# ECO 372: Introduction to Econometrics

Spring 2025

## Lecture 3: Linear Regression with One and Multiple Regressors and Assumptions of Classical Linear Regression


**Sakib Bin Amin, Ph.D.**

Associate Professor in Economics

# Outline

Our objectives will be to explore:

- Introduction to Regression and General Notations
- Population Models Vs. Sample Models
- Linear Regression Model With One Regressor
- The Ordinary Least Squares Estimator (OLS) Estimator: Derivation
- Beyond OLS Derivation
- Example of OLS with One Regressor
- Assumptions for OLS Estimator
- Multivariate Regression
- Example of OLS with Two Regressor

# Introduction to Regression

- The term regression was introduced by Francis Galton
- Regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter
- Although regression analysis deals with the dependence of one variable on other variables, it does not necessarily imply causation
- Closely related to but conceptually very much different from regression analysis is correlation analysis
  - The correlation coefficient measures this strength of (linear) association
  - In regression, we estimate or predict the average value of one variable on the basis of the fixed values of other variables. We may want to know whether we can predict the average score on a statistics examination by knowing a student's score on a mathematics examination

# General Notations

- Before we proceed to a formal analysis of regression theory, let us dwell briefly on the matter of terminology and notation.
- In the literature the terms dependent variable and explanatory variable are described variously.

| **Dependent variable** | **Explanatory variable** |
|---|---|
| ⇕ | ⇕ |
| Explained variable | Independent variable |
| ⇕ | ⇕ |
| Predictand | Predictor |
| ⇕ | ⇕ |
| **Regressand** | **Regressor** |
| ⇕ | ⇕ |
| Response | Stimulus |
| ⇕ | ⇕ |
| Endogenous | Exogenous |
| ⇕ | ⇕ |
| Outcome | Covariate |
| ⇕ | ⇕ |
| Controlled variable | Control variable |

# Population Models Vs. Sample Models

- Regression modelled with population data is know and population regression.
- Population regression, by default should always be consistent.
- However, it is not possible to use the population as data given its volume.
- So, to draw a sample from the population and obtain inference about the population through chosen parameters.
- The sample has to be representative and has to be collected/drawn from the population
- Then we will be needing an estimator that allows us | given the sample to compute estimates for the unknown parameters of the population
  - *Note that An estimator is a function that contains the sample values as arguments*
- Once we have an estimator and observe a sample, we can compute estimates (=numerical values) for the unknown quantities
- To estimate the unknown parameters, there exists different estimators that differ with respect to statistical properties.

# Linear Regression Model With One Regressor

A linear regression model with a single regressor can be written as follows

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Here,

$i$=runs over all observations, $i$={1, 2, 3,…,N}

$Y_i$=dependent variable

$X_i$= independent variable

$\beta_0$= intercept of the regression line

$\beta_1$= slope of the regression line

$u_i$= error term= $Y_i - \beta_0 + \beta_1 X_i$

$\beta_0 + \beta_1 X_i$= regression function

With which method can we estimate the unknown $\beta_0$ and $\beta_1$ ?

# The Ordinary Least Squares Estimator (OLS) Estimator: Derivation

- The Ordinary Least Squares estimator is frequently abbreviated as OLS estimator
- The OLS estimator goes back to C.F. Gauss (1777-1855).
- It is derived by choosing the values of the unknown parameters such that sum of squared residual is minimised.
- The squared of error for ith observation is $u_i^2 = (Y_i - \beta_0 + \beta_1 X_i)^2$. Then sum of squared residual is $\sum_{i=1}^{N}(Y_i - \beta_0 - \beta_1 X_i)^2 = W$
- This univariate optimisation involves taking the derivative and setting equal to 0. We change the notations of the parameters a bit to call them estimated parameters. $\beta_0$ and $\beta_1$ are now $\widehat{\beta_0}$ and $\widehat{\beta_1}$, respectively

$$\frac{\partial W}{\partial \widehat{\beta_0}} = \sum_{i=1}^{N} -2(Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i) = 0 \quad (1)$$

$$\frac{\partial W}{\partial \widehat{\beta_1}} = \sum_{i=1}^{N} -2X_i(Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i) = 0 \quad (2)$$

# OLS Estimator: Derivation

- A bit of algebra from equation (1) will lead us to:

$$\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$$

- From equation (2) we can find:

$$\sum_{i=1}^{N} 2X_iY_i - \widehat{\beta_0}X_i - \widehat{\beta_1}{X_i}^2 = 0$$

- Plug in the value of $\widehat{\beta_0}$ and get:

$$\sum_{i=1}^{N} 2X_iY_i - \left(\bar{Y} - \widehat{\beta_1}\bar{X}\right)X_i - \widehat{\beta_1}{X_i}^2 = 0$$

- After some algebraic manipulation we can find $\widehat{\beta_1}$ as follow:

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

Note: Follow the handout for the complete derivation.

