

## אירוע ניסוי

המחקר להלן נועד לאמוד את השפעת הדיוור הישיר כחלק ממסע פרסום שנועד להניע את הציבור להתנהגות אחראית בנושא מניעת מחלות של סרטן עור הנובע מנזקי קרינה.

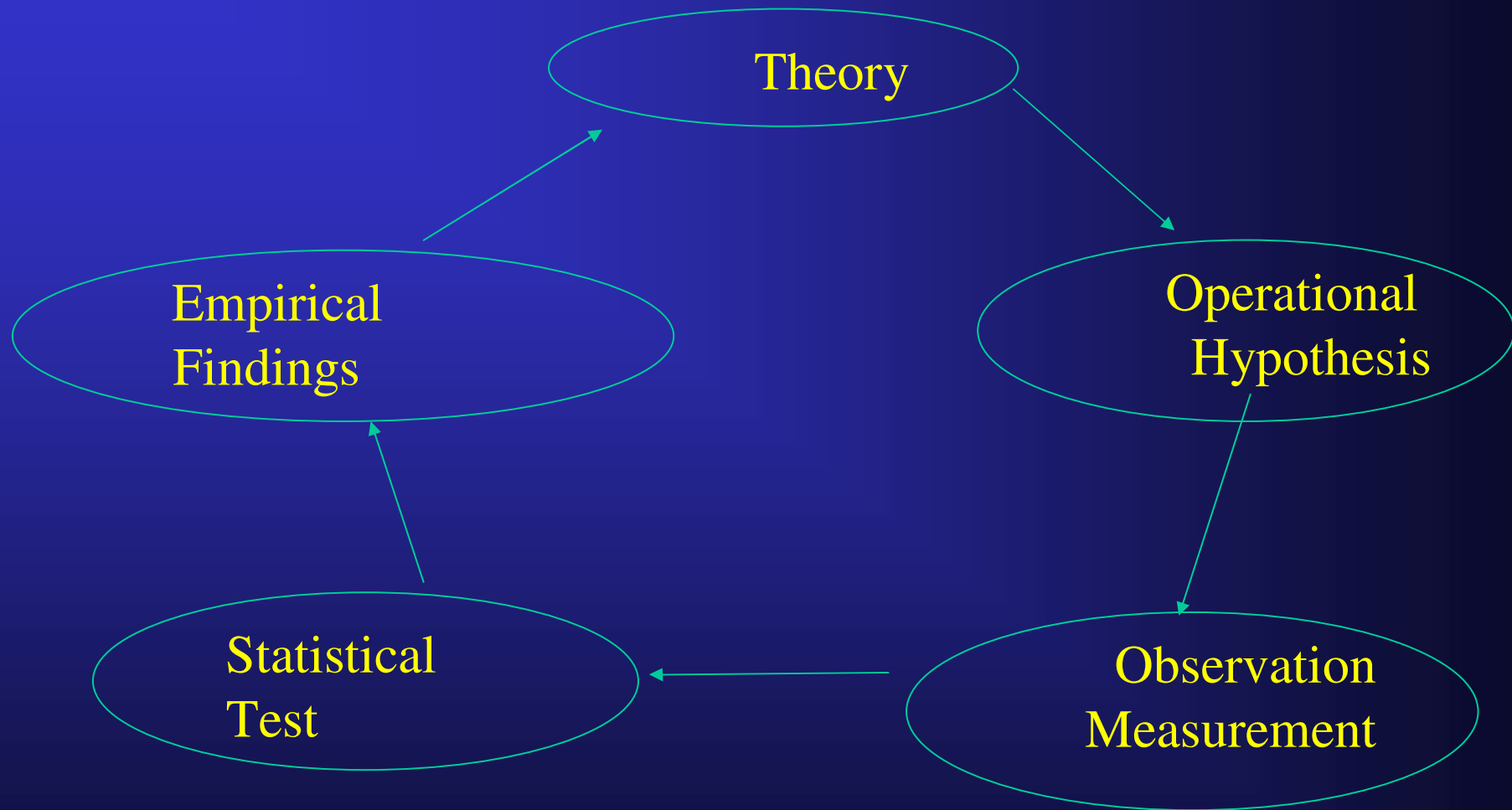
הנשאלים חולקו לשלוש קבוצות אקראיות בנות 1000 כל אחת. קבוצה אחת שימשה כקבוצת ביקורת ולא טופלה כלל. קבוצה שנייה קיבלה דיוור מוכוון הפחדה. קבוצה שלישית קיבלה דיוור מוכוון עובדות (ציטוט מאמרים, נתונים וכדומה).

ארבעה שבועות לאחר הדיוור התקשרו אל הנשאלים כדי להעריך את רמת הידע שלהם לגבי הגורמים, הסימפטומים ואמצעי המניעה של סרטן עור.

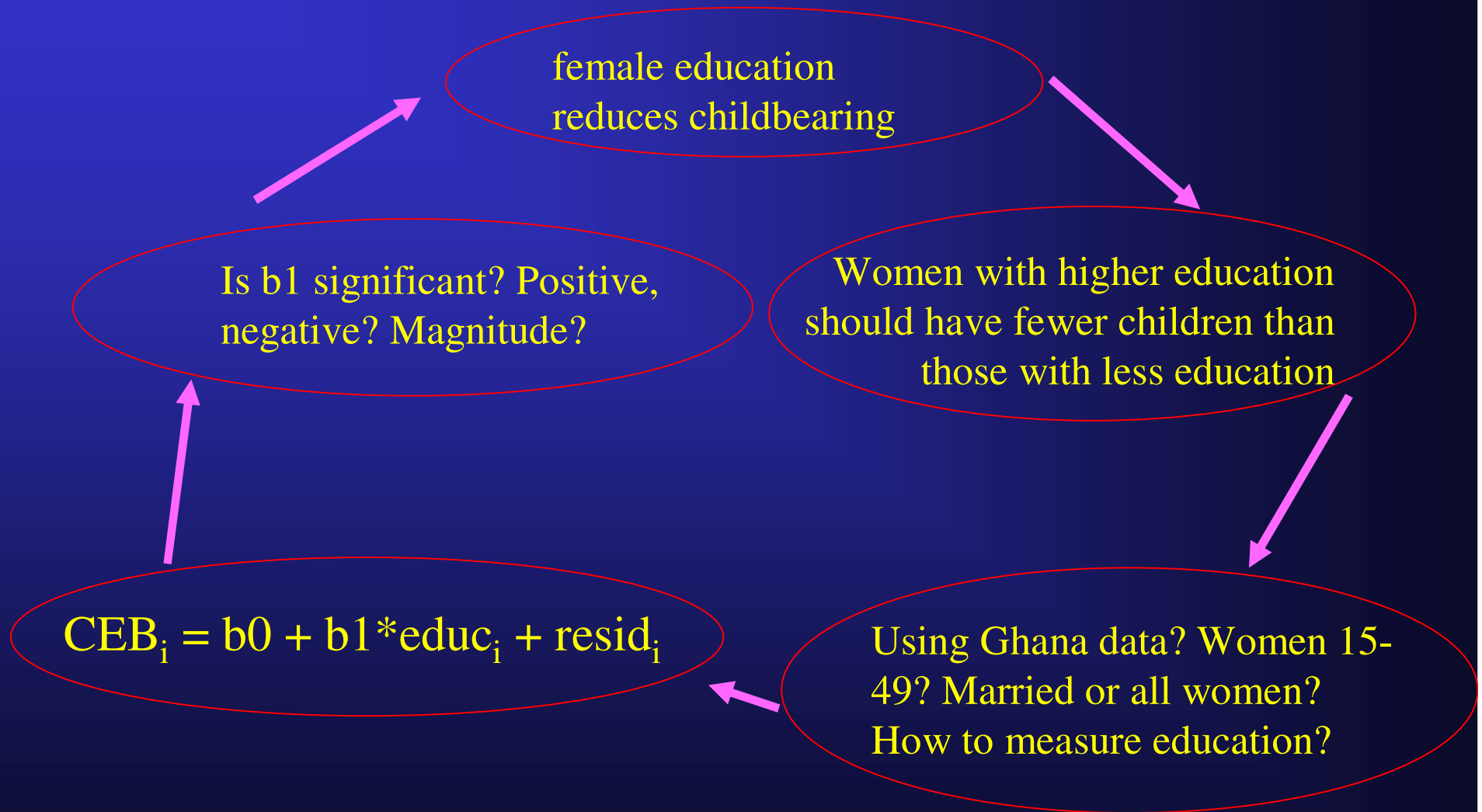
## אירוע ניסוי - שאלות

- 1) מהו המשתנה הבלתי תלוי?
- 2) כיצד הוא תופעל?
- 3) מהו המשתנה התלוי?
- 4) כיצד הוא נמדד?
- 5) אילו משתנים אחרים יכולים לערער על תקיפות הממצאים?
- 6) כיצד הם נשלטו?
- 7) איך היית אתה שולט עליהם?
- 8) כיצד היית מנתח את התוצאות?
- 9) איך היית מעצב את הניסוי?

# The traditional *scientific* approach



# Example of our approach



# Correlation

- Pearson product moment correlation coefficient:

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

- Call this  $r$ :  $-1 \leq r \leq +1$
- perfect negative correlation:  $-1$
- perfect positive correlation:  $+1$

## Very simple examples

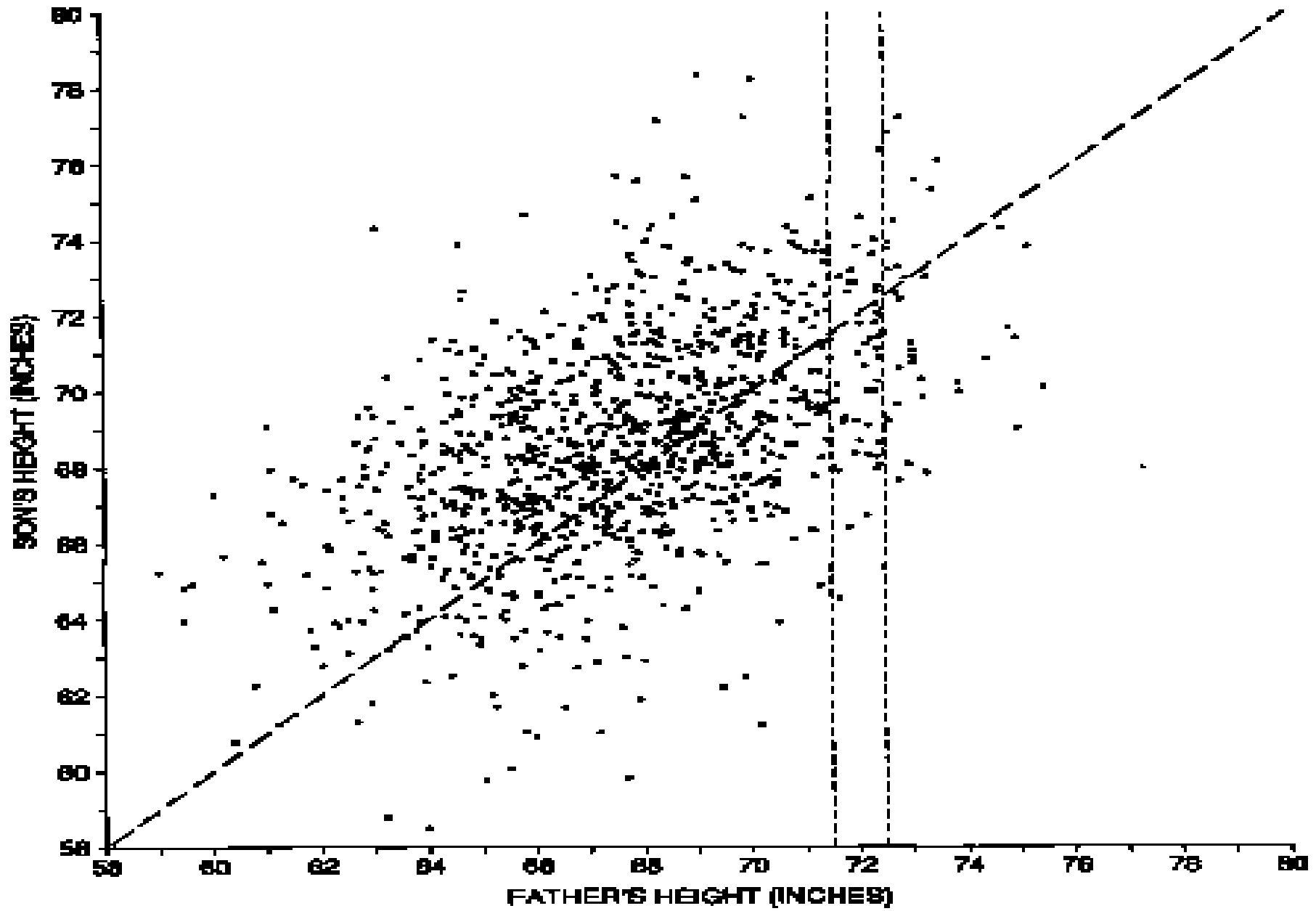
- First, an example to calculate covariance to do “by hand” or using excel:

student	grade	workhrs
1	50	1
2	65	2
3	55	3
4	80	4
5	75	5

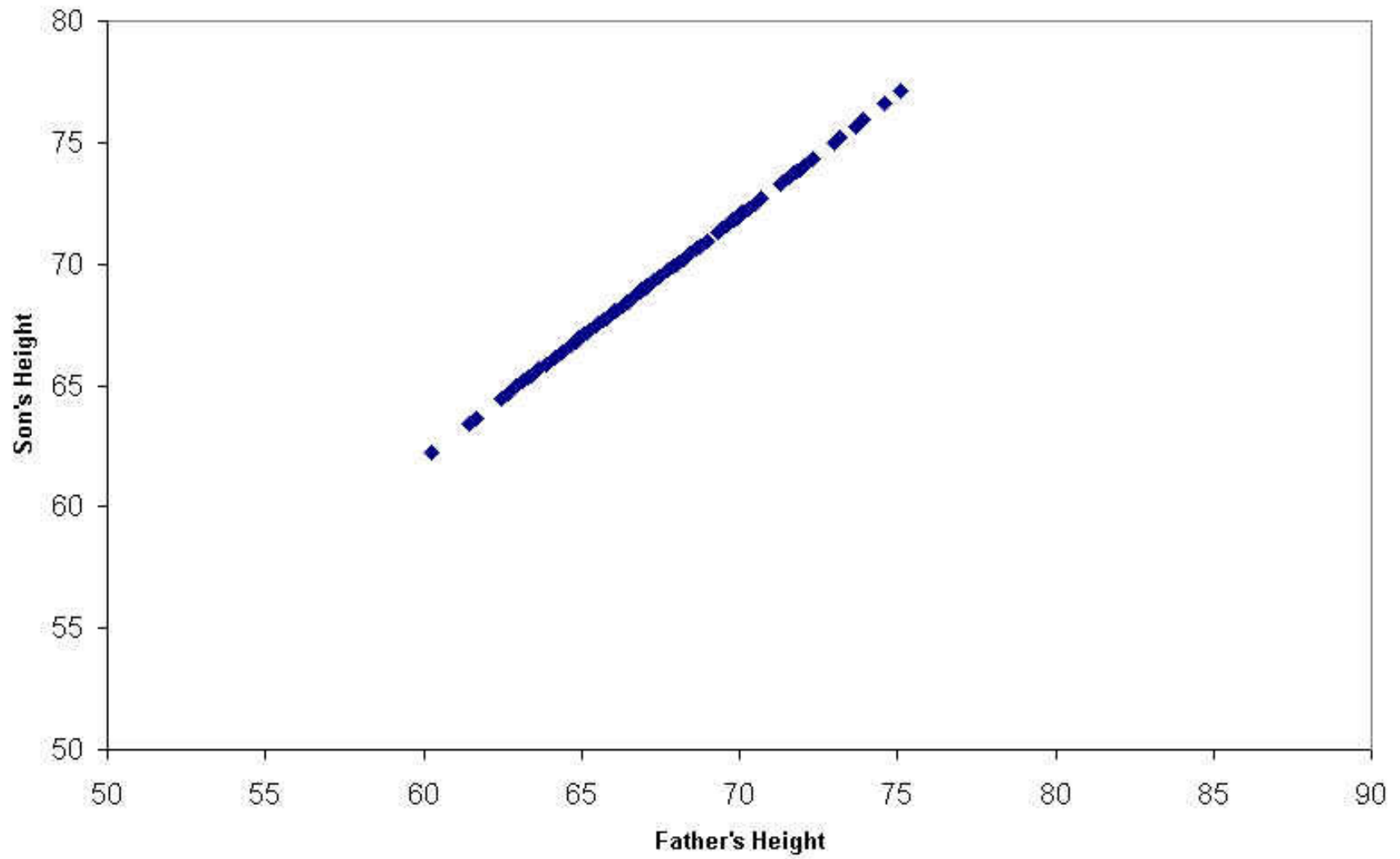
- look at the data in scatterplot
- Calculate  $r$

## Real historical empirical question: The analysis of heredity

- Statisticians in Victorian England concerned with quantifying heredity.
- Question is whether the height of a father determines the height of the son.
- Data on 1,078 father-son heights based on Pearson's data from last century.
- Galton's main question: Is there a relationship between father and son heights?
- Data organized as list with two columns. Hard to read or interpret
- Each point gives us father-son pair heights
- Use scatter diagram to examine data. (see handout)
- Is heredity the only possible explanation?



Height of Fathers and Sons



## Relating height and weight

We have data for 988 men between the ages of 18 and 24  
(from the Health and Nutrition Examination Survey)

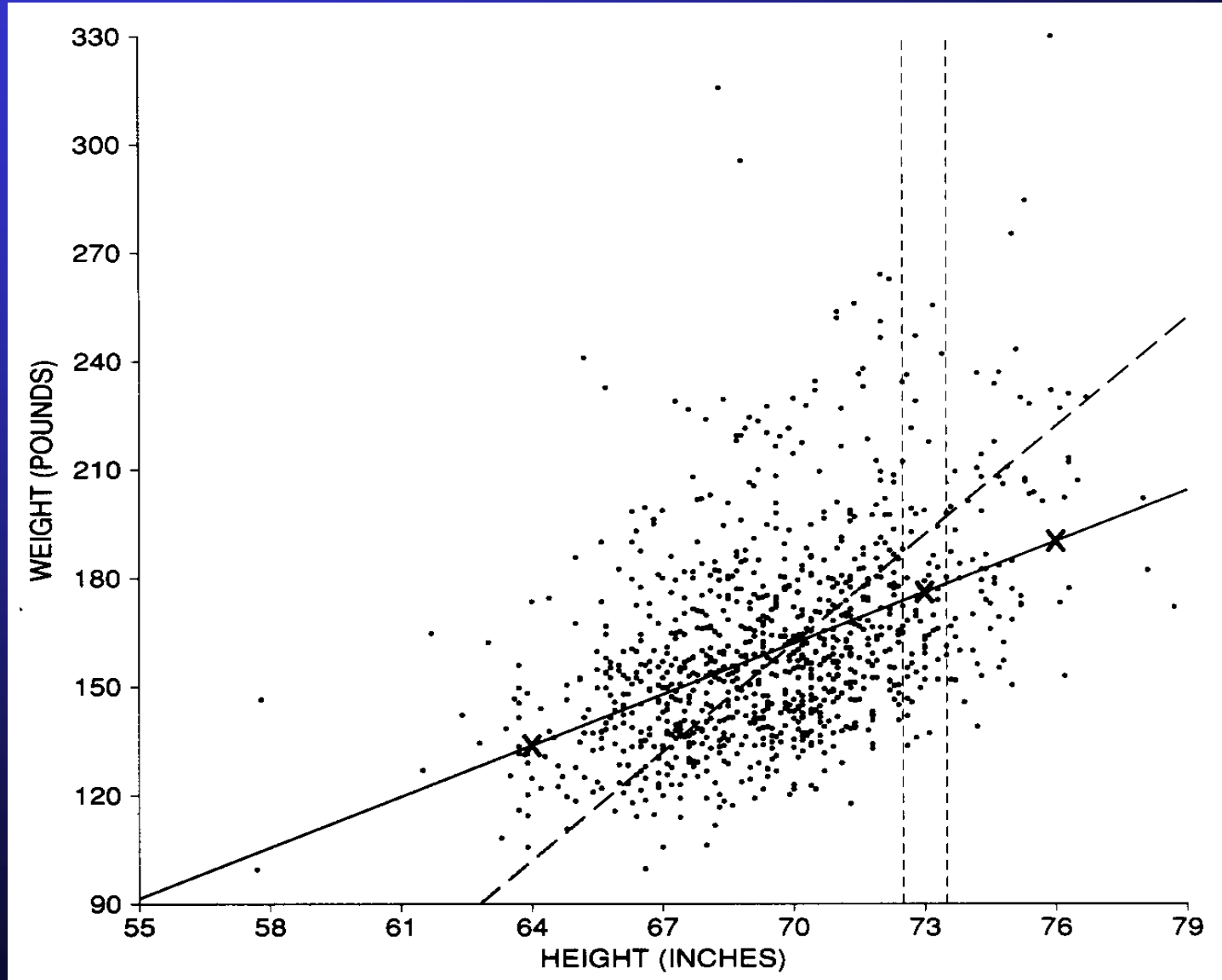
Height: mean=70 inches, SD=3 inches

Weight: mean=162 pounds, SD=30 pounds

Correlation,  $r = 0.47$

Scale of graph chosen so that one SD for height or weight  
covers same distance.

**Figure.** Scatter diagram for the heights and weights of 988 men ages 18-24 from the Health and Nutrition Examination Survey from the US.



# Leaving the world of statistics: the straight line

- How to relate the values in  $Y$  to values in  $X$
- The simplest method is a straight line:  
 $Y=a+bX$
- slope,  $b=dY/dX$
- intercept: value of  $Y$  when  $X=0$
- Why is this important?
  - Allows us a way to predict and test values of  $Y$  for values of  $X$ .

# Describing the education-wage relationship

- Wages= $Y$
- Education= $X$

$$Y=4+1.3*X$$

- What does this say about?
  - Wages for someone with 0 education?
  - Affect of 1 year of education?
  - Earnings of people with 0 versus 16 years?
    - » 4 versus 24.8

# A statistical model for the linear relationship

- A statistical model includes “unknown” error term, e;

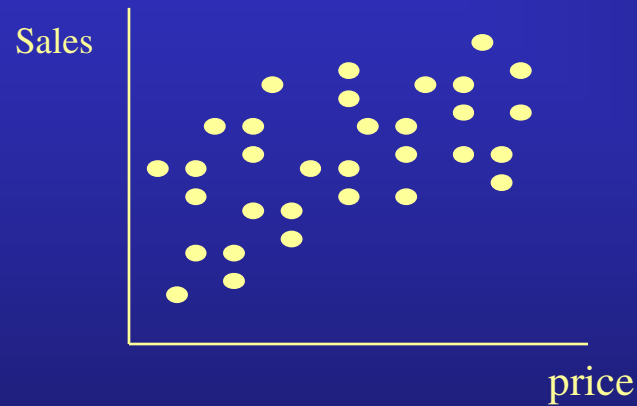
$$Y = a + b * X + e$$

- Always remember - e includes everything else - not just random error but all other factors that influence Y.
- Our task is to find the coefficients (a and b) to the line that BEST fit the data

- WHAT DOES IT MEAN “BEST”?

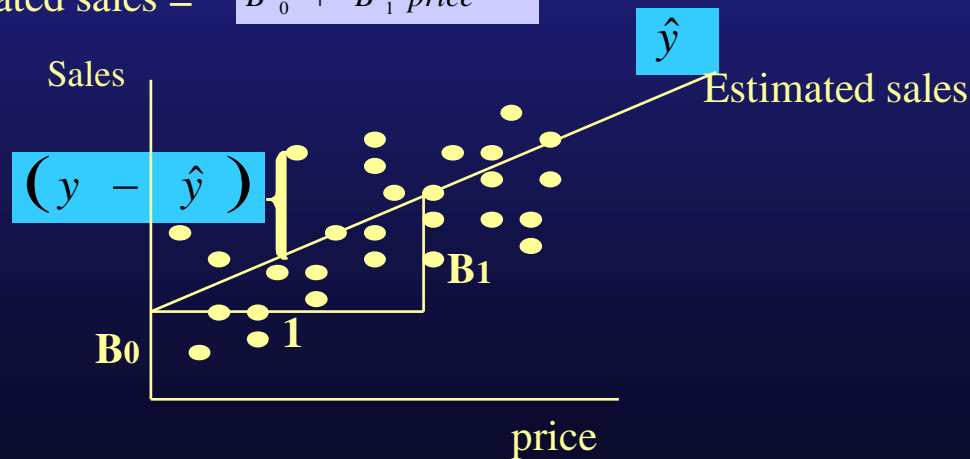
# Linear Regression

## The Basic Concept



- Sales - dependent variable or criterion variable (y).
- Price - Independent variable or predictor variable (x).

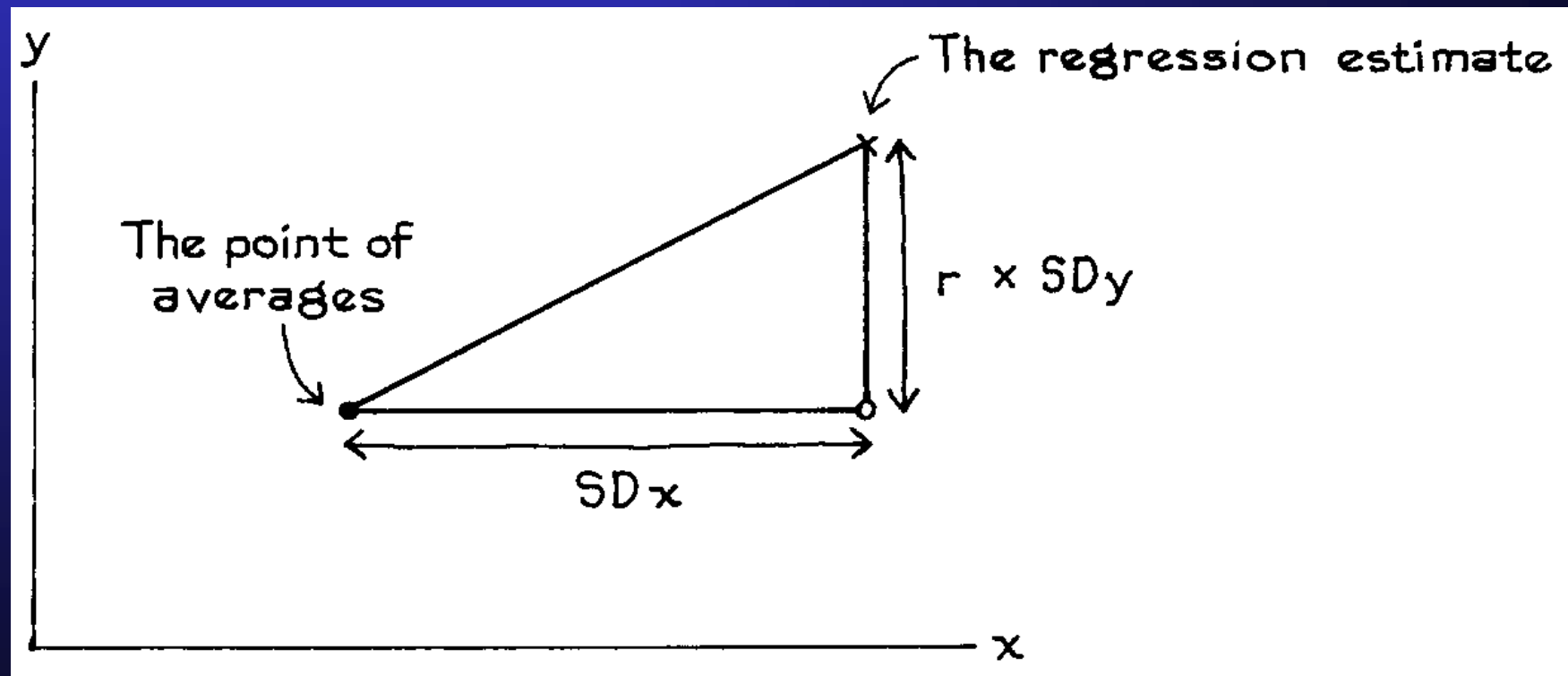
Estimated sales =  $B_0 + B_1 \text{ price}$



