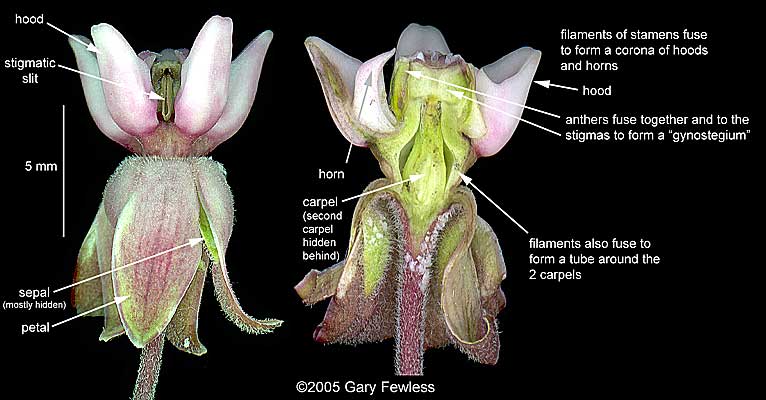
**Dataset #4: Correlation and Regression**

In this Dataset, you will be learning how to use several kinds of correlation and regression approaches using JASP (<https://jasp-stats.org/>). The JASP guide is linked on the website if you need more explanation of menus and options. This dataset will be using data from four published studies whose data has been archived in the Dryad Digital Repository or another database.

One is a study of floral traits and pollinator visitation in a species of milkweed (*Asclepias syriaca*). The citation is below. Read at least the abstract of the study to understand what the data represents, as well as the README file accompanying the data on Dryad (second link below). The figure below may help you visualize what the variables mean. A simplified version of the data is titled “Asclepias”.

La Rosa RJ, Conner JK (2017) Floral function: effects of traits on pollinators, male and female pollination success, and female fitness across three species of milkweeds (Asclepias). American Journal of Botany 104(1): 150-160. <https://doi.org/10.3732/ajb.1600328>   
README file: <https://doi.org/10.5061/dryad.n81g1>



The second dataset comes from a study of the drivers of tree mortality in comparison to size and age. The citation is below. Read at least the abstract of the study to understand what the data represents. A simplified version of the data with just one tree species is titled “tree\_mortality”.

Cailleret M, Bigler C, Bugmann H, Camarero JJ, Cufar K, Davi H, Mészáros I, Minunno F, Peltoniemi M, Robert E, Suarez M, Tognetti R, Martinez-Vilalta J (2016) Towards a common methodology for developing logistic tree mortality models based on ring-width data. Ecological Applications 26(6): 1827-1841. <https://doi.org/10.1890/15-1402.1>

Dryad Link: <https://doi.org/10.5061/dryad.1bv6n>

To import each of these files into JASP when needed, simply navigate to Open and select the appropriate CSV file. Usually default settings are appropriate.  
  
Before proceeding with any data inspection or analysis in JASP, you must ALWAYS check that the imported variables are correctly coded as either ‘nominal’, ‘ordinal’, or continuous (referred to in JASP as ‘scale’). This can be done using the list of variables on the top of the data table, clicking on the variable and changing the type. This is critical to correct analysis.

Using the provided datasets, answer the following questions:

(1) Open the “Asclepias” dataset. You will be examining correlations among six measured floral traits – *gyn\_width, hood\_length, hood\_height, horn\_reach, slit\_length,* and *gap\_width*. Inspect the distributions of these six variables using the Descriptives menu. Do they look approximately normally distributed based on histograms or Q-Q plots? Note whether each variable appears normally distributed or not.  
  
Now proceed to the Regression > Correlation menu. Here you can perform either Pearson (parametric) or Spearman (nonparametric) correlations, as well as visualize correlations and perform assumption checks (e.g., bivariate normality). Start with calculating pairwise Pearson correlations for all combinations of the six variables. Do any pairs fail the bivariate normality assumption? Excluding those pairwise correlations which fail the assumption check, report a correlation matrix and scatterplots for all pairwise correlations.  
  
For those pairwise correlations that fail the assumption of bivariate normality, perform a Spearman rank correlation and report the results along with a scatterplot.

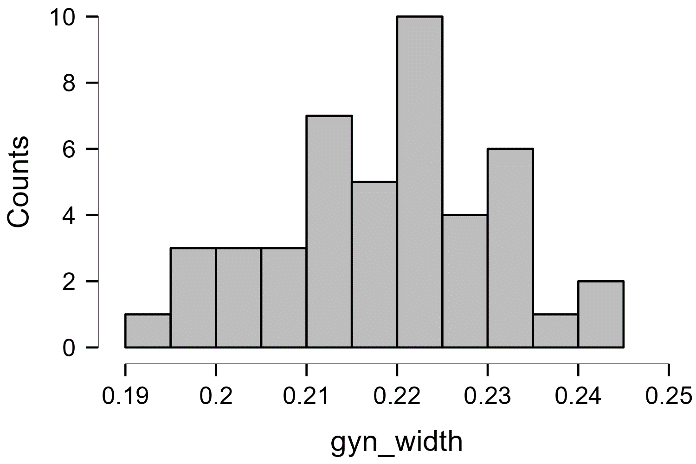
**Results**

**Descriptive Statistics**

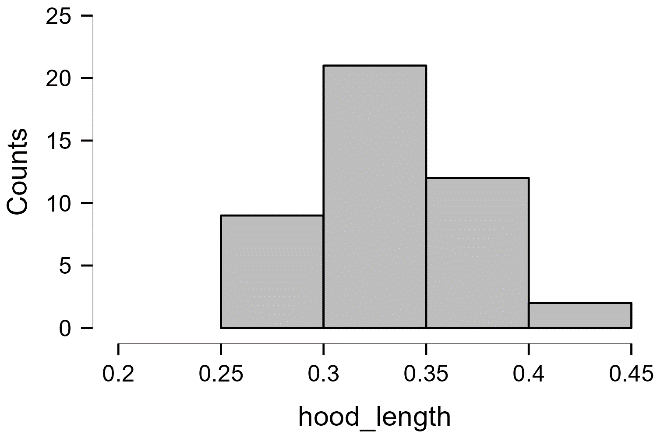
| **Descriptive Statistics** | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **gyn\_width** | | **hood\_length** | | **hood\_height** | | **horn\_reach** | | **slit\_length** | | **gap\_width** | |
| Valid |  | 45 |  | 45 |  | 45 |  | 45 |  | 45 |  | 45 |  |
| Missing |  | 0 |  | 0 |  | 0 |  | 0 |  | 0 |  | 0 |  |
| Mean |  | 0.219 |  | 0.324 |  | 0.508 |  | 0.204 |  | 0.190 |  | 0.061 |  |
| Std. Deviation |  | 0.012 |  | 0.042 |  | 0.044 |  | 0.034 |  | 0.012 |  | 0.012 |  |
| Minimum |  | 0.193 |  | 0.240 |  | 0.424 |  | 0.121 |  | 0.159 |  | 0.027 |  |
| Maximum |  | 0.241 |  | 0.436 |  | 0.604 |  | 0.286 |  | 0.210 |  | 0.084 |  |
|  | | | | | | | | | | | | | |

