R Programming For Natural Resource Professionals

Week 12-13

ANOVA and regression in R

The week ahead…

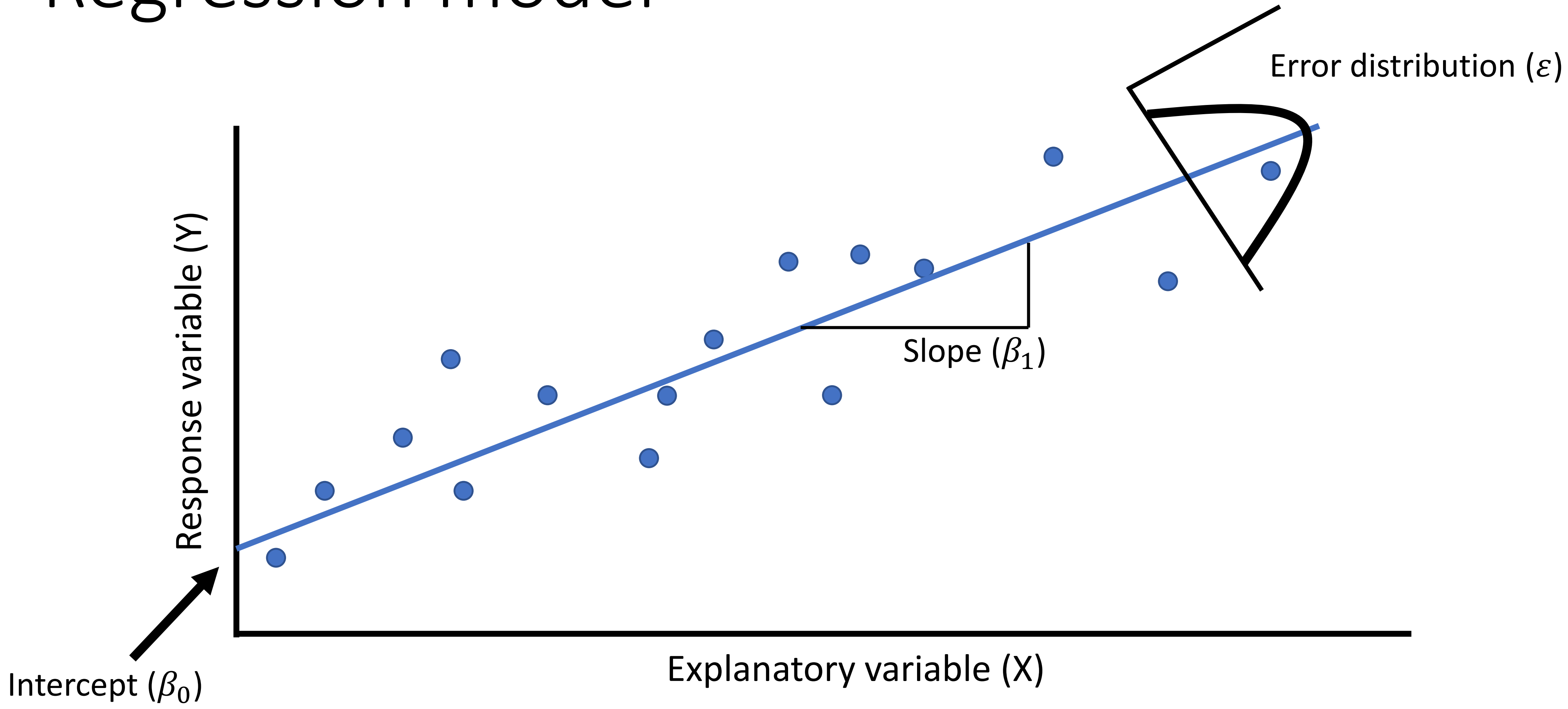
A(nother) syllabus change

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



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

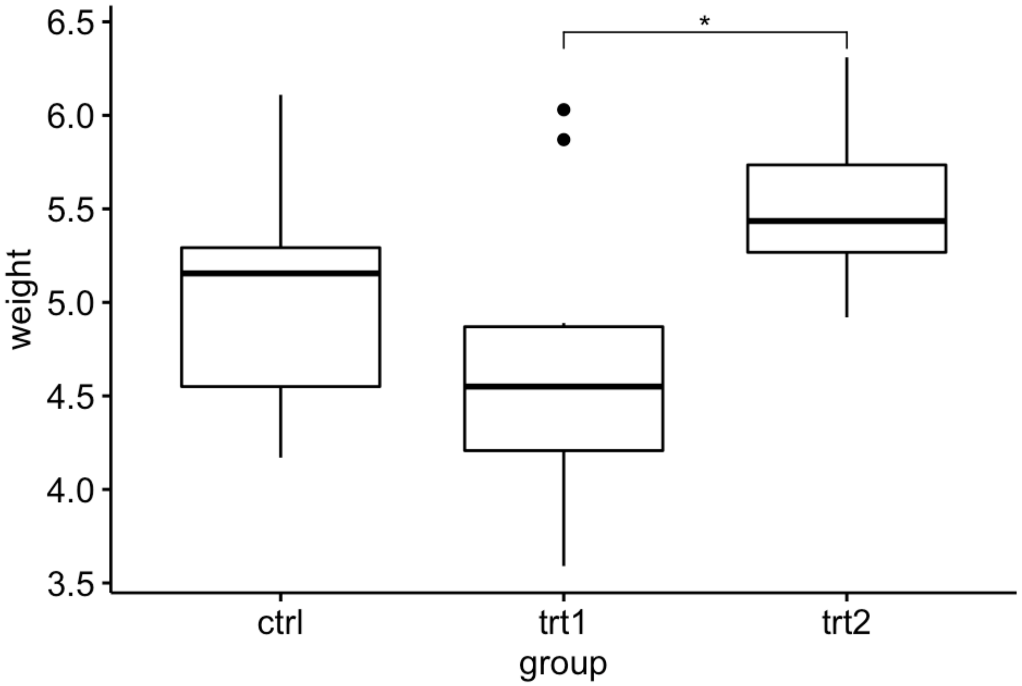
Goals of constructing models

