SMI606: Week 5

Logistic Regression

Rhiannon Williams

Sheffield Methods Institute, the University of Sheffield

rswilliams2@sheffield.ac.uk

Sign in

Link

Learning objectives: what will I learn?

By the end of this week you will:

- Understand logistic variables
- Be able to run and interpret logistic regression models in R
- Be able to find and interpret log odds

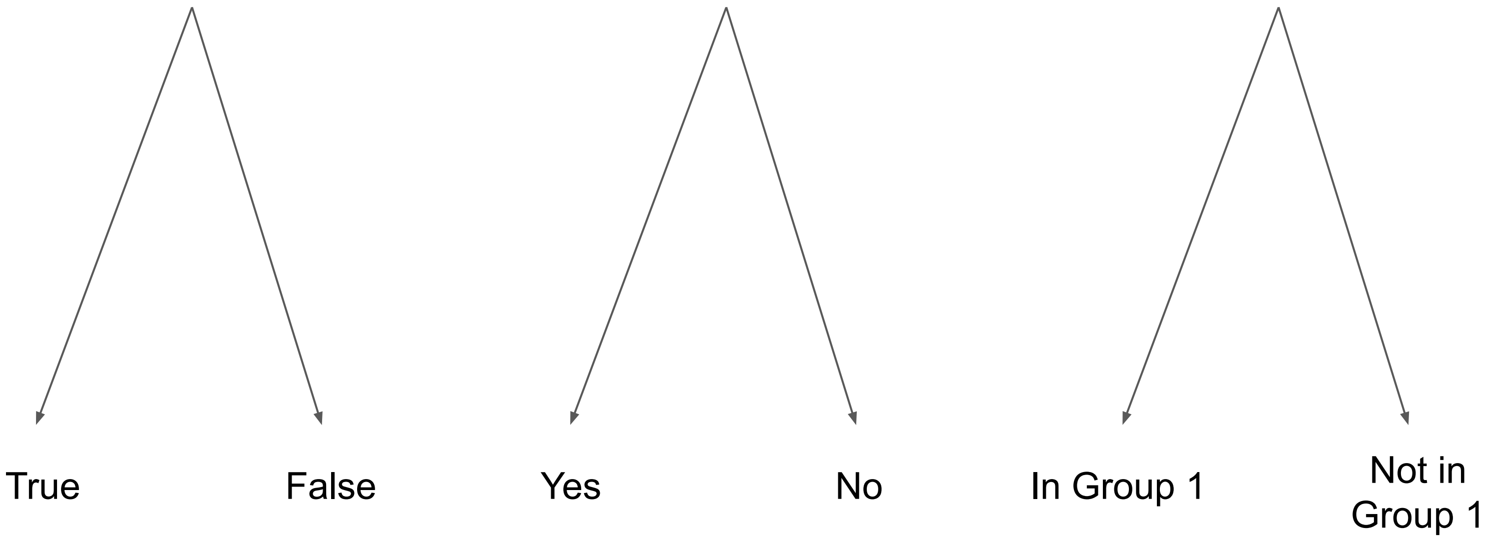
Learning objectives: how does this week fit into my course?

- This week's material builds on your bivariate and multiple linear regression learning. Logistic regression expands the types of research questions you can explore.

- Logistic regression is one of the possible approaches you can apply to your assignment for this module.

What are logistic variables?

A logistic/binary/dichotomous variable has only **two outcomes**.



What are logistic variables?

Share some examples of logistic variables on the Jamboard

Logistic variables in action

Research question: What characteristics are associated with a person's ability to meet their housing costs?

Data: Understanding Society longitudinal survey

Dependent variable: In the last twelve months, have you ever found yourself behind with your rent? [1 = Yes, 0 = No]

How do logistic variables work in R?

