Week 11:
Beyond Multiple Linear Regression

Phase III

Recap so far

- Linear Regression: Relationship between a continuous dependent variable and a numerical or categorical independent variable.
- Multiple Linear Regression: Relationship between a dependent variable and multiple numerical and/or categorical independent variables.

What we have not covered

- Relationships between dependent and independent variables are not constant; instead, they depend on other factors (a third variable).
- Different dependent variable types (e.g. categorical, count data, etc).

Key ideas

- We can use **interactions** between independent variables when we suspect a relationship between a dependent and independent variable depends on another variable.
- We can use a **logistic regression** when the dependent variable is categorical and binary.
- Logistic regression is one of a family of models called **Generalized Linear Models**, which are often applied when the assumptions of linear regression fail.

Interactions

Extending the regression model

- We saw that we could use regression to model the linear relationship between a dependent variable, Y, and an independent variables X:

$$Y_i = \alpha + \beta \cdot X_i + \epsilon_i$$

- The dependence of *Y* on *X* is encoded on *β.*
- But what if this *β* depends on another variable?
- We can model this effect by using an **interaction term.**

The beer-goggles effect

➢ The following dataset was collected to study the effect of consumed alcohol on our subjective perception of physical attractiveness.

Alcohol	None		2 Pints		4 Pints	
Gender	Female	Male	Female	Male	Female	Male
	65	50	70	45	55	30
	70	55	65	60	65	30
	60	80	60	85	70	30
	60	65	70	65	55	55
	60	70	65	70	55	35
	55	75	60	70	60	20
	60	75	60	80	50	45
	55	65	50	60	50	40
Total	485	535	500	535	460	285
Mean	60.625	66.875	62.50	66.875	57.50	35.625
Variance	24.55	106.70	42.86	156.70	50.00	117.41

Field, A. P. (2023). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

The beer-goggles effect

➢ We fit a linear regression to study this effect (alcohol intake→attractiveness).

```
Call:
lm(formula = attractiveness ~ alcohol, data = dat.attractiveness %>%
    filter(alcohol != "2 Pints"))

Residuals:
    Min       1Q  Median      3Q     Max
-26.562  -8.750   1.250   8.438  23.438

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      63.750      2.944  21.652  < 2e-16 ***
alcohol4 Pints  -17.187      4.164  -4.128 0.000268 ***
```

The beer-goggles effect

- We fit a linear regression to study this effect (alcohol intake→attractiveness).
- Large alcohol intakes (> 2 pints) makes us significantly choose less attractiveness mates!

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      63.750      2.944  21.652  < 2e-16 ***
alcohol4 Pints  -17.187      4.164  -4.128 0.000268 ***
```

The beer-goggles effect

- We fit a linear regression to study this effect (alcohol intake→attractiveness).
- Large alcohol intakes (> 2 pints) makes us significantly choose less attractiveness mates!
- Is this effect different between men and women?

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      63.750      2.944  21.652  < 2e-16 ***
alcohol4 Pints  -17.187      4.164  -4.128 0.000268 ***
```

The beer-goggles effect: gender differences

- To answer this question say we estimate the regression model that includes both alcohol and gender variables:

```
Call:
lm(formula = attractiveness ~ alcohol + gender, data = dat.attractiveness %>%
    filter(alcohol != "2 Pints"))

Residuals:
     Min       1Q   Median       3Q      Max 
-22.6562  -7.6562  -0.4687   6.2500  20.1563 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     59.844      3.446  17.368  < 2e-16 ***
alcohol4 Pints -17.187      3.979  -4.320 0.000167 ***
genderFemale     7.812      3.979   1.964 0.059234 .  
```

The beer-goggles effect: gender differences

➢ To answer this question say we estimate the regression model that includes both alcohol and gender variables:

```
Call:
lm(formula = attractiveness ~ alcohol + gender, data = dat.attractiveness %>%
    filter(alcohol != "2 Pints"))

Residuals:
     Min       1Q   Median       3Q      Max 
-22.6562  -7.6562  -0.4687   6.2500  20.1563 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      59.844      3.446  17.368  < 2e-16 ***
alcohol4 Pints  -17.187      3.979  -4.320 0.000167 ***
genderFemale      7.812      3.979   1.964 0.059234 .  
```

The beer-goggles effect: gender differences

- To answer this question say we estimate the regression model that includes both alcohol and gender variables:

```
Call:
lm(formula = attractiveness ~ alcohol + gender, data = dat.attractiveness %>%
    filter(alcohol != "2 Pints"))

Residuals:
     Min       1Q   Median       3Q      Max 
-22.6562  -7.6562  -0.4687   6.2500  20.1563 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      59.844      3.446  17.368  < 2e-16 ***
alcohol4 Pints  -17.187      3.979  -4.320 0.000167 ***
genderFemale      7.812      3.979   1.964 0.059234 .  
```

- However, this does not answer our question (Why?).

