# Quantitative Methods

# Data DNA

# Data DNA

- The analyst needs to comprehend the Data DNA before designing data analysis because:
- DNA type **determines** to a large extent the choice of appropriate econometric method.
- **source of variation**
- **format of measurement**
- As you will see there is a wide variety of data types, which means that an analyst must possess in her arsenal many weapons.

# source of variation

- **cross-sectional** (atemporal): data represent a single point in time (or various time points have been artificially collapsed on a single point. Essentially a snapshot. $Y_i$, $X_i$
- Typical cases:
- a survey based on a random sample of firms or individuals.
- Thus, the variation **stems** from differences **between** cross sectional units.

# source of variation cont

- **time series:** data are collected sequentially in time (and most of the time with a fixed periodicity), $Y_t$, $X_t$
- The cross-sectional dimension is unity. Hence, the data track a single cross-sectional unit over time.
- For example: a firm, a household, a country.
- Thus, the variation **stems** from differences **within** the cross sectional unit, reflecting the effect of passage of time.

# source of variation cont

- **Panel:** data for various cross-sectional units are collected repeatedly in time.
- For example, follow: a set of firms, or a set of households, or a set of countries over a number of time periods. $Y_{it}$, $X_{it}$
- Thus, the overall variation **stems** from:
- (i) differences **between** cross sectional units, and
- (ii) differences **within** every cross sectional unit, reflecting the effect of passage of time.

# format of measurement

- **A: continuous** (roe, gdp growth),
- **B: discrete**
- B1: **dichotomous** (default vs. non-default, export vs. non-export, dividend vs. no dividend),
- **B2: ordered** (life satisfaction, loan interest rates down, same, up),
- **B3: categorical** (choice of transportation mode),
- **B4: count** (number of terrorist events per unit of time, number of defaults per unit of time, number of plants per unit of time)

# A few examples

- Data for US commercial banks in 2009 (See excel file us_banks_example_1)
- Source of variation: cross-sectional (approx. 8000 banks across the US in 2009).

# Variables analyzed

- **Size1:** *Logarithm of Total Assets*
- *Size2: Logarithm of Number of Employees*
- **Risk1:** *Risk Weighted Assets / Total Assets*
- *risk2: Provisions for Loan Losses / Total Loans*
- **(Profitability)**
- ***Nim****: (Total Interest Income - Total Interest Expense) / Total Assets*
- ***Noi****: Net Operating Income / Total Assets*
- ***Default***: indicator variable showing if a bank defaulted in that year
- *Capital Adequacy*
- Asset Concentration Hierarchy

# Asset Concentration Hierarchy (specgrp)

- An indicator of an institutions' primary specialization in terms of asset concentration. **(Groups are mutually exclusive):**

  **1 - International Specialization** Institutions with assets greater than $10 billion and more than 25 percent of total assets in foreign offices.
  **2 - Agricultural Specialization** Banks with agricultural production loans plus real estate loans secured by farmland in excess of 25 percent of total loans and leases.
  **3 - Credit-card Specialization** Institutions with credit-card loans plus securitized receivables in excess of 50 percent of total assets plus securitized receivables.
  **4 - Commercial Lending Specialization** Institutions with commercial and industrial loans, plus real estate construction and development loans, plus loans secured by commercial real estate properties in excess of 25 percent of total assets.
  **5 - Mortgage Lending Specialization** Institutions with residential mortgage loans, plus mortgage-backed securities, in excess of 50 percent of total assets.
  **6 - Consumer Lending Specialization** Institutions with residential mortgage loans, plus credit-card loans, plus other loans to individuals, in excess of 50 percent of total assets.
  **7 - Other Specialized < $1 Billion** Institutions with assets less than $1 billion and with loans and leases are less than 40 percent of total assets.
  **8 - All Other < $1 Billion** Institutions with assets less than $1 billion that do not meet any of the definitions above, they have significant lending activity with no identified asset concentrations.
  **9 - All Other > $1 Billion** Institutions with assets greater than $1 billion that do not meet any of the definitions above, they have significant lending activity with no identified asset concentrations.

# What about format of measurement?
# A dichotomous and a categorical variable

```
. tab default
```

| default | Freq. | Percent | Cum. |
|---|---|---|---|
| 0 | 8,003 | 98.24 | 98.24 |
| 1 | 143 | 1.76 | 100.00 |
| Total | 8,146 | 100.00 | |

```
. tab specgrp
```

| specgrp | Freq. | Percent | Cum. |
|---|---|---|---|
| 1 | 4 | 0.05 | 0.05 |
| 2 | 1,564 | 19.54 | 19.59 |
| 3 | 23 | 0.29 | 19.88 |
| 4 | 4,450 | 55.60 | 75.48 |
| 5 | 765 | 9.56 | 85.04 |
| 6 | 80 | 1.00 | 86.04 |
| 7 | 289 | 3.61 | 89.65 |
| 8 | 770 | 9.62 | 99.28 |
| 9 | 58 | 0.72 | 100.00 |
| Total | 8,003 | 100.00 | |

# An ordered variable *capitalization*

$$CAP_{i,t} = \begin{cases} 0, \text{ if bank } i \text{ is critically undercapitalized in year } t \\ 1, \text{ if bank i is significantly undercapitalized in year } t \\ 2, \text{ if bank } i \text{ is undercapitalized in year } t \\ 3, \text{ if bank } i \text{ is adequately capitalized in year } t \\ 4, \text{ if bank } i \text{ is well capitalized in year } t \end{cases}$$

| group(__000 006) | Freq. | Percent | Cum. |
|---|---|---|---|
| 0 | 19 | 0.24 | 0.24 |
| 1 | 73 | 0.93 | 1.17 |
| 2 | 92 | 1.17 | 2.34 |
| 3 | 172 | 2.19 | 4.53 |
| 4 | 7,509 | 95.47 | 100.00 |
| Total | 7,865 | 100.00 | |

