Introduction to linear regression

A/B TESTING IN R

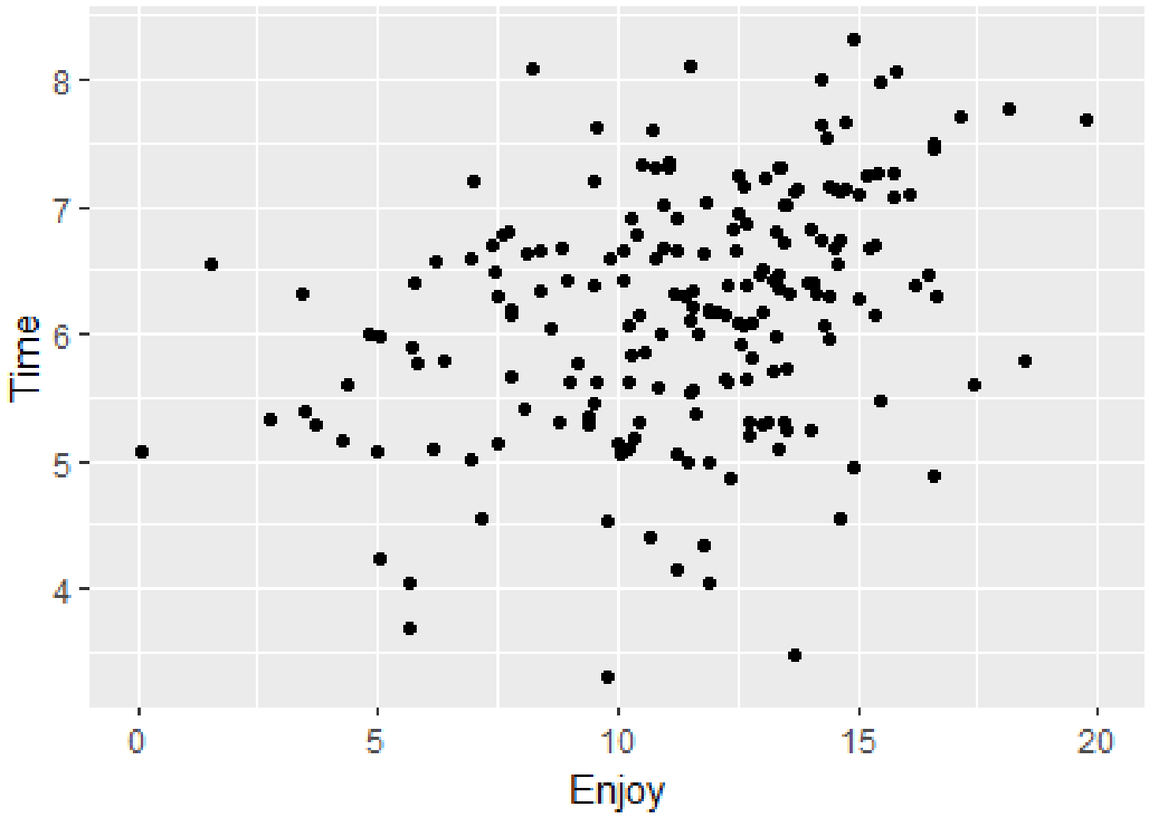


Lauryn Burleigh
Data Scientist

datacamp

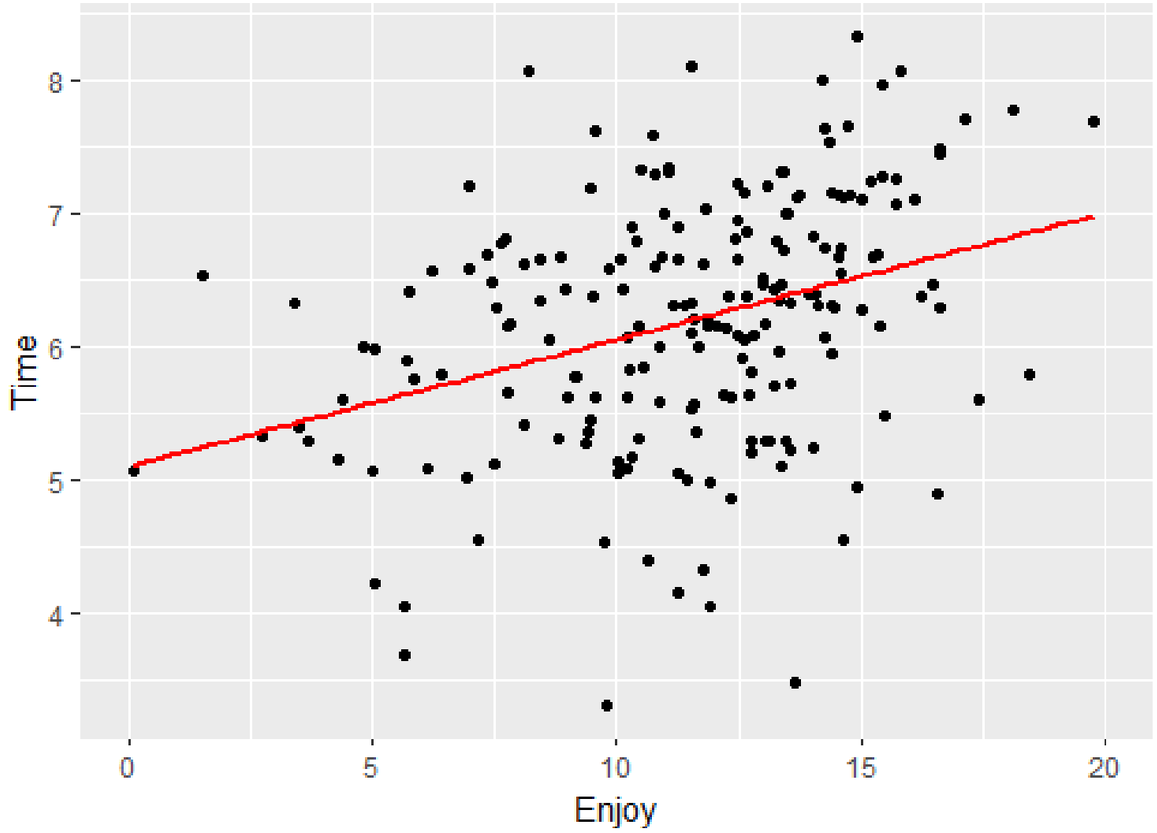
Regression

- A/B design - time to eat Cheese or Pepperoni pizza
- Regression - factors that impact eating time
  - Dependent variable - time to eat
  - Independent variables - enjoyment, hunger, visual appeal



