```r
# Reading in data
data <- read.csv('mammals.csv')

# Cleaning data to remove NA rows in dreaming and non_dreaming columns
cleaned_data <- data %>%
  filter(!is.na(dreaming), !is.na(non_dreaming))
```

Q3.1 Independent and Identically Distributed

1. IID Assumption

The assumption of independent and identically distributed (IID) data relies on the design and sampling methodology of the study itself. In this case, the researchers analyzed mammals in laboratory conditions, while ecological factors such as predation and sleeping place exposure were inferred from field observations in the literature. These factors were quantified using an ordinal scale, with 1 being the lowest and 5 being the highest for each. However, the methodology for selecting species in the study was not explicitly detailed.

The study appears to draw heavily from Zepelin and Rechtschaffen (1974), which also derived data from previous literature. Comparing the taxonomic Orders of mammals studied suggests a similar sample composition, raising questions about potential dependencies in the data. While non-dreaming and dreaming sleep were measured independently, the possibility of dependencies between mammals in the study (e.g., predator-prey relationships) cannot be ignored. For instance, if a rabbit and its predator wolf are both included, their predation and exposure scores are likely correlated, potentially influencing their dreaming and non-dreaming sleep measures. This interdependence challenges the independence condition of IID.

Additionally, the identically distributed assumption may be compromised due to unknown sampling procedures and potential spatial clustering effects. For example, if mammals were sampled from similar ecological regions, their behaviors might exhibit regional dependencies. Temporal clustering, on the other hand, is less of a concern, as evolutionary habits of predator-prey relationships are unlikely to change significantly over short timescales.

We must take note that the consequences of non-IID sampling would indicate that our findings hold no guarantees about the population. Instead, findings may reflect specific clusters or regions, which deviates from the research question. Addressing non-IID sampling could involve acquiring new, independent data or redesigning the sampling process to avoid the clustering effects. If that isn't possible, statistical adjustments, such as clustered standard errors, can account for dependencies and provide more accurate uncertainty estimates under the current data-generating process.

References

- Zepelin, H, and A Rechtschaffen. "Mammalian sleep, longevity, and energy metabolism." Brain, behavior and evolution vol. 10,6 (1974): 425-70. doi:10.1159/000124330
- Allison, T, and D V Cicchetti. "Sleep in mammals: ecological and constitutional correlates." Science (New York, N.Y.) vol. 194,4266 (1976): 732-4. doi:10.1126/science.982039

Q3.2 No Perfect Collinearity

Perfect colinearity is easy to spot as the model will either not work or drop a feature:

```r
# Creating the model
model_1 <- lm(brain_wt ~ dreaming + non_dreaming, data = cleaned_data)
coeftest(model_1)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   775.996    243.047  3.1928 0.002572 **
## dreaming       68.285     76.228  0.8958 0.375130
## non_dreaming  -83.099     30.036 -2.7666 0.008189 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In our case, neither the dreaming nor non-dreaming features are dropped, so there is no perfect colinearity.

To assess near perfect colinearity instead, regressions will have large standard errors on colinear features, which does appear to be the case here. This could potentially mean the dreaming and non-dreaming variables are related and are trying to balance each other out. We can check for the correlation between those two variables:

```r
# Checking for colinearity
cor(cleaned_data$dreaming, cleaned_data$non_dreaming, use = "complete.obs")
```

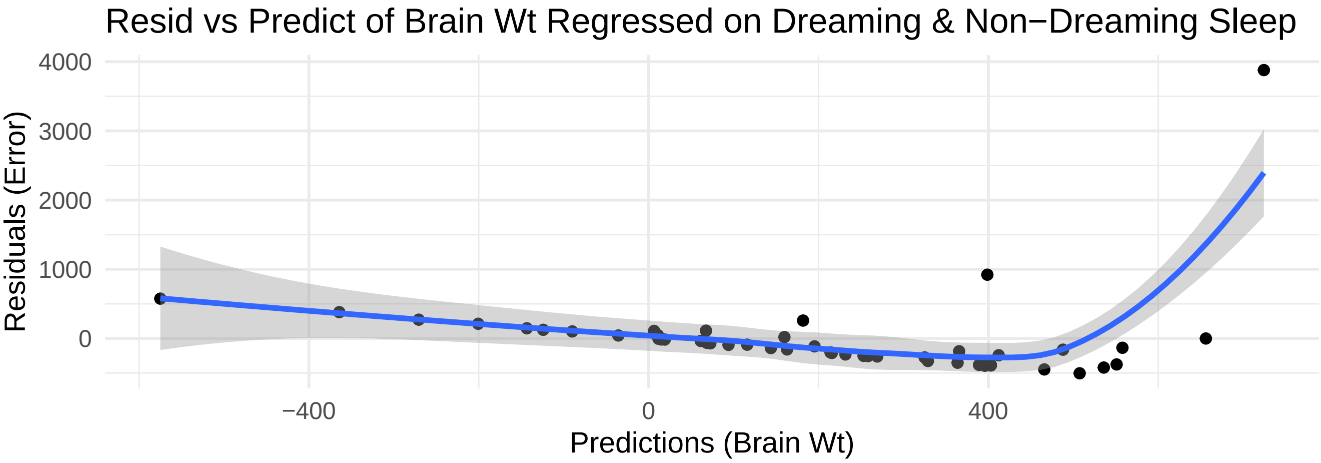
```
## [1] 0.5142539
```

However, we see that they are not strongly correlated.

Therefore, with the data given, we have strong evidence to suggest that the no perfect colinearity assumption can be met. While the dreaming and non-dreaming variables may be related in some way, they each provide us with more information we can use in the model. If we want to place extra emphasis on the large standard errors on colinear features and want to be extra safe, we can use less data, but keep the information (by doing some form of dimension reduction like PCA or factor analysis).

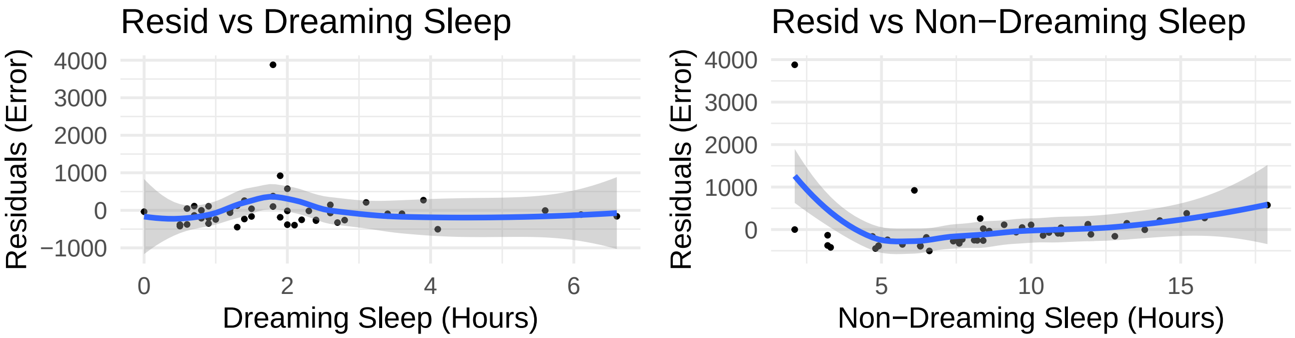
Q3.3 Linear Conditional Expectation

To check for this assumption, we can plot a residuals vs prediction plot to see whether the residuals are linear across the predictions.



This plot shows strong evidence that there is some non-linear relationship that is not being captured by our model, as the residuals are both non-linear and non-zero. We note that there is one glaring outlier at the upper right corner of the plot, which greatly affects our model, and removing it may give stronger evidence for validating this assumption.

We can also try looking at residuals against both the dreaming and non-dreaming variables:



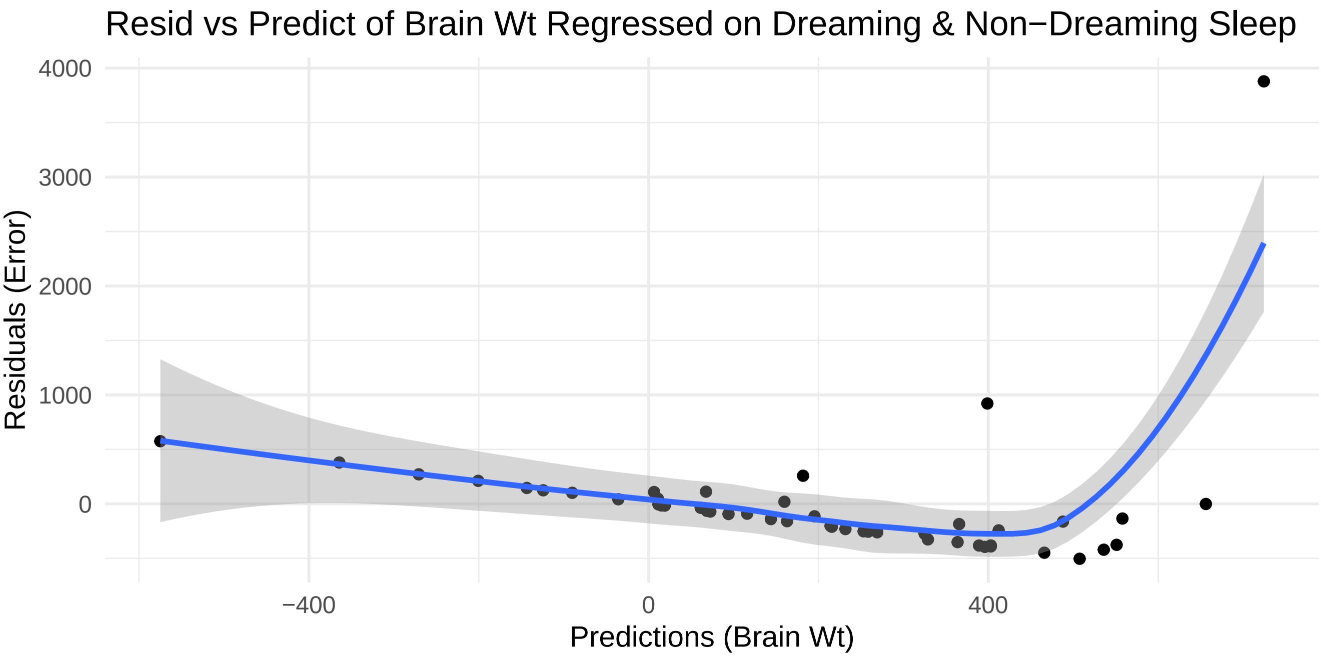
