ST 512 Project

Body Sub-site Predicts Microbiome Diversity in Humans

Connor Draney, Sunni Patton, Kelly Shannon, Michael Sieler

06 March, 2022

Contents

Analysis Overview

Research Question

Can we predict alpha diversity scores based on sex, body sub-site or the interaction of these terms?

Hypothesis

The human body is comprised of multiple body sites with unique environmental conditions, we hypothesis these areas will manifest unique alpha diversity scores between body sub-sites. Furthermore, we also hypothesize that these areas will differ based on sex and the interaction of body sub-site and sex.

Statistical analysis

All statistical analyses and visualizations were conducted in R (v 4.0.5) unless otherwise specified. Data was obtained from the Human Microbiome Project (https://www.hmpdacc.org/micro_analysis/microbiome_analyses.php). After removal of samples of a read depth of less than 1,000 the minimum read depth of 1,003, NAs were removed at the Phylum level and 292 ASVs were identified. Additionally, we filtered for samples that were processed at JCCI, not repeated samples from an individual, and not marked as misslabeled or contaminated. Alpha-diversity was calculated using estimate_richness function (Phyloseq v 1.38.0). Alpha-scores were calculated using Shannon, Simpson, Phylogenetic, and Richness indices. Using case influence statistics, we removed any unusual observations that fell beyond standard cutoff values for Cook's Distance, leverage and standardized residuals. These data were then analyzed using linear models (LMs) to build models in order to determine the best predictors of alpha-diversity. Two-way ANOVA was used to assess the models.

Results

We obtained 16S rRNA gene sequence data to investigate how the human microbiome diversity may be influenced by body sub-site or sex. To assess diversity, we used linear models (LMs) to identify parameters, such as body sub-site, that best explained the variation in microbiome diversity, as measured by Shannon, Simpson, Phylogenetic and Richness Diversity Indices. The results from our ANOVA tests:

- Shannon:
  - The main effect of sex is statistically significant and small ($F(1, 488) = 12.31$, $p < .001$; 95% CI [6.98e-03, 1.00]). The main effect of HMPbodysubsite is statistically significant and large ($F(15, 488) = 96.72$, $p < .001$; CI [0.72, 1.00]). The interaction between sex and HMPbodysubsite is statistically significant and medium ($F(12, 488) = 2.74$, $p = 0.001$; 95% CI [0.01, 1.00]).
- Simpson:
  - The main effect of sex is statistically not significant and very small ($F(1, 479) = 2.19$, $p = 0.139$; 95% CI [0.00, 1.00]). The main effect of HMPbodysubsite is statistically significant and large ($F(15, 479) = 62.57$, $p < .001$; 95% CI [0.62, 1.00]). The interaction between sex and HMPbodysubsite is statistically not significant and small ($F(12, 479) = 0.92$, $p = 0.532$; 95% CI [0.00, 1.00])
- Phylogenetic:
  - The main effect of sex is statistically significant and very small ($F(1, 500) = 3.95$, $p = 0.047$; 95% CI [4.65e-05, 1.00]). The main effect of HMPbodysubsite is statistically significant and large ($F(15, 500) = 110.86$, $p < .001$; 95% CI [0.74, 1.00]). The interaction between sex and HMPbodysubsite is statistically significant and small ($F(12, 500) = 2.55$, $p = 0.003$; 95% CI [0.01, 1.00])

- Richness:
  - The main effect of sex is statistically significant and very small (F(1, 500) = 3.95, p = 0.047; 95% CI [4.65e-05, 1.00]). The main effect of HMPbodysubsite is statistically significant and large (F(15, 500) = 110.86, p < .001; 95% CI [0.74, 1.00]). The interaction between sex and HMPbodysubsite is statistically significant and small (F(12, 500) = 2.55, p = 0.003; 95% CI [0.01, 1.00])

Summary:

- Overall, our results suggest that the main effect of body sub-site and sex are statistically significant across all indics, except for in the Simposon's index sex is not statistically significant. Furthermore, the interaction of body sub-site and sex is statistically significant in all but Simpson's index.

Our data analysis and associated files can be found here: https://github.com/sielerjm/ST512_Project

Import Data

We obtained data in the form of a phyloseq data object from the Human Genome Project (HMP) that was previously processed for taxanomic identification of microbial organisms. We then cleaned and processed the data further for statistical analysis.

```
if(redo.analysis$redo.importData == T){

  # Download Data
  # temp_test = tempfile()
  # test_url = "http://joey711.github.io/phyloseq-demo/HMPv35.RData"
  # download.file(test_url, destfile = paste0(data.path, "/Raw/HMPv35.RData"))

  load(paste0(data.path, "/Raw/HMPv35.RData"))
  ps.unclean <- HMPv35
  rm(HMPv35)  # remove this obj from global env

  # Save
  save(ps.unclean, file = paste0(saveObj.path, "/ps-unclean.RData"))

} else {

  # Load
  load(file = paste0(saveObj.path, "/ps-unclean.RData"))

}
```

Clean sample data

The 4743 samples in the HMP were obtained from multiple individuals (2555 males and 2188 females), processed by several locations, and in some cases repeated samples were taken from the same individual. We subset the data to ensure independence by limiting samples from one location and only the first visit. Additionally, we removed samples noted as "Mislabeled" or "Contaminated". After removing samples using these criteria, we were left with 609 samples.

```
if(redo.analysis$redo.cleanData == T){

  # Find the center with largest counts for the following conditions
  sample.data.frame(ps.unclean) %>%
    count(RUNCENTER,
          Mislabeled,
          Contaminated,
          visitno) %>%
    arrange(-n)

  # Subset based on parameters
  ps.cleaned <- subset_samples(ps.unclean,  # physeq object
                               Mislabeled == F &  # Remove mislabeled
                               Contaminated == F &  # remove contaminated
                               visitno == 1 & # first visit, so no repeat sampling of individuals, ind
                               RUNCENTER == "JCVI"  # Same center, independence
                               )
```

```
  # Remove columns, if rownames are already sample names, no need to have an extra sample column
  sample_data(ps.cleaned) <- sample_data(ps.cleaned)[, c(-1)]  # [rows, cols]

  # Check to see how many samples dropped from original dataset
  print(paste0("Dropped ",
               nrow(sample_data(ps.unclean)) - nrow(sample_data(ps.cleaned)),
               " samples after cleaning.")
              )

  rm(ps.unclean)  # remove this obj from global env

  # Save
  save(ps.cleaned, file = paste0(saveObj.path, "/ps-cleaned.RData"))

} else {

  # Load
  load(file = paste0(saveObj.path, "/ps-cleaned.RData"))

}
```

Rarefaction

We rarefied the data to control for uneven sampling efforts Sanders, H. L. (1968), Willis, A.D. (2019).

```
if(redo.analysis$redo.rarefy == T){

  rarefaction.minimum <- 1000
  min.smpl.size <- min(sample_sums(ps.cleaned)[sample_sums(ps.cleaned) >= rarefaction.minimum])

  summary(sample_sums(ps.cleaned))

  # Rarefy

  ps.rar <- {
    ps.rar <- ps.cleaned
    ps.rar <- rarefy_even_depth(
        physeq = ps.rar,
        sample.size = min.smpl.size,
        trimOTUs = TRUE,
        rngseed = 42
        )
    rename.NA.taxa(ps.rar)
  }


  # Save
  save(ps.rar, file = paste0(saveObj.path, "/ps-rar.RData"))

} else {
```

```
  # Load
  load(file = paste0(saveObj.path, "/ps-rar.RData"))

}
```

```
# Make a list of phyloseq objects, dataframes and datatables
ps.list <- list(RAR = list(ps.all = ps.rar,
                           df.all = sample.data.frame(ps.rar),
                           dt.all = sample.data.table(ps.rar)
                           ))

# Add observation numbers to data
ps.list[["RAR"]][["df.all"]] <- dplyr::mutate(ps.list[["RAR"]][["df.all"]], obs_num = row_number(), .be
ps.list[["RAR"]][["dt.all"]] <- dplyr::mutate(ps.list[["RAR"]][["dt.all"]], obs_num = row_number(), .be


# head(ps.list$RAR$df.all)
```

Alpha-diversity

We calculated alpha diversity scores, a measure of number of unique organisms within a single sample, using the indices Shannon, Simpson, Phylogenetic, and Richness (Whittaker 1960). Each calculates alpha-diversity using slightly different mathematical approaches to measure evenness (distribution of organisms) and/or richness (number of organismss). If differences in results are seen between indices, this can reveal insights into which kinds of organisms are present (e.g., common vs rare).

```
# Select which alpha measures we want to analyze
methods.alpha <- c("Shannon", "Simpson", # Non-phylogenetic measures, add additional measures here
                   "Phylogenetic", "Richness") %>%
                  purrr::set_names()  # Set's names list elements in "alpha.methods"

if(redo.analysis$redo.alphaDiv == T){

  # Calculate alpha scores, save to list
  ps.list$RAR[["alphaScore"]] <- alpha_base(ps.list$RAR$ps.all,  # Phyloseq object
                                  methods.alpha,  # list of alpha methods
                                  "Sample",
                                  phylo.div = T
                                  )


  # Add alpha scores to data table
  ps.list$RAR[["dt.all.alpha"]] <- ps.list$RAR$dt.all[ps.list$RAR$alpha, on = "Sample"] %>% setkeyv("Sa

  # Melt data table
  ps.list$RAR[["dt.all.alpha.melt"]] <- melt_to_datatable_2(datatable1 = ps.list$RAR$dt.all,
                                                            datatable2 = ps.list$RAR$dt.all.alpha,
                                                            vars = methods.alpha,
                                                            var.name = "Alpha.Metric",
                                                            samp.name = "Sample",
                                                            val.name = "Alpha.Score"
```

```
                                                                    )
  # Save
  save(ps.list, file = paste0(saveObj.path, "/ps-list.RData"))

} else {

  # Load
  base::load(file = paste0(saveObj.path, "/ps-list.RData"))

}
```

Fit linear model

We fit the data using linear model to predict alpha-diversity score as a function of body sub-site, sex, or their interaction.

