Week Ten

Last Week

- Probability Distributions for Count Data
- Count Regression
- Exams (Last - Last Week)

This Week: Generalized Linear Models for Count Data

Today:

- Take Home Exam Recap
- Activity:
  - Model exploration for count data
- Thursday: Lab

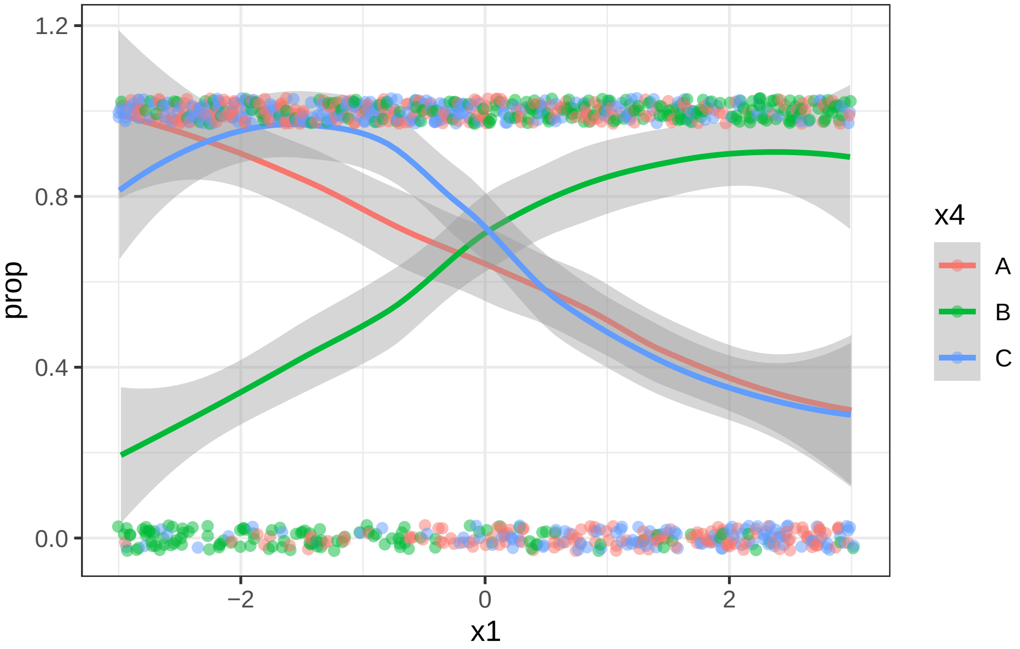
