Eco 213: Basic Data Analysis and Econometrics
Lecture 2 : Ordinary Least Squares
February 16, 2019

Dr. Garima Malik
Department of Economics
Shiv Nadar University

Outline

- Linear & Non-linear Relationships
- Fitting a line using OLS
- Inference in Regression
- Omitted Variables & R^2
- Types of Regression Analysis
- Properties of OLS Estimates
- Assumptions of OLS

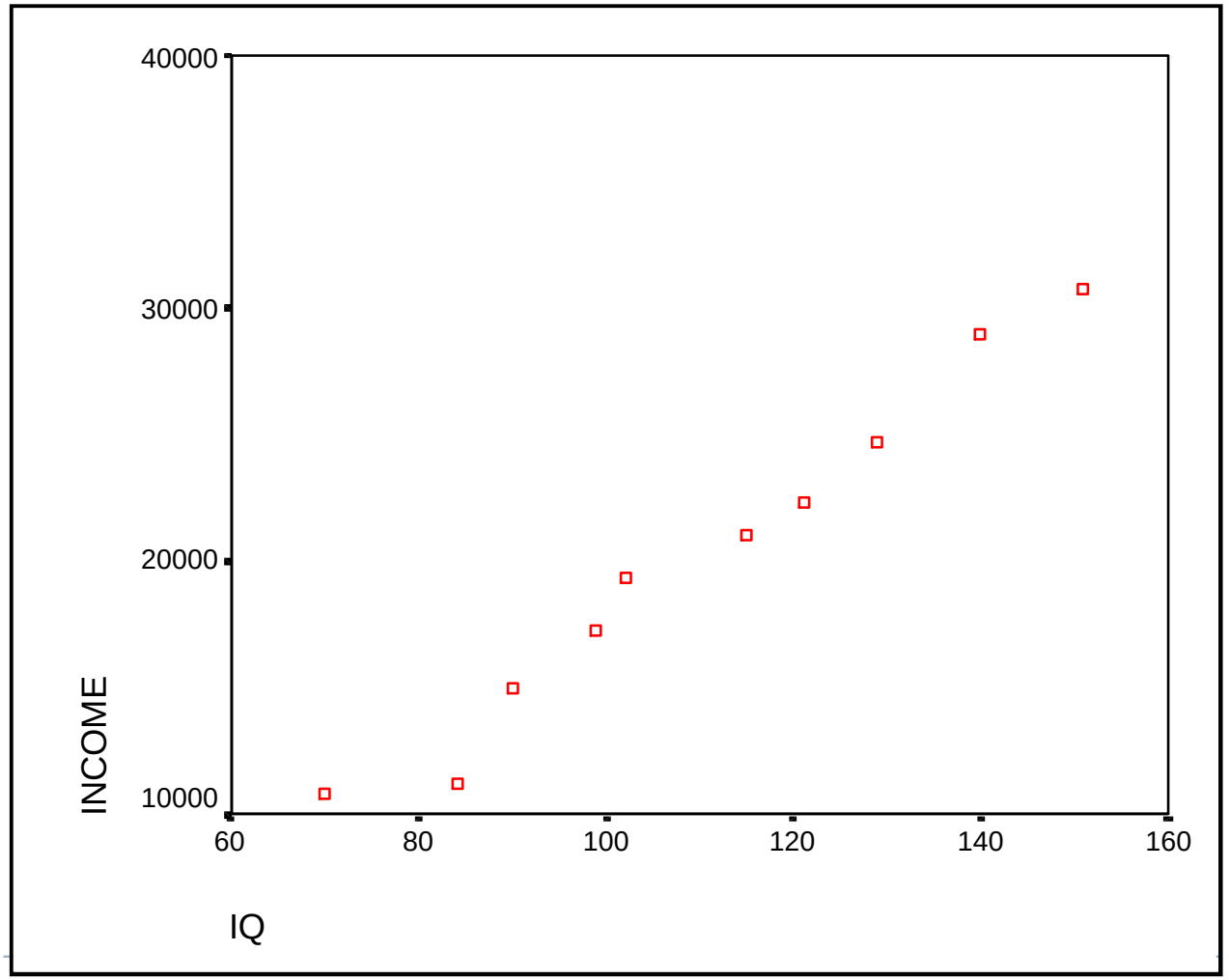
Linear & Non-linear relationships between variables

- Often of greatest interest in social science is investigation into relationships between variables:
  - is social class related to political perspective?
  - is income related to education?
  - is worker alienation related to job monotony?
- We are also interested in the direction of causation, but this is more difficult to prove empirically:
  - our empirical models are usually structured assuming a particular theory of causation

Relationships between scale variables

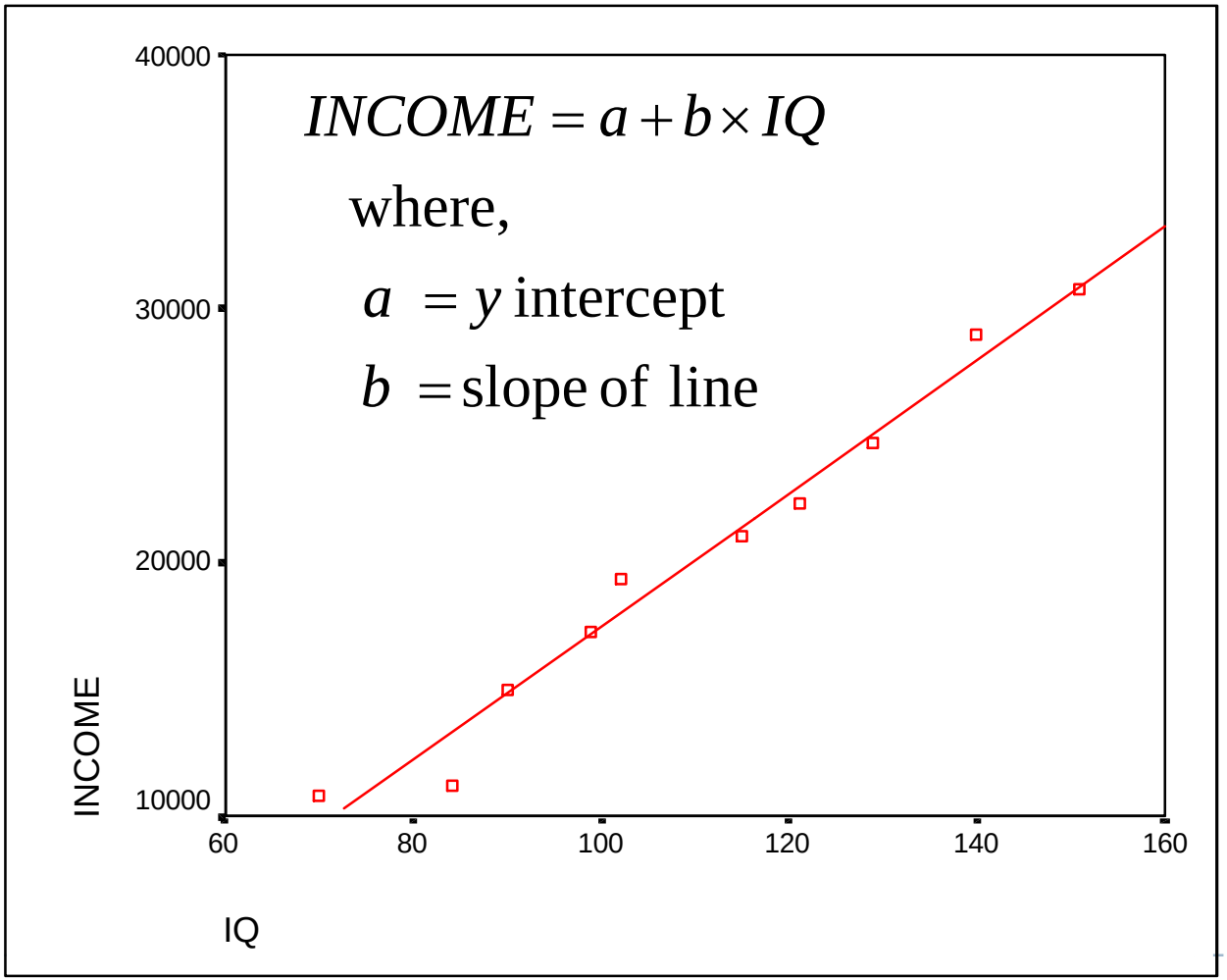
- The most straight forward way to investigate evidence for relationship is to look at scatter plots:
  - traditional to:
    - put the *dependent* variable (I.e. the "effect") on the vertical axis
      - or "*y* axis"
    - put the *explanatory* variable (I.e. the "cause") on the horizontal axis
      - or "*x* axis"

Scatter plot of IQ and Income:



We would like to find the line of best fit:

$$\hat{y} = a + bx$$

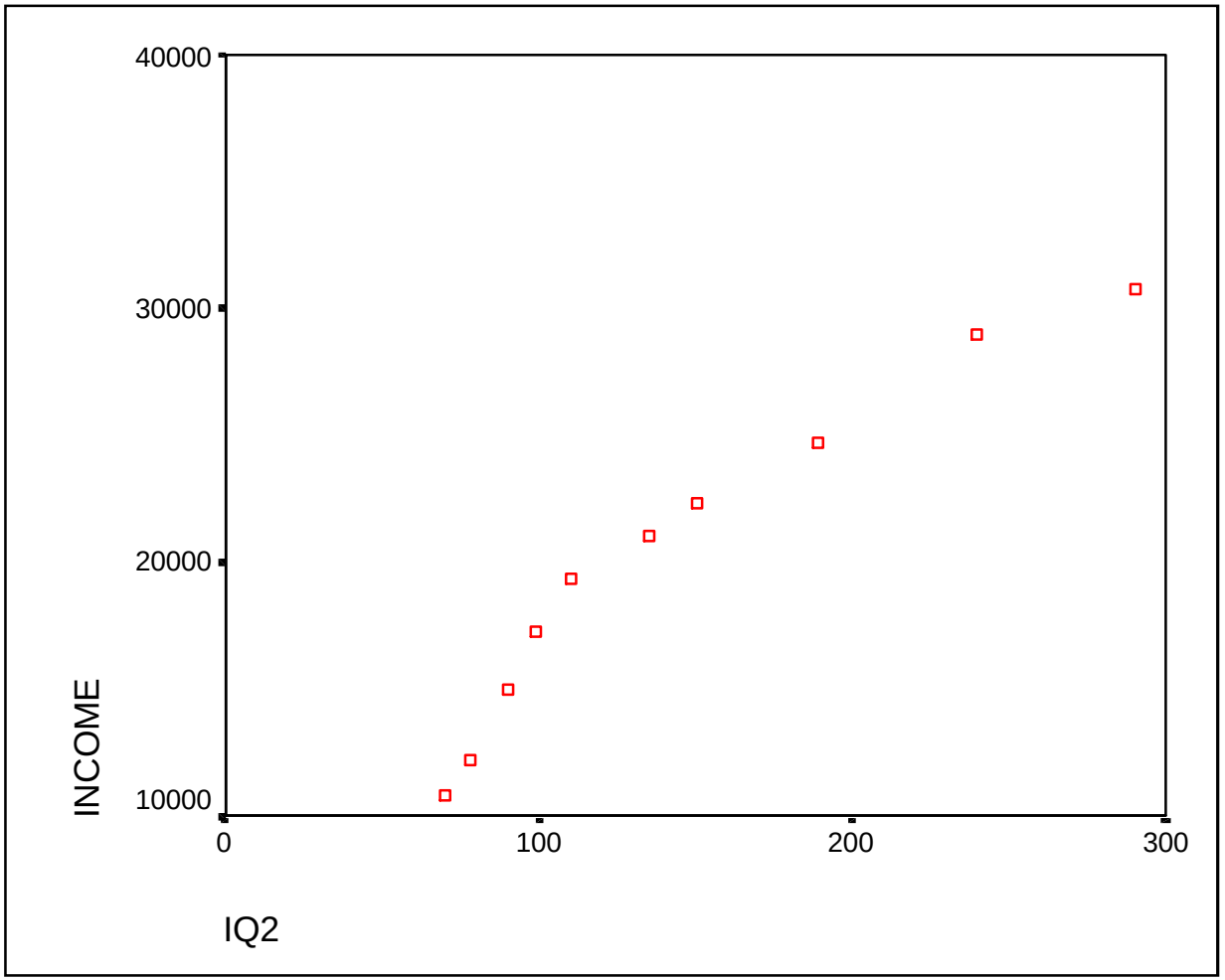


What does the output mean?

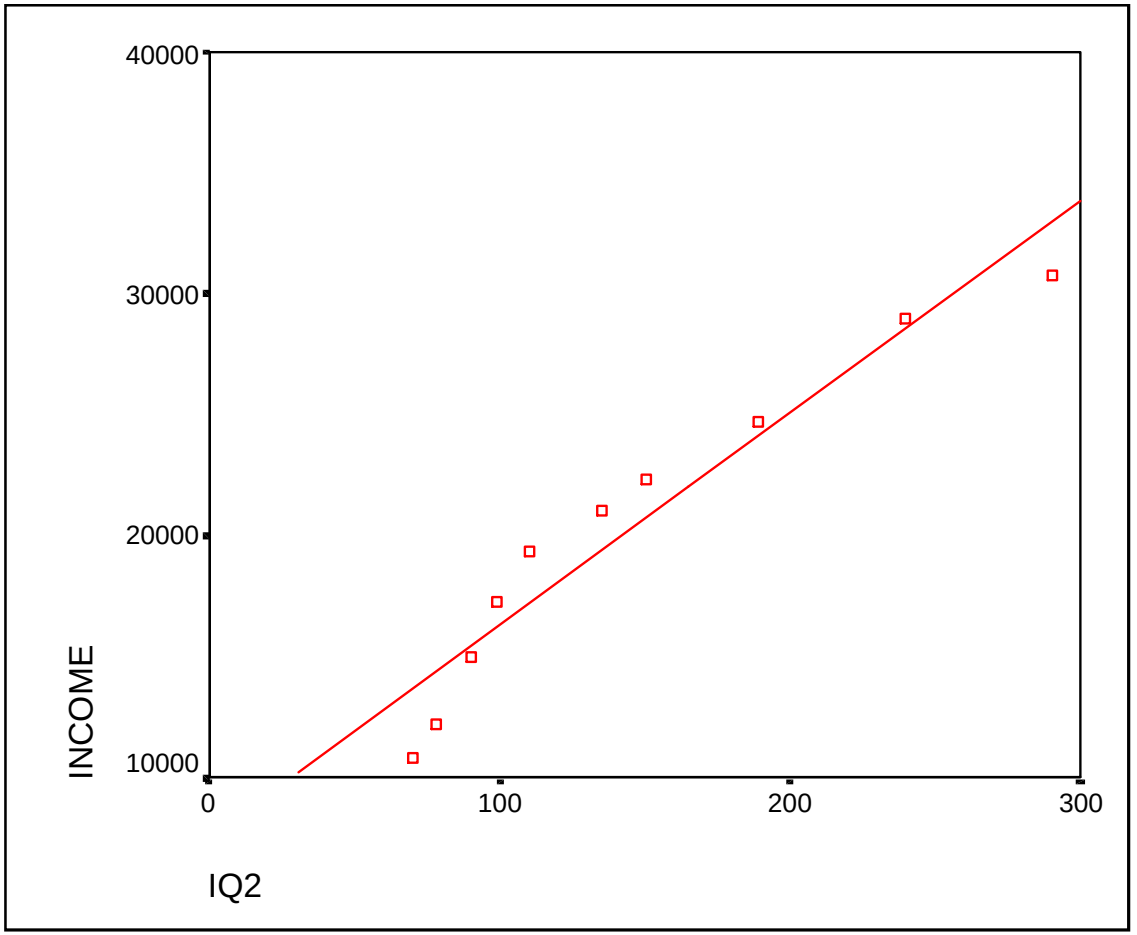
Model		B
1	(Constant)	-8236.836
	IQ	258.523

