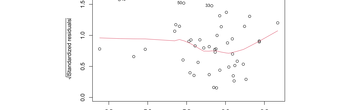
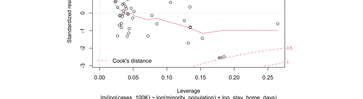
# W203.3: Lab 2: Regression to Study the Spread of Covid-19

Jeremy Fisher, Isabella Rodgers, Thomas Lungin
Joel Ferguson, Kevin Hodge

## Loading covid-19 csv file and setting up dataframes

```
load_update_date = "10-16-2020"
# load_update_date = "10-23-2020"

# Loading and preprocessing data
# Merge of Kaiser Hospitalizations, Jhu's Global Deaths, Jhu's Global Confirmed, and
# Merge of Kaiser Hospitalizations, Jhu's Global and Kaiser Deaths, and Jhu's Global
# Merge of Kaiser Hospitalizations, Jhu's Global and Kaiser Deaths, and Jhu's Global
# Merge of Kaiser Hospitalizations, Jhu's Global and Kaiser Deaths, and Jhu's Global

## Kaiser Hospitalizations
kaiser_hosp_file = "./data/kaiser_hospitalizations.csv"
kaiser_hosp = read.csv(kaiser_hosp_file)
kaiser_hosp$date = as.Date(kaiser_hosp$date)
kaiser_hosp$ca_hosp = kaiser_hosp$ca_hosp
kaiser_hosp$ca_icu = kaiser_hosp$ca_icu
kaiser_hosp$ca_vent = kaiser_hosp$ca_vent
kaiser_hosp$ca_hosp_per_icu = kaiser_hosp$ca_hosp / kaiser_hosp$ca_icu
kaiser_hosp$ca_icu_per_vent = kaiser_hosp$ca_icu / kaiser_hosp$ca_vent
kaiser_hosp$ca_hosp_per_vent = kaiser_hosp$ca_hosp / kaiser_hosp$ca_vent
kaiser_hosp$ca_icu_per_hosp = kaiser_hosp$ca_icu / kaiser_hosp$ca_hosp
kaiser_hosp$ca_vent_per_hosp = kaiser_hosp$ca_vent / kaiser_hosp$ca_hosp
kaiser_hosp$ca_vent_per_icu = kaiser_hosp$ca_vent / kaiser_hosp$ca_icu

## Global Deaths
jhu_deaths_file = "./data/jhu_global_deaths.csv"
jhu_deaths = read.csv(jhu_deaths_file)
jhu_deaths$date = as.Date(jhu_deaths$date)
jhu_deaths$ca_deaths = jhu_deaths$ca_deaths
jhu_deaths$ca_deaths_per_day = jhu_deaths$ca_deaths / jhu_deaths$ca_population
jhu_deaths$ca_deaths_per_million = jhu_deaths$ca_deaths / jhu_deaths$ca_population * 1000000
jhu_deaths$ca_deaths_per_hosp = jhu_deaths$ca_deaths / kaiser_hosp$ca_hosp
jhu_deaths$ca_deaths_per_icu = jhu_deaths$ca_deaths / kaiser_hosp$ca_icu
jhu_deaths$ca_deaths_per_vent = jhu_deaths$ca_deaths / kaiser_hosp$ca_vent
jhu_deaths$ca_deaths_per_hosp_per_day = jhu_deaths$ca_deaths / kaiser_hosp$ca_hosp / jhu_deaths$ca_days
jhu_deaths$ca_deaths_per_icu_per_day = jhu_deaths$ca_deaths / kaiser_hosp$ca_icu / jhu_deaths$ca_days
jhu_deaths$ca_deaths_per_vent_per_day = jhu_deaths$ca_deaths / kaiser_hosp$ca_vent / jhu_deaths$ca_days

## Global Confirmed
jhu_confirmed_file = "./data/jhu_global_confirmed.csv"
jhu_confirmed = read.csv(jhu_confirmed_file)
jhu_confirmed$date = as.Date(jhu_confirmed$date)
jhu_confirmed$ca_confirmed = jhu_confirmed$ca_confirmed