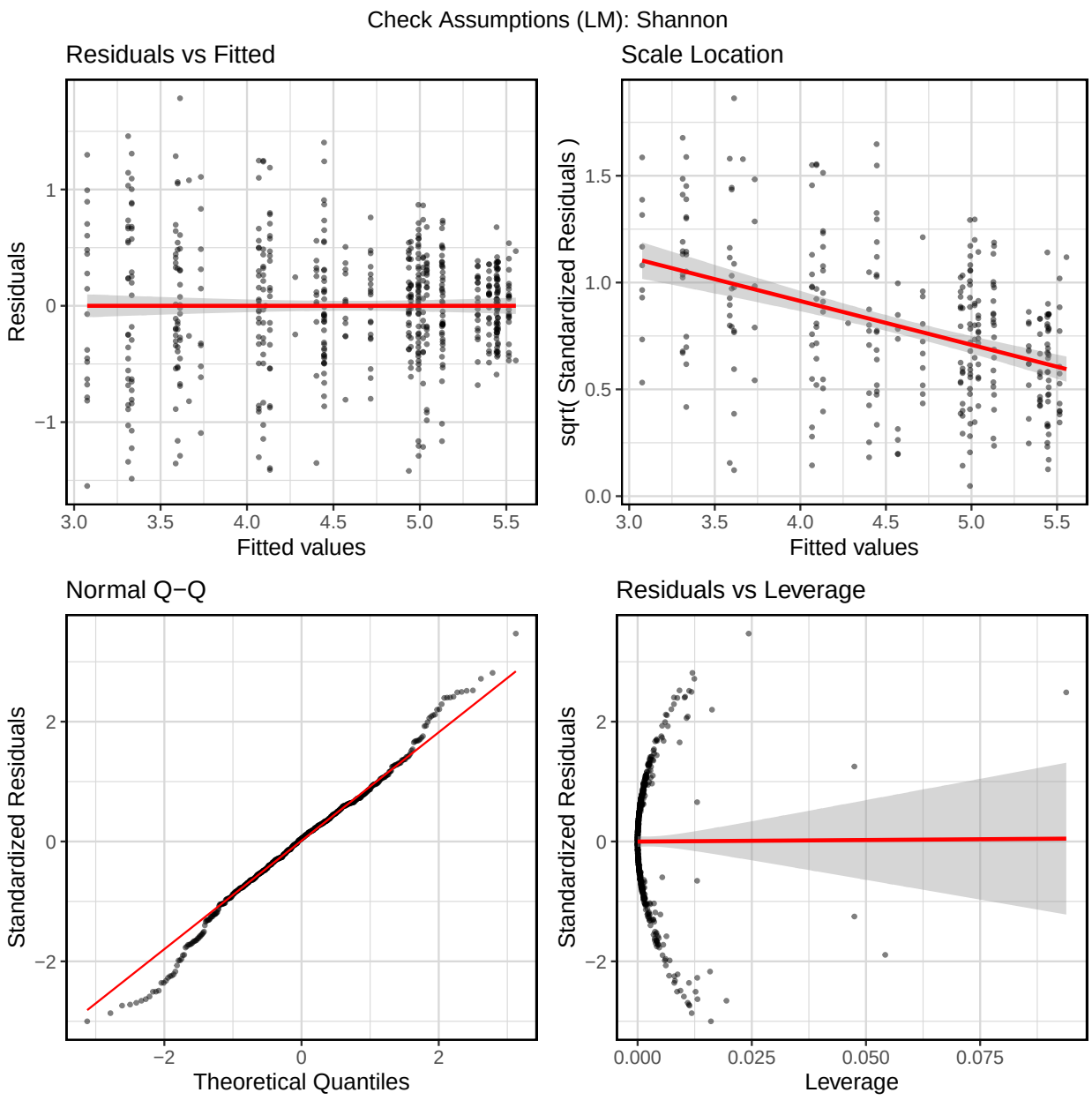
```
caseInfStats <- list()

data <- ps.list$RAR[["dt.all.alpha.melt"]]
# data <- dplyr::mutate(data, obs_num = row_number(), .before = 1)

caseInfStats[["mod.unref"]] <- lapply(methods.alpha, function(alpha){
  lm( formula = "Alpha.Score ~ sex*HMPbodysubsite",
       data = subset(data, Alpha.Metric == alpha)
      )
})
```

Check assumptions (unrefined model)

To assess if the assumptions of linear model regression were met, we visually inspected residuals vs fitted values, standardized residuals vs quantiles, standardized residuals vs fitted values and residuals vs leverage for each diversity index. In general, Residuals vs fitted appears slightly heteroscedasticity, Q-Q plots show that the data slightly deviates from the diagonal line indicating that the data may be non-normal, Scale-location plots show that std. residuals are negatively associated with fitted values indicating heteroscedasticity, and there are several points with high leverage, but none that appear to have too high. The plots associated with the Simpson's index appear to be significantly heteroscedastic and non-normal.

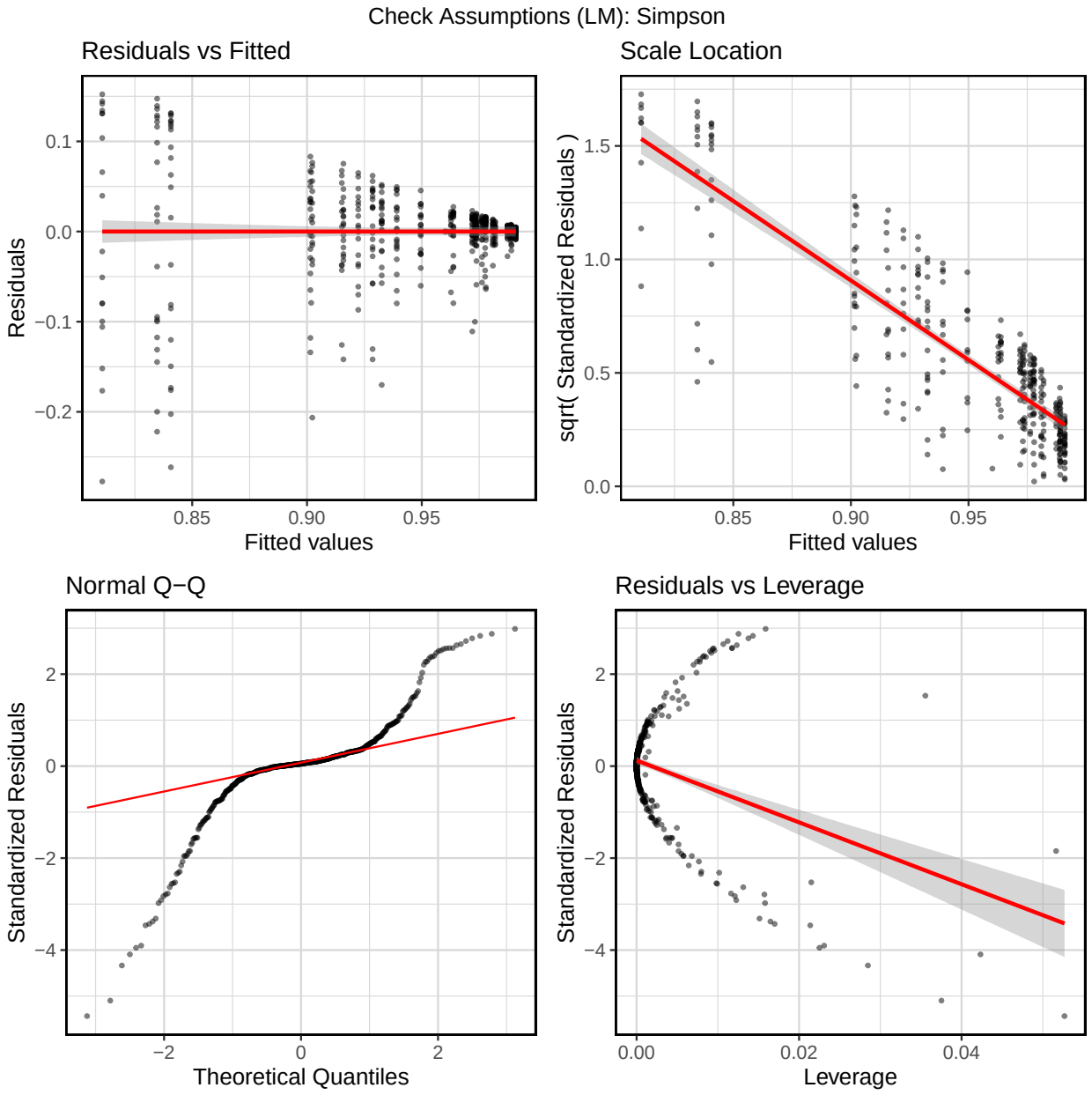
Shannon

```
data <- caseInfStats[["mod.unref"]]
index <- "Shannon"
check_assump_plots(data, index)
```


Check Assumptions (LM): Shannon
Residuals vs Fitted
Residuals
Fitted values
Scale Location
sqrt(Standardized Residuals)
Fitted values
Normal Q−Q
Standardized Residuals
Theoretical Quantiles
Residuals vs Leverage
Standardized Residuals
Leverage

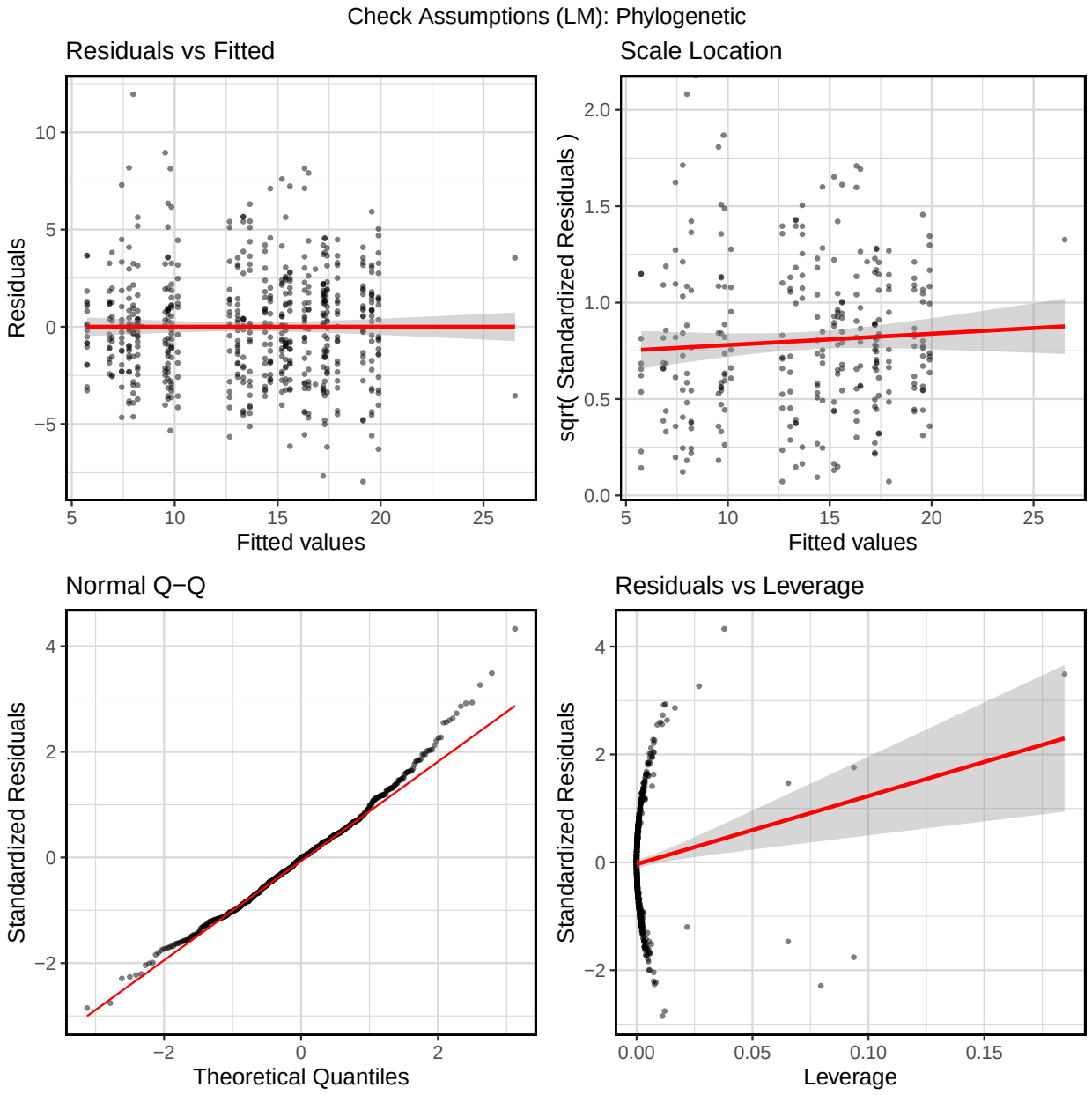
Simpson

```
data <- caseInfStats[["mod.unref"]]
index <- "Simpson"
check_assump_plots(data, index)
```



