431 Lab 04 Sketch and Grading Rubric

Instructor: Dr. Thomas E. Love   Lab Author: Mr. Wyatt P. Bensken

Due: 2021-10-11 | Last Edit: 2021-10-12 15:24:53

Contents

1 Loading Packages

```
library(palmerpenguins)
library(janitor)
library(broom)
library(tidyverse)
```

2 Learning Objectives

1. Critically evaluate a news story on scientific literature, incorporating the principles of probability and odds
2. Given data, work through a linear regression model including visualizing the data, building the model, predicting an outcome on new points.

3 Packages and Functions

In Lab 04 we were hoping you would become more familiar with the following packages and functions:

Packages:

- `tidyverse`
- `palmerpenguins`
- `broom`
- `janitor` (optional)

Functions:

- `%>%`
- `filter()`
- `mutate()`
- `sample_n()`
- `anti_join()`
- `nrow()`
- `tabyl()`
- `lm()`
- `summary()`
- `confint()`
- `ggplot()`
- `augment()`
- `summarise()`

4 The Data

In Questions 6-8, this lab again uses the the `penguins` data (note: use the `penguins` tibble, and not the `penguins_raw` tibble for this Lab) contained in the `palmerpenguins` package in R. The complete citation is . . .

Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/. doi: 10.5281/zenodo.3960218.

Additional information on the data are provided by Allison Horst at the github site linked above. In particular, you'll find a nice cartoon of the three species of penguin contained in the data and a detailed description of the bill measurements that are worth your time.

5 Part A: News Story and Research Article (Questions 1-5, 70 points)

Find a headline from the internet related to health or medicine that describes the findings of a study published on January 1, 2017 or later. Then find the study being referred to in PUBMED. Use the formula for updating your opinions about health news developed in this article by Jeff Leek, along with the abstract and full contents of the published study to complete Questions 1-5. While it won't be necessary to prepare any R code to respond to Questions 1-5, we think it will be good practice for you to prepare your response in R Markdown anyway.

5.1 Question 1 (5 points)

Specify the URL where we can see the headline and news story describing the findings of the study. Feel free to use `bit.ly` or a related tool online to produce a shortened URL for this purpose. Specify the reference completely, including the names of the author(s) of the news story, and its full title, and source.

5.1.1 Grading Rubric

- Award 5 points if they have a working link that meets the specifications.
- Award no more than 2 points if they don't have the authors, title and source, and so forth, or if the link doesn't work.

5.2 Question 2 (5 points)

Specify a URL where we can see at least the abstract of the complete study. Again, shortened URLs are fine. Give the complete reference to the study, as well, including the authors, full title, journal name and so forth.

5.2.1 Grading Rubric

- Award 5 points if they have a working link that meets the specifications.
- Award no more than 2 points if they don't have the authors, title and source, journal so forth, or if the link doesn't work.

5.3 Question 3 (15 points)

Describe, in a few sentences, your original opinion (gut feeling) related to the conclusions of the study as summarized in the headline and news article, first in terms of a probability statement, and then calculate the appropriate odds, remembering to convert statements about probabilities to statements about odds. Provide some motivation for your internal prior probability, describing your relevant personal experiences or other factors that drove your gut feeling.

Remember, if X is an event, and `Pr(X)` is the probability that X occurs, and `odds(X)` are the odds that X occurs, then

```
Pr(X) = odds(x) / (1 + odds(x))
```

and

```
odds(X) = Pr(X) / (1 - Pr(X)).
```

5.3.1 Grading Rubric

Full 15 points awarded requires:

- Probability statement relating to gut feeling
- Calculate odds related to gut feeling
- Motivation/explanation of gut feeling
- Complete sentences with correct spelling and grammar
- Lose 3 points for each missing or problematic piece, down to a minimum of 0

5.4 Question 4 (30 points)

> Evaluate the study in terms of the six specifications proposed by Jeff Leek in this article at FiveThirtyEight when evaluating study support. Be sure to specify your conclusion about **each** of the six specifications, and provide direct quotes and summarize the evidence from the abstract or paper to address the issues raised and justify your conclusions. We want to see a clear, motivated conclusion about each of the six specifications, as well as direct quotes and evidence summaries to address the issues raised and justify conclusions. We suggest you use a different subheading for each of the six specifications so it's easy for us to see your conclusions in each case.

