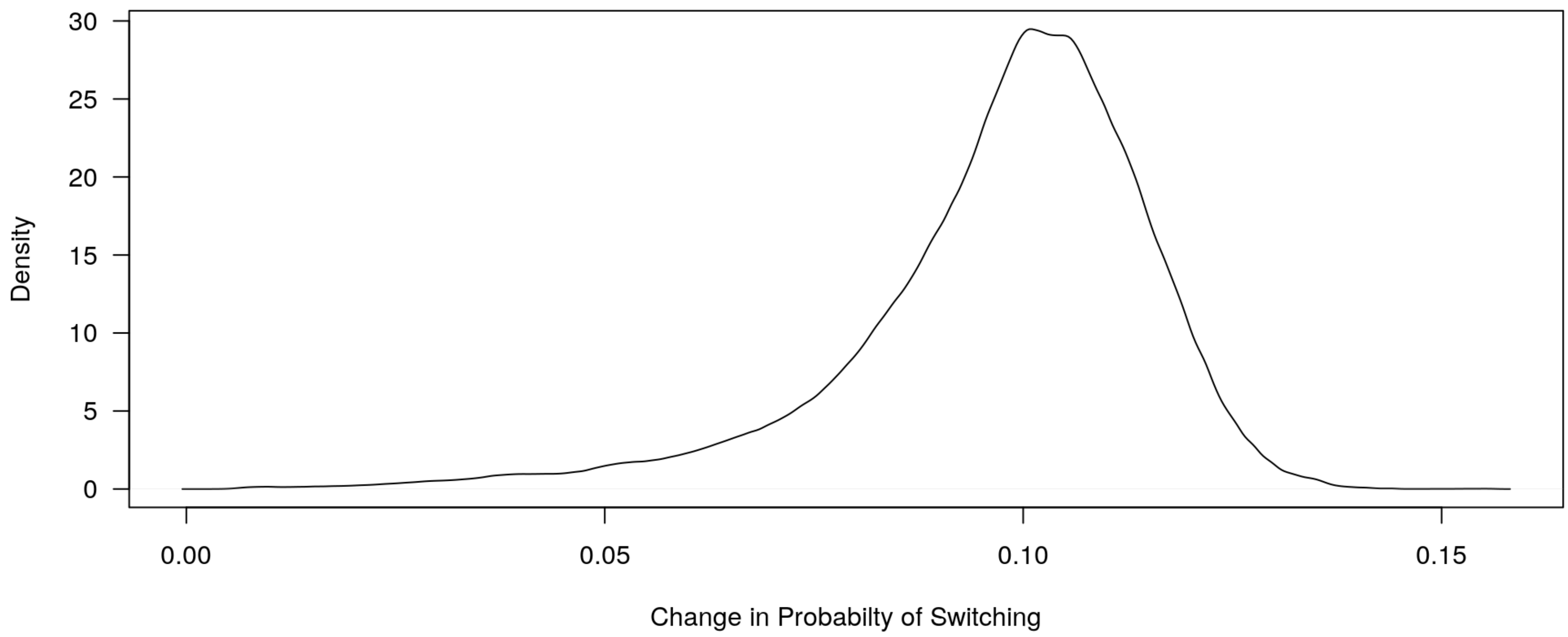
Nonlinear Plot

```
plot_nonlinear(post) # coloring is in log-odds units
```



Plotting the Effect of an Increase in Arsenic

```
mu_0 <- posterior_epred(post)
df <- wells; df$arsenic <- df$arsenic + 1
mu_1 <- posterior_epred(post, newdata = df)
plot(density(mu_1 - mu_0), main = "", xlab = "Change in Probabilty of Switching")
```



A Binomial Model for Romney vs Obama in 2012

```r
poll <- readRDS("GooglePoll.rds") # WantToWin is coded as 1 for Romney and 0 for Obama
library(dplyr)
collapsed <- filter(poll, !is.na(WantToWin)) %>%
             group_by(Region, Gender, Urban_Density, Age, Income) %>%
             summarize(Romney = sum(grepl("Romney", WantToWin)), Obama = n() - Romney) %>%
             na.omit
```

```r
post <- stan_glm(cbind(Romney, Obama) ~ ., data = collapsed, family = binomial(link = "probit"),
                 QR = TRUE, init_r = 0.25)
```

```r
print(post, digits = 2)
```

```
...
##                        Median MAD_SD
## (Intercept)            -0.33   0.09
## RegionNORTHEAST        -0.09   0.06
## RegionSOUTH             0.19   0.04
## RegionWEST             -0.09   0.05
## GenderMale              0.24   0.04
## Urban_DensitySuburban  -0.13   0.06
## Urban_DensityUrban     -0.32   0.06
```

```
## Age25-34                0.07   0.06
## Age35-44                0.33   0.07
## Age45-54                0.52   0.06
## Age55-64                0.53   0.06
## Age65+                  0.83   0.06
## Income25,000-49,999    -0.07   0.05
## Income50,000-74,999    -0.04   0.05
## Income75,000-99,999    -0.06   0.09
## Income100,000-149,999   0.11   0.18
## Income150,000+          0.49   0.58
...
```