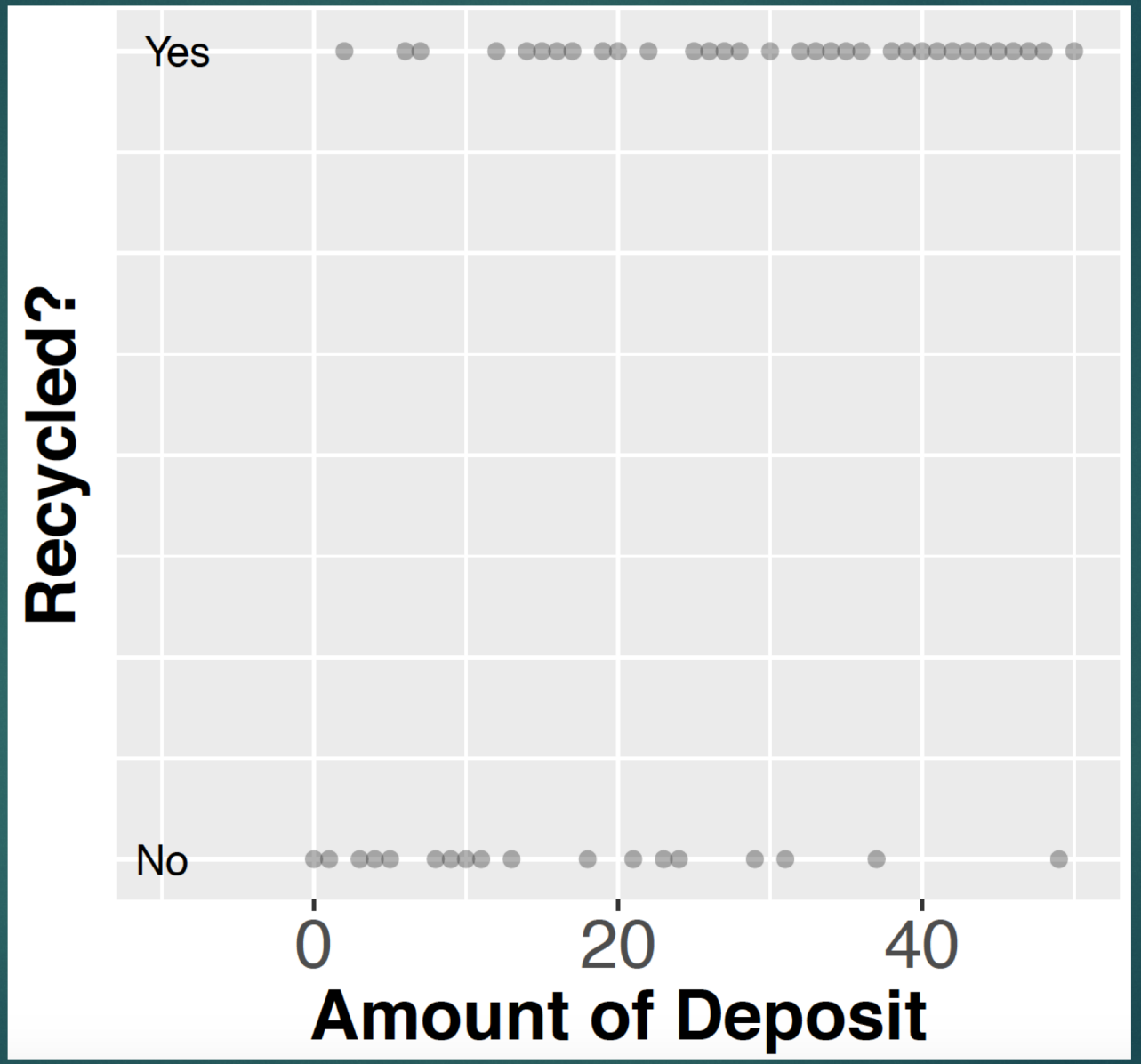
Introduction to Logistic Regression

QAC 201

In an effort to increase recycling, an environmental research organization tested out deposit amounts from 0 cents to 50 cents. Each item was then recorded as whether or not it was recycled (yes=1, no=0).

