If you're interested in following along with the data example from today's lecture (not required!), please go to:

https://rstudio.cloud/content/4211518

If you haven't used RStudio Cloud before, you will need to:

- Create an account with RStudio (probably easiest to do this through gmail?).
- Install the tidyverse and tableone packages.
- Save a permanent copy of the project, if you want to be able to return to it later. (Should be top right on your screen, next to the bright red "TEMPORARY COPY" words.)

Again, this is NOT required to be able to follow the lecture. Just if you'd find it helpful to your learning.

# Logistic Regression

## BDSI 2022

**Elizabeth Chase, June 28, 2022**

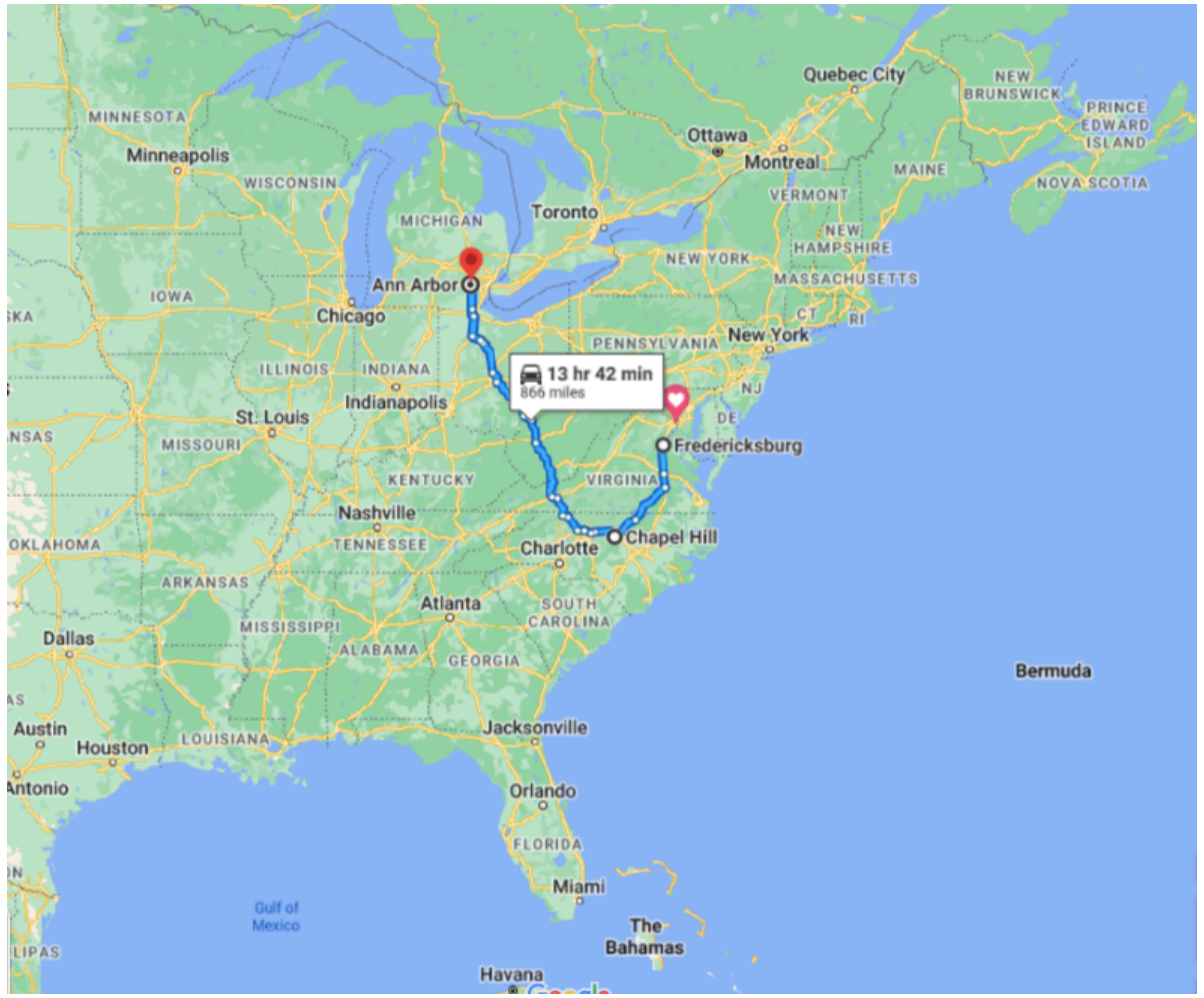
# Learning Objectives

By the end of this lecture, students will be able to:

- Identify when logistic regression can and should be used.
- Understand the structure of a logistic regression model and its high-level connections to linear models.
- Comfortably interpret odds, odds ratios, log odds, and log odds ratios.
- Be able to fit a logistic regression in R and interpret its output.
- Recognize separation of a logistic regression, and know some simple solutions to it.

# A Bit About Me

- Just finished the 5th year of my PhD in biostatistics
- Graduated from UNC-Chapel Hill in 2017
- Originally from Fredericksburg, VA
- I work with Jeremy Taylor and Phil Boonstra on Bayesian methods for complex longitudinal data analysis
- ecchase@umich.edu




# ICPSR Data

- Inter-university Consortium for Political and Social Research (ICPSR) has a ton of great datasets, many of which are publicly accessible (even if you don't go to U-M)!