# OLS Estimator: Beyond Derivation

- Generally, we call it predicted value of Y or Y-hat $(\hat{Y}_i)$ & $\hat{Y}_i = \widehat{\beta_0} + \widehat{\beta_1} X_i$. Also known as regression line. The residual $\hat{u}_i$ is the difference between $Y_i$ and $\hat{Y}_i$: $\hat{u}_i = Y_i - \hat{Y}_i$
- When the sample is desirable and OLS assumption (we will see later), then estimated sample parameters will reflect population parameters and sample regression line will coincide (at least reach close if not coinside) with the population regression line.
- ESS - Explained sum of Squares: $\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$
- TSS - Total Sum of Squares (ESS + SSR): $\sum_{i=1}^{N}(Y_i - \bar{Y})^2$
- $R^2$ measures the share of variation in Y due to the variation in X: $R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$
  - Ranges between 0 and 1. In general, the $R^2$ does not take on the extreme value of 0 or 1 but falls somewhere in between.
  - $R^2$ near 1 indicates that the regressor is good at predicting the variable of interest .
  - Adjusted $R^2$ is a corrected goodness of fit : $\text{Adj. } R^2 = 1 - \frac{1-R^2(N-1)}{N-K-1}$.
- Standard deviation of the regression error $\hat{u}_i$: $SER = \sqrt{SSR/_{n-2}}$

# OLS Estimator: Example

- Objective: We want to check whether experience has any effect on wage.
- So, our single regressor linear model can be expressed as following:

$$Wage_i = \alpha_0 + \beta_1 EXP_i + u_i$$

STATA Commands

webuse regsmpl
reg ln_wage ttl_exp

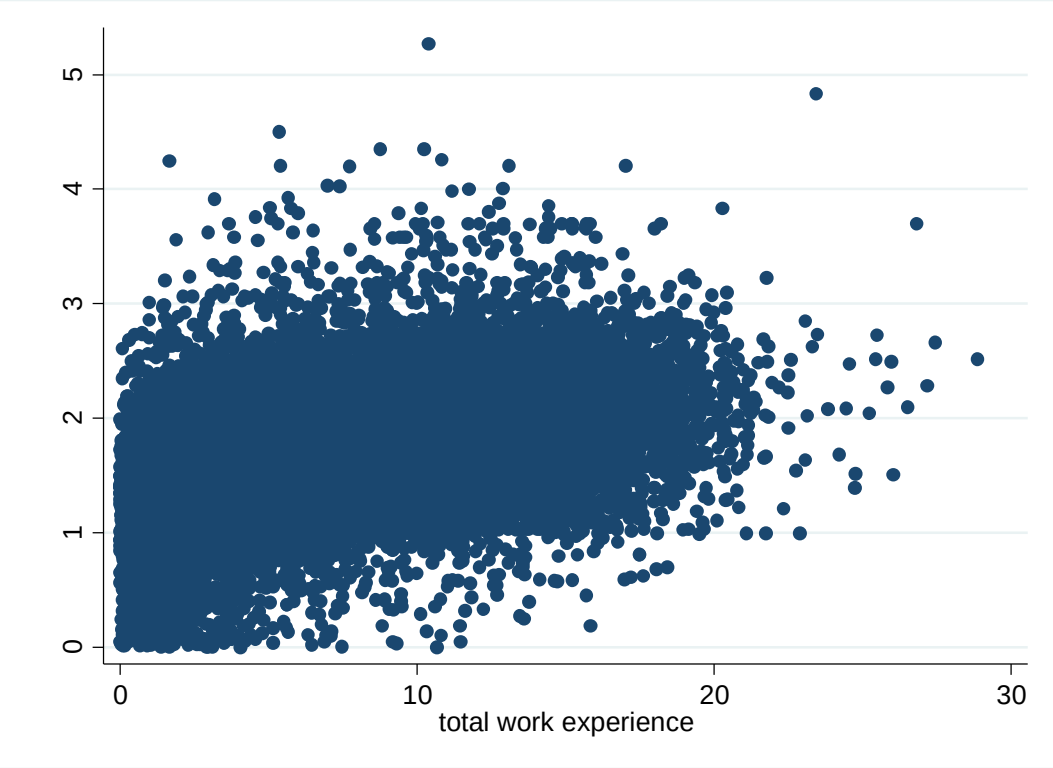
```
. reg ln_wage ttl_exp
```

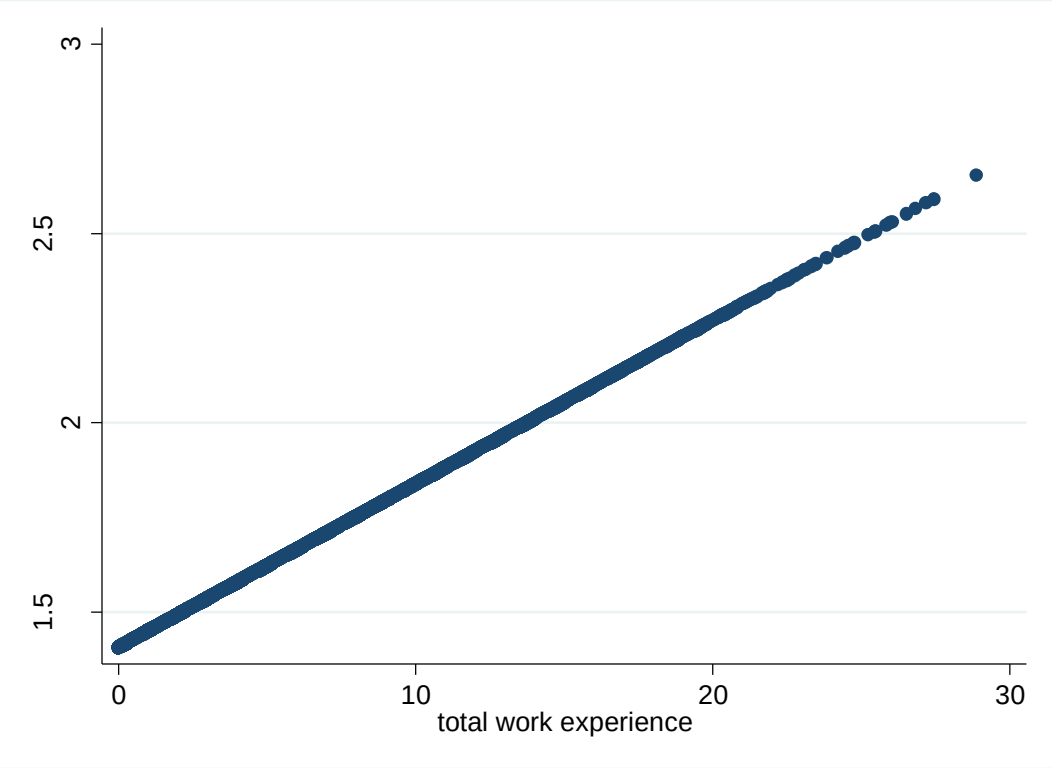
| Source | SS | df | MS | Number of obs | = | 28,534 |
|---|---|---|---|---|---|---|
| | | | | F(1, 28532) | = | 6110.45 |
| Model | 1150.37005 | 1 | 1150.37005 | Prob > F | = | 0.0000 |
| Residual | 5371.51384 | 28,532 | .188262787 | R-squared | = | 0.1764 |
| | | | | Adj R-squared | = | 0.1764 |
| Total | 6521.88388 | 28,533 | .228573367 | Root MSE | = | .43389 |

| ln_wage | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ttl_exp | .0431613 | .0005522 | 78.17 | 0.000 | .042079 | .0442435 |
| _cons | 1.406646 | .0042866 | 328.15 | 0.000 | 1.398244 | 1.415048 |

# OLS Estimator: Example

- Now let us see actual and fitted values side by side:




STATA Commands:

scatter ln_wage ttl_exp

predict ln_wage_hat

twoway (lfit ln_wage_hat ttl_exp ) (scatter ln_wage_hat ttl_exp) or scatter ln_wage_hat ttl_exp

# OLS Estimator: Assumptions

- The OLS model which we have used is follows some assumptions widely known as the Classical Linear Regression Model (CLRM) assumptions.
- We usually make the following set of assumptions about the $u_i$, which originates from the idea of normality.

