Computer exercise 2: solutions

Anna Lindgren

8 April 2019

```
sleep <- read.delim("../data/sleep.txt")
sleep$Danger <- factor(sleep$Danger, levels = c(1, 2, 3),
                       labels = c("low", "medium", "high"))
```

Exercise 2: Sleep in Mammals

Danger

(a) Danger frequencies

```
with(sleep, table(Danger))
#> Danger
#>    low medium   high
#>     17     24     16
```

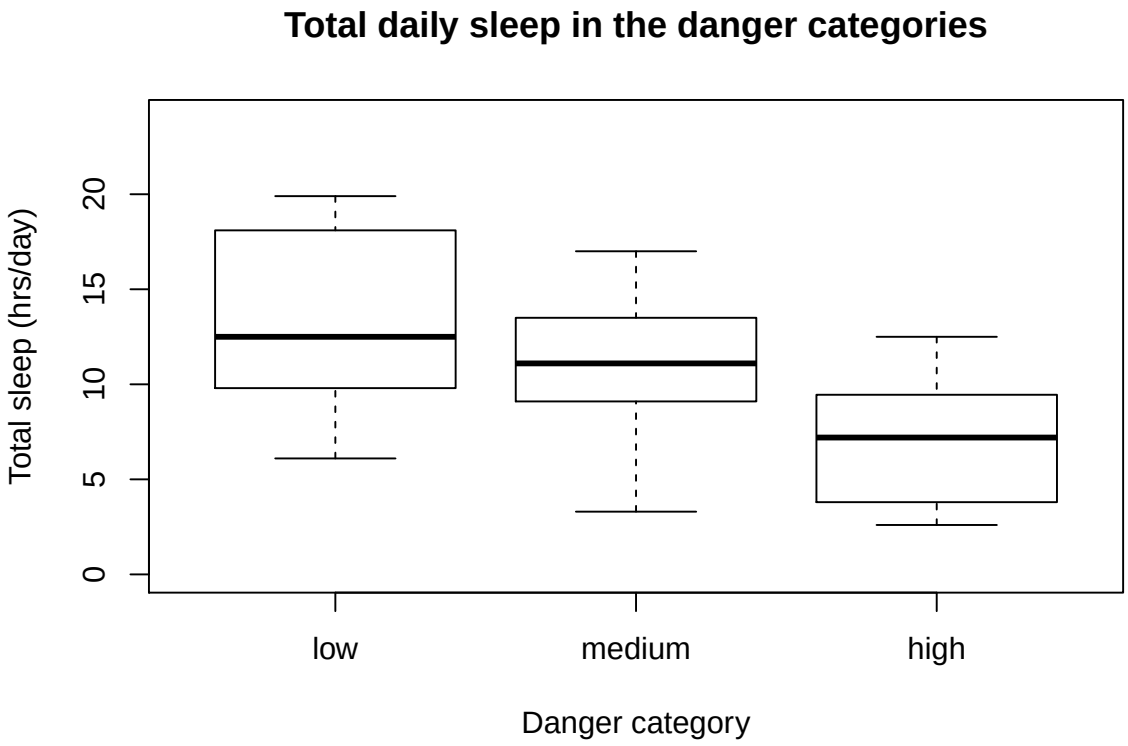
Medium danger is the most frequent category.

(b) Average sleep in different danger groups

```
aggregate(TotalSleep ~ Danger, data = sleep, FUN = "mean")
#>   Danger TotalSleep
#> 1    low   13.38235
#> 2 medium   11.15000
#> 3   high    6.73750
```

The more danger the less sleep. Seems reasonable.

```
ytext = "Total sleep (hrs/day)"
xtext.d = "Danger category"
with(sleep, plot(TotalSleep ~ Danger,
                 ylim = c(0, 24),
                 ylab = ytext,
                 xlab = xtext.d,
                 main = "Total daily sleep in the danger categories"))
```



The amount of daily sleep decreases with increasing danger. The medians in the box plots agree fairly well with the corresponding means. There is also a large variation within each group.

