Homework 4

Charles Ancel

9/21/2023

Homework Instructions

Make sure to add your name to the header of the document. When submitting the assignment on Gradescope, be sure to assign the appropriate pages of your submission to each Exercise.

For questions that require code, please create or use the code chunk directly below the question and type your code there. Your knitted pdf will then show both the code and the output, so that we can assess your understanding and award any partial credit.

For written questions, please provide your answer after the indicated *Answer* prompt.

You are encouraged to knit your file as you work, to check that your coding and formatting are done so appropriately. This will also help you identify and locate any errors more easily.

Homework Setup

We'll use the following packages for this homework assignment.

```
library(ggplot2)
```

Exercise 1: Formatting [5 points]

The first five points of the assignment will be earned for properly formatting your final document. Check that you have:

- included your name on the document
- properly assigned pages to exercises on Gradescope
- select **page 1 (with your name)** and this page for this exercise (Exercise 1)
- all code is printed and readable for each question
- generated a pdf file

Exercise 2: Mammalian Sleep Model [23 points]

We'll use the `msleep` dataset from the `ggplot2` package for this assignment. At first glance of the `msleep` data, you may notice some missing values encoded as NAs. For this question, we will use the sleep (`sleep_total`) and bodyweight of an animal, which have no missing values.

```
?msleep
```

part a

I wonder about how the amount of sleep required by an animal changes based on the bodyweight of an animal. For example, do animals who weigh more require more sleep. What would be the primary purpose of fitting a model like this? **Bold your answer** below by surrounding your selection with two asterisks "**".

(a) predicting an observation **(b) explaining a structure/system**

part b

Fit a linear model to estimate the sleep required based on the bodyweight. Print the summary of this linear model. Then, write the fitted model based on this output. Be sure to use proper notation.

```
sleep_model <- lm(sleep_total ~ bodywt, data=msleep)

summary(sleep_model)
```

```
##
## Call:
## lm(formula = sleep_total ~ bodywt, data = msleep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7008 -2.3787 -0.4268  3.2732  9.1731
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.7269205  0.4773797  22.470  < 2e-16 ***
## bodywt      -0.0017647  0.0005971  -2.956  0.00409 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.254 on 81 degrees of freedom
## Multiple R-squared:  0.09735,    Adjusted R-squared:  0.08621
## F-statistic: 8.736 on 1 and 81 DF,  p-value: 0.004085
```

Answer: - The fitted linear model is:

$$\text{sleep\_total} = 10.7269 - 0.0017647 \times \text{bodywt}$$

- β_0 (Intercept): 10.7269205
- β_1 (Coefficient of bodywt): -0.0017647
- The t-value for the intercept is 22.470 and is highly significant (p-value $<$ 2e-16).
- The t-value for the body weight is -2.956 and is also significant (p-value = 0.00409).
- The coefficient of determination (R^2) is 0.09735. This means that approximately 9.735% of the variability in total sleep can be explained by the body weight of the animals.

part c

Interpret the intercept for this model.

Answer:

Intercept (β_0) is 10.7269. It means for an animal with a body weight of 0, the predicted total sleep is about 10.7269 hours. However, an animal with a weight of 0 isn't realistic.

part d

Is the intercept for this model meaningful? Is it reliable? Explain.

Answer: Given the intercept is 10.7269, it represents the predicted sleep for an animal with a body weight of 0.

Meaningfulness: The intercept isn't meaningful in a biological context, as animals can't have a weight of 0.

Reliability: While statistically significant, the intercept's real-world applicability is limited due to the aforementioned reason.

part e

Interpret the slope for this model.

Answer:

part e: Interpret the slope for the model.

The slope (coefficient of `bodywt`) is -0.0017647.

Interpretation: For every 1 unit increase in body weight, the predicted total sleep decreases by approximately 0.0017647 hours.

part f

Report and interpret the coefficient of determination for this relationship.

```
r_squared <- summary(sleep_model)$r.squared
r_squared
```

```
## [1] 0.09735061
```

Answer: Multiple R^2 is 0.09735.

Interpretation: Approximately 9.735% of the variability in total sleep can be explained by the body weight of the animals.

part g

Based on part f, calculate the correlation.

```
r <- sqrt(r_squared)

if (coef(sleep_model)["bodywt"] < 0) {
  r <- -r
}
r
```

```
## [1] -0.3120106
```

Answer: The value -0.3120106 indicates a weak negative linear relationship between `bodywt` and `sleep_total`. This means that as body weight increases, the total sleep tends to decrease slightly, consistent with the negative slope we observed in the regression model.