**Distribution Plots**

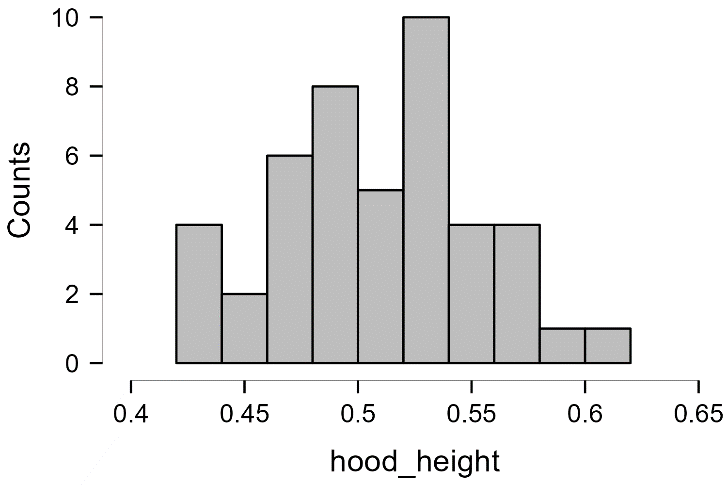
**gyn\_width**



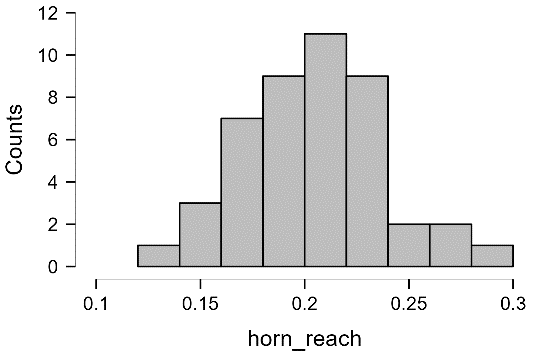
**hood\_length**



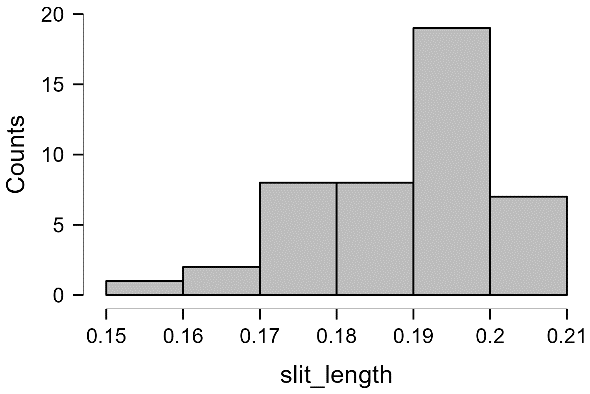
**hood\_height**



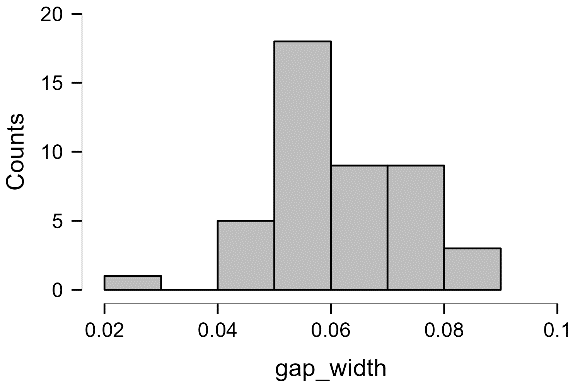
**horn\_reach**



**slit\_length**

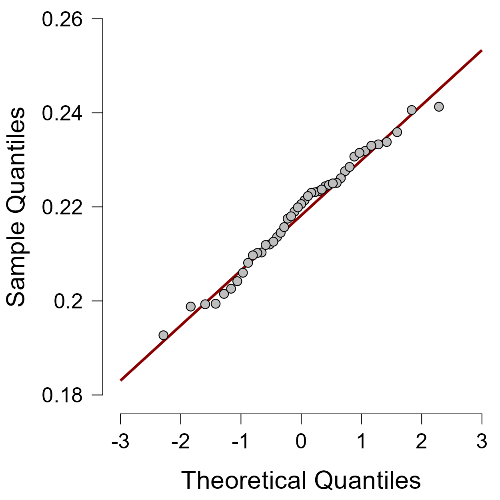


**gap\_width**

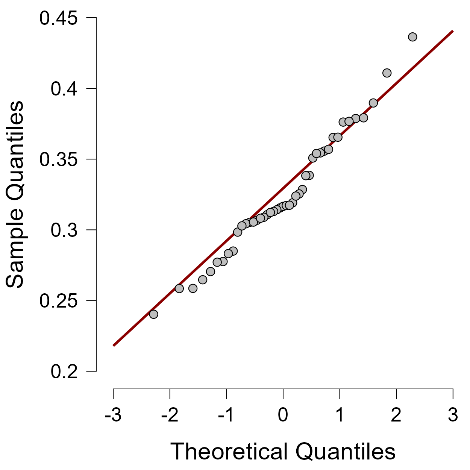


**Q-Q Plots**

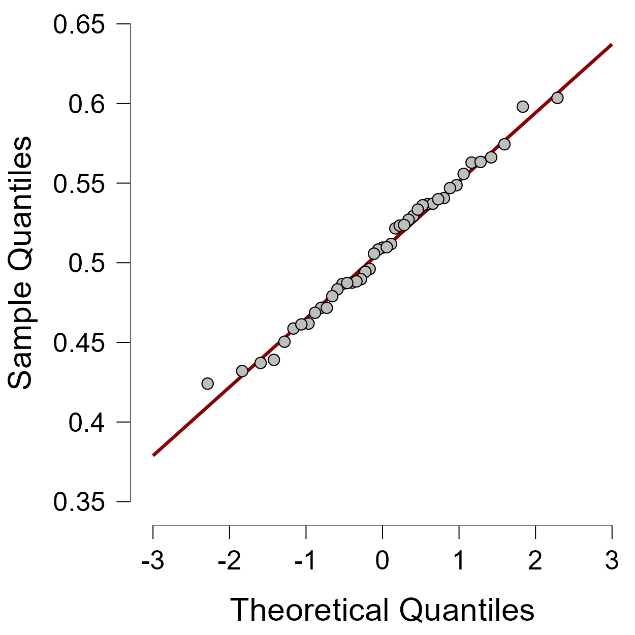
**gyn\_width**



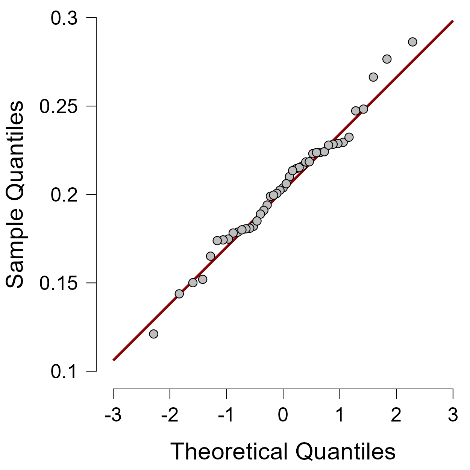
**hood\_length**



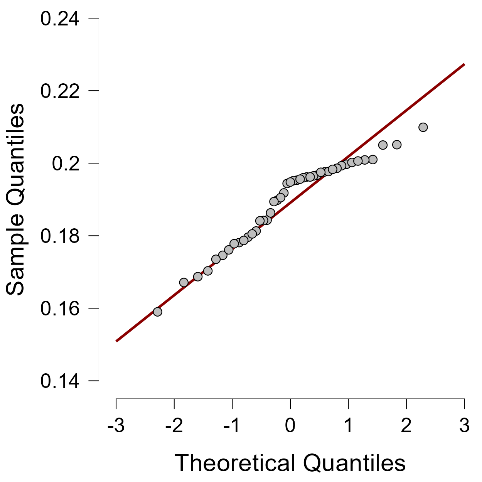
**hood\_height**



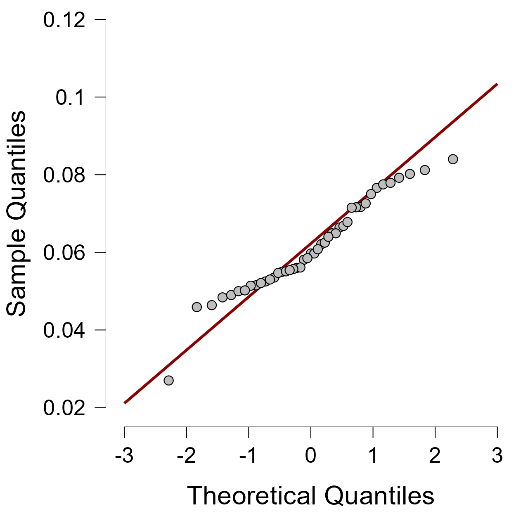
**horn\_reach**



**slit\_length**



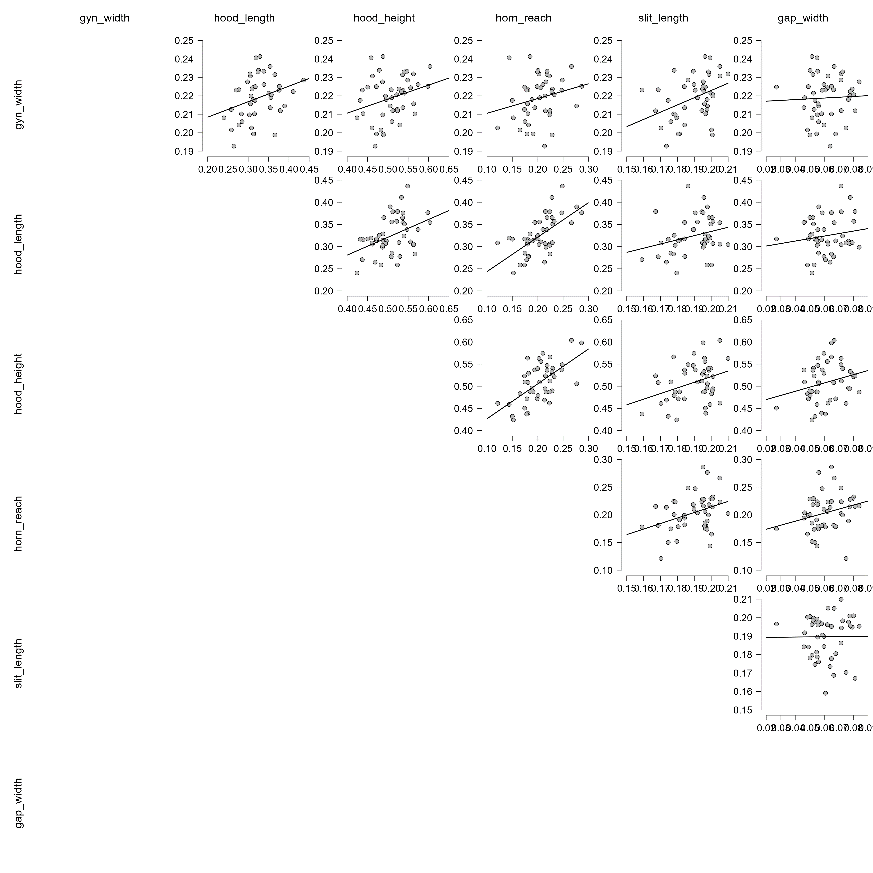
**gap\_width**



**Correlation**

| **Pearson's Correlations** | | | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | |  | | **gyn\_width** | | **hood\_length** | | **hood\_height** | | **horn\_reach** | | **slit\_length** | | **gap\_width** | |
| 1. gyn\_width |  | Pearson's r |  | — |  |  |  |  |  |  |  |  |  |  |  |
|  |  | p-value |  | — |  |  |  |  |  |  |  |  |  |  |  |
| 2. hood\_length |  | Pearson's r |  | 0.299 | \* | — |  |  |  |  |  |  |  |  |  |
|  |  | p-value |  | 0.046 |  | — |  |  |  |  |  |  |  |  |  |
| 3. hood\_height |  | Pearson's r |  | 0.280 |  | 0.418 | \*\* | — |  |  |  |  |  |  |  |
|  |  | p-value |  | 0.062 |  | 0.004 |  | — |  |  |  |  |  |  |  |
| 4. horn\_reach |  | Pearson's r |  | 0.228 |  | 0.626 | \*\*\* | 0.610 | \*\*\* | — |  |  |  |  |  |
|  |  | p-value |  | 0.132 |  | 4.335×10-6 |  | 8.689×10-6 |  | — |  |  |  |  |  |
| 5. slit\_length |  | Pearson's r |  | 0.386 | \*\* | 0.262 |  | 0.340 | \* | 0.346 | \* | — |  |  |  |
