Chapter 12: Linear Regression

November 30, 2009

12.1 Introduction

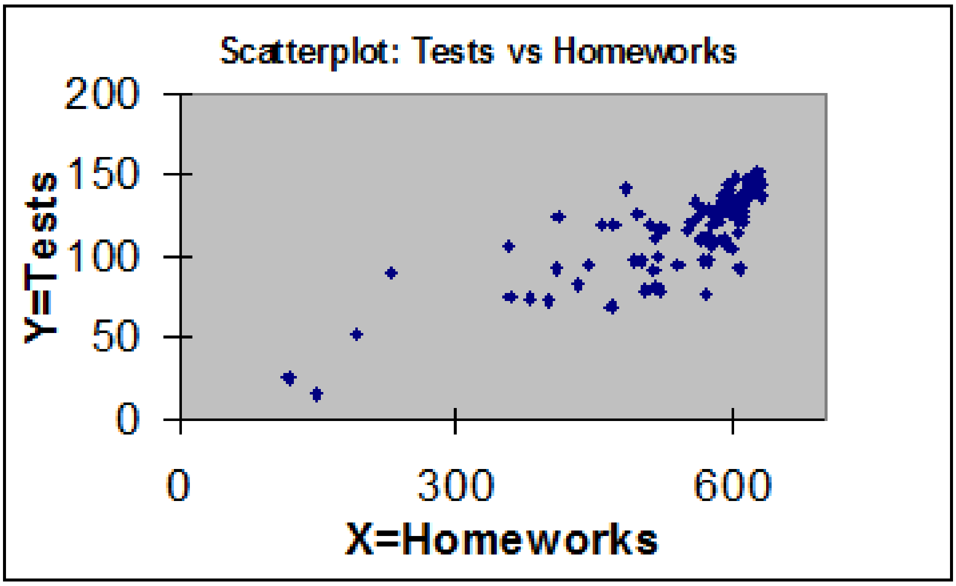
In linear regression, to explain values of a continuous *response variable* Y we use a *continuous explanatory variable* X.

We will have pairs of observations of two numerical variables (X, Y):
$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Examples:

- X = concentration, Y = rate of reaction,
- X = weight, Y = height,
- X = total Homework score to date, Y = total score on Tests to date.

They are represented by points on the *scatterplot*.



Two Contexts

1. Y is an observed variable and the values of X are specified by the experimenter.
2. Both X and Y are observed variables.

If the experimenter *controls* one variable, it is usually labeled X and called the explanatory variable.

The response variable is the Y.

When X and Y are both only *observed*, the distinction between explanatory and response variables is somewhat arbitrary, but must be made as their roles are different in what follows.

12.2 The Fitted Regression Line

Equation for the Fitted Regression Line

This is the "closest" line to the points of the scatterplot.

We consider Y a linear function of X plus a random error.

We will first need some notation to describe the *influence* of X on Y:

- The following are as usual:

$$SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad SS_y = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$s_x = \sqrt{\frac{SS_x}{n-1}} \qquad s_y = \sqrt{\frac{SS_y}{n-1}}$$

- One new quantity is the sum of products:

$$\mathrm{SP}_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.$$

- We consider a linear model: $Y = \beta_0 + \beta_1 X +$ random error.
- β_0 is called the intercept and β_1 is called the slope.

We only have a sample so we will *estimate* β_0 and β_1:

- We estimate β_1 by

$$b_1 = \frac{SP_{xy}}{SS_x}$$

- We estimate β_0 by

$$b_0 = \bar{y} - b_1\bar{x}.$$

The line $y = b_0 + b_1 x$ is the "best" straight line though the data. It is also known as the "least-squares line". (Explanations will be given later.)

We will call it the fitted regression line.

