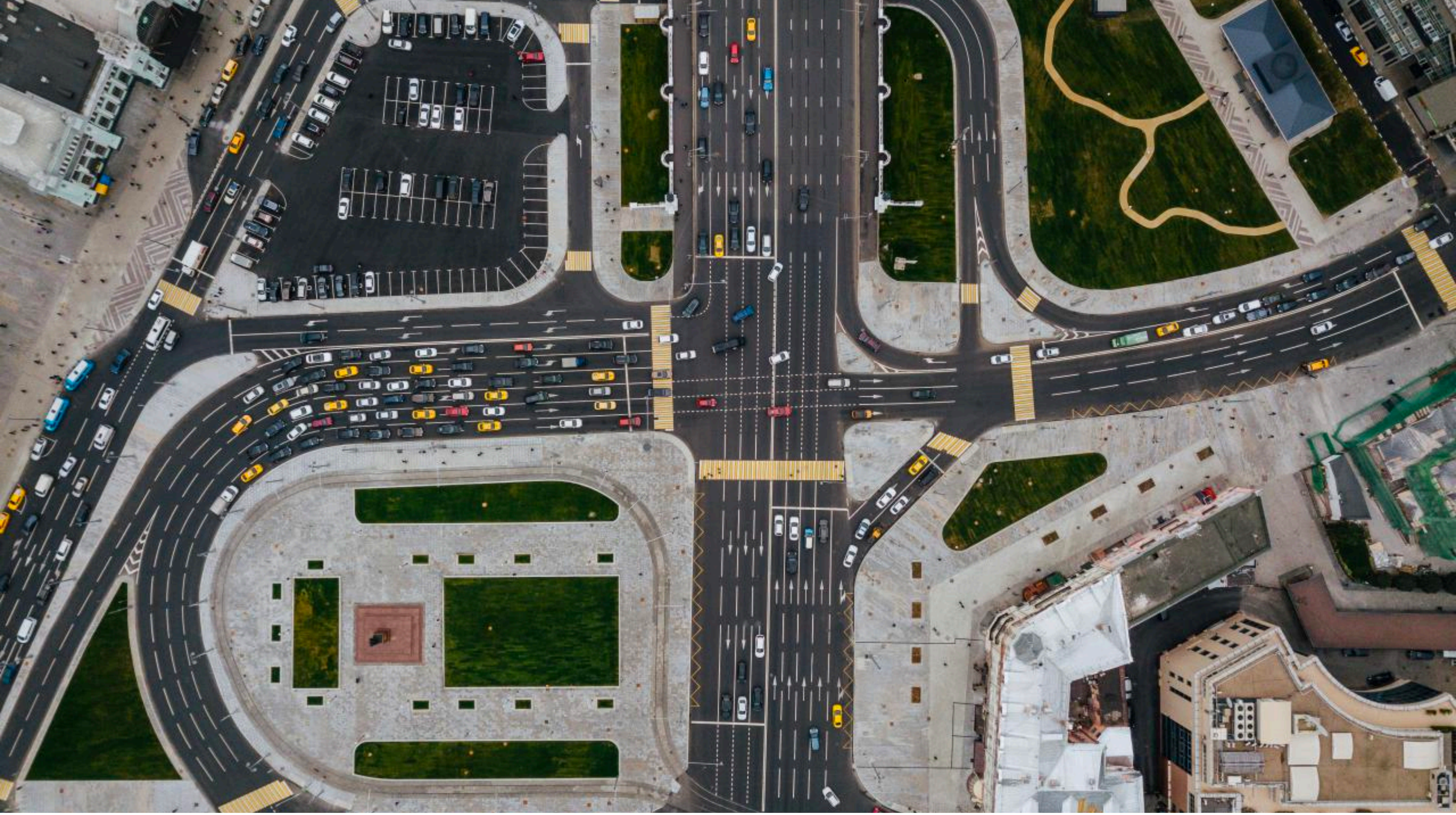


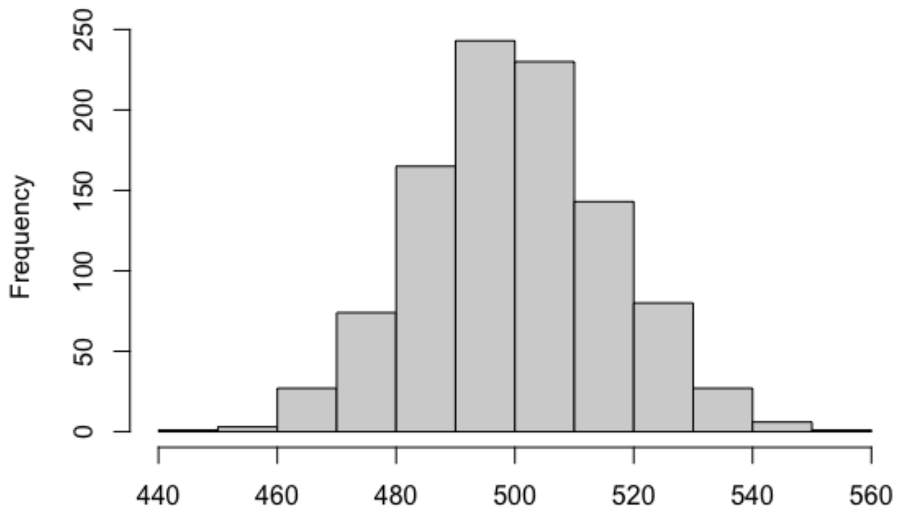
Interaction analysis

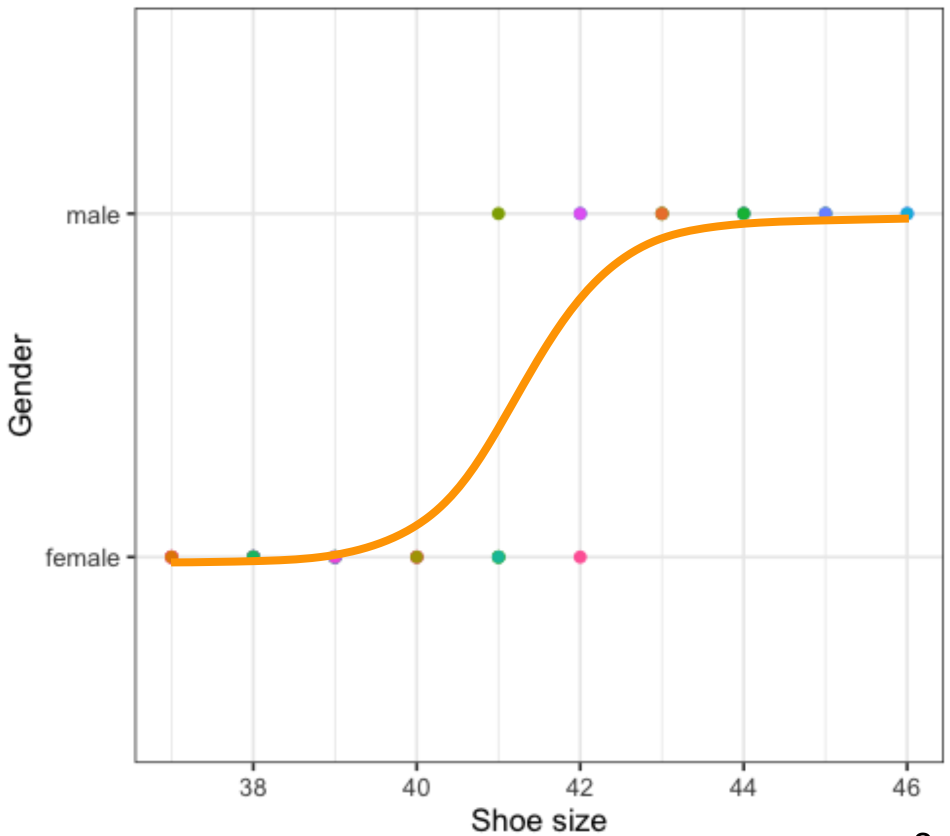
Methods 1, E2021 - Lecture 11
Tuesday 23/11/2021
Fabio Trecca


QUIZ
TIME

Recap: Bernoulli trials and the binomial distribution

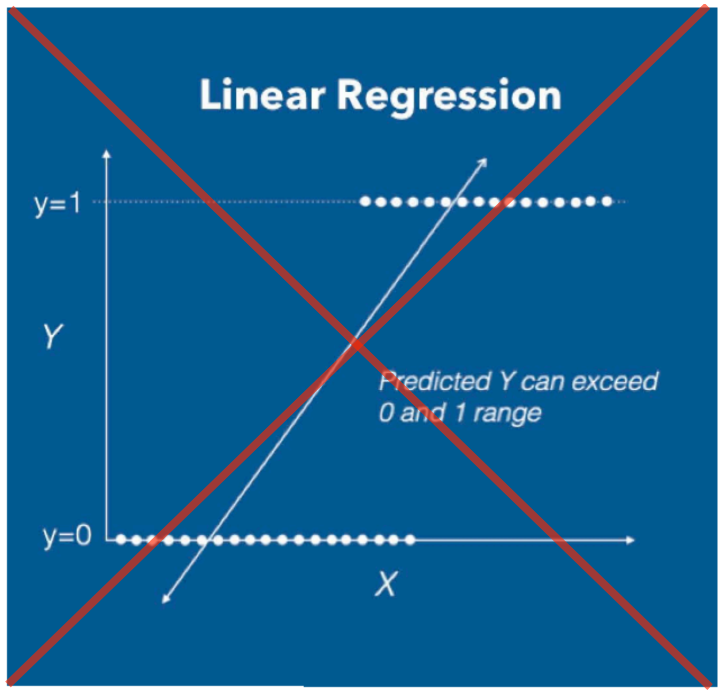
- $p + q = 1$
- $p = 1 - q$
- $q = 1 - p$
- $odds = \dfrac{p}{1-p}$
- $P_x = \binom{n}{k} p^x q^{n-x}$
- conditional probability = $P(Y_i | X_i)$

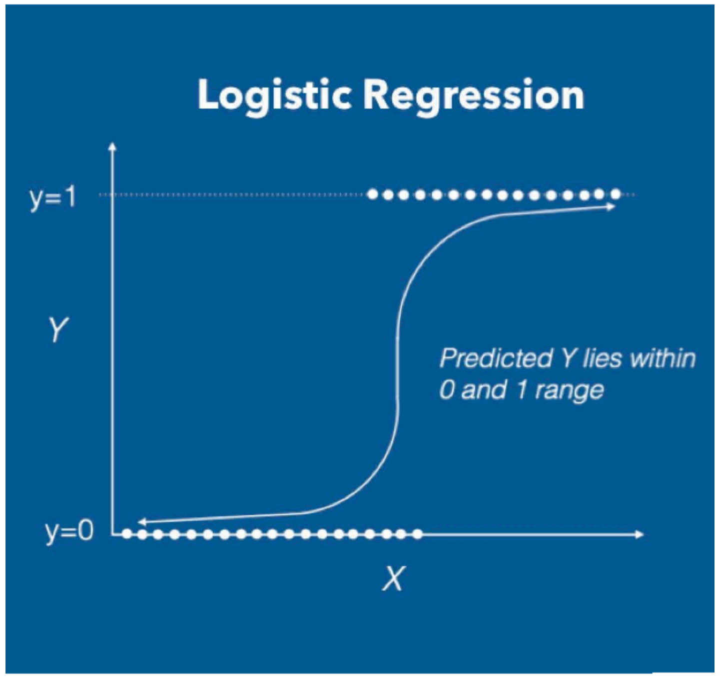




Recap: Linear vs logistic regression

- **Linear regression**:
  - gives the predicted mean value of an outcome variable at a particular value of a predictor variable
- **Logistic regression**:
  - gives the conditional probability that an outcome variable equals one at a particular value of a predictor variable
  - conditional probability = $P(Y_i \mid X_i)$





