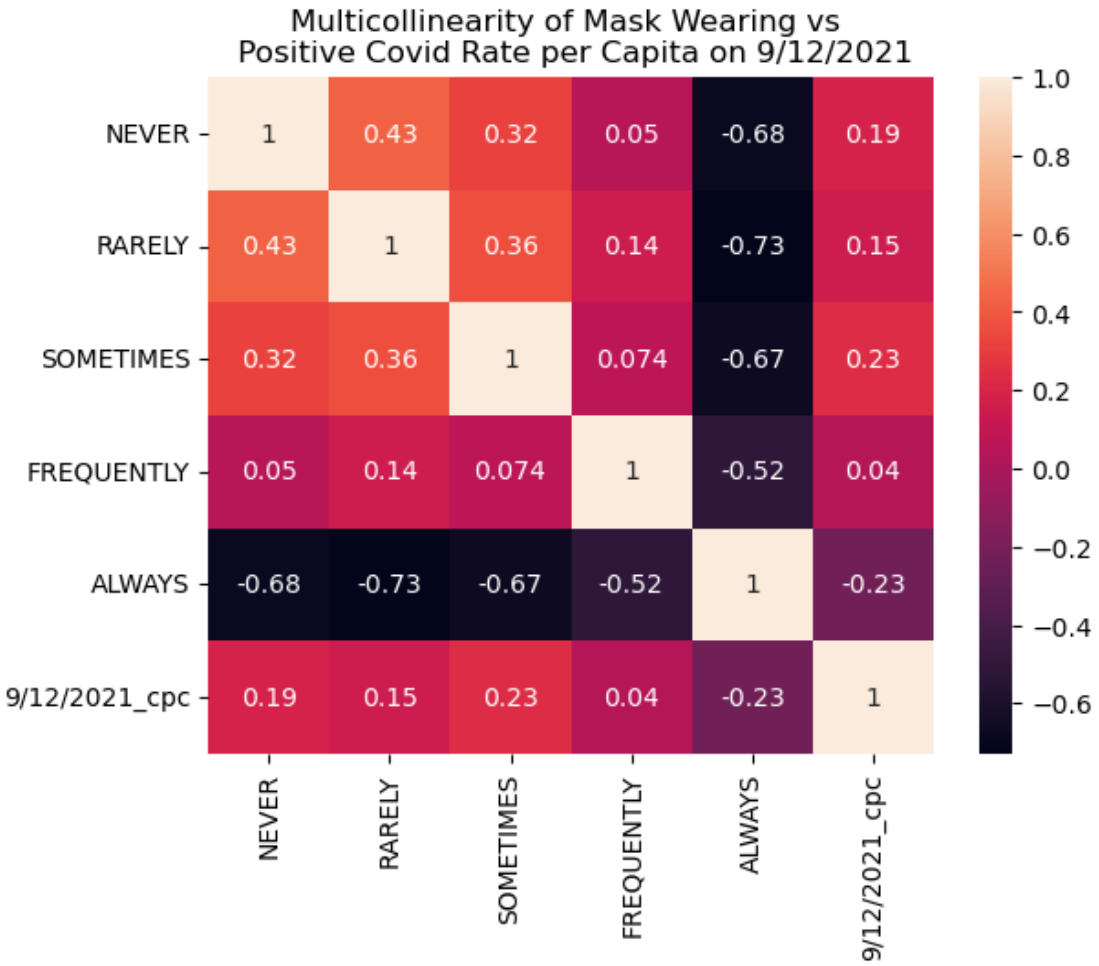

0.0.1 Question 1c

Our goal is to use county-wise mask usage data to predict the number of COVID-19 cases per capita on September 12th, 2021 (i.e., the column `9/12/2021_cpc`). But before modeling, let's do some EDA to explore the multicollinearity in these features, and then we will revisit this in question.

Create a visualization that shows the pairwise correlation between each combination of columns in `mask_data`. For 2-D visualizations, consider Seaborn's heatmap. Remember to title your plot.