Extending the regression model: Interaction effects

- In order to address this question, we need to include an **interaction term between** both **independent variables**.

- This can be easily achieved by just adding to model a term that goes like $(X_1 \cdot X_2)$. That is:

$$Y_i = \alpha + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \beta_3 \cdot (X_{i1} \cdot X_{i2}) + \epsilon_i$$

Extending the regression model: Interaction effects

- In order to address this question, we need to include an **interaction** term **between** both **independent variables**.

- This can be easily achieved by just adding to model a term that goes like $(X_1 \cdot X_2)$. That is:

Main effect

$$Y_i = \alpha + \boxed{\beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2}} + \beta_3 \cdot (X_{i1} \cdot X_{i2}) + \epsilon_i$$

Extending the regression model: Interaction effects

- In order to address this question, we need to include an **interaction** term **between** both **independent variables**.

- This can be easily achieved by just adding to model a term that goes like $(X_1 \cdot X_2)$. That is:

Interaction effect

$$Y_i = \alpha + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \boxed{\beta_3 \cdot (X_{i1} \cdot X_{i2})} + \epsilon_i$$

Extending the regression model: Interaction effects

- In order to address this question, we need to include an **interaction** term **between** both **independent variables**.

- This can be easily achieved by just adding to model a term that goes like $(X_1 \cdot X_2)$. That is:

$$Y_i = \alpha + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \beta_3 \cdot (X_{i1} \cdot X_{i2}) + \epsilon_i$$

- In R:

```
lm(DV~IV1 + IV2 + IV2:IV1, data)

lm(DV~IV1*IV2, data)
```

The beer-goggles effect: gender differences

Regression model: Attractiveness ~ $\alpha + \beta_1 \cdot \text{Alcohol} + \beta_1 \cdot \text{gender} + \beta_3 \cdot (\text{gender} \cdot \text{Alcohol})$

Interpretation:

α: expected attractiveness for men (gender=0) that had no alcohol.

β_1: change in attractiveness from consuming 4 pints when gender = 0 (men).

β_2: Difference in attractiveness between genders that had no alcohol.

β_3: Difference in attractiveness between genders that had no alcohol, to the difference between genders that had alcohol.

The beer-goggles effect: gender differences

```
Call:
lm(formula = attractiveness ~ alcohol * gender, data = dat.attractiveness %>%
    filter(alcohol != "2 Pints"))

Residuals:
    Min      1Q  Median      3Q     Max
-16.875  -5.625  -0.625   5.156  19.375

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                   66.875      3.055  21.890  < 2e-16 ***
alcohol4 Pints               -31.250      4.320  -7.233 7.13e-08 ***
genderFemale                  -6.250      4.320  -1.447    0.159
alcohol4 Pints:genderFemale   28.125      6.110   4.603 8.20e-05 ***
```

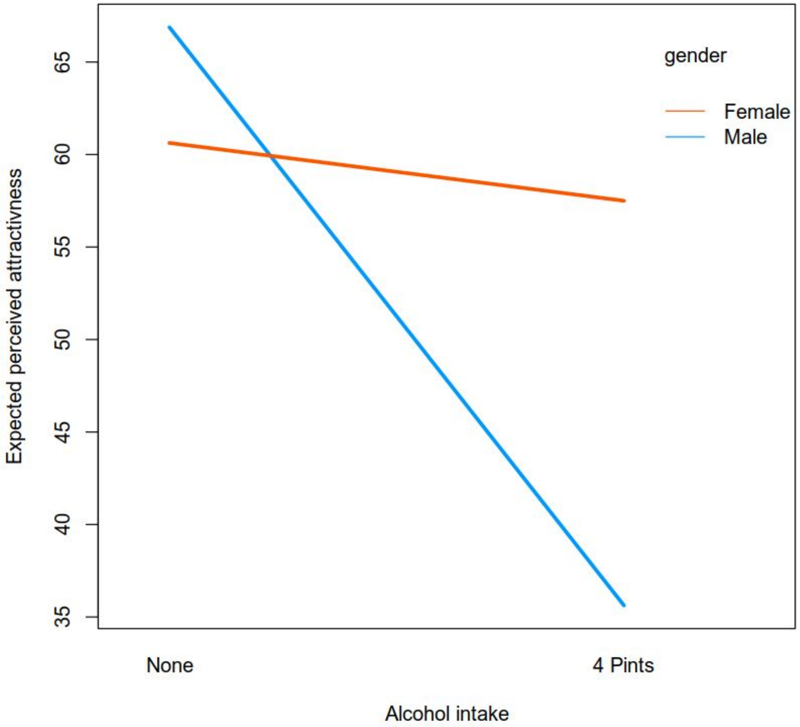
The beer-goggles effect: gender differences

```
Call:
lm(formula = attractiveness ~ alcohol * gender, data = dat.attractiveness %>%
    filter(alcohol != "2 Pints"))

Residuals:
    Min      1Q  Median      3Q     Max
-16.875  -5.625  -0.625   5.156  19.375

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                   66.875      3.055  21.890  < 2e-16 ***
alcohol4 Pints               -31.250      4.320  -7.233 7.13e-08 ***
genderFemale                  -6.250      4.320  -1.447    0.159
alcohol4 Pints:genderFemale   28.125      6.110   4.603 8.20e-05 ***
```