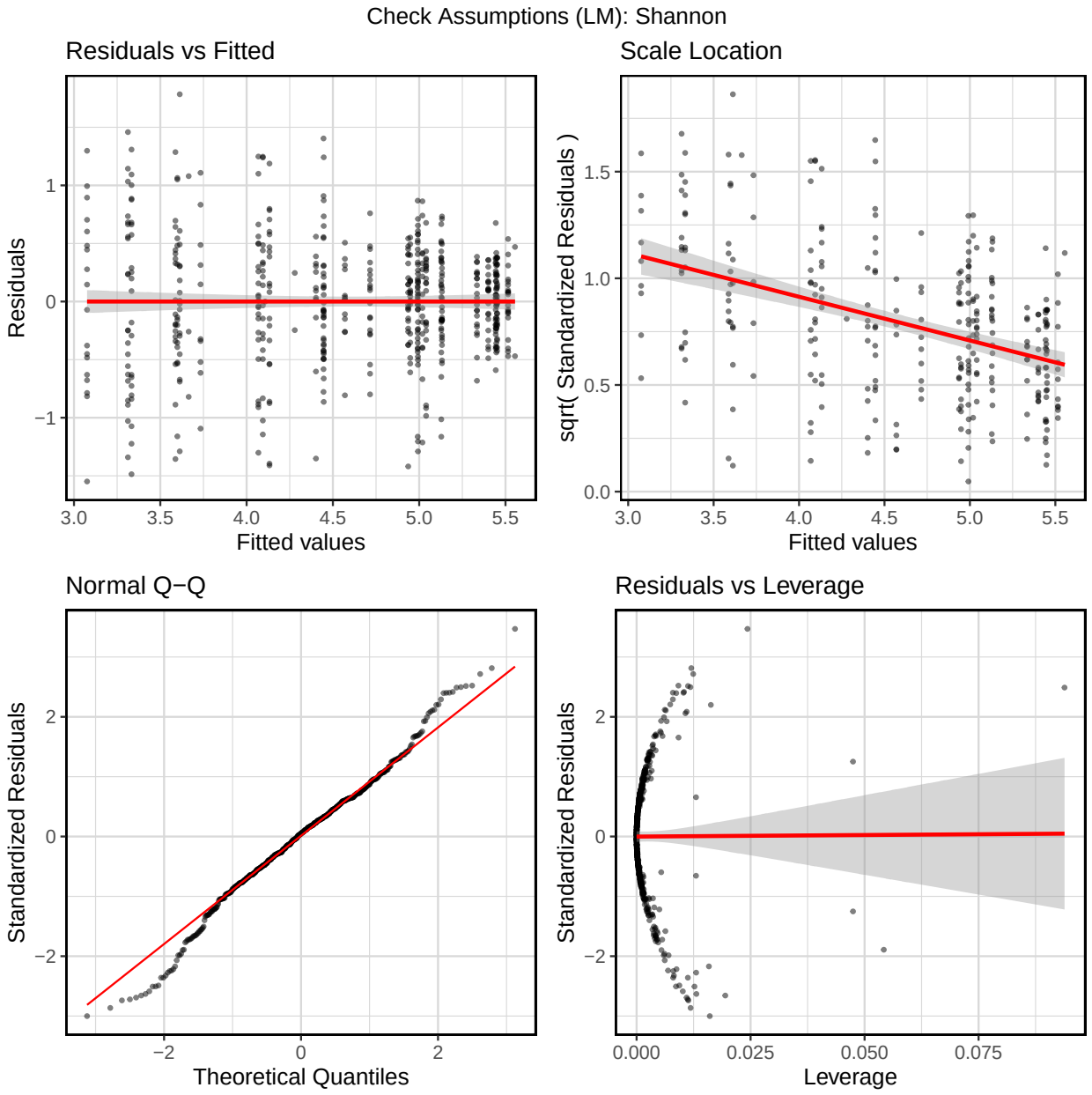
Phylogenetic

```
data <- caseInfStats[["mod.unref"]]
index <- "Phylogenetic"
check_assump_plots(data, index)
```



Richness

```
data <- caseInfStats[["mod.unref"]]
index <- "Shannon"
check_assump_plots(data, index)
```



Remove unusual observations

The previous plots revealed that the data is largely consistent, but there are some inconsistencies in variance and non-normality for some of the alpha indices. To correct for this, we applied case influence statistics to remove any unusually influential and leveraged samples. We removed any samples using cutoffs of 2p/n (p is the number of unknown parameters in the model) for leverage, 1 for Cook's Distance and above 2 or below -2 for standardized residuals.

```
caseInfStats <- list()

data <- ps.list$RAR[["dt.all.alpha.melt"]]
# data <- dplyr::mutate(data, obs_num = row_number(), .before = 1)

caseInfStats[["mod.unref"]] <- lapply(methods.alpha, function(alpha){
  lm( formula = "Alpha.Score ~ sex*HMPbodysubsite",
       data = subset(data, Alpha.Metric == alpha)
      )
})

# Fortify data for plotting
caseInfStats[["dataFort.unref"]] <- lapply(methods.alpha, function(alpha){
  fortify(caseInfStats$mod.unref[[alpha]], subset(data, Alpha.Metric == alpha))
})

# Rename some column names
lapply(methods.alpha, function(alpha){
  setnames(caseInfStats$dataFort.unref[[alpha]],
           old=c(".hat", ".cooksd", ".stdresid"),
           new=c("Lev", "CooksD", "StdResid"))
})


# Unrefinded model

## Plot

caseInfStats[["unref.plot"]] <- lapply(names(methods.alpha), function(alpha){
  qplot(obs_num, value, data = reshape::melt(caseInfStats$dataFort.unref[[alpha]][, c("obs_num","Lev",
    id.vars = "obs_num")) +
    geom_point(aes(color = variable)) +
    facet_grid(variable ~ ., scale = "free_y") +
    labs(title = paste0("Case-influence statistics plot: Unrefined model (", alpha,")")) +
    scale_color_brewer(palette = "Dark2") +
    theme(legend.position = "none") + scale_x_continuous(breaks = scales::breaks_pretty(10))

})

names(caseInfStats[["unref.plot"]]) <- names(methods.alpha)
```

From the original 562 samples, we removed 45, 54, 33, and 31 unusual observations in the Shannon, Simpson, Phylogenetic and Richness indices, respectively.

Fit refined model

After removing the unusual samples, we again fit the data to a linear model to assess body sub-site, sex, or their interaction's ability to predict alpha-diversity score.

```
# Refined model

### Statisticians use rough cutoffs of 2p/n (p is the number of unknown parameters in the model) for le
#     1 for Cook's Distance and above 2 or below -2 for standardized residuals.
#     Observations falling outside these ranges warrant further attention.

caseInfStats[["cutoff.lev"]] <- lapply(methods.alpha, function(alpha){
  cutOff.lev <- (2 * length(caseInfStats$mod.unref[[alpha]][["coefficients"]]) /
                   nrow(caseInfStats$dataFort.unref[[alpha]]))
})


caseInfStats[["dataFort.sub"]] <- lapply(methods.alpha, function(alpha){
  caseInfStats$dataFort.unref[[alpha]]  %>%
    dplyr::filter(Lev < caseInfStats[["cutoff.lev"]][[alpha]]) %>%  # Leverage Cut off
    dplyr::filter(StdResid < 2 & StdResid > -2 ) %>%  # StdResid Cut Off
    dplyr::select(obs_num:Alpha.Score)  # Removes the old case statistic influence data
})


# make refined model
caseInfStats[["mod.ref"]] <- lapply(methods.alpha, function(alpha){
  lm( formula = "Alpha.Score ~ sex*HMPbodysubsite",
       data = caseInfStats[["dataFort.sub"]][[alpha]]
      )
})


# Fortify data for plotting
caseInfStats[["dataFort.ref"]] <- lapply(methods.alpha, function(alpha){
  fortify(caseInfStats$mod.ref[[alpha]], caseInfStats[["dataFort.sub"]][[alpha]])
})

# Rename some column names
lapply(methods.alpha, function(alpha){
  setnames(caseInfStats$dataFort.ref[[alpha]],
           old=c(".hat", ".cooksd", ".stdresid"),
           new=c("Lev", "CooksD", "StdResid"))
})


# Refinded model Plot

caseInfStats[["ref.plot"]] <- lapply(names(methods.alpha), function(alpha){
  tmp.data <- caseInfStats[["dataFort.ref"]]
  qplot(obs_num, value, data = reshape::melt(tmp.data[[alpha]][, c("obs_num","Lev", "CooksD", "StdResid
    id.vars = "obs_num")) +
    geom_point(aes(color = variable)) +
    facet_grid(variable ~ ., scale = "free_y") +
```

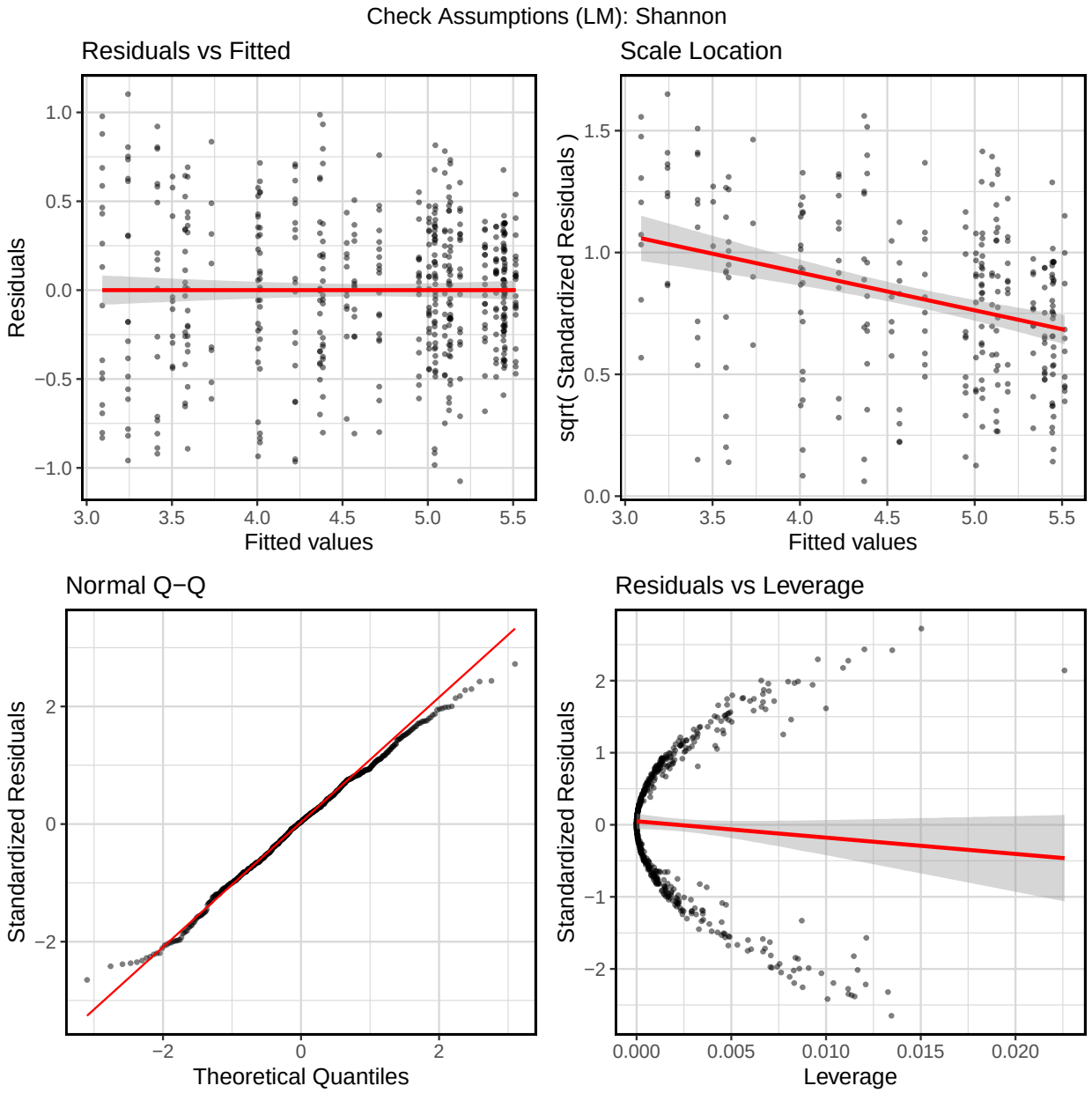
```
    labs(title = paste0("Case-influence statistics plot: Refined model (", alpha,")")) +
    scale_color_brewer(palette = "Dark2") +
    theme(legend.position = "none") + scale_x_continuous(breaks = scales::breaks_pretty(10))

})

names(caseInfStats[["ref.plot"]]) <- names(methods.alpha)
```

Check assumptions (refined model)

After removing the unusual observations, the plots for each indice appear to have been improved. However, the Simpson's data appears to not meet the assumptions of normality.