a. Dependent Variable: INCOME

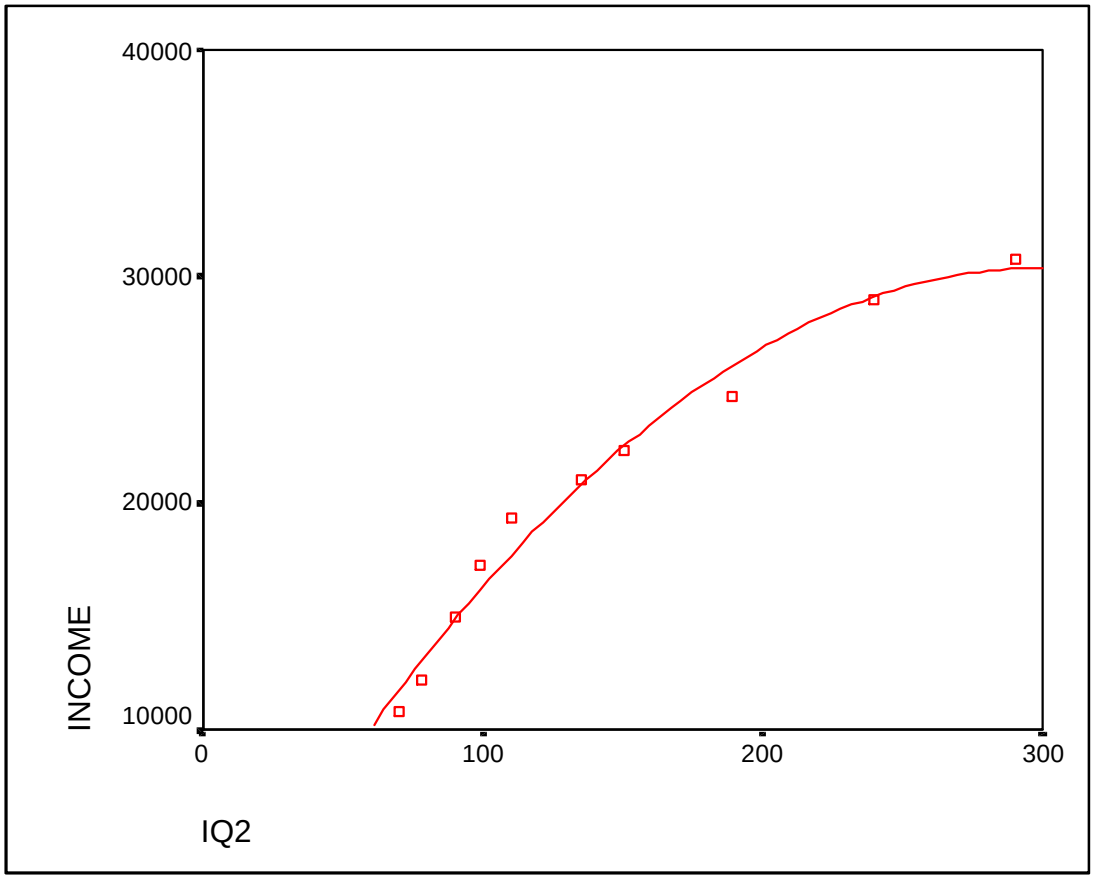
Sometimes the relationship appears non-linear:



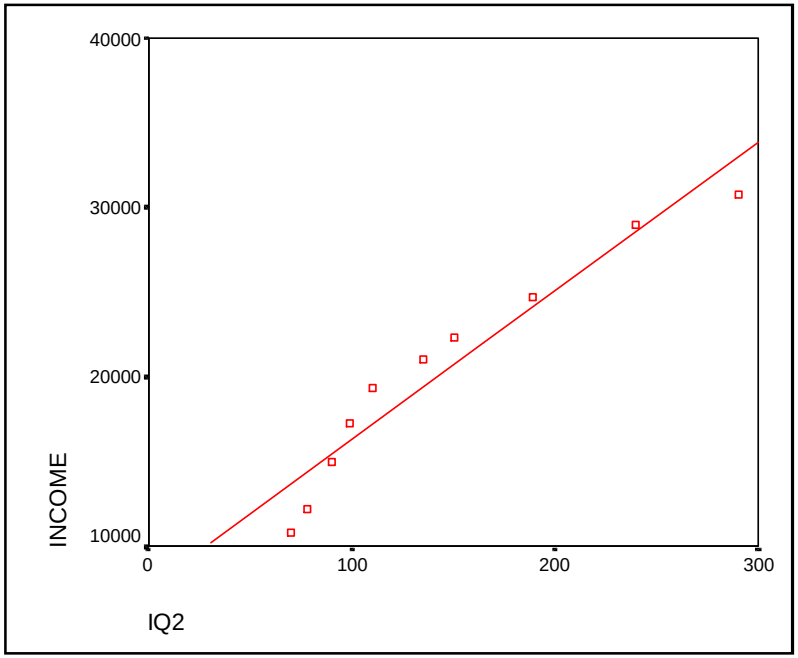
... and so a *straight* line of best fit is not always very satisfactory:

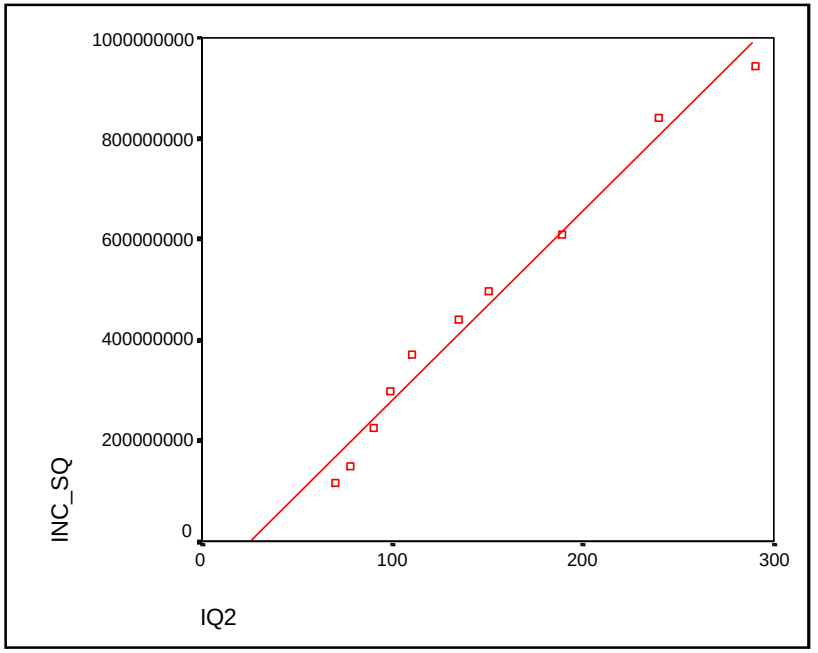


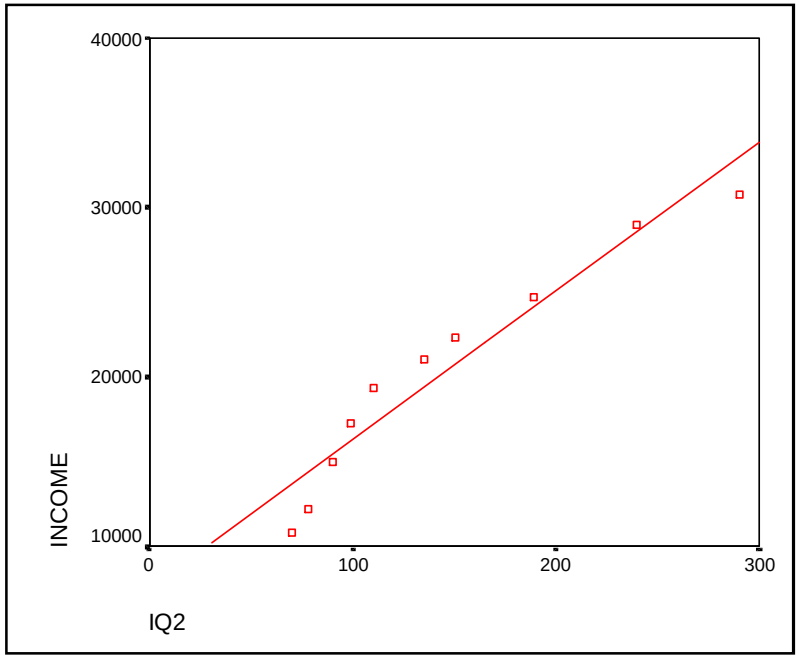
Could try a *quadratic* line of best fit:

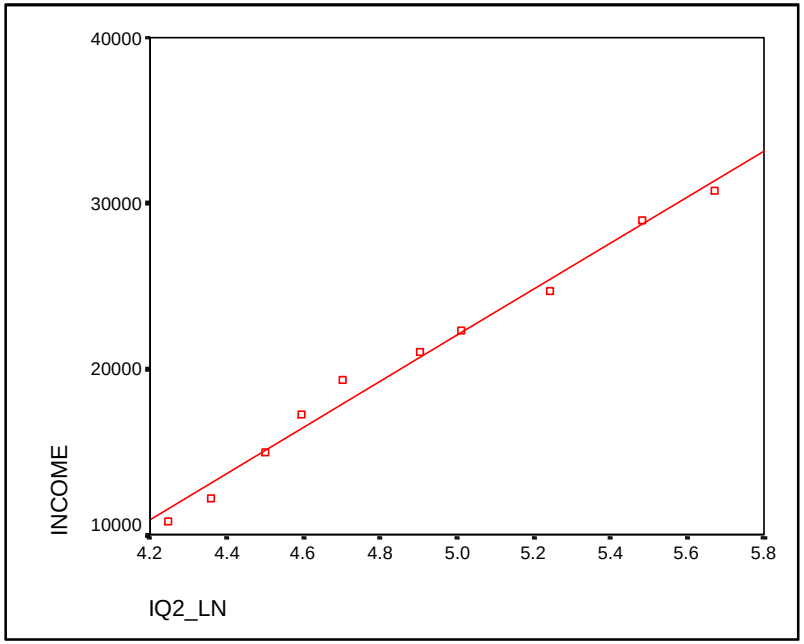


But we can simulate a non-linear relationship by first transforming one of the variables:

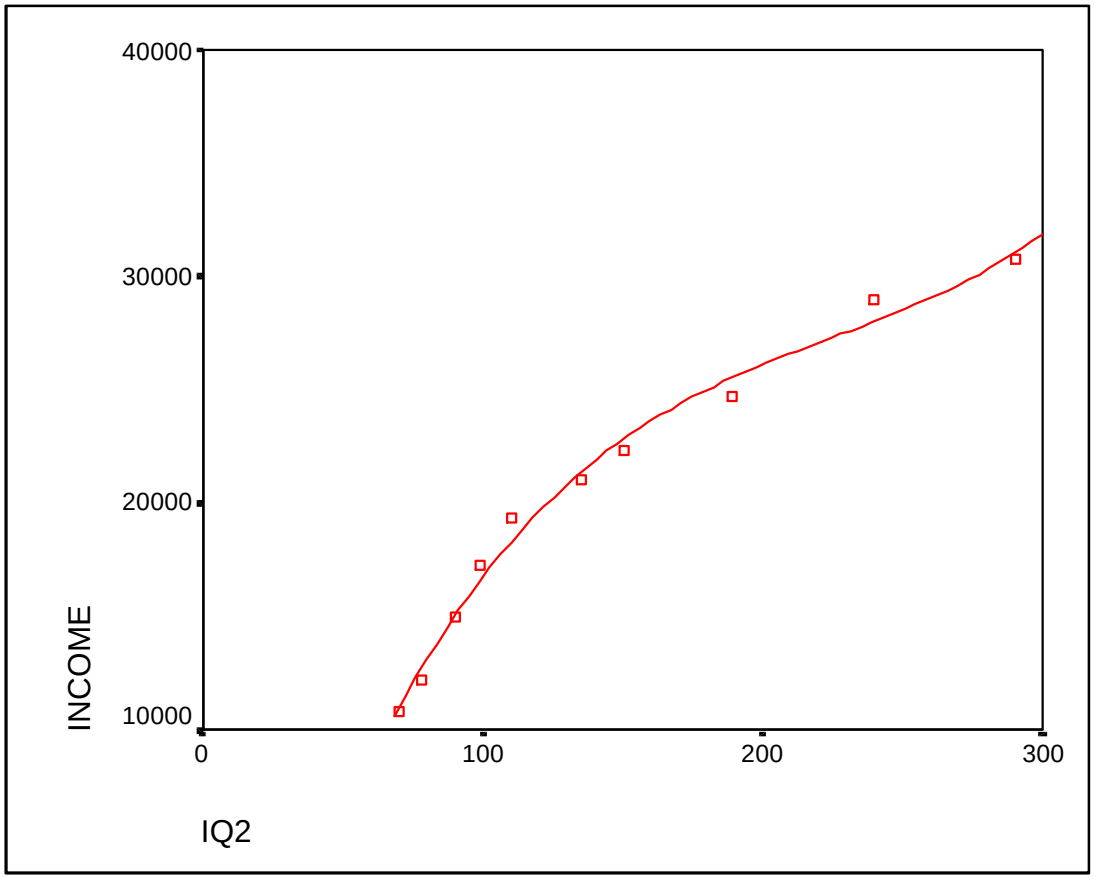




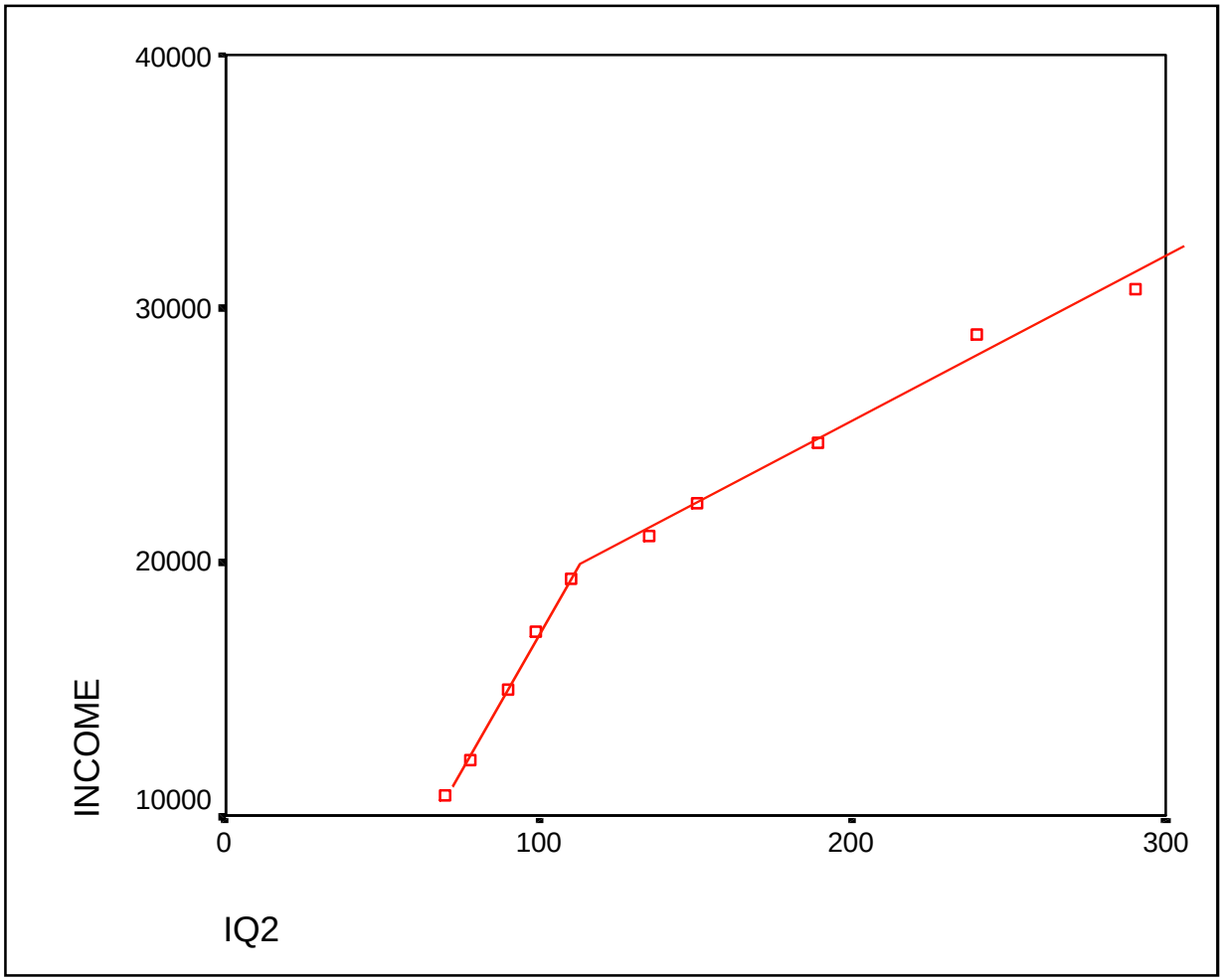

40000
30000
20000
10000
INCOME
0
100
200
300
IQ2


40000
30000
20000
10000
INCOME
4.2
4.4
4.6
4.8
5.0
5.2
5.4
5.6
5.8
IQ2_LN

… or a *cubic* line of best fit: (overfitted?)



Or could try two linear lines: “structural break”



Inference in Regression: Hypothesis tests on the slope coefficient:

- Regressions are usually run on samples, so what can we say about the *population* relationship between *x* and *y*?
- Repeated samples would yield a range of values for estimates of $b \sim N(\beta, s_b)$
  - I.e. *b* is normally distributed with mean = β = population mean = value of *b* if regression run on population
- If there is no relationship in the population between *x* and *y*, *then* $\beta = 0$, & this is our H_0

What does the standard error mean?

