Answer Sketch and Grading Rubric for Homework I

431 Staff and Professor Love

*Due **2019-12-01** at 5 PM. Last Edited 2019-10-22 16:29:32*

Contents

Question 1

is an essay. We don't provide sketches for essay questions.

Initial R Setup for Questions 2-6

Here's the R setup we used, and we'll read in the data set, as was suggested.

```
knitr::opts_chunk$set(comment = NA)

library(here); library(janitor); library(magrittr)
library(patchwork); library(broom); library(tidyverse)

hwI_plasma <- read_csv(here("data", "hwI_plasma.csv")) %>%
```

```
    mutate_if(is.character, as.factor) %>%
    mutate(subj_ID = as.character(subj_ID))
```

Partitioning into training/test samples

Later, we'll need both a training sample and a test sample. We'll get those with this code...

```
set.seed(2019431)
hwI_training <- hwI_plasma %>% sample_n(240)
hwI_test <- anti_join(hwI_plasma, hwI_training,
                      by = "subj_ID")
```

Question 2 (15 points)

Use the `hwI_training` data frame to plot the distribution of the outcome of interest, which is `betaplasma`, and then plot the logarithm of `betaplasma`. Specify which of the two distributions better matches the desirable qualities of an outcome variable in a regression model. Whichever choice you make about the outcome – either `betaplasma` or `log(betaplasma)` – stick with it for the rest of this homework.

Answer 2

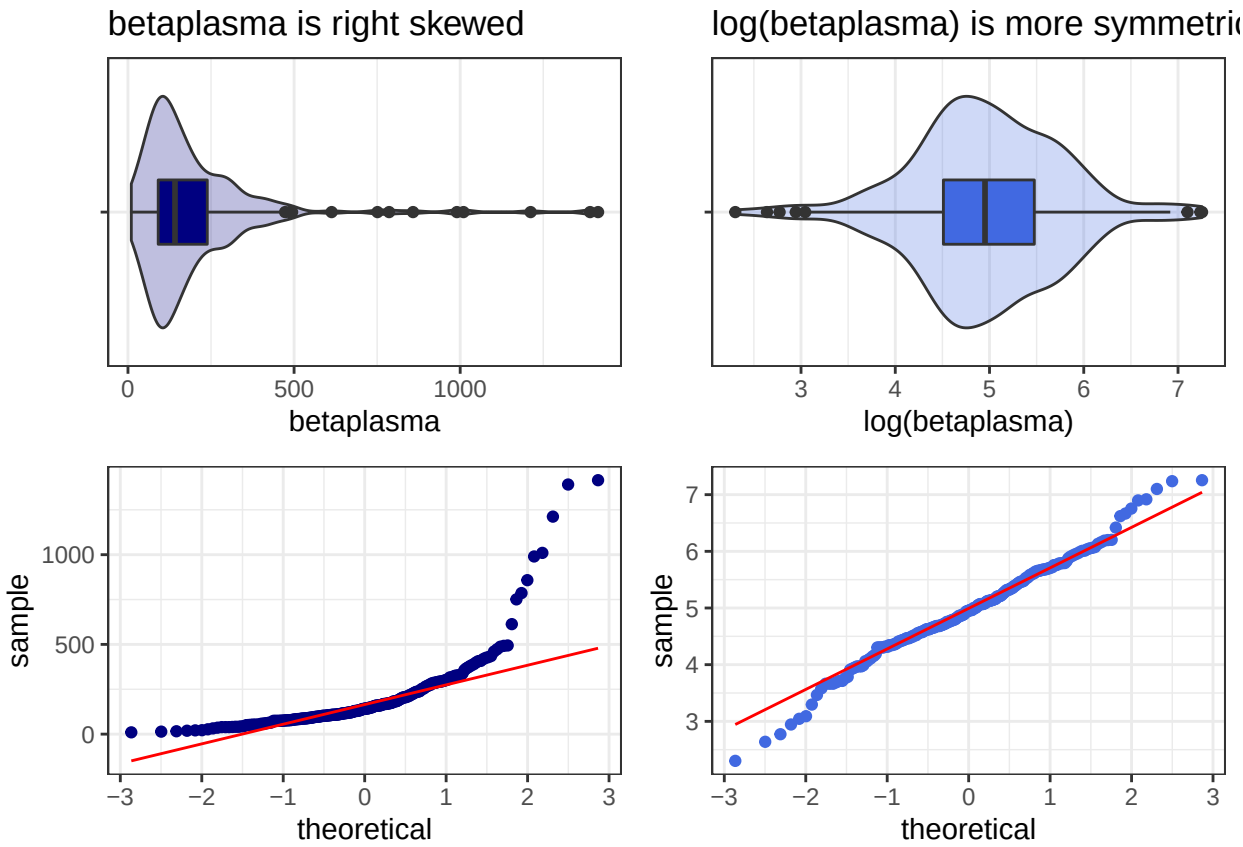
```
p1 <- ggplot(hwI_training, aes(x = "", y = betaplasma)) +
  geom_violin(fill = "navy", alpha = 0.25) +
  geom_boxplot(width = 0.25, fill = "navy") +
  coord_flip() +
  theme_bw() +
  labs(title = "betaplasma is right skewed", x = "")

p2 <- ggplot(hwI_training, aes(x = "", y = log(betaplasma))) +
  geom_violin(fill = "royalblue", alpha = 0.25) +
  geom_boxplot(width = 0.25, fill = "royalblue") +
  coord_flip() +
  theme_bw() +
  labs(title = "log(betaplasma) is more symmetric", x = "")

p3 <- ggplot(hwI_training, aes(sample = betaplasma)) +
  geom_qq(col = "navy") +
  geom_qq_line(col = "red") +
  theme_bw()

p4 <- ggplot(hwI_training, aes(sample = log(betaplasma))) +
  geom_qq(col = "royalblue") +
  geom_qq_line(col = "red") +
  theme_bw()

p1 + p2 + p3 + p4 + plot_layout(nrow = 2)
```



