Lab 6: Regression

Robert Kubinec, Alison Yao

11/4/2021

Introduction

Please complete the prompts and code chunks in this Rmarkdown file for Lab 6. Please ensure the file knits and produces a report that appears correct before submitting via a Git push to the repo created for you.

The data for this lab comes from a research group studying some of the factors that predicted COVID-19 mortality. It can be downloaded directly from a website using the `read_csv` command from the `readr` package (this function can also take a filename from a website):

```
covid_data <- read_csv("https://wzb-ipi.github.io/corona/df_full.csv")
```

```
## New names:
## * `` -> ...1
```

```
## Rows: 17120 Columns: 141
```

```
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr    (7): geoid2, country, continent, region, scode, weeknumber, forcats::...
## dbl  (129): ...1, month, day, year, elapsed, population_2019, cases, deaths,...
## lgl    (3): metro_area, iso_3166_2_code, census_fips_code
## date   (2): date, date_rep
```

```
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(covid_data)
```

```
## Rows: 17,120
## Columns: 141
## $ ...1                                <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11…
## $ geoid2                              <chr> "ABW", "ABW", "ABW", "ABW", "ABW"…
## $ date                                <date> 2019-12-30, 2020-01-06, 2020-01-…
## $ month                               <dbl> 12, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3,…
## $ day                                 <dbl> 30, 6, 13, 20, 27, 3, 10, 17, 24,…
## $ year                                <dbl> 2019, 2020, 2020, 2020, 2020, 202…
## $ elapsed                             <dbl> -1, 6, 13, 20, 27, 34, 41, 48, 55…
## $ country                             <chr> "Aruba", "Aruba", "Aruba", "Aruba…
## $ continent                           <chr> "America", "America", "America", …
## $ population_2019                     <dbl> NA, 106766, 106766, 106766, 10676…
## $ cases                               <dbl> 0, 50, 14, 28, 5, 3, 0, 1, 0, 0, …
## $ deaths                              <dbl> 0, 0, 0, 0, 2, 0, 0, 1, 0, 0, 0, …
## $ date_rep                            <date> 2019-12-30, 2020-01-06, 2020-01-…
## $ cases_cum                           <dbl> 0, 50, 64, 92, 97, 100, 100, 101,…
## $ deaths_cum                          <dbl> 0, 0, 0, 0, 2, 2, 2, 3, 3, 3, 3, …
## $ deaths_cum_log                      <dbl> 0.000000, 0.000000, 0.000000, 0.0…
## $ deaths_cum_l7                       <dbl> NA, 0, 0, 0, 0, 2, 2, 2, 3, 3, 3,…
## $ deaths_cum_g7                       <dbl> NA, NA, NA, NA, NA, NA, 0.0000000…
## $ region                              <chr> "Latin America & Caribbean", "Lat…
## $ gov_effect                          <dbl> 1.058392, 1.058392, 1.058392, 1.0…
## $ trade                               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ ineq                                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ gdp_pc                              <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ pop_tot                             <dbl> 0.105845, 0.105845, 0.105845, 0.1…
## $ older_m                             <dbl> 6093, 6093, 6093, 6093, 6093, 609…
## $ older_f                             <dbl> 8252, 8252, 8252, 8252, 8252, 825…
## $ air_travel                          <dbl> 12.52191, 12.52191, 12.52191, 12.…
## $ fdi                                 <dbl> 135529099, 135529099, 135529099, …
## $ pop_density                         <dbl> 588.0278, 588.0278, 588.0278, 588…
## $ urban                               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ pop_below14_2018                    <dbl> 17.80245, 17.80245, 17.80245, 17.…
## $ migration_share                     <dbl> 34.7621, 34.7621, 34.7621, 34.762…
## $ oil                                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ life_exp_2017                       <dbl> 76.01, 76.01, 76.01, 76.01, 76.01…
## $ soc_insur_cov                       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ soc_contrib                         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ soc_safety                          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ polity                              <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ gini                                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ elf_epr                             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ rq_polarization                     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ count_powerless                     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ share_powerless                     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ media_critical                      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ journal_harass                      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ health_equality                     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ property_rights                     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ transparent_law                     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ bureaucracy_corrupt                 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ pos_gov_lr                          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ polar_rile                          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ trust_people                        <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ trust_gov                           <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ electoral_pop                       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ federal_ind                         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
```

```
## $ woman_leader                <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ checks_veto                 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ polariz_veto                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ dist_senate                 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ dist_presid                 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ dist_parlm                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ dist_anyelec                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ elect_pressure              <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ infections_mers             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ infections_sars             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ infections_ebola            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ infection                   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
## $ vdem_libdem                 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ al_etfra                    <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ al_religfra                 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ fe_etfra                    <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ vdem_mecorrpt               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ share_health_ins            <dbl> 99.2, 99.2, 99.2, 99.2, 99.2, 99.…
## $ pandemic_prep               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ resp_disease_prev           <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ detect_index                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ doctors_pc                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ hosp_beds_pc                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ literacy_rate               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ healthcare_qual             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ acc_sanitation              <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ health_exp_pc               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ hdi                         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ health_index                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ respond_index               <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ scode                       <chr> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ state_fragility             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ effect                      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ legit                       <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ seceff                      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ secleg                      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ poleff                      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ polleg                      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ ecoeff                      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ ecoleg                      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ soceff                      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ socleg                      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ pr                          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ share_older                 <dbl> 13.55284, 13.55284, 13.55284, 13.…
## $ pop_tot_log                 <dbl> -2.24578, -2.24578, -2.24578, -2.…
## $ pop_density_log             <dbl> 6.376774, 6.376774, 6.376774, 6.3…
## $ distancing_bin              <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ lockdown_bin                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ lockdown_n                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ distancing_n                <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ metro_area                  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ iso_3166_2_code             <lgl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ census_fips_code            <lgl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ retail                      <dbl> NA, NA, NA, NA, NA, NA, NA, -1, -…
## $ grocery                     <dbl> NA, NA, NA, NA, NA, NA, NA, -1, -…
## $ parks                       <dbl> NA, NA, NA, NA, NA, NA, NA, 16, 2…
## $ transit                     <dbl> NA, NA, NA, NA, NA, NA, NA, -4, -…
## $ work                        <dbl> NA, NA, NA, NA, NA, NA, NA, -20, …
```

```
## $ residential                            <dbl> NA, NA, NA, NA, NA, NA, NA, 6, 18…
## $ mobility_index                         <dbl> NA, NA, NA, NA, NA, NA, NA, -2.0,…
## $ stringency                             <dbl> NA, 0.00, 0.00, 0.00, 0.00, 0.00,…
## $ C1_School.closing                      <dbl> NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
## $ C2_Workplace.closing                   <dbl> NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
## $ C3_Cancel.public.events                <dbl> NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
## $ C4_Restrictions.on.gatherings          <dbl> NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
## $ C5_Close.public.transport              <dbl> NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
## $ C6_Stay.at.home.requirements           <dbl> NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
## $ C7_Restrictions.on.internal.movement   <dbl> NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
## $ C8_International.travel.controls       <dbl> NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
## $ H1_Public.information.campaigns        <dbl> NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
## $ temp_cumul                             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ precip_cumul                           <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ weeknumber                             <chr> "52_2019", "1_2020", "2_2020", "3…
## $ excess_deaths_weekly                   <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ excess_deaths_last_obs                 <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ excess_deaths_cum                      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ excess_deaths_cum_log                  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ deaths_cum_per_million                 <dbl> 0.00000, 0.00000, 0.00000, 0.0000…
## $ deaths_cum_per_million_log             <dbl> 0.000000, 0.000000, 0.000000, 0.0…
## $ excess_deaths_cum_per_million          <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ excess_deaths_cum_per_million_log      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, N…
## $ `forcats::fct_explicit_na(geoid2)`     <chr> "ABW", "ABW", "ABW", "ABW", "ABW"…
## $ relative_start                         <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, …
## $ elapsed_rel                            <dbl> -7, 0, 7, 14, 21, 28, 35, 42, 49,…
## $ relative_start_d                       <dbl> 160, 160, 160, 160, 160, 160, 160…
## $ elapsed_rel_d                          <dbl> -161, -154, -147, -140, -133, -12…
```