We want to see a clear, motivated conclusion about each of the six specifications (each being worth 5 points), as well as direct quotes and evidence summaries to address the issues raised and justify conclusions.

The six specifications we are looking for are:

1. Was the study a clinical study in humans?
2. Was the outcome of the study something directly related to human health like longer life or less disease? Was the outcome something you care about, such as living longer or feeling better?
3. Was the study a randomized, controlled trial (RCT)?
4. Was it a large study - at least hundreds of patients?
5. Did the treatment have a major impact on the outcome?
6. Did predictions hold up in at least two separate groups of people?

5.4.1 Grading Rubric

- Students should receive 4 points for successfully addressing each of the 6 components above, plus a bonus of an additional 6 points if they successfully complete all 6 parts. Successful completion means the student's response includes complete sentences with correct spelling and grammar, using quotations or paraphrasing the complete study appropriately.

- Students should score no higher than 2/4 for any part where they don't successfully generate a good response.

5.5 Question 5 (15 points)

> Incorporate the study support assessment into a Bayes' Rule calculation to obtain the final odds you should now be willing to give to the headline, and specify this value in terms of a probability statement, as well. Then react to the final conclusion specified by this approach in a few sentences. How does your subjective posterior probability that the headline is true match up with the formula's conclusions? Do you feel that the formulaic approach has yielded an appropriate conclusion for you in this case? Why or why not?

5.5.1 Grading Rubric

Grading here is broken up into two pieces.

First, up to 6 points are awarded for the calculations.

- We wanted to see correct calculations of both the odds and probability in light of the prior probability established in Question 3 and the answers to the specifications described in Question 4. For 6 points they should provide **odds** and **probability** using Bayes' Rule, responding with **complete sentences with correct spelling and grammar**, losing 2 points for each missing item in bold, down to a minimum of 0.

Next, up to 9 points are awarded for the reaction.

- For the additional 9 points, we then wanted to see you specify your subjective feeling about what the probability *should* be, and then match that up with the result of the calculation.
- The student should reach a logical conclusion and summarize his/her reaction on whether or not this approach was accurate for assessing their initial gut feeling vs. their final conclusion after reading the article/study. No right or wrong answer, but using both logic and complete sentences are necessary.
- Give 7 or less if they don't manage to do all of these pieces.

6 Part B: Palmer Penguins (Questions 6-8, 30 points)

6.1 A Note on sample_n() versus slice_sample() (Added 2021-10-12)

It was brought to our attention that while the slides and notes have used `slice_sample()`, the answer sketch used a slightly outdated approach of `sample_n()`. These two functions, as best as we can tell, produce two identical datasets. Either of these should be appropriate to use, although `slice_sample()` is preferred.

6.2 Question 6 (5 points)

> Start with the 333 penguins in the `penguins` tibble who have complete data, and create two tibbles. The first should be called `pen_train` and should contain a random sample of exactly 200 of the 333 penguins with complete data, while the second tibble, called `pen_test` should contain the other 133 penguins with complete data. Use `4312021` to set your seed for random sampling so that we all obtain the same sample. Use R code to obtain the samples, and then provide code which demonstrates that your two samples contain exactly 200 and exactly 133 penguins. Comment on the code you present as necessary so that we can clearly see that you understand what is being done in each step.