Coefficients [a]

Model		Unstandardized Coefficients	
		B	Std. Error
1	(Constant)	-8236.836	1250.461
	IQ	258.523	11.089

a. Dependent Variable: INCOME

Hypothesis test on *b*:

- (1) $H_0: \beta = 0$

  (I.e. slope coefficient, if regression run on population, would = 0)

  $H_1: \beta \neq 0$
- (2) $\alpha = 0.05$ or 0.01 etc.

$$t = \frac{b - \beta}{s_b} = \frac{b - 0}{s_b} = \frac{b}{s_b}$$

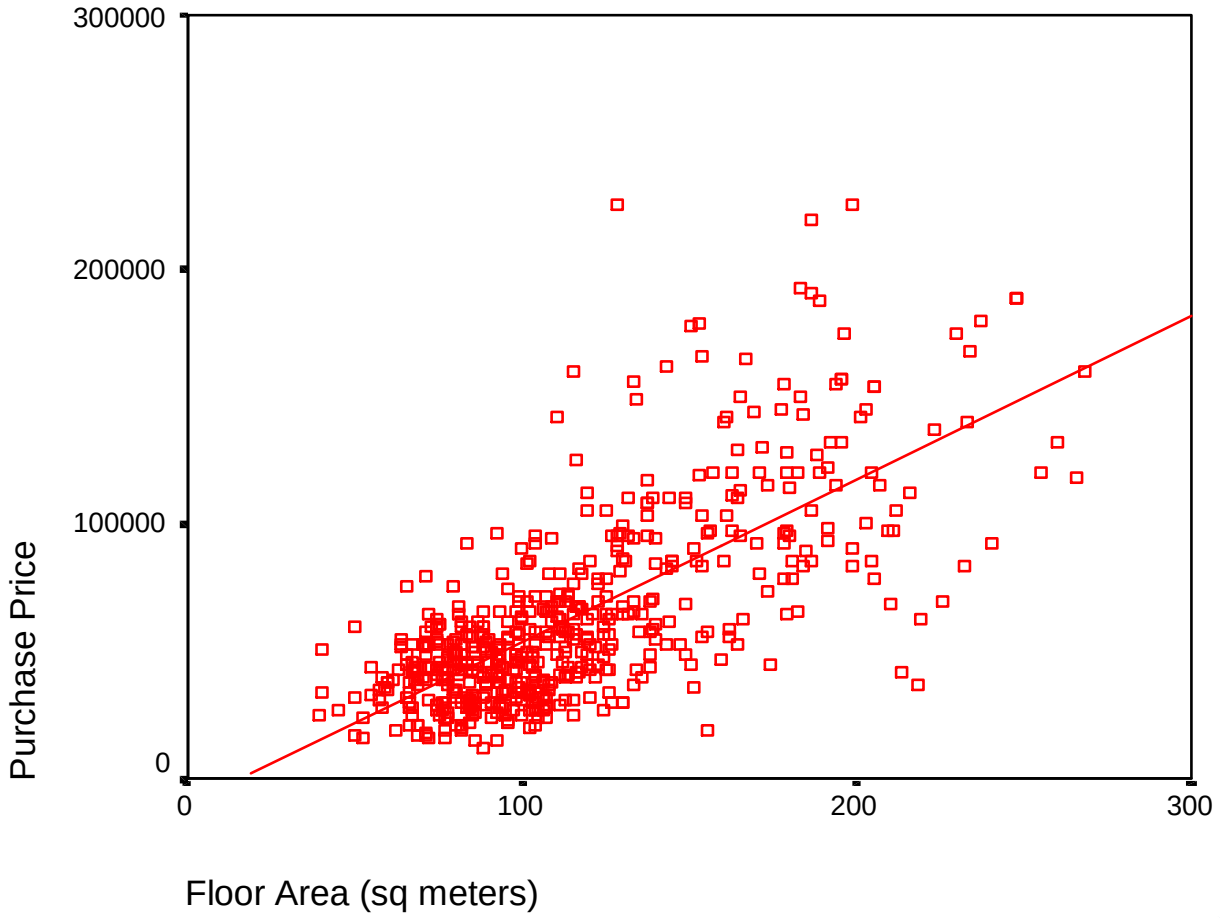
$$df = n - 2$$

- (3) Reject H_0 iff $P < \alpha$
  - (N.B. Rule of thumb: $P < 0.05$ if $t_c \geq 2$, and $P < 0.01$ if $t_c \geq 2.6$)

Omitted Variables & R^2

Q/ is floor area the only factor?

How much of the variation in Price does it explain?



R-square

- R-square tells you how much of the variation in *y* is explained by the explanatory variable *x*
  - $0 < R^2 < 1$ (NB: you want R^2 to be near 1).
  - If more than one explanatory variable, use Adjusted R^2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.721[a]	.519	.519	26925.18

a. Predictors: (Constant), Floor Area (sq meters)

Example: 2 explanatory variables

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-23817.3	3623.761		-6.573	.000
	Floor Area (sq meters)	507.271	30.722	.572	16.511	.000
	Number of Bathrooms	24951.769	3401.282	.254	7.336	.000

a. Dependent Variable: Purchase Price

Model Summary

R	R Square	Adjusted R Square
.750[a]	.562	.560

Types of regression analysis:

- *Univariate regression*: one explanatory variable
  - what we've looked at so far in the above equations
- *Multivariate regression*: >1 explanatory variable
  - more than one equation on the RHS
- *Log-linear regression & log-log regression*:
  - taking logs of variables can deal with certain types of non-linearities & useful properties (e.g. elasticities)
- *Categorical dependent variable regression*:
  - dependent variable is dichotomous -- observation has an attribute or not
    - e.g. unemployed or not etc.

Properties of OLS estimators

- OLS estimates of the slope and intercept parameters have been shown to be BLUE (provided certain assumptions are met):
  - Best
  - Linear
  - Unbiased
  - Estimator

  - Proofs

- "**Best**" in that they have the minimum variance compared with other estimators (i.e. given repeated samples, the OLS estimates for α and β vary less between samples than any other sample estimates for α and β).
- "**Linear**" in that a straight line relationship is assumed.
- "**Unbiased**" because, in repeated samples, the mean of all the estimates achieved will tend towards the population values for α and β.
- "**Estimates**" in that the true values of α and β cannot be known, and so we are using statistical techniques to arrive at the best possible assessment of their values, given the information available.

Assumptions of OLS

For estimation of *a* and *b* to be BLUE and for regression inference to be correct:

1. Equation is correctly specified:
   - Linear in parameters (can still transform variables)
   - Contains all relevant variables
   - Contains no irrelevant variables
   - Contains no variables with measurement errors
2. Error Term has zero mean
3. Error Term has constant variance

- 4. Error Term is not autocorrelated
  - I.e. correlated with error term from previous time periods
- 5. Explanatory variables are fixed
  - observe normal distribution of *y* for repeated fixed values of *x*
- 6. No linear relationship between RHS variables
  - I.e. no “multicollinearity”

Confidence Intervals for regression coefficients

- Population slope coefficient CI:

$$\beta = b \pm t_i SE_b$$

- Rule of thumb:

$$\beta \approx b \pm 2 \times SE_b$$

e.g. regression of floor area on number of bathrooms, CI on slope:

$$\beta = 64.6 \pm 2 \times 3.8$$

$$= 64.6 \pm 7.6$$

95% CI = $(57, 72)$

Coefficient of determination

Coefficient of determination is a measure in the regression analysis that shows the explanatory power independent variables (regressors) in explaining the variation on dependent variable (regressand). The total variation on the dependent variable can be decomposed as following:

$$Var(y_i) = \sum_i [y_i - \bar{y}]^2 = \sum_i [(\hat{y}_i - \bar{y}) + \hat{e}_i]^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i \hat{e}_i^2 + 2\sum_i (\hat{y}_i - \bar{y})e_i$$

$$Var(y_i) = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i \hat{e}_i^2$$ For T observations and K explanatory variables

[Total variation] = [Explained variation] + [Residual variation]

df = T-1 K-1 T-K

Coefficient of determination

$$1 = \frac{\sum (y_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} + \frac{\sum \hat{e}_i^2}{\sum (y_i - \bar{y})^2} = R^2 + \left(1 - R^2\right)$$

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}; \quad 0 \le R^2 \le 1$$

Total variation = explained variation + unexplained variation

R-SQUARE = 0.9807 R-SQUARE ADJUSTED = 0.9775

Testing for a Statistically Significant Regression

H_o: There is no relationship between Y and X.

H_A: There is a relationship between Y and X.

Which of two competing models is more appropriate?

Linear Model : $Y = \beta_0 + \beta_1 X + \varepsilon$

Mean Model : $Y = \mu + \varepsilon$

We look at the sums of squares of the prediction errors for the two models and decide if that for the linear model is **significantly smaller** than that for the mean model.

Concerns:

- **The proposed functional relationship will not fit exactly, i.e. something is either wrong with the data (errors in measurement), or the model is inadequate (errors in specification).**
- **The relationship is not truly known until we assign values to the parameters of the model.**

The possibility of errors into the proposed relationship is acknowledged in the functional symbolism as follows:

$$Y = f(X) + \varepsilon$$

ε is a random variable representing the result of both errors in model specification and measurement.

