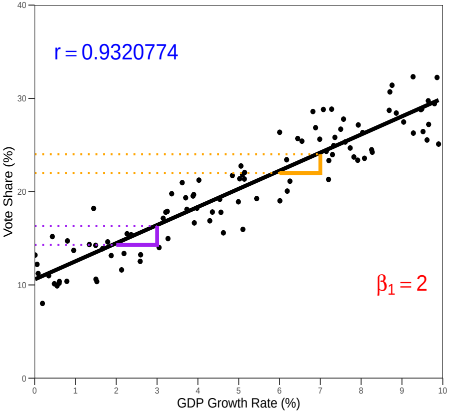
OLS Regression

Hui Zhou
Assistant Professor, Ph.D.,
Department of Political Science
Saint Louis University

May 8, 2024

Comparing Correlation and Regression

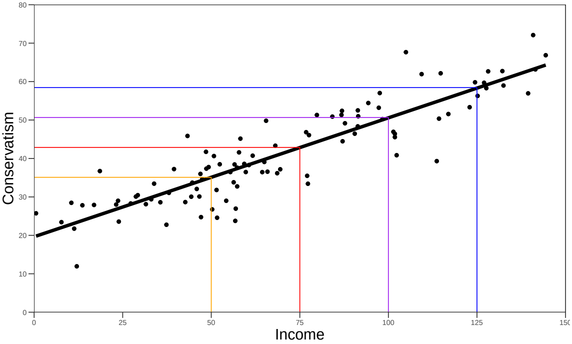


- A correlation coefficient tells us how closely two variables move together. Generally speaking, correlation between two variables will be high if scatter points are close to the fitted regression line.
- A regression coefficient speaks to the slope of the regression line. It tells us by how much Y will change if X goes up by one unit.
- In the meanwhile, a regression also allows you to make predictions about Y given a value of X.

What Is a Simple Linear Regression?

The regression model that uses a straight-line (linear) relationship to predict a numerical dependent variable Y from a *single* numerical independent variable X.

- Technically speaking, a linear model means the change in Y remains constant given a stable change in X



Specifying a Simple Linear Regression Model

- Suppose we are interested in the relationship between GDP growth rate and vote share for the incumbent. We can specify the following *population regression model*

$$Vote_i = \beta_0 + \beta_1 Growth_i + \mu_i$$

  1. β_0 is called intercept. It is the expected value of vote share when growth rate X equals 0.
  2. β_1 is called slope. It measures how much the change in vote share Y will be given a one-unit change in GDP growth.
  3. μ_i is called the error term or disturbance term that is not explained by the regression model.

- Both β_0 and β_1 are called parameters of the population regression model.

Marginal Effect in a Linear Model

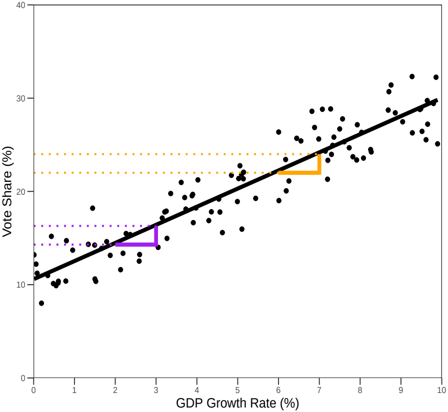


Figure 1: A hypothetical relationship between growth and voting

- The slope coefficient β_1 is called the marginal effect.
- It stands for the effect of X on Y given a one-unit change in X.
- Take the left-hand side picture as an example. If the GDP growth rate goes up by 1 percentage point, the vote share for the incumbent will increase by 2 percentage points.
- This marginal effect remains unchanged regardless of the baseline.

