

# Applied Regression Modeling: A Business Approach

## Chapter 2: Simple Linear Regression Sections 2.1–2.3

by Iain Pardoe

<b>2.1 Probability model for X and Y</b>	<b>2</b>
Simple linear regression model for X and Y . . . . .	2
Possible relationships between X and Y . . . . .	3
Straight-line model . . . . .	4
HOMES2 data . . . . .	5
Simple linear regression model equation . . . . .	6
<b>2.2 Least squares criterion</b>	<b>7</b>
Least squares criterion . . . . .	7
Estimating the model . . . . .	8
Estimated equation . . . . .	9
Computer output . . . . .	10
<b>2.3 Model evaluation</b>	<b>11</b>
Model evaluation . . . . .	11
Evaluating fit numerically . . . . .	12
Regression standard error, $s$ . . . . .	13
Regression standard error interpretation . . . . .	14
Coefficient of determination, $R^2$ . . . . .	15
Calculating $R^2$ . . . . .	16
Interpreting $R^2$ . . . . .	17
$R^2$ examples . . . . .	18
Correlation . . . . .	19
Correlation examples . . . . .	20
Slope parameter, $b_1$ . . . . .	21
Hypothesis test for $b_1$ . . . . .	22
Computer output and slope test illustration . . . . .	23
Slope confidence interval . . . . .	24

### Simple linear regression model for $X$ and $Y$

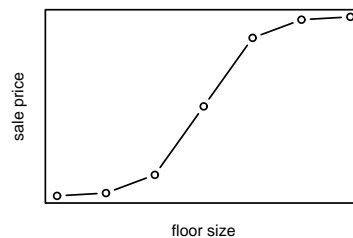
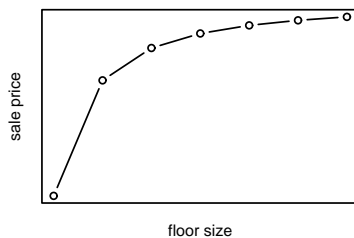
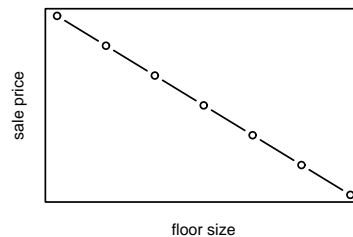
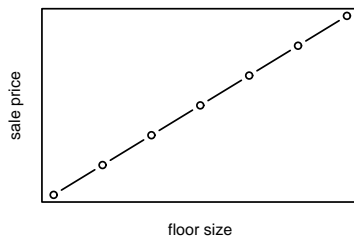
- $Y$  is a quantitative *response* variable (a.k.a. dependent, outcome, or output variable).
- $X$  is a quantitative *predictor* variable (a.k.a. independent or input variable, or covariate).
- Two variables play different roles, so important to identify which is which and define carefully, e.g.:
  - $Y$  is sale price, in \$ thousands;
  - $X$  is floor size, in thousands of square feet.
- How much do we expect  $Y$  to change by when we change the value of  $X$ ?
- What do we expect the value of  $Y$  to be when we set the value of  $X$  at 2?
- Note: *association* (observational data) not *causation* (experimental data).

© Iain Pardoe, 2006

2 / 24

### Possible relationships between $X$ and $Y$

Which factors might lead to the different relationships?