The error term:

Another way to emphasize

$$\varepsilon = Y - f(X)$$

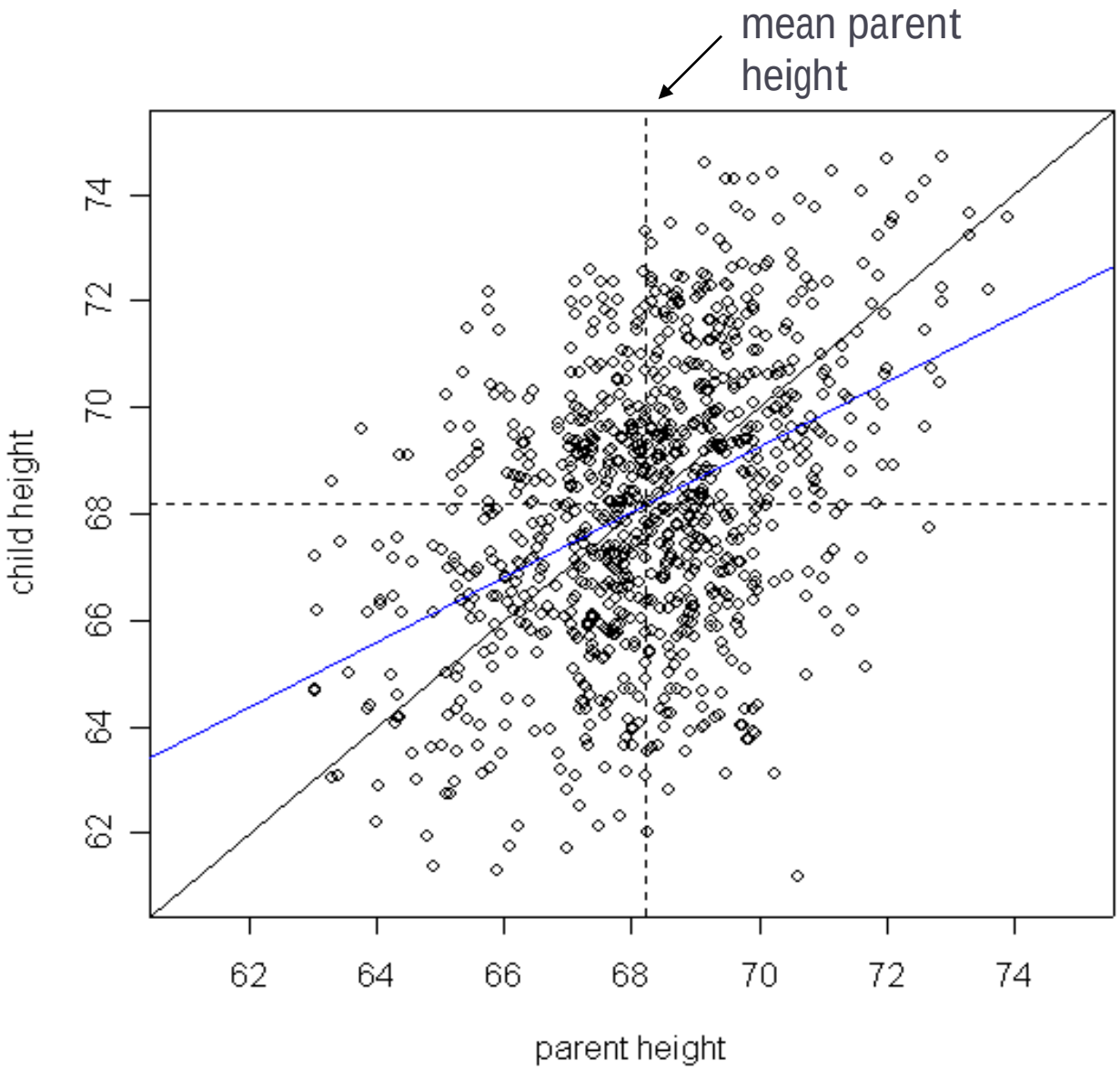
or, emphasizing that f(X) depends on unknown parameters.

$$Y = f(X \mid \beta_0, \beta_1) + \varepsilon$$

What if we don't know the functional form of the relationship?

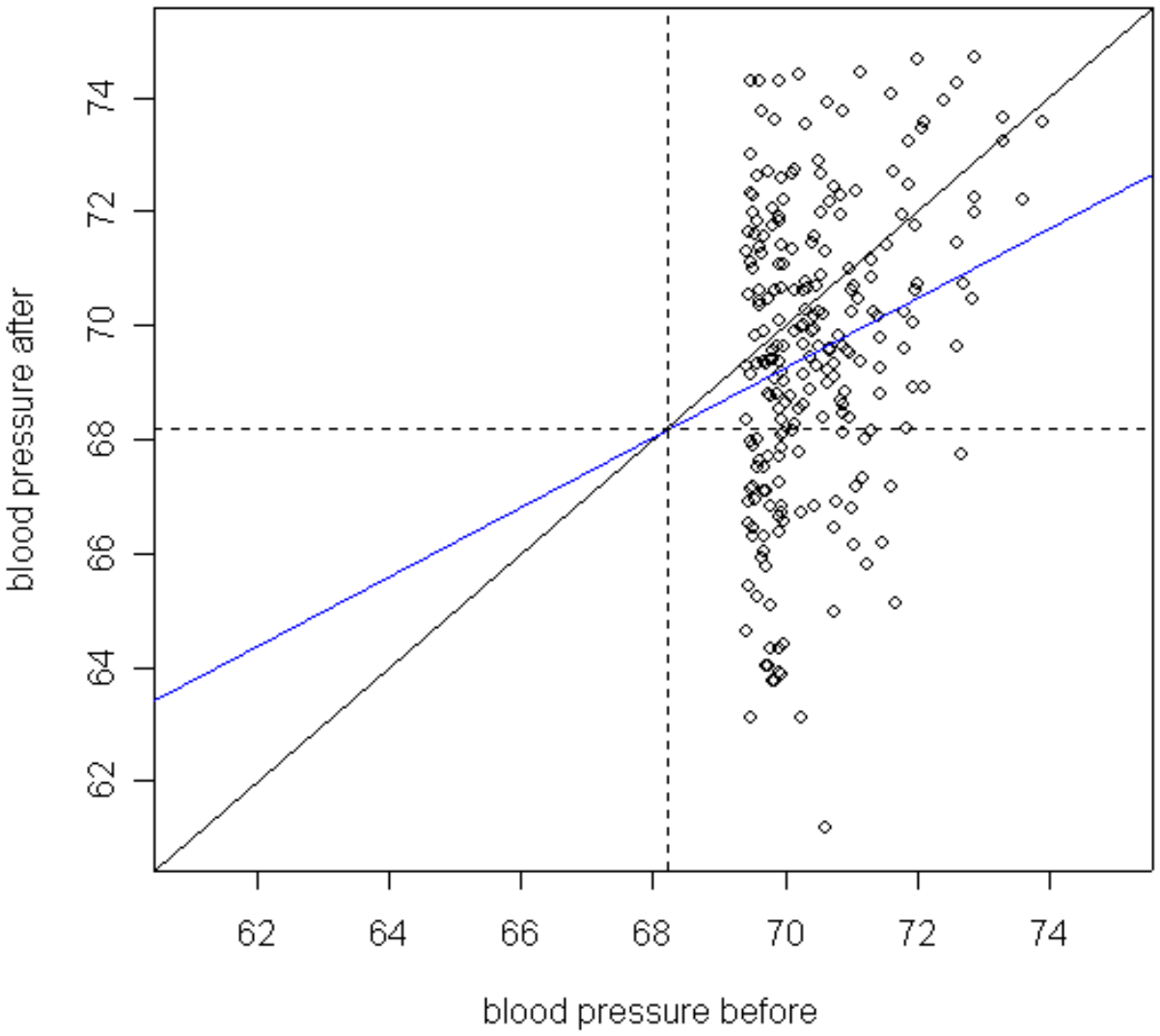
- Look at a scatter plot of the data for suggestions.
- Hypothesize about the nature of the underlying process. Often the hypothesized processes will suggest a functional form.

Regression to the Mean: Height Data



Data: 952 parent-child pairs of heights. Parent height is average of the two parents. Women's heights have been adjusted to make them comparable to men's.