In the code below we first create a tibble from the built-in data, and then filter that tibble to the 333 complete cases. I've also created a unique ID for each penguin which is just its row number. This will help us in splitting the data.

```
penguins <- palmerpenguins::penguins

penguins <- penguins %>%
  filter(complete.cases(.)) %>%
  mutate(id = row_number())
```

Next, we'll split our data into a training and test sample.

```
set.seed(4312021)

pen_train <- sample_n(penguins, 200)
pen_test <- anti_join(penguins, pen_train, by = "id")
```

Finally, we run some quick summaries on our samples to demonstrate how many penguins there are in each. We could simply run `nrow()` on each to see the number of rows.

```
nrow(pen_train)
```

```
[1] 200
```

```
nrow(pen_test)
```

```
[1] 133
```

Alternatively, we could create a nice `tabyl()`, say of species, which includes the totals.

```
pen_train %>%
  tabyl(species) %>%
    adorn_totals("row") %>%
    adorn_pct_formatting() %>%
    knitr::kable(align = 'lrr', caption = "Training Sample")
```

Table 1: Training Sample

species	n	percent
Adelie	86	43.0%
Chinstrap	43	21.5%
Gentoo	71	35.5%
Total	200	100.0%

And here are the results within the test sample.

```
pen_test %>%
  tabyl(species) %>%
    adorn_totals("row") %>%
    adorn_pct_formatting() %>%
    knitr::kable(align = 'lrr', caption = "Test Sample")
```

Table 2: Test Sample

species	n	percent
Adelie	60	45.1%
Chinstrap	25	18.8%
Gentoo	48	36.1%
Total	133	100.0%

6.2.1 Grading Rubric

If a student successfully completes this step they should receive all 5 points.

- If there are any issues they should receive a maximum of 2 points.
- If the student does not walk through (in English sentences) their code, they should score 3/5

6.3 Question 7 (15 points)

> Build a linear model to examine the relationship between body mass in grams (the outcome of interest) and bill length in millimeters (our predictor) using ordinary least squares and the `pen_train` data set you created in Question 6. Call this model `model1`.
>
> Create an attractive and thoughtfully labeled plot (including an appropriate title) that shows the association of bill length (placed on the horizontal x axis) and body mass (on the vertical y axis) for the 200 penguins in your training sample.
>
> Add appropriate smooth curves (using both lm and loess) to the plot. Also please add the actual regression equation (including the coefficients, rounded to two decimal places) to the plot in a clear way.
>
> Then write a couple of sentences interpreting what this figure tells you about the relationship between `bill_length_mm` and `body_mass_g`.

First, we'll build our linear model.

```r
model1 <- lm(body_mass_g ~ bill_length_mm, data = pen_train)
```

We could use `tidy()` and `glance()` from the `broom` package to obtain summaries of the coefficients, and some of the key summaries of quality of fit.

```r
tidy(model1, conf.int = TRUE, conf.level = 0.95)
```

```
# A tibble: 2 x 7
  term           estimate std.error statistic  p.value conf.low conf.high
  <chr>             <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)       459.      385.       1.19 2.35e- 1   -301.      1219.
2 bill_length_mm     85.3      8.62      9.90 4.96e-19     68.3      102.
```

```r
glance(model1) %>%
  select(r.squared, adj.r.squared, sigma, AIC, BIC, nobs)
```

```
# A tibble: 1 x 6
  r.squared adj.r.squared sigma   AIC   BIC  nobs
      <dbl>         <dbl> <dbl> <dbl> <dbl> <int>
1     0.331         0.328  665. 3171. 3181.   200
```

Alternatively, after building the model, `model1`, we can use `summary()` to see our regression results. Notably, we see the estimate, standard error, t value, and p-value for both the intercept, but more importantly our predictor. There are helpful notations for significance, but we'll ignore those for now. We are also given information on the residual standard error, our degrees of freedom, the F-statistic, and the p-value for the model. Importantly, we are also given the multiple R^2 and the adjusted R^2. As you'll remember from class and the course notes, the R^2 represents the proportion of variance in the outcome that the model explains. This means that `bill_length_mm` explains 32.8% of the variation in `body_mass_g`. We can use this information to write our regression equation as: $\textit{body\_mass\_g} = 458.762 + 85.334(\textit{bill\_length\_mm})$.

```r
summary(model1)
```

```
Call:
lm(formula = body_mass_g ~ bill_length_mm, data = pen_train)

Residuals:
     Min       1Q   Median       3Q      Max 
-1708.11  -496.72    35.66   460.72  1642.82 

Coefficients:
```