# The slope of the regression line

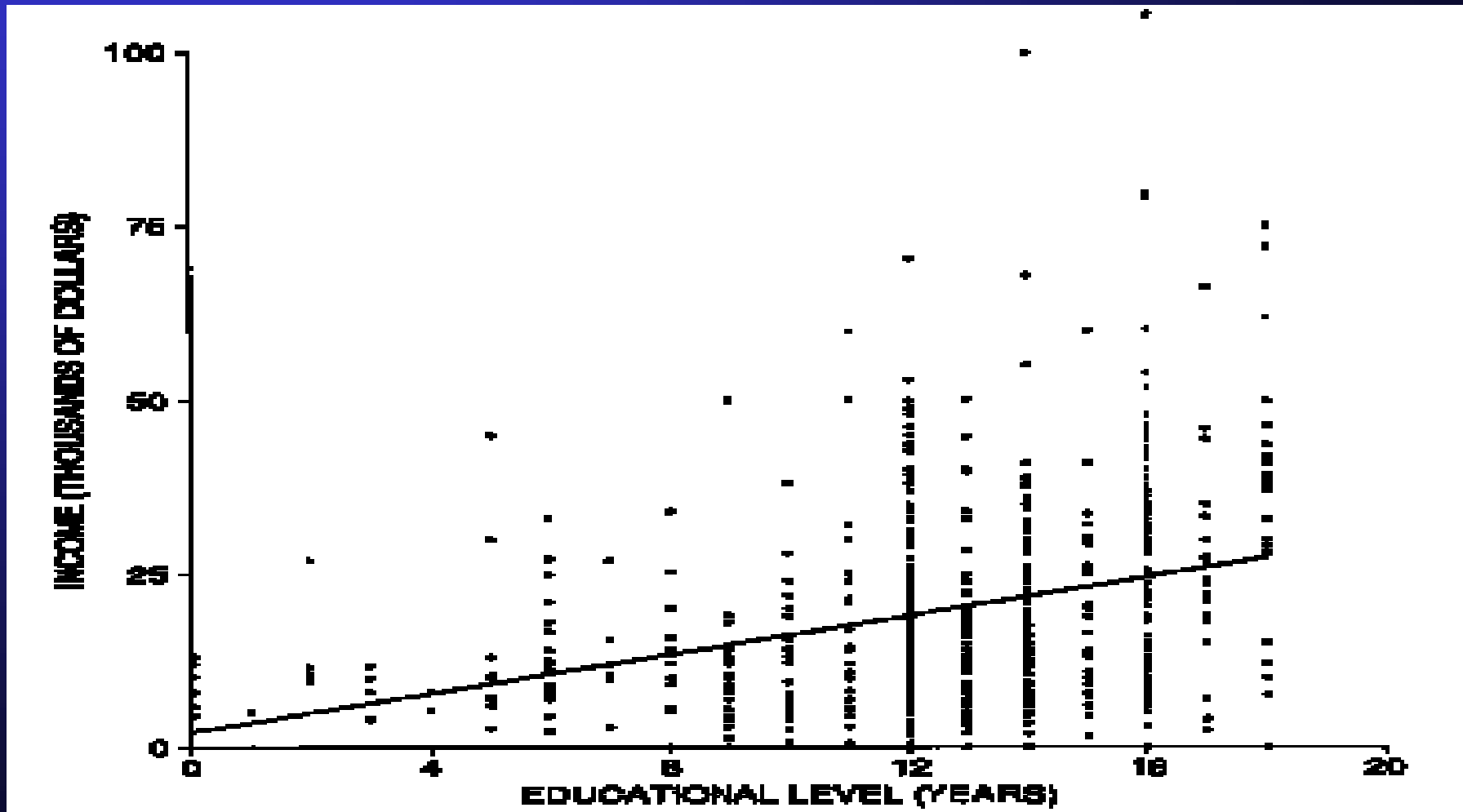
The rule for relating the heights and weights is,

**On average, each increase of 1 SD in X there is an increase of only  $r$  SDs in Y**



# Regression Example

Figure 1 shows the relationship between income and education, for a representative sample of 637 California men age 25-29 in 1988. average education is 12.5 years, SD is 4 years; average income is \$19,700, SD is \$16,000,  $r = 0.35$ .



## The standard approach for finding the best line

- No line will ever be perfect in social research.
- How do we judge between different lines with different errors?
- Favorite and simplest approach is *least squares criterion*
  - choose coefficients that make sum of squared prediction errors as small as possible.
  - Which deviation do we minimize?
  - Why do we square them?

# Estimation of the straight line parameters

- 1) Graphical eye ball: Pick the line that looks best. The problems appear when there are more than one predictor in the model.
- 2) Find a line that optimizes some measure of a good fit. The most common is the least squares estimation in which the sum of the squared differences between the predicted criterion and the observed data is minimized.

$$\left( \sum (y - \hat{y})^2 \right) \longrightarrow \min$$



$$B_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$
$$B_0 = \bar{y} - B_1\bar{x}$$

Where  $\bar{x}$ ,  $\bar{y}$  are averages of the sample

The standard error of estimates is a typical measure of the deviation of the predictors from the actual values of the dependent variable (analogous to standard deviation).

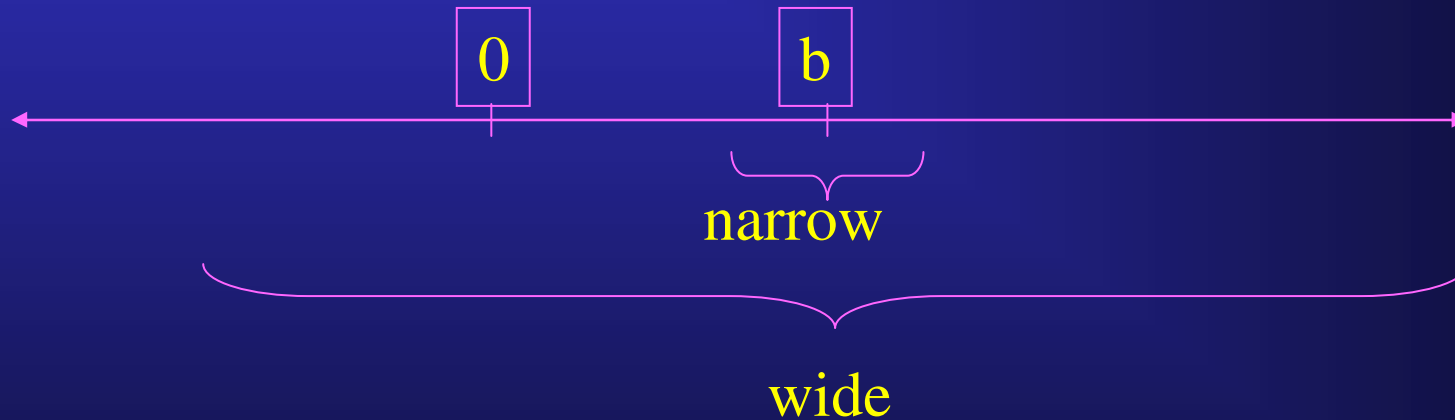
$$S_{y,x} = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}} = \sqrt{\frac{\sum (y - \bar{y})^2 - B_1 \sum (x - \bar{x})(y - \bar{y})}{n-2}}$$

# Testing coefficient significance

- What “confidence” do we have in value we obtain?
- OR, how likely are these values to be simply due to chance?
- The  $b$ 's are estimated from sampled data, they are estimates like the sample mean with properties.
- We can draw confidence intervals for any sample estimate and we do this here.
- Then we will use statistical testing to determine the likelihood of result by chance

# Testing coefficient significance

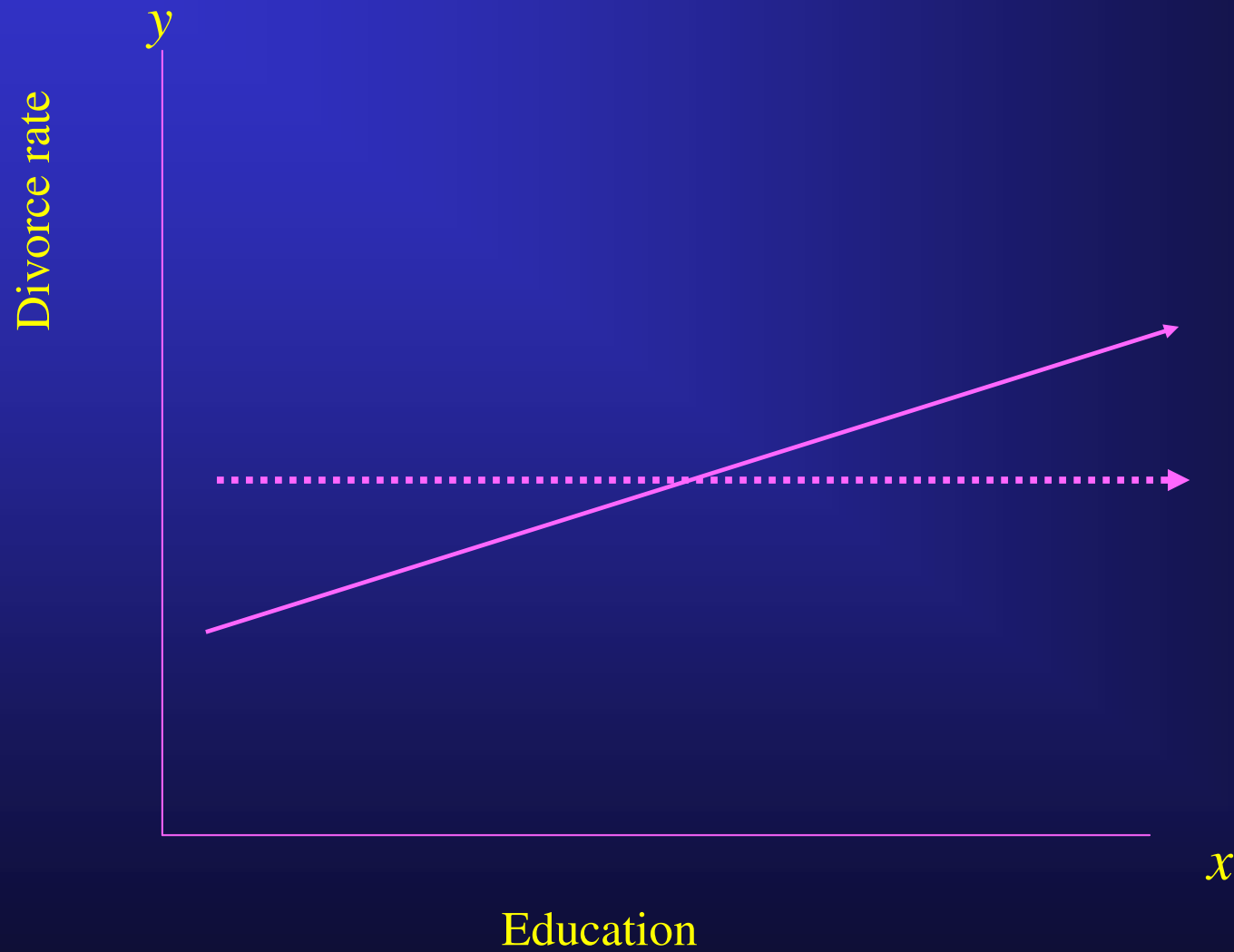
- Remember confidence intervals? The boundaries within which we BELIEVE our values are likely to fall.



# Testing coefficient significance

- Hypothesis testing:
  - Define null hypothesis as  $B=c$  where  $c$  is usually but not always 0
  - Define alternative hypothesis as  $B \neq c$
  - Normally do 2-sided test except where we have strong theoretical basis for a priori of coefficient.
- Example:
  - Effect of advertising budget on sales rate. Want to test whether lower budget leads to lower sales rates.

# Testing coefficient significance

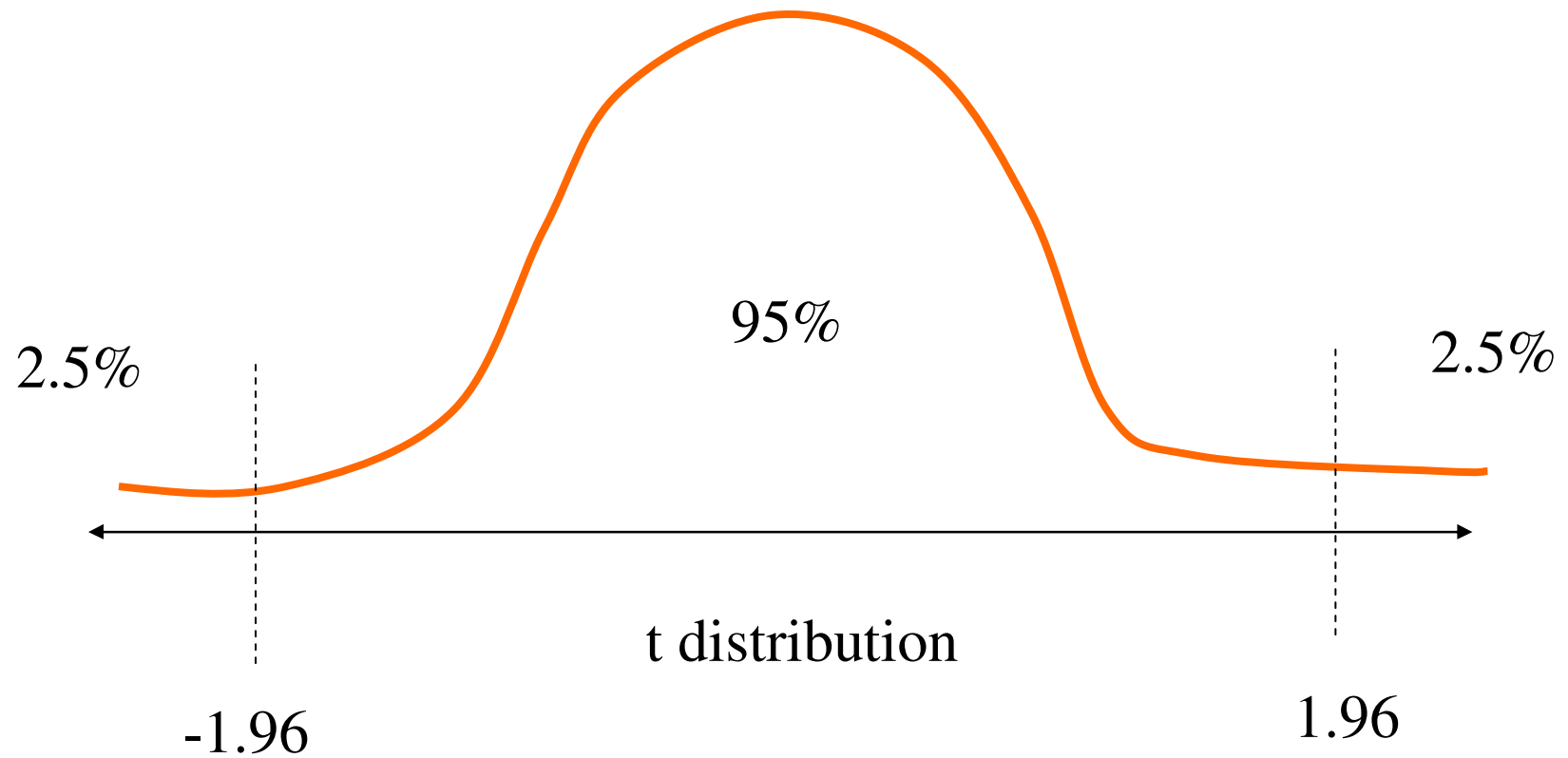


## Testing coefficient significance

$$SE(b) = \frac{SD(y)}{SD(x)} \cdot \sqrt{\frac{1 - r_{xy}^2}{n - 2}}$$

$$t_c = \frac{b - H_0(b)}{SE(b)}$$

# Testing coefficient significance



# The coefficient of determination

The correlation coefficient  $R^2$  measures the closeness of the relation between the predicted and the actual values of the dependent variable.