Sample Regression Model

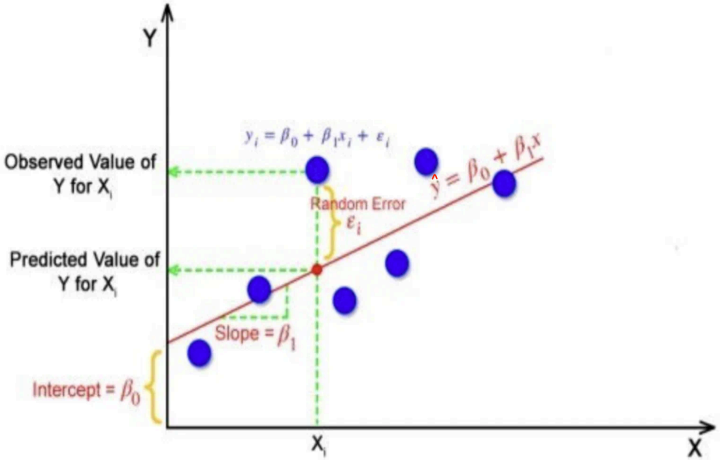
- However, we will never be able to figure out the parameters β_0 and β_1, unless we can collect the data on the entire population. As a result, we use a *sample regression model* to make inferences about the population.

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

- $\hat{\beta}_0$ refers to the estimated intercept from the sample. We add a hat to distinguish it from the population parameter β_0.
- $\hat{\beta}_1$ is the estimated slope based on a sample. Similarly, we use a hat to distinguish it from the population parameter β_1.
- Both $\hat{\beta}_0$ and $\hat{\beta}_1$ are called *sample regression coefficients*.
- $\hat{\mu}_i$ is the error term based on the sample. It is the difference between observed y_i and predicted $\hat{y}$.

$$\begin{aligned} E(y_i|x_i) = \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ y_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\mu}_i \\ \implies \hat{\mu}_i &= y_i - \hat{y}_i \end{aligned}$$

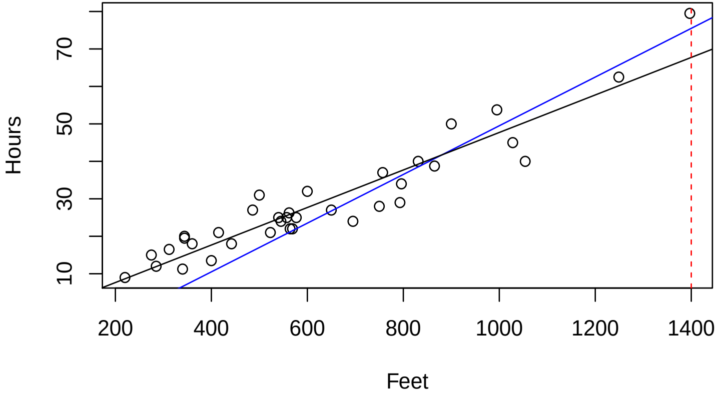
Understanding Residuals



- All predicted values, $\hat{y}$, are located on the regression line.
- The vertical height of a point stands for the observed value y_i.
- Their difference is called a random error, prediction error, or residual.

The Ordinary Least Squares (OLS) Method

- Estimating a simple linear regression model is like picking a straight line. Which line should you pick?
- We need to have a rule of thumb, which is determined by error size.

