CS 147:
Computer Systems Performance Analysis
Linear Regression Models

2015-06-15

CS147

CS 147:
Computer Systems Performance Analysis
Linear Regression Models

Overview

Overview
What is a (good) model?
Estimating Model Parameters
Allocating Variation
Confidence Intervals for Regressions
Parameter Intervals
Prediction Intervals
Verifying Regression

What Is a (Good) Model?

- For correlated data, model predicts response given an input
- Model should be equation that fits data
- Standard definition of “fits” is *least-squares*
  - Minimize squared error
  - Keep mean error zero
  - Minimizes variance of errors

What Is a (Good) Model?

- For correlated data, model predicts response given an input
- Model should be equation that fits data
- Standard definition of “fits” is *least-squares*
  - Minimize squared error
  - Keep mean error zero
  - Minimizes variance of errors

Least-Squared Error

- If $\hat{y} = b_0 + b_1 x$ then error in estimate for x_i is $e_i = y_i - \hat{y}_i$
- Minimize Sum of Squared Errors (SSE)

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

- Subject to the constraint

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

Least-Squared Error

- If $\hat{y} = b_0 + b_1 x$ then error in estimate for x_i is $e_i = y_i - \hat{y}_i$
- Minimize Sum of Squared Errors (SSE)

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

- Subject to the constraint

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

Estimating Model Parameters

- Best regression parameters are

$$b_1 = \frac{\sum x_i y_i - n\overline{x}\,\overline{y}}{\sum x_i^2 - n\overline{x}^2} \qquad b_0 = \overline{y} - b_1\overline{x}$$

  where

$$\overline{x} = \frac{1}{n}\sum x_i \qquad \overline{y} = \frac{1}{n}\sum y_i$$

- Note that book may have errors in these equations!

Estimating Model Parameters
- Best regression parameters are

$$b_1 = \frac{\sum x_i y_i - n\overline{x}\,\overline{y}}{\sum x_i^2 - n\overline{x}^2} \qquad b_0 = \overline{y} - b_1\overline{x}$$

where

$$\overline{x} = \frac{1}{n}\sum x_i \qquad \overline{y} = \frac{1}{n}\sum y_i$$

- Note that book may have errors in these equations!

Parameter Estimation Example

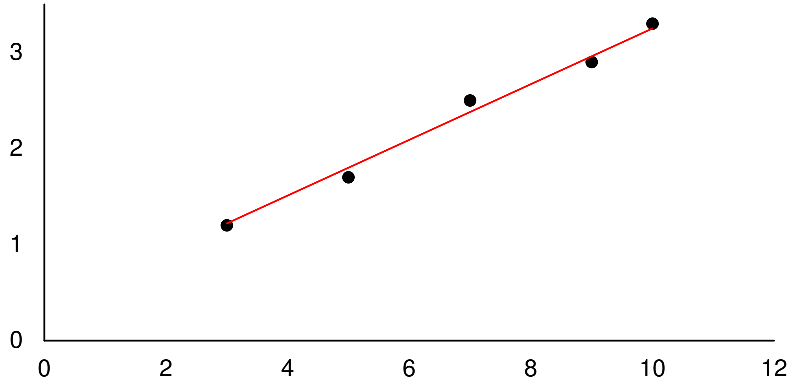
- Execution time of a script for various loop counts:

Loops	3	5	7	9	10
Time	1.2	1.7	2.5	2.9	3.3

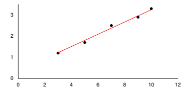
- $\overline{x} = 6.8$, $\overline{y} = 2.32$, $\sum xy = 88.54$, $\sum x^2 = 264$
- $b_1 = \dfrac{88.54 - 5(6.8)(2.32)}{264 - 5(6.8)^2} = 0.29$
- $b_0 = 2.32 - (0.29)(6.8) = 0.35$

Parameter Estimation Example
- Execution time of a script for various loop counts:
Loops 3 5 7 9 10
Time 1.2 1.7 2.5 2.9 3.3
- $\overline{x} = 6.8$, $\overline{y} = 2.32$, $\sum xy = 88.54$, $\sum x^2 = 264$
- $b_1 = \dfrac{88.54 - 5(6.8)(2.32)}{264 - 5(6.8)^2} = 0.29$
- $b_0 = 2.32 - (0.29)(6.8) = 0.35$

Graph of Parameter Estimation Example



Graph of Parameter Estimation Example



Allocating Variation

Analysis of Variation (ANOVA):

