Technical Appendix: The Limitations of Regression Models

by Carolina Reid, PhD, Faculty Research Advisor

March 2020

TERNER CENTER FOR HOUSING INNOVATION
UC BERKELEY

In the two reports we released today, *The Hard Costs of Construction: Recent Trends in Labor and Materials Costs for Apartment Buildings in California* and *The Costs of Affordable Housing Production: Insights from California's 9% Low Income Housing Tax Credit Program,* we present regression models that lend insights into what factors are associated with the high costs of development in California. The strength of a regression model is that it allows us to better compare apples to apples – how do two projects in the same place with similar characteristics differ on costs, and what accounts for those differences? Because the costs of development are so dependent on when, where, and how a building is built, it is important to try and minimize those differences in exploring what is happening with costs in California.

However, all regression models are limited across two important dimensions. The first is known as "unobservable" or "omitted" variable bias. We don't have information on every difference between two projects – for example, we don't know if one project was located on a site with a gas tank underground that needed to be removed, or if it entailed a CEQA challenge, or if there was a major change order in the middle of construction (all circumstances that our interviews pointed to as important drivers of costs). In this way, our models are just partial explanations of total development costs.

The second problem is "endogeneity." This refers to the causal direction of the variables included in the model. For example, we find that elevators are associated with higher development costs, but it could be that projects that are more expensive require elevators: they could be on a small, oddly shaped infill lot that requires a taller building and thus an elevator. In this way, endogeneity and omitted variables work together – if we can't fully control for whether a property is on an oddly shaped infill lot, we may incorrectly attribute costs to the elevator when the elevator is just a byproduct of the design of the building and where it is located.

These limitations are exacerbated when you have a small sample, because to build an accurate regression model, you need multiple projects that vary along all the dimensions you're testing. So for example, if you only have one senior project in San Francisco that has prevailing wage, an elevator, was awarded tax credits in 2016, and included sustainable building techniques and had impact fees, there's no other similar project to compare it with to establish whether another characteristic – e.g., the number of sources –increases the costs of development. Because our sample is relatively small in statistical terms—we only have 626 projects for which we have data on all the variables in our model—we are limited in how many different dimensions of each project we can test at the same time. If you keep adding variables to a model, you will only by virtue of the small sample size make certain characteristics insignificant, even when they do matter. This is why researchers use other diagnostics to determine model fit and ensure that the final model is the "best" fit for the data, even if it is partial or incomplete.

We have tried throughout the report to be transparent about the limitations of the analysis, and our main recommendations relate to the additional research and data analysis that is needed to identify the right set of policy options that can reduce costs and at the same time produce high quality affordable housing. At the Terner Center, one value we hold very strongly is data

informed decision-making and transparency in our analysis, as well as the fact that all research is partial and can be improved upon.

That said, one area where there is considerable methodological debate regarding the correct way to model development costs is around prevailing wage. We should say at the outset that **we absolutely agree that we cannot say that prevailing wages are the *cause* of higher development costs.** But we also believe that our research demonstrates a relationship between projects that require prevailing wage and higher costs based on all the models we ran, as well as on the qualitative data we collected from interviews with affordable housing developers, construction managers, and general contractors. This does not mean we are in favor of removing prevailing wages – as we point out in the reports, paying living wages and investing in a trained, unionized labor force are policy choices with public benefits, as are decisions to invest in projects in higher opportunity neighborhoods or subsidize more sustainable building techniques to mitigate climate change.

To bring more transparency to this issue, we present here a series of robustness checks related to the model we present in *The Costs of Affordable Housing Production: Insights from California's 9% Low Income Housing Tax Credit Program*. Building a model entails making choices – we seek here to make the reasoning for our choices transparent, even though of course other researchers may make different determinations. It also gives us an opportunity to talk about important areas for additional data and research needs.

Decision #1: Our study focuses on "total development costs", as opposed to focusing on construction costs specifically and removing land from the total project costs.

Because we are interested in the overall costs of development, and what is driving up the amount of subsidy that is needed for each unit of affordable housing, we made the decision to focus on total development costs and not just construction costs in this paper. However, the way land is treated in LIHTC 9% applications varies, meaning that our models may be capturing variation due to this land variable as opposed to real differences in cost drivers.

