# Homework 06 Solutions

## 1 Problem 1

```
library(tidyverse)
library(kableExtra)
maskdata <- read.csv("../maskdata.csv")
```

### 1.1 Part a

The first two code sections produce summary information and EDA as encouraged in the assignment.

```
library(knitr)
table_data <- maskdata %>%
  mutate(study = factor(study),
         allo.wm = factor(study),
         mask.class = factor(mask.class),
         virus = factor(virus),
         breath.num = factor(breath.num),
         female = factor(female,
                         labels = c("male", "female")),
         `log10(nasal)` = log10(nasal),
         `log10(throat)` = log10(throat),
         `log10(aerosols)` = log10(aerosols),
         `log10(droplets)` = log10(droplets)
         ) %>%
  select(-c(nasal, throat, aerosols, droplets))

options(knitr.kable.NA = '')
table_data %>%
  select(female, study, allo.wm, mask.class, virus, breath.num,
         cough.num, `log10(nasal)`, `log10(throat)`, `log10(aerosols)`,
         `log10(droplets)`) %>%
  summary() %>%
  knitr::kable(format = "latex",
               caption = "Summary of variables in mask experiment.") %>%
  kableExtra::row_spec(row = 0, bold = T) %>%
  kableExtra::kable_styling(latex_options =c("scale_down"))
```

We observe NA values in each of the viral load variables, the homework and project document mention that all these variables (recorded in copies of the virus per ml) have a value of 2 if the viral load was measured as 0, and that the NAs are associated with measurements not being taken or being lost. Table 2 shows that most of these variables see very few NA values recorded. There also doesn't seem to be much relationship between when the NAs occur in these variables.

```
na_table <- maskdata %>%
  dplyr::select(nasal, throat, aerosols, droplets) %>%
```

Table 1: Summary of variables in mask experiment.

| | female | study | allo.wm | mask.class | virus | breath.num | cough.num | log10(nasal) | log10(throat) | log10(aerosols) | log10(droplets) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | male :49 | cold:81 | cold:81 | both:26 | cold:54 | 1:88 | Min. : 0.00 | Min. : 2.389 | Min. :1.531 | Min. :0.301 | Min. :0.3010 |
| | female:65 | flu :33 | flu :33 | ctrl:43 | flu :43 | 2:26 | 1st Qu.: 0.00 | 1st Qu.: 5.532 | 1st Qu.:2.878 | 1st Qu.:0.301 | 1st Qu.:0.3010 |
| | | | | mask:45 | hcov:17 | | Median : 3.00 | Median : 6.514 | Median :3.854 | Median :0.301 | Median :0.3010 |
| | | | | | | | Mean : 9.57 | Mean : 6.412 | Mean :4.150 | Mean :1.150 | Mean :0.6758 |
| | | | | | | | 3rd Qu.:11.00 | 3rd Qu.: 7.345 | 3rd Qu.:5.359 | 3rd Qu.:2.245 | 3rd Qu.:0.3010 |
| | | | | | | | Max. :99.00 | Max. :10.097 | Max. :8.059 | Max. :5.276 | Max. :3.5185 |
| | | | | | | | | NA's :6 | NA's :32 | NA's :4 | NA's :8 |

Table 2: Number of NAs in each viral load column.