$$R = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2} \sqrt{\sum(y - \bar{y})^2}} = \frac{\sum xy - n\bar{x}\bar{y}}{nS_x S_y}$$

More directly:

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

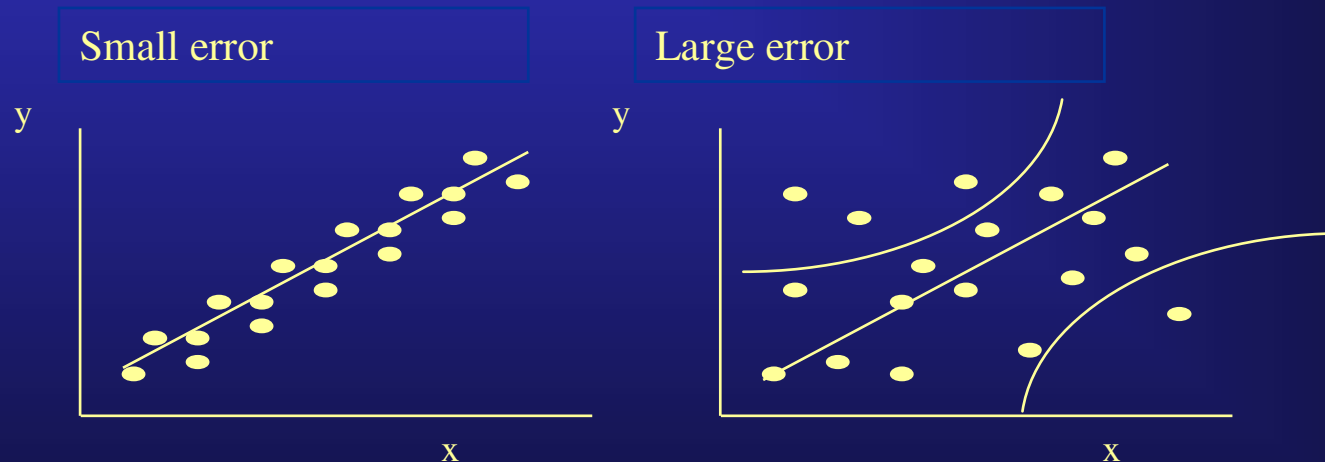
$R^2 = 1 - \text{unexplained variance in } y / \text{total variance in } y =$   
 $\text{percent of the variance in } y \text{ explained by } x$

# Interpretation of the regression results

$B_0$  - most of the cases has no meaning or not relevant (is there a zero price?)

$B_1$  - The slope of the regression line is interpreted as the amount y would increase if x is increased in one unit. In case of sales Vs. price the slope is expected to be negative in most of the cases.

**The standard Error of estimates** - An important use of regression is forecasting. The standard error is a measure for the range around the line of the data points:

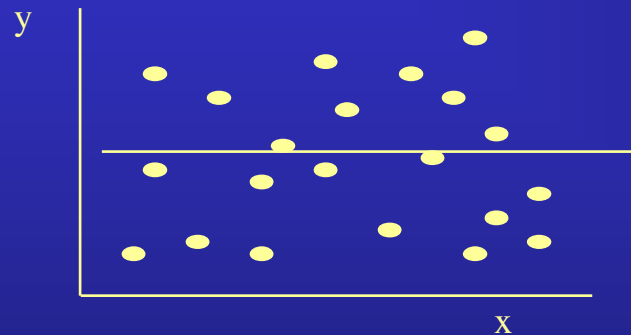


The prediction is different in these two cases even if the slope is the same, for a determined confidence level a forecasting value ( $y'$ ) within a range can be computed (for each  $x'$ ):

$$y'(x') = \hat{y} \pm t_{n-2, \alpha} s_{y,x} \sqrt{1 + \frac{1}{n} + \frac{(x' - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

# Interpretation

**The coefficient of determination ( $R^2$ )** - This is the index of a fit between the predicted and the actual values of the dependent variable. 1 indicates a perfect correlation, meaning that  $x$  is a perfect predictor of  $y$ . 0 means that there is no fit at all and  $x$  can not predict anything.



Home assignment: Use the following data about Motor vehicles registration (in USA, Taking from [1]) to create a regression model. What is a the confidence level of this model?

<u>Year</u>	<u>Registrations (millions)</u>
1	63.2
2	65.8
3	68.8
4	71.7
5	74.9
6	77.8
7	80.0
8	83.2

# Multiple Regression

The basic concepts - using several predictors:

$$y = B_0 + B_1x_1 + B_2x_2 + \dots + B_kx_k + e$$

## Estimations

•The coefficients are derived by using computer programs, the most common method is the “least squares”

•The squared error of the estimate is

$$S_{y,x} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - k}}$$

•Coefficient of determination is:  $R^2 = 1 - \frac{S_{y,x}^2}{S_y^2}$  = 1-unexplained var/ Total Var = percent of the total variance in y explained by x

Q: Why not run a regression for each predictor?

# Interpretation

The regression coefficient  $B_i$  is interpreted as the amount by which  $y$  will change if  $x_i$  is changed by one unit and all the other predictors remain constant.

$$\text{Sales} = 2 + 1.5 (\text{advertising budget}) - 0.5 (\text{price})$$

The standard error of estimate is interpreted as before.

$R^2$  is a measure for the closeness between the predictors and the dependent variable. For a special case with one predictor it is equal to the correlation coefficient. **Since  $R^2$  is based on the data used to construct the model additional tests are recommended (e.g. “split half” etc.)**

Assume error of 10, and advertising budget of 100\$ price of 50\$.

- What will be a “perfect prediction?”
- What would be the confidence interval if we assume a large sample (high confidence)?.

Recall that

$$C.I = y \pm t_{n-k, \alpha} (S_{y,x})$$

# The confidence of the model

The B's are subject to error. We assume that the errors are normally distributed, hence, we can estimate the "true value of each Bi:  $B_i \pm t_{n-k, \alpha} [STD (B_i)]$

True Bi = Estimated value of

The important question is whether the true Bi is significantly different from zero. In other words does the ith variable really influences the independent variable?

The computer programs use either an F test or a t test. Failing to be significantly different from 0 means that variable i may have no influence, It is recommended to remove this variable unless the prediction power of the regression equation is reduced dramatically.

The statistics involve for the hypothesis that Bi=0:

t-test statistic  $\frac{B_i - 0}{S_{B_i}}$

F test statistic  $\left( \frac{B_i}{S_{B_i}} \right)^2$

# The confidence of the model

The tests on  $B_i$  coefficients address the question whether a particular variable improves prediction. A test on  $R^2$  address the question of whether the predictors as a group are significantly related to the dependent variable. The test is based on Anova (F test):

$$\frac{\frac{\text{Explained S.S}}{k}}{\frac{\text{Unexplained S.S}}{n-k-1}} = \frac{\frac{\sum(y-\bar{y})^2 - \sum(y-\hat{y})^2}{k}}{\frac{\sum(y-\bar{y})^2}{n-k-1}} = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}} \approx F_{k,n-k-1}$$

**When the F test on R is insignificant the entire regression is essentially worthless.**

Adjusted  $R^2$ :

The more independent variables you have the higher  $R^2$  will be. A better criterion that takes the number of the independent variables into account exists:

$$\text{Adjusted } R^2 = 1 - \left(1 - \text{unadjusted } R^2\right) \frac{n-1}{n-k-1}$$

## Which of the the predictors are the most useful predictors?

Below are the basic approaches that are used to answer this question:

- **The absolute size of the regression coefficient**: Larger coefficient is more important. The problem with this approach is that the scales may be different.
- **The Beta Coefficient**: This coefficient would have been obtained if the regression had been performed on standardized variables (mean 0 STD 1).  $\text{Beta} = B(\text{STD of independent variable}) / (\text{STD dependent Variable})$ . Higher beta implies for higher influence
- **Elasticity**: The percent of change in the dependent variable for 1% change in the predictor.

$$\text{Elasticity} = \frac{\Delta y / y}{\Delta x / x} = \frac{\Delta y \times x}{\Delta x \times y} = B \frac{x}{y} \xrightarrow{\text{linearty}} \text{Elasticity} = B \frac{\bar{x}}{\bar{y}}$$

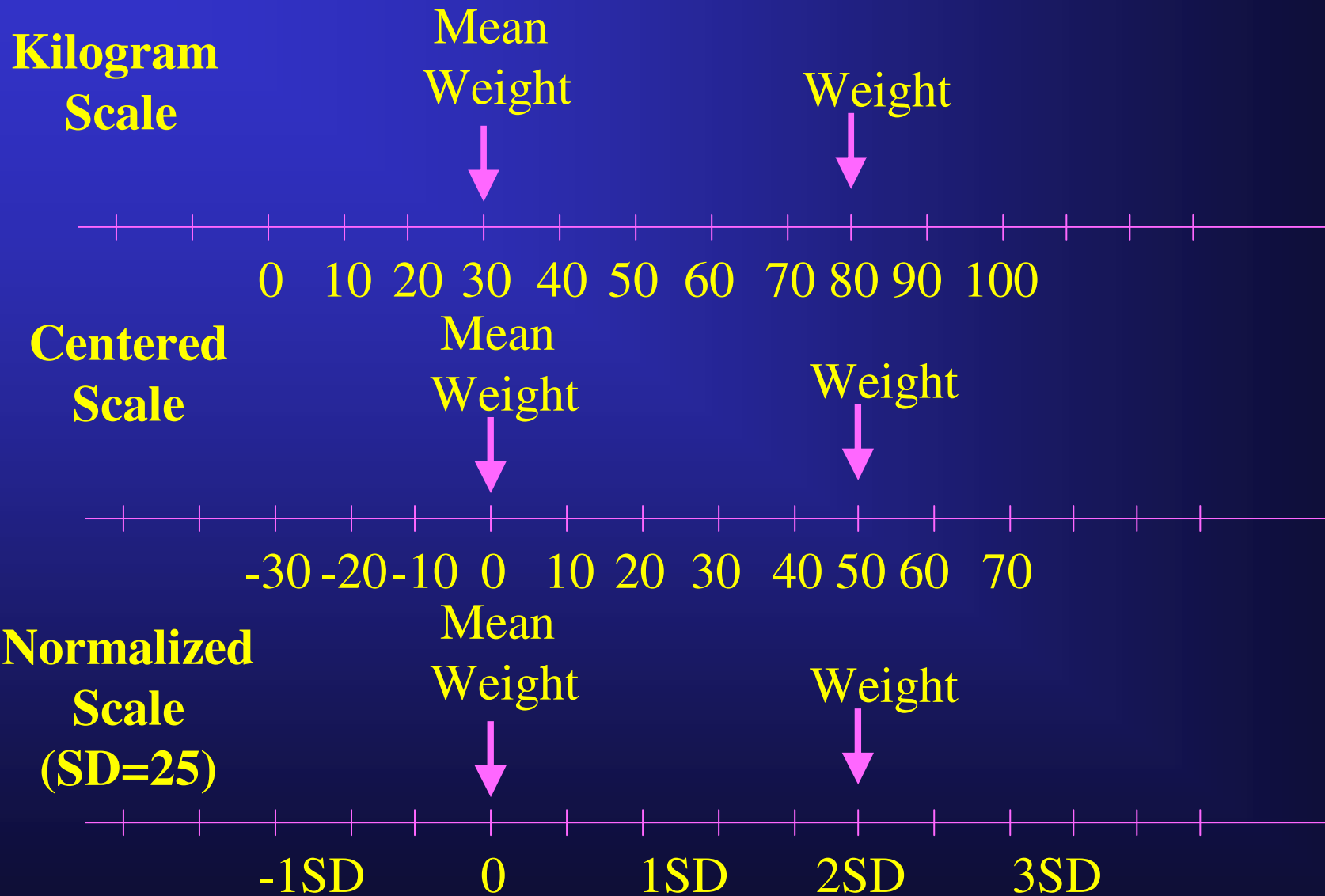
- **The marginal significance of the variable**

# Standardization

- How do we deal with scale dependency?
  - We “normalize”
  - How to “normalize”?
  - The standard normalization.
    - $z = (x - \text{mean}(x)) / \text{sd}(x)$
    - first we center, then we adjust to distribution spread.
    - What is mean? What is sd?
- Divide by  $SD(x)$  and  $SD(y)$ :

$$SD(x) = \sqrt{\frac{\sum_{i=1}^N (x - \bar{x})^2}{N - 1}}$$

# Normalization – A child sample



# Issues in using regression

## Multicollinearity

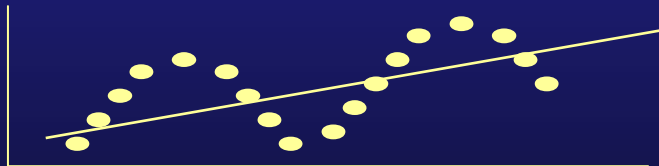
A common phenomenon of strong interrelations among the independent variables. This does not violate any assumptions in the model but it makes the estimates of the regression coefficients unreliable (read an illustration in [1]). Collinearity means that the predictors are correlated (in the same order of magnitude as the correlation with the dependent variable) and this may cause uncertainty in the estimation of the regression coefficients.