- If no regression, best guess of y is $\overline{y}$
- Observed values of y differ from $\overline{y}$, giving rise to errors (variance)
- Regression gives better guess, but there are still errors
- We can evaluate quality of regression by allocating sources of errors

Allocating Variation

Analysis of Variation (ANOVA):

- If no regression, best guess of y is $\overline{y}$
- Observed values of y differ from $\overline{y}$, giving rise to errors (variance)
- Regression gives better guess, but there are still errors
- We can evaluate quality of regression by allocating sources of errors

The Total Sum of Squares

Without regression, squared error is

$$
\begin{aligned}
\mathrm{SST} &= \sum_{i=1}^{n}(y_i-\overline{y})^2 = \sum_{i=1}^{n}(y_i^2 - 2y_i\overline{y} + \overline{y}^2) \\
&= \left(\sum_{i=1}^{n} y_i^2\right) - 2\overline{y}\left(\sum_{i=1}^{n} y_i\right) + n\overline{y}^2 \\
&= \left(\sum_{i=1}^{n} y_i^2\right) - 2\overline{y}(n\overline{y}) + n\overline{y}^2 \\
&= \left(\sum_{i=1}^{n} y_i^2\right) - n\overline{y}^2 \\
&= \mathrm{SSY} - \mathrm{SS0}
\end{aligned}
$$

The Total Sum of Squares

Without regression, squared error is

$$
\begin{aligned}
\mathrm{SST} &= \sum_{i=1}^{n}(y_i-\overline{y})^2 = \sum_{i=1}^{n}(y_i^2 - 2y_i\overline{y} + \overline{y}^2) \\
&= \left(\sum_{i=1}^{n} y_i^2\right) - 2\overline{y}\left(\sum_{i=1}^{n} y_i\right) + n\overline{y}^2 \\
&= \left(\sum_{i=1}^{n} y_i^2\right) - 2\overline{y}(n\overline{y}) + n\overline{y}^2 \\
&= \left(\sum_{i=1}^{n} y_i^2\right) - n\overline{y}^2 \\
&= \mathrm{SSY} - \mathrm{SS0}
\end{aligned}
$$

The Sum of Squares from Regression

- Recall that regression error is

$$\mathrm{SSE} = \sum e_i^2 = \sum (y_i - \overline{y})^2$$

- Error without regression is SST (previous slide)
- So regression explains $\mathrm{SSR} = \mathrm{SST} - \mathrm{SSE}$
- Regression quality measured by *coefficient of determination*

$$\mathrm{R}^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\mathrm{SST} - \mathrm{SSE}}{\mathrm{SST}}$$

The Sum of Squares from Regression
- Recall that regression error is
$$\mathrm{SSE} = \sum e_i^2 = \sum (y_i - \overline{y})^2$$
- Error without regression is SST (previous slide)
- So regression explains $\mathrm{SSR} = \mathrm{SST} - \mathrm{SSE}$
- Regression quality measured by *coefficient of determination*
$$\mathrm{R}^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\mathrm{SST} - \mathrm{SSE}}{\mathrm{SST}}$$

Evaluating Coefficient of Determination

- Compute SST $= (\sum y^2) - n\bar{y}^2$
- Compute SSE $= \sum y^2 - b_0 \sum y - b_1 \sum xy$
- Compute $\mathrm{R}^2 = \dfrac{\mathrm{SST} - \mathrm{SSE}}{\mathrm{SST}}$

Evaluating Coefficient of Determination
- Compute SST $= (\sum y^2) - n\bar{y}^2$
- Compute SSE $= \sum y^2 - b_0 \sum y - b_1 \sum xy$
- Compute $\mathrm{R}^2 = \dfrac{\mathrm{SST} - \mathrm{SSE}}{\mathrm{SST}}$

Example of Coefficient of Determination

For previous regression example:

Loops	3	5	7	9	10
Time	1.2	1.7	2.5	2.9	3.3

- $\sum y = 11.60$, $\sum y^2 = 29.79$, $\sum xy = 88.54$, $n\bar{y}^2 = 5(2.32)^2 = 26.9$
- $\text{SSE} = 29.79 - (0.35)(11.60) - (0.29)(88.54) = 0.05$
- $\text{SST} = 29.79 - 26.9 = 2.89$
- $\text{SSR} = 2.89 - 0.05 = 2.84$
- $\text{R}^2 = (2.89 - 0.05)/2.89 = 0.98$

