

# R Programming For Natural Resource Professionals

Week 12-13

ANOVA and regression in R

The week ahead...

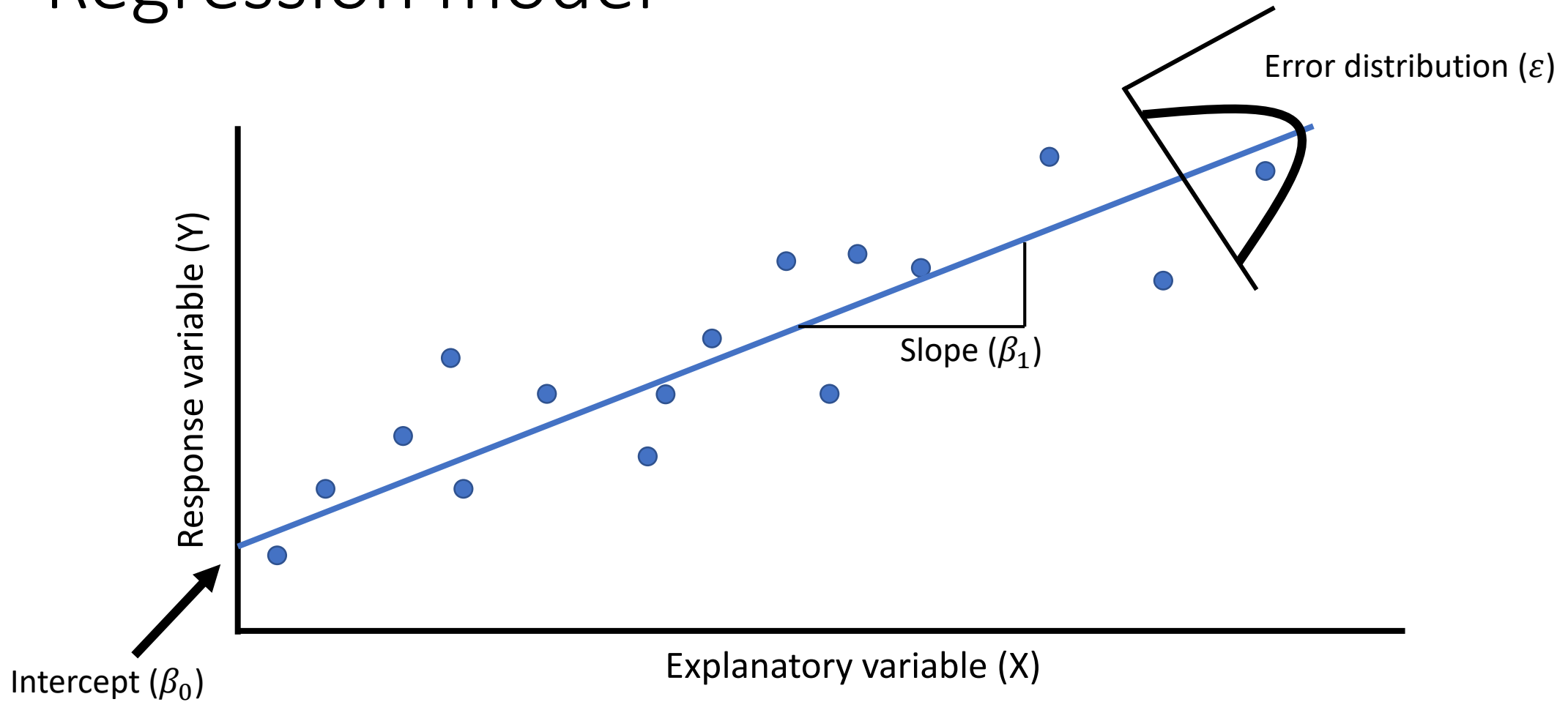
# A(nother) syllabus change

Two of the homework assignments will be larger in scope and will serve as a midterm and final. The final homework ~~will require data analysis of the students' own data~~ and a short write up formatted like a scientific paper. If the student does not have their own data, a dataset will be provided.

# Learning objectives for this week

1. Perform and interpret regression
  1. T-test
  2. ANOVA
  3. Linear regression
2. Assess model assumptions (i.e., model validation)
3. Model comparison

# Regression model



$$Y = \beta_0 + \beta_1 + \varepsilon$$

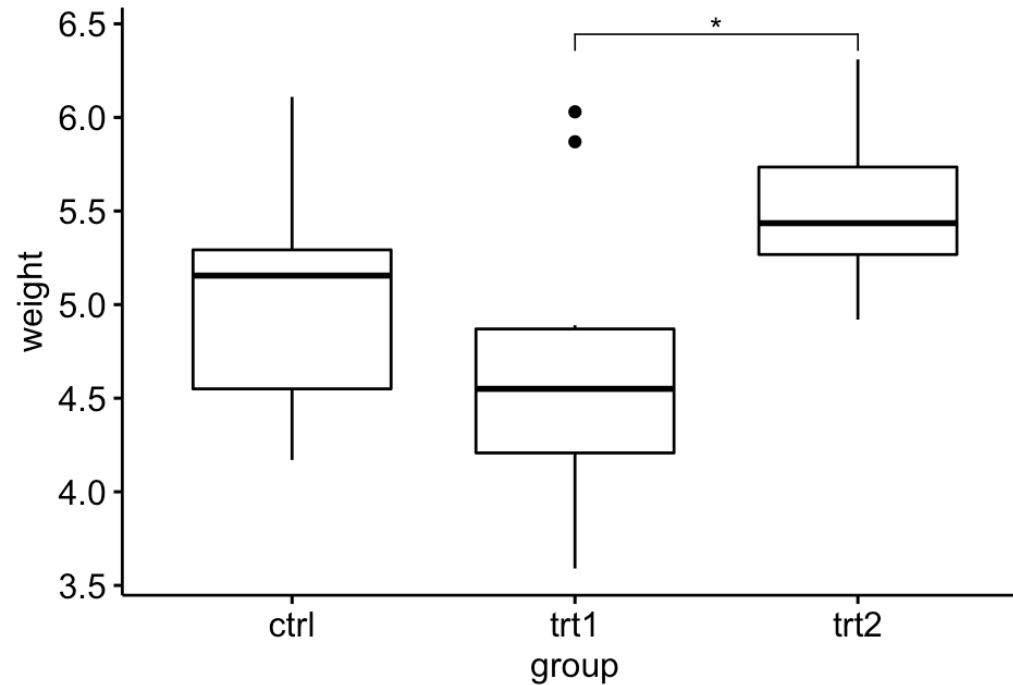
## Goals of constructing models

1. Parameter estimation: What parameter values best fit the data?
  - Model fitting
2. Inference: How certain are the estimates the model produces?
  - Assessing goodness-of-fit
3. Adequacy: Is the model the right choice?
  - Model selection