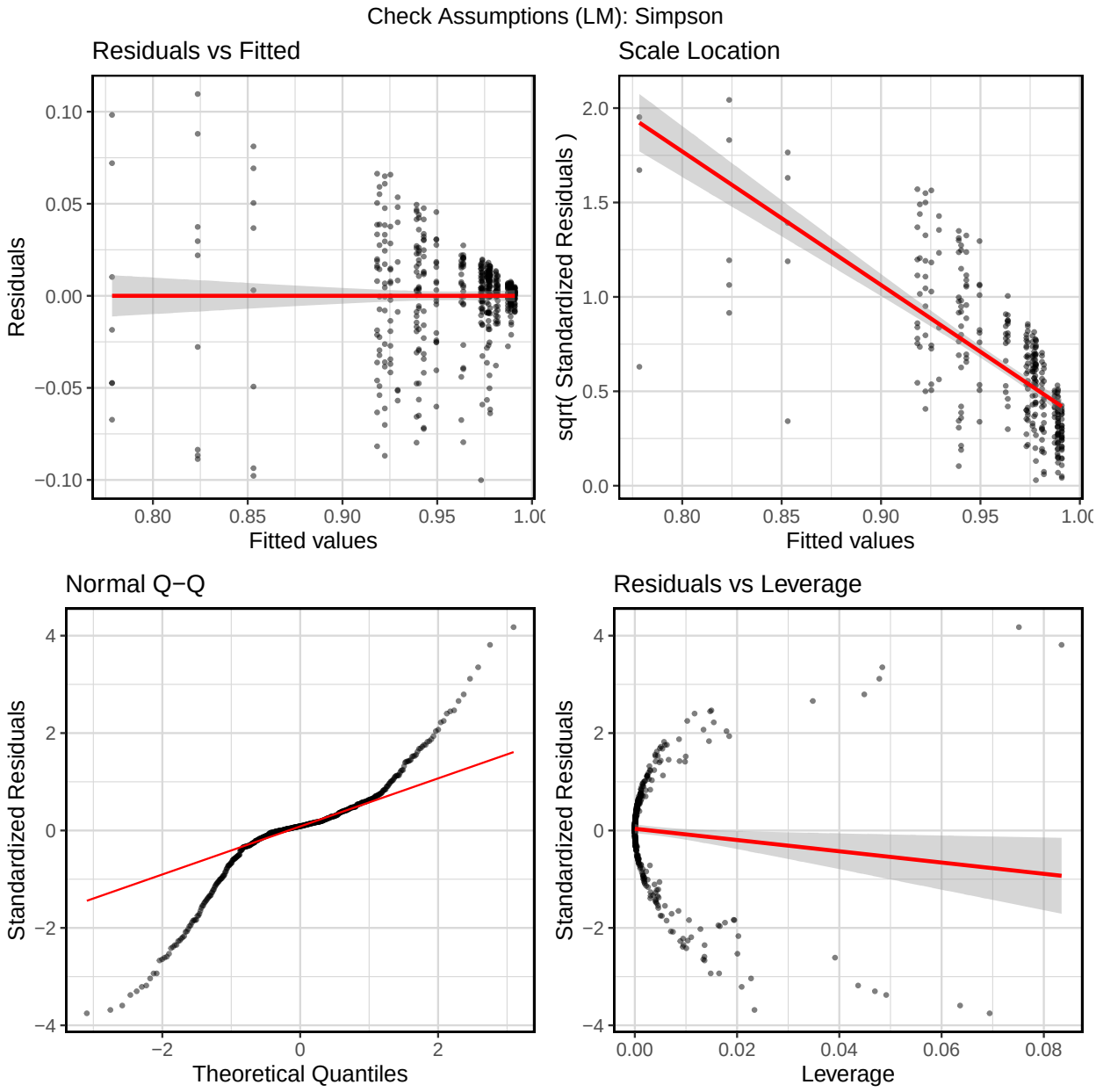
Shannon

```
data <- caseInfStats[["mod.ref"]]
index <- "Shannon"
check_assump_plots(data, index)
```



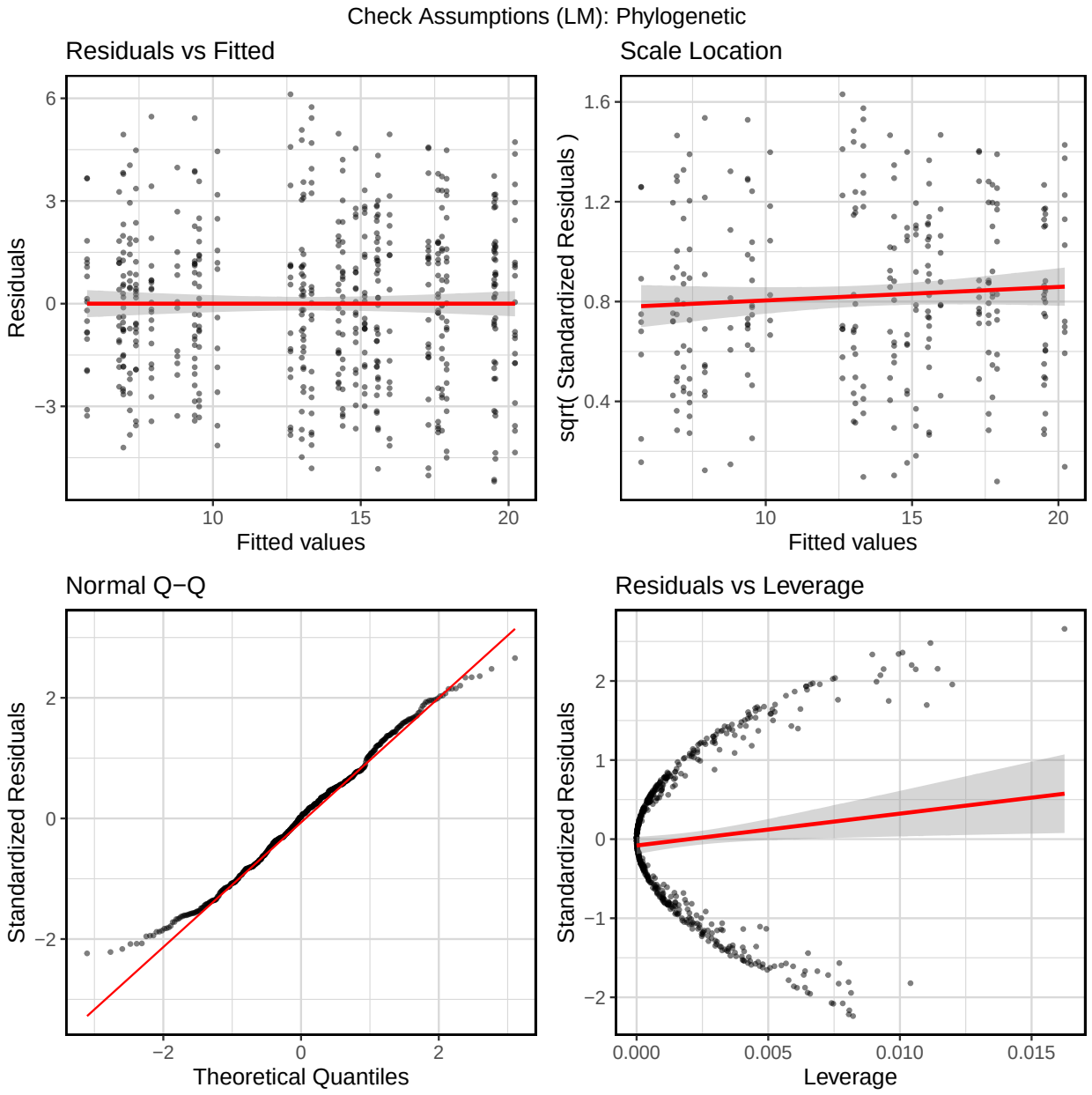
Simpson

```
data <- caseInfStats[["mod.ref"]]
index <- "Simpson"
check_assump_plots(data, index)
```



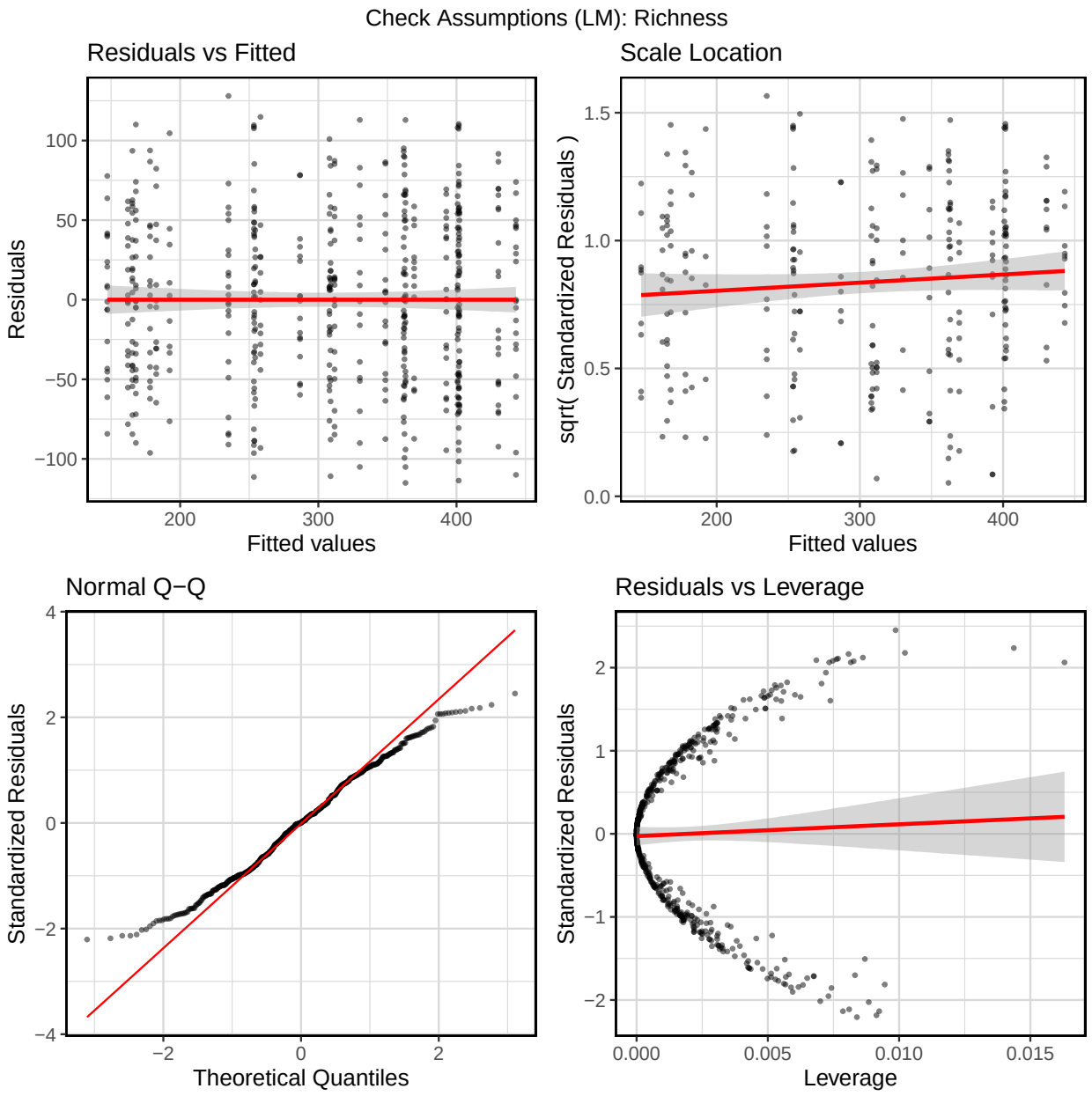
Phylogenetic

```
data <- caseInfStats[["mod.ref"]]
index <- "Phylogenetic"
check_assump_plots(data, index)
```



