Applied Data Science
fall 2017
5004.002
Session 2: Bi-variate linear regression

Instructor: Prof. Stanislav Sobolevsky
Course Assistants: Tushar Ahuja, TBD

Supervised learning

Data/input	Labels/output	Dependence
x_1	y_1	$y = f(x)$
x_2	y_2	$y = w_1 x + w_0$
$\dots$	$\dots$	
x_N	y_N	

Linear Model - motivation

Motivation:

- simple
- easy to interpret
- often sufficient
- serve as a baseline

Examples:

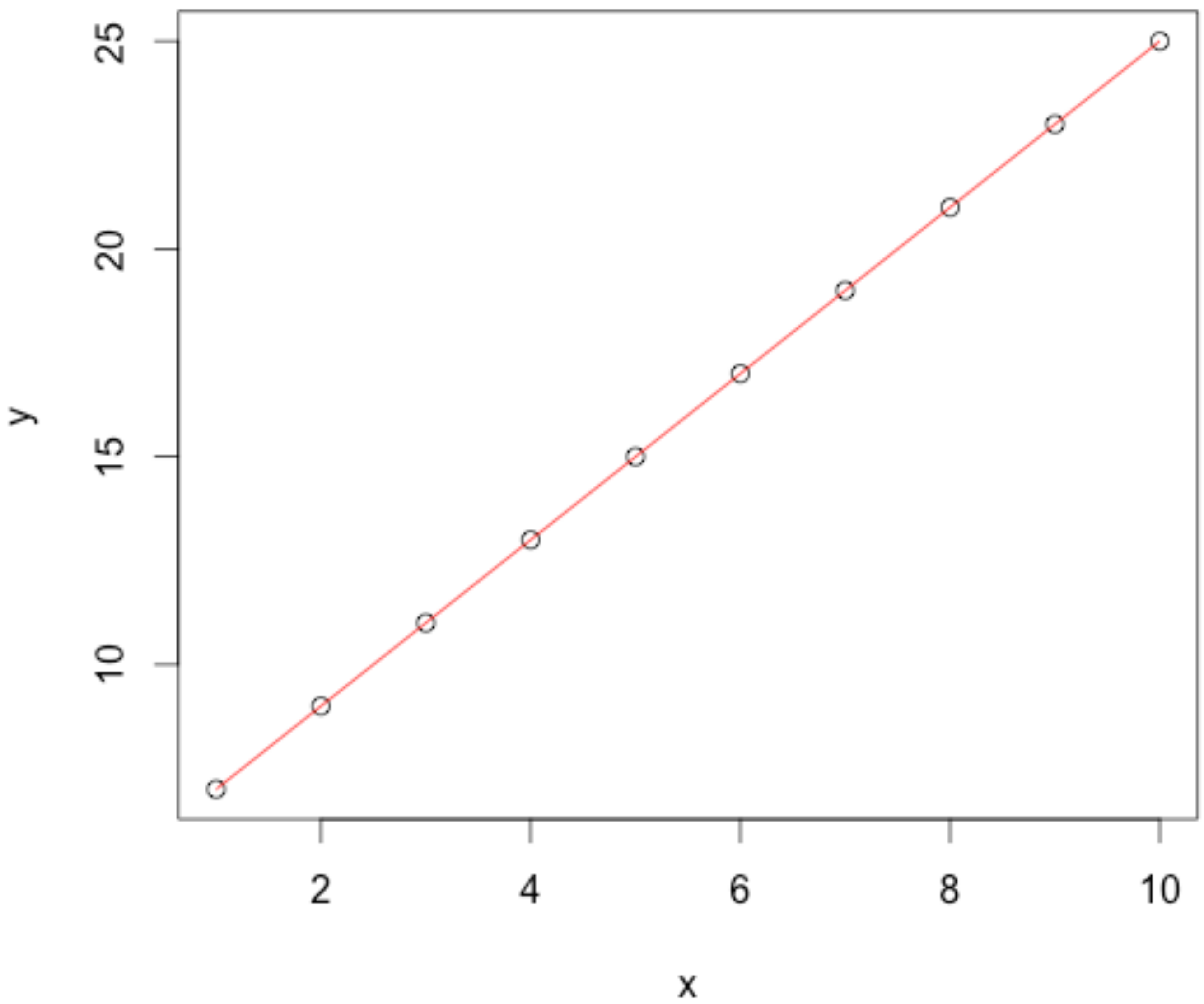
- House price depending on size
- Vehicle emission depending on speed
- Energy usage depending on building size, occupancy, T
- Average income depending on the education level
- Taxi usage depending on temperature
- Urban income, crime, innovation vs population

