Linear Regression with One Regressor

Chapter 4

Learning objectives

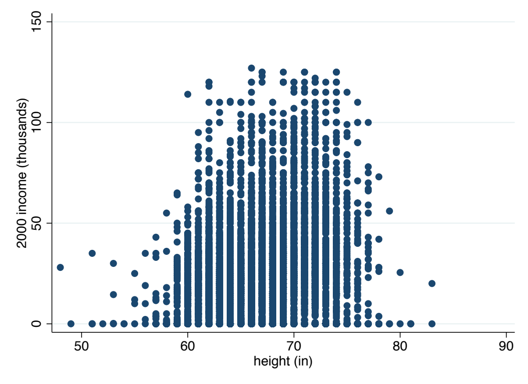
- Set up appropriate equations to estimate relationship between two variables using OLS
- Interpret intercept and slope coefficients for simple linear regression
- Define and calculate residuals
- Calculate measures of fit, including R^2, ESS, TSS, SSR, and SER
- Understand underlying assumptions for estimation of β_0 and β_1

Linear Regression

Overview of linear models

$\boxed{\text{What is the relationship between height and income?}}$

Overview of linear models



Overview of linear models

Several tools to determine the *linear* relationship between two variables:

- Scatter plots (visual)
- Covariance/correlation coefficient

Regression analysis

We use regression analysis to...

- Predict the value of a dependent variable based on the value of at least one independent variable.
- Explain relationship between changes in independent variable and changes in dependent variable.

Dependent variable: Variable we wish to explain (endogenous variable)
Independent variable: Variable we use to explain dependent variable (exogenous variable)

Definition of the simple regression model

- We can relate y to x with the **simple linear regression model**:

$$y = \beta_0 + \beta_1 x + u,$$

- Assume true in population of interest.

Components of population model

$$y = \beta_0 + \beta_1 x + u$$

- u: **error term** or disturbance. Other factors that might affect y
- β_0: **intercept parameter**
- β_1: **slope parameter**

Our goal: get good estimates of β_0 and β_1

Changes in x, holding u fixed

Ceteris paribus: Holding all other things equal

$$y = \beta_0 + \beta_1 x + u,$$

all other factors that affect y are in u. We want to know how y changes when x changes, *holding u fixed*.

Changes in x, holding u fixed

- Let Δ denote "change."
- Holding u fixed means $\Delta u = 0$. So

$$\begin{aligned}\Delta y &= \beta_1 \Delta x + \Delta u \\ &= \beta_1 \Delta x \text{ when } \Delta u = 0.\end{aligned}$$

- This equation effectively defines β_1 as a slope, with restriction $\Delta u = 0$.

How does height affect income?

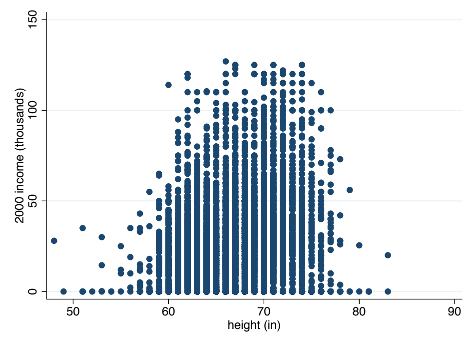
Example 1 (Height and Income)

$$income = \beta_0 + \beta_1 height + u$$

where u contains somewhat "nebulous" factors

$$\Delta income = \beta_1 \Delta height \text{ when } \Delta u = 0$$

Example: Relationship between height and income



- Data from 2000 NSLY on height (in inches) and annual income (in thousands)
- Estimate a regression line - use Stata because $n = 12,016$

Deriving OLS

Deriving the ordinary least squares estimates

- Given data on x and y, how can we estimate the population parameters, β_0 and β_1?

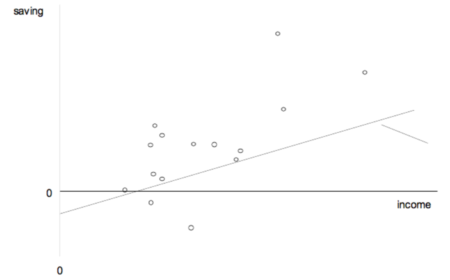
Deriving the ordinary least squares estimates

- Plug any observation into the population equation:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

where the i subscript indicates a particular observation.

