HW10

Chapters 11 and 12

Your Name Here

The code below just loads some packages and makes it so that enough digits are printed that you won't get confused by rounding errors.

```
library(dplyr) # functions like summarize
library(ggplot2) # for making plots
library(gridExtra)
library(GGally)
library(readr)
library(car)

options("pillar.sigfig" = 10) # print 10 significant digits in summarize output
```

Problem 1: Galapagos (Adapted from Sleuth3 12.20)

Quote from book:

> The data [read in below] come from a 1973 study. (Data from M. P. Johnson and P. H. Raven, "Species Number and Endemism: The Galapagos Archipelago Revisited," *Science* 179 (1973): 893-5.) The number of species on an island is known to be related to the island's area. Of interest is what other variables are also related to the number of species, after island area is accounted for.

The data set includes the following variables:

- Island: a character vector indicating the island
- Total: total number of observed species
- Native: number of native species
- Area: area (km^2)
- Elev: elevation (m)
- DistNear: distance from nearest island (km)
- DistSc: distance from Santa Cruz (km)
- AreaNear: area of nearest island (km^2)

In this analysis, our response variable is `Native`, the number of native species. You will use `Area`, `Elev`, `DistNear`, `DistSc`, and `AreaNear` as possible explanatory variables.

```
galapagos <- read_csv("http://www.evanlray.com/data/sleuth3/ex1220_galapagos.csv")
```

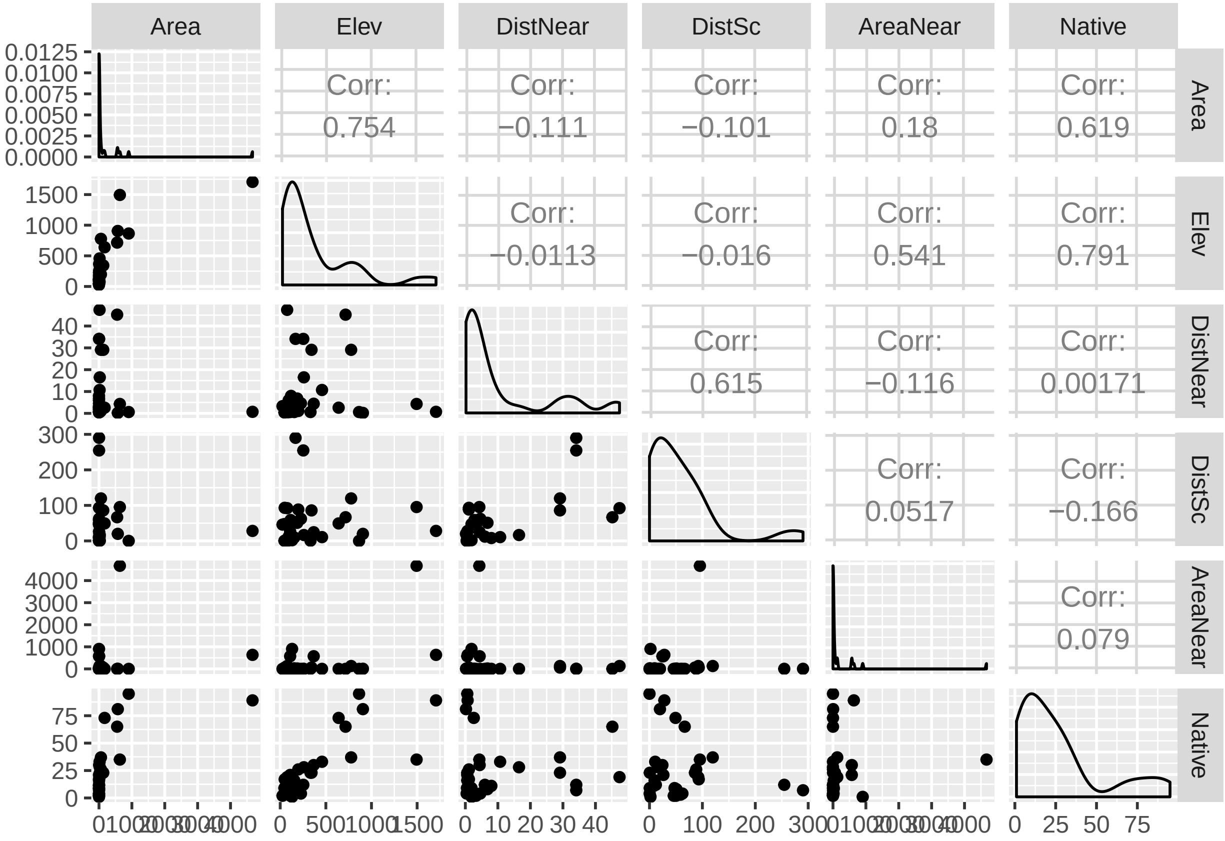
```
## Parsed with column specification:
## cols(
##   Island = col_character(),
##   Total = col_double(),
##   Native = col_double(),
##   Area = col_double(),
##   Elev = col_double(),
##   DistNear = col_double(),
##   DistSc = col_double(),
##   AreaNear = col_double()
## )
```

```
head(galapagos)
```

```
## # A tibble: 6 x 8
##   Island       Total Native  Area  Elev DistNear DistSc AreaNear
##   <chr>        <dbl>  <dbl> <dbl> <dbl>    <dbl>  <dbl>    <dbl>
## 1 Baltra          58     23 25.09   332      0.6    0.6     1.84
## 2 Bartolome       31     21  1.24   109      0.6   26.3   572.33
## 3 Caldwell         3      3  0.21   114      2.8   58.7     0.78
## 4 Champion        25      9  0.1     46      1.9   47.4     0.18
## 5 Coamano          2      1  1.05   130      1.9    1.9   903.82
## 6 Daphne Major    18     11  0.34   119      8      8       1.84
```

