Berkeley Economic Review

UC Berkeley's Premier Undergraduate Economics Journal

Regression Workshop

March 20, 2024

What we'll go over

- What is regression?
- Types of variables
- Log transformations
- Bias, error, things to look out for
- Causality
- Examples

What is Linear Regression?

- **Linear Regression is a way to predict something, an outcome (think income, your grade in MATH 1B) given some inputs (Parental income, Your grade in MATH 1A)**

$$E\,[Y|X]$$

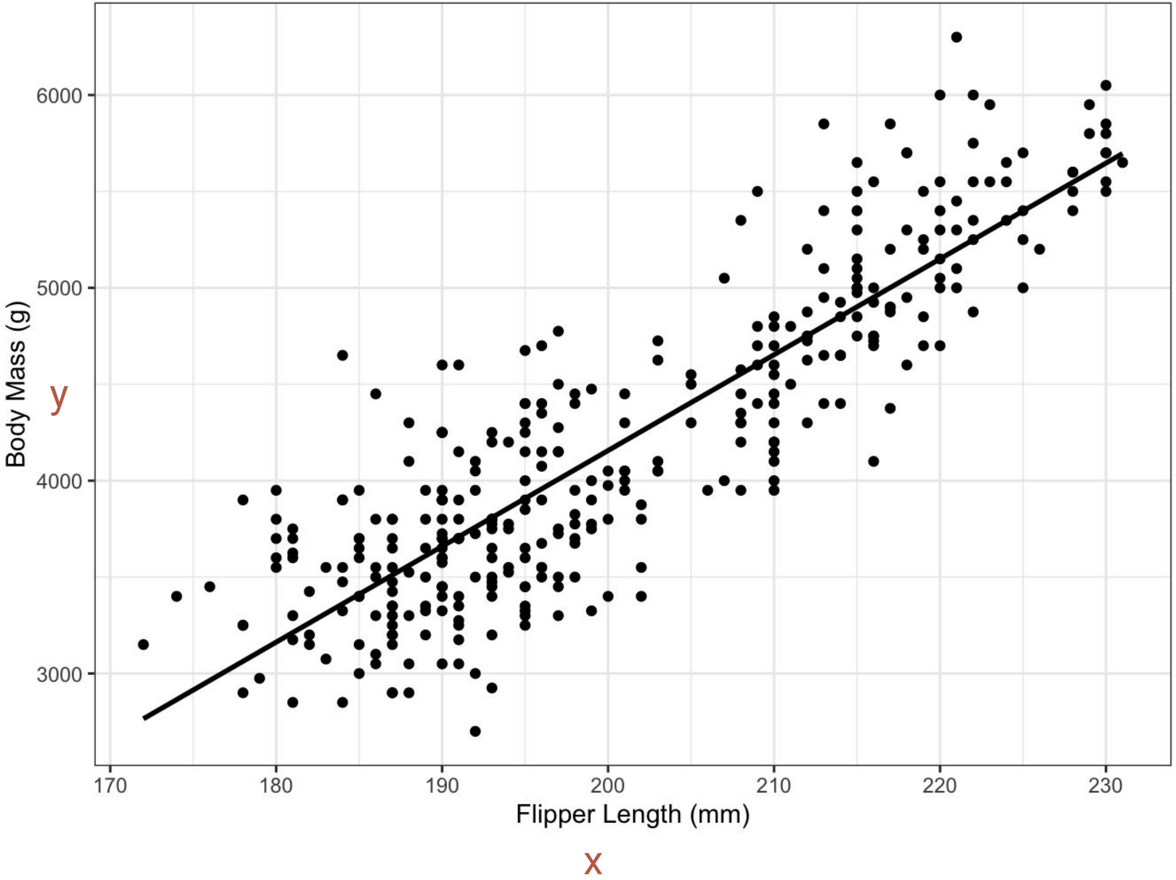
Read as “The expected value of Y given X”

Examples:

- Expected income given your years of education
- Expected grade on ECON 1 midterm given hours of studying