Taking the logarithm of `betaplasma` improves the fit of a Normal distribution to the data, and we will adopt that transformation of our outcome in the remainder of this work.

Question 3 (10 points)

Use the `hwI_training` data frame to build a model for your outcome (as decided in Question 2) using four predictors: `age`, `sex`, `bmi`, and `fiber`. Call that model `model_04`.

Summarize `model_04` and write a sentence or two to evaluate it. Be sure you describe the model's R^2 value. Also, be sure to interpret the model's residual standard error, in context, specifying appropriate units of measurement.

Answer 3

```
model_04 <- lm(log(betaplasma) ~ age + sex + bmi + fiber,
               data = hwI_training)

summary(model_04)
```

```
Call:
lm(formula = log(betaplasma) ~ age + sex + bmi + fiber, data = hwI_training)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-2.48014 -0.36849 -0.04136  0.41227  2.03317

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.982622   0.292037  17.062  < 2e-16 ***
age          0.011576   0.003218   3.597 0.000392 ***
sexM        -0.433006   0.134824  -3.212 0.001504 **
bmi         -0.035402   0.007839  -4.516 9.96e-06 ***
fiber        0.030768   0.008543   3.602 0.000386 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7178 on 235 degrees of freedom
Multiple R-squared:  0.1929,    Adjusted R-squared:  0.1792
F-statistic: 14.04 on 4 and 235 DF,  p-value: 2.739e-10
```

Key points we're hoping you will make:

- `model_04` accounts for 19.3% of the vatiation in the log of `betaplasma`. That's not a great result, in most settings.
- The residual standard error of the model is about 0.72, and this implies that about 95% of the prediction errors (residuals) made by the model predicting log(`betaplasma`) within the data set should be between -1.44 and 1.44, and that virtually all residuals should be between -2.16 and +2.16. Since the overall range of the data on the log scale is about 3-7, that's not a very impressive performance.
- The model finds statistically detectable incremental effects of each of the four predictors (age, sex, bmi and fiber) using any of our usual α levels.

Question 4 (10 points)

For your `model_04`, what is the estimated effect of being female, rather than male, on your outcome, holding everything else (age, bmi and fiber) constant. Provide and interpret a 95% confidence interval for that effect on your outcome.