© Iain Pardoe, 2006

3 / 24

## Straight-line model

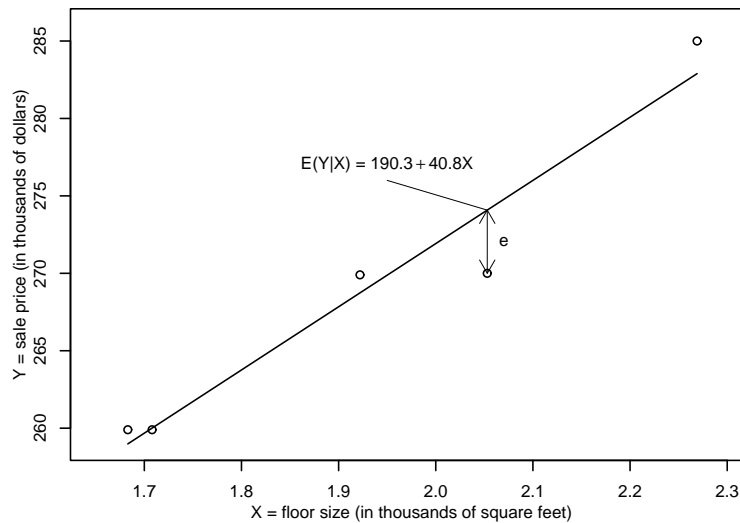
- Simple linear regression models straight-line relationships (like upper two plots on last slide).
- Suppose sale price is (on average) \$190,300 plus 40.8 times floor size.
  - $E(Y|X_i) = 190.3 + 40.8X_i$ ,  
where  $E(Y|X_i)$  means “the expected value of  $Y$  given that  $X$  is equal to  $X_i$ ”.
- Individual sale prices can deviate from this expected value by an amount  $e_i$  (called a “random error”).
  - $Y_i|X_i = 190.3 + 40.8X_i + e_i$  ( $i = 1, \dots, n$ ).
  - $Y_i|X_i = \text{deterministic part} + \text{random error}$ .
- Error,  $e_i$ , represents variation in  $Y$  due to factors other than  $X$  which we haven't measured, e.g., lot size, # beds/baths, age, garage, schools.

© Iain Pardoe, 2006

4 / 24

## HOMES2 data

$Y$	259.9	259.9	269.9	270.0	285.0
$X$	1.683	1.708	1.922	2.053	2.269

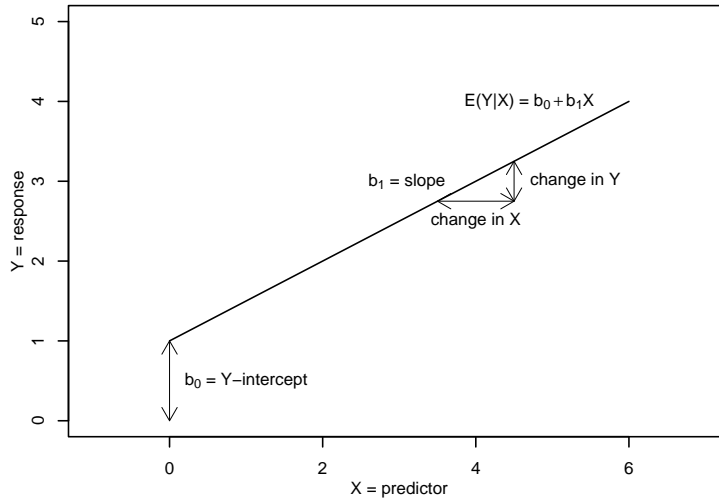


© Iain Pardoe, 2006

5 / 24

## Simple linear regression model equation

Population:  $E(Y|X) = b_0 + b_1X$ .



© Iain Pardoe, 2006

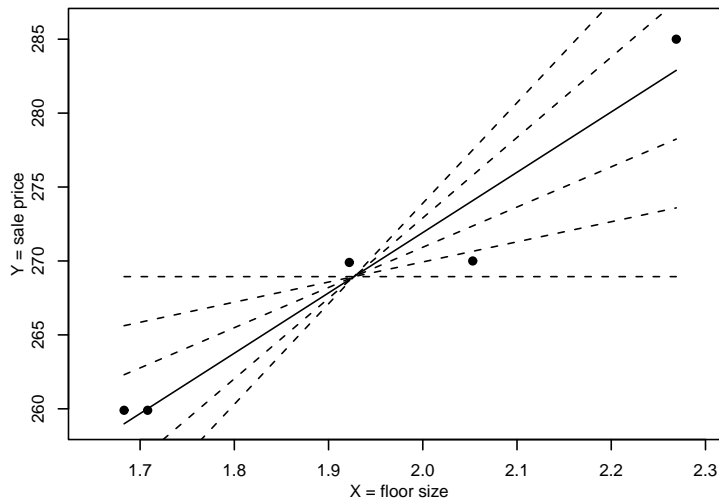
6 / 24

## 2.2 Least squares criterion

7 / 24

### Least squares criterion

Which line fits the data best?