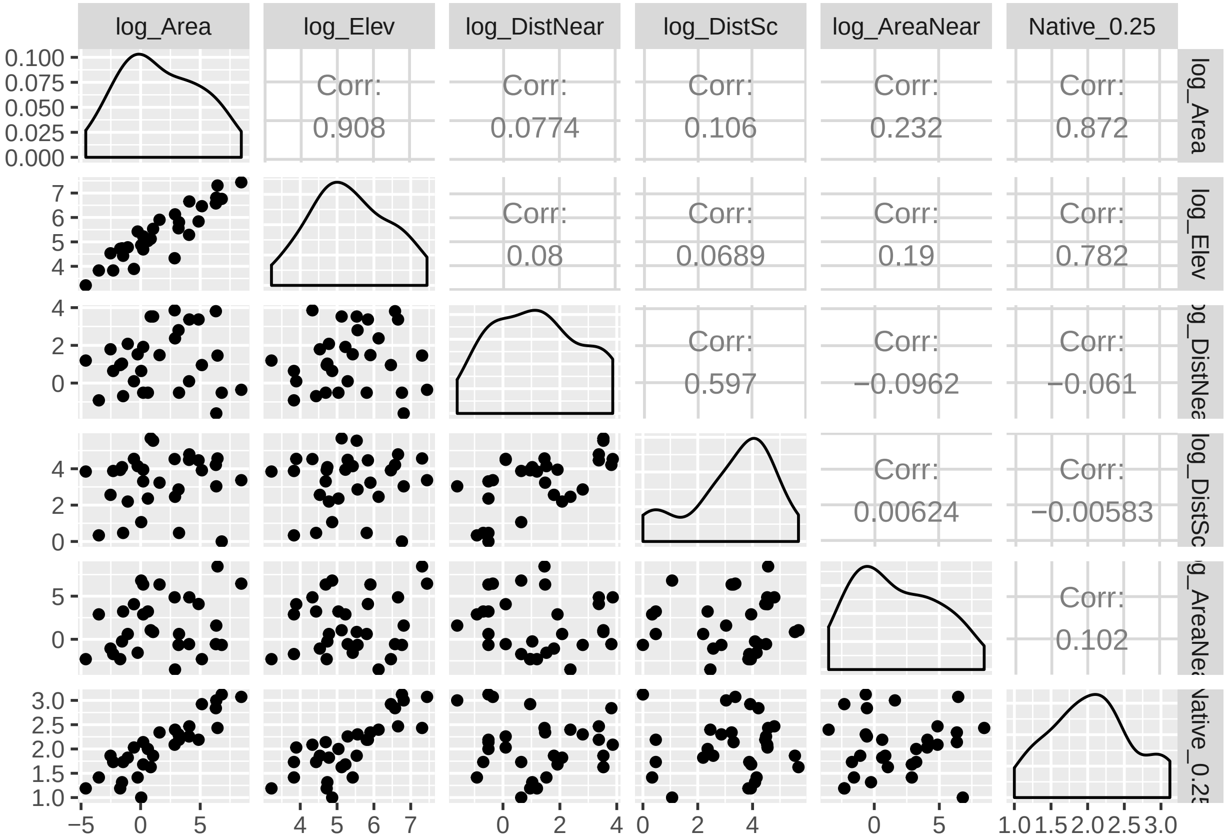
(a) Make a pairs plot of the data including only the variables you will use in your analysis (with the response variable last).

```
ggpairs(galapagos %>% select(Area, Elev, DistNear, DistSc, AreaNear, Native))
```



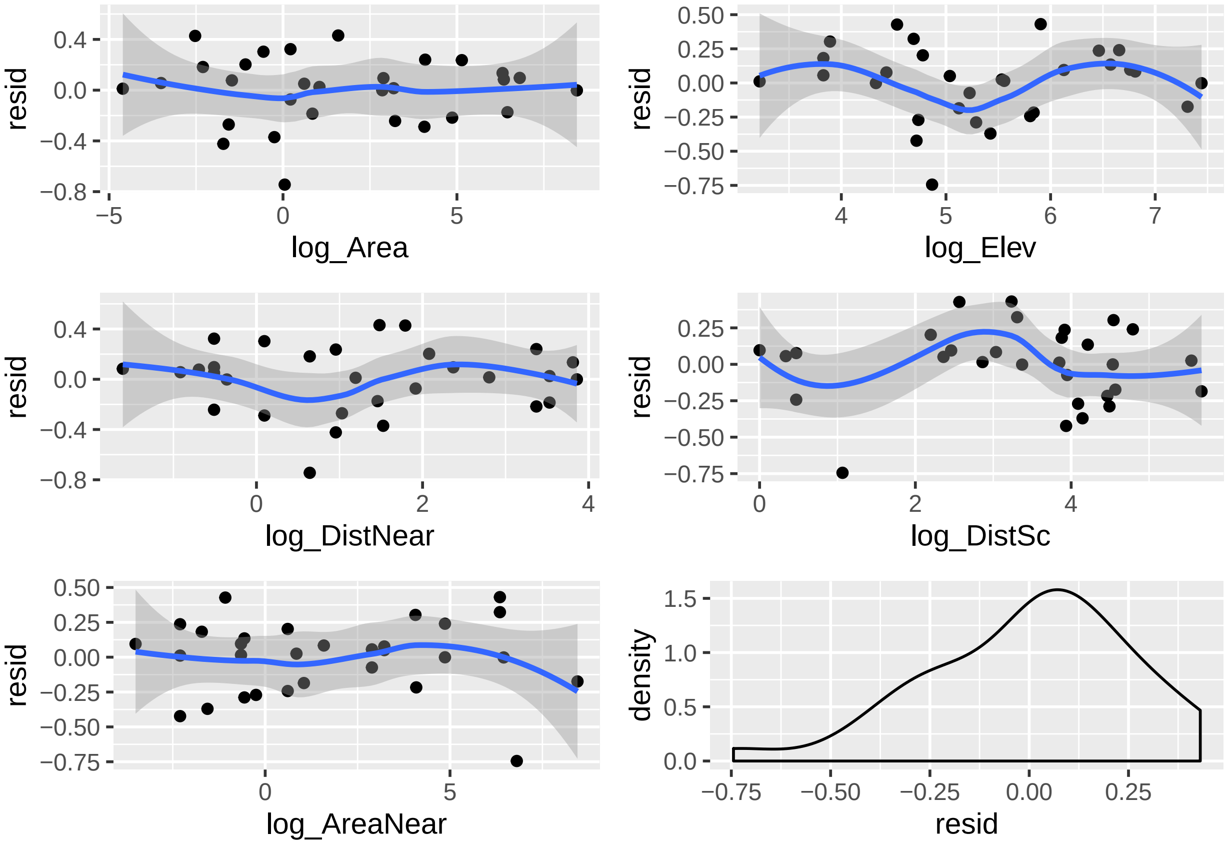
(b) Identify a set of transformations for all variables in the model so that the regression conditions appear to be fairly well satisfied. In doing this, consider pairs plots of the transformed data and plots of residuals vs explanatory variables in a regression model including all transformed explanatory variables. (You should have both of these types of plots.) After this step, you should feel fairly confident that any models you fit will either have approximately linear relationships among transformed variables, or know how you will handle non-linearity by adding polynomial terms in the model. You should also feel feel confident that the variance of residuals is fairly constant across values of explanatory variables. Note that `DistSc` includes some 0 values. A common trick in cases like this is to add 1 to the observed values of that variable before doing transformations (this means that things like a log transformation are an option).

```r
galapagos_transformed <- galapagos %>%
  transmute(
    log_Area = log(Area),
    log_Elev = log(Elev),
    log_DistNear = log(DistNear),
    log_DistSc = log(DistSc + 1),
    log_AreaNear = log(AreaNear),
    Native_0.25 = Native^0.25
  )

ggpairs(galapagos_transformed)
```