Exercise 3: Hand Calculations [30 points]

We've used `R` to generate the summary statistics for the `msleep` dataset so far. Now let's take a moment and confirm some of these calculations "by hand" (still using `R` to perform the calculations).

part a

Calculate $\bar{x}$, $\bar{y}$, S_{xx}, and S_{xy} for the msleep dataset. Clearly label and print your results.

```
x_bar <- mean(msleep$bodywt, na.rm = TRUE)
y_bar <- mean(msleep$sleep_total, na.rm = TRUE)

S_xx <- sum((msleep$bodywt - x_bar)^2, na.rm = TRUE)
S_xy <- sum((msleep$bodywt - x_bar) * (msleep$sleep_total - y_bar), na.rm = TRUE)

x_bar
```

```
## [1] 166.1363
```

```
y_bar
```

```
## [1] 10.43373
```

```
S_xx
```

```
## [1] 50767575
```

```
S_xy
```

```
## [1] -89590.99
```

part b

Calculate the estimates for β_0 and β_1, using the values calculated in part a. Clearly label and print your results. Compare these estimates to what you found in Exercise 2b.

```
beta_1 <- S_xy / S_xx
beta_0 <- y_bar - beta_1 * x_bar

beta_0
```

```
## [1] 10.72692
```

```
beta_1
```

```
## [1] -0.001764729
```

Comparison:

From the manual calculations (part b of Exercise 3): 1. β_0 (Intercept): 10.72692 2. β_1 (Coefficient of `bodywt`): -0.001764729

From the linear model in Exercise 2b: 1. β_0 (Intercept): 10.7269205 2. β_1 (Coefficient of `bodywt`): -0.0017647

Comparison: The values of β_0 and β_1 from the manual calculations match very closely with the values from the linear regression model in Exercise 2b. This is expected, as both methods aim to estimate the same linear relationship between `bodywt` and `sleep_total`.

In summary, the manually calculated estimates for the intercept and slope are consistent with the results obtained from the linear regression model.

part c

Calculate the residuals for the dataset. You can use any built-in `R` function to generate the fitted values of the dataset, but you should not use a built-in function to calculate the residuals. No need to print all of the residuals, but please do print the first few residuals.

```r
y_hat <- beta_0 + beta_1 * msleep$bodywt

residuals <- msleep$sleep_total - y_hat

head(residuals)
```

```
## [1]  1.461316  6.273927  3.675462  4.173113 -5.668083  3.679874
```

part d

Calculate the SSR, SSE, and SST for the model. You may use anything that you have calculated in the earlier parts of this question. Clearly label and print these three values.

```r
SSR <- sum((y_hat - y_bar)^2)
SSE <- sum(residuals^2)
SST <- sum((msleep$sleep_total - y_bar)^2)

SSR
```

```
## [1] 158.1038
```

```r
SSE
```

```
## [1] 1465.962
```

```r
SST
```

```
## [1] 1624.066
```

part e

Given the formulas for these quantities: - SST represents the total variability in the response variable (`sleep_total`). - SSR represents the variability in the response variable that is explained by the predictor (`bodywt`). - SSE represents the variability in the response that is not explained by the predictor; it's the error or the unexplained variability.

The relationship between them is:

$$SST = SSR + SSE$$

This is evident from the results provided: 1624.066 (SST) 158.1038 (SSR) + 1465.962 (SSE)

The relationship indicates that the total variability in the response can be partitioned into the variability that's explained by the model (SSR) and the variability that's unexplained (SSE). ## part f

Calculate the coefficient of determination from the values calculated in part d. How does this compare to what you found in Exercise 2d (based on Exercise 2b)?

```r
R_squared_calculated <- SSR / SST

R_squared_calculated
```

```
## [1] 0.09735061
```

Comparison:

The calculated R^2 from **part f** is 0.09735061.

From Exercise 2f, the R^2 value was 0.09735.

The two values match, which is expected since both methods aim to estimate the proportion of the total variation in the response variable (`sleep_total`) that is explained by the predictor (`bodywt`).

