# 2. Model assumptions

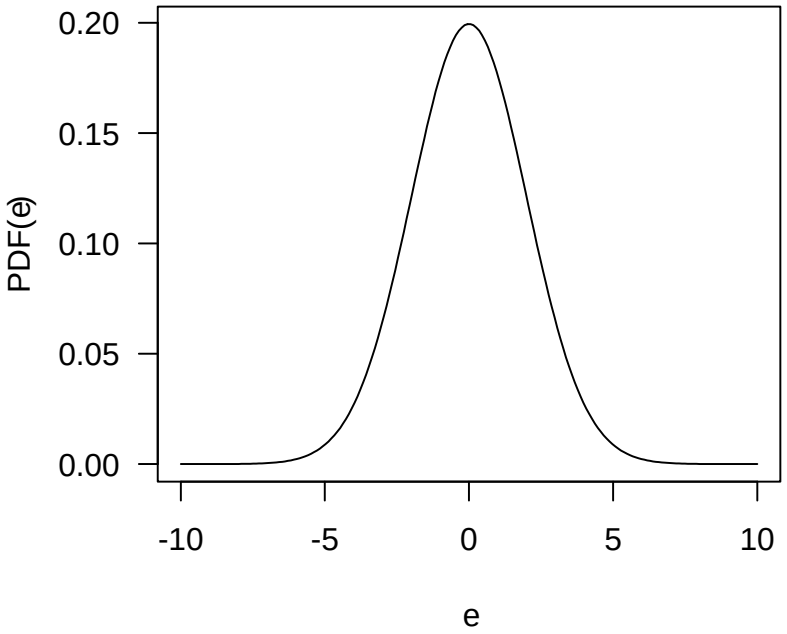
- Know what the assumptions of linear models are & and the problems that arise when they are violated
- Understand how to assess whether they are being violated
- Learn how to correct/deal with these violations (ongoing – more detail will be given later on in the “Extensions” part of the course)

# Are these results real?

- There are a number of assumptions that these linear models make
- If these assumptions are not met, your results are not valid!

# Assumptions of LMs

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ $\epsilon_i \sim N(0, \sigma^2)$



The assumptions of your model are about the

# RESIDUALS/ERRORS

**NOT the raw data!**

# Normality of errors

- After you've fit your model, do the RESIDUALS (NOT raw data) appear to follow a Gaussian distribution?
- Why violating this is bad:
  - Skewed data will alter your error for your estimates, making testing whether they are different from zero difficult
- Why violating this is not so bad:
  - Most models are very robust to (moderate) departures from normality
- How to fix:
  - Check for outliers – a few very large data points highly influential
  - Use different distribution
  - Transform
  - Tests of normality (KS, Shapiro-Wilk) very conservative; graphical inspection better

# Variance homogeneity

- Is the variance in your residuals the same across the whole range? Is the variance among any grouping variables the same?
- Why violating this is bad:
  - Minor violations probably okay, but larger heteroscedasticy can seriously impact your errors on your parameter estimates making any hypothesis testing invalid.
- How to fix:
  - Make sure to include proper predictors
  - Add weights to predictors/grouping variables
  - Variance-stabilizing transformations
  - Statistical tests (Bartletts) very sensitive to non-normality; graphical inspection better

# Linearity and additivity

- Is there a straight line relationship between your predictors and response? Are the effects of the predictors on your response additive?
- Why violating this is bad:
  - Pretty serious violation that will make your parameter estimates and predictions completely inaccurate
- How to fix:
  - Check that not missing an important predictor.
  - Non-linear transformation (e.g. log)
  - Move into a different modeling world (e.g. additive)

# Independence

- Do the residuals from your model follow any pattern either over time or in relation to predictor variables?
- Why violating this is bad:
  - If your data are inherently grouped together (e.g. individuals, plots), and you don't account for this you are artificially inflating your degrees of freedom (pseduo-replication) and violating all assumptions of hypothesis testing
  - If your data is auto-correlated then this will impact your estimates of your parameter estimate errors.

- How to fix:
  - Add lags in data (temporal)
  - **Include random effects for groups**

The assumptions of your model are about the

# How to check model assumptions

# Good R practices

- **Annotate, annotate, annotate**
  - Your worst collaborator is you 6 months ago
- Name variables and models something meaningful (‘data’ sucks)
- Do not overwrite data files/models/variables
- Write code in independent chunks
- Do not save workspace in R Studio
  - Tools -> Global options -> uncheck ‘restore workspace’ and ‘never’ on dropdown

# What am I doing here? What data set is this? What project is this?

```
fix1 <- lmer(resp1 ~ sl + obs +  f1*t1 + f1*I(t1n^2)
             + (1|ID),
             data = data, na.action = na.omit, REML = T)

plot(fix1)

colnames(model.matrix(fix1))

Fixed <- fixef(fix1)[2]*model.matrix(fix1)[,2] + fixef(fix1)[3]*model.matrix(fix1)[,3] +
  fixef(fix1)[4]*model.matrix(fix1)[,4] + fixef(fix1)[5]*model.matrix(fix1)[,5] +
  fixef(fix1)[6]*model.matrix(fix1)[,6] + fixef(fix1)[7]*model.matrix(fix1)[,7] +
  fixef(fix1)[8]*model.matrix(fix1)[,8]

varF <- var(Fixed)

varF/(varF + VarCorr(fix1)$Fish.ID[1] + attr(VarCorr(fix1), "sc")^2)

(varF + VarCorr(fix1)$Fish.ID[1] )/
  (varF + VarCorr(fix1)$Fish.ID[1] + (attr(VarCorr(fix1), "sc")^2))
```

# How to check model assumptions

- Go to R to demonstrate checking model assumptions with the clam data and loyn data

# Species richness on marine beaches

- We collected data on species richness at 9 different beaches. Each beach varied in its level of tidal exposure (likely an important predictor of richness). Within each beach, we sampled at 5 different sites – each site varied at its tidal level (another important predictor).

- What is our model?

# Species richness on marine beaches

```
Call:
lm(formula = Richness ~ factor(Exposure) + NAP, data = rikz)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5702 -1.7124 -0.6704  0.8728 13.0686

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          12.945      1.407   9.198 1.61e-11 ***
factor(Exposure)10   -5.381      1.567  -3.434  0.00137 **
factor(Exposure)11   -8.821      1.568  -5.625 1.48e-06 ***
NAP                  -2.717      0.476  -5.708 1.13e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.134 on 41 degrees of freedom
Multiple R-squared:  0.6345,    Adjusted R-squared:  0.6078
F-statistic: 23.73 on 3 and 41 DF,  p-value: 4.586e-09
```

# What are degrees of freedom?

- Number of observations in the data that are free to vary when estimating parameters

# Why are degrees of freedom important?

- Because they give you power!
- The value of the different test-statistics (t, F, etc) will be determined by your df
- The more degrees of freedom you have, the better you will be able to find an effect (if there is an effect to be found)

# Model assumptions

- Know what the assumptions of linear models are & and the problems that arise when they are violated
- Understand how to assess whether they are being violated
- Learn how to correct/deal with this (ongoing – more detail will be given later on in the "Extensions" part of the course)

FURTHER READING on model assumptions:

- Section 2.3 in Zuur
- Jacqmin-Gadda et al. 2007. Robustness of the linear mixed model to misspecified error distribution. *Comp Stats & Data Analysis* 51