Visualization: Interaction plots

In R:

```
interaction.plot(
x.factor=dat.attractiveness$alcohol,
trace.factor=dat.attractiveness$gender,
response=dat.attractiveness$attractiveness
)
```



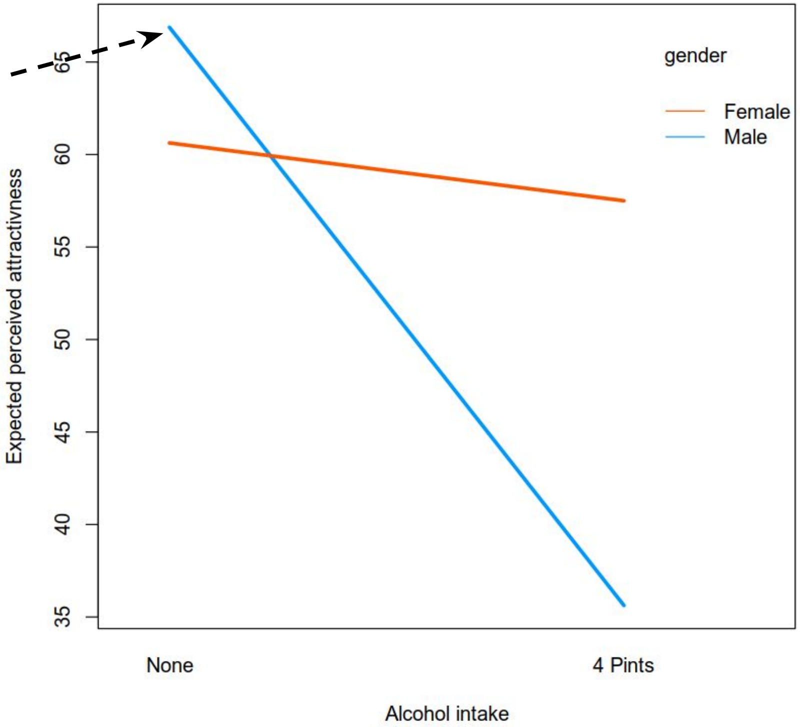
Visualization: Interaction plots

Attractiveness (no alcohol, male):
66.87 - 31.25*0 - 6.25*0 + 28.13*(0*0) = 66.87

Attractiveness (4 pints, Male):
66.87 - 31.25*1 - 6.25*0 + 28.13*(1*0) = 35.62

Attractiveness (no alcohol, Female):
66.87 - 31.25*0 - 6.25*1 + 28.13*(0*0) = 60.62