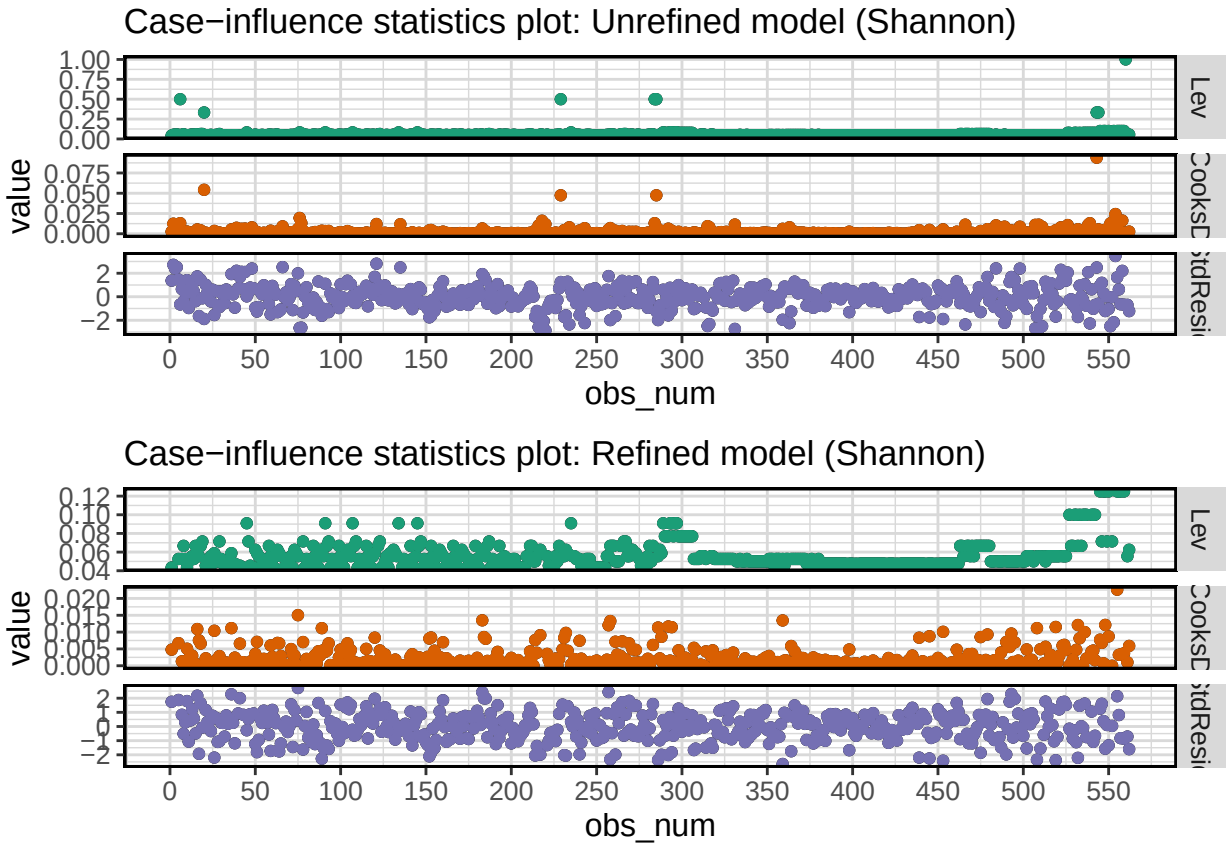
Richness

```
data <- caseInfStats[["mod.ref"]]
index <- "Richness"
check_assump_plots(data, index)
```

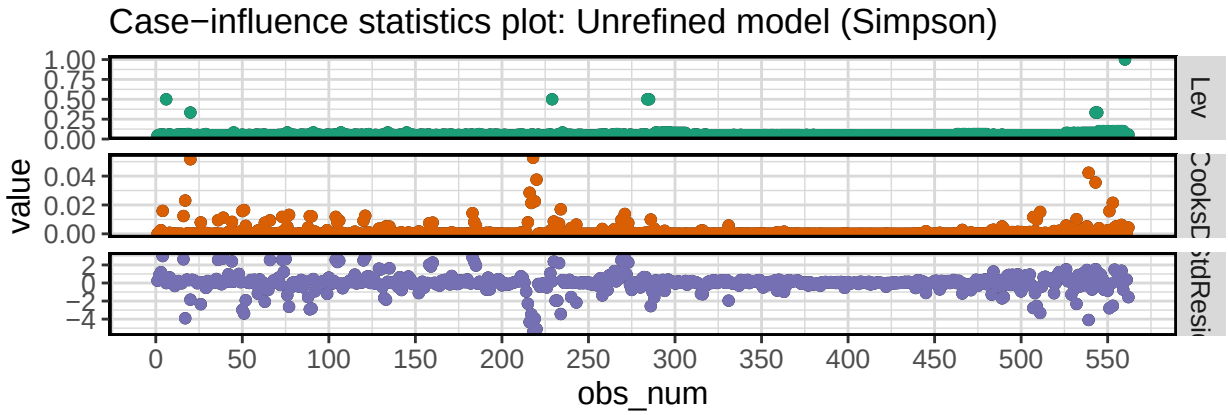


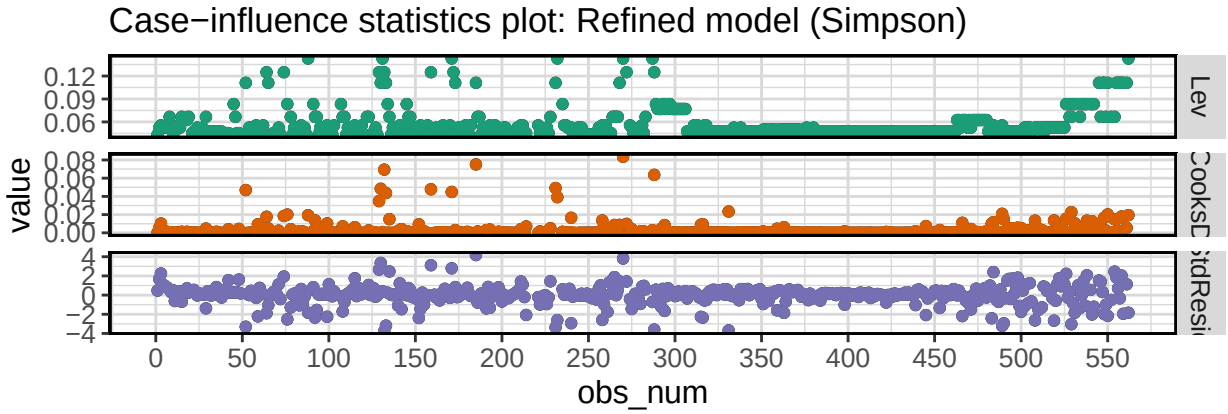
Plot unrefined and refined models

```
## $Shannon
```

