Planeamiento Avanzado en Cadenas de Aprovisionamiento

centrum.pucp.edu.pe

centrum PUCP ESCUELA PARA LOS BUENOS NEGOCIOS

UNIDAD 2: SUPPLY CHAIN FORECASTING

Sesión 4
Regresión Lineal

Market Fulfillment Centers (MFC)

Delivery-as-a-Service

Walmart GoLocal

Forecasting

In general, we have observations

(x_i, y_i) ← **the ith observation is a pair of numbers**

Our data looks like:

x	y	i
12.0	192	1
12.0	160	2
5.0	155	3
5.0	120	4
7.0	150	5
13.0	175	6
4.0	100	7
12.0	165	8

The plot enables us to see the relationship between x and y.

Covariance

Consider two variables, X and Y.

The concept of covariance asks:

Is Y larger (or smaller) when X is larger ?

We measure this using something called covariance s_{xy}

Covariance > 0	Larger X	⟺	Larger Y
Covariance < 0	Larger X	⟺	Smaller Y

Planeamiento Avanzado de Cadenas de Aprovisionamiento

Here is the actual formula but most people never calculate covariance by hand………

The sample covariance between x and y is:

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

What are the units of covariance ?

In this example, we look at the relationship between team payroll and team performance in Major League Baseball using data from the 2010 season (for a total of 30 teams).

The variables of interest:

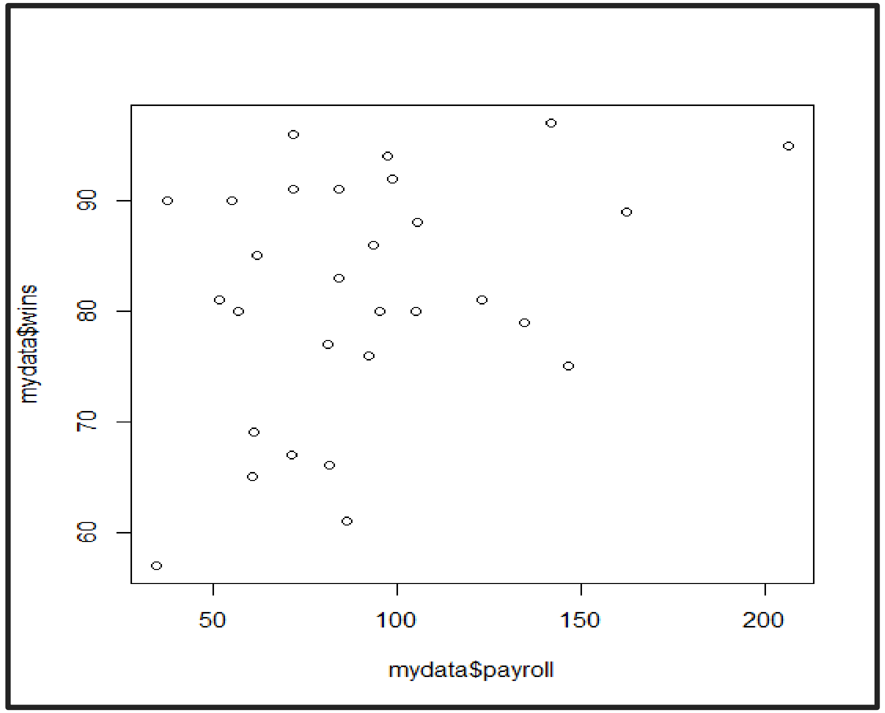
Payroll team payroll (in millions of dollars)

Wins number of games out of 162 that the team won.

```
mydata=read.csv("https://goo.gl/SsfWgg")
```

The Data

team	payroll	wins	winpct
New York Yankees	206.3	95	0.58642
Boston Red Sox	162.4	89	0.54938
Chicago Cubs	146.6	75	0.46296
Philadelphia Philli	141.9	97	0.59877
New York Mets	134.4	79	0.48765
Detroit Tigers	122.9	81	0.5
Chicago White Sox	105.5	88	0.54321
Los Angeles Angels	105	80	0.49383
San Francisco Gian	98.6	92	0.5679
Minnesota Twins	97.6	94	0.58025
Los Angeles Dodge	95.4	80	0.49383
St. Louis Cardinals	93.5	86	0.53086
Houston Astros	92.4	76	0.46914
Seattle Mariners	86.5	61	0.37654
Atlanta Braves	84.4	91	0.56173
Colorado Rockies	84.2	83	0.51235
Baltimore Orioles	81.6	66	0.40741
Milwaukee Brewer	81.1	77	0.47531
Tampa Bay Rays	71.9	96	0.59259
Cincinnati Reds	71.8	91	0.56173
Kansas City Royals	71.4	67	0.41358
Toronto Blue Jays	62.2	85	0.52469
Washington Natior	61.4	69	0.42593
Cleveland Indians	61.2	69	0.42593
Arizona Diamondb	60.7	65	0.40123
Florida Marlins	57	80	0.49383
Texas Rangers	55.3	90	0.55556
Oakland Athletics	51.7	81	0.5
San Diego Padres	37.8	90	0.55556
Pittsburgh Pirates	34.9	57	0.35185



Would you say the covariance is positive, negative or zero?

Beware of Interpreting Covariance

- Covariance depends on the units!

	payroll	wins	winpct
payroll	1461.5032644	154.7241379	0.955087269
wins	154.7241379	121.1034483	0.747552151
winpct	0.9550873	0.7475522	0.004614519

Only the **sign** of covariance matters

Making Size Matter

- Does a covariance of 154.72 imply a strong or weak relationship ?
- Solution: The correlation coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

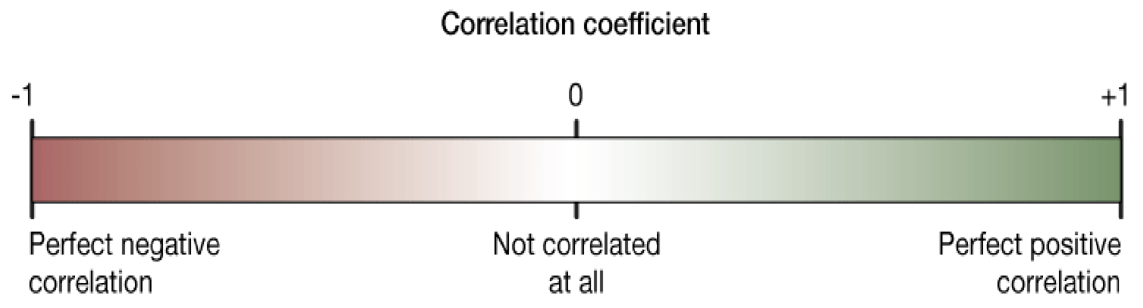
covariance

Standard deviation of x

Standard deviation of y

The Correlation

- A numerical summary of the strength of a linear relationship between two variables.
- Correlations are bound between –1 and 1.
- Sign: direction of the relationship (+ or -)
- Absolute value: strength of the relationship. Example: -0.6 is a stronger relationship tan +0.4



Correlation in R

```
> cor(mydata[,-1])
          payroll      wins    winpct
payroll 1.0000000 0.3677731 0.3677731
wins    0.3677731 1.0000000 1.0000000
winpct  0.3677731 1.0000000 1.0000000
```

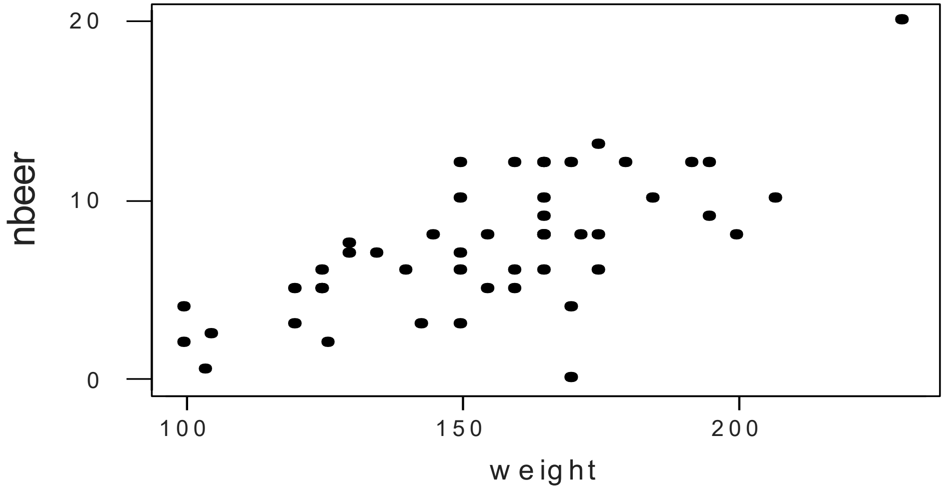
What is the correlation of Payroll with Payroll or WinPct with WinPct ?