Attractiveness (4 pints, female):
66.87 - 31.25*1 - 6.25*1 + 28.13*(1*1) = 57.50



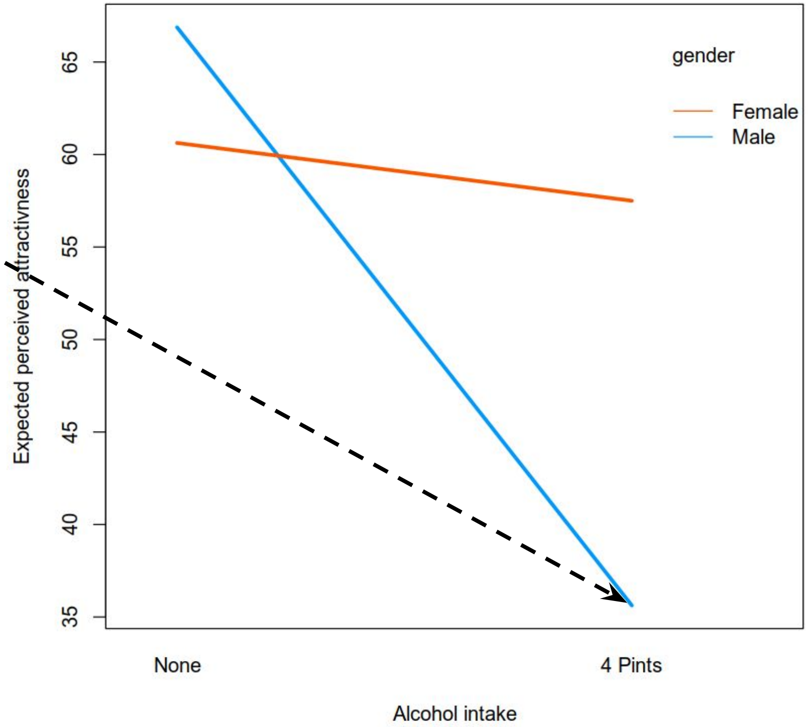
Visualization: Interaction plots

Attractiveness (no alcohol, male):
66.87 - 31.25*0 - 6.25*0 + 28.13*(0*0) = 66.87

Attractiveness (4 pints, male):
66.87 - 31.25*1 - 6.25*0 + 28.13*(1*0) = 35.62

Attractiveness (no alcohol, female):
66.87 - 31.25*0 - 6.25*1 + 28.13*(0*0) = 60.62

Attractiveness (4 pints, female):
66.87 - 31.25*1 - 6.25*1 + 28.13*(1*1) = 57.50



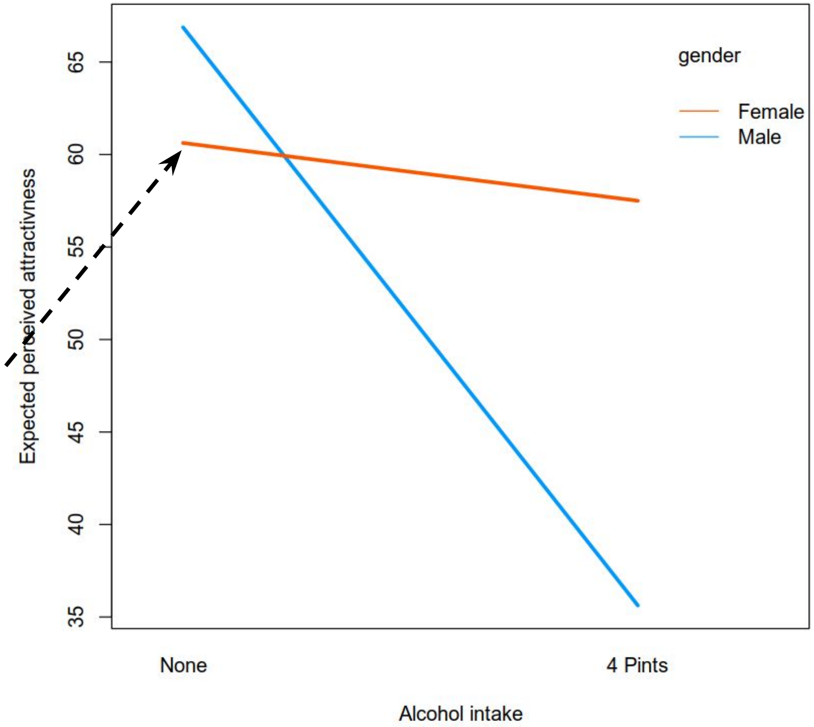
Visualization: Interaction plots

Attractiveness (no alcohol, Male):
66.87 - 31.25*0 - 6.25*0 + 28.13*(0*0) = 66.87

Attractiveness (4 pints, Male):
66.87 - 31.25*1 - 6.25*0 + 28.13*(1*0) = 35.62

Attractiveness (no alcohol, female):
66.87 - 31.25*0 - 6.25*1 + 28.13*(0*0) = 60.62