As a robustness check, we re-ran our model excluding land acquisition costs from the cost per unit metric. (Table A.1). We find similar results on prevailing wage – the model suggests that projects with prevailing wages still have higher costs, even after taking out land costs. However, other interesting differences emerge – in particular, we find that the "High Poverty and Segregation" variable loses significance, perhaps lending credence to the idea that it is land and not construction costs that influence the cost of building in a higher income neighborhood. Impact fees and the number of sources also become insignificant, suggesting that these components of costs are more important to overall development costs than to hard costs alone.

Table A.1: Does it Make a Difference if we Take Land Acquisition out of Development Costs?

Dependent variable: Development Costs Excluding Land Acquisition

Variables	$2019	
Project Size (Number of Units)	-1,199	***
Year Awarded Funding (Compared to Projects Built in 2008 and 2009)		
2010 to 2014	-9,742	
2015 to 2019	72,797	***
Type of Development (Compared to Senior Projects)		
Permanent Supportive Housing	33,030	**
Family Housing	87,376	***
Geography (Compared to Inland California)		
Bay Area	99,270	***
Los Angeles	16,214	*
Rural Counties	306	
Opportunity Category (Compared to other Opportunity Categories)		
High Poverty and Segregation Tract	736	
Project Characteristics		
Project Includes Prevailing Wage	51,383	***
Project Includes Structural Parking	51,296	***
Project Includes Elevator	44,125	***
Project Includes Sustainable Building Materials	7,530	
Project Includes Development Fees	11,844	
Each Funding Source	1,103	
Intercept	228,747	***
Adjusted R-squared	0.444	
N	609	

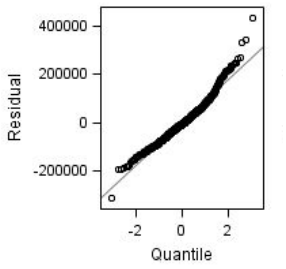
Source: California LIHTC 9% Projects, 2008 – 2019. All dollar amounts adjusted for inflation.
Notes: *** p < 0.001, ** p < .01, * p < .10 (indicates the significance of the result – estimates without stars are not significantly different from the comparison group).

A bigger concern for us in using this suggested variable as our outcome variable is that our model fit diagnostics are worse. While the Adjusted R-squared is one measure of fit (in the broadest terms, this model explains less (44.4%) of the variation in costs than our model of total development costs (52.6%), Figure A.1 shows that the model in the paper does a better (if

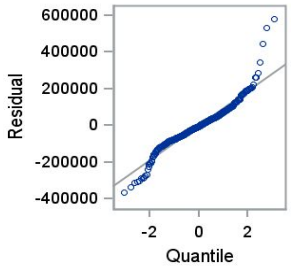
imperfect) job of identifying the factors associated with low and high cost projects, with larger residuals especially at the tails of the distribution.

Figure A.1: Fit Diagnostics for Total Development Costs per Unit v. Excluding Land Costs

Total Development Costs per Unit



Total Development Costs Excluding Land Costs per Unit



Decision #2: Rather than including a separate control for each year a project was awarded funding, we rely on aggregate 3-year dummies to capture variation over time.

Certainly, projects costs change over time, and there's a lot of variation each year in the factors that might be influencing development costs that fall into that category of "unobservables" or variables we don't have in our datasets. We control for the changing value of money—in other words, inflation--using the CPI-All Urban Consumers inflation factors prior to running any descriptive statistics and our model. While there are construction cost inflation indices available, we chose not to use those because part of our goal is to show that construction costs have gone up faster than general inflation – in effect, using construction cost inflators will "subtract away" the rising cost of construction, precisely what we're trying to understand.

The second way researchers address unobserved time trends is by adding a variable for every year in the dataset that controls for things that happened that year – these are referred to as year "dummies" or year "fixed effects." When we began our analysis, we included a dummy for each year. This is the "ideal" model. The problem is, the more variables you add, the more likely it is that you won't find enough projects to compare within that year that have all the other characteristics we're interested in. Our model fit diagnostics get worse, rather than better, by adding more variables. This is why we decided to group together projects into multiple, similar year increments, to still capture time trends but reduce the overall number of variables in the model.

In Table A.2, we present our model with full year controls. It shows that many of the year dummies are insignificant, and that including these variables reduces our ability to understand what is happening with supportive housing (since there are fewer supportive projects in every year, we lose that needed variation). Yet overall, the findings are consistent with the model we

include in the paper, although the magnitude of effects and significance for some of the variables change.