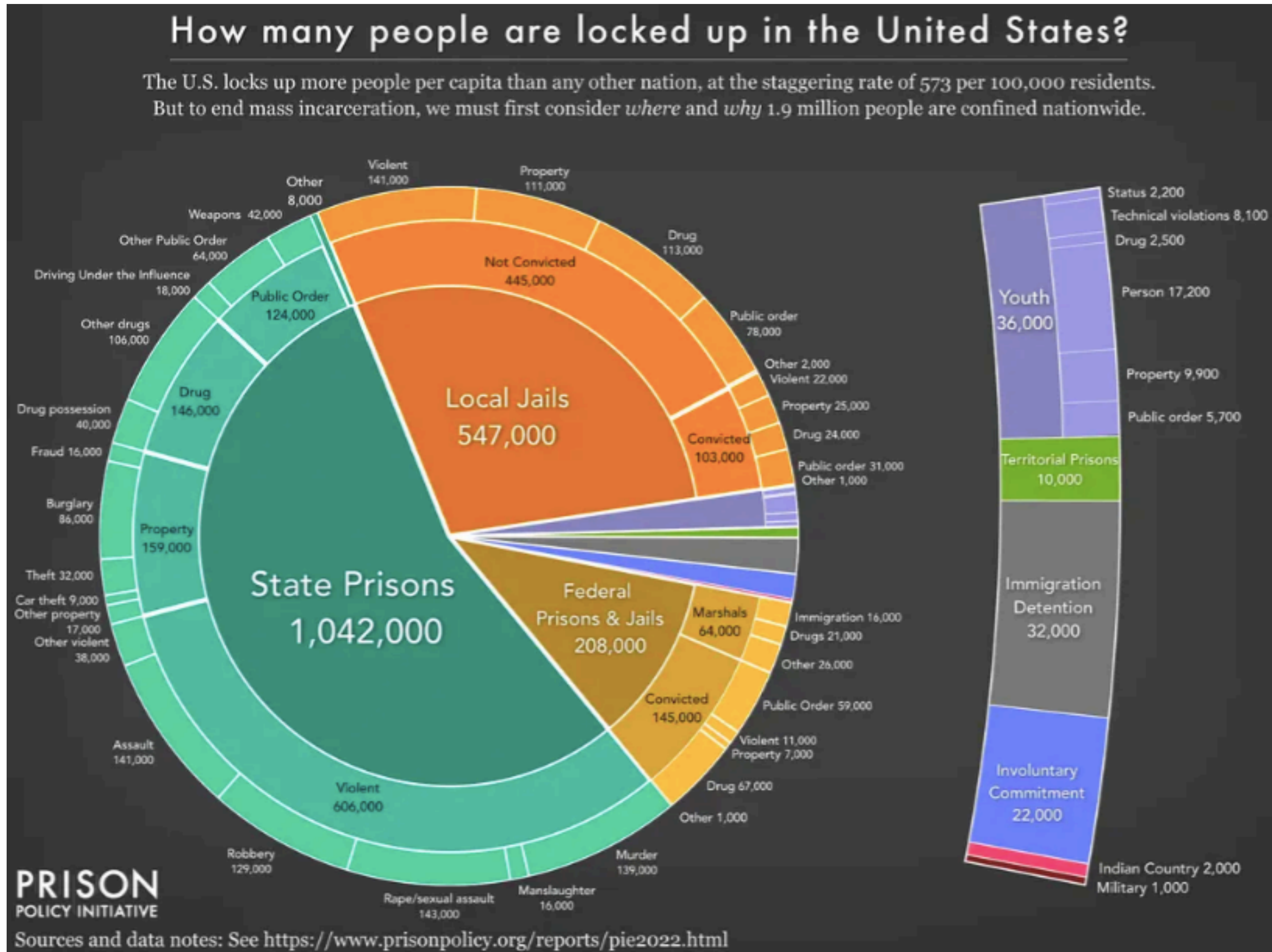
- Highly recommend checking it out if you need a dataset for a class project or are just curious.

- https://www.icpsr.umich.edu/web/pages/

- Today's dataset comes from ICPSR.

# Motivating Question

- We will analyze data from the 2016 Survey of Prison Inmates, United States
- The U.S. imprisons a lot of people.
  - This survey covers the **1,502,671** adult U.S. prisoners in 2016, held at **2,001** unique prisons.
  - Excludes local jails and prisons operated exclusively by the U.S. military, Immigration & Customs Enforcement, U.S. Marshals, and American Indian correctional authorities.

# Motivating Question

- Every 5 years, the Bureau of Justice Statistics conducts a representative sample of all of its prisoners ages 18+ in state or federal correctional facilities:
  - Stage 1: sampled 364 prisons, stratified by:
    - Sex housed
    - State vs. federal facility
    - State
  - Stage 2: sample 24,848 prisoners within sampled prisons from Stage 1.
- Response rate on the prisoner level was 70.0% and 98.4% on the prison level.
- Interviewed in-person by trained interviewers.

# Motivating Question

- We’re going to be exploring predictors of depression among prisoners.
- Question asked: How often during the past 30 days did you feel so depressed that nothing could cheer you up? Responses:
    - Depressed: replied “all of the time” or “most of the time”
    - Not depressed: replied “some of the time”, “a little of the time”, or “none of the time”
- Predictor of interest: expected year of release

prison_data_analysis.R

```r
library(tidyverse)
library(tableone)

#First, we load the data:
load("~/Desktop/BDSI/prison_data.RData")

#We're going to filter to prisoners who have a release date:
prison_dat <- filter(prison_dat, release_year > 0)

#We generate some descriptives on key predictors, stratified by our outcome of
#interest--whether the prisoner is frequently depressed:
table_com <- CreateTableOne(vars=c("hispanic", "white", "black", "am_indian",
                                   "asian", "hi_pi", "other_race",
                                   "sex_birth", "sex_orientation", "education",
                                   "offense_type", "authority_held", "have_release_date",
                                   "have_children", "diag_depression", "counseling_prison",
                                   "seen_doctor_prison", "job_training_prison",
                                   "education_prison", "work_assignment",
                                   "work_assignment_req"),
                            strata=c("freq_depressed"),
                            data=prison_dat,
                            argsNormal = list(NULL),
                            argsNonNormal = list(var.equal = TRUE))

test <- print(table_com, contDigits=1, printToggle=FALSE, showAllLevels = TRUE)
```

| | Not Depressed | Depressed |
|---|---|---|
| **Year of Release (mean (SD))** | 2020.3 (7.1) | 2021.1 (8.6) |

**$p < 0.001$**

# Linear Regression

Denote $y_i = 1$ if prisoner $i = 1,...,n$ is depressed and $y_i = 0$ otherwise. Let $x_i$ be prisoner $i$'s expected year of release.

Then our linear regression model is:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

where $\epsilon_i \sim N(0,\sigma^2)$ is some normally distributed error.

Assumptions:

# Linear Regression

Denote $y_i = 1$ if prisoner $i = 1,\ldots,n$ is depressed and $y_i = 0$ otherwise. Let $x_i$ be prisoner $i$'s expected year of release.