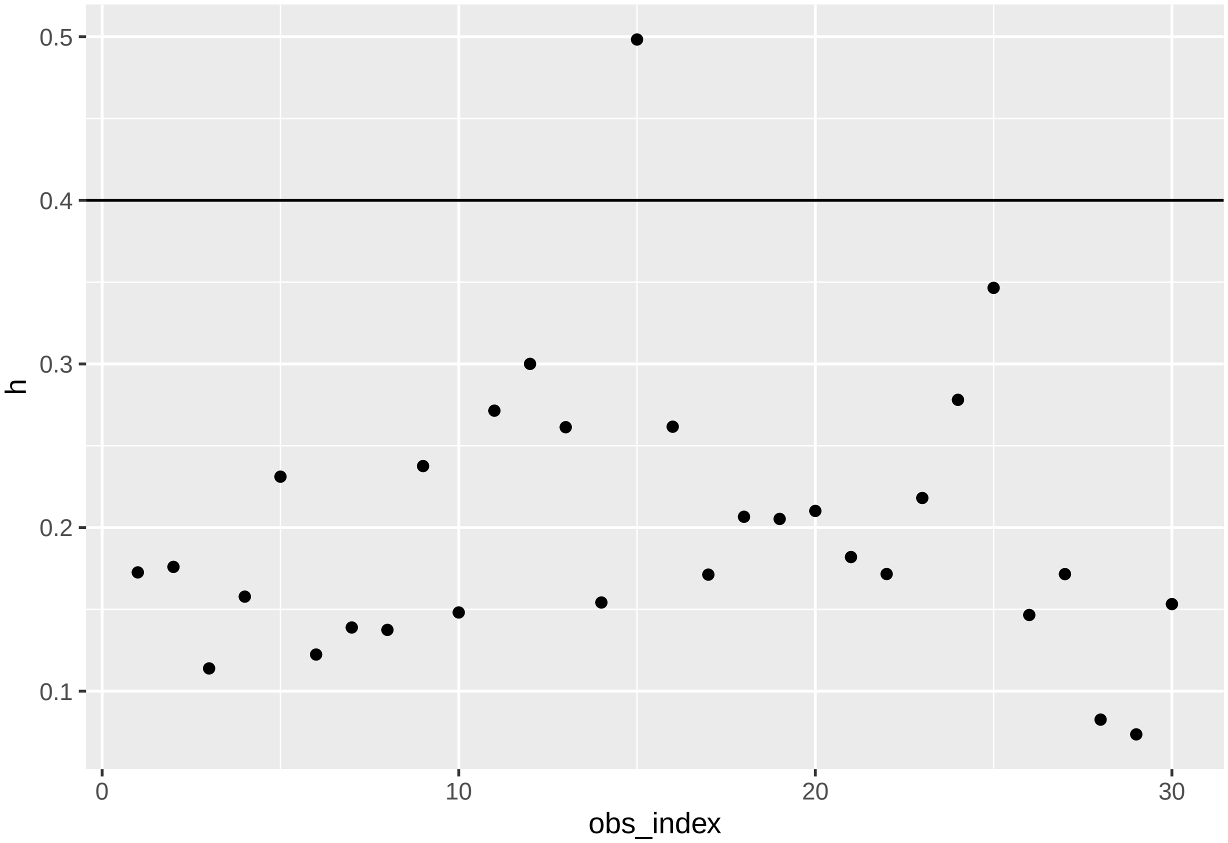
```r
lm_full <- lm(Native_0.25 ~ log_Area + log_Elev + log_DistNear + log_DistSc + log_AreaNear,
  data = galapagos_transformed)
galapagos_transformed <- galapagos_transformed %>%
  mutate(
    resid = residuals(lm_full)
  )

p1 <- ggplot(data = galapagos_transformed, mapping = aes(x = log_Area, y = resid)) +
  geom_point() +
  geom_smooth()
p2 <- ggplot(data = galapagos_transformed, mapping = aes(x = log_Elev, y = resid)) +
  geom_point() +
  geom_smooth()
p3 <- ggplot(data = galapagos_transformed, mapping = aes(x = log_DistNear, y = resid)) +
  geom_point() +
  geom_smooth()
p4 <- ggplot(data = galapagos_transformed, mapping = aes(x = log_DistSc, y = resid)) +
  geom_point() +
  geom_smooth()
p5 <- ggplot(data = galapagos_transformed, mapping = aes(x = log_AreaNear, y = resid)) +
  geom_point() +
  geom_smooth()
p6 <- ggplot(data = galapagos_transformed, mapping = aes(x = resid)) +
  geom_density()
grid.arrange(p1, p2, p3, p4, p5, p6)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

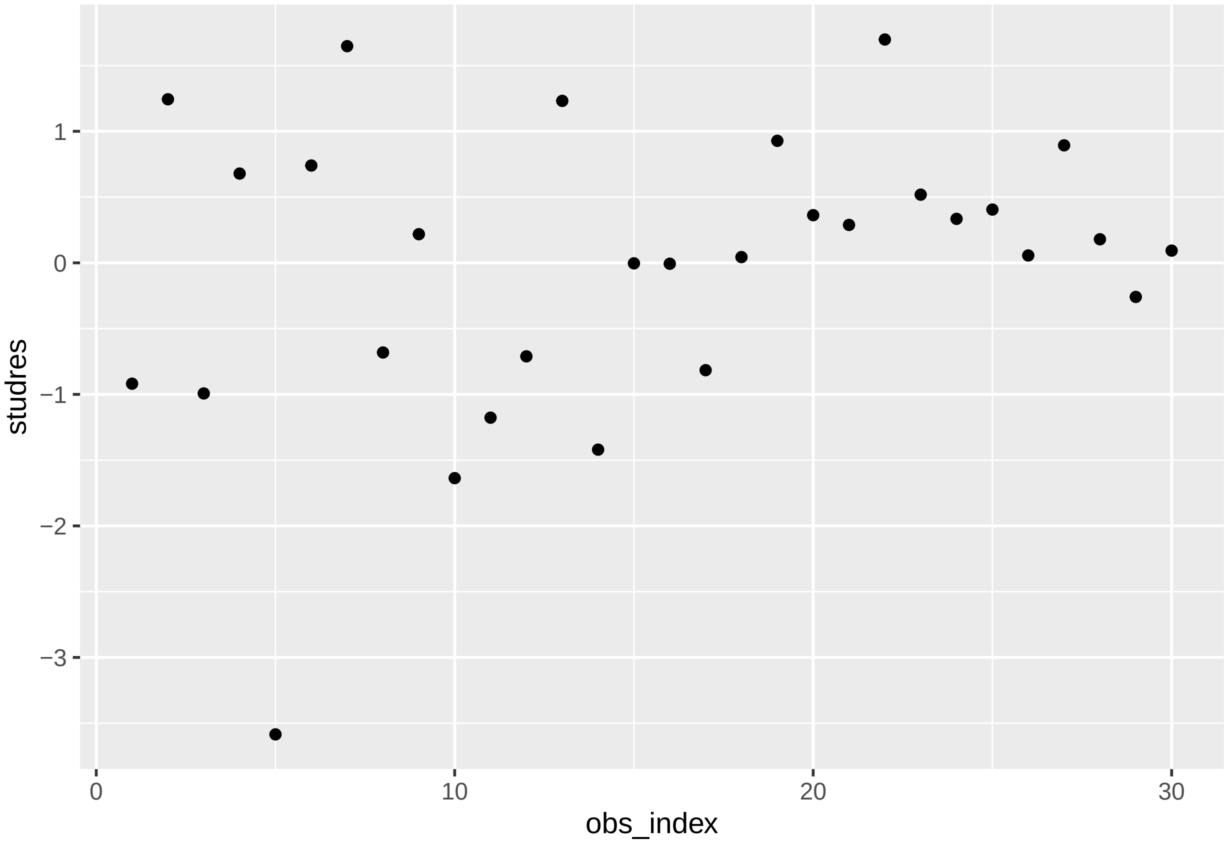


(c) Check for influential observations, outliers or high leverage observations.

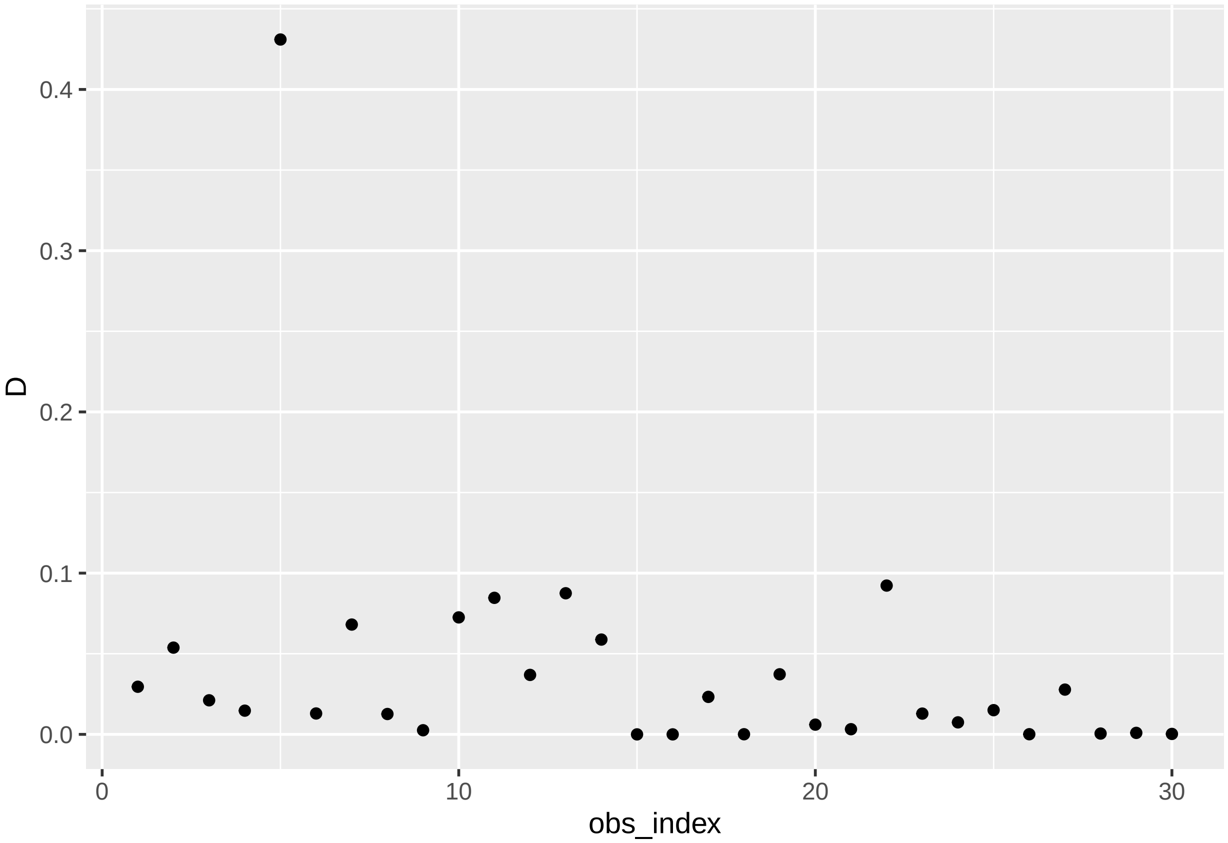
```
galapagos_transformed <- galapagos_transformed %>%
  mutate(
    obs_index = row_number(),
    h = hatvalues(lm_full),
    studres = rstudent(lm_full),
    D = cooks.distance(lm_full)
  )