Hint: You should be plotting 36 values corresponding to the pairwise correlations of the six columns in `mask_data`. You may optionally set `annot=True`, but it isn't necessary.

```
In [9]: sns.heatmap(data=mask_data.corr(), annot=True)
        plt.title('Multicollinearity of Mask Wearing vs \n Positive Covid Rate per Capita on 9/12/2021')
```


Multicollinearity of Mask Wearing vs
Positive Covid Rate per Capita on 9/12/2021
NEVER
RARELY
SOMETIMES
FREQUENTLY
ALWAYS
9/12/2021_cpc
1 0.43 0.32 0.05 -0.68 0.19
0.43 1 0.36 0.14 -0.73 0.15
0.32 0.36 1 0.074 -0.67 0.23
0.05 0.14 0.074 1 -0.52 0.04
-0.68 -0.73 -0.67 -0.52 1 -0.23
0.19 0.15 0.23 0.04 -0.23 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

0.0.2 Question 1d

Describe the trends and takeaways visible in the visualization of pairwise correlations you plotted in part (c). Specifically, what does the correlation between pairs of features (i.e., mask usage categories) look like? What does the correlation between mask usage categories and COVID-19 cases per capita look like?

The correlation between the ALWAYS wear mask feature and all the other mask frequency features is relatively strongly negative. Meanwhile, the NEVER, RARELY, and SOMETIMES mask usage categories seem to be more weakly positively correlated with each other. the FREQUENTLY category is relatively pretty weakly correlated to the other mask usage categories, except for with ALWAYS where it shows a relatively strong negative correlation (albeit not as strong as the other mask usage categories with ALWAYS).

The ALWAYS mask usage category also has the only negative correlation with covid 19 cases per capita. The SOMETIMES, ALWAYS, RARELY, and NEVER mask usage categories all have relatively similar levels of linear correlation with covid cases per capita, while the FREQUENTLY descriptor has the least correlation.

0.0.3 Question 1e

If we were to build a linear regression model (with an intercept term) using all five mask usage columns as features, what problem would we encounter?

We see strong colinearity between the ALWAYS mask usage category and all the other categories, which makes sense. The ALWAYS mask usage category is most likely not linearly independent from the other categories, as an increase in ALWAYS masked individuals would also lead to a decrease in all the other categories, and since ALWAYS masked and NEVER masked are the least subjective mask usage categories.

Since the ALWAYS mask usage feature is colinear with other features, it would lead to issues when bootstrapping to determine confidence intervals for whether or not that singular feature is relevant to the linear model (theta != 0)

Additionally, one of the mask features can be calculated from the rest of the mask features (by subtracting 1 from the sum of the rest), making the 5 mask usage columns not linearly independent. This could pose problems when using all the columns along with a bias column if attempting to solve for theta analytically (not with gradient descent)

0.0.4 Question 2b

To visualize the model performance from part (a), let's make the following two visualizations: 1. The observed values vs. the predicted values on the test set. 2. The residuals plot (note: in multiple linear regression, the residual plot has the residuals plotted against the predicted values).

In both plots, the predicted values should be on the x-axis so that it is easy to make comparisons.

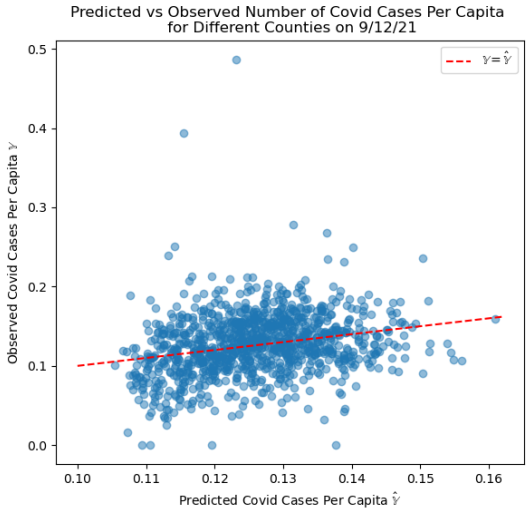
Note: * We've used `plt.subplot` (documentation) so that you can view both visualizations side-by-side. For example, `plt.subplot(121)` sets the plottable area to the first column of a 1x2 plot grid; you can then call `Matplotlib` and `Seaborn` functions to plot that area, before the next `plt.subplot(122)` area is set. * **Remember to add a guiding line to both plots where $\hat{Y} = Y$, i.e., where the residual is 0**. `plt.plot` and `plt.axhline` might be helpful here! * Please add descriptive titles and axis labels for your plots!