Example: Let X be the total score on our Homeworks to date (in points) and Y be the total score on Tests (in points). The following summary statistics were obtained:

$$n = 99$$

$$\bar{x} = 546.76 \qquad \bar{y} = 117.07$$

$$SS_x = 990098.2 \qquad SS_y = 62442.5$$

$$SP_{xy} = 199201.7$$

We obtain

$$b_1 = SP_{xy}/SS_x = 1999201.7/990098.2 =$$

and

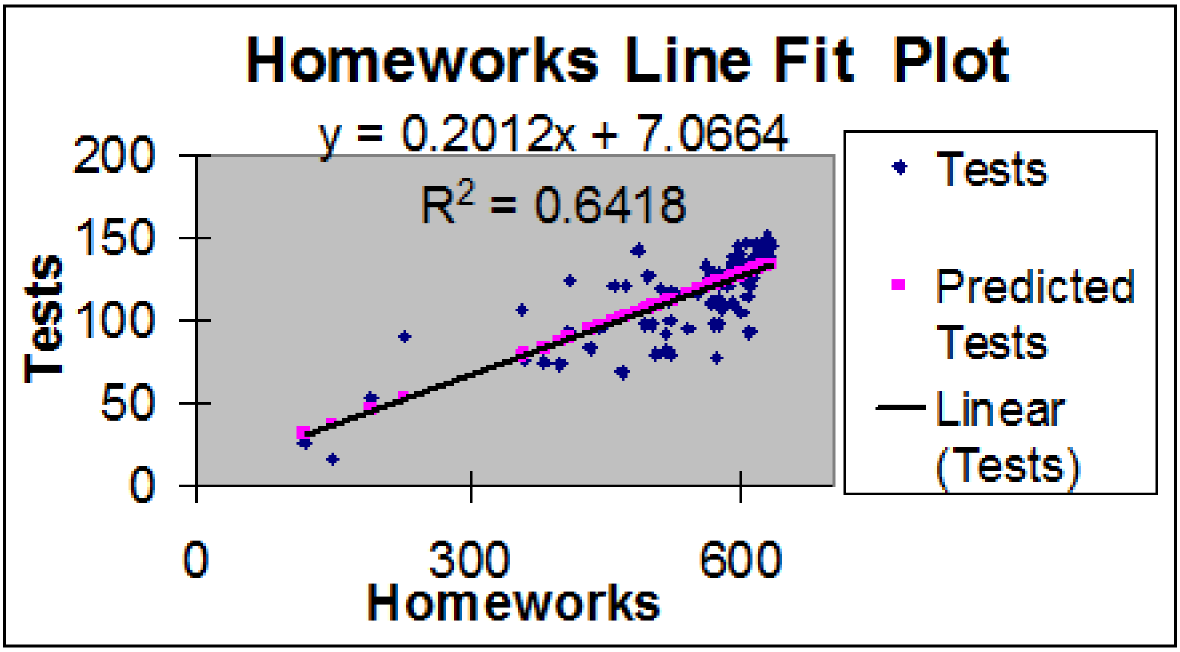
$$b_0 = \bar{y} - b_1 \bar{x} =$$

Note: use many significant digits of b_1 to calculate b_0.

The fitted regression line is

$$\text{Tests} = 7.065 + 0.2012 * \text{Homeworks}.$$

Here the plot for our data with "predicteds" and regression line:



[Discussion]How do we interpret slope and intercept in linear equations? Consider, e.g., $F = 32 + 1.88C$, and the above equation..

[This is related to Problem 12.5 (c) on your last homework.]

Predicteds and Residual Sum of Squares

For each value of x_i in the sample there is a value of y predicted by the fitted regression line.

- We denote it $\hat{y}_i = b_0 + b_1 x_i$.
- For example: for Homeworks=546.76 (the average total score on homeworks), the predicted Tests are... . Comment on this!
- Predicted $\hat{y}_i$ is usually not the same as the observed y for that x (i.e. y_i for x_i).
- The difference between the observed and predicted value is called the residual:
  - residual $= y_i - \hat{y}_i$.
- For example, one person had a score of 609 on Homeworks, and the the person accumulated 133 on Tests. Calculate the residual. .