# ANOVA

Common use: Explanatory variable is more than two categories

*Do the means of more than two independent samples differ?*



# ANOVA assumptions

- 1. Normality:** Model residuals are approximately normally distributed.
- 2. Homogeneity of variances:** Both samples have approximately the same variance.
- 3. Random sampling:** Samples were obtained using a random sampling method.
- 4. Independence:** The observations in one sample are independent of the observations in the other sample.



# Violated ANOVA assumptions

## **One-way ANOVA: Kruskal-Wallis ANOVA**

`kruskal.test(response ~ predictor, data = dat)`

## **Two-way ANOVA: variable transformations**

`log(var)`, `sqrt(var)`, etc.

*[Nuances beyond the scope of this course]*

# One-way ANOVA

Common use: Determine whether differences exist between the means of three or more independent (unrelated) samples.

```
lm(response ~ predictor, data = dat) %>%  
  anova() %>%  
  tidy()
```

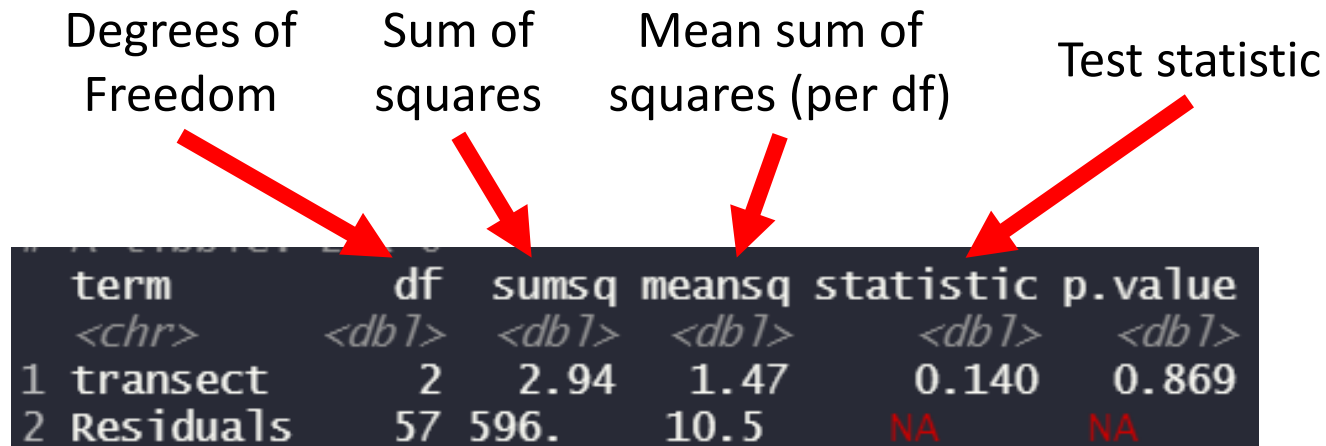


Diagram illustrating the components of the ANOVA table output, with red arrows pointing from labels to the corresponding columns:

- Degrees of Freedom points to the **df** column.
- Sum of squares points to the **sumsq** column.
- Mean sum of squares (per df) points to the **meansq** column.
- Test statistic points to the **statistic** column.

|   | term      | df    | sumsq | meansq | statistic | p.value |
|---|-----------|-------|-------|--------|-----------|---------|
|   | <chr>     | <dbl> | <dbl> | <dbl>  | <dbl>     | <dbl>   |
| 1 | transect  | 2     | 2.94  | 1.47   | 0.140     | 0.869   |
| 2 | Residuals | 57    | 596.  | 10.5   | NA        | NA      |

# One-way ANOVA: summary() output

```
lm(response ~ predictor, data = dat) %>%  
summary()
```

First group is termed  
'intercept' and becomes  
reference group  
(transectR1). Others are  
relative to that reference.

```
Call:  
lm(formula = leaf1area ~ transect, data = .)  
  
Residuals:  
    Min       1Q   Median       3Q      Max   
-6.9911 -2.5478  0.1162  2.4588  7.7210   
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)      
(Intercept)   9.8788     0.7232  13.660  <2e-16 ***  
transectR2    -0.5129     1.0228   -0.501    0.618   
transectR6    -0.4078     1.0228   -0.399    0.692   
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 3.234 on 57 degrees of freedom  
Multiple R-squared:  0.0049,    Adjusted R-squared:  -0.03002  
F-statistic: 0.1403 on 2 and 57 DF,  p-value: 0.8694
```

t-test evaluates whether  
coefficients are significantly  
different from zero

Assessment of model fit

Overall model p-value

# Two-way ANOVA

Used to determine the effect of two categorical predictor variables on a continuous response variable

```
lm(response ~ predictor1 + predictor2, data = dat) %>%  
  anova() %>%  
  tidy()
```

# Two-way ANOVA: Post-hoc analysis

*Post hoc*: Latin phrase meaning “after this.” Used to describe follow-up analyses.