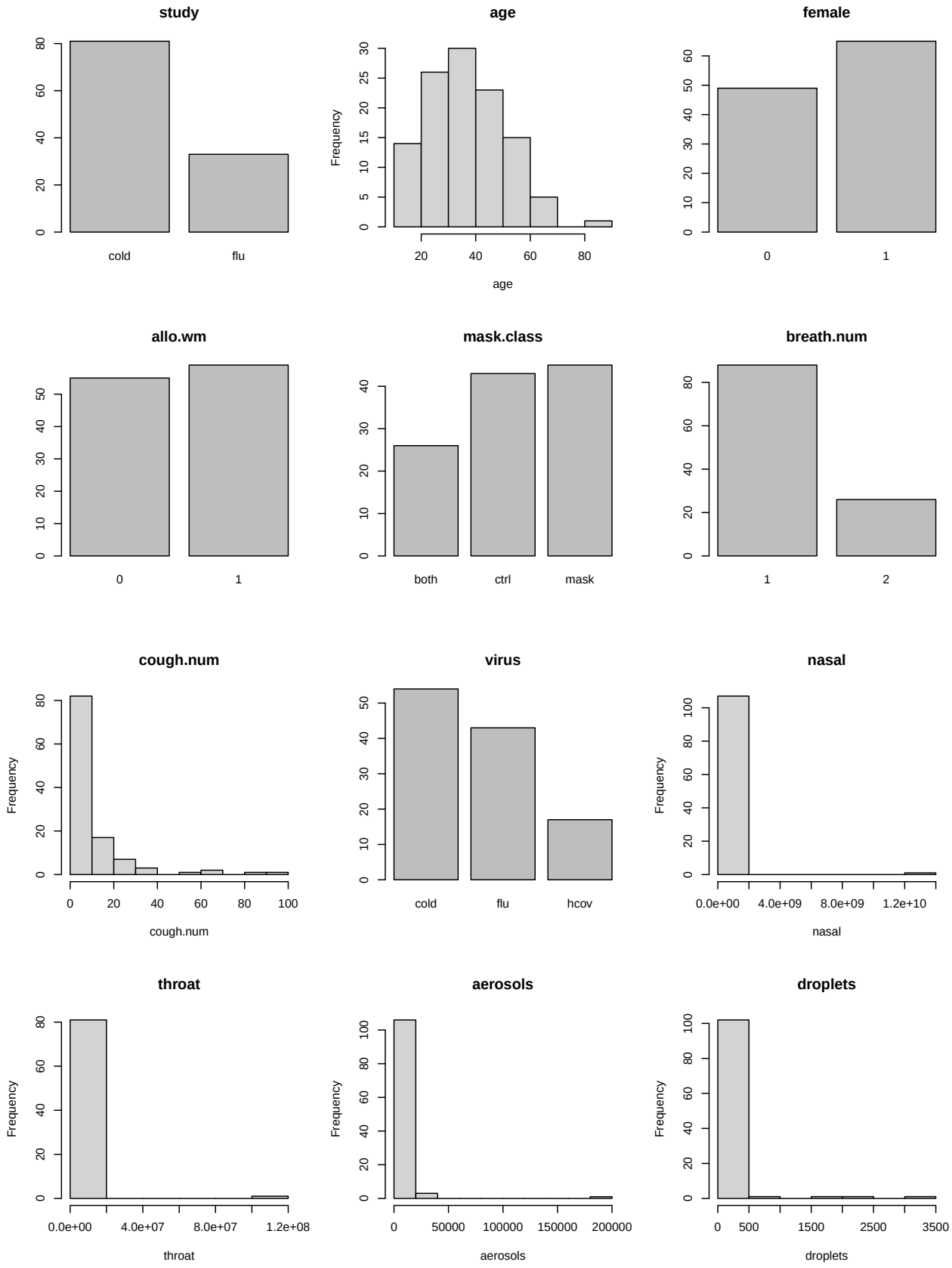
| | nasal | throat | aerosols | droplets | rows with any NA |
|---|---|---|---|---|---|
| not NA | 108 | 82 | 110 | 106 | 75 |
| is NA | 6 | 32 | 4 | 8 | 39 |

```
  sapply(function(col) table(is.na(col)))

any_na <- sum(apply(maskdata, 1, function(row) any(is.na(row))))
na_table <- cbind(na_table, `rows with any NA` = c(nrow(maskdata) - any_na, any_na))
rownames(na_table) <- c("not NA", "is NA")

na_table %>%
  knitr::kable(caption = "Number of NAs in each viral load column.") %>%
  kableExtra::kable_styling()
```

```
par(mfrow = c(4,3))
for (c_name in colnames(maskdata)){
  if(length(unique(maskdata[,c_name])) <= 3){
     c_vals <- names(table(maskdata[,c_name]))
     c_counts <- table(maskdata[,c_name])
     barplot(names.arg = c_vals,height= c_counts, main = c_name)
  }else {
    hist(maskdata[,c_name], main = c_name, xlab = c_name)
  }
}
```




It's highly encouraged to examine the data more deeply and understand how the variables relate. In the

Table 3: Split of studies (year) and types of virus

| allo.wm | both | ctrl | mask |
|---|---|---|---|
| **0** | 13 | 42 | 0 |
| **1** | 13 | 1 | 45 |

Table 4: Allocated to wear a mask vs mask class recorded

| study | cold | flu | hcov |
|---|---|---|---|
| **cold** | 36 | 17 | 12 |
| **flu** | 4 | 18 | 1 |

following EDA, we examine the interaction between variables, this helps us prepare for answering the questions of interesting in the later subsections.

The relationship between `allo.wm` and `mask.class`, as presented in Table 3, is very interesting, and digging into the data a bit more, it appears that even though individuals were allocated to a "wear mask" or "don't wear mask" class, but also had the opportunity to breath both with a mask and without a mask (in 2 seperate 30 minute breathing sessions). It's hard to rectify how an individual with `allo.wm = 1` (which means they were assigned to wear the mask), is listed as `mask.class = "ctrl"`. It's also unclear what the viral load `aerosol` and `droplets` data actually contains for individuals with `mask.class = "both"`. As such, we're going to use the `mask.class` feature to split of the groups (and ignore individuals that gave samples for both types as we are unclear what the data associated with that individual means). We also explored potentials for individuals with `mask.class = "both"` to have 2 rows (for both options), but that is not the case.

```
# table(maskdata$allo.wm, maskdata$mask.class)
maskdata %>%
  group_by(allo.wm, mask.class) %>%
  summarize(total = n()) %>%
  pivot_wider(names_from = mask.class, values_from = total, values_fill = 0) %>%
  knitr::kable(caption = "Split of studies (year) and types of virus") %>%
  kable_styling() %>%
  column_spec(1,bold = T, border_right = T)
```