- We observe y_i and x_i, but not u_i.



Deriving the ordinary least squares estimates

We choose $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to minimize the mean squared error:

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

Deriving the ordinary least squares estimates

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\text{Sample Covariance}(x_i, y_i)}{\text{Sample Variance}(x_i)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Deriving the ordinary least squares estimates

Sample variance of the x_i cannot be zero, which only rules out the case where each x_i is the same value.

However, this is very rare!

Deriving the ordinary least squares estimates

- Define a **fitted value** for each data point i as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

  We have n of these. It is the value we predict for y_i given that x has taken on the value x_i.

- The mistake we make is the **residual**:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i,$$

  and we have n residuals.

Example: height and income

```
.
.
. reg income height_in

      Source |       SS       df       MS              Number of obs =   12016
-------------+------------------------------           F(  1, 12014) =  247.83
       Model |  125382.214     1  125382.214           Prob > F      =  0.0000
    Residual |  6078127.43 12014  505.920379           R-squared     =  0.0202
-------------+------------------------------           Adj R-squared =  0.0201
       Total |  6203509.64 12015  516.313745           Root MSE      =  22.493

------------------------------------------------------------------------------
 income_2000 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 height_inch |   .7949441   .0504963    15.74   0.000     .6959632     .893925
       _cons |  -36.61049   3.388627   -10.80   0.000    -43.25275   -29.96823
------------------------------------------------------------------------------
```

Example: height and income

$$\widehat{income} = -36.61 + 0.79 \; height$$
$$n = 12016$$

- How much is an additional inch of height worth?
- What is the predicted income for someone who is six feet tall?
- Consider person 898, who is 64 inches tall and earned 21k in 2000. What is her residual?

Measures of Fit

Goodness-of-fit

We define the <u>total</u> sum of squares, <u>estimated</u> sum of squares, and <u>residual</u> sum of squares:

$$y_i = \hat{y}_i + \hat{u}_i$$

$$\begin{aligned} TSS &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\ ESS &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\ SSR &= \sum_{i=1}^{n} \hat{u}_i^2 \end{aligned}$$

Properties of OLS on any Sample of Data

- Assuming $TSS > 0$, we can define the fraction of the total variation in y_i that is explained by x_i (or the OLS regression line) as

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

- Called the **R-squared** of the regression.

$$0 \leq R^2 \leq 1$$

Do not fixate on R^2. Having a "high" R-squared is neither necessary nor sufficient to infer causality.

Standard error of the regression (SER)

We can estimate the variance of the regression

$$\boxed{\hat{\sigma}^2 = s_e^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2} = \frac{SSR}{n-2}}$$

- Divide by $n-2$ because we've used up two d.f: one on $\hat{\beta}_0$ and one on $\hat{\beta}_1$.
- We call $s_e = \sqrt{s_e^2}$ the **standard error of the regression** (SER)

Assumptions

Three least squares assumptions

1. Zero conditional mean: $E[u_i|X_i] = 0$
   - Holds in RCT setting - we try to approximate this
   - Same as saying that u_i and X_i are uncorrelated
2. X_i, Y_i are i.i.d.
3. Large outliers are unlikely (finite kurtosis)

Under these three assumptions, $\hat{\beta}_1$ is an unbiased estimator of β_1.

Zero conditional mean

- x and u have distributions in the population.
- For example, if $x = height$ then, in principle, we could figure out its distribution in the population of adults over, say, 30 years old.
- Suppose u is gender (or childhood nutrition, or SES, or confidence, etc.). Assuming we can measure u, it also has a distribution in the population.
- We must restrict how u and x relate to each other *in the population*.

$E(u) = 0$

- First, we make a simplifying assumption that is without loss of generality: the average, or expected, value of u is zero in the population:

$$E(u) = 0$$

  where $E(\cdot)$ is the expected value (or averaging) operator.
- Normalizing "nutrition," or "ability," to be zero in the population should be harmless. It is.