**Tukey Honest Significant Differences:** Evaluates combinations of categories within variables.

```
model <- aov(response ~ predictor1 + predictor2, data = dat)
TukeyHSD(model)
```

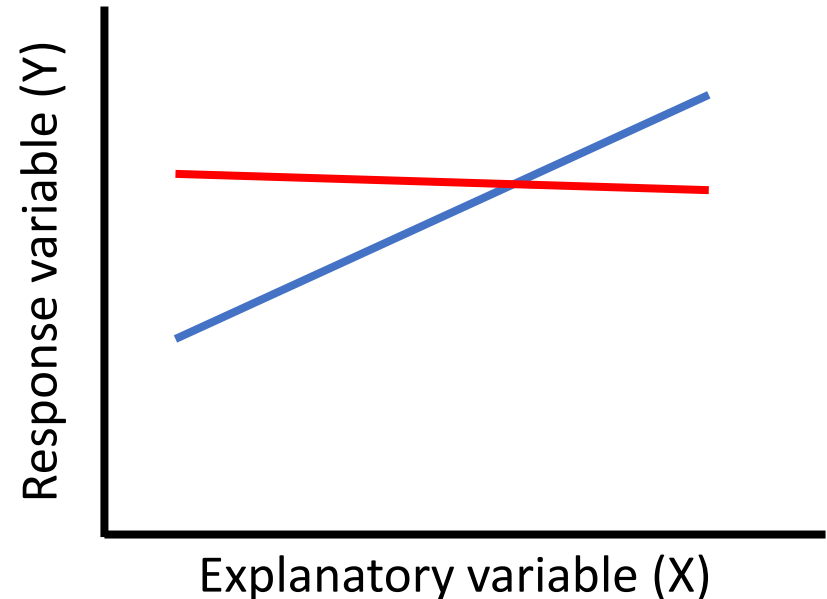
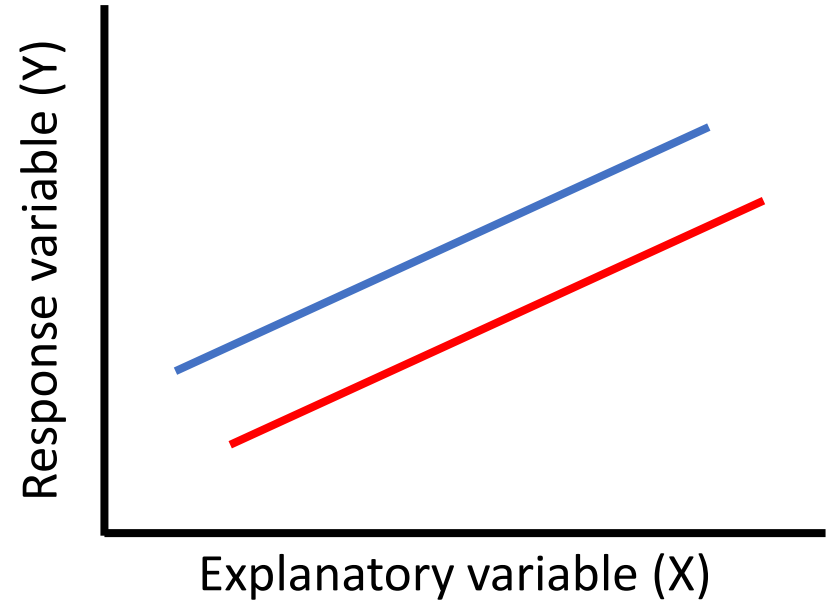
# Modeling interactions

Additive model: Slopes are fixed among predictor variables. They affect the response variable in the same way.

- R syntax: `var1 + var2`

Model with interaction: Slopes allowed to vary among predictor variables. They can uniquely affect the response variable.

- R syntax: `var1*var2` or `var1:var2`



# Modeling interactions

Only include an interaction if you hypothesize that one is present.

To test statistical support:

```
model1 <- lm(response ~ predictor1+predictor2, data = dat)
```

```
model2 <- lm(response ~ predictor1*predictor2, data = dat)
```

```
anova(model1, model2)
```

```
Analysis of Variance Table

Model 1: mean.Hobs ~ basin + habitat
Model 2: mean.Hobs ~ basin * habitat
  Res.Df    RSS Df Sum of Sq    F Pr(>F)
1     38 0.0040055
2     36 0.0039314  2  7.4114e-05 0.3393 0.7145
```

Null hypothesis: There is no difference between the models

# Simple linear regression

- How does one continuous variable depend on another continuous variable?
- With regression, we'll switch to checking most assumptions based on residuals instead of raw data
- So, first make your model.

```
model <- dat %>%  
  lm(response ~ predictor, data = .)
```



# Simple linear regression

- `Broom::tidy()` to view results as a tibble (e.g., `model %>% tidy()`)
- `summary()` to view the base R presentation (e.g., `model %>% summary()`).

```
# A tibble: 2 x 5
  term      estimate std.error statistic  p.value
<chr>      <dbl>      <dbl>      <dbl>    <dbl>
1 (Intercept)  479.         54.8         8.74 2.72e-15
2 year        -0.195        0.0283      -6.88 1.23e-10
```

```
Call:
lm(formula = ice_duration ~ year, data = .)

Residuals:
    Min       1Q   Median       3Q      Max
-68.750  -8.844   0.915  11.821  47.700

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  479.2000    54.8283   8.740 2.72e-15 ***
year         -0.1946     0.0283  -6.878 1.23e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.31 on 163 degrees of freedom
(2 observations deleted due to missingness)
Multiple R-squared:  0.2249,    Adjusted R-squared:  0.2202
F-statistic: 47.31 on 1 and 163 DF,  p-value: 1.234e-10
```