Then our linear regression model is:

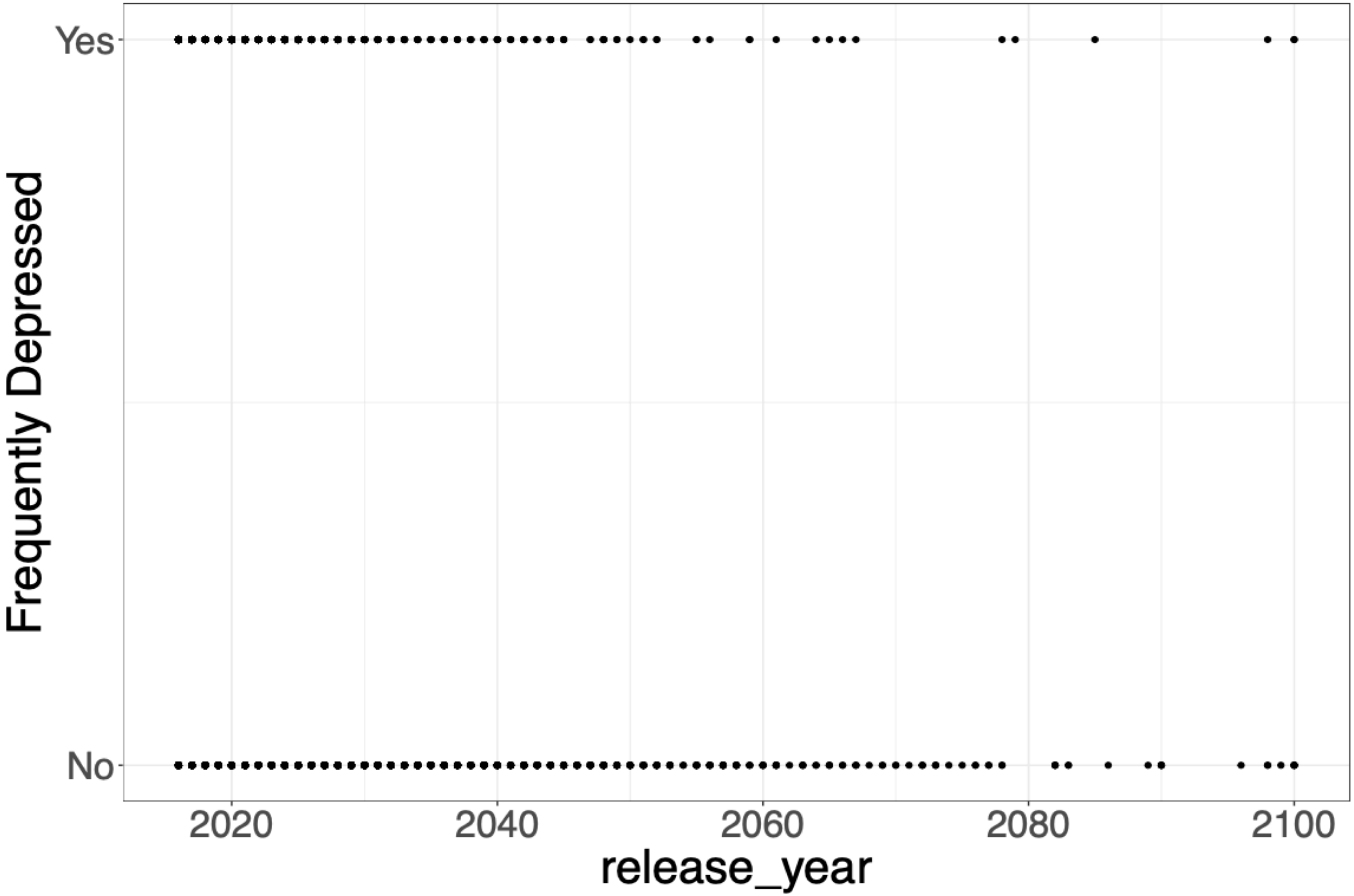
$$y_i = \alpha + \beta x_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ is some normally distributed error.

Assumptions:

- Linear: $\alpha + \beta x_i$ form is appropriate
- Independence: $y_i, y_j,\ i \neq j$ do not influence each other
- Homoscedastic: $\sigma^2$ is the same for everyone
- Normally distributed

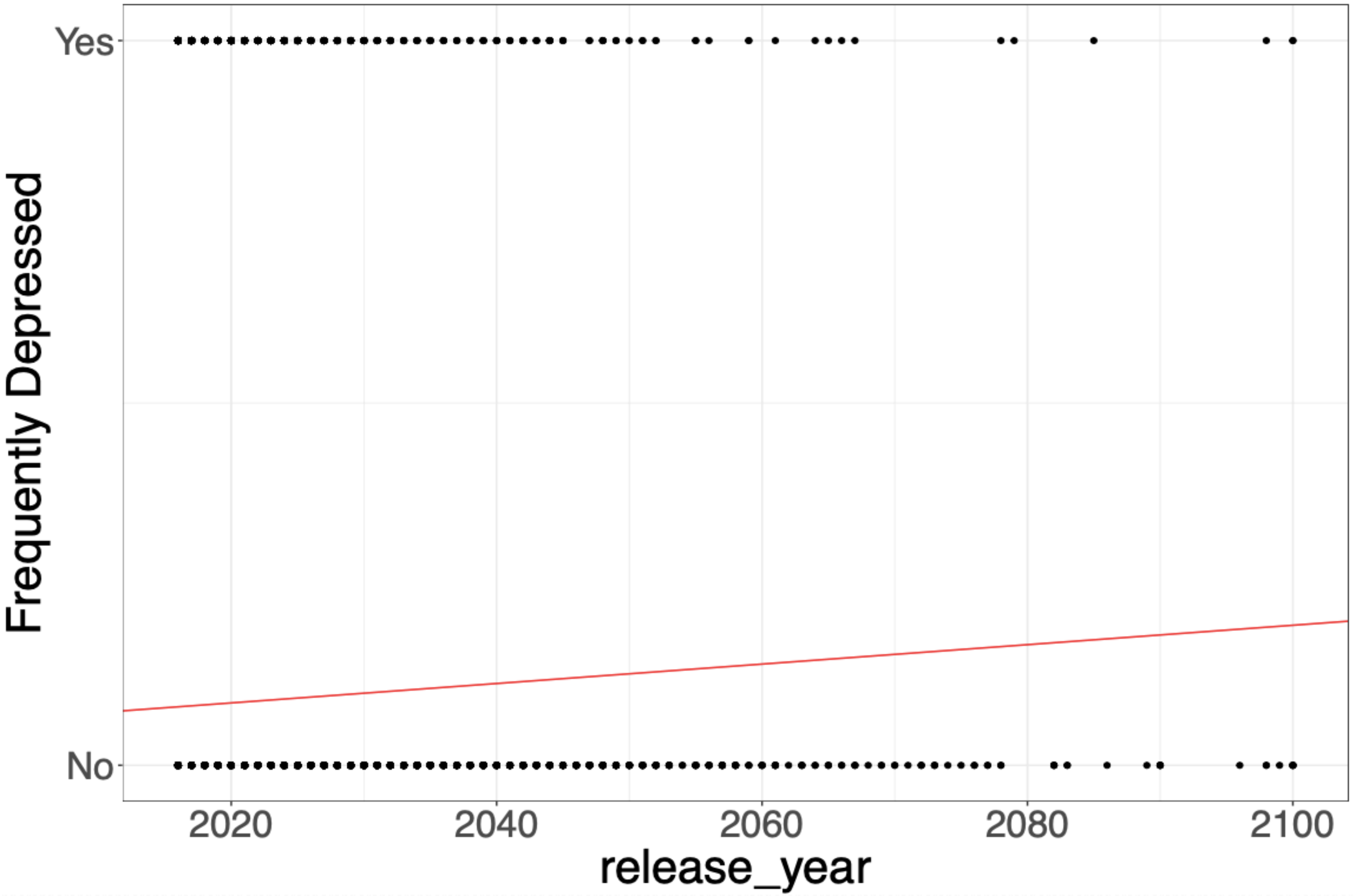
```r
#We try fitting a linear model.
#Let's look at the relationship between expected year of release and
#being depressed:
depressed_release_year <- ggplot(data = prison_dat) +
  geom_point(aes(x = release_year, y = as.numeric(freq_depressed=="Yes"))) +
  theme_bw() + ylab("Frequently Depressed") +
  scale_y_continuous(breaks = c(0, 1), labels = c("No", "Yes"))

depressed_release_year

