[Re] A New Approach to Estimating the Production Function for Housing

Aniruddha Dutta, Mansi Rathod, Mayur Gupta, Pradeep Prabhakar, Sreeja Vishaly Manohar

14 March, 2020

A reference implementation of:

Epple, Dennis, Brett Gordon, and Holger Sieg. 2010. *“A New Approach to Estimating the Production Function for Housing.”* American Economic Review, 100 (3): 905-24.DOI: 10.1257/aer.100.3.905

# Introduction

The purpose of the paper is to do a replication of the paper “A New Approach to Estimating the Production Function for Housing” (Epple, Gordon, and Sieg [2010](#ref-epple2010new)) that estimates the production function for housing. Based on two factors - the observed variation in land prices and housing values per unit of land, the original paper has provided an algorithm to identify the housing supply function per unit of land. This in turn is used to derive an estimate of the underlying production function. The approach yielded plausible estimates for the price elasticity of housing supply per unit of land, based on data from Allegheny County in Pennsylvania. In the paper,we propose a reference implementation by adding two additional models to replicate an estimation function that relates land price and home value per unit land and produce the corresponding tables and plots.Estimating housing production functions is challenging, as the quantity and price per unit of the housing services are not observed by the econometrician. Replicating the study helps us to understand how the underlying production function is estimated by treating prices and quantities of housing services as latent variables, without relying on strong functional form assumptions.

# Methods

The authors in the paper(Epple, Gordon, and Sieg [2010](#ref-epple2010new)) proposed a flexible estimation procedure based on duality theory to estimate the production functions with minimal functional form assumptions. The underlying assumption is that the production function exhibits constant returns to scale and that there exists a monotonic relationship that captures a relationship at equilibrium between the value of housing per unit of land and the price of land . Thus estimating the function becomes the first step in the process. The parametric functions used for the purpose is OLS and the paper has used multiple transformations on data to produce linear, log linear and polynomial(quadratic and cubic) models to best identify the relationship.

Since, the function forms the base of further analysis and calculation, we decided to replicate the study of these models to test the robustness and the assumptions behind the models. For this purpose, we followed a two-variant approach. The first variant is to use a different model for replication and the second is to use a different loss function to calculate the estimates in the model.

For the first case, we decided to use Generalized Linear Model with different distribution families. Using different families allows us to verify the condition of normality of error in the models used by authors. We noticed that the gaussian distribution family with log independent and dependent variable produced a line that fitted the best with the data and the corresponding coefficients were very similar to the log linear model used in the paper.For the loss function, we wrote a function to implement log linear regression with gradient descent loss function. The gradient descent loss function provides a more flexible approach because of the presence of hyperparameters *learning rate* and *number of iteration*. Since the approach is not used in the paper, we ran a simulation study in order to determine the hyperparameters that delivered values of coefficients very similar to that of the models in the paper(Epple, Gordon, and Sieg [2010](#ref-epple2010new))

Based on the equilibrium relationship ,the authors proposed that there exists a differential equation that characterizes the normalized housing supply function . For this purpose, the base paper used Ordinary Differential Equations (using ‘odesolve’ package) for calculating the supply function, for parametric, semi-nonparametric and nonparametric estimates of . However, since ‘odesolve’ has been removed from the CRAN repository and the more recent package ‘deSolve’ completely supersedes it, we have used ‘deSolve’ for our replication.

The paper(Epple, Gordon, and Sieg [2010](#ref-epple2010new)) also estimated using instrumental variable estimator suggested by Hausmann, Newey, Ichimura, and Powell (1991) (Hausman et al. [1991](#ref-hausman1991identification)). In this approach,the estimates for higher order polynomials are realized using a linear model. For our replication,we replaced the OLS based HNIP estimator with a bayesian regression function. Bayesian model is used as it is more flexible to further model development and can directly model posteriors of derived/calculated quantities. ‘Bolstad’ package provides a bayes.lm function which has been used to get the estimates for the cubic model. Upon analysis, we can conclude that the replicated estimator produced coefficients that are similar to the actual paper.

## Data

The dataset used by the paper (Epple, Gordon, and Sieg [2010](#ref-epple2010new)) comes from Allegheny County Web site (Pennsylvania state) maintained by the Office of Property assessments, which is based on the appraisals conducted by Sabre Systems and uses land value appraisals as well as property appraisals. Using geocoding, property assignment to travel zones and subsetting only for functioning residential properties, the count came down to 358,677 properties. Of these properties,the authors used a subsample of housing units that were built after 1995 for the estimation procedure which was 6362 houses. The data captures some important metrics for the houses such as the price of land, value per unit land, plot area, travel time and its geo-location (latitude and longitude). We will be using the same data for our replication in this paper. The descriptive statistics of the dataset with 6362 residential properties is shown in Table 1.