The Residual Sum of Squares is defined as

$$\mathrm{SS(resid)} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- It can be calculated more easily by

$$\mathrm{SS(resid)} = SS_y - SP_{xy}^2 / SS_x$$

We can now be more specific: the fitted regression line is (by definition) the line, which minimizes SS(resid) among all possible straight lines. Hence "the best".

For our data we have

$$\begin{aligned}\mathrm{SS(resid)} &= 62442.51 - (199201.7)^2/990098.2 \\ &= \end{aligned}$$

Residual Standard Deviation

The residual standard deviation is defined as

$$s_{Y|X} = \sqrt{\frac{\mathrm{SS(resid)}}{n-2}}$$

- This quantity describes the variability of the residuals, or, which is the same, the (vertical) variability of $y_i's$ around the regression line. Similarly, we use s_Y to describe the variability of $y_i's$ around $\bar{y}$.
- For "nice" data sets, we expect roughly 68% of the observed y's to be within $\pm s_{Y|X}$ of the regression line and roughly 95% of the observed y to be within $\pm 2s_{Y|X}$ of the regression line.

For our data this is

$$s_{Y|X} = \sqrt{\frac{22364.3}{99-2}} =$$

[Interpretation] What is the SD of Tests among students who scored about 600?
[Compare problem 12.30.]

12.3 Parametric Interpretation of Regression

The linear model is

$$Y = \beta_0 + \beta_1 X + \text{random error}$$

What is this random error?

- For a fixed (given) value of X, we will think of Y as a random variable with mean $\mu_{Y|X} = \beta_0 + \beta_1 X$ and standard deviation denoted $s_{Y|X}$. Here $Y|X$ is to express the dependence on values of X.
- We also assume normality: $Y \sim N(\mu_{Y|X}, \sigma_{Y|X})$.
- We can write this as
$$Y = \beta_0 + \beta_1 X + N(0, \sigma_{Y|X}).$$

In most of what follows we make the assumptions:

1. $\sigma_{Y|X}$ is the same for all values of X.
1. The random errors are independent normal random variables with mean 0 and SD $\sigma_{Y|X}$.
   - ▶ $\sigma_{Y|X}$ is estimated by our $s_{Y|X}$.

<u>Estimation in the Linear Model</u>

The Random Sub-sampling model: For each observed pair (x, y), we regard the value of y as having been sampled at random from the conditional population of Y values associated with the $X = x$.

[Picture]

12.4 Statistical Inference Concerning β_1

The standard error of b_1 is given by

$$\text{SE}_{b_1} = \frac{s_{Y|X}}{\sqrt{SS_x}}.$$

- Note that the standard error gets smaller as:
  - $s_{Y|X}$ gets small (observations close to line)
  - sample gets larger (SS_x gets bigger)
  - the x's are more spread out (SS_x gets bigger).

For our data we find that

$$\text{SE}_{b_1} = 15.18/\sqrt{990098.2} =$$

Confidence intervals for β_1

These are constructed in the usual way:

$$b_1 \pm t(n-2)_{\alpha/2}\text{SE}_{b_1}.$$

- Note that the degrees of freedom are $df = n - 2$.
- For our midterm data $t(\mathbf{97})_{0.025} = 1.985$ (from Excel) so that the 95% C.I. for β_1 is

  $$0.2012 \pm 0.01526 * 1.985 = (0.1709, 0.2315)\,.$$

  Interpretation?

[Compare problem 12.22 (a).]

Hypothesis Tests about β_1

We can also do a t-test with b_1.

We will assume that the linear model is true: $Y = \beta_0 + \beta_1 X + N(0, \sigma)$. We want to see if there is evidence that $\beta_1 \neq 0$. This may be stated as "X having (nonzero) effect on Y within the linear model".

Details:

[change (X) and (Y) to variable names]

Does Y influence X within the linear model?

Let β_1 be the slope of the linear regression of (Y) on (X).