Bi-variate Linear Model

$$y \sim x \qquad \{(x_i, y_i), i = 1..N\}$$

$$y = w_1 x + w_0$$

$$y = 2x + 5$$



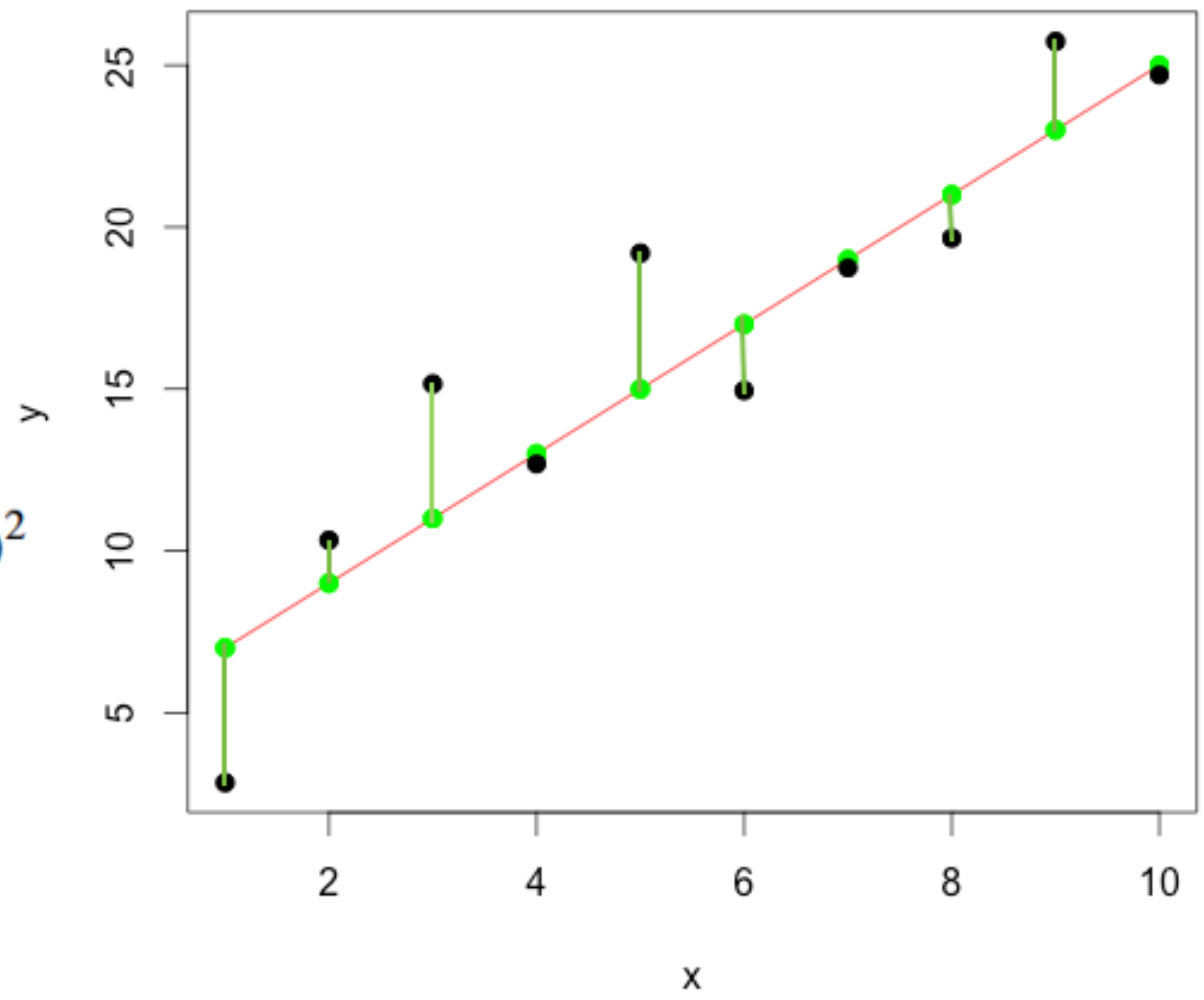
Linear Model

$$y = w_1 x + w_0 + \varepsilon$$

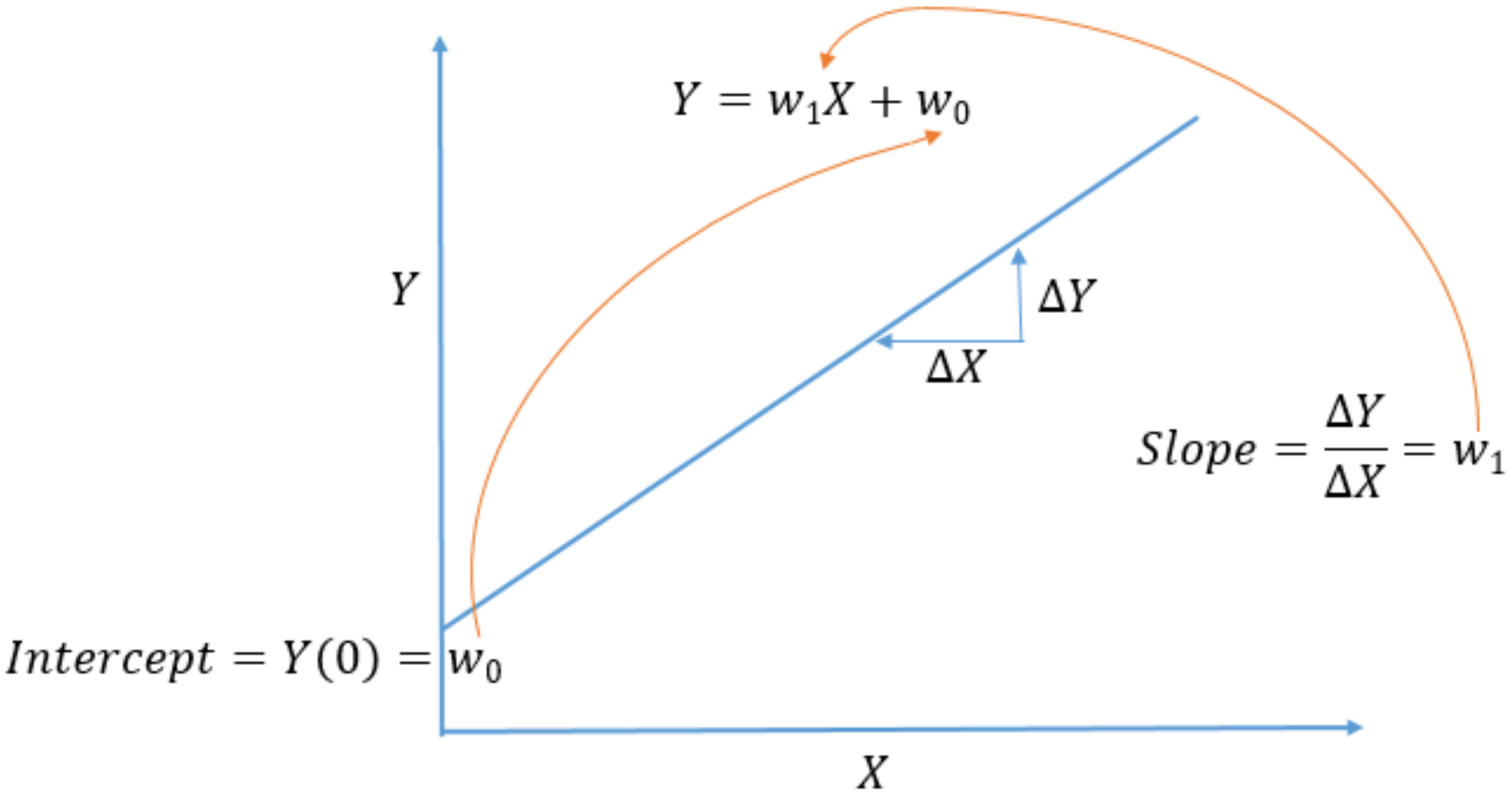
$$\varepsilon_i = y_i - w_1 x_i - w_0$$

$$RSS(w) = \sum_i \varepsilon_i^2 = \sum_i (y_i - w_1 x_i - w_0)^2$$

$$\hat{w} = argmin_w RSS(w)$$



Linear Model Coefficients - slope coefficient and intercept

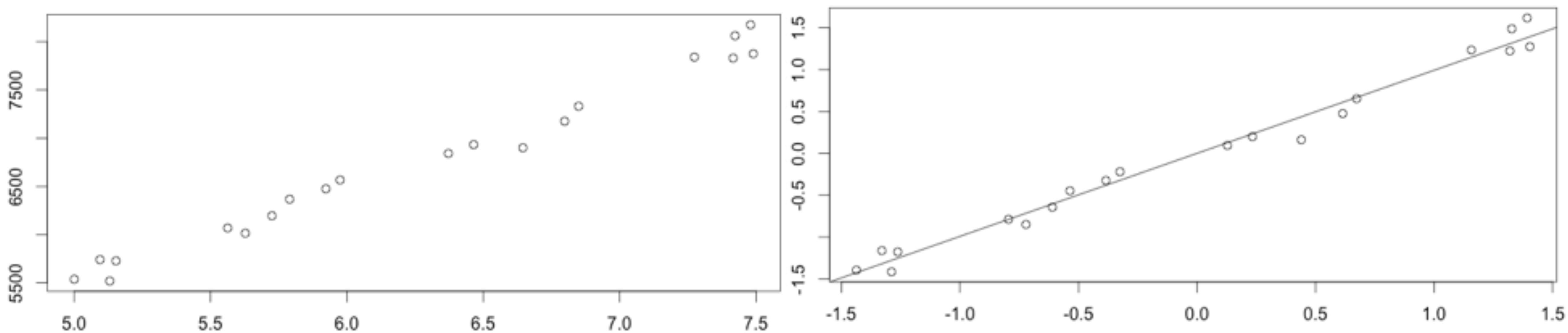


Linear Model - normalization

$$x := x - E[X] \qquad y := y - E[Y]$$

$$y = w_1 x + \varepsilon$$

$$x := x/std[X] \qquad y := y/std[Y]$$



Linear Model - basic fitting approach

$$RSS(w) = \sum_i \varepsilon_i^2 = \sum_i (y_i - w_1 x_i - w_0)^2$$

$$\hat{w} = argmin_w RSS(w)$$

$$\begin{cases} \frac{\partial RSS(\hat{w})}{\partial w_1} = 0, \\ \frac{\partial RSS(\hat{w})}{\partial w_0} = 0. \end{cases} \qquad \begin{cases} \sum_i 2x_i(y_i - \hat{w}_1 x_i - \hat{w}_0) = 0, \\ \sum_i 2(y_i - \hat{w}_1 x_i - \hat{w}_0) = 0, \end{cases}$$