© Iain Pardoe, 2006

7 / 24

## Estimating the model

- Population:  $E(Y | X) = b_0 + b_1X$ .
- Sample:  $\hat{Y} = \hat{b}_0 + \hat{b}_1X$  (estimated model).
- Obtain  $\hat{b}_0$  and  $\hat{b}_1$  by finding best fit line (least squares line).
- Mathematically, minimize sum of squared errors:

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^n \hat{e}_i^2 \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^n (Y_i - \hat{b}_0 - \hat{b}_1X_i)^2. \end{aligned}$$

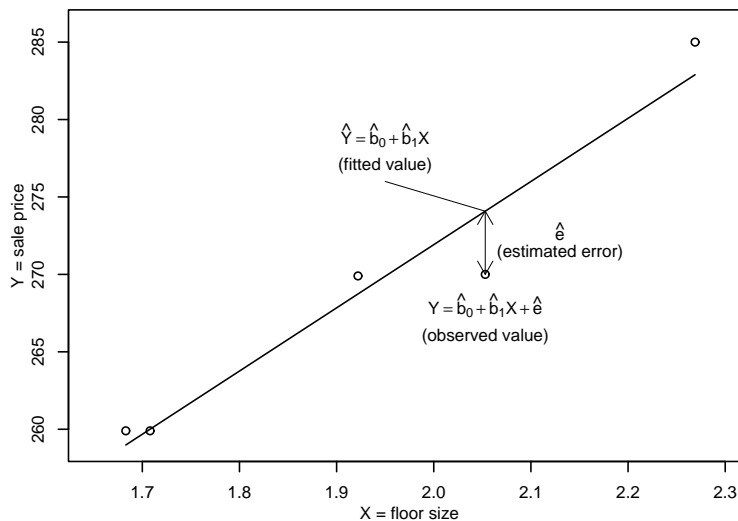
- Can use calculus (partial derivatives), but we'll use computer software to find  $\hat{b}_0$  and  $\hat{b}_1$ .

© Iain Pardoe, 2006

8 / 24

## Estimated equation

Sample:  $\hat{Y} = \hat{b}_0 + \hat{b}_1X$ .



© Iain Pardoe, 2006

9 / 24

## Computer output

		Parameters <sup>a</sup>			
Model		Estimate	Std. Error	t-stat	Pr(> t )
1	(Intercept)	190.318	11.023	17.266	0.000
	X	40.800	5.684	7.179	0.006

<sup>a</sup> Response variable: Y.

- Estimated equation:  $\hat{Y} = \hat{b}_0 + \hat{b}_1 X = 190.3 + 40.8X$ .
- We expect  $Y = \hat{b}_0$  when  $X = 0$ , but *only* if this makes sense and we have data close to  $X = 0$  (not the case here).
- We expect  $Y$  to change by  $\hat{b}_1$  when  $X$  increases by one unit, i.e., we expect sale price to increase by \$40,800 when floor size increases by 1000 sq. feet.
- For this example, more meaningful to say we expect sale price to increase by \$4080 when floor size increases by 100 sq. feet.