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()
```

Degrees of Freedom

Sum of squares

Mean sum of squares (per df)

Test statistic

```
  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.

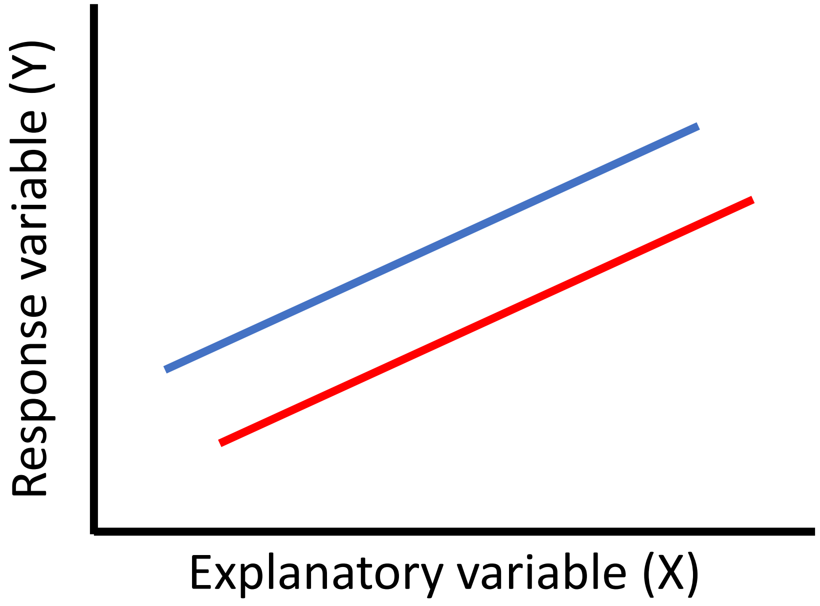
Tukey Honest Significant Differences: Evaluates combinations of categories within variables.