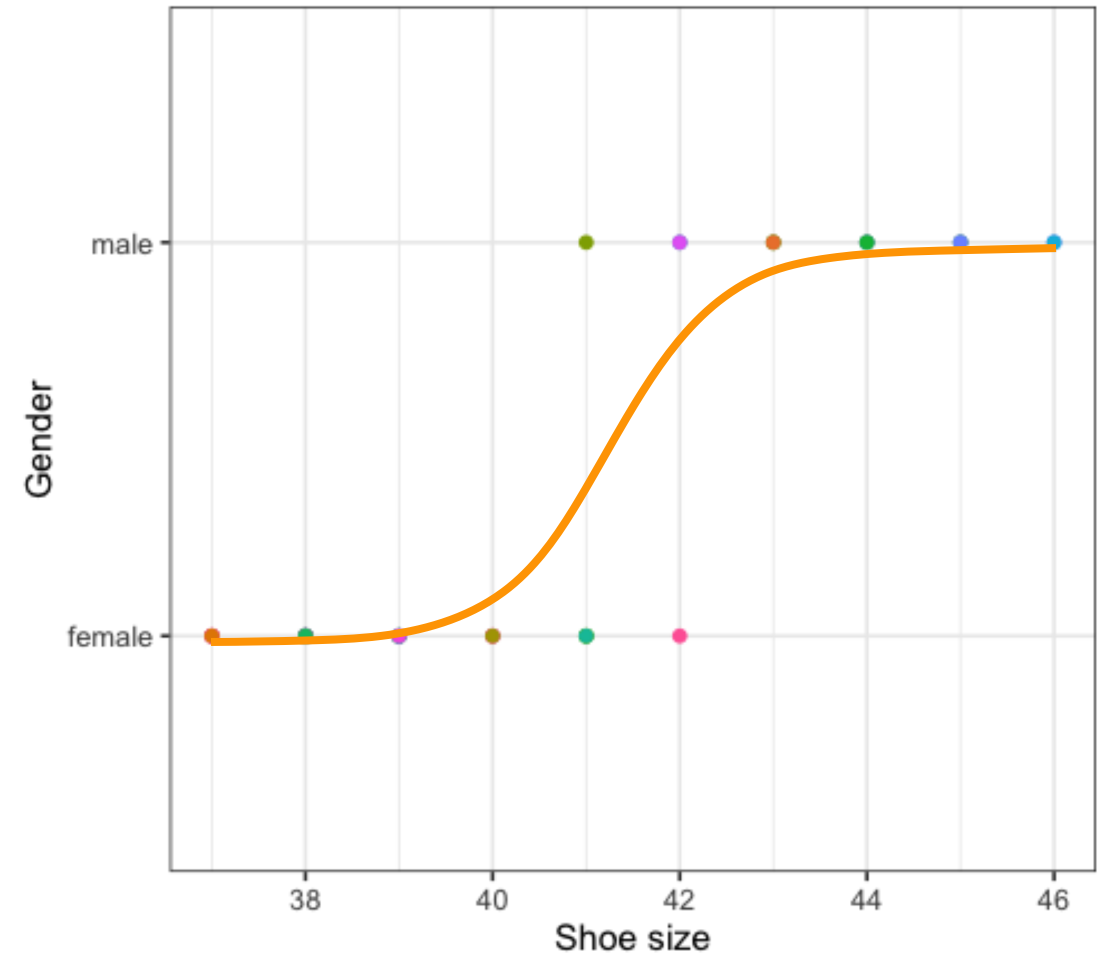
Recap: Logistic regression equation

- $$P(Y_i) = \frac{e^{(\beta_0 + \beta_1 X_i + \varepsilon_i)}}{1 + e^{(\beta_0 + \beta_1 X_i + \varepsilon_i)}}$$

- where $\beta_0 + \beta_1 X_i + \varepsilon_i$ = simple linear regression

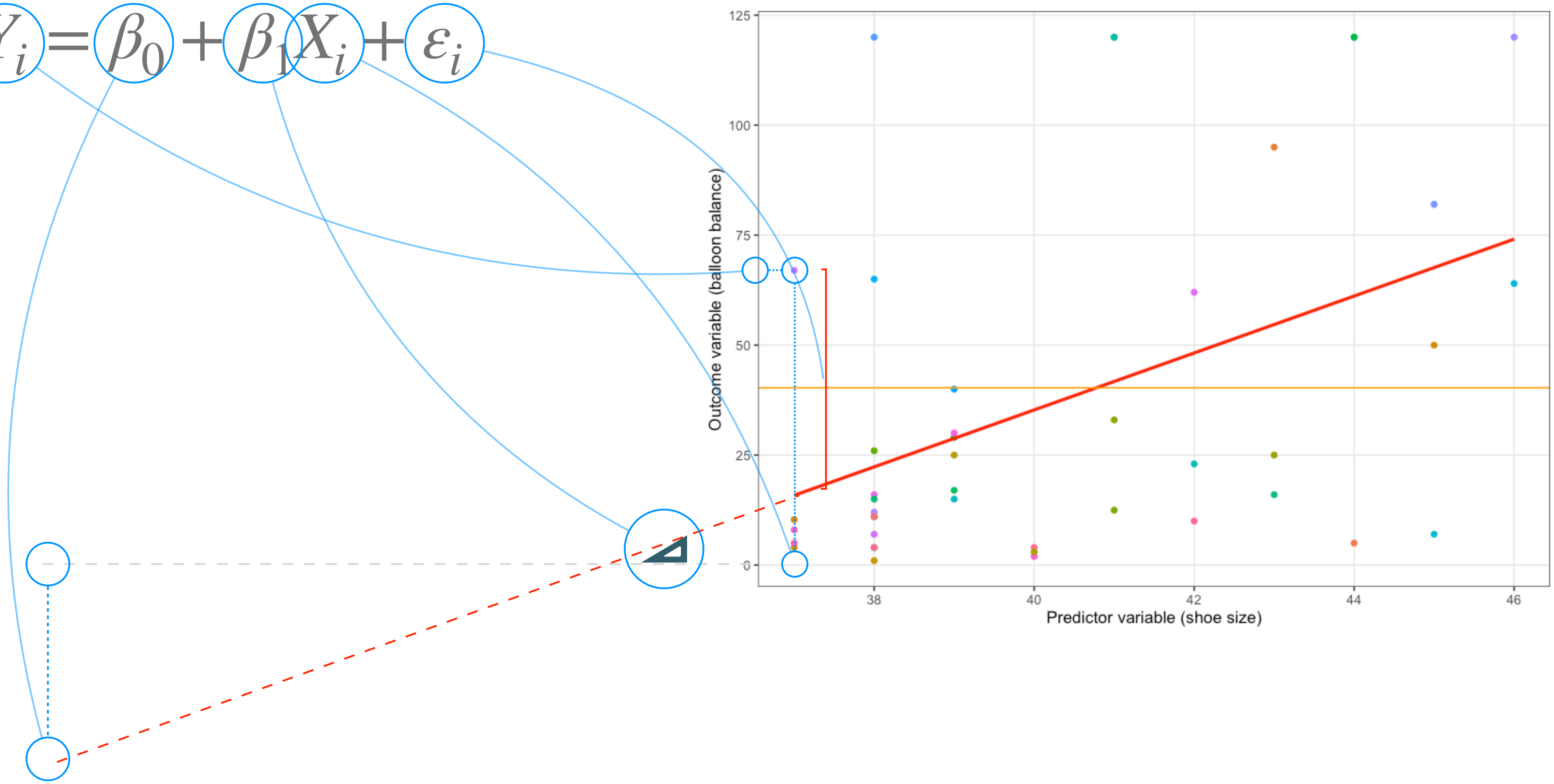
Recap: Interpreting the output of logistic regression (2)

```
> summary(glm(gender_N ~ shoesize, family = binomial(link = logit),
data))

Call:
glm(formula = gender_N ~ shoesize, family = binomial(link = logit),
    data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.66609  -0.05431  -0.01526   0.01807   1.81759

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -105.570     49.470  -2.134   0.0328 *
shoesize       2.540      1.192   2.131   0.0331 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 55.0432  on 43  degrees of freedom
Residual deviance:  8.8056  on 42  degrees of freedom
AIC: 12.806

Number of Fisher Scoring iterations: 8


> boot::inv.logit(-105.57+(36*2.54))
[1] 0

> boot::inv.logit(-105.57+(41*2.54))
[1] 0.1930987

> boot::inv.logit(-105.57+(43*2.54))
[1] 0.9746673

> boot::inv.logit(-105.57+(46*2.54))
[1] 0.9999873
```



