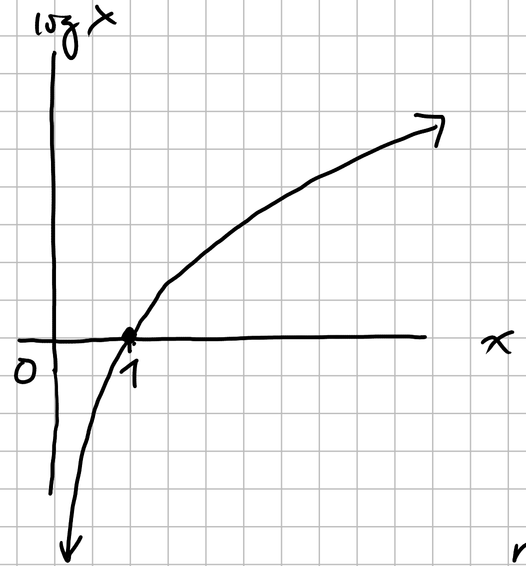
impact of transformations on visualizations & numerical summaries

$X_t \to \log X_t$



$\log_b(1) = 0$ regardless of the base $b>0$

$\log_b 0$ undefined $b>0$

$\overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$ $\qquad$ $\frac{1}{n}\sum_{t=1}^{n} \log X_t$

Are these two related?

$$\log\left(\frac{1}{n}\sum_{t=1}^{n} X_t\right) \neq \frac{1}{n}\sum_{t=1}^{n} \log X_t$$

→ log (average) ≠ average (log)

Question: numerical summary
transformation

transform (numerical summary)
$\stackrel{?}{=}$ numerical summary (transform)

✓ 1 2 3 4 5 original scale

$\log_e 1$ $\quad \log_e 2$ $\quad \boxed{\log_e 3}$ $\quad \log_e 4$ $\quad \log_e 5$ transform using logs $\qquad$ order is preserved

How to undo log

$$\log_b x = y$$

$$b^{\log_b x} = b^y$$

$$\| $$

$$x$$

$$b^y = x$$

$3 \longrightarrow \log_e 3$

$\longleftarrow \exp(\log_e 3)$

$\| $

$e^{\log_e 3}$

NOTE In R, 1e02 $*10^2$

don't confuse with base e

Why make a big deal?

In practice, you might not have access to the original data.

$X_t \longrightarrow X_t - \overline{X}$ centering

$X_t \longrightarrow \dfrac{X_t - \overline{X}}{sd}$ standardizing

$X_t \longrightarrow X_t^2$

---

data $\longrightarrow$ visualization / summaries

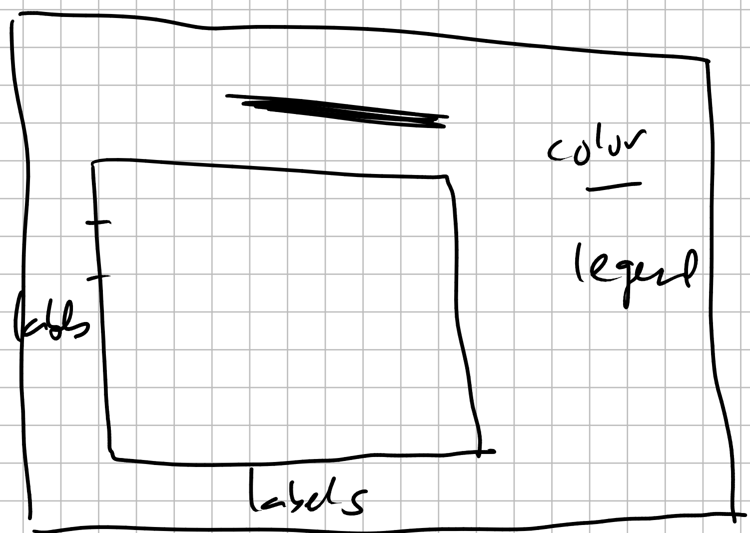
transform $\downarrow$ reduce

transformed data $\longrightarrow$?

monetary amounts   base 10 might be a good idea

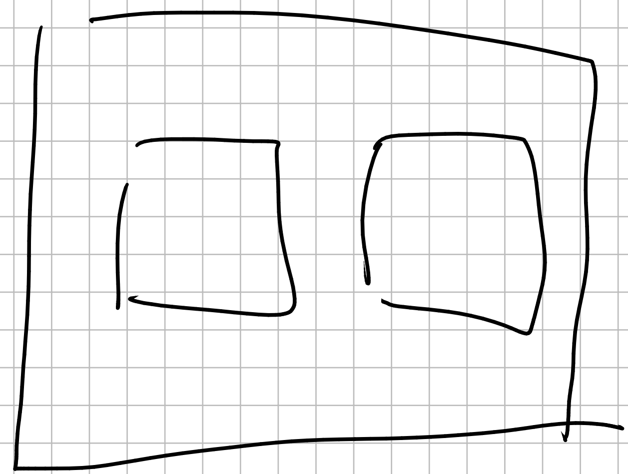
$$\log_{10} 100 = 2$$
$$\log_{10} 1000 = 3$$
$$\log_{10} 10000 = 4$$
$$\log_{10} 1{,}000{,}000 = 6$$

Scatterplot ← visualization involving two variables



WYSIWYG

Canvas



lm(TotalCompMil ~ NetSalesMil, data = execComp)

lm(execComp$TotalCompMil ~ execComp$NetSalesMil)

dependent variables
independent variables } misleading descriptions.