Table 4 helps us see that that the study year and the type of virus isn't independent, we that only 1 individual in our data set in the `flu` study year had coronavirus.

```
# table(maskdata$study, maskdata$virus)
maskdata %>%
  filter(mask.class != "both") %>%
  group_by(study, virus) %>%
  summarize(total = n()) %>%
  pivot_wider(names_from = virus, values_from = total, values_fill = 0) %>%
  knitr::kable(caption = "Allocated to wear a mask vs mask class recorded") %>%
  kable_styling() %>%
  column_spec(1,bold = T, border_right = T)
```

Finally, we observed that the variable `breath.num` equals 2 if `mask.class = "both"` and otherwise it equals 1. Given this observation, we won't be using this variable due to the above discussion of `mask.class` vs `allo.wm`.

## 1.2 Part b

```
aerosols_base_lm <- lm(log10(aerosols-1) ~ mask.class,
                       data = maskdata[maskdata$mask.class != "both",])
droplets_base_lm <- lm(log10(droplets-1) ~ mask.class,
                       data = maskdata[maskdata$mask.class != "both",])
```

### 1.2.1 aerosol model

```
summary(aerosols_base_lm)
```

```
##
## Call:
## lm(formula = log10(aerosols - 1) ~ mask.class, data = maskdata[maskdata$mask.class !=
##     "both", ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3540 -0.7275 -0.5311  0.8657  3.9225
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.3540     0.2218   6.104 3.28e-08 ***
## mask.classmask  -0.8229     0.3065  -2.685  0.00878 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.403 on 82 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.0808, Adjusted R-squared:  0.06959
## F-statistic: 7.208 on 1 and 82 DF,  p-value: 0.008781
```

The above model's coefficients suggest that wearing a mask reduces the percentage of viral load in aerosol released in 30 minutes of breathing by 82.29%. This statistically significant result. For both of the models in this section, we are basically doing a 1 way ANOVA (with a single binary covariate), as such the diagnostic plots won't make that much sense. We decided to use $log_{10}(\text{viral load} - 1)$ as the viral load variables are highly skewed, and the associated paper we are comparing against also examined these loads in the log scale[1].

### 1.2.2 droplets model

```
summary(droplets_base_lm)
```

```
##
## Call:
## lm(formula = log10(droplets - 1) ~ mask.class, data = maskdata[maskdata$mask.class !=
##     "both", ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6749 -0.6749 -0.2218 -0.2218  2.8435
```

[1]Given that the minimum value (when no viral load was actually recorded) was 2, we did a $log(x-1)$ transform so those values would be transformed to 0.

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.6749     0.1449   4.659 1.23e-05 ***
## mask.classmask  -0.4531     0.2036  -2.225   0.0289 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9276 on 81 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.05761,    Adjusted R-squared:  0.04597
## F-statistic: 4.951 on 1 and 81 DF,  p-value: 0.02885
```

The above model's coefficients suggest that wearing a mask reduces the percentage of viral load in droplets released in 30 minutes of breathing buy 45.31%. This statistically significant result.

## 1.3 Part c

In order to only make 2 models, but still be able to answer the question of the efficiency of masks conditional on virus type we build models of the type

$$\log_{10}(\text{viral load}) \sim \text{virus} + \text{mask.class} : \text{virus} - 1 \ .$$

This allows us to examine the efficiency of the masks per virus type using the interaction term.

```
aerosols_interact_lm <- lm(log10(aerosols - 1) ~ virus + mask.class:virus - 1,
                       data = maskdata[maskdata$mask.class != "both",])

droplets_interact_lm <- lm(log10(droplets - 1) ~ virus + mask.class:virus - 1,
                       data = maskdata[maskdata$mask.class != "both",])
```

### 1.3.1 aerosol model