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

Modeling interactions

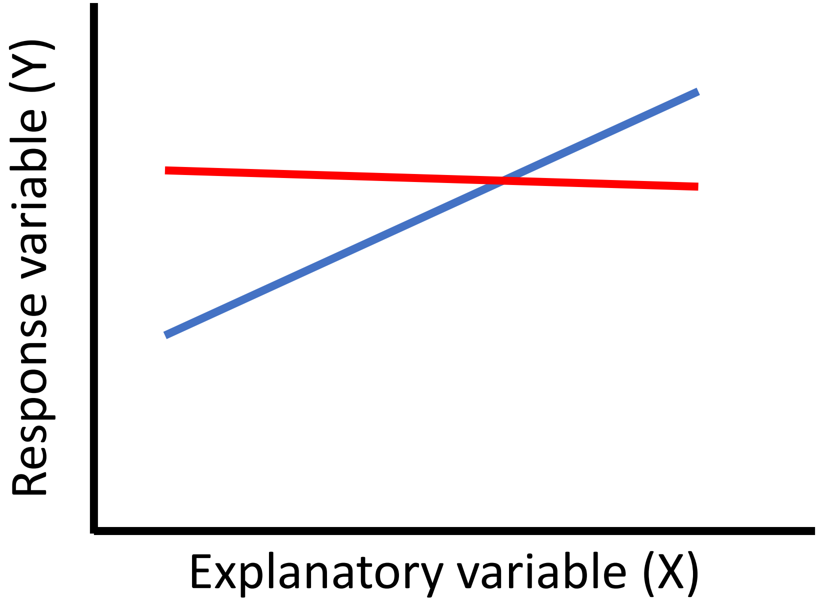
Additive model: Slopes are fixed among predictor variables. They affect the response variable in the same way.

- 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 <u>residuals</u> instead of <u>raw data</u>

- 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

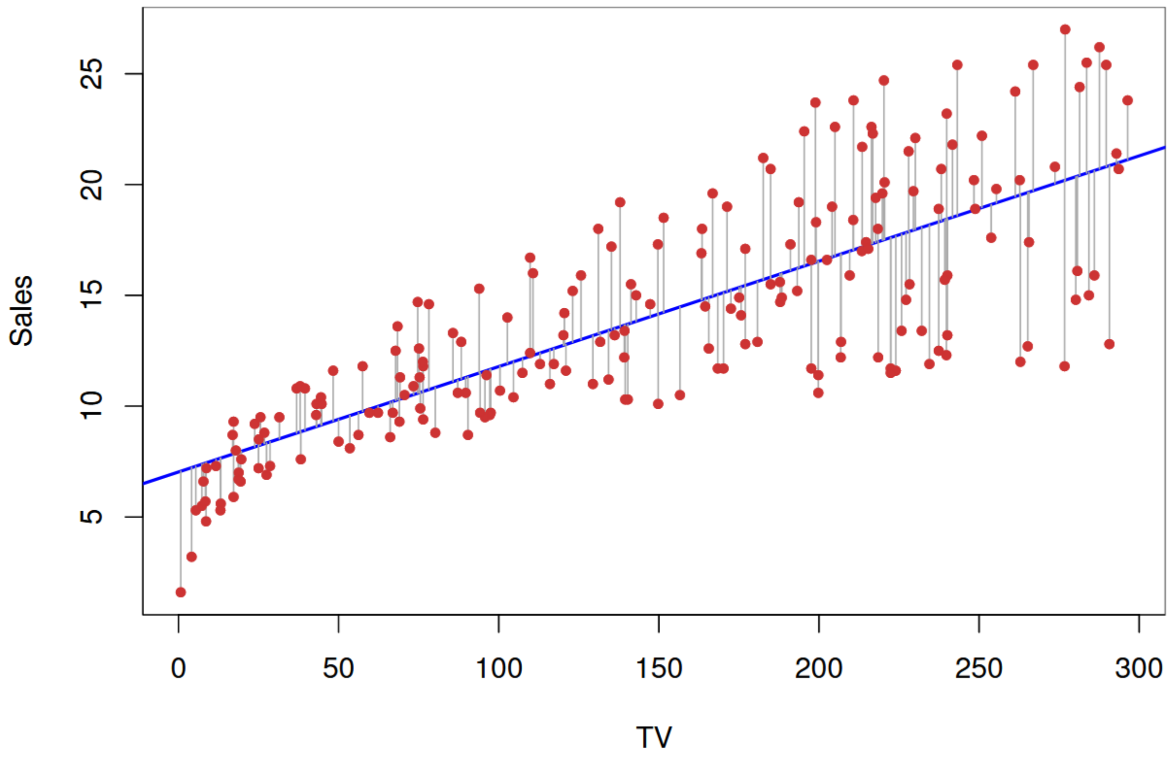
F-statistic comparing model performance vs. noise

