Stat 21 Homework 5 - Solution

Your name here Collaborators: [list any collaborators here]

Due: Oct 27th, by noon ET

Problem 4

Let's consider the data set called *msleep* which is contained in the R package `ggplot2`.

```
library(ggplot2)
head(msleep)
```

```
## # A tibble: 6 x 11
##   name  genus vore  order conservation sleep_total sleep_rem sleep_cycle awake
##   <chr> <chr> <chr> <chr> <chr>              <dbl>     <dbl>       <dbl> <dbl>
## 1 Chee~ Acin~ carni Carn~ lc                  12.1      NA        NA      11.9
## 2 Owl ~ Aotus omni  Prim~ <NA>                17         1.8      NA       7
## 3 Moun~ Aplo~ herbi Rode~ nt                  14.4       2.4      NA       9.6
## 4 Grea~ Blar~ omni  Sori~ lc                  14.9       2.3       0.133   9.1
## 5 Cow   Bos   herbi Arti~ domesticated         4         0.7       0.667  20
## 6 Thre~ Brad~ herbi Pilo~ <NA>                14.4       2.2       0.767   9.6
## # ... with 2 more variables: brainwt <dbl>, bodywt <dbl>
```

This data set looks at the amount of time spent sleeping for different mammals and records other factors such as brain and body weight of these animals. Suppose we are interested in the total amount of sleep an animal gets (variable name `sleep_total`) as predicted by the total body weight of the animal (variable name `bodywt`).

a) There is a qualitative variable named *order* in this data set. Fit two separate linear regression models for the animals of *order* "Carnivora" and of *order* "Primates". Report the estimated regression equations and print the summary of the two linear models.

b) What is the estimate of the variance of the random error for each of these two models?

c) Do you think we should combine the data and just fit a single linear regression model to both *orders* Carnivora and Primates? Justify your answer with 1-2 sentences and (possibly) a supporting plot.

Solution

```
carnivora <- msleep %>% filter(order=="Carnivora") %>% select(sleep_total,bodywt)
primates <- msleep %>% filter(order=="Primates") %>% select(sleep_total,bodywt)

SLR_carn <- lm(sleep_total~bodywt, data = carnivora)
  SLR_carn %>% summary
```

```
##
## Call:
## lm(formula = sleep_total ~ bodywt, data = carnivora)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0892 -1.3180  0.6448  2.4067  3.9322
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.15297    1.43235   6.390 7.93e-05 ***
## bodywt       0.01670    0.01751   0.954    0.363
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.517 on 10 degrees of freedom
## Multiple R-squared:  0.08337,    Adjusted R-squared:  -0.008289
## F-statistic: 0.9096 on 1 and 10 DF,  p-value: 0.3627
```

```
SLR_prim <- lm(sleep_total~bodywt, data = primates)
  SLR_prim %>% summary
```

```
##
## Call:
## lm(formula = sleep_total ~ bodywt, data = primates)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5390 -1.0032 -0.4876  0.0190  5.9085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.11268    0.72943  15.235 3.01e-08 ***
## bodywt      -0.04414    0.02941  -1.501    0.164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.094 on 10 degrees of freedom
## Multiple R-squared:  0.1838, Adjusted R-squared:  0.1022
## F-statistic: 2.252 on 1 and 10 DF,  p-value: 0.1643
```

(a) For full credit the student must show their R code that they used to create two subset of the data (one for each order - primates or carnivora). They must also show the R output of the summaries for each of the two linear models. (It's ok if their code doesn't match mine exactly but the summary for their linear models should match the above output exactly.) It is important that they do not include an error term in the estimated regression equation and that they note that the model is for the average (or mean) time slept.