(c) Linear model with danger categories

```
model.danger <- lm(TotalSleep ~ Danger, data = sleep)
summary(model.danger)
#>
#> Call:
#> lm(formula = TotalSleep ~ Danger, data = sleep)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -7.8500 -2.8500 -0.3824  2.6500  6.5176
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   13.3824     0.9483  14.111  < 2e-16 ***
#> Dangermedium  -2.2324     1.2395  -1.801   0.0773 .
#> Dangerhigh    -6.6449     1.3619  -4.879 9.83e-06 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 3.91 on 54 degrees of freedom
#> Multiple R-squared:  0.3138, Adjusted R-squared:  0.2884
#> F-statistic: 12.35 on 2 and 54 DF,  p-value: 3.839e-05
```

```
x0 <- data.frame(Danger = c("low", "medium", "high"))
pred.danger <- cbind(x0, fit = predict(model.danger, x0))
pred.danger
#>   Danger      fit
```

```
#> 1    low 13.38235
#> 2 medium 11.15000
#> 3   high  6.73750
```

The predictions are the same as the category averages. They are computed as

$$\hat{Y}_{\text{low}} = \hat{\beta}_0 = 13.3823529$$

$$\hat{Y}_{\text{medium}} = \hat{\beta}_0 + \hat{\beta}_1 = 13.3823529 + (-2.2323529) = 11.15$$

$$\hat{Y}_{\text{high}} = \hat{\beta}_0 + \hat{\beta}_2 = 13.3823529 + (-6.6448529) = 6.7375$$

(d) F-test for Danger

The Danger is the only variable in the model we can use a global F-test.

```
anova(model.danger)
#> Analysis of Variance Table
#>
#> Response: TotalSleep
#>           Df Sum Sq Mean Sq F value    Pr(>F)
#> Danger     2 377.54 188.769  12.347 3.839e-05 ***
#> Residuals 54 825.60  15.289
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the P-value $= 3.8394844 \times 10^{-5} < \alpha = 0.05$ we can reject H_0: $\beta_1 = \beta_2 = 0$. Yes, we need the danger variable.

We could also look at the last line in the summary:

```
summary(model.danger)
#>
#> Call:
#> lm(formula = TotalSleep ~ Danger, data = sleep)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -7.8500 -2.8500 -0.3824  2.6500  6.5176
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   13.3824     0.9483  14.111  < 2e-16 ***
#> Dangermedium  -2.2324     1.2395  -1.801   0.0773 .
#> Dangerhigh    -6.6449     1.3619  -4.879 9.83e-06 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 3.91 on 54 degrees of freedom
#> Multiple R-squared:  0.3138, Adjusted R-squared:  0.2884
#> F-statistic: 12.35 on 2 and 54 DF,  p-value: 3.839e-05
```

(e) Danger parameter confidence intervals

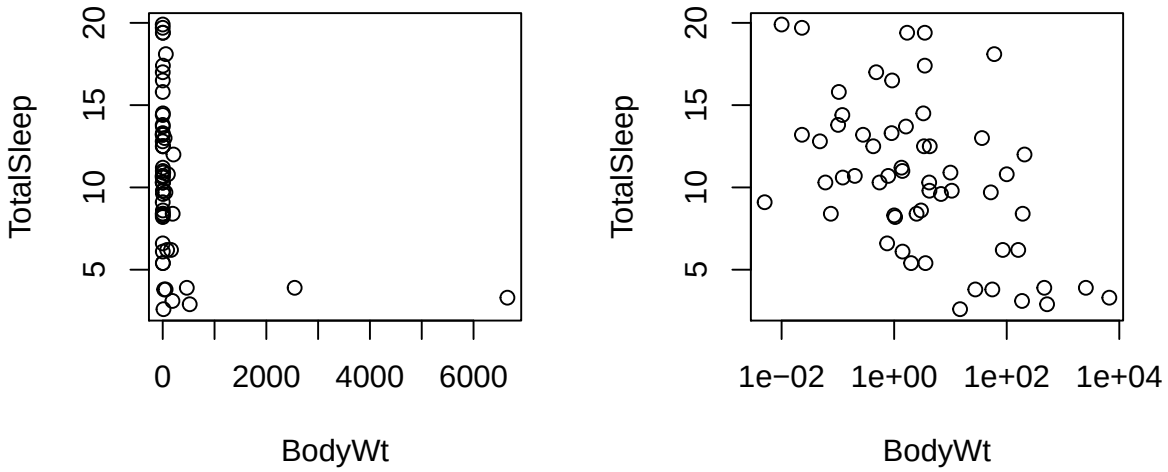
```
confint(model.danger)
#>                   2.5 %    97.5 %
#> (Intercept)  11.481046 15.283660
#> Dangermedium -4.717423  0.252717
#> Dangerhigh   -9.375398 -3.914308
```

Note that the interval for β_1, Dangermedium, covers zero, in agreement with the non-significant result in the summary.

Danger and body weight

(f) Linear function of weight