“A logistic regression model has a dependent variable that is dichotomous, having only **0 and 1 as coded values**.” (Schumacker, 2014)

1	Yes, True, In group
0	No, False, Not in group

How do logistic variables work in R?

respondent	age	payment_problems
0000001	21	yes
0000002	34	no
0000003	27	yes
0000004	67	yes
0000005	42	no
...	...	...

How do logistic variables work in R?

```
# read in the data
total <- read.csv(file = "survey_data.csv")
```

For each respondent, our data records we have a categorical variable telling us whether they have had housing problem payments or not. We can recode this as a logistic variable, where having problems = 1 and not having problems = 0.

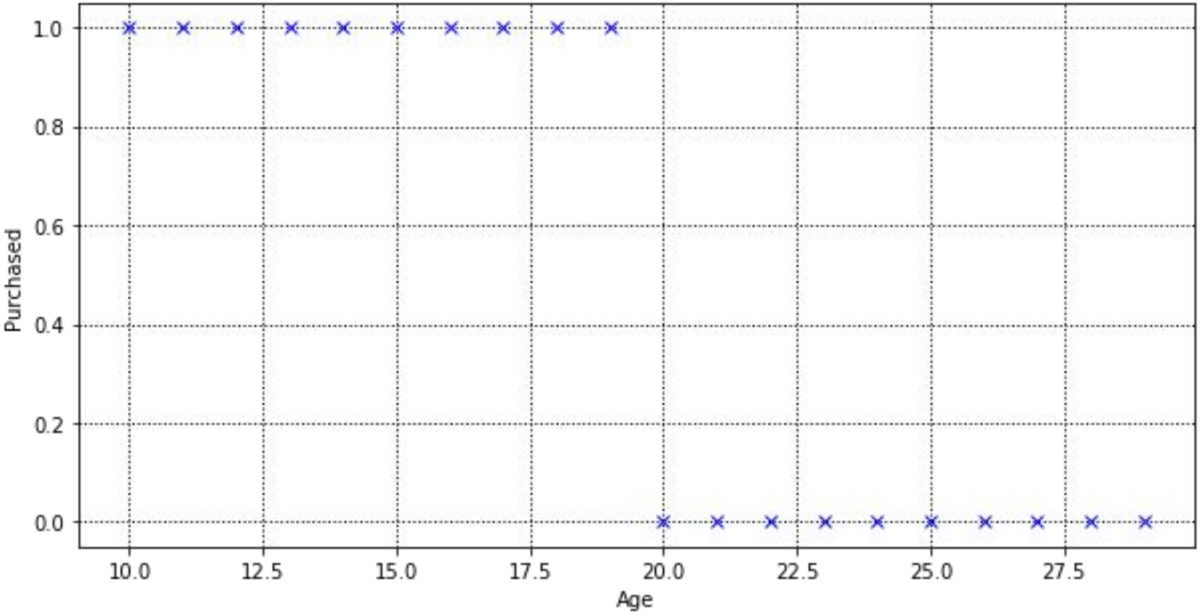
```
# set up the logistic variable

total<- mutate(total, outcome = if_else(payment_problems=="yes", 1, 0))
```

Why don’t we just use linear regression for logistic variables?

If we try to use linear regression...