**Assumption 1: Error is normally distrusted with zero mean**

- Conditional distribution of $u_i$ given $X_i$ has a mean of 0
- This assumption is a formal mathematical statement about the other factors contained in $u_i$ and asserts that these other factors are unrelated to $X_i$.
- This condition is also widely known as the Orthogonality Condition.

$$E(u_i|X_i) = 0$$

# OLS Estimator: Assumptions

**Assumption 2: Variance is constant across the Sample**

- The error variance remains close to constant across the sample.
- This is needed for statistical inference. If variance tends to be changing abruptly, then the inferences looses its power of prediction (think of as in the average is changing randomly in abrupt manner, how could you come to a conclusion in predicting effect of experience on wage?).

$$Var(u_i|X_i) = E(u_i^2) = \sigma^2$$

**Assumption 3: Covariance across error terms is zero and observations are i.i.d**

- The co-movement of error terms indicate that sample is not randomly assigned
- Absence of randomness violates population-sample ideology.

$$cov(u_i, u_j) = E(u_i, u_j) = 0$$

# OLS Estimator: Assumptions

**Assumption 4: Large outliers are unlikely**

- Observations with  values are far outside the usual range of the data—are unlikely.
- Large outliers can make OLS regression results misleading.
- Another way to state this assumption is variables having high kurtosis.

# Linear Regression Model With Multiple Regressor

- Recall our previous one regressor linear model.
- We investigated the effect of experience on wage of individuals with a cross-section data.
- Could this have produced a misleading estimate of the causal effect of experience on wage?
- Broadly speaking, wage can depend on many aspects other than just experience.
- This leads us to the concept of regression with multiple predictors, which we generally call Multivariate regression.

# Multivariate Regression

- Multiple regression model with 2 explanatory variables:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

Multiple regression model with K explanatory variables:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

$i$=runs over all observations, $i$={1, 2, 3,…,N}

$Y_i$=dependent variable

$X_i$= independent variable

$\beta_0$= intercept of the regression line

$\beta_1, \ldots, \beta_k$= slopes/predictors

$u_i$= error term

Interpretation: $\beta_k = \frac{\Delta Y}{\Delta X_k}$, Holding $X_1, X_2, \ldots, X_{k-1}$ constant.

# Multivariate Regression

- Let us take the previous example and extend it.
- Recall, We wanted to check whether experience has any effect on wage. Now we want to extend our analysis by adding an extra variable.
- Idea is to check whether impact of experience remains same or changes coupled with whether we can find any new economic intuition for the new variable.
- We take the last model and add education as the new predictor.

$$Wage_i = \alpha_0 + \beta_1 EXP_i + \beta_2 EDU_i + u_i$$

STATA Commands:

```
webuse regsmpl, clear
reg ln_wage ttl_exp grade
scatter ln_wage grade
scatter ln_wage ttl_exp
```

