

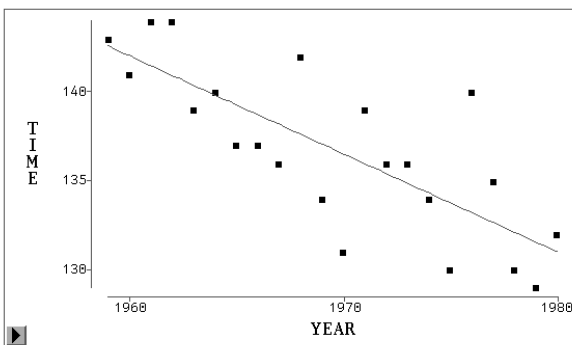
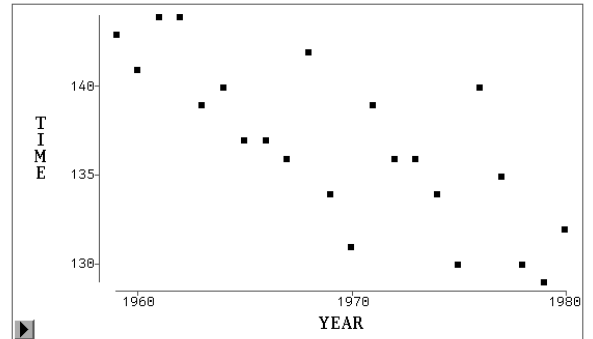
**STAT:2010**  
**Statistical Methods and**  
**Computing**

**Linear Regression, continued**

Lecture 6  
 Feb. 1-3, 2016

Kate Cowles  
 374 SH, 335-0727  
 kate-cowles@uiowa.edu

**Another example: Men's winning**  
**times in the Boston Marathon, 1959-**  
**80**



$$\hat{y} = 1221.05 - 0.5505x$$

What does this equation tell us?

**Prediction using an estimated regres-**  
**sion line**

Example: What is the predicted PCH for a country with PCGDP = \$20,000?

$$\begin{aligned}\hat{y} &= -465.7 + 0.0968(20000) \\ &= 2401.70\end{aligned}$$

What is the predicted winning time for the Boston marathon in 1965?

## How well does the regression line predict the response variable?

- The **coefficient of determination** or  $R^2$ 
  - When there is only 1 explanatory variable,  $R^2 = r^2$  — the square of the correlation coefficient  $r$  between the response variable and the explanatory variable
  - the proportion of the variability among the observed values of the response variable that is explained by the linear regression
- Example: in the OECD health care expenditures data,  $R^2 = 0.764$ 
  - 76.4% of the variability in per capita health care expenditures is explained by PCGDP

## Residuals

- A **residual** is the difference between an observed value and a predicted value of the response variable.

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

- There is a residual for each data point.
- The residual for the  $i$ th observation will be positive if the observed value lies above the estimated regression line.
- The mean of the residuals from a least-squares fit is always 0

## Notation

Recall:

- $y_i$  is the observed value of the response variable for subject  $i$
- $\hat{y}_i$  is the value predicted by the regression line for subject  $i$

$$\hat{y}_i = a + bx_i$$

## Example: the OECD health care expenditures data

$$\hat{y} = -465.7 + 0.0968x$$

- The predicted score for the US, for which  $x_1 = 30,514$  dollars is

$$\hat{y}_1 = -465.7 + 0.0968(30514) = 2488$$

- The actual value of PCH for the US is \$3898.
- The residual for the US is positive because the data point lies above the regression line.

$$r_i = 3898 - 2488 = 1410$$

## Residual plots

- A residual plot is a scatterplot of the regression residuals against the predicted values of the response variable.
- Residual plots help
  - assess fit of a regression line
  - look for violations of the assumptions of linear regression and for problematic data points

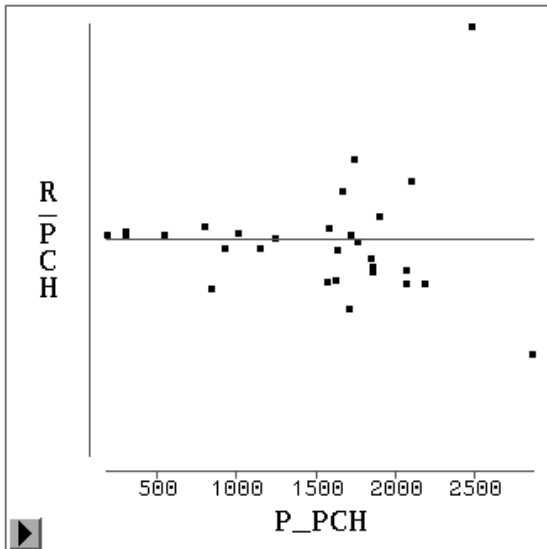
- individual points with large residuals
  - outliers in the vertical direction
  - these points are not well described by the regression equation
- individual points that are extreme in the horizontal direction (unusual values of explanatory variable)
  - These may be influential observations.

