
0.0.1 Question 1c

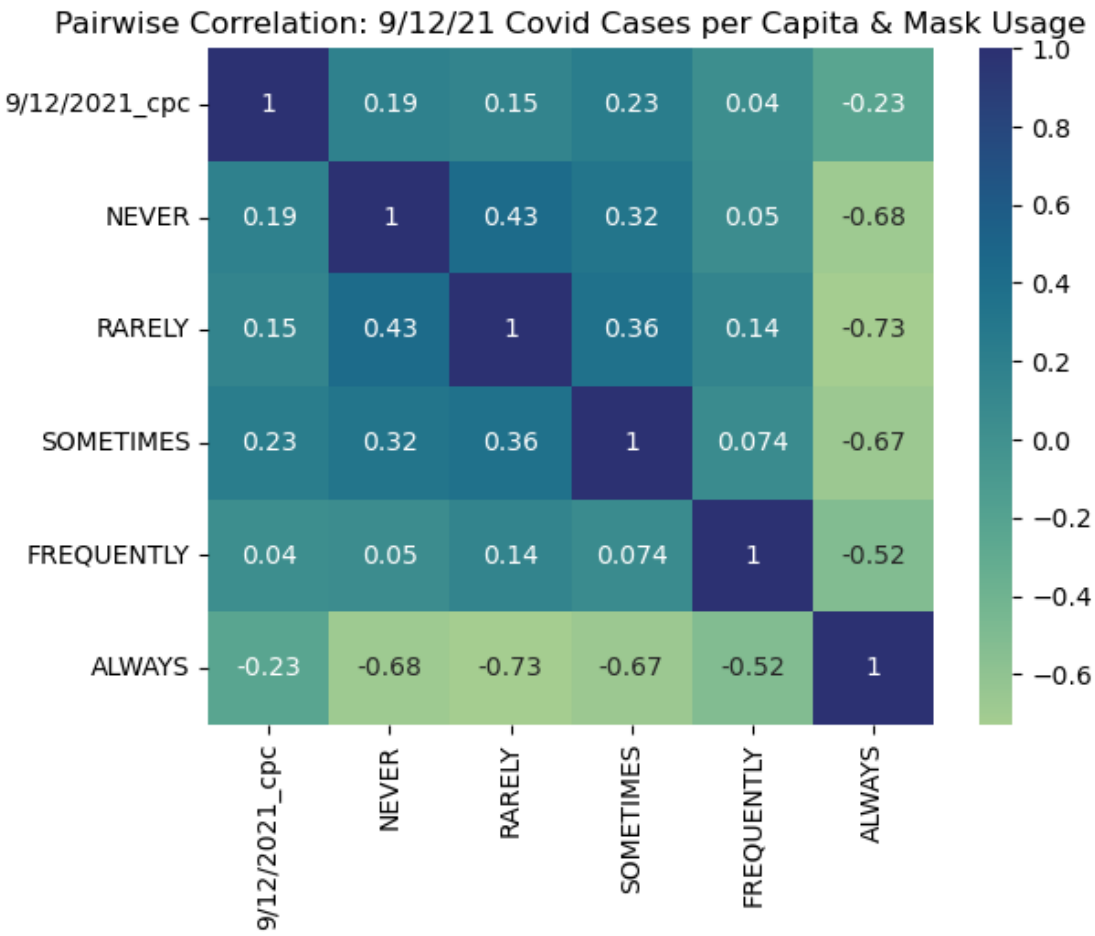
Our goal is to use county-wise mask usage data to predict the number of COVID-19 cases per capita on September 12th, 2021 (i.e., the column `9/12/2021_cpc`). But before modeling, let's do some EDA to explore the multicollinearity in these features, and then we will revisit this in question.

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 [35]: sns.heatmap(mask_data.corr(), annot = True, cmap="crest")
         plt.title('Pairwise Correlation: 9/12/21 Covid Cases per Capita & Mask Usage')
```