# Cont, and a continuous variable *size*

```
. sum size1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       size1 |      8003    12.03081    1.353498   8.021519   21.24158

. sum size1, det

                            size1
-------------------------------------------------------------
      Percentiles      Smallest
 1%     9.441483       8.021519
 5%     10.15305        8.08985
10%     10.51124       8.090587       Obs                8003
25%     11.14812       8.125039       Sum of Wgt.        8003

50%     11.89168                      Mean           12.03081
                        Largest       Std. Dev.      1.353498
75%     12.73329       20.22883
90%     13.63482       20.88608       Variance       1.831958
95%     14.35278        21.1749       Skewness       1.079542
99%     16.35735       21.24158       Kurtosis       6.349429
```

# Sample cross-properties: correlation

```
. pwcorr size1 risk1 cap chargeoffs , sig

             |    size1    risk1     cap2 charge~s
-------------+------------------------------------
       size1 |   1.0000
             |
             |
       risk1 |   0.1573   1.0000
             |   0.0000
             |
        cap2 |  -0.0675  -0.0294   1.0000
             |   0.0000   0.0092
             |
  chargeoffs |   0.1947   0.0516  -0.3876   1.0000
             |   0.0000   0.0000   0.0000
             |
```

# Time series

- So far we saw cross-sectional data of different measurement formats
- Now we will see a few times series, focusing on continuous formats
- The data belong to the same dataset and correspond to **State Street Bank and Trust Company** covering the period 2001-2009

# descriptives

```
. sum risk1, det

                            risk1
-------------------------------------------------------------
      Percentiles      Smallest
 1%     53.98246       53.98246
 5%     53.98246       57.81529
10%     53.98246       58.36076       Obs                   9
25%     58.36076       58.68864       Sum of Wgt.           9

50%     59.59463                      Mean           60.17669
                        Largest       Std. Dev.       4.24402
75%     61.25119       60.80703
90%     69.73478       61.25119       Variance       18.01171
95%     69.73478        61.3554       Skewness       1.068434
99%     69.73478       69.73478       Kurtosis       4.277709

. sum nim3, det

                            nim3
-------------------------------------------------------------
      Percentiles      Smallest
 1%     2.332453       2.332453
 5%     2.332453       2.476847
10%     2.332453       2.480285       Obs                   9
25%     2.480285       2.538968       Sum of Wgt.           9

50%     2.581661                      Mean           2.615039
                        Largest       Std. Dev.      .2158093
75%      2.60105       2.592585
90%     2.996401        2.60105       Variance       .0465737
95%     2.996401       2.935103       Skewness       .7928834
99%     2.996401       2.996401       Kurtosis       2.512343

. tab cap

group(__000 |
       006) |      Freq.     Percent        Cum.
------------+-----------------------------------
          4 |          8      100.00      100.00
------------+-----------------------------------
      Total |          8      100.00
```

# ...and cross moments

```
. pwcorr size1 risk1 nim3, sig

             |    size1    risk1     nim3
-------------+---------------------------
       size1 |   1.0000
             |
             |
       risk1 |  -0.5429   1.0000
             |   0.1309
             |
        nim3 |  -0.6282   0.6762   1.0000
             |   0.0700   0.0455
```

# Panel Data

# What and Why?

- **What:**
- Panel data are a form of longitudinal data, involving regularly repeated observations on the same individuals
- **Individuals** may be people, households, firms, countries, etc
- **Repeat observations** are typically different time periods
- **Why**:
- Repeated observations on individuals allow for possibility of isolating effects of unobserved differences between individuals
  - We can study dynamics
  - The ability to make causal inference is enhanced by temporal ordering

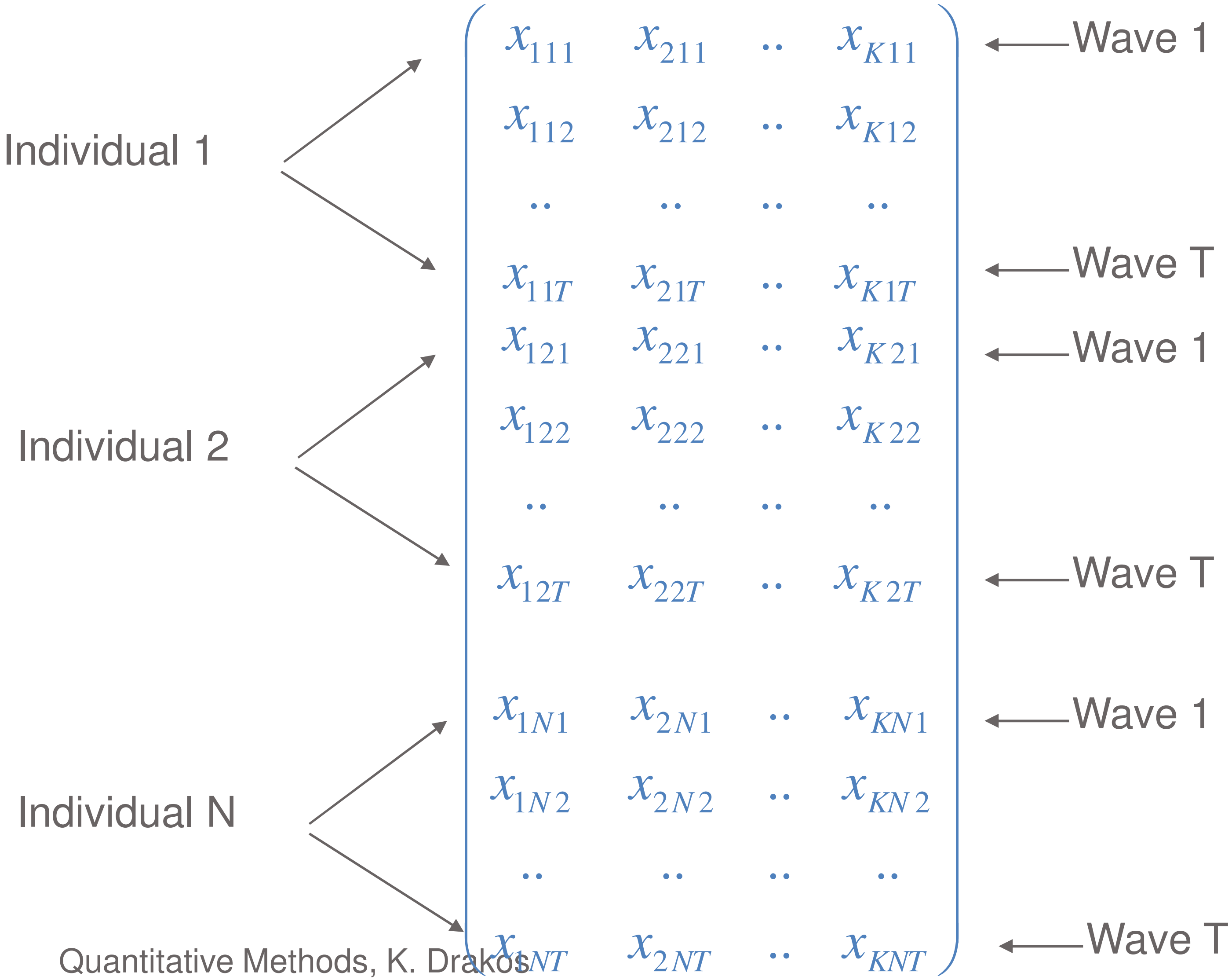
# BUT don't expect too much...