Answer 4

I prefer to do this with `tidy` from the `broom` package, although `confint(model_04)` would also work.

```
tidy(model_04, conf.int = TRUE) %>%
  knitr::kable(digits = 2)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	4.98	0.29	17.06	0	4.41	5.56
age	0.01	0.00	3.60	0	0.01	0.02
sexM	-0.43	0.13	-3.21	0	-0.70	-0.17
bmi	-0.04	0.01	-4.52	0	-0.05	-0.02
fiber	0.03	0.01	3.60	0	0.01	0.05

Our model finds the estimated effect of being Male, rather than being Female, explicitly, but since `sex` in this data set is a binary variable, we can just reverse the sign of our estimate from `sexM` to obtain the estimate for `sexF`. Another available option would be to adjust the order of the levels for the `sex` factor (using `fct_relevel`) so as to directly estimate `sexF` instead of `sexM`.

Our `model_04` estimates the effect of being Female, rather than Male, on log(`betaplasma`) as an increase of 0.43. The 95% confidence interval is (0.17, 0.70).

- So, if we have two subjects of the same age, bmi and fiber, but different sex, then the female subject is estimated to have a log(`betaplasma`) value that is 0.43 larger than the male, and our 95% confidence interval for this difference is (0.17, 0.70), indicating a statistically detectable positive effect at the 5% level.

Question 5 (15 points)

Now use the `hwI_training` data frame to build a new model for your outcome (as decided in Question 2) using the following 10 predictors: `age`, `sex`, `smoking`, `bmi`, `vitamin`, `calories`, `fat`, `fiber`, `alcohol`, and `cholesterol`. Call that model `model_10`.

Compare `model_10` to `model_04` in terms of **adjusted** R^2, and residual standard error. Which model performs better on these summaries, in the training sample?

Answer 5

```
model_10 <- lm(log(betaplasma) ~ age + sex + smoking + bmi +
                 vitamin + calories + fat + fiber + alcohol +
                 cholesterol,
               data = hwI_training)
```

```
temp1 <- glance(model_04) %>%
  mutate(modelname = "model_04") %>%
  select(modelname, adj.r.squared, sigma)

temp2 <- glance(model_10) %>%
  mutate(modelname = "model_10") %>%
  select(modelname, adj.r.squared, sigma)

bind_rows(temp1, temp2) %>% knitr::kable(digits = 3)
```

modelname	adj.r.squared	sigma
model_04	0.179	0.718
model_10	0.210	0.704

The model with 10 predictors has a larger adjusted R^2 and a smaller residual standard error. Each of these suggests that `model_10` fits the data more effectively within our training sample than does `model_04`.

Another way to say this is that regarding *in-sample* prediction accuracy, we choose `model_10` over `model_04`.

Question 6 (20 points)

Use the code provided in Section 14 of the Project Study B Example to calculate and then compare the prediction errors made by the two models (`model_10` and `model_04`) you have generated. You should:

- Calculate the prediction errors in each case, then combine the results from the two models, tweaking the code and descriptions found in Section 14 of the Project Study B Example.

  - **HINT**: If you chose to transform the outcome variable back in Question 2, then you will need to estimate the predictions here back on the original scale of `betaplasma`, rather than on the logarithmic scale. That involves making predictions on the log scale, and then back-transforming them with the `exp` function before calculating the residuals and eventually the summary statistics.
- Visualize the prediction errors in each model, using the code in Section 14.2 of the Project Study B Example.
- Form the table comparing the model predictions, using the code in Section 14.3 of the Project Study B Example. Compare the models in terms of MAPE, MSPE and maximum prediction error.

Based on your results, what conclusions do you draw about which model (`model_10` or `model_04`) is preferable? Is this the same conclusion you drew in Question 5?

Answer 6

For full credit, you should estimate the predictions on the original scale of `betaplasma`, rather than on the logarithmic scale. That involves making predictions on the log scale, and then back-transforming them with the `exp` function before calculating the residuals and then the summary statistics.

Calculate the prediction errors

```
test_mod_04 <- test_mod_04 <- augment(model_04, newdata = hwI_test) %>%
  mutate(modelname = "model_04",
         .predictedbetaplasma = exp(.fitted),
         .resid = betaplasma - .predictedbetaplasma)

test_04 <- test_mod_04 %>%
  select(subj_ID, modelname, betaplasma, .predictedbetaplasma, .resid)

head(test_04, 2)
```

```
# A tibble: 2 x 5
  subj_ID modelname betaplasma .predictedbetaplasma .resid
  <chr>   <chr>          <dbl>                <dbl>  <dbl>
1 S-1006  model_04          35                117.   -82.5
2 S-1016  model_04          51                 61.8  -10.8
```