Next Week: Multicategory Regression

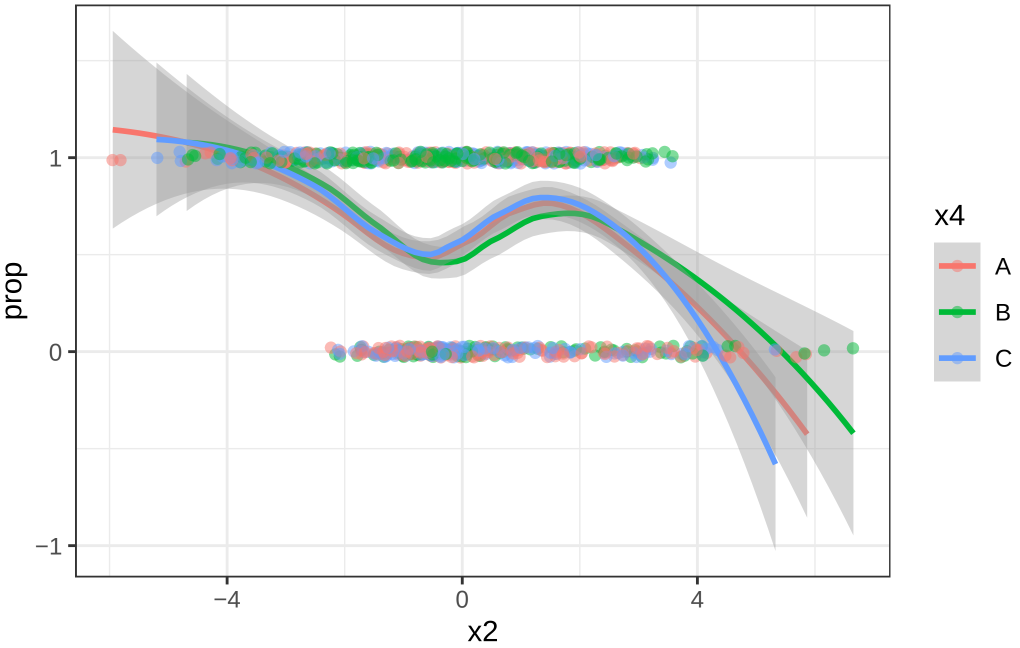
Part II of Exam

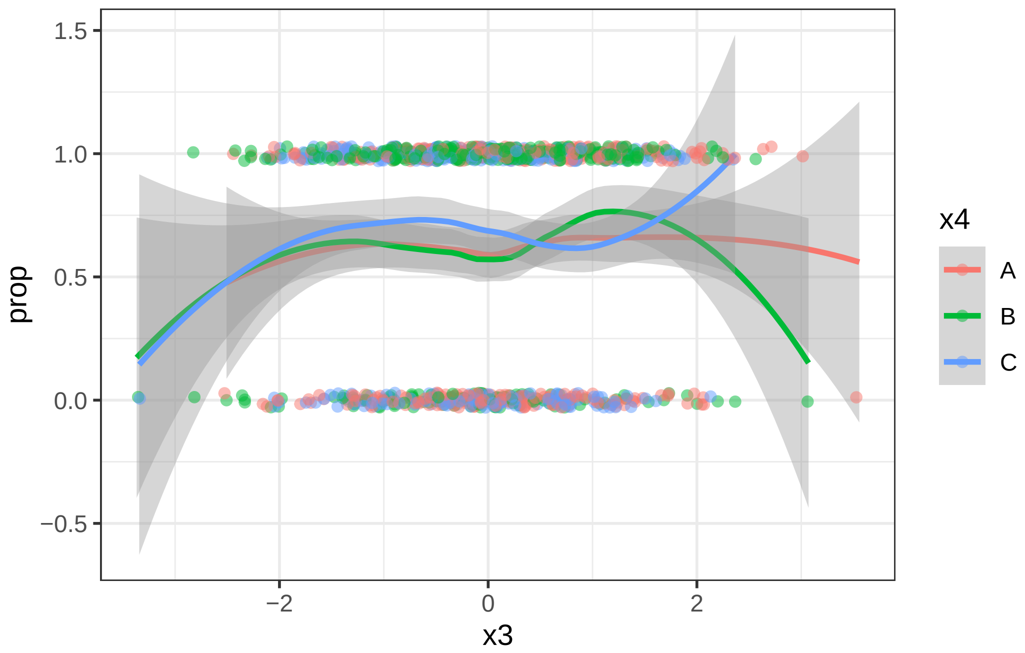
The second part of the exam will involve model fitting with logistic regression. Use the `midterm_data` and note that y is a single binary variable.