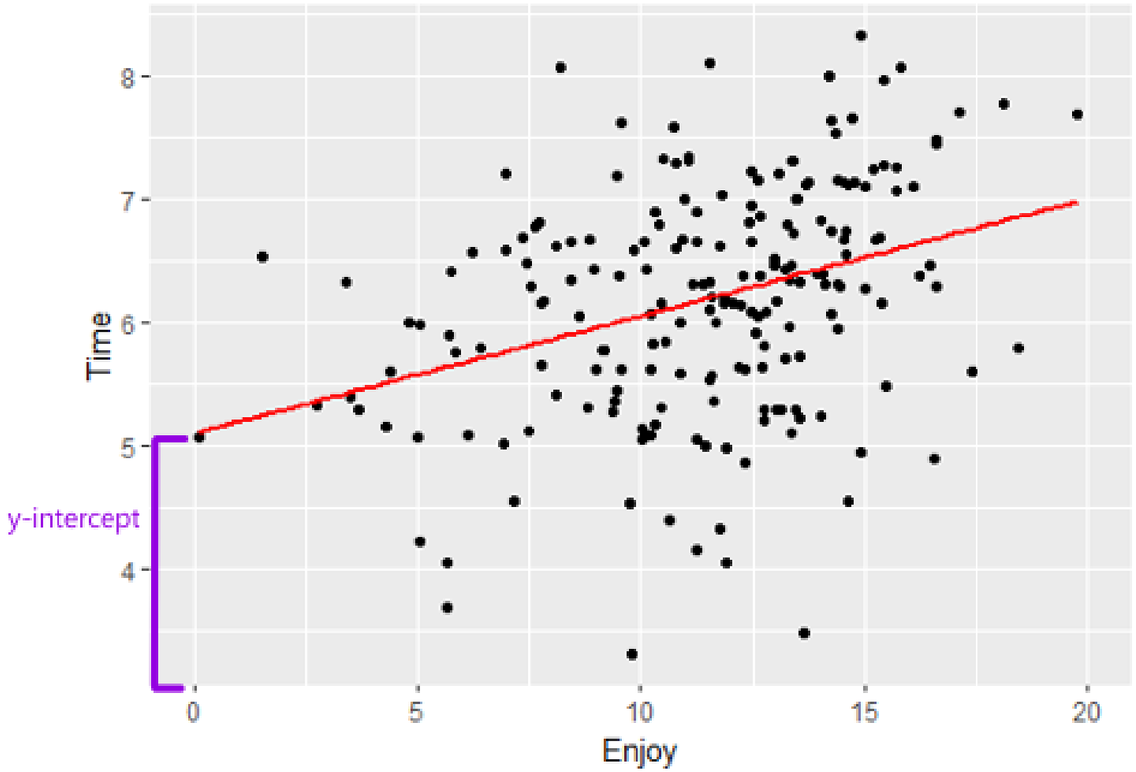
Regression line

- Regression line - prediction of y



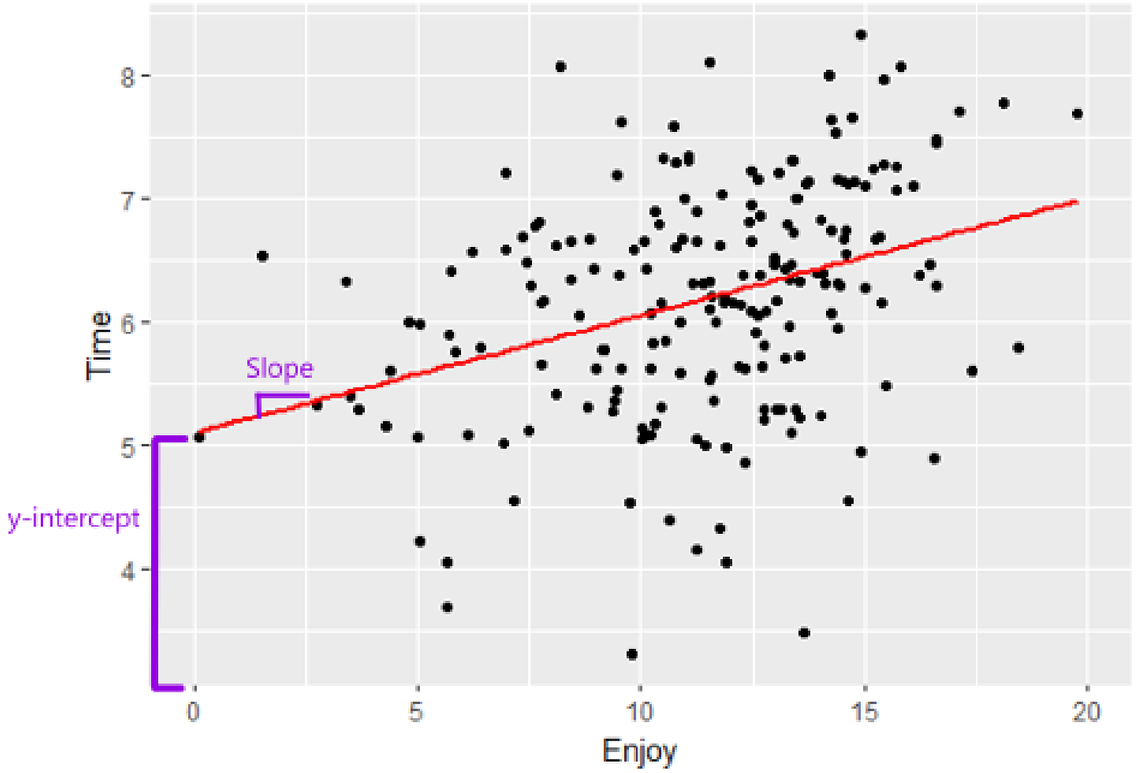
Regression line

- Regression line - prediction of y
- $\hat{y} = \beta_0 + \beta_1 X_1 + \varepsilon$
  - β_0 - y-intercept