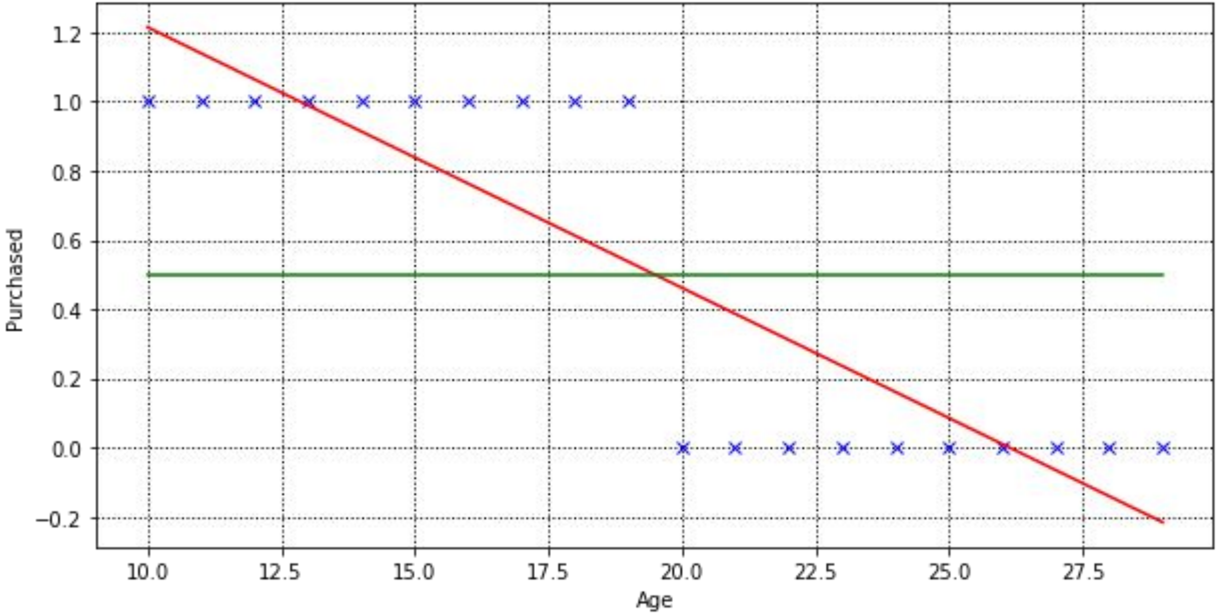
Let’s say we have a dataset of customers telling us their age and whether they made a purchase (1 = made a purchase, 0 = did not make a purchase).



https://www.analyticsvidhya.com/blog/2020/12/beginners-take-how-logistic-regression-is-related-to-linear-regression/

If we try to use linear regression...

Now let's add a regression line in red. **What's gone wrong with our graph?**



https://www.analyticsvidhya.com/blog/2020/12/beginners-take-how-logistic-regression-is-related-to-linear-regression/

If we try to use linear regression...

By using a linear regression model on our logistic dependent variable, we've produced results on our regression line that go above 1 or below 0.

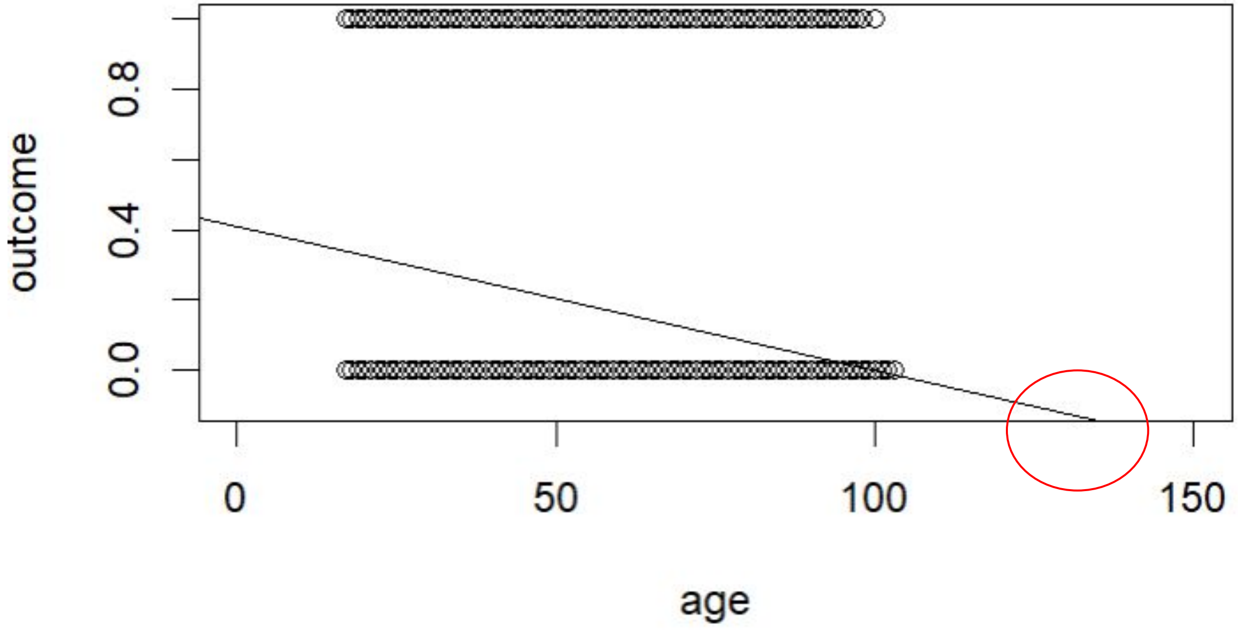
For a logistic variable, these results aren't possible, so our model isn't very helpful!