Linear Model - basic approach

$$\begin{cases} \sum_i 2x_i(y_i - \hat{w}_1 x_i - \hat{w}_0) = 0, \\ \sum_i 2(y_i - \hat{w}_1 x_i - \hat{w}_0) = 0, \end{cases} \qquad \begin{cases} \hat{w}_1 \left(\sum_i (x_i)^2\right) + \hat{w}_0 \left(\sum_i x_i\right) = \sum_i x_i y_i, \\ \hat{w}_1 \left(\sum_i x_i\right) + N\hat{w}_0 = \sum_i y_i, \end{cases}$$

$$\left(\sum_i (x_i)^2 - \left(\sum_i x_i\right)^2 /N\right) \hat{w}_1 = \sum_i x_i y_i - \left(\sum_i y_i\right)\left(\sum_i x_i\right)/N$$

$$\hat{w}_1 = \frac{\sum_i x_i y_i - \left(\sum_i y_i\right)\left(\sum_i x_i\right)/N}{\sum_i (x_i)^2 - \left(\sum_i x_i\right)^2 /N} \qquad \hat{w}_0 = \frac{\sum_i y_i - \hat{w}_1\left(\sum_i x_i\right)}{N}$$

Linear Model - basic approach

$$\hat{w}_1 = \frac{\sum_i x_i y_i - \left(\sum_i y_i\right)\left(\sum_i x_i\right)/N}{\sum_i (x_i)^2 - \left(\sum_i x_i\right)^2 /N}$$

$$\hat{w}_0 = \frac{\sum_i y_i - \hat{w}_1 \left(\sum_i x_i\right)}{N}$$

$$\hat{w}_1 = \frac{\frac{\sum_i x_i y_i}{N} - \frac{\sum_i y_i}{N}\frac{\sum_i x_i}{N}}{\frac{\sum_i (x_i)^2}{N} - \left(\frac{\sum_i x_i}{N}\right)^2}$$

$$E[X] = \frac{\sum_i x_i}{N} \qquad E[Y] = \frac{\sum_i y_i}{N}$$

$$var[X] = E[(X - E[X])^2]$$

$$= E[X^2] - 2E[X]^2 + E[X]^2 = E[X^2] - E[X]^2$$

$$\hat{w}_1 = \frac{E[XY] - E[X]E[Y]}{E[X^2] - E[X]^2} = \frac{E[(X - E[X])(Y - E[Y])]}{var[X]}$$

$$\hat{w}_0 = E[Y] - \hat{w}_1 E[X]$$

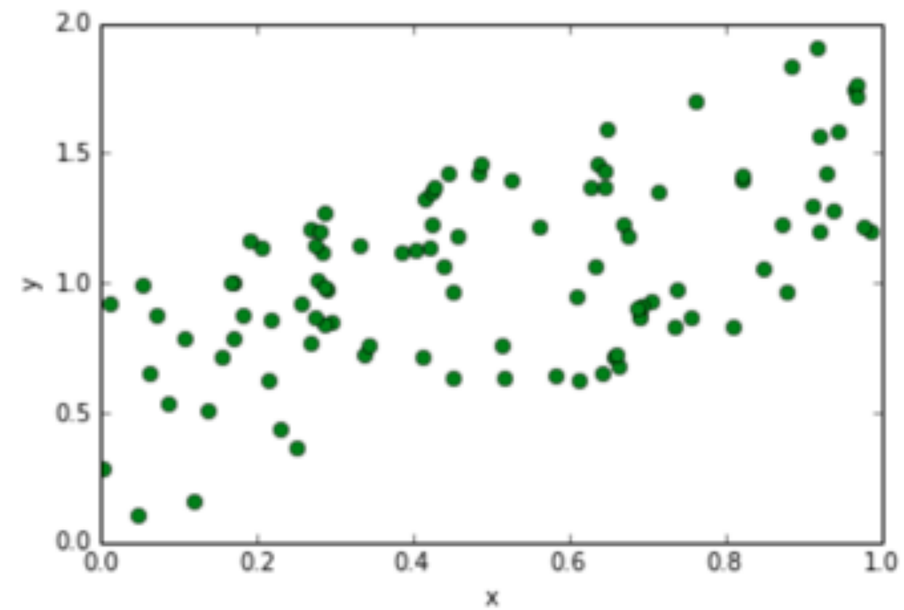
Correlation

Covariance:

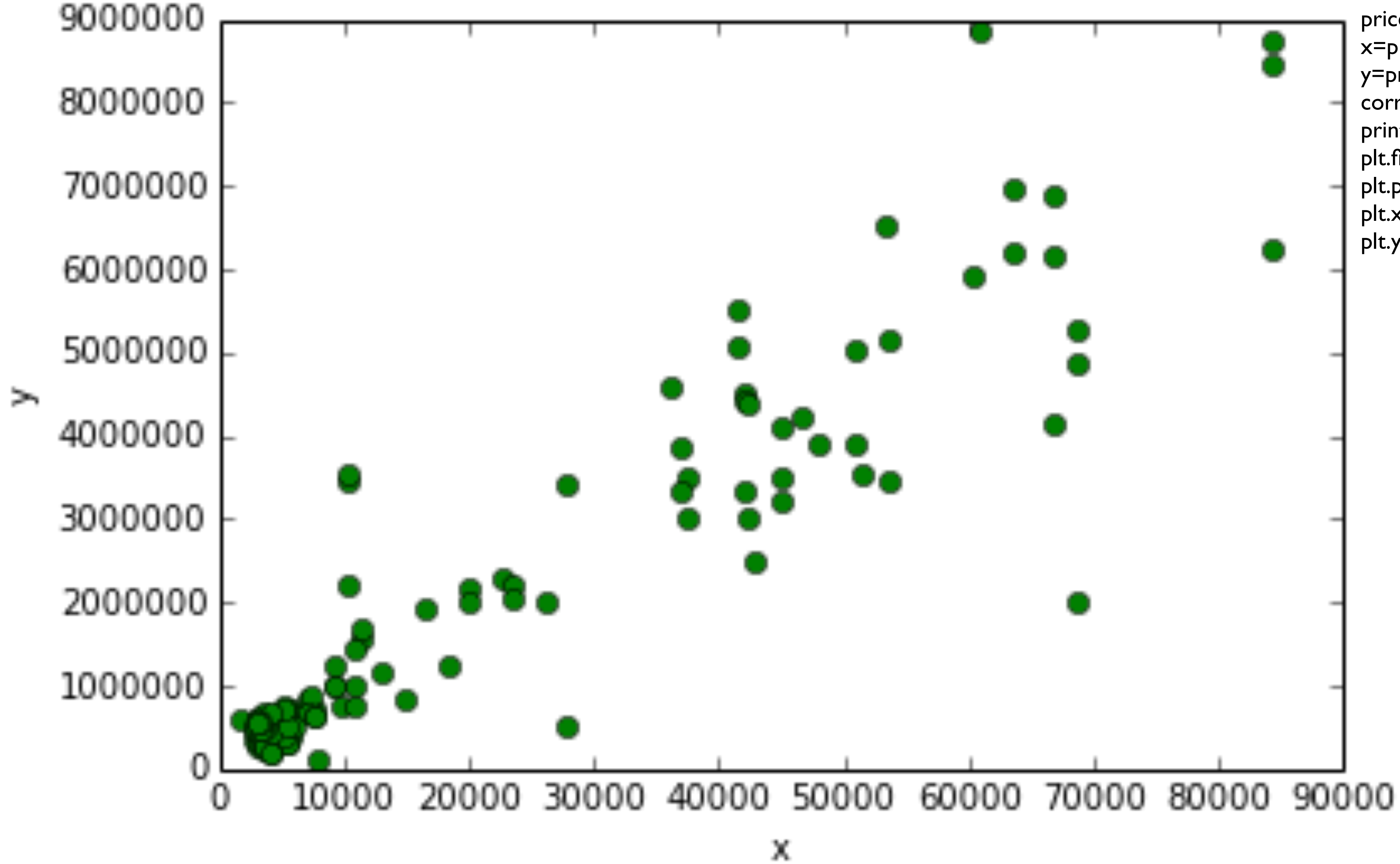
$$cov(X,Y) = E\left[(X - E[X])\,(Y - E[Y])\right]$$

Pearson's correlation coefficient:

$$corr(X,Y) = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)}$$



Correlation - house price vs size




```
import numpy as np
from scipy.stats.stats import pearsonr
prices = np.loadtxt("NYC_RE_10466_multi.csv",delimiter=",")
x=prices[:,0]
y=prices[:,1]
corr=pearsonr(x,y)[0]
print('Correlation={0}'.format(corr))
plt.figure()
plt.plot(x,y,'og')
plt.xlabel('x')
plt.ylabel('y')
```

Correlation=0.926477987l4

Linear Model - basic approach, continued

$$\hat{w}_1 = \frac{E[XY] - E[X]E[Y]}{E[X^2] - E[X]^2} = \frac{E[(X - E[X])(Y - E[Y])]}{var[X]}$$

$$\hat{w}_1 = \frac{cov(X,Y)}{var[X]} = corr(X,Y)\frac{std[Y]}{std[X]}$$

$$std[X] = std[Y] = 1: \quad \hat{w}_1 = corr(X,Y)$$

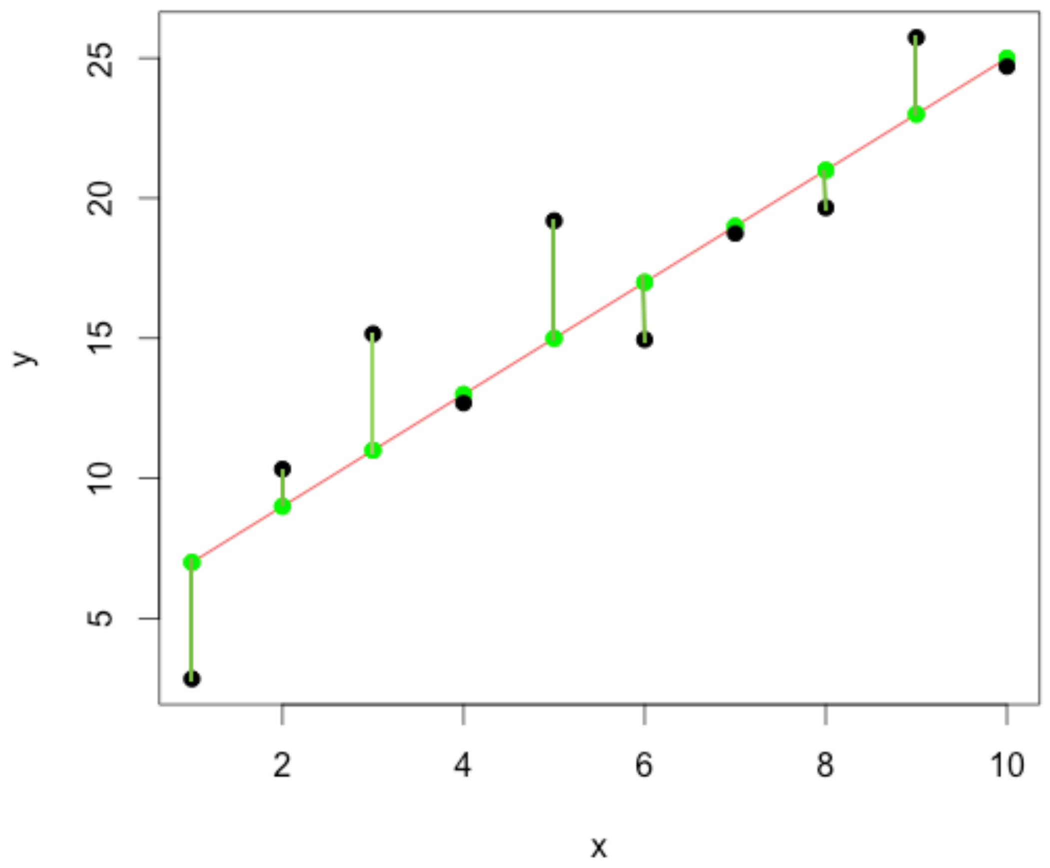
$$E[X] = E[Y] = 0: \quad \hat{w}_0 = E[Y] - \hat{w}_1 E[X] = 0$$