Example of Coefficient of Determination
For previous regression example:
Loops 3 5 7 9 10
Time 1.2 1.7 2.5 2.9 3.3
$\sum y = 11.60$, $\sum y^2 = 29.79$, $\sum xy = 88.54$, $n\bar{y}^2 = 5(2.32)^2 = 26.9$
SSE = 29.79 − (0.35)(11.60) − (0.29)(88.54) = 0.05
SST = 29.79 − 26.9 = 2.89
SSR = 2.89 − 0.05 = 2.84
$\text{R}^2 = (2.89 - 0.05)/2.89 = 0.98$

Standard Deviation of Errors

- Variance of errors is SSE divided by degrees of freedom
  - DOF is $n - 2$ because we've calculated 2 regression parameters from the data
  - So variance (*mean squared error*, MSE) is $\mathrm{SSE}/(n-2)$
- Standard deviation of errors is square root: $s_e = \sqrt{\dfrac{\mathrm{SSE}}{n-2}}$ (minor error in book)

Standard Deviation of Errors
- Variance of errors is SSE divided by degrees of freedom
  - DOF is $n - 2$ because we've calculated 2 regression parameters from the data
  - So variance (*mean squared error*, MSE) is $\mathrm{SSE}/(n-2)$
- Standard deviation of errors is square root: $s_e = \sqrt{\dfrac{\mathrm{SSE}}{n-2}}$ (minor error in book)

Checking Degrees of Freedom

Degrees of freedom always equate:

- SS0 has 1 (computed from $\overline{y}$)
- SST has $n-1$ (computed from data and $\overline{y}$, which uses up 1)
- SSE has $n-2$ (needs 2 regression parameters)
- So
$$\begin{array}{ccccc} \text{SST} & = & \text{SSY} - \text{SS0} & = & \text{SSR} + \text{SSE} \\ n-1 & = & n \quad -1 & = & 1 \quad + (n-2) \end{array}$$

Checking Degrees of Freedom

Degrees of freedom always equate:

- SS0 has 1 (computed from $\overline{y}$)
- SST has $n-1$ (computed from data and $\overline{y}$, which uses up 1)
- SSE has $n-2$ (needs 2 regression parameters)
- So
$$\begin{array}{ccccc} \text{SST} & = & \text{SSY} - \text{SS0} & = & \text{SSR} + \text{SSE} \\ n-1 & = & n \quad -1 & = & 1 \quad + (n-2) \end{array}$$

Example of Standard Deviation of Errors

- For regression example, SSE was 0.05, so MSE is $0.05/3 = 0.017$ and $s_e = 0.13$
- Note high quality of our regression:
  - $R^2 = 0.98$
  - $s_e = 0.13$
  - Why such a nice straight-line fit?

Example of Standard Deviation of Errors
- For regression example, SSE was 0.05, so MSE is $0.05/3 = 0.017$ and $s_e = 0.13$
- Note high quality of our regression:
  - $R^2 = 0.98$
  - $s_e = 0.13$
  - Why such a nice straight-line fit?

Confidence Intervals for Regressions

- Regression is done from a single population sample (size n)
  - Different sample might give different results
  - True model is $y = \beta_0 + \beta_1 x$
  - Parameters b_0 and b_1 are really means taken from a population sample

Confidence Intervals for Regressions

- Regression is done from a single population sample (size n)
  - Different sample might give different results
  - True model is $y = \beta_0 + \beta_1 x$
  - Parameters b_0 and b_1 are really means taken from a population sample

Calculating Intervals for Regression Parameters

- Standard deviations of parameters:

$$
\begin{aligned}
s_{b_0} &= s_e \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum x^2 - n\overline{x}^2}} \\
s_{b_1} &= \frac{s_e}{\sqrt{\sum x^2 - n\overline{x}^2}}
\end{aligned}
$$

- Confidence intervals are $b_i \mp t_{[1-\frac{\alpha}{2};n-2]} s_{b_i}$
- Note that t has $n-2$ degrees of freedom!

2015-06-15

CS147
└ Confidence Intervals for Regressions
　└ Parameter Intervals
　　└ Calculating Intervals for Regression Parameters

Calculating Intervals for Regression Parameters

- Standard deviations of parameters:

$$
\begin{aligned}
s_{b_0} &= s_e \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum x^2 - n\overline{x}^2}} \\
s_{b_1} &= \frac{s_e}{\sqrt{\sum x^2 - n\overline{x}^2}}
\end{aligned}
$$

- Confidence intervals are $b_i \mp t_{[1-\frac{\alpha}{2};n-2]} s_{b_i}$
- Note that t has $n-2$ degrees of freedom!