Age	Predicted Y Value
10	$\boxed{1.21428571}$
15	0.83834586
19	0.53759398
20	0.46240602
25	0.08646617
30	$\boxed{-0.28947368}$

https://www.analyticsvidhya.com/blog/2020/12/beginners-take-how-logistic-regression-is-related-to-linear-regression/

The same thing happens in our housing example!

```
linear_model <- lm(outcome ~ age, data = total)
plot(outcome ~ age, total, xlim=c(0,150), ylim=c(-0.1, 1))
abline(linear_model)
```



Sub-zero outcomes!

Instead we use logistic regression for logistic dependant variables

Differences between linear and logistic regression

$$\bar{y} \quad = \quad b_0 \quad + \quad b_1 x_1 \quad \ldots$$

Predicted outcome

Intercept (the value of Y when all predictor values are 0

Predictor variable coefficients

We’re using the same structure, but the way the regression weighting is generated is different.

Differences between linear and logistic regression

Linear regression	Logistic regression
Linearity	
Homoscedasticity	
Effect of outliers	
Normality of residuals	
Effect of multicollinearity	

Differences between linear and logistic regression

Linear regression	Logistic regression
Linearity	Not needed
Homoscedasticity	Not needed
Effect of outliers	Effect of outliers
Normality of residuals	Not needed
Effect of multicollinearity	Effect of multicollinearity

Differences between linear and logistic regression

Linear regression uses the **least squares criterion**. It selects the regression weights to minimise the sum of squared errors. Logit regression uses **maximum likelihood estimation**. It uses an iterative process to build a statistical model where the observed data is most probable.

How does the weighting work in logistic regression?

$$\bar{y} \quad = \quad b_0 \quad + \quad b_1 x_1 \quad \ldots = \quad \mathbf{log\ (p\ /\ 1\text{-}p)}$$

P is the probability that the outcome is 1, so 1 - p is the probability that the outcome is 0.

How does the weighting work in logistic regression?

$$Y_i = a + b_1X_1 + \cdots + b_jX_j + e_i \quad = \quad \log\ (p\ /\ 1\text{-}p)$$

P is the probability that the outcome is 1, so 1 - p is the probability that the outcome is 0.

P / 1-P is the **odds ratio**.

If we roll a 6 sided die, the odds that our result will be three is 1 in 6, or 1 / 6. The odds that is won't be three is 1 - 1 / 6, or 5 / 6. This is equal to p/(1-p) = (1/6)/(5/6) = 20%.

How does the weighting work in logistic regression?

$$Y_i = a + b_1X_1 + \cdots + b_jX_j + e_i \quad = \quad \mathbf{\log (p / 1\text{-}p)}$$

P is the probability that the outcome is 1, so 1 - p is the probability that the outcome is 0.

P / 1-P is the **odds ratio**.

log(p / 1-p) is the **log odds.**

By taking the logarithm of the odds ratio, we get a normal distribution and shrink extreme values.