$$y \sim corr(X,Y)x$$

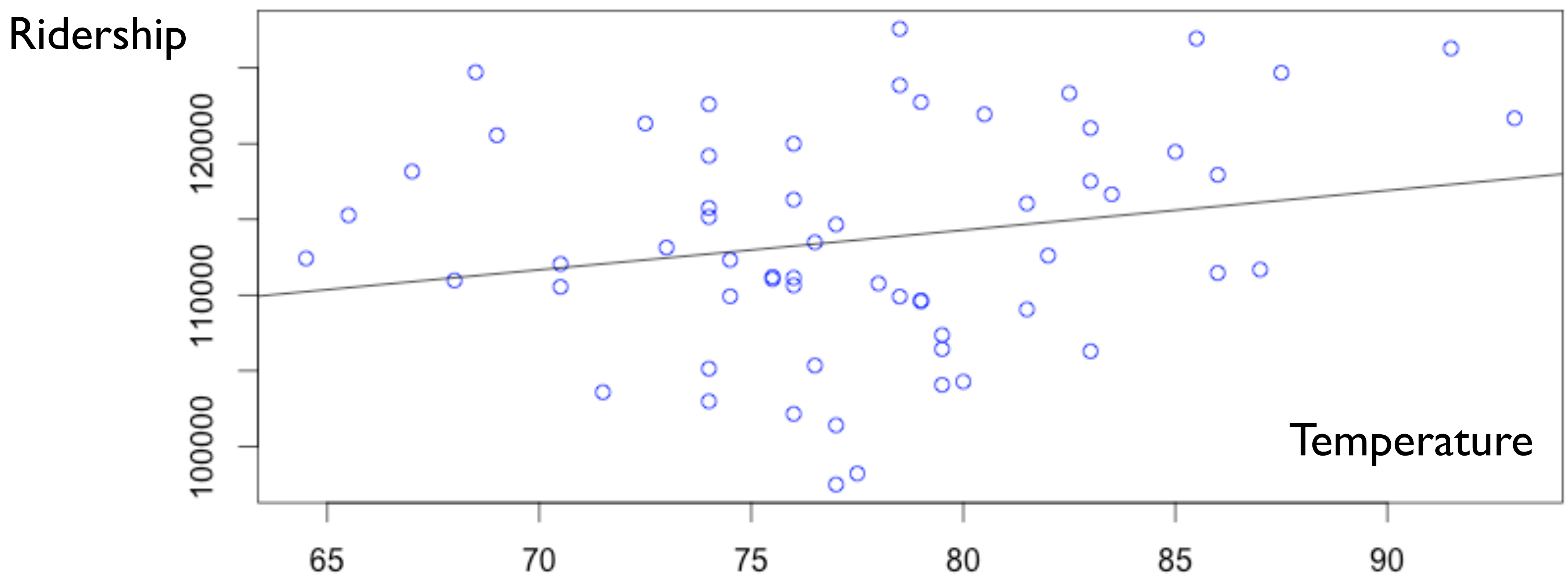
Linear Model - R-squared

$$R^2 = 1 - \frac{RSS}{\sum_i (y_i - \bar{y})^2} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2},$$