© Iain Pardoe, 2006

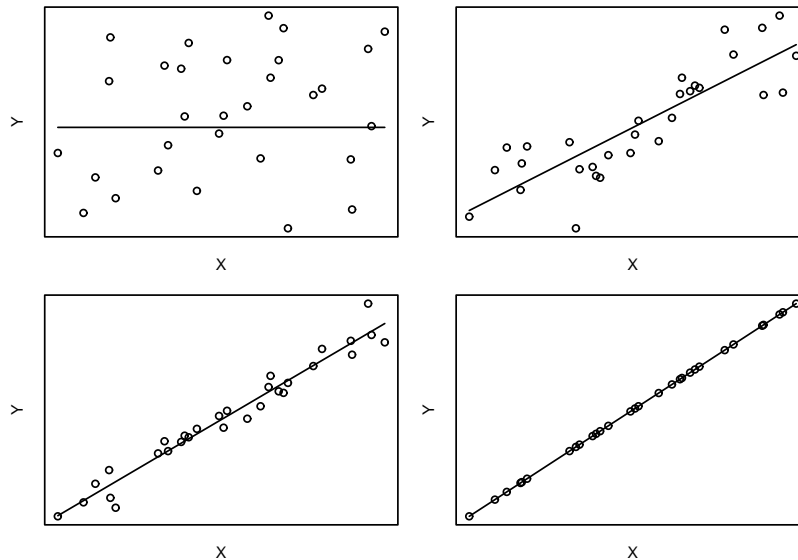
10 / 24

## 2.3 Model evaluation

11 / 24

### Model evaluation

How well does the model fit each dataset?



© Iain Pardoe, 2006

11 / 24

## Evaluating fit numerically

Three methods:

- How close are the actual observed  $Y$ -values to the model-based fitted values,  $\hat{Y}$ ?
  - Calculate the *regression standard error*,  $s$ .
- How much of the variability in  $Y$  have we been able to explain with our model?
  - Calculate the *coefficient of determination*,  $R^2$ .
- How strong is the evidence of a straight-line relationship between  $Y$  and  $X$ ?
  - Estimate and test the *slope parameter*,  $b_1$ .

© Iain Pardoe, 2006

12 / 24

## Regression standard error, $s$

### Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.972 <sup>a</sup>	0.945	0.927	2.7865

<sup>a</sup> Predictors: (Intercept), X.

- Regression standard error,  $s$ , estimates the std. dev. of the simple linear regression random errors:

$$s = \sqrt{\frac{\text{SSE}}{n - 2}}$$

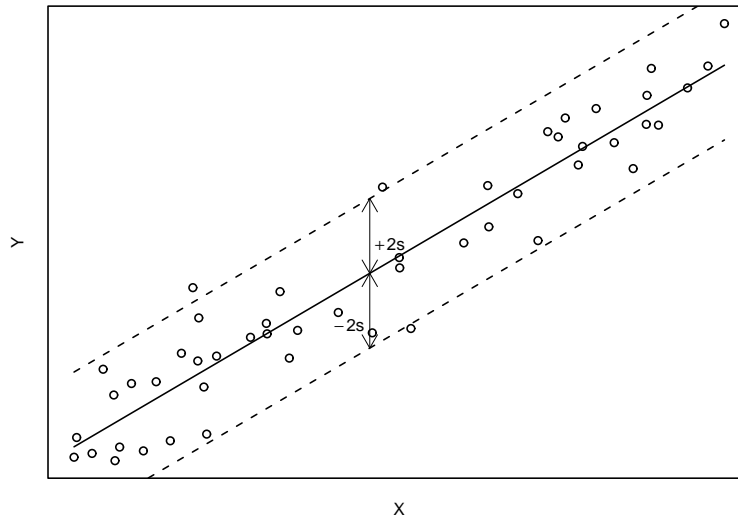
- Unit of measurement for  $s$  is the same as unit of measurement for  $Y$ .
- Approximately 95% of the observed  $Y$ -values lie within plus or minus  $2s$  of their fitted  $\hat{Y}$ -values.
- Since  $2s = 5.57$ , we can expect to predict an unobserved sale price from a particular floor size to within approx.  $\pm \$5570$  (at a 95% confidence level).