Detection: Test the correlations between the variables, large standard error in the coefficients and common sense about possible relations.

Cure: Reduce variables, Factor analysis, stepwise regression (careful!!!), getting more data (why?), identify possible problems before collecting data.

## Autocorrelation

Autocorrelation occurs when the errors are correlated in a serial manner. This is typical when a time series data is obtained but a cycling phenomenon is ignored.



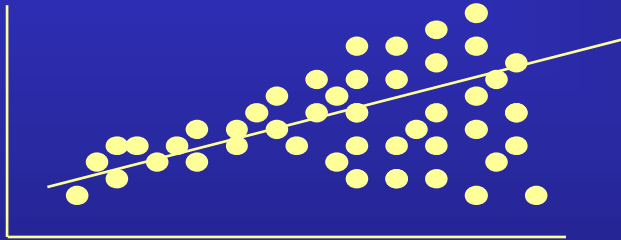
Detection: By plotting the data or some statistical tests (e.g. Durbin-Watson).

Cure: Add variable to remove the cycle or non linearity in the data, use dummy variables.

# Issues in using regression

## Heteroscedasticity

When the error is related to the size of an independent variable - the larger the value of  $x$  the larger is the error.



Use normalized measures.

## Outliers:

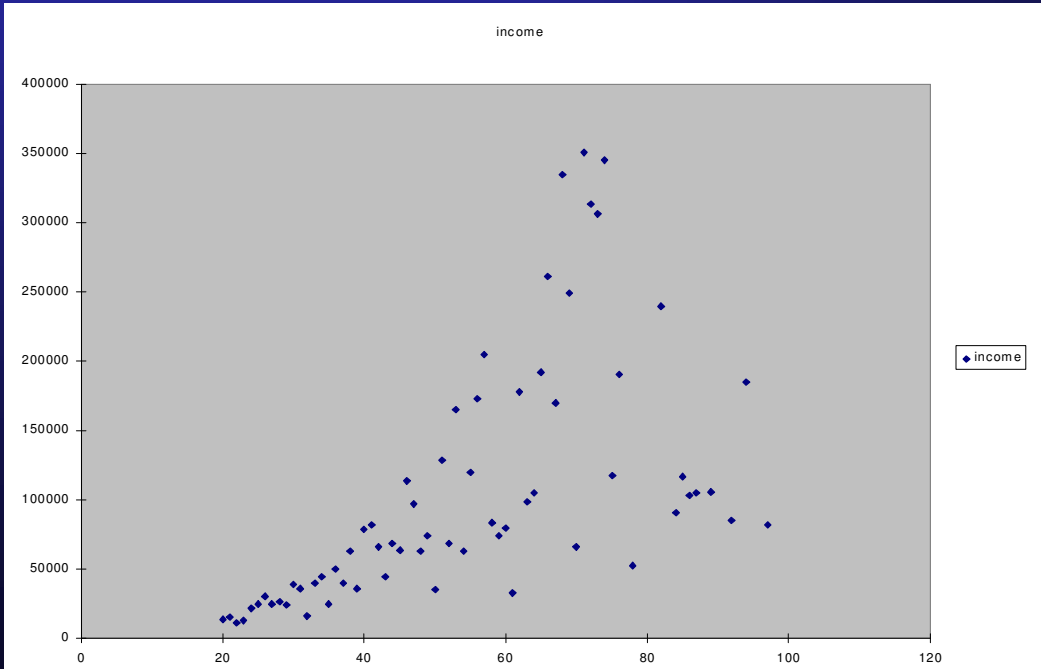
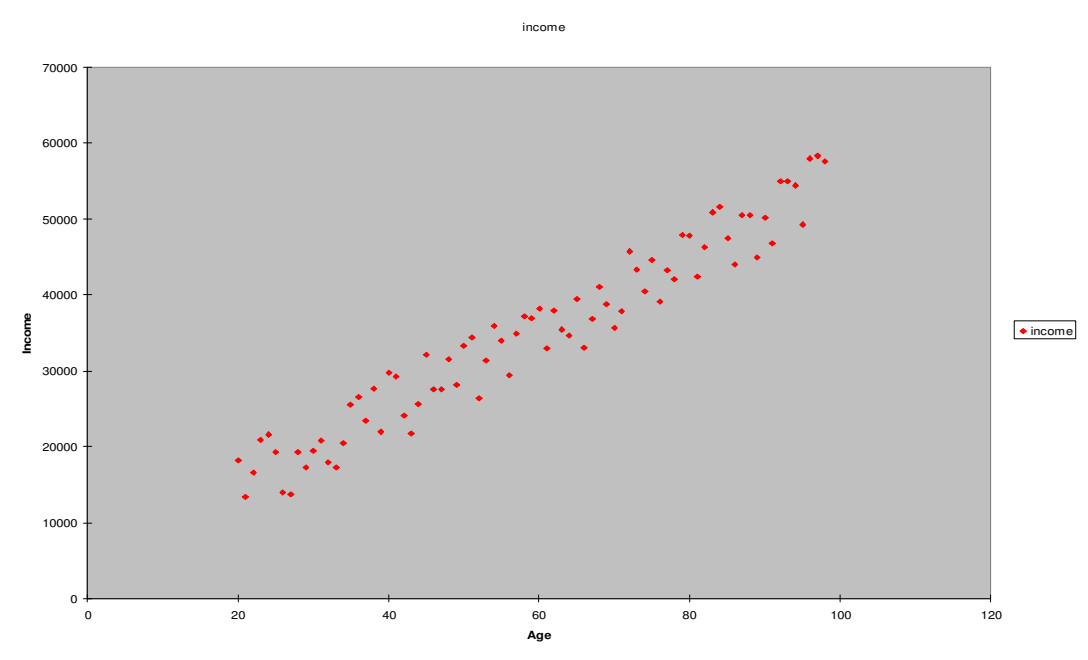
Commonsense is welcomed as long as it is sensible - you have to find a logical explanation before you remove or manipulate an outlier. Do not forget that you may find a new predictor when you analyze an outlier.

## Omitted variables

## Relevant Range

# Homoskedasticity

- The variance of the error in the linear model is constant
- Variance of the error constant for values of  $X$ 's.
- Advantage is easy to check: visually or with statistical tests
- two graphs follow: one with uniform scatter and the other with non-uniform scatter



# Consequences of heteroskedasticity

- Inefficiency:
  - least squares doesn't give minimum SE's.
  - reason is OLS gives equal weight to all obs rather than less weight to high variance obs
  - potential solution is weighted least squares
- Biased standard errors:
  - SE's of betas will be wrongly estimated
  - leads to bias in testing and confidence intervals

# Testing for heteroskedasticity

- Many potential tests
  1. Graphical
  2. Statistical tests. Basic idea involves regressing squared  $e$ 's on  $x$ 's or  $\hat{y}$ .
    1. Breusch-Pagan Test
    2. White Test
- Solution: weighted least squares OR live with less efficient model

# Dummy Variables

- Why are dummy variables so important?
  - They allow us to control for and analyze effect of dichotomous variables like gender and any other yes/no variable
  - They allow us to control for and analyze effect of unordered categorical variable such as race, religion, nationality.
  - They allow to control and analyze effect of ordered categorical variables such as social class, education level (primary, secondary, post-secondary), and any categorical variable.
  - They provide a flexible method to analyze non-linearities of continuous variables.

# Dummy Variables

- How do dummy variables work?
  - Suppose  $Y$  is income,  $X_1$  is education and we run regression of  $Y = b_0 + b_1X_1 + e$ .
  - Suppose we are concerned that men and women have different relationships between  $X_1$  and  $Y$ . We could run different regressions on men and women but this won't tell us whether the differences are statistically significant. What we want is to incorporate a variable for gender into the regression:

# Dummy Variables

–  $Y = b_0 + b_1X_1 + b_2*male + e$

where male=0 if female and male=1 if male.

This means that the results will be:

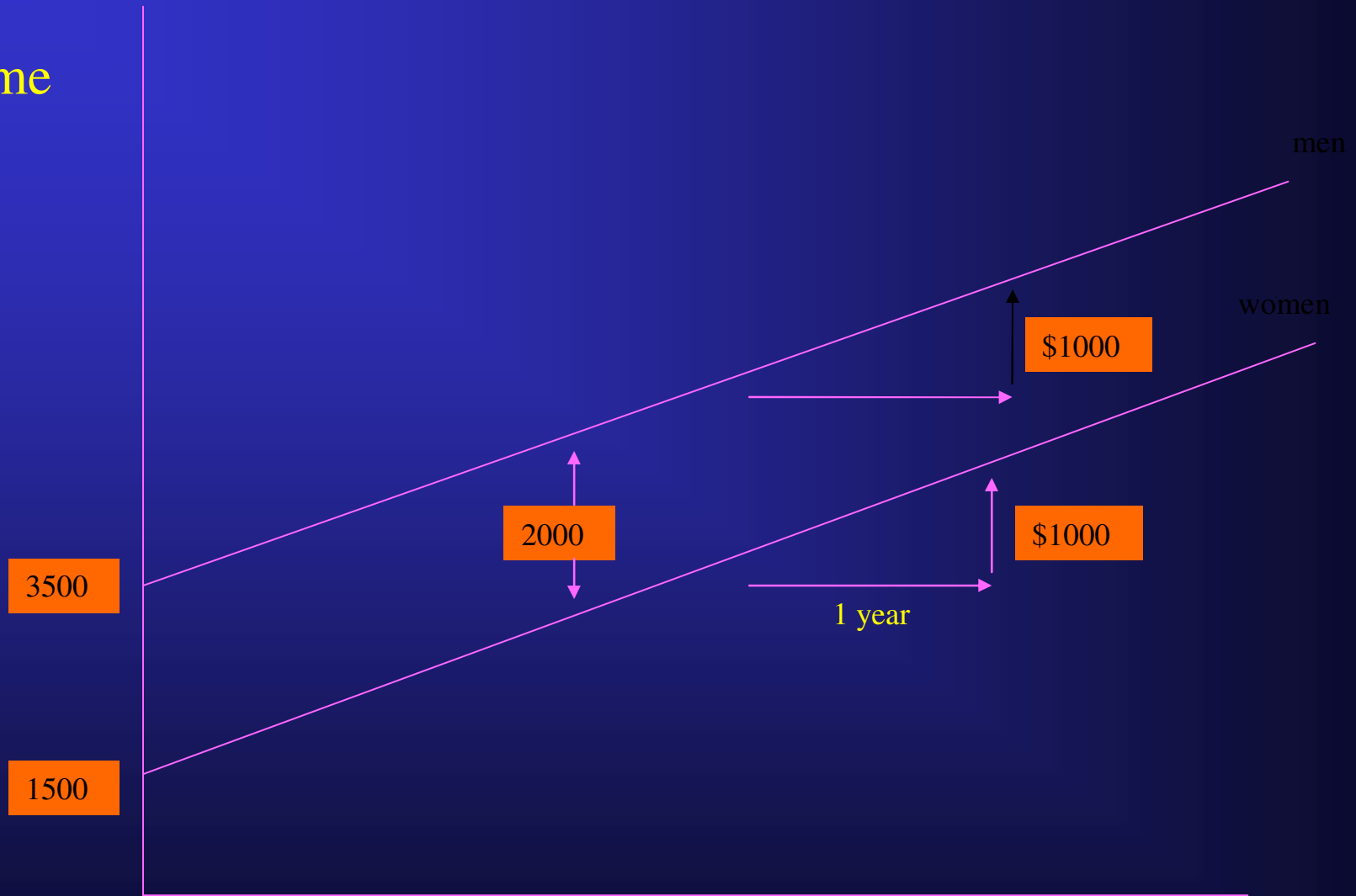
For women:  $Y = b_0 + b_1*educ$

For men:  $Y = b_0 + b_1*educ + b_2*male$

Suppose  $b_0=1500$ ,  $b_1=1000$ , and  $b_2=2000$ .

# Graphing the role of dummy variables

Income



Education

# Dummy Variables

$$Y = b_0 + b_1 * \text{primary} + b_2 * \text{mid} + b_3 * \text{sec} + e$$

$$\text{height}(\text{none}) = b_0$$

$$\text{height}(\text{prim}) = b_0 + b_1$$

$$\text{height}(\text{mid}) = b_0 + b_2$$

$$\text{height}(\text{sec}) = b_0 + b_3$$

- Why don't we add dummy for none?
  - We called one category the omitted/reference.
  - Because if none is also a variable, then summing the four variables will lead to ONE which is impossible. Thus one category must be dropped!

# Graphing the role of dummy variables

Income



Experience

## The use of Categorical Variables (Dummy Variables)

A lot of data in marketing is categorical in its nature (sex, occupation, city), such nominal scales has to be incorporated. One way is to run a separated regression for each categorical value (e.g. each city). The other approach is using dummy variables. The assumption is that the other variables' effect is the same in each value of the categorical variable while the constant changes.

### Illustrative Example

Consider the data on quarterly fuel oil shipments to UK in 1964-1966 (taken from [1]).

<u>Quarter</u>	<u>Year</u>	<u>Sales</u>
1	1964	210
2	“	120
3	“	140
4	“	260
1	1965	220
2	“	125
3	“	145
4	“	270
1	1966	215
2	“	128
3	“	149
4	“	275

Q: Describe the seasonal effect and its influence on a regression model

# A regression model for fuel oil shipments

A regression equation with dummy variables (value equals 0 or 1):

$$\text{shipments} = B_0 + B_1(\text{time}) + B_2(\text{winter}) + B_3(\text{spring}) + B_4(\text{summer})$$

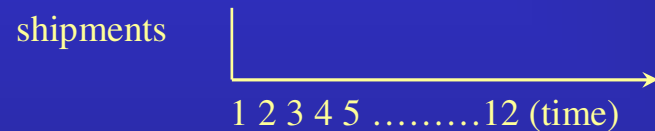
The Fall is not included because if all the independent dummy variables will be included they will be perfectly multicollinear. This redundancy will lead to “explosion” of the calculations. In general, if a categorical variable has c categories a c-1 dummy variables are needed. Here is the seasonal dummy variables:

## Dummy Variables

<u>Shipments</u>	<u>Time</u>	<u>Winter</u>	<u>Spring</u>	<u>Summer</u>	<u>Fall</u>
210	1	1	0	0	0
120	2	0	1	0	0
140	3	0	0	1	0
260	4	0	0	0	1
220	5	1	0	0	0
.....					
274	12	0	0	0	1

## Questions

- Plot The shipments Vs. time without using dummy variables



- Compute the coefficients in the dummy variables regression equation.
- Plot again a graph of shipments for each season
- What is the time effect? What is the seasonal effect?
- What is your estimation for the shipments in 1967? What are the limitations of your prediction?
- Repeat the process of regression with dummy variables only this time use the year as a predictor (time = 1,2,3 instead of 1,2,3...12). How does this change your results? What is your conclusion?