jhu_confirmed$ca_confirmed_per_day = jhu_confirmed$ca_confirmed / jhu_confirmed$ca_days
jhu_confirmed$ca_confirmed_per_million = jhu_confirmed$ca_confirmed / jhu_confirmed$ca_population * 1000000
jhu_confirmed$ca_confirmed_per_hosp = jhu_confirmed$ca_confirmed / kaiser_hosp$ca_hosp
jhu_confirmed$ca_confirmed_per_icu = jhu_confirmed$ca_confirmed / kaiser_hosp$ca_icu
jhu_confirmed$ca_confirmed_per_vent = jhu_confirmed$ca_confirmed / kaiser_hosp$ca_vent
jhu_confirmed$ca_confirmed_per_hosp_per_day = jhu_confirmed$ca_confirmed / kaiser_hosp$ca_hosp / jhu_confirmed$ca_days
jhu_confirmed$ca_confirmed_per_icu_per_day = jhu_confirmed$ca_confirmed / kaiser_hosp$ca_icu / jhu_confirmed$ca_days
jhu_confirmed$ca_confirmed_per_vent_per_day = jhu_confirmed$ca_confirmed / kaiser_hosp$ca_vent / jhu_confirmed$ca_days

## Merge Kaiser Hospitalizations, Jhu's Global and Kaiser Deaths, and Jhu's Global Confirmed
jhu_deaths_kaiser_hosp = merge(jhu_deaths, kaiser_hosp, by=c("ca", "date"))
jhu_deaths_kaiser_hosp = jhu_deaths_kaiser_hosp[order(jhu_deaths_kaiser_hosp$ca, jhu_deaths_kaiser_hosp$date),]
jhu_deaths_kaiser_hosp = jhu_deaths_kaiser_hosp[,c("ca", "date", "ca_deaths", "ca_deaths_per_day", "ca_deaths_per_million", "ca_deaths_per_hosp", "ca_deaths_per_icu", "ca_deaths_per_vent", "ca_deaths_per_hosp_per_day", "ca_deaths_per_icu_per_day", "ca_deaths_per_vent_per_day")]
jhu_deaths_kaiser_hosp = jhu_deaths_kaiser_hosp[complete.cases(jhu_deaths_kaiser_hosp),]

## Merge Kaiser Hospitalizations, Jhu's Global and Kaiser Deaths, and Jhu's Global Confirmed
jhu_deaths_kaiser_hosp_confirmed = merge(jhu_deaths_kaiser_hosp, jhu_confirmed, by=c("ca", "date"))
jhu_deaths_kaiser_hosp_confirmed = jhu_deaths_kaiser_hosp_confirmed[order(jhu_deaths_kaiser_hosp_confirmed$ca, jhu_deaths_kaiser_hosp_confirmed$date),]
jhu_deaths_kaiser_hosp_confirmed = jhu_deaths_kaiser_hosp_confirmed[,c("ca", "date", "ca_deaths", "ca_deaths_per_day", "ca_deaths_per_million", "ca_deaths_per_hosp", "ca_deaths_per_icu", "ca_deaths_per_vent", "ca_deaths_per_hosp_per_day", "ca_deaths_per_icu_per_day", "ca_deaths_per_vent_per_day", "ca_confirmed", "ca_confirmed_per_day", "ca_confirmed_per_million", "ca_confirmed_per_hosp", "ca_confirmed_per_icu", "ca_confirmed_per_vent", "ca_confirmed_per_hosp_per_day", "ca_confirmed_per_icu_per_day", "ca_confirmed_per_vent_per_day")]
jhu_deaths_kaiser_hosp_confirmed = jhu_deaths_kaiser_hosp_confirmed[complete.cases(jhu_deaths_kaiser_hosp_confirmed),]

## Merge Kaiser Hospitalizations, Jhu's Global and Kaiser Deaths, and Jhu's Global Confirmed
```