Regression to the Mean is a Powerful Effect

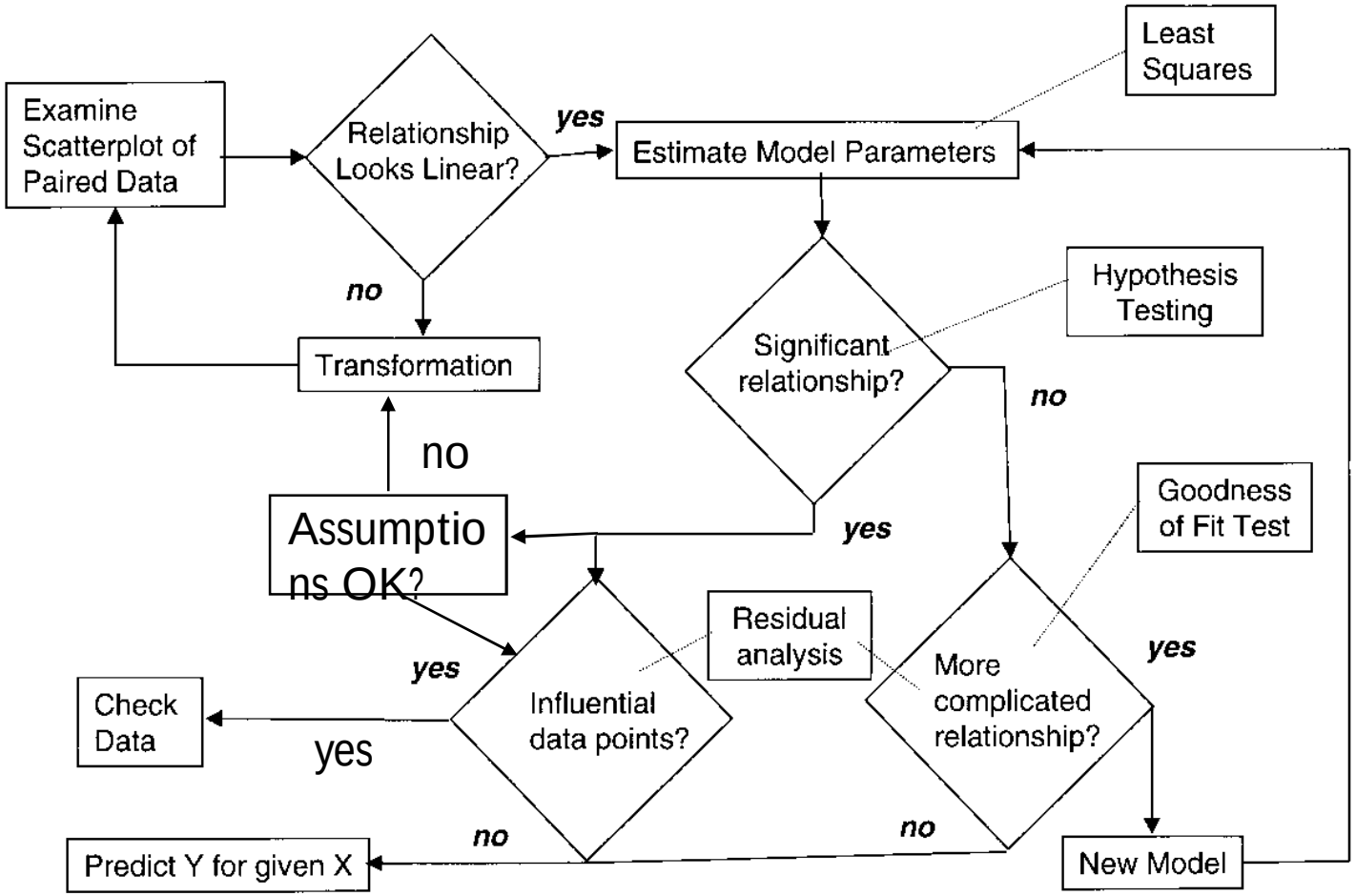


Same data, but suppose response is now blood pressure (bp) before & after (day 1, day 2).

If we track only those with elevated bp before (above 3rd quartile) , we see an amazing improvement, even though no treatment took place

This is the regression effect at work. If it is not recognized and taken into account, misleading results and biases can occur.

How is a Simple Linear Regression Analysis done? A Protocol



11-16

Steps in a Regression Analysis

1. Examine the scatterplot of the data.
    - Does the relationship look linear?
    - Are there points in locations they shouldn't be?
    - Do we need a transformation?
2. Assuming a linear function looks appropriate, estimate the regression parameters.
    - How do we do this? (Method of Least Squares)
3. Test whether there really is a statistically significant linear relationship. Just because we assumed a linear function it does not follow that the data support this assumption.
    - How do we test this? (F-test for Variances)
4. If there is a significant linear relationship, estimate the response, Y, for the given values of X, and compute the residuals.
5. Examine the residuals for systematic inadequacies in the linear model as fit to the data.
    - Is there evidence that a more complicated relationship (say a polynomial) should be considered; are there problems with the regression assumptions? (Residual analysis).
    - Are there specific data points which do not seem to follow the proposed relationship? (Examined using influence measures**).**

Simple Linear Regression - Example and Theory

SITUATION: A company that repairs small computers needs to develop a better way of providing customers typical repair cost estimates. To begin this process, they compiled data on repair times (in minutes) and the number of components needing repair or replacement from the previous week. The data, sorted by number of components are as follows:

i	Number of components x_i	Repair time y_i
1	1	23
2	2	29
3	4	64
4	4	72
5	4	80
6	5	87
7	6	96
8	6	105
9	8	127
10	8	119
11	9	145
12	9	149
13	10	165
14	10	154

Paired Observations (x_i, y_i)

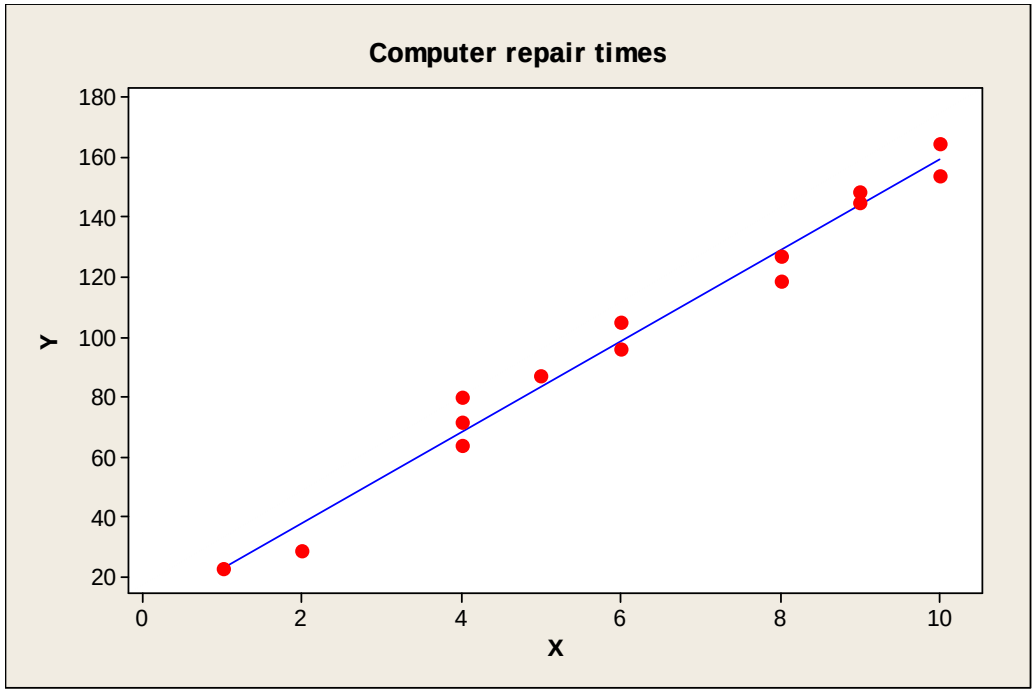
Assumed Linear Regression Model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$\text{for } i = 1, 2, \ldots, n$$

Estimating the regression parameters

Objective: Minimize the difference between the observation and its prediction according to the line.



$$\varepsilon_i = y_i - \hat{y}_i$$

$$= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$\hat{y}_i = \text{predicted y value when x} = x_i$$

We want the line which is best for all points. This is done by finding the values of β_0 and β_1 which minimizes some sum of errors. There are a number of ways of doing this. Consider these two:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} |\varepsilon_i|$$

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \varepsilon_i^2 \quad \longleftarrow$$

Sum of squared residuals

The method of least squares produces estimates with statistical properties (e.g. sampling distributions) which are easier to determine.

Regression => least squares estimation

$\hat{\beta}_0 \quad \hat{\beta}_1$ Referred to as least squares estimates.

Normal Equations

Calculus is used to find the least squares estimates.

$$E(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial E}{\partial \beta_0} = 0$$

$$\frac{\partial E}{\partial \beta_1} = 0$$

Solve this system of two equations in two unknowns.

Note: The parameter estimates will be functions of the data, hence they will be statistics.