$$R^2 = corr(x, y)^2$$

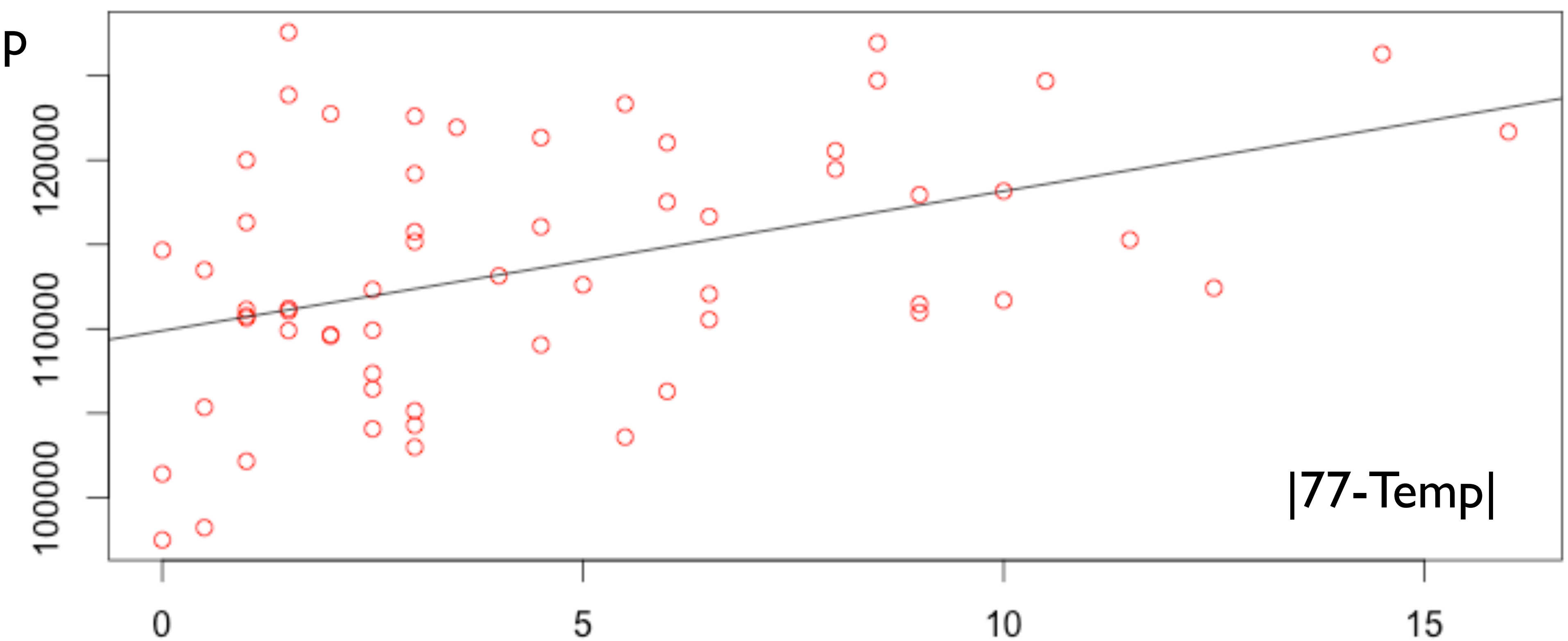


Non-linear dependence: Taxi ridership vs temperature



Correlation 21.1%

Non-linear dependence



Correlation 42.7%

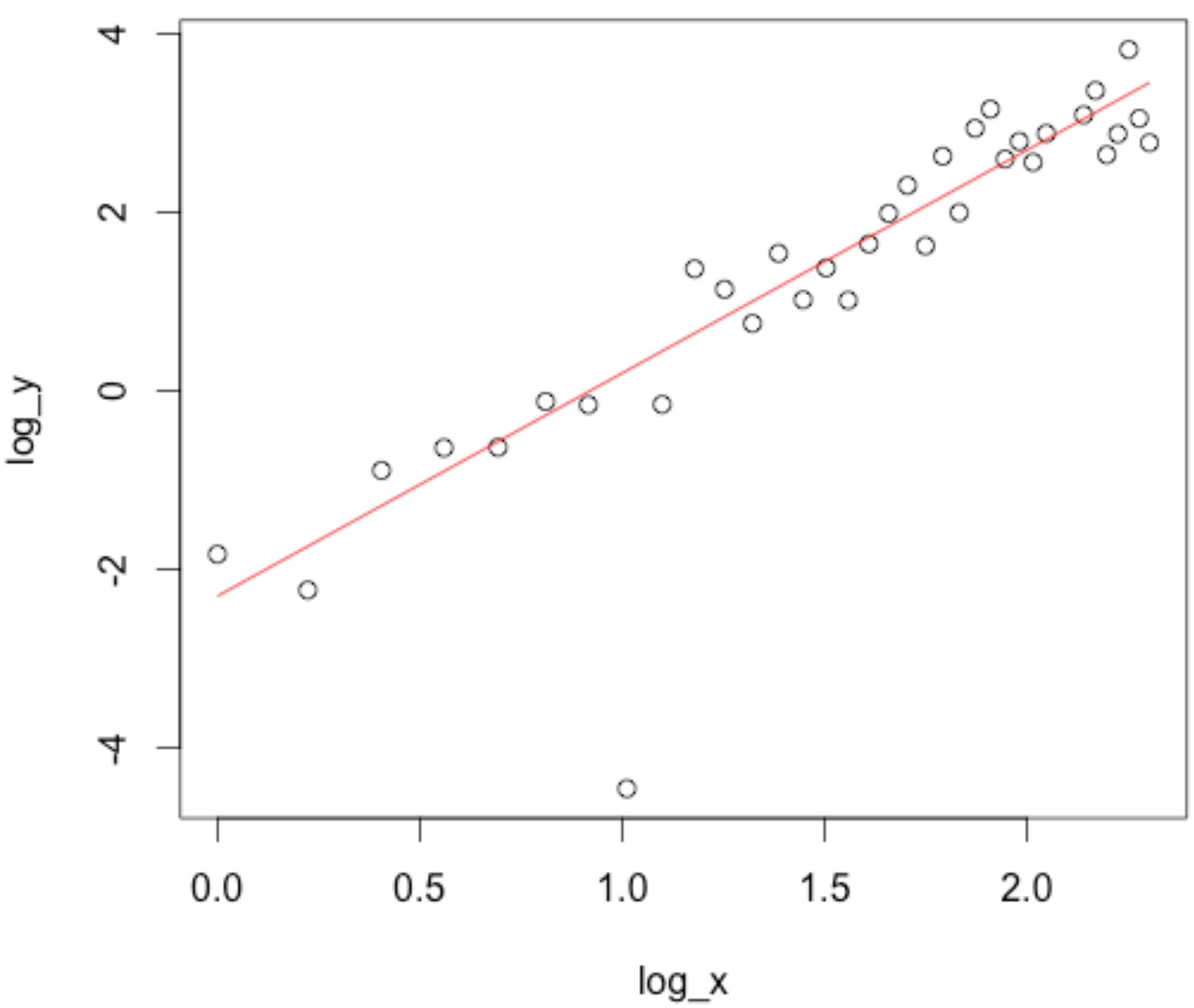