Example of Parameter Confidence Intervals

- Recall $s_e = 0.13$, $n = 5$, $\sum x^2 = 264$, $\overline{x} = 6.8$
- So
$$s_{b_0} = 0.13\sqrt{\frac{1}{5} + \frac{(6.8)^2}{264-5(6.8)^2}} = 0.16$$
$$s_{b_1} = \frac{0.13}{\sqrt{264-5(6.8)^2}} = 0.004$$
- Using 90% confidence level, $t_{0.95;3} = 2.353$
- Thus, $b0$ interval is $0.35 \mp 2.353(0.16) = (-0.03, 0.73)$
  - Not significant at 90%
- And b_1 is $0.29 \mp 2.353(0.004) = (0.28, 0.30)$
  - Significant at 90% (and would survive even 99.9% test)

2015-06-15

CS147
└Confidence Intervals for Regressions
└Parameter Intervals
└Example of Parameter Confidence Intervals

Example of Parameter Confidence Intervals
- Recall $s_e = 0.13$, $n = 5$, $\sum x^2 = 264$, $\overline{x} = 6.8$
- So $s_{b_0} = 0.13\sqrt{\frac{1}{5} + \frac{(6.8)^2}{264-5(6.8)^2}} = 0.16$, $s_{b_1} = \frac{0.13}{\sqrt{264-5(6.8)^2}} = 0.004$
- Using 90% confidence level, $t_{0.95;3} = 2.353$
- Thus, $b0$ interval is $0.35 \mp 2.353(0.16) = (-0.03, 0.73)$
  - Not significant at 90%
- And b_1 is $0.29 \mp 2.353(0.004) = (0.28, 0.30)$
  - Significant at 90% (and would survive even 99.9% test)

Confidence Intervals for Predictions

- Previous confidence intervals are for *parameters*
  - How certain can we be that the parameters are correct?
- Purpose of regression is *prediction*
  - How accurate are the predictions?
  - Regression gives mean of predicted response, based on sample we took

Confidence Intervals for Predictions

- Previous confidence intervals are for *parameters*
  - How certain can we be that the parameters are correct?
- Purpose of regression is *prediction*
  - How accurate are the predictions?
  - Regression gives mean of predicted response, based on sample we took

Predicting *m* Samples

- Standard deviation for *mean* of future sample of m observations at x_p is

$$s_{\hat{y}_{mp}} = s_e\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{\sum x^2 - n\overline{x}^2}}$$

- Note deviation drops as $m \to \infty$
- Variance minimal at $x = \overline{x}$
- Use t-quantiles with $n - 2$ DOF for calculating confidence interval

Predicting *m* Samples

- Standard deviation for *mean* of future sample of m observations at x_p is

$$s_{\hat{y}_{mp}} = s_e\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{\sum x^2 - n\overline{x}^2}}$$

- Note deviation drops as $m \to \infty$
- Variance minimal at $x = \overline{x}$
- Use t-quantiles with $n - 2$ DOF for calculating confidence interval

Example of Confidence of Predictions

- Using previous equation, what is predicted time for a single run of 8 loops?
- Time $= 0.35 + 0.29(8) = 2.67$
- Standard deviation of errors $s_e = 0.13$

$$s_{\hat{y}_{1,8}} = 0.13\sqrt{1 + \frac{1}{5} + \frac{(8-6.8)^2}{264 - 5(6.8)^2}} = 0.14$$

- 90% interval is then $2.65 \mp 2.353(0.14) = (2.34, 3.00)$

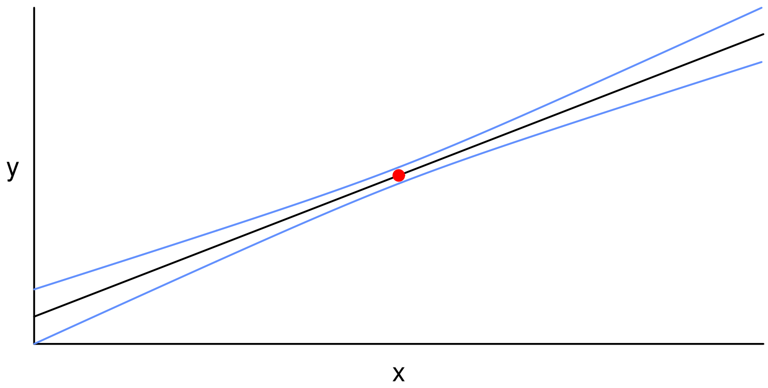
Example of Confidence of Predictions

- Using previous equation, what is predicted time for a single run of 8 loops?
- Time $= 0.35 + 0.29(8) = 2.67$
- Standard deviation of errors $s_e = 0.13$

$$s_{\hat{y}_{1,8}} = 0.13\sqrt{1 + \frac{1}{5} + \frac{(8-6.8)^2}{264 - 5(6.8)^2}} = 0.14$$