© Iain Pardoe, 2006

13 / 24

## Regression standard error interpretation

CLT: 95% of  $Y$ -values lie within  $\pm 2s$  of regression line.

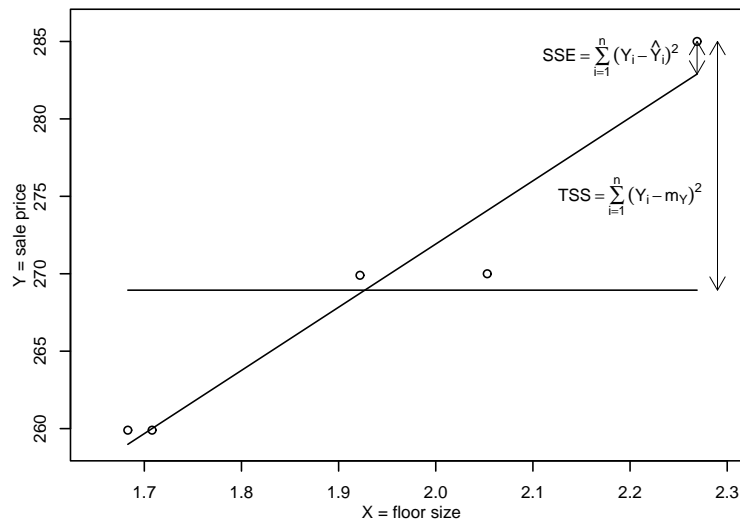


© Iain Pardoe, 2006

14 / 24

## Coefficient of determination, $R^2$

Measures of variation for simple linear regression.



© Iain Pardoe, 2006

15 / 24



## Calculating R<sup>2</sup>

- Without model, estimate  $Y$  with sample mean  $m_Y$ .
- With model, estimate  $Y$  using fitted  $\hat{Y}$ -value.
- How much do we reduce our error when we do this?
- Total error without model:  
 $TSS = \sum_{i=1}^n (Y_i - m_Y)^2$ , variation in  $Y$  about  $m_Y$ .
- Remaining error with model:  
 $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ , unexplained variation.
- Proportional reduction in error:  $R^2 = \frac{TSS - SSE}{TSS}$ .
- Home prices example:  $R^2 = \frac{423.4 - 23.3}{423.4} = 0.945$ .
- 94.5% of the variation in sale price (about its mean) can be explained by a straight-line relationship between sale price and floor size.

© Iain Pardoe, 2006

16 / 24

## Interpreting R<sup>2</sup>

### Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.972 <sup>a</sup>	0.945	0.927	2.7865

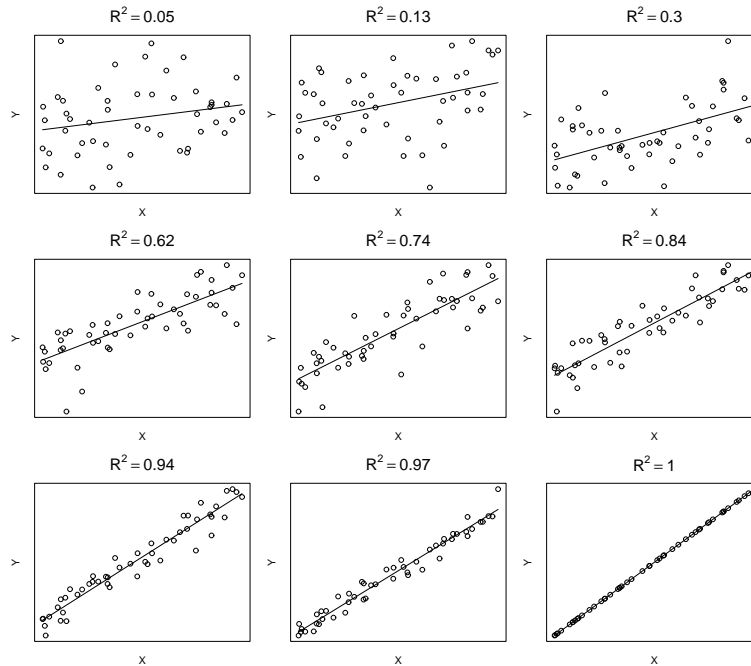
<sup>a</sup> Predictors: (Intercept), X.