Sums of Squares

Let:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2$$

$$= \sum_{i=1}^{n}(x_i^{\,2}) - \frac{1}{n}\left(\sum_{i-1}^{n} x_i\right)^2$$

Sums of squares of x.

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$= (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2$$

$$= \sum_{i=1}^{n}(y_i^{\,2}) - \frac{1}{n}\left(\sum_{i-1}^{n} y_i\right)^2$$

Sums of squares of y.

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$= (x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_i - \bar{y})$$

$$= \sum_{i=1}^{n}(x_i y_i) - \frac{1}{n}\left(\sum_{i-1}^{n} x_i\right)\left(\sum_{i-1}^{n} y_i\right)$$

Sums of cross products of x and y.

Parameter estimates:

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$$

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

Easy to compute with a spreadsheet program.
Easier to do with a statistical analysis package.

Example:

$$\hat{\beta}_1 = 7.71$$

$$\hat{\beta}_0 = 15.20$$

$$\hat{y}_i = 15.20 + 7.71 x_i$$ Prediction

Sums of Squares About the Mean (TSS)

Sum of squares about the mean: sum of the prediction errors for the null (mean model) hypothesis.

$$TSS = S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

TSS is actually a measure of the variance of the responses.

Residual Sums of Squares

Sum of squares for error: sum of the prediction errors for the alternative (linear regression model) hypothesis.

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

SSE measures the variance of the residuals, the part of the response variation that is not explained by the model.

Regression Sums of Squares

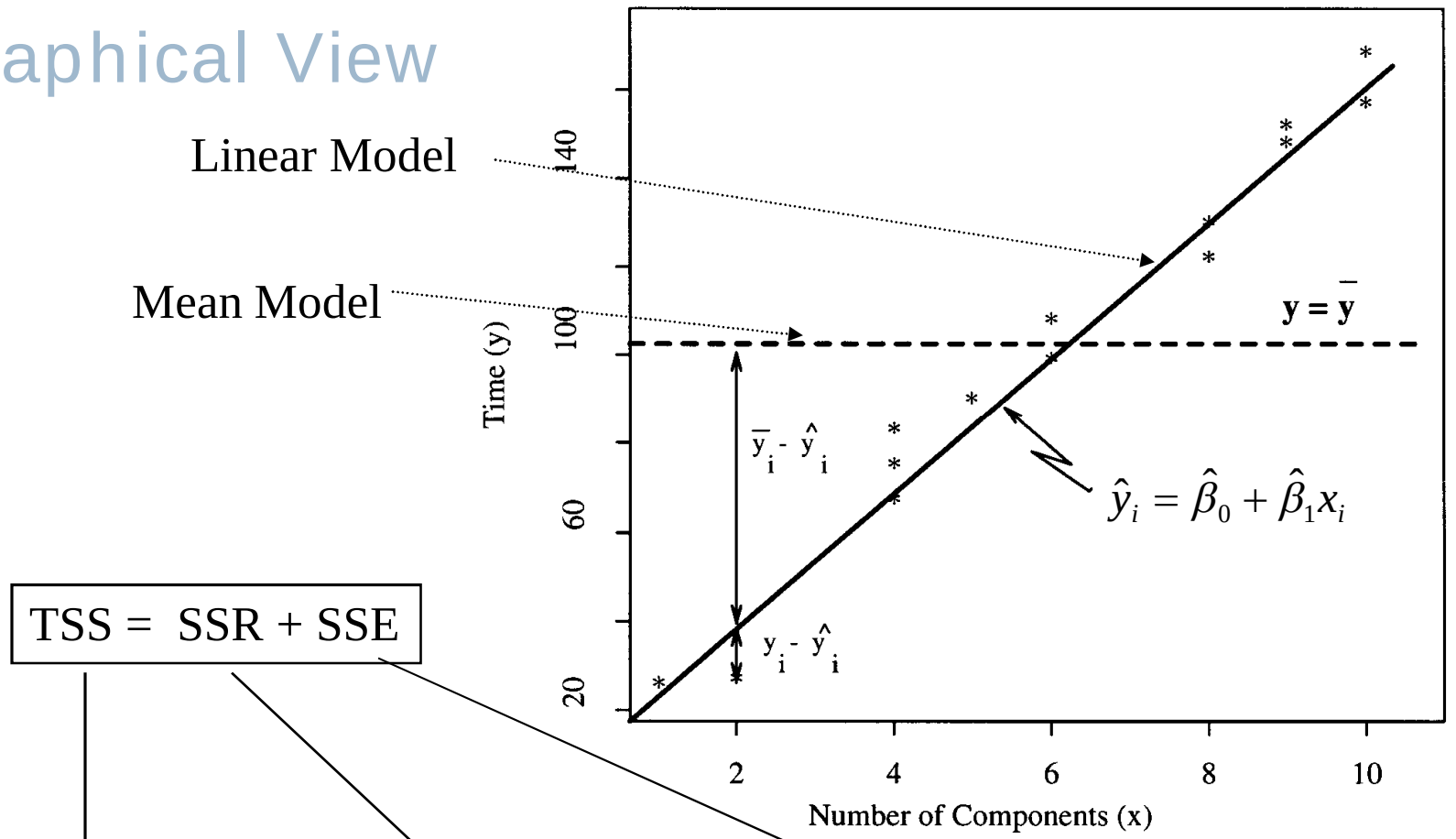
Sum of squares due to the regression: difference between TSS and SSE, i.e. SSR = TSS – SSE.

$$SSR = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} (\bar{y} - \hat{y}_i)^2$$

SSR measures how much variability in the response is explained by the regression.

Graphical View



TSS = SSR + SSE

Total variability in y-values = Variability accounted for by the regression + Unexplained variability

$$TSS = SSR + SSE$$

Total variability in y-values	=	Variability accounted for by the regression	+	Unexplained variability

regression model fits well

Then SSR approaches TSS and SSE gets small.

regression model adds little

Then SSR approaches 0 and SSE approaches TSS.

Mean Square Terms

Mean Square Total

Sample variance of the response, y:

$$\hat{\sigma}_{\mathrm{T}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{\mathrm{TSS}}{n-1} = \mathrm{MST}$$

$$\hat{\sigma}_{\mathrm{R}}^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \frac{\mathrm{SSR}}{1} = \mathrm{MSR}$$

Regression Mean Square:

Residual Mean Square

$$\hat{\sigma}_{\varepsilon}^2 = \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{\mathrm{SSE}}{n-2} = \mathrm{MSE}$$

F Test for Significant Regression

Both MSE and MSR measure the same underlying variance quantity under the assumption that the null (mean) model holds.

$$\sigma_R^2 \approx \sigma_\varepsilon^2$$

Under the alternative hypothesis, the MSR should be much greater than the MSE.

$$\sigma_R^2 > \sigma_\varepsilon^2$$

Placing this in the context of a test of variance.

$$F = \frac{\sigma_R^2}{\sigma_\varepsilon^2} = \frac{\text{MSR}}{\text{MSE}}$$

Test Statistic

F should be near 1 if the regression is *not significant*, i.e. H_0: mean model holds.

Formal test of the significance of the regression.

H_0: No significant regression fit.

H_A: The regression explains a significant amount of the variability in the response.

or

The slope of the regression line is significant.

or

X is a significant predictor of Y.

Test Statistic: $$F = \frac{MSR}{MSE}$$

Reject H_0 if: $$F > F_{1,n-2,\alpha}$$

Where α is the probability of a type I error.

Analysis of Variance Table

We summarize the computations of this test in a table.

Source	Sums of Squares SSQ	Degrees of Freedom DF	Mean Squares MS	F
Regression	SSR	1	MSR	$F = \frac{MSR}{MSE}$
Error	SSE	n-2	MSE	
Total	SSM	n-1		

TSS

Parameter Standard Error Estimates

Under the assumptions for regression inference, the least squares estimates themselves are random variables.

1. $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n$ are independent of each other.
2. The ε_i are normally distributed with mean zero and have common variance σ^2.

Using some more calculus and mathematical statistics we can determine the distributions for these parameters.

$$\hat{\beta}_0 \mapsto \mathrm{N}\left(\beta_0, \sigma^2 \frac{\sum x_i^2}{nS_{XX}}\right) \qquad \hat{\beta}_1 \mapsto \mathrm{N}\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right)$$

Testing regression parameters

The estimate of σ^2 is the mean square error: MSE

important

$$\hat{\sigma}^2 = MSE$$

Test H_0: $\beta_1=0$:

$$t_{\beta_1} = \frac{\hat{\beta}_1 - 0}{\sqrt{MSE / S_{XX}}}$$

Reject H_0 if:

$$\left| t_{\beta_1} \right| > t_{n-2,\alpha/2}$$

(1-α)100% CI for β_1:

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} \sqrt{\frac{MSE}{S_{XX}}}$$