We also explored how the day_at_home_days variable is linked to the minority variable. While there is a positive relationship between the two, it seems statistically not significant and generally weak as indicated by the R2 of the regression below.

```
mod_home_min <- lm(log_stay_home_days ~ log(minority_population), data_df)
coefficient_home_min, nice = round4)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.00000    0.00000 1.0e+06  <2e-16 ***
## log(minority_population)  0.00000    0.27995 0.0000    1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mod_home_min)$r.squared
```

```
## [1] 1.8852767e-17
```

```
summary(mod_home_min)$adj.r.squared
```

```
## [1] 0.0440352
```

**Model 3**

Our last model includes several more variables that capture additional "intervention variables" including face coverings, tests, and business closures. Our interest in face coverings ultimately comes from the research cited in our literature review in U.S. society. The significance of face covering mandates isn't considerably high (better' log_face_covered_days' for 'mn_enforced?TRUE') though including the interaction term does improve the adjusted R2 of our model. The negative coefficient helps show the inherent nature of this subject - with face covering mandates, there may be a decrease in the rate of cases.

A limitation of our model is in regards to face covering mandates:

1. How much our society productively utilizes these face covering mandates?
2. How are these mandates enforced in the same way for every state?
3. How is it enforced?
4. Finally, does a state ultimately get more adoption from its population to wear face coverings when an official mandate exists?

Without data to help answer these questions, it may be difficult to understand the true scope of face coverings in our society.

For days since business closure mandates, we again see strong significance with low cases. The interesting nature of our model is how business closure and stay-at-home mandates both present significance when included together, though the nature of the relationship between the two is not clear. Further research is needed to understand this relationship.

We also included the population percentage between 0 and 12 people who are younger than 20 in this model, and this variable had a statistically significant p-value. Newsu media as well as in-depth research into COVID-19 have hinted at the fact that younger people may have a lower risk to develop complications or the disease compared to those whom are older. Due to this information, it is entirely possible that younger generations follow CDC guidelines and lockdown mandates more frequently. This can further expose a general population and create further spread potentially even to higher risk people. Another model may be that younger people are testing positive for COVID-19, but may have less severe symptoms or even asymptomatic. The coefficient for this variable is a high 2.26, which may display that states with younger populations may have a much greater case rate.

Tests per 100K is a state has a decent significance with case rate. This may present a reverse causality metric, in which a state's case rate increases which then compels a state government to increase testing.

$$LogCasesRate = \beta_0 + \beta_1 * LogMinorityPopulation + \beta_2 * LogPopulation < 12years old + \beta_3 * LogStayHomeDaysBusClosure + \beta_4 * LogFaceCoveredDays + \beta_5 * LogTests + \beta_6 * LogBusinessClosedDays$$

```
model3 <- lm(log(cases_100K) ~ log(minority_population) + log_stay_home_days
             + log(population_under_20) + log_face_covered_days + fm_enforced + log(tests)
             + log_business_closed_days, data = main_df)
coefficient(model3, nice = round4)
```

```
##
## t test of coefficients:
##
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.161968  2.213074 0.7319 0.465614
## log(minority_population)  0.000000  0.000000 0.0000 1.000000
## log_stay_home_days       -0.120511  0.050015 -2.4098 0.016993 *
## log(population_under_20)  2.261495  0.575564  3.9303 0.142899
## log_face_covered_days    -0.000000  0.000000 -0.0000 1.000000
## fm_enforcedTRUE          -0.575542  1.231240 -0.4675 0.641654
## log(tests)               0.072521  0.078729  0.9211 0.358812
## log_business_closed_days -0.000000  0.000000 -0.0000 1.000000
## log_face_covered_days:fm_enforcedTRUE  0.161473  0.206440  0.7820 0.435942
##
## (Intercept)                 ***
## log(minority_population)    **
## log_stay_home_days          .
## log(population_under_20)
## log_face_covered_days
## fm_enforcedTRUE
## log(tests)
## log_business_closed_days    **
## log_face_covered_days:fm_enforcedTRUE
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model3)$r.squared
```

```
## [1] 0.7950372
```

```
summary(model3)$adj.r.squared
```

```
## [1] 0.7408678
```

```
sd_business_kim <- pol %>%
  select(c(.,
           geom_binary(business = log_business_closed_days))
```

```
sd_business_scatter <- pol %>%
  ggplot(.) +
  aes(x = log_business_closed_days, y = log(cases_100K)) +
  geom_point() +
  stat_smooth()
```

```
sd_business_kim / sd_business_scatter
```