Table A.2: Model Results including Individual Year Dummies

Dependent variable: Total Development Costs		
Variables	**$2019**	
Project Size (Number of Units)	-1,149	***
Year Awarded Funding (Compared to Projects Built in 2008)		
2009	940	
2010	7,204	
2011	-30,055	
2012	-55,528	**
2013	-37,378	*
2014	-9,776	
2015	22,811	
2016	71,443	***
2017	32,916	
2018	60,894	**
2019	59,353	**
Type of Development (Compared to Senior Projects)		
Permanent Supportive Housing	18,673	
Family Housing	85,388	***
Geography (Compared to Inland California)		
Bay Area	144,117	***
Los Angeles	67,295	***
Rural Counties	-16,737	
Opportunity Category (Compared to other Opportunity Categories)		
High Poverty and Segregation Tract	-14,803	*
Project Characteristics		
Project Includes Prevailing Wage	55,733	***
Project Includes Structural Parking	43,671	***
Project Includes Elevator	21,917	*
Project Includes Sustainable Building Materials	26,402	***

Project Includes Development Fees	11,401	
Each Funding Source	5,476	**
Intercept	228,747	***
Adjusted R-squared	0.5531	
N	626	

Source: California LIHTC 9% Projects, 2008 – 2019. All dollar amounts adjusted for inflation.
Notes: *** p < 0.001, ** p < .01, * p < .10 (indicates the significance of the result – estimates without stars are not significantly different from the comparison group).

We faced a similar constraint with our geographic controls. Ideally, we would include a dummy for every county or TCAC region, to control for unobservables across places. (In the best case scenario, we would control for differences at the census tract level.) But that requires a much larger sample. This is why we grouped TCAC's regions into four broader categories – the Bay Area, the Los Angeles coastal region, rural counties, and then the state's inland areas. In Table A.3, we present the full model results using both year and TCAC region dummies, for both the Total Development Costs and Total Development Costs minus land costs.

Again, variables change in magnitude and significance, showing how dependent a model is on what you include. It also shows that more variables don't necessarily lead to better model fit, as our adjusted R-squared value is lower than in the model we report in the paper. (The adjusted R-squared is just one metric of model fit, and in our opinion, one of the least important. We also tested for multicollinearity and heteroscedasticity of error terms in all of our model building.)

Table A.3: Model Results including Individual Year Dummies and TCAC Regions

Dependent variable: Total Development Costs	**Dependent variable: Total Development Costs**		**Dependent variable: Total Development Costs Excluding Land**	
Variables	**$2019**			
Project Size (Number of Units)	-992	***	-1,102	***
Year Awarded Funding (Compared to Projects Built in 2008)				
2009	1,389		-39,269	*
2010	-1,093		-27,404	
2011	-24,045		-42,577	*
2012	-54,038	*	-55,177	*
2013	-32,449		-46,777	*
2014	-16,615		-35,025	
2015	17,713		32,414	
2016	66,169	**	58,135	*
2017	27,693		11,420	
2018	52,876	**	54,447	*

2019	50,750	*	63,681	**
Type of Development (Compared to Senior Projects)				
Permanent Supportive Housing	22,598	*	33,173	**
Family Housing	89,296	***	83,956	***
TCAC Region (Compared to San Francisco)				
Capital North	-58,471	**	-28,084	
Central Coast	-28,711		-20,862	
Central	-109,088	***	-81,705	***
Inland Empire	-41,558	*	-26,489	
Los Angeles	2,587		-25,922	
North East Bay	65,521	***	58,366	**
Orange	-15,161		-34,377	
San Diego	6,022		-28,121	
South West Bay	64,196	**	45,146	*
Opportunity Category (Compared to other Opportunity Categories)				
High Poverty and Segregation Tract	-8,570		7,265	
Project Characteristics				
Project Includes Prevailing Wage	53,574	***	52,539	***
Project Includes Structural Parking	42,747	***	54,251	***
Project Includes Elevator	26,653	**	34,586	**
Project Includes Sustainable Building Materials	21,519	**	10,019	
Project Includes Development Fees	13,005		9,123	
Each Funding Source	6,561	***	605	
Intercept	324,323	***	302,853	***
Adjusted R-squared	0.5174		0.4484	
N	626		609	

Source: California LIHTC 9% Projects, 2008 – 2019. All dollar amounts adjusted for inflation.
Notes: *** p < 0.001, ** p < .01, * p < .10 (indicates the significance of the result – estimates without stars are not significantly different from the comparison group).