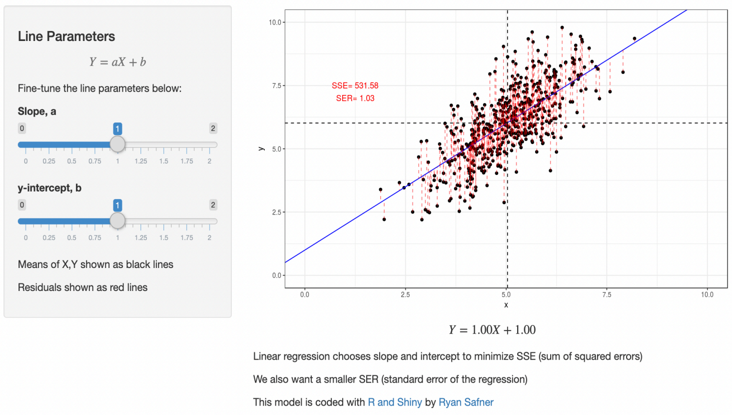


Predicted Values and Regression Line

1. Blue line: $y_i = -15.5 + 0.065x_i + \hat{\mu}_i$
   - $\sum_{i=1}^{i=n} {\mu_i}^2 = \sum_{i=1}^{i=n} (y_i - \hat{y}_i)^2 = 1993.153$
2. Black line: $y_i = -2.37 + 0.05x_i + \hat{\mu}_i$
   - $\sum_{i=1}^{i=n} {\mu_i}^2 = \sum_{i=1}^{i=n} (y_i - \hat{y}_i)^2 = 860.7186$
3. The error size associated with the black line is smaller. Thus, we should choose the black line.
4. We can predict the dependent variable to get $\hat{y}_i$ based on any given values of x_i using the formula $\hat{y}_i = -2.37 + 0.05x_i$
   - When $x_i = 0, \hat{y}_i = -2.37$
   - When $x_i = 200, \hat{y}_i = 7.63$
   - When $x_i = 600, \hat{y}_i = 27.63$
   - When $x_i = 1000, \hat{y}_i = 47.63$
   - When $x_i = 1400, \hat{y}_i = 67.63$
5. These predicted values will eventually form the regression line.

Ordinary Least Squares (OLS)

- Is it possible to find a line that produces the least amount of errors?
- Yes! The Ordinary Least Squares (OLS) method allows us to always find a model that minimizes the sum of squared errors.
- See a simulation by a Shiny dashboard.



Estimating Coefficients Using OLS

- A two-variable (bivariate) OLS regression model comes with two coefficients: $\hat{\beta}_1$ (slope) and $\hat{\beta}_0$ (intercept).

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{137992.8}{2755433} = 0.05$$

- Once we calculated the slope coefficient, $\hat{\beta}_1$, the intercept coefficient $\hat{\beta}_0$ can be derived as follows:

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

- Computing the regression coefficients manually

$$\hat{\beta}_0 = 28.95833 - 0.05 \times 625.5556 = -2.37$$

- Finally, we get the regression model

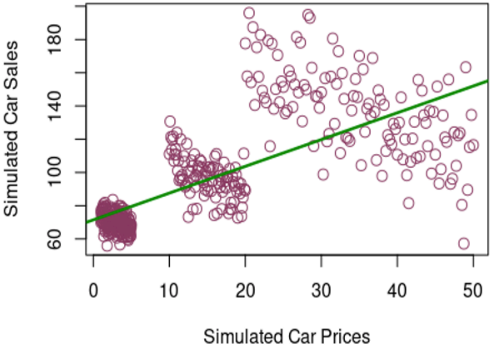
$$y_i = -2.37 + 0.05x_i + \hat{\mu}_i$$

Interpretating Regression Results

$$E(y_i|x_i) = \hat{y}_i = -2.37 + 0.05x_i$$

- The intercept coefficient is -2.37, suggesting that the mean of the dependent variable Labor Hours is -2.37, when the independent variable is set at 0. Note that this statement does not make sense because the dependent variable will never be negative.
- More important is the slope coefficient, which represents a marginal effect of X on Y, denoting the change in Y given a one unit change in X.
- Standard statement: A one-unit increase in the amount of furniture to be moved is *associated with* an increase of 0.05 labor hours to get the job done.
- This effect cannot be interpreted as causal unless some very restrictive assumptions are satisfied.

Taking Confounders into Account



- A confounding variable results in a false relationship between X and Y. Such a variable is also simplified as a *confounder*.
- A confounder exists in the context of not only categorical variables (Simpson's Paradox) but also numerical variables.

- Confounders must be considered in empirical analysis.
- To control for confounders, we need to estimate a multiple regression model.

From a Simple Regression to a Multiple Regression

1. A multiple regression extends the simple linear regression model by including multiple independent variables in a model and assuming a straight-line or linear relationship between each independent variable and the dependent variable.

