HW2

HW2

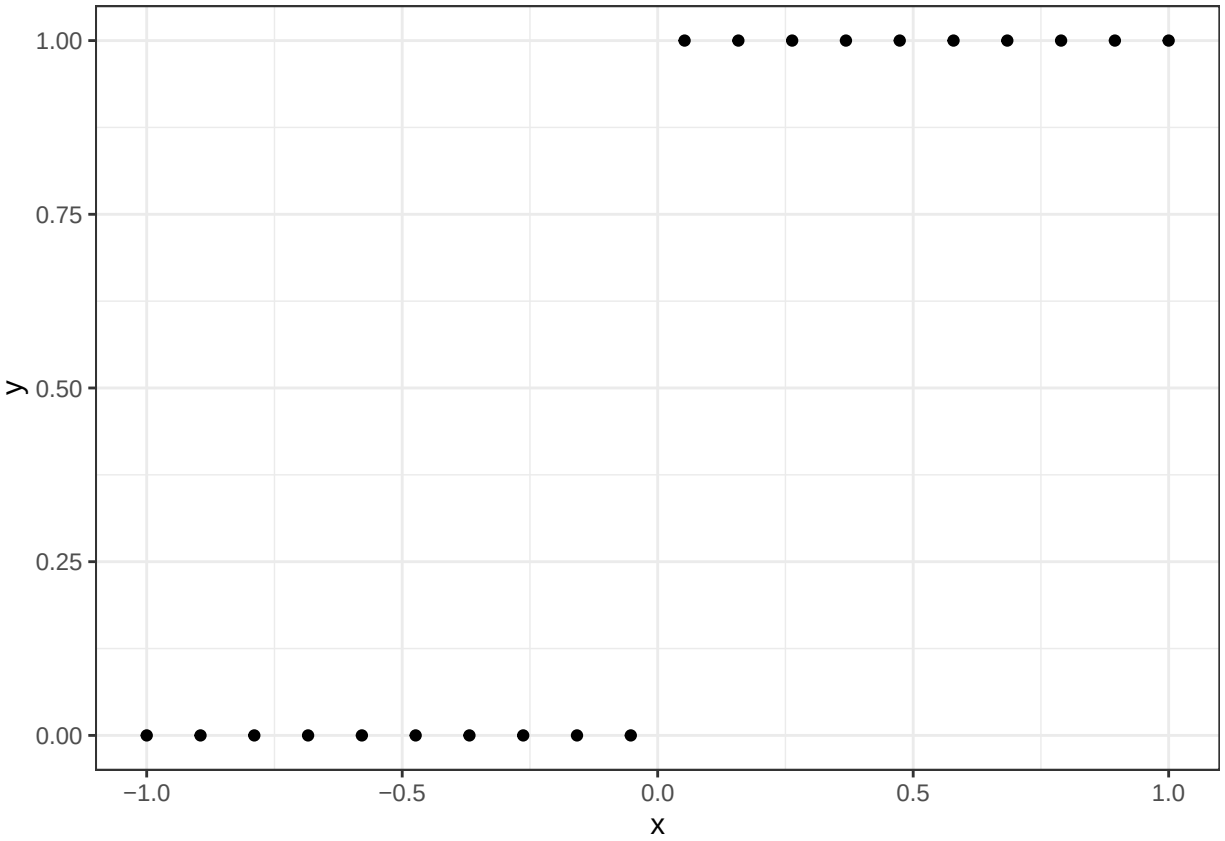
Q1. (4 points)

With binary regression, "separation" is a common problem. This occurs when a continuous predictor is perfectly separated with all zeros below a certain point and all zeros above a certain point. See the simulated data below for an example.

```
x <- seq(-1, 1, length.out = 20)
y <- rep(c(0,1), each = 10)

df_sep <- tibble(x=x, y=y)

df_sep %>% ggplot(aes(y=y, x=x)) + geom_point() + theme_bw()
```



Using both `glm` and `stan_glm` with a probit link to fit the data. Identify the differences in the model output and discuss why they might differ.

Q2. (4 points)

Revisiting STAT505 HW 9, Q3.

```
candy <- read_csv("https://math.montana.edu/ahoegh/teaching/stat446/candy-data.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   competitorname = col_character(),
##   chocolate = col_double(),
##   fruity = col_double(),
##   caramel = col_double(),
##   peanutyalmondy = col_double(),
##   nougat = col_double(),
##   crispedricewafer = col_double(),
##   hard = col_double(),
##   bar = col_double(),
##   pluribus = col_double(),
##   sugarpercent = col_double(),
##   pricepercent = col_double(),
##   winpercent = col_double()
## )
```

Use the candy dataset and a probit regression model to analyze the relationship between one or two predictors in the dataset and the outcome (whether the candy contains chocolate). Describe how each input affects Pr[chocolate = 1].

Q3. (6 points)

Using a dataset with Yelp scores in Madison, Wisconsin, model the probability of 1-star, 2-star, 3-star, 4-star, and 5-star reviews as a function of the two included neighborhoods. Summarize the model parameters and make a graphic / table (preferably a graphic) to display your results.

```
set.seed(01312021)
yelp_biz <- read_csv("https://math.montana.edu/ahoegh/teaching/stat532/data/yelp_biz_info.csv") %>%
  filter(neighborhood %in% c("South Campus",'Williamson - Marquette'))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   business_id = col_character(),
##   name = col_character(),
##   neighborhood = col_character(),
##   address = col_character(),
##   city = col_character(),
##   state = col_character(),
##   postal_code = col_double(),
##   latitude = col_double(),
##   longitude = col_double(),
##   categories = col_character()
## )
```

```
yelp_reviews <- read_csv("https://math.montana.edu/ahoegh/teaching/stat532/data/yelp_biz_reviews.csv")
```

```
##
```

```
## -- Column specification --------------------------------------------------------
## cols(
##   review_id = col_character(),
##   user_id = col_character(),
##   business_id = col_character(),
##   stars = col_double(),
##   date = col_date(format = "")
## )
```

```
yelp_comb <- yelp_reviews %>%
  right_join(yelp_biz, by = 'business_id') %>%
  mutate(stars = factor(stars))
```