In both of these plots, we see comparatively stronger evidence of linear residuals across the entire range of x; however, the outlier mentioned previously still persists and pulls the model away. If we're able to create another model, we may try a log transform on brain_wt to curtail the effect of the outlier. However, there doesn't seem to be a linear relationship between the residuals and the predictions. Therefore, this condition is not met.

The consequence of this is that while we have still fit the best linear predictor, the estimated coefficient does not match the relationship in the data. We might be able to use nonlinear models that are more efficient, or we can add more complexity to improve the prediction (and inference). However, the assumption of linear conditional expectation is not met in this current model.

Q3.4 Constant Error Variance

Without even needing to do a test, because the linear conditional expectation assumption is not met, this assumption is not relevant and therefore not satisfied. This is because regardless of whether the model looks homoskedastic or heteroskedastic, it could simply just be due to a poorly specified model. Consequently, tests for homoskedasticity may not be informative, as they cannot distinguish between these scenarios.

Nevertheless, we can still examine the the variance of residuals using either an ocular test or a format test, such as as the Breusch-Pagan test. Starting with an eye test, we see in the residuals vs. predictions plot above that the residuals seem to vary along the predicted values (especially influenced by the outlier). Excluding this outlier, the residuals appear to have relatively constant variance, suggesting that the errors may be homoskedastic under these conditions.



Using the Breusch-Pagan test to further validate:

```
bptest(model_1)
```

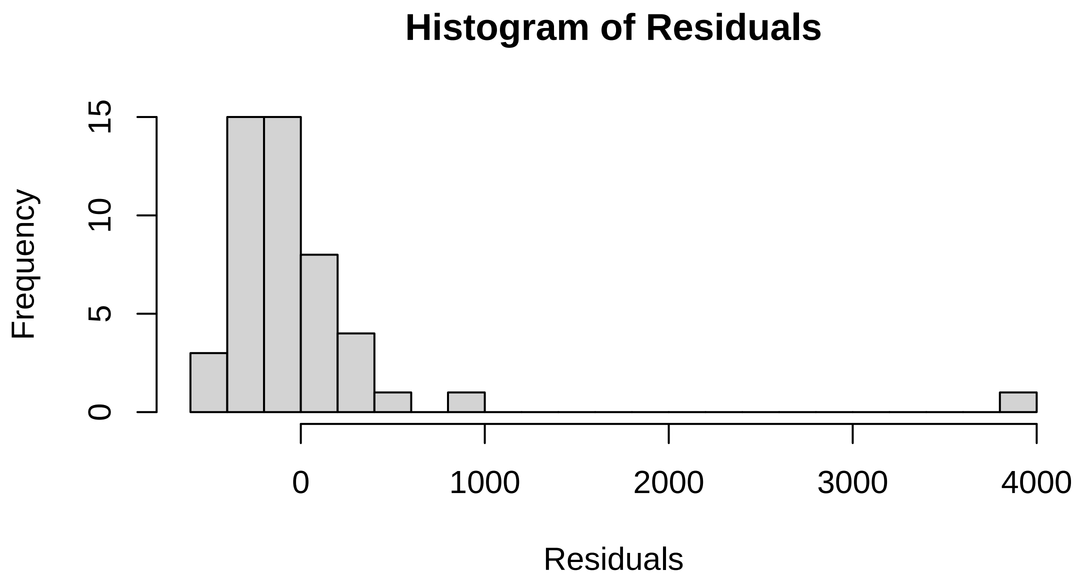
```
##
##  studentized Breusch-Pagan test
##
## data:  model_1
## BP = 4.44, df = 2, p-value = 0.1086
```

We actually see that this model does not appear to suffer from heteroskedastic error variance, with a p-value > 0.05. However, we must take note again that this may instead be a model misspecification problem because linear conditional expectation is not met, and therefore this assumption is not relevant here. Nevertheless, if we wanted to ensure we circumvent this issue, we can simply use robust standard errors instead.

Q3.5 Normal Distribution of Errors

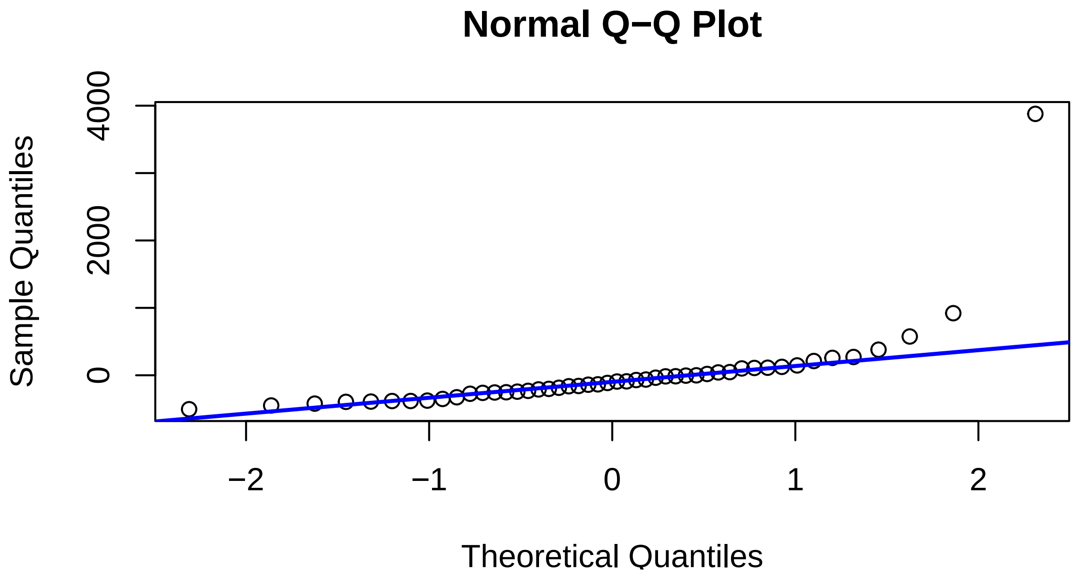
We can easily check whether the errors are normally distributed using either a histogram or a qqplot. Starting with a histogram:

```r
hist(model_1$residuals,
     breaks = 30,
     main = "Histogram of Residuals",
     xlab = "Residuals")
```



We can see that while it's roughly normal on the left, the outlier at ~4000 skews the data right. Looking at the qq-plot, we see something similar:

```r
qqnorm(model_1$residuals)
qqline(model_1$residuals, col = "blue", lwd = 2)
```



Once again, we see that while normality generally holds, the outlier at the extreme end skews the distribution. This outlier makes it challenging to fully evaluate the assumption of normally distributed errors, but overall, it appears that this assumption is not satisfied.