```
test_mod_10 <- test_mod_10 <- augment(model_10, newdata = hwI_test) %>%
  mutate(modelname = "model_10",
         .predictedbetaplasma = exp(.fitted),
         .resid = betaplasma - .predictedbetaplasma)

test_10 <- test_mod_10 %>%
  select(subj_ID, modelname, betaplasma, .predictedbetaplasma, .resid)

head(test_10, 2)
```

```
# A tibble: 2 x 5
  subj_ID modelname betaplasma .predictedbetaplasma .resid
  <chr>   <chr>          <dbl>                <dbl>  <dbl>
1 S-1006  model_10          35                125.   -90.5
2 S-1016  model_10          51                 71.4  -20.4
```

```
test_comp <- union(test_04, test_10) %>%
  arrange(subj_ID, modelname)
```

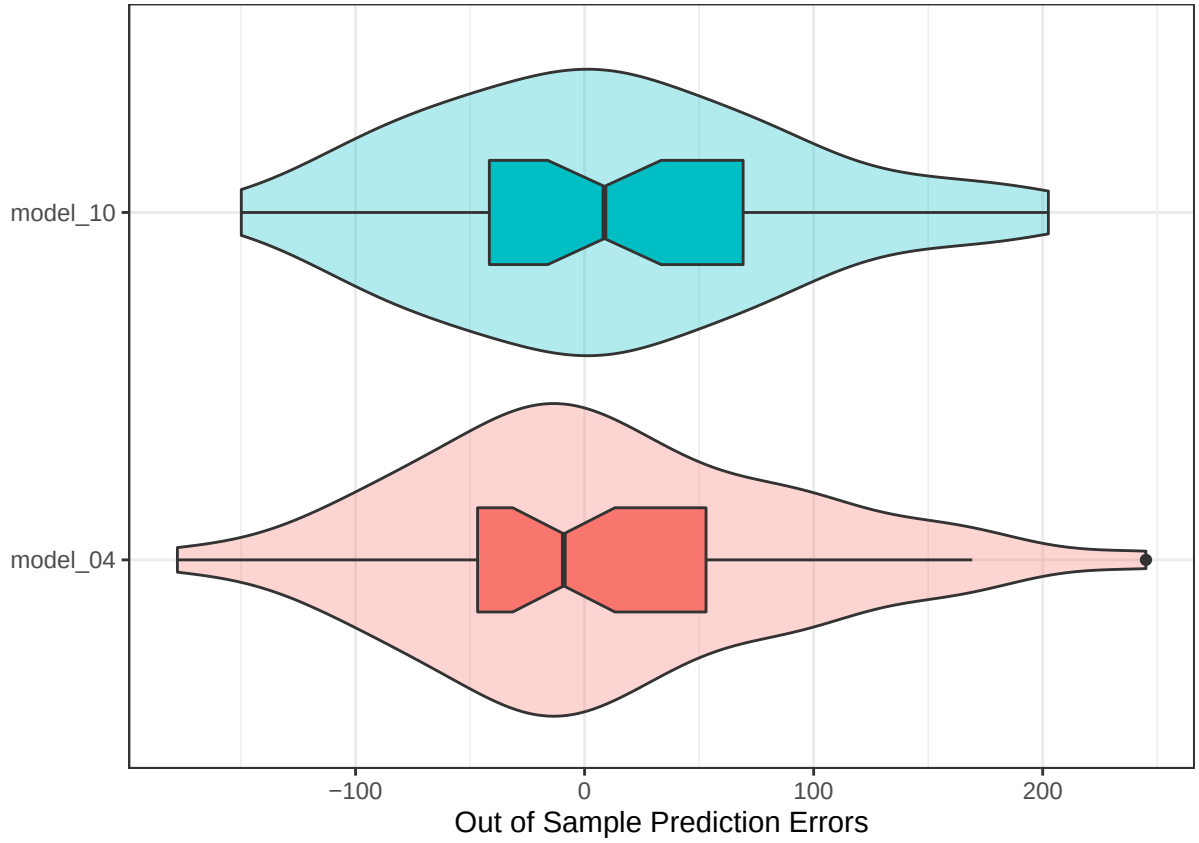
```
head(test_comp,4)
```

```
# A tibble: 4 x 5
  subj_ID modelname betaplasma .predictedbetaplasma .resid
  <chr>   <chr>          <dbl>                <dbl>  <dbl>
1 S-1006  model_04          35                117.   -82.5
2 S-1006  model_10          35                125.   -90.5
3 S-1016  model_04          51                 61.8  -10.8
4 S-1016  model_10          51                 71.4  -20.4
```

Visualize the prediction errors

Boxplots would work. I'll show a facetted set of histograms instead.