```
summary(aerosols_interact_lm)
```

```
##
## Call:
## lm(formula = log10(aerosols - 1) ~ virus + mask.class:virus -
##     1, data = maskdata[maskdata$mask.class != "both", ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6559 -0.8063 -0.5214  0.7607  4.2576
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## viruscold                   1.5232     0.3242   4.698 1.11e-05 ***
## virusflu                    1.0188     0.3649   2.792  0.00658 **
## virushcov                   1.6559     0.5770   2.870  0.00528 **
## viruscold:mask.classmask   -0.7753     0.4649  -1.668  0.09936 .
## virusflu:mask.classmask    -0.4974     0.4881  -1.019  0.31133
## virushcov:mask.classmask   -1.6559     0.7863  -2.106  0.03843 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.413 on 78 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.3696, Adjusted R-squared:  0.3211
## F-statistic: 7.622 on 6 and 78 DF,  p-value: 1.868e-06
```

Interestingly, when we start focusing on the efficacy of masks conditional on the virus type, the evidence is a bit more murky. The above model's coefficients suggest that wearing a mask when you have a **cold** reduces the percentage of viral load in aerosol released in 30 minutes of breathing by 77.53% (this is **not** statistically significant result). Similarly when you have the **flu**, wearning a mask reduces the precentage of viral load in aerosol released in 30 minutes of breathing by 49.74% (this is also **not** statistically significant result). If you have a coronavirus, mask wearing is associated with a reduction of your viral load in aerosol released in 30 minutes by 165.59% (this **is** a statistically significant result).

If we interpret the question as *"Is the efficacy of masks different between the different virus types"*, then we'd be looking to see if $\beta_{\text{mask:cold}} = \beta_{\text{mask:flu}} = \beta_{\text{mask:coronavirus}}$. To do this type of test we'd actually want to look at a slightly different model, namely

$$\log_{10}(\text{viral load}) \sim \text{virus} + \text{mask.class} + \text{mask.class} : \text{virus} - 1 \ .$$

```
aerosols_interact_lm2 <- lm(log10(aerosols - 1) ~ virus + mask.class +
                              mask.class:virus - 1,
   data = maskdata[maskdata$mask.class != "both",])
```

We can then examine amount of variability explained by the interaction term to determine if the efficacy for masks is different across the virus types. In the following summary, we see that the associated p-value for this question is not significant (0.459), suggesting that we cannot conclude that efficacy of masks is different across virus strands.

```
anova(aerosols_interact_lm2)
```

```
## Analysis of Variance Table
##
## Response: log10(aerosols - 1)
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## virus              3  74.851 24.9503 12.4917 9.404e-07 ***
## mask.class         1  13.349 13.3489  6.6833    0.0116 *
## virus:mask.class   2   3.140  1.5698  0.7859    0.4593
## Residuals         78 155.794  1.9974
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

### 1.3.2 droplets model

```
summary(droplets_interact_lm)
```

```
##
## Call:
## lm(formula = log10(droplets - 1) ~ virus + mask.class:virus -
##     1, data = maskdata[maskdata$mask.class != "both", ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -0.7419 -0.5850  0.0000  0.0000  2.9334
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## viruscold                   0.7223     0.2072   3.485 0.000815 ***
## virusflu                    0.5850     0.2393   2.445 0.016789 *
## virushcov                   0.7419     0.3784   1.961 0.053529 .
## viruscold:mask.classmask   -0.1401     0.3109  -0.451 0.653467
## virusflu:mask.classmask    -0.5850     0.3201  -1.827 0.071510 .
## virushcov:mask.classmask   -0.7419     0.5156  -1.439 0.154269
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9268 on 77 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.2686, Adjusted R-squared:  0.2116
## F-statistic: 4.714 on 6 and 77 DF,  p-value: 0.0003893
```

A similar translation of this summary table can be made. We leave this to the reader.

```
droplets_interact_lm2 <- lm(log10(droplets - 1) ~ virus + mask.class +
                              mask.class:virus - 1,
   data = maskdata[maskdata$mask.class != "both",])
anova(droplets_interact_lm2)
```

```
## Analysis of Variance Table
##
## Response: log10(droplets - 1)
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## virus             3 19.472  6.4906  7.5563 0.0001701 ***
## mask.class        1  3.562  3.5620  4.1469 0.0451483 *
## virus:mask.class  2  1.259  0.6295  0.7329 0.4838491
## Residuals        77 66.140  0.8590
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## 1.4 Part d

As mentioned in the EDA in part a, we won't include # of breaths as it is only 2 when with `mask.class == "both"` (else it is 1), and we're not looking at the data where `mask.class == "both"`.

Our model focuses on inference (focusing on potential collaborators' interest in the best estimate of the impact of masks on reducing the viral load across these populations). We include `age` and `female` features to account for demographic covariates that might impact an individual's production viral load production (which interestingly seems to have significant coefficients). We include also include nasal and throat recorded viral load, as it is natural to assume that an individual's viral load produced by breathing is associated with the amount of viral load found in the nose and throat. And finally, we include `study` to account for any measurement / experiment differences across the studies.