- Variation between firms (or people) usually far exceeds variation over time for a firm

  $\Rightarrow$ a panel with $T$ waves doesn't give $T$ times the information of a cross-section

- Variation over time may not exist for some important variables or may be inflated by measurement error
- We still need very strong assumptions to draw clear inferences from panels: sequencing in time does *not* necessarily reflect causation

# The Basic Data Structure



# Review of Probability and Statistics

**Empirical problem:** trading activity and stock volatility

- Research question: What is the effect on monthly stock volatility (or some other frequency) of increasing trading activity by 10 trades or by 100 thousand euros?
- We must use data to find out

# The Data Set

Average high-low price range for all stocks listed in ASE ($n$ = 426) during January 2008.

Variables:

- Average (based on daily data) high-low price range for each stock (proxy for stock return intraday volatility)
- Number of trades (trading extensive margin)
- Average size of trade = euro volume / number of trades (trading intensive margin)

# Review of Statistical Theory

1. **The probability framework for statistical inference**
2. Estimation
3. Testing
4. Confidence Intervals

**The probability framework for statistical inference**

(a) Population, random variable, and distribution
(b) Moments of a distribution (mean, variance, standard deviation, covariance, correlation)
(c) Conditional distributions and conditional means
(d) Distribution of a sample of data drawn randomly from a population: $Y_1,\ldots, Y_n$

# (a) Population, random variable, and distribution

***Population***

- The group or collection of all possible entities of interest (stocks)
- We will think of populations as infinitely large ($\infty$ is an approximation to "very big")

***Random variable Y***

- Numerical summary of a random outcome (average stock HLR, stock NTRAD)

# *Population distribution of Y*

- The probabilities of different values of $Y$ that occur in the population, for ex. $\Pr[Y = 650]$ (when $Y$ is discrete)
- or: The probabilities of sets of these values, for ex.

  $\Pr[640 \leq Y \leq 660]$ (when $Y$ is continuous).

# (b) Moments of a population distribution: mean, variance, standard deviation, covariance, correlation

***mean*** = expected value (expectation) of $Y$

= $E(Y)$

= $\mu_Y$

= long-run average value of $Y$ over repeated realizations of $Y$

***variance*** = $E(Y - \mu_Y)^2$

= $\sigma_Y^2$

= measure of the squared spread of the distribution

***standard deviation*** = $\sqrt{\text{variance}} = \sigma_Y$

# Moments, ctd.

$$\textbf{\textit{skewness}} = \frac{E\left[\left(Y-\mu_Y\right)^3\right]}{\sigma_Y^3}$$

= measure of asymmetry of a distribution

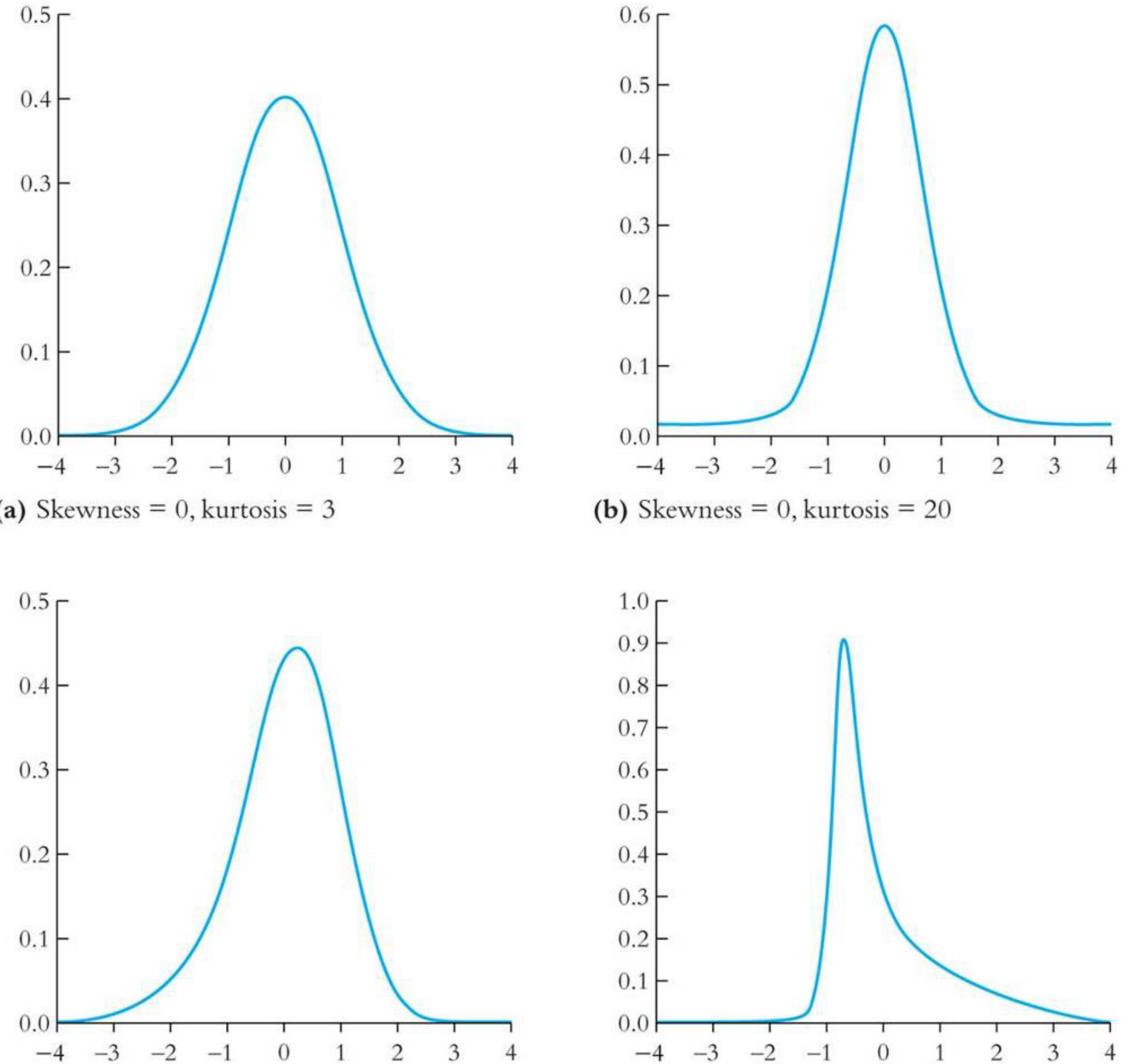
- *skewness* = 0: distribution is symmetric
- *skewness* > (<) 0: distribution has long right (left) tail