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
```

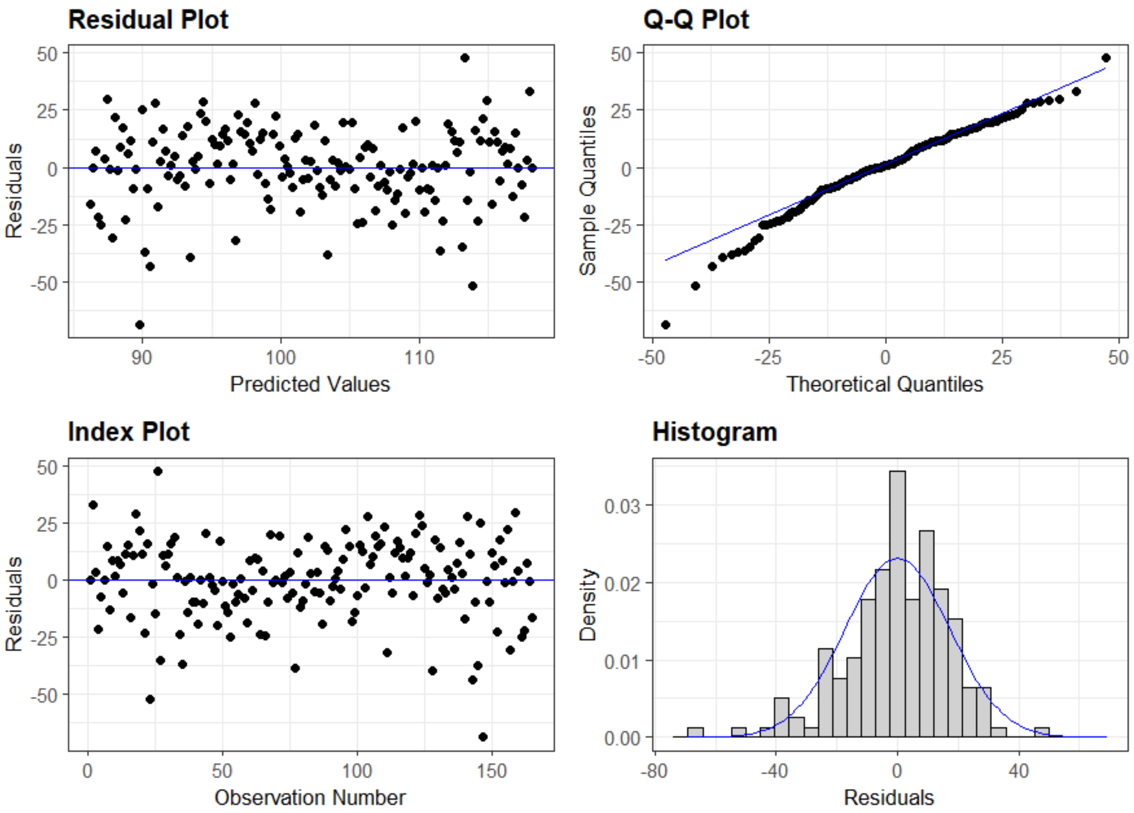
Simple linear regression

- 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 (ε).



Simple linear regression

- Plot residuals to assess assumptions
  - Access residuals using broom::augment()
- Four generally useful plots
  - qq plot
  - Index plot
  - Residual plot
  - Histogram

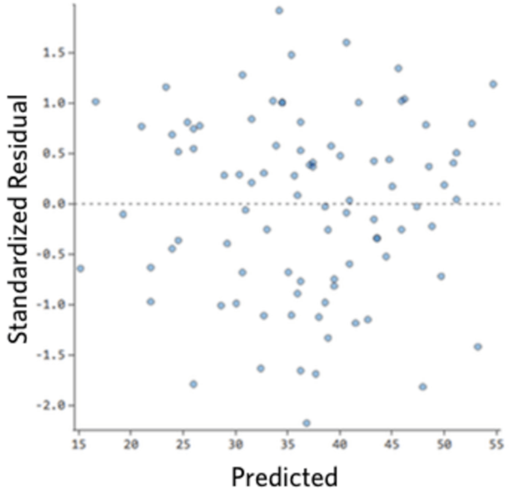


- Can be made all at once using ggResidpanel::resid_panel()
  - model %>% resid_panel()

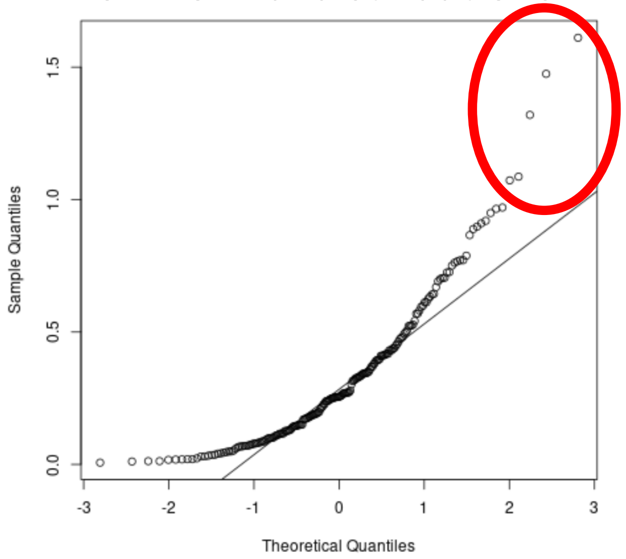
Simple linear regression

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

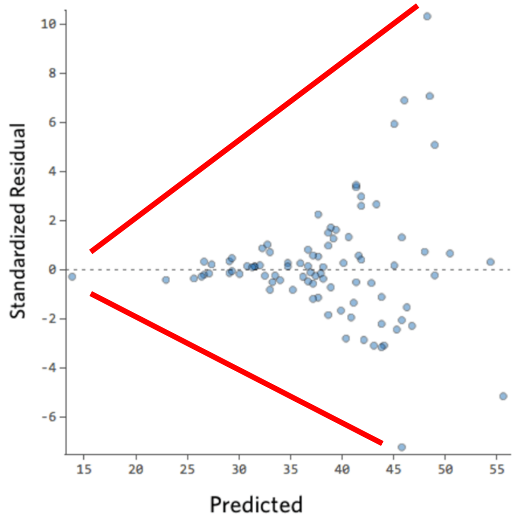