#>   
#> Table 1 - Descriptive statistics of Residential data  
#> ===================================================================  
#> Statistic Mean Median St. Dev. Min Max   
#> -------------------------------------------------------------------  
#> Value per unit of land, v 21.44 14.29 26.91 0.15 366.62   
#> Price of land, p 3.32 2.28 3.86 0.05 41.75   
#> Lot area (sq. ft.) 26,756.10 15,507 52,196.96 540 1,207,483  
#> Travel time (minutes) 29.12 30 9.47 1 59   
#> -------------------------------------------------------------------

# Results

For our replication study, we have the results for the following variants introduced in the original paper:  
1. Excluded the top 1 and bottom 1 percentile records from the residential data based on value of unit per land. This is based on the claims by the author that the results were robust to the presence of extreme values.  
2. Added two models to the OLS based estimates- GLM log-linear model and a log-linear model based on gradient descent loss function.  
3. Replicated the supply and production function estimates for generalized log-linear model along with the original models.  
4. Replicated the plots of supply function and production function with 95% CI bands for the generalized log-linear model  
5. Replicated the instrumental variable estimator model(suggested by Hausman, Newey, Ichimura, and Powell (Journal of Econometrics, 1991)(Hausman et al. [1991](#ref-hausman1991identification)) for third order polynomial using Bayesian regression.

### Regression estimates and plots

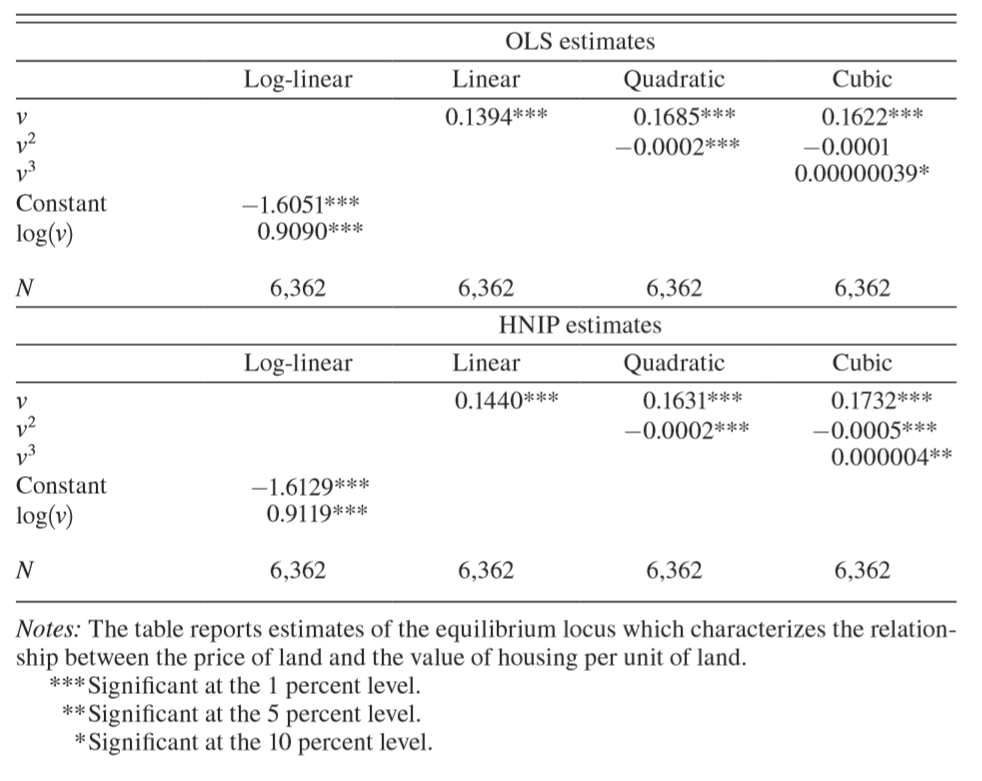


Figure 1: Estimates of equilibrium locus(Epple, Gordon, and Sieg [2010](#ref-epple2010new))

The original paper estimates the r(v) functions using OLS for log-linear, linear, quadratic and cubic models. We obtained similar results using gaussian log-linear model and gradient descent log linear model in addition to the four models. For both of these two models, they fit the main features of the residential data substantially well and their performance is similar to the earlier four models. All the p-values were calculated using heteroskedasticity-robust standard errors and they were significant at 1% level.