Rule of Thumb

Magnitude of r	Interpretation
.00-.20	Very weak
.20-.40	Weak to moderate
.40-.60	Medium to substantial
.60-.80	Very Strong
.80-1.00	Extremely Strong

The correlation corresponding to the scatterplot we looked at earlier is:

```
Correlation of nbeer and weight = 0.692
```



Lifehack

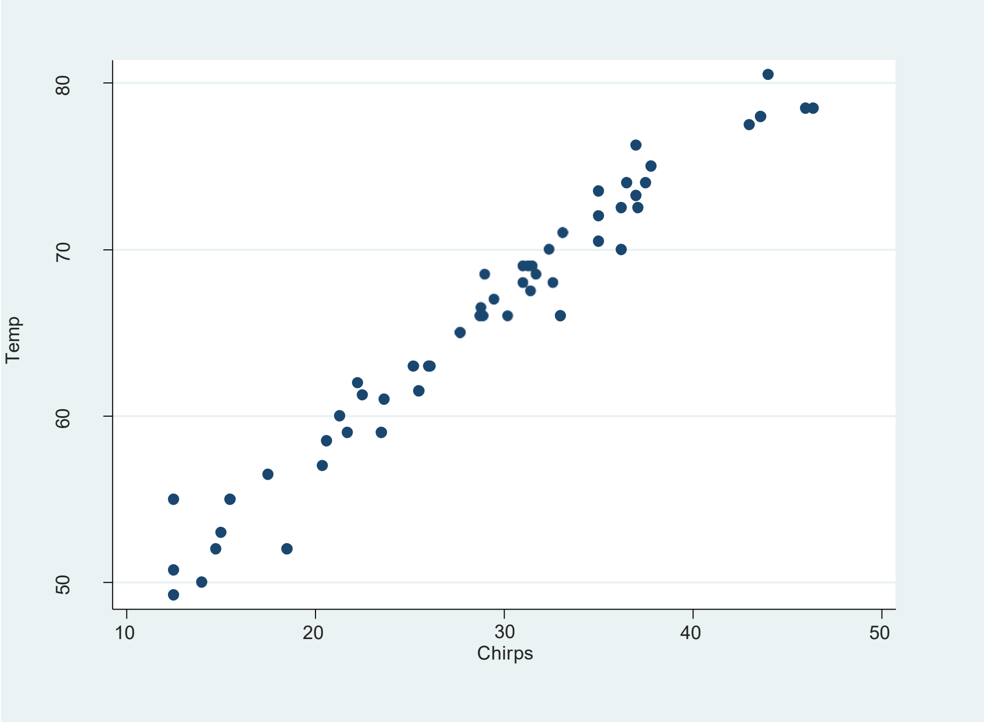
Home › Lifehack › How To Tell The Temperature With ...

How To Tell The Temperature With Cricket Chirps

The "cricket chirps – temperature" correlation first appeared in 1897 when physicist Amos Dolbear noticed that you can pretty accurately determine the number of cricket calls by using the outdoor temperature (the reversed idea). Since then, people have been using many ways to get the temperature by the number of chirps within a certain time interval but thankfully, science has finally given us the "golden formula". *(the article continues after the ad)*

Cricket Data

- X= number of chirps per 15 nds
- Y = Temperature



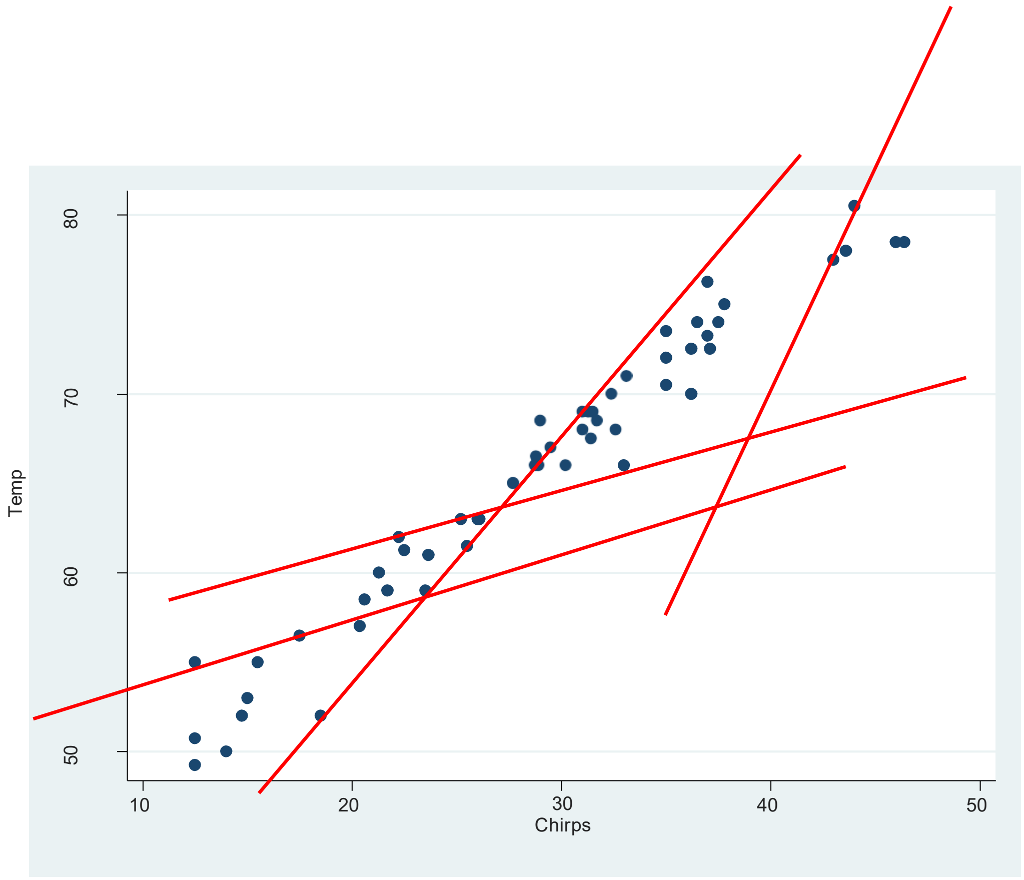
Data from http://blog.globe.gov/sciblog/2007/10/05/measuring-temperature-using-crickets/

How to Model?

- We will fit a line to the data set.
- This is what regression does-it relates a Y variable to an X variable.
- There are many ways to fit a line to data, though one method is the most popular (but not always the best method).

Which Two Points?

- Two points define a line, but which two points (and thus which line?)



Pause: The Equation of a Line

English words for the French word montant

amount, figure, rising, sum

- Most Americans have been brainwashed

$$Y = mX + b$$

- (allegedly in France they use y=sx+b)

- As adults, we will now use the notation

$$Y = b_0 + b_1 X$$

https://www.math.duke.edu//education/webfeats/Slope/Slopederiv.html

Notation for Our Line

- We need to be able to distinguish between our observed Y values, and the Y values that our line produces.

- So given a slope and intercept, we produce what is called the fitted line:

$$\hat{Y}_i = b_0 + b_1 X_i$$

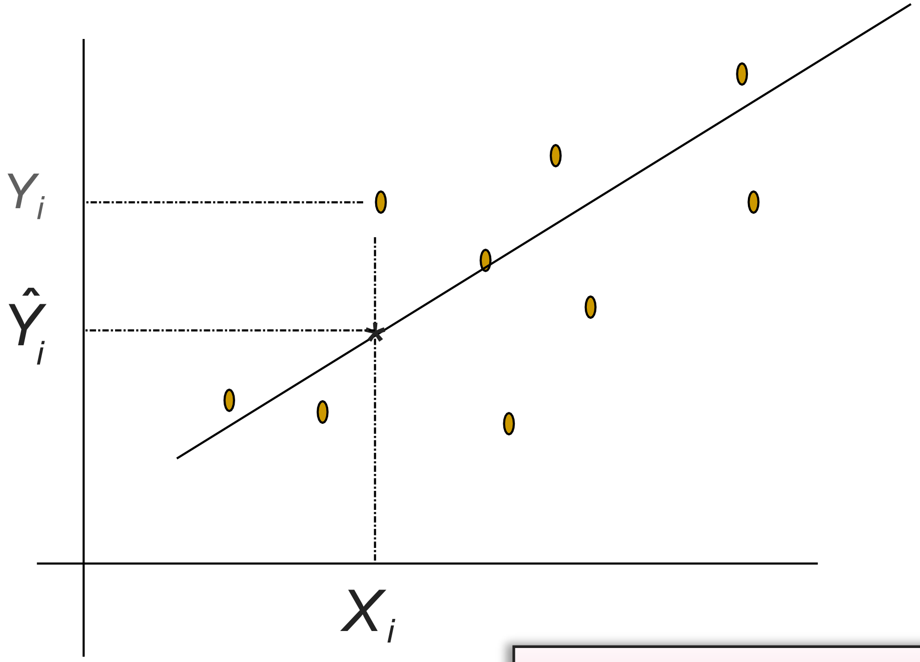
Pause: To Fit a Line to Data

- Fitting a line to data means to find "good" values of b_0 and b_1
- We define our fitting error as

$$e_i = Y_i - b_0 - b_1 X_i = Y_i - \hat{Y}_i$$

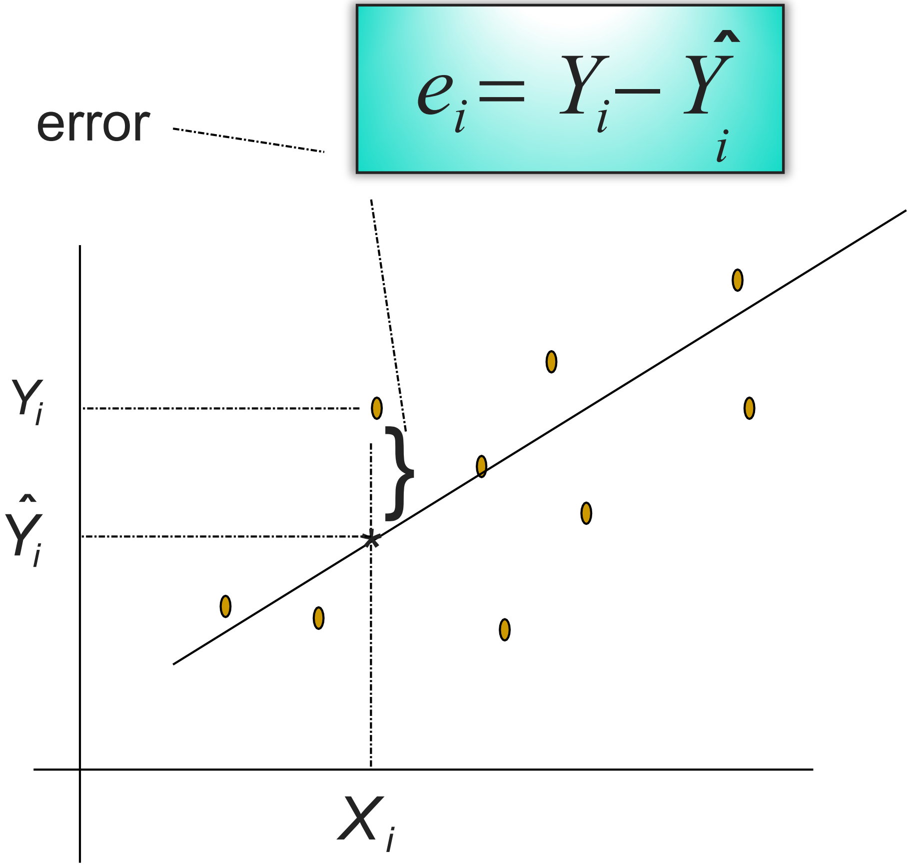
- Ideally, we want all the errors to be zero. Is this always possible?
- So we need a **criterion function**

Observed versus Fitted Values



$$\hat{Y}_i = b_0 + b_1 X_i$$

The Errors



Criterion Function

- The most popular method of fitting a line to data is called the **least-squares method**, and involves solving the following problem

$$\min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

- This can be solved in R, or, because it is a continuous criterion function, calculus can be used to find the solution.