ggplot(data = galapagos_transformed, mapping = aes(x = obs_index, y = h)) +
  geom_hline(yintercept = 2 * 6 / nrow(galapagos_transformed))+
  geom_point()
```



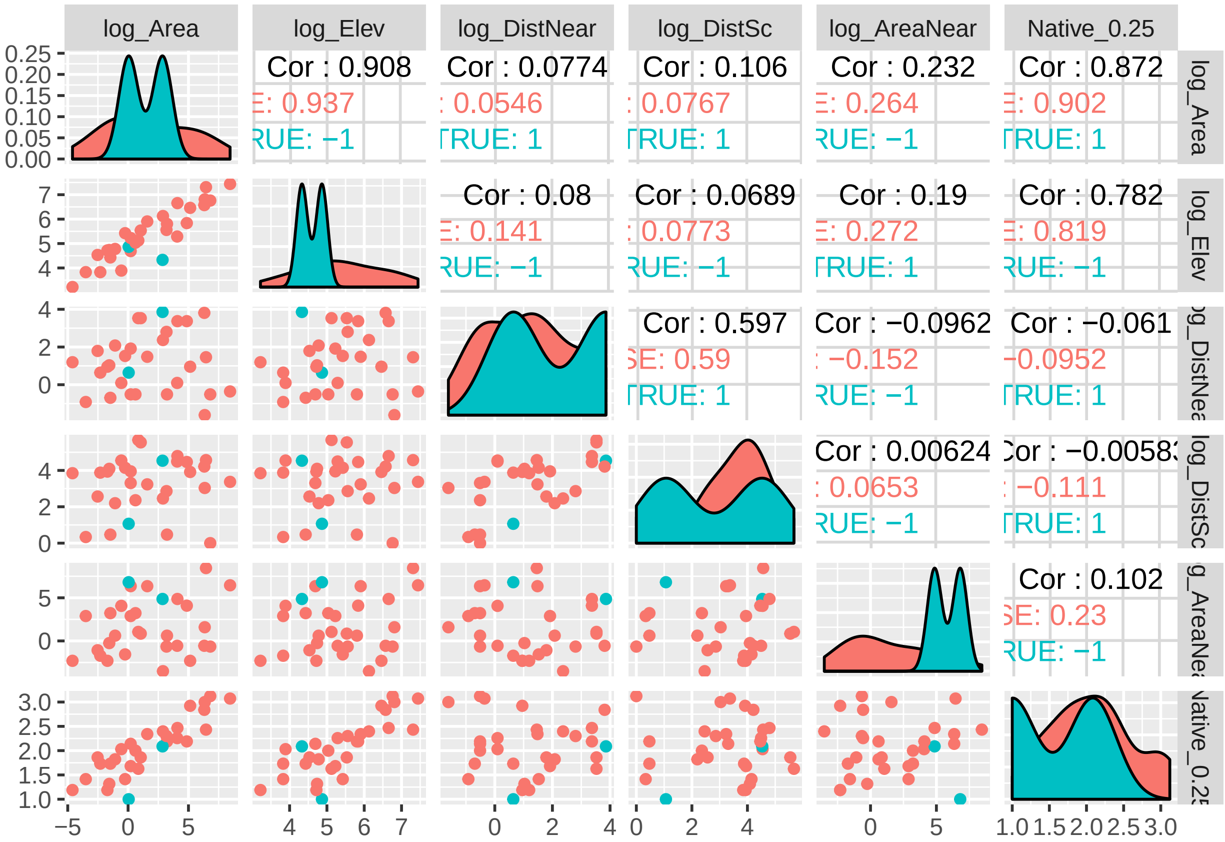
```
ggplot(data = galapagos_transformed, mapping = aes(x = obs_index, y = studres)) +
  geom_point()
```



