Hypothesis testing and basic linear models
Week 4, Lecture 08

Richard E.W. Berl

Spring 2019

Regression

New day, new data set…

This one is listed as "S5 Table" at the following link: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0150798#sec020

Download the CSV and place it in your `/data` folder.

The source is:

Dinkins, J. B., et al. (2016). Microhabitat conditions in Wyoming's sage-grouse core areas: Effects on nest site selection and success. PLOS ONE, 11(3), e0150798. doi: 10.1371/journal.pone.0150798

Descriptions of the variables are scattered around the paper and in the four other identically-titled supporting tables.

```
grouse = read.csv("./data/journal.pone.0150798.s005.CSV", header=T)
```

```
str(grouse)
```

```
## 'data.frame':    1747 obs. of  53 variables:
##  $ Type              : Factor w/ 2 levels "Nest","Random": 1 1 1 1 1 1 1 1 1 1 ...
##  $ YEAR              : int  2008 2008 2008 2008 2008 2008 2008 2008 2008 2008 ...
##  $ UniqueID          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Fate              : Factor w/ 2 levels "fail","hatch": 1 2 1 2 1 2 2 1 2 2 ...
##  $ COX_TIME          : int  22 28 24 28 18 20 26 15 8 18 ...
##  $ annual_prec_30year: num  349 349 349 386 268 ...
##  $ CORE_AREA         : Factor w/ 2 levels "NO","YES": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Shrub_10M         : num  38.9 30.8 60.4 34.3 61 ...
##  $ ARTR_10M          : num  19.6 0 60.4 17 61 ...
##  $ Shrub_H_10M       : num  47.3 53.5 36.5 34 51.8 ...
##  $ ARTR_H_10M        : num  41.7 0 36.5 26.2 51.8 ...
##  $ VO                : num  45 6.38 40 35 65.5 ...
##  $ AnGrass_10M       : num  1.67 0 0.4 6.95 0 ...
##  $ PerGrass_10M      : num  26.9 23.7 12.7 17.2 10.6 ...
##  $ ResGrass_10M      : num  7.38 6.77 5.03 6.43 7.11 ...
##  $ FoodF_10M         : num  2.75 2.69 15.21 5.36 14.98 ...
##  $ NFoodF_10M        : num  4.172 3.683 5.017 0.678 3.406 ...
##  $ Bground_10M       : num  12.22 23.99 25.65 7.92 17.98 ...
##  $ Cactus_10M        : num  0 0 4.17 0 0 ...
##  $ BioCrust_10M      : num  0.4 8.194 0.678 0.831 3.344 ...
##  $ Rock_10M          : num  1.2 16.77 4.29 14.98 19.53 ...
##  $ Litter_10M        : num  38.9 48.7 53.1 61.5 42.6 ...
##  $ PerGrass_H_10M    : num  28.9 38.2 64.3 35.1 28.4 ...
##  $ ResGrass_H_10M    : num  15.8 21.3 15.7 26.7 17.2 ...
##  $ Shrub_5m          : num  49.1 39.2 65.8 45.1 70.9 ...
##  $ Artr_5m           : num  28.3 0 65.8 27.3 70.9 ...
##  $ Shrub_1m          : num  68.9 54.8 67.8 43.6 61 ...
##  $ Artr_1m           : num  53.3 0 67.8 14.3 61 ...
##  $ AnGrass_1M        : num  0 0 0.72 0 0 ...
##  $ PerGrass_1M       : num  34.2 28.6 18.6 14.1 14.8 ...
##  $ ResGrass_.1M      : num  11.85 7.85 7.84 0.94 11.85 ...
##  $ FoodF_1M          : num  1.33 4.23 24.25 6.63 1.22 ...
##  $ NFoodF_1M         : num  0 6.02 6.02 0.61 3.12 0.72 3.01 0.61 3.73 0.61 ...
##  $ BGround_1M        : num  10.2 29.1 31 10.6 18.6 ...
##  $ Cactus_1M         : num  0 0 7.51 0 0 ...
##  $ BioCrust_1M       : num  0.11 4.23 0.61 0.762 3.01 ...
##  $ Rock_1M           : num  0.83 16.54 7.73 19.68 20.13 ...
##  $ Litter_1M         : num  26.9 52.5 40.5 67.5 48 ...
##  $ AnGrass_3M        : num  3.76 0 0 15.64 0 ...
##  $ PerGrass_3M       : num  17.68 17.68 5.29 20.93 5.42 ...
##  $ ResGrass_3M       : num  1.8 5.42 1.52 13.29 1.18 ...
##  $ FoodF_3M          : num  4.525 0.762 3.9 3.763 32.175 ...
##  $ NFoodF_3M         : num  9.387 0.762 3.763 0.762 3.763 ...
##  $ Bground_3M        : num  14.68 17.68 18.91 4.53 17.16 ...
##  $ Cactus_3M         : num  0 0 0 0 0 ...
##  $ BioCrust_3M       : num  0.762 13.15 0.762 0.9 3.763 ...
##  $ Rock_3M           : num  1.66 17.05 0 10.29 18.77 ...
##  $ Litter_3M         : num  53.9 43.8 68.8 53.9 35.8 ...
##  $ Gap_0.5M          : int  6 6 5 6 5 6 6 3 3 5 ...
##  $ Gap_1M            : int  2 1 1 1 1 2 5 1 1 2 ...
##  $ Gap_2M            : int  1 1 1 1 0 0 1 0 0 0 ...
##  $ Gap_3M            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ X                 : logi  NA NA NA NA NA NA ...
```

