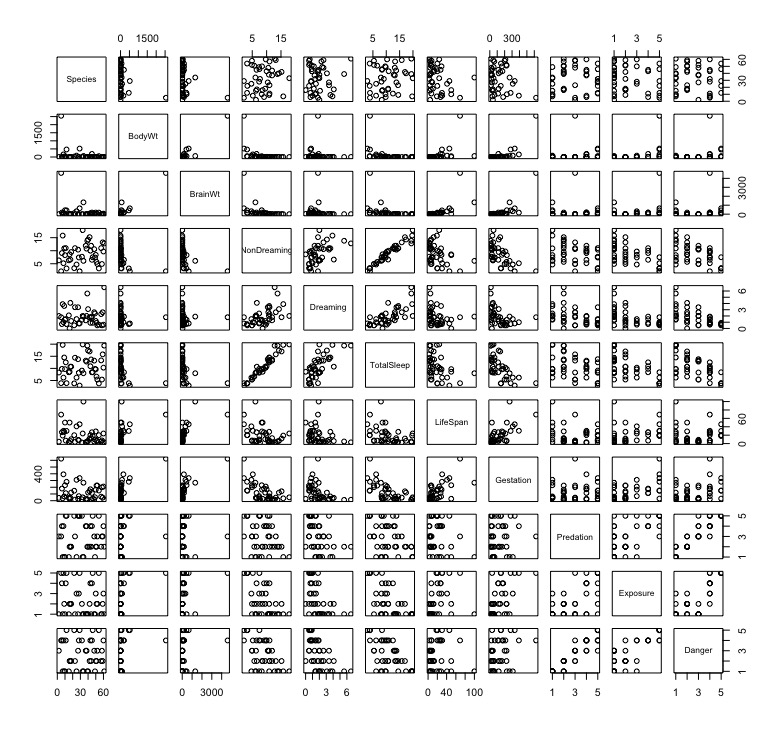
**Questions**

**This homework concerns the mammals dataset from the library openintro. This dataset contains some missing values (denoted NA). Exclude all observations with an NA in one of the variables with the command: mammals2 <- mammals[rowSums(is.na(mammals)) == 0,] This command saves the data subset to a new dataframe, mammals2. We shall do all computations on mammals2.**

1. **Do a pairwise plot of all variables. For each plot, briefly describe what you see.**



From pairwise plots, we can generally see if there is obvious trend for any combinations. The most obvious linear relationship could be observed from the plot of NonDreaming vs TotalSleep. Other positive relation graphs from top-left to bottom-right are **Brainwt vs Bodywt, Dreaming vs Total sleep,** Lifespan vs brainwt**, Gestation vs Bodywt, Gestation vs brainwt,** Gestation vs lifespan, Exposure vs Predation, **Danger vs Predation**, and **Danger vs Exposure.** Above are plots with positive correlations and the correlation values are at least 0.6. The relationship of Dreaming vs nondreaming and Exposure vs Gestation are certainly positive but a little questionable for the exact relationship. Their correlation value is between 0.5 to 0.6.

Many of them also have negative correlation. The plot with obvious negative relationship are Gestation vs Nondreaming, Gestation vs TotalSleep, and Exposure vs Totalsleep, which has correlation value lower than -0.6. The plots have less obvious but still can see the trend are Exposure vs Nondreaming, Exposure vs Dreaming, Danger vs Nondreaming, and Danger vs TotalSleep.

Other graphs have dots spreading rather randomly or nonlinear trend.

For the variable TotalSleep, we could observe that all other variable except Species have some relationship to some degree. And these variables except NonDreaming and Dreaming are all having a negative slope or inverse relationship.

1. **We want to fit a linear regression model that can be used to predict TotalSleep. Explain why Dreaming, NonDreaming and Species are BAD variables to include in this regression model.**

When I try to do a regression model TotalSleep ~ NonDreaming + Dreaming, the result coefficient of NonDreaming and Dreaming are both 1.00 with an extremely small coefficient, which means this is almost a perfect fit with formula TotalSleep = NonDreaming + Dreaming. This make sense because the sum of nondreaming sleeping time and dreaming sleeping time is how we calculate the total sleeping time.

Another example is a regression model TotalSleep ~ NonDreaming + Dreaming + LifeSpan that also gives us extremely small coefficient for intercept and LifeSpan and 1.00 for NonDreaming and Dreaming. Hence even adding another variable is not going to influence the plot.