```
ggplot(data = galapagos_transformed, mapping = aes(x = obs_index, y = D)) +
  geom_point()
```

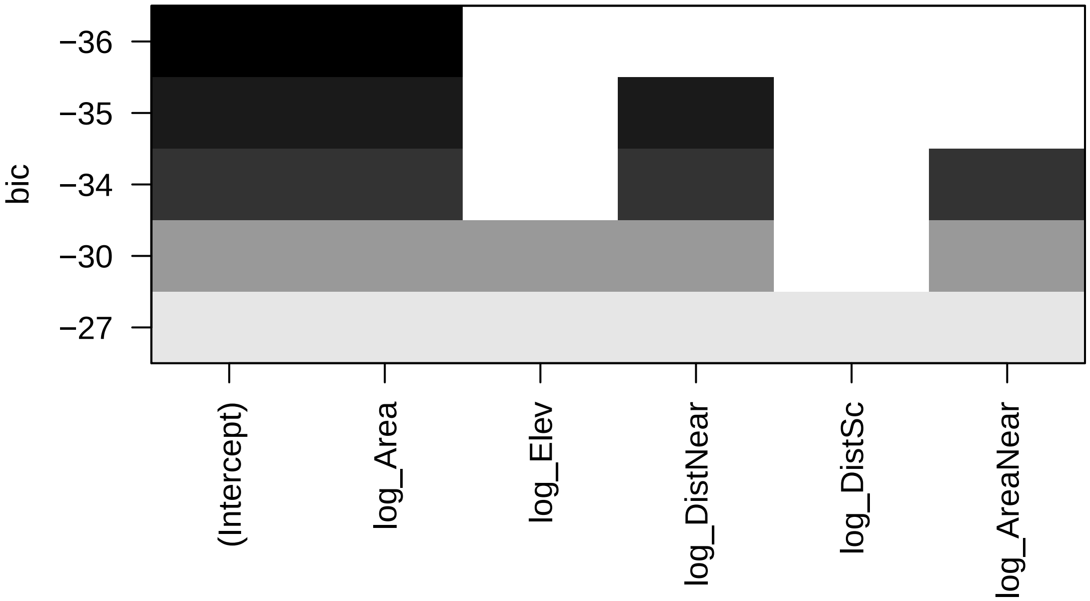


```
galapagos_transformed <- galapagos_transformed %>%
  mutate(suspicious = obs_index %in% c(5, 15))

ggpairs(galapagos_transformed, mapping = aes(color = suspicious), columns = 1:6)
```



(d) Use all subsets regression to identify a set of models with similar ability to model these data well, based on your transformed variables. If necessary, perform this step both with and without the outliers or influential observations included.