For these analyses, we'll convert YEAR to a factor (categorical variable).

```
grouse$YEAR = as.factor(grouse$YEAR)
```

Logistic regression

```
?glm
```

Does one variable predict another (positively or negatively), when the outcome is nominal?

"Does the amount of litter within 10 meters of a site predict whether a greater sage-grouse chooses to nest there?"

```
grouseGLM = glm(Type ~ Litter_10M, data=grouse, family="binomial")
summary(grouseGLM)
```

```
##
## Call:
## glm(formula = Type ~ Litter_10M, family = "binomial", data = grouse)
##
## Deviance Residuals:
##    Min       1Q   Median       3Q      Max
## -1.303   -1.124   -1.018    1.216    1.467
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.399534   0.165774    2.41 0.015948 *
## Litter_10M  -0.013329   0.004038   -3.30 0.000965 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2415.1  on 1746  degrees of freedom
## Residual deviance: 2404.1  on 1745  degrees of freedom
## AIC: 2408.1
##
## Number of Fisher Scoring iterations: 4
```

Multiple logistic regression

Do any of these variables predict another (positively or negatively), when the outcome is nominal?

```
grouseGLM2 = glm(Fate ~ annual_prec_30year + Rock_3M + Cactus_3M + BioCrust_3M,
                data=grouse[grouse$Type == "Nest",], family="binomial")
summary(grouseGLM2)
```

```
##
## Call:
## glm(formula = Fate ~ annual_prec_30year + Rock_3M + Cactus_3M +
##     BioCrust_3M, family = "binomial", data = grouse[grouse$Type ==
##     "Nest", ])
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.430  -1.129  -1.018   1.220   1.405
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          0.0941507  0.2658020   0.354   0.7232
## annual_prec_30year -0.0009114  0.0007560  -1.206   0.2280
## Rock_3M              0.0168827  0.0076407   2.210   0.0271 *
## Cactus_3M            0.0124873  0.0354564   0.352   0.7247
## BioCrust_3M         -0.0075332  0.0141608  -0.532   0.5947
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1280.7  on 924  degrees of freedom
## Residual deviance: 1273.7  on 920  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 1283.7
##
## Number of Fisher Scoring iterations: 4
```

Mixed effects model

Do any of these variables predict another (positively or negatively), when I also have one or more variables that describe a subset of the data I could have collected?

```
install.packages("lme4")
```

```
library(lme4)
```

```
## Loading required package: Matrix
```

```
?glmer
```

```
grouseME = glmer(Type ~ Litter_10M + Rock_10M + (1 | YEAR),
                 data=grouse, family="binomial")
summary(grouseME)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Type ~ Litter_10M + Rock_10M + (1 | YEAR)
##    Data: grouse
##
##      AIC      BIC   logLik deviance df.resid
##   2398.4   2420.2  -1195.2   2390.4     1743
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.1829 -0.9623 -0.6985  1.0081  1.7305
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  YEAR   (Intercept) 0.05516  0.2349
## Number of obs: 1747, groups:  YEAR, 7
##
## Fixed effects:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.396723   0.230900   1.718  0.08577 .
## Litter_10M  -0.012274   0.004345  -2.825  0.00473 **
## Rock_10M    -0.003847   0.008646  -0.445  0.65631
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##            (Intr) Lt_10M
## Litter_10M -0.843
## Rock_10M   -0.562  0.323
```

Analysis of variance (ANOVA)

One-way (1 Nominal Variable and 1 Measurement Variable)

Do the means of a variable differ by group?