|  |  | p-value |  | 0.009 |  | 0.082 |  | 0.022 |  | 0.020 |  | — |  |  |  |
| 6. gap\_width |  | Pearson's r |  | 0.042 |  | 0.157 |  | 0.252 |  | 0.253 |  | 0.008 |  | — |  |
|  |  | p-value |  | 0.783 |  | 0.302 |  | 0.094 |  | 0.093 |  | 0.960 |  | — |  |
|  | | | | | | | | | | | | | | | |
| \* p < .05, \*\* p < .01, \*\*\* p < .001 | | | | | | | | | | | | | | | |

**Correlation plot**



**Descriptive Statistics**

| **Descriptive Statistics** | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **gyn\_width** | | **hood\_length** | | **hood\_height** | | **horn\_reach** | | **slit\_length** | | **gap\_width** | |
| Valid |  | 45 |  | 45 |  | 45 |  | 45 |  | 45 |  | 45 |  |
| Missing |  | 0 |  | 0 |  | 0 |  | 0 |  | 0 |  | 0 |  |
| Mean |  | 0.219 |  | 0.324 |  | 0.508 |  | 0.204 |  | 0.190 |  | 0.061 |  |
| Std. Deviation |  | 0.012 |  | 0.042 |  | 0.044 |  | 0.034 |  | 0.012 |  | 0.012 |  |
| Minimum |  | 0.193 |  | 0.240 |  | 0.424 |  | 0.121 |  | 0.159 |  | 0.027 |  |
| Maximum |  | 0.241 |  | 0.436 |  | 0.604 |  | 0.286 |  | 0.210 |  | 0.084 |  |
|  | | | | | | | | | | | | | |

(2) Continue using the “Asclepias” dataset. You will now be conducting linear regression, first conducting simple linear regression (ordinary least squares) with just two traits. Navigate to Regression > Linear Regression. Build a linear regression model where *hood height* is the dependent variable and *horn reach* is the independent variable. Before interpreting output, first check assumptions and report both residual-versus-predicted and quantile-quantile plots for standardized residuals. Do residuals appear homoscedastic and normally distributed, as is required for linear regression?  
  