```
aerosols_interact_lm <- lm(log10(aerosols) ~ virus + mask.class:virus + age +
                             female + log10(nasal) + log10(throat) + study - 1,
                       data = maskdata[maskdata$mask.class != "both",])

summary(aerosols_interact_lm)
```

```
##
```

```
## Call:
## lm(formula = log10(aerosols) ~ virus + mask.class:virus + age +
##     female + log10(nasal) + log10(throat) + study - 1, data = maskdata[maskdata$mask.class !=
##     "both", ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0577 -0.8504 -0.1688  0.6667  2.7485
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## viruscold                  -0.05285    0.99063  -0.053  0.95767
## virusflu                   -0.83462    1.11837  -0.746  0.45892
## virushcov                  -0.14371    1.34366  -0.107  0.91524
## age                         0.02424    0.01130   2.146  0.03666 *
## female                     -1.13966    0.33913  -3.361  0.00148 **
## log10(nasal)                0.16240    0.15278   1.063  0.29283
## log10(throat)               0.19730    0.13021   1.515  0.13589
## studyflu                   -0.07756    0.40246  -0.193  0.84794
## viruscold:mask.classmask -1.18717    0.51245  -2.317  0.02458 *
## virusflu:mask.classmask  -0.36244    0.50164  -0.723  0.47327
## virushcov:mask.classmask -2.11400    0.85098  -2.484  0.01631 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.249 on 51 degrees of freedom
##   (26 observations deleted due to missingness)
## Multiple R-squared:  0.6384, Adjusted R-squared:  0.5604
## F-statistic: 8.185 on 11 and 51 DF,  p-value: 5.121e-08
```

Under this model, it appears that we can reject hypothesis that masks don't reduce the amount of viral load for those with a cold or a flu (and the estimated impact is 118.72 and 211.4% respectively - aka pretty high.

Our data has a lot of left censured variables (basically all the viral load recordings), and the diagnostic plots in Figures 1 and 2 still highlight such things. We also examined added variable plots to understanding if additional patterns between variables and residuals where present but didn't find any. VIF analysis (when accounting for subspace structure in the data for the categorical variables) the generalized VIF values are not too worrisome related to the inflation of standard errors for the $\beta$ values. Additionally, given the small size of the data set, and the nature of the variables there wasn't too much imagined could be done / variable that could be removed in the situation.

```
plot(x = aerosols_interact_lm$fitted.values,
     y = (aerosols_interact_lm$residuals + aerosols_interact_lm$fitted.values),
     ylab = "log10(aerosol viral load)",
     xlab = "estimated log10(aerosol viral load)")
```

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

# 2 Problem 2