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

   - You can add as many variables as you want provided that the estimation is allowed by the dataset
     - key independent variable vs. control variable
   - There is a linear relationship between each independent variable and the dependent variable

2. Both β_1 and β_2 are called partial regression coefficients, which correspond to the change in Y given a one-unit change in X provided all other variables are held constant.

The OLS Method and Regressions

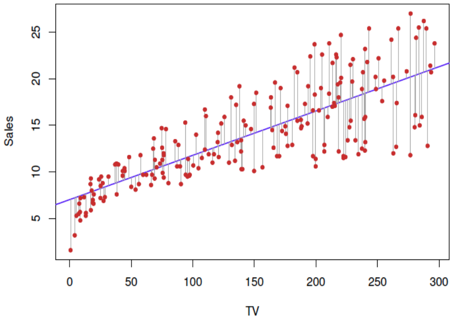


Figure 2: OLS with only one predictor

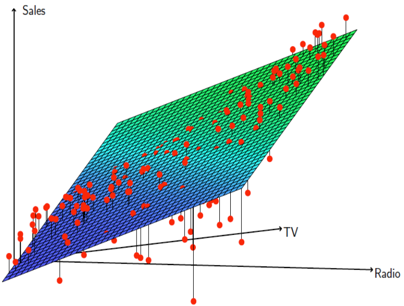


Figure 3: OLS with two predictors

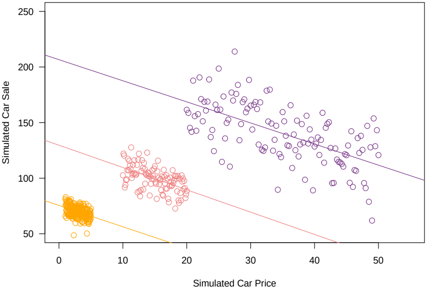
- No matter whether we use a simple regression or multiple regression, the goal of the OLS estimation is always to minimize the sum of squared residuals $\sum_{i=1}^{i=n} \hat{u}_i^2$.

Reconsider the Car Sales Question

$$Sale_i = \hat{\beta}_0 + \hat{\beta}_1 \times Price_i + \hat{\beta}_2 \times Era_i + \hat{u}_i$$

Table 1: Car sales and car prices

	Dependent Variable	
	Simulated Car Sales	
	(1)	(2)
Simulated Prices	1.716***	−1.906***
	(0.078)	(0.125)
factor(Era)1980		52.673***
		(2.165)
factor(Era)2010		131.427***
		(4.273)
Constant	71.268***	75.495***
	(1.618)	(0.975)
Observations	423	423
R^2	0.536	0.858
Adjusted R^2	0.535	0.857
Residual Std. Error	23.023 (df = 421)	12.760 (df = 419)
F Statistic	486.972*** (df = 1; 421)	845.606*** (df = 3; 419)
Note:		*p<0.1; **p<0.05; ***p<0.01



- Controlling for the Era variable allows us to estimate a regression within each level of Era.
- The overall effect of Price on Sale is a weighted average of the three slope coefficients in each Era.

Thank you for listening!

Appendix

1. Regression and correlation
2. Model specification
3. Regression assumptions
4. Inference-based and prediction-based modeling
5. Standard error of the regression
6. Estimating a linear regression in R

Regression and correlation

- Correlation coefficient

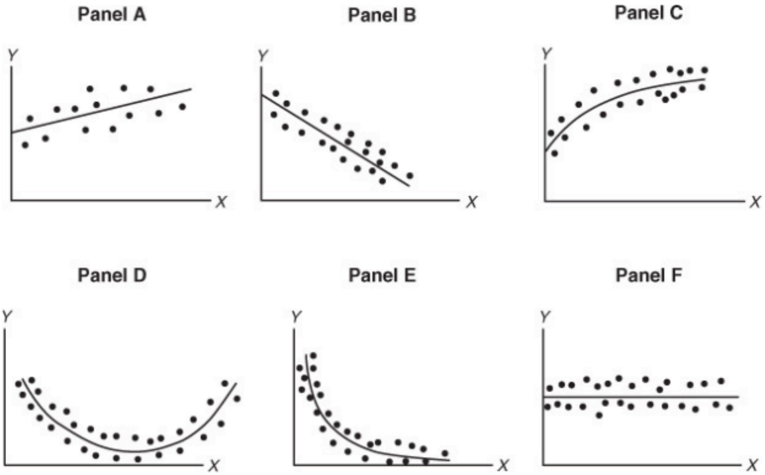
$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