$H_0 : \beta_1 = 0$; there is zero linear influence of (Y) on (X).

$H_A : \beta_1 \neq 0$; there is a non-zero influence of (Y) on (X).

[H_A could be directional]

Use a non-directional t-test. $t_s = b_1/SE_{b_1}$ has a t-distribution with $n-2$ degrees of freedom under H_0.

Critical value is $t(n-2)_{\alpha/2}$. Reject H_0 if $|t_s| > t(n-2)_{\alpha/2}$.
[Compare 12.22 (b).]

Tests vs Homeworks example:

Is there a nonzero linear influence of Homeworks on Tests' score?

Let β_1 be the slope of the linear regression of Tests on Homeworks.

$H_0 : \beta_1 = 0$; there is zero linear linear influence of Homeworks on Tests.

$H_A : \beta_1 \neq 0$; there is a nonzero linear influence of Homeworks on Tests.

Use a non-directional t-test. $t_s = b_1/SE_{b_1}$ has a t-distribution with $n-2 = 97$ degrees of freedom under H_0.

Test at $\alpha = 0.05$. Critical value is $t(97)_{0.025} = 1.985$. Reject H_0 is $|t_s| > 1.985$.

$t_s =$ so

These data provide evidence at the 0.05 significance level that there is a (positive) linear influence of Homeworks' perpormance on Tests' results.
[Discussion] Is this useful to predict, summarize, model?

12.5 The Correlation Coefficient

Definition:

$$r = \frac{SP_{xy}}{\sqrt{SS_x\, SS_y}}.$$

It measures the strength and direction of the linear association between Y and X. In our example:

$$r = 199201.7/\sqrt{990098.2 * 62442.5} =$$

(Positive, moderately strong correlation between Homeworks and Tests.)
It is worthwhile to consider r^2 and its relationship to variability of the response variable Y.

The Coefficient of Determination, r^2

The quantity **SS(total)** $= SS_y$ measures the total variability in the y's.

The difference between SS(total) and SS(resid) is called the **sum of squares regression** or **SS(reg)**. It measures the variability in $y_i's$ which is due to the regression model (variability of $\hat{y}_i's$):

$$\text{SS(reg)} = \sum(\hat{y}_i - \bar{y})^2.$$

These sums of squares are related in the following (Pythagorean) way:

$$\mathrm{SS(total)} = \mathrm{SS(reg)} + \mathrm{SS(resid)}.$$

This makes it easy to calculate from the previous quantities.

The **coefficient of determination** is defined by

$$r^2 = \frac{\mathrm{SS(reg)}}{\mathrm{SS(total)}}.$$

It can be interpreted as the fraction (in quadratic terms) of total variation in Y that is "accounted for" or "explained" by the regression.

From the relationship of the sums of squares above, we also have

$$r^2 = 1 - \frac{\mathrm{SS(resid)}}{\mathrm{SS(total)}}.$$

<u>Our data:</u>

We have $SS(total) = \mathbf{SS_y} = 62442.5$.

Therefore

$$\mathrm{SS(reg)} = 62442.5 - 22364.3 =$$

Therefore,

$$r^2 = \frac{40078.2}{62442.5} =$$

[Compare with $r = 0.80115$.]

Therefore, 64.2% of the variation in Tests' scores is explained by the regression on Homeworks' scores. [Interpretations.]

[Compare Problem 12.28 (b) and 12.30.]

Comments:

- $0 \leq r^2 \leq 1$.
- $r^2 = 1$ if and only if all of the sample data points lie on a line.
- if $r^2 = 0$ then 0% of the variation in Y is explained by variation in X (and $t_s = 0$ also).

<u>Comments on Correlation Coefficient</u>

The **correlation coefficient**, r, is the square root of r^2 multiplied by the sign of b_1.

It is related to b_1 as follows:

$$b_1 = r\frac{s_Y}{s_X} = r\sqrt{SS_y/SS_x}.$$

This is sometimes used to calculate b_1. [Verify in our example.]