# Simple linear regression

- Additional model performance stats available using `broom::glance()`

Residual  
standard  
deviation

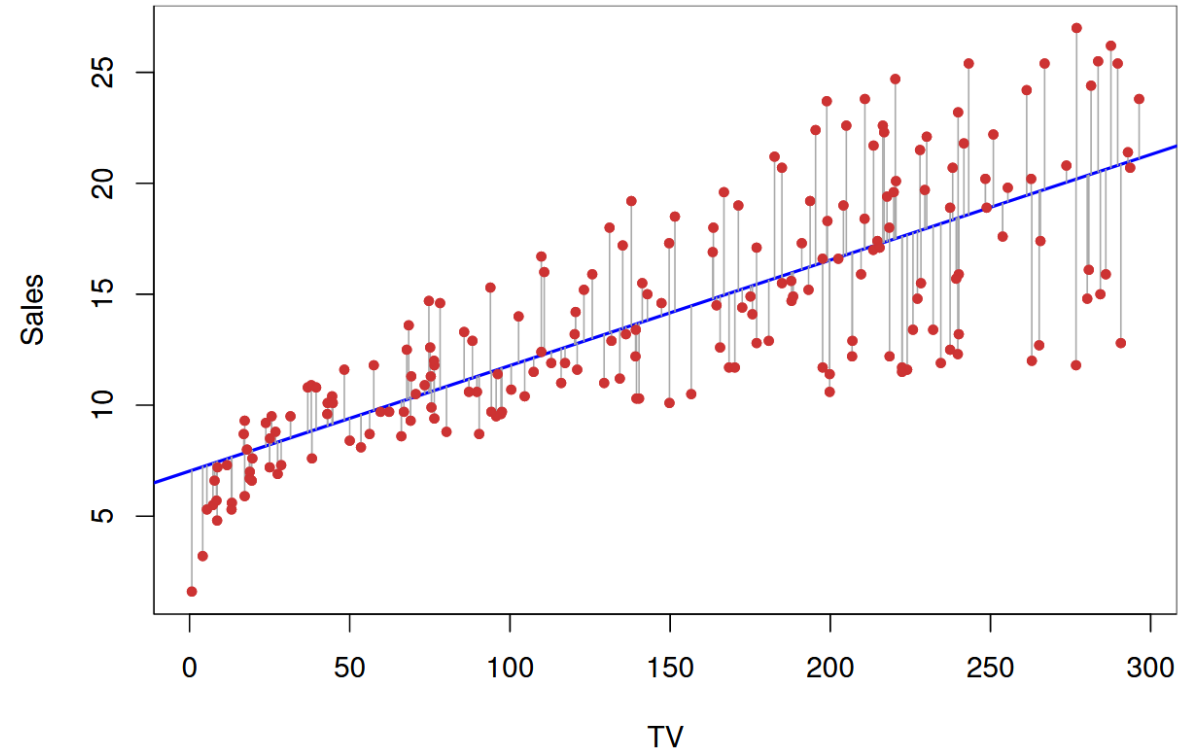
F-statistic comparing  
model performance vs.  
noise

