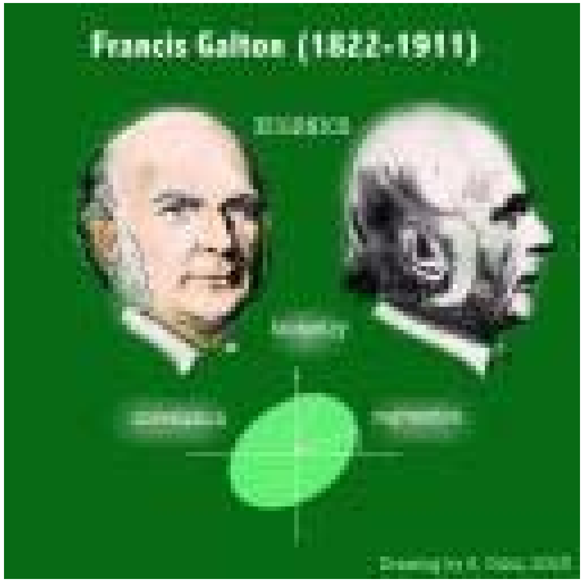
Who Invented Regression?

1. **Who invented regression?**
2. **Omitted Variables and Multivariate Regression**
3. **Omitted Variable Bias (OVB)**
4. **Experiments vs. OVB**
5. **R^2**

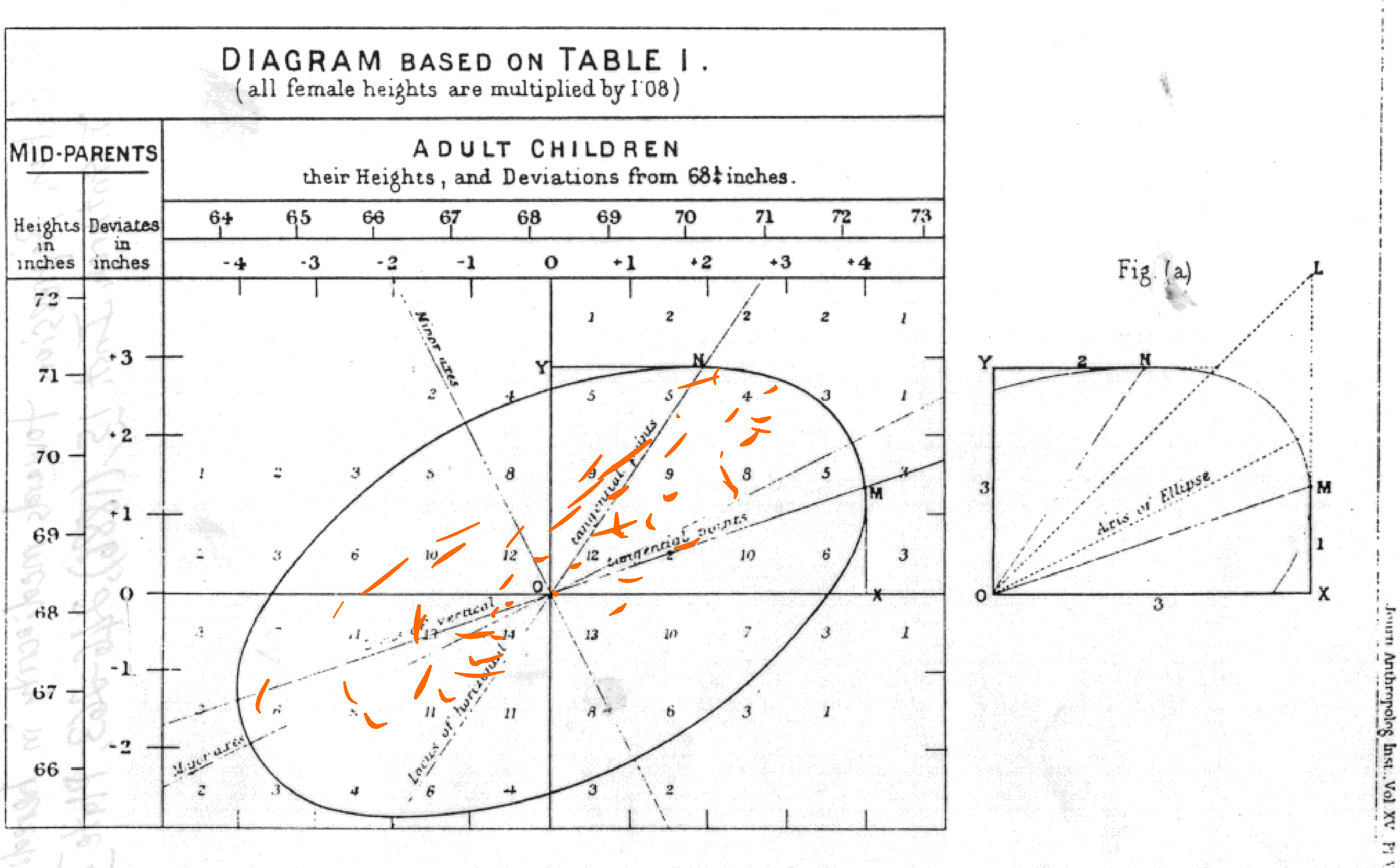
1. Who invented regression?