#>   
#> Table 2 - Estimates of Equilibrium locus  
#> =============================================================================  
#> OLS and GLM estimates   
#> ----------------------------------------------------------------  
#> Log-Linear Linear Quadratic Cubic Generalized-Log-Linear  
#> -----------------------------------------------------------------------------  
#> logv 0.919\*\*\* 0.919\*\*\*   
#> (0.005) (0.005)   
#>   
#> v 0.153\*\*\* 0.163\*\*\* 0.172\*\*\*   
#> (0.001) (0.001) (0.002)   
#>   
#> v2 -0.0001\*\*\* -0.001\*\*\*   
#> (0.00002) (0.0001)   
#>   
#> v3 0.00000\*\*\*   
#> (0.00000)   
#>   
#> Constant -1.632\*\*\* -1.632\*\*\*   
#> (0.014) (0.014)   
#>   
#> -----------------------------------------------------------------------------  
#> Observations 6,234 6,234 6,234 6,234 6,234   
#> =============================================================================  
#> Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01  
#> Gradient\_descent.Log.linear  
#> Constant -1.631195  
#> log(v) 0.918491

As observed in Table 2, while the base paper had slope to be as 0.909 and the intercept coefficient as -1.605 for the log-linear model in Figure 1, we obtained similar results with (0.919,-1.632) and (0.918,-1.631) as (slope, intercept) for our GLM model and gradient descent models respectively.

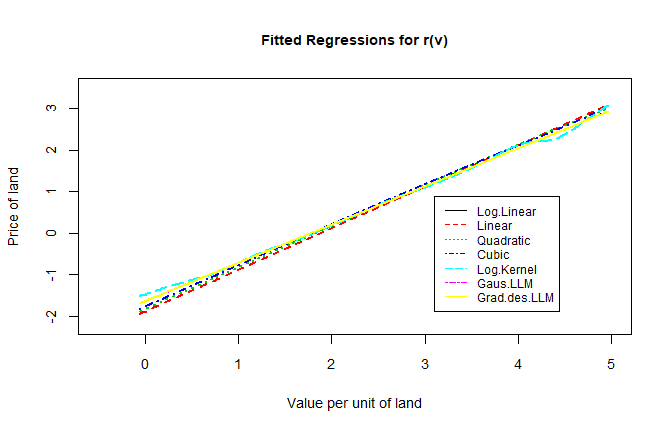


Figure 2: Plot of fitted models

For effective comparison of all the six models used, we have also enclosed a plot as seen in Figure 2. We find all the models to fit the data very well. We were able to establish the monotonicity condition of to be satisfied for the two replication models as well under all the polynomial estimation cases for the range of values of v observed in the data.

### Supply and production function plots

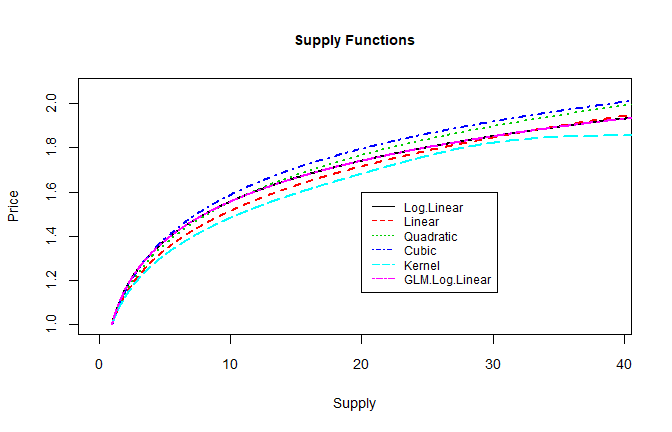


Figure 3: Supply functions by model

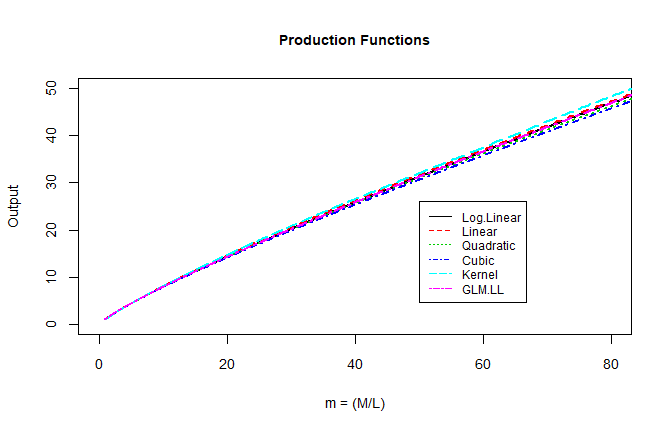


Figure 4: Production functions by model

As seen in Figures 3 and 4, we find that the supply functions and the corresponding production functions of all the base models as well as the GLM log-linear model have similar shapes.

