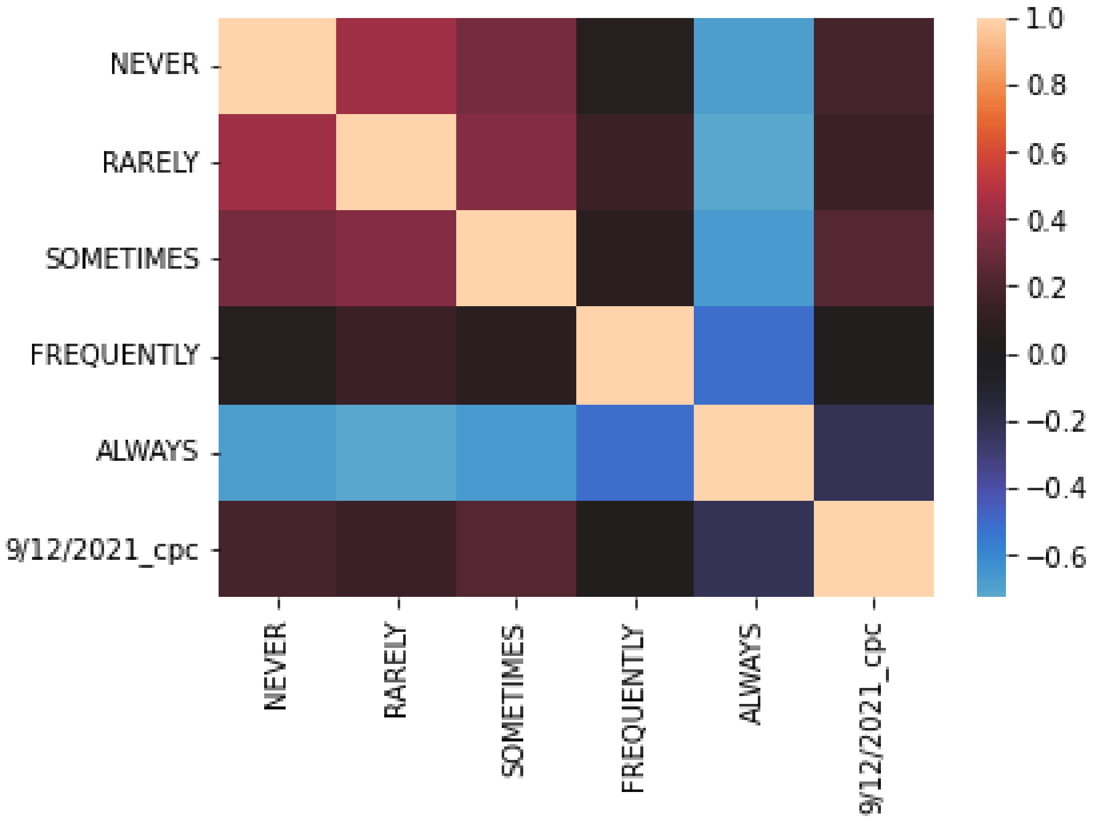
0.0.1 Question 2c

In our first model, we will use county-wise mask usage data to predict the number of COVID-19 cases on September 12th, 2021 (i.e., the column `9/12/2021_cpc`). Create a visualization that shows the pairwise correlation between each combination of columns in `mask_data`. For 2-D visualizations, consider Seaborn's heatmap.

Hint: You should be plotting 36 values corresponding to the pairwise correlations of the six columns in `mask_data`.

```
In [45]: sns.heatmap(mask_data.corr(),center = 0)
```

```
Out[45]: <AxesSubplot:>
```



0.0.2 Question 2d

(1) Describe the trends and takeaways visible in the visualization of pairwise correlations you plotted in Question 2c.

(2) Consider the following linear regression model

$$\hat{y} = \theta^T x,$$

where $\hat{y}$ is the predicted number of COVID-19 cases per capita on 9/12/2021 and x is the five mask usage features. Comment on the quality of predictions and interpretability of features if we fit this linear model to the data.

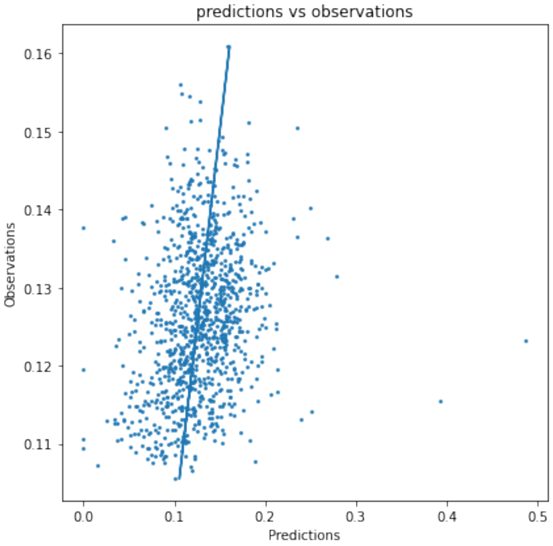
1) The more frequently people are wearing mask, the smaller the correlation to covid cases is. Which means putting mask on have a positive effect on stoping the spread of covid, and people who seldomly wearing mask are more likely to have covid.
2) From the correlation heatmap above, we could see the relationship of each features to covid cases are diagonally symmetric, thus transposing the matrix of features should not have any influence on final prediction.