```
beauty <- read_csv("../ProfEvaltnsBeautyPublic.csv") %>%
  dplyr::select(-profevaluation, -profnumber, -multipleclass,
```

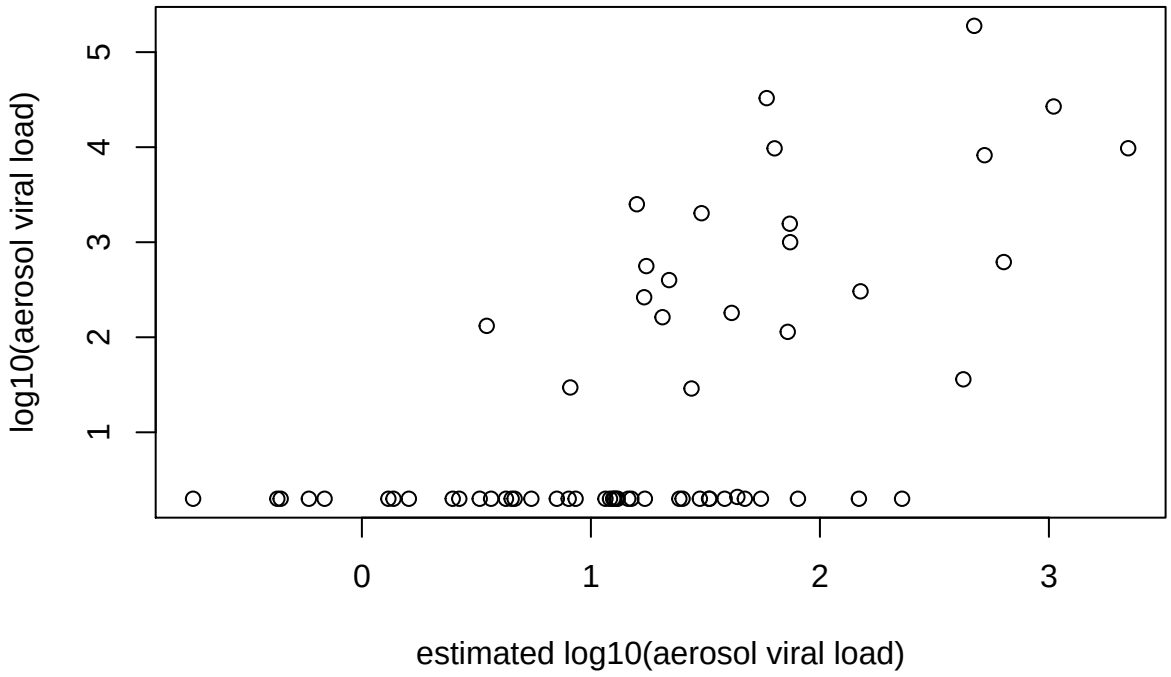



Figure 1: True vs Estimated log10(aerosol viral load).

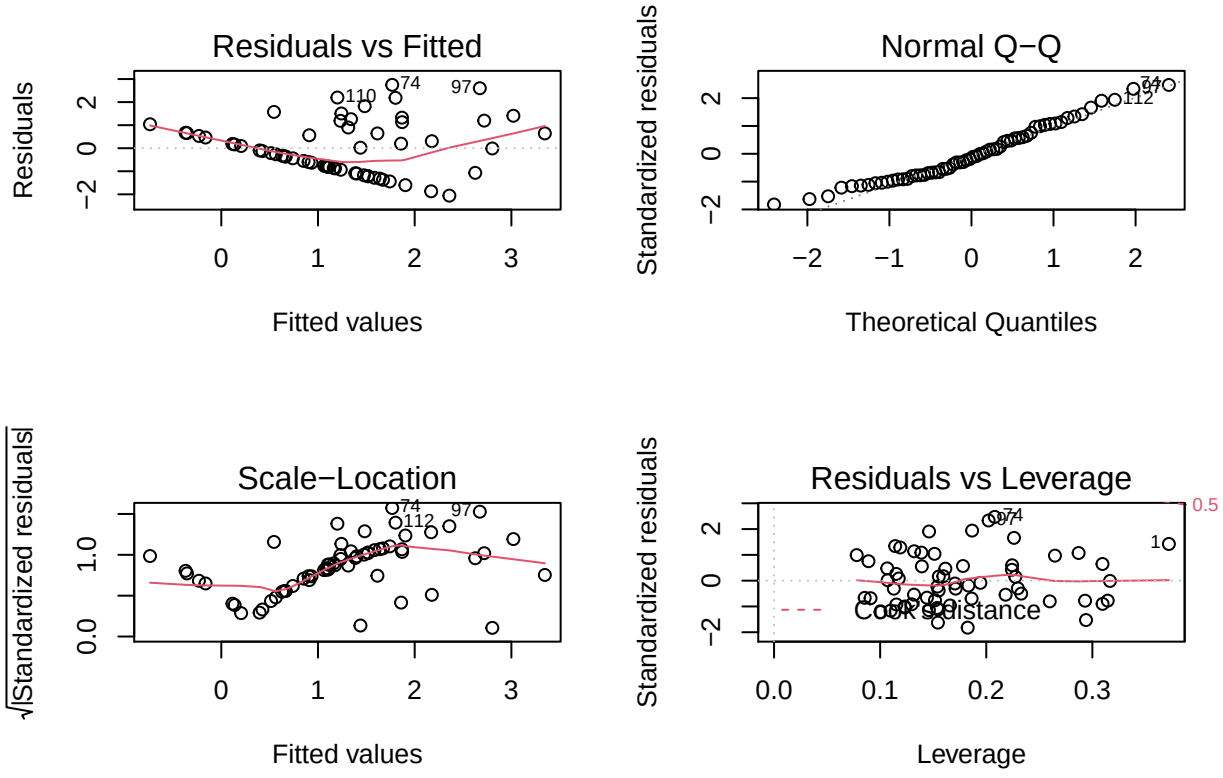



Figure 2: Diagnostics for final aerosol viral load model.

```
                -starts_with("class"))
```

## 2.1 Part a

```
library(leaps)
all_subsets <- regsubsets(courseevaluation ~.,
                          data = beauty,
                          nvmax = 10, intercept = T,
                          method = "exhaustive")

summ <- summary(all_subsets)

best_model_index <- which.min(summ$bic)

dropped <- summ$outmat %>% apply(2, function(col) {mean(col == " ") == 1})
all_subset_info <- cbind(summ$outmat[,!dropped], BIC = summ$bic)
rownames(all_subset_info) <- 1:nrow(all_subset_info)
all_subset_info %>% data.frame %>%
  rownames_to_column(var = "num variables") %>%
  knitr::kable() %>%
  kableExtra::kable_styling(latex_options = "scale_down") %>%
  row_spec(best_model_index, background = "gray")
```

| num variables | minority | beautyf2upper | beautyfupperdiv | btystdave | btystdf2u | btystdfu | female | nonenglish | onecredit | percentevaluating | blkandwhite | BIC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | * | | | -14.0820700135937 |
| 2 | | | | * | | | | | * | | | -30.0187608382119 |
| 3 | | | * | | | | * | | * | | | -42.0422726206511 |
| 4 | | | * | | | | * | | * | * | | -52.6352707419912 |
| 5 | * | | * | | | | * | | * | * | | -61.8604866501093 |
| 6 | | | * | | | | * | * | * | * | * | -68.7273809819797 |
| 7 | * | | * | | | | * | * | * | * | * | -68.168278525645 |
| 8 | | * | * | | * | | * | * | * | * | * | -67.592998887792 |
| 9 | * | * | * | | * | | * | * | * | * | * | -67.4303455606005 |
| 10 | * | * | * | | * | * | * | * | * | * | * | -64.9400133420853 |