- $-1 \leq r \leq 1$.
- if $r = \pm 1$ then all of the data lie on a line.

[some pictures; see also page 556 in textbook]

Inference about the Correlation

Bivariate Random Sampling Model: Each pair (x_i, y_i) can be regarded as having been sampled from a population of (x, y).

In the bivariate random sampling model, the sample correlation coefficient r estimates the population correlation coefficient ρ (rho).

Due to the relationships

$$b_1 = r\frac{s_y}{s_x}; \qquad \beta_1 = \rho\frac{\sigma_Y}{\sigma_X}$$

testing $H_0 : \rho = 0$ is the same as testing $H_0 : \beta_1 = 0$.

We also have that

$$t_s = \frac{b_1}{SE_{b_1}} = r\frac{s_Y}{s_X}\frac{\sqrt{SS_x}}{s_{Y|X}} = r\sqrt{\frac{n-2}{1-r^2}}\,.$$

Thus, rather than testing "for linear influence" we may, and will gladly, perform tests for nonzero correlation. This is a simpler calculation and interpretation. [Do this below for our example.]

[Compare Problem 12.33.]

12.6 Guidelines

Like any statistical procedure, there are a number of potential dangers when using linear regression. We will discuss a few here.

Least-squares regression will fit a straight line through *any* set of data, even if the linear pattern is inappropriate (e.g. curvilinearity).

- A scatter plot of your data is a simple way to visually assess if your data have a nonzero linear trend/correlation.
- After you fit a regression, a plot of the fitted values ($\hat{y}_i$'s) vs. the residuals can reveal problems as well — is their a pattern?
- A normal probability plot of the residuals can reveal problems about the normality assumptions.

[Residual plot etc. for our example.]

Example:

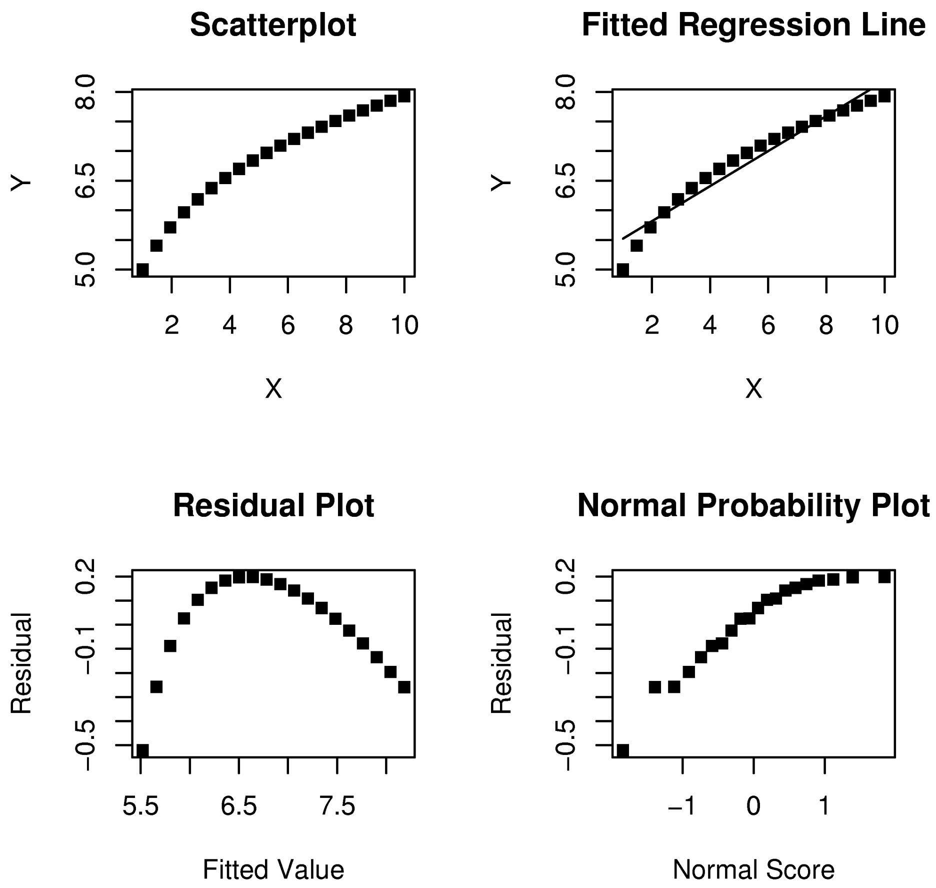
Here is an example where relation between X and Y is very clear but certainly not linear.

