Math 430: Lecture 8a

Logistic regression

Professor Catalina Medina

Examples of Outcome Variables

Specify variable type (think about the possible values)

- Temperature continuous - linear regression
- Presence of Alzheimer's binary *logistic regression!
- Rent continuous - linear regression
- Computer operating system categorical
- Daily number of customers discrete
- Income class categorical ordinal

Which could we model with **linear regression**?

Identify the response variable type:

- A car insurance company is trying to predict if an applicant will have a car crash in the next 5 years, based on their income.

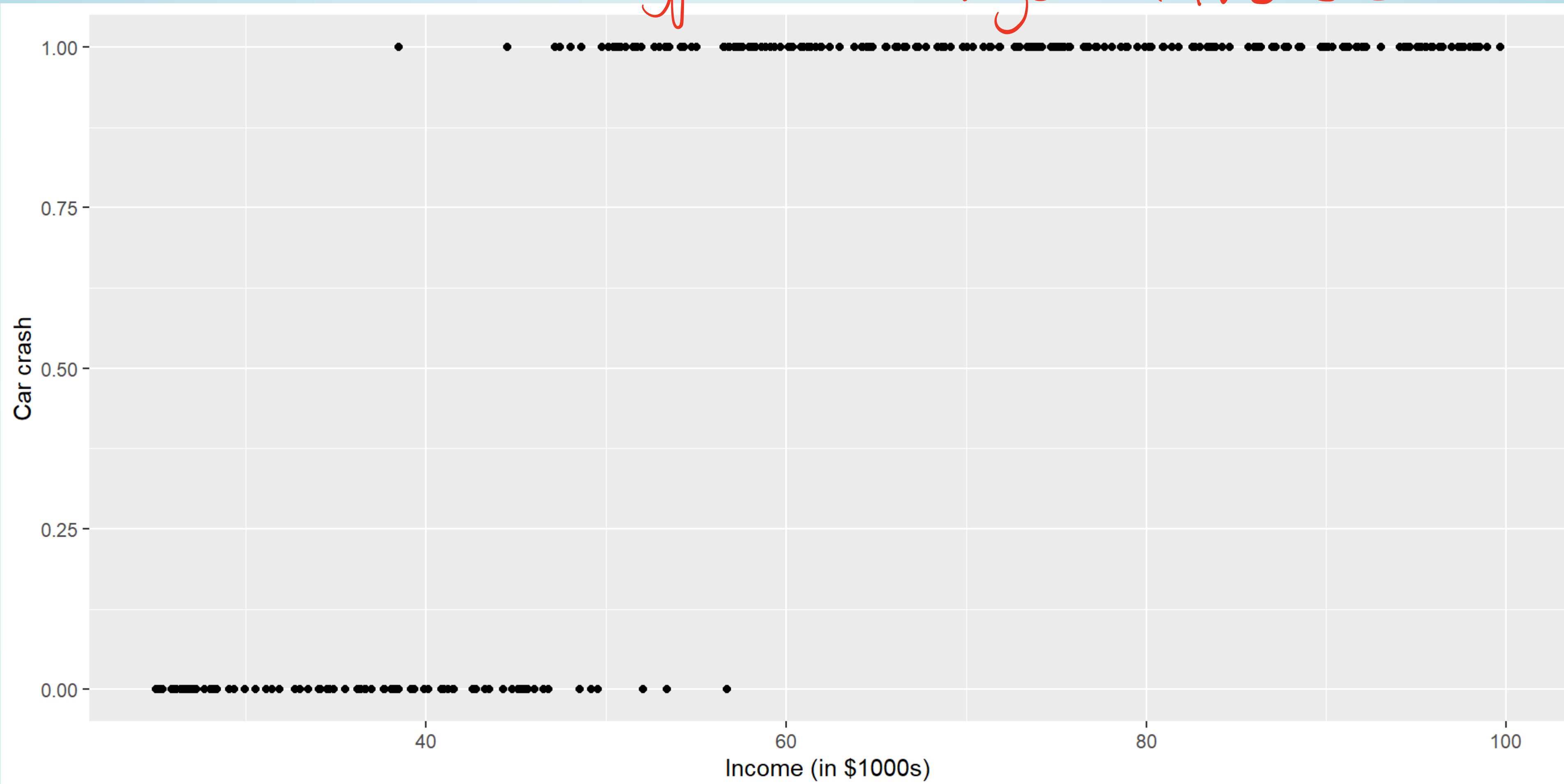
response
binary (categorical with only 2 levels)