```
par(mfrow = c(1, 2))
with(sleep, plot(TotalSleep ~ BodyWt))
with(sleep, plot(TotalSleep ~ BodyWt, log = "x"))
```



The body weight variable is very skewed (Elephants are heavy!) and it is difficult to see any linear relationship. By taking the logarithm of the weight we get something much more reasonable. When using body weight as it is, we would assume that adding 1kg to the body weight would give the same decrease in the amount of sleep, regardless of the size of the species. But 1kg is a huge change in a small animal but unnoticable in large animals (adding 1kg to a 6+ ton elephant would not be expected to give any noticable change in the amount of sleep).

When taking the log of the body weight we assume that relative changes in weight, e.g. doubling the weight, would have a fixed effect on the amount of sleep needed.

(g) Log weight and danger

We first fit the linear model $\text{TotalSleep}_i = \beta_0 + \beta_1 \ln(\text{BodyWt}_i) + \epsilon_i$:

```
model.bodywt <- lm(TotalSleep ~ log(BodyWt), data = sleep)
summary(model.bodywt)
#>
#> Call:
#> lm(formula = TotalSleep ~ log(BodyWt), data = sleep)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -6.7034 -2.6740 -0.3211  2.2137  9.9039
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  11.4396     0.5577  20.511  < 2e-16 ***
```

```
#> log(BodyWt)  -0.7922     0.1713  -4.625 2.31e-05 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 3.969 on 55 degrees of freedom
#> Multiple R-squared:  0.2801, Adjusted R-squared:  0.267
#> F-statistic: 21.39 on 1 and 55 DF,  p-value: 2.313e-05
```

We then fit the linear model $\text{TotalSleep}_i = \beta_0 + \beta_1 \ln(\text{BodyWt}_i) + \beta_2 \text{Danger}_{\text{medium},i} + \beta_3 \text{Danger}_{\text{high},i} + \epsilon_i$:

```
model.bodywtdanger <- lm(TotalSleep ~ log(BodyWt) + Danger, data = sleep)
summary(model.bodywtdanger)
#>
#> Call:
#> lm(formula = TotalSleep ~ log(BodyWt) + Danger, data = sleep)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -7.8533 -2.0739 -0.1159  2.5558  6.6844
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   14.1858     0.8536  16.619  < 2e-16 ***
#> log(BodyWt)   -0.6766     0.1629  -4.154 0.000119 ***
#> Dangermedium  -3.1729     1.1100  -2.858 0.006073 **
#> Dangerhigh    -5.4722     1.2269  -4.460  4.3e-05 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 3.428 on 53 degrees of freedom
#> Multiple R-squared:  0.4824, Adjusted R-squared:  0.4531
#> F-statistic: 16.46 on 3 and 53 DF,  p-value: 1.102e-07
```

Do we need the danger variable?

```
anova(model.bodywt, model.bodywtdanger)
#> Analysis of Variance Table
#>
#> Model 1: TotalSleep ~ log(BodyWt)
#> Model 2: TotalSleep ~ log(BodyWt) + Danger
#>   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
#> 1     55 866.2
#> 2     53 622.8  2     243.4 10.356 0.0001597 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the P-value $= 1.5974402 \times 10^{-4} < 0.05 = \alpha$ we should reject H_0: $\beta_2 = \beta_3 = 0$. Yes, we still need the danger variable.