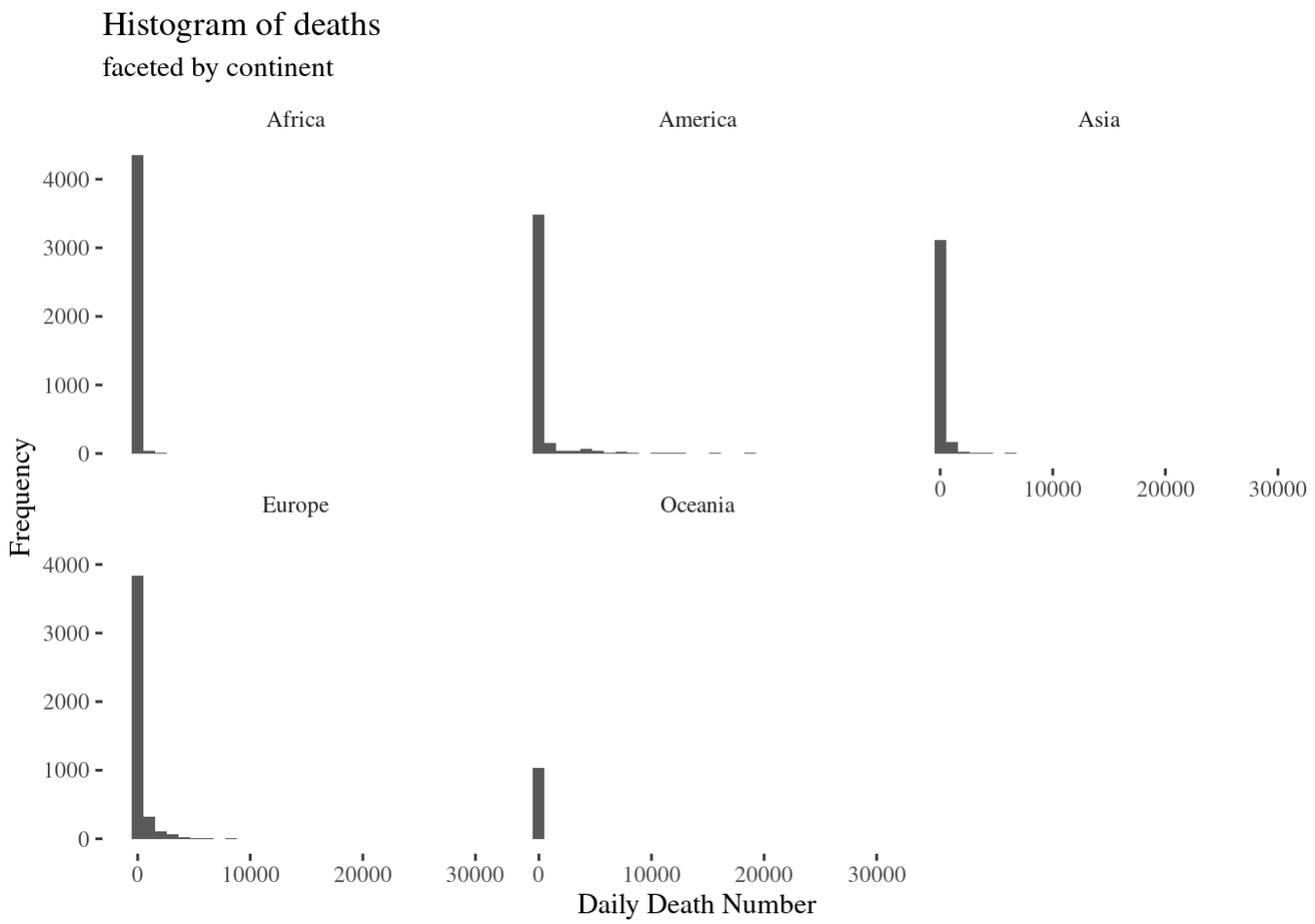
As can be seen, this dataset has a lot of different columns available. For our purposes we will focus on the column `deaths`, which records the daily reported COVID-19 deaths by country.