If this regression model meets the assumptions for linear regression, report the regression equation (y=mx+b), R2 and p-value for this regression. What does this regression seem to mean biologically? Interpret the slope coefficient.

**Linear Regression**

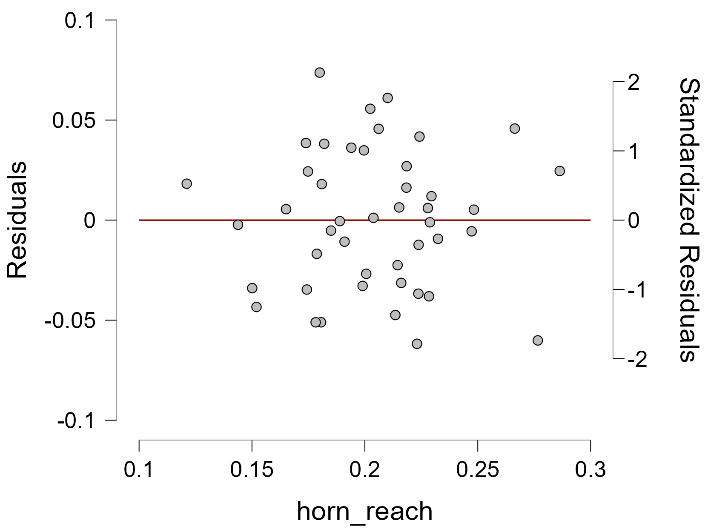
| **Model Summary - hood\_height** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | | **R** | | **R²** | | **Adjusted R²** | | **RMSE** | |
| H₀ |  | 0.000 |  | 0.000 |  | 0.000 |  | 0.044 |  |
| H₁ |  | 0.610 |  | 0.372 |  | 0.357 |  | 0.035 |  |
|  | | | | | | | | | |

| **ANOVA** | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | |  | | **Sum of Squares** | | **df** | | **Mean Square** | | **F** | | **p** | |
| H₁ |  | Regression |  | 0.031 |  | 1 |  | 0.031 |  | 25.468 |  | 8.689×10-6 |  |
|  |  | Residual |  | 0.053 |  | 43 |  | 0.001 |  |  |  |  |  |
|  |  | Total |  | 0.084 |  | 44 |  |  |  |  |  |  |  |
|  | | | | | | | | | | | | | |
| *Note.*  The intercept model is omitted, as no meaningful information can be shown. | | | | | | | | | | | | | |

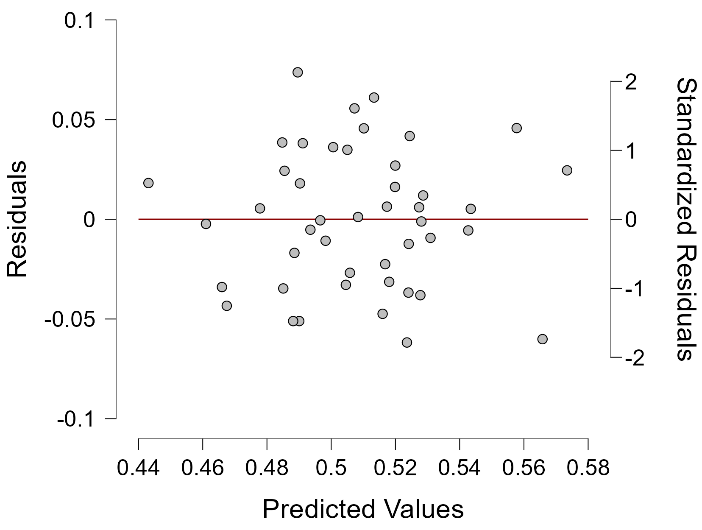
| **Coefficients** | | | | | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | **95% CI** | | | |
| **Model** | |  | | **Unstandardized** | | **Standard Error** | | **Standardized** | | **t** | | **p** | | **Lower** | | **Upper** | |
| H₀ |  | (Intercept) |  | 0.508 |  | 0.007 |  |  |  | 78.120 |  | 7.672×10-49 |  | 0.495 |  | 0.522 |  |
| H₁ |  | (Intercept) |  | 0.348 |  | 0.032 |  |  |  | 10.755 |  | 9.056×10-14 |  | 0.282 |  | 0.413 |  |
|  |  | horn\_reach |  | 0.789 |  | 0.156 |  | 0.610 |  | 5.047 |  | 8.689×10-6 |  | 0.474 |  | 1.104 |  |
|  | | | | | | | | | | | | | | | | | |