$$R \sim X, X = |77 - T|$$

Power law scaling

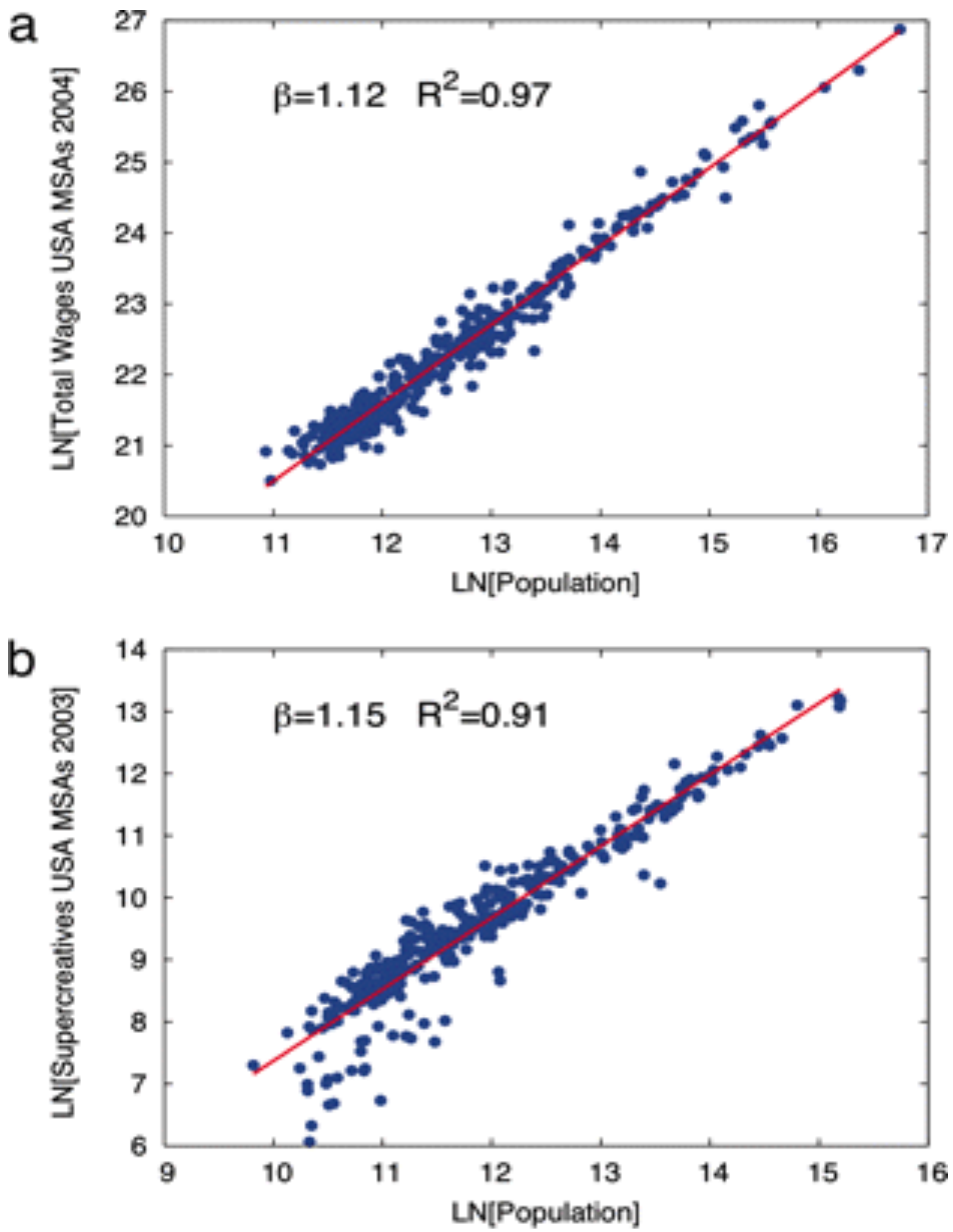
$$y \sim px^q$$

$$log(y) \sim q \cdot log(x) + log(p)$$

$$w_1 = q,\ w_0 = log(p)$$



Power law scaling