Planeamiento Avanzado de Cadenas de Aprovisionamiento

Using R´s Least Squares Function

```
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5235 -0.8901  0.2048  1.0205  3.8273 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 40.02525    0.74414   53.79   <2e-16 ***
x            0.89180    0.02471   36.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

How wrong are we???????

The most popular criterion for **fitting a line** is called the *least squares method.* This method says to

Find b_0 and b_1 ← These two values define a line

that makes this sum as small as possible

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

The farther away a point is from the estimated line, the more serious the error. By squaring the errors, we “penalize” large residuals so that we can avoid them.

The values of b_0 and b_1 which minimize the residual sum of squares are:

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

These formulas can be derived using calculus- we pass.

These formulas are the intercept and slope for the “best fitting line”.

Example

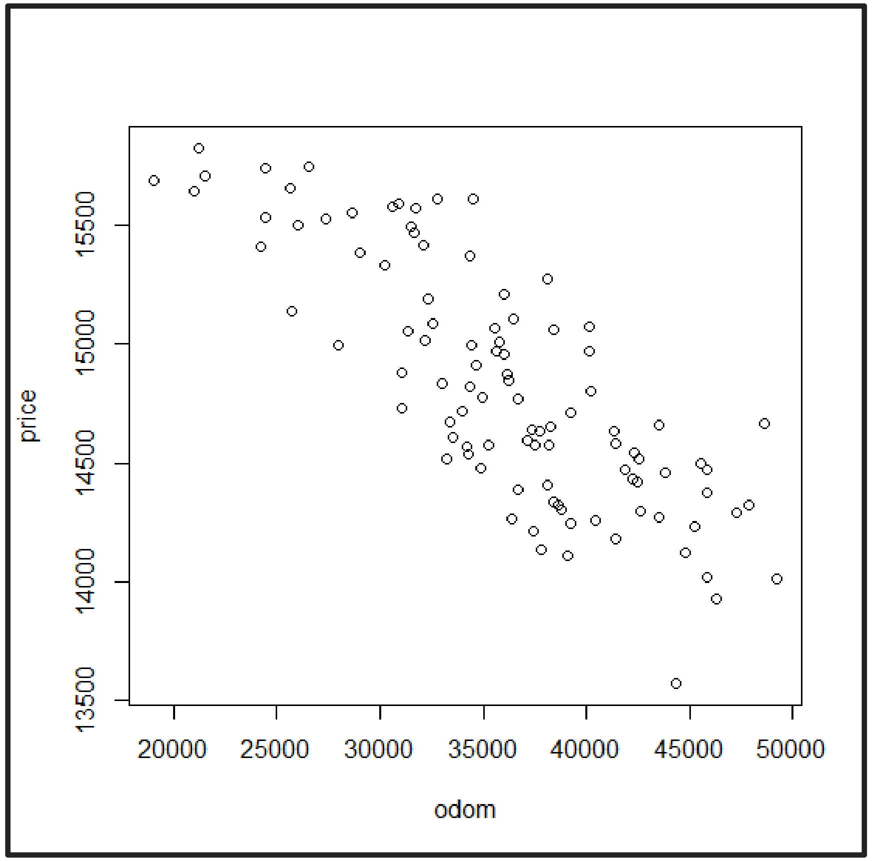
- Suppose we want to **predict** the sale price of used Honda Accords.
- Many factors influence the price of a used car; model year, condition, transmission type, 2 or 4 door, color, mileage, how badly owner wants to sell, etc....
- We will choose just the variable mileage and see if price can be predicted from the mileage of the car.

Setting up everything in R

```
> fname="http://people.fas.harvard.edu/~mparzen/stat139/accordprices.csv"
> mydata=read.csv(fname)

> names(mydata)
 [1] "Price"     "Odometer" "Color"     "X"         "X.1"       "X.2"
 [7] "X.3"       "X.4"      "X.5"       "X.6"       "X.7"

> price=mydata$Price
> odom=mydata$Odometer
```

Scatter Plot of Car Data

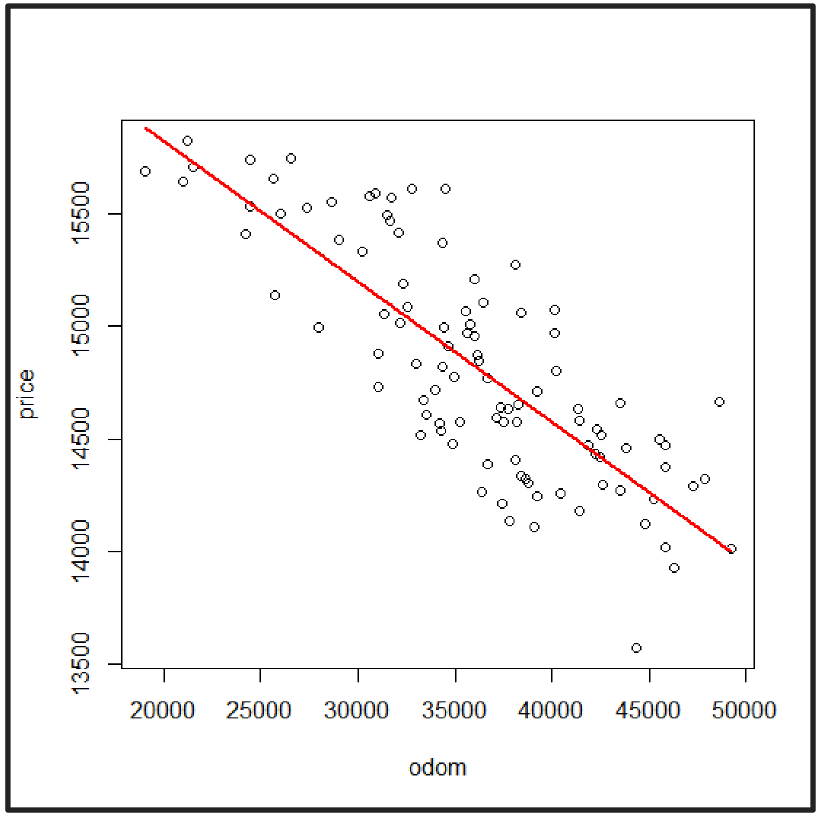
What's going on?

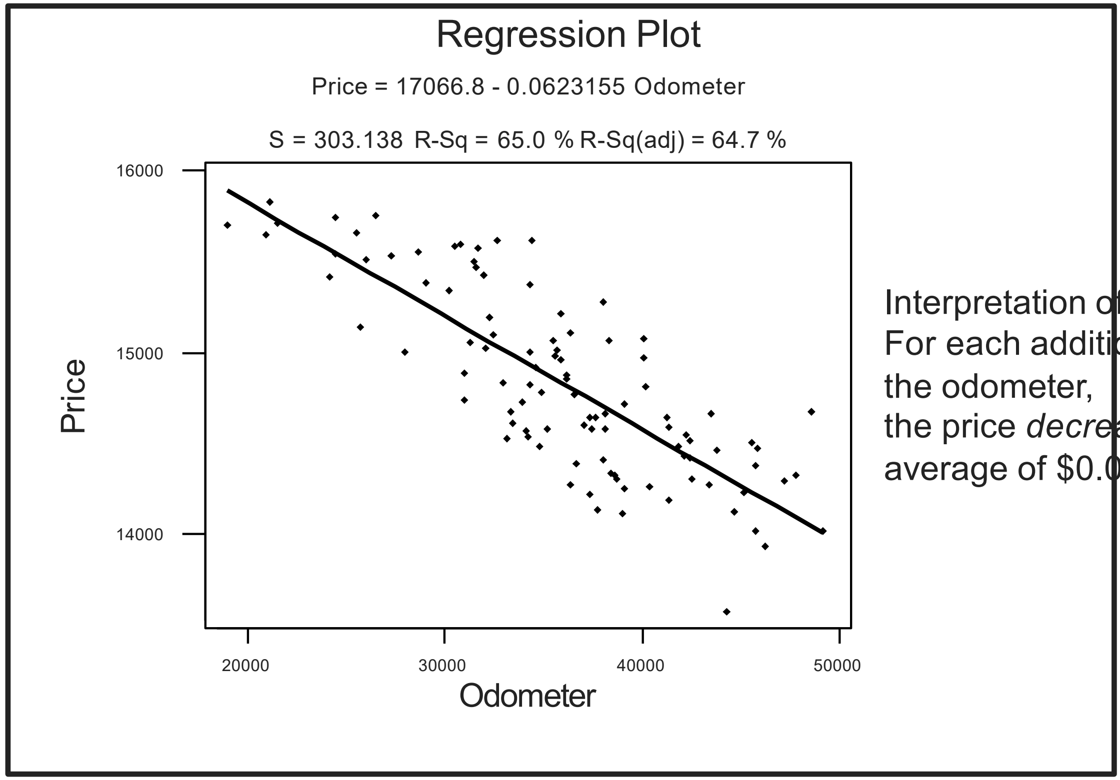


Performing Regression in R

```
> library(car)
> fit=lm(price~odom)
> summary(fit)

Call:
lm(formula = price ~ odom)

Residuals:
     Min        1Q      Median         3Q        Max
-730.32 -235.01           1.31     187.75     691.25

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.707e+04  1.690e+02  100.97   <2e-16 ***
odom        -6.232e-02  4.618e-03  -13.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 303.1 on 98 degrees of freedom
Multiple R-squared:  0.6501,    Adjusted R-squared:  0.6466
F-statistic: 182.1 on 1 and 98 DF,  p-value: < 2.2e-16

> plot(odom,price)
> regLine(fit)
```

Fitted Line Plot in R





Interpretation of the slope:
For each additional mile on the odometer,
the price *decreases* by an average of $0.062

Do not interpret the intercept as cars that have not been driven cost $17066.8

Properties of the Residuals and Fitted Values

The residuals and fitted values obtained from the least squares line have special properties.

Let's go back to the Accord data and check
them out.

Obtaining the residuals and fits

```
> fit=lm(price~odom)
> resids=residuals(fit)
> fits=fitted(fit)
> cbind(odom,price,fits,resids)
   odom price     fits      resids
1 37388 14636 14736.91 -100.914999
2 44758 14122 14277.65 -155.649930
3 45833 14016 14210.66 -194.660791
4 30862 15590 15143.59  446.414196
```

X **Y** $\hat{Y}$ e

Planeamiento Avanzado de Cadenas de Aprovisionamiento

What can R tell us about the residuals?