Estimated regression equation for order Carnivora:

$$\text{average time slept} = 9.91527 + 0.0167(\text{body weight})$$

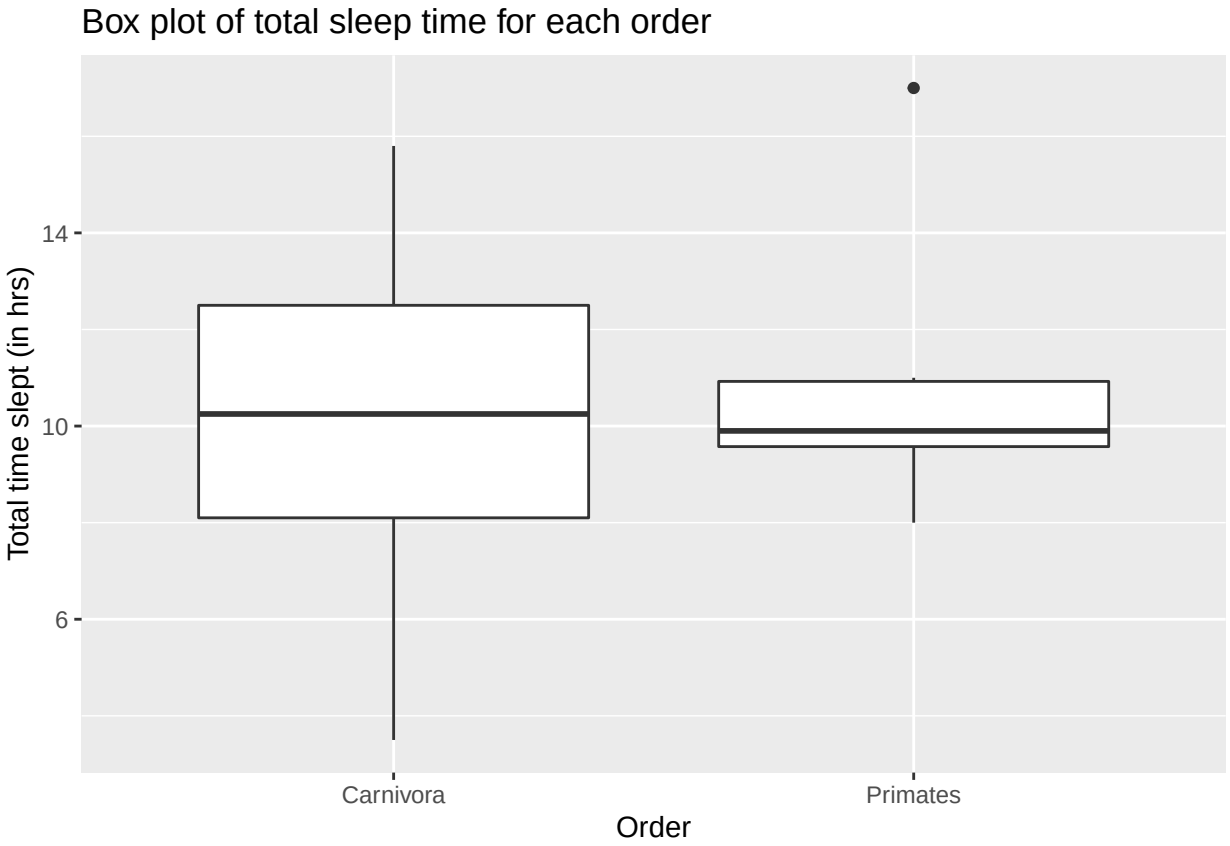
Estimated regression equation for order Primates:

$$\text{average time slept} = 11.11268 - 0.04414(\text{body weight})$$

(b) Estimate for the error variance in the model for order Carnivora: $\hat{\sigma}^2 = 3.517^2$

Estimate for the error variance in the model for order Primates: $\hat{\sigma}^2 = 2.094^{2}$‘

```
msleep2 <- msleep %>% mutate(order_cat = factor(order)) %>%
                      filter((order=="Carnivora")|(order=="Primates"))
ggplot(msleep2, aes(x=order, y=sleep_total)) +
  geom_boxplot() +
  labs(title="Box plot of total sleep time for each order", x="Order", y="Total time slept (in hrs)")
```



(c) They do not need to include the box plot to get full credit. To get full credit they must note that the estimated variance of the errors is different in each of the two models above. In SLR, we need to assume that the variance of the errors is constant but that does not seem to be the case based on part (b) (and/or on the box plot above). Therefore, it would not be a good idea to fit a single SLR predicting the amount of time slept for both of these animal orders at the same time.