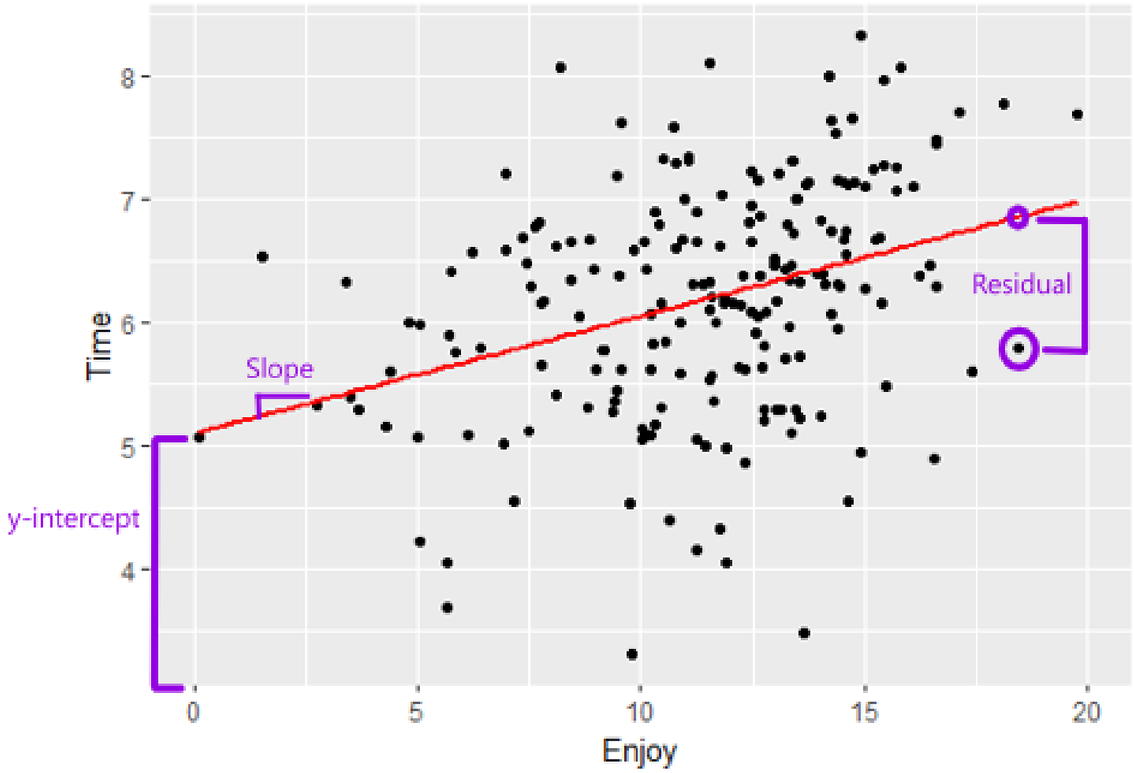
Regression line

- Regression line - prediction of y
- $\hat{y} = \beta_0 + \beta_1 X_1 + \varepsilon$
  - β_0 - y-intercept
  - β_1 - slope



Regression line

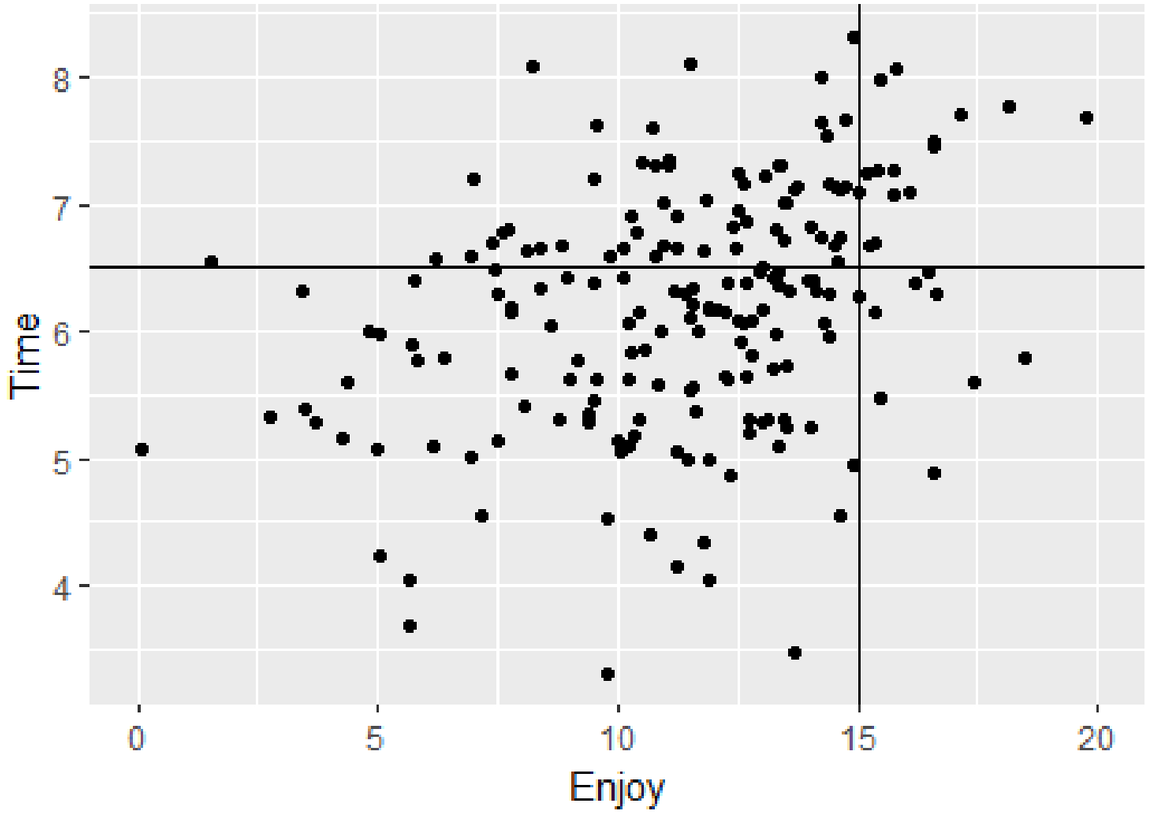
- Regression line - prediction of y
- $\hat{y} = \beta_0 + \beta_1 X_1 + \varepsilon$
  - β_0 - y-intercept
  - β_1 - slope
  - ε - error
- Reduce error
  - Multiple regression - 2+ independent variables
  - $\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$



Predicting data

```
yhat <- 5.32 + 0.08*(15)
```

```
ggplot(pizza, aes(x = Enjoy,
                  y = Time)) +
  geom_point() +
  geom_hline(yintercept = yhat) +
  geom_vline(xintercept = 15)
```



Regression considerations

- Correlation does not imply causation
- Assess reasonable variables
- Decisions/actions data lead to
- Quality data



Let's practice!