Histogram of COVID-19 Deaths

Please make a histogram (`geom_histogram`) with `ggplot2` for the `deaths` column. Also, use the `facet_wrap` function to facet the plot by the `continent` column. (Hint: if you forgot how these functions work, use the help command `?` plus the command name to read examples). To make your plot look pretty, add the `+ theme_tufte()` command at the end of the ggplot function call, and be sure to add appropriate legends/labels.

```
covid_data %>%
ggplot(aes(x = deaths)) +
  facet_wrap(. ~ continent) +
  geom_histogram() +
  theme_tufte() +
  labs(
    title = 'Histogram of deaths',
    subtitle = 'faceted by continent',
    x = 'Daily Death Number',
    y = 'Frequency'
  )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



You might notice that the number of deaths are spread out. You can change the x axis to use multiples of 10 by adding the `+ scale_x_log10()` to the ggplot2 call. Do that in the chunk below to compare with the plot above:

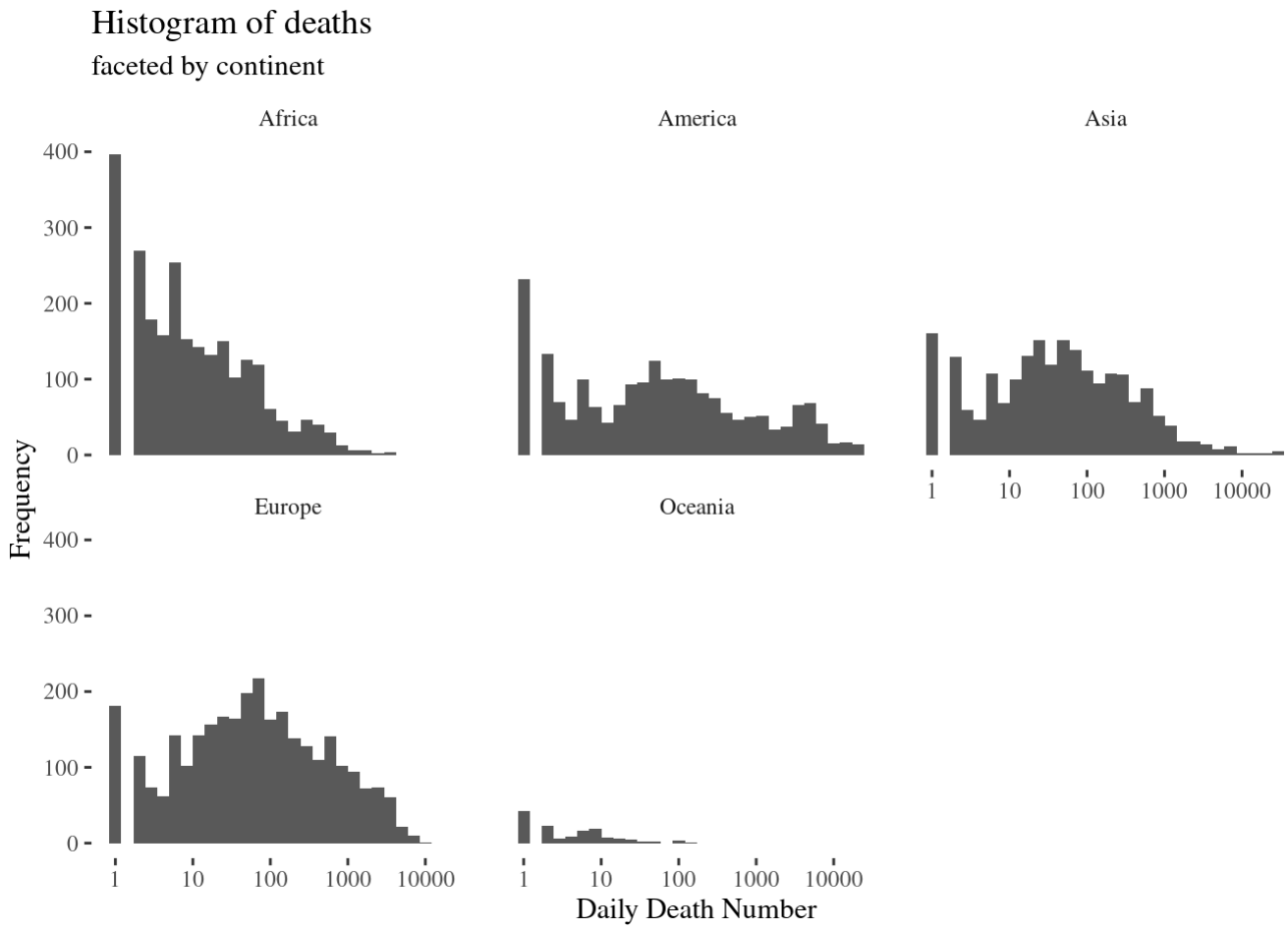
```
covid_data %>%
ggplot(aes(x = deaths)) +
  facet_wrap(. ~ continent) +
  geom_histogram() +
  scale_x_log10() +
  theme_tufte() +
  labs(
    title = 'Histogram of deaths',
    subtitle = 'faceted by continent',
    x = 'Daily Death Number',
    y = 'Frequency'
  )
```