At sites with greater sage-grouse nests that hatched, was the mean percentage of shrub cover within 10 meters different each year?

```
grouseAOV = aov(Shrub_10M ~ YEAR,
                data=grouse[grouse$Type=="Nest" & grouse$Fate=="hatch",])
anova(grouseAOV)  # summary() works fine too
```

```
## Analysis of Variance Table
##
## Response: Shrub_10M
##            Df Sum Sq Mean Sq F value    Pr(>F)
## YEAR        6  15405 2567.54  12.224 1.004e-12 ***
## Residuals 436  91578  210.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

But this didn't really answer our question. Year matters, but which years are different?

Multiple comparisons (or "post hoc" tests)

```
TukeyHSD(grouseAOV)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Shrub_10M ~ YEAR, data = grouse[grouse$Type == "Nest" & grouse$Fate == "hatch", ]
##
## $YEAR
##                 diff        lwr        upr     p adj
## 2009-2008   2.7579304  -6.432867 11.9487275 0.9741339
## 2010-2008  -2.8070536 -12.467867  6.8537600 0.9780228
## 2011-2008  -4.1223571 -12.453266  4.2085521 0.7650092
## 2012-2008 -14.7485280 -23.511609 -5.9854467 0.0000184
## 2013-2008 -12.1394848 -20.435750 -3.8432193 0.0003626
## 2014-2008 -11.3941530 -19.473315 -3.3149913 0.0007016
## 2010-2009  -5.5649840 -14.819829  3.6898613 0.5615406
## 2011-2009  -6.8802874 -14.736806  0.9762314 0.1302659
## 2012-2009 -17.5064583 -25.819847 -9.1930695 0.0000000
## 2013-2009 -14.8974151 -22.717189 -7.0776414 0.0000006
## 2014-2009 -14.1520833 -21.741134 -6.5630324 0.0000012
## 2011-2010  -1.3153035  -9.716819  7.0862116 0.9992513
## 2012-2010 -11.9414744 -20.771707 -3.1112421 0.0014072
## 2013-2010  -9.3324311 -17.699595 -0.9652674 0.0177261
## 2014-2010  -8.5870994 -16.739048 -0.4351509 0.0314754
## 2012-2011 -10.6261709 -17.977751 -3.2745909 0.0004540
## 2013-2011  -8.0171277 -14.805512 -1.2287438 0.0092588
## 2014-2011  -7.2717959 -13.793068 -0.7505238 0.0177729
## 2013-2012   2.6090432  -4.703255  9.9213412 0.9400773
## 2014-2012   3.3543750  -3.710647 10.4193970 0.7983131
## 2014-2013   0.7453318  -5.731624  7.2222881 0.9998742
```

Notice the *p*-values have been "adjusted." We'll come back to why soon.

Two-way (2 Nominal Variables and 1 Measurement Variable)

Does the amount of bare ground cover differ by core/non-core area or by nesting/non-nesting site?

```
grouseAOV2 = aov(Bground_10M ~ CORE_AREA + Type,
                 data=grouse)
summary(grouseAOV2)
```

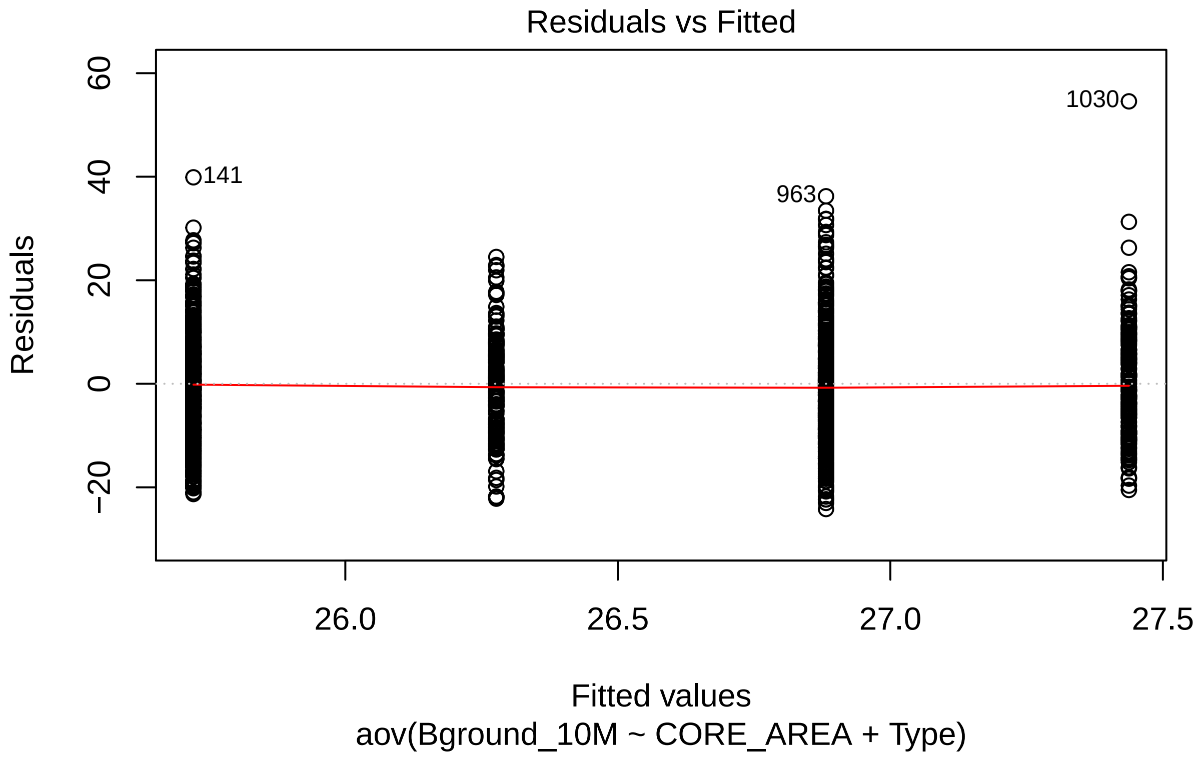
```
##                 Df Sum Sq Mean Sq F value Pr(>F)
## CORE_AREA        1    152   152.1   1.563 0.2115
## Type             1    579   578.9   5.946 0.0149 *
## Residuals     1744 169804    97.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