```
library(leaps)
candidate_models <- regsubsets(Native_0.25 ~ log_Area + log_Elev + log_DistNear + log_DistSc + log_AreaN
  data = galapagos_transformed)
plot(candidate_models)
```



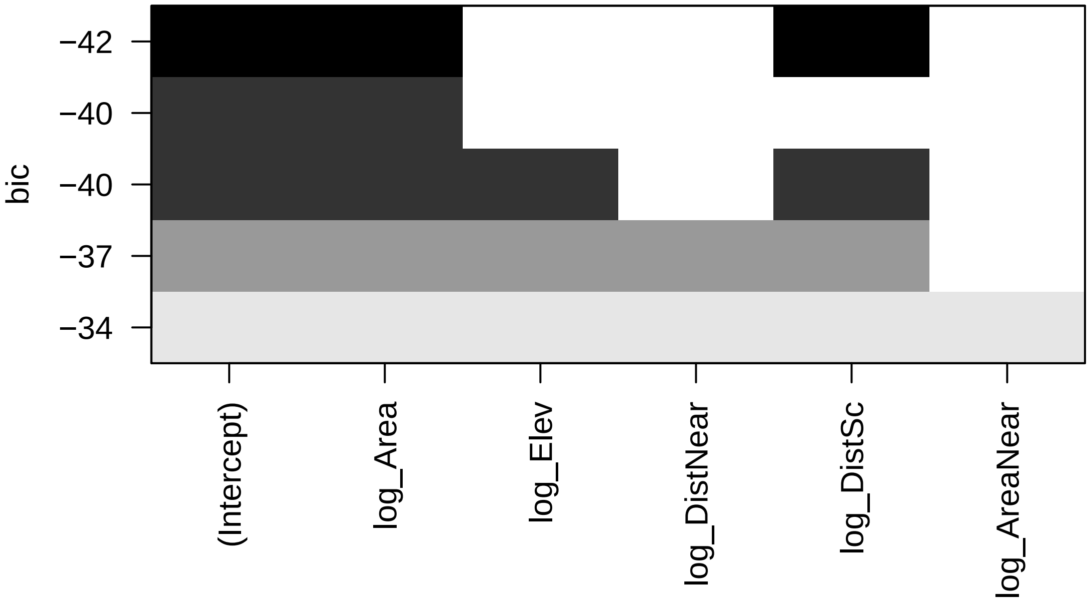
```
summary(candidate_models)
```

```
## Subset selection object
## Call: regsubsets.formula(Native_0.25 ~ log_Area + log_Elev + log_DistNear +
##     log_DistSc + log_AreaNear, data = galapagos_transformed)
## 5 Variables  (and intercept)
##              Forced in Forced out
## log_Area         FALSE      FALSE
## log_Elev         FALSE      FALSE
## log_DistNear     FALSE      FALSE
## log_DistSc       FALSE      FALSE
## log_AreaNear     FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          log_Area log_Elev log_DistNear log_DistSc log_AreaNear
## 1  ( 1 ) "*"      " "      " "          " "        " "         
## 2  ( 1 ) "*"      " "      "*"          " "        " "         
## 3  ( 1 ) "*"      " "      "*"          " "        "*"         
## 4  ( 1 ) "*"      "*"      "*"          " "        "*"         
## 5  ( 1 ) "*"      "*"      "*"          "*"        "*"         
```

```
summary(candidate_models)$bic
```

```
## [1] -36.15035 -34.91195 -33.51806 -30.22219 -26.89620
```

```
candidate_models2 <- regsubsets(Native_0.25 ~ log_Area + log_Elev + log_DistNear + log_DistSc + log_Are
  data = galapagos_transformed %>% filter(!suspicious))
plot(candidate_models2)
```



```r
summary(candidate_models2)
```

```
## Subset selection object
## Call: regsubsets.formula(Native_0.25 ~ log_Area + log_Elev + log_DistNear +
##     log_DistSc + log_AreaNear, data = galapagos_transformed %>%
##     filter(!suspicious))
## 5 Variables  (and intercept)
##              Forced in Forced out
## log_Area         FALSE      FALSE
## log_Elev         FALSE      FALSE
## log_DistNear     FALSE      FALSE
## log_DistSc       FALSE      FALSE
## log_AreaNear     FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          log_Area log_Elev log_DistNear log_DistSc log_AreaNear
## 1  ( 1 ) "*"      " "      " "          " "        " "
## 2  ( 1 ) "*"      " "      " "          "*"        " "
## 3  ( 1 ) "*"      "*"      " "          "*"        " "
## 4  ( 1 ) "*"      "*"      "*"          "*"        " "
## 5  ( 1 ) "*"      "*"      "*"          "*"        "*"
```

```r
summary(candidate_models2)$bic
```

```
## [1] -40.33967 -42.38225 -40.04481 -36.84316 -33.51199
```

When all observations are included, the following three models have roughly similar performance:

Model 1: log Area, log DistNear, and log AreaNear as explanatory variables

Model 2: log Area and log DistNear as explanatory variables

Model 3: log Area as the only explanatory variable

When one outlier and one high leverage observation are omitted, the following three models have roughly similar performance:

Model 1: log Area, log DistSc, and log Elev as explanatory variables

Model 2: log Area and log DistSc as explanatory variables

Model 3: log Area as the only explanatory variable

(e) Obtain the model fits for all models you identified in part (d) as explaining the data about as well as each other, and print the model summaries.

```
fit1 <- lm(Native_0.25 ~ log_Area + log_DistNear + log_AreaNear, data = galapagos_transformed)
summary(fit1)
```

```
##
## Call:
## lm(formula = Native_0.25 ~ log_Area + log_DistNear + log_AreaNear,
##     data = galapagos_transformed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72395 -0.19768  0.05207  0.17589  0.42334
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.89972    0.07354  25.834  < 2e-16 ***
## log_Area      0.15514    0.01572   9.872 2.77e-10 ***
## log_DistNear -0.05232    0.03292  -1.589    0.124
## log_AreaNear -0.02200    0.01640  -1.341    0.191
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.279 on 26 degrees of freedom
## Multiple R-squared:  0.7921, Adjusted R-squared:  0.7681
## F-statistic: 33.02 on 3 and 26 DF,  p-value: 5.091e-09
```

```
fit2 <- lm(Native_0.25 ~ log_Area + log_DistNear, data = galapagos_transformed)
summary(fit2)
```

```
##
## Call:
## lm(formula = Native_0.25 ~ log_Area + log_DistNear, data = galapagos_transformed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84087 -0.19373  0.04474  0.21246  0.46082
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.86380    0.06949  26.820  < 2e-16 ***
## log_Area      0.15005    0.01547   9.696 2.74e-10 ***
## log_DistNear -0.04713    0.03317  -1.421    0.167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2831 on 27 degrees of freedom
## Multiple R-squared:  0.7777, Adjusted R-squared:  0.7613
## F-statistic: 47.24 on 2 and 27 DF,  p-value: 1.524e-09
```

```
fit3 <- lm(Native_0.25 ~ log_Area, data = galapagos_transformed)
summary(fit3)
```

```
##
## Call:
## lm(formula = Native_0.25 ~ log_Area, data = galapagos_transformed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81591 -0.15749  0.05753  0.23017  0.42722
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.80867    0.05869  30.818  < 2e-16 ***
## log_Area     0.14834    0.01571   9.445 3.34e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2882 on 28 degrees of freedom
## Multiple R-squared:  0.7611, Adjusted R-squared:  0.7526
## F-statistic: 89.21 on 1 and 28 DF,  p-value: 3.342e-10
```

