Homework-3-Q2

BIOS507 Spring 2025 | Dr Lukemire | Elizabeth Nemeti Due: March 3 2025

Problem 2.

A nutrition study was carried out to try to better understand the relationship between protein consumption at breakfast and energy levels throughout the day. A sample of size 20 was collected, and for each individual in the study the researchers measured:

- Protein consumption at breakfast on a continuous scale ranging from 1 to 25.
- A energy score (self-rated) ranging from 1 to 100

The data can be found in the nutrition_protein_dataset.xlsx. Note that this is an excel file instead of a csv. It can be read into R using the read_excel function. This requires you to load the readxl library.

```
data_path = "/Users/elizabethnemeti/Documents/GitHub/BIOS507-Coursework/Homeworks/Homework-3/"
data_file <- file.path(data_path, "nutrition_protein_dataset.xlsx")
nutrition_data <- read_excel(data_file)

head(nutrition_data)
```

```
## # A tibble: 6 x 2
##   Protein Energy_Score
##     <dbl>        <dbl>
## 1    1.62         5.35
## 2    2.16        10.5
## 3    4.09        10.2
## 4    5.02        18.5
## 5    5.45        18.2
## 6    6.52        26.9
```

```
str(nutrition_data)
```

```
## tibble [20 x 2] (S3: tbl_df/tbl/data.frame)
##  $ Protein     : num [1:20] 1.62 2.16 4.09 5.02 5.45 ...
##  $ Energy_Score: num [1:20] 5.35 10.54 10.17 18.48 18.25 ...
```

predictor variable (x) -> Protein response variable (y) -> Energy_Score

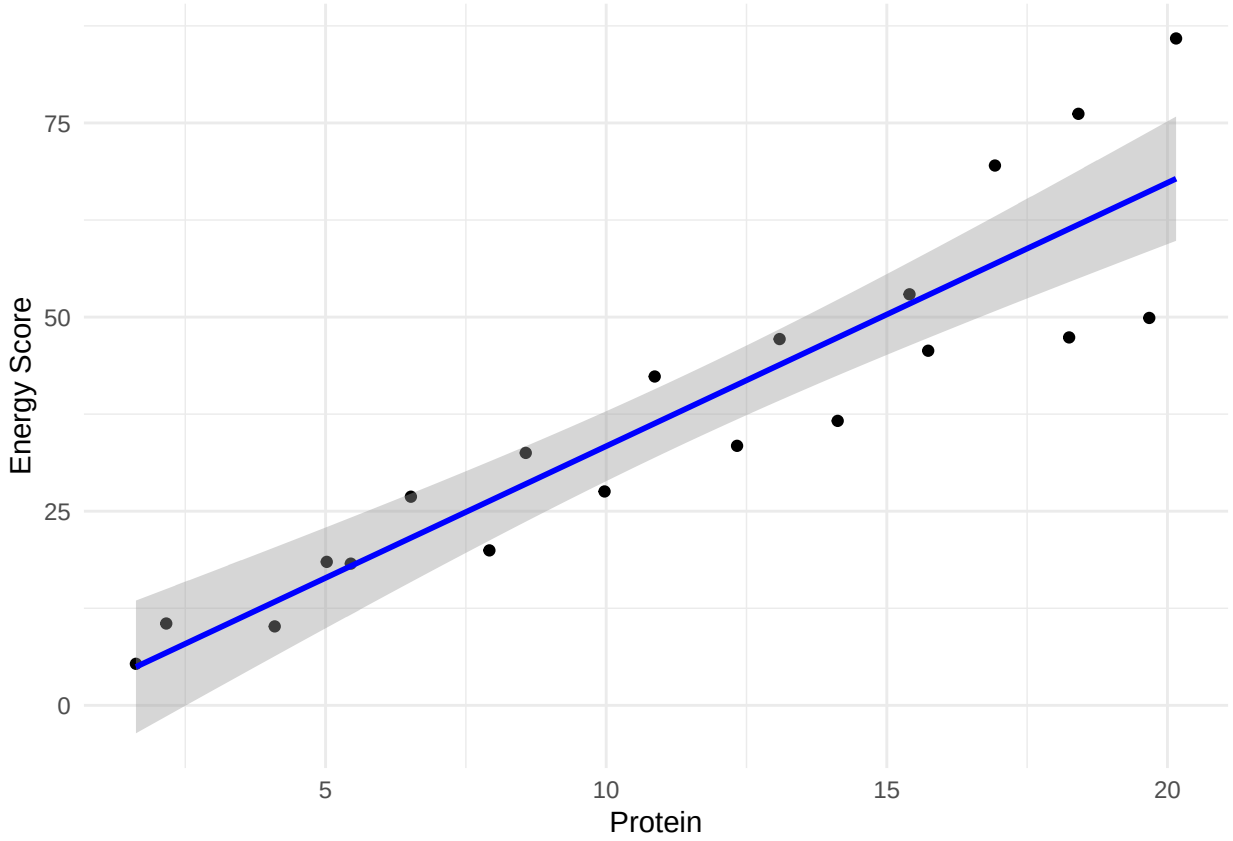
1. Analyze the data using a simple linear regression model and draw conclusions about the association between protein consumption and energy levels. Be sure to carry out all of the usual steps in the analysis.

```
summary(nutrition_data)
```

```
##     Protein        Energy_Score
##  Min.   : 1.621   Min.   : 5.348
##  1st Qu.: 6.253   1st Qu.:19.587
##  Median :11.598   Median :35.028
##  Mean   :11.315   Mean   :37.834
##  3rd Qu.:16.034   3rd Qu.:48.008
##  Max.   :20.154   Max.   :85.880
```

```r
nutrition_data_scatterplot <- ggplot(
  nutrition_data,
  aes(x = Protein, y = Energy_Score)) +
  geom_point() +
  geom_smooth(method = "lm", color = "blue") +
  theme_minimal() +
  xlab("Protein") +
  ylab("Energy Score")
nutrition_data_scatterplot
```