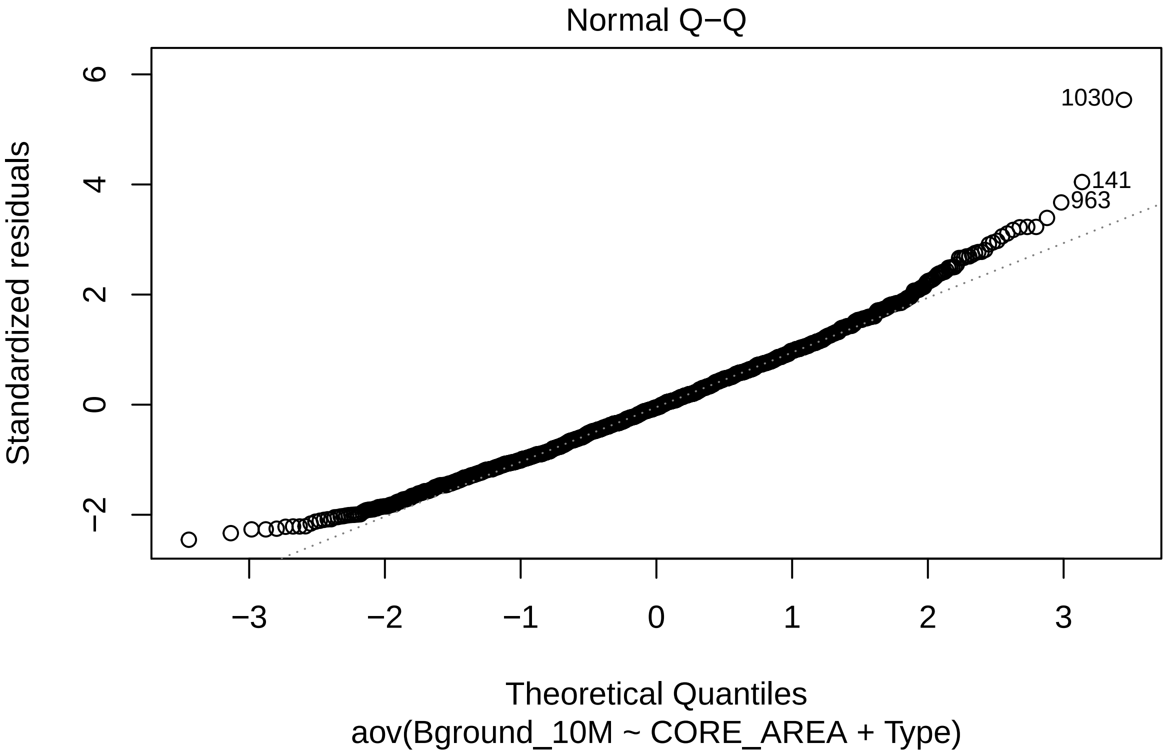
Nested ANOVA is a different type of analysis, when one variable is nested within another (e.g. plots and subplots within each plot). Look into it for more detail if you have this kind of data.