A/B TESTING IN R

Linear regression

A/B TESTING IN R

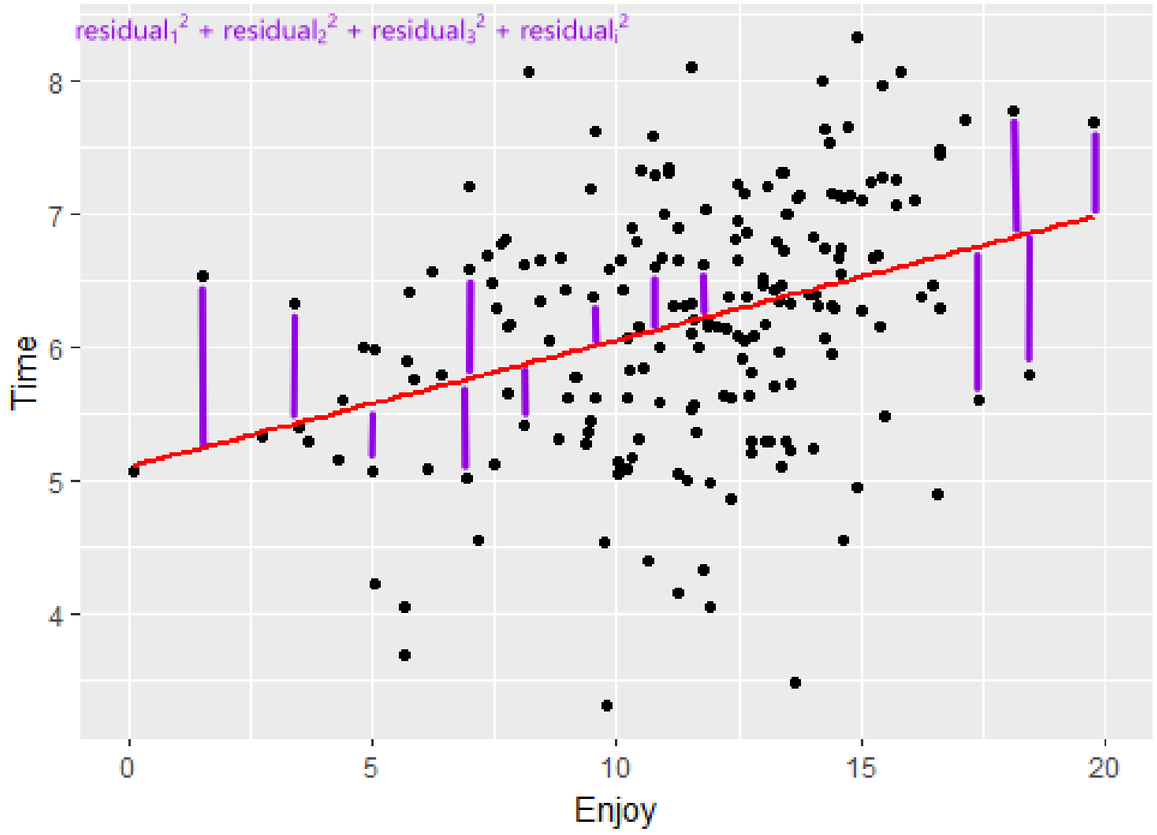


Lauryn Burleigh
Data Scientist

datacamp

Least squares

- Residual sum of squares
  - Each point residual - squared, then summed
- Line of best fit - smallest sum of squares
- Error - mean square error
  - Sum of squares / N



Linear regression model

```
linear <- lm(Time ~ Enjoy,
             data = Pizza)
summary(linear)
```

```
Call:
lm(formula = Time ~ Enjoy, data = pizza)

Residuals:
     Min       1Q   Median       3Q      Max
-2.89270 -0.59857  0.04758  0.67764  2.12600

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.31964    0.19886  26.750  < 2e-16 ***
Enjoy        0.07707    0.01672   4.608 7.26e-06 ***

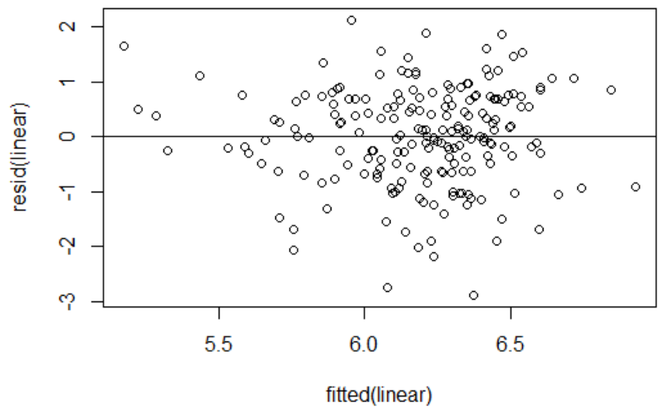
Residual standard error: 0.8947 on 198 degrees of freedom
Multiple R-squared:  0.09687,
Adjusted R-squared:  0.09231
F-statistic: 21.24 on 1 and 198 DF,  p-value: 7.262e-06
```

Assessing assumptions

Homoscedasticity

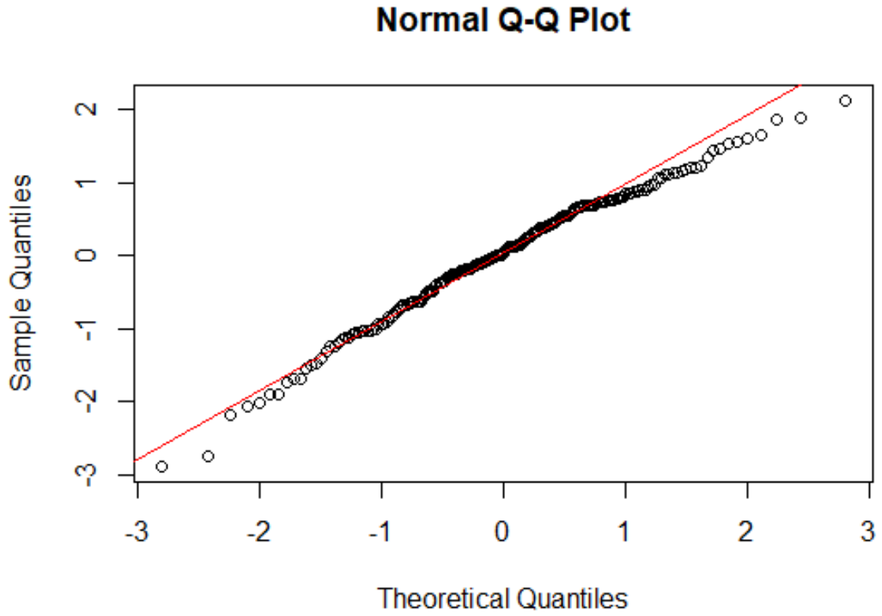
- Consistent variance

```
plot(fitted(linear), resid(linear));
abline(0,0)
```