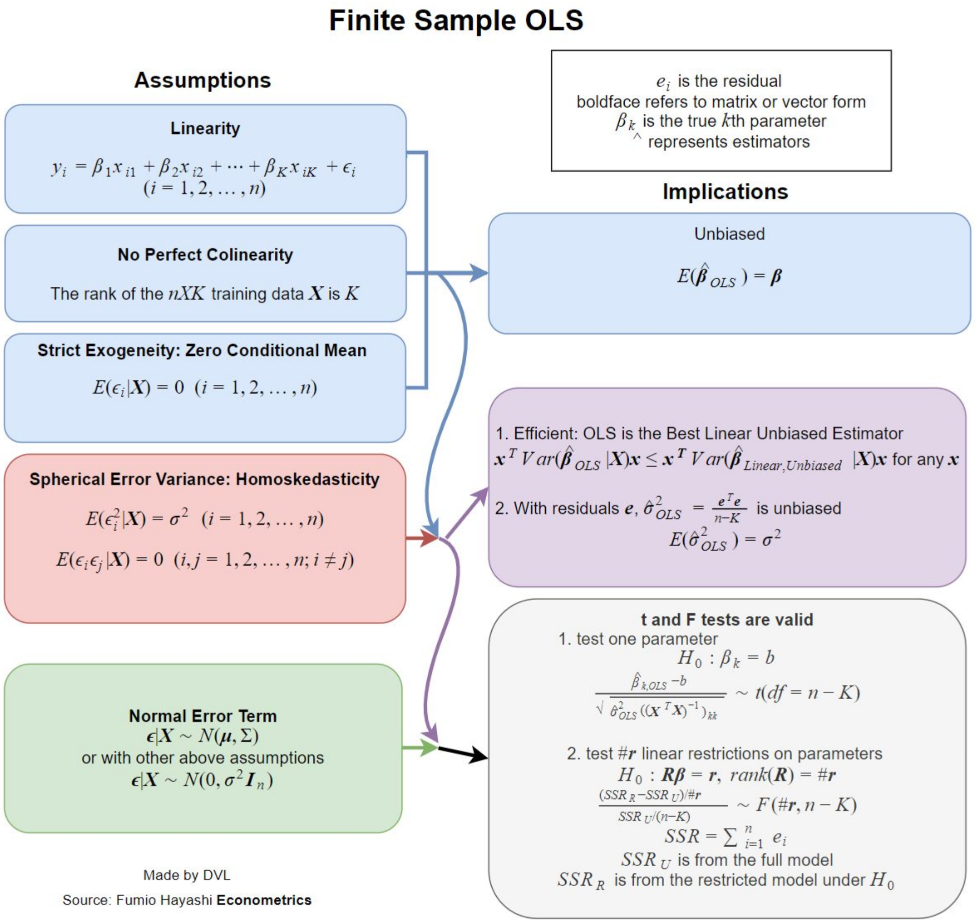
$E[Y|X] = a + bX$

On the right hand side of the equation, a is the intercept, while b is the effect on Y associated with a one unit increase in X.



Assumptions

The most common regression is known as the OLS regression. These regressions have a set of assumptions which have to be met for accurate results.



Types of Variables (2 we'll focus on)

Continuous: takes on a real number value

- Household spending on food,
- Number of parking spaces available at a given time,
- Number of people unemployed in a certain county

Dummy: Only takes the value of 0 or 1

- Immigrant status (0 for native, 1 for immigrant)
- Employed (0 for no, 1 for yes)
- HS Graduate (0 for no, 1 for yes)
- Male (0 for not male, 1 for male)

Panel Data

The most common form of data we will use will be in panel form. Each individual unit will have observations overtime.