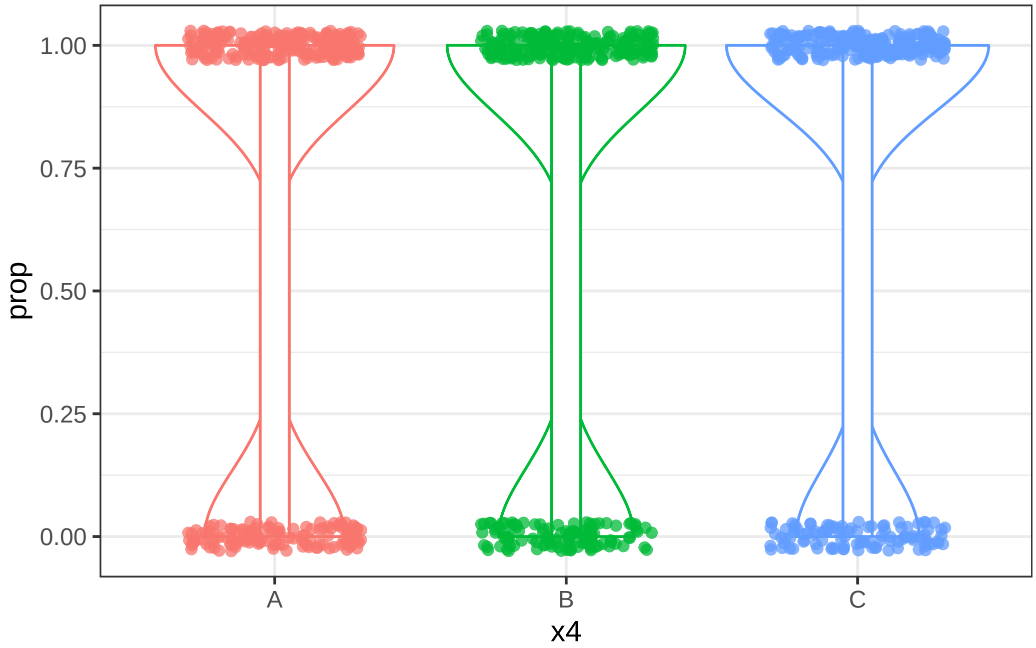
```r
set.seed(10062025)
n <- 1000
x1 <- seq(-3, 3, length.out = n)
x2 <- rnorm(n, sd = 2)
x3 <- rnorm(n)
x4 <- sample(c('A','B','C'), size = n, replace = T)

dat_tibble <- tibble(x1, x2, x3, x4)
X_mat <- model.matrix(~ x1 + x2 + I(x2^2) + I(x2 ^3) + x3 + x4 + x1:x4)

beta0 <- 0
beta1 <- -1
beta2 <- 1
beta2_sq <- .6
beta2_cube <- -.3
beta3 <- 0
beta4b <- 0
beta4c <- 0
beta1_4b <- 2
beta1_4c <- 0

beta_vec <- c(beta0, beta1, beta2, beta2_sq, beta2_cube, beta3, beta4b, beta4c, beta1_4b, be

pi <- invlogit(X_mat %*% beta_vec)

y <- rbinom(n, 1, pi)
midterm_data <- tibble(y = y, x1, x2, x3, x4)
```

4. (4 points) Create a set of EDA figures to explore the relationship between the response (success out of 1 trial) and the potential covariates.


1.2
0.8
0.4
0.0
prop
−2
0
2
x1
x4
A
B
C


1
0
−1
prop
−4
0
4
x2
x4
A
B
C





5. (4 points) Summarize your findings in the figures

Which variables and combinations of variables to you think are important?

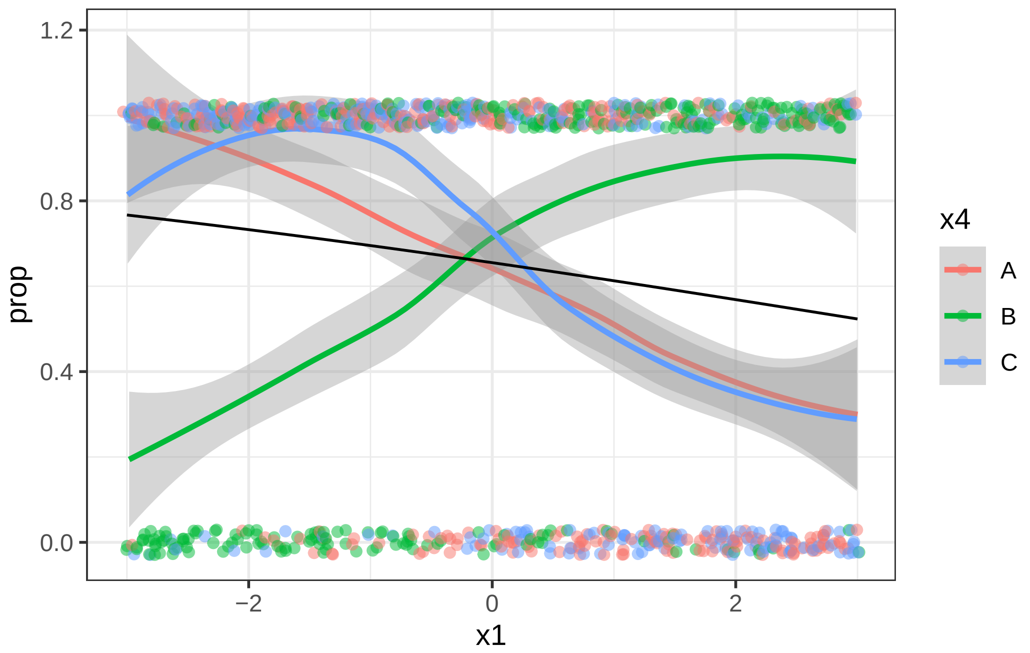
- x_1 tends to have a decreasing relationship with the success probability; however, the relationship is increasing for group B in x_4
- x_2 appears to have a non-linear relationship with the success probability but is relatively consistent across groups of x_4
- x_3 might have a weak relationship with success probability - potentially quadratic although there is a lot of uncertainty in model fits in the tails of x_3.
- x_4 doesn't seem to be particularly meaningful alone

6. (4 points) Using residual diagnostics and AIC fit a series of models.