```
> summary(resids)
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-730.300 -235.000     1.306     0.000   187.700   691.200
```

Hmm. The mean of the residuals is 0.

What does that imply about the sum of the residuals?

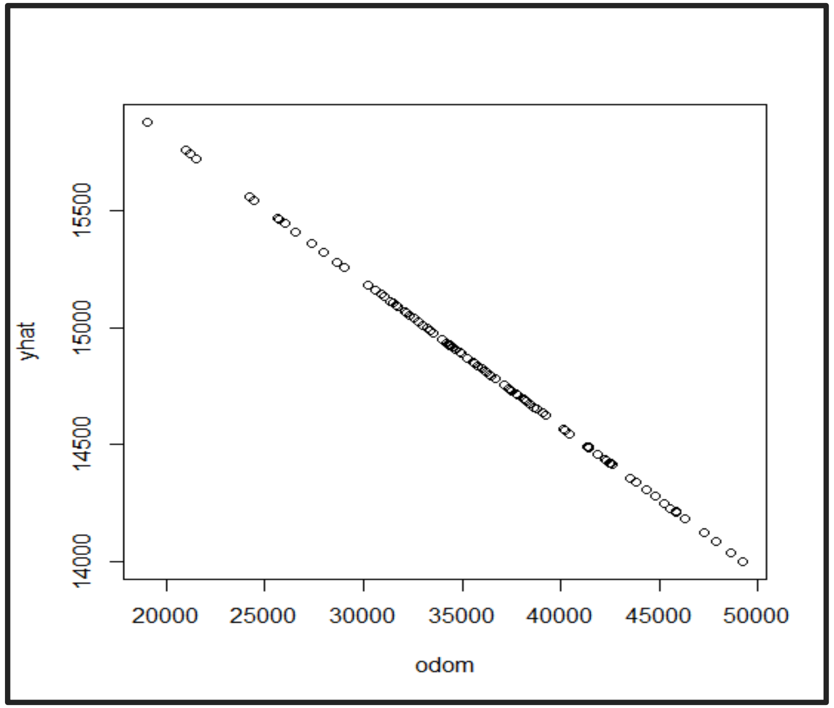
Does this make sense?

What can R tell us about the residuals?

Let´s check out these "yhat" values

```
> yhat=fits
> plot(odom,yhat)
>
```

Plot of $\hat{Y}$ versus odometer



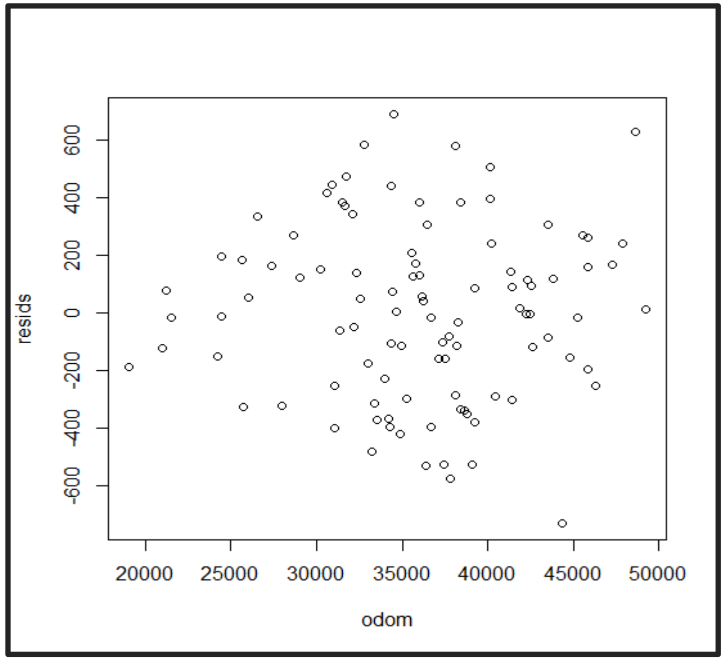
Is there a linear relationship between yhat and *X*?

$$corr(\hat{Y}, X) = ?$$

Planeamiento Avanzado de Cadenas de Aprovisionamiento

Let's get a handle on these "e" things.

Plot of *e* versus odometer



Is there a linear relationship between *e* and *X*?

$$corr(e, X) = ?$$

Planeamiento Avanzado de Cadenas de Aprovisionamiento

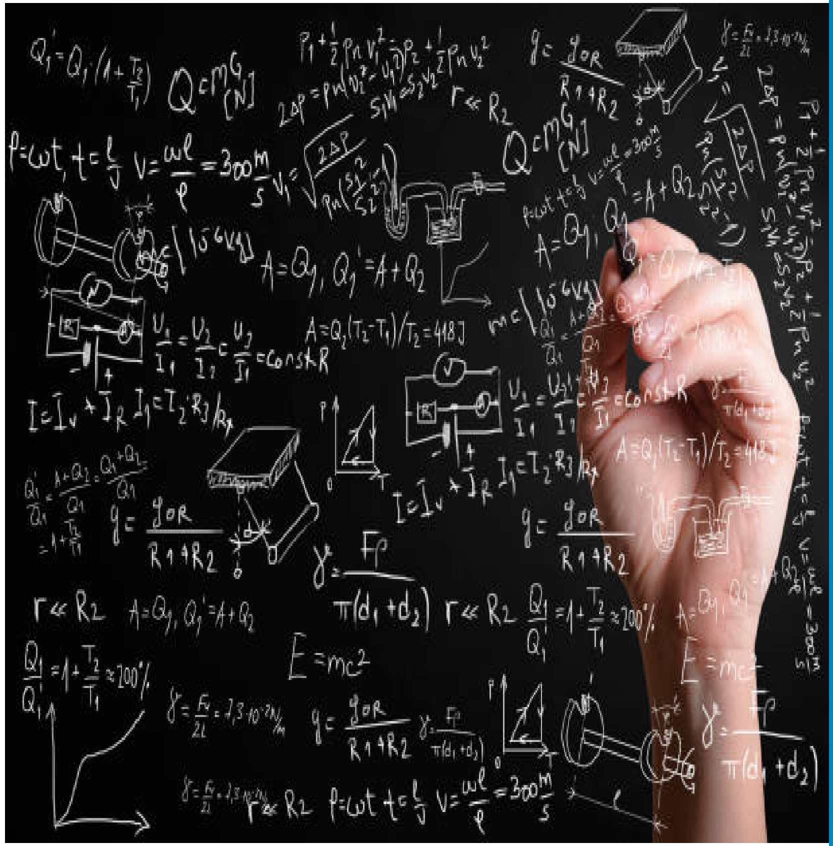
Basic Algebra

$$Y = \hat{Y} + (Y - \hat{Y})$$

or equivalently

$$Y = \hat{Y} + e$$

this is an important decomposition of Y



To summarize:

We have the decomposition of our observation

$$Y = \hat{Y} + e$$

Related to X
$[corr(\hat{Y}, X) = 1]$

Unrelated to X
$[corr(e, X) = 0]$

So, $$Var(Y) = Var(\hat{Y}) + Var(e)$$

or,

$$\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \frac{1}{n-1}\sum_{i=1}^{n} e_i^2$$

or,

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n} e_i^2$$

total sum of squares **SST**

regression ss **SSR**

error ss **SSE**

$$Var(Y) = Var(\hat{Y}) + Var(e)$$

SST = SSR + SSE

Decomposing information

The Accord Data Again

```
> summary(fit)
Call:
lm(formula = price ~ odom)