Attractiveness (4 pints, female):
66.87 - 31.25*1 - 6.25*1 + 28.13*(1*1) = 57.50



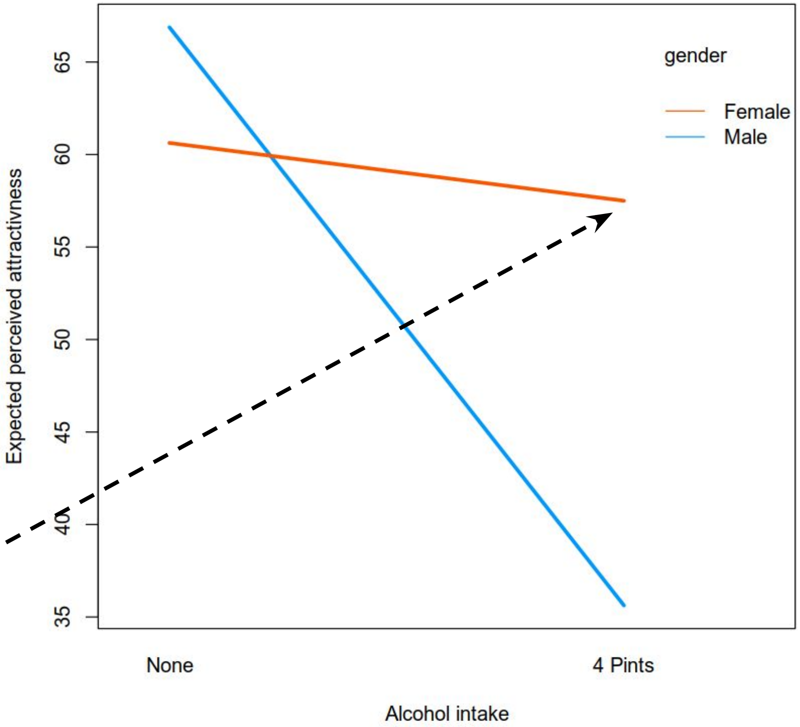
Visualization: Interaction plots

Attractiveness (no alcohol, male):
66.87 - 31.25*0 - 6.25*0 + 28.13*(0*0) = 66.87

Attractiveness (4 pints, male):
66.87 - 31.25*1 - 6.25*0 + 28.13*(1*0) = 35.62

Attractiveness (no alcohol, female):
66.87 - 31.25*0 - 6.25*1 + 28.13*(0*0) = 60.62

Attractiveness (4 pints, female):
66.87 - 31.25*1 - 6.25*1 + 28.13*(1*1) = 57.50



Visualization: Interaction plots

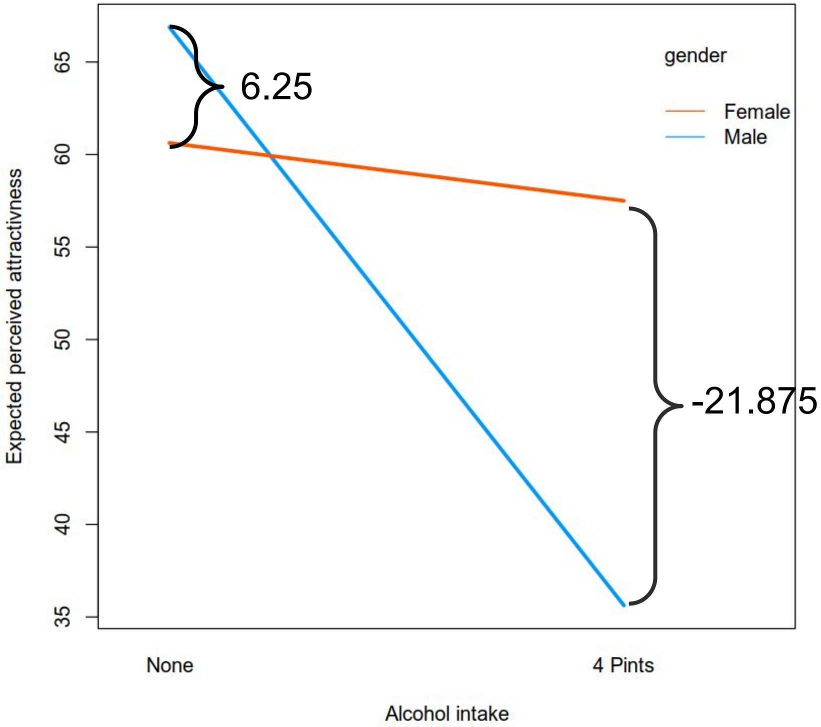
Attractiveness (No alcohol, Male):
66.87 - 31.25*0 - 6.25*0 + 28.13*(0*0) = 66.87

Attractiveness (4 Pints, Male):
66.87 - 31.25*1 - 6.25*0 + 28.13*(1*0) = 35.62

Attractiveness (No alcohol, Female):
66.87 - 31.25*0 - 6.25*1 + 28.13*(0*0) = 60.62