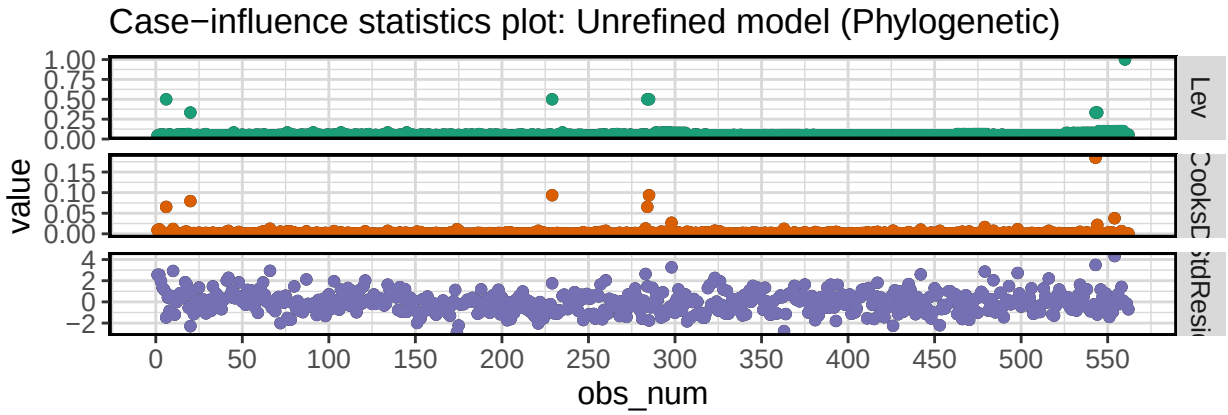


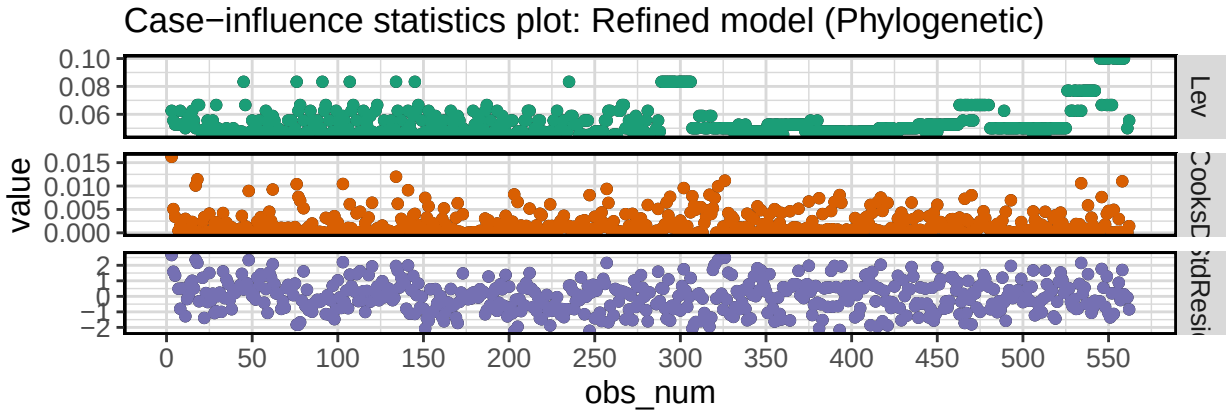

```
##
## $Simpson
```



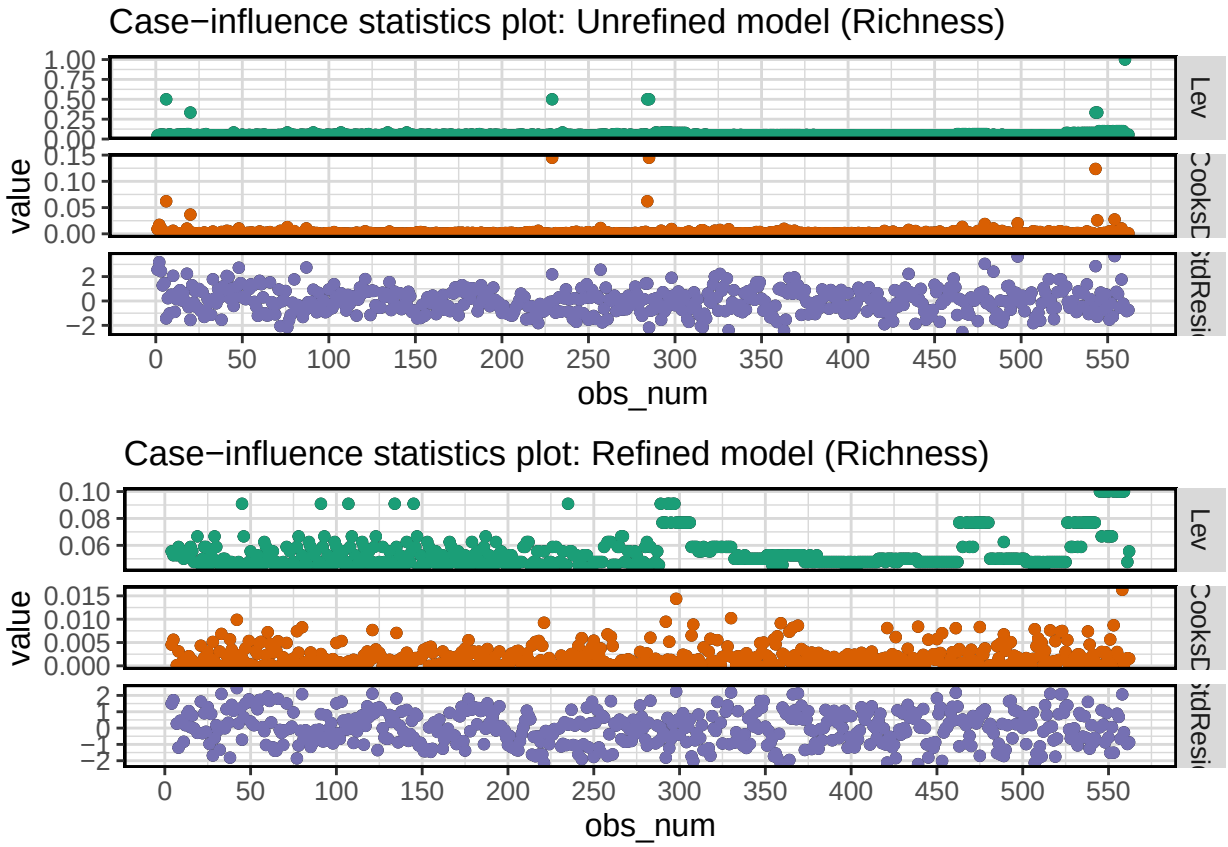



```
##
## $Phylogenetic
```


Case−influence statistics plot: Unrefined model (Phylogenetic)
value
obs_num
Lev
CooksD
StdResi


Case−influence statistics plot: Refined model (Phylogenetic)
value
obs_num
Lev
CooksD
StdResi

```
##
## $Richness
```


Case−influence statistics plot: Unrefined model (Richness)
value
obs_num
Lev
CooksD
StdResi
Case−influence statistics plot: Refined model (Richness)
value
obs_num
Lev
CooksD
StdResi

Linear model results

Shannon

```
index <- "Shannon"

Anova(caseInfStats[["mod.ref"]][[index]], type = 2) %>% report()
```

```
## The ANOVA suggests that:
##
##   - The main effect of sex is statistically significant and small (F(1, 488) = 12.31, p < .001; Eta2
##   - The main effect of HMPbodysubsite is statistically significant and large (F(15, 488) = 96.72, p
##   - The interaction between sex and HMPbodysubsite is statistically significant and medium (F(12, 48
##
## Effect sizes were labelled following Field's (2013) recommendations.
```

```
Anova(caseInfStats[["mod.ref"]][[index]], type = 2)
```

```
## Anova Table (Type II tests)
##
## Response: Alpha.Score
##                      Sum Sq  Df F value    Pr(>F)
## sex                   2.139   1 12.3088 0.0004925 ***
## HMPbodysubsite      252.107  15 96.7150 < 2.2e-16 ***
## sex:HMPbodysubsite    5.719  12  2.7425 0.0013018 **
## Residuals            84.804 488
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
caseInfStats[["mod.ref"]][[index]] %>% report_table()
```

```
## Parameter                                                | Coefficient |         95% CI | t(488) |
## ----------------------------------------------------------------------------------------------------
## (Intercept)                                              |        4.52 | [ 4.28,  4.77] |  36.00 |
## sex [male]                                               |        0.05 | [-0.29,  0.38] |   0.26 |
## HMPbodysubsite [Attached Keratinized gingiva]            |       -0.51 | [-0.82, -0.20] |  -3.25 |
## HMPbodysubsite [Buccal mucosa]                           |       -0.52 | [-0.85, -0.19] |  -3.14 |
## HMPbodysubsite [Hard palate]                             |       -0.16 | [-0.47,  0.16] |  -0.98 |
## HMPbodysubsite [Left Retroauricular crease]              |       -0.95 | [-1.28, -0.62] |  -5.64 |
## HMPbodysubsite [Mid vagina]                              |       -1.11 | [-1.43, -0.79] |  -6.88 |
## HMPbodysubsite [Palatine Tonsils]                        |        0.61 | [ 0.31,  0.91] |   3.97 |
## HMPbodysubsite [Posterior fornix]                        |       -1.43 | [-1.75, -1.11] |  -8.78 |
## HMPbodysubsite [Right Retroauricular crease]             |       -0.93 | [-1.26, -0.61] |  -5.64 |
## HMPbodysubsite [Saliva]                                  |        0.92 | [ 0.62,  1.22] |   6.01 |
## HMPbodysubsite [Stool]                                   |        0.42 | [ 0.11,  0.74] |   2.65 |
## HMPbodysubsite [Subgingival plaque]                      |        0.93 | [ 0.63,  1.23] |   6.06 |
## HMPbodysubsite [Supragingival plaque]                    |        0.92 | [ 0.61,  1.23] |   5.84 |
## HMPbodysubsite [Throat]                                  |        0.52 | [ 0.22,  0.82] |   3.37 |
## HMPbodysubsite [Tongue dorsum]                           |        0.19 | [-0.12,  0.50] |   1.21 |
## HMPbodysubsite [Vaginal introitus]                       |       -1.28 | [-1.60, -0.97] |  -8.04 |
```

```
## sex [male] * HMPbodysubsite [Attached Keratinized gingiva] |        0.16 | [-0.27,  0.59] |   0.74 |
## sex [male] * HMPbodysubsite [Buccal mucosa]                 |        0.34 | [-0.10,  0.77] |   1.51 |
## sex [male] * HMPbodysubsite [Hard palate]                   |        0.69 | [ 0.25,  1.13] |   3.06 |
## sex [male] * HMPbodysubsite [Left Retroauricular crease]    |        0.11 | [-0.39,  0.60] |   0.43 |
## sex [male] * HMPbodysubsite [Palatine Tonsils]              |       -0.14 | [-0.55,  0.28] |  -0.65 |
## sex [male] * HMPbodysubsite [Right Retroauricular crease]   |       -0.13 | [-0.61,  0.34] |  -0.55 |
## sex [male] * HMPbodysubsite [Saliva]                        |        0.03 | [-0.39,  0.44] |   0.13 |
## sex [male] * HMPbodysubsite [Stool]                         |        0.13 | [-0.30,  0.56] |   0.60 |
## sex [male] * HMPbodysubsite [Subgingival plaque]            |       -0.16 | [-0.58,  0.26] |  -0.76 |
## sex [male] * HMPbodysubsite [Supragingival plaque]          |       -0.09 | [-0.52,  0.33] |  -0.42 |
## sex [male] * HMPbodysubsite [Throat]                        |        0.10 | [-0.32,  0.52] |   0.46 |
## sex [male] * HMPbodysubsite [Tongue dorsum]                 |        0.25 | [-0.18,  0.67] |   1.14 |
##                                                             |             |                |        |
## AIC                                                         |             |                |        |
## BIC                                                         |             |                |        |
## R2                                                          |             |                |        |
## R2 (adj.)                                                   |             |                |        |
## Sigma                                                       |             |                |        |
```

Simpson

```
index <- "Simpson"

Anova(caseInfStats[["mod.ref"]][[index]], type = 2) %>% report()
```

```
## The ANOVA suggests that:
##
##   - The main effect of sex is statistically not significant and very small (F(1, 479) = 2.19, p = 0.
##   - The main effect of HMPbodysubsite is statistically significant and large (F(15, 479) = 62.57, p <
##   - The interaction between sex and HMPbodysubsite is statistically not significant and small (F(12,
##
## Effect sizes were labelled following Field's (2013) recommendations.
```

```
Anova(caseInfStats[["mod.ref"]][[index]], type = 2)
```

```
## Anova Table (Type II tests)
##
## Response: Alpha.Score
##                      Sum Sq  Df F value Pr(>F)
## sex                 0.00170   1  2.1934 0.1393
## HMPbodysubsite      0.72782  15 62.5740 <2e-16 ***
## sex:HMPbodysubsite 0.00851  12  0.9150 0.5315
## Residuals           0.37143 479
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
caseInfStats[["mod.ref"]][[index]] %>% report_table()
```

Parameter	Coefficient	95% CI	t(479)
(Intercept)	0.96	[0.95, 0.98]	119.76
sex [male]	0.01	[-0.01, 0.03]	1.00
HMPbodysubsite [Attached Keratinized gingiva]	-0.02	[-0.04, 0.00]	-2.24
HMPbodysubsite [Buccal mucosa]	-0.04	[-0.06, -0.02]	-3.89
HMPbodysubsite [Hard palate]	-0.01	[-0.03, 0.01]	-1.28
HMPbodysubsite [Left Retroauricular crease]	-0.04	[-0.06, -0.02]	-3.47
HMPbodysubsite [Mid vagina]	-0.11	[-0.13, -0.08]	-8.63
HMPbodysubsite [Palatine Tonsils]	0.02	[0.00, 0.04]	1.84
HMPbodysubsite [Posterior fornix]	-0.18	[-0.21, -0.16]	-13.92
HMPbodysubsite [Right Retroauricular crease]	-0.04	[-0.07, -0.02]	-4.12
HMPbodysubsite [Saliva]	0.03	[0.01, 0.05]	2.63
HMPbodysubsite [Stool]	0.02	[0.00, 0.04]	1.85
HMPbodysubsite [Subgingival plaque]	0.03	[0.01, 0.05]	2.84
HMPbodysubsite [Supragingival plaque]	0.03	[0.01, 0.05]	2.75
HMPbodysubsite [Throat]	0.01	[-0.01, 0.03]	1.43
HMPbodysubsite [Tongue dorsum]	1.05e-03	[-0.02, 0.02]	0.10
HMPbodysubsite [Vaginal introitus]	-0.14	[-0.16, -0.11]	-11.33
sex [male] * HMPbodysubsite [Attached Keratinized gingiva]	-8.50e-03	[-0.04, 0.02]	-0.60
sex [male] * HMPbodysubsite [Buccal mucosa]	5.67e-03	[-0.02, 0.03]	0.40
sex [male] * HMPbodysubsite [Hard palate]	0.01	[-0.02, 0.04]	0.90

```
## sex [male] * HMPbodysubsite [Left Retroauricular crease]   |   -7.27e-03 | [-0.04,  0.02] |  -0.45 |
## sex [male] * HMPbodysubsite [Palatine Tonsils]             |       -0.02 | [-0.04,  0.01] |  -1.13 |
## sex [male] * HMPbodysubsite [Right Retroauricular crease]  |   -9.88e-03 | [-0.04,  0.02] |  -0.64 |
## sex [male] * HMPbodysubsite [Saliva]                       |       -0.01 | [-0.04,  0.02] |  -0.78 |
## sex [male] * HMPbodysubsite [Stool]                        |       -0.02 | [-0.04,  0.01] |  -1.06 |
## sex [male] * HMPbodysubsite [Subgingival plaque]           |       -0.01 | [-0.04,  0.01] |  -1.05 |
## sex [male] * HMPbodysubsite [Supragingival plaque]         |       -0.01 | [-0.04,  0.01] |  -0.91 |
## sex [male] * HMPbodysubsite [Throat]                       |       -0.01 | [-0.04,  0.02] |  -0.75 |
## sex [male] * HMPbodysubsite [Tongue dorsum]                |   -1.73e-03 | [-0.03,  0.03] |  -0.12 |
##                                                           |             |                |        |
## AIC                                                       |             |                |        |
## BIC                                                       |             |                |        |
## R2                                                        |             |                |        |
## R2 (adj.)                                                 |             |                |        |
## Sigma                                                     |             |                |        |
```