explanatory
continuous

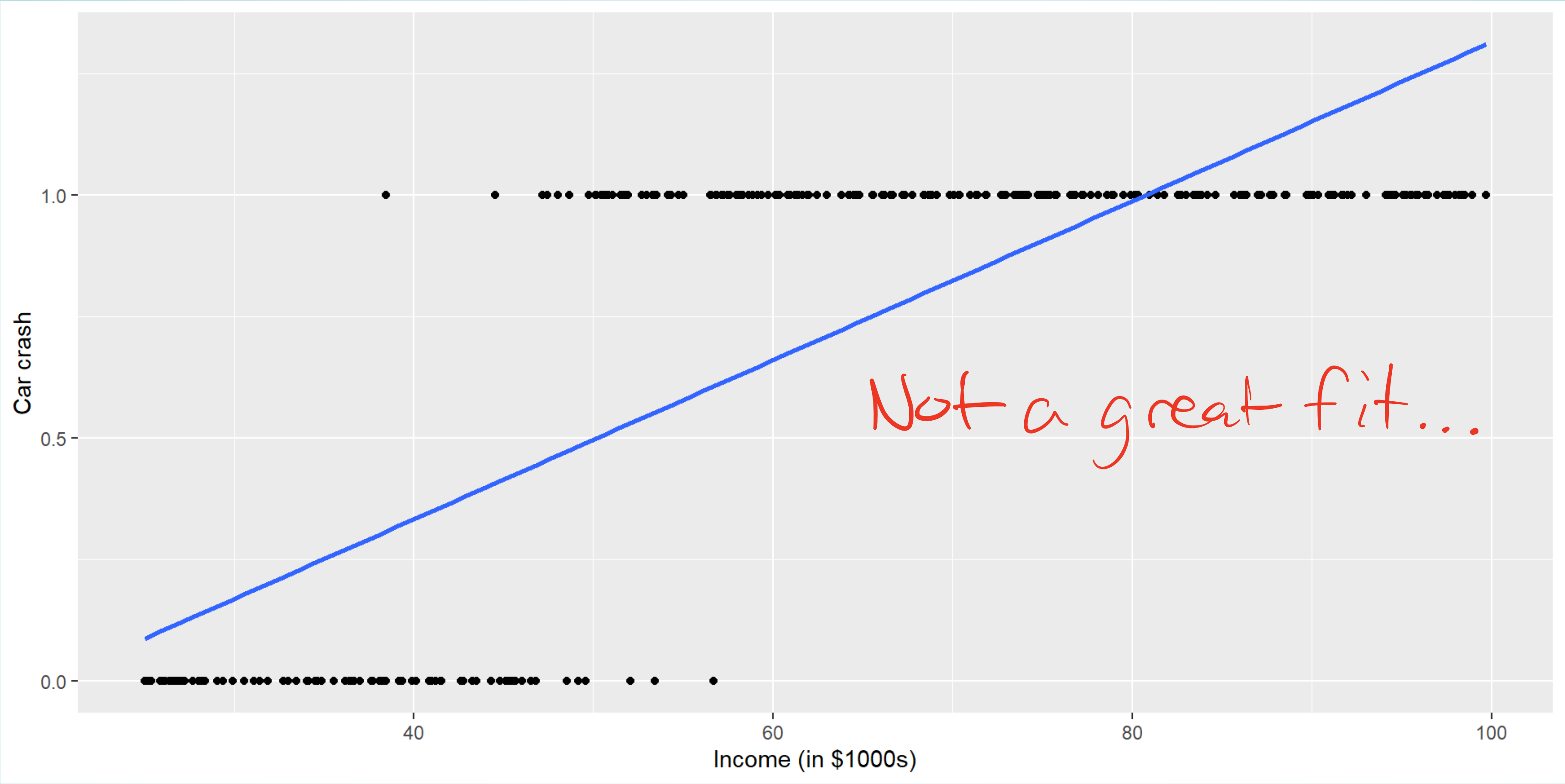
Can we model using linear regression anyways?

Let's simulated data...

Scatterplot is not the ideal plot type for a categorical & numeric variables



What does it mean to fit a line to this data?



We don't get an error, but ...