Stats for  
comparing model  
performance

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC    BIC deviance df.residual  nobs
  <dbl>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl> <dbl> <dbl>   <dbl>    <int> <int>
1   0.225     0.220   17.3    47.3 1.23e-10     1 -704. 1413. 1423.  48857.    163   165
```

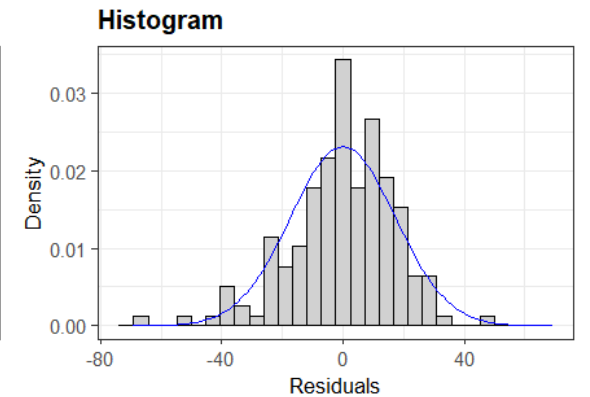
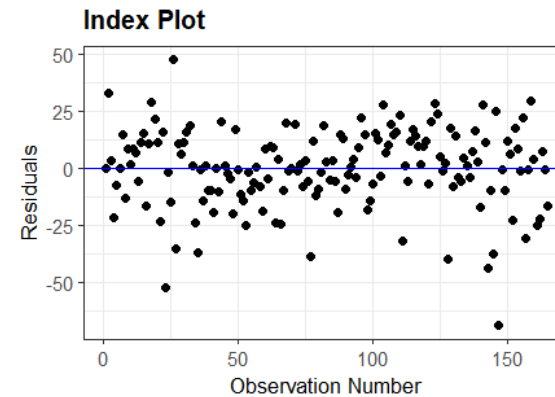
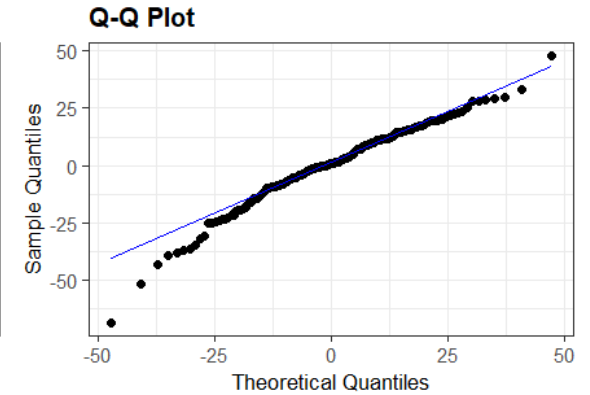
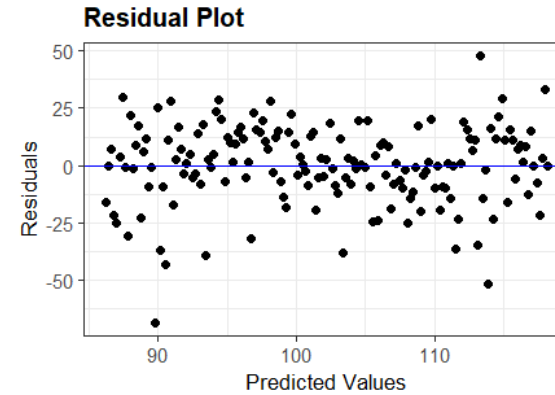
# Simple linear regression

- Plot residuals to assess assumptions
- Defining ‘residuals’
  - When fitting a model, the goal is to minimize the sum of residuals’ values.
  - Residuals are the numerical realization of the model’s error term ( $\epsilon$ ).



# Simple linear regression

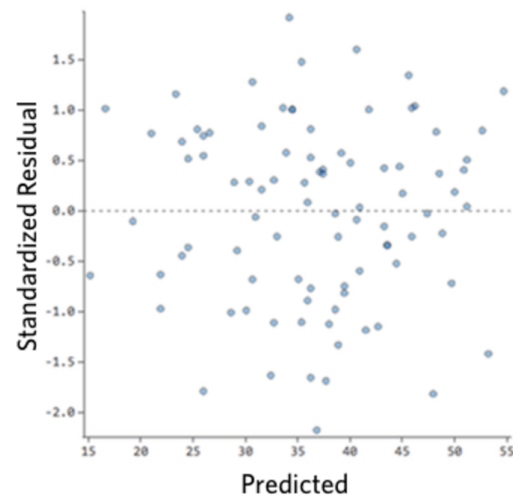
- Plot residuals to assess assumptions
  - Access residuals using `broom::augment()`
- Four generally useful plots
  - qq plot
  - Index plot
  - Residual plot
  - Histogram
- Can be made all at once using `ggResidpanel::resid_panel()`
  - `model %>% resid_panel()`



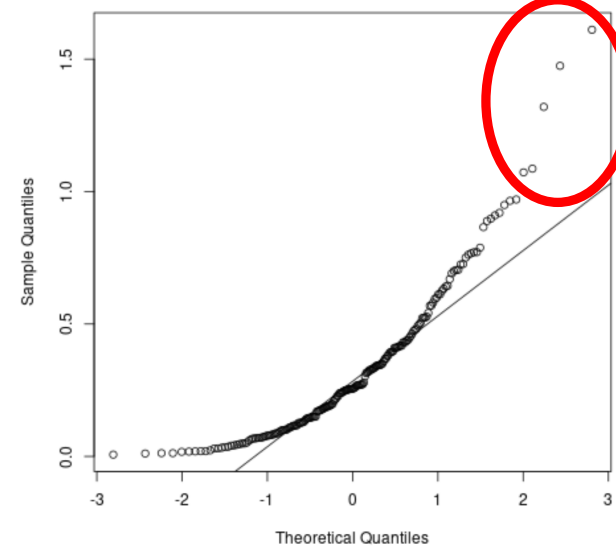
# Simple linear regression

- Assumptions of linear regression
  - Normal distribution of residuals
  - Homogeneity of variance among variables
  - No outliers
  - Linear relationship between variables

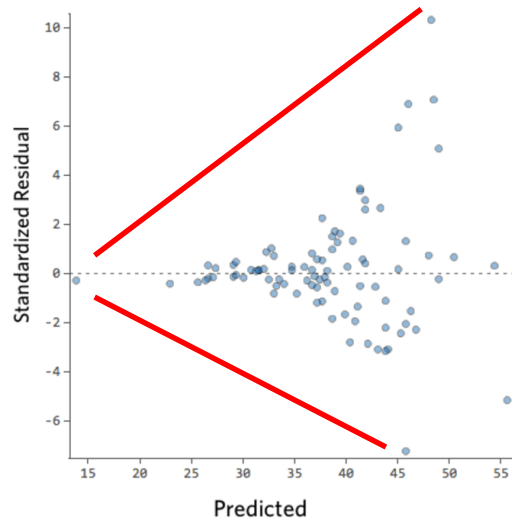
No obvious issues



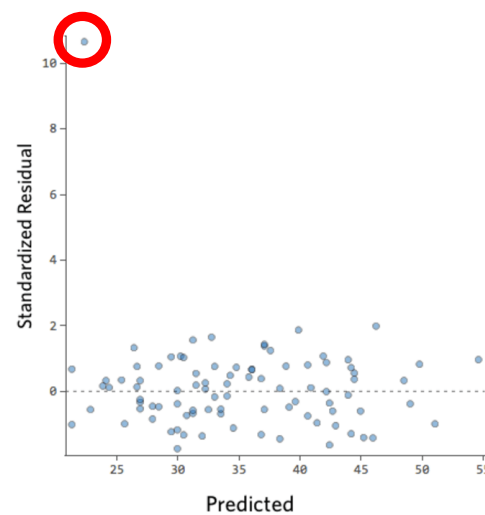
Non-normal distributions



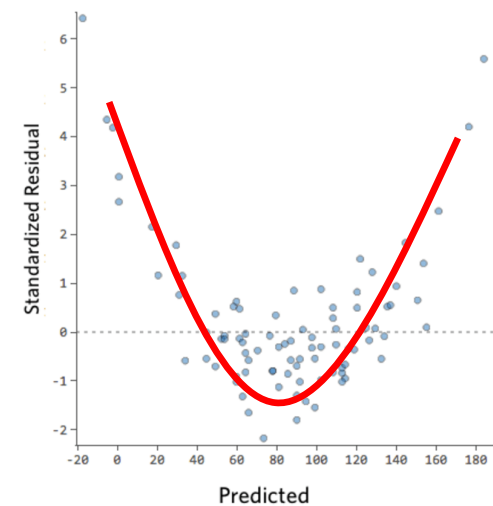
Heterogeneous variance



Presence of outliers



Nonlinear relationships



# Simple linear regression

## What to do about violated assumptions?

- Most common fix is to transform (log, square root, etc.)
  - Start with the predictor variables
  - Can also transform response variable or both predictor and response
  - No clear consensus on whether to plot transformed or non-transformed data
- Possible to drop outliers with proper rationale
- Use a data simulation approach to test your hypothesis (more next week)

# What to do with extra course period next week?

- <https://forms.gle/ZrrErz1wF8sz673j8>



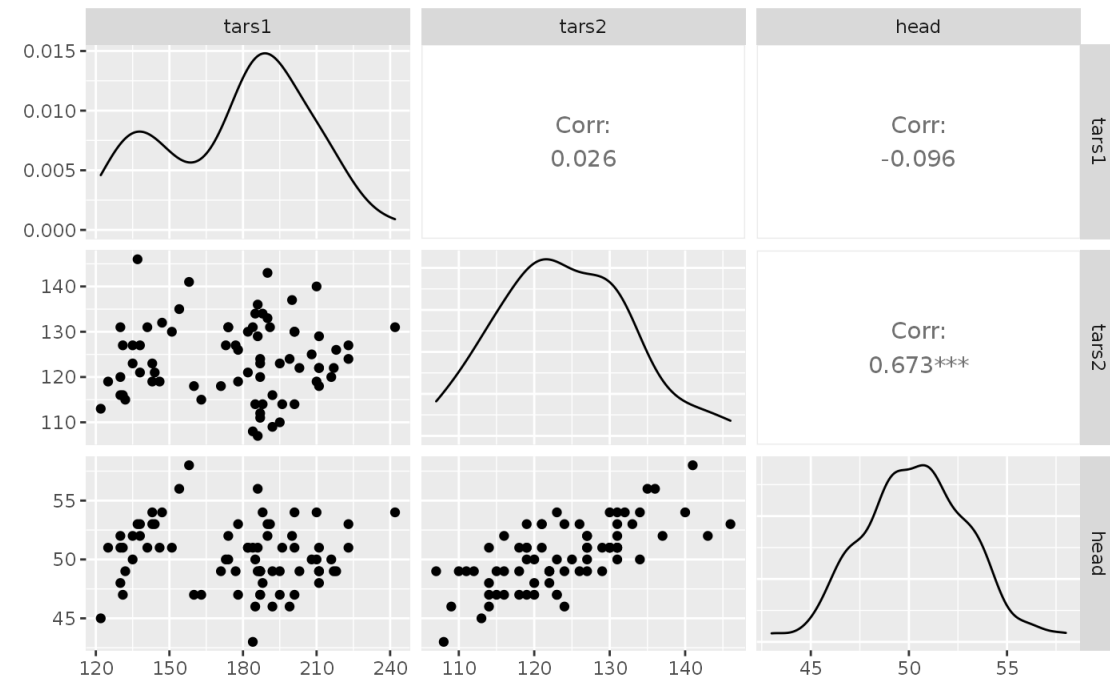
# Multiple linear regression

- How does one continuous variable depend on a set of other variables?