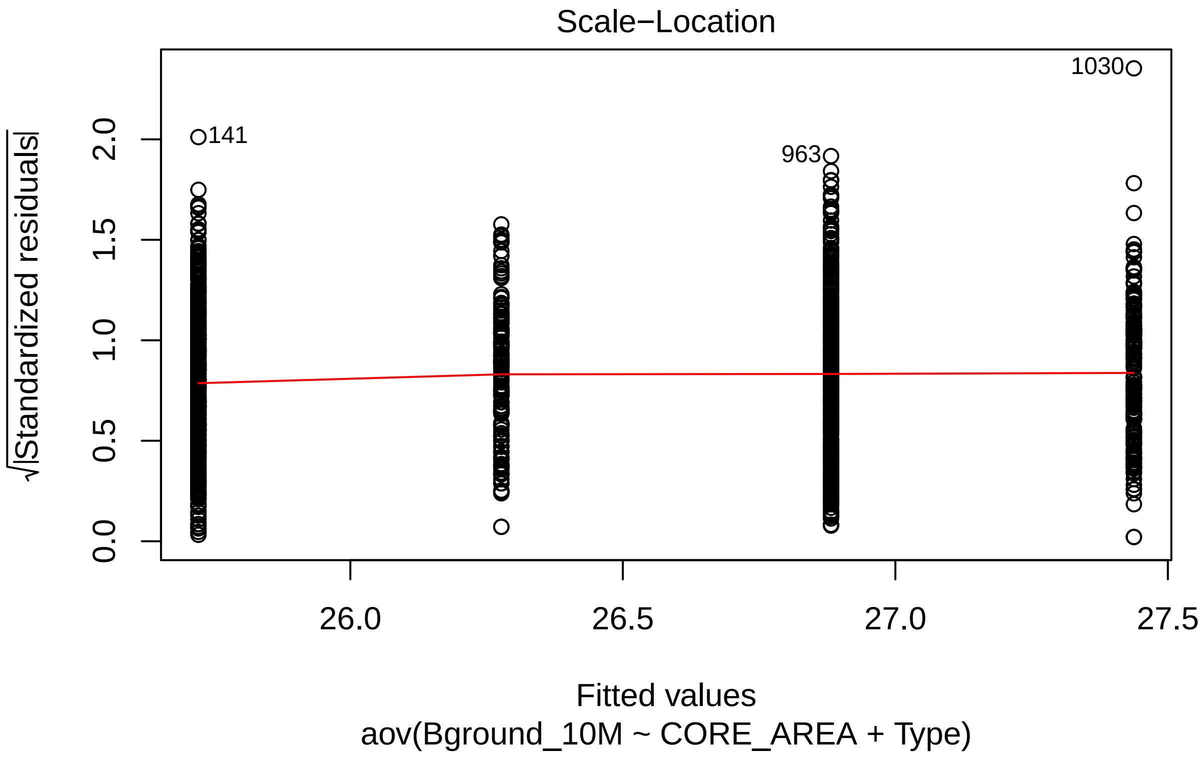
Quantile-quantile plots

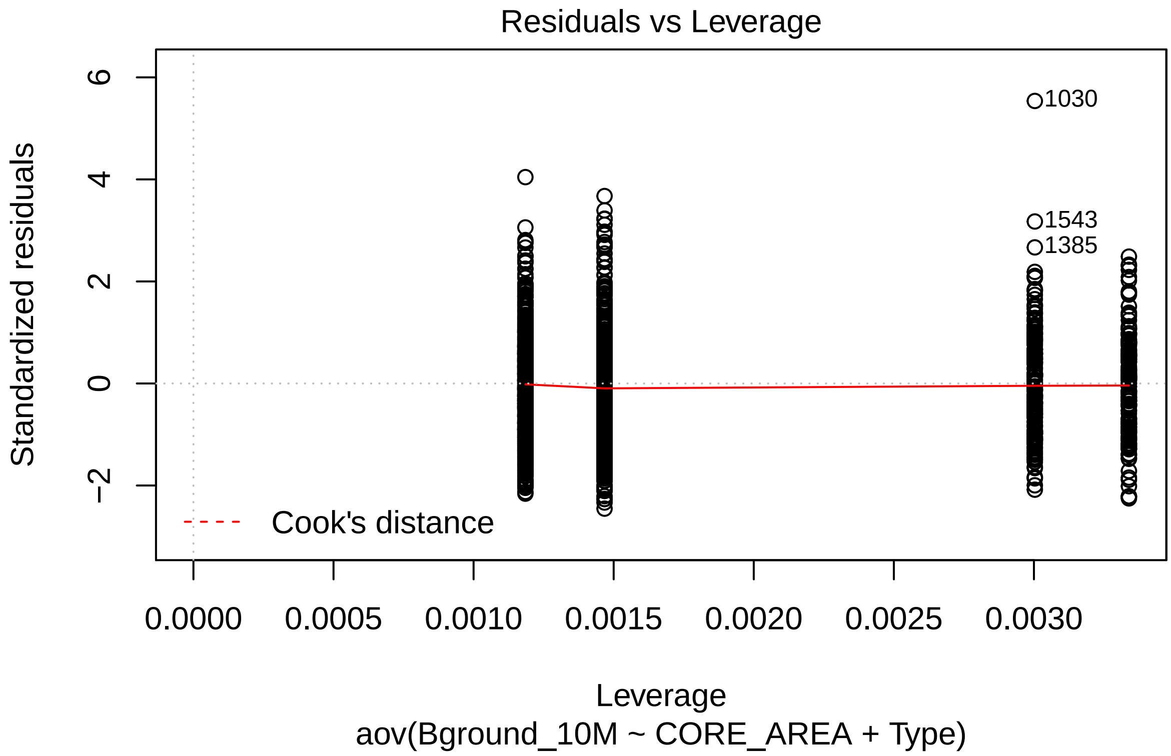
Check for normality, one of the assumptions of ANOVA.

```
plot(grouseAOV2)
```