Normality

```
qqnorm(resid(linear));
qqline(resid(linear), col = "red")
```



Making predictions

```
Enjoy  <- 12
topredict <-  data.frame(Enjoy)
predict(linear, newdata = topredict)
```

```
       1
6.244452
```

```
Enjoy  <- c(12, 14)
topredict <-  data.frame(Enjoy)
predict(linear, newdata = topredict)
```

```
       1        2
6.244452 6.398587
```

Including groups

```
grplinear <- lm(Time ~ Enjoy + Topping,
                data = Pizza)
summary(grplinear)
```

```
Call:
lm(formula = Time ~ Enjoy + Topping,
data = pizza)

Residuals:
     Min       1Q   Median       3Q      Max
-2.87771 -0.51529  0.03993  0.68685  2.19460

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.53891    0.24823  22.314  < 2e-16 ***
Enjoy            0.06656    0.01815   3.668 0.000315 ***
ToppingCheese -0.20159    0.13729  -1.468 0.143606

Residual standard error: 0.8921 on 197 degrees of freedom
Multiple R-squared:  0.1066,
Adjusted R-squared:  0.09758
F-statistic: 11.76 on 2 and 197 DF,  p-value: 1.499e-05
```

```
Enjoy  <- c(12, 14)
Topping <- "Cheese"
topredict <-  data.frame(Enjoy, Topping)
predict(grplinear, newdata = topredict)
```

```
       1        2
6.136022 6.269139
```

Let's practice!

A/B TESTING IN R

Logistic regression

A/B TESTING IN R

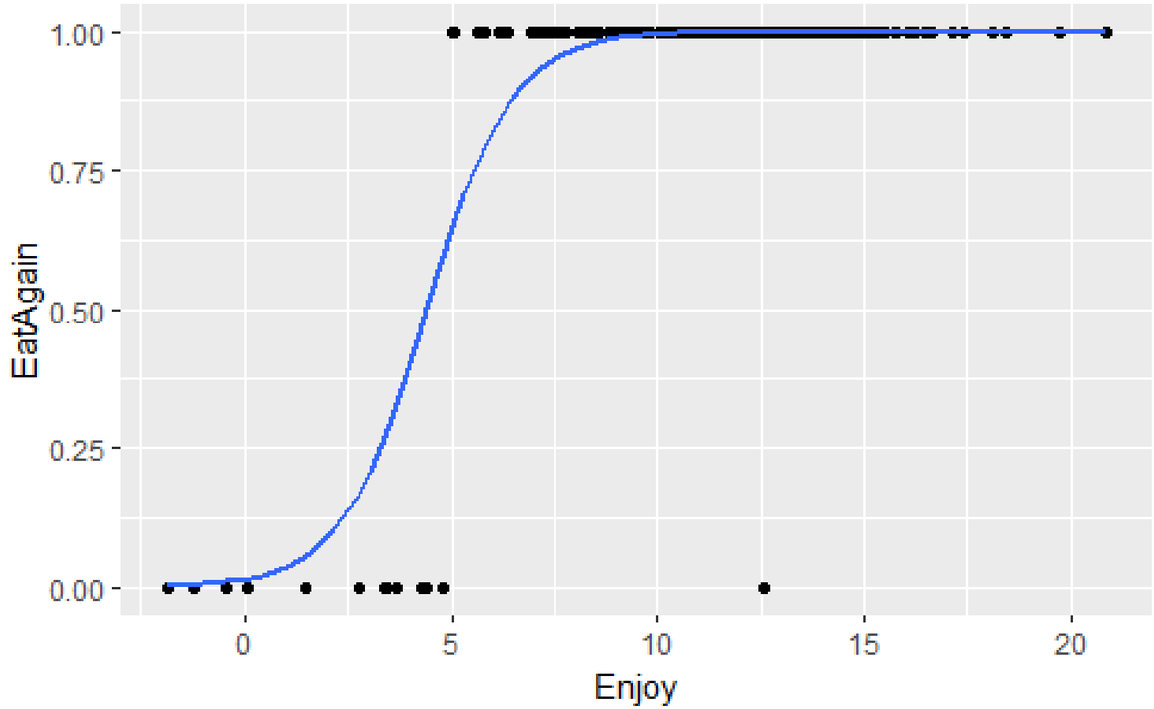


Lauryn Burleigh
Data Scientist

datacamp

Log in logistic

- Binary
  - Eat pizza again - yes/no
- Derive probability of category
- Predicted dependent variable
  - Odds = P / 1-P
  - Linear to independent variable
- $\log(P / 1\text{-}P) = \beta_0 + \beta_1 X_1$
  - error within log odds



Logistic regression model

```
logistic <- glm(EatAgain ~ Enjoy,
                data = Pizza,
                family = binomial)
summary(logistic)
```

```
chival <- logistic$null.deviance - logistic$deviance
dfval <- logistic$df.null - logistic$df.residual
pchisq(q = chival, df = dfval, lower.tail = FALSE)
```

```
[1] 7.472441e-16
```

```
Call:
glm(formula = EatAgain ~ Enjoy, family = binomial,
data = pizza)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9677   0.0116   0.0297   0.0885   0.9446

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.2361     1.4757  -2.871 0.004097 **
Enjoy         0.9611     0.2533   3.794 0.000148 ***

Null deviance: 96.204  on 199  degrees of freedom
Residual deviance: 31.199  on 198  degrees of freedom
AIC: 35.199
Number of Fisher Scoring iterations: 9
```

Predicting data

```
Enjoy  <- c(12, 14)
topredict <-  data.frame(Enjoy)
predict(logistic, topredict,
        type = "response")
```

```
        1         2
0.9993229 0.9999009
```

Including groups

```
grplogistic <- glm(EatAgain ~ Enjoy +
                   Topping,
                   data = pizza,
                   family = binomial)
summary(grplogistic)
```

```
Call:
glm(formula = EatAgain ~ Enjoy + Topping, family = binomial,
    data = pizza)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.8557   0.0058   0.0222   0.0862   0.8887

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -9.4498     3.8806  -2.435  0.01489 *
Enjoy             1.3402     0.4118   3.255  0.00114 **
ToppingCheese   3.4652     2.0499   1.690  0.09095 .

Null deviance: 96.204  on 199  degrees of freedom
Residual deviance: 28.232  on 197  degrees of freedom
AIC: 34.232

Number of Fisher Scoring iterations: 9
```

Let's practice!