$$\textbf{\textit{kurtosis}} = \frac{E\left[\left(Y-\mu_Y\right)^4\right]}{\sigma_Y^4}$$

= measure of mass in tails

= measure of probability of large values

- *kurtosis* = 3: normal distribution
- kurtosis > 3: heavy tails ("***leptokurtotic***")



(a) Skewness = 0, kurtosis = 3

(b) Skewness = 0, kurtosis = 20

(c) Skewness = –0.1, kurtosis = 5

(d) Skewness = 0.6, kurtosis = 5

Q

# 2 random variables: joint distributions and covariance

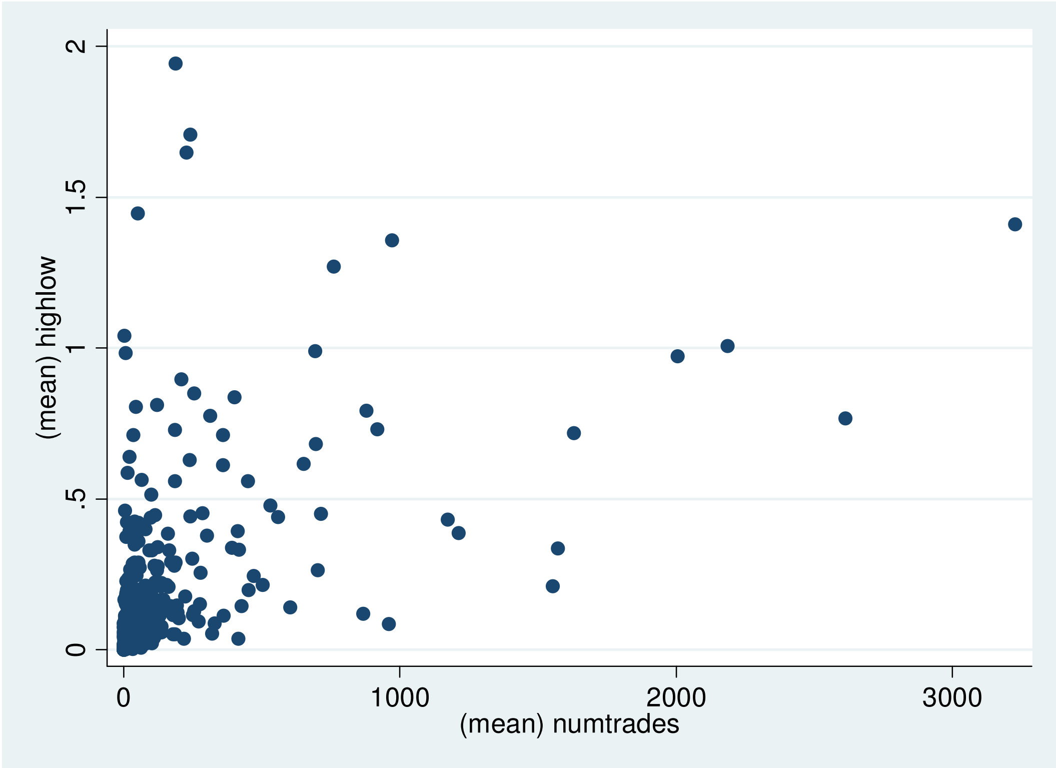
- Random variables $X$ and $Z$ have a ***joint distribution***
- The ***covariance*** between $X$ and $Z$ is

$$\text{cov}(X,Z) = E[(X - \mu_X)(Z - \mu_Z)] = \sigma_{XZ}$$

- The covariance is a measure of the linear association between $X$ and $Z$; its units are units of $X \times$ units of $Z$
- $\text{cov}(X,Z) > 0$ means a positive relation between $X$ and $Z$
- If $X$ and $Z$ are independently distributed, then $\text{cov}(X,Z) = 0$ (but not vice versa!!)
- The covariance of a r.v. with itself is its variance:

$$\text{cov}(X,X) = E[(X - \mu_X)(X - \mu_X)] = E[(X - \mu_X)^2] = \sigma_X^2$$

# The covariance between HLR and NTRAD is positive:



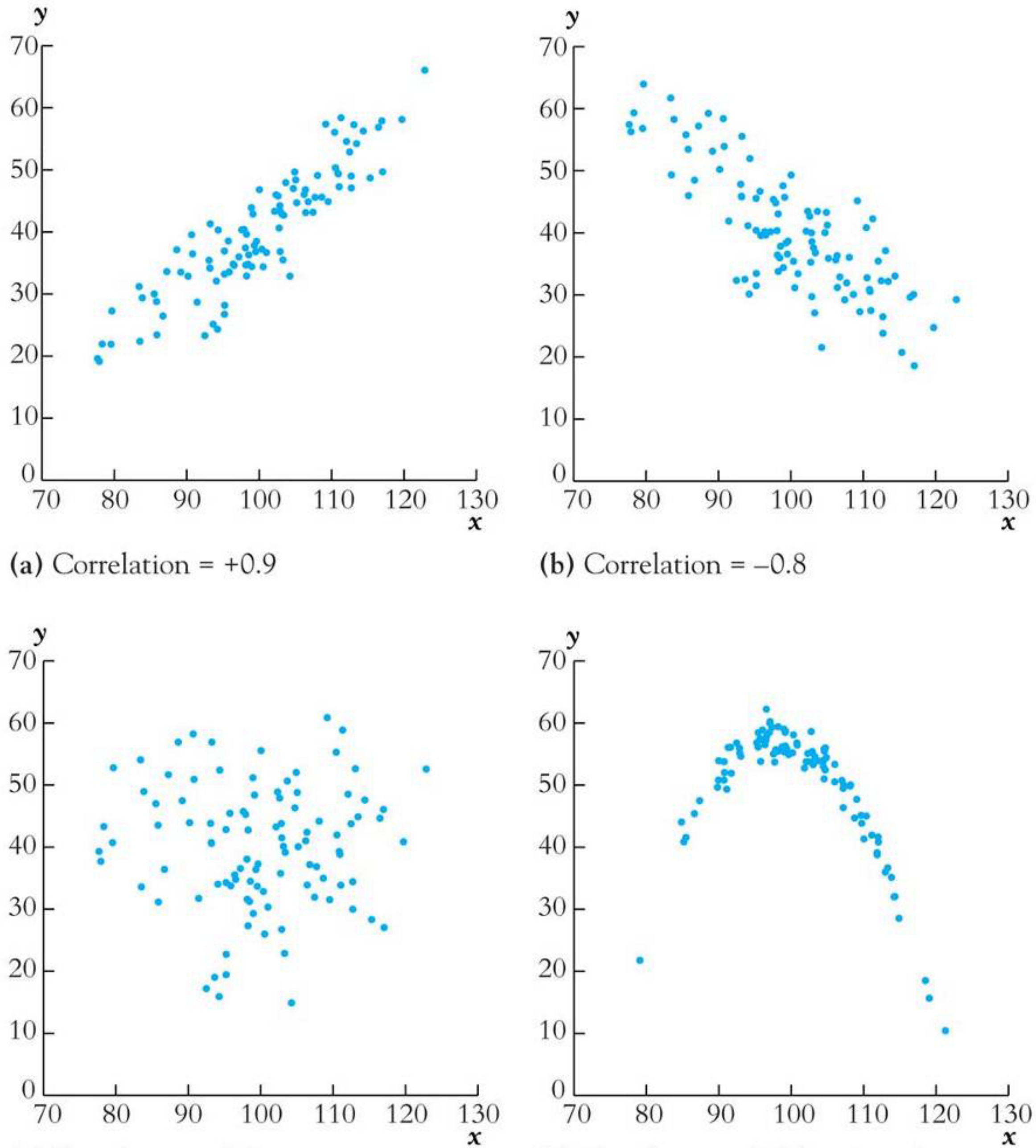

so is the ***correlation***…

The *correlation coefficient* is defined in terms of the covariance:

$$\text{corr}(X,Z) = \frac{\text{cov}(X,Z)}{\sqrt{\text{var}(X)\,\text{var}(Z)}} = \frac{\sigma_{XZ}}{\sigma_X \sigma_Z} = r_{XZ}$$

- $-1 \leq \text{corr}(X,Z) \leq 1$
- $\text{corr}(X,Z) = 1$ means perfect positive linear association
- $\text{corr}(X,Z) = -1$ means perfect negative linear association
- $\text{corr}(X,Z) = 0$ means no linear association

# *The correlation coefficient measures linear association*




(a) Correlation = +0.9

(b) Correlation = –0.8

(c) Correlation = 0.0

(d) Correlation = 0.0 (quadratic)

# Linear Regression with One Regressor

# Linear Regression with One Regressor

- Linear regression allows us to estimate, and make inferences about, *population* slope coefficients. Ultimately our aim is to estimate the causal effect on $Y$ of a unit change in $X$ – but for now, just think of the problem of fitting a straight line to data on two variables, $Y$ and $X$.

The problems of statistical inference for linear regression are, at a general level, the same as for estimation of the mean or of the differences between two means. Statistical, or econometric, inference about the slope entails:

- Estimation:
  - How should we draw a line through the data to estimate the (population) slope (answer: ordinary least squares).
  - What are advantages and disadvantages of OLS?
- Hypothesis testing:
  - How to test if the slope is zero?
- Confidence intervals:
  - How to construct a confidence interval for the slope?

# Linear Regression: Some Notation and Terminology

The ***population regression line*:**

$$\text{HLR} = \beta_0 + \beta_1 * NTRAD$$

$\beta_1$ = slope of population regression line

$$= \frac{\Delta(HLR)}{\Delta(NTRAD)}$$

= change in volatility for a unit change in the number of trades

- *Why are $\beta_0$ and $\beta_1$ "population" parameters*?
- We would like to know the population value of $\beta_1$.
- We don't know $\beta_1$, so must estimate it using data.