- R<sup>2</sup> measures the proportion of variation in  $Y$  (about its mean) that can be explained by a straight-line relationship between  $Y$  and  $X$ .
- If  $TSS = SSE$  then  $R^2 = 0$ : using  $X$  to predict  $Y$  hasn't helped and we might as well use  $m_Y$  to predict  $Y$  regardless of the value of  $X$ .
- If  $SSE = 0$  then  $R^2 = 1$ : using  $X$  allows us to predict  $Y$  perfectly (with no random errors).
- Such extremes rarely occur and usually  $R^2$  lies between zero and one, with higher values of  $R^2$  corresponding to better fitting models.

© Iain Pardoe, 2006

17 / 24

## R<sup>2</sup> examples



© Iain Pardoe, 2006

18 / 24

## Correlation

### Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
1	0.972 <sup>a</sup>	0.945	0.927	2.7865

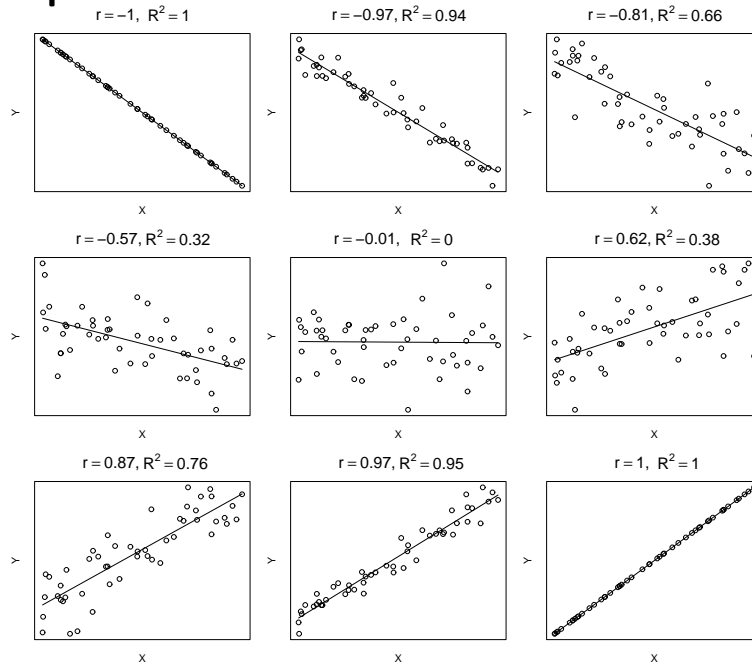
<sup>a</sup> Predictors: (Intercept), X.

- Correlation coefficient,  $r$ , measures the strength and direction of linear association between  $Y$  and  $X$ :
  - $r \approx -1$  indicates a negative linear relationship;
  - $r \approx +1$  indicates a positive linear relationship;
  - $r \approx 0$  indicates no *linear* relationship.
- Simple linear regression:  $\sqrt{R^2} =$  absolute value of  $r$  ("multiple R" above).
- But, correlation is less useful than  $R^2$  in *multiple* linear regression.