- For this dataset:
  - $r^2 = 0.9476$, $r = 0.973$
  - The p-value for testing $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ is less than 0.000000001.
  - The fitted regression line is

$$Y = 5.22 + 0.2959X.$$

Linear model may obscure the real nature of the data (but may also be used as the first approximation).

- The Plots...



Outliers

An **outlier** is a point that is unusually far from the fitted regression line (that is, has an unusually high residual).

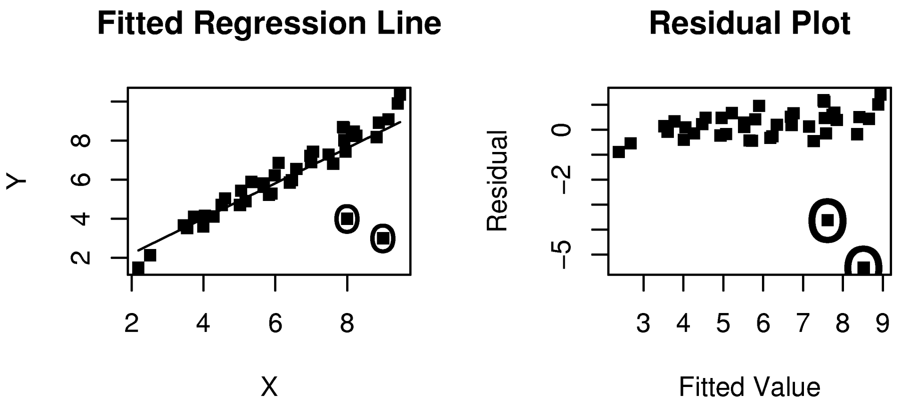
Outliers can distort regression analysis in 2 ways:

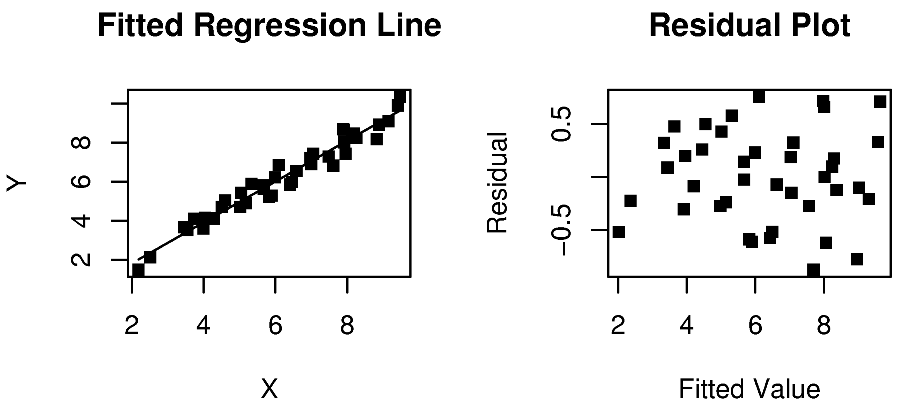
1. They inflate $s_{Y|X}$ and reduce r.
2. They can unduly influence the regression line.

Example:

1. In the following example, the first two plots show the fitted regression and residual plot in the presence of 2 outliers (circled).
   - Fitted regression line is $Y = 0.45 + 0.89X$.
   - $r^2 = 0.7784$.
   - The p-value for testing $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ is less than 0.00001 (i.e. very small).