1. **Treat Predation, Exposure and Danger as numericals. Run model1, the linear regression model with TotalSleep vs BodyWt, BrainWt, LifeSpan, Gestation, Predation, Exposure and Danger. Clearly show the R command that you use, and include the R’s model summary.**

R command:

> model1.1 <- lm(TotalSleep ~ Gestation + Predation + Danger,data = mammals2)

> summary(model1)

Result:

Call:

lm(formula = TotalSleep ~ BodyWt + BrainWt + LifeSpan + Gestation +

Predation + Exposure + Danger, data = mammals2)

Residuals:

Min 1Q Median 3Q Max

-6.2292 -1.8823 -0.1445 1.8914 5.9885

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 17.1091251 1.3363885 12.803 1.47e-14 \*\*\*

BodyWt 0.0047024 0.0059218 0.794 0.43266

BrainWt -0.0009979 0.0035541 -0.281 0.78059

LifeSpan -0.0145760 0.0462766 -0.315 0.75471

Gestation -0.0188108 0.0069799 -2.695 0.01086 \*

Predation 2.3151350 1.0926906 2.119 0.04150 \*

Exposure 0.5844391 0.6845807 0.854 0.39924

Danger -4.5375726 1.3567624 -3.344 0.00202 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.039 on 34 degrees of freedom

Multiple R-squared: 0.6548, Adjusted R-squared: 0.5837

F-statistic: 9.213 on 7 and 34 DF, p-value: 2.398e-06

R suggest that the intercept, Gestation, Predation, and Danger are significant

1. **Write down the equation that R gives you. Interpret all the coefficients and the p-values associated with the coefficients. Report the R2 and adjusted R2 of your model. What are the meaning of these values?**

Formula given by R is:

Since all of the variable are numerical, the coefficient for each variable simply means how adding one unit of the variable would affect the hours of total sleeping time.

R-squared is 0.6548, and adjusted R-squared is 0.5837. This information means that many dots on the graph is still far from the predicted linear model. The fit is moderate but not close to a good fit.

1. **Treat Predation, Exposure and Danger as categoricals. Run model2, the linear regression model with TotalSleep vs BodyWt, BrainWt, LifeSpan, Gestation, Predation, Exposure and Danger. Clearly show the R command that you use, and include the R’s model summary.**

I firstly convert the class of Predation, Exposure, and Danger to ‘factor’ and put it in another dataset naming it mammals2.1, then I run the regression:

> model1.2 <- lm(TotalSleep ~ BodyWt + BrainWt + LifeSpan + Gestation + Predation + Exposure + Danger,data = mammals2.1)

> summary(model1.2)

Result of summary from R:

Call:

lm(formula = TotalSleep ~ BodyWt + BrainWt + LifeSpan + Gestation +

Predation + Exposure + Danger, data = mammals2.1)

Residuals:

Min 1Q Median 3Q Max

-5.8430 -1.3825 -0.0377 0.9234 6.6646

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 15.048825 1.654848 9.094 2.1e-09 \*\*\*

BodyWt 0.005295 0.006233 0.849 0.40368

BrainWt -0.001824 0.003783 -0.482 0.63386

LifeSpan 0.002857 0.052728 0.054 0.95722

Gestation -0.019381 0.008277 -2.342 0.02747 \*

Predation2 4.611332 2.325738 1.983 0.05849 .

Predation3 6.876712 3.596627 1.912 0.06740 .

Predation4 9.985318 4.365802 2.287 0.03092 \*

Predation5 9.471563 4.715789 2.008 0.05551 .

Exposure2 -0.777955 1.679608 -0.463 0.64724

Exposure3 -1.017955 2.534686 -0.402 0.69138

Exposure4 0.791349 3.326964 0.238 0.81393

Exposure5 1.081892 5.102923 0.212 0.83381

Danger2 -5.842000 2.403106 -2.431 0.02256 \*

Danger3 -11.217196 3.529768 -3.178 0.00392 \*\*

Danger4 -12.621907 4.940129 -2.555 0.01709 \*

Danger5 -18.079357 6.739015 -2.683 0.01276 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