The data appeared as follows:


Yes
No
Recycled?
0
20
40
Amount of Deposit

Terminology

- Odds = $\frac{p}{1-p}$
- If we say the odds of getting a job offer before graduation for college students is 3. That means that it is 3 times more likely for a student to graduate from college with a job offer than it is for a student to graduate from college without a job offer. (Another way of saying it would be to say the probability of a job offer is 75%).
- If we say that the odds of getting a job offer before graduation for college students is 0.5, then that would mean that it is half as likely for a student to graduate from a college with a job offer than to not graduate with a job offer. (Another way of saying it would be to say that the probability of a job offer is 33.3%)

A logistic regression will allow us to test this relationship:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.00730    0.61284  -1.644  0.10024
deposit      0.07054    0.02478   2.847  0.00441 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$$log\left(\frac{P(Returned)}{P(Not\ Returned)}\right) = -1.00730 + 0.07054(Deposit)$$

$e^{0.07054}=1.0731$

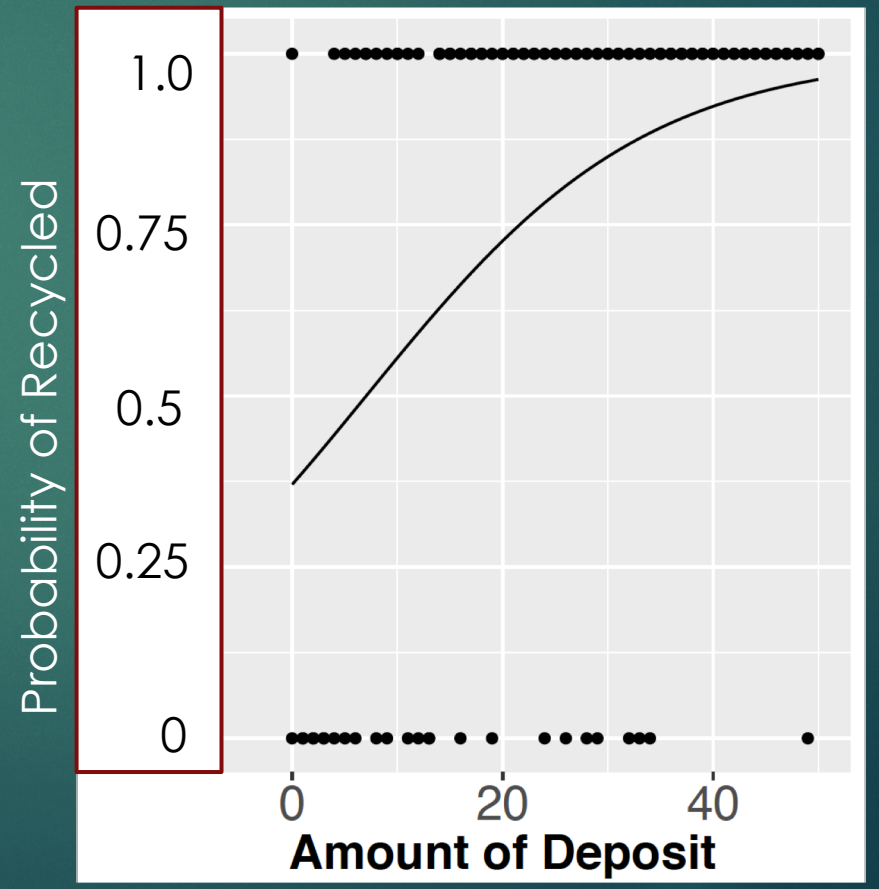
A logistic regression reveals that there is a significant association between deposit amount and recycling (OR=1.0731, p = 0.004).