2. In the second example, the two outliers were removed before fitting the line.
   - Fitted regression line is $Y = 0.06 + 0.98X$.
   - $r^2 = 0.9674$.
   - The p-value for testing $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ is less than 0.00001 (i.e. very small).





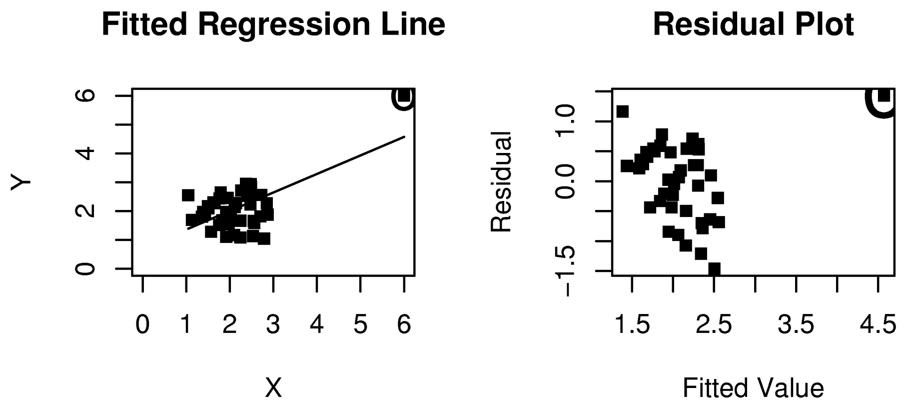
Influential Points

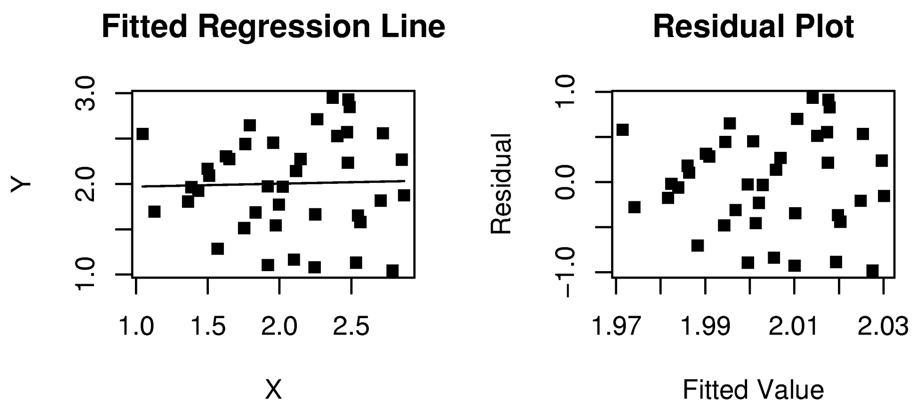
An **influential point** is a point whose presence changes very much the outcome of regression.

- A point which is far the majority of the data in the x direction may have a large effect on the regression analysis.

Example:

- In this example there is a point at $(6,6)$ which is very influential for the linear regression.
  1. In the first two plots, the influential point at $(6,6)$ leads to the following regression:
     - $Y = .71 + 0.32X$
     - $r^2 = 0.376$
     - The p-value for testing $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ is less than 0.0002 (i.e. very small).
  2. After removal, the regression line is
     - $Y = 1.94 + 0.032X$
     - $r^2 = 0.00085$
     - The p-value for testing $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ is greater than 0.5.





What should you do with these points?

- An *arbitrary* removal of data points is not recommended.
- You need to figure out the nature of each unusual observation:
  - Was it recorded incorrectly?
  - Does it belong to the population we want to study?
- Statistical software has regression diagnostics that can help identify outliers, influential points and other problems.

Dangers of Extrapolation

- While your data may provide evidence of a linear relationship between Y and X, this relationship may not hold outside the range of X values actually observed.