Attractiveness (4 Pints Female):
66.87 - 31.25*1 - 6.25*1 + 28.13*(1*1) = 57.50

$\beta_3 = 6.25 - (-21.875) = 28.125$



Recap: interaction terms

- We can use **interactions** between independent variables when we suspect a relationship between a dependent and independent variable depends on another variable.
- The inclusion of an interaction term depends on the research problem: is it theoretically founded? Is it reasonable based on our real-word experience?
- Why not always include an interaction term? Doing so increases the number of independent variables in the regression model, which might be counterproductive due to the increase of complexity.

Logistic regression

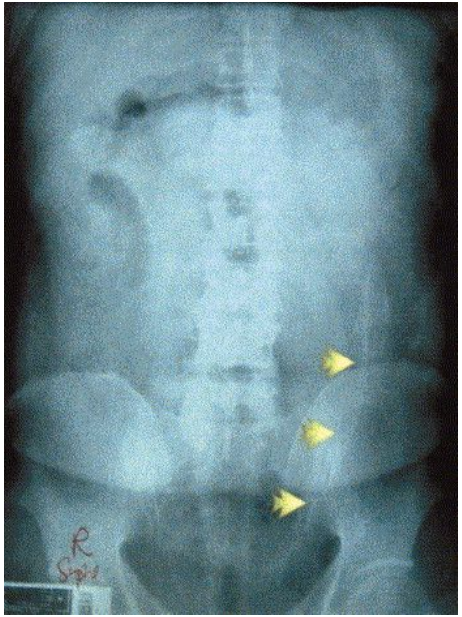
Categorical variable as the dependent variable

- So far we have been working with a **dependent variable** that is continuous, but what if we have a **binary categorical** variable instead?
- Logistic regression is a statistical method for describing this kind of relationships.
- Logistic regression works with odds, which are simply the ratio of the probability of an event occurring, $\hat{p}$, to the probability of the event not occurring $(1 - \hat{p})$:

$$odds = \frac{\hat{p}}{1 - \hat{p}}$$

Eels and constipation cure

This example is based on a research paper (Lo, Wong, Leung, Law, & Yip, 2004) that reported the case of a 50-year-old man who presented himself at the Accident and Emergency Department (ED for the Americans) with abdominal pain. A physical examination revealed peritonitis so they took an X-ray of the man's abdomen. The X-ray revealed the shadow of an eel. On further questioning, the patient admitted an eel was inserted into the rectum in an attempt to relieve constipation.



Research question: Do really eels in the rectum help cure constipation?

Field, A. P. (2023). Discovering statistics using R and RStudio (2nd ed.). London: Sage.

Eels and constipation cure: data

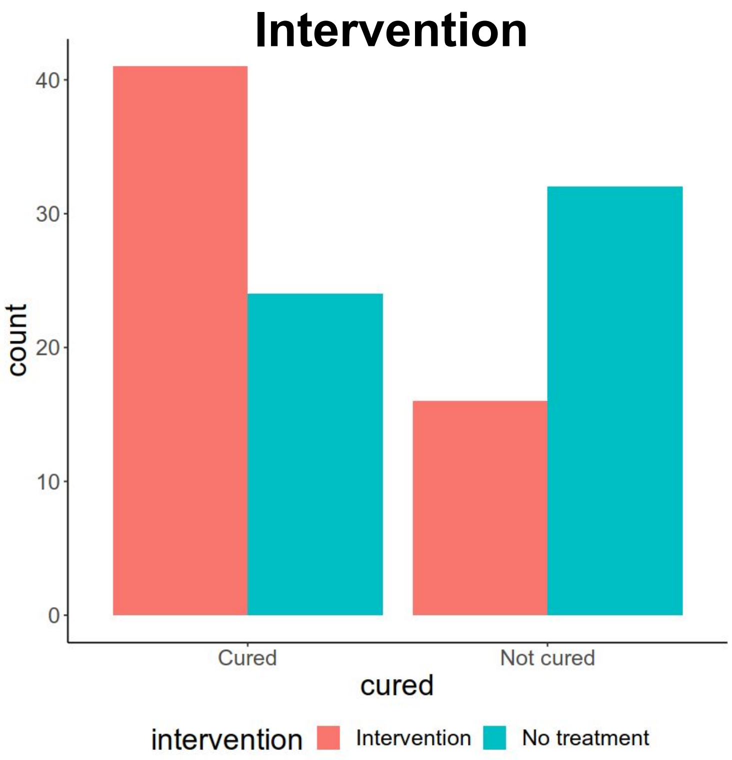
id	cured	intervention	duration
<chr>	<chr>	<chr>	<int>
ga442	Not cured	No treatment	7
y024o	Not cured	No treatment	7
9k65h	Not cured	No treatment	6
lqx1y	Cured	No treatment	8
p7415	Cured	Intervention	7
3vmpg	Cured	No treatment	6