- In the simple linear regression of y on x, we typically refer to x as the
  - Independent Variable, or
  - Right-Hand Side Variable, or
  - Explanatory Variable, or
  - Regressor, or
  - Covariate, or
  - Control Variables

# The Population Linear Regression Model – general notation

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1,\dots, n$$

- $X$ is the ***independent variable*** or ***regressor***
- $Y$ is the ***dependent variable***
- $\beta_0$ = ***intercept***
- $\beta_1$ = ***slope***
- $u_i$ = the regression ***error***
- The regression error consists of omitted factors, or possibly measurement error in the measurement of $Y$. In general, these omitted factors are other factors that influence $Y$, other than the variable $X$

# The Ordinary Least Squares Estimator

*How can we estimate $\beta_0$ and $\beta_1$ from data?*

Recall that $\bar{Y}$ was the least squares estimator of $\mu_Y$: $\bar{Y}$ solves,

$$\min_m \sum_{i=1}^{n} (Y_i - m)^2$$

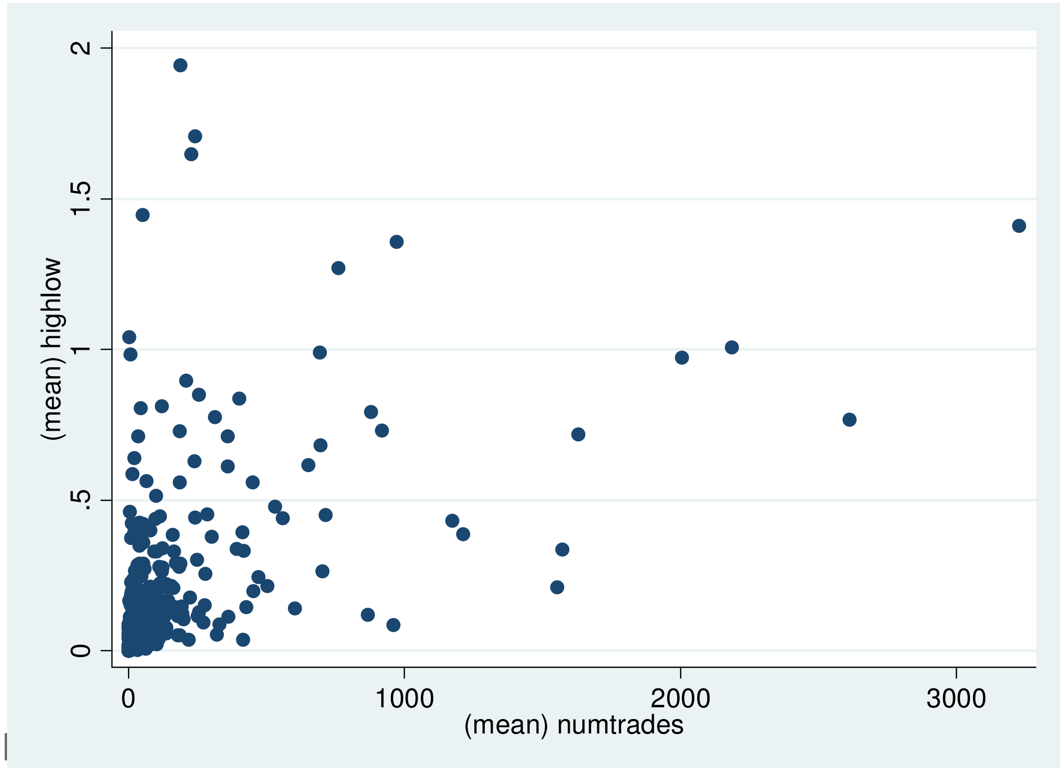
By analogy, **we will focus on the least squares ("*ordinary least squares*" or "*OLS*") estimator of the unknown parameters $\beta_0$ and $\beta_1$**, which solves,

$$\min_{b_0, b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

# Mechanics of OLS

The population regression line: HLR = $\beta_0 + \beta_1 * NTRAD$

$$\beta_1 = \frac{\Delta(HLR)}{\Delta(NTRAD)} = ??$$



# The OLS estimator solves:

$$\min_{b_0, b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

- The OLS estimator minimizes the average squared difference between the actual values of $Y_i$ and the prediction ("predicted value") based on the estimated line.
- This minimization problem can be solved using calculus
- **The result is the OLS estimators of $\beta_0$ and $\beta_1$.**

# Summary of OLS slope estimate

- The slope estimate is the sample covariance between *x* and *y* divided by the sample variance of *x*
- If *x* and *y* are positively correlated, the slope will be positive
- If *x* and *y* are negatively correlated, the slope will be negative
- Only need *x* to vary in our sample
- Intuitively, OLS is fitting a line through the sample points such that the sum of squared residuals is as small as possible, hence the term least squares
- The residual, $\hat{u}$, is an estimate of the error term, u, and is the difference between the fitted line (sample regression function) and the sample point

# THE OLS ESTIMATOR, PREDICTED VALUES, AND RESIDUALS

The OLS estimators of the slope $\beta_1$ and the intercept $\beta_0$ are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \tag{4.7}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}. \tag{4.8}$$

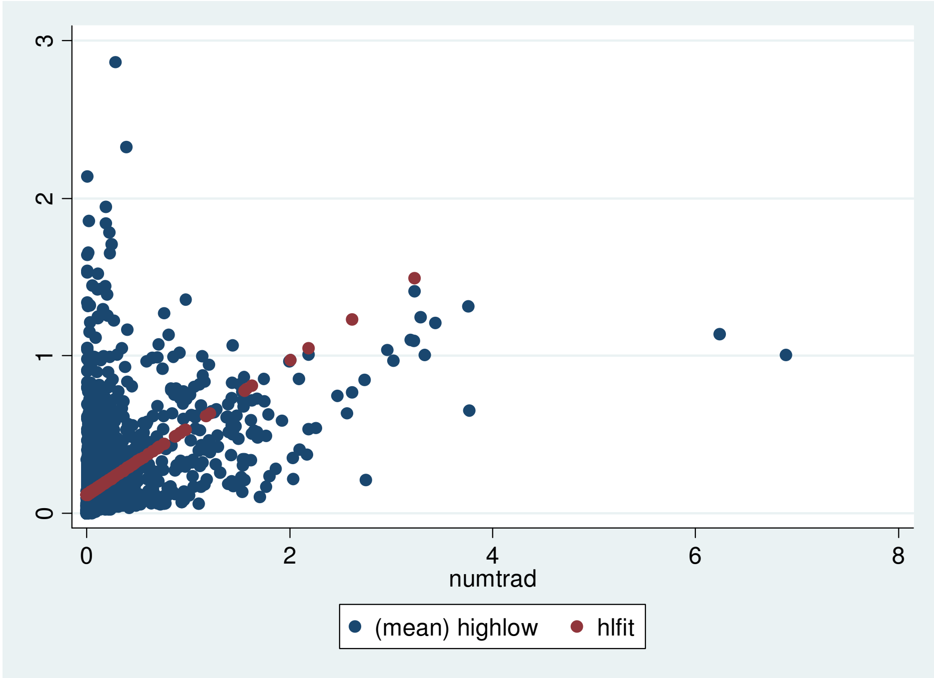
The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,\ i = 1, \ldots, n \tag{4.9}$$

$$\hat{u}_i = Y_i - \hat{Y}_i,\ i = 1, \ldots, n. \tag{4.10}$$

The estimated intercept ($\hat{\beta}_0$), slope ($\hat{\beta}_1$), and residual ($\hat{u}_i$) are computed from a sample of $n$ observations of $X_i$ and $Y_i$, $i = 1, \ldots, n$. These are estimates of the unknown true population intercept ($\beta_0$), slope ($\beta_1$), and error term ($u_i$).