Linear regression assumes

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_p X_{p,i} + \epsilon_i \text{ and } \epsilon_i \sim N(0, \sigma^2)$$

so

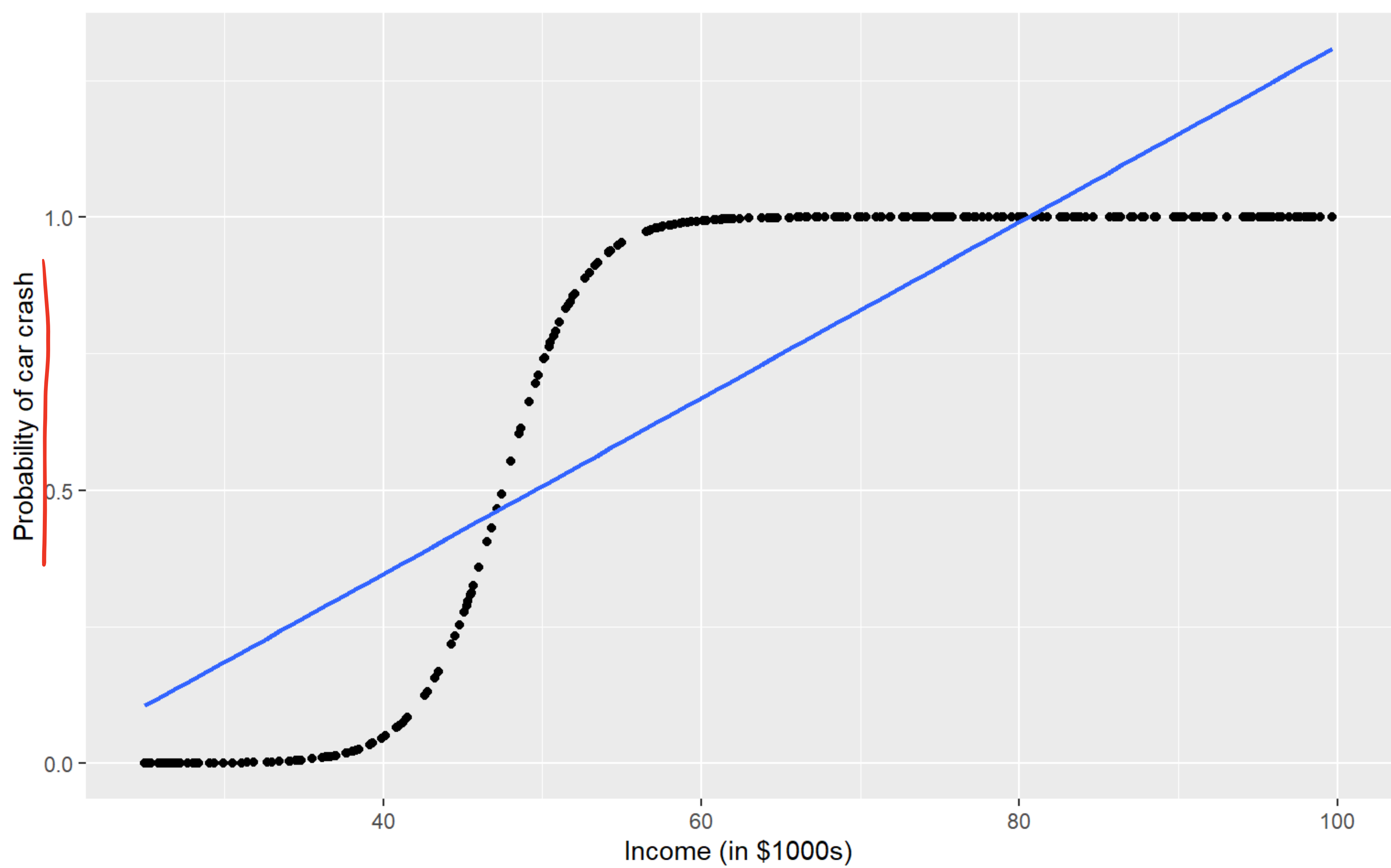
$$E[Y_i | X_{1,i}, \ldots, X_{p,i}] = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_p X_{p,i}$$

What is the expectation of a binary variable?

Bernoulli distribution is good for binary variables

$Y_i | X_i \sim \text{Bernoulli}(\pi_i)$ probability of "success"

$\Rightarrow$ Predicted values will actually be probabilities!