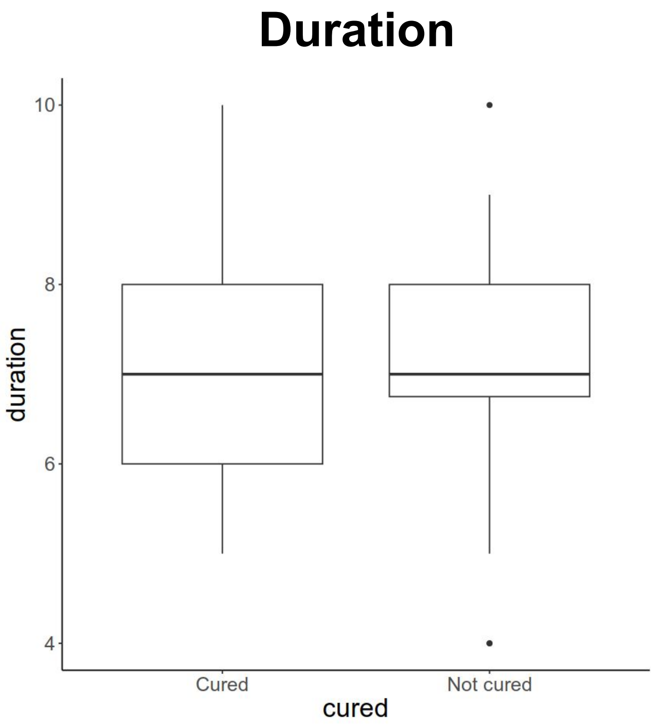
Dependent variable: cured (Cured, or Not cured).

Independent variable: Intervention i.e. eel up the anus (Intervention, or No treatment);

Independent variable: Duration (the number of days before treatment that the patient was constipated).

Eels and constipation cure: descriptive statistics





Eels and constipation cure

- It seems clear that intervention has an effect whether patients are cured. Duration is less clear, although we would need to test this explicitly. How can we do all this?

- We can not set cured to 1 and not cured to 0 and run a usual multiple linear regression model: this would yield values between 1 and 0, which is not appropriate.

- <u>One way</u>: treat cured and not cured categories as heads and tails arising from flipping a coin, and try to estimate the fairness of the coin using a transformation of a linear model of the independent variables.

Generalized Linear Model

- A general way of addressing this is by extending the linear regression model for any kind of dependent variable: the **generalized linear model** (GLM).
- Logistic regression is just one example of this type of model.
- Generalized linear models have the following three characteristics:

1. A probability distribution describing the dependent variable.
2. A linear model: $\eta_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$.
3. A link function relating the linear model to the parameter of the dependent variable distribution (e.g. a mean, a proportion, etc): $g(\theta) = \eta$ or $\theta = g^{-1}(\eta)$.

The linear regression model

The multiple linear regression model is just another case of a Generalized Linear Model:

1. A probability distribution describing the dependent variable.

$$Y \sim \mathcal{N}(\mu, \sigma^2))$$

2. A linear model:

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}.$$

3. A link function: the identity function (i.e. $\theta \equiv \mu$).

$$\mu = \eta$$

Logistic regression

- Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical independent variables.
- We assume that a binomial distribution produce the dependent variable, and we therefore want to model the probability of success, *p*, for a given set of independent variables (i.e. whether a coin comes up heads).
- The logistic model gets specified with a reasonable link function. The most commonly used for Logistic regression is the logit function.

$$\operatorname{logit}(p) \equiv \log\left(\frac{p}{1-p}\right), \text{for } 0 \leq p \leq 1$$

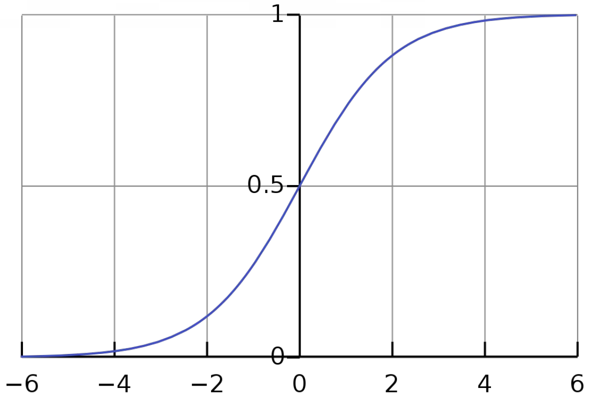
Properties of the logit function

➢ The **logit** function takes a value between 0 and 1 and maps it to a value between −∞ and ∞.