```
model <- dat %>%  
  lm(response ~ predictor1 + predictor2 + predictor3..., data = .)
```

# Multiple linear regression

- A new assumption to check: multicollinearity
  - Predictor variables must be independent of each other
- Visual approach: `GGally::ggpairs`
- Semi-quantitative approach: `car::vif`
  - Performed on model itself
  - Conservation cut-off: 2.5
  - Some argue for cut-offs up to 10.0



# Multiple linear regression

```
model %>% broom::tidy()
```

```
# A tibble: 3 x 5
  term      estimate std.error statistic  p.value
<chr>      <dbl>      <dbl>      <dbl>    <dbl>
1 (Intercept)  54.2         4.78        11.3 1.69e-24
2 salary      0.0222      0.00543      4.08 5.95e- 5
3 walks       1.02       0.113        9.00 4.76e-17
```

```
model %>% broom::glance()
```

```
A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC    BIC deviance df.residual  nobs
  <dbl>      <dbl>    <dbl>      <dbl>    <dbl>   <dbl> <dbl> <dbl> <dbl>      <int>    <int>
1  0.384      0.380    35.5        81.2 4.08e-28     2 -1311. 2630. 2644. 328436.     260     263
```

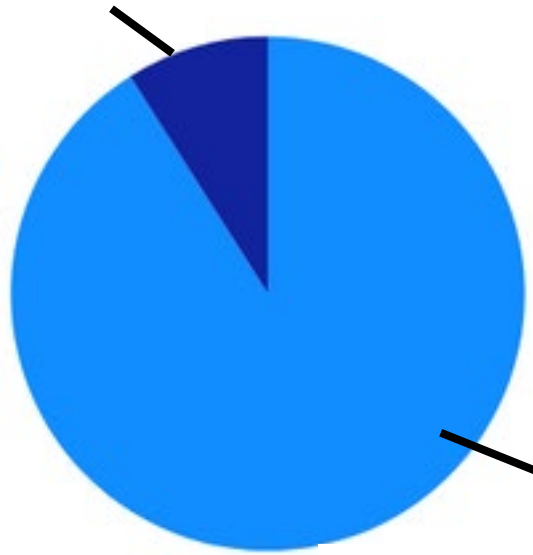
# ANCOVA

- Purpose: Provides a means of controlling for non-target variation in data when it is not possible to control for it in the experiment itself.
- ANCOVA = Analysis of Covariation
  - ANOVA + regression
- Mix of continuous and categorical predictor variables
- Continuous variables referred to as *covariates*.
- Assumes parallel slopes across groups (i.e., must use an additive model)

# ANOVA vs. ANCOVA variance partitioning

ANOVA variance

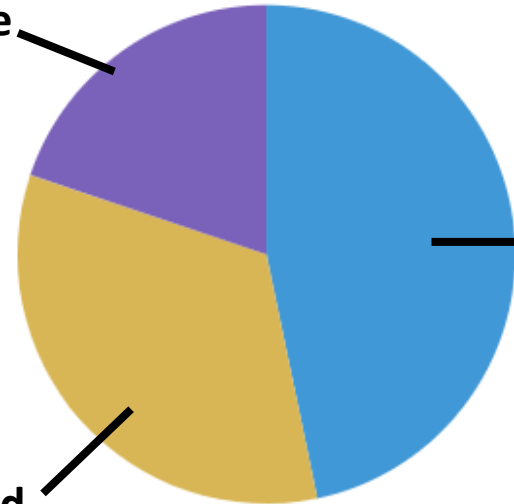
Variation explained by  
predictor variable



Unexplained  
variation

ANCOVA error variance

Variation explained  
by predictor variable



Unexplained  
variation

Variation explained  
by covariates

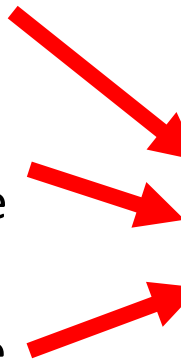
# ANCOVA example

