Week 5

Advay Vyas

5/6/25

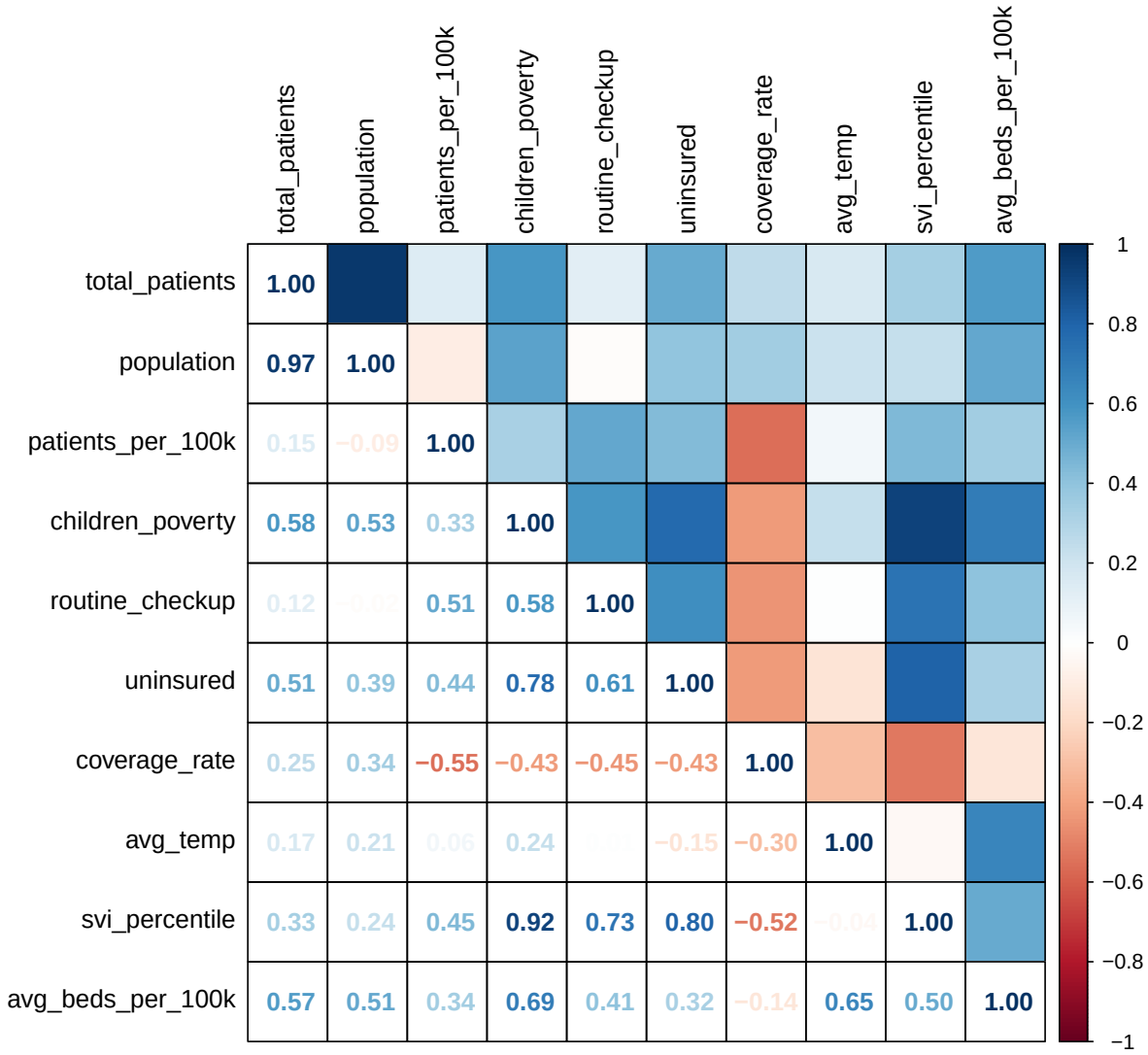
Contents

Introduction

This week, we are going to conduct multiple linear regression and try to incorporate interaction variables and ANOVA for some additional modeling boosts. The (tentative) indicator variables are flu vaccination coverage rate and uninsured % (from Week 4), temperature (from Week 3), Social Vulnerability Index (SVI) and beds per capita (from Week 2). Since this is my last week until the break ends, I wanted to conduct a sort of final project that would try to tie everything together and hopefully show some real trends from the data that I've web scraped and cleaned.