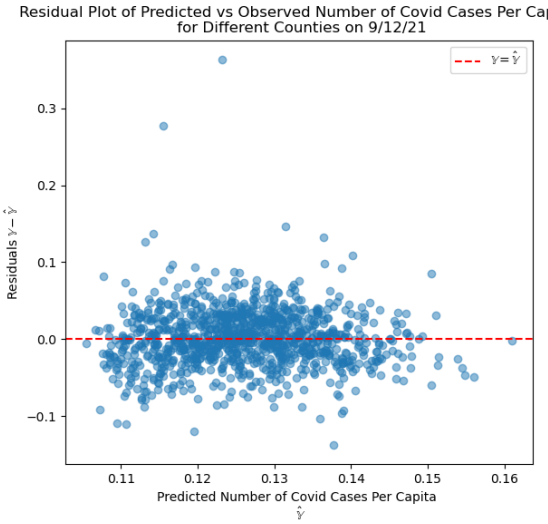
In [42]:
```
plt.figure(figsize=(12,6))      # do not change this line
plt.subplot(121)                # do not change this line
# 1. plot observations vs. predictions
plt.scatter(x=Y_test_pred, y=Y_test, alpha=0.5)
plt.xlabel('Predicted Covid Cases Per Capita $\hat{\mathbb{Y}}$')
plt.ylabel('Observed Covid Cases Per Capita $\mathbb{Y}$')
plt.title('Predicted vs Observed Number of Covid Cases Per Capita \n for Different Counties on

plot_vals = np.linspace(0.1, 0.162)
plt.plot(plot_vals, plot_vals, c='red', linestyle='--', label='$\mathbb{Y} = \hat{\mathbb{Y}}$
plt.legend()

plt.subplot(122)               # do not change this line
# 2. plot residual plot
plt.title('Residual Plot of Predicted vs Observed Number of Covid Cases Per Capita \n for Diff
plt.xlabel('Predicted Number of Covid Cases Per Capita \n $\hat{\mathbb{Y}}$')
plt.ylabel('Residuals $\mathbb{Y} - \hat{\mathbb{Y}}$');
plt.scatter(x=Y_test_pred, y=Y_test-Y_test_pred, alpha=0.5)
plt.axhline(y=0, c='red', linestyle='--', label='$\mathbb{Y} = \hat{\mathbb{Y}}$')
plt.legend()

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


Predicted vs Observed Number of Covid Cases Per Capita
for Different Counties on 9/12/21
$y = \hat{y}$
Observed Covid Cases Per Capita y
Predicted Covid Cases Per Capita $\hat{y}$
0.0
0.1
0.2
0.3
0.4
0.5
0.10
0.11
0.12
0.13
0.14
0.15
0.16


Residual Plot of Predicted vs Observed Number of Covid Cases Per Capita
for Different Counties on 9/12/21
$y = \hat{y}$
Residuals $y - \hat{y}$
Predicted Number of Covid Cases Per Capita
$\hat{y}$
−0.1
0.0
0.1
0.2
0.3
0.11
0.12
0.13
0.14
0.15
0.16

0.0.5 Question 2c

Describe what the plots in part (b) indicate about this linear model. In particular, are the predictions good?

The predictions are reasonable, as there is a reasonable linear trend for the predicted vs observed covid cases per capita graph. Additionally, the residuals seem randomly distributed across the Y = Y_hat line.

0.0.6 Question 3d

Interpret the confidence intervals above for each of the θ_i, where θ_0 is the intercept term, and the remaining θ_i's are parameters corresponding to mask usage features. How does this relate to your observations in `q1d`, and what does that indicate about the usefulness of all the features for your model?

Hint: The lecture or course notes discussing collinearity might be a useful starting point when approaching this question.

Thetas 1 through 5 are the various mask usage features, and we mentioned in Q1d that they show multicollinearity. This means that when making inferences about what features are important for our model, we cannot inherently change one feature independently while holding the others constant (if ALWAYS masked goes up, the other mask usage features *must* go down). Therefore, perhaps using only one of these features for the model makes the most sense. We cannot draw any conclusions about the intercept term.

0.0.7 Question 4b

Comment on the ratio `ratio`, 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?

Note: The Bias-Variance decomposition from the lecture is:

$$\text{model risk} = \sigma^2 + (\text{model bias})^2 + \text{model variance}$$

where σ^2 is the observation variance, or "irreducible error".

The ratio of the model variance to the model risk is quite small, only 0.006397075186207152. The model variance is not the dominating term for this bias variance dcomposition, as the model variance is much smaller than the model risk. It appears that the model bias and the observation variance dominate the model risk

---

0.0.8 Question 4d

Propose a method of reducing the mean square error using the insights gained from the bias-variance decomposition above.

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

We can add more features to increase the variance of our model, but simultaneously decreasing the mean squared error of our modell.