```
## Warning in self$trans$transform(x): NaNs produced
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 7374 rows containing non-finite values (stat_bin).
```



Compared to the previous graph, this one looks more normally distributed and the differences in the bar heights are smaller.

Based on these histograms, which continents seem to have the highest numbers of daily COVID-19 deaths?

Europe. The mean of the log(x) distribution seems to be the biggest (around 100), then the x value must have been the biggest.

Regression Model: Univariate

We will first look at associations between COVID-19 deaths and a column/variable called `pandemic_prep` in the dataset. This variable is a measure of how prepared a country was for pandemics in general before COVID-19 hit.

Enter in the correct formula arguments to the `brm` function to predict `deaths` as the outcome with `pandemic_prep` as the independent (right-hand side) variable. Then run the code chunk to fit the model. (Hint: check the `brm` command help page for examples).

```
deaths_univ_mod <- brm(formula = deaths ~ pandemic_prep,
                       data=covid_data,
                       refresh=0)
```

```
## Warning: Rows containing NAs were excluded from the model.
```

```
## Compiling Stan program...
```

```
## Trying to compile a simple C file
```

```
## Running /opt/R/4.1.0/lib/R/bin/R CMD SHLIB foo.c
## gcc -I"/opt/R/4.1.0/lib/R/include" -DNDEBUG   -I"/cloud/lib/x86_64-pc-linux-gnu-li
brary/4.1/Rcpp/include/"  -I"/cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppEigen/inc
lude/"  -I"/cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppEigen/include/unsupported"
-I"/cloud/lib/x86_64-pc-linux-gnu-library/4.1/BH/include" -I"/cloud/lib/x86_64-pc-lin
ux-gnu-library/4.1/StanHeaders/include/src/"  -I"/cloud/lib/x86_64-pc-linux-gnu-libra
ry/4.1/StanHeaders/include/"  -I"/cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppParal
lel/include/"  -I"/cloud/lib/x86_64-pc-linux-gnu-library/4.1/rstan/include" -DEIGEN_N
O_DEBUG  -DBOOST_DISABLE_ASSERTS  -DBOOST_PENDING_INTEGER_LOG2_HPP  -DSTAN_THREADS  -
DBOOST_NO_AUTO_PTR  -include '/cloud/lib/x86_64-pc-linux-gnu-library/4.1/StanHeaders/
include/stan/math/prim/mat/fun/Eigen.hpp'  -D_REENTRANT -DRCPP_PARALLEL_USE_TBB=1   -
I/usr/local/include   -fpic  -g -O2  -c foo.c -o foo.o
## In file included from /cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppEigen/includ
e/Eigen/Core:88,
##                  from /cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppEigen/includ
e/Eigen/Dense:1,
##                  from /cloud/lib/x86_64-pc-linux-gnu-library/4.1/StanHeaders/inclu
de/stan/math/prim/mat/fun/Eigen.hpp:13,
##                  from <command-line>:
## /cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppEigen/include/Eigen/src/Core/util/M
acros.h:628:1: error: unknown type name ‘namespace’
##   628 | namespace Eigen {
##       | ^~~~~~~~~
## /cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppEigen/include/Eigen/src/Core/util/M
acros.h:628:17: error: expected ‘=’, ‘,’, ‘;’, ‘asm’ or ‘__attribute__’ before ‘{’ to
ken
##   628 | namespace Eigen {
##       |                 ^
## In file included from /cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppEigen/includ
e/Eigen/Dense:1,
##                  from /cloud/lib/x86_64-pc-linux-gnu-library/4.1/StanHeaders/inclu
de/stan/math/prim/mat/fun/Eigen.hpp:13,
##                  from <command-line>:
## /cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppEigen/include/Eigen/Core:96:10: fat
al error: complex: No such file or directory
##    96 | #include <complex>
##       |          ^~~~~~~~~
## compilation terminated.
## make: *** [/opt/R/4.1.0/lib/R/etc/Makeconf:168: foo.o] Error 1
```