# More issues to consider

- Causality Vs. correlation
- Simultaneous causality
- Regression and ANOVA
- Interactions
- Non linearity

# Non-linear relationship

- $Y = a + b * X$
- Have we specified the form of  $X$ ?
- Suppose we replace  $X$  with  $X^2$ 
  - Then, we can rewrite:  $Z=X^2$  and  
 $Y = a + b * Z$
- Suppose we replace  $X$  with  $\log(X)$ 
  - Then, we can rewrite:  $Z=\log(X)$  and  
 $Y = a + b * Z$
- Relationship still LINEAR

# Multiplicative model and comments

Life are not linear, yet in most of the marketing cases linearity assumption is a good start. Other functions and filters can be used to add accuracy: Polynomial smoothings, spline, etc. In some of the cases analytical skill can help. For example the following multiplicative model can be transformed to a linear equation and a regression can be applied:

$$y = B_0 + x_1^{B_1} x_2^{B_2} \Rightarrow \log y = \log B_0 + B_1 \log x_1 + B_2 \log x_2$$

By setting:

$$\begin{aligned} y' &= \log y \\ x_1' &= \log x_1 \\ x_2' &= \log x_2 \end{aligned}$$

$$y' = A_0 + A_1 x_1' + A_2 x_2'$$

where

$$B_0 = \log^{-1} A_0$$

$$B_1 = A_1$$

$$B_2 = A_2$$

a regression equation can be obtained and the parameters can be computed:

# Using the linear model to examine non-linear relationships

- Recall, the linear regression model does NOT require the X's to be measured linearly
- We frequently examine non-linear relationships. Why?
  - There may be a theoretical reason why we expect a non-linear relationship between the Y and X variable
  - Examination of the relationship may show that a linear approximation is poor.
- Always begin by LOOKING at the data

# Using the linear model to examine non-linear relationships

- Suppose we want to examine effect of age on income
- View plot.
- Income increases with age but also appears to decline at higher ages. The plot also ignores the effect of schooling.
- A better plot to tell whether there is a non-linear relationship involves controlling for the schooling effect: we call this a partial residual plot

# Using the linear model to examine non-linear relationships

- Perhaps most common transformation is the natural logarithm (ln)

$$\ln(y) = b_0 + b_1x_1 + b_2x_2 + e \quad \text{or}$$

$$y = \exp(b_0 + b_1x_1 + b_2x_2 + e)$$

- Useful if all the y values are non-negative and their predictions should be non-negative
- Particularly useful when variable has a long positive tail
- Variables like income, assets, and wages are particularly prone to be positively skewed. These are often helped by taking log or semi-log:

$$Y = b_0 + b_1 \cdot \log(x_1) + e$$

# Using the linear model to examine non-linear relationships

- If plots insufficient, we fit non-linear terms.
- Easiest is to add quadratic (squared) term:

$$y = b_0 + b_1x_1 + b_2x_2 + e$$

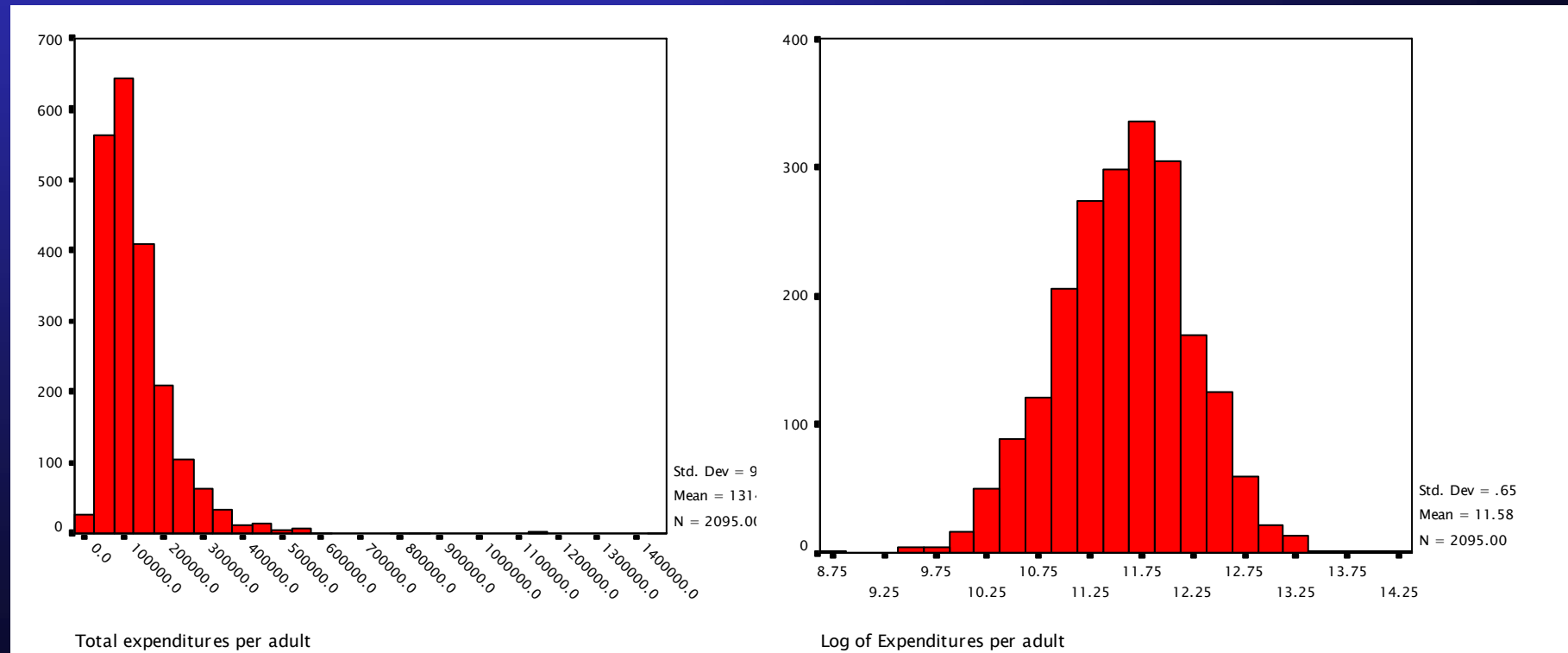
$$y = b_0 + b_1x_1 + b_2x_2 + b_3(x_2^2) + e$$

$$\text{income} = b_0 + b_1*\text{educ} + b_2*\text{age} + b_3*\text{age}^2 + e$$

- Is squared term significant? Implications?
- Plot the relationship at the mean education level.
- Useful to do JOINT F-test on  $b_2$   $b_3$

# Using the linear model to examine non-linear relationships

- Regression will try and explain the dominant feature of the distribution in our data. If distribution is highly skewed, we may not find the coefficients very helpful.



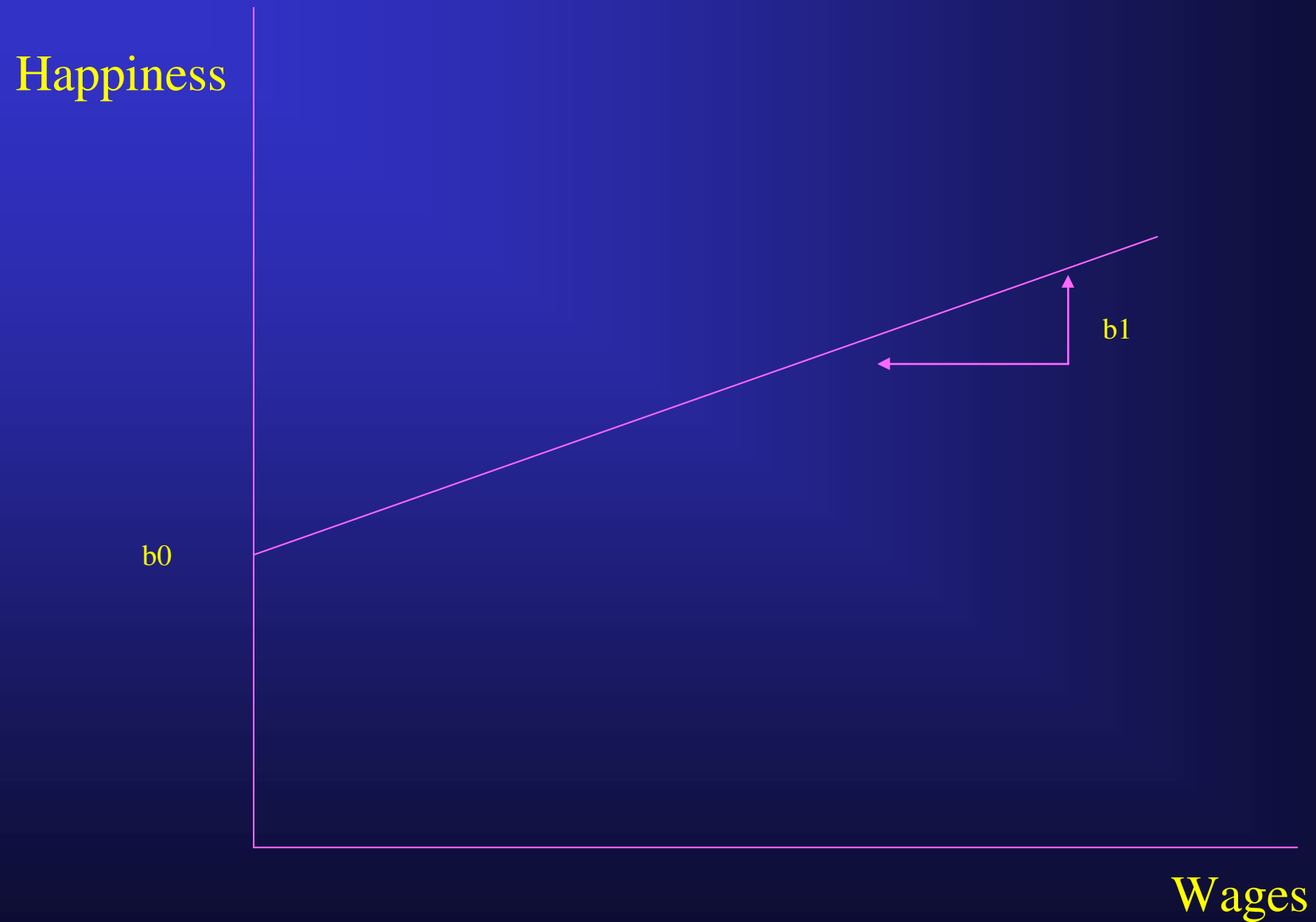
# Interaction Effects

- The basic  $Y = a + b_1X_1 + b_2X_2 + e$
- Interaction is another form of non-linearity because the effect of  $x_1$  is constant but varies for different levels of  $x_2$
- Examples:
  - effect of education on wages may differ by IQ, gender, race,
- New Model:
- $Y = a + b_1X_1 + b_2X_2 + b_3(X_1 * X_2) + e$

## Dummy variables

- Let's use an example of how wages affect people's happiness.
  - Suppose happiness,  $H$ , only depends on wages,  $W$ :  
$$H = b_0 + b_1W + e$$
  
t-test on  $b_1$  is whether  $H$  varies with  $W$

# Graphing the role of dummy variables



# Dummy variables

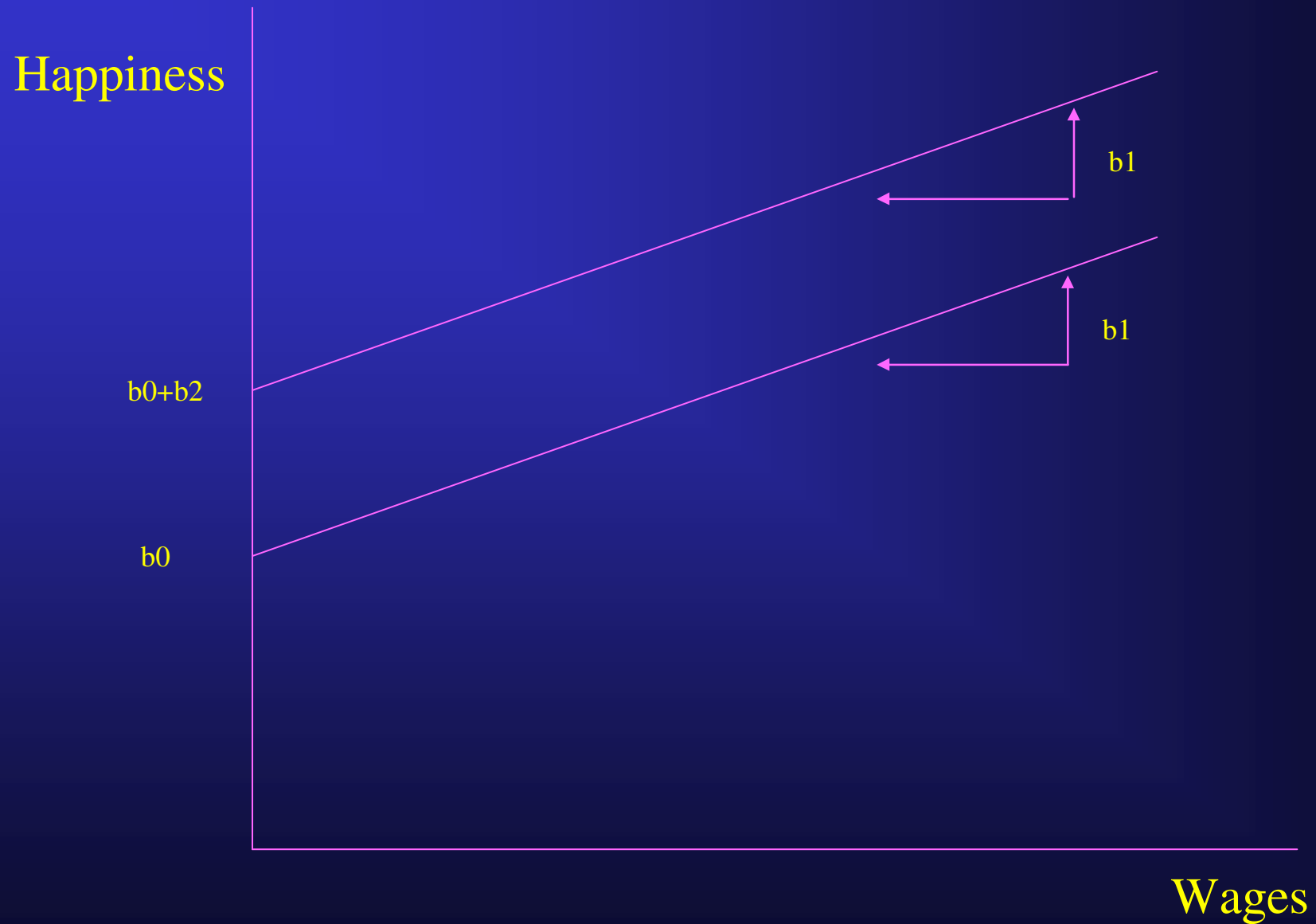
- What if we believe wages may affect happiness differently for men and women? Then we may introduce dummy for male, M:

$$H = b_0 + b_1W + b_2M + e$$

The t-test of  $b_1$  is still whether H varies with W controlling for M,

The t-test on  $b_2$  is whether H varies between women and men: constant gender effect

# Graphing the role of dummy variables



# Dummy variable interactions

- Rather than assuming different intercepts and similar slopes for men and women, we might assume that their slopes are different and similar constants!