Do we need the body weight?

```
anova(model.danger, model.bodywtdanger)
#> Analysis of Variance Table
#>
#> Model 1: TotalSleep ~ Danger
#> Model 2: TotalSleep ~ log(BodyWt) + Danger
#>   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
#> 1     54 825.6
#> 2     53 622.8  1     202.8 17.258 0.0001194 ***
#> ---
```

```
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

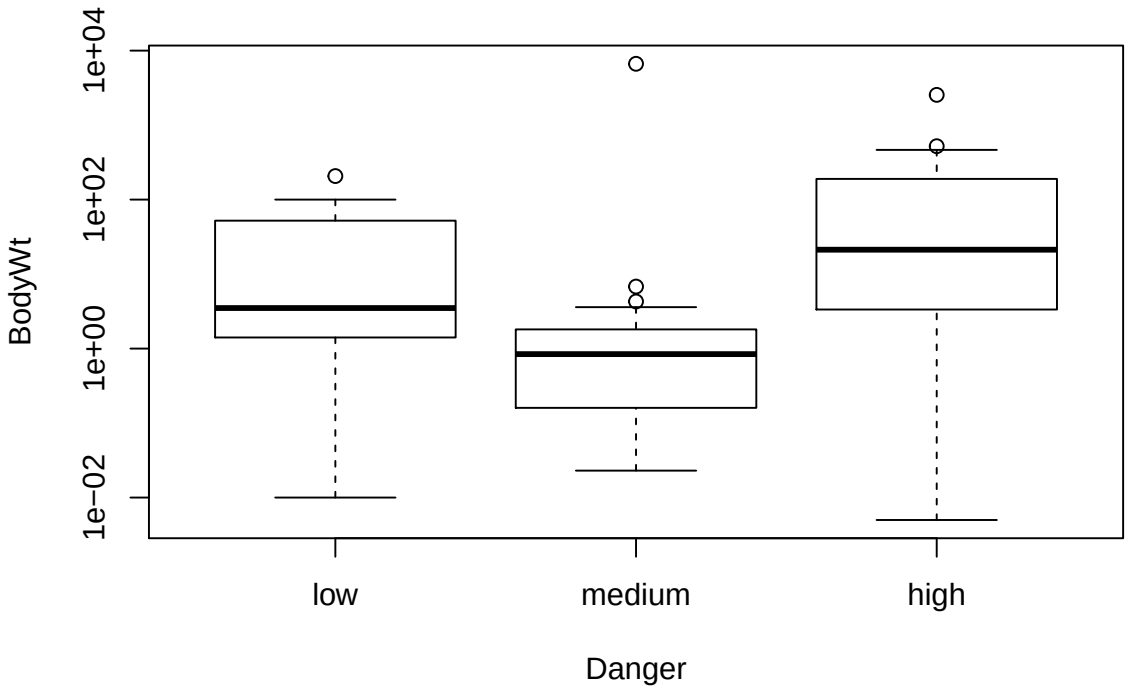
Since the P-value $= 1.193597 \times 10^{-4} < 0.05 = \alpha$ we should reject H_0: $\beta_1 = 0$. Yes, we still need the body weight.

(h) New danger parameter confidence intervals

```
confint(model.danger)
#>                   2.5 %    97.5 %
#> (Intercept)  11.481046 15.283660
#> Dangermedium -4.717423  0.252717
#> Dangerhigh   -9.375398 -3.914308
confint(model.bodywtdanger)
#>                   2.5 %     97.5 %
#> (Intercept)  12.473654 15.8978733
#> log(BodyWt)  -1.003242 -0.3499164
#> Dangermedium -5.399272 -0.9464831
#> Dangerhigh   -7.933106 -3.0112907
```

We can note that when we add log body weight, the interval for medium danger no longer covers zero.