```
## Start sampling
```

```
summary(deaths_univ_mod)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: deaths ~ pandemic_prep
##    Data: covid_data (Number of observations: 14800)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Population-Level Effects:
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept      -660.66     31.54  -722.05  -597.65 1.00     3857     2899
## pandemic_prep    22.55      0.72    21.12    23.96 1.00     3918     3333
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma  1238.91      7.33  1224.95  1253.32 1.00     4346     2959
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Based on the output from the summary command, you should be able to see the value of the beta coefficient for the `pandemic_prep` variable (i.e., the slope). In the following sentence, fill in the blanks to interpret the association for the variables:

For a 1 unit increase in `pandemic_prep`, the number of COVID-19 deaths increased by 22.5490474 on average (i.e. the most likely estimate), with an uncertainty interval from 21.1153338 to 23.9593958.

It looks like `pandemic_prep` actually increased COVID-19 deaths. Do you think this association is causal? Why or why not?

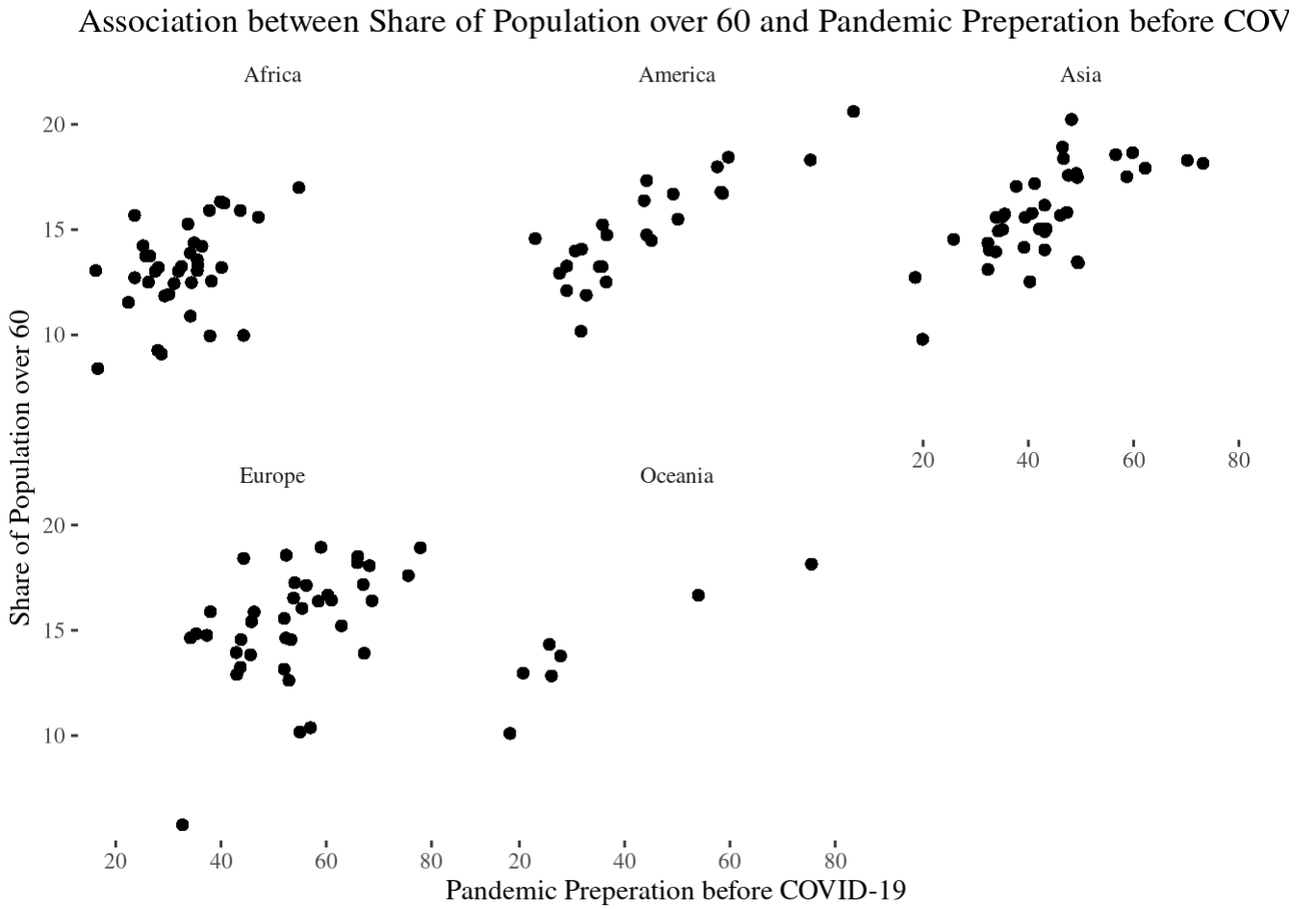
No, it is not. One should not say that `pandemic_prep` increased COVID-19 deaths. Correlation does not necessarily imply causality, so the only thing we are certain is that there is an association. Instead, one should say that the increase of `pandemic_prep` is associated with an increase in COVID-19 deaths. Also, from simple logic, it does not make sense that preparation causes covid deaths.