A/B TESTING IN R

Plotting regression

A/B TESTING IN R

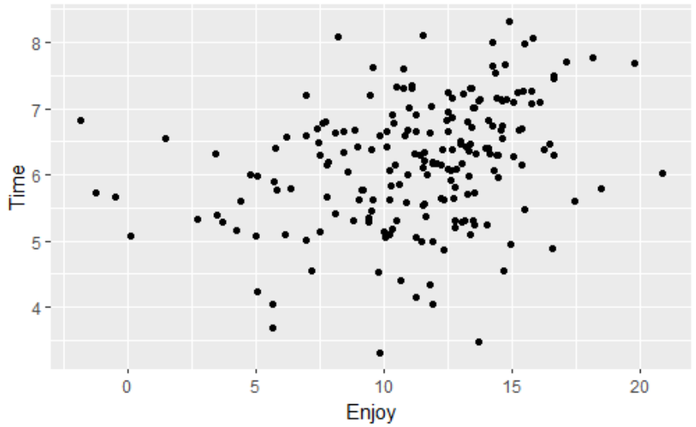


Lauryn Burleigh
Data Scientist

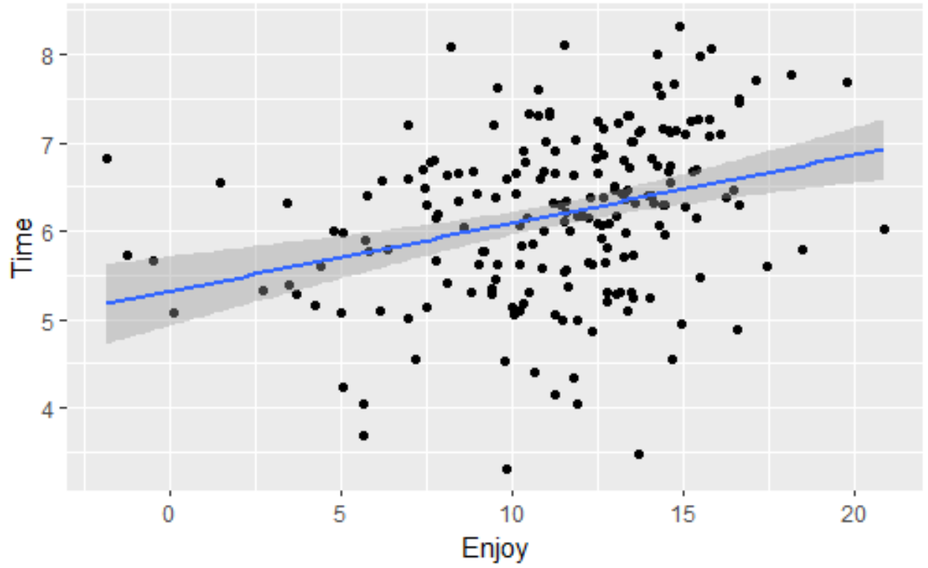
datacamp

Linear line

```
cor.test(~ Enjoy + Time, data = pizza,
         method = "pearson")
library(ggplot2)
ggplot(pizza, aes(x = Enjoy,
                  y = Time)) +
  geom_point()
```



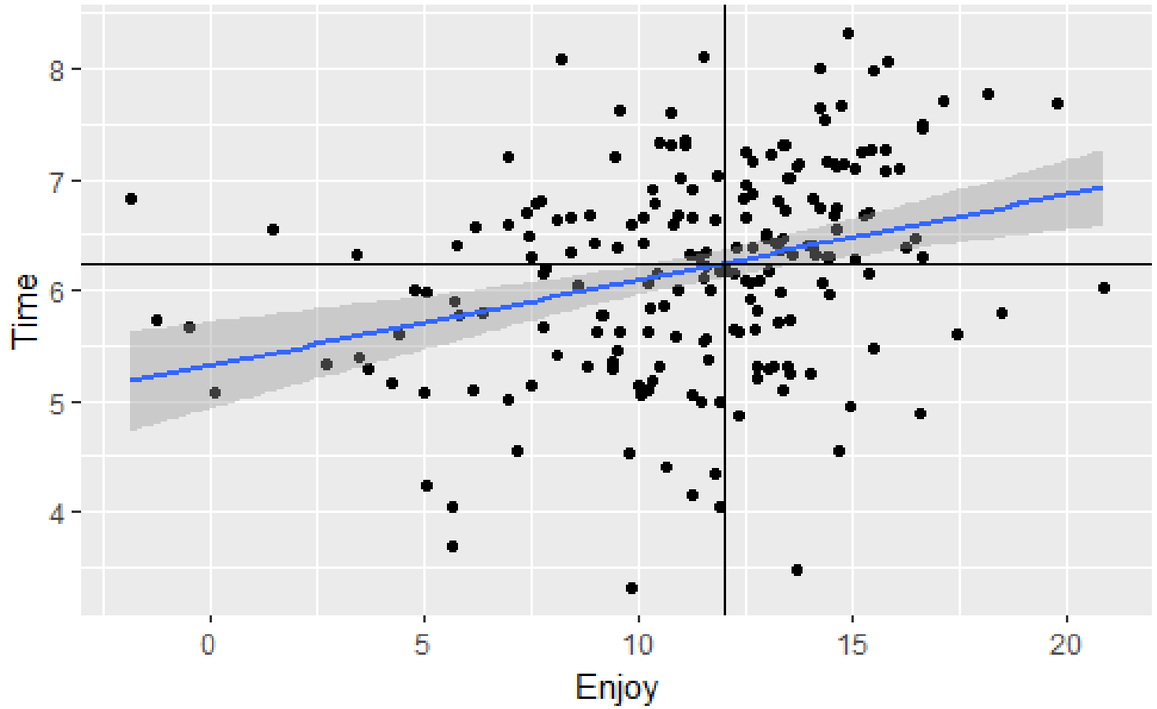

```
ggplot(pizza, aes(x = Enjoy,
                  y = Time)) +
  geom_point() +
  geom_smooth(method = "lm")
```