```
with(sleep, plot(BodyWt ~ Danger, log = "y"))
```



As seen in the plot, body weight and danger are not independent. The medium danger species tend to be smaller than either low or high danger species. Adding body weight in the model allows us to separate the two conflicting effects. Species living in medium danger sleep less that species *with the same body weight* living in low danger. But this effect was blurred since the medium danger species were often smaller than the low danger species, and smaller species sleep more.

(i) Residuals

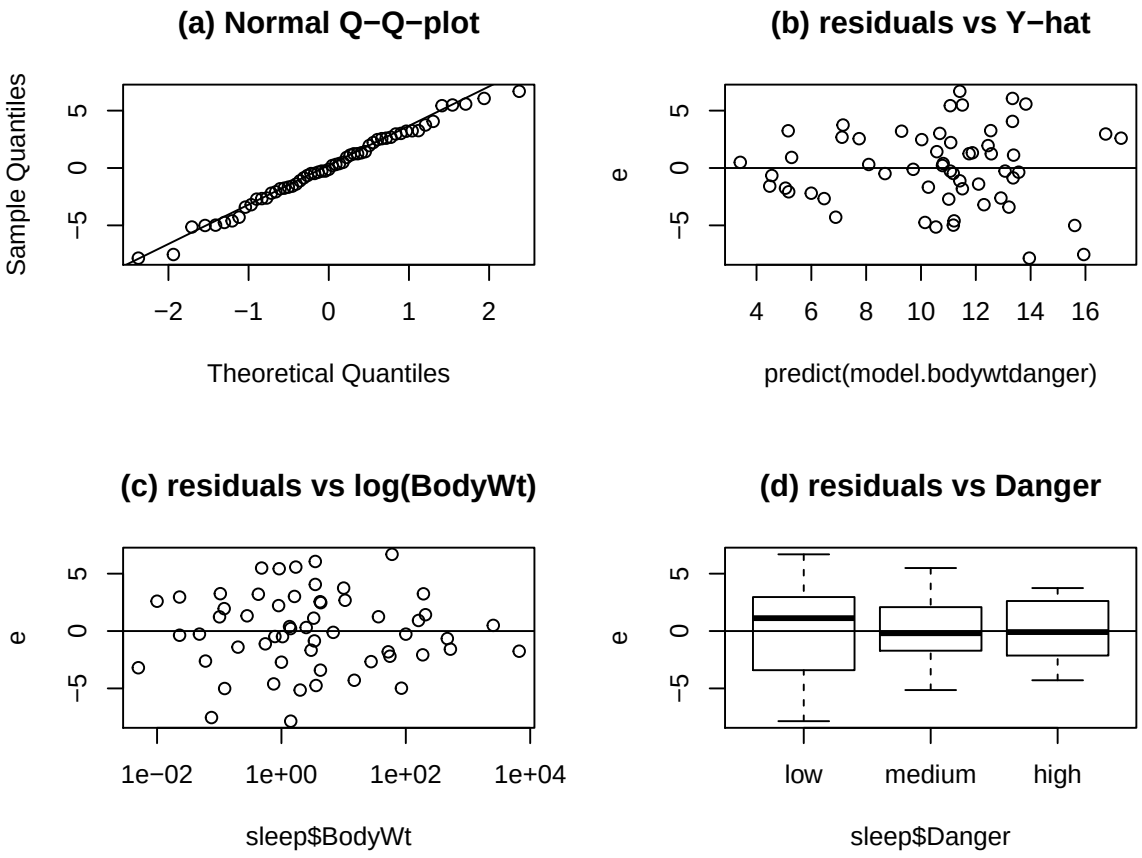
```r
par(mfrow = c(2, 2))
e <- model.bodywtdanger$residuals

qqnorm(e, main = "(a) Normal Q-Q-plot")
qqline(e)

plot(e ~ predict(model.bodywtdanger), main = "(b) residuals vs Y-hat")
abline(h = 0)

plot(e ~ sleep$BodyWt, log = "x", main = "(c) residuals vs log(BodyWt)")
abline(h = 0)

plot(e ~ sleep$Danger, main = "(d) residuals vs Danger")
abline(h = 0)
```



The Q-Q-plot in Fig (a) shows that the residuals are close to normally distributed.

In Fig (b) we see a tendency towards increasing variance with increasing $\hat{Y}$.

In Fig (c) we do not see anyting worrying. There is no systematic tendencies and the variance seems constant over ln(BodyWt).

In Fig (d) we see that the variance is largest among the low danger species and decreasing with increasing danger. This could explain the behaviour in Fig (b) since the low danger species are the ones who tend to sleep more and thus have a higher predicted amount of sleep.

Predictions

(j) A new species