### Supply and production functions with 95% confidence band for GLM Log-linear model

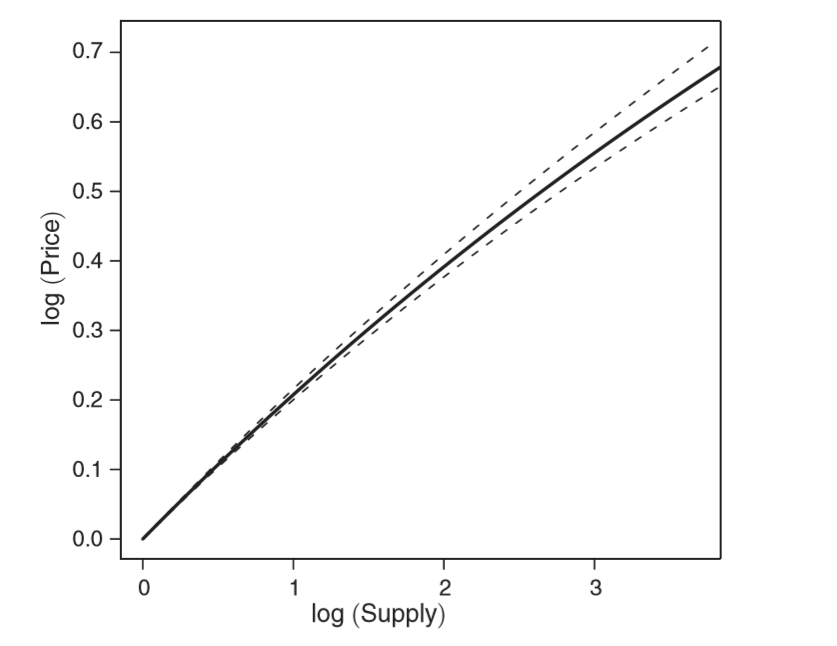


Figure 5: Log-Linear supply function with 95 % confidence band(Epple, Gordon, and Sieg [2010](#ref-epple2010new))

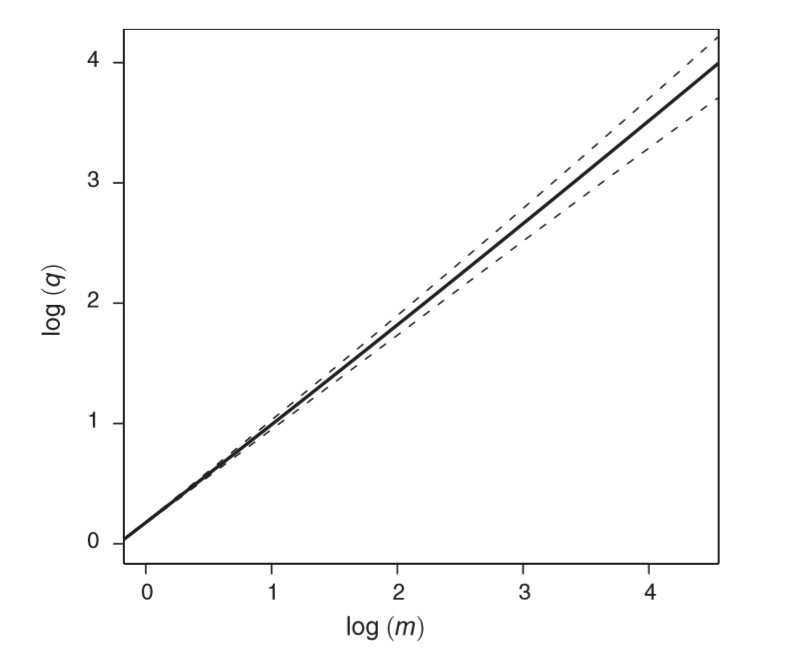


Figure 6: Log-Linear production function with 95 % confidence band(Epple, Gordon, and Sieg [2010](#ref-epple2010new))

Using simulation, standard errors are calculated based on which the corresponding 95% confidence interval bands are estimated for the supply and production functions of GLM Log-Linear model.