A new correlation matrix

I've summed up the previous weeks (beds from weeks 1-2, svi from week 2, city health from week 3-4, flu vaccination from week 3-4, and temperature from week 3) into one large correlation matrix. We are primarily interested in correlations with the patients variable and we see that coverage_rate, routine_checkup, uninsured, SVI percentile, and avg_beds seem to have a significant correlation. We know from last week that the insurance correlation isn't really substantive, so let's move on and ignore that when we build our linear regression.



	total_patients	population	patients_per_100k	children_poverty	routine_checkup	uninsured	coverage_rate	avg_temp	svi_percentile	avg_beds_per_100k
total_patients	1.00									
population	0.97	1.00								
patients_per_100k	0.15	−0.09	1.00							
children_poverty	0.58	0.53	0.33	1.00						
routine_checkup	0.12	−0.02	0.51	0.58	1.00					
uninsured	0.51	0.39	0.44	0.78	0.61	1.00				
coverage_rate	0.25	0.34	−0.55	−0.43	−0.45	−0.43	1.00			
avg_temp	0.17	0.21	0.06	0.24	0.03	−0.15	−0.30	1.00		
svi_percentile	0.33	0.24	0.45	0.92	0.73	0.80	−0.52	−0.04	1.00	
avg_beds_per_100k	0.57	0.51	0.34	0.69	0.41	0.32	−0.14	0.65	0.50	1.00

The new model

A quick aside: due to having a very small dataset (only 8 variables), the 95% confidence intervals for a lot of statistics will include 0 (be statistically insignificant). It's important to note the trends here instead of statistical significance since that is hard to nail down with a small sample like ours.

Adding flu vaccination coverage rate

Table 1: Model 1

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	141.407	53.000	2.668	0.037	11.720	271.093
coverage_rate	-2.196	1.352	-1.625	0.155	-5.504	1.112

Our first model is simply between patients and flu vaccination coverage rate, and our estimate here is -2.196, so with every percent increase in coverage_rate, the amount of flu patients per 100,000 decreases by about 2. Looks good! We also get an R^2 value of 0.3055, which is good for our first variable.

Adding routine checkup coverage percentage

Table 2: Model 2

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-39.473	217.555	-0.181	0.863	-598.718	519.771
coverage_rate	-1.601	1.546	-1.035	0.348	-5.576	2.374
routine_checkup	2.189	2.550	0.858	0.430	-4.365	8.743

Next, we had the routine_checkup variable which cuts the coverage_rate variable estimate slightly and the intercept quite a lot. Still statistically insignificant, and this time with a "more centered about 0" 95% CI. However, our R^2 value stayed quite similar, jumping up by about 0.1 to 0.3947.

Adding temperature data

Table 3: Model 3

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-9.998	288.909	-0.035	0.974	-812.138	792.143
coverage_rate	-1.716	1.828	-0.939	0.401	-6.791	3.359
routine_checkup	2.106	2.872	0.733	0.504	-5.867	10.080
avg_temp	-0.276	1.476	-0.187	0.861	-4.374	3.821

The temperature data brings the intercept closer to 0, leaves coverage_rate and routine_checkup virtually unchanged and contributes a tiny effect per degree in Fahrenheit. The R^2 value reflects this, with a virtually identical 0.4.

Adding SVI data

Table 4: Model 4

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-17.018	355.575	-0.048	0.965	-1148.617	1114.580
coverage_rate	-1.763	2.263	-0.779	0.493	-8.965	5.439
routine_checkup	2.264	4.320	0.524	0.637	-11.485	16.013
avg_temp	-0.295	1.734	-0.170	0.876	-5.815	5.225
svi_percentile	-1.779	31.308	-0.057	0.958	-101.416	97.858

Next, the social vulnerability index percentile is added to the linear regression. While the estimate changes slightly to accommodate for this new variable, we can see that since SVI itself is on a scale from 0 to 1, it really has no substantial effect. Our R^2 value is henceforth 0.4006.

Adding beds per 100,000

Table 5: Model 5

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	653.503	370.954	1.762	0.220	-942.582	2249.587
coverage_rate	-6.241	2.434	-2.564	0.124	-16.713	4.232
routine_checkup	0.896	2.839	0.316	0.782	-11.321	13.114
avg_temp	-6.470	2.913	-2.221	0.156	-19.003	6.063
svi_percentile	-69.484	35.713	-1.946	0.191	-223.147	84.178
avg_beds_per_100k	0.240	0.105	2.295	0.149	-0.210	0.691

While adding beds per 100,000, now we are getting somewhere! This variable shakes up a lot of the previous variables and looks to be incredibly predictive of the flu patients per 100,000 (about every 5 beds to a new patient, roughly). The R^2 value takes a leap towards 0.835!

ANOVA

R^2 and adjusted R^2

Table 6: R^2 table

Variable	R^2	Adjusted R^2
coverage_rate	0.3055	0.1898
routine_checkup	0.3947	0.1526
avg_temp	0.4000	-0.0500
svi_percentile	0.4006	-0.3986
avg_beds_per_100k	0.8350	0.4225

The above table shows how R^2 value develops over each variable being added the difference between R^2 and adjusted R^2, which takes into consideration value relative to the sample size. Therefore, we can see a clear penalty assessed to the addition of temperature and SVI - they aren't useful. Meanwhile, we see a boost from coverage_rate and beds followed by routine_checkup as well. It's important to note here that while our original R^2 values showed that coverage_rate had an "impact" of about 0.3, the adjusted values say it is slightly less; vice versa for routine_checkups, albeit small.