Residuals:
     Min          Median      3Q      Max
                1

     Q
-730.32 -235.01     1.31   187.75   691.25
(Intercept)     Estimate Std. Error t value Pr(>|t|)
Coefficients:1.707e+04    1.690e+02  100.97   <2e-16 ***
odom          -6.232e-02   4.618e-03  -13.49   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 303.1 on 98 degrees of freedom
Multiple R-squared:  0.6501,    Adjusted R-squared:  0.6466
F-statistic: 182.1 on 1 and 98 DF,  p-value: < 2.2e-16
```

the R squared value

R2 Criticism

The Race (3): Coefficient of Determination?

R^2 is often called the "coefficient of determination." The result (or cause) of this unfortunate terminology is that the R^2 statistic is sometimes interpreted as a measure of the influence of X on y. Others consider it to be a measure of the fit between the statistical model and the true model. A high R^2 is considered to be proof that the correct model has been specified or that the theory being tested is correct. A higher R^2 in one model is taken to mean that that model is better.

All these interpretations are wrong. R^2 is a measure of the spread of points around a regression line, and it is a poor measure of even that (Achen, 1982). Taking all variables as deviations from their means, R^2 can

Worse, however, is that there is no statistical theory behind the R^2 statistic. Thus, R^2 is not an estimator because there exists no relevant population parameter. All calculated values of R^2 refer only to the particular sample from which they come. This is clear from the standardized coefficient example in preceding paragraphs, but it is more graphically

Q: But do you really want me to stop using R^2? After all, my R^2 is higher than that of all my friends and higher than those in all the articles in the last issue of the *APSR!*

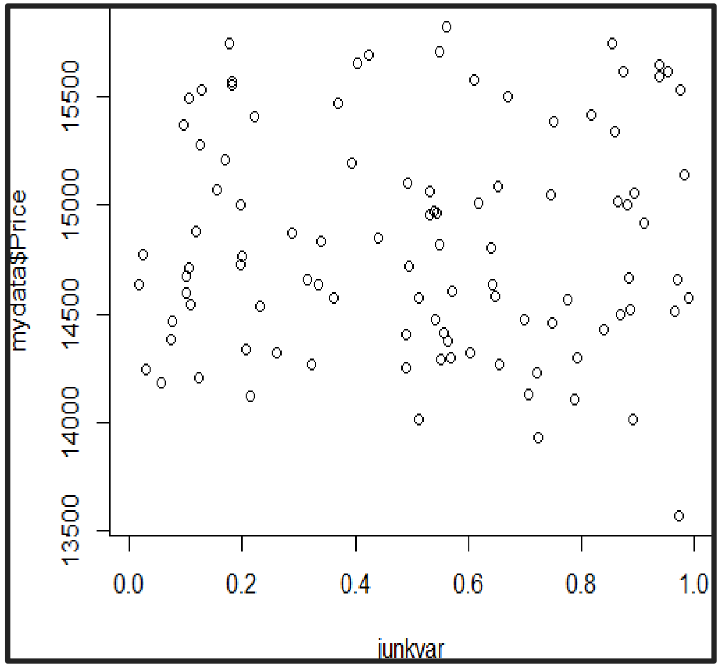
A: If your goal is to get a big R^2, then your goal is not the same as that for which regression analysis was designed. The purpose of regression analysis and all of parametric statistical analyses is to estimate interesting population parameters (regression coefficients in this case). *The best regression model usually has an* R^2 *that is lower than could be obtained otherwise.*

Planeamiento Avanzado de Cadenas de Aprovisionamiento

Example: Adding Junk to a Model

- We can generate random data in R

```
junkvar=runif(length(mydata$Price))
 plot(junkvar,mydata$Price)
```



There is no relationship between price and this junk variable.

Original Model

```
> summary(fit)

Call:
lm(formula = Price ~ Odometer)

Residuals:
   Min      1Q  Median      3Q     Max 
-730.3 -235.0     1.3   187.7   691.2 

Coefficients:
               Estimate  Std. Error t value Pr(>|t|)    
(Intercept) 17066.76607   169.02464   101.0   <2e-16 ***
Odometer       -0.06232     0.00462   -13.5   <2e-16 ***
---
Signif. codes:              0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 303 on 98 degrees of freedom
  (139 observations deleted due to missingness)
Multiple R-squared:  0.65,      Adjusted R-squared:  0.647 
F-statistic:  182 on 1 and 98 DF,  p-value: <2e-16
```

However, what happens to R2?

```
> fit=lm(mydata$Price~mydata$Odometer+junkvar)
> summary(fit)

Coefficients:
                    Estimate  Std. Error t value Pr(>|t|)
(Intercept)      17045.57912   177.03317   96.28   <2e-16 ***
mydata$Odometer     -0.06235     0.00464  -13.44   <2e-16 ***
junkvar             43.44618   103.13553    0.42     0.67
---
Residual standard error: 304 on 97 degrees of freedom  (139 observations
  deleted due to missingness)
Multiple R-squared:  0.651,     Adjusted R-squared:  0.644
F-statistic: 90.4 on 2 and 97 DF,  p-value: <2e-16
```

- The value of R-squared went up, even though this isn't a better model!
- **Looking towards the next lecture, there is info on this output that tells us we don't need junkvar in the model**

Example

For example, we know that there isn't an **exact relationship** between mileage of a car and its price (how do we know this, by the way?)

price = \$17067-\$.06 (odometer)

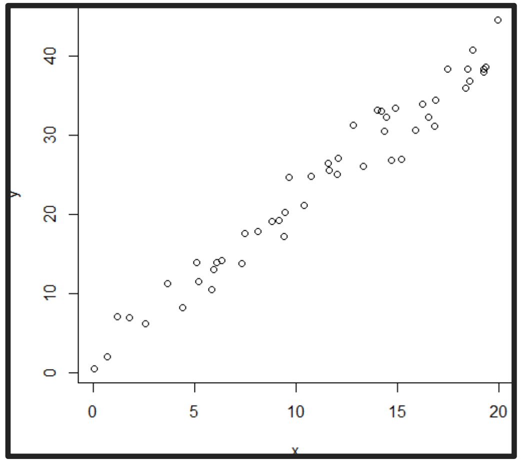
That is, not every Accord with 30000 miles will sell for \$15267. Some will sell for more, and some houses will sell for less.

A more realistic statement is that

Average Car Price = \$17067-\$.06(odometer) This is a main point about regression: we model the **average of something** rather than the something itself.

When things are right

Consider the data:



this plot looks like the kind of data our model is meant to describe.

Always plot Y vs X!

As a further check we examine the residuals.

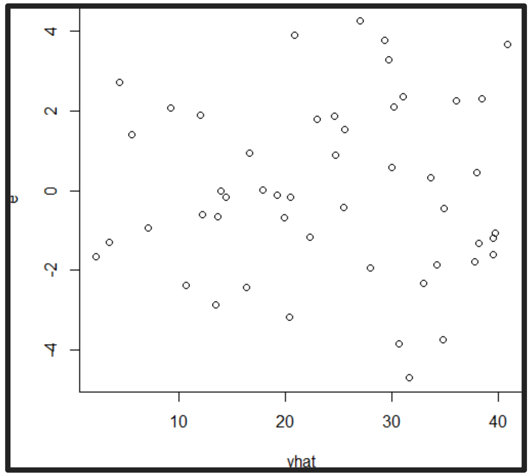
Obtaining Residuals in R

- We need the residuals, fitted values and standardized residuals

```
> fit=lm(y~x)
> e=residuals(fit)
> yhat=fitted(fit)
> sres=rstudent(fit)
```

Plot Residuals versus Yhat