Probability of car crash
1.0
0.5
0.0
40
60
80
100
Income (in $1000s)

What is our expecation for people with an income of $45,000?

```r
crash_model <- lm(car_crash ~ income, data = sim_crash_data)

new_data <- data.frame(income = 45)

predict(crash_model, newdata = new_data)
```

```
        1 
0.4151667 
```

We get an ok prediction...

What is our expecation for people with an income of $90,000?

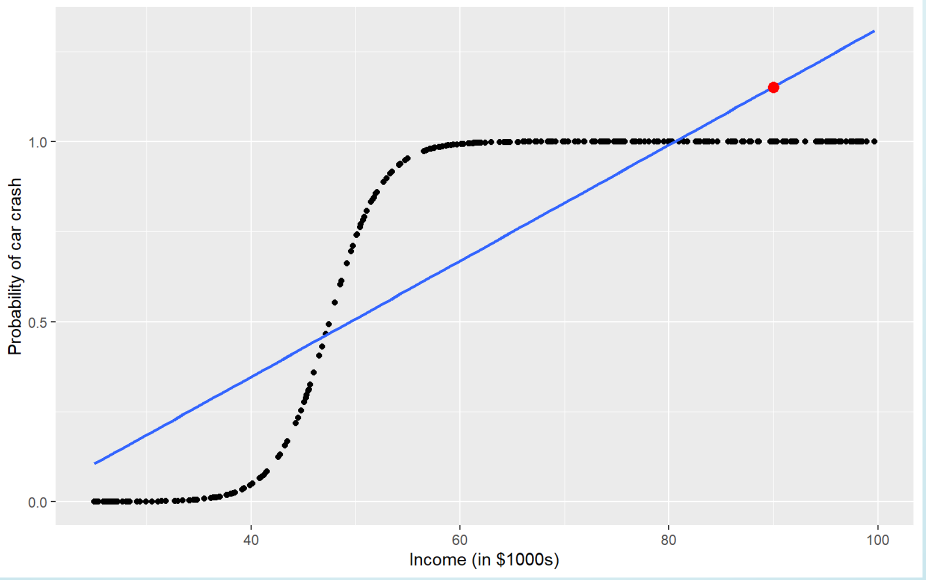
```r
crash_model <- lm(car_crash ~ income, data = sim_crash_data)

new_data <- data.frame(income = 90)

predict(crash_model, newdata = new_data)
```

```
       1
1.151587
```

invalid probability!
Probability must be between 0 & 1



We can fit a linear model for a binary response, but:

- the line is really fitting the $E[Y_i|X_i]$ $= \pi_i$, the probability
- the range of $E[Y_i|X_i] = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_p X_{p,i}$ is $(-\infty, \infty)$
- we may get invalid predictions

We can use a logistic model instead!

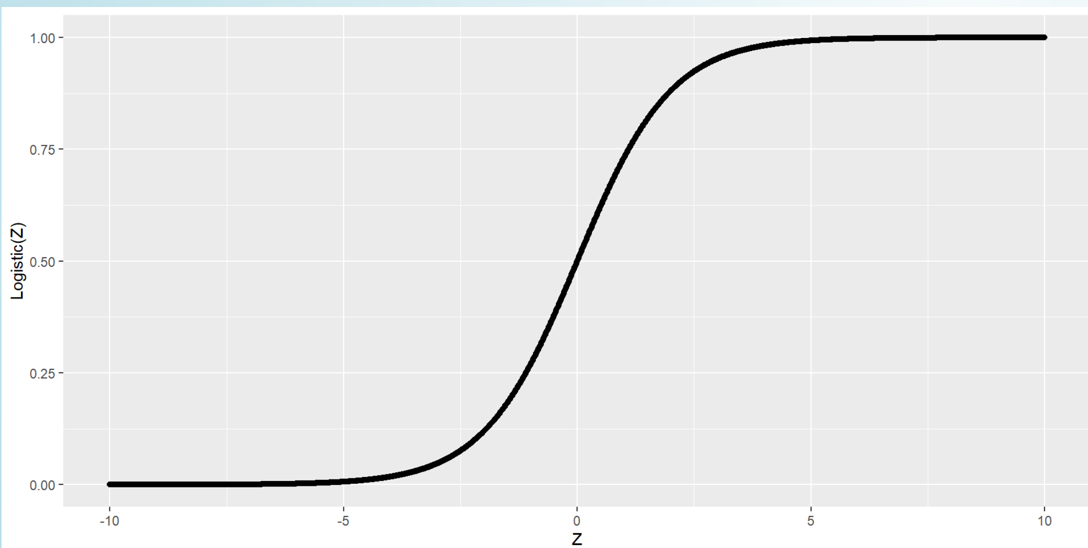
Logistic function

The logistic function ranges from 0 to 1

$$\text{logistic}(Z) = \frac{e^Z}{1+e^Z} = \frac{e^z}{e^z(e^{-z}+1)} = \frac{1}{e^{-z}+1}$$