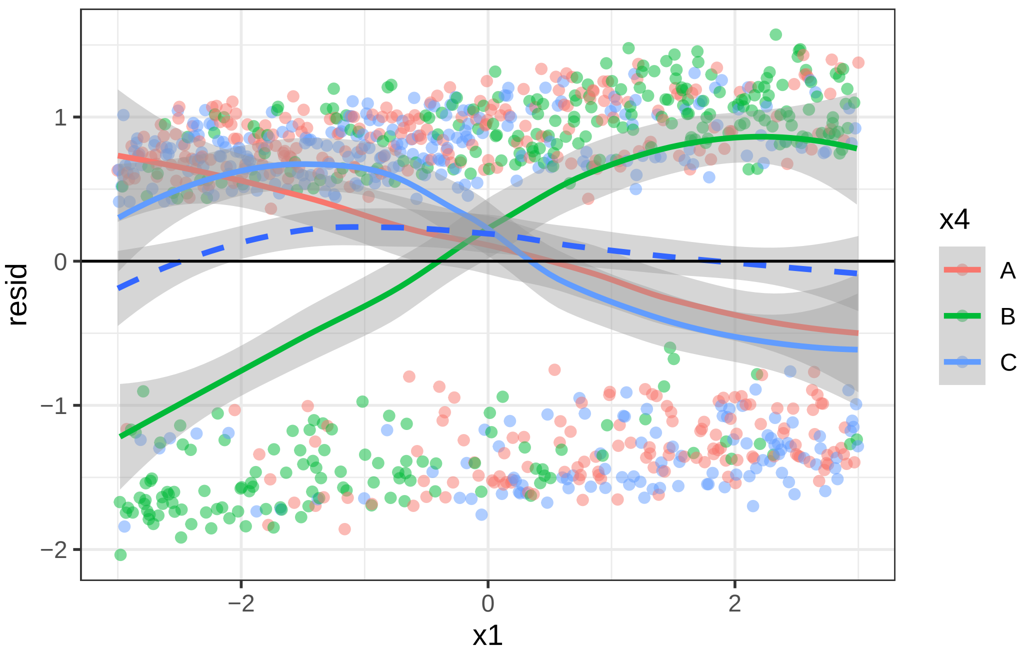
You don't need to print out all of these results, but include a written summary of models you explored. You are welcome to use bullet points for this section.

```r
start_model <- glm(y ~ x1 + x2 + x3 + x4, data = midterm_data, family = binomial)
```

```r
fit_data <- tibble(x1 = midterm_data$x1,
                   y = midterm_data$y,
                   x4 = midterm_data$x4,
                   prop = invlogit(coef(start_model)[[1]] + coef(start_model)[[2]] * x1))

midterm_data |>
  mutate(prop = y )|>
  ggplot(aes(y = prop, x = x1, color = x4)) +
  geom_smooth(method = 'loess', formula = 'y ~ x') +
  geom_jitter(alpha = .5, height = .03, width = .03) +
  theme_bw() +
  geom_line(data = fit_data, color = 'black')
```



```
tibble(x1 = midterm_data$x1,
       resid =rstandard(start_model),
       x4 = midterm_data$x4) |>
  ggplot(aes(y = resid, x = x1, color = x4)) +
  geom_jitter(alpha = .5) +
  geom_smooth(method = 'loess', formula = 'y ~ x') +
  geom_smooth(method = 'loess', formula = 'y ~ x', inherit.aes = F, aes(y=resid, x = x1), lin
  theme_bw() +
  geom_hline(yintercept = 0)
```



7. (4 points) Graphically summarize the final model you selected

Include estimated model fits for all parameters or combinations of parameters included in your model.

8. (4 points) Written summary the final model you selected

Describe the final model you selected and discuss how each variable (or combination of variables) impact the probability of success.