```
model <- dat %>%  
  lm(response ~ predictor + covariate, data = .)
```

Intercept = reference factor level

Parameter estimates for other  
factor levels, relative to reference

Importance of covariate



# A tibble: 4 x 5

|   | term<br><chr> | estimate<br><dbl> | std.error<br><dbl> | statistic<br><dbl> | p.value<br><dbl> |
|---|---------------|-------------------|--------------------|--------------------|------------------|
| 1 | (Intercept)   | -0.353            | 0.848              | -0.417             | 6.79e- 1         |
| 2 | groupgrp2     | -0.546            | 0.177              | -3.09              | 3.61e- 3         |
| 3 | groupgrp3     | -2.87             | 0.176              | -16.4              | 1.36e-19         |
| 4 | pretest       | 0.987             | 0.0491             | 20.1               | 7.55e-23         |

# Model selection

- Goal: choose the mix of predictor variables that most parsimoniously explain the response variable.
- Parsimony: the idea that variables are often related and behave in the simplest or most economical way.
  - Occam's Razor: Of two competing theories, the simpler explanation is to be preferred.

# Model selection

• **Akaike information criterion (AIC)**, a measure of the goodness fit of an estimated statistical model

- Bayes factor

• **Bayesian information criterion (BIC)**, also known as the Schwarz information criterion, a statistical criterion for model selection

- Bridge criterion (BC), a statistical criterion that can attain the better performance of AIC and BIC despite the appropriateness of model specification.<sup>[4]</sup>

- Cross-validation

• **Deviance information criterion (DIC)**, another Bayesian oriented model selection criterion

- False discovery rate

- Focused information criterion (FIC), a selection criterion sorting statistical models by their effectiveness for a given focus parameter

- Hannan–Quinn information criterion, an alternative to the Akaike and Bayesian criteria

- Kashyap information criterion (KIC) is a powerful alternative to AIC and BIC, because KIC uses Fisher information matrix

• **Likelihood-ratio test**

- Mallows's  $C_p$

- Minimum description length

- Minimum message length (MML)

- PRESS statistic, also known as the PRESS criterion

- Structural risk minimization

- Stepwise regression

- Watanabe–Akaike information criterion (WAIC), also called the widely applicable information criterion

- **Extended Bayesian Information Criterion (EBIC)** is an extension of ordinary Bayesian information criterion (BIC) for models with high parameter spaces.

- **Extended Fisher Information Criterion (EFIC)** is a model selection criterion for linear regression models.



# Model selection

## Two main strategies

- Forward optimization: start with **no** predictors and keep adding variables until the most parsimonious model is identified.
- Backward optimization: start with **all** the predictors and remove variables, starting with the least statistically significant ones, until the most parsimonious model is identified.
- Rule of thumb: Forward optimization best for large number of variables, otherwise use backward optimization.

# Model selection

- For backward optimization, first construct a 'global' model
  - Include all plausible predictor variables.
  - Do not simply include all predictors. Introduces spurious results and (likely) multicollinearity.

# Model selection

- MASS::stepAIC
- Lowest AIC value = most parsimonious
  - Doesn't mean it is a good model

```
model <- dat %>%
```

```
  lm(response ~ - [all predictors being considered], data = .)
```

```
stepAIC(model, direction = ["forward" or "reverse" or "both"])
```

# Model selection

AIC value  
with variable  
inclusion

Selected  
model

```
Step:  AIC=423.72
Life_expectancy ~ Alcohol + BMI + GDP + Adult_Mortality

              Df Sum of Sq    RSS    AIC
<none>                 2726.8 423.72
+ percentage_expenditure 1     10.96 2715.8 425.16
+ Population              1      0.00 2726.8 425.72
- GDP                    1    161.61 2888.4 429.72
- BMI                    1    280.86 3007.7 435.35
- Alcohol                1    468.79 3195.6 443.77
- Adult_Mortality        1   3037.60 5764.4 525.77

Call:
lm(formula = Life_expectancy ~ Alcohol + BMI + GDP + Adult_Mortality,
    data = lifeExp2014)