```
var_names <- colnames(all_subset_info)[all_subset_info[best_model_index,] == "*"]

all_subsets_model <- lm(as.formula(paste("courseevaluation ~",
                                         paste0(var_names, collapse = " + "))),
                        data = beauty)

summary(all_subsets_model)
```

```
##
## Call:
## lm(formula = as.formula(paste("courseevaluation ~", paste0(var_names,
##     collapse = " + "))), data = beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.97607 -0.30533  0.04791  0.37959  1.04118
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.355273   0.113922  29.452  < 2e-16 ***
## beautyfupperdiv   0.058282   0.012559   4.641 4.54e-06 ***
```

```
## female            -0.254260   0.048290  -5.265 2.16e-07 ***
## nonenglish        -0.388186   0.097125  -3.997 7.49e-05 ***
## onecredit          0.500846   0.099146   5.052 6.35e-07 ***
## percentevaluating  0.005502   0.001415   3.889 0.000115 ***
## blkandwhite        0.251170   0.063629   3.947 9.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.495 on 456 degrees of freedom
## Multiple R-squared:  0.2143, Adjusted R-squared:  0.204
## F-statistic: 20.73 on 6 and 456 DF,  p-value: < 2.2e-16
```

## 2.2 Part b

We'll show forward regression.

```
forward_step <- step(lm(courseevaluation ~1, data = beauty),
  scope  = as.formula(paste("courseevaluation ~",
                      paste0(names(beauty)[names(beauty) != "courseevaluation"],
                             collapse = " + "))),
                      direction = "forward", trace = 0)

summary(forward_step)
```

```
##
## Call:
## lm(formula = courseevaluation ~ onecredit + btystdave + percentevaluating +
##     female + minority + blkandwhite + nonenglish + beautyfupperdiv +
##     btystdfu + btystdf2u + tenuretrack + formal + age, data = beauty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86588 -0.29589  0.04495  0.33963  1.06133
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.836838   1.639295  -0.510 0.609962
## onecredit          0.503749   0.107838   4.671 3.96e-06 ***
## btystdave         -0.167839   0.075372  -2.227 0.026456 *
## percentevaluating  0.004850   0.001451   3.342 0.000902 ***
## female            -0.302259   0.052308  -5.778 1.41e-08 ***
## minority          -0.155301   0.074788  -2.077 0.038412 *
## blkandwhite        0.285158   0.069381   4.110 4.70e-05 ***
## nonenglish        -0.381223   0.104535  -3.647 0.000297 ***
## beautyfupperdiv    0.939194   0.323894   2.900 0.003918 **
## btystdfu          -1.770340   0.656800  -2.695 0.007294 **
## btystdf2u          0.092159   0.042790   2.154 0.031790 *
## tenuretrack       -0.132773   0.062020  -2.141 0.032828 *
## formal             0.143150   0.069643   2.055 0.040410 *
## age               -0.004442   0.002837  -1.566 0.118145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4867 on 449 degrees of freedom
```

```
## Multiple R-squared:  0.2523, Adjusted R-squared:  0.2306
## F-statistic: 11.65 on 13 and 449 DF,  p-value: < 2.2e-16
```

### 2.3 Part c

```
library(glmnet)
cv_lasso_selection <- cv.glmnet(x = as.matrix(beauty[,names(beauty) != "courseevaluation"]),
          y = beauty$courseevaluation, alpha = 1)
```

```
lasso <- glmnet(x = as.matrix(beauty[,names(beauty) != "courseevaluation"]),
                y = beauty$courseevaluation,
                alpha = 1, lambda = cv_lasso_selection$lambda.1se)


coef_lasso <- as.matrix(coef(lasso))
coef_lasso_df <- data.frame(columns = names(coef_lasso[coef_lasso[,1] != 0,]),
         beta = coef_lasso[coef_lasso[,1] != 0,])
rownames(coef_lasso_df) <- NULL

coef_lasso_df
```

```
##              columns          beta
## 1        (Intercept)  3.558843e+00
## 2           minority -3.562730e-02
## 3      beautyf2upper  3.663638e-03
## 4    beautyfupperdiv  2.828790e-02
## 5          btystdf2u  8.129368e-06
## 6             female -1.061187e-01
## 7           fulldept  7.330144e-02
## 8         nonenglish -1.531938e-01
## 9          onecredit  3.355938e-01
## 10 percentevaluating  3.137356e-03
## 11       blkandwhite  1.104157e-01
```

### 2.4 Part d