As deposit amount increases by 1 cent, the odds that a bottle is returned is expected to increase by a factor of 1.0731.

Another way of expressing the same equation is:

$$P(Returned) = \frac{\exp(-1.00730 + 0.07054(Deposit))}{1 + \exp(-1.00730 + 0.07054(Deposit))}$$

Which graphically looks like:



Now suppose we want to see how material type (plastic or aluminum) seems to relates to the likelihood it is recycled, in addition to deposit amount.

A new model is run with material type and deposit amount as predictors and with whether it was recycled as the response variable.

The output is as follows:

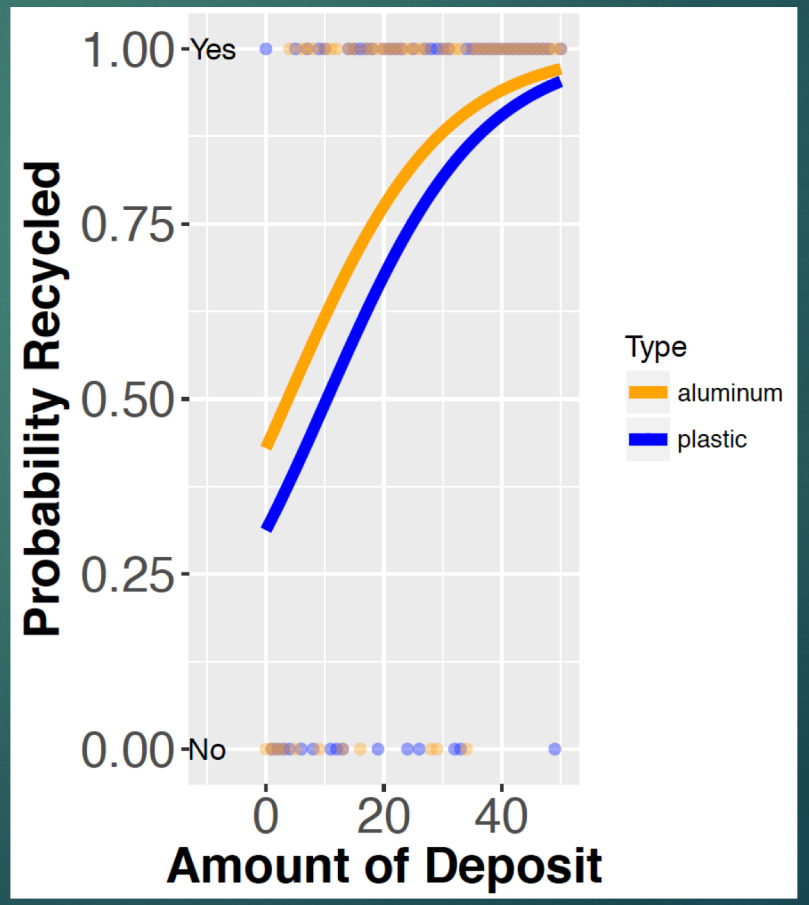