```
fit1a <- lm(Native_0.25 ~ log_Area + log_DistSc + log_AreaNear, data = galapagos_transformed %>% filter
summary(fit1a)
```

```
##
## Call:
## lm(formula = Native_0.25 ~ log_Area + log_DistSc + log_AreaNear,
##     data = galapagos_transformed %>% filter(!suspicious))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36159 -0.15993  0.00683  0.14532  0.36227
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.068e+00  1.095e-01  18.879 6.61e-16 ***
## log_Area      1.471e-01  1.337e-02  11.004 7.36e-11 ***
## log_DistSc   -6.729e-02  2.989e-02  -2.251   0.0338 *
## log_AreaNear -1.513e-05  1.477e-02  -0.001   0.9992
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2347 on 24 degrees of freedom
## Multiple R-squared:  0.846,  Adjusted R-squared:  0.8267
## F-statistic: 43.94 on 3 and 24 DF,  p-value: 6.651e-10
```

```
fit2a <- lm(Native_0.25 ~ log_Area + log_DistSc, data = galapagos_transformed %>% filter(!suspicious))
summary(fit2a)
```

```
##
## Call:
## lm(formula = Native_0.25 ~ log_Area + log_DistSc, data = galapagos_transformed %>%
##     filter(!suspicious))
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36155 -0.15995  0.00682  0.14536  0.36234
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.06787    0.10681   19.36  < 2e-16 ***
## log_Area     0.14711    0.01265   11.63 1.39e-11 ***
## log_DistSc  -0.06729    0.02926   -2.30   0.0301 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.23 on 25 degrees of freedom
## Multiple R-squared:  0.846,  Adjusted R-squared:  0.8337
## F-statistic: 68.66 on 2 and 25 DF,  p-value: 6.994e-11
```

```
fit3a <- lm(Native_0.25 ~ log_Area, data = galapagos_transformed %>% filter(!suspicious))
summary(fit3a)
```

```
##
## Call:
## lm(formula = Native_0.25 ~ log_Area, data = galapagos_transformed %>%
##     filter(!suspicious))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41110 -0.18423  0.04419  0.22051  0.37839
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.84874    0.05213   35.47  < 2e-16 ***
## log_Area     0.14488    0.01361   10.64 5.67e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2482 on 26 degrees of freedom
## Multiple R-squared:  0.8134, Adjusted R-squared:  0.8062
## F-statistic: 113.3 on 1 and 26 DF,  p-value: 5.674e-11
```

(f) Summarize what your analysis has to say about the association of each of the explanatory variables in the data set with the response, after accounting for the explanatory variables in your models. Indicate which of your findings are consistent across the various models considered and which depend on the details of your analysis.

All of the models with low BIC showed very strong evidence of a positive association between an island's area and the number of native species found on the island, among islands similar to those in this study. This result held whether or not one high leverage observation and one outlier were included. If those two observations were removed, there was some weak evidence of a negative association between distance from Santa Cruz and the number of native species in the population of islands similar to those in this study, after accounting for the area of the island; this finding is not reliable since it depends on the removal of those two observations. After accounting for the size of the island, there was not evidence of an association between any of the other explanatory variables and the number of native species on the island.

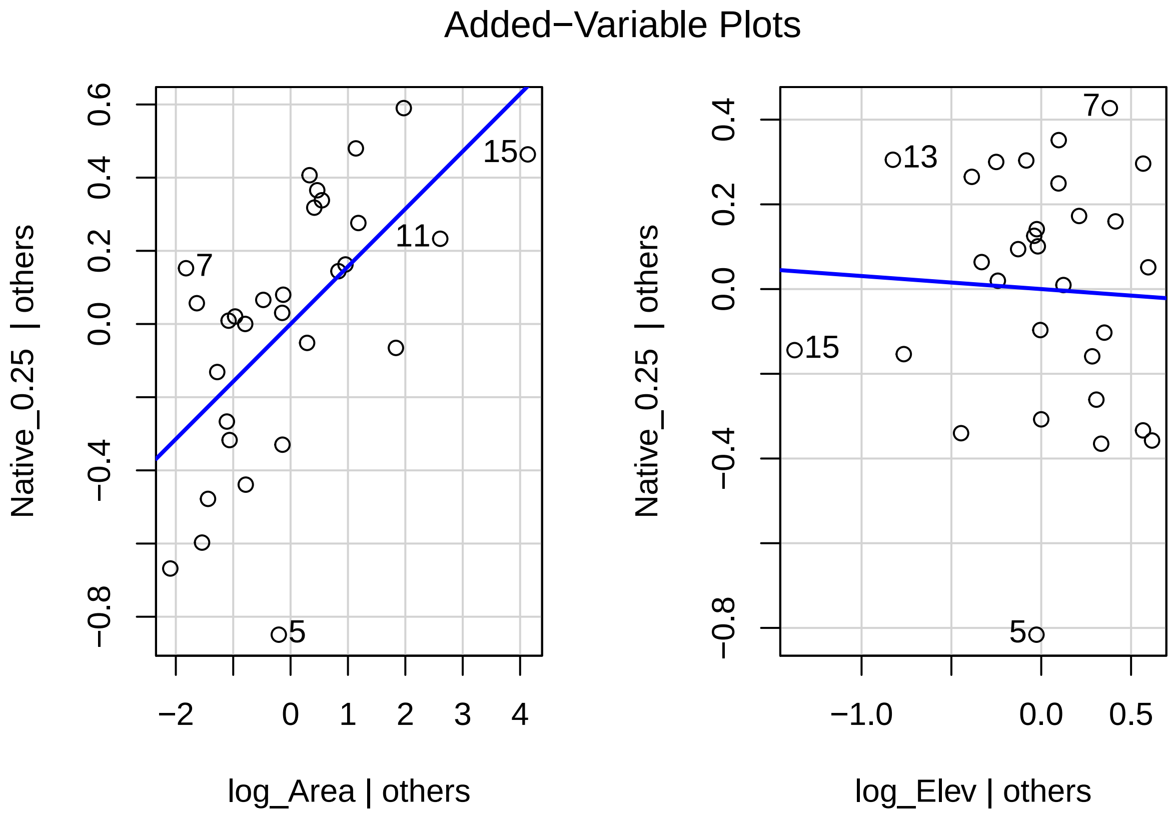
(g) In this part we'll think through what's going on in a model that includes only your transformed `Elev` variable and your transformed `Area` variable.

i. Fit a model that has your (potentially transformed) `Native` as the response and your (potentially transformed) `Elev` and `Area` variables as explanatory variables. Print the model summary and also use the `avPlots` function to create added variables plots for these variables.