Coefficients:
(Intercept)      Alcohol          BMI          GDP  Adult_Mortality
  7.327e+01    5.298e-01    7.651e-02    7.259e-05   -4.820e-02
```

AIC value  
with variable  
exclusion

# Repeated measures ANOVA

- Used to analyze time series data
- Sometimes referred to as analysis of longitudinal data
- Example experimental design



| Unit | Treatment | Time<br>1 | Time<br>2 | Time<br>3 | Time<br>4 |
|------|-----------|-----------|-----------|-----------|-----------|
| 1    | A         |           |           |           |           |
| 2    | A         |           |           |           |           |
| 3    | A         |           |           |           |           |
| 4    | B         |           |           |           |           |
| 5    | B         |           |           |           |           |
| 6    | B         |           |           |           |           |
| 7    | C         |           |           |           |           |
| 8    | C         |           |           |           |           |
| 9    | C         |           |           |           |           |

# Repeated measures ANOVA

- Hypothetical research question
  - 9 woodlots
    - 3 with selective cutting
    - 3 with understory burning
    - 3 with no action
  - Which had greatest growth after four years?

| woodlot | treatment | year<br>1 | year<br>2 | year<br>3 | year<br>4 |
|---------|-----------|-----------|-----------|-----------|-----------|
| 1       | Cutting   | 42        | 54        | 57        | 68        |
| 2       | Cutting   | 38        | 42        | 49        | 54        |
| 3       | Cutting   | 51        | 55        | 57        | 63        |
| 4       | Burning   | 45        | 48        | 55        | 59        |
| 5       | Burning   | 39        | 43        | 48        | 51        |
| 6       | Burning   | 50        | 54        | 59        | 62        |
| 7       | None      | 48        | 50        | 47        | 46        |
| 8       | None      | 51        | 55        | 56        | 58        |
| 9       | None      | 41        | 44        | 43        | 48        |

# Repeated measures ANOVA

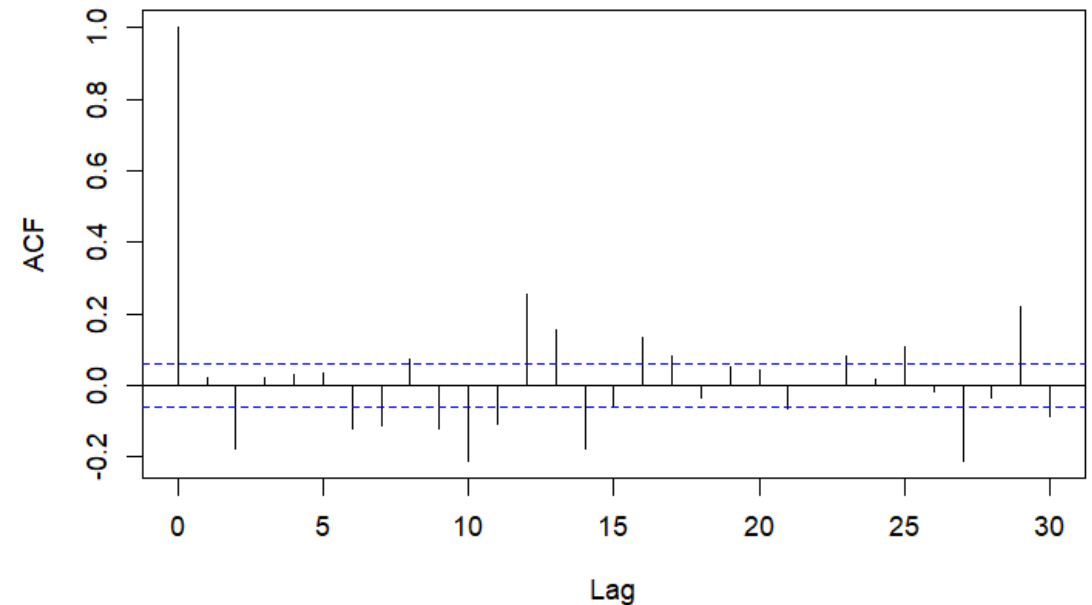
- Violates ANOVA assumption of independence of observations.
- Need to model that correlation structure within your model.

| woodlot | treatment | year<br>1 | year<br>2 | year<br>3 | year<br>4 |
|---------|-----------|-----------|-----------|-----------|-----------|
| 1       | Cutting   | 42        | 54        | 57        | 68        |
| 2       | Cutting   | 38        | 42        | 49        | 54        |
| 3       | Cutting   | 51        | 55        | 57        | 63        |
| 4       | Burning   | 45        | 48        | 55        | 59        |
| 5       | Burning   | 39        | 43        | 48        | 51        |
| 6       | Burning   | 50        | 54        | 59        | 62        |
| 7       | None      | 48        | 50        | 47        | 46        |
| 8       | None      | 51        | 55        | 56        | 58        |
| 9       | None      | 41        | 44        | 43        | 48        |

# Repeated measures ANOVA

- Evaluate the amount of temporal correlation present in your data using `stats::acf()`.
- Peaks outside blue dotted lines is evidence of temporal correlation

```
model <- lm(response ~ predictor1 + predictor2...,  
             data = dat)  
acf(residuals(model))
```





# Repeated measures ANOVA

- A generalized least squares model allows correlation structure to be modeled
  - nlme::gls()

```
model <- gls(response ~ predictor1 + predictor2...,  
             data = .,  
             correlation = [correlation structure function])
```

# Repeated measures ANOVA

- Modeling correlation structures
  - See `?nlme::corClasses`
  - Correct correlation structure depends on characteristics of the data
  - `corARMA` is most adaptable so most broadly useful for repeated measures

# Repeated measures ANOVA

- [corARMA](#)
  - Autocorrelation-moving average correlation structure
  - Requires a specification of where correlation exists
    - 'form' argument
    - Syntax is a one-sided formula (e.g., `form = ~1|var`)
    - For repeated measures, 'var' is the variable that is correlated through time
  - Requires fitting two parameters
    - $p$  is the autoregressive order (i.e., how strong is the temporal correlation)
    - $q$  is the moving average of the autoregressive order (i.e., how much does the temporal correlation vary throughout the dataset?)
    - These parameters need to be optimized

# Repeated measures ANOVA

- Optimizing corARMA's  $p$  and  $q$  parameters
  - Generate a set of candidate models using different values for  $p$  and  $q$ 
    - Optimal values usually range from 0-5, but that isn't necessarily a limit
  - Use AIC() to compare the candidate models and choose the parameterization with the lowest AIC value

# Repeated measures ANOVA

- Run model with optimized parameters and view results using `summary()`

```
Generalized least squares fit by REML
Model: mass ~ measurementInterval + pop
Data: gobyDat
      AIC      BIC    logLik
1128.203 1163.076 -557.1013

Correlation Structure: ARMA(2,1)
Formula: ~1 | ind
Parameter estimate(s):
      Phi1      Phi2      Theta1
1.76936549 -0.77918590 -0.03140105

Coefficients:
              Value Std.Error   t-value p-value
(Intercept)   9.141783  0.8158344  11.205440  0.0000
measurementInterval -0.054650  0.0137763  -3.966937  0.0001
popMSK         2.153335  1.1107884   1.938565  0.0528

Correlation:
              (Intr) msrmnI
measurementInterval -0.270
popMSK              -0.681  0.000

Standardized residuals:
      Min      Q1      Med      Q3      Max
-1.7348940 -1.0292700 -0.7551568 -0.2498638  3.4339898

Residual standard error: 5.014439
Degrees of freedom: 1080 total; 1077 residual
```

} Values pertaining to model selection

} Parameter estimates describing correlation structure

} **Model parameter estimates**

} Correlation measurements

} Distribution of residuals and standard errors