$$\lim_{z\to\infty} \frac{1}{e^{-z}+1} = \lim_{z\to\infty} \frac{1}{\frac{1}{e^z}+1} = \frac{1}{\frac{1}{\infty}+1} = 1$$

$$\lim_{z\to-\infty} \frac{1}{e^{-z}+1} = \frac{1}{e^{\infty}+1} = \frac{1}{\infty} = 0$$



We like it because it's bounded!

Logistic regression: form 1

Let $Y_i|X_{1,i},\ldots,X_{p,i} \sim Bernoulli(\pi_i)$, which means $E[Y_i|X_{1,i},\ldots,X_{p,i}] = \pi_i$

The logistic regression model is

$$E[Y_i|X_{1,i},\ldots,X_{p,i}] = \pi_i = \frac{e^{\beta_0+\beta_1 X_{1,i}+\ldots+\beta_p X_{p,i}}}{1+e^{\beta_0+\beta_1 X_{1,i}+\ldots+\beta_p X_{p,i}}}$$

Logistic Response Function Rewritten

$$\pi_i = \frac{e^{\beta_0+\beta_1 X_{1,i}+\ldots+\beta_p X_{p,i}}}{1+e^{\beta_0+\beta_1 X_{1,i}+\ldots+\beta_p X_{p,i}}}$$

$\beta_0+\beta_1 X_{1,i}+\ldots+\beta_p X_{p,i} = L$

Solve for L

$$\pi_i = \frac{e^L}{1+e^L}$$

$$\pi_i(1+e^L) = e^L$$

$$\pi_i + \pi_i e^L = e^L$$

$$\pi_i = e^L - \pi_i e^L$$

$$\pi_i = e^L(1-\pi_i)$$

$$\frac{\pi_i}{1-\pi_i} = e^L$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = L = \beta_0+\beta_1 X_{1i}+\ldots+\beta_p X_{pi}$$

*Note statisticians use natural log $\ln(x)$ but commonly write $\log(x)$

Logistic regression: form 2

Through some algebra last slide we can rewrite the model as

$$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_p X_{p,i}$$

- Form 1 is useful for understanding why we use the logistic function
- Form 2 is useful for interpretation

Logistic regression: form 2

odds

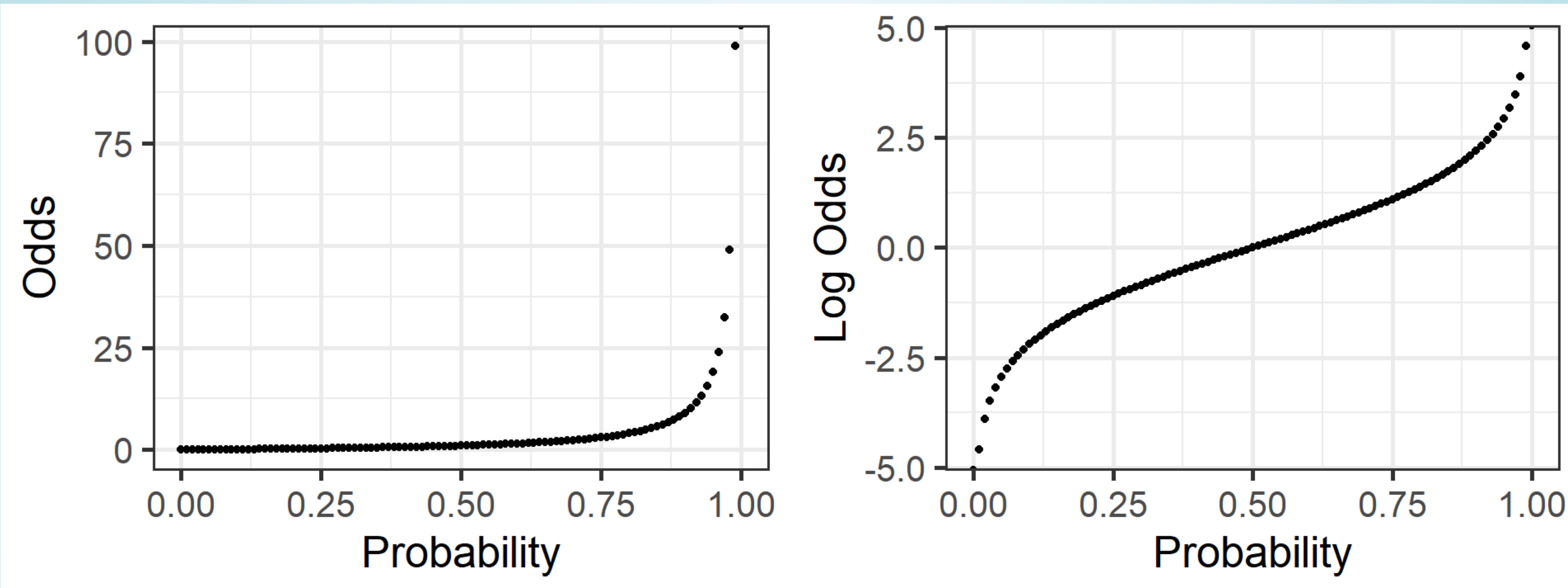
$$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_p X_{p,i}$$