For more on logarithm: https://www.sheffield.ac.uk/mash/mathematics/logs

Interpreting the results

This means that when we apply logistic regression to our data, the results are produced as **log odds.**

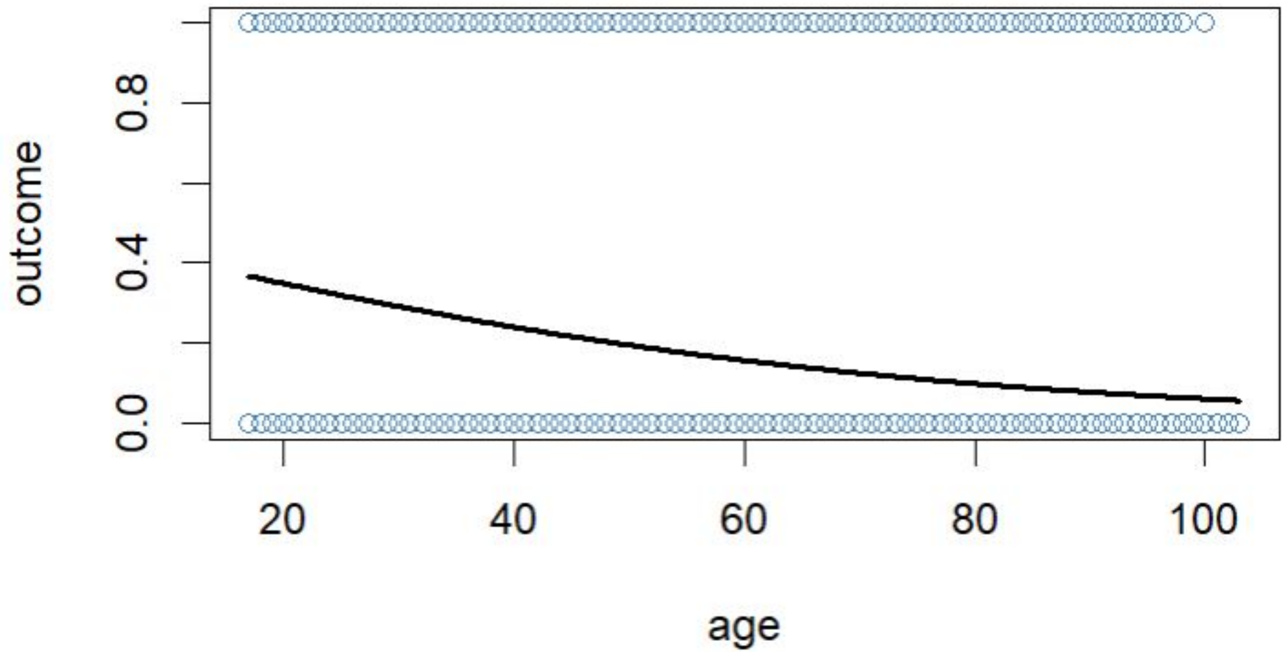
To get them back into **odds** we need to reverse the logarithm process. The reverse of logarithm is **exponentiating**.

$$\text{Odds} = \exp(\text{log odds}) - 1$$

You can exponentiate using the exp function on a calculator (or Google!). In R you can run the function exp( ).

Running a logistic regression model in R

logit_model_1 = glm((outcome) ~ age, family=binomial, data = total)



```
Call:
glm(formula = (outcome) ~ age, family = binomial, data = seminar_data)

Deviance Residuals:
    Min        1Q    Median        3Q       Max
-0.9572  -0.7583  -0.6281  -0.4502   2.3680

Coefficients:
              Estimate Std. Error z value             Pr(>|z|)
(Intercept) -0.0926609  0.0330006  -2.808              0.00499 **
age         -0.0264863  0.0007208 -36.744 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 50430  on 48284  degrees of freedom
Residual deviance: 48958  on 48283  degrees of freedom
AIC: 48962

Number of Fisher Scoring iterations: 4
```

