Answer Sketch for Homework 8

431 Staff and Professor Love

'Due 2018-11-30, version 2018-11-30

Contents

1 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.

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

library(broom); library(tidyverse)

hw8_plasma <- read_csv("hw8_plasma.csv") %>%
    mutate_if(is.character, funs(as.factor(.))) %>%
    mutate(subj_ID = as.character(subj_ID))
```

Dealing with the Errant 0 value for `betaplasma`

I eventually realized that there was one subject (S-1065) with an implausible `betaplasma` value of 0.

```
hw8_plasma %>% arrange(betaplasma) %>%
  select(subj_ID, betaplasma) %>%
  head(., 3)
```

```
# A tibble: 3 x 2
  subj_ID betaplasma
  <chr>        <int>
1 S-1065           0
2 S-1042          14
3 S-1192          16
```

I'll change that 0 value in `betaplasma` to 10, and then proceed.

```
hw8_plasma <- hw8_plasma %>%
  mutate(betaplasma = replace(betaplasma, betaplasma == 0, 10))

hw8_plasma %>% arrange(betaplasma) %>%
  select(subj_ID, betaplasma) %>%
  head(., 3)
```

```
# A tibble: 3 x 2
  subj_ID betaplasma
  <chr>        <dbl>
1 S-1065          10
2 S-1042          14
3 S-1192          16
```

Other options available to you were:

- to delete that observation (with something like `hw8_plasma <- hw8plasma %>% filter(betaplasma != 0))`)
- to add 1 to every `betaplasma` before taking the log

1.1 Partitioning into training/test samples

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

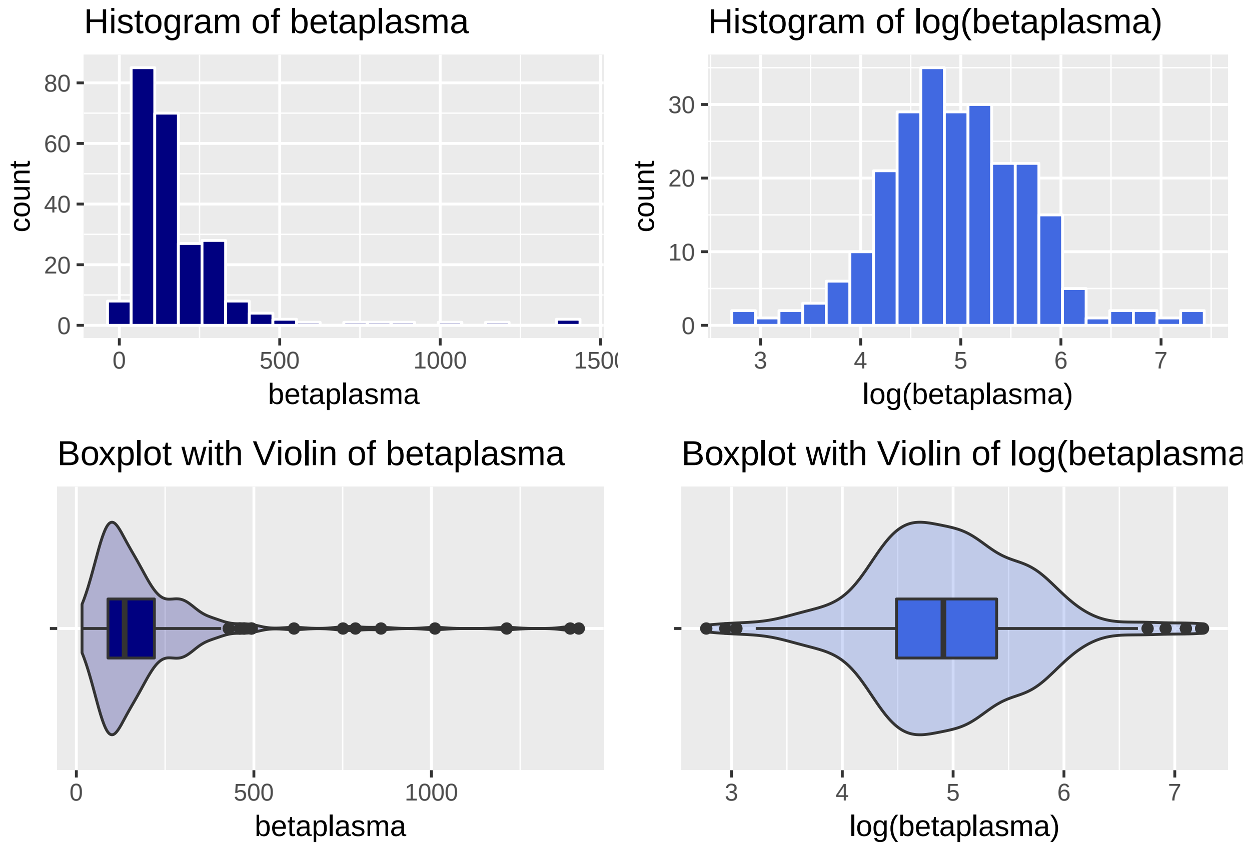
```
set.seed(431008)
hw8_training <- hw8_plasma %>% sample_n(240)
hw8_test <- anti_join(hw8_plasma, hw8_training,
                      by = "subj_ID")
```

2 Question 2 (15 points)

Use the `hw8_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 as to which outcome (`betaplasma` or `log(betaplasma)`), stick with it for the rest of this homework.

2.1 Answer 2