Decision #3: The model omits important variables that can influence costs, such as project area "market" wages, architecture and engineering costs per square foot, and developer type.

There are a lot of additional variables we would like to control for in our models, and we continue to enter and clean data to be able to extend this analysis in future reports. For example, we are very curious to understand which environmental requirements add most to costs – is it materials like bamboo flooring, or is it energy conservation measures? Right now, we treat all sustainable building techniques equally, which is far from ideal. Similarly, we'd like to understand which cities place the most onerous parking requirements on their affordable projects, even when those projects are located in neighborhoods well served by public transit.

While data on average area market wages are available at the county level, we chose not to include them here because of a similar concern about endogeneity – depending on how many jobs are shop jobs in an area, the average construction wage will be influenced by the union wage. In other words, the average market wage in an area is not independent from the prevailing wage variable. Regarding developer type and architecture and engineering costs, these are variables we are still cleaning and do not have confidence in the data quality to present in our current paper.

Decision #4: Econometric techniques, such as propensity score modelling, can help to address concerns over endogeneity, and may be better specifications than the linear regression model presented in the paper.

Recognizing that establishing causality is difficult, particularly in simple linear regression models, economists have developed new statistical techniques to try and establish stronger causal links between various input and outcomes of interest. One method is called "propensity score matching (PSM)." While it is a popular technique, it is very sensitive to the approach used to "match" observations, and is falling out of disfavor with many economists. While the statistical reasons are too detailed to get into here, a simple way of thinking about it is that if you're limiting your analysis to projects that are perfectly matched (or weighting the data accordingly), you're changing the sample so much that you're no longer actually modelling the messy reality you're trying to understand.[1]

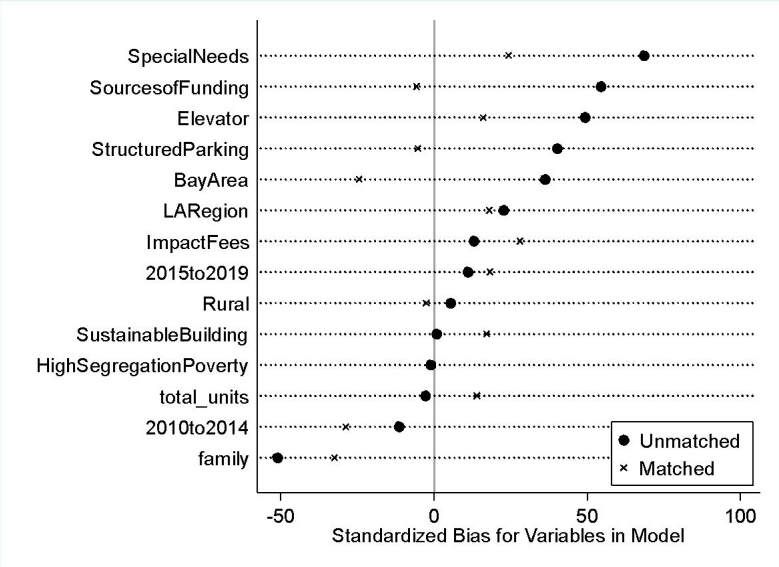
As we note in the introduction, we are not trying to establish causality with our models, but rather, explore the relationships between various aspects of development projects and costs. Still, we thought it would be interesting to explore how a PSM model trying to identify differences in prevailing wage versus non prevailing wage projects might lead to different results than a simple linear regression.

Importantly, it is true that prevailing wage projects are different from non-prevailing wage projects. Figure A.2 presents a visual representation of this. The dark circles show that in the

[1] This has been referred to as the PSM paradox: if one's data are so imbalanced that making valid causal inferences from it without heavy modeling assumptions is impossible, then PSM will reduce imbalances but the resulting data are not very useful for causal inference by any method. https://gking.harvard.edu/files/gking/files/psnot.pdf

"unmatched" data, prevailing wage projects are much more common when the project is supportive housing or has more sources of funding. The figure also shows that while matching improves the comparability of prevailing wage and non-prevailing wage projects (The x marks are closer to the 0 line than the dark circles), there is still some bias in the matched pairs. It also shows the limitations of a small sample – the fact that it is hard to balance matched pairs for the Bay Area, for example, is because we don't have enough similar projects within the Bay Area that we can find two projects that are similar on all dimensions except for prevailing wage.

Figure A.2: Bias in Matched Pairs, Prevailing v. Non-Prevailing Wage Projects



Source: California LIHTC 9% Projects, 2008 – 2019.