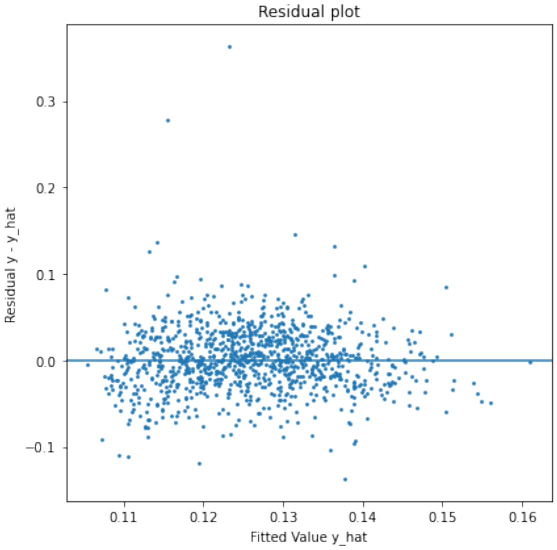
0.0.3 Question 3b

Visualize the model performance from part (a) by plotting two visualizations: (1) the predictions vs observations on the test set and (2) the residuals for the test set.

Some notes: * We've used `plt.subplot` (documentation) so that you can view both visualizations side-by-side. For example, `plt.subplot(121)` sets the plottable area to the first column of a 1x2 plot grid; you can then call Matplotlib and Seaborn functions to plot that area, before the next `plt.subplot(122)` area is set. * Remember to add a guiding line to both plot where $\hat{y} = y$, i.e., where the residual is 0. * Remember to label your axes.

```
In [48]: plt.figure(figsize=(12,6))      # do not change this line

         plt.subplot(121)                # do not change this line
         # (1) predictions vs observations
         plt.scatter(y_test, linear_model.predict(X_test),s=3, alpha=1)
         plt.plot(linear_model.predict(X_test),linear_model.predict(X_test))
         plt.xlabel('Predictions')
         plt.ylabel('Observations')
         plt.title('predictions vs observations')

         plt.subplot(122)               # do not change this line
         # (2) residual plot
         plt.scatter(linear_model.predict(X_test),y_test-linear_model.predict(X_test),s=3, alpha=1)
         # plt.plot(linear_model.predict(X_test),linear_model.predict(X_test))
         plt.axhline(0)
         plt.xlabel('Fitted Value y_hat')
         plt.ylabel('Residual y - y_hat')
         plt.title('Residual plot')



         plt.tight_layout()             # do not change this line
```


predictions vs observations
0.16
0.15
0.14
0.13
0.12
0.11
Observations
0.0
0.1
0.2
0.3
0.4
0.5
Predictions


Residual plot
0.3
0.2
0.1
0.0
−0.1
Residual y - y_hat
0.11
0.12
0.13
0.14
0.15
0.16
Fitted Value y_hat

0.0.4 Question 3c

Describe what the plots in part (b) indicates about this linear model. Justify your answer.

The left plot of part b shows the relationship between predictions are observations. We could see a linear pattern from this plot, which means our model is working. The right plot is the residual of our model, with points represents the residual for each prediction values. We could see residuals are roughly evenly distributed above/below x-axis, which indicates a reasonable prediction.

0.0.5 Question 4d

Interpret the confidence intervals above for each of the θ_i, where θ_0 is the intercept term and the remaining θ_i for $i > 0$ are parameters corresponding to mask usage features. What does this indicate about our data and our model?

Describe a mathematical reason why this could be happening.

Hint: Take a look at the design matrix!

This means we have 95% of confidence to say our thetas are in above intervals. For theta 1-5, since they represents the frequency of wearing mask, and the sum of these five features are one, they are obviously correlated, which explains why the confidence intervals of them are close to each other.

0.0.6 Question 5b

Comment on the ratio `prop_var`, which is the proportion of the expected square error on the data point captured by the model variance. Is the model variance the dominant term in the bias-variance decomposition? If not, what term(s) dominate the bias-variance decomposition?

Justify your answer.

Since prop_var is very small(0.003), the variance is not the dominant term. Instead, the Model Bias should be the dominant term.

0.0.7 Question 5d

Propose a solution to reducing the mean square error using the insights gained from the bias-variance decomposition above. Please show all quantities and work that informs your analysis.

Assume that the standard bias-variance decomposition used in lecture can be applied here.

As we see from 5a and 5c, the mean_mse drop from 0.03 to 0.0013. This is because in 5a we are only predicting on one single points with multiple models, while we are predicting multiple points(250 points) with multiple models in 5c. By applying this multi-prediction, we are actually lowering the variance and bias at the same time. Since rmse is related with both model variance and bias, only dropping in both could effectively lower mean_rmse.

0.0.8 Question 6c

Compare the RMSE of our improved model with an extra feature with the intercept term removed with the RMSE obtained in the model from Question 3a.

Comment on what you would *expect* to happen if you repeated the multicollinearity and bias-variance analyses on this new model using bootstrapping. Specifically, what would you expect to happen with this new model bias?

Hint: If you wish, you may want to carry out this analysis by adding a cell below this. Please delete it afterwards and note that you *may* run into memory issues if you run it too many times!

the bias would not be zero since there is no intercept term. but the rmse is getting smaller.