# Application to the Volatility – *number of trades* data (per 1000 trades)




$$\text{Estimated slope } = \hat{\beta}_1 = 0.426$$

$$\text{Estimated intercept} = \hat{\beta}_0 = 0.117$$

$$\text{Estimated regression line: } \widehat{HLR} = 0.117 + 0.426^{*}\text{NTRAD}$$

# *Interpretation of the estimated slope and intercept*

$$\widehat{HLR} = 0.117 + 0.426 * \mathrm{NTRAD}$$

- Stocks with thousand more trades, on average, have high-low ranges that are 0.426 points higher.
- That is, $\dfrac{\Delta(HLR)}{\Delta(NTRAD)} = 0.426$
- The intercept (taken literally) means that, according to this estimated line, stocks with zero trades would have a (predicted) high-low range of 0.117
- This interpretation of the intercept makes no sense – because no trading means no volatility!!! – here, the intercept is not economically meaningful.

# Predicted values & residuals:

One of the stocks in the data set is ΑΕΓΕΚ, for which HLR = 0.052 and NTRAD = 0.320

predicted value: $\hat{Y}_{\text{ΑΕΓΕΚ}} = 0.117 + 0.426 * 0.320 = 0.2532$

residual: $\hat{u}_{\text{ΑΕΓΕΚ}} = 0.0528 - 0.2532 = -0.20052$

# OLS regression: STATA output

```
. reg highlow  numtrad if month==1

      Source |       SS       df       MS              Number of obs =     426
-------------+------------------------------           F(  1,   424) =  147.81
       Model |  7.85250502     1  7.85250502           Prob > F      =  0.0000
    Residual |  22.5248062   424  .053124543           R-squared     =  0.2585
-------------+------------------------------           Adj R-squared =  0.2568
       Total |  30.3773112   425  .071476026           Root MSE      =  .23049

------------------------------------------------------------------------------
     highlow |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     numtrad |     .42659   .0350876    12.16   0.000     .3576226    .4955573
       _cons |   .1173006   .0119639     9.80   0.000     .0937846    .1408166
------------------------------------------------------------------------------
```

(we'll discuss the rest of this output later)

# Measures of Fit

A natural question is how well the regression line "fits" or explains the data. There are two regression statistics that provide complementary measures of the quality of fit:

- The ***regression*** $\boldsymbol{R^2}$ measures the fraction of the variance of *Y* that is explained by *X*; it is unitless and ranges between zero (no fit) and one (perfect fit)

- The ***standard error of the regression*** (***SER***) measures the magnitude of a typical regression residual in the units of *Y*.

# More terminology

We can think of each observation as being made up of an explained part, and an unexplained part,

$y_i = \hat{y}_i + \hat{u}_i$ We then define the following:

$\sum (y_i - \bar{y})^2$ is the total sum of squares (SST)

$\sum (\hat{y}_i - \bar{y})^2$ is the explained sum of squares (SSE)

$\sum \hat{u}_i^2$ is the residual sum of squares (SSR)

Then $\text{SST} = \text{SSE} + \text{SSR}$

# Proof that SST = SSE + SSR

$$\sum (y_i - \bar{y})^2 = \sum [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2$$

$$= \sum [\hat{u}_i + (\hat{y}_i - \bar{y})]^2$$

$$= \sum \hat{u}_i^2 + 2\sum \hat{u}_i (\hat{y}_i - \bar{y}) + \sum (\hat{y}_i - \bar{y})^2$$

$$= \text{SSR} + 2\sum \hat{u}_i (\hat{y}_i - \bar{y}) + \text{SSE}$$

and we know that $\sum \hat{u}_i (\hat{y}_i - \bar{y}) = 0$

**The *regression* $R^2$** is the fraction of the sample variance of $Y_i$ "explained" by the regression.

$Y_i = \hat{Y}_i + \hat{u}_i$ = OLS prediction + OLS residual

$\Rightarrow$ sample var ($Y$) = sample var($\hat{Y}_i$) + sample var($\hat{u}_i$) (*why?*)

$\Rightarrow$ total sum of squares = "explained" SS + "residual" SS

*Definition of* $R^2$:

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{\hat{Y}})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- $R^2 = 0$ means $ESS = 0$
- $R^2 = 1$ means $ESS = TSS$
- $0 \leq R^2 \leq 1$
- For regression with a single $X$, $R^2$ = the square of the correlation coefficient between $X$ and $Y$

# The Standard Error of the Regression (SER)

The *SER* measures the spread of the distribution of $u$. The *SER* is (almost) the sample standard deviation of the OLS residuals:

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}})^2}$$

$$= \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

(the second equality holds because $\bar{\hat{u}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0$).