Air Travel and Preparedness

We will next look at a variable that correlates with `pandemic_prep` and could also affect `deaths`. This variable is `air_travel`, or the measure of the average total number of flights into a given country over time. First, to examine the association visually, construct a scatterplot (`geom_point`) with `ggplot2` where `pandemic_prep` is on the x axis and `air_travel` is on the y axis. You should also facet this plot by `continent` using `facet_wrap`. To make your plot look pretty, add the `+ theme_tufte()` command at the end of the ggplot function call, and be sure to add appropriate legends/labels.

```
covid_data %>%
ggplot(aes(x = pandemic_prep, y = air_travel)) +
  geom_point() +
  facet_wrap(. ~ continent) +
  theme_tufte() +
  labs(
    title = 'Association between Share of Population over 60 and Pandemic Preperation
before COVID-19',
    x = 'Pandemic Preperation before COVID-19',
    y = 'Share of Population over 60'
  )
```

```
## Warning: Removed 5440 rows containing missing values (geom_point).
```



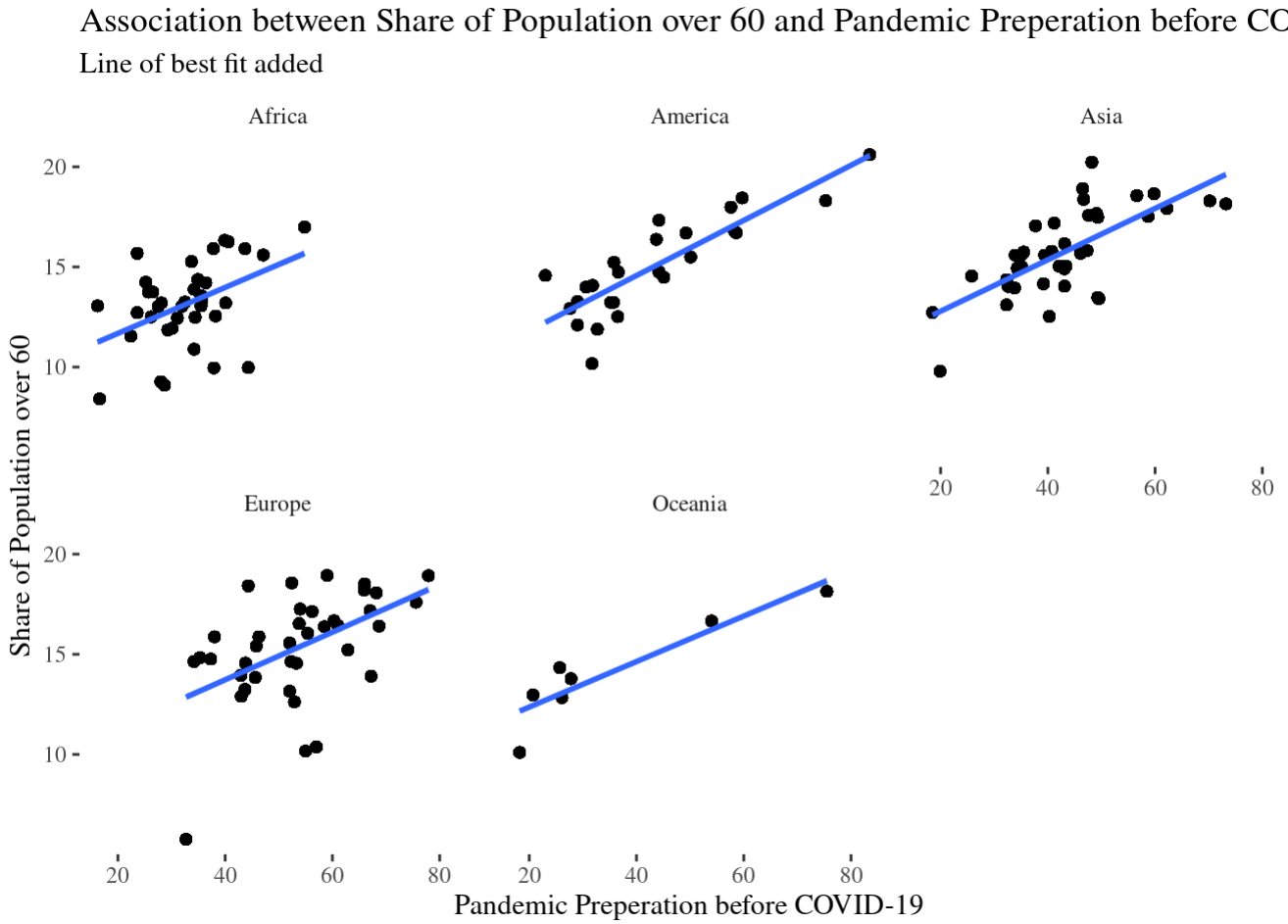
Now do the same plot except add the `stat_smooth(method="lm")` command to add a line of best fit to the relationship:

```
covid_data %>%
ggplot(aes(x = pandemic_prep, y = air_travel)) +
  geom_point() +
  facet_wrap(. ~ continent) +
  stat_smooth(method="lm") +
  theme_tufte() +
  labs(
    title = 'Association between Share of Population over 60 and Pandemic Preperation
before COVID-19',
    subtitle = 'Line of best fit added',
    x = 'Pandemic Preperation before COVID-19',
    y = 'Share of Population over 60'
  )
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 5440 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 5440 rows containing missing values (geom_point).
```