- Francis Galton,
  - climatologist,
  - gentleman explorer
  - social scientist



Heredity and Height

"regression" to the mean



2. Omitted Variables and Omitted Variable Bias

- **What if you left out an important variable?**
- **Many interesting relationships have more than 2 dimensions**

  GRE prep course example

  Coffee example

  Problem set and exam example
- **We need more variables..**

  “multivariate” regression

2. OLS Multivariate regression

The OLS Estimators, Predicted Values, and Residuals in the Multiple Regression Model

The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are the values of $b_0, b_1, \ldots, b_k$ that minimize the sum of squared prediction mistakes $\underbrace{\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2}_{\sum e_i^2}$. The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki},\ i = 1, \ldots, n, \text{ and} \tag{5.11}$$

$$\hat{u}_i = Y_i - \hat{Y}_i,\ i = 1, \ldots, n. \tag{5.12}$$

The OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ and residual $\hat{u}_i$ are computed from a sample of n observations of $(X_{1i}, \ldots, X_{ki}, Y_i)$, $i = 1, \ldots, n$. These are estimators of the unknown true population coefficients $\beta_0, \beta_1, \ldots, \beta_k$ and error term, u_i.

Look familiar? Same criterion with more variables.

2. Properties of OLS estimators in Multivariate Regression

- Consistent
- Unbiased
- Approximately N(.) in large samples
- Same first order conditions (for 2 or more X's)

$$\sum_{i=1}^{N} e_i = 0$$

$$\sum_{i=1}^{N} X_{1i} e_i = 0$$

$$\sum_{i=1}^{N} X_{2i} e_i = 0$$

First order conditions for multivariate regression

$$\frac{\partial e_i^2}{\partial b_0} = \frac{\partial e_i^2}{\partial e_i} \times \frac{\partial e_i}{\partial b_0} = 2e_i\,(-1)$$

$$\underset{\{b_0, b_1, b_2 \ldots b_k\}}{\text{Min}} \sum e_i^2 \;, \quad e_i = (Y_i - \hat{Y}_i)$$

$$= (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i} - b_3 X_{3i} - \ldots - b_k X_{ki})$$

k+1 1st order conditions..

$$0 = \frac{\partial \sum e_i^2}{\partial b_0} = \frac{\partial \sum e_i^2}{\partial b_0} = \frac{\partial e_1^2}{\partial b_0} + \frac{\partial e_2^2}{\partial b_0} + \ldots \frac{\partial e_N^2}{\partial b_0} = -2e_1 - 2e_2 - 2e_3 - 2e_4 \ldots - 2e_N$$

$$= -2\sum e_i \iff \boxed{\sum e_i = 0} \quad \#1$$

$$0 = \frac{\partial \sum e_i^2}{\partial b_1} = \frac{\partial e_1^2}{\partial b_1} + \frac{\partial e_2^2}{\partial b_1} + \frac{\partial e_3^2}{\partial b_1} + \ldots \frac{\partial e_N^2}{\partial b_1} = -2e_1 X_{11} - 2e_2 X_{12} - 2e_3 X_{13} \ldots - 2e_N X_N$$

$$= -2\sum_i e_i X_{1i} \iff \boxed{\sum e_i X_{1i} = 0} \quad \#2$$

$$\vdots$$

$$0 = \frac{\partial \sum e_i^2}{\partial b_k} = -2\sum e_i X_{ki} \iff \boxed{\sum e_i X_{ki} = 0} \quad \#k{+}1$$

If the k+1 equations are not linearly dependent we can solve for $b_0, b_1, b_2 \ldots b_k$.