Reporting a logistic regression model in R

```
Call:
glm(formula = (outcome) ~ age, family = binomial, data = seminar_data)

Deviance Residuals:
    Min        1Q    Median        3Q       Max
-0.9572  -0.7583  -0.6281  -0.4502   2.3680

Coefficients:
              Estimate Std. Error z value            Pr(>|z|)
(Intercept) -0.0926609  0.0330006  -2.808             0.00499 **
age         -0.0264863  0.0007208 -36.744 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 50430  on 48284  degrees of freedom
Residual deviance: 48958  on 48283  degrees of freedom
AIC: 48962

Number of Fisher Scoring iterations: 4
```

- **AIC: comparing model fit**
- p-values (Pr(>|t|)): whether the associations are statistically significant.
- Intercept/slope (Estimate): The strength and direction of the relationship
  - Direction
  - Effect size
  - Confidence intervals: use confint(model)

Reporting a logistic regression model in R

```
Call:
glm(formula = (outcome) ~ age, family = binomial, data = seminar_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9572  -0.7583  -0.6281  -0.4502   2.3680

Coefficients:
              Estimate Std. Error z value             Pr(>|z|)
(Intercept) -0.0926609  0.0330006  -2.808              0.00499 **
age         -0.0264863  0.0007208 -36.744 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 50430  on 48284  degrees of freedom
Residual deviance: 48958  on 48283  degrees of freedom
AIC: 48962

Number of Fisher Scoring iterations: 4
```

- AIC: comparing model fit
- **p-values** (Pr(>|t|)): whether the associations are statistically significant.
- Intercept/slope (Estimate): The strength and direction of the relationship
  - Direction
  - Effect size
  - Confidence intervals: use confint(model)

Reporting a logistic regression model in R

```
Call:
glm(formula = (outcome) ~ age, family = binomial, data = seminar_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9572  -0.7583  -0.6281  -0.4502   2.3680

Coefficients:
              Estimate Std. Error z value             Pr(>|z|)
(Intercept) -0.0926609  0.0330006  -2.808              0.00499 **
age         -0.0264863  0.0007208 -36.744 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 50430  on 48284  degrees of freedom
Residual deviance: 48958  on 48283  degrees of freedom
AIC: 48962

Number of Fisher Scoring iterations: 4
```

- AIC: comparing model fit
- p-values (Pr(>|t|)): whether the associations are statistically significant.
- Intercept/slope (**Estimate**): The strength and direction of the relationship
  - Direction
  - Effect size
  - Confidence intervals: use confint(model)

Interpreting log odds as odds

```
Coefficients:
              Estimate
(Intercept) -0.0926609
age         -0.0264863
---
```

Odds = exp(log odds) - 1

$\exp(-0.0264863) = 0.97$

Odds = $1-0.97 = -0.026$

-0.026 -> 2.6% decrease

An age increase in 1 year is associated with a 2.6% decrease of the likelihood of housing payment problems.

Running a logistic regression model in R

health_condition: does the respondent have a health condition (1) or not (0)?
benefit_group: is the respondent in the new (1) or old (0) benefit system?

```
# add explanatory variables

logit_model_2 = glm((outcome) ~ age + health_condition + benefit_group,
family=binomial, data = total)

summary(logit_model_2)
```

```
Call:
glm(formula = (outcome) ~ age + health_condition + benefit_group,
    family = binomial, data = seminar_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0298  -0.7573  -0.6284  -0.4491   2.3859

Coefficients:
                   Estimate Std. Error z value              Pr(>|z|)
(Intercept)      -0.1245068  0.0341081  -3.650              0.000262 ***
age              -0.0271632  0.0007742 -35.084 < 0.0000000000000002 ***
health_condition  0.0964513  0.0239085   4.034             0.0000548 ***
benefit_group     0.1323283  0.0347291   3.810              0.000139 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 50430  on 48284  degrees of freedom
Residual deviance: 48928  on 48281  degrees of freedom
AIC: 48936

Number of Fisher Scoring iterations: 4
```

```
Coefficients:
                  Estimate :
(Intercept)      -0.1245068
age              -0.0271632
health_condition  0.0964513
benefit_group     0.1323283
---
```