- Left side is called the *log odds* or *logit*

Probability and Odds

What are odds? $\text{Odds} = \frac{\text{Prob success}}{\text{Prob failure}} = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1-P(Y=1)} = \frac{\pi}{1-\pi}$

Relationship between probability and odds



As probability increases so do odds and log odds!

Parameter Interpretation: Solving for β_1

$$\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1}$$

Now we have a model that connects $E[Y|X] = \pi$ to our data, but how do we interpret the parameters, β_j's?

Recall for linear regression: $E[Y|X] = \beta_0 + \beta_1 X$

$$E[Y_i|X_i = x+1] - E[Y_i|X_i = x] = \beta_0 + \beta_1(x+1) - (\beta_0 + \beta_1 x)$$
$$= \cancel{\beta_0} + \cancel{\beta_1 x} + \beta_1 \cancel{-\beta_0} \cancel{-\beta_1 x}$$
$$= \beta_1.$$

Note if you had more X's they would be held constant so they would cancel out and we would get the same.

Parameter Interpretation: Solving for β_1

$$\log\left(\frac{E[Y_i|X_i]}{1 - E[Y_i|X_i]}\right) = \beta_0 + \beta_1 X_i$$

Let

$\pi_a = E[Y_i | X_i = x+1]$ and $\pi_b = E[Y_i | X_i = x]$

Then

$$\log\left(\frac{\pi_a}{1-\pi_a}\right) - \log\left(\frac{\pi_b}{1-\pi_b}\right) = \beta_0 + \beta_1(x+1) - (\beta_0 + \beta_1 x)$$

$$\log\left(\frac{\pi_a}{1-\pi_a} \Big/ \frac{\pi_b}{1-\pi_b}\right) = \cancel{\beta_0} + \cancel{\beta_1 x} + \beta_1 - \cancel{\beta_0} - \cancel{\beta_1 x}$$

$$\log\left(\frac{\pi_a}{1-\pi_a} \Big/ \frac{\pi_b}{1-\pi_b}\right) = \beta_1$$

"log odds ratio" is not very interpretable. Therefore we will commonly interpret a transformation of β_1.

Parameter Interpretation: Explaining β_1

Odds Ratio

$$e^{\beta_1} = \frac{\pi_a}{1-\pi_a} \Big/ \frac{\pi_b}{1-\pi_b}$$

we can think of this as

$$\frac{\pi_a}{1-\pi_a} = e^{\beta_1} \frac{\pi_b}{1-\pi_b}$$

We typically interpret the odds ratio e^{β_1} instead of log odds ratio β_1

If $e^{\beta_1} = 1 \Rightarrow \beta_1 = 0$

There is no change in odds, no change in probability, associated with a 1 unit increase in X

If $e^{\beta_1} < 1 \Rightarrow \beta_1 < 0$

There is a decrease in odds, decrease in probability associated with a 1 unit increase in X

If $e^{\beta_1} > 1 \Rightarrow \beta_1 > 0$

There is an increase in odds, increase in probability, associated with a 1 unit increase in X

Parameter Estimation: MLE

We want the Maximum Likelihood Estimate (MLE) for the β_j's... why?