IN STATA:
regress y x1 x2..xk, robust

3. Omitted Variable "Bias"

e.g. y - SAT score
x_1 - SAT prep.
x_2 - SAT ability

- Short regression

$$y = b_0^s + \underline{b_1^s} x_1 + e^S \quad \text{(SR)}$$

- Long regression

$$y = b_0^L + b_1^L x_1 + b_2^L x_2 + e^L \qquad \text{(LR)}$$

- Claim:

$$b_1^s = b_1^L + b_2^L b_{21} ,$$

$$\frac{dY}{dx_1} = \left.\frac{\partial Y}{\partial x_1}\right|_{x_2} + \frac{\partial Y}{\partial x_2}\frac{dx_2}{dx_1}$$

b_{21} is slope of a regression of x_2 on x_1

$$b_{21} = \frac{\Sigma(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)}{\Sigma(x_{1i}-\bar{x}_1)^2}$$

Omitted variable bias formula - derivation

$$b_1^s = \frac{\sum (X_{1i}-\bar{X}_1)(Y_i-\bar{Y})}{\sum (X_{1i}-\bar{X}_1)^2} = \frac{\sum (X_{1i}-\bar{X}_1)Y_i}{\sum (X_{1i}-\bar{X}_1)^2} = \frac{\sum (X_{1i}-\bar{X}_1)[\overset{1}{b_0^L} + \overset{2}{b_1^L X_{1i}} + \overset{3}{b_2^L X_{2i}} + \overset{4}{e_i}]}{\sum (X_{1i}-\bar{X}_1)^2}$$

$$= \overset{1}{0} + b_1^L \overset{2}{\frac{\sum (X_{1i}-\bar{X}_1)X_{1i}}{\sum (X_{1i}-\bar{X}_1)(X_{1i}-\bar{X}_1)}} + b_2^L \overset{3}{\frac{\sum (X_{1i}-\bar{X}_1)(X_{2i}-\bar{X}_2)}{\sum (X_{1i}-\bar{X}_1)^2}} + \frac{\sum (X_{1i}-\bar{X}_1)e_i^L}{\sum (X_{1i}-\bar{X}_1)^2}$$

$$b_1^s = b_1^L + b_2^L b_{21}$$

$$\sum (X_{1i}-\bar{X}_1)e_i^L$$

$$= \sum X_{1i} e_i^L - \cancel{\sum \bar{X}_1 e_i^L}^{\,0}$$

4. Why experiments eliminate OVB

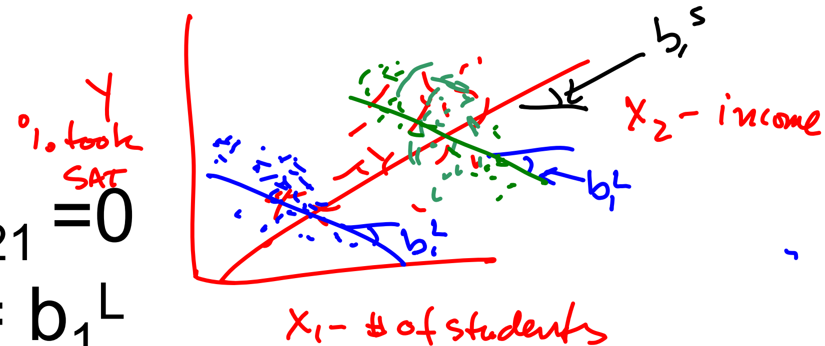
(s) $Y = b_0^s + b_1^s X_1 + e^s$

(L) $Y = b_0^L + b_1^L X_1 + b_2^L X_2 + e^L$

(+) (-) (+) (+)

$$b_1^s = b_1^L + b_2^L b_{21} \, ,$$

= 0



- So there's no OVB if $b_{21} = 0$
  i.e., $b_{21} = 0$ implies $b_1^s = b_1^L$

.. Which you can guarantee if you design an experiment in which X_1 is uncorrelated with other X's (omitted variables).

Random assignment of X_1 is sure to do that.

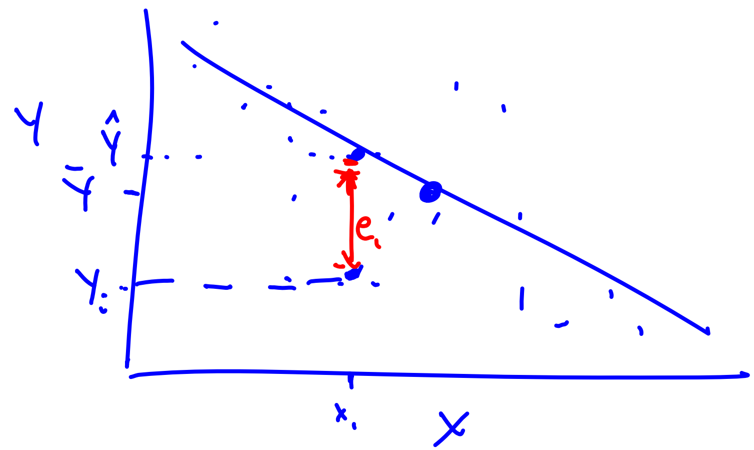
.. Back to examples to demonstrate

5. R^2 – How much Variation Explained?

- How much of the variation in Y did we explain with the regression line?