```
Out[35]: Text(0.5, 1.0, 'Pairwise Correlation: 9/12/21 Covid Cases per Capita & Mask Usage')
```


Pairwise Correlation: 9/12/21 Covid Cases per Capita & Mask Usage
9/12/2021_cpc
NEVER
RARELY
SOMETIMES
FREQUENTLY
ALWAYS
1
0.19
0.15
0.23
0.04
-0.23
0.19
1
0.43
0.32
0.05
-0.68
0.15
0.43
1
0.36
0.14
-0.73
0.23
0.32
0.36
1
0.074
-0.67
0.04
0.05
0.14
0.074
1
-0.52
-0.23
-0.68
-0.73
-0.67
-0.52
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 Question 1c. 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 graphic shows a strong negative association between wearing a mask at all times ("ALWAYS") and any other mask usage characteristic. In other words, when "ALWAYS" characteristic is high, other characteristics of mask usage are low. There is a strong positive correlation between the "RARELY" and "NEVER" columns. Lower mask usage, specifically "NEVER," appears to have a slight positive correlation with higher cases per capita. In contrast, higher mask usage, such as "ALWAYS," shows a minor negative connection with cases per capita, suggesting a potential decrease in cases with high mask usage.

0.0.3 Question 1e

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

When constructing a linear regression model using all five mask usage columns as features, a problem may arise due to the high collinearity between columns like “RARELY” and “NEVER.” This collinearity can make it challenging to select an appropriate model for linear regression.

0.0.4 Question 2b

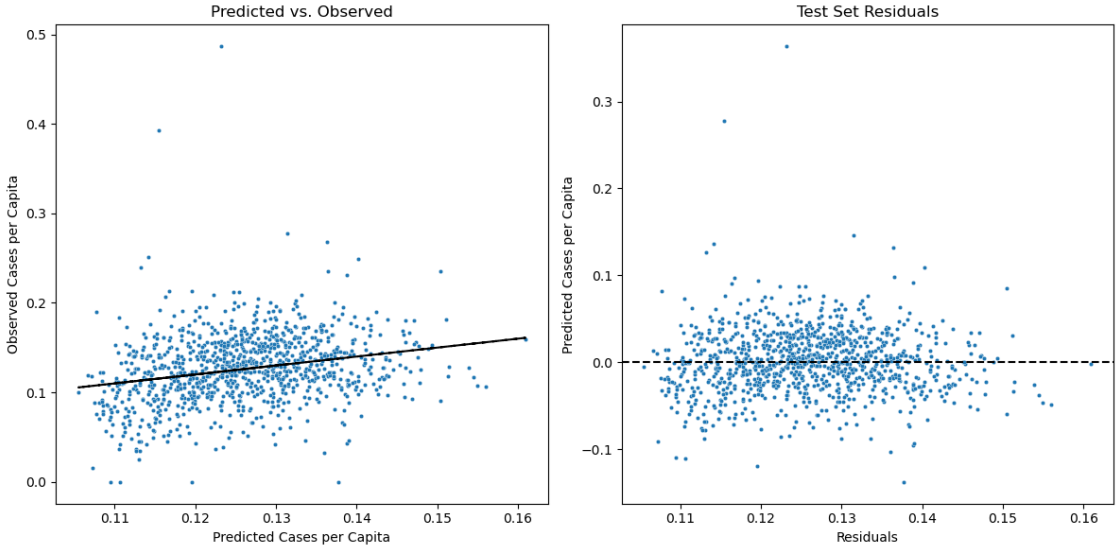
To visualize the model performance from part (a), let's make the following two visualizations: 1. The predicted values vs. observed values on the test set, 2. The residuals plot. (Note: in multiple linear regression, the residual plot has the residuals plotted against the predicted values).

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**. * Please add descriptive titles and axis labels for your plots!

In [36]:
```
plt.figure(figsize=(12,6))      # do not change this line
plt.subplot(121)                # do not change this line
# 1. plot predictions vs. observations
sns.scatterplot(x = Y_test_pred, y = Y_test, s=10)
plt.plot(Y_test_pred, Y_test_pred, c='k', ls='--')
plt.title('Predicted vs. Observed')
plt.xlabel('Predicted Cases per Capita')
plt.ylabel('Observed Cases per Capita')

plt.subplot(122)
# 2. plot residual plot
sns.scatterplot(x=Y_test_pred, y=(Y_test - Y_test_pred), s=10)
plt.axhline(y=0, c='k', ls='--')
plt.title('Test Set Residuals')
plt.xlabel('Residuals')
plt.ylabel('Predicted Cases per Capita')

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


Predicted vs. Observed
Observed Cases per Capita
Predicted Cases per Capita
0.0
0.1
0.2
0.3
0.4
0.5
0.11
0.12
0.13
0.14
0.15
0.16
Test Set Residuals
Predicted Cases per Capita
Residuals
−0.1
0.0
0.1
0.2
0.3

0.0.5 Question 2c

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

The first plot shows that the model is neither overpredicting or underpredicting, but there's a relatively high variability as most of the predicted values are ~0.1 away from the actual values. We can observe a trend in the residual plot, where the predicted residual points closely follow the distribution of the actual values. This indicate that the linear model might not be appropriate in make predictions.

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. What does this indicate about our data and our model?

Describe a reason why this could be happening.

Hint: Take a look at the design matrix, heatmap, and response from Question 1!

The confidence intervals, which range from around -1 to 2.7, indicate that some characteristics of mask usage would not significantly affect the frequency of COVID-19 cases. The heat map shows how each one is connected to the others. Due to the fact that one of Cheese's variables cannot be changed while the others remain constant, this collinearity makes it difficult for us to understand how the variables are actually related to the prediction.

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:

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

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

Because variance only contributes to less than 0.5%, we conclude that the variance is not the dominant affecting the model's performance. The primary factor affecting the model are its bias and the irreducible error.

0.0.8 Question 4d

Propose a solution to reduce the mean square error using the insights gained from the bias-variance decomposition above.

Assume that the standard bias-variance decomposition used in Lecture 19 can be applied here.

The MSE is a combination of the model's variance and the square of its bias. Given that the variance contributes relatively less to the MSE, it means that model bias contributes more. Therefore, we can reduce the model bias to reduce MSE.