```
plot(yhat,e)
```



$$Y = \hat{Y} + e$$

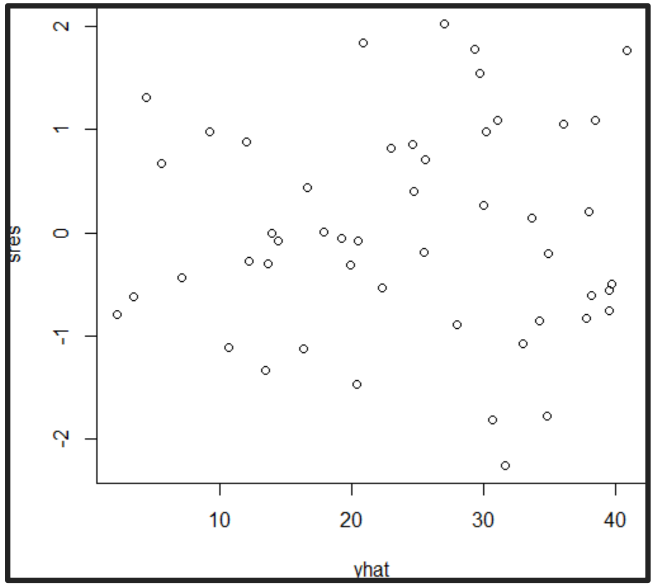
This is the way a residual plotlooks when the model fits the data:

No obvious pattern!!!!!

residsunrelated to X!!!!!!

Planeamiento Avanzado de Cadenas de Aprovisionamiento

(or) Plot standardized residuals vs Yhat



no obvious pattern!!!!!

resids unrelated to X!!!!!!

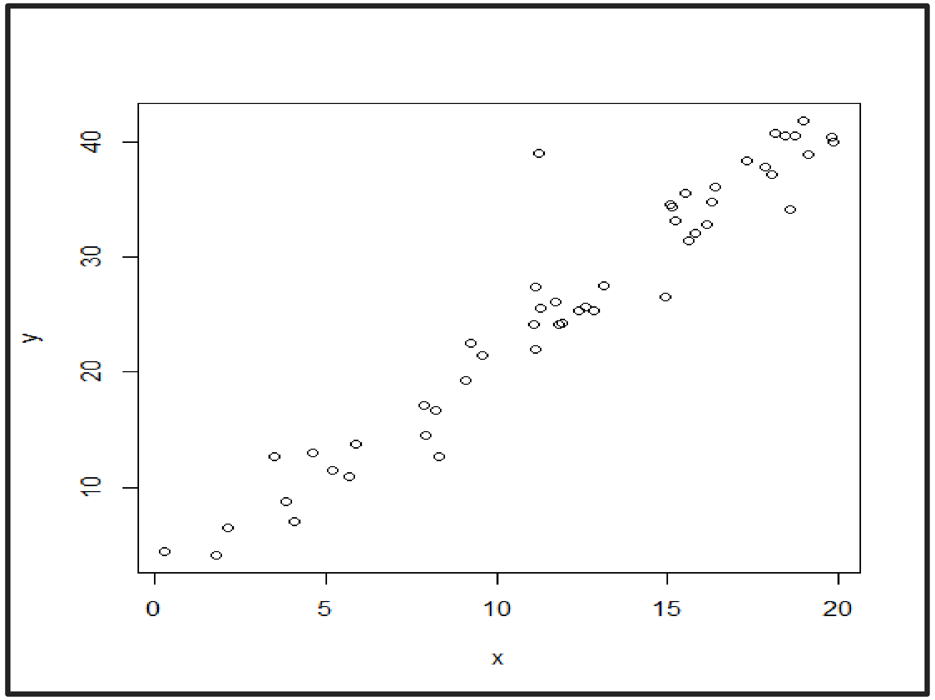
standardized resids between -2 and +2!!!!!!

Outliers

Sometimes we get a point which is unusual- different from all the rest, in that the deviation away from the line seems particularly large. We call these funny points **outliers** (because it sounds better than "funny points").

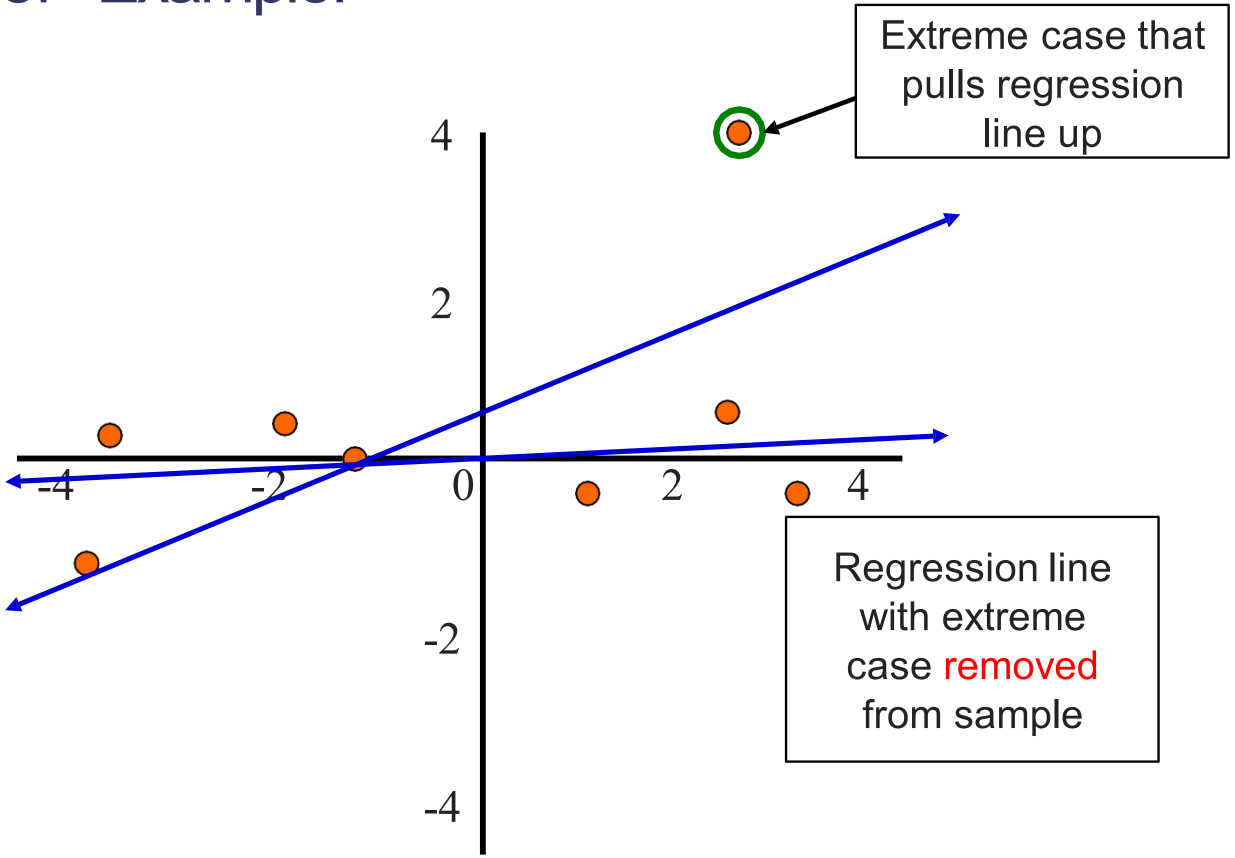
Consider the data set

There seems to be one funny point !!



Outliers can dramatically change the line

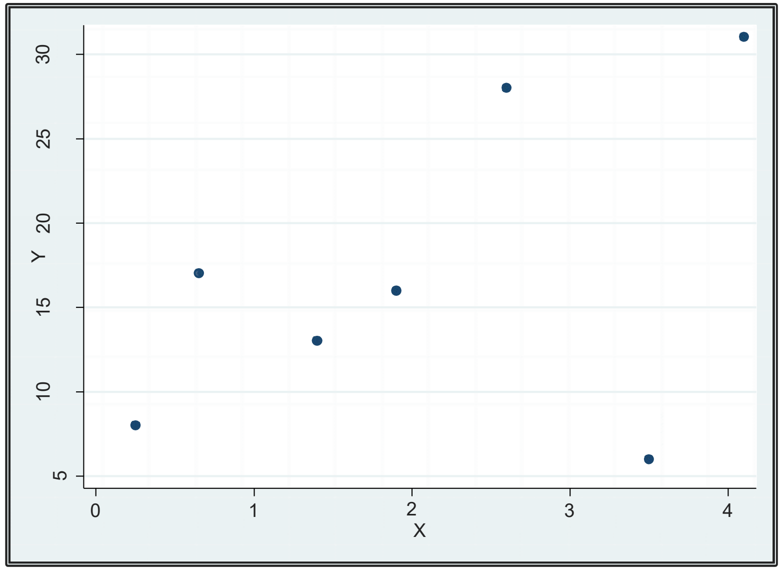
- Outlier Example:



Example: Study time and student achievement

- X variable: Average # hours spent studying per day
- Y variable: Score on reading test

Case	X	Y
1	2.6	28
2	1.4	13
3	.65	17
4	4.1	31
5	.25	8
6	1.9	16
7	3.5	6



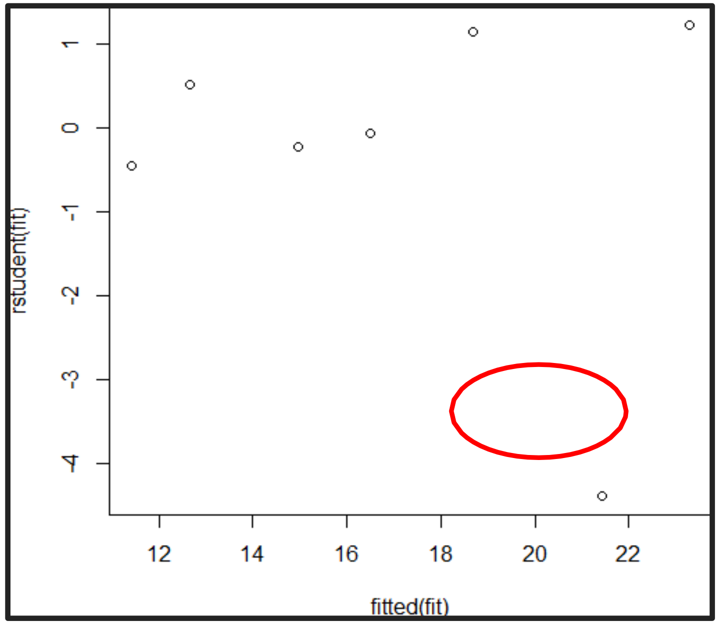
Regression Output

```
> fit=lm(y~x)
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
      1          2         3     4         5         6        7
       9.3274    -     4.3355  7.7058    -3.4320    -0.5158 -15.4456
           1.9753
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    10.662      6.402   1.665    0.157
x               3.081      2.617   1.177    0.292

Residual standard error: 9.162 on 5 degrees of freedom
Multiple R-squared:  0.217,     Adjusted R-squared:  0.0604
F-statistic: 1.386 on 1 and 5 DF,  p-value: 0.2921
```

Do you need X in the model?

Diagnostic Plot

```
plot(fitted(fit),rstudent(fit))
```

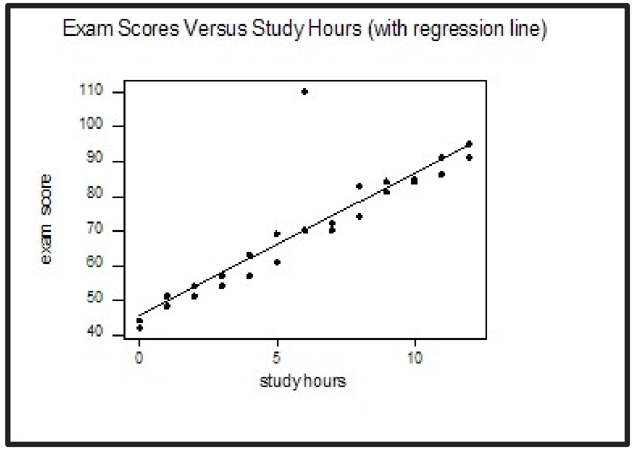


Planeamiento Avanzado de Cadenas de Aprovisionamiento

Remove the outlier

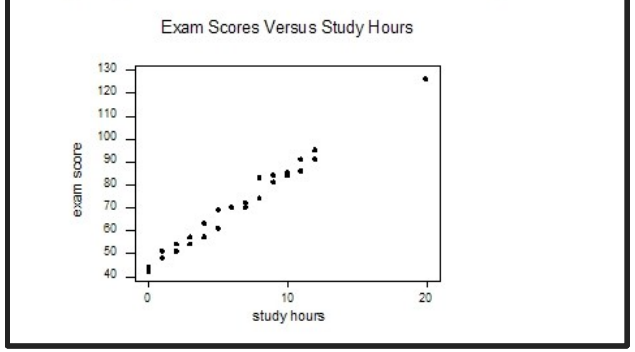
```
> fit=lm(y[-7]~x[-7])
> summary(fit)

Call:
lm(formula = y[-7] ~ x[-7])