```


Yes
No
Frequently Depressed
2020
2040
2060
2080
2100
release_year

```r
#We fit the model:
lin_model <- lm((freq_depressed=="Yes") ~ release_year,
                data = prison_dat)

summary(lin_model)
```

```
Call:
lm(formula = (freq_depressed == "Yes") ~ release_year, data = prison_dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.19322 -0.08741 -0.08205 -0.08071  0.91929 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -2.6195124  0.6523416  -4.016 5.96e-05 ***
release_year  0.0013394  0.0003229   4.148 3.37e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.281 on 14528 degrees of freedom
Multiple R-squared:  0.001183,	Adjusted R-squared:  0.001114 
F-statistic: 17.21 on 1 and 14528 DF,  p-value: 3.37e-05
```

```r
#We plot predictions from that model:
depressed_release_year + geom_abline(slope = coef(lin_model)[2],
                                     intercept = coef(lin_model)[1], col = "red")

```



# An Alternative to Linear Regression

A different way to write our linear regression model is:

$$y_i | x_i \sim N(\alpha + \beta x_i, \sigma^2)$$

What's a more appropriate way to describe the distribution of $y_i$?

# An Alternative to Linear Regression

A different way to write our linear regression model is:

$$y_i | x_i \sim N(\alpha + \beta x_i, \sigma^2)$$

What's a more appropriate way to describe the distribution of $y_i$?

$$y_i | x_i \sim Bernoulli(\pi(x_i))$$

where $\pi(x_i)$ denotes the probability that $y_i = 1$, conditional on $x_i$.

# Modeling $\pi(x_i)$

So our new task is coming up with an appropriate way to model $\pi(x_i)$.

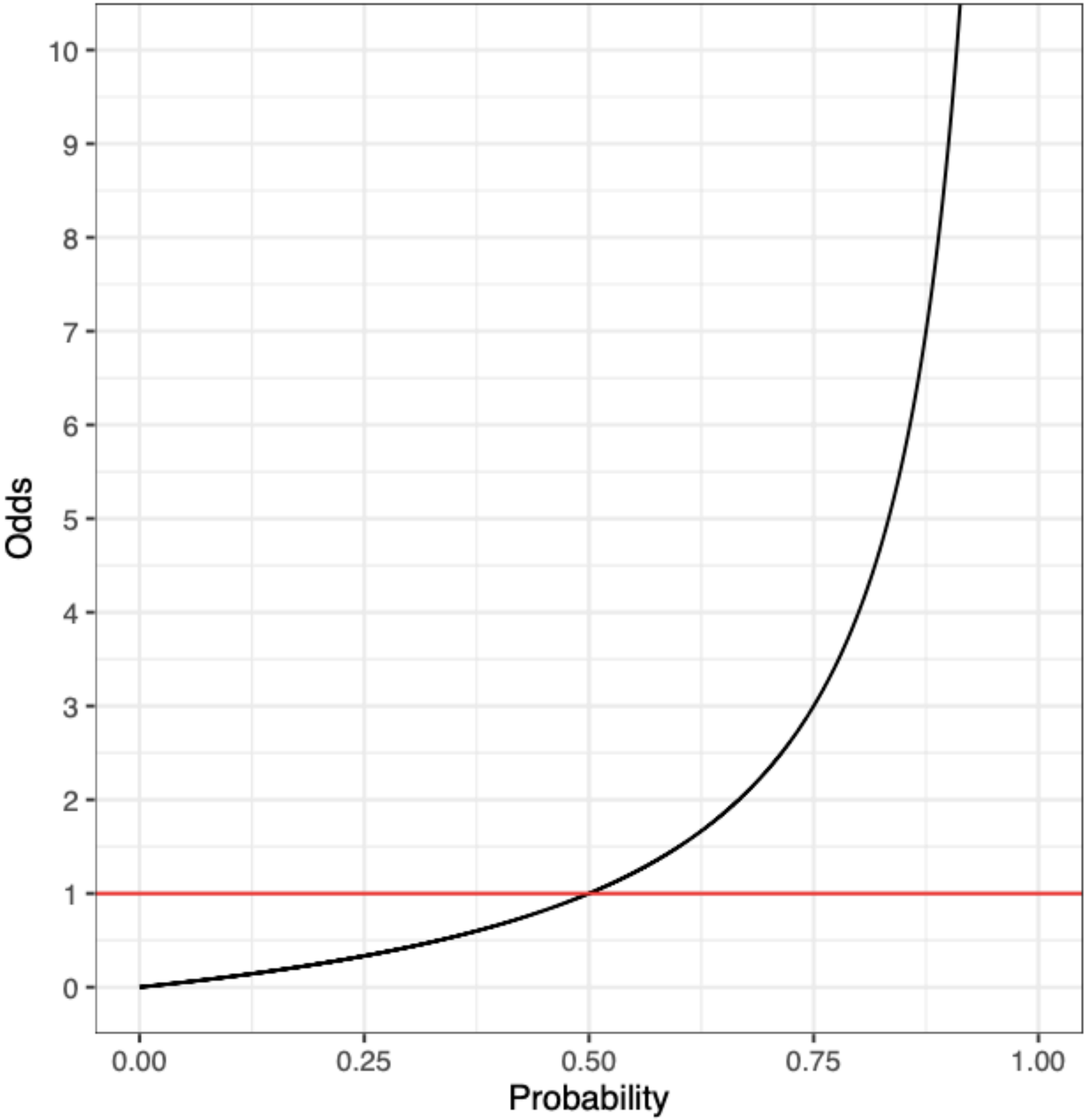
- Bounded between 0 and 1, so $\pi(x_i) = \alpha + \beta x_i$ is not going to cut it.
- What if we come up with some transformation $g(\pi(x_i))$ so that $g(\pi(x_i))$ has range $(-\infty, \infty)$?
- The logistic function is the most popular choice, which has the form:

$$g(\pi(x_i)) = logit(\pi(x_i)) = ln(\frac{\pi(x_i)}{1 - \pi(x_i)})$$

# Modeling $\pi(x_i)$

What the heck is $logit(\pi(x_i)) = ln(\frac{\pi(x_i)}{1-\pi(x_i)})$ ?!