Country[1] USA

	Country	Year	Share	LogShare	GDP	TRADE	INDIFF	MARKET_CAP	REER	FEMT	State	lag1
1	USA	1979	3.1759153	4.1759245	3.1759153	.77261806	1.59	9.48	.00445952	.	1	.
2	USA	1980	-.24461139	4.0775374	-.24461139	.78292836	1.02	11.94	.00064887	.	1	4.175925
3	USA	1981	2.5948932	4.0673159	2.5948932	.79809797	.18	12.99	.00537308	.	1	4.077538
4	USA	1982	-1.9105607	4.0943446	-1.9105607	.76272027	-1.09	13.89	.00870641	.	1	4.067316
5	USA	1983	4.6327104	4.0741419	4.6327104	.80325587	-1.46	17.9	.00357724	.	1	4.094345
6	USA	1984	7.2585845	4.0430513	7.2585845	.94100115	-.33	23.64	.00565254	.	1	4.074142
7	USA	1985	4.2392679	4.0324692	4.2392679	.98991969	-.42	29.65	-.00613548	.	1	4.043051
8	USA	1986	3.511871	4.0217739	3.511871	1.0710836	-.13	39.62	-.00843238	.	1	4.032469
9	USA	1987	3.4618128	4.0412953	3.4618128	1.1558425	.75	52.11	-.01123901	.	1	4.021774
10	USA	1988	4.2036331	3.9963642	4.2036331	1.2597821	.81	46.42	-.00267794	.	1	4.041296
11	USA	1989	3.6802641	3.9589066	3.6802641	1.3549662	.53	38.21	.00357157	.	1	3.996364
12	USA	1990	1.9192262	3.9019727	1.9192262	1.4358428	.6	36.88	-.00536581	.	1	3.958907
13	USA	1991	.00698681	3.9199912	.00698681	1.4791796	-.21	33.87	-.00103514	.	1	3.901973
14	USA	1992	3.5825867	4.0127729	3.5825867	1.582294	-.2	35.78	.00340714	.	1	3.919991
15	USA	1993	2.7458979	4.0000339	2.7458979	1.6779405	.07	43.46	.00071032	.	1	4.012773
16	USA	1994	4.0379162	4.0448041	4.0379162	1.8536463	.36	48.6	-.00082982	.	1	4.000034
17	USA	1995	2.7187	4.0768282	2.7187	2.0215726	.47	57.77	.00139911	265.79932	1	4.044804
18	USA	1996	3.79616	4.1268601	3.79616	2.1924917	.62	75.19	.00150017	.	1	4.076828
19	USA	1997	4.486791	4.1758968	4.486791	2.4717969	.29	94.5	.00774395	.	1	4.12686
20	USA	1998	4.449766	4.2381666	4.449766	2.6531153	.24	118.83	-.00072948	383.35755	1	4.175897
21	USA	1999	4.6853276	4.2662421	4.6853276	2.8338913	.72	161.11	-.00061011	.	1	4.238167
22	USA	2000	4.0921108	4.2728698	4.0921108	3.1489921	1.1	236.88	.0059525	.	1	4.266242
23	USA	2001	.97614185	4.276397	.97614185	3.02022	.74	241.17	.00304624	272.5819	1	4.27287

Statistical Significance

When doing a regression analysis, it is important to look out for different levels of statistical significance. One way this is done is through hypothesis testing and "P-values"

Put simply, a P-Value for a regression is the chance that the association between the explanatory variable and the outcome variable was actually just by chance.

P-values are typically calculated using a t-test. In research, statistical significance at the 5% level ($|t| >= 1.96$ aka $p < 0.05$) is the gold standard.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 4.020e+06
Date:                Sun, 07 Jul 2019   Prob (F-statistic):          2.83e-239
Time:                        04:03:37   Log-Likelihood:                -146.51
No. Observations:                 100   AIC:                             299.0
Df Residuals:                      97   BIC:                             306.8
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.3423      0.313      4.292      0.000       0.722       1.963
x1            -0.0402      0.145     -0.278      0.781      -0.327       0.247
x2            10.0103      0.014    715.745      0.000       9.982      10.038
==============================================================================
Omnibus:                        2.042   Durbin-Watson:                   2.274
Prob(Omnibus):                  0.360   Jarque-Bera (JB):                1.875
Skew:                           0.234   Prob(JB):                        0.392
Kurtosis:                       2.519   Cond. No.                         144.
==============================================================================
```

Confidence Intervals

Regression results will also produce confidence intervals. A 95% confidence interval means that there is a 95% chance the true coefficient lies within the interval.

The confidence intervals can be used to identify the range of values that the regression coefficients of interest can most likely take.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 4.020e+06
Date:                Sun, 07 Jul 2019   Prob (F-statistic):          2.83e-239
Time:                        04:03:37   Log-Likelihood:                -146.51
No. Observations:                 100   AIC:                             299.0
Df Residuals:                      97   BIC:                             306.8
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.3423      0.313      4.292      0.000       0.722       1.963
x1            -0.0402      0.145     -0.278      0.781      -0.327       0.247
x2            10.0103      0.014    715.745      0.000       9.982      10.038
==============================================================================
Omnibus:                        2.042   Durbin-Watson:                   2.274
Prob(Omnibus):                  0.360   Jarque-Bera (JB):                1.875
Skew:                           0.234   Prob(JB):                        0.392
Kurtosis:                       2.519   Cond. No.                         144.
==============================================================================
```

Consider:

How would we model this?