This means incorporating an **interaction** between gender,  $M$ , and wages,  $W$  on  $H$ .

$$H = b_0 + b_1W + b_2(W*M) + e$$

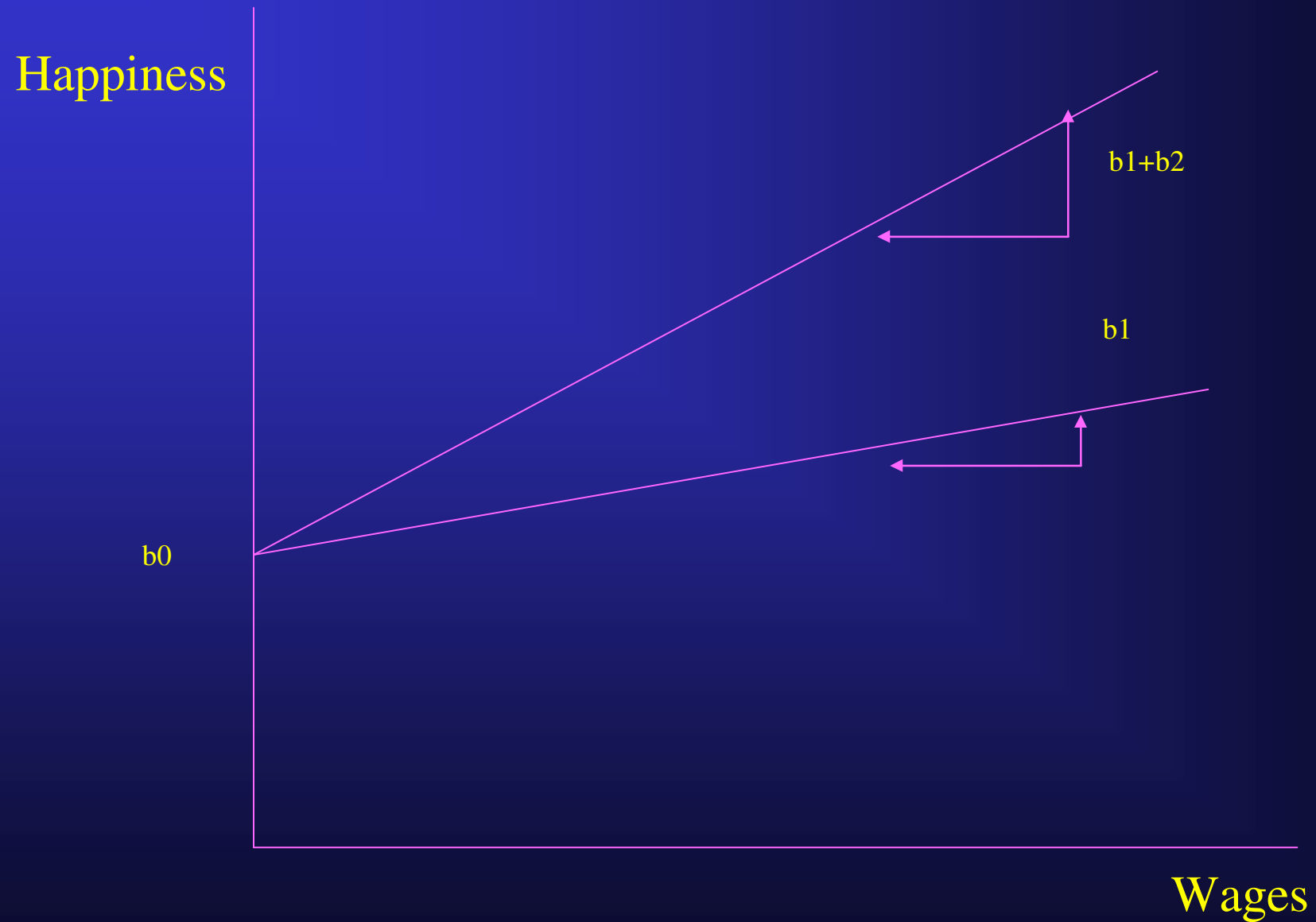
The t-test on  $b_2$  tells us whether there is a change in the relationship between  $W$  and  $H$  for men and women. For men ( $M=1$ ),

$$H = b_0 + b_1W + b_2(W*1) = b_0 + (b_1+b_2)*W$$

for women ( $M=0$ )

$$H = b_0 + b_1W + b_2(W*0) = b_0 + b_1*W$$

# Graphing the role of dummy variables



## Dummy variables continued

- We may want to test whether both slope and intercept vary between men and women:

$$H = b_0 + b_1W + b_2M + b_3(W*M) + e$$

$b_1$  gives pure wage effect,  $b_2$  gives gender effect, and then  $b_3$  tells us whether the wage effect varies between gender. Two slopes and two intercepts means we are drawing two lines

$$\text{Men: } H = (b_0+b_2) + (b_1+b_3)*W$$

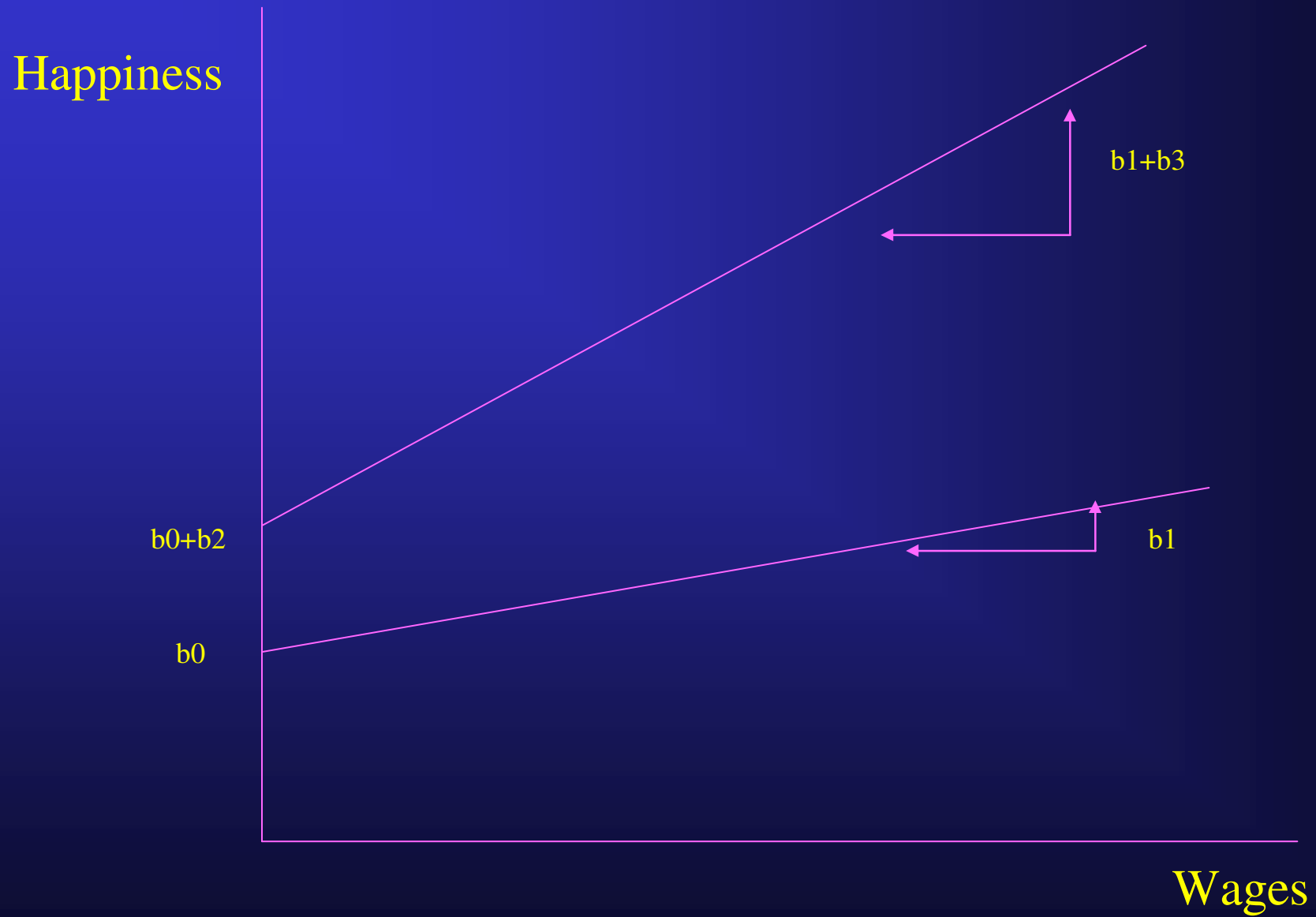
$$\text{Women: } H = b_0 + b_1*W$$

$b_2$  t-test tells us overall gender effect

$b_1$  t-test tells us overall wage effect

$b_3$  t-test tells us whether wage effect varies by gender

# Graphing the role of dummy variables



# Dummy variable interactions

- This last option is most general unless we want to run two separate regressions.
- Final option is to run separate regressions on men and women. This also means two separate lines and the coefficients will be similar.
- Preferable if possible to run them together b/c then you can test for differences rather than assume they exist. If strong theoretical reasons to separate, this also important
- Difference is that variance to change for men and women separately, thus one regression is more efficient

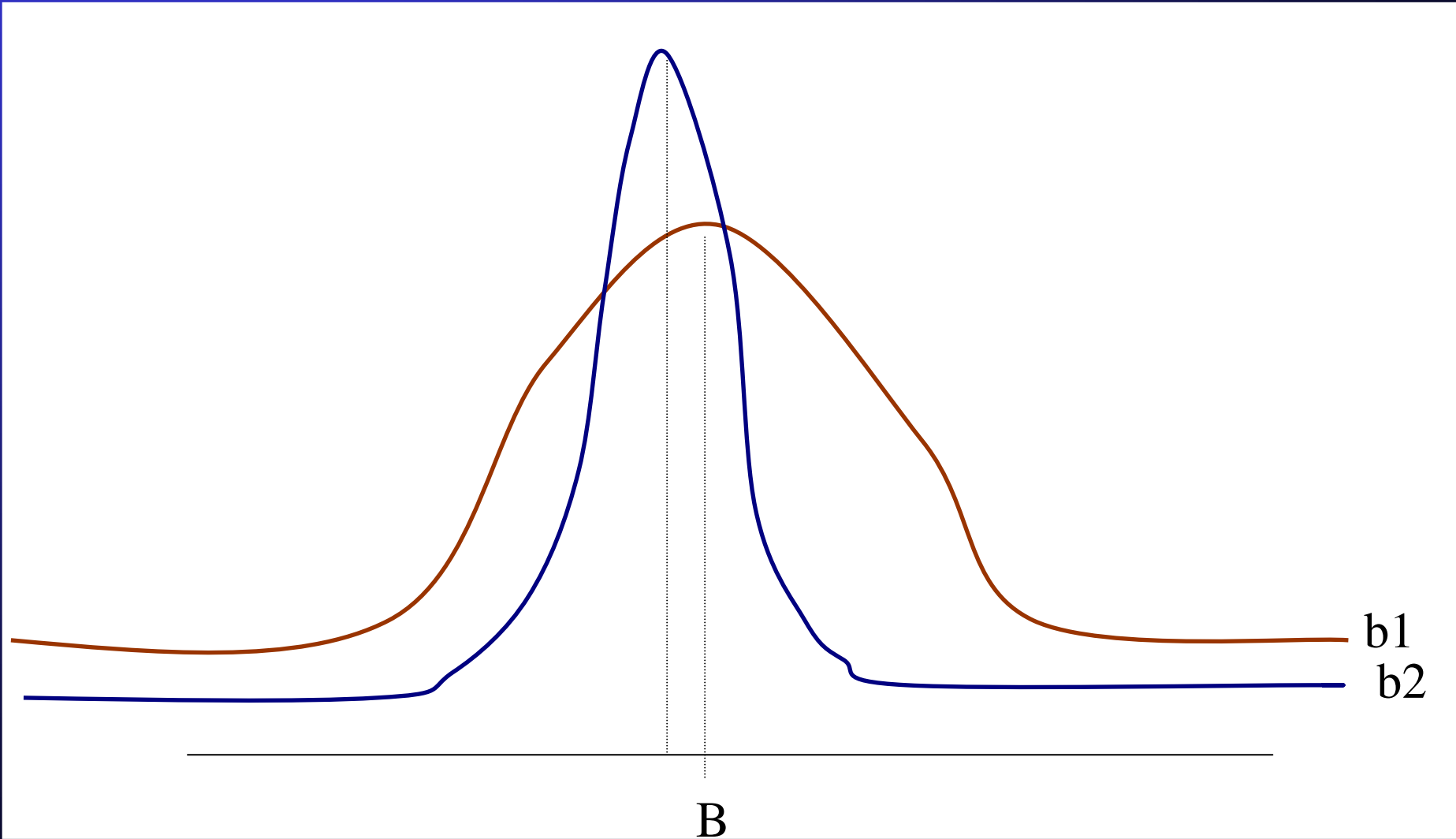
# Assessing performance of the method

- Bias
  - This means there is a TRUE value which we are trying to estimate.
  - An estimator is unbiased if there is no systematic tendency for it to produce estimates that are too high or too low.
  - Thus we hope to have a method which **ON AVERAGE** produces estimates of the TRUE value which are not bigger or smaller.

