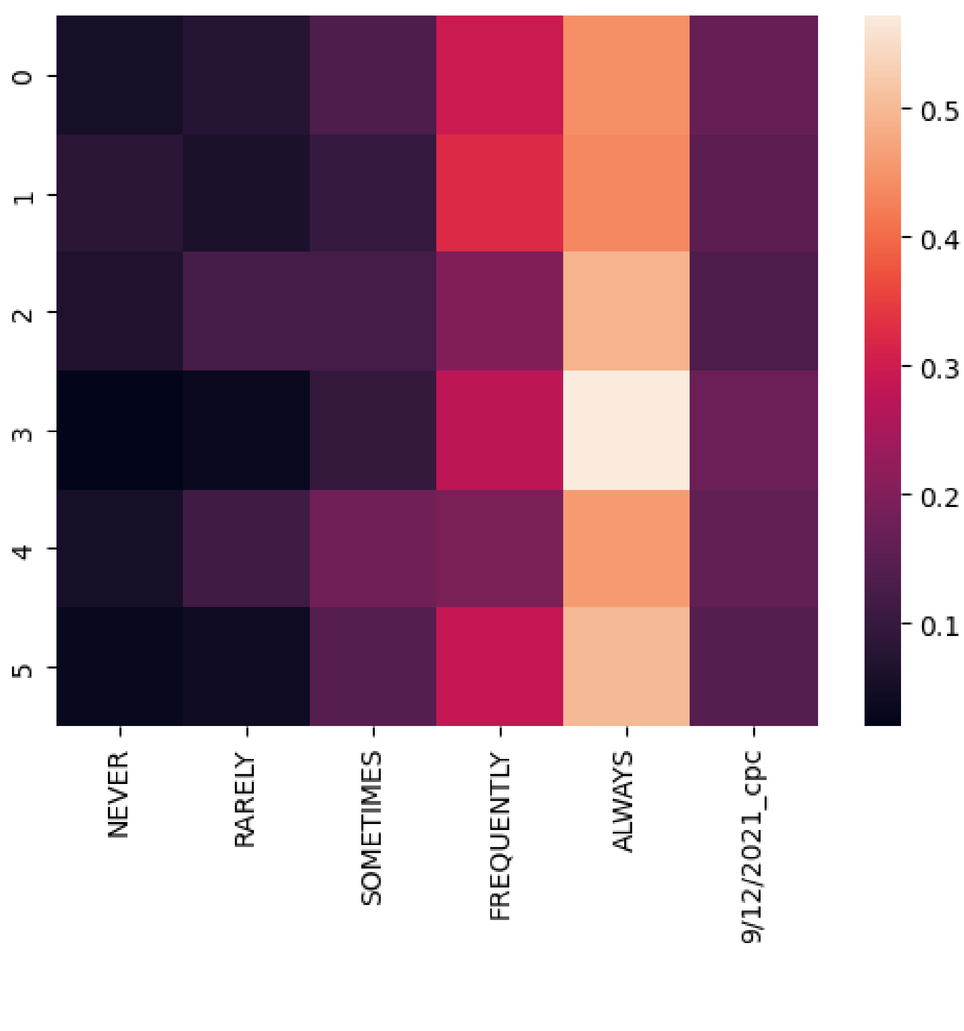
0.0.1 Question 2c

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 [98]: sns.heatmap(mask_data.head(6))
```

```
Out[98]: <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.

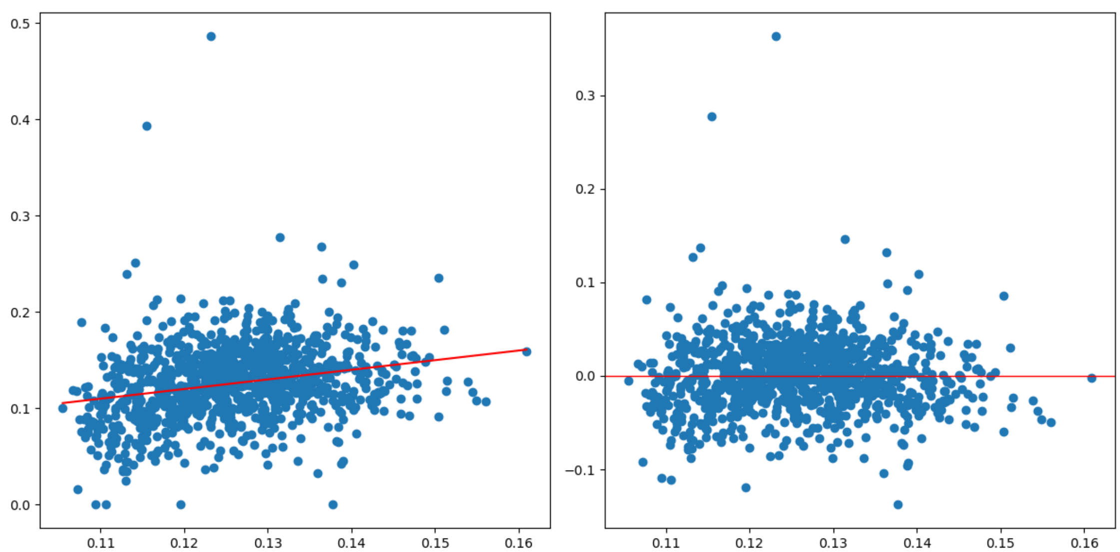
Type your answer here, replacing this text.

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 [124]:
```
plt.figure(figsize=(12,6))      # do not change this line
lg = np.linspace(np.min(y_predicted), np.max(y_predicted), 100)
plt.subplot(121)                # do not change this line
# (1) predictions vs observations
plt.scatter(y_predicted, y_test);
plt.plot(lg, lg, c='r')
plt.subplot(122)               # do not change this line
# (2) residual plot
plt.scatter(y_predicted, y_test-y_predicted);
plt.axhline(0, c='r', linewidth=1)
plt.tight_layout()             # do not change this line
```



0.0.4 Question 3c

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

Type your answer here, replacing this text.

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!

Type your answer here, replacing this text.

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.

Type your answer here, replacing this text.

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.

Type your answer here, replacing this text.

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!

Type your answer here, replacing this text.