Interpretation in some books:

↑ net sales by 1 million $ ⟶ ↑ CEO compensation by 170 $

implicitly thinking of the same CEO!

counterfactual outcomes

conditions are required to have this "nicer" interpretation!

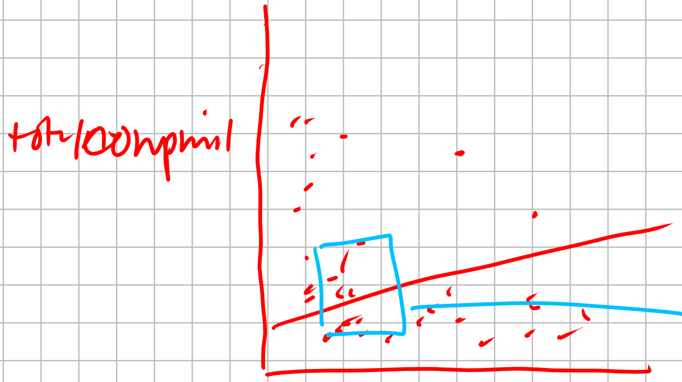
indexes → t

firms / CEOs

t		NetSalesMil$_t$	TotalCompMil$_t$	$\widehat{\text{TotalCompMil}}_t$
1	1	1.5		$3.77 + 0.0017(1.5)$
2	1			
3	1			
:	:			
n	1			

different from each other!

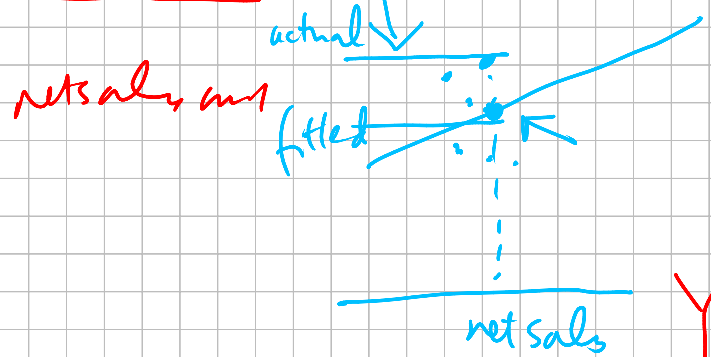
Y-hat

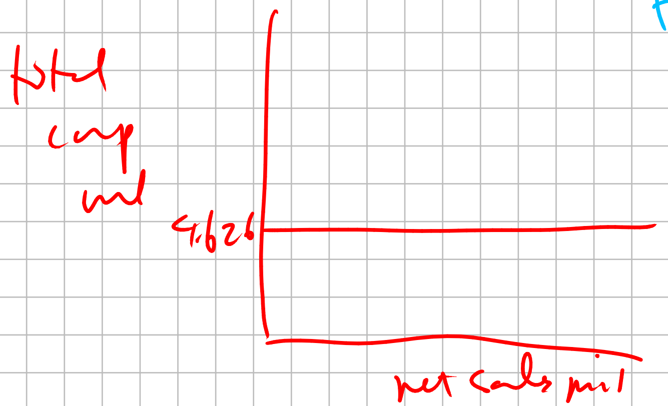


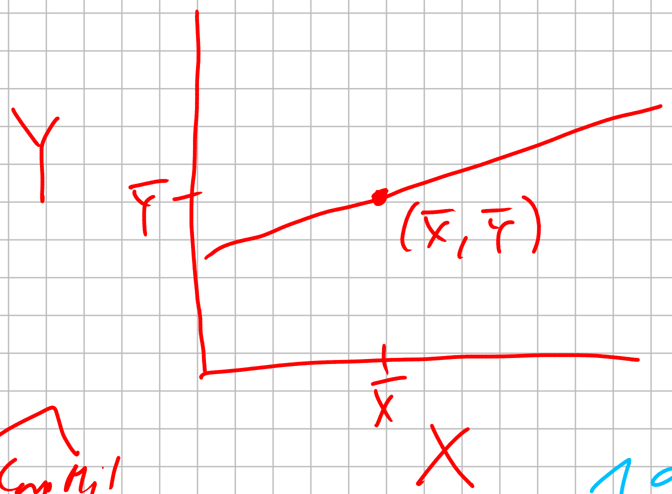
regression line

lm( )

$\text{totalCompMil}_t = 3.77 + 0.00017\ \text{NetSalesMil}_t$







$\hat{Y}_t = \text{intercept} + \text{slope} * X_t$

1.90×10^{-16} very very very close to 0

t	NetSalesMil	TotalCompMil	$\widehat{\text{TotalCompMil}}$
1			✓
2			✓
3			✓
⋮	⋮		⋮
n			

→ fitted values

fitted(lm(TotalCompMil ~ NetSalesMil, data = eCSub))

Sum of residuals for lm() = 0

$\Rightarrow$ average of residuals from lm() = 0

lm()

residuals sum up to zero

mean(fitted values) = mean(regressand)

$Y_t \to$ TotalCompMil$_t$

$X_t \to$ NetSalesMil

lm()

$$\min \boxed{\sum_{t=1}^{n}} \underbrace{\left(Y_t - \boxed{\hat{Y}_t} \right)}_{\text{residual}}{}^{2} \quad \text{squared}$$

sum

O L S

ordinary least squares

$$\min_{\text{intercept, slope}} \sum_{t=1}^{n} \left(Y_t - \left(\overset{?}{\text{intercept}} + \overset{?}{\text{slope}} * X_t \right) \right)^2$$