Phylogenetic

```
index <- "Phylogenetic"

Anova(caseInfStats[["mod.ref"]][[index]], type = 2) %>% report()
```

```
## The ANOVA suggests that:
##
##   - The main effect of sex is statistically significant and very small (F(1, 500) = 3.95, p = 0.047;
##   - The main effect of HMPbodysubsite is statistically significant and large (F(15, 500) = 110.86, p
##   - The interaction between sex and HMPbodysubsite is statistically significant and small (F(12, 500
##
## Effect sizes were labelled following Field's (2013) recommendations.
```

```
Anova(caseInfStats[["mod.ref"]][[index]], type = 2)
```

```
## Anova Table (Type II tests)
##
## Response: Alpha.Score
##                     Sum Sq  Df  F value    Pr(>F)
## sex                   22.3   1   3.9495  0.047430 *
## HMPbodysubsite      9391.5  15 110.8586 < 2.2e-16 ***
## sex:HMPbodysubsite  172.8  12   2.5503  0.002794 **
## Residuals          2823.9 500
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
caseInfStats[["mod.ref"]][[index]] %>% report_table()
```

```
## Parameter                                                   | Coefficient |         95% CI | t(500) |
## ------------------------------------------------------------------------------------------------------
## (Intercept)                                                 |       10.15 | [ 8.80, 11.50] |  14.80 |
## sex [male]                                                  |       -1.36 | [-3.26,  0.55] |  -1.40 |
## HMPbodysubsite [Attached Keratinized gingiva]               |       -0.77 | [-2.44,  0.91] |  -0.90 |
## HMPbodysubsite [Buccal mucosa]                              |        2.47 | [ 0.69,  4.25] |   2.72 |
## HMPbodysubsite [Hard palate]                                |        4.10 | [ 2.36,  5.84] |   4.63 |
## HMPbodysubsite [Left Retroauricular crease]                 |       -2.96 | [-4.77, -1.15] |  -3.22 |
## HMPbodysubsite [Mid vagina]                                 |       -2.76 | [-4.45, -1.07] |  -3.20 |
## HMPbodysubsite [Palatine Tonsils]                           |        5.43 | [ 3.74,  7.12] |   6.32 |
## HMPbodysubsite [Posterior fornix]                           |       -4.41 | [-6.15, -2.67] |  -4.98 |
## HMPbodysubsite [Right Retroauricular crease]                |       -3.18 | [-4.96, -1.40] |  -3.51 |
## HMPbodysubsite [Saliva]                                     |        9.40 | [ 7.71, 11.09] |  10.93 |
## HMPbodysubsite [Stool]                                      |        2.85 | [ 1.09,  4.61] |   3.18 |
## HMPbodysubsite [Subgingival plaque]                         |        9.36 | [ 7.69, 11.04] |  10.98 |
## HMPbodysubsite [Supragingival plaque]                       |        7.59 | [ 5.85,  9.33] |   8.57 |
## HMPbodysubsite [Throat]                                     |        4.98 | [ 3.31,  6.66] |   5.84 |
## HMPbodysubsite [Tongue dorsum]                              |        2.91 | [ 1.19,  4.63] |   3.32 |
## HMPbodysubsite [Vaginal introitus]                          |       -2.23 | [-3.94, -0.53] |  -2.57 |
## sex [male] * HMPbodysubsite [Attached Keratinized gingiva]  |        1.51 | [-0.88,  3.90] |   1.24 |
## sex [male] * HMPbodysubsite [Buccal mucosa]                 |        3.56 | [ 1.10,  6.03] |   2.84 |
## sex [male] * HMPbodysubsite [Hard palate]                   |        3.08 | [ 0.57,  5.59] |   2.41 |
```

```
## sex [male] * HMPbodysubsite [Left Retroauricular crease]   |        1.12 | [-1.57,  3.82] |   0.82 |
## sex [male] * HMPbodysubsite [Palatine Tonsils]             |        1.32 | [-1.10,  3.73] |   1.07 |
## sex [male] * HMPbodysubsite [Right Retroauricular crease]  |        1.21 | [-1.37,  3.80] |   0.92 |
## sex [male] * HMPbodysubsite [Saliva]                       |        2.02 | [-0.38,  4.42] |   1.65 |
## sex [male] * HMPbodysubsite [Stool]                        |        1.69 | [-0.76,  4.14] |   1.35 |
## sex [male] * HMPbodysubsite [Subgingival plaque]           |       -0.25 | [-2.63,  2.13] |  -0.21 |
## sex [male] * HMPbodysubsite [Supragingival plaque]         |        0.90 | [-1.53,  3.32] |   0.73 |
## sex [male] * HMPbodysubsite [Throat]                       |        3.84 | [ 1.44,  6.24] |   3.14 |
## sex [male] * HMPbodysubsite [Tongue dorsum]                |        2.68 | [ 0.27,  5.09] |   2.18 |
##                                                           |             |                |        |
## AIC                                                       |             |                |        |
## BIC                                                       |             |                |        |
## R2                                                        |             |                |        |
## R2 (adj.)                                                 |             |                |        |
## Sigma                                                     |             |                |        |
```

Richness

```
index <- "Richness"

Anova(caseInfStats[["mod.ref"]][[index]], type = 2) %>% report()
```

```
## The ANOVA suggests that:
##
##   - The main effect of sex is statistically significant and small (F(1, 502) = 20.18, p < .001; Eta2
##   - The main effect of HMPbodysubsite is statistically significant and large (F(15, 502) = 91.18, p <
##   - The interaction between sex and HMPbodysubsite is statistically significant and small (F(12, 502)
##
## Effect sizes were labelled following Field's (2013) recommendations.
```

```
Anova(caseInfStats[["mod.ref"]][[index]], type = 2)
```

```
## Anova Table (Type II tests)
##
## Response: Alpha.Score
##                      Sum Sq  Df F value    Pr(>F)
## sex                   57658   1  20.181 8.752e-06 ***
## HMPbodysubsite      3907814  15  91.185 < 2.2e-16 ***
## sex:HMPbodysubsite    70147  12   2.046   0.01906 *
## Residuals           1434248 502
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
caseInfStats[["mod.ref"]][[index]] %>% report_table()
```

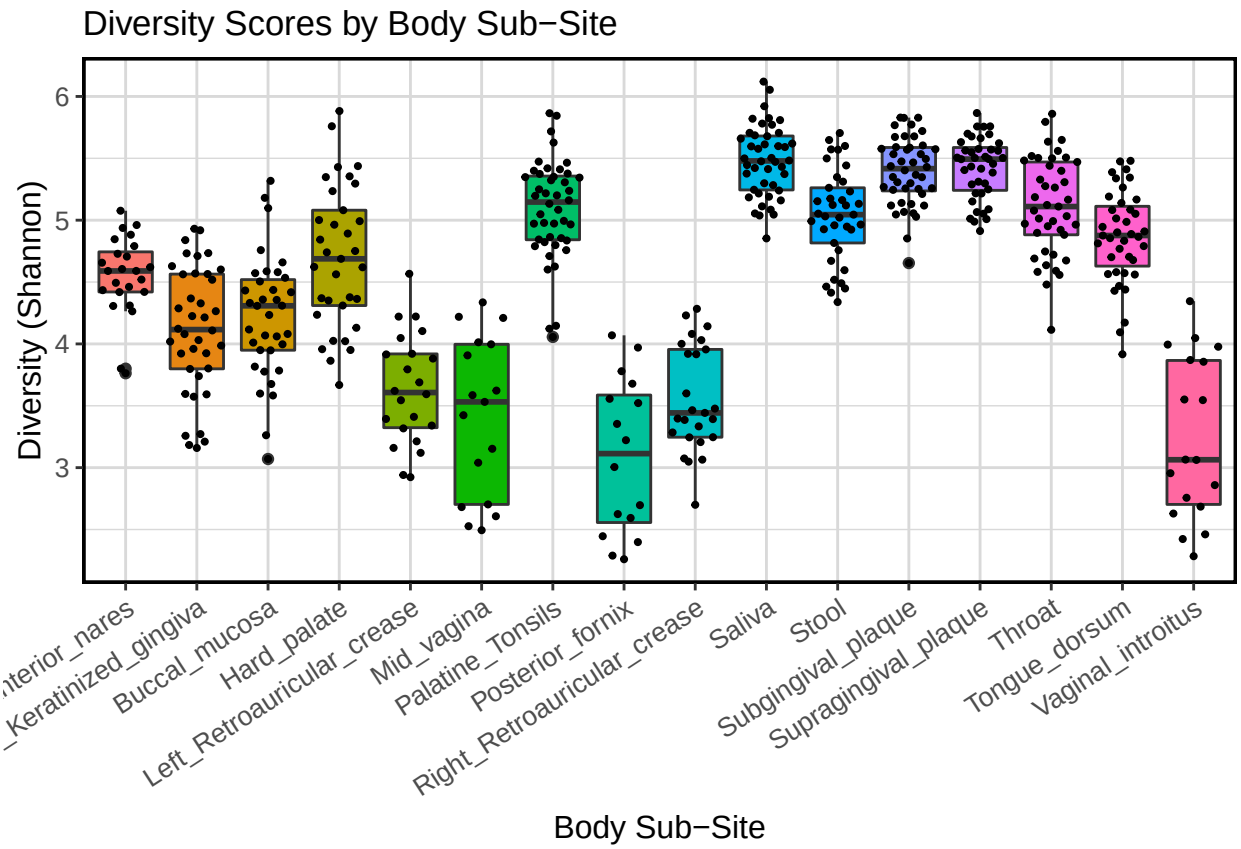
```
## Parameter                                                | Coefficient |           95% CI | t(502)
## ---------------------------------------------------------------------------------------------------
## (Intercept)                                              |      254.36 | [ 222.70, 286.03] |  15.78
## sex [male]                                               |        3.79 | [ -39.23,  46.81] |   0.17
## HMPbodysubsite [Attached Keratinized gingiva]            |      -19.36 | [ -58.14,  19.42] |  -0.98
## HMPbodysubsite [Buccal mucosa]                           |       -0.68 | [ -41.81,  40.46] |  -0.03
## HMPbodysubsite [Hard palate]                             |       32.40 | [  -8.24,  73.04] |   1.57
## HMPbodysubsite [Left Retroauricular crease]              |      -92.16 | [-133.85, -50.48] |  -4.34
## HMPbodysubsite [Mid vagina]                              |      -88.91 | [-127.69, -50.13] |  -4.50
## HMPbodysubsite [Palatine Tonsils]                        |      107.49 | [  68.41, 146.58] |   5.40
## HMPbodysubsite [Posterior fornix]                        |     -107.09 | [-147.28, -66.90] |  -5.23
## HMPbodysubsite [Right Retroauricular crease]             |      -76.13 | [-116.76, -35.49] |  -3.68
## HMPbodysubsite [Saliva]                                  |      175.97 | [ 136.88, 215.06] |   8.85
## HMPbodysubsite [Stool]                                   |       54.53 | [  14.33,  94.72] |   2.67
## HMPbodysubsite [Subgingival plaque]                      |      147.38 | [ 108.88, 185.87] |   7.51
## HMPbodysubsite [Supragingival plaque]                    |      146.53 | [ 106.74, 186.32] |   7.24
## HMPbodysubsite [Throat]                                  |      108.00 | [  69.22, 146.78] |   5.47
## HMPbodysubsite [Tongue dorsum]                           |       53.69 | [  13.90,  93.48] |   2.65
## HMPbodysubsite [Vaginal introitus]                       |      -86.41 | [-125.50, -47.32] |  -4.34
## sex [male] * HMPbodysubsite [Attached Keratinized gingiva] |       14.59 | [ -39.05,  68.23] |   0.53
## sex [male] * HMPbodysubsite [Buccal mucosa]              |       54.27 | [  -1.33, 109.87] |   1.91
## sex [male] * HMPbodysubsite [Hard palate]                |       78.83 | [  20.97, 136.69] |   2.68
```

```
## sex [male] * HMPbodysubsite [Left Retroauricular crease]   |       26.41 | [ -34.33,  87.15] |   0.85
## sex [male] * HMPbodysubsite [Palatine Tonsils]             |       -2.55 | [ -56.42,  51.31] |  -0.09
## sex [male] * HMPbodysubsite [Right Retroauricular crease]  |        0.67 | [ -57.20,  58.53] |   0.01
## sex [male] * HMPbodysubsite [Saliva]                       |        8.93 | [ -45.18,  63.03] |   0.31
## sex [male] * HMPbodysubsite [Stool]                        |       17.38 | [ -38.41,  73.17] |   0.61
## sex [male] * HMPbodysubsite [Subgingival plaque]           |      -12.91 | [ -66.35,  40.53] |  -0.47
## sex [male] * HMPbodysubsite [Supragingival plaque]         |       -2.59 | [ -56.96,  51.78] |  -0.09
## sex [male] * HMPbodysubsite [Throat]                       |       35.48 | [ -18.68,  89.63] |   1.29
## sex [male] * HMPbodysubsite [Tongue dorsum]                |       36.71 | [ -17.91,  91.32] |   1.31
##                                                            |             |                   |
## AIC                                                        |             |                   |
## BIC                                                        |             |                   |
## R2                                                         |             |                   |
## R2 (adj.)                                                  |             |                   |
## Sigma                                                      |             |                   |
```

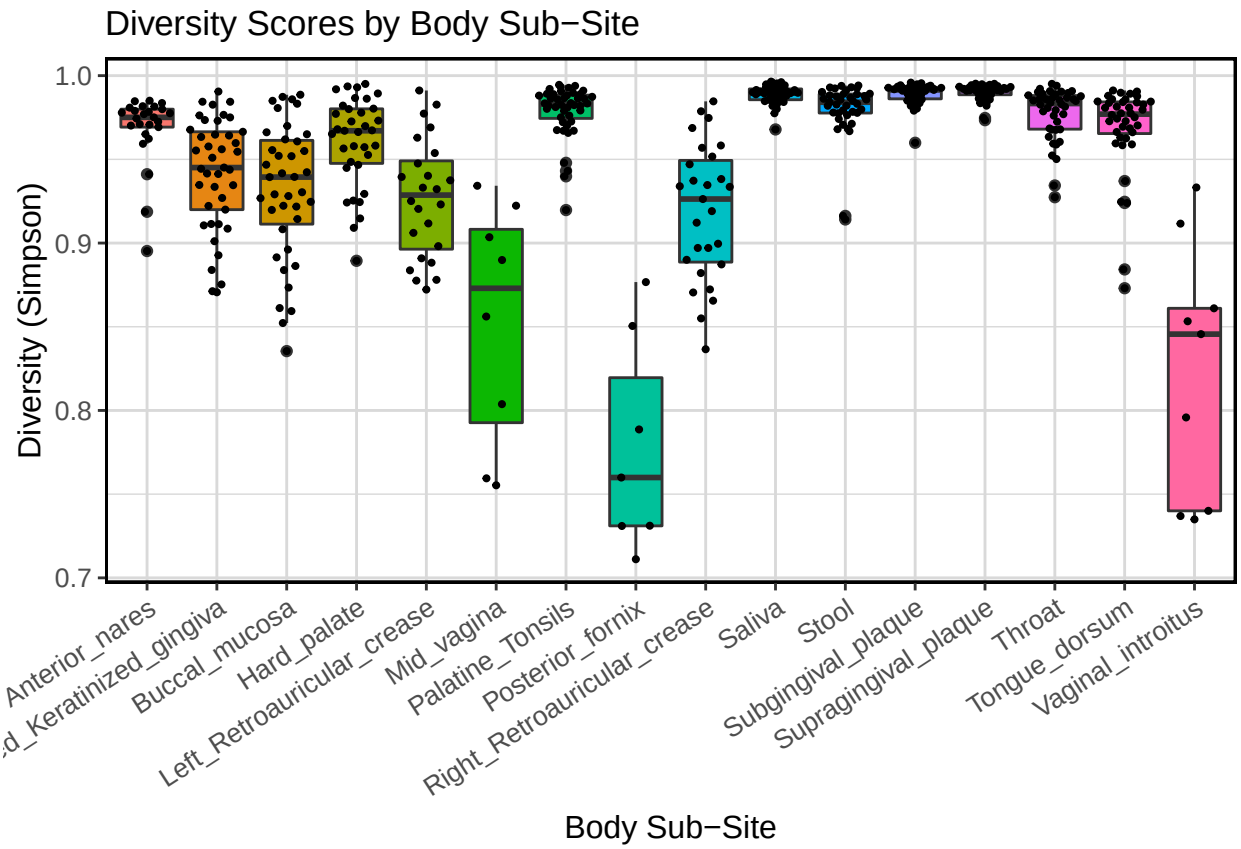
Plots

```
# Create an empty plot list
plots <- list()

if(redo.analysis$redo.plots == T){

  plots[["AlphaDiversity"]][["Sex.bodySubSite"]]  <- lapply(names(methods.alpha), function(alpha){
    tmp.data <- na.omit(as.data.frame(caseInfStats[["dataFort.ref"]][[alpha]]) )
    ggplot(tmp.data, aes(x = HMPbodysubsite, y=Alpha.Score)) +
        geom_boxplot(aes(fill = HMPbodysubsite)) +
        ggbeeswarm::geom_quasirandom(size = 0.75) +  # spaces the dots out nicely
        facet_grid(. ~ .) +  # (Y-axis ~ X-axis)
        theme(legend.position = "none",
              axis.text.x = element_text(angle = 33, hjust = 1, vjust=1)
              ) +
        labs(
          title = "Diversity Scores by Body Sub-Site",
          # caption = "",
          y = paste0("Diversity (", alpha,")"),
          x = "Body Sub-Site"
        )
  })

  names(plots[["AlphaDiversity"]][["Sex.bodySubSite"]]  ) <- names(methods.alpha)


  # Save
  save(object = plots , file = paste0(saveObj.path, "/plots-alpha-sex-bodysubsite.RData"))

} else {

  # Load
  load(file = paste0(saveObj.path, "/plots-alpha-sex-bodysubsite.RData") )

}

plots
```

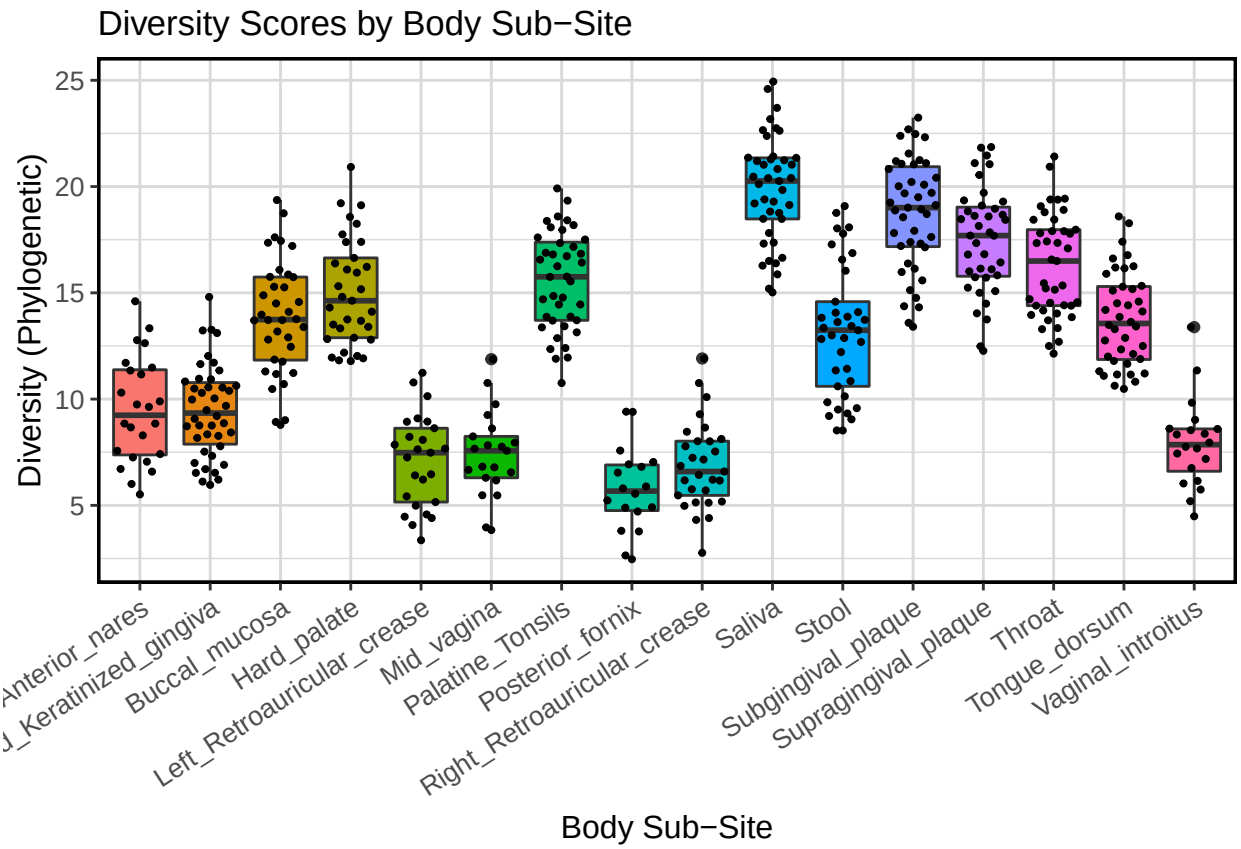
```
## $AlphaDiversity
## $AlphaDiversity$Sex.bodySubSite
## $AlphaDiversity$Sex.bodySubSite$Shannon
```




```
##
## $AlphaDiversity$Sex.bodySubSite$Simpson
```




```
##
## $AlphaDiversity$Sex.bodySubSite$Phylogenetic
```




```
##
## $AlphaDiversity$Sex.bodySubSite$Richness
```


Diversity Scores by Body Sub−Site
Diversity (Richness)
500
400
300
200
100
Anterior_nares
ed_Keratinized_gingiva
Buccal_mucosa
Hard_palate
Left_Retroauricular_crease
Mid_vagina
Palatine_Tonsils
Posterior_fornix
Right_Retroauricular_crease
Saliva
Stool
Subgingival_plaque
Supragingival_plaque
Throat
Tongue_dorsum
Vaginal_introitus
Body Sub−Site