- Inside of the parentheses, we have the odds: $\frac{\pi(x_i)}{1-\pi(x_i)}$
- The odds range from $[0, \infty)$.


Odds
10
9
8
7
6
5
4
3
2
1
0
0.00
0.25
0.50
0.75
1.00
Probability

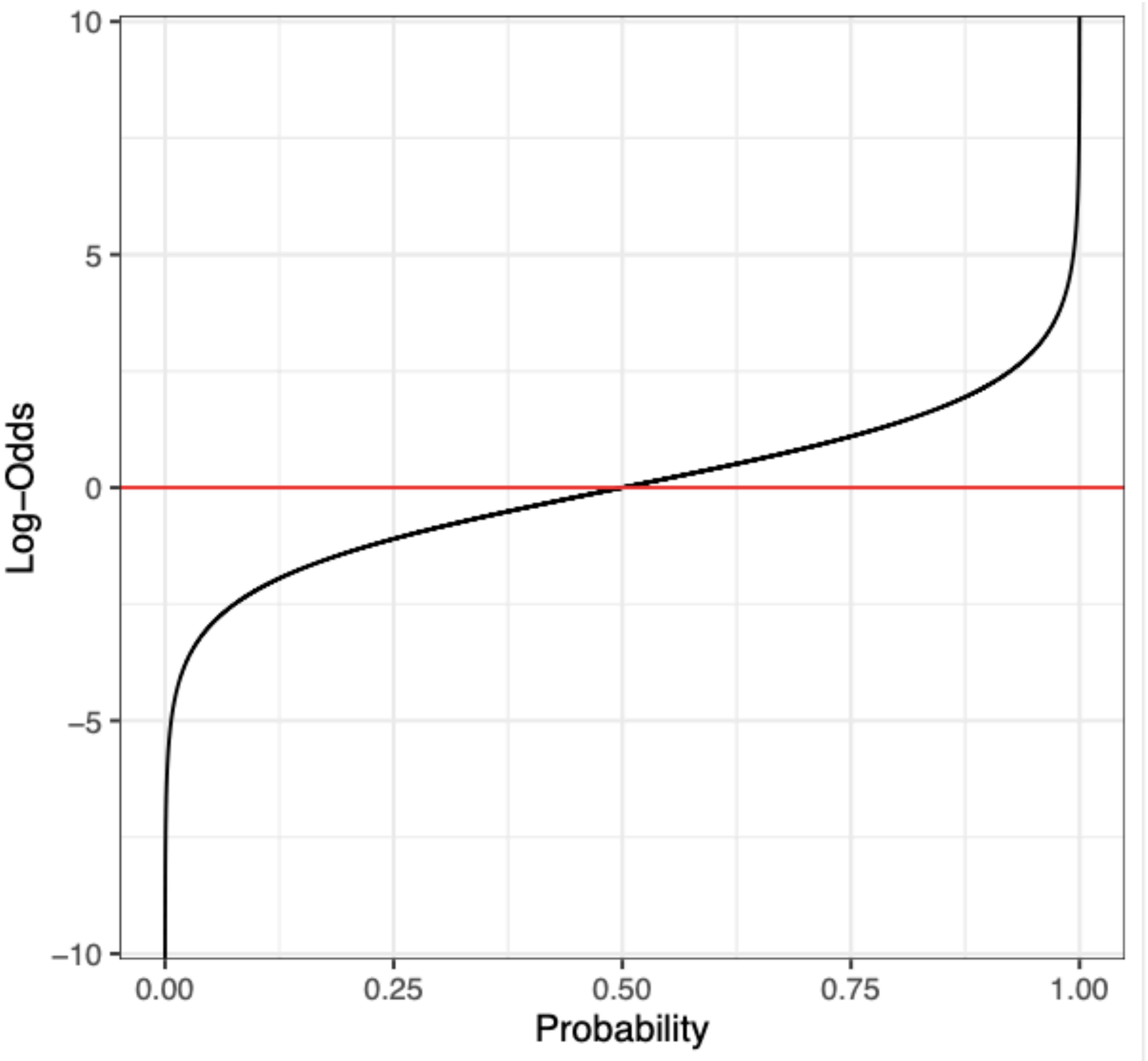
# Modeling $\pi(x_i)$

Then we take the log of the odds, and because of how the log works:

$$ln((0,\infty)) \rightarrow (-\infty, \infty)$$

So now we can use our standard linear form!