# Multivariate Regression: Example

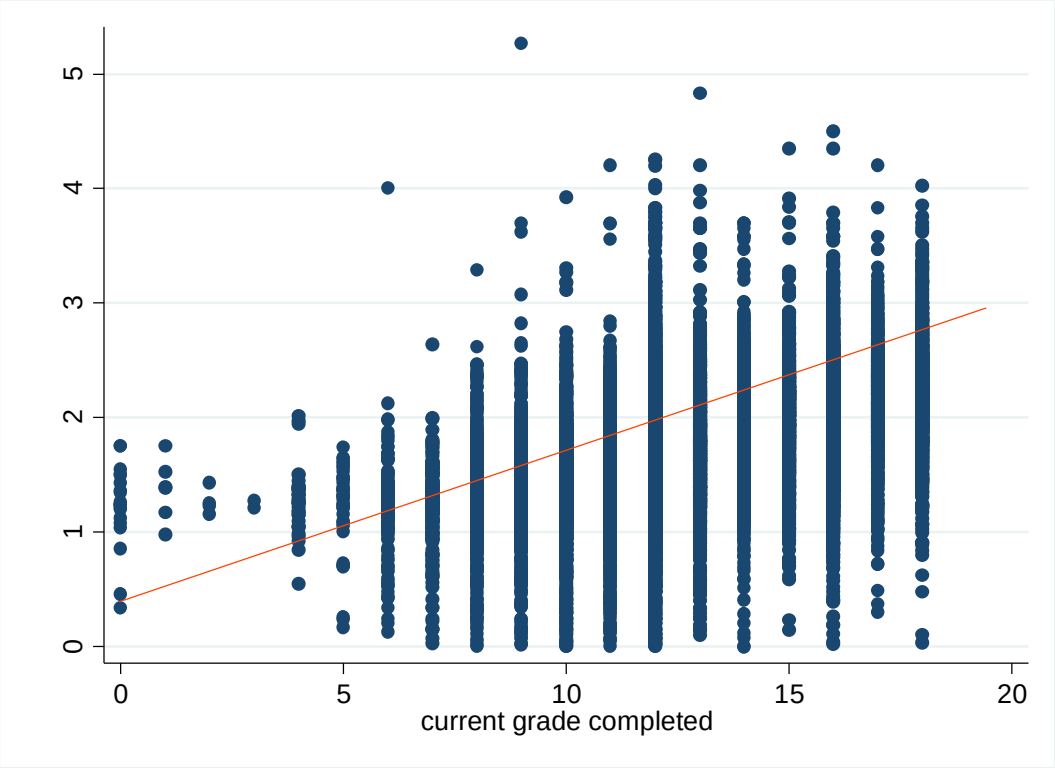
| Source | SS | df | MS | Number of obs | = | 28,532 |
|---|---|---|---|---|---|---|
| | | | | F(2, 28529) | = | 6089.03 |
| Model | 1951.0667 | 2 | 975.53335 | Prob > F | = | 0.0000 |
| Residual | 4570.6764 | 28,529 | .160211588 | R-squared | = | 0.2992 |
| | | | | Adj R-squared | = | 0.2991 |
| Total | 6521.7431 | 28,531 | .228584455 | Root MSE | = | .40026 |

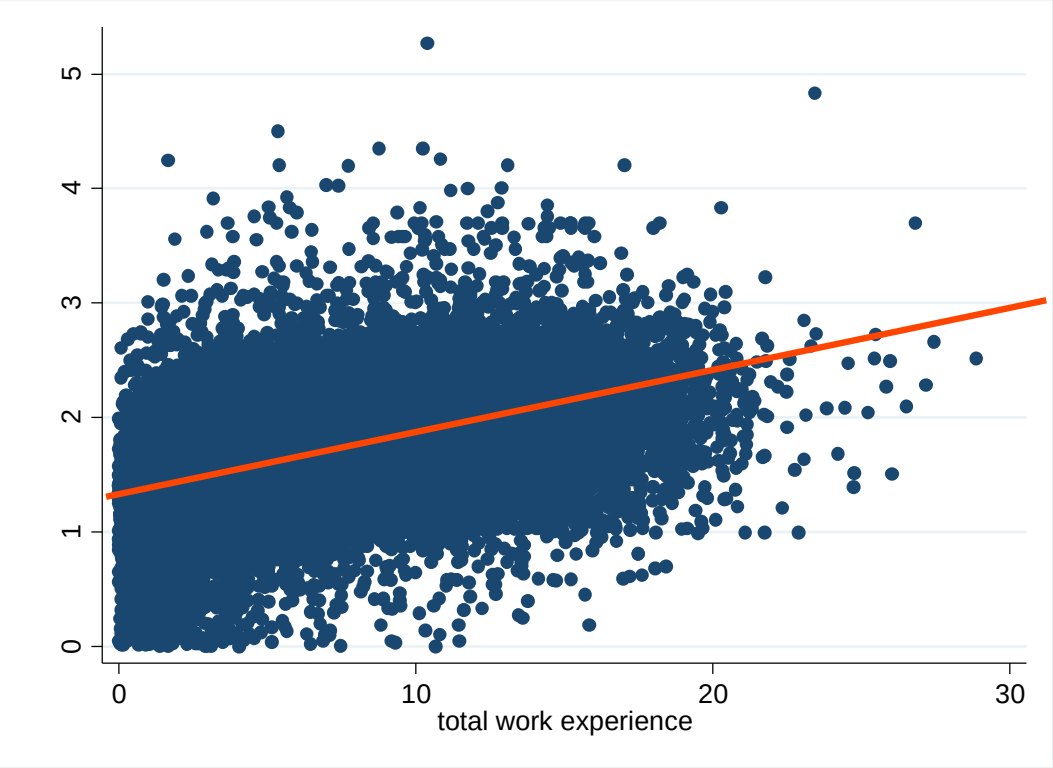
| ln_wage | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| ttl_exp | .0344367 | .0005242 | 65.69 | 0.000 | .0334092 | .0354642 |
| grade | .0741641 | .0010493 | 70.68 | 0.000 | .0721075 | .0762208 |
| _cons | .5314258 | .0129982 | 40.88 | 0.000 | .5059487 | .5569029 |