Odds = exp(log odds) - 1

Variable	**Log odds**	**exp(log odds)**	**exp(log odds) - 1**	**Odds %**
Age	-0.0271632	0.97	-0.027	2.7% decrease
Health condition				
Benefit group				

```
Coefficients:
                   Estimate 
(Intercept)       -0.1245068
age               -0.0271632
health_condition   0.0964513
benefit_group      0.1323283
---
```

Odds = exp(log odds) - 1

Variable	**Log odds**	**exp(log odds)**	**exp(log odds) - 1**	**Odds %**
Age	-0.0271632	0.97	-0.027	2.7% decrease
Health condition	0.0964513			
Benefit group				

```
Coefficients:
                 Estimate
(Intercept)     -0.1245068
age             -0.0271632
health_condition 0.0964513
benefit_group    0.1323283
---
```

Odds = exp(log odds) - 1

Variable	**Log odds**	**exp(log odds)**	**exp(log odds) - 1**	**Odds %**
Age	-0.0271632	0.97	-0.027	2.7% decrease
Health condition	0.0964513	1.101		
Benefit group				

```
Coefficients:
                  Estimate
(Intercept)      -0.1245068
age              -0.0271632
health_condition  0.0964513
benefit_group     0.1323283
---
```

Odds = exp(log odds) - 1

Variable	**Log odds**	**exp(log odds)**	**exp(log odds) - 1**	**Odds %**
Age	-0.0271632	0.97	-0.027	2.7% decrease
Health condition	0.0964513	1.101	0.101	
Benefit group				

```
Coefficients:
                    Estimate :
(Intercept)        -0.1245068
age                -0.0271632
health_condition    0.0964513
benefit_group       0.1323283
---
```

Odds = exp(log odds) - 1

Variable	Log odds	exp(log odds)	exp(log odds) - 1	Odds %
Age	-0.0271632	0.97	-0.027	2.7% decrease
Health condition	0.0964513	1.101	0.101	10.1% increase
Benefit group				

```
Coefficients:
                    Estimate
(Intercept)       -0.1245068
age               -0.0271632
health_condition   0.0964513
benefit_group      0.1323283
---
```

Odds = exp(log odds) - 1

Variable	**Log odds**	**exp(log odds)**	**exp(log odds) - 1**	**Odds %**
Age	-0.0271632	0.97	-0.027	2.7% decrease
Health condition	0.0964513	1.101	0.101	10.1% increase
Benefit group	0.1323283			

```
Coefficients:
                   Estimate
(Intercept)      -0.1245068
age              -0.0271632
health_condition  0.0964513
benefit_group     0.1323283
---
```

Odds = exp(log odds) - 1

Variable	**Log odds**	**exp(log odds)**	**exp(log odds) - 1**	**Odds %**
Age	-0.0271632	0.97	-0.027	2.7% decrease
Health condition	0.0964513	1.101	0.101	10.1% increase
Benefit group	0.1323283	1.14	0.14	14% increase