Adjusting the intercept

- The presence of β_0 in

$$y = \beta_0 + \beta_1 x + u$$

  allows us to assume $E(u) = 0$. If the average of u is different from zero, we just adjust the intercept, leaving the slope the same. If $\alpha_0 = E(u)$ then we can write

$$y = (\beta_0 + \alpha_0) + \beta_1 x + (u - \alpha_0),$$

  where the new error, $u - \alpha_0$, has a zero mean.
- New intercept is $\beta_0 + \alpha_0$. But slope, β_1, has not changed.

Definition of the simple regression model

KEY QUESTION: How do we need to restrict the dependence between u and x?

- We could assume u and x **uncorrelated** in the population:

$$Corr(x, u) = 0$$

- Zero correlation actually works for many purposes, but it implies only that u and x are not **linearly** related. Ruling out only linear dependence can cause problems with interpretation and makes statistical analysis more difficult.

Definition of the simple regression model

- An better assumption involves the mean of the error term for each slice of the population determined by values of x:

$$E(u|x) = E(u), \text{ all values } x,$$

  where $E(u|x)$ means "the expected value of u given x."
- We say u is **mean independent** of x.
- How realistic is this?

Definition of the simple regression model

- Suppose u is "ability" and x is years of education. We need, for example,

$$E(ability|x = 8) = E(ability|x = 12) = E(ability|x = 16)$$

so that the average ability is the same in the different portions of the population with an 8^{th} grade education, a 12^{th} grade education, and a four-year college education.

Zero conditional mean assumption

- Combining $E(u|x) = E(u)$ (the substantive assumption) with $E(u) = 0$ (a normalization) gives

$$E(u|x) = 0, \text{ all values } x$$

- Called the **zero conditional mean assumption**

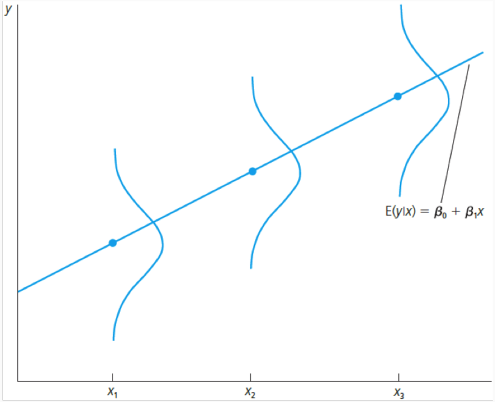
Zero conditional mean assumption

- Because the expected value is a linear operator, $E(u|x) = 0$ implies

$$E(y|x) = \beta_0 + \beta_1 x + E(u|x) = \beta_0 + \beta_1 x,$$

which shows the **population regression function** is a linear function of x.

Definition of the simple regression model



Definition of the simple regression model

- The straight line in the previous graph is the PRF, $E(y|x) = \beta_0 + \beta_1 x$. The conditional distribution of y at three different values of x are superimposed.
- For a given value of x, we see a range of y values: remember, $y = \beta_0 + \beta_1 x + u$, and u has a distribution in the population.

Sampling distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$

- Recall the CLT tells us that as $n \to \infty$, $\bar{X} \sim N(\mu, \sigma^2_{\bar{x}})$
- If three assumptions, hold the sampling distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$ are normal!
- Because estimators get closer and closer to true values (variances go to 0), they are consistent
- Because of CLT, as $n \to \infty$, $\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$
  - Usually, we're quite happy with $n > 100$

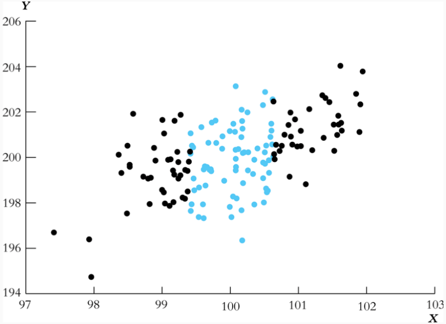
Sampling distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$

For large n, $\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{var[(X_i - \mu_X)u_i]}{var(X_i)^2}$$

$$\text{Larger variance in } X \rightarrow \text{smaller variance in } \beta_1$$

$$\text{Smaller variance in } u \rightarrow \text{smaller variance in } \beta_1$$



Conclusion

4.1 The Linear Regression Model

- Overview of linear models
- Components of population model
- Simple regression model

4.2 Estimating the Coefficients of the Linear Regression Model

4.3 Measures of Fit

4.4/4.5 Assumptions and Sampling Distributions