```r
p1 <- ggplot(hw8_training, aes(x = betaplasma)) +
  geom_histogram(bins = 20,
                 col = "white", fill = "navy") +
  labs(title = "Histogram of betaplasma")

p2 <- ggplot(hw8_training, aes(x = log(betaplasma))) +
  geom_histogram(bins = 20,
                 col = "white", fill = "royalblue") +
  labs(title = "Histogram of log(betaplasma)")

p3 <- ggplot(hw8_training, aes(x = "", y = betaplasma)) +
  geom_violin(fill = "navy", alpha = 0.25) +
  geom_boxplot(width = 0.25, fill = "navy") +
  coord_flip() +
  labs(title = "Boxplot with Violin of betaplasma",
       x = "")

p4 <- ggplot(hw8_training, aes(x = "", y = log(betaplasma))) +
  geom_violin(fill = "royalblue", alpha = 0.25) +
  geom_boxplot(width = 0.25, fill = "royalblue") +
  coord_flip() +
  labs(title = "Boxplot with Violin of log(betaplasma)",
       x = "")

gridExtra::grid.arrange(p1, p2, p3, p4, nrow = 2)
```



Clearly, 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.

3 Question 3 (10 points)

Use the `hw8_training` data frame to build a model for your outcome (as decided in Question 2) using the following 4 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.

3.1 Answer 3

```
model_04 <- lm(log(betaplasma) ~ age + sex + bmi + fiber,
               data = hw8_training)

summary(model_04)
```

```

Call:
lm(formula = log(betaplasma) ~ age + sex + bmi + fiber, data = hw8_training)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8343 -0.3337 -0.0705  0.4397  2.0942 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.775231   0.277362  17.217  < 2e-16 ***
age          0.011056   0.003023   3.657 0.000315 ***
sexM        -0.506428   0.136586  -3.708 0.000261 ***
bmi         -0.028348   0.007351  -3.856 0.000149 ***
fiber        0.033439   0.008001   4.179 4.13e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6618 on 235 degrees of freedom
Multiple R-squared:  0.1937,	Adjusted R-squared:   0.18 
F-statistic: 14.12 on 4 and 235 DF,  p-value: 2.442e-10
```

Key points we're hoping you will make:

- `model_04` accounts for 19.4% 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 0.66, 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 -1.32 and 1.32, and that virtually all residuals should be between -1.98 and +1.98. 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 significant incremental effects of each of the four predictors (age, sex, bmi and fiber.)

4 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.

4.1 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)	4.78	0.28	17.22	0	4.23	5.32
age	0.01	0.00	3.66	0	0.01	0.02
sexM	-0.51	0.14	-3.71	0	-0.78	-0.24
bmi	-0.03	0.01	-3.86	0	-0.04	-0.01
fiber	0.03	0.01	4.18	0	0.02	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 0.51. The 95% confidence interval is (0.24, 0.78).

- 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 0.51 larger than the male, and our 95% confidence interval for this difference is (0.24, 0.78) points, thus indicating a statistically significant effect at the 5% level.

5 Question 5 (15 points)

Now use the hw8_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?

5.1 Answer 5

```
model_10 <- lm(log(betaplasma) ~ age + sex + smoking + bmi +
                 vitamin + calories + fat + fiber + alcohol +
                 cholesterol,
               data = hw8_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	0.180	0.662
model_10	0.202	0.653

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`.

6 Question 6 (20 points)

Use the code provided in the Project Study 2 Demonstration (section 14) 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, following section 14.1 of the Project Study 2 Demonstration.
  - **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 of the Demo Project.
- Form the table comparing the model predictions, using the code in section 14.3. 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?

6.1 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.