© Iain Pardoe, 2006

19 / 24

## Correlation examples

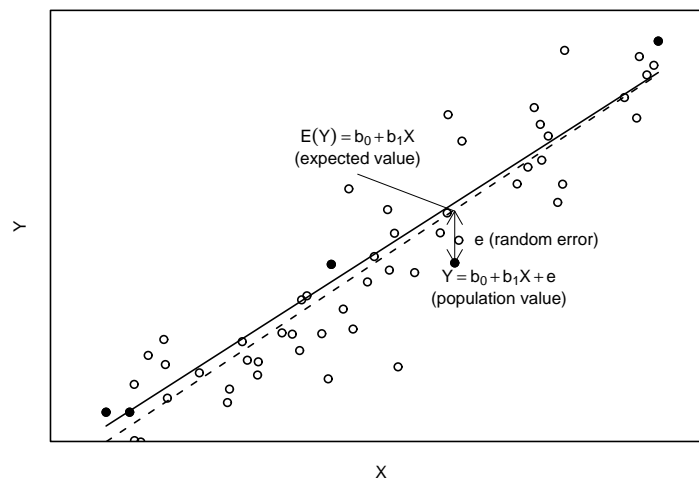


© Iain Pardoe, 2006

20 / 24

## Slope parameter, $b_1$

Infer from sample slope about the population slope.



© Iain Pardoe, 2006

21 / 24

## Hypothesis test for $b_1$

- Recall univariate t-statistic =  $\frac{m_Y - E(Y)}{s_Y / \sqrt{n}} \sim t_{n-1}$ .
- Here, slope t-statistic =  $\frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} \sim t_{n-2}$ .
- NH:  $b_1 = 0$  versus AH:  $b_1 \neq 0$ .
- t-statistic =  $\frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{40.8 - 0}{5.684} = 7.18$ .
- Significance level = 5%.
- Critical value is 3.182 (97.5<sup>th</sup> percentile of  $t_3$ ).
- Since t-statistic (7.18) is between 5.841 and 10.215, p-value is between 0.01 and 0.002.
- Since t-statistic (7.18) > critical value (3.182) and p-value < signif. level, reject NH in favor of AH.
- In other words, the sample data favor a nonzero slope (at a significance level of 5%).
- Exercise: do an upper tail test for this example.

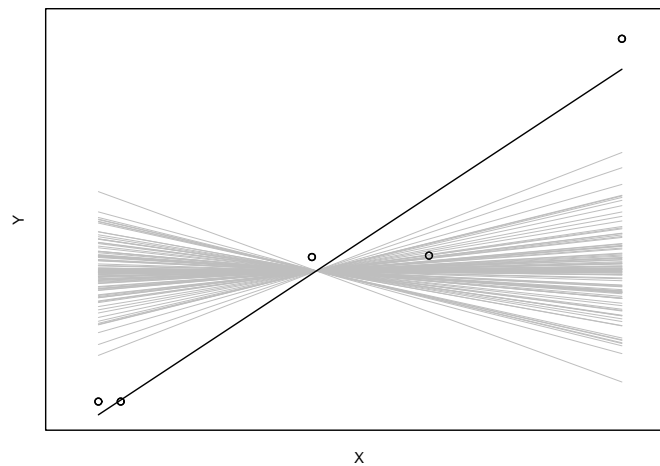
© Iain Pardoe, 2006

22 / 24

## Computer output and slope test illustration

Parameters <sup>a</sup>					
Model		Estimate	Std. Error	t-stat	Pr(> t )
1	(Intercept)	190.318	11.023	17.266	0.000
	X	40.800	5.684	7.179	0.006

<sup>a</sup> Response variable: Y.



© Iain Pardoe, 2006

23 / 24

## Slope confidence interval

- Calculate a 95% confidence interval for  $b_1$ .
- 97.5<sup>th</sup> percentile of  $t_3$  is 3.182.
- $\hat{b}_1 \pm 97.5^{\text{th}} \text{ percentile}(s_{\hat{b}_1}) = 40.8 \pm 3.182 \times 5.684 = 40.8 \pm 18.1 = (22.7, 58.9)$ .
- Loosely speaking: based on this dataset, we are 95% confident that the population slope,  $b_1$ , is between 22.7 and 58.9.
- More precisely: if we were to take a large number of random samples of size 5 from our population of homes and calculate a 95% confidence interval for each, then 95% of those confidence intervals would contain the (unknown) population slope.
- Exercise: calculate a 90% confidence interval for  $b_1$ .