Lab scientist wants to find the impact of steroids on rat longevity

↓

They assign treatment and control randomly, 100 each

↓

Rats with steroids on average lived 6 months less than rats without steroids, with a p-value less than 0.05

↓

They arrive at the conclusion that steroids cause rats' lifespans to decrease on average by 6 months

Consider:

How would we model this?

Economist wants to find the impact of education on income

They get survey data, 100 people with bachelor's degree, 100 people without bachelor's degree (HS Diploma)

On average, those with a bachelor's degree earn 30,000 more per year than those without

They arrive at the conclusion that having a bachelor's degree **causes** someone to make 30,000 more

Consider:

Economist wants to find the impact of education on income

They get survey data, 100 people with bachelor's degree, 100 people without bachelor's degree

How would we model this?

On average, those with a bachelor's degree earn 30,000 more per year than those without

They arrive at the conclusion that having a bachelor's degree **causes** someone to make 30,000 more

This is wrong

Why?

Omitted Variable Bias

Along with checking that the assumptions are reasonably met, coefficient estimates can still be inaccurate if there are variables that are not taken into account.

Variables which are related to the independent variables are an explanation for the dependent variable may be included within the error term below. This is where controls (or dummies) come into play.

Dependent Variable (Response Variable)

Independent Variables (Predictors)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$$

Y intercept

Slope Coefficient

Error Term

Causality

Even if we use control variables and dummies to absorb bias from the error, it is many times difficult to say something is causal.

This is where causal models come into play. They help to better narrow down the causal effect of a independent variable. Many times this is done by mathematically removing the error that is related to the independent variable in some way.

Most Common Examples:

- Fixed Effects Model (FE)
- Difference-in-Differences Model (DID)
- Method of Instrumental Variables (IV)
- Regression Discontinuity Design (RDD)

Fixed Effects Model

The Fixed Effects Model can be thought of as comparing the effect of an independent variable within specific groups.

This helps to eliminate the effects of the differences between these groups.

The most common ways to estimate the Fixed Effects Model is through demeaning or creating dummy variables for each group.

Common Examples of groups:

- State or Country
- Time (month, year)

Dummy Variable Method

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \gamma_3 D3_i + \cdots + \gamma_n Dn_i + u_{it}.$$

Demeaning Method

$$\boxed{Y_{it}} = \beta_1 + \beta_2 X_{2it} + \cdots + \alpha_i + \delta_t + u_{it}$$

$$\boxed{\bar{Y}_i} = \beta_1 + \beta_2 \bar{X}_{2i} + \cdots + \alpha_i + \bar{\delta}_t + \bar{u}_i$$

$$\boxed{\left(Y_{it} - \bar{Y}_i\right)} = \beta_2 \left(X_{2it} - \bar{X}_{2i}\right) + \cdots + \left(\delta_t - \bar{\delta}_t\right) + \left(u_{it} - \bar{u}_i\right)$$

Log Transformation

What if we have non-linear relationships?

Log transformations help us interpret non-linear relationships as linear

Works well for relationships that are multiplicative or exponential

Eg;

$$Y = aX^{\beta}$$ Multiplicative (1)

or

$$Y = ae^{\beta X}$$ Exponential (2)

After log-transformation:

(1) $$ln(Y) = ln(a) + \beta ln(X)$$

(2) $$ln(Y) = ln(a) + \beta X$$

Log Transformation: Interpretation

We have a few options for applying natural logs

$$\text{Log-Log: } ln(Y) = \beta_0 + \beta_1 ln(X) + \epsilon$$

$$\text{Lin-Log: } Y = \beta_0 + \beta_1 ln(X) + \epsilon$$

$$\text{Log-Lin: } ln(Y) = \beta_0 + \beta_1 X + \epsilon$$

$$\text{Log-Log: } ln(Y) = \beta_0 + \beta_1 ln(X) + \epsilon$$

The coefficient Beta_1 represents the elasticity of Y to X.
A 1% increase in X results in a Beta_1 percent change in Y

$$\text{Lin-Log: } Y = \beta_0 + \beta_1 ln(X) + \epsilon$$

The coefficient Beta_1 represents the absolute change in Y from a percent change in X
A 1% increase in X results in Beta_1 increase in Y
Useful for when you expect non-constant returns to scale

$$\text{Log-Lin: } ln(Y) = \beta_0 + \beta_1 X + \epsilon$$

The coefficient Beta_1 represents a percent change in Y from an absolute change in X
A 1 unit increase results in Beta_1 percent change in Y