- Regression coefficient in a simple linear regression

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

- It follows that (1) a regression coefficient has the same sign as a covariance or correlation coefficient; (2) the size of a regression coefficient is determined by both the covariance between x and y and the variation in x.

Model specification

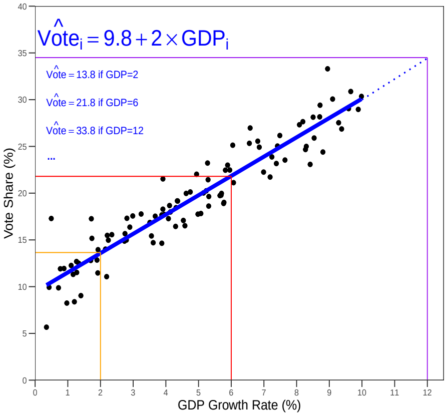


Regression assumptions

- In order for regression models to be a valid representation of the underlying population, some assumptions must be satisfied.
  1. Linearity: the parameters we are estimating by OLS are linear.
  2. Nonexistence of endogeneity (or no specification bias): all variables that affect y and x have been included in the model.
  3. Zero conditional mean: the expectation of the residuals is zero.
  4. Homoscedasticity: the variance of the error term is constant.
  5. No autocorrelation: there is no correlation between residuals.
  6. Multicollinearity: no exact collinearity or near collinearity between independent variables.
  7. Normality: errors are independent of Xs and normally distributed.
  8. Influential data points: nonexistence of influential points.
- The first six assumptions are typically referred to as the *Gauss–Markov assumptions*.
- If the normality assumption is also counted, then we will have classical linear model (CLM) assumptions (Wooldridge, 2014).

Inference-based and prediction-based modeling

- OLS can handle both inference-based modeling and prediction-based modeling.
- After estimating a regression model using OLS, you can interpret the effect of a variable using marginal effect (β_1).
- Additionally, you can also predict Y given different values of X. **The predicted values of Y are just located on the regression line**.



- However, OLS might not be the optimal method if the interest is solely in yielding accurate predictions.

Standard error of the regression

- Here, we use a similar concept, standard error, to capture our uncertainty in the estimated regression coefficients.
- Unseen variance of the population regression stochastic component, μ_i is estimated based on sample residual terms:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{\mu_i}^2}{n-2}$$

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \hat{\mu_i}^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}}$$

- $n - 2$ is the degrees of freedom. Since we have two parameters to estimate, the degrees of freedom is $n - 2$.
- The standard error of the regression measures the standard deviation of the differences between predicted y from observed y. The smaller the $\hat{\sigma}$, the better the model fit.

Estimating a linear regression in R

```
> reg1 <- lm(Hours~Feet, data=mydata)
# estimate the regression
> summary(reg1)
# output results
Call:
lm(formula = Hours ~ Feet, data = mydata)

Residuals:
Min        1Q     Median      3Q       Max
-10.4149  -3.4293   0.2115   3.3329  11.9075

Coefficients:
Estimate   Std. Error  t value   Pr(>|t|)
(Intercept) -2.369660  2.073261    -1.143    0.261
Feet         0.050080  0.003031    16.522    <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.031 on 34 degrees of freedom
Multiple R-squared:  0.8892,   Adjusted R-squared:  0.886
F-statistic:   273 on 1 and 34 DF,  p-value: < 2.2e-16
```