```r
x0 <- data.frame(BodyWt = 30, Danger = "medium")
cbind(x0, pred = predict(model.bodywtdanger, x0, interval = "prediction"))
#>   BodyWt Danger pred.fit pred.lwr pred.upr
#> 1     30 medium 8.711707 1.596217  15.8272
```

(k) Homo Sapiens

```r
x0 <- data.frame(Species = "Homo Sapiens", BodyWt = 62, Danger = "low")
cbind(x0, pred = predict(model.bodywtdanger, x0, interval = "confidence"))
#>         Species BodyWt Danger pred.fit pred.lwr pred.upr
#> 1 Homo Sapiens     62    low 11.39343 9.469113 13.31775
```

(l) Marmot vs Vervet

We have the expected values for the two species as

$$E(Y_{\text{Marmot}}) = \beta_0 + \beta_1 \cdot \ln 4$$

and

$$E(Y_{\text{Vervet}}) = \beta_0 + \beta_1 \cdot \ln 4 + \beta_3.$$

The difference is

$$E(Y_{\text{Marmot}}) - E(Y_{\text{Vervet}}) = \beta_0 + \beta_1 \cdot \ln 4 - (\beta_0 + \beta_1 \cdot \ln 4 + \beta_3) = -\beta_3$$

and the confidence interval is simply $-I_{\beta_3}$ hours:

```r
ci.l <- -confint(model.bodywtdanger)["Dangerhigh", c(2, 1)]
ci.l
#>   97.5 %    2.5 %
#> 3.011291 7.933106
```

(m) Marmot vs Homo Sapiens

We have the expected values for the two species as

$$E(Y_{\text{Marmot}}) = \beta_0 + \beta_1 \cdot \ln 4$$

and

$$E(Y_{\text{Homo Sapiens}}) = \beta_0 + \beta_1 \cdot \ln 62.$$

The difference is

$$E(Y_{\text{Marmot}}) - E(Y_{\text{Homo Sapiens}}) = \beta_0 + \beta_1 \cdot \ln 4 - (\beta_0 + \beta_1 \cdot \ln 62) = \beta_1(\ln 4 - \ln 62) = -\ln \frac{62}{4} \cdot \beta_1$$

and the confidence interval is simply $-\ln \frac{62}{4} \cdot I_{\beta_1}$ hours:

```r
ci.m <- -log(62/4) * confint(model.bodywtdanger)["log(BodyWt)", c(2, 1)]
ci.m
#>    97.5 %     2.5 %
#> 0.9590649 2.7497261
```

(n) Vervet vs Homo Sapiens

We now have the difference as

$$E(Y_{\text{Vervet}}) - E(Y_{\text{Homo Sapiens}}) = \beta_0 + \beta_1 \cdot \ln 4 + \beta_3 - (\beta_0 + \beta_1 \cdot \ln 62) = \beta_1(\ln 4 - \ln 62) + \beta_3 = \ln \frac{4}{62} \cdot \beta_1 + \beta_3.$$

```r
xx <- matrix(c(0, log(4/62), 0, 1), nrow = 1)
preddif <- xx %*% coefficients(model.bodywtdanger)
```

The estimate is $\ln \frac{4}{62} \cdot \hat{\beta}_1 + \hat{\beta}_3 = -3.6178028$ with variance

$$(\ln \frac{4}{62})^2 V(\hat{\beta}_1) + V(\hat{\beta}_3) + 2 \ln \frac{4}{62} Cov(\hat{\beta}_1, \hat{\beta}_3)$$

```r
s <- summary(model.bodywtdanger)$sigma
XtXinv <- summary(model.bodywtdanger)$cov.unscaled
Vxx <- xx %*% XtXinv %*% t(xx)

# preddif and Vxx are 1 x 1 matrices so we need to pick the scalar value out first
# if we want to calculate both limits at the same time with +/- as c(-1, 1)
# which cannot be multiplied on to a matrix. R "feature".
preddif[1, 1] + qt(1 - 0.05 / 2, nrow(sleep) - 4) * s * sqrt(Vxx[1, 1]) * c(-1, 1)
#> [1] -6.4234243 -0.8121813
```

The Vervet is expected to sleep less than Homo Sapiens.