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

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
fm_kim <- pol %>%
  ggplot(.) +
  geom_binary(fm_enforced = log_face_covered_days)
```

```
fm_scatter <- pol %>%
  ggplot(.) +
  aes(x = log_face_covered_days, y = log(cases_100K), color = fm_enforced) +
  geom_point() +
  stat_smooth()
```

```
fm_kim / fm_scatter
```

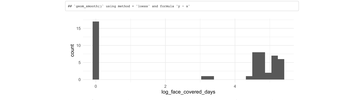
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
fm_kim / fm_scatter_2
```

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

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```




## 3. Limitations of your Model

We are working with a state-level dataset, therefore we have 51 observations (when we include Washington DC). This can be considered a small sample, and the asymptotic properties that we used as the CLT to get to the properties of OLS estimators most likely do not apply. For small samples, the key assumptions underpinning the Classic Linear Models are:

1. Is our sample IID? As we discussed in the Data section above, we do not believe that our sample is iid. Observations across groups of states are likely correlated with each other due to spatial dependence and across states due to varying methodologies.
2. Linearity in parameters? It is hard for us to tell if our relationship is linear.
3. Absence of perfect collinearity? (tested at times, but to be more fully discussed in the final report)
4. Homoscedastic error term? From the model 1' graph, there seem to be some relationship between the residuals and the independent variable which leads to a heteroskedastic hypothesis (but to be done in the final report)
5. Zero conditional mean? We believe that model 2' fits the G-C-Q test indicated the residuals follow a normal distribution relatively closely (to be done for model 1 and 2 in the final report)

## 4. A Regression Table

Our three models are summarized below:

```
sm.model1 = coefficient(model1, nice = round4)[,2, "Std. Error"]
sm.model2 = coefficient(model2, nice = round4)[,2, "Std. Error"]
sm.model3 = coefficient(model3, nice = round4)[,2, "Std. Error"]

stargazer(model1, model2, model3, type = "latex", omit.stat = "F",
          dep.var.labelsinclude = FALSE,
          column.sep.width = "3pt", # to remove excess space
          font.size = "tiny", # to make font size smaller
          se = list(NULL, sm.model2, sm.model3),
          caption = "Relationship between Covid-19 cases and key variables in the U.S.")
```

```
##
## Table 1: Relationship between Covid-19 cases and key variables in the U.S.
## ==========================================================================
##                                   Dependent variable:
## ---------------------------------------------------------------------------
##                                       log(cases_100K)
## ---------------------------------------------------------------------------
##                                  (1)          (2)          (3)
## ---------------------------------------------------------------------------
## log(minority_population)       0.000***     0.000***     0.000***
##                               (0.129)      (0.076)      (0.050)
## ---------------------------------------------------------------------------
## log_stay_home_days                         -0.320***    -0.120**
##                                            (0.127)      (0.050)
## ---------------------------------------------------------------------------
## log(population_under_20)                                 2.261**
##                                                         (0.960)
## ---------------------------------------------------------------------------
## log_face_covered_days                                   -0.000
##                                                         (0.000)
## ---------------------------------------------------------------------------
## fm_enforced                                             -0.576
##                                                         (0.667)
## ---------------------------------------------------------------------------
## log(tests)                                              0.274***
##                                                         (0.073)
## ---------------------------------------------------------------------------
## log_business_closed_days                                -0.000
##                                                         (0.000)
## ---------------------------------------------------------------------------
## log_face_covered_days:fm_enforced                        0.161
##                                                         (0.178)
## ---------------------------------------------------------------------------
## Constant                       0.000***     0.000***     0.162***
##                               (0.054)      (0.000)      (0.489)
## ---------------------------------------------------------------------------
## Observations                     50           50           50
## R2                              0.206        0.429        0.795
## Adjusted R2                     0.206        0.394        0.741
## Residual Std. Error             0.076 (df = 49) 0.061 (df = 42) 0.045 (df = 36)
## ===========================================================================
## Note:                                   *p<0.05; **p<0.01; ***p<0.001
```