Normal Q−Q
Standardized residuals
6
4
2
0
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
aov(Bground_10M ~ CORE_AREA + Type)
1030
141
963


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
141
963
1030
26.0
26.5
27.0
27.5
Fitted values
aov(Bground_10M ~ CORE_AREA + Type)



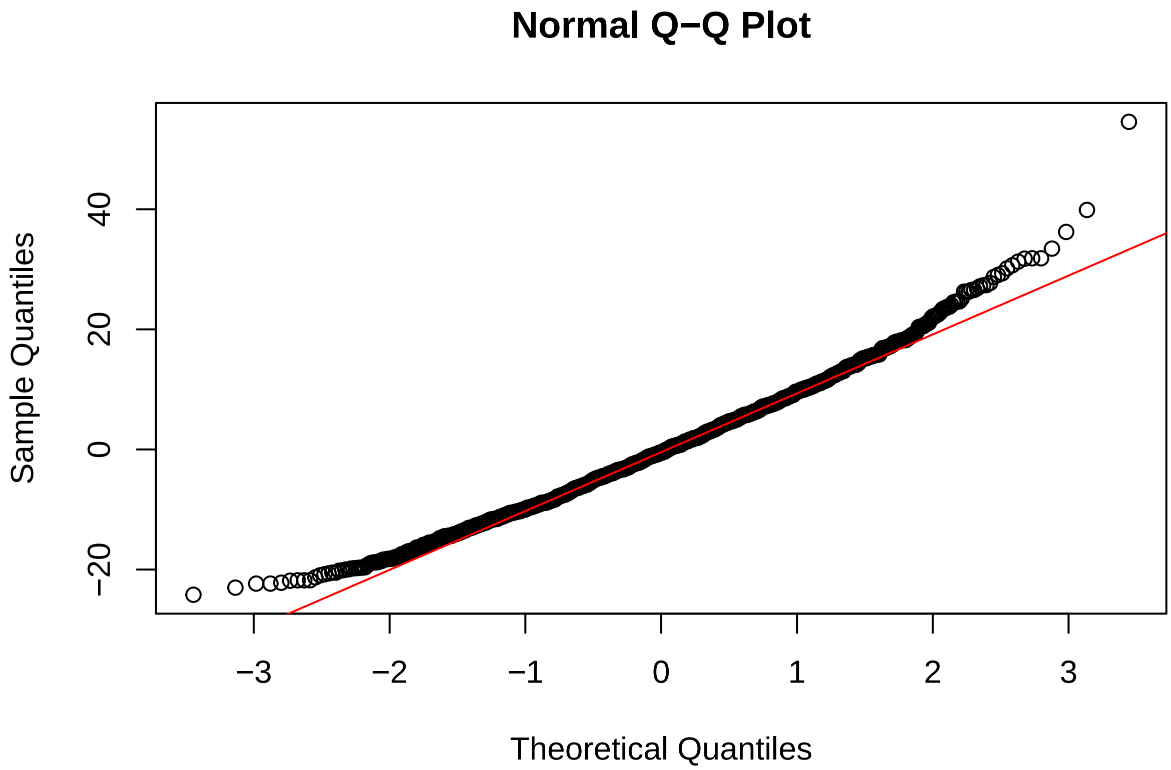
If we just want the Q-Q plot, we can also use `qqnorm()`.

```
?qqnorm
```

```
names(grouseAOV2)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "contrasts"     "xlevels"       "call"          "terms"
## [13] "model"
```

```
qqnorm(grouseAOV2$residuals)
qqline(grouseAOV2$residuals, col="red")
```