Visualising logistic regression predictions

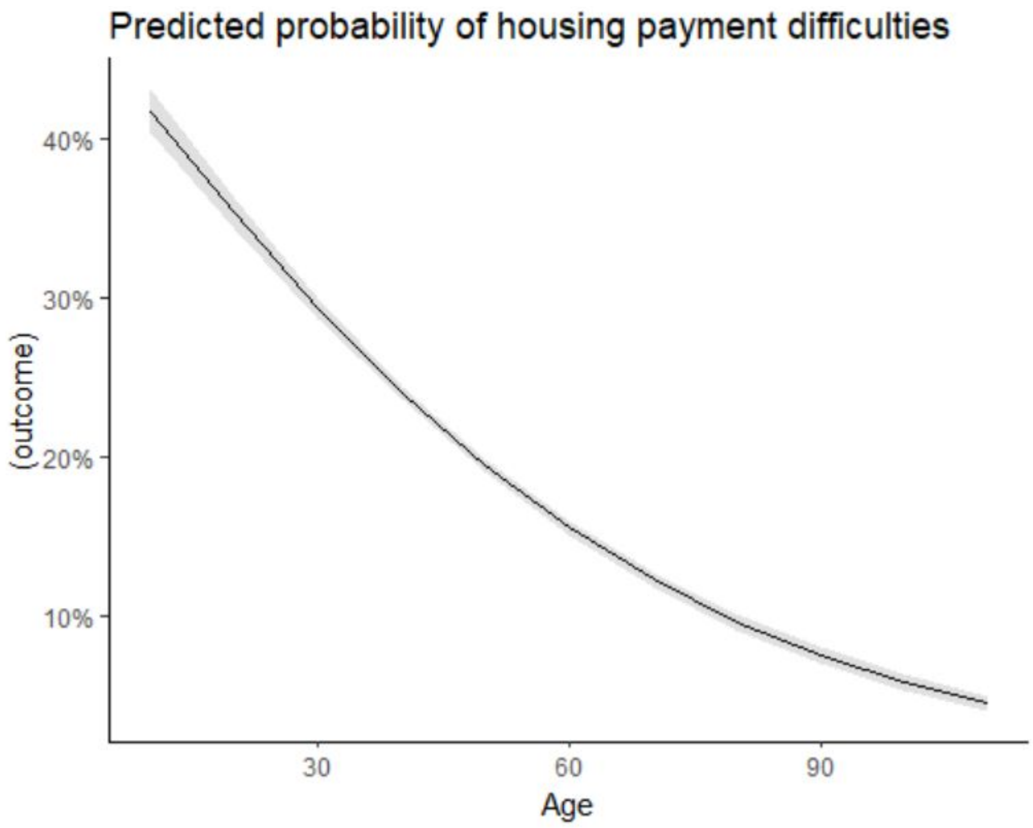
```
ggeffect(logit_model_2, terms
= "age") %>%

  plot() +

  ggtitle("Predicted probability
of housing payment
difficulties") +

  xlab("Age") +

  theme_classic()
```



Testing logistic regression predictions

```
# add predictions to data
seminar_data <- seminar_data %>%
  mutate(
    predictions = predict(logit_model_2, type = "response", newdata =
seminar_data))

# change predictions to binary
seminar_data <- seminar_data %>%
  mutate(
    predictions = case_when(is.na(predictions) ~ NA_real_,
                            predictions >= 0.5 ~ 1,
                            TRUE ~ 0))
```

Testing logistic regression predictions

```
# add predictions to data
seminar_data <- seminar_data %>%
  mutate(
    predictions = predict(logit_model_2, type = "response", newdata =
seminar_data))

# change predictions to binary
seminar_data <- seminar_data %>%
  mutate(
    predictions = case_when(is.na(predictions) ~ NA_real_,
                            predictions >= 0.5 ~ 1,
                            TRUE ~ 0))
```

Accuracy: 0.78
78% of the predictions were accurate

Summary

- Logistic regression expands the types of research questions and dependant variables we can explore in our analysis, allowing us to analyse binary outcomes.
- Logistic regression also lets us analyse datasets that don't fit the requirements for linear regression, such as non-linear data.
- We can build a logistic regression model in R using the command glm(x ~ y, family=binomial).
- When interpreting logistic regression models, we convert the coefficient or log odds to odds, making them easier to interpret.

R Exercise

This week, we are going to apply logistic regression to US Census and the Southern Poverty Law Center data on active hate groups in the USA.

- Download the `week-9-r-exercises.zip` file from Blackboard and open the .Rproj folder and .Rmd file.