From a high-level, the stargazer output shows robustness across all our models. In that coefficients' signs are consistent and significance levels remain high. Model 1 shows that the minority population variable is a significant factor in explaining case rate. Model 2 shows that a significant parameter in our first model and may show the impact that autocorrelation constructs may have on the exposure of COVID. Furthermore, with the introduction of days since stay-home mandates, the model improves significantly. The negative correlation coefficient and its significance level suggests that stay-at-home mandates do have a negative impact on case rate, controlling for other factors. Model 3 shows that may also describe how governors successfully use stay-at-home mandates to "drive down the curve". In our third model, tests significance presents both potential forward and reverse causality. Forward-wise, governments will find cases at a faster rate when testing capacity is increased. Reverse-wise, governments will increase testing when cases spike. Interestingly, both causalities together allow for a population to have greater visibility for the virus.

Our model shows that stay-at-home mandates drive the enforcement of a fabricated mandate is essential for the 'face_covered' days to be meaningful per data. Through this interaction term, in comparison to the coefficients of both 'fm_enforced' and 'log_face_covered_days' being negative, actually has a positive correlation coefficient, which may imply that governors create a mandate only after cases spike to a local region. Another consideration for why our face covering variables have insignificance in comparison to our stay-at-home variable may be due to the inherent nature of when these practices were used in U.S. For example, stay at home mandates were made in the beginning of the pandemic whereas face covering mandates were made later. This may be a result of the information available to the public and scientists at the time of the timeframe. Consequently, cases may have risen despite the creation of face mask mandates due to the ending of stay at home mandates in certain states. Another possibility is that face mask mandates may simply not be effective in controlling the spread of COVID.

Finally, in our third model, we also have a variable that represents business closures. Per data, this variable may be very consistent with our stay at home mandate variable. Thus, there may exist a certain collinearity with these two parameters that may need to be further explored. Our model shows that business closures have a negligible impact.

Another important note is that we cannot be together this a majority of our variables due to the underlying heteroskedasticity of our variables. As can be seen above, the minority population variable with cases rate is a perfect example. The variability of residuals changes per a given minority population percentage. The nature of this can be due to the fact that when a population is exposed enough to COVID, despite any characteristics of the population, it can spread exponentially.

## 5. Discussion of Omitted Variables

The first omitted variable we analyzed was the *stay_home_days* which we have available data and could further tests its effects. The $\hat{\delta}_{OVB}$ coefficient in model 3, that represents the explanation of the case rate given the population of minorities is lower than $\hat{\delta}_{OLS}$ in model 2 because of the bias from the omitted variable *stay_home_days*. As seen in the summaries for model 1, model 2 and the regression of *stay_home_days* given *minority_population*:

If $\hat{\beta}_1 < 0$ and $\alpha_1 > 0$ then the omitted variable bias is $\hat{\beta}_1\alpha_1 < 0$ and $\hat{\beta}_1 > 0$ then the OLS coefficient on minority will be scaled toward zero (less positive) being statistical significance.

That explains that although a higher proportion of minority population has a positive effect in the case rate in model 1, it is underestimated in model 1 as it does not capture the negative effect of staying at home.