Correcting for multiple comparisons

```
myT1 = t.test(grouse$ARTR_10M[grouse$Fate == "hatch"],
              grouse$ARTR_10M[grouse$Fate == "fail"])
myT1
```

```
##
##  Welch Two Sample t-test
##
## data:  grouse$ARTR_10M[grouse$Fate == "hatch"] and grouse$ARTR_10M[grouse$Fate == "fail"]
## t = -1.0064, df = 915.69, p-value = 0.3145
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.9666624  0.9553825
## sample estimates:
## mean of x mean of y
##  28.73331  29.73895
```

```
myT2 = t.test(grouse$Shrub_H_10M[grouse$Fate == "hatch"],
              grouse$Shrub_H_10M[grouse$Fate == "fail"])
myT2
```

```
##
##  Welch Two Sample t-test
##
## data:  grouse$Shrub_H_10M[grouse$Fate == "hatch"] and grouse$Shrub_H_10M[grouse$Fate == "fail"]
## t = -1.3535, df = 922.3, p-value = 0.1762
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.310085  0.607891
## sample estimates:
## mean of x mean of y
##  39.15406  40.50515
```

```
myT3 = t.test(grouse$FoodF_10M[grouse$Fate == "hatch"],
              grouse$FoodF_10M[grouse$Fate == "fail"])
myT3
```

```
##
##  Welch Two Sample t-test
##
## data:  grouse$FoodF_10M[grouse$Fate == "hatch"] and grouse$FoodF_10M[grouse$Fate == "fail"]
## t = -1.2016, df = 914.83, p-value = 0.2298
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.7924151  0.1905721
## sample estimates:
## mean of x mean of y
##  5.032824  5.333745
```

```
myT4 = t.test(grouse$Rock_10M[grouse$Fate == "hatch"],
              grouse$Rock_10M[grouse$Fate == "fail"])
myT4
```

```
##
##  Welch Two Sample t-test
##
## data:  grouse$Rock_10M[grouse$Fate == "hatch"] and grouse$Rock_10M[grouse$Fate == "fail"]
## t = 0.96158, df = 913.23, p-value = 0.3365
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4000636  1.1686939
## sample estimates:
## mean of x mean of y
##  8.663488  8.279172
```

```
myT5 = t.test(grouse$BioCrust_10M[grouse$Fate == "hatch"],
              grouse$BioCrust_10M[grouse$Fate == "fail"])
myT5
```

```
##
##  Welch Two Sample t-test
##
## data:  grouse$BioCrust_10M[grouse$Fate == "hatch"] and grouse$BioCrust_10M[grouse$Fate == "fail"]
## t = -0.066447, df = 917.71, p-value = 0.947
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.4730400  0.4420572
## sample estimates:
## mean of x mean of y
##  3.297912  3.313403
```

If the probability of a false positive is 0.05 for each one of these tests…

Then the probably of *at least one* false positive in five tests is…

```
1 - (1 - 0.05)^5
```

```
## [1] 0.2262191
```

To correct it, we need to adjust our *p*-values.

```
myPVals = c(myT1$p.value, myT2$p.value, myT3$p.value, myT4$p.value,
            myT5$p.value)
myPVals
```

```
## [1] 0.3144759 0.1762128 0.2298311 0.3365148 0.9470365
```

```
?p.adjust
```

```
p.adjust(myPVals, method="bonferroni")
```

```
## [1] 1.0000000 0.8810638 1.0000000 1.0000000 1.0000000
```