The modelling yields interesting insights into prevailing wage projects. Prevailing wage is much more common on supportive and special needs housing than on family and senior units; in addition, the likelihood that prevailing wage is required goes up if a project has more sources of funding. As Table A.4 shows, other variables, such as time period or other project characteristics—other than the presence of an elevator—do not have a significant association with whether a project is prevailing wage. Region does matter however; Los Angeles and the East and West Bay are similar to San Francisco when it comes to requiring prevailing wage projects, while other parts of the state are less likely to require prevailing wage on LIHTC 9% projects. It also points to how tricky it is to disentangle causality: these coastal areas are also the most expensive, so they may need more sources of funding, which may in turn trigger prevailing wage.

Table A.4: Logistic Regression Predicting the Likelihood that a Project Requires Prevailing Wage

	Coefficient on Likelihood Project is Prevailing Wage	
Intercept	-0.8625	
Total Units	-0.00122	
Project Type		
Supportive Housing	2.0777	***
Family Housing	0.1311	
Number of Sources	0.3046	***
Year of Credit Award		
2009	0.3229	
2010	-0.6135	
2011	0.0871	
2012	0.3948	
2013	-0.4265	
2014	-0.2536	
2015	-0.2632	
2016	-0.6476	
2017	-0.7031	
2018	-0.7888	
2019	0.4504	
TCAC Region (Omitted: San Francisco)		
Capital North	-1.5172	**
Central Coast	-0.8025	*
Central	-1.2876	**
Inland Empire	-1.7881	***
Los Angeles	-0.5582	
North East Bay	-0.3916	
Orange	-2.0961	***
San Diego	-1.4835	**
South West Bay	-0.0559	
Project Characteristics		
High Poverty and Segregation Tract	-0.0178	
Project Includes Development Fees	0.1055	
Project Includes Structural Parking	0.2125	
Project Includes Elevator	0.7477	**
Project Includes Sustainable Building Materials	-0.00067	

It is important to note that the model with all of these variables violates some of the assumptions of PSM, including balancing requirements and the conditions of "common support." Instead, we're going to use the specification of the simplified model in the report in our PSM specifications, collapsing the year and geography dummies into categories (with the exception of model e in Table A.5 which includes the full list of variables).

There are multiple ways of modelling propensity score matching, each is going to give different results. In Table A.5, we present the results from multiple different modelling approaches, showing the estimated effect of prevailing wage on per unit development costs using these different approaches.

Table A.5: Estimates of Impact of Prevailing Wage using Propensity Score Matching

Matching Method	Effect of Prevailing Wage on Per Unit Development Costs	
a) Nearest Neighbor	36,472	*
b) Radius Matching	77,926	***
c) Kernel Matching	51,135	***
d) Stratification Matching	53,880	***
e) teffects with common support (extended model with all year and region dummies)	66,077	***
f) teffects with common support (simplified model with bucketed year and regions)	39,291	*

All of these models present different dollar amounts for the relationship between prevailing wage and total development costs. This points to how difficult it is to estimate the "right" answer – it really depends on what you put in a model and how you choose to execute it. This is why we refrain from making a claim about causality when it comes to any single variable in our models – we're instead seeking to identify associations between variables and total development costs. However, the fact that all of these specifications show at least a positive correlation between prevailing wage and higher costs leads us to have confidence in the general relationship we report in our papers.

Conclusion

Our hope by providing this appendix is to show that modelling is an imperfect science, and that it is possible to make lots of modelling decisions that can change the outcomes of an analysis. By being transparent about our choices and why, we hope to contribute to greater transparency in what goes into development costs. It is important to emphasize that these checks are not meant to prove that prevailing wage has a causal impact on development costs – our honest assessment is that with the data we have and the techniques that we have used, it is not possible to make a causal determination. We also want to reiterate that the 9% LIHTC program is just one subset of affordable housing developments; in all research, what is in your sample, including what is included and what years it covers, is going to influence the outcomes of the analysis. This is why our estimates in our two reports also vary from one another – they are based on different years and different samples of properties.

That said, we stand by our analysis and findings. We hope that policymakers and other stakeholders will use our research as it was intended: to begin a dialogue about how we could reduce costs that will allow us to build more high quality affordable housing while not detracting from our policy goals of economic stability and environmental sustainability, and to point to the value of using data rather than ideology to drive policy decisions.