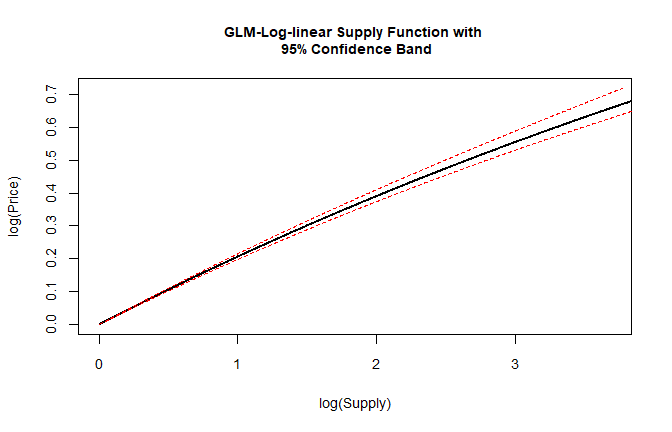


Figure 7: Generalized log-linear supply function with 95 % confidence band

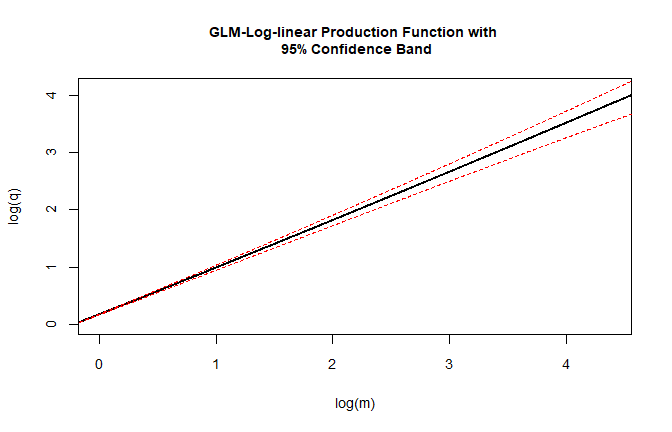


Figure 8: Generalized log-linear production function with 95 % confidence band

As observed, we were able to replicate the supply and production functions with a 95% confidence interval band for GLM log-linear model as seen in Figures 7 and 8 and are similar to the plots produced in the original paper, as seen in Figures 5 and 6 .

### Replication of non-linear regressor estimation using HNIP

#>   
#> Table 3 - HNIP estimates for third order Polynomial  
#> ======================  
#> Cubic mean Cubic SD  
#> ----------------------  
#> v 0.183899 0.013881  
#> v2 -0.000777 0.000283  
#> v3 0.000006 0.000003  
#> ----------------------

The function replicated using bayesian regression for third order polynomial is similar to the one estimated in the original paper using OLS based Linear Regression model. The original paper produced a mean of 0.1732, –0.0005 and 0.000004 for V, V^2 and V^3 coefficients respectively and the replicated bayesian model produced similar means of 0.183899,-7.7710^{-4} and 610^{-6} as seen in Table 3.

# Conclusion

By introducing new models and removing outliers from the Allegheny residential data,we were able to successfully replicate the following key claims from the paper (Epple, Gordon, and Sieg [2010](#ref-epple2010new)) 1) by considering price and quantities as latent variables, the observed variation in land prices and housing values per unit of land is sufficient to identify the housing supply function per unit of land and 2) Given the supply function, the underlying production function for housing can be estimated, without strong functional form assumptions. While the team faced some difficulties during the replication process due to a couple of outdated packages, the codes and the data from the original paper performed as intended and made our replication study a lot easier. Compared to the models used in (Epple, Gordon, and Sieg [2010](#ref-epple2010new)) for estimating the , our Gaussian Generalized Log Linear Model and Gradient Descent Log Linear Model had similar performances and the Generalized log linear model had similar supply and production function estimates.

# References

Epple, Dennis, Brett Gordon, and Holger Sieg. 2010. “A New Approach to Estimating the Production Function for Housing.” *American Economic Review* 100 (3): 905–24.

Hausman, Jerry A, Whitney K Newey, Hidehiko Ichimura, and James L Powell. 1991. “Identification and Estimation of Polynomial Errors-in-Variables Models.” *Journal of Econometrics* 50 (3): 273–95.