```
ggplot(test_comp, aes(x = modelname, y = .resid,
                      fill = modelname)) +
  geom_violin(alpha = 0.3) +
  geom_boxplot(width = 0.3, notch = TRUE) +
  guides(fill = FALSE) +
  coord_flip() +
  theme_bw() +
  labs(x = "", y = "Out of Sample Prediction Errors")
```



Form the table comparing predictions on MAPE, MSPE and max error

```
test_comp %>%
  group_by(modelname) %>%
  summarize(n = n(),
            MAPE = mean(abs(.resid)),
            MSPE = mean(.resid^2),
            max_error = max(abs(.resid)))
```

```
# A tibble: 2 x 5
  modelname     n  MAPE  MSPE max_error
  <chr>     <int> <dbl> <dbl>     <dbl>
1 model_04     50  65.5 7194.      245.
2 model_10     50  63.7 6501.      202.
```

Our conclusion from the table is that `model_10` shows better (i.e. smaller) results on MAPE, MSPE and maximum prediction error.

Another way to say this is that regarding *out-of-sample* prediction accuracy, we again choose `model_10` over `model_04`, as we did in response to Question 5.

Grading Rubric

Question 1 (30 points)

The specifications for the essay are:

1. Length is between 200 and 400 words.
2. English is correctly used throughout, with no typographical, grammatical or syntax errors.
3. A key idea is identified and clearly stated that actually appears in *The Signal and the Noise*.
4. An accurate and properly cited quote from the book is provided that is relevant to the identified key idea.
5. The context for the quote within Silver's book is described in the student's essay.
6. The essay clearly specifies how the idea in the book has changed their way of thinking about something which is explained in the essay.
7. The essay is clearly written, in general.
8. The essay is interesting to read.

- Award 29-30 points to all essays which meet all 8 of those specifications. I assume there will be 5-6 such essays, but if there are as many as ten, that's still OK.
- All other essays that answer the question in an appropriate way meeting 6-7 of the specifications should receive a score between 25 and 28.
- All essays which meet 4-5 of those specifications should receive a score between 20 and 24.
- Essays which meet less than 4 specifications should receive a score no higher than 19.

For students scoring 24 and below, please indicate in the comments which of the first 7 specifications they failed to meet.

Question 2 (15 points)

- 10 points for developing an appropriate, attractive set of plots of the relevant distributions, that are appropriately labeled. A single plot for each (`betaplasma` and `log(betaplasma)`) is all that is required.
- 5 additional points for coming to the correct conclusion regarding the logical choice of outcome transformation.

Question 3 (10 points)

- 3 points for specifying the R^2 value, and not, for instance, the adjusted R^2.
- 3 more for specifying and interpreting the residual standard error.
- 4 more for evaluating the model as accounting for a statistically significant, but not especially large amount of the variation in the outcome of interest (and getting that outcome right.)

Question 4 (10 points)

- 6 points for a correct estimate of the female effect.
- 4 more for a correct estimate of the 95% confidence interval, properly explained.
- If they in fact show the male effect size and CI without realizing it or if they misinterpret how to go from male to female, then that's a maximum of 7 points on the question.
- If they show the male effect size, and label it correctly as the male effect size, then that's a maximum of 9 points on the question.

Question 5 (15 points)

- 5 points for fitting the model correctly
- 5 more for specifying the adjusted R^2 and residual standard error correctly.
- 5 more for identifying the better model on these in-sample summaries.

Question 6 (20 points)

- 5 points for calculating prediction errors,
- with an additional 5 points for correctly doing the transformation.
- 3 points for an appropriate visualization of whatever errors they wound up with
- 3 points for an appropriate table summarizing whatever errors they wound up with
- 4 points for a correct conclusion in light of their table (whatever it says), and reference back to the decision in Question 5.