**Residuals vs. Covariates**

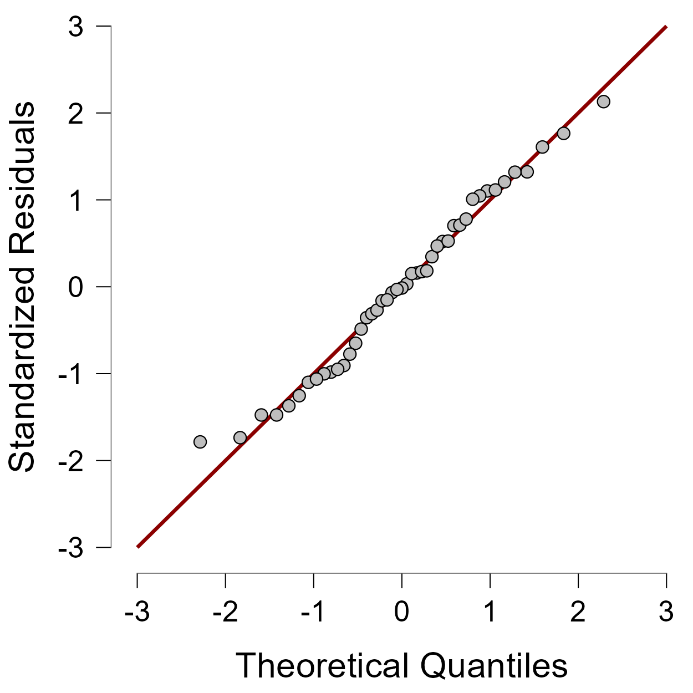
**Residuals vs. horn\_reach**



**Residuals vs. Predicted**



**Q-Q Plot Standardized Residuals**



(3) Continue using the “Asclepias” dataset. You will now be conducting multiple regression! For this species, you will be predicting the number of *fruits* produced by a plant as a function of the other variables assessed (floral traits, pollinator visitation rates, etc). To do this, we will use stepwise regression and model selection to build the best model to predict fruit production. Before you do so, you need to check for multicollinearity. To do this, navigate to Regression > Correlation. Include all continuous variables, and generate pairwise Pearson correlation coefficients for all variable combinations, as well as a scatterplot matrix. Inspect the scatterplot matrix to check for multicollinearity among variables, and report it below. Are any variables collinear?  
  
Now that that’s done, navigate to Regression > Linear Regression. Place *fruits* as the dependent variable, and all of the other continuous variables as predictor variables as predictor variables under ‘covariates’. The Method permits you to specify ‘enter’ (manual entry of predictors), ‘forward’ (forward stepwise), ‘backward’ (reverse stepwise), or ‘stepwise’(same as forward but with a removal test performed at each stage to remove the least predictive variable). First try ‘enter’(manual entry) to run a full model with all predictor variables. Report the R2, adjusted R2 (penalized for the number of variables) p-value of the full model, and coefficients table. Note which predictor variables have coefficients that are not different from zero using the estimated 95% confidence interval.  
  
Now try manually adding and removing variables to build a model that has the highest R2 with the fewest variables. Notice how the variable statistics change when you include or exclude variables. Once you feel comfortable with how adding and removing variables affects the model, try automatic stopping-rule methods like ‘forward’, ‘reverse’, and ‘stepwise’. Don’t assume that the automated Forward/Backward search will find the best model!  
  
Once you think you’ve found the best model, check assumptions and report both residual-versus-predicted and quantile-quantile plots for standardized residuals. Do residuals appear homoscedastic and normally distributed, as is required for linear regression?   
  
If you are happy with your best model in terms of assumptions being met, report the regression equation (y=mx+b), R2 and p-value of the ANOVA (a test of absolute fit) for this regression, and the coefficient table with 95% confidence intervals. In a few sentences, assess how good this multiple regression is, and interpret what each term in the regression equation means biologically using the coefficients.

**Linear Regression**

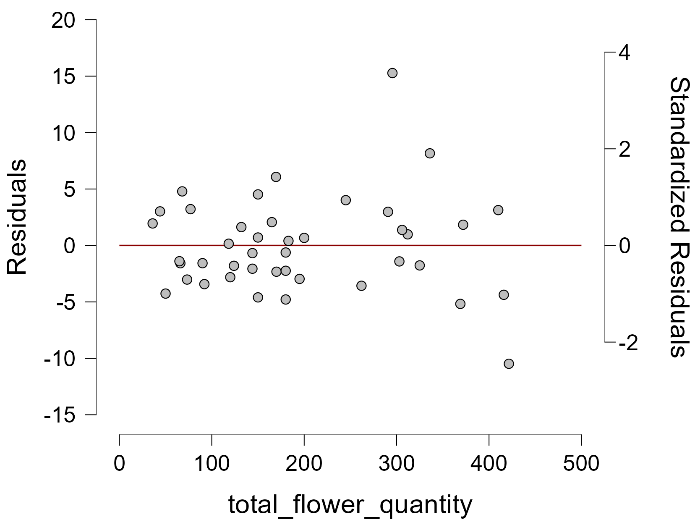
| **Model Summary - fruits** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | | **R** | | **R²** | | **Adjusted R²** | | **RMSE** | |
| 1 |  | 0.000 |  | 0.000 |  | 0.000 |  | 5.605 |  |
| 2 |  | 0.487 |  | 0.237 |  | 0.218 |  | 4.956 |  |
| 3 |  | 0.645 |  | 0.416 |  | 0.386 |  | 4.391 |  |
|  | | | | | | | | | |

| **ANOVA** | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | |  | | **Sum of Squares** | | **df** | | **Mean Square** | | **F** | | **p** | |
| 2 |  | Regression |  | 305.297 |  | 1 |  | 305.297 |  | 12.428 |  | 0.001 |  |
|  |  | Residual |  | 982.608 |  | 40 |  | 24.565 |  |  |  |  |  |
|  |  | Total |  | 1287.905 |  | 41 |  |  |  |  |  |  |  |
| 3 |  | Regression |  | 535.925 |  | 2 |  | 267.962 |  | 13.897 |  | 2.775×10-5 |  |
|  |  | Residual |  | 751.980 |  | 39 |  | 19.282 |  |  |  |  |  |
|  |  | Total |  | 1287.905 |  | 41 |  |  |  |  |  |  |  |
|  | | | | | | | | | | | | | |
| *Note.*  The intercept model is omitted, as no meaningful information can be shown. | | | | | | | | | | | | | |