$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

The *SER*:

- has the units of *u*, which are the units of *Y*
- measures the average "size" of the OLS residual (the average "mistake" made by the OLS regression line)
- The ***root mean squared error*** (*RMSE*) is closely related to the *SER*:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}$$

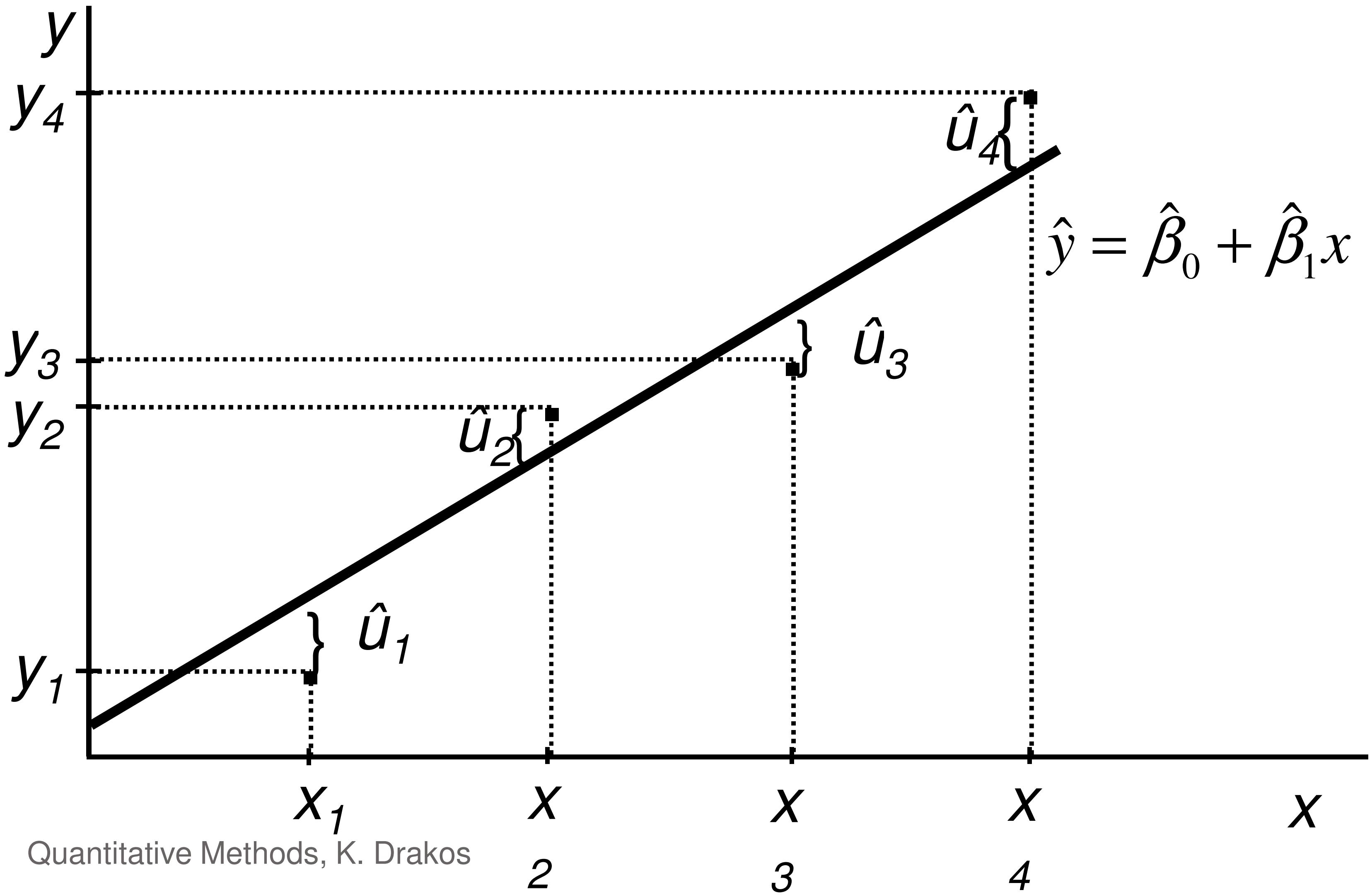
This measures the same thing as the *SER* – the minor difference is division by $1/n$ instead of $1/(n–2)$.

*Technical note*: why divide by $n$–2 instead of $n$–1?

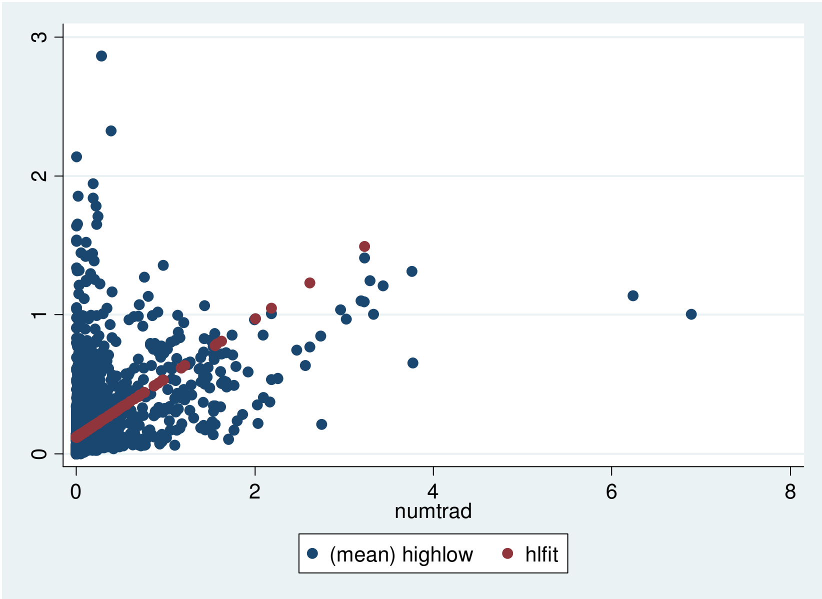
$$SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

- Division by $n$–2 is a "degrees of freedom" correction – just like division by $n$–1 in $s_Y^2$, except that for the *SER*, two parameters have been estimated ($\beta_0$ and $\beta_1$, by $\hat{\beta}_0$ and $\hat{\beta}_1$), whereas in $s_Y^2$ only one has been estimated ($\mu_Y$, by $\bar{Y}$).
- When $n$ is large, it makes negligible difference whether $n$, $n$–1, or $n$–2 are used – although the conventional formula uses $n$–2 when there is a single regressor.

# Sample regression line, sample data points and the associated estimated error terms



# Example of the $R^2$ and the $SER$




$$\widehat{HLR} = 0.117 + 0.426*\text{NTRAD}$$

$$R^2 = .2585, \; SER = 0.2299$$

*NTRAD explains only 25% of the variation in high-low ranges. Does this make sense? Does this mean the NTRAD is unimportant for volatility?*