Residuals:
   1         2         3         4        5        6
      4.6798 -3.4467      4.8492 -  -1.8597  -3.3107
                            0.9119
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     8.428      3.019    2.791    0.0492 *
x[-7]           5.728      1.359    4.215    0.0135 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.259 on 4 degrees of freedom
Multiple R-squared:  0.8163,    Adjusted R-squared:  0.7703
F-statistic: 17.77 on 1 and 4 DF,  p-value: 0.01353
```

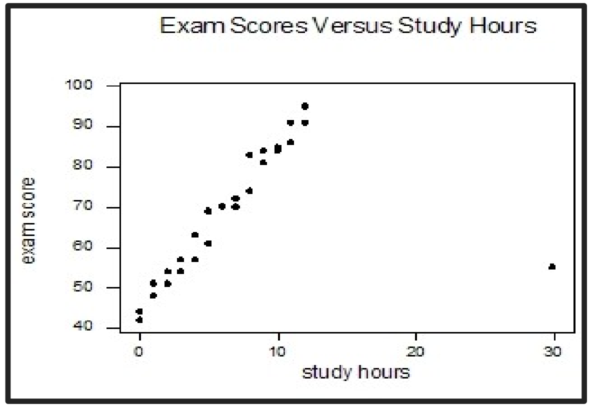
There is now a relationship! The outlier was hiding the linear relationship. Naughty outlier!



Note: Not all outliers are bad

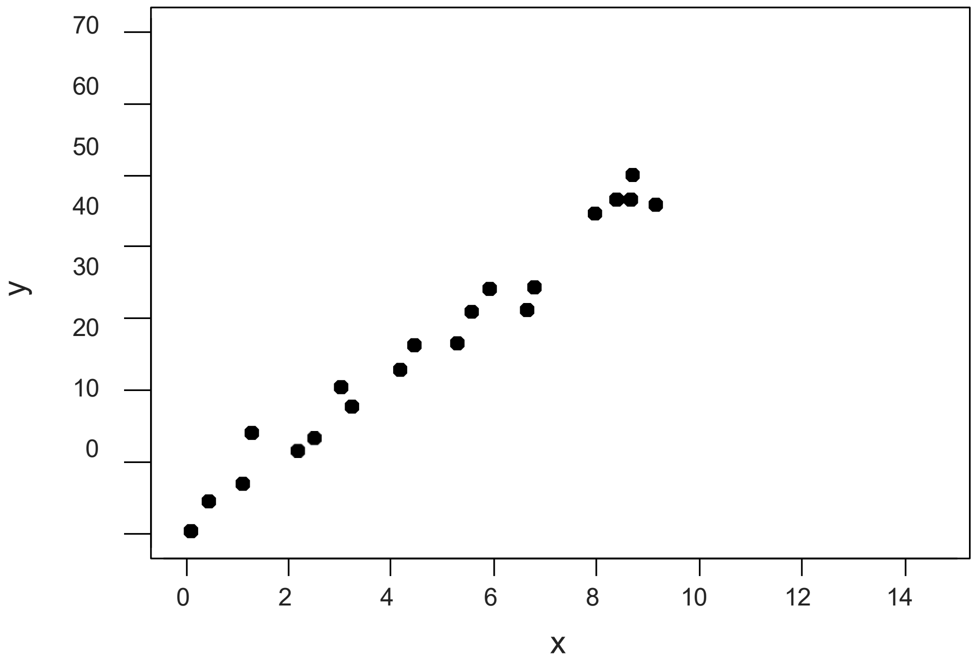
Example of an outlier that doesn't have a large residual:



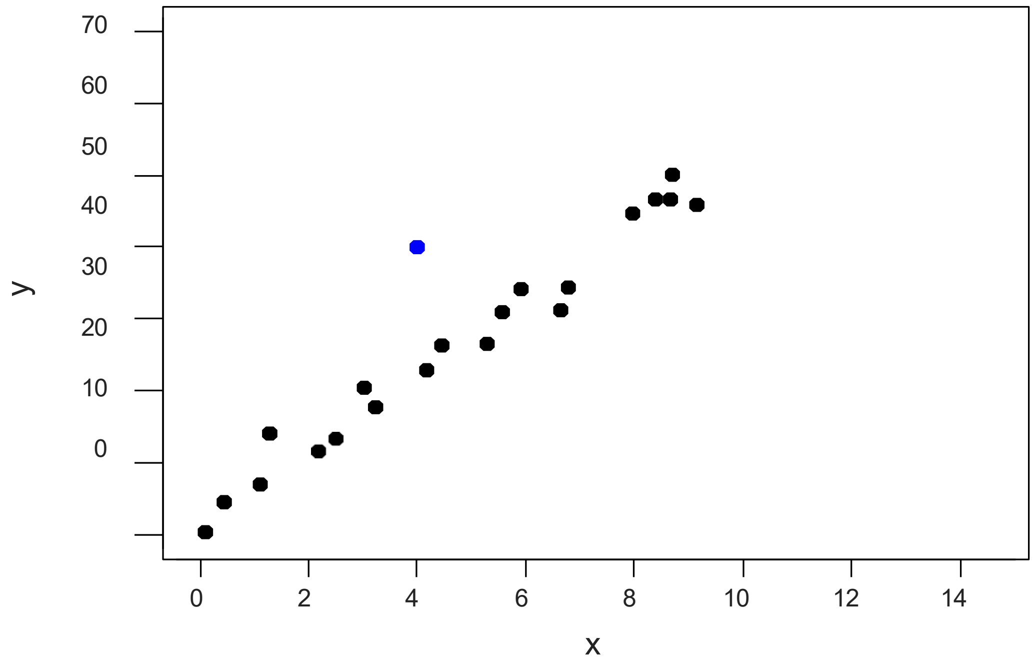


Influential observations are the worst.

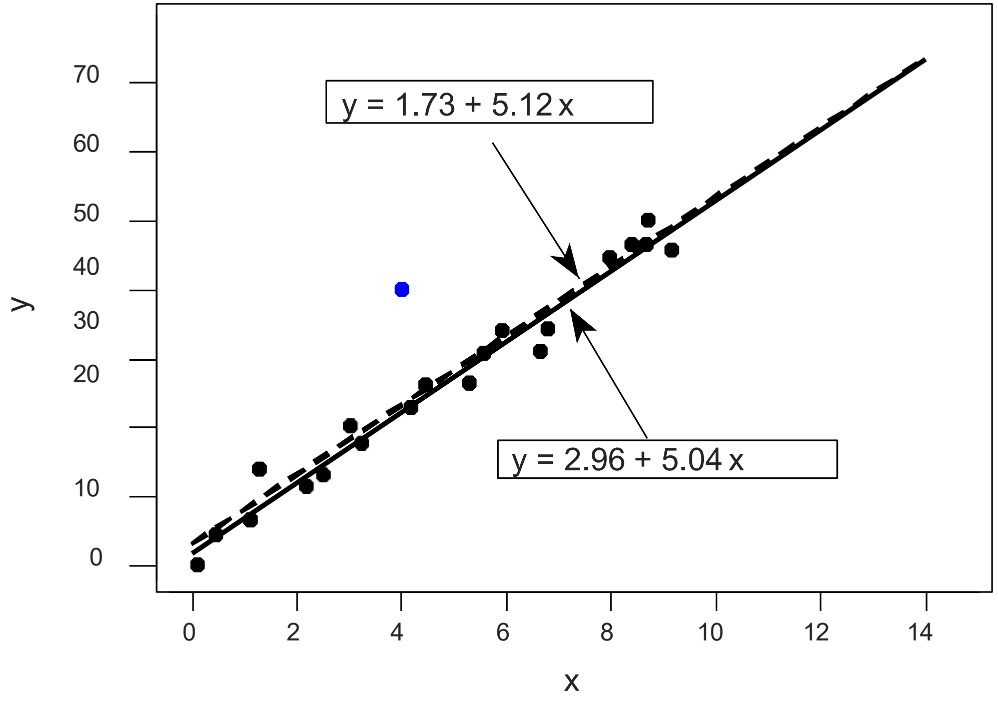
No outliers?



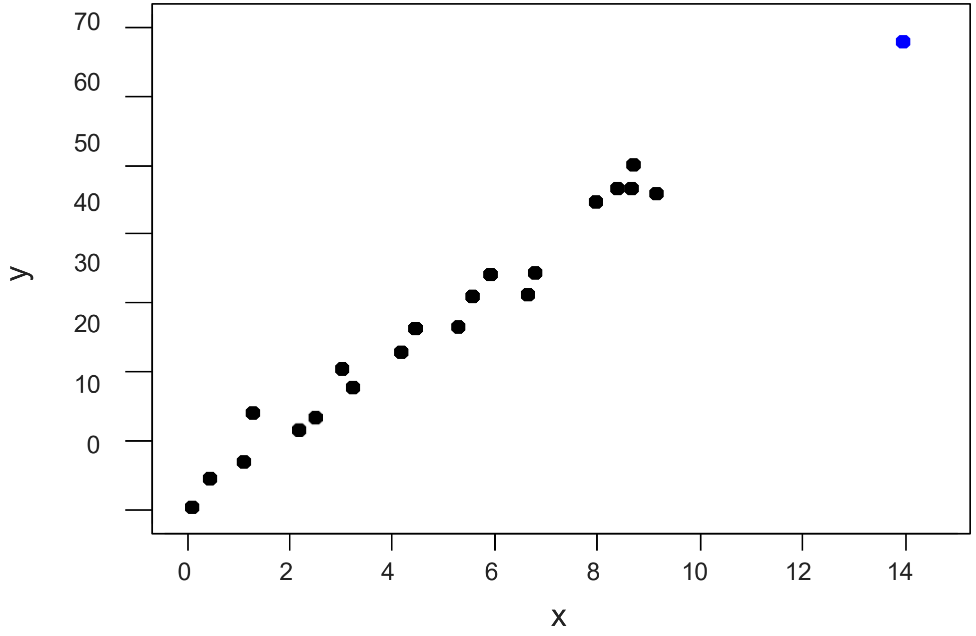
An outlier? Influential?



An outlier? Influential?

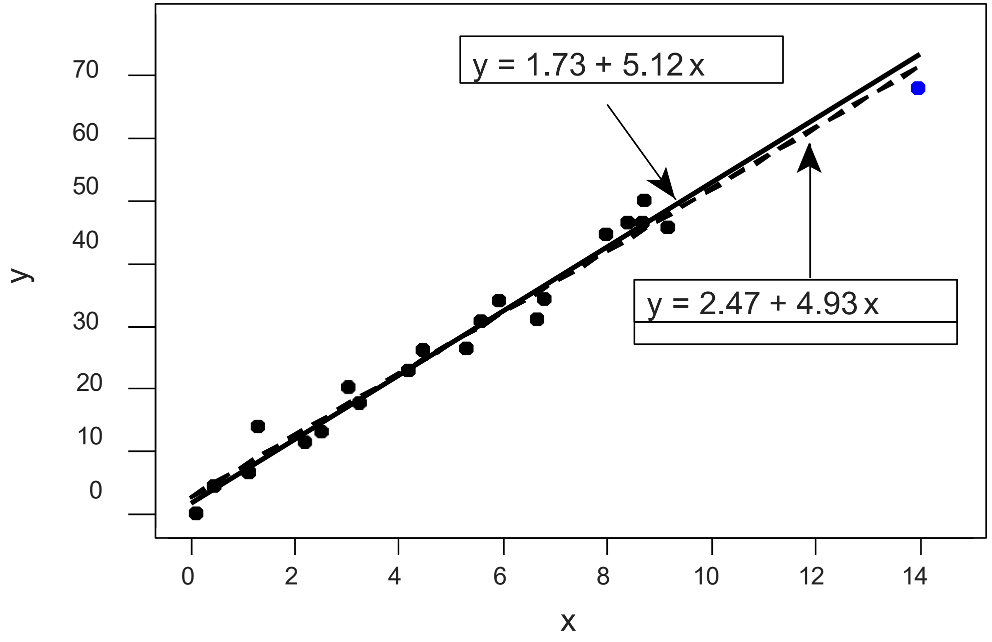


An outlier? Influential?

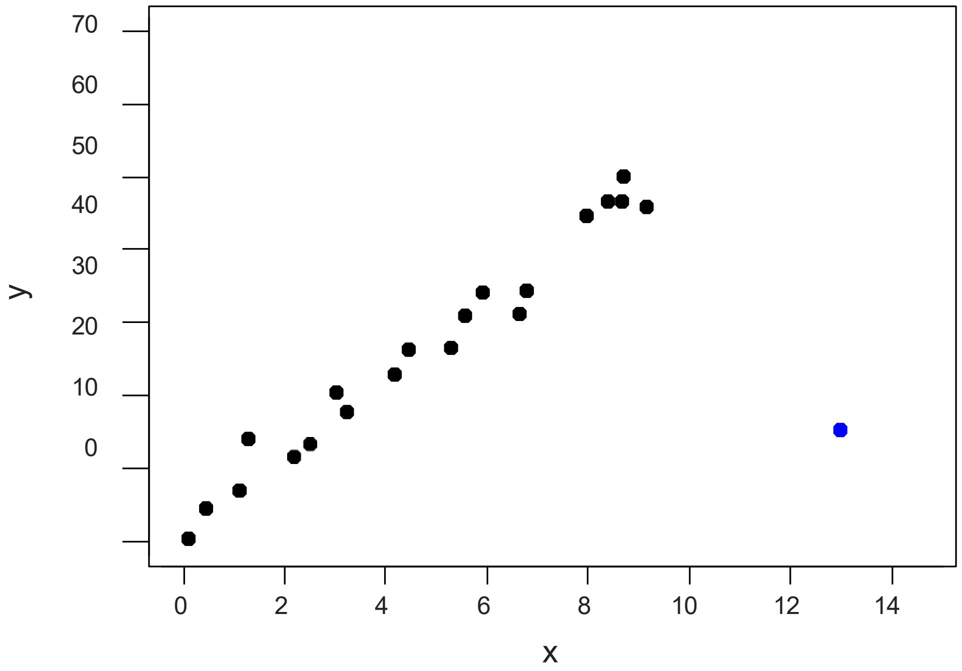


Planeamiento Avanzado de Cadenas de Aprovisionamiento

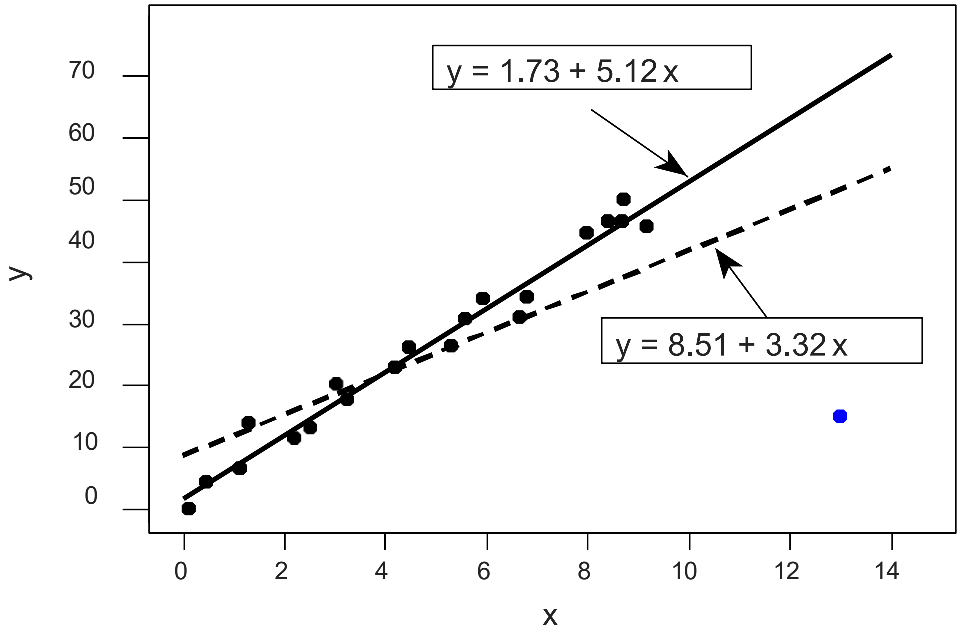
An outlier? Influential?



An outlier? Influential?



An outlier? Influential?



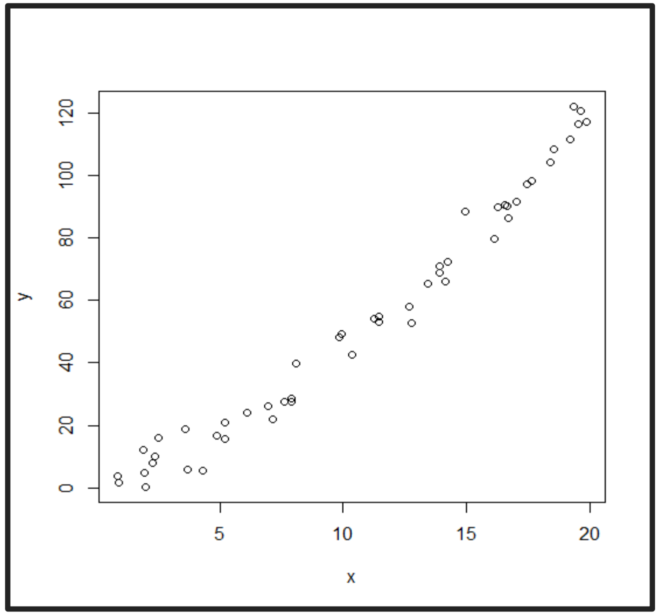