- The MLE aims to find estimates $\hat{\theta} = \{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1}\}$ that maximize the likelihood function over the parameter space with a given data set.
- The likelihood function $L(\theta|y)$ measures the goodness of fit of a statistical model to a sample of data for given values of the unknown parameters θ.
- The Maximum Likelihood Estimators have desirable asymptotic (large sample size) properties

Parameter Estimation: MLE for Logistic Regression

We want to find estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that maximize $L(\theta|y)$

- For simple linear regression we can directly solve for the maximum likelihood estimators and they happen to be identical to the least square estimators
- For logistic regression we can't directly calculate the estimators, we have to use optimization algorithms. Luckily those are implemented for us!

Inference With Logistic Regression

In linear regression we used a t-test to test for statistically significant features.

Beware in logistic regression there are different tests that can be used (we will talk about these)

- Likelihood Ratio Test
- Wald Test

we will talk about these in a different lecture

Know which one your software is using!

Classification / Prediction With Logisitic Regression

How do we convert the predicted probabilities to values of Y (e.g., 0 vs 1 or TRUE vs FALSE)?

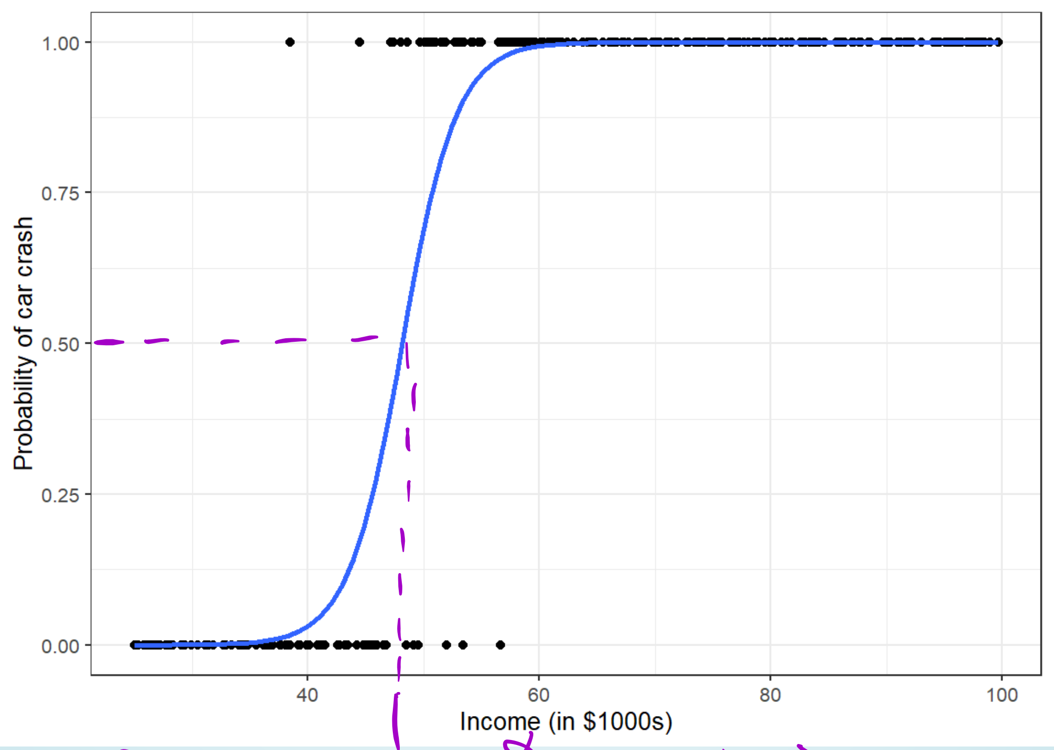
The cutoff should be a domain knowledge-based decision. The natural cutoff with no context would be 0.5

```
ggplot(sim_crash_data, aes(y = car_crash, x = income)) +
    geom_point() +
    geom_smooth(
      method = glm,
      se = FALSE,
      method.args = list(family = "binomial")
    ) +
    theme_bw(base_size = 10) +
    labs(y = "Probability of car crash", x = "Income (in $1000s)")
```

this is how we can plot our logistic regression

Classification / Prediction With Logisitic Regression

With a given cut off I want to make a prediction (0 or 1), let's use a cutoff of 0.5