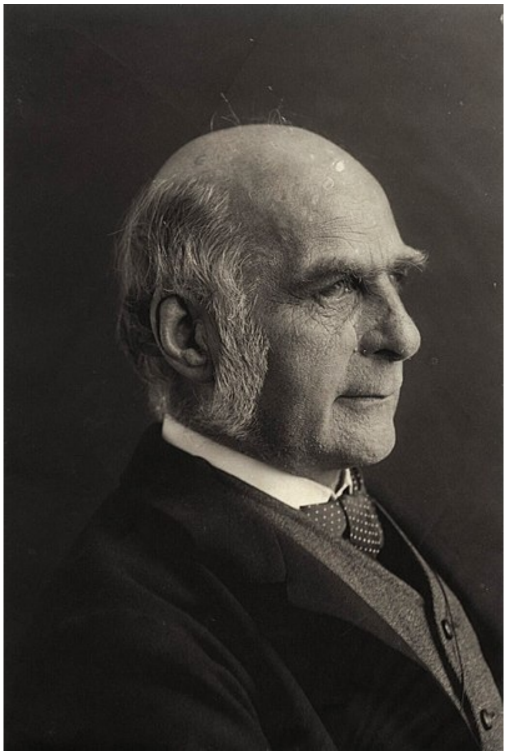
```
p.adjust(myPVals, method="fdr")
```

```
## [1] 0.4206435 0.4206435 0.4206435 0.4206435 0.9470365
```

And now a quick diversion into the *alternative* history of statistics, where we talk about how…

Galton, Pearson, and Fisher were terrible people

Sir Francis Galton



Source: National Portrait Gallery, London

Charles Darwin's half-cousin

Contributions:

- Correlation
- Regression
- Standard deviation
- Questionnaires

and…

- *Eugenics*

"My proposal is to make the encouragement of the Chinese settlements at one or more suitable places on the East Coast of Africa a par of our national policy, in the belief that the Chinese immigrants would not only maintain their position, but that they would multiply and their descendants supplant the inferior Negro race. I should expect the large part of the African seaboard, now sparsely occupied by lazy, palavering savages… might in a few years be tenanted by industrious, order loving Chinese… average negroes possess too little intellect, self-reliance, and self-control to make it possible for them to sustain the burden of any respectable form of civilization without a large measure of external guidance and support."

Source: *Africa for the Chinese: To the Editor of The Times (1873)*

It's worth thinking about how these ideas continue in our own field (using different language) in the thinking and intent behind both local and international conservation practices.

Karl Pearson

Galton Professor of Eugenics, University College London

Source: National Portrait Gallery, London

Contributions:

- Correlation coefficient

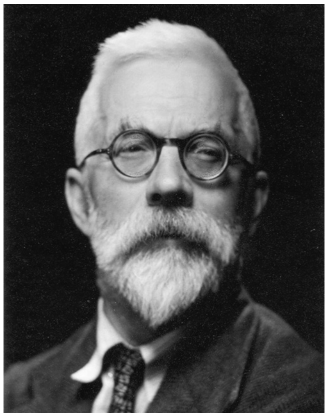
- *p*-value
- Hypothesis testing
- Chi-squared test
- Principal component analysis
- Histogram

"History shows me one way, and one way only, in which a high state of civilization has been produced, namely, the struggle of race with race, and the survival of the physically and mentally fitter race. If you want to know whether the lower races of man can evolve a higher type, I fear the only course is to leave them to fight it out among themselves, and even then the struggle for existence between individual and individual, between tribe and tribe, may not be supported by that physical selection due to a particular climate on which probably so much of the Aryan's success depended."

Source: *National Life from the Stand-point of Science: An Address Delivered at Newcastle (1901)*

Sir Ronald A. Fisher

Galton Professor of Eugenics, University College London



Source: University of Adelaide Library

Contributions:

- Null hypothesis
- Analysis of variance (ANOVA)
- Maximum likelihood
- Experimental randomization
- Modern synthesis of biology

"In one respect the theory of selection by climate and disease appears to possess an advantage over that of race mixture. If the latter were the only agency at work, the disappearance of the ruling class would be accompanied by a permanent improvement of the natives. The effect of successive conquests should accumulate; so that we should expect that a people, such as the Egyptians, should be reasonably far advanced towards the type of a ruling race. The reverse appears to be the case. The effect of the selective influence of climate and disease, on the other hand, would appear to undo completely the racial benefits of an invasion."

Source: *The Genetical Theory of Natural Selection (1930)*

IMPORTANT NOTE:

I vehemently oppose the abhorrent notions expressed in the above quotations. They represent the worst misuses of social science–social Darwinism and scientific racism–and have been used to justify centuries of colonialism and oppression. The excuse that they were "products of their time" is inadequate, because they were among the most prominent proponents of these ideas and lent their authority and legitimacy to them. Moreover, they developed and incorrectly applied a number of their methods specifically in attempts to support their biased views on race (and see the "rejoinder" at the end of Galton's letter for evidence that not everyone thought this way, even in Galton's time). Their advancement of racist "science" directly influenced the rise of white supremacism in Nazi Germany and elsewhere, and its legacy continues today with the recent resurgence of so-called "race realism." These men have done enduring damage to science and to generations of people and do not deserve to be celebrated.

(pdf / Rmd)