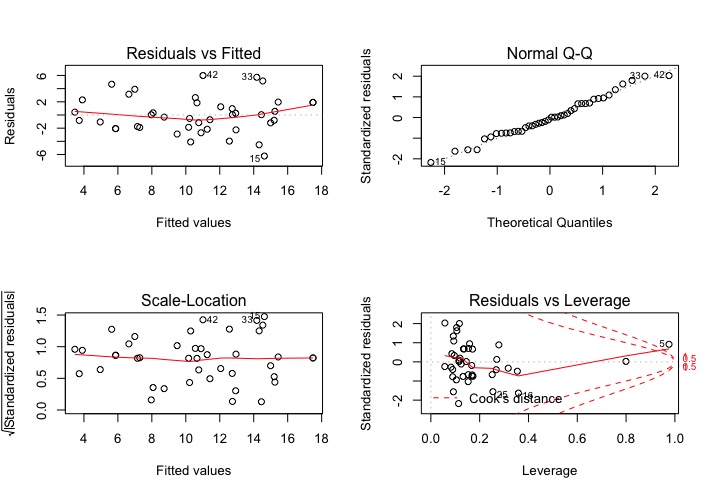
Residual standard error: 3.082 on 25 degrees of freedom

Multiple R-squared: 0.7389, Adjusted R-squared: 0.5718

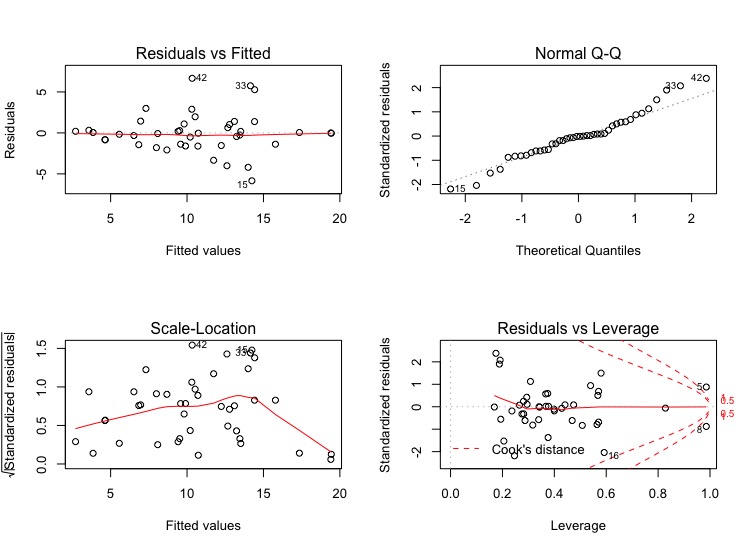
F-statistic: 4.422 on 16 and 25 DF, p-value: 0.0004694

1. **Compare model1 and model 2: comment on the coefficients and the diagnostic plots (Say which, if any, of the (a) independence (no mean trend) (b) normal distribution and (c) constant variance assumptions are violated.)**

**Diagnostic plot for Model1**



**Diagnostic plot for Model2**



Both model1 and model2 holds the assumption of (a) independence (no mean trend) and (b) normal distribution very well. The (c) constant variance assumption has no problem in model1 but not for model2. However, this could be caused by limited many data at higher fitted values. The data with less fitted values does have random distributed square root of standardized residuals.

1. **Do variable selection with the stepAIC command, starting with model1. Call this model1.AIC. Compare model1.AIC against model1: comment on the coefficients and the diagnostic plots.**

Model1.AIC has four variables and an intercept, whereas model1 has seven variables. For each variable, model1.AIC has higher p-value, which means each variable included make more difference than that in model1. Although the multiple r-squared of model1.AIC is lower than that of model1 by 0.015, the adjusted r-square of model1.AIC is higher than that of model1 by 0.02.

The diagnostic graph of model1.AIC shows that the all assumption holds, which is the same as model1.

1. **Do variable selection with the stepAIC command, starting with model2. Call this model2.AIC. Compare model2.AIC against model2: comment on the coefficients**

Variable selected in model2.AIC are BodyWt, Gestation, Predation, and Danger. Compare to model2, BrainWt, Lifespan, and Exposure are eliminated. From the variables’ coefficients, we notice the variables excluded are having high p-values, which means these variables are not affecting the model very much and lower the adjusted r-squared at the same time. And what’s interesting is that the value of coefficient for the variables does not change much from model2 to model2.AIC. For R-squared value, similar to how model1.AIC is different from model1, model2.AIC also have a lower multiple r-square but higher adjusted r-squared value.