- 90% interval is then $2.65 \mp 2.353(0.14) = (2.34, 3.00)$

Prediction Confidence



Verifying Assumptions Visually

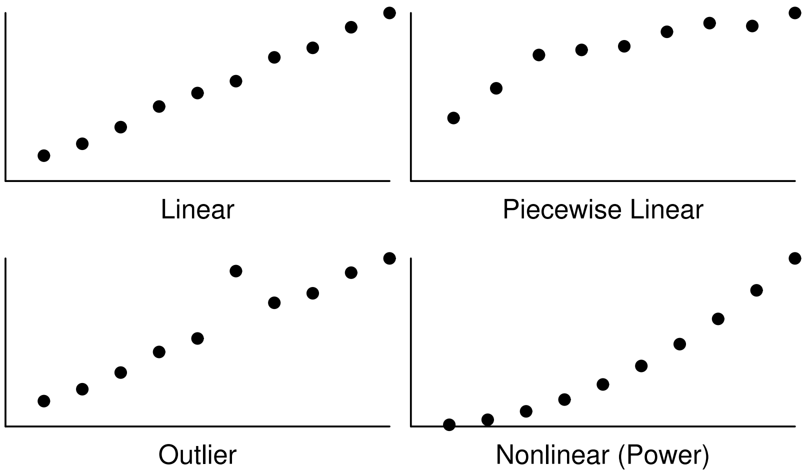
- Regressions are based on assumptions:
  - Linear relationship between response y and predictor x
    - Or nonlinear relationship used in fitting
  - Predictor x nonstochastic and error-free
  - Model errors statistically independent
    - With distribution $N(0, c)$ for constant c
- If assumptions violated, model misleading or invalid

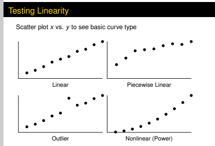
Verifying Assumptions Visually

- Regressions are based on assumptions:
  - Linear relationship between response y and predictor x
    - Or nonlinear relationship used in fitting
  - Predictor x nonstochastic and error-free
  - Model errors statistically independent
    - With distribution $N(0, c)$ for constant c
- If assumptions violated, model misleading or invalid

Testing Linearity

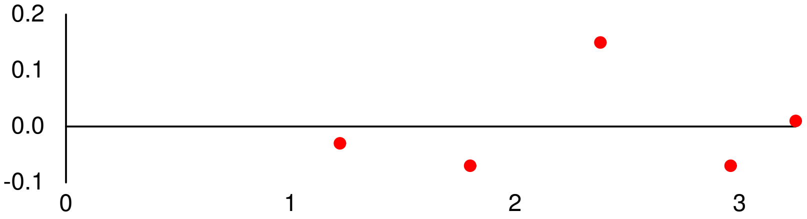
Scatter plot x vs. y to see basic curve type





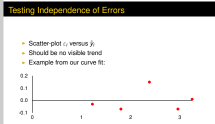
Testing Independence of Errors

- Scatter-plot ε_i versus $\hat{y}_i$
- Should be no visible trend
- Example from our curve fit:



2015-06-15

CS147
└Verifying Regression
└Testing Independence of Errors



More on Testing Independence

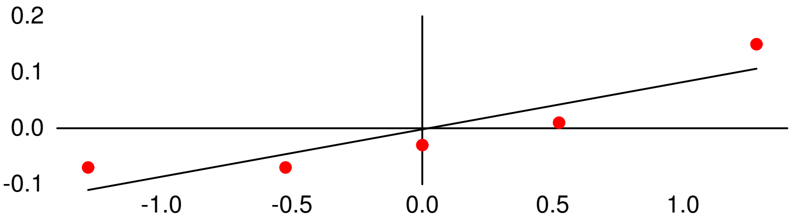
- May be useful to plot error residuals versus experiment number
  - In previous example, this gives same plot except for x scaling
- No foolproof tests
  - “Independence” test really *disproves* particular **de**pendence
  - Maybe next test will show different dependence!