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



Does a linear relationship appear appropriate? -> Yes

Assumed regression model for Y -> Energy Score = ß0 + ß1(Protein) + (epsilon)

Assumptions about (epsilon): - average error should be 0 - errors should follow a normal distribution (bell curve) - there should be no pattern in the errors, we should see homoscedasticity

```r
nutrition_model <- lm(Energy_Score ~ Protein, data = nutrition_data)
summary(nutrition_model)
```

```
##
## Call:
## lm(formula = Energy_Score ~ Protein, data = nutrition_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.309  -6.561   0.813   4.322  18.059
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5503     4.5793  -0.120    0.906
## Protein       3.3924     0.3601   9.421 2.22e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.351 on 18 degrees of freedom
## Multiple R-squared:  0.8314, Adjusted R-squared:  0.822
## F-statistic: 88.76 on 1 and 18 DF,  p-value: 2.217e-08
```

ß0 (intercept) = -0.5503 (aka Y when Protein is 0) ß1 (slope) = 3.3924 (aka for each unit increase in Protein, Energy Score increases by 3.3924)

Estimated model -> E[Energy_Score] = (-0.5503) + (3.3924) x (Protein)

```
anova(nutrition_model)
```

```
## Analysis of Variance Table
##
## Response: Energy_Score
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Protein    1 7761.0  7761.0  88.763 2.217e-08 ***
## Residuals 18 1573.8    87.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(nutrition_model)$r.squared
```

```
## [1] 0.8314029
```

R squared being 0.8314029 means 83.14% of the variation in Energy Score can be explained by Protein and it's therefore a good predictor, remaining % would have to be explained by other factors beyond Protein.

H0: 1 = 0 -> Protein does not affect Energy Score HA: 1!= 0 -> Protein does affect Energy Score = 0.05

Two tailed test, as we're testing if the slope is different from 0 in either direction, not just one. Therefore, (alpha)/2 = 0.025, and the test statistic is the t-statistic.

df residual for t test = 18

```
# t-value via manual t-test
# t = ß1hat/standard error of ß1hat
t = 3.3924184/0.3600746
t
```

```
## [1] 9.421432
```

```
summary(nutrition_model)$coefficients
```

```
##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -0.5502902  4.5793351 -0.1201681 9.056811e-01
## Protein      3.3924184  0.3600746  9.4214323 2.216730e-08
```

```
# = 9.4214323 matches our t value from the summary table

# critical t-value
qt(0.025, df = 18, lower.tail = TRUE)  # get -ve val
```

```
## [1] -2.100922
```

```r
qt(0.025, df = 18, lower.tail = FALSE)  # get +ve val
```

```
## [1] 2.100922
```

```r
# = ±2.100922
```

To decide on whether to reject the null hypothesis, we need to check if |t| > critical value.

|9.4214323| > 2.100922, therefore we reject the null hypothesis.

Now, to get our 95% CI:

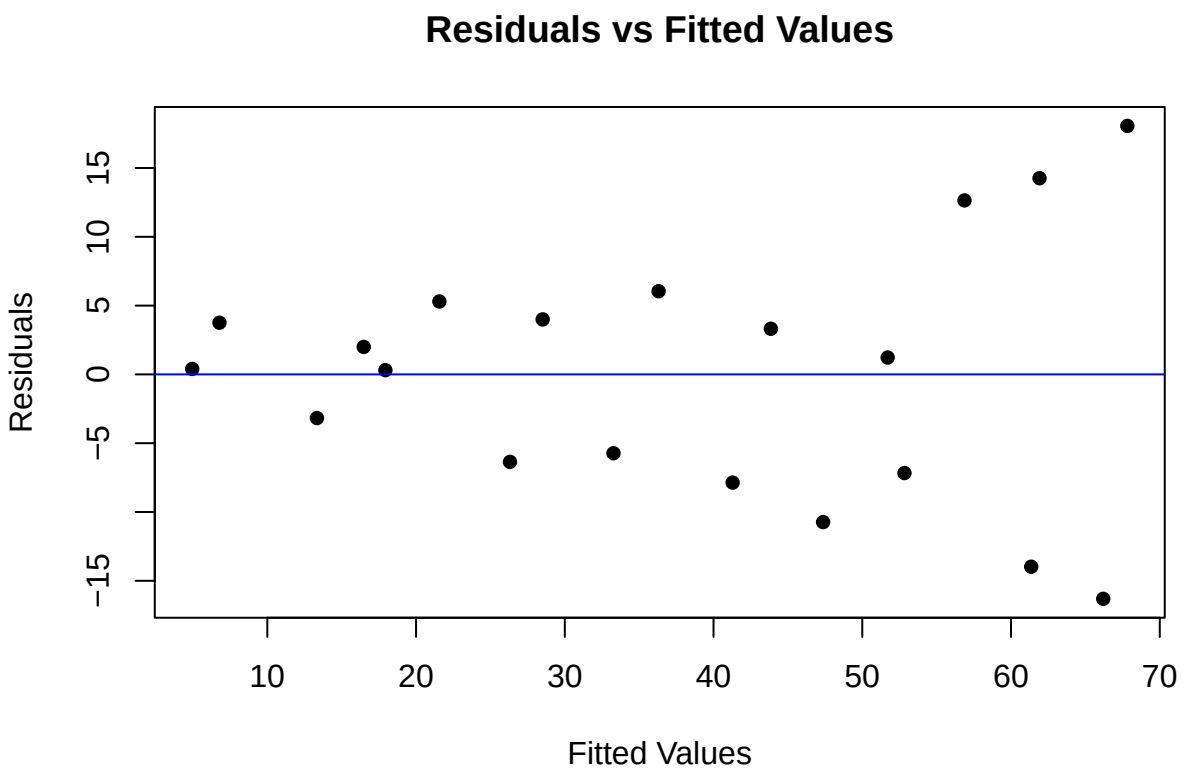
```r
confint(nutrition_model)
```

```
##                 2.5 %   97.5 %
## (Intercept) -10.17112 9.070536
## Protein       2.63593 4.148907
```

In the context of the nutrition study, since the p-value (2.217e-08) is much smaller than = 0.05, we have strong evidence to reject the H0, and that Protein influences Energy Score. The R squared value being 0.83, further supports that Protein influences the participants' Energy Score. We are 95% confident that the interval 2.64-4.15 units contains the true slope (1), whereby for each additional unit of Protein, we see an increase of about 3.39 units in Energy Score.

2. Create a plot of the residuals vs fitted values. What do you notice?

```r
plot(nutrition_model$fitted.values, residuals(nutrition_model),
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residuals vs Fitted Values",
     pch = 16)
abline(h = 0, col = "blue")
```



Looking at the fitted values vs residuals, most points center around 0. However, there is visible fanning

towards the larger fitted values, indicating heteroscedasticity where there is higher variability of errors toward the larger values (being the higher Energy Scores). What this could mean is that for lower Energy scores, this model does a better prediction than for higher Energy Scores.

3. How might this be affecting your conclusions? Specifically comment on the point estimate and the confidence interval from Part (1).

Our 95% CI for ß1 is 2.64-4.15 units, however, since the model appears to predict less accurately for higher Energy Scores, thereby indicating heteroscedasticity, our CI may be too narrow. This is because when heteroscedasticity violates the assumption of constant variance, leading to incorrect calculation of standard errors. Here, if we underestimated the range of the CI, we could be overconfident in our estimate, and the true slope (ß1) may fall outside the interval. Our estimated effect of Protein on Energy Score is 3.39 units per additional unit of Protein, however, this estimate may be slightly incorrect if heteroscedasticity is not addressed, as it affects the accuracy of standard errors and CIs.

4. Interpret what you saw in the residuals vs fitted plot in terms of protein and energy levels (instead of residuals and fitted values). Provide some guess as to why this might happen. Note that there is no single "correct answer" here.

After observing the residuals vs fitted plot and seeing that the data points fan out toward the higher values, it could mean the model predicts Energy Scores more accurately for lower values but performs worse for higher values. One possible reason for this is that, at lower energy levels, protein intake might play a more direct role, but as energy levels increase, additional predictors might need to be considered that contribute to energy increase. For example, sleep, sunlight, movement, vitamins, minerals, recovery, and mental state all contribute to energy levels. Additionally, we are unaware of the participants genetic and metabolic makeups, therefore we cannot infer how much energy each participant was able to metabolize from the protein at breakfast, or if the meals were proportionate to the age, sex, weight and height of each participant.