Points to consider:

- Coefficients (any changes?)
- Coefficient Standard Errors and P-Val
- Model fit parameters (any changes?)

# Multivariate Regression: Example



# The Role of Control Variables in Multivariate Regression

- We replace the first OLS assumption 1 with conditional mean independence
- Consider we have 2 regressors where $X_{1\,i}$ is the variable of interest $X_{2\,i}$ is the s the control variable
- Then conditional mean independence implies

$$E(u_i|X_{1\,i}, X_{2\,i}) = E(u_i|X_{2\,i})$$

- Given this $X_{1\,i}$ treated as if it were randomly assigned. Including $X_{2\,i}$ makes $X_{1\,i}$, uncorrelated with the error term, so that OLS can estimate the casual effect of $X_{1\,i}$ on $Y_i$
- Not using controls (when immensely needed) may lead to biased and inconsistent results.

# Rule of Thumb for Multivariate Regression Analysis

- First, have a base specification - the variables of interest and control variables (if needed) that theory would suggest.
- Second, develop a list of alternative specifications (with additional variables, non-linear models, etc.).
- The idea is to compare the estimate from your base case to other specifications.
- If the estimate changes a lot, then your original specification was probably incorrect.
- An example is provided in the next slide

Table 2: Impact of Outages on Electrification Decisions

| | *Full sample* | | | *Permanent households* | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Outages frequency | -0.0789** | -0.0583** | -0.0571* | -0.0921** | -0.0879** | -0.0902* |
| | (0.0350) | (0.0294) | (0.0328) | (0.0447) | (0.0431) | (0.0509) |
| (log) Month expenditure | | 0.129*** | 0.125*** | | 0.0587*** | 0.0538*** |
| | | (0.00736) | (0.00828) | | (0.00836) | (0.0103) |
| Head of hh primary educ | | 0.207*** | 0.208*** | | 0.189*** | 0.206*** |
| | | (0.0147) | (0.0159) | | (0.0299) | (0.0379) |
| Robust/permanent wall | | 0.228*** | 0.236*** | | 0.0248 | 0.0580*** |
| | | (0.0150) | (0.0168) | | (0.0192) | (0.0215) |
| Hh size | | -0.0240*** | -0.0188*** | | -0.0146*** | -0.00409 |
| | | (0.00278) | (0.00301) | | (0.00480) | (0.00517) |
| (log) Months in dwelling | | -0.0152*** | -0.0128*** | | -0.00640 | -0.00937 |
| | | (0.00337) | (0.00379) | | (0.00537) | (0.00706) |
| Distance grid (no Garissa) | | | 0.00603* | | | 0.00879* |
| | | | (0.00356) | | | (0.00471) |
| Informal connections | | | 0.0647*** | | | 0.0281 |
| | | | (0.0210) | | | (0.0260) |
| Street lights | | | 0.0699*** | | | 0.0717*** |
| | | | (0.0188) | | | (0.0189) |
| Distance plant (no Garissa) | | | -0.00130 | | | 0.00102 |
| | | | (0.00179) | | | (0.00323) |
| Slum | | | -0.0879*** | | | -0.0889*** |
| | | | (0.0228) | | | (0.0239) |
| Urban | | | 0.112*** | | | 0.167*** |
| | | | (0.0241) | | | (0.0557) |
| City dummies | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Province dummies | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 13,102 | 13,016 | 10,121 | 4,177 | 4,153 | 2,603 |
| $R^2$ | 0.048 | 0.209 | 0.228 | 0.032 | 0.088 | 0.124 |

Standard errors clustered at the Enumeration Area (EA) in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Source:
Bajo-Buenestado, R. (2021). The effect of blackouts on household electrification status: evidence from Kenya. *Energy Economics*, 94, 105067.