```
fit_both <- lm(Native_0.25 ~ log_Area + log_Elev, data = galapagos_transformed)
summary(fit_both)
```

```
##
## Call:
## lm(formula = Native_0.25 ~ log_Area + log_Elev, data = galapagos_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8167 -0.1842  0.0616  0.2340  0.4390
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.95900    0.58985   3.321 0.002579 **
## log_Area     0.15724    0.03821   4.115 0.000327 ***
## log_Elev    -0.03081    0.12025  -0.256 0.799759
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2931 on 27 degrees of freedom
## Multiple R-squared:  0.7617, Adjusted R-squared:  0.744
## F-statistic: 43.15 on 2 and 27 DF,  p-value: 3.903e-09
```

```
avPlots(fit_both)
```



ii. Fit a model that has (potentially transformed) `Elev` as the response and (potentially transformed) `Area` as the only explanatory variable. Add the residuals from this model to your data set with transformed variables.

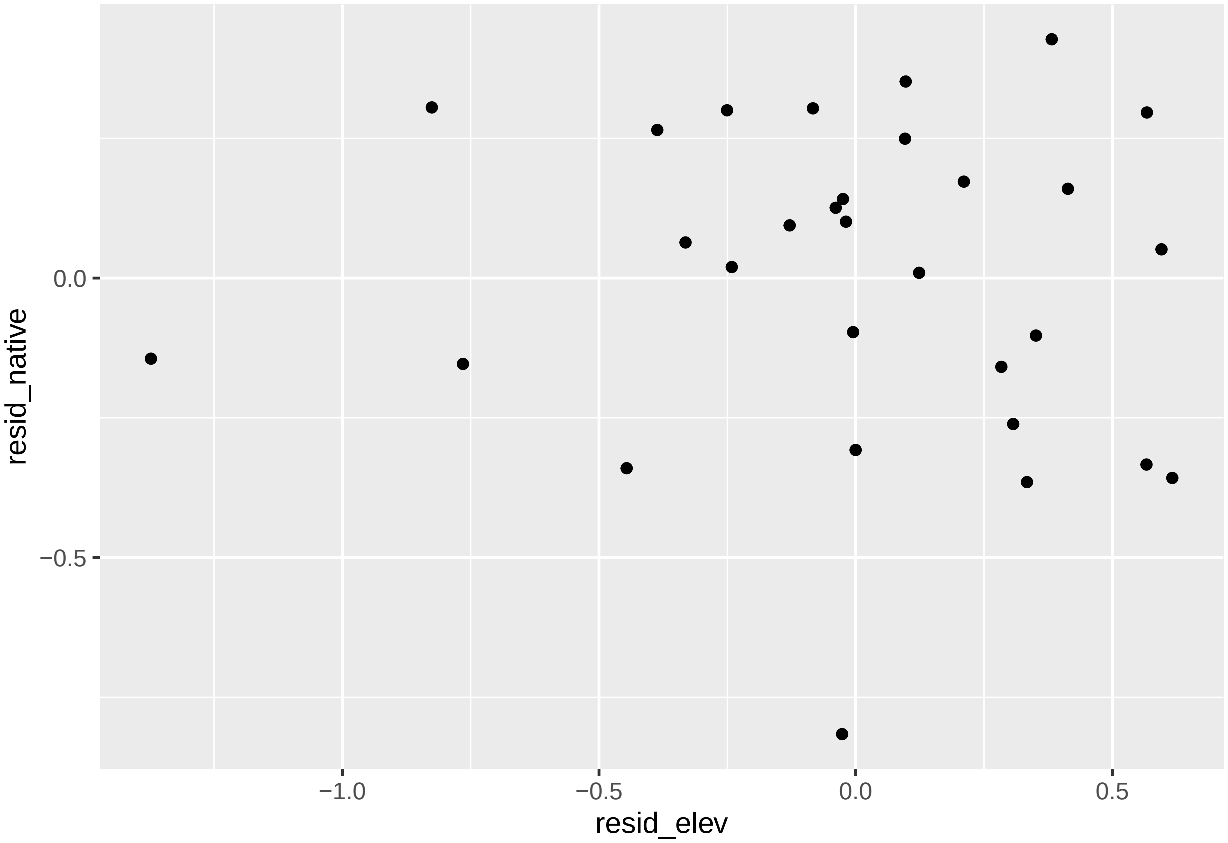
```
fit_elev <- lm(log_Elev ~ log_Area, data = galapagos_transformed)
galapagos_transformed <- galapagos_transformed %>%
  mutate(
    resid_elev = residuals(fit_elev)
  )
```

iii. Fit a model that has (potentially transformed) `Native` as the response and (potentially transformed) `Area` as the only explanatory variable. Add the residuals from this model to your data set with transformed variables.

```
fit_native <- lm(Native_0.25 ~ log_Area, data = galapagos_transformed)
galapagos_transformed <- galapagos_transformed %>%
  mutate(
    resid_native = residuals(fit_native)
  )
```

iv. Make a plot that has the residuals from part ii on the horizontal axis and the residuals from part iii on the vertical axis. Compare this plot to the added variable plot for `Elev` from part i.

```
ggplot(data = galapagos_transformed, mapping = aes(x = resid_elev, y = resid_native)) +
  geom_point()
```



This plot matches the added variable plot above.

v. Fit a linear model that has the residuals from part iii as the response and the residuals from part ii as the explanatory variable. Print out the model summary. Compare the coefficient estimate for the slope to the coefficient estimate for `Elev` from your model in part i.

```
av_fit <- lm(resid_native ~ resid_elev, data = galapagos_transformed)
summary(av_fit)
```

```
##
## Call:
## lm(formula = resid_native ~ resid_elev, data = galapagos_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8167 -0.1842  0.0616  0.2340  0.4390
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.528e-17  5.255e-02   0.000    1.000
## resid_elev  -3.081e-02  1.181e-01  -0.261    0.796
##
## Residual standard error: 0.2878 on 28 degrees of freedom
## Multiple R-squared:  0.002425,   Adjusted R-squared:  -0.0332
## F-statistic: 0.06806 on 1 and 28 DF,  p-value: 0.7961
```

The coefficient estimate from this fit is the same as the coefficient estimate for elevation in the fit from part i.