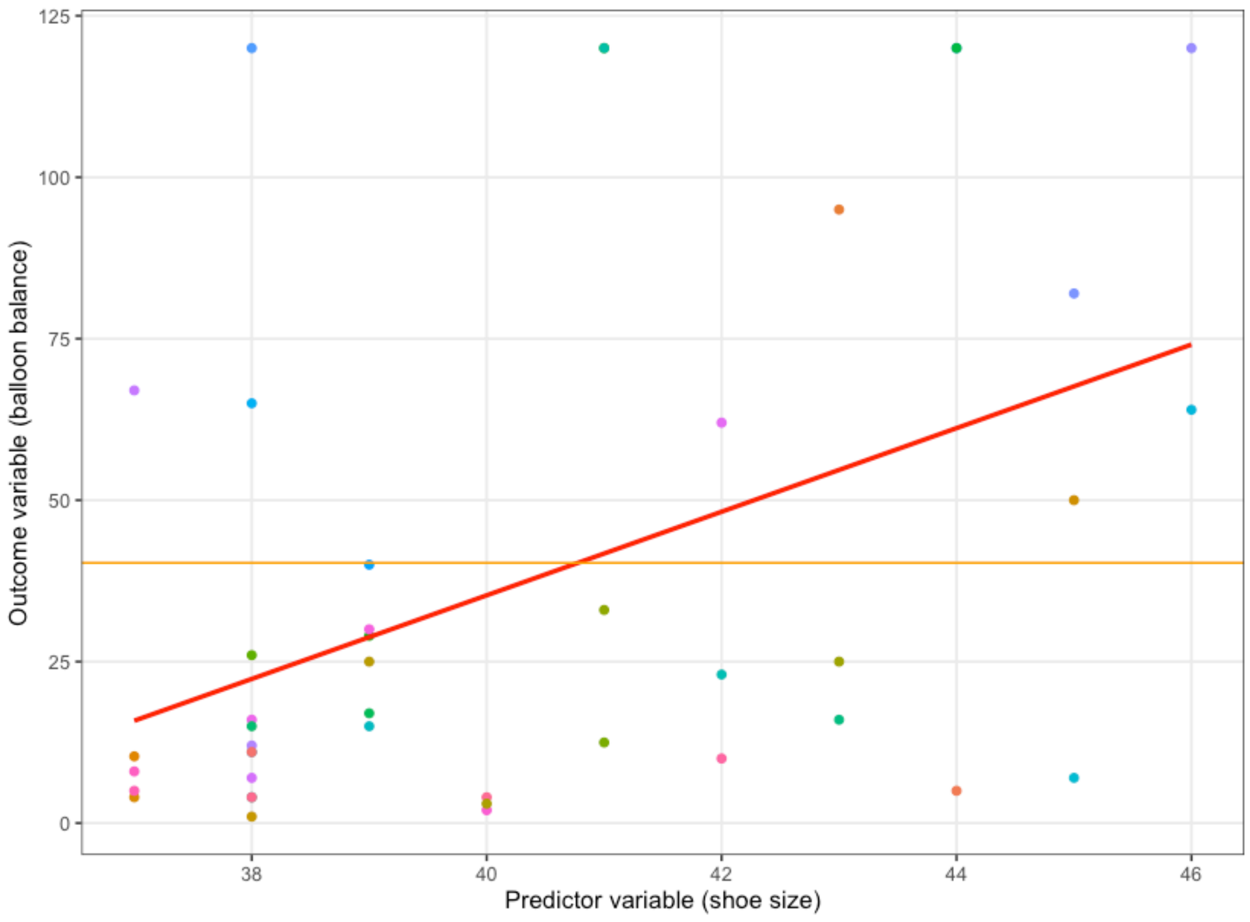
Recap: Interpreting the linear regression formula (1)

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$



Recap: Interpreting the linear regression formula (2)

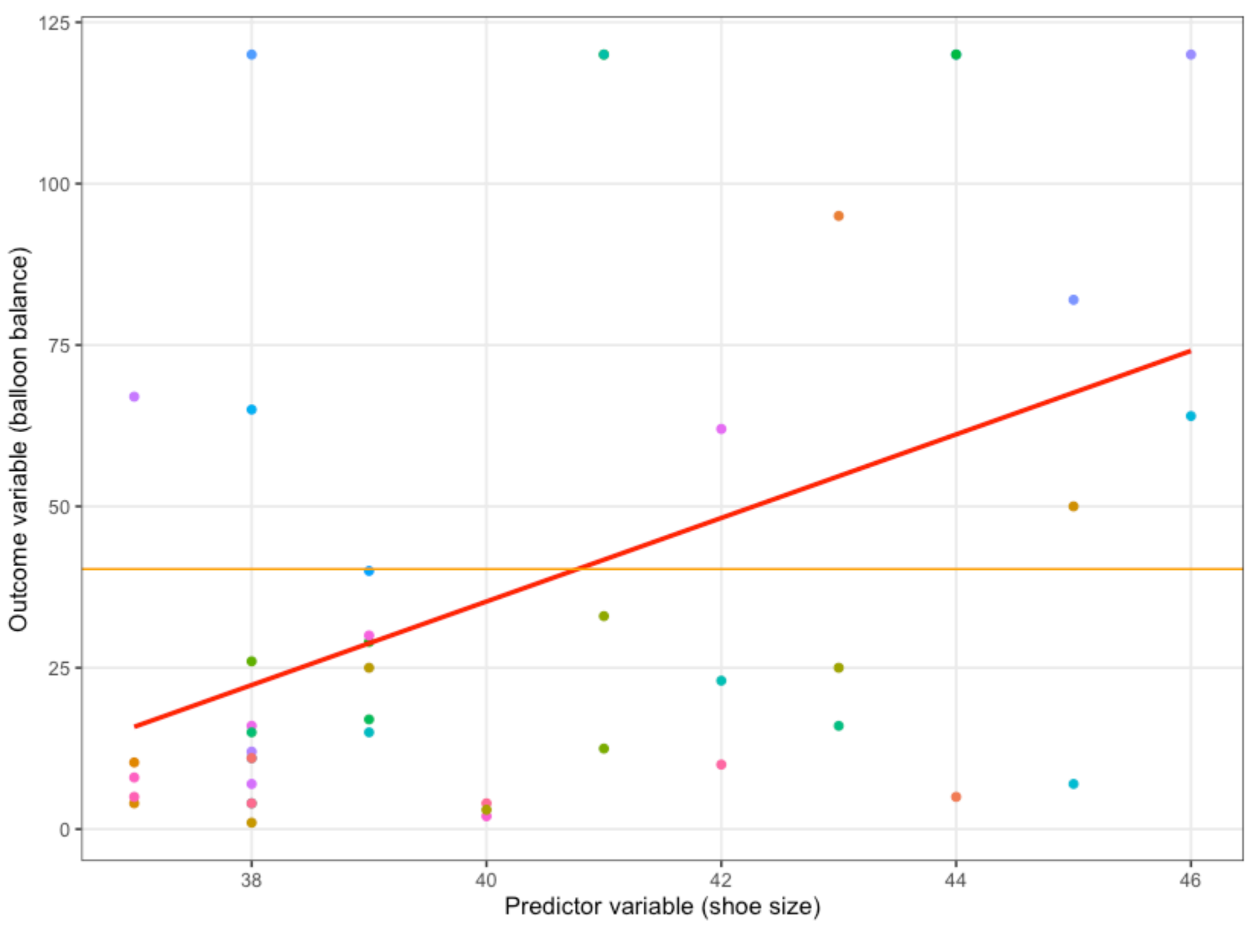
- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i =$
- $= -223.612 + 6.472X_i + \varepsilon_i$



Recap: Interpreting the linear regression formula (3)

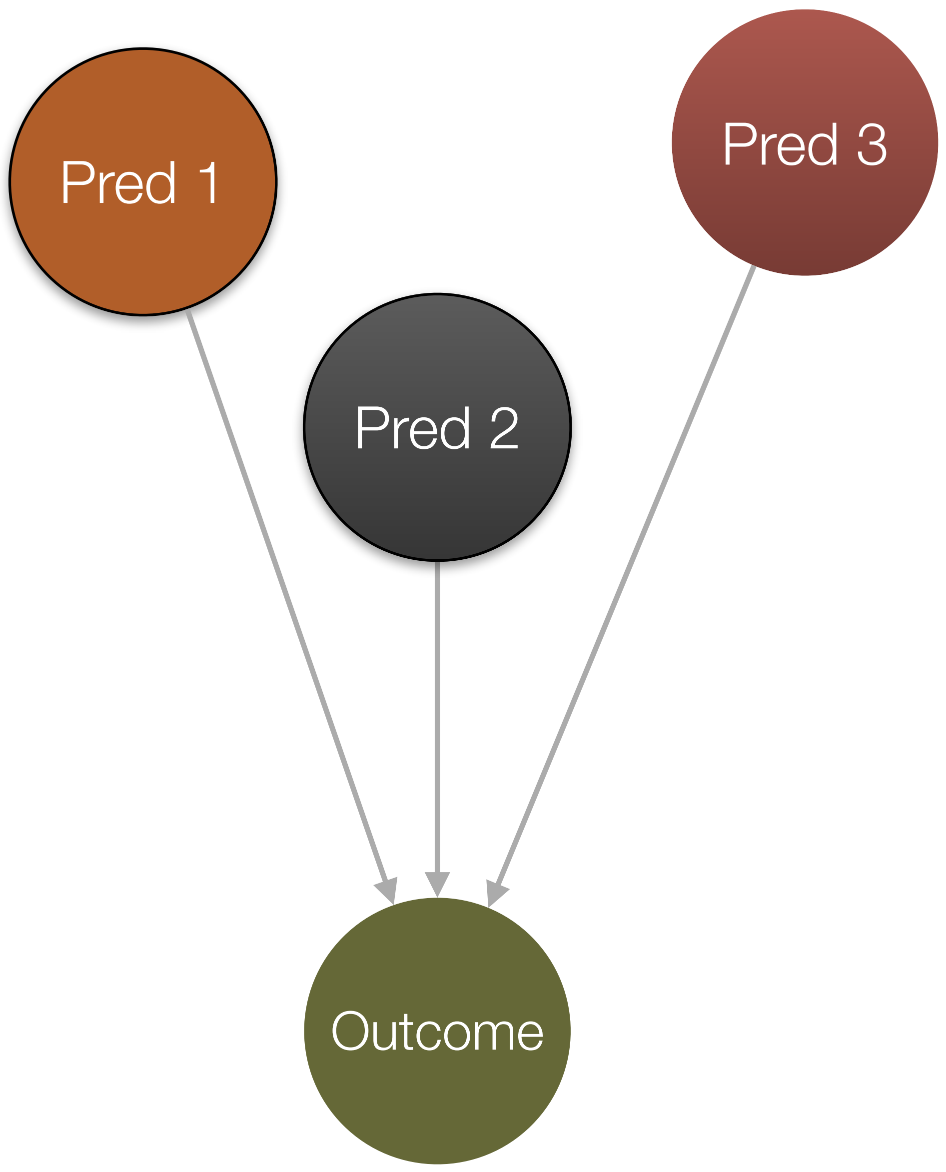
- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i =$
- $= -223.612 + 6.472X_i + \varepsilon_i$

- $X_{Fabio} = 44$
- $Y_{Fabio} = -223.612 + 6.472 \times 44 + \varepsilon_{Fabio}$
- $Y_{Fabio} = 61.146$