Linear prediction

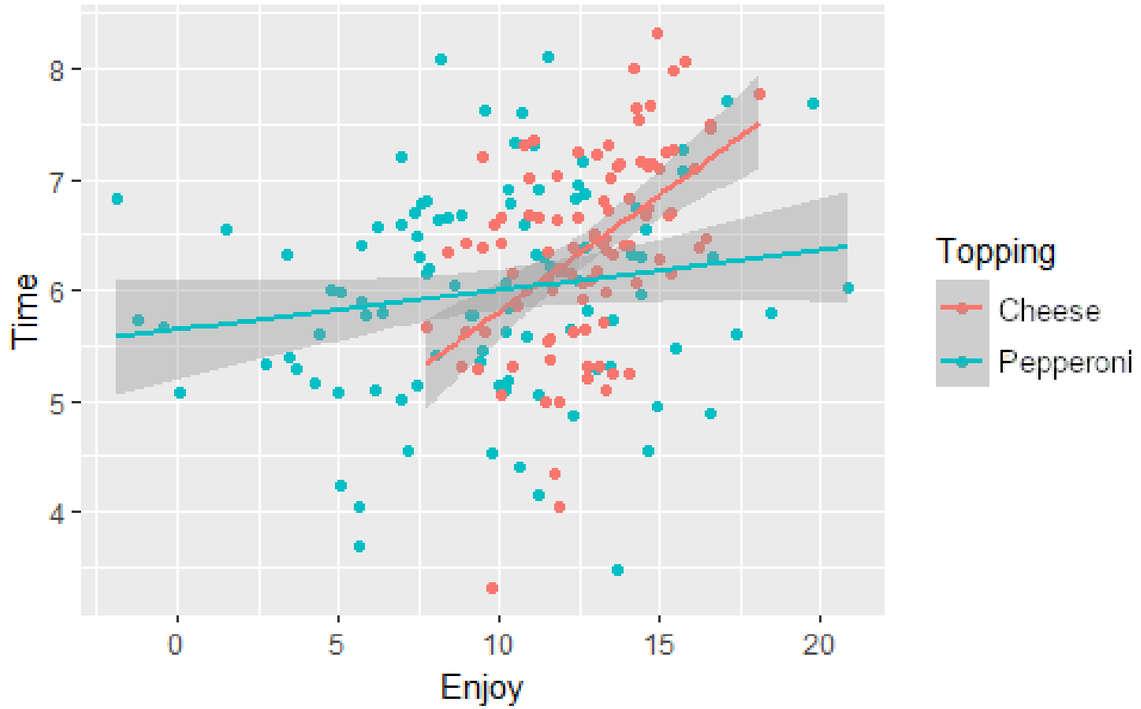
```
linear <- lm(Time ~ Enjoy,
                 data = Pizza)
Enjoy  <- 12
topredict <-  data.frame(Enjoy)
prediction <- predict(linear,
                      newdata = topredict)

ggplot(pizza, aes(x = Enjoy, y = Time)) +
  geom_point() +
  geom_smooth(method = "lm") +
  geom_hline(yintercept = prediction) +
  geom_vline(xintercept = Enjoy)
```



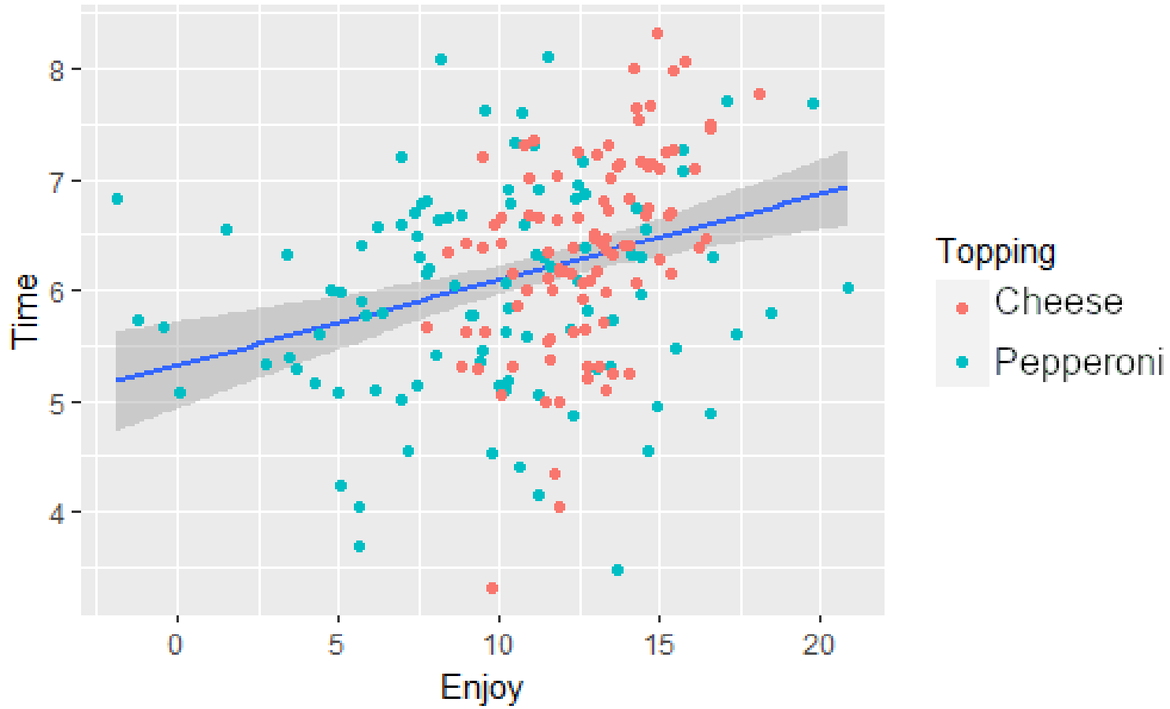
Denoting groups

```
ggplot(pizza, aes(x = Enjoy, y = Time,
                  color = Topping)) +
  geom_point() +
  geom_smooth(method = "lm")
```