➢ **Logit** is the **inverse** of the **logistic** function:

$$\mathrm{logit}^{-1}(p) \equiv \left(\frac{1}{1+e^{-x}}\right), \text{for} -\infty < x < \infty$$

The **logistic (inverse logit)** function takes a value between −∞ and ∞, and maps it to a value between 0 and 1.



The logistic regression model

The logistic regression model is again another case of a Generalized Linear Model.

1. A probability distribution describing the dependent variable.

$$Y \sim Binom(p, N)$$

2. A linear model:

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}.$$

3. A link function: the logit function.

$$\text{logit}\ (p) = \eta \qquad (\textit{here } \theta \equiv p)$$

Modelling

In R we fit a GLM in the same was as a linear model except using **glm** instead of **lm**. We must also specify the type of GLM to fit using the *family* argument.

```
Call:
glm(formula = cured ~ intervention + duration, family = binomial(),
    data = eel.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6025  -1.0572   0.8107   0.8161   1.3095

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.234660   1.220563  -0.192  0.84754
interventionIntervention  1.233532   0.414565   2.975  0.00293 **
duration                 -0.007835   0.175913  -0.045  0.96447
```

Modelling

In R we fit a GLM in the same was as a linear model except using **glm** instead of **lm**. We must also specify the type of GLM to fit using the *family* argument.

```
Call:
glm(formula = cured ~ intervention + duration, family = binomial(),
    data = eel.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6025  -1.0572   0.8107   0.8161   1.3095

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.234660   1.220563  -0.192  0.84754
interventionIntervention  1.233532   0.414565   2.975  0.00293 **
duration                 -0.007835   0.175913  -0.045  0.96447
```

Estimated logistic regression model

```
Call:  glm(formula = cured ~ intervention + duration, family = binomial(),
    data = eel.dat)

Coefficients:
            (Intercept)  interventionIntervention                  duration
              -0.234660                  1.233532                 -0.007835
```

Model: $P(\text{cured}) = \dfrac{1}{1 + \exp(0.234 - 1.233 \cdot \text{intervention} + 0.007 \cdot \text{duration})}$

Interpretation: Odds

Probability of cured with intervention (intervention=1) and no days of treatment:

$$P(\text{cured}) = \frac{1}{1 + e^{0.234 - 1.233 \cdot 1 + 0.007 \cdot 0}} = 0.731$$

$$P(\text{not cured}) = 1 - P(\text{cured}) = 0.269$$

$$\text{odds} = \frac{0.731}{0.269} = 2.717$$

Interpretation: Odds

Probability of cured with no intervention (intervention=0) and no days of treatment:

$$P(\text{cured}) = \frac{1}{1 + e^{0.234 - 1.233 \cdot 0 + 0.007 \cdot 0}} = 0.441$$

$$P(\text{not cured}) = 1 - P(\text{cured}) = 0.559$$

$$\text{odds} = \frac{0.441}{0.559} = 0.789$$

Interpretation: Odds-ratio

- We can calculate the odds-ratio (OR), as the the proportionate change in odds by dividing the odds after a unit change in the independent variable by the odds before that change:

$$\mathrm{OR} = \frac{2.717}{0.789} = 3.443$$

- The odds of a patient who is treated being cured are 3.443 times higher than those of a patient who is not treated.

Inference

summary(glm(cured ~ intervention + duration, data=eel.dat, family = binomial()))

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.234660   1.220563  -0.192  0.84754
interventionIntervention  1.233532   0.414565   2.975  0.00293 **
duration                 -0.007835   0.175913  -0.045  0.96447
```

- ➢ No hypothesis test for the whole model.
- ➢ Inference on individual coefficients through z-tests.
- ➢ The calculation of the standard errors is tricky (beyond the scope of this course).

Recap

1. When **dependent variables** are **categorical and binary**, we can use **logistic regression.**
2. Logistic regression is one of a family of models called **Generalized Linear Models** that often apply when the assumptions of linear regression fail.
3. This is an extremely general statistical method that you'll be able to use in almost all of the cases you're likely to work with.