No obvious issues



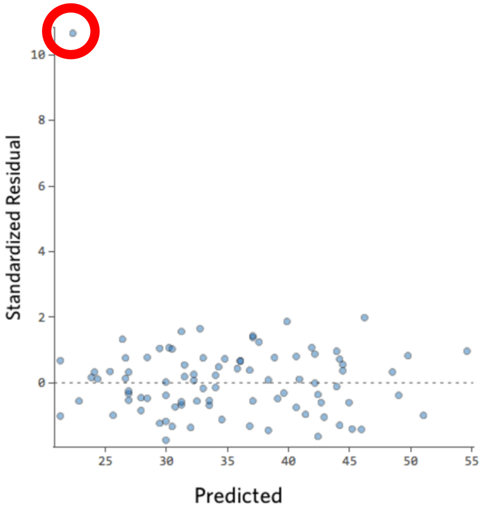
Non-normal distributions



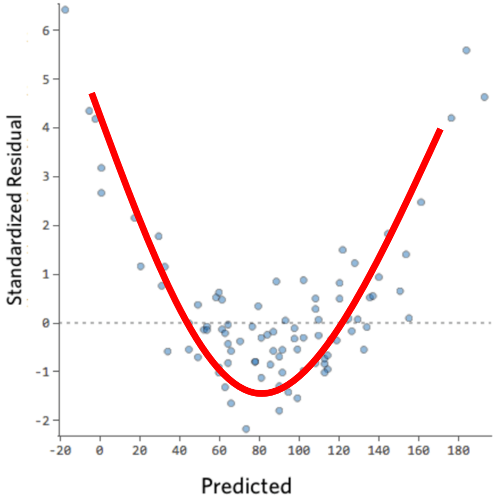
Heterogeneous variance



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)

Multiple linear regression

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

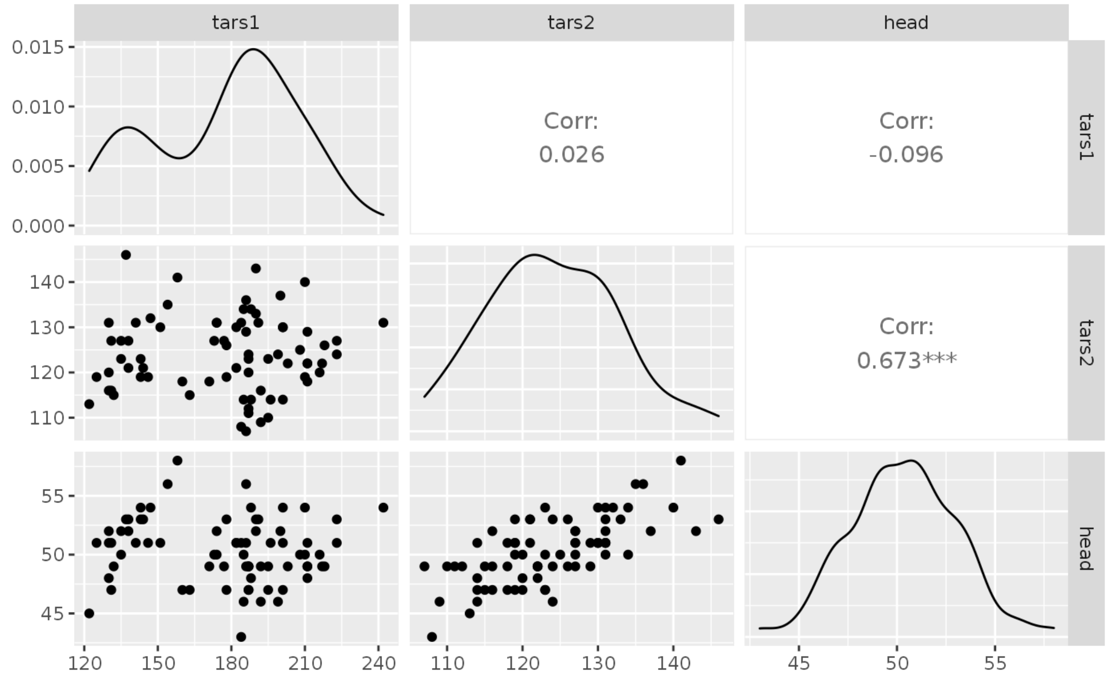
```
model <- dat %>%
    lm(response ~ predictor1 + predictor2 + predictor3…, data = .)
```

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>    <dbl>       <int> <int>
    0.384         0.380  35.5      81.2 4.08e-28     2 -1311. 2630. 2644.  328436.         260   263
```

Model selection

- Goal: choose the mix of predictor variables that most parsimoniously explain the response variable.
- Parsimony: the scientific principle that things are usually connected or behave in the simplest or most economical way, especially with reference to alternative evolutionary pathways.

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.[4]
- 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

- First, consider which predictor variables you have reason to believe will affect the response variable
  - 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

Selected model

AIC value with variable inclusion

AIC value with variable exclusion

```
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
```