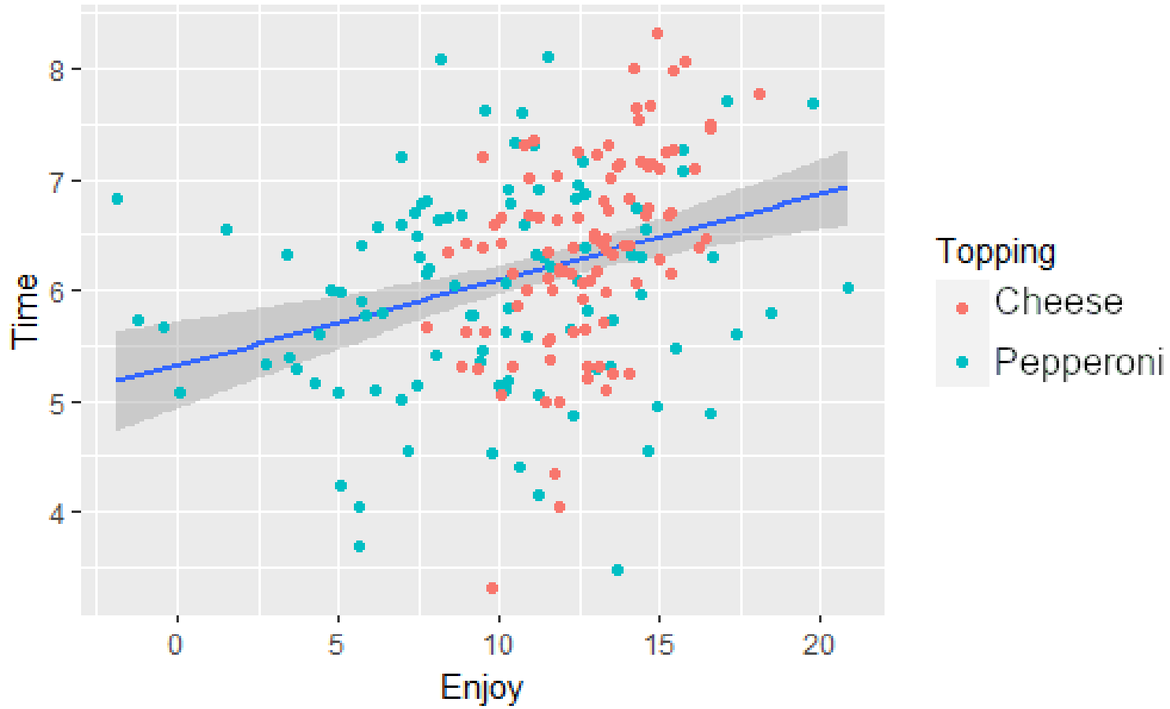
Denoting groups

```
ggplot(pizza, aes(x = Enjoy,
                  y = Time)) +
  geom_point(aes(colour = Topping)) +
  geom_smooth(method = "lm")
```



Logistic line

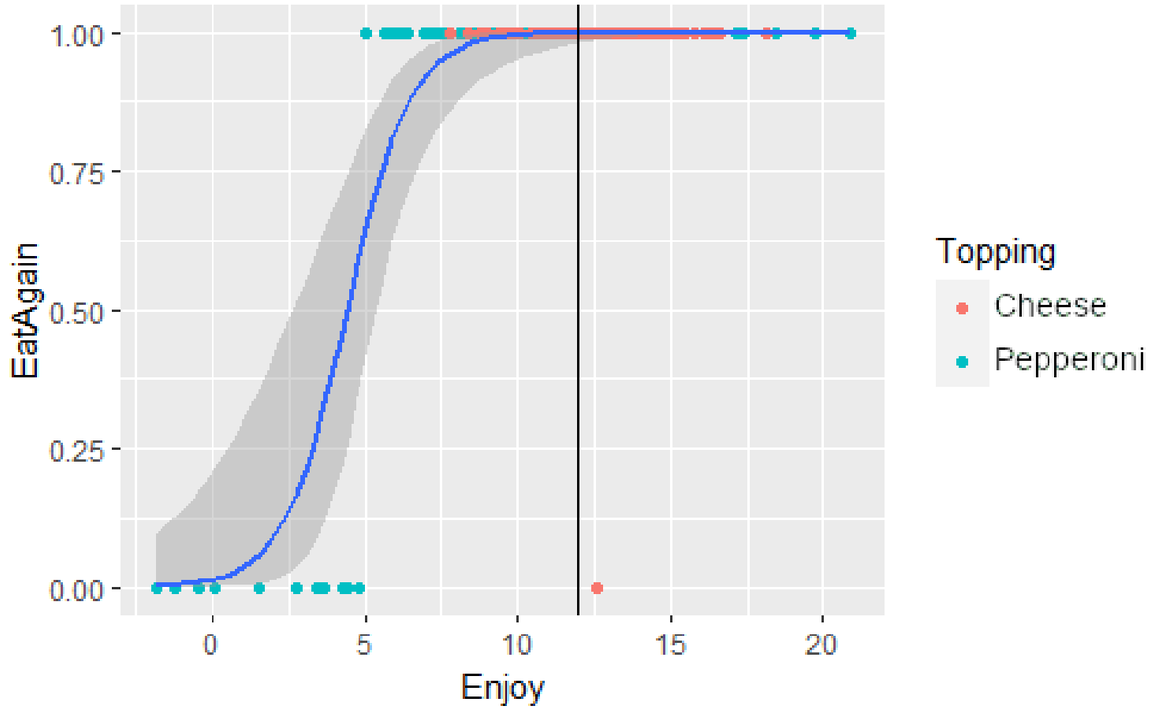
```
ggplot(pizza, aes(x = Enjoy,
                  y = EatAgain)) +
  geom_point(aes(colour = Topping)) +
  geom_smooth(method = "glm",
              method.args =
                list(family =
                     "binomial"))
```



Logistic prediction

```
Enjoy  <- 12

ggplot(pizza2, aes(x = Enjoy,
                   y = EatAgain)) +
  geom_point(aes(colour = Topping)) +
  geom_smooth(method = "glm",
              method.args =
              list(family =
                   "binomial")) +
  geom_vline(xintercept = Enjoy)
```



Let's practice!

A/B TESTING IN R

Wrap-up

A/B TESTING IN R



Lauryn Burleigh
Cognitive Neuroscientist

How far you've come

- A/B design

	Comparing Groups	Associating Variables	Predicting Data
Parametric	t-test `t.test()`	Pearson Correlation `cor.test()`	Linear Regression `lm()`
Non-Parametric	Mann-Whitney U test `wilcox.test()`	Spearman Correlation `cor.test()`	Logistic Regression `glm()`
	Chi-Square test of Independence `chisq.test()`		
	Fisher's Exact `fisher.test()`		

- Data visualization: `ggplot()`
- Data presentation

Further learning

DataCamp Courses

- Introduction to Regression in R
- Experimental Design in R
- Foundations of Inference in R

Congratulations!

A/B TESTING IN R