## Things to watch for in a residual plot

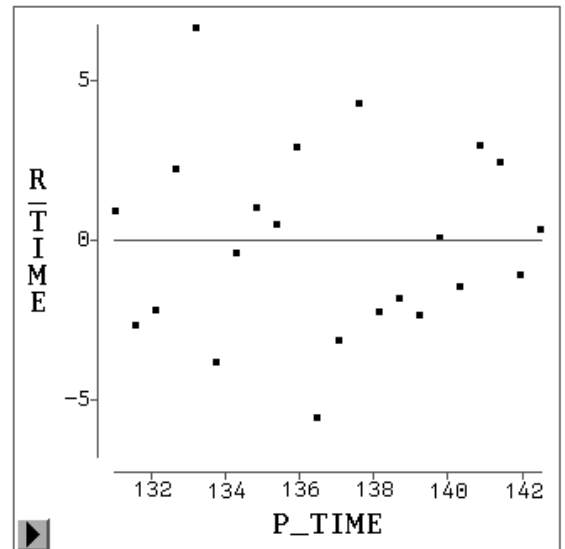
- a random scatter of points
  - This is what you *want* to see.
- A curved pattern
  - indicates that the relationship between the response variable and the explanatory variable is *not* linear
  - violation of an assumption
- increasing or decreasing spread around the zero line
  - indicates violation of the assumption that  $\sigma$  is the same in all the subpopulations

## Idealized patterns in residual plots

### The residual plot for the OECD health care expenditures data



### The residual plot for the Boston marathon data



### Outliers and influential observations

- Outlier: an observation that lies outside the overall pattern of the other observations.
- Influential observation: an observation is *influential* for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the  $x$  direction (have unusual values of the explanatory variable) are often influential in computing the least-squares regression line.

- In the OECD data, the US and Luxemburg are both outliers.
- The US is influential.
  - With the US included in the analysis,  $R^2 = 0.764$ . If the US is deleted,  $R^2$  increases to 0.846.

## Facts about least-squares regression

- Keep straight which is the response variable and which is the explanatory variable. If they are switched, a different regression line results.
- The correlation coefficient  $r$  and the slope  $b$  of the regression line are closely related.
  - They always have the same sign (both positive, or both negative, or both zero).
  - The slope of the regression line is

$$b = r \frac{s_y}{s_x}$$

This means that a change of one standard deviation in  $x$  corresponds to a change of  $r$  standard deviations in  $y$ .

- How large or small the slope is does *not* indicate how strong the relationship between the response variable and the explanatory variable is.

- The least-squares regression line always passes through the point  $(\bar{x}, \bar{y})$ .
- The square of the correlation coefficient,  $r^2$ , is the fraction of the variation in the values of  $y$  that is explained by the least-squares regression line.
  - How much better are we able to predict  $y$  because we know  $x$ ?

– The magnitude of the slope depends on the units in which we measure both variables.

\* Example: If we measured the winning times in the Boston marathon in hours instead of minutes, the slope would be -0.0092 instead of -0.55, but the relationship between winning time and year of race would be the same!

– The correlation coefficient  $r$  is needed to quantify the strength of the relationship.

- But the correlation coefficient is not enough to enable us to *predict* the value of a response variable if we know the value of an explanatory variable.
  - For prediction, we need the regression equation.

## Caveats about regression and correlation

- It usually doesn't make sense to try to use the regression equation to predict for values of the explanatory variable outside the range of observed data.
- Correlations based on averaged data are usually too large to be applicable to individuals.
  - Example: Correlation between national female literacy rates and national infant mortality rates in countries in Latin America
- Lurking variables
  - one or more variables that have an important effect on the relationship among variables under study but that are not considered in the study