$$R^2 = \frac{\sum_{i=1}^{N} (\hat{y} - \bar{y})^2}{\sum_{i=1}^{N} (y - \bar{y})^2}$$

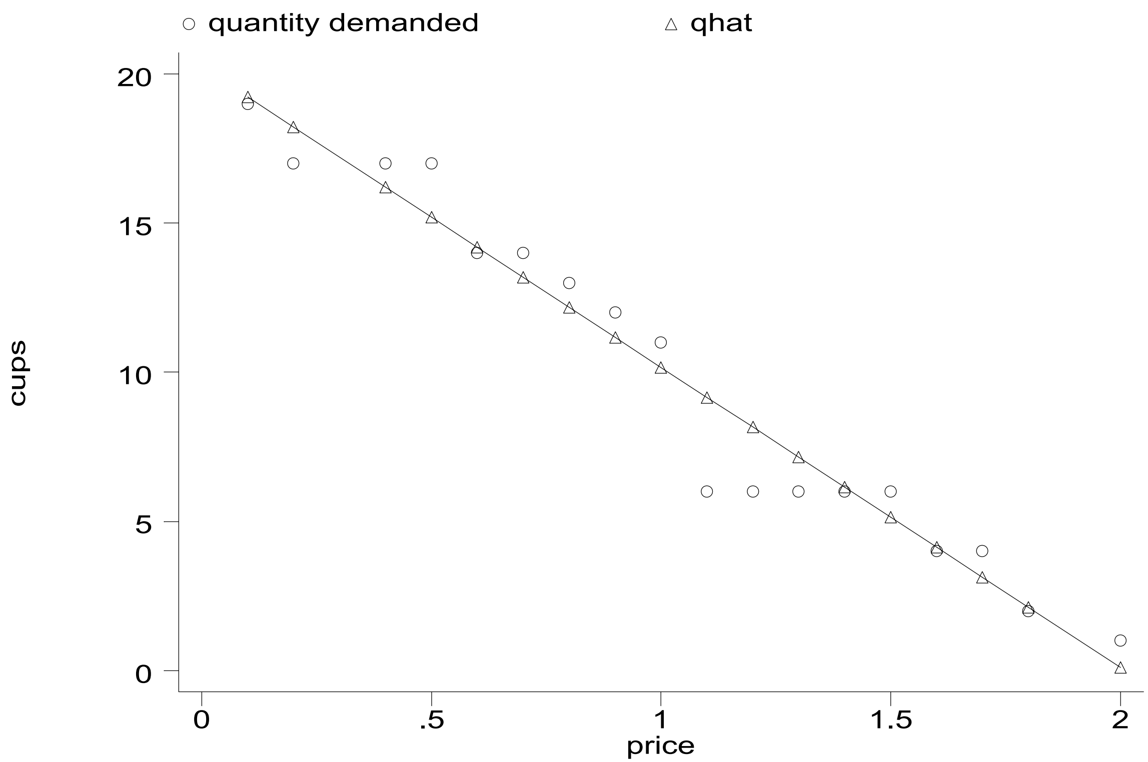
$$= 1 - \frac{\sum_{i=1}^{N} e^2}{\sum_{i=1}^{N} (y - \bar{y})^2}$$



CLAIM: $\Sigma(Y_i - \bar{Y}) = \Sigma(\hat{Y}_i - \bar{Y}) + \Sigma e^2$

total = on the line + off the line

Coffee Demand – High R^2



E.g. Coffee Demand – high R^2

- p is the price of coffee,
- q is the quantity (in cups)

```
. reg q p, robust

Regression with robust standard errors                 Number of obs =      23
                                                       F(  1,    21) =   28.20
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.7349
                                                       Root MSE      =  4.0549

------------------------------------------------------------------------------
             |               Robust
           q |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           p |  -6.246766   1.176301    -5.31   0.000    -8.693018   -3.800513
       _cons |    17.5064   1.822797     9.60   0.000     13.71568    21.29711
------------------------------------------------------------------------------
```

Eg. Wage Regression - Low R^2

*** Lhwage is log(hourly wage), ed is years of education**

```
regress lhwage ed, robust

Regression with robust standard errors                 Number of obs =   13743
                                                       F(  1, 13741) = 1795.40
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1185
                                                       Root MSE      =   .5083

------------------------------------------------------------------------------
             |               Robust
      lhwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ed |   .0704563   .0016628    42.37   0.000     .0671969    .0737156
       _cons |   .9852746   .0238393    41.33   0.000     .9385464    1.032003
------------------------------------------------------------------------------
```