Exercise 4: Inference for the Mammal Sleep Model [21 points]

part a

For the Chinchilla (bodyweight of 0.420), calculate the predicted total amount of sleep and the corresponding residual.

```
chinchilla_bodywt <- 0.420
predicted_sleep <- beta_0 + beta_1 * chinchilla_bodywt

predicted_sleep
```

```
## [1] 10.72618
```

```
chinchilla_observed_sleep <- msleep$sleep_total[msleep$name == "Chinchilla"]

chinchilla_residual <- chinchilla_observed_sleep - predicted_sleep

chinchilla_residual
```

```
## [1] 1.773821
```

Answer:

part b

Calculate the 80% confidence interval for the slope of this model. *Note: the multiplier (t^*) is 1.292 for this situation.* Make sure to show your setup for the calculation. Report the 80% confidence interval below. Based on this confidence interval, is -0.002 a plausible value for the slope predicting sleep from body weight for all mammals?

```
std_error_slope <- summary(sleep_model)$coefficients["bodywt", "Std. Error"]

alpha <- 0.2
df <- nrow(msleep) - 2
t_critical <- qt(1 - alpha/2, df)

margin_error <- t_critical * std_error_slope

lower_bound <- beta_1 - margin_error
upper_bound <- beta_1 + margin_error

c(lower_bound, upper_bound)
```

```
## [1] -0.0025361978 -0.0009932593
```

Answer: The interval is:

$$-0.0025361978 \leq \beta_1 \leq -0.0009932593$$

Given this interval, the value -0.002 falls within the range. This means that -0.002 is a plausible value for the slope when predicting sleep from body weight for all mammals, based on the 80% confidence interval.

part c

Suppose that a previous study had found that for each increase in the body weight by 1 kg, the estimated average sleep time decreased by -0.01, on average.

Might this same relationship be true for all mammals, including those in our data? Or is there evidence that this relationship for mammals is different from the one for reptiles?

In other words, use a t-test to test:

- $H_0 : \beta_1 = -0.01$
- $H_1 : \beta_1 \neq -0.01$

Calculate and report the value of your test statistic. Then, based on the size of your test statistic, anticipate what the decision for the statistical test would be with a significance level of $\alpha = 0.05$.

```
t_statistic <- (beta_1 - (-0.01)) / std_error_slope

p_value <- 2 * (1 - pt(abs(t_statistic), df))

c(t_statistic, p_value)
```

```
## [1] 13.7928  0.0000
```

Answer: Here's the interpretation:

1. t-statistic: 13.7928
2. p-value: 0.0000 (rounded, so it's very close to 0)

The results indicate that the relationship between body weight and sleep duration for mammals is statistically different from the hypothesized value for reptiles. The p-value is very close to 0, so we reject the null hypothesis.

part d

If a news article used this data to make the claim "Gaining weight will make you sleep less!" would you agree with that conclusion? Why or why not?

There is not an objectively right answer for this question. Consider all of the information that we've compiled and calculations that we've performed to help guide your response.

Answer: The data indicates a negative association between body weight and sleep duration for mammals. However, the relationship is weak, and correlation doesn't imply causation. While weight might influence sleep to some extent, it's not the sole factor. Claiming "Gaining weight will make you sleep less!" oversimplifies the findings and doesn't account for other potential influencing factors.

part e

For the default hypothesis test for the intercept, report the following:

- The null and alternative hypotheses (in symbols)
- The value of the test statistic
- The p-value of the test

Be sure to report this information in text below. No need to calculate any values directly for this part; observing and identifying appropriate information from R output is sufficient.

Answer: For the default hypothesis test for the intercept, the information you need can be extracted from the output of the `lm()` function, specifically from the summary of the linear model.

1. **The null and alternative hypotheses (in symbols) for the intercept**:

   - $H_0 : \beta_0 = 0$
   - $H_1 : \beta_0 \neq 0$

   This means that the null hypothesis assumes there's no effect (intercept is zero), while the alternative hypothesis posits that there is an effect (intercept is different from zero).

2. **The value of the test statistic**:
    - You'll find this value in the summary output of the linear model under the "Coefficients" section, specifically in the row for "(Intercept)" and the column "t value."
3. **The p-value of the test**:
    - Similarly, this value can be found in the summary output of the linear model under the "Coefficients" section, in the row for "(Intercept)" and the column "Pr(>|t|)."

From the output you provided for Exercise 2b:

- Test statistic value for the intercept: 22.470
- p-value for the intercept: < 2e-16 (which is essentially 0)

Now looking at our hypothesis and responding to the question: - Null Hypothesis H_0: $\beta_0 = 0$ - Alternative Hypothesis H_1: $\beta_0 \neq 0$ - Test statistic value for the intercept: 22.470 - p-value for the intercept: < 2e-16 (essentially 0)

Given the extremely low p-value, there's strong evidence against the null hypothesis, suggesting the intercept is different from zero.