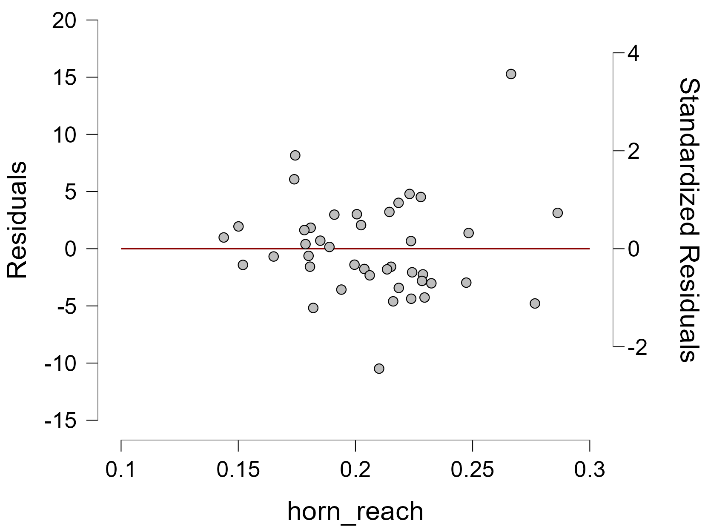
| **Coefficients** | | | | | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | | | | | | | **95% CI** | | | |
| **Model** | |  | | **Unstandardized** | | **Standard Error** | | **Standardized** | | **t** | | **p** | | **Lower** | | **Upper** | |
| 1 |  | (Intercept) |  | 5.952 |  | 0.865 |  |  |  | 6.883 |  | 2.412×10-8 |  | 4.206 |  | 7.699 |  |
| 2 |  | (Intercept) |  | 1.200 |  | 1.550 |  |  |  | 0.774 |  | 0.444 |  | -1.933 |  | 4.332 |  |
|  |  | total\_flower\_quantity |  | 0.024 |  | 0.007 |  | 0.487 |  | 3.525 |  | 0.001 |  | 0.010 |  | 0.038 |  |
| 3 |  | (Intercept) |  | -13.920 |  | 4.582 |  |  |  | -3.038 |  | 0.004 |  | -23.188 |  | -4.651 |  |
|  |  | total\_flower\_quantity |  | 0.023 |  | 0.006 |  | 0.465 |  | 3.795 |  | 5.021×10-4 |  | 0.011 |  | 0.036 |  |
|  |  | horn\_reach |  | 74.156 |  | 21.442 |  | 0.424 |  | 3.458 |  | 0.001 |  | 30.786 |  | 117.526 |  |
|  | | | | | | | | | | | | | | | | | |
| *Note.*  The following covariates were considered but not included: poll\_visits\_per10min\_noflies, display\_flowers\_1day. | | | | | | | | | | | | | | | | | |

**Residuals vs. Covariates**

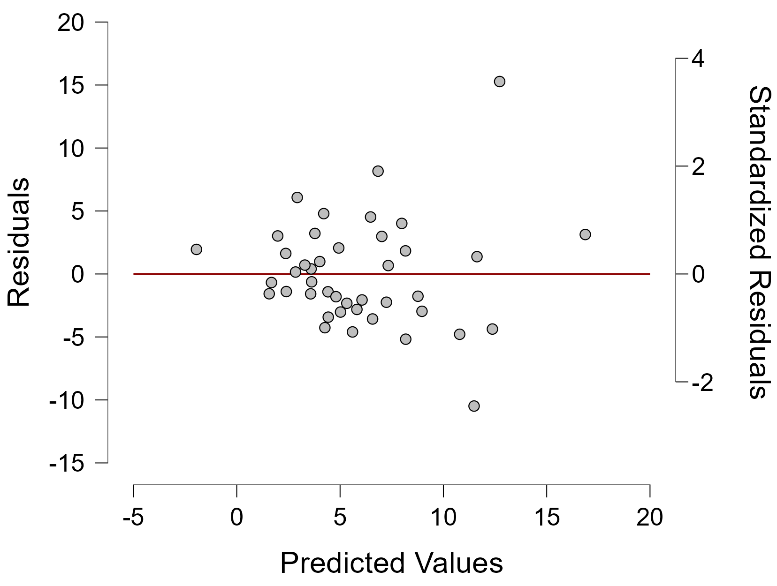
**Residuals vs. total\_flower\_quantity**



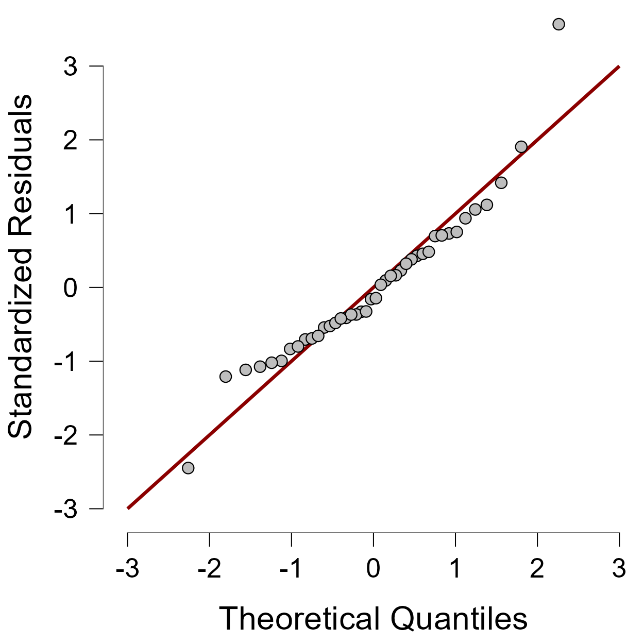
**Residuals vs. horn\_reach**



**Residuals vs. Predicted**



**Q-Q Plot Standardized Residuals**



(4) Now, open the “tree\_mortality” dataset. You are going to be assessing whether age or diameter at breast height (DBH, in cm) predicts whether a silver fir tree is dead or alive (categorical data) in the subsequent year. To do this, navigate to Regression > Logistic Regression and select *status* as the response variable and use *age* and *DBH* and site as predictors. This will run a logistic regression of a model with all three predictors. A Wald chi-Square goodness-of-fit test tells you whether the overall model (and each variable) explains significant variation in mortality. An AIC value is generated that can let you compare your model to a null model as well as any nested model (model with a subset of terms relative to the full model). A confusion matrix can also show you the performance of the model in assigning observations correctly versus incorrectly, and calculate sensitivity (% true positives) and specificity (% true negatives) of the model.

|  |
| --- |
| For the full model, is this model any good at prediction based on the Wald chi-squared and/or AIC relative to a null model? If so, which predictor variable(s) are the best predictors of mortality? For the continuous variables, interpret the biology of their directionality.  Now, try other models either manually or using one of the other Methods (forward, backward, stepwise). Report the best model of mortality you can build based on AIC, chi-squared, and confusion matrix results. |

**Logistic Regression**

| **Model Summary - status** | | | | | | | | | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | | **Deviance** | | **AIC** | | **BIC** | | **df** | | **Χ²** | | **p** | | **McFadden R²** | | **Nagelkerke R²** | | **Tjur R²** | | **Cox & Snell R²** | |
| H₀ |  | 684.247 |  | 686.247 |  | 690.493 |  | 515 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| H₁ |  | 579.927 |  | 609.927 |  | 673.619 |  | 501 |  | 104.320 |  | 6.661×10-16 |  | 0.152 |  | 0.249 |  | 0.188 |  | 0.183 |  |
|  | | | | | | | | | | | | | | | | | | | | | |

| **Coefficients** | | | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | **Wald Test** | | | | | |
|  | | **Estimate** | | **Standard Error** | | **z** | | **Wald Statistic** | | **df** | | **p** | |
| (Intercept) |  | 0.069 |  | 0.939 |  | 0.073 |  | 0.005 |  | 1 |  | 0.942 |  |
| age |  | -0.008 |  | 0.004 |  | -2.033 |  | 4.134 |  | 1 |  | 0.042 |  |
| DBH |  | 0.069 |  | 0.012 |  | 5.666 |  | 32.102 |  | 1 |  | 1.463×10-8 |  |
| site (Canalicchio) |  | -2.968 |  | 0.853 |  | -3.478 |  | 12.100 |  | 1 |  | 5.043×10-4 |  |
| site (Issole2) |  | -0.187 |  | 0.893 |  | -0.209 |  | 0.044 |  | 1 |  | 0.834 |  |
| site (Lopet) |  | -0.496 |  | 1.125 |  | -0.441 |  | 0.195 |  | 1 |  | 0.659 |  |
| site (Paco\_Ezpela\_High) |  | -0.832 |  | 1.045 |  | -0.796 |  | 0.634 |  | 1 |  | 0.426 |  |
| site (Paco\_Ezpela\_Low) |  | -1.720 |  | 0.962 |  | -1.789 |  | 3.199 |  | 1 |  | 0.074 |  |
| site (Ravnik) |  | -1.905 |  | 0.841 |  | -2.264 |  | 5.127 |  | 1 |  | 0.024 |  |
| site (Ventoux\_Dvx1) |  | -0.871 |  | 0.922 |  | -0.945 |  | 0.892 |  | 1 |  | 0.345 |  |
| site (Ventoux\_Dvx2) |  | -0.979 |  | 0.894 |  | -1.095 |  | 1.199 |  | 1 |  | 0.273 |  |
| site (Ventoux\_Dvx3) |  | -0.758 |  | 0.910 |  | -0.833 |  | 0.694 |  | 1 |  | 0.405 |  |
| site (Ventoux\_TC) |  | -0.445 |  | 0.797 |  | -0.559 |  | 0.312 |  | 1 |  | 0.576 |  |
| site (Vesubie3) |  | -0.906 |  | 0.884 |  | -1.024 |  | 1.049 |  | 1 |  | 0.306 |  |
| site (Vesubie6) |  | 0.416 |  | 1.308 |  | 0.318 |  | 0.101 |  | 1 |  | 0.750 |  |
|  | | | | | | | | | | | | | |
| *Note.*  status level 'LIVING' coded as class 1. | | | | | | | | | | | | | |

| **Factor Descriptives** | | | |
| --- | --- | --- | --- |
| **site** | | **N** | |
| Bistra |  | 19 |  |
| Canalicchio |  | 66 |  |
| Issole2 |  | 27 |  |
| Lopet |  | 9 |  |
| Paco\_Ezpela\_High |  | 10 |  |
| Paco\_Ezpela\_Low |  | 12 |  |
| Ravnik |  | 32 |  |
| Ventoux\_Dvx1 |  | 20 |  |
| Ventoux\_Dvx2 |  | 30 |  |
| Ventoux\_Dvx3 |  | 25 |  |
| Ventoux\_TC |  | 227 |  |
| Vesubie3 |  | 26 |  |
| Vesubie6 |  | 13 |  |
|  | | | |

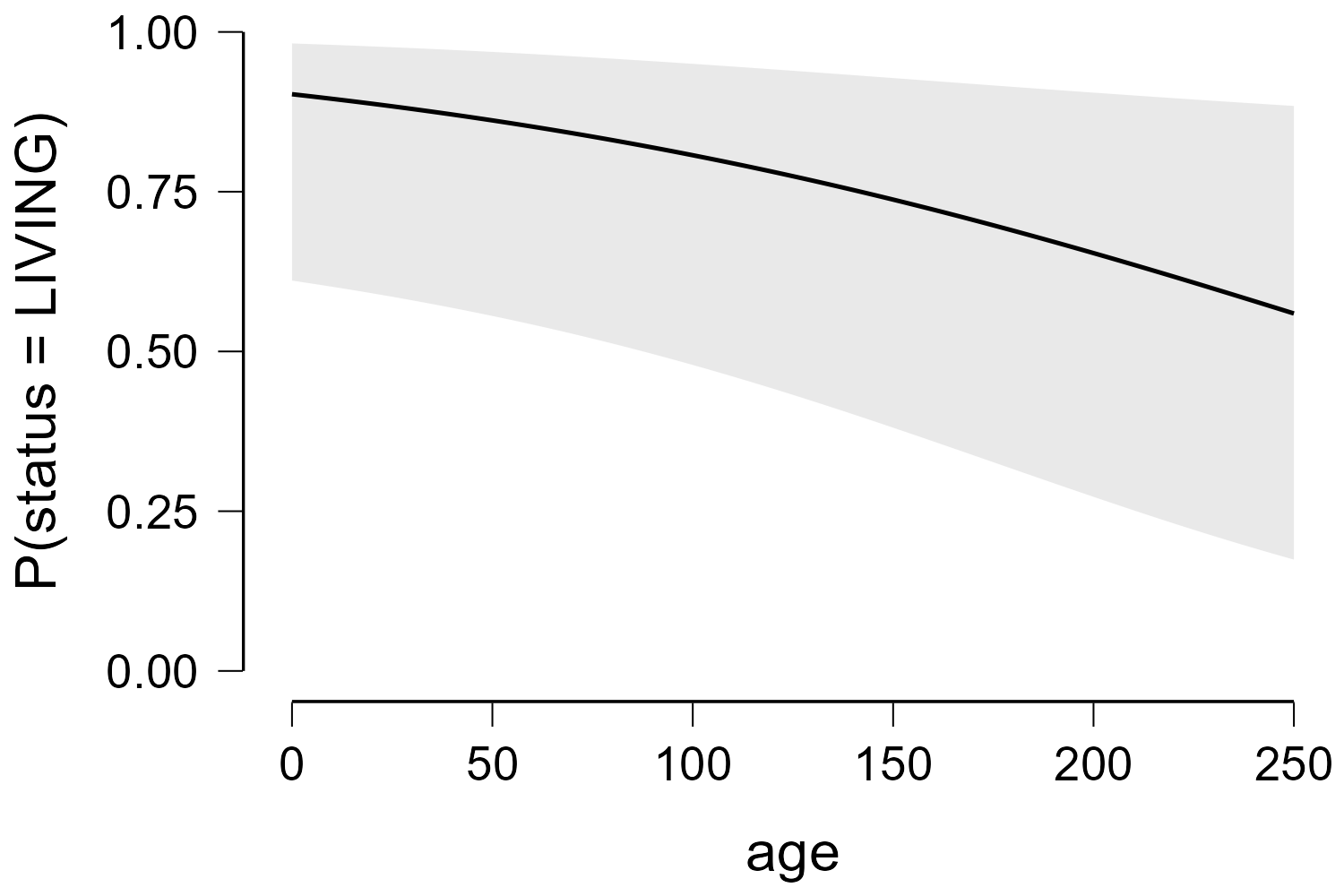
**Performance Diagnostics**

| **Confusion matrix** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Predicted** | | | |  | |
| **Observed** | | **DEAD** | | **LIVING** | | **% Correct** | |
| DEAD |  | 67 |  | 128 |  | 34.359 |  |
| LIVING |  | 16 |  | 305 |  | 95.016 |  |
| Overall % Correct |  |  |  |  |  | 72.093 |  |
|  | | | | | | | |
| *Note.*  The cut-off value is set to 0.5 | | | | | | | |

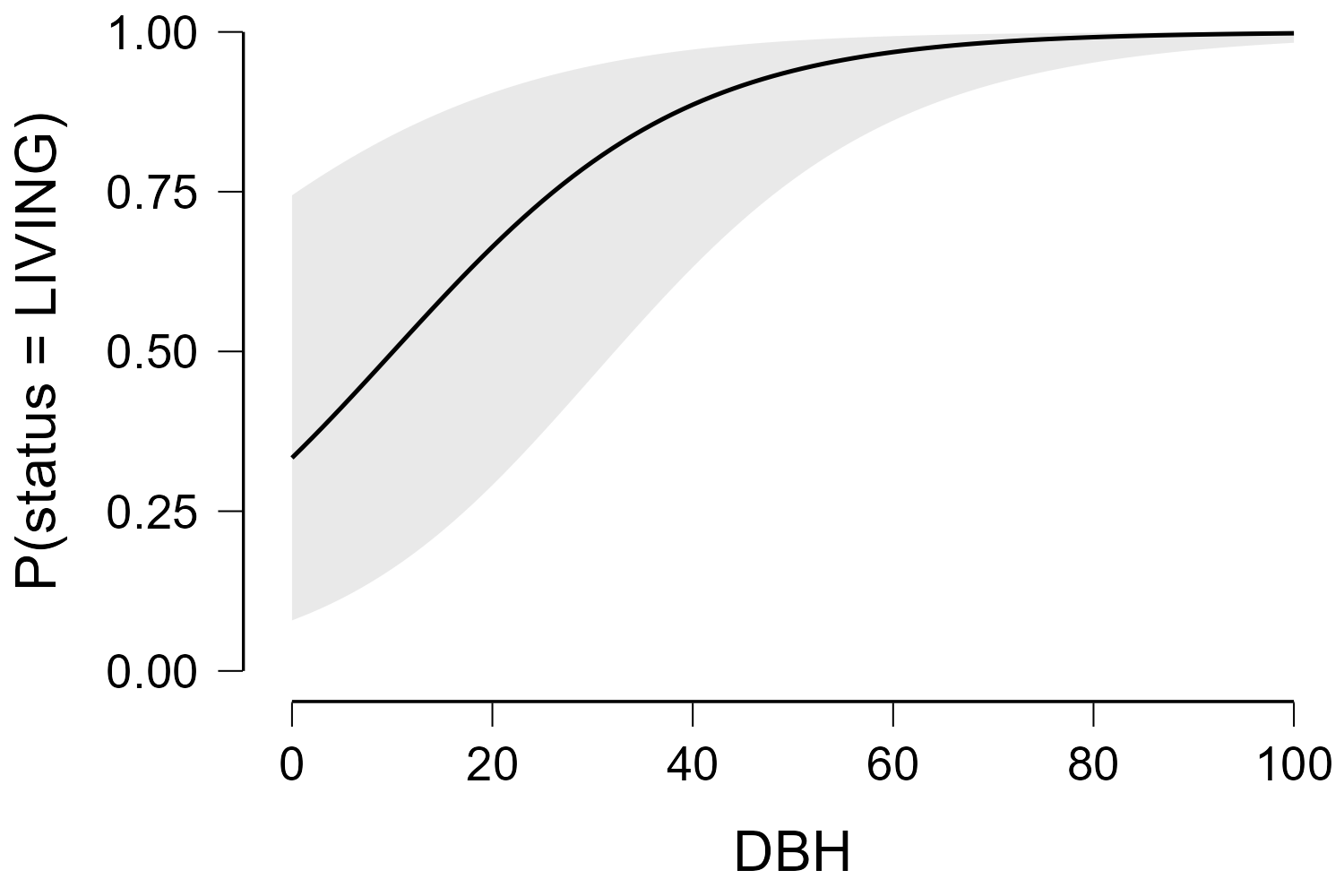
| **Performance metrics** | | | |
| --- | --- | --- | --- |
|  | | **Value** | |
| Sensitivity |  | 0.950 |  |
| Specificity |  | 0.344 |  |
|  | | | |

**Estimates plots**

**age**



**DBH**



**site**