6.1.1 Calculate the prediction errors

```
test_mod_04 <- test_mod_04 <- augment(model_04, newdata = hw8_test) %>%
  mutate(modelname = "model_04",
         .predictedbetaplasma = exp(.fitted),
         .resid = betaplasma - .predictedbetaplasma)
```

```r
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          67                 91.4  -24.4
2 S-1012  model_04          41                182.  -141.
```

```r
test_mod_10 <- test_mod_10 <- augment(model_10, newdata = hw8_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          67                 73.8   -6.83
2 S-1012  model_10          41                154.  -113.
```

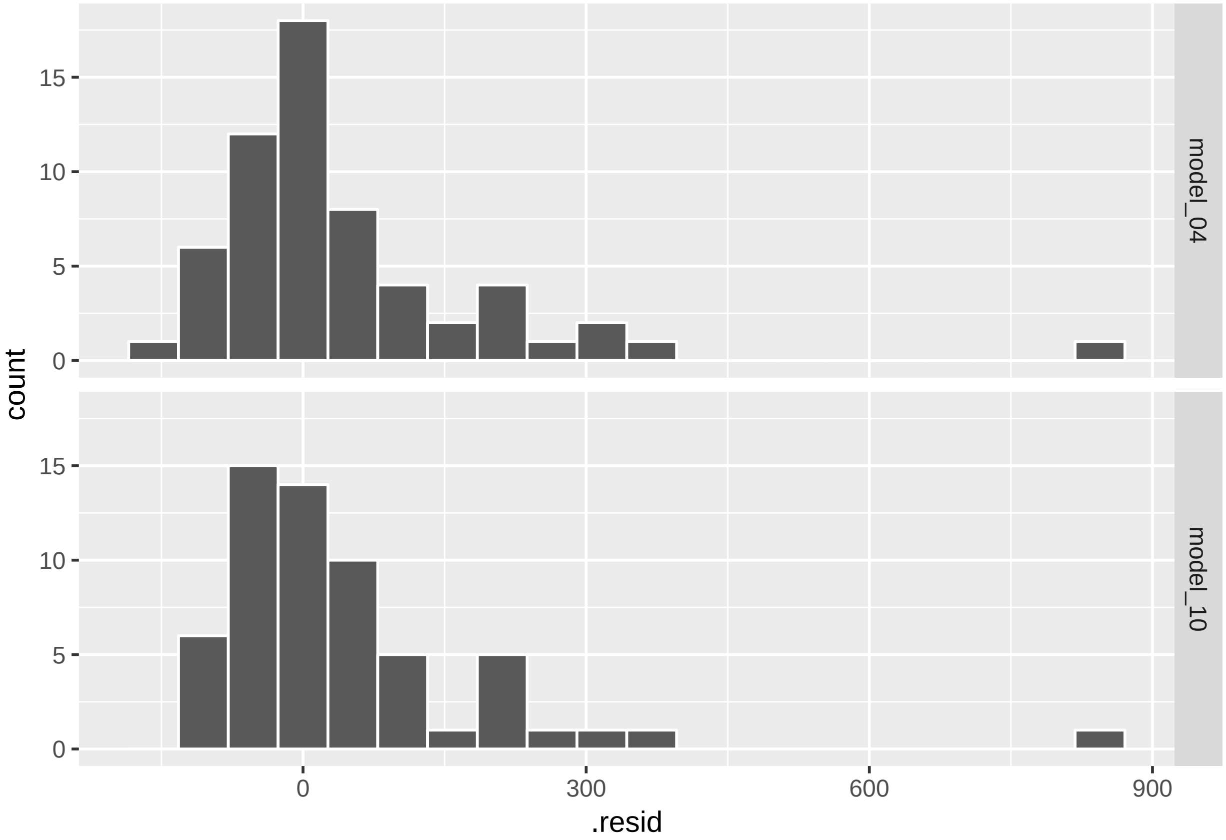
```r
test_comp <- union(test_04, test_10) %>%
  arrange(subj_ID, modelname)

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          67                 91.4  -24.4
2 S-1006  model_10          67                 73.8   -6.83
3 S-1012  model_04          41                182.  -141.
4 S-1012  model_10          41                154.  -113.
```

6.1.2 Visualize the prediction errors

We've used boxplots in class (for instance, Class 24), so I'll show a facetted set of histograms instead, here.

```r
ggplot(test_comp, aes(x = .resid)) +
  geom_histogram(bins = 20, col = "white") +
  facet_grid (modelname ~ .)
```



6.1.3 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     60  94.9 25842.      861.
2 model_10     60  91.3 24459.      852.
```

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.