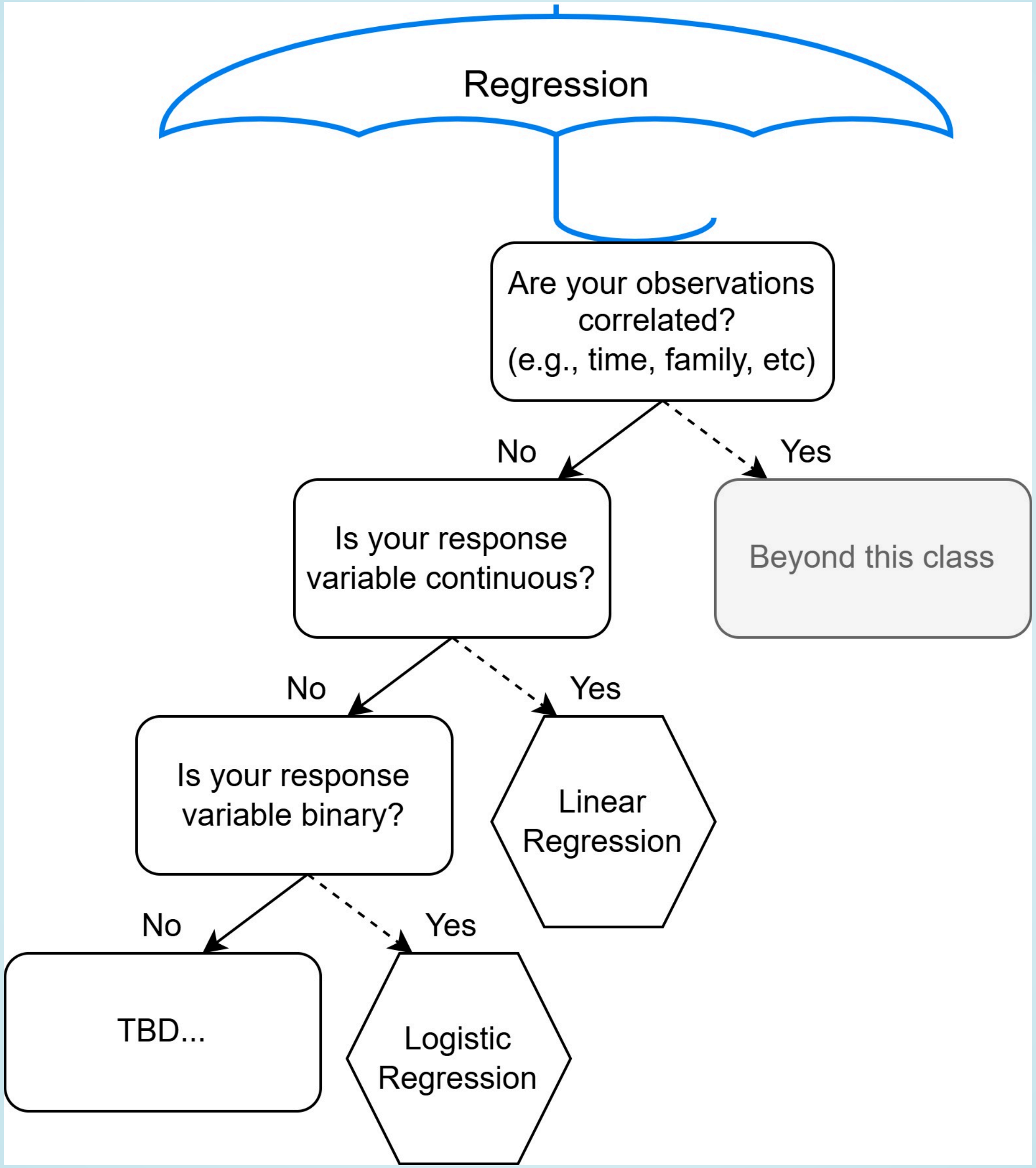
If income < \$48,000 ⇒ predict crash
If income ≥ \$48,000 ⇒ predict no crash

Confusion Matrix

For linear regression we mostly just use RSE

There are many considerations for how "well" our logistic model is predicting

Confusion Matrix		Predicted	
		Positive (1)	Negative (0)
Actual	Positive (1)	$n_{1,1}$ **True Positive**	$n_{1,2}$ **False Negative**
	Negative (0)	$n_{2,1}$ **False Positive**	$n_{2,2}$ **True Negative**


Regression
Are your observations correlated? (e.g., time, family, etc)
No
Yes
Is your response variable continuous?
Beyond this class
No
Yes
Is your response variable binary?
Linear Regression
No
Yes
TBD...
Logistic Regression