Model 1 $\hat{\beta}_1$ and $\hat{\beta}_2$ coefficients:

```
summary.lm(model1)
```

```
##
## Call:
## lm(formula = log(cases_100K) ~ log(minority_population), data = main_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.130667 -0.031575  0.007798  0.079861  0.200000
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.000000   0.054000       0        1
## log(minority_population)  0.129077   0.027995  4.611  3.2e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Residual standard error: 0.0763 on 50 degrees of freedom
##  (2 observations deleted due to missingness)
## Multiple R-squared:  0.2056, Adjusted R-squared:  0.2002
## F-statistic:  19.3 on 1 and 50 DF,  p-value: 0.00034
```

Model 2 $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\gamma}_1$ coefficients:

```
summary.lm(model2)
```

```
##
## Call:
## lm(formula = log(cases_100K) ~ log(minority_population) + log_stay_home_days,
##     data = main_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.190127 -0.041024  0.000002  0.051042  0.200000
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.000000   0.054000       0        1
## log(minority_population)  0.076000   0.050000  1.520    0.136
## log_stay_home_days       -0.320000   0.126900 -2.521    0.015 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Residual standard error: 0.0608 on 49 degrees of freedom
##  (2 observations deleted due to missingness)
## Multiple R-squared:  0.4293, Adjusted R-squared:  0.4059
## F-statistic: 18.34 on 2 and 49 DF,  p-value: 0.0000
```

Model for omitted variable *stay_home_days* $\hat{\alpha}_0$ and $\hat{\alpha}_1$ coefficients:

```
summary.lm(sed_home_min)
```

```
## Call:
## lm(formula = log_stay_home_days ~ log(minority_population), data = main_df)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.00000 -0.00000  0.00000  0.00000  0.00000
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.000000   0.000000       0        1
## log(minority_population)  0.000000   0.279950  0.0000    1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Residual standard error: 0.0000 on 50 degrees of freedom
##  (2 observations deleted due to missingness)
## Multiple R-squared:  0.000,  Adjusted R-squared:  -0.0204
## F-statistic: 0.000 on 1 and 50 DF,  p-value: 1.000
```

```
power('Omitted Variable Bias on minority:', -0.12092 * 0.3000)
```

```
## [1] "Omitted Variable Bias on minority: -0.036276"
```

```
power('Approximation for minority coefficient when stay home days is omitted:', -0.12092 + 0.3000 * 0.036276)
```

```
## [1] "Approximation for minority coefficient when stay home days is omitted: -0.109117"
```

For our first report, we will investigate 2 additional omitted variables (either from model 3 or variables not present in data set).

## 6. Conclusion

Our research explored the relationship between Covid cases in a State and various variables. To account for the different levels of population in the various States, we decided to work with the number of cases per 100K citizens. Using data as of 10/30/2020 in our first model, we uncovered a statistically significant relationship between the log of the ratio of cases per 100k of the sample and the fraction of minority residents in a State. Our second model, which controlled for the number of days since stay-at-home mandates were implemented, uncovered the impact of the policy on the number of cases on cases increase in a given sample. Our third model evaluated the impact on the first and second models of the inclusion of other population characteristics (presence of youth between the age of 20 and other potential public policy interventions (e.g., length of time since a mask mandate came in effect).

While we uncovered a statistically significant relationship between the proportion of a state's minority population and case prevalence, as well as highlighted the effectiveness of stay-at-home orders, we believe our model is statistically significant relationship for face masks. We hesitate to draw any definitive conclusions on the data point given the nature of the operationalization of the variable (a face mask mandate does not mean that everyone wears a face mask).

Our analysis also raises more questions than it answers at this stage. If that are the underlying causal forces among minority populations in a state? (2) What is the economic impact of stay at home mandates? How is the economic burden distributed across members of society? (3) Given the heterogeneity across states, what are effective ways to study causal and reverse causal channels?