η^2

Table 7: η^2 table

Variable	η^2	Confidence	Lower	Upper
coverage_rate	0.3055	0.95	0	1
routine_checkup	0.0892	0.95	0	1
avg_temp	0.0053	0.95	0	1
svi_percentile	0.0006	0.95	0	1
avg_beds_per_100k	0.4344	0.95	0	1

In this table above, we summarize the effect size changes (η^2) and this is basically the R^2 except it isn't cumulative. The same variables look prominent here as well. We are also assuming (I think) that the confidence interval including 0 is due to a lack in sample size to predict on.

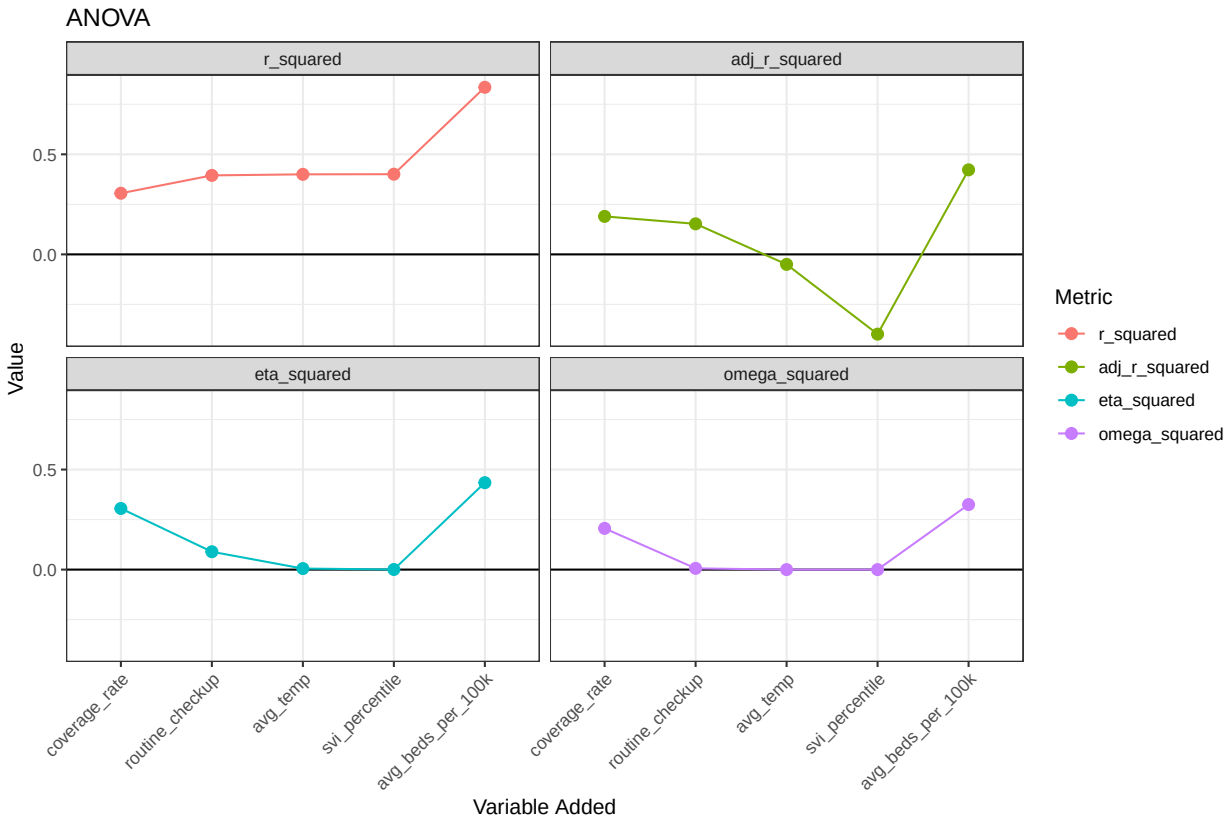
ω^2

Table 8: ω^2 table

Variable	ω^2	Confidence	Lower	Upper
coverage_rate	0.2060	0.95	0	1
routine_checkup	0.0062	0.95	0	1
avg_temp	0.0000	0.95	0	1
svi_percentile	0.0000	0.95	0	1
avg_beds_per_100k	0.3250	0.95	0	1

In this table, we find the ω^2 values which is a different kind of effect size metric that adjusts for bias and is more conservative. We see that clearly reflected in the values, where avg_beds_per_100k leads the pack followed by coverage_rate and a nonexistent routine_checkup contribution.

Graph of effect size metrics by variable added



Conclusion

Week 5

The important variables I think we have are avg_beds_per_100k, coverage_rate, and routine_checkup (in that order). I think the dashboard already has the coverage rate included so maybe including the hospital beds in that city and their routine checkup % could help fine tune each city's individual predictions. It's important to keep in mind this statistical analysis was inherently narrow - I only looked at 8 different Texan cities at one specific quarter of 2022 (the flu season "spike") and collected data based off of that. Still, I think my findings could be useful in some context and I hope they are!

Looking forward

What a ride! For how short this was, it was honestly great. Thank you so much to the Meyers Lab and my mentor Dr. Dongah Kim! I'm excited to start working with more advanced tools in the fall and contributing to that flu patient dashboard.