Planeamiento Avanzado de Cadenas de Aprovisionamiento

Nonlinearity

Another key assumption is that *Y* is a linear function of *X*.

What happens when this assumption fails ? Consider the data plotted below:

There is some nonlinearity evident in the plot !!



We run the regression and obtain the standardized residuals

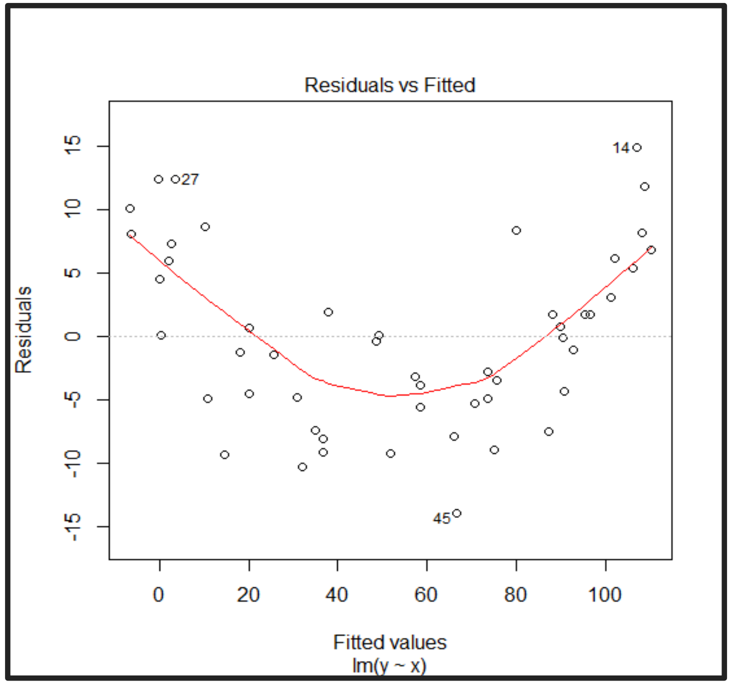
```
> fit=lm(y~x)
> sumary(fit)
Error: could not find function "sumary"
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-13.8924  -4.9015  -0.2035   5.8075  14.8862

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.8471     2.0254  -5.849 4.26e-07 ***
x             6.1471     0.1644  37.396  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.044 on 48 degrees of freedom
Multiple R-squared:  0.9668,    Adjusted R-squared:  0.9661
F-statistic:  1398 on 1 and 48 DF,  p-value: < 2.2e-16
```

Note that R^2 is pretty high.

As a diagnostic, we plot the residuals versus X

```
plot(fit,which=1)
```

there should be no relationship between the resids and X!!!!



The nonlinearity is even more evident in the residual plot !! What is wrong with fitting a linear regression to this data?

Planeamiento Avanzado de Cadenas de Aprovisionamiento

THANK YOU!!

Section for questions

Carlos Mariño, Ph.D.
Director de Investigación
Profesor
(511) 626 7100 ext 7200
cmarino@pucp.pe
www.centrum.pucp.edu.pe