Exercise 5: Assumptions for the Mammal Sleep Model [21 points]

part a

For the fitted model to be appropriate and for the inference procedures from Exercise 4 to be valid, certain assumptions must be met. First, write out the four assumptions.

Answer: For the fitted linear regression model to be appropriate and for the inference procedures to be valid, the four primary assumptions are:

1. **Linearity**: The relationship between the independent variable(s) and the dependent variable is linear.
2. **Independence**: The residuals are independent. In other words, the residuals from one prediction have no effect on the residuals from another prediction.
3. **Homoscedasticity (Constant Variance)**: The residuals have constant variance at every level of the independent variable(s).
4. **Normality**: The residuals are approximately normally distributed.
5. Linearity: The relationship between the predictor(s) and the response is linear.
6. Independence: Observations are independent of each other.
7. Homoscedasticity: The variance of the residuals is constant across all levels of the independent variable(s).
8. Normality: The residuals are approximately normally distributed.

part b

Which of these four assumptions cannot be checked with a plot?

Answer: Out of the four assumptions, the **Independence** assumption is the one that cannot be directly checked with a plot. While scatter plots and residual plots can provide insights into linearity, homoscedasticity, and normality (e.g., through Q-Q plots), the independence of observations is often based on the study design and domain knowledge.

The assumption of Independence cannot be directly checked with a plot.

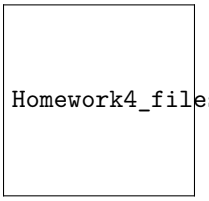
part c

Create the two plots (it's ok if you end up with 4 plots) that we can use to check our assumptions for the mammal sleep data. Interpret these two plots. Be sure to specify if the assumptions seem reasonable and describe what you see that supports your conclusions.

```
par(mfrow=c(2,2))

plot(sleep_model$fitted.values, residuals(sleep_model),
     main="Residuals vs. Fitted", xlab="Fitted values", ylab="Residuals")
abline(h=0, col="red")

qqnorm(residuals(sleep_model), main="Normal Q-Q Plot")
qqline(residuals(sleep_model))
```



Answer:

1. **Residuals vs. Fitted Values**:
   - The residuals are scattered without a clear pattern around zero, suggesting a linear relationship between predictors and response.
   - The consistent spread across the fitted values indicates homoscedasticity, meaning the variance remains constant.
2. **Normal Q-Q Plot**:
   - Most data points align with the diagonal, implying residuals are approximately normally distributed.
   - Some deviations at the tails hint at slight non-normality, but these are minor.

In summary, the plots suggest the linear regression assumptions are predominantly satisfied for the mammal sleep data, with minor concerns about perfect normality.

part d

There are two additional statistical tests that can be used to help assess our assumptions for the linear model. Perform these two tests here. Then, based just on these two tests, assess the corresponding assumptions.

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(stats)

bp_test <- bptest(sleep_model)

shapiro_test <- shapiro.test(residuals(sleep_model))

list(Breusch_Pagan_Test = bp_test, Shapiro_Wilk_Test = shapiro_test)
```

```
## $Breusch_Pagan_Test
##
##  studentized Breusch-Pagan test
##
## data:  sleep_model
## BP = 0.20069, df = 1, p-value = 0.6542
##
##
## $Shapiro_Wilk_Test
##
##  Shapiro-Wilk normality test
##
## data:  residuals(sleep_model)
## W = 0.97894, p-value = 0.1908
```

Answer:

1. **Breusch-Pagan Test** (Homoscedasticity):
   - Test Statistic (BP): 0.20069
   - p-value: 0.6542

   The Breusch-Pagan test suggests that the residuals have constant variance across levels of the independent variable, as the p-value of 0.6542 is greater than 0.05. This means the assumption of homoscedasticity is met.

2. **Shapiro-Wilk Test** (Normality of Residuals):
   - Test Statistic (W): 0.97894
   - p-value: 0.1908

   The Shapiro-Wilk test indicates that the residuals are approximately normally distributed. The p-value of 0.1908 (greater than 0.05) suggests no significant deviation from normality.

In conclusion, based on these two tests, the assumptions of homoscedasticity and normality for the mammal sleep data linear regression model appear to be satisfied.