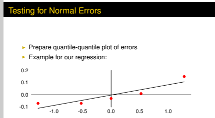
More on Testing Independence

- May be useful to plot error residuals versus experiment number
  - In previous example, this gives same plot except for x scaling
- No foolproof tests
  - “Independence” test really *disproves* particular **de**pendence
  - Maybe next test will show different dependence!

Testing for Normal Errors

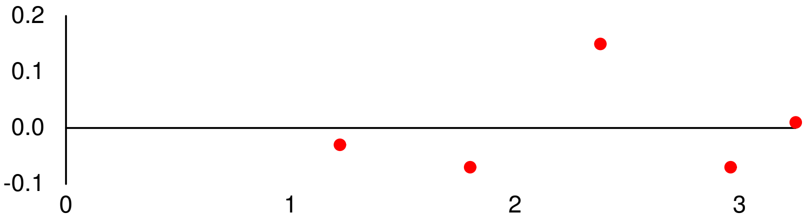
- Prepare quantile-quantile plot of errors
- Example for our regression:





Testing for Constant Standard Deviation

- Tongue-twister: *homoscedasticity*
- Return to independence plot
- Look for trend in spread
- Example:



Testing for Constant Standard Deviation
- Tongue-twister: *homoscedasticity*
- Return to independence plot
- Look for trend in spread
- Example:

Linear Regression Can Be Misleading

- Regression throws away some information about the data
  - To allow more compact summarization
- Sometimes vital characteristics are thrown away
  - Often, looking at data plots can tell you whether you will have a problem

Linear Regression Can Be Misleading

- Regression throws away some information about the data
  - To allow more compact summarization
- Sometimes vital characteristics are thrown away
  - Often, looking at data plots can tell you whether you will have a problem

Example of Misleading Regression

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Example of Misleading Regression

What Does Regression Tell Us?

- **Exactly the same thing for each data set!**
- $n = 11$
- Mean of $y = 7.5$
- $y = 3 + 0.5x$
- Standard error of regression is 0.118
- All the sums of squares are the same
- Correlation coefficient = 0.82
- $R^2 = 0.67$

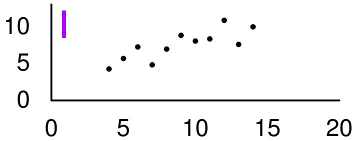
What Does Regression Tell Us?

- **Exactly the same thing for each data set!**
- $n = 11$
- Mean of $y = 7.5$
- $y = 3 + 0.5x$
- Standard error of regression is 0.118
- All the sums of squares are the same
- Correlation coefficient = 0.82
- $R^2 = 0.67$

Now Look at the Data Plots

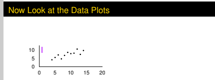
Now Look at the Data Plots

Now Look at the Data Plots

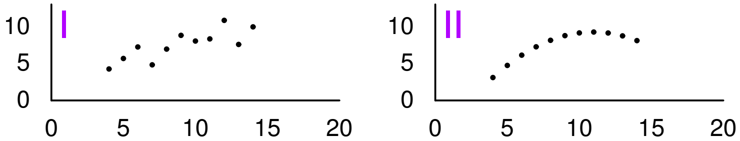


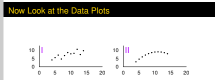
CS147
Verifying Regression
Now Look at the Data Plots

2015-06-15

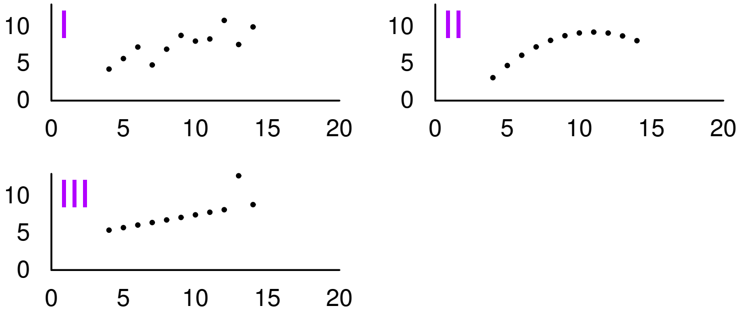


Now Look at the Data Plots



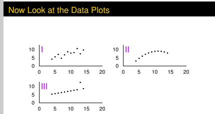


Now Look at the Data Plots



2015-06-15

CS147
└Verifying Regression
└Now Look at the Data Plots



Now Look at the Data Plots