1. **Which model amongst the above 4 is the best? (Give a brief justi- fication). For the better model, summarize the relationship between TotalSleep and other attributes of a mammal.**

From the view of adjusted r-square, model2.AIC is a better model among these four. Indeed, the variables choosen by stepAIC are the same from both model1 and model2. The only difference between them is that predation and danger are viewed as categorical variables.mo

1. **The species Homo Sapiens has the following attributes: BodyWt = 75, BrainWt = 1.4, LifeSpan = 77, Gestation = 268, Predation = 2, Exposure = 2, Danger = 2. Use your model** **to predict TotalSleep for this species. Is your prediction reasonable? Explain why or why not.**

According to the summary given by R from model2.AIC, the formula to predict TotalSleep is:

14.735164 + 75\*0.002921+268\*(-0.021244)+5.023594-6.375342 =7.909099 hours.

This prediction is reasonable because human total sleep hours should be around eight.

**(Bonus) Ngoc played with this data and obtained an adjusted R2 of 0.6848 and a reasonable fit (with no dropping of any observations). Present a model that has an adjusted R2 at least as good. For your model, shows the diagnostic plot, and provide interpretations for the coefficients.**

> modelbonus <- lm(TotalSleep ~ I(BodyWt^-0.2) + I(BrainWt^-5) + I(log(LifeSpan)) + I(Gestation) + Predation + Danger, data = mammals2.1)

> summary(modelbonus)

Result of summary:

Call:

lm(formula = TotalSleep ~ I(BodyWt^-0.2) + I(BrainWt^-5) + I(log(LifeSpan)) +

I(Gestation) + Predation + Danger, data = mammals2.1)

Residuals:

Min 1Q Median 3Q Max

-4.8042 -0.8907 0.0277 1.1705 5.4667

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.783e+00 2.819e+00 0.987 0.331714

I(BodyWt^-0.2) 5.232e+00 1.079e+00 4.851 3.84e-05 \*\*\*

I(BrainWt^-5) -5.560e-04 1.762e-04 -3.156 0.003712 \*\*

I(log(LifeSpan)) 1.601e+00 6.618e-01 2.419 0.022083 \*

I(Gestation) -1.291e-02 4.450e-03 -2.901 0.007024 \*\*

Predation2 9.852e+00 1.928e+00 5.109 1.87e-05 \*\*\*

Predation3 1.218e+01 2.924e+00 4.166 0.000254 \*\*\*

Predation4 1.221e+01 3.401e+00 3.591 0.001199 \*\*

Predation5 1.163e+01 3.682e+00 3.159 0.003686 \*\*

Danger2 -9.331e+00 1.671e+00 -5.584 5.01e-06 \*\*\*

Danger3 -1.310e+01 2.625e+00 -4.992 2.59e-05 \*\*\*

Danger4 -1.128e+01 3.352e+00 -3.365 0.002170 \*\*

Danger5 -1.505e+01 3.823e+00 -3.937 0.000474 \*\*\*

---

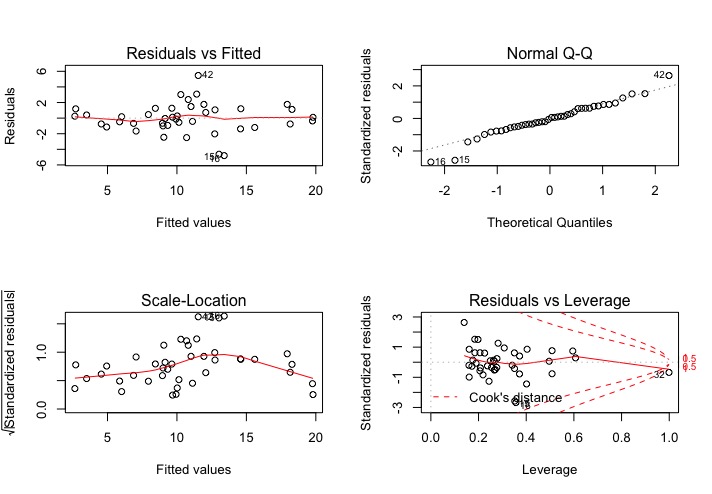
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.236 on 29 degrees of freedom

Multiple R-squared: 0.8405, Adjusted R-squared: 0.7745

F-statistic: 12.74 on 12 and 29 DF, p-value: 1.652e-08

Diagnostic plot for Modelbonus



The formula to predict total sleep with the coefficient R gives us in the summary is:

Coefficients for , , and simply means when , , , and increase 1 unit, then corresponding predicted total sleeping time will increase or decrease by their coefficients. Since Predation and Danger are categorical variables, then when given predation or danger is in certain category, only the corresponding coefficient in such category will be useful, and all other category variable equals zero. For example, if the animal predation level is four, then is the default variable, which is included in intercept at the beginning, so if the animal fall into the first category, all predation variable equals to one.