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     458.762    385.391   1.190    0.235
bill_length_mm   85.334      8.619   9.901   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 664.5 on 198 degrees of freedom
Multiple R-squared:  0.3312,    Adjusted R-squared:  0.3278
F-statistic: 98.03 on 1 and 198 DF,  p-value: < 2.2e-16
```

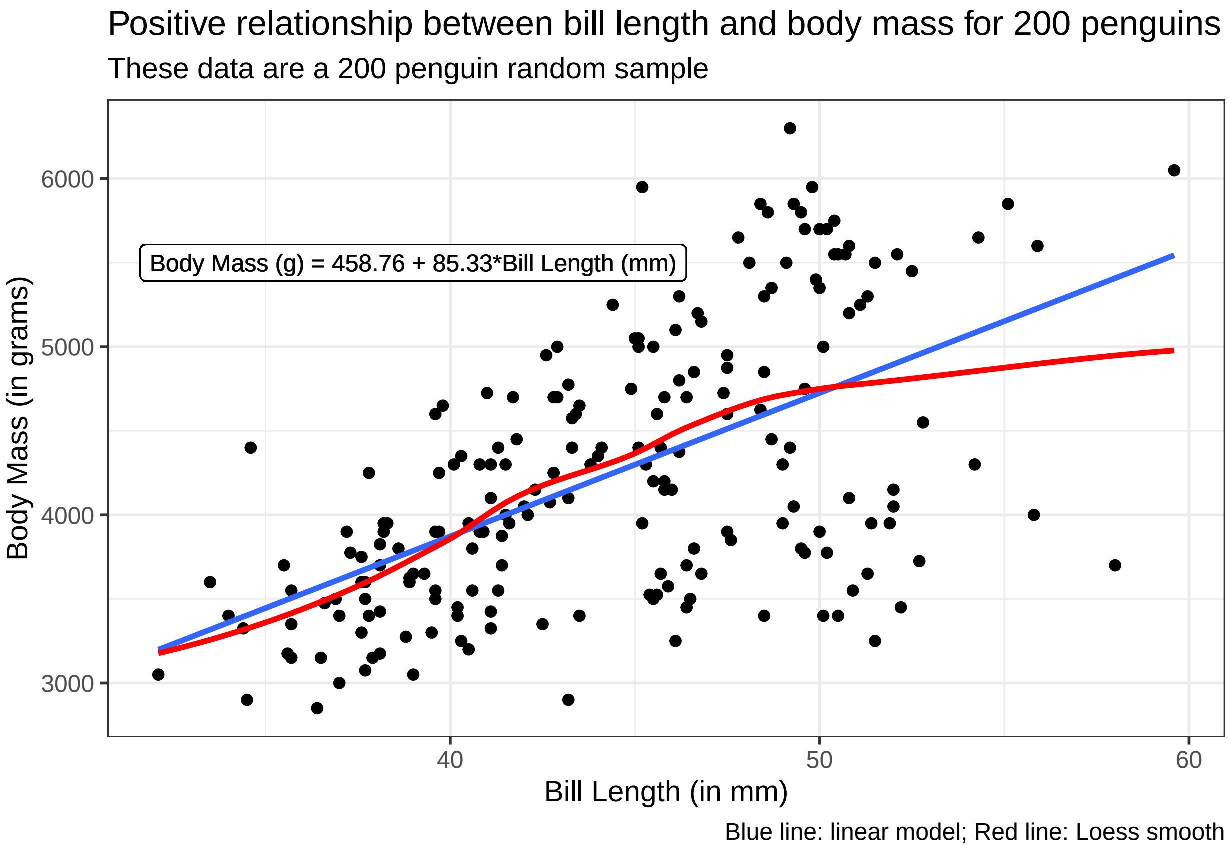
The `summary()` function gives us the estimates, but it'll also be important to get the confidence interval for these estimates. We can use `confint()`, which defaults to a 95% confidence interval. Putting this together with our summary we can say that for each increase of 1mm in bill length we would expect an increase in body mass of 85.3g (95% CI: 68.3, 102.3).