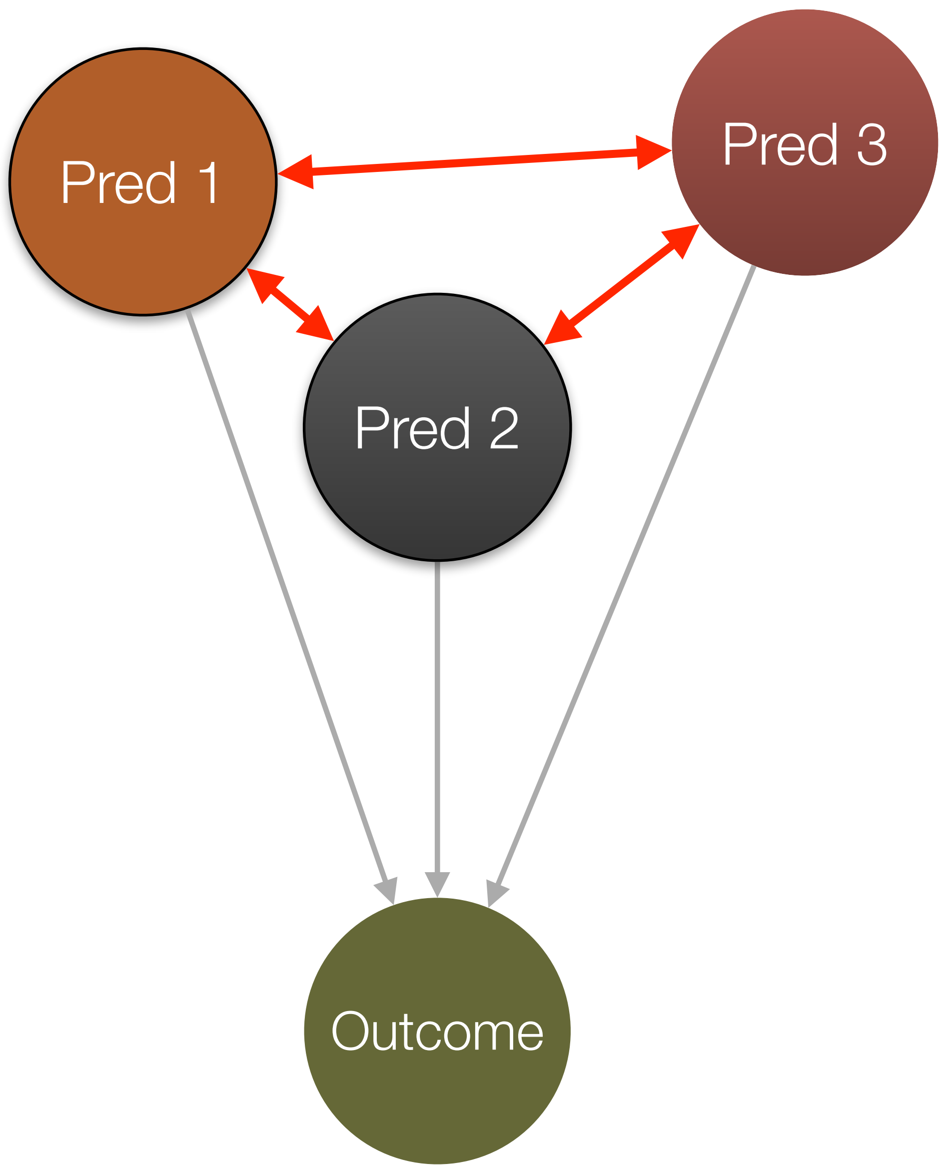
Interaction analysis

- Until now we have only looked at the effect of one predictor at the time
- Interaction analysis allows us to see how 2+ predictor variables influence the outcome variable **together**



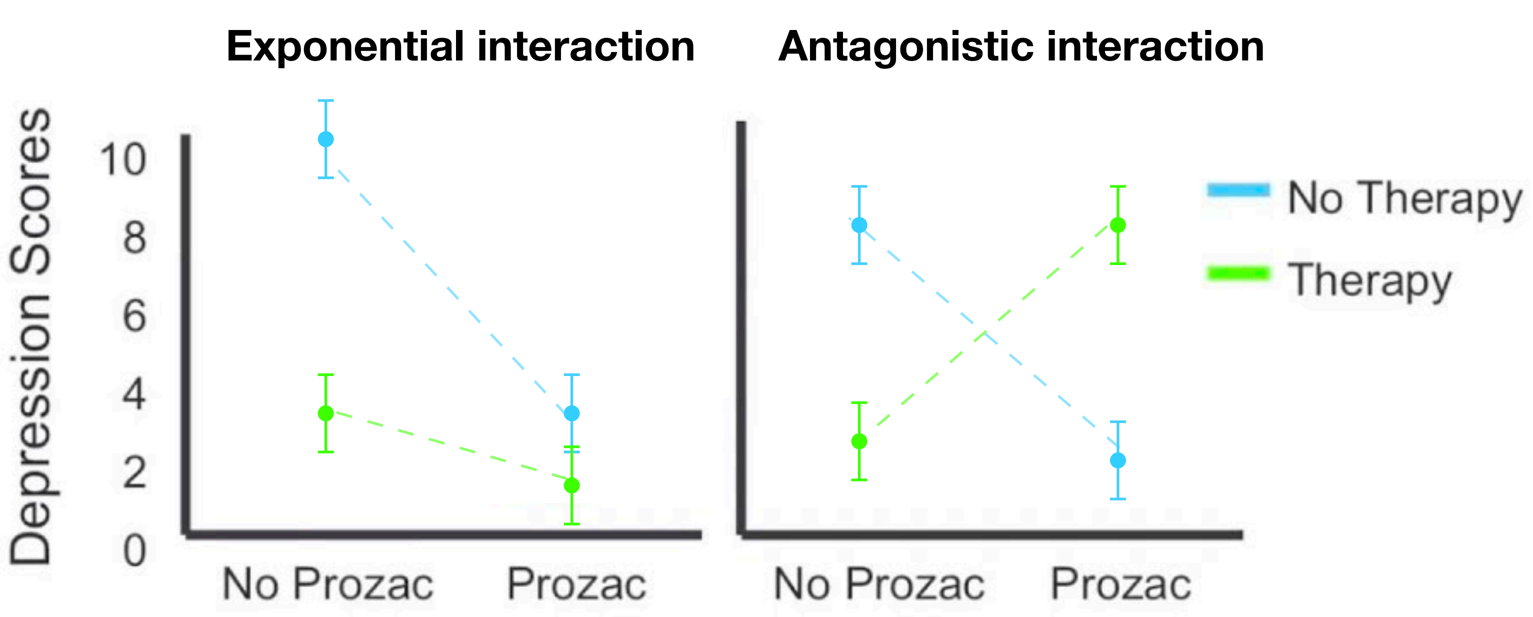
Interaction analysis

- Until now we have only looked at the effect of one predictor at the time
- Interaction analysis allows us to see how 2+ predictor variables influence the outcome variable **together**



Main effects vs Interaction effects

- Main effect: the effect of a predictor on an outcome variable
- Interaction effect: the effect of one predictor on an outcome variable as a function of another predictor variable



Interaction effects in linear regression

- $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3(x_{1i} \times x_{2i}) + \varepsilon_i$

Interaction effects in linear regression

- $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 (x_{1i} \times x_{2i}) + \varepsilon_i$

  intercept

Interaction effects in linear regression

- $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 (x_{1i} \times x_{2i}) + \varepsilon_i$

intercept

slope for the first predicto
r

Interaction effects in linear regression

- $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3(x_{1i} \times x_{2i}) + \varepsilon_i$

intercept

slope for the first predicto r

slope for the second predictor

Interaction effects in linear regression

- $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 (x_{1i} \times x_{2i}) + \varepsilon_i$

intercept

slope for the first predicto r

slope for the second predictor

slope for the multiplicative effect between the two predictors

Three types of two-way interactions

- Two continuous predictors
- Two categorical predictors
- One continuous and one categorical predictor
- High-order interactions: three-way interactions and above can be very hard to interpret and are generally avoided

Two categorical predictors (1)

- **Does the effect of one categorical predictor depend on the levels of another categorical predictor?**
- Eg: sex (male vs female), condition (1 vs 2), continuous response
- 2*2 categorical design
- Since the lines cross, we know that this is an antagonistic interaction

Main effect of sex

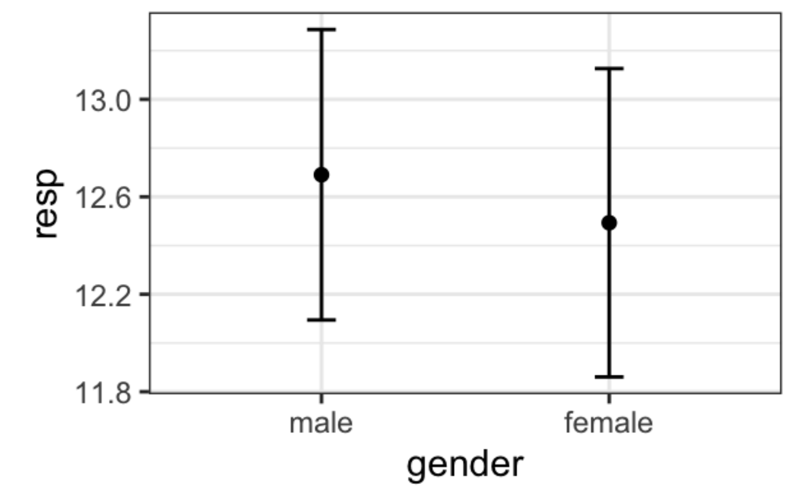
	sex	resp
1	male	12.69
2	female	12.49

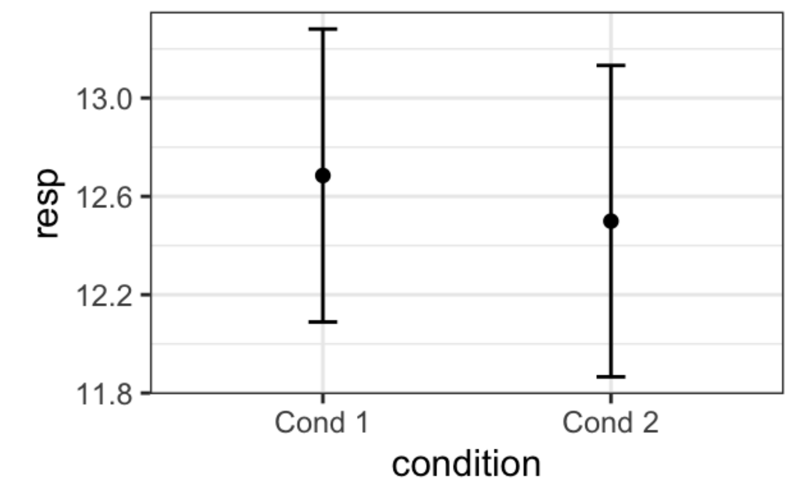
Main effect of condition

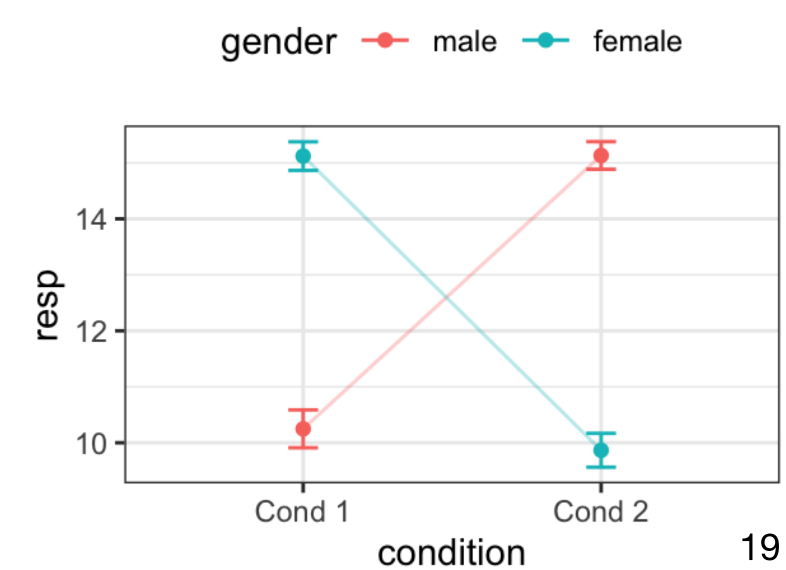
	condition	resp
1	Cond 1	12.68
2	Cond 2	12.49