```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.28241    0.50344  -0.561 0.574830
deposit      0.07621    0.02011   3.789 0.000151 ***
sizeplastic -0.50520    0.50700  -0.996 0.319033
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

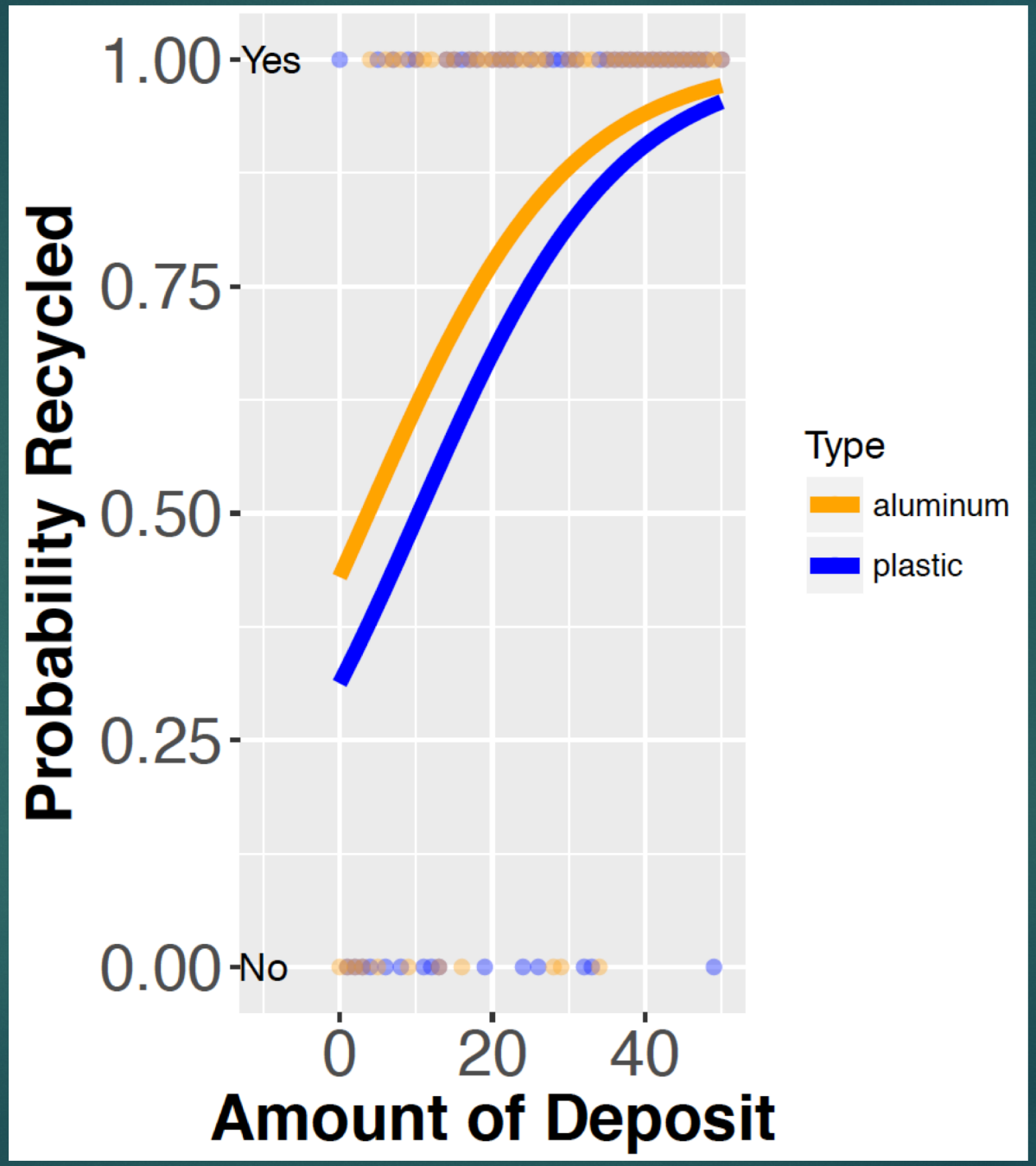
$$\log\left(\frac{P(Returned)}{P(Not\ Returned)}\right) = -0.28241 + 0.07621(Deposit) - 0.50520(Plastic)$$

Again another way of expressing the same equation is:

$$P(Returned) = \frac{\exp(-0.28241 + 0.07621(Deposit) - 0.50520(Plastic)}{1 + \exp(-0.28241 + 0.07621(Deposit) - 0.50520(Plastic))}$$

Which graphically looks like:




1.00
Yes
0.75
0.50
0.25
0.00
No
Probability Recycled
0
20
40
Amount of Deposit
Type
aluminum
plastic