```
#basically coding up a document term matrix function for some reason

dtm_coef <- function(names_list){
  all_words <- unique(unlist(names_list))
  matrix_info <- matrix(nrow = length(names_list), ncol = length(all_words))

  r_idx <- 1
  for (col_names in names_list){
    matrix_info[r_idx,] <- all_words %in% col_names
    r_idx <- r_idx + 1
  }

  colnames(matrix_info) <- all_words
  rownames(matrix_info) <- names(names_list)

  return(matrix_info)
```

```
}

variables_in_model <- list(
  "best subset" = names(coef(all_subsets_model)),
  "lasso" = coef_lasso_df$columns,
  "forward stepwise" = names(forward_step$coefficients),
  "homework 05, 2c" = c(names(beauty[,names(beauty) != "courseevaluation"]),
                        "profevaluation", "(Intercept)")
)
```

```
t(dtm_coef(variables_in_model)) %>%
  data.frame(check.names = F) %>%
  rownames_to_column(var = "features") %>%
  mutate(`best subset` = ifelse(`best subset`,
                                cell_spec(`best subset`, "latex", color = "black"),
                                cell_spec(`best subset`, "latex", color = "red")),
         `lasso` = ifelse(`lasso`,
                                cell_spec(`lasso`, "latex", color = "black"),
                                cell_spec(`lasso`, "latex", color = "red")),
         `forward stepwise` = ifelse(`forward stepwise`,
                                cell_spec(`forward stepwise`, "latex", color = "black"),
                                cell_spec(`forward stepwise`, "latex", color = "red")),
         `homework 05, 2c` = ifelse(`homework 05, 2c`,
                                cell_spec(`homework 05, 2c`, "latex", color = "black"),
                                cell_spec(`homework 05, 2c`, "latex", color = "red"))) %>%
  knitr::kable(format = "latex", caption = "Features include in each model.",
               escape = F) %>%
  kableExtra::kable_styling(latex_options = c("scale_down"),)
```

The number of features (excluding the intercept) in our models increase from 6 in the best subset model, lasso with 10, forward stepwise regression selecting 14 and our full model with 30 features in homework 05. If we take out the lasso model these models are nested inside each other. **It's hard to evaluate prediction models as we left nothing to evaluate them on.** We observe that the lasso had actually higher training mse than the best subset model - but that's not to crazy because the penalty means that we're get more training error allowed.

Table 5: Features include in each model.

| features | best subset | lasso | forward stepwise | homework 05, 2c |
|---|---|---|---|---|
| (Intercept) | TRUE | TRUE | TRUE | TRUE |
| beautyfupperdiv | TRUE | TRUE | TRUE | TRUE |
| female | TRUE | TRUE | TRUE | TRUE |
| nonenglish | TRUE | TRUE | TRUE | TRUE |
| onecredit | TRUE | TRUE | TRUE | TRUE |
| percentevaluating | TRUE | TRUE | TRUE | TRUE |
| blkandwhite | TRUE | TRUE | TRUE | TRUE |
| minority | FALSE | TRUE | TRUE | TRUE |
| beautyf2upper | FALSE | TRUE | FALSE | TRUE |
| btystdf2u | FALSE | TRUE | TRUE | TRUE |
| fulldept | FALSE | TRUE | FALSE | TRUE |
| btystdave | FALSE | FALSE | TRUE | TRUE |
| btystdfu | FALSE | FALSE | TRUE | TRUE |
| tenuretrack | FALSE | FALSE | TRUE | TRUE |
| formal | FALSE | FALSE | TRUE | TRUE |
| age | FALSE | FALSE | TRUE | TRUE |
| tenured | FALSE | FALSE | FALSE | TRUE |
| beautyflowerdiv | FALSE | FALSE | FALSE | TRUE |
| beautym2upper | FALSE | FALSE | FALSE | TRUE |
| beautymlowerdiv | FALSE | FALSE | FALSE | TRUE |
| beautymupperdiv | FALSE | FALSE | FALSE | TRUE |
| btystdfl | FALSE | FALSE | FALSE | TRUE |
| btystdm2u | FALSE | FALSE | FALSE | TRUE |
| btystdml | FALSE | FALSE | FALSE | TRUE |
| btystdmu | FALSE | FALSE | FALSE | TRUE |
| didevaluation | FALSE | FALSE | FALSE | TRUE |
| lower | FALSE | FALSE | FALSE | TRUE |
| students | FALSE | FALSE | FALSE | TRUE |
| btystdvariance | FALSE | FALSE | FALSE | TRUE |
| btystdavepos | FALSE | FALSE | FALSE | TRUE |
| btystdaveneg | FALSE | FALSE | FALSE | TRUE |
| profevaluation | FALSE | FALSE | FALSE | TRUE |