$$logit(\pi(x_i)) = \alpha + \beta x_i$$

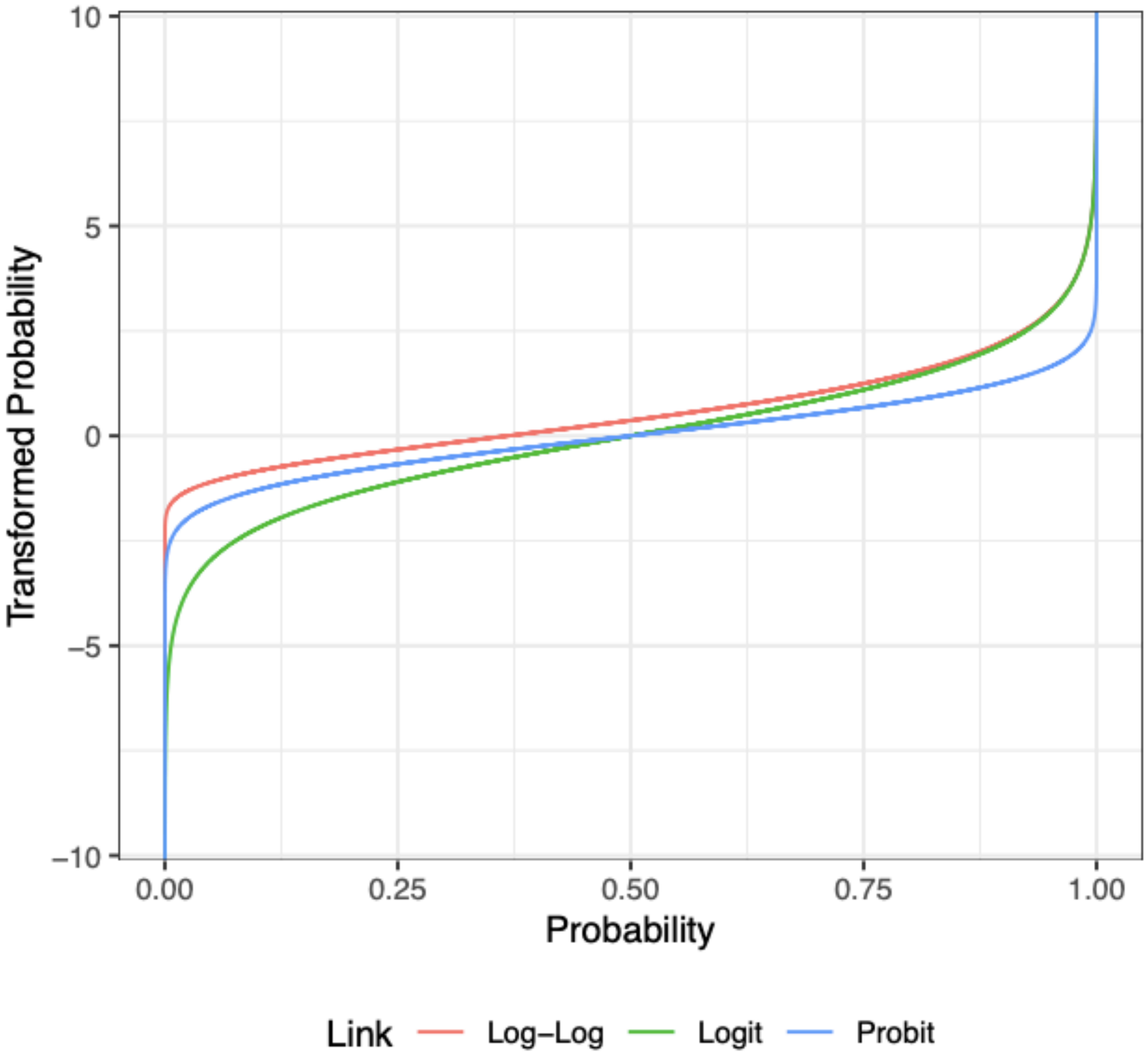


# Side Note

We don't have to use the logit link! Although it's a popular choice and works well, any transformation from

$$[0,1] \to (-\infty, \infty)$$

will do the trick. Here are some other popular choices.



# Logistic Regression Model

Thus our logistic regression model takes the form:

$$ln(\frac{\pi(x_i)}{1 - \pi(x_i)}) = \alpha + \beta x_i$$

$$y_i \sim Bern(\pi(x_i))$$

Note that $E(y_i | x_i) = \pi(x_i)$ because of the properties of the Bernoulli distribution, so we could also write our model as:

$$ln(\frac{E(y_i | x_i)}{1 - E(y_i | x_i)}) = \alpha + \beta x_i$$

# Logistic Regression Model

Still need to make the assumptions:

- Linear form is appropriate for modeling $logit(\pi(x_i))$
- Independence between observations
- $y_i \sim Bernoulli(\pi(x_i))$

```r
#Fit a logistic regression:
logistic_model <- glm((freq_depressed=="Yes") ~ release_year,
                      data = prison_dat, family = "binomial")

summary(logistic_model)
```

```
Call:
glm(formula = (freq_depressed == "Yes") ~ release_year, family = "binomial",
    data = prison_dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7130  -0.4264  -0.4150  -0.4123   2.2395

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -30.804776   6.918468  -4.453 8.49e-06 ***
release_year  0.014078   0.003423   4.113 3.91e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8555.0  on 14529  degrees of freedom
Residual deviance: 8539.8  on 14528  degrees of freedom
AIC: 8543.8

Number of Fisher Scoring iterations: 5
```

# Interpreting a Logistic Regression Model

Let's consider what happens with a one-unit change in $x_i$:

# Interpreting a Logistic Regression Model

We just found that

$$exp(\beta) = \frac{\pi(x_i + 1)/(1 - \pi(x_i + 1))}{\pi(x_i)/(1 - \pi(x_i))}$$

which we recognize as:

$$exp(\beta) = \frac{odds(x_i + 1)}{odds(x_i)}$$

We refer to this quantity as the **odds ratio.**

# Interpreting a Logistic Regression Model

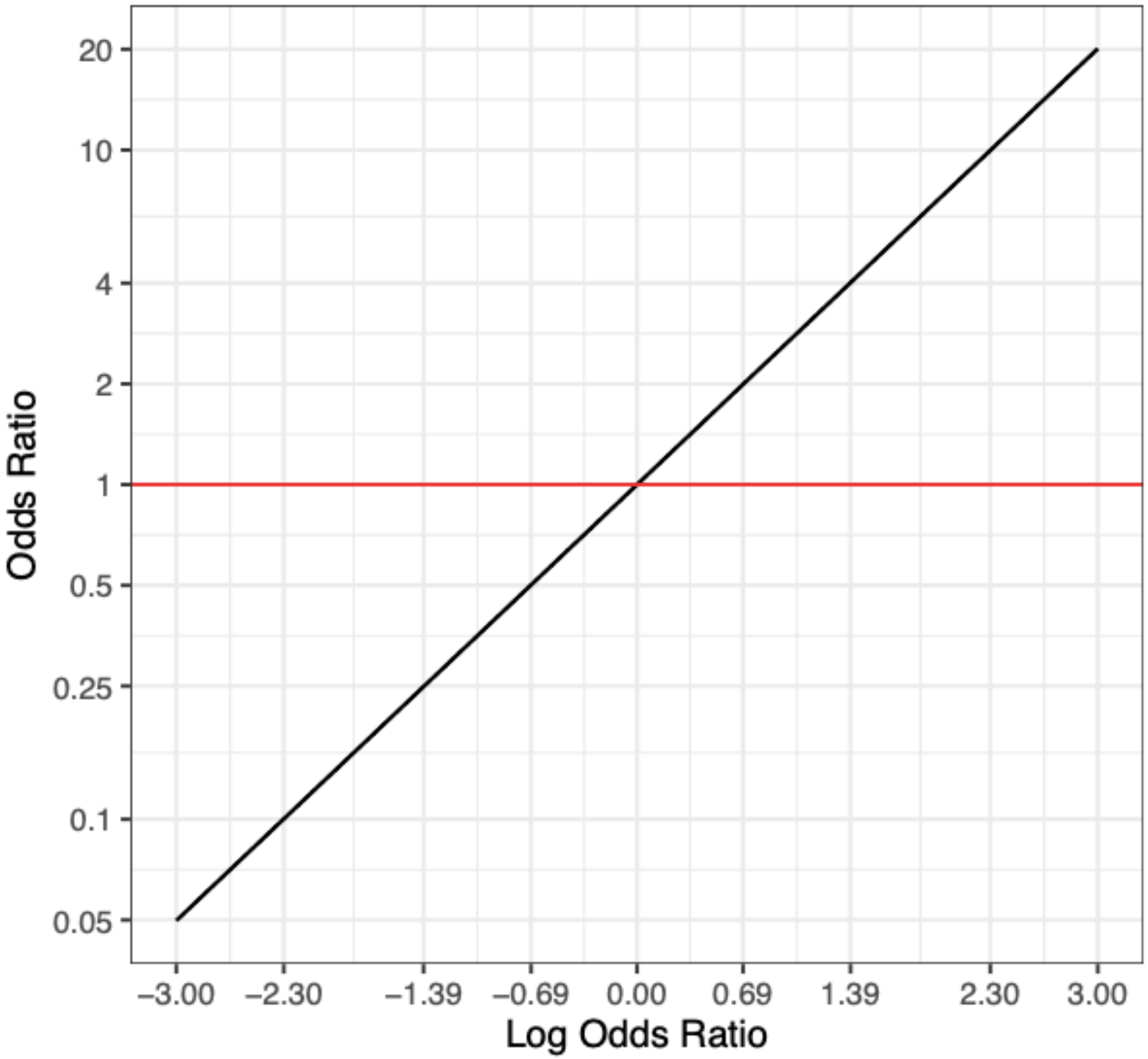
- Odds ratios are the ratio between the odds of some outcome happening in group 1 relative to the odds of some outcome happening in group 0.
- Related to the risk ratio (the ratio between the risk of some outcome happening in group 1 relative to the risk of some outcome happening in group 0) but not the same.
    - This relationship actually depends on the prevalence of the outcome in group 0.
- Difficult to intuit.

# Interpreting a Logistic Regression Model

- Sample interpretation for continuous predictor: Suppose $\hat{\beta} = 1.0$. Then $exp(\hat{\beta}) \approx 2.7$. We would interpret this:
    - The odds of being depressed increase 2.7 times for every unit increase of this predictor. So as this predictor increases, the odds of being depressed increases..
- Sample interpretation for categorical predictor: Suppose $\hat{\beta} = 1.0$. Then $exp(\hat{\beta}) \approx 2.7$. We would interpret this:
    - The odds of being depressed are 2.7 times higher in the treatment group than in the control group. So individuals in the treatment group are more likely to be depressed than individuals in the control group.
- Let's interpret $\hat{\beta}$ from our data example.

```r
#Fit a logistic regression:
logistic_model <- glm((freq_depressed=="Yes") ~ release_year,
                      data = prison_dat, family = "binomial")

summary(logistic_model)
```

```
Call:
glm(formula = (freq_depressed == "Yes") ~ release_year, family = "binomial",
    data = prison_dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7130  -0.4264  -0.4150  -0.4123   2.2395

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -30.804776   6.918468  -4.453 8.49e-06 ***
release_year  0.014078   0.003423   4.113 3.91e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8555.0  on 14529  degrees of freedom
Residual deviance: 8539.8  on 14528  degrees of freedom
AIC: 8543.8

Number of Fisher Scoring iterations: 5
```


Odds Ratio
20
10
4
2
1
0.5
0.25
0.1
0.05
–3.00
–2.30
–1.39
–0.69
0.00
0.69
1.39
2.30
3.00
Log Odds Ratio

# Interpreting a Logistic Regression Model

Note that if:

$$ln(\frac{\pi(x_i)}{1 - \pi(x_i)}) = \alpha + \beta x_i$$

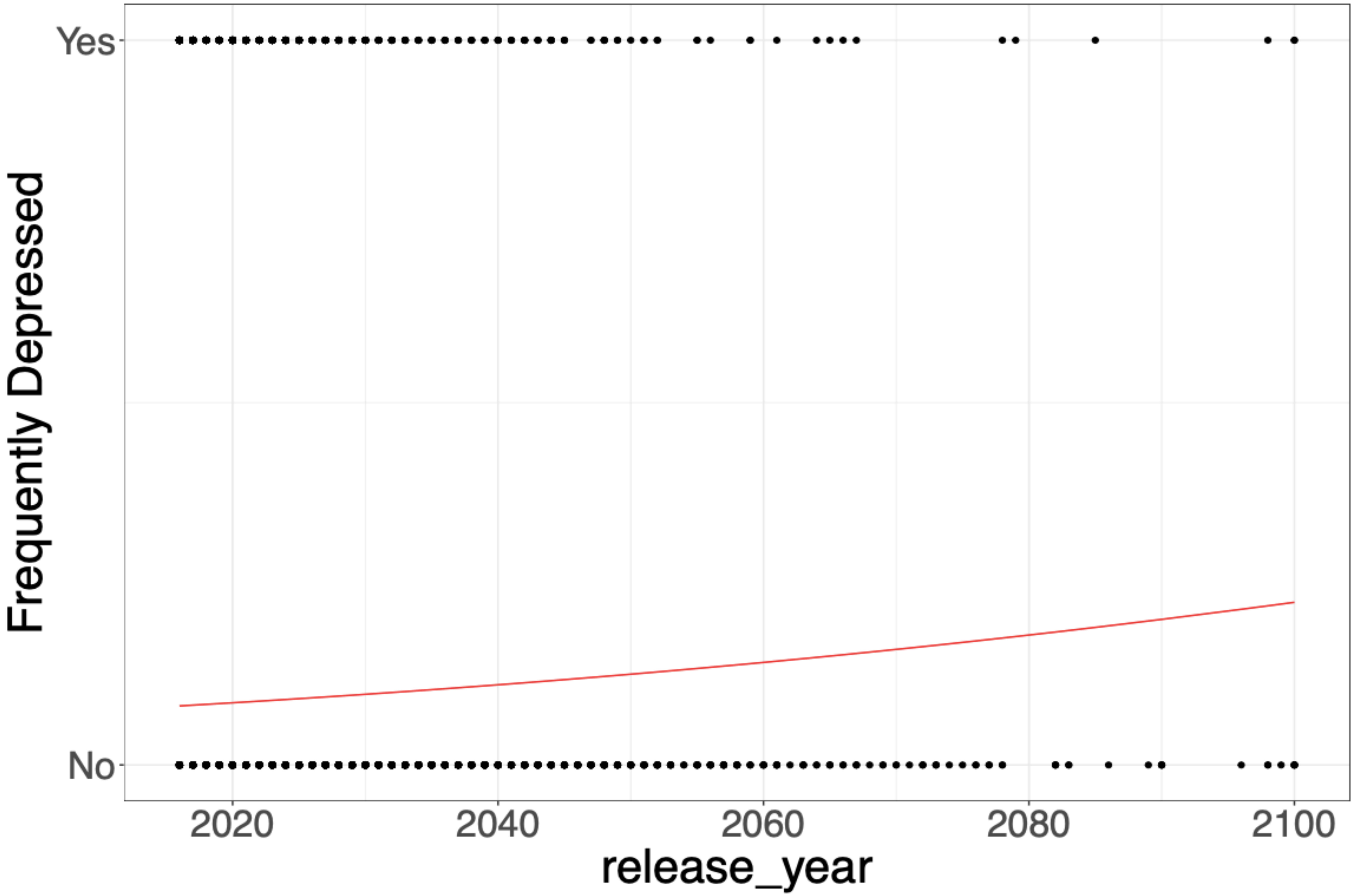
Then:

$$\pi(x_i) = \frac{exp(\alpha + \beta x_i)}{1 + exp(\alpha + \beta x_i)}$$

Let's interpret $\hat{\alpha}$ from our model.

Now let's predict the probability of being depressed for a prisoner with release date 2100.

```r
prison_dat$logistic_preds1 <- predict(logistic_model, type = "response")

depressed_release_year + geom_line(data = prison_dat,
                                   aes(x = release_year, y = logistic_preds1),
                                   color = "red")

```



# Inference with Logistic Regression Models

We can get confidence intervals and conduct testing from a logistic regression model similarly to how we would for a linear regression model.

It's important to do everything on the original model scale (i.e. the log scale).

Is $\hat{\beta}$ significant?

$$\hat{Z} = \frac{\hat{\beta}}{sd(\hat{\beta})}$$

Then compare $\hat{Z}$ to a normal cumulative distribution function to get the p-value, same as usual. Or just read it off the model output from R.  :)

# Inference with Logistic Regression Models

95% confidence interval for $\hat{\beta}$:

$$(\hat{\beta} - 1.96 * sd(\hat{\beta}), \hat{\beta} + 1.96 * sd(\hat{\beta}))$$

95% confidence interval for the odds ratio, or $exp(\hat{\beta})$:

$$(exp[\hat{\beta} - 1.96 * sd(\hat{\beta})], exp[\hat{\beta} + 1.96 * sd(\hat{\beta})])$$

Note that exponentiation should always be the LAST thing you do!

Let's calculate a 95% confidence interval for $exp(\hat{\beta})$ from our model.

# Multiple Logistic Regression

We can also consider more than just a single predictor:

$$ln(\frac{\pi(x_{1i}, x_{2i}, \ldots, x_{pi})}{1 - \pi(x_{1i}, x_{2i}, \ldots, x_{pi})}) = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

We interpret each predictor as the effect of changing that sole predictor while holding all others constant.

Can add interactions, quadratic terms, … etc.

# More Examples

- Let’s add whether the subject has done any job training during their imprisonment and practice interpreting that.
- Let’s add job training, education, and an interaction between them.
- Let’s try adding on whose authority the subject is imprisoned and see how that affects things.

```r
#Fit a multiple logistic regression:
logistic_model2 <- glm((freq_depressed=="Yes") ~ release_year + # fill in some other predictors!!,
                       data = prison_dat, family = "binomial")

summary(logistic_model2)
```

```r
#Generate predictions from that regression:
prison_dat$logistic_preds2 <- predict(logistic_model2, type = "response")
```

# Separation in Logistic Regression

- We see evidence of separation when we get extremely large estimates of the odds ratio and standard deviation for some predictor.
- Usually happens for one of two reasons:
    - The predictor has a true, extremely large relationship with the outcome, such that all or almost all of the individuals in one of the categories have the same outcome.
    - One of the categories has a very small sample size, and as a result all or almost all of the individuals in one of the categories have the same outcome.
- Especially likely to occur if you have lots of interactions between categorical predictors.

# Separation in Logistic Regression

- If your model is separated, you cannot trust any of the findings: the model has not converged.
- A couple of solutions:
    - If you think the problem is caused by small sample size (and not a legitimate extremely strong relationship), you can try excluding the problematic predictor, merging categories in a sensible way to increase sample sizes, or dropping that category.
    - If you think the problem is caused by a legitimate strong relationship, you can try either adding a couple of augmented data rows or using Firth's correction. This is beyond the scope of this lecture, but it's good to be aware. See "More Reading" below.

# Limitations of This Analysis

- Didn’t address within-prison correlation
- Didn’t include the sampling weights
- Need to do way more descriptive/exploratory analysis before feeling confident in these results!
- Maybe shouldn’t exclude prisoners who don’t yet have a release date.
- Probably other problems.

# More Reading

- Logistic regression + statistics:
    - Highly recommend Alan Agresti's *Categorical Data Analysis* (3rd edition)
    - The classic McCullagh and Nelder *Generalized Linear Models* (2nd edition)
    - More information on separation:
        - Mansournia et al., "Separation inn Logistic Regression: Causes, Consequences, and Control," *AJE* 2017: https://academic.oup.com/aje/article/187/4/864/4084405?login=true
        - Heinze, "A comparative investigation of methods for logistic regression with separated or nearly separated data," *Statistics in Medicine* 2006: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2687
- U.S. prison system:
    - More statistics and some common myths about the U.S. prison system: https://www.prisonpolicy.org/reports/pie2022.html
    - Some history on the evolution of the U.S. prison system and its connections to slavery and racial injustice: https://www.nytimes.com/interactive/2019/08/14/magazine/prison-industrial-complex-slavery-racism.html
    - The full dataset from ICPSR (so you don't have to trust my janky data cleaning): https://www.icpsr.umich.edu/web/ICPSR/studies/37692/summary

# Contact Info

Please feel free to reach out with questions or concerns about this lecture, statistics, grad school, life, etc.!

ecchase@umich.edu