On the whole, how would you describe the relationship between `pandemic_prep` and `air_travel` ?

They are positively correlated across all the continents.

Think about the well-known relationship that COVID-19 mortality is much higher for older people. How could `air_travel` be a confounding variable that could explain the odd association between `pandemic_prep` and `deaths` ?

Countries with higher air travel have a higher risk of imported pandemics, which cause them to be better prepared. Also, countries with a higher air travel are more susceptible to COVID-19, therefore causing more deaths. Therefore, `air_travel` creates a spurious correlation between `pandemic_prep` and `deaths` .

Control with Regression

To test our theory that `air_travel` might be a confounder that could explain the strange association between `pandemic_prep` and `deaths` , let's do another regression model except we will include both `air_travel` *and* `pandemic_prep` as right-hand side variables (join with the `+` sign) with `deaths` as the outcome/left-hand side variable.

```
covid_bivar_mod <- brm(formula= deaths ~ air_travel + pandemic_prep,
                       data=covid_data,
                       refresh=0)
```

```
## Warning: Rows containing NAs were excluded from the model.
```

```
## Compiling Stan program...
```

```
## Trying to compile a simple C file
```

```
## Running /opt/R/4.1.0/lib/R/bin/R CMD SHLIB foo.c
## gcc -I"/opt/R/4.1.0/lib/R/include" -DNDEBUG   -I"/cloud/lib/x86_64-pc-linux-gnu-li
brary/4.1/Rcpp/include/"  -I"/cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppEigen/inc
lude/"  -I"/cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppEigen/include/unsupported"
-I"/cloud/lib/x86_64-pc-linux-gnu-library/4.1/BH/include" -I"/cloud/lib/x86_64-pc-lin
ux-gnu-library/4.1/StanHeaders/include/src/"  -I"/cloud/lib/x86_64-pc-linux-gnu-libra
ry/4.1/StanHeaders/include/"  -I"/cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppParal
lel/include/"  -I"/cloud/lib/x86_64-pc-linux-gnu-library/4.1/rstan/include" -DEIGEN_N
O_DEBUG  -DBOOST_DISABLE_ASSERTS  -DBOOST_PENDING_INTEGER_LOG2_HPP  -DSTAN_THREADS  -
DBOOST_NO_AUTO_PTR  -include '/cloud/lib/x86_64-pc-linux-gnu-library/4.1/StanHeaders/
include/stan/math/prim/mat/fun/Eigen.hpp'  -D_REENTRANT -DRCPP_PARALLEL_USE_TBB=1   -
I/usr/local/include   -fpic  -g -O2  -c foo.c -o foo.o
## In file included from /cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppEigen/includ
e/Eigen/Core:88,
##                  from /cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppEigen/includ
e/Eigen/Dense:1,
##                  from /cloud/lib/x86_64-pc-linux-gnu-library/4.1/StanHeaders/inclu
de/stan/math/prim/mat/fun/Eigen.hpp:13,
##                  from <command-line>:
## /cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppEigen/include/Eigen/src/Core/util/M
acros.h:628:1: error: unknown type name ‘namespace’
##   628 | namespace Eigen {
##       | ^~~~~~~~~
## /cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppEigen/include/Eigen/src/Core/util/M
acros.h:628:17: error: expected ‘=’, ‘,’, ‘;’, ‘asm’ or ‘__attribute__’ before ‘{’ to
ken
##   628 | namespace Eigen {
##       |                 ^
## In file included from /cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppEigen/includ
e/Eigen/Dense:1,
##                  from /cloud/lib/x86_64-pc-linux-gnu-library/4.1/StanHeaders/inclu
de/stan/math/prim/mat/fun/Eigen.hpp:13,
##                  from <command-line>:
## /cloud/lib/x86_64-pc-linux-gnu-library/4.1/RcppEigen/include/Eigen/Core:96:10: fat
al error: complex: No such file or directory
##    96 | #include <complex>
##       |          ^~~~~~~~~
## compilation terminated.
## make: *** [/opt/R/4.1.0/lib/R/etc/Makeconf:168: foo.o] Error 1
```

```
## Start sampling
```

```
summary(covid_bivar_mod)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: deaths ~ air_travel + pandemic_prep
##    Data: covid_data (Number of observations: 11680)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Population-Level Effects:
##              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     -1991.92     74.57 -2140.03 -1847.83 1.00     4341     3597
## air_travel      126.35      6.38   113.69   139.04 1.00     3250     2994
## pandemic_prep    10.64      1.16     8.39    12.89 1.00     3304     2757
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma  1364.57      8.98  1347.48  1382.84 1.00     4095     2798
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Now fill out the sentence again about the relationship between `pandemic_prep` and `deaths`:

For a 1 unit increase in `pandemic_prep`, the number of COVID-19 deaths increased by 10.6409529 on average (i.e. the most likely estimate), with an uncertainty interval from 8.3916342 to 12.8919031.

Compared to your earlier estimate, including `air_travel` reduced the size of the association by about one-half, or about 50% decrease. There is still a puzzling association in that countries with higher levels of pandemic preparedness experienced more COVID-19 deaths, but the original association we looked at can be partly explained by the inclusion of air travel as a *control* variable.

Why might the level of air travel in a country be a possible explanation for why that country is more prepared for pandemics and also more likely to experience COVID-19 deaths?

A higher level of air travel means that the average total number of flights into a given country is higher. Then, the country might be more prepared for any pandemics because more flights have a higher risk of spreading pandemics. Also, countries with a higher air travel are more susceptible to COVID-19, therefore causing more deaths.

Posterior Predictions

Finally, we will use our fitted model to examine how many COVID-19 deaths a country might experience given different levels of pandemic preparedness and air travel. To do so we will use the `posterior_epred` function from the `brms` package to calculate draws/samples for our experiments.

First, fill in the code in the chunk below to find the predicted number of COVID-19 deaths for a country with the *minimum* (hint: `min` function) pandemic preparedness score and the *maximum* (hint: `max` function) air travel. (Another hint: you may need to include the `na.rm=T` argument to these functions given the presence of `NA` values).

To create our predictions, we will pass a `tibble` with these values to the `posterior_epred` function:

```
# use an appropriate function to create these columns in the
# underlines below

predict_data <- tibble(pandemic_prep = min(covid_data$pandemic_prep, na.rm=T),
                       air_travel = max(covid_data$air_travel, na.rm=T))


mod_pred <- posterior_epred(covid_bivar_mod,
                            newdata=predict_data)

# we want to convert if from a matrix to a tibble

mod_pred <- as_tibble(mod_pred)
```

```
## Warning: The `x` argument of `as_tibble.matrix()` must have unique column names if
`.name_repair` is omitted as of tibble 2.0.0.
## Using compatibility `.name_repair`.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
summary(mod_pred$V1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   484.5   743.3   783.8   784.0   826.2  1024.4
```

Based on the summary of the posterior draws shown above, what is the most likely (mean/median) value for the prediction? What is a plausible uncertainty interval for the predicted COVID-19 deaths (i.e. 1st and 3rd quartiles)?

The most likely mean is 784.0273035 and the most likely medium is 783.7790521. A plausible uncertainty interval is from 743.2855865 to 826.2334135.

Now we'll do the same thing, except we'll calculate predicted COVID-19 deaths for the *maximum* pandemic preparedness and *minimum* air travel:

```
# use an appropriate function to create these columns in the
# underlines below

predict_data <- tibble(pandemic_prep = max(covid_data$pandemic_prep, na.rm=T),
                       air_travel = min(covid_data$air_travel, na.rm=T))


mod_pred <- posterior_epred(covid_bivar_mod,
                            newdata=predict_data)

# we want to convert if from a matrix to a tibble

mod_pred <- as_tibble(mod_pred)

summary(mod_pred$V1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -717.88 -440.25 -375.68 -375.75 -312.65   27.04
```

Do you notice anything odd about the model's prediction? Why might this be a strange prediction for COVID-19 deaths?

The negative statistics are odd, because it is impossible for deaths to be negative.

What happened here is called an issue of *model fit*. There aren't any countries in the data with those specific values for air travel and preparedness (i.e. countries with a ton of pandemic preparedness but no air travel), so our model made a prediction far outside of the data. This is called an extrapolation, and it can be dangerous when we use a model beyond its original scope.

As a final exercise, using the `mod_pred` tibble, make a density plot (`geom_density`) to visualize our uncertainty in terms of the samples we have of the model prediction. Use the `fill` argument to the `geom_density` function to set the interior of the density curve to a pleasing color, and set the `alpha` argument to a value less than 1 to permit the density to be partially transparent. Be sure to add appropriate legends/labels.

```
mod_pred %>%
  ggplot(aes(x = V1)) +
  geom_density(fill = 'skyblue',
               alpha = 0.6) +
  labs(
    title = 'Density plot of model prediction uncertainty',
    x = 'Predicted Value',
    y = 'Density'
  )
```