```
confint(model1)
```

```
                    2.5 %    97.5 %
(Intercept)    -301.23567 1218.7591
bill_length_mm   68.33779  102.3295
```

Next, we'll build our figure. As we can see in the figure below, as bill length increases as does the body mass of a penguin. Overall, it looks like a fairly strong, positive, relationship between these two variables. We have also included in our figure both a linear model line (in blue) and a Loess smooth line (in red). We have also added our regression equation to the figure. Overall, this figure suggests there is a positive relationship wherein as the bill length increases as does the body mass. The curves we've added suggest that a linear model is largely appropriate but perhaps some important variation and outliers exist at high values of bill length. Although this could also be partially due to the smaller number of data points.

```
ggplot(data = pen_train, aes(x = bill_length_mm, y = body_mass_g)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, formula = y ~ x) +
  geom_smooth(method = "loess", se = FALSE, formula = y ~ x, col = "red") +
  geom_label(x = 39, y = 5500, size = 3, alpha = 0.2,
             label = "Body Mass (g) = 458.76 + 85.33*Bill Length (mm)") +
  labs(y = "Body Mass (in grams)",
       x = "Bill Length (in mm)",
       title = "Positive relationship between bill length and body mass for 200 penguins",
       subtitle = "These data are a 200 penguin random sample",
       caption = "Blue line: linear model; Red line: Loess smooth") +
  theme_bw()
```



6.3.1 Grading Rubric

- Each part of this question (building the model, building the figure, and interpreting the figure) is each worth 5 points.
- If the student has successfully built the model, including running summary, they should receive all 5 points.
- If the figure includes two lines, the regression equation, and appropriate axes and titles, they should receive an additional 5 points.
  - As always deduct 2 points for issues with axes labels or titles.
- Finally if the have ample text to walk through what they have done and the interpretation from their figure they should receive an additional 5 points.
  - If their conclusion is unclear or wrong deduct up to 3 points.
- If the student built the model or the figure using the entire data set, or the test data, they should receive no more than 8/15.

6.4 Question 8 (10 points)

> Use the `model1` you created in Question 7 to predict the data in `pen_test` and summarize the results by specifying the root mean squared prediction error, and the mean and maximum absolute prediction error you obtain across the 133 penguins in `pen_test`. Place the results in an attractive and clear table. Then describe the units of measurement for your root mean squared prediction error result in a sentence.

The first thing we'll need to do is build predictions for our new model. To start, we'll use the `augment()` function.

```
model1_test <- augment(model1, newdata = pen_test) %>%
  mutate(model = "Test Data")
```

We'll now create a variable which is the difference between our fitted (predicted) body mass in grams, and the observed.

```
model1_test <- model1_test %>%
  mutate(res_body_mass_g = body_mass_g - .fitted)
```

We can now use `summarise()` to obtain our root mean squared prediction error (RMSPE), mean absolute prediction error (MAPE), and maximum absolute prediction error (maxAPE).

```
model1_test %>%
  summarise(MAPE = mean(abs(res_body_mass_g)),
            maxAPE = max(abs(res_body_mass_g)),
            RMSPE = sqrt(mean(res_body_mass_g^2))) %>%
  knitr::kable()
```

MAPE	maxAPE	RMSPE
516.7006	1760.91	631.6267

We can see from these results that we have an RMSPE of 631.63 grams. Importantly, the RMSPE has the same units of measure as our original outcome which was, in this case, grams. We can say that the root mean squared prediction error for the body mass of our 133 penguins in the test data was 631.6 grams.

6.4.1 Grading Rubric

- If the student has successfully fit the model to the test data, and calculated the MAPE, maxAPE, and RMSPE, with an appropriate interpretation they should receive all 10 points.
- If there are issues in fitting the model or calculating the MAPE, maxAPE, or RMSPE, the student should lose up to 7 points.
- If the student incorrectly interprets the units of measure for the RMSPE, they should lose 3 points.