# Assessing performance of the method

- Efficiency
  - This is a measure of how much variation in the estimate there is around the TRUE value.
  - Thus, prefer a method which usually produces estimates that are close (+ or -) rather than far from the TRUE value.

# How to assess the performance of the method: bias and efficiency



## The classical assumptions that underlie the use of OLS regression

1. linear functional form
2. Mean independence term
3. homoskedasticity
4. non-autocorrelation
5. normality of the error term

# Linearity of functional form

- The basic model:

$$Y = a + bX + e$$

- Why is it called a linear model?
- could be the equation for a line.
- more importantly, manner in which coefficients (b) and (a) and error enter the equation.
  - thus, Y and X can be transformed in any way.

$$\log(Y) = a + b (\log X) + e$$

- Non-linear:  $y = (1/(a+bx)) + e$

# Mean independence

- Unbiasedness

$$E(e_i)=0$$

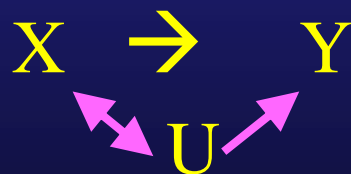
- $e_i$  is the sum of many different unobserved influences - we assume they sum to 0
- concern that  $e_i$  not centered around zero would allow us to simply add the difference to the constant term and recenter  $e_i$  around 0.
- More generally, mean independence necessary for ceteris paribus:

# No association between e and X

- How do we know we are learning about effect of x on y holding other factors constant?
- $wage = b_0 + b_1 * educ + e$
- Where are all other factors?
- Reliable estimates of  $b_0$  and  $b_1$  from a random sample when we introduce assumptions on association between e and x
- If e and x are associated, can't say that we have found effect of x holding other factors constant
- Correlation is one possibility but ...
- Better to assume that  $E(e|x)=0$

## No association between $e$ and $X$

- Consider what this means in wage example
- What might be included in  $e$ ?
- Implication is that average  $e$  constant regardless of  $x$ :  
 $E(e|x)=E(e)=0$
- State an example of when this might be wrong!
- When not true, may be evidence of a spurious association between  $x$  and  $y$ .



# No association between e and X

- Serious concerns:
  - severe bias in b's
  - frequent problem: omitted X's, reverse causation, measurement error
  - and difficult to test
- Unlikely problem with experimental data
- Very difficult to overcome with survey data

# Real and predicted values

- Remember:

$$Y = a + bX + e$$

$$\hat{Y} = a + bX$$

$$Y - \hat{Y} = e$$

- Exercise: How do fertility rates depend on infant mortality rates? Use world95.dta
- Stata: Regress Y X
- Stata: Use coefficients and “generate” function to calculate the yhat for each case
- Stata: Use predict command to predict yhat directly after regress command
- Compare 2 types of predictions

# Three ways to predict in Stata

1. Find the coefficient after running a regression and then use the “generate” command:

```
generate yhat = 0.7 + 0.4*X
```

2. Use the generate command directly with the coefficients after a regression:

```
gen yhat = 0.7 + _b[educ]*X
```

3. Use Stata’s direct prediction capability:

```
predict yhat
```

The error equals  $y - \text{yhat}$

# Non-autocorrelation

- consider the error term:
  - a different value for each individual
  - sum of unobserved effects
- why should errors for any two people be correlated?
  - sampled from same village
  - live in the same household
- Consequence is also inefficiency and SE's which tend to be biased down.
- fear of mistaken significance unless corrected

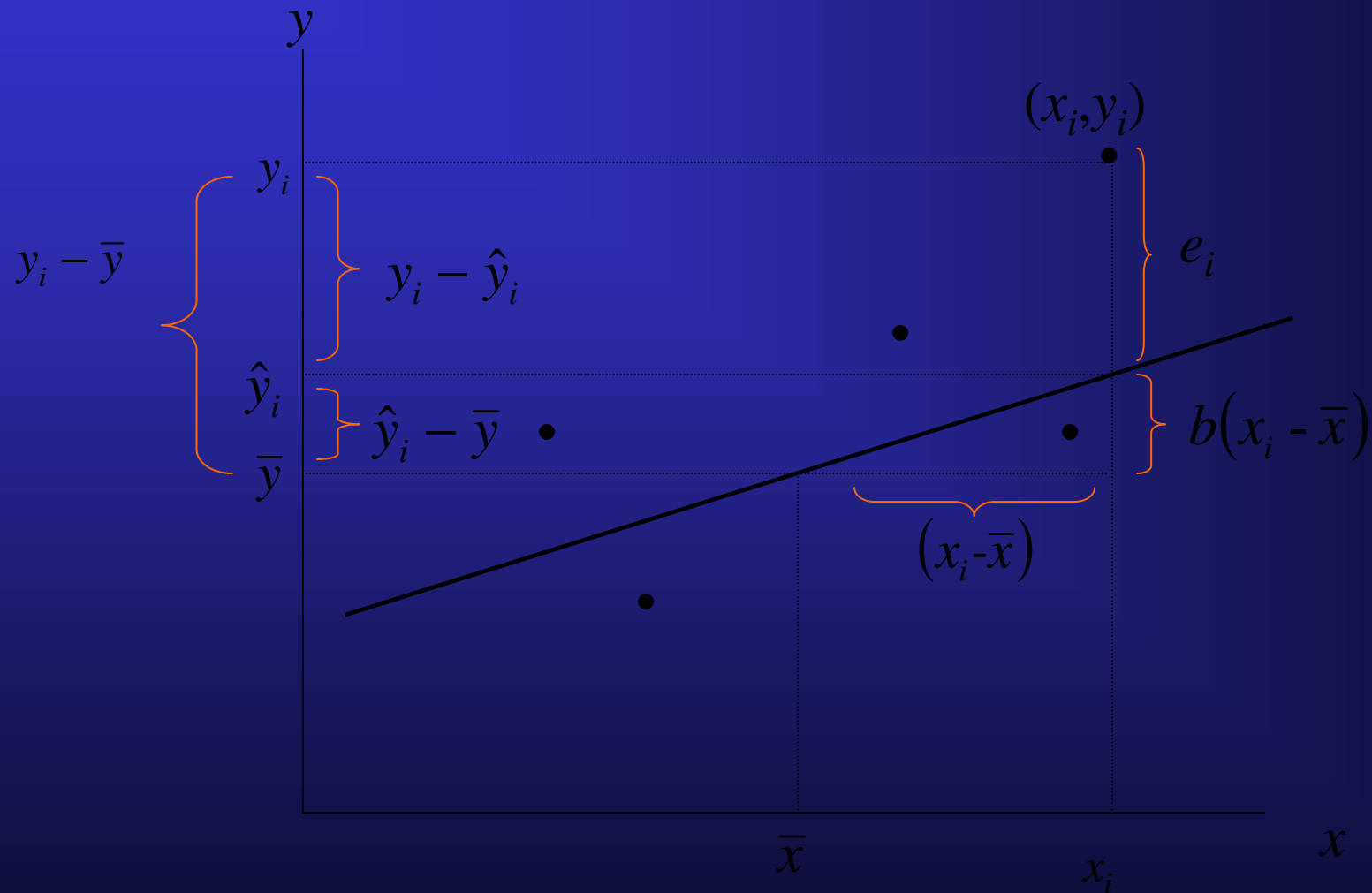
# Normality Assumption

- most confusing assumptions
- least important of major assumptions
- only necessary for statistical testing
- refers only to distribution of error term which is unobserved and NOT to the individual variables
- We will see that the distributions of  $X$ 's and  $Y$  may affect results but don't need to be normal

# Normality Assumption

- if not normal it will be normal in large samples because of central limit theorem
- if small samples it will matter for testing
- what is small?
- Smaller than 100-200 usually.
  - Then we should examine residuals.
  - Statistical tests usually require large samples which is when we don't need them!
  - Option: tighten critical value! For example: 1%

# What quantity do we minimize?



# An important relationship

- Define 3 measures:

- **TSS** is total sum of square errors and equals difference between  $\bar{y}$  and  $y_i$ 's

$$\text{TSS} = \sum (y_i - \bar{y})^2$$

- **RSS** is regression sum of squares and equals difference between  $\hat{y}$  and  $\bar{y}$ .

$$\text{RSS} = \sum (\hat{y} - \bar{y})^2$$

- **ESS** is error sum of squares and equals difference between  $y_i$  and  $\hat{y}$

$$\text{ESS} = \sum (y_i - \hat{y})^2$$

# What the line explains?

Percentage of total variation in  $\bar{y}$  which can be attributed to linear relation with X

$$TSS = RSS + ESS$$

so (dividing by TSS)

$$RSS/TSS = 1 - ESS/TSS$$

or

$$R^2 = RSS/TSS$$

Calculate  $R^2$  by finding  $\hat{y}$  for regression of TFR on IMR. First calculate regression equation in SPSS. Use Excel for all other calculations!

# The F-test

- F-test is powerful because of flexibility
- Does *not* rely standard errors like t-test
- F-test uses residual sum of squares of the entire regression
- Basic idea is that F-test tells us the effect of imposing *restrictions* on the model
- It provides us with a statistical test of the significance of the change in the residual sum of squares

# The F-test and restrictions

- We often talk about *restrictions* or a *restricted* model. The *restricted* model is always relative to the *unrestricted* model.
- For example: Suppose we have the following model:  $Y = b_0 + b_1X_1 + b_2X_2$ . We could consider this an unrestricted model. And we could then also consider a variation where we assume  $b_2=0$  and  $b_1=0$ . This is the *restricted* model.
- In fact, this is the test of whether the regression model explains anything or not.

- Alternatively, the F-test compares *unrestricted* model to a different *restricted* model:

$$Y = b_0 + b_1X_1$$

Here we only assume  $b_2=0$ .

This is similar to a t-test since it tells us how much the residuals of the model change when we drop only one variable.

- Remember *restrictions* can only make residuals of model bigger - never smaller!
- Since adding variables always makes residuals smaller, we need to also account for change in degrees of freedom -- like with adjusted  $R^2$

# The general formula for the F-test

$$F_{m, n-k} = \frac{ESS_R - ESS_U}{ESS_U} \cdot \frac{n - k_u}{m}$$

- $ESS_{U,R}$  = residual sum squares of unrestricted and restricted models
- $n$  = sample size
- $k$  = number of estimated coefficients in unrestricted model
- $m$  = number of restrictions

## Intuition behind the F-test formula

- $RSS_U < RSS_R$  because a model with less variables can never explain MORE
- Suppose dropping the variable has a SMALL and then LARGE effect. What is the effect?
- Thus, first term tells us proportional change in RSS due to dropping variables!
- The second term allows us to translate this into a test statistic based on the F-distribution
- The F-statistic is distributed with  $m, n-k$  degrees of freedom.  $M$  for numerator (columns in F-table) and  $n-k$  for denominator (rows in table)

# The F-test and the F-distribution

- Thus, the F-test gives us a number and we compare it to the number for an F-distribution to see whether it is  $>$  or  $<$  than the critical value where  $m$  is col df and  $n-k$  is row df.

<http://www.statsoft.com/textbook/stathome.html>

- Stata always gives the F value for a regression for the test where ALL coefficients are restricted to be equal to 0 against the unrestricted model. This test is almost always accepted.
- More interesting to run test using:  
  . Test x1 x2 x3 ... xN

# Examples with the F-test

- Test effect of dropping all variables using F-test
- Test effect of dropping two education categories
  - What are the values in the equation?
  - $M=2$
  - $n=35$
  - $k=4$
  - $F=?$

# What the line doesn't explain

- Useful because of ease of interpretation
- May be low value even if  $X$  is good explanatory variable if lots of variation in  $Y$ 
  - tends to be low in cross-section data, higher with time series
- May be inappropriate if non-linear relationship

## מטלת פרויקט (חובה)

בנה מודל של רגרסיה ליניארית הכולל 3 משתנים בלתי תלויים (מנבאים) לפחות ומשתנה תלוי אחד. השתמש בנתונים קיימים, במידה והנך נדרש לנתונים שאינם קיימים לאחד המשתנים ניתן לבנות שאלון ולבצע הערכה לנתונים אילו (בתאום עם המתרגל). המשתנה התלוי יהיה מדד כלשהו להתנהגות שוק רצויה בהקשר של הפרויקט (הכנסות, מכירות, פניות טלפון וכדומה).

הערך את מידת הביטחון במודל

ערוך ניתוח רגישות ובנה המלצה לקומבינציה חדשה של משתנים מנבאים שיביאו לערך טוב יותר של המשתנה התלוי.

## שיטות תצפית

לא מוסווה	מוסווה	
מחקרי Rating (הצבת מכשיר בבית הלקוח). פאנל קבע בסופר.	שיטות "פסיכו- פיזיקליות".	מובנה (יודעים מה מחפשים)
מחקרים "אנתרופולוגים" על צרכנים	שיטת Fisher Price בגני הילדים, "גרבולוגיה", לקוח מדומה, וודיאוכרט בסופר, מראות חד כיווניות	לא מובנה

## שיטות תישאול (תקשורת)

לא מוסווה	מוסווה	
שאלון, בדיקות מתוחכמות יותר (למשל שימוש בביפרים למחקר על ניצול זמן).	נדיר יחסית	מובנה
מחקרים איכותניים: ראיונות עומק, קבוצות מיקוד.	שיטות השלכה, ניתוח תוכן של השלמות משפטים, פירוש קריטורות וסיפורים.	לא מובנה

# The F-test for the regression fit

- based on F-distribution which is distribution for ratio of two chi-square variables which are sums of normal distributed variables
- we use this standard distribution to create a test of the fit of the model
- gives us a statistic to use to test for ratio of explained variance divided by unexplained variance. The F-statistic:

$$F_{1,N-2} = [RSS/1] / [ESS/(N-2)]$$

- subscripts denote degrees of freedom
- minimum and maximum of F?
- Example: Find F in regression of income on age.