Sex * Condition

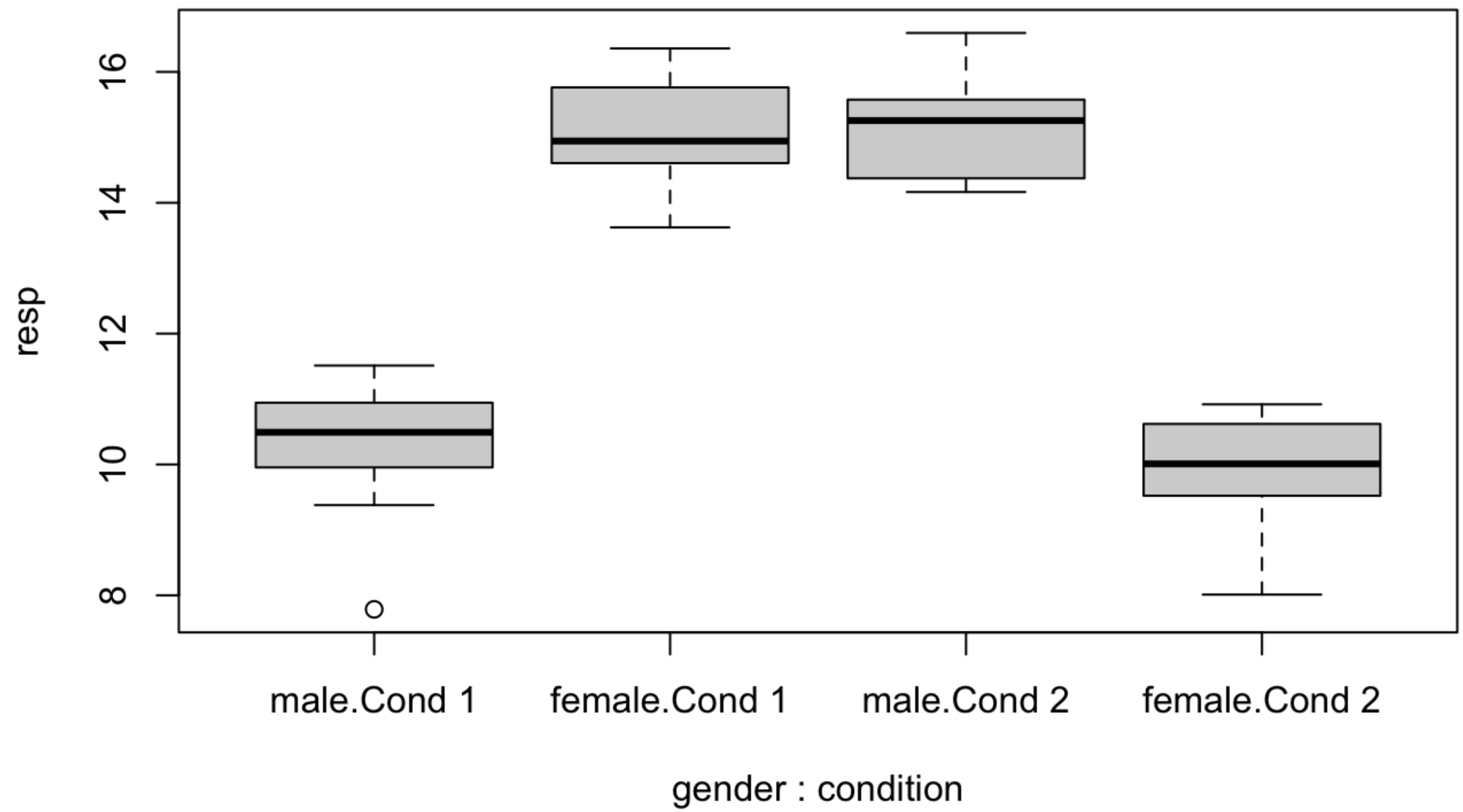
	Cond 1	Cond 2
male	10.24	15.13
female	15.12	9.86







Two categorical predictors (2)



Testing the interaction in R

- `summary(lm(resp ~ condition * gender, data = data))`

```
Call:
lm(formula = resp ~ condition * gender, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.4635 -0.4570  0.1093  0.6152  1.4631 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                   10.2488     0.2881   35.57  < 2e-16 ***
conditionCond 2                4.8834     0.4074   11.98 3.98e-14 ***
genderfemale                   4.8719     0.4074   11.96 4.26e-14 ***
conditionCond 2:genderfemale -10.1378     0.5762  -17.59  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.9111 on 36 degrees of freedom
Multiple R-squared:  0.8961,	Adjusted R-squared:  0.8874 
F-statistic: 103.5 on 3 and 36 DF,  p-value: < 2.2e-16
```

Interpreting the output

- β_0 **(Intercept)**: the value of Y when condition and gender are in their reference (baseline) level
- β_1 **condition**: the change in Y when condition goes from “Cond 1” to “Cond 2” in the reference level of gender (“male”)
- β_2 **gender**: the change in Y when gender goes from “male” to “female” in the reference level of condition (“Cond 1”)
- β_4 **condition*gender**: the additional effect of “Cond 2” for females
- $\beta_1 + \beta_4$: the simple effect of “Cond 2” for females

Simple effects ("marginal effects") in a categorical by categorical interaction (1)

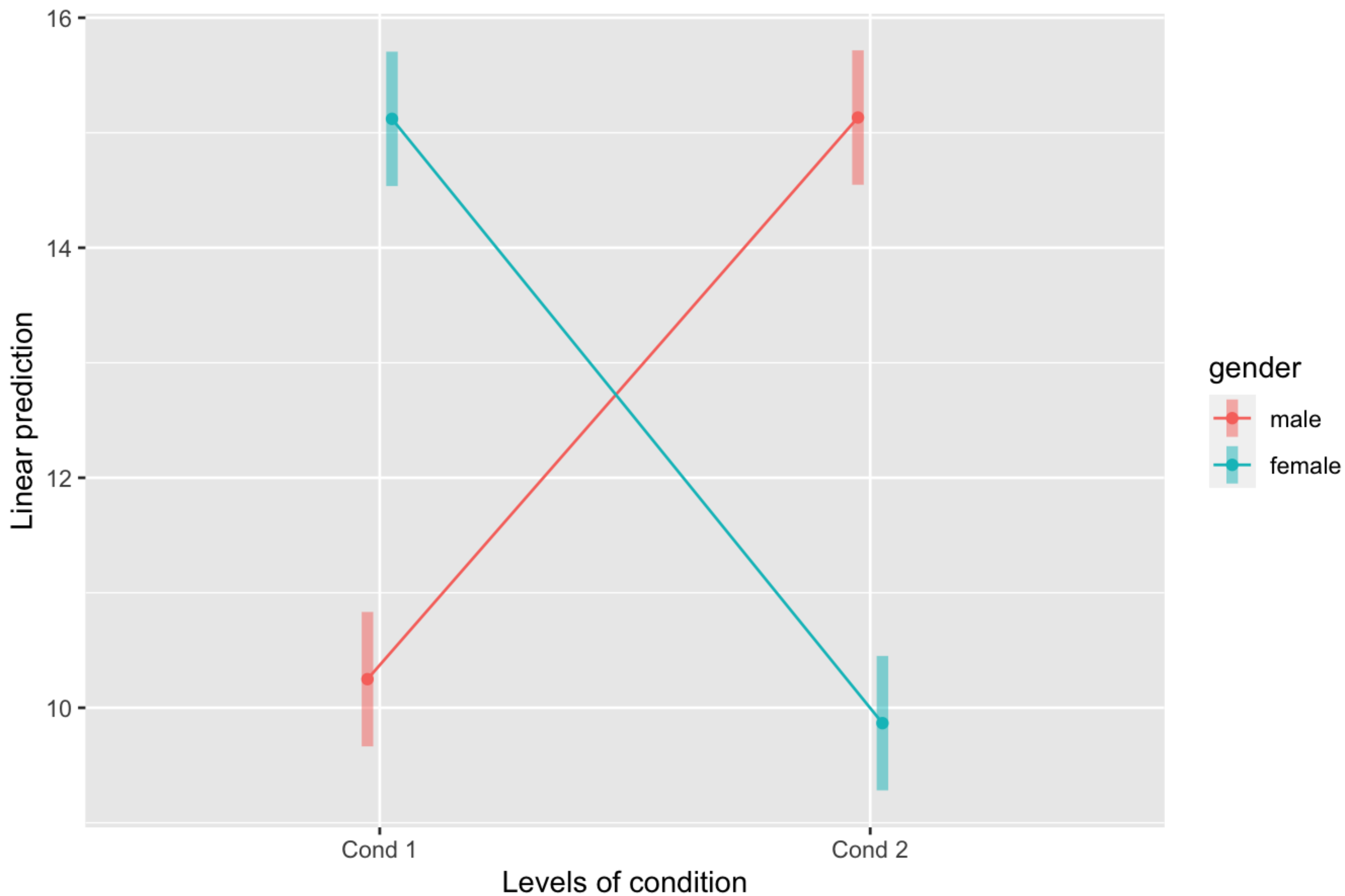
- `library(emmeans)`
- `model <- lm(resp ~ condition * gender, data = data)`
- `emmeans(model, ~ condition * gender)`

```
 gender condition emmean    SE df lower.CL upper.CL
 male   Cond 1     10.25 0.288 36     9.66     10.8
 female Cond 1     15.12 0.288 36    14.54     15.7
 male   Cond 2     15.13 0.288 36    14.55     15.7
 female Cond 2      9.87 0.288 36     9.28     10.5

Confidence level used: 0.95
```

- **Note**: These are not descriptive statistics for the raw data, but rather the model predictions (i.e., the data points that fall on the regression lines estimated by the model)

Simple effects ("marginal effects") in a categorical by categorical interaction (2)

- `library(emmeans)`
- `emmip(model, gender ~ condition, CIs = TRUE)`

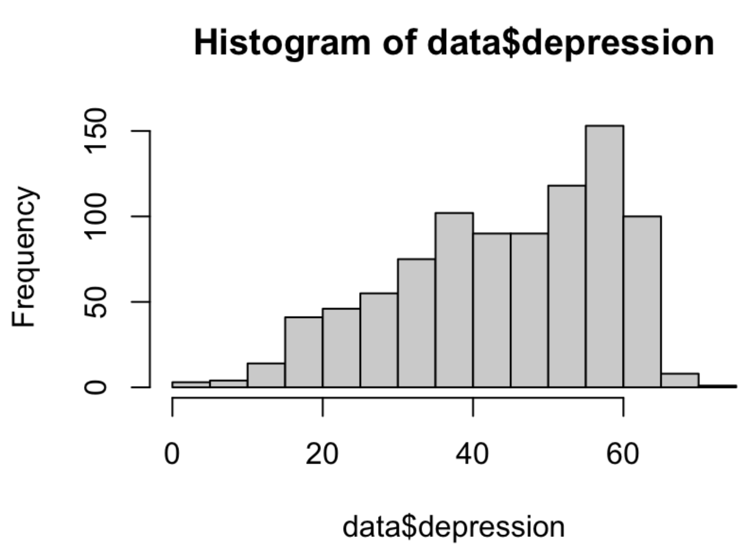


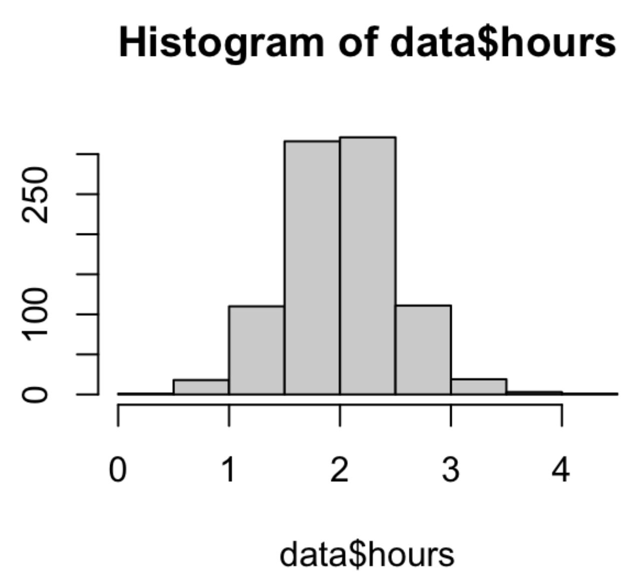
Two continuous predictors (2)

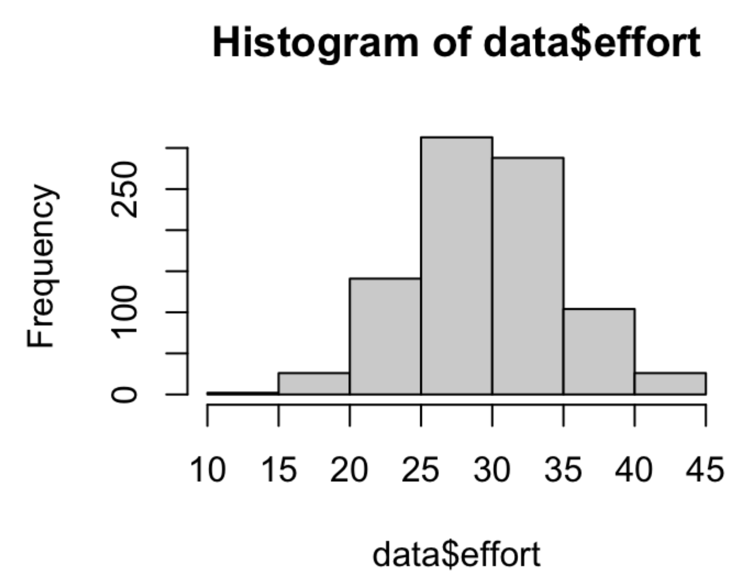
- **Does the effect of one continuous predictor depend on the levels of another continuous predictor?**

- Eg: amount of HIIT (hours), effort, depression scale

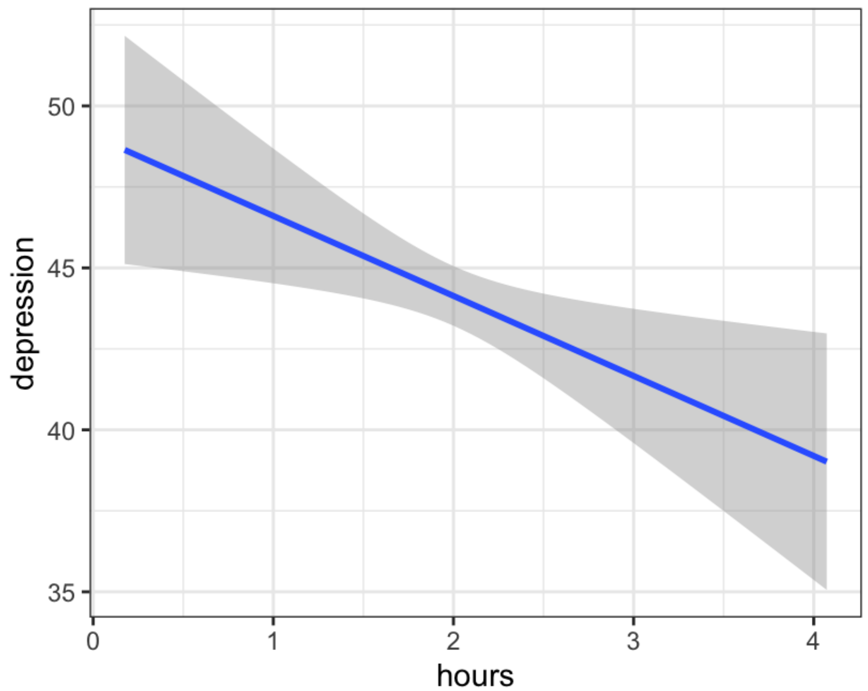
```
> summary(data$depression)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   34.10   46.27   44.13   55.89   71.29

> summary(data$hours)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1751  1.6764  2.0051  2.0024  2.3375  4.0722

> summary(data$effort)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.95   26.26   29.63   29.66   33.10   44.08
```

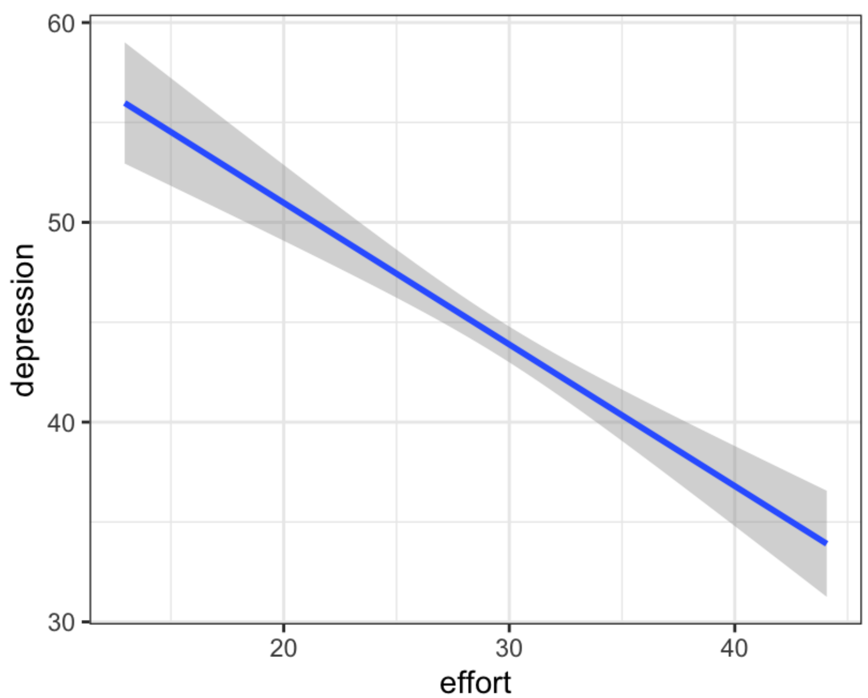


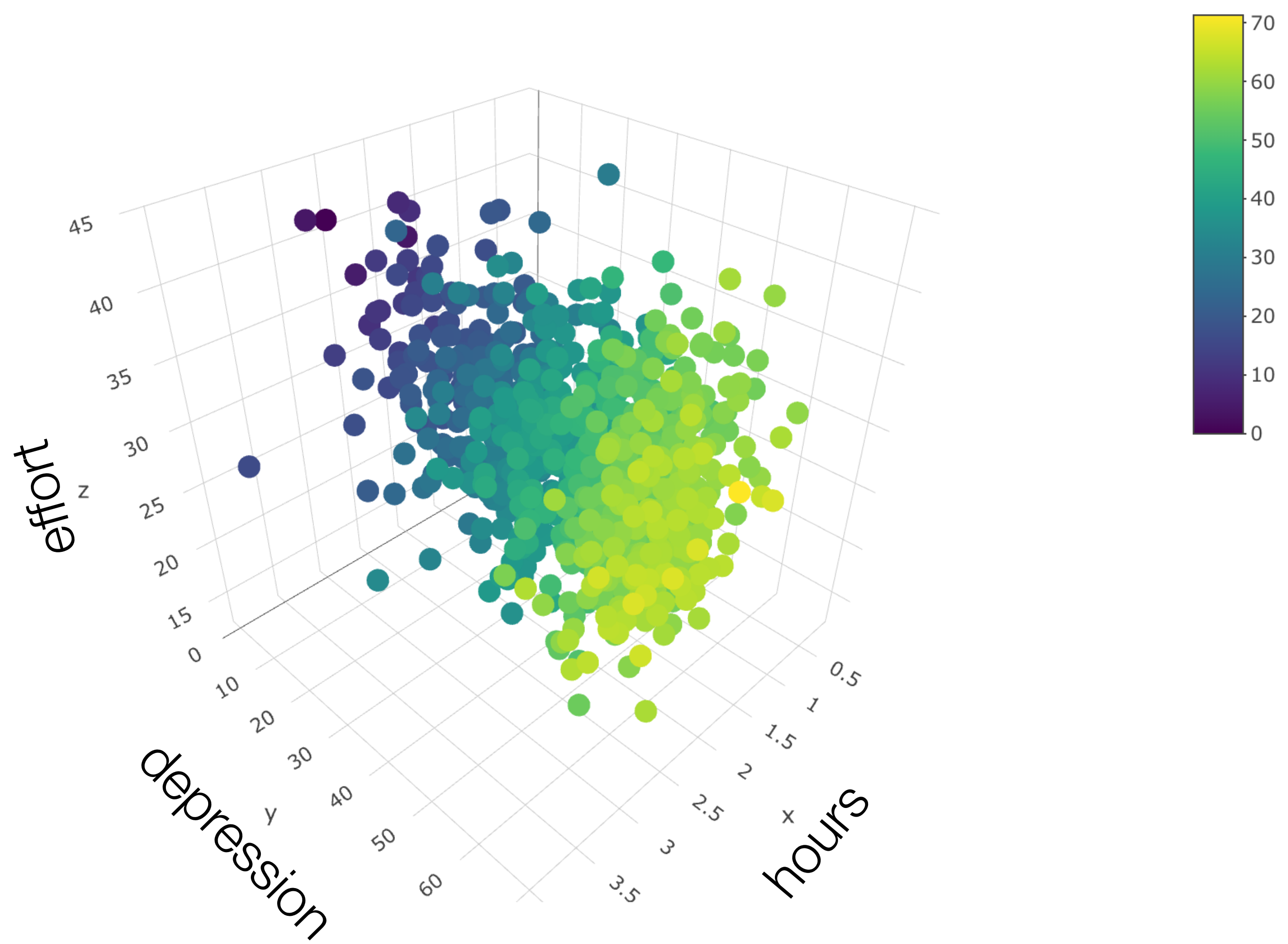




Two continuous predictors (3)







Testing the interaction in R

- `summary(lm(depression ~ hours * effort, data))`

```
Call:
lm(formula = depression ~ hours * effort, data = data)

Residuals:
   Min     1Q Median     3Q    Max 
-34.51 -11.13   1.78  10.60  29.52 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  46.35182   11.60362   3.995 7.01e-05 ***
hours         9.37568    5.66392   1.655   0.0982 .  
effort        0.08028    0.38465   0.209   0.8347    
hours:effort -0.39335    0.18750  -2.098   0.0362 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.56 on 896 degrees of freedom
Multiple R-squared:  0.07818,	Adjusted R-squared:  0.07509 
F-statistic: 25.33 on 3 and 896 DF,  p-value: 9.826e-16
```

Interpreting the output

- β_0 **(Intercept)**: The predicted value of Y when hours = 0 and effort = 0
- β_1 **(hours)**: The change in Y when hours increases by 1 unit and effort = 0
- β_2 **(effort)**: The change in Y when effort increases by 1 unit and hours = 0
- β_3 **(hours*effort)**: The change in the slope of hours for every one unit increase in effort (or vice versa)

Simple effects (“marginal effects”) in a continuous by continuous interaction (1)

- To estimate marginal effects, we need to look at the slope of **hours** for specific levels of **effort**

- We code **high effort** as effort that is 2SD > median, **medium effort** as the median of effort, and **low effort** as effort that is 2SD < median

```
high_effort <- round(median(data$effort) + 2*sd(data$effort), digits = 2)
medium_effort <- round(median(data$effort), digits = 2)
low_effort <- round(median(data$effort) - 2*sd(data$effort), digits = 2)

mylist <- list(effort=c(low_effort,medium_effort,high_effort))

emtrends(model, ~effort, var=“hours",at=mylist)
```

```
 effort hours.trend    SE  df lower.CL upper.CL
   19.4        1.76 2.164 896    -2.48    6.012
   29.6       -2.28 0.915 896    -4.08   -0.483
   39.9       -6.33 2.105 896   -10.46   -2.196

Results are averaged over the levels of: hours
Confidence level used: 0.95
```

Simple effects (“marginal effects”) in a continuous by continuous interaction (2)

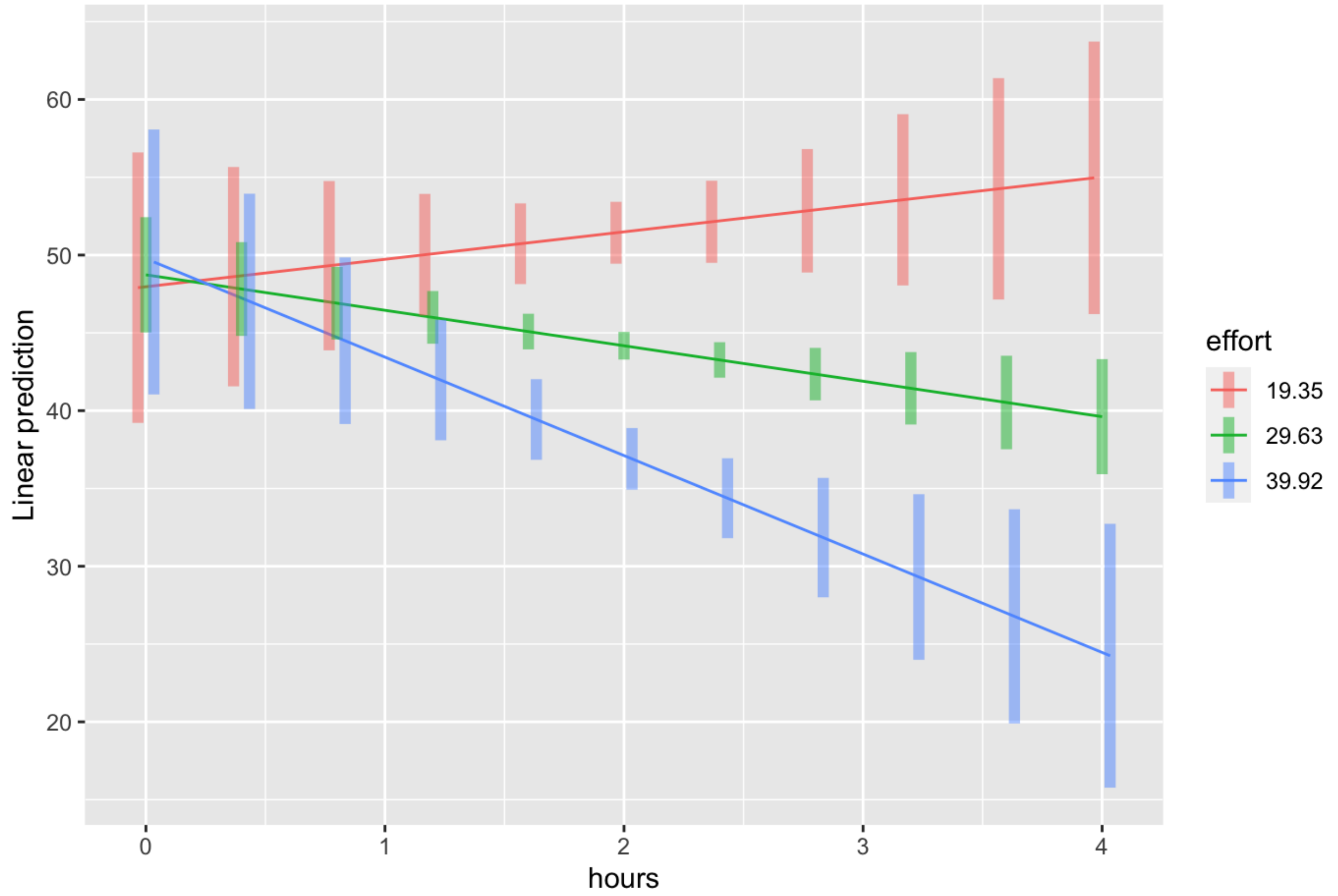
- For plotting, we want **hours** to be on the x-axis and **effort** to be separate the lines

- For **hours** we create a sequence of values to span a reasonable range of hours (we use a sequence of 0.4 increments)

- **effort** is coded as earlier

Simple effects (“marginal effects”) in a continuous by continuous interaction (3)

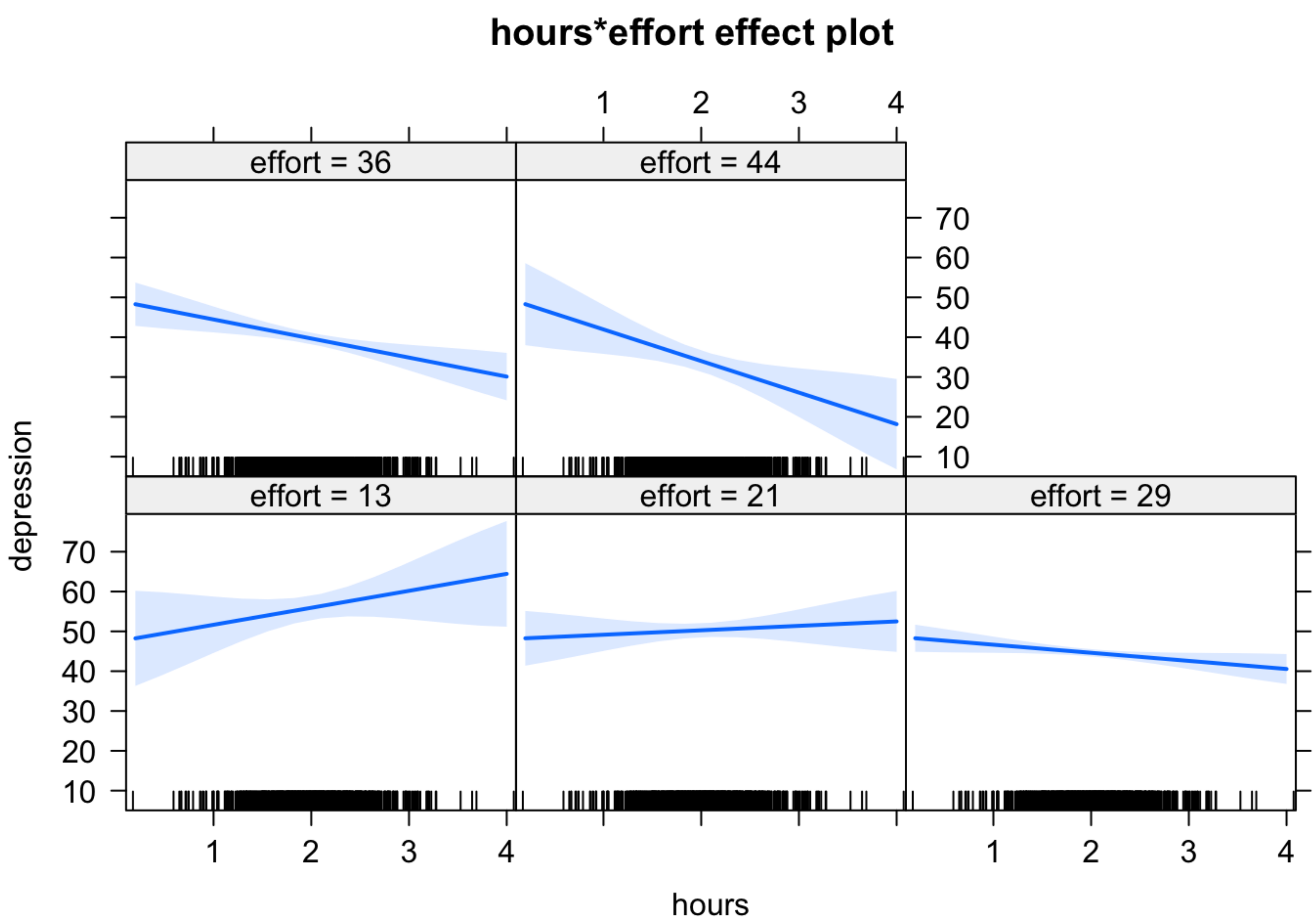
```
mylist <- list(hours=seq(0,4,by=0.4),effort=c(low_effort,medium_effort,high_effort))
emmip(model,effort~hours,at=mylist, CIs=TRUE)
```



Note: the lines cross each other, so this is also an antagonistic interaction

An alternative method (easier but not as nice)

```
plot(allEffects(model))
```



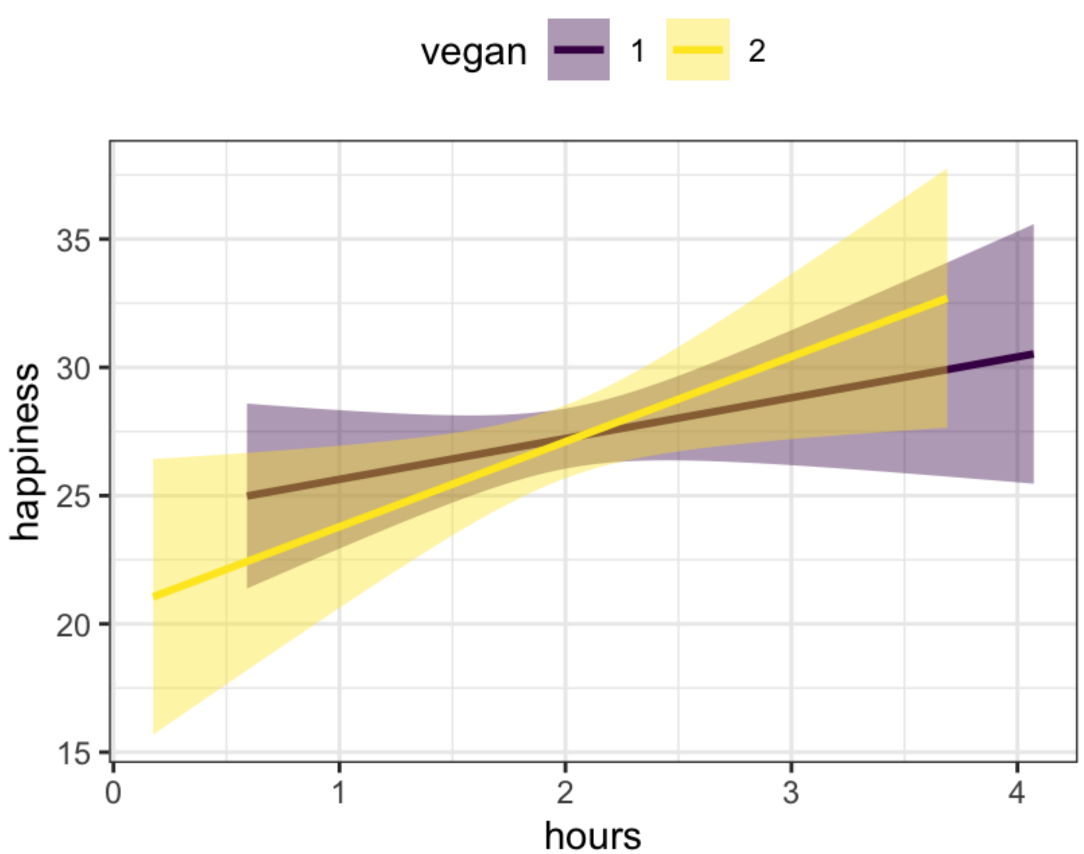
Is there a significant difference between the slopes?

```
> emtrends(model, pairwise ~effort, var="hours",at=mylist, adjust="none")

$emtrends
 effort hours.trend    SE  df lower.CL upper.CL
   19.4        1.76 2.164 896    -2.48    6.012
   29.6       -2.28 0.915 896    -4.08   -0.483
   39.9       -6.33 2.105 896   -10.46   -2.196

Results are averaged over the levels of: hours
Confidence level used: 0.95

$contrasts
 contrast      estimate   SE  df t.ratio p.value
 19.35 - 29.63     4.04 1.93 896   2.098  0.0362
 19.35 - 39.92     8.09 3.86 896   2.098  0.0362
 29.63 - 39.92     4.05 1.93 896   2.098  0.0362

Results are averaged over the levels of: hours
```

One continuous and one categorical predictor

- **Does the effect of a continuous predictor depend on the levels of a categorical predictor?**
- Eg: amount of HIIT (hours), vegan diet, depression scale
- 1 = vegan
- 0 = non-vegan

```
> table(data$vegan)

  1   2 
450 450 

> summary(data$happiness)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00   15.39   25.02   27.16   37.19   71.29 

> summary(data$hours)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.1751  1.6764  2.0051  2.0024  2.3375  4.0722 
```



Testing the interaction in R

```
> summary(lm(happiness ~ hours * vegan, data = data))

Call:
lm(formula = happiness ~ hours * vegan, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-27.118 -11.350  -1.963  10.001  42.376 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    24.044      2.805   8.571   <2e-16 ***
hours           1.591      1.352   1.177    0.240    
vegan2         -3.571      3.915  -0.912    0.362    
hours:vegan2    1.724      1.898   0.908    0.364    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.06 on 896 degrees of freedom
Multiple R-squared:  0.008433,	Adjusted R-squared:  0.005113 
F-statistic:  2.54 on 3 and 896 DF,  p-value: 0.05523
```

Interpreting the output

- β_0 **(Intercept)**: The predicted value of Y when hours = 0 and vegan = 0 (reference level)
- β_1 **(hours)**: The change in Y when hours increases by 1 unit and vegan = 0
- β_2 **(vegan)**: The change in Y when vegan goes from 0 to 1 and hours = 0
- β_3 **(hours*vegan)**: The change in the slope of hours when vegan goes from 0 to 1

Simple effects ("marginal effects") in a continuous by categorical interaction (1)

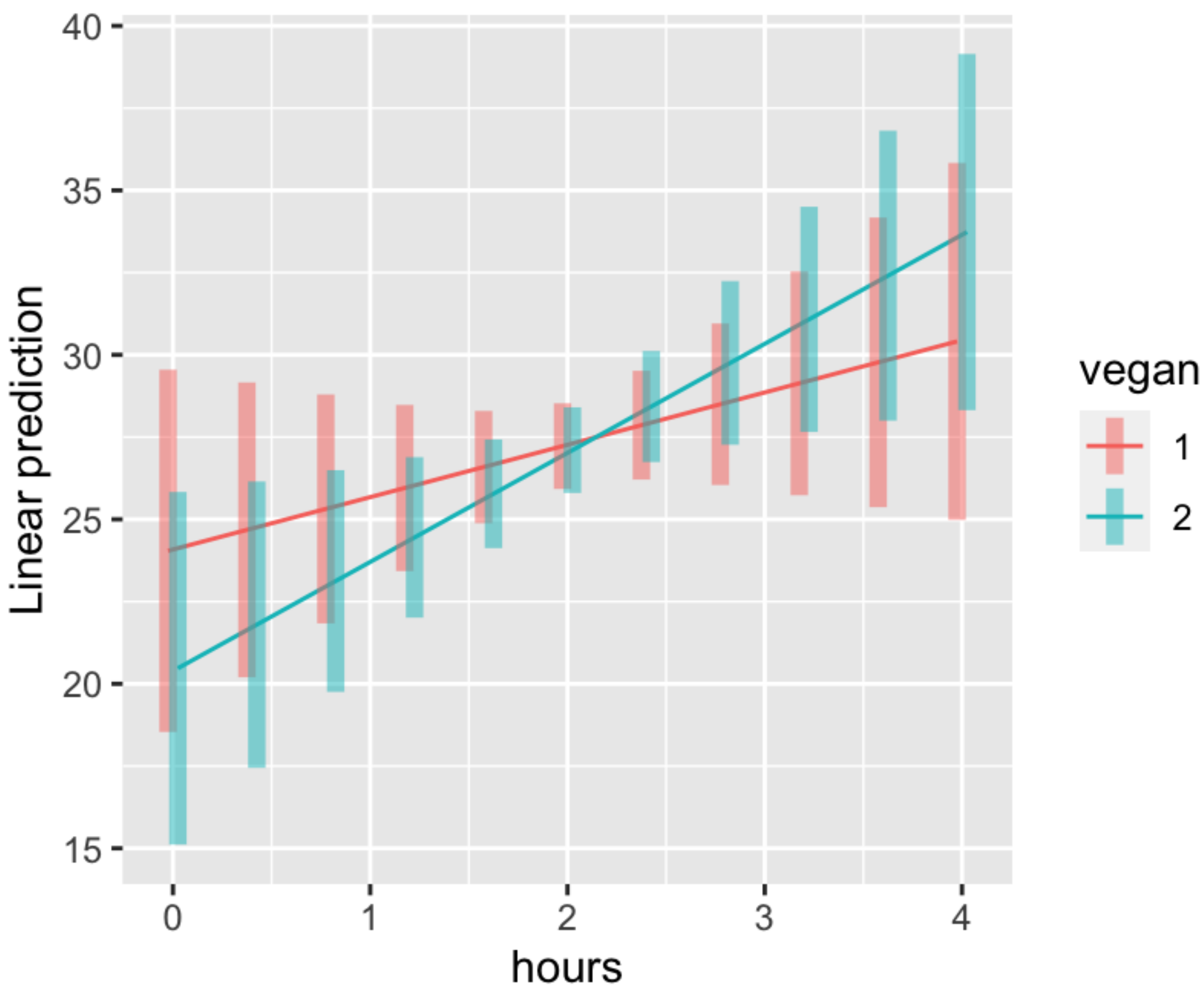
```
> emtrends(model, ~ vegan, var="hours")

 vegan hours.trend   SE  df lower.CL upper.CL
 1            1.59 1.35 896   -1.063     4.25
 2            3.32 1.33 896    0.702     5.93

Confidence level used: 0.95
```

```
> emtrends(model, pairwise ~ vegan, var="hours")

$emtrends
 vegan hours.trend   SE  df lower.CL upper.CL
 1            1.59 1.35 896   -1.063     4.25
 2            3.32 1.33 896    0.702     5.93

Confidence level used: 0.95

$contrasts
 contrast estimate  SE  df t.ratio p.value
 1 - 2       -1.72 1.9 896  -0.908  0.3639
```

Simple effects ("marginal effects") in a continuous by continuous interaction (2)

- To plot the data, we use the same "trick" as before by picking some specific values of **hours**

```
mylist <- list(hours = seq(0,4, by = 0.4))
emmip(model, vegan ~ hours, at = mylist, CIs = TRUE)
```



Take-home message

- Regression is a very powerful tool
- It allows us to measure the effect of predictor variables on an outcome variable
- Sometimes, multiple predictors may interact with each other
- This means that the effect of one predictor on the outcome variable is different for different levels of another predictor
- Interaction analysis helps us unearth these interactive effects
- We rarely go beyond a three-way interaction because of interpretability issues

A word on the textbook

- The book is your friend… use it!
- Methods 2 will be math-heavy
- This book does a decent job at giving you the basic mathematical notions that will prepare you for what comes next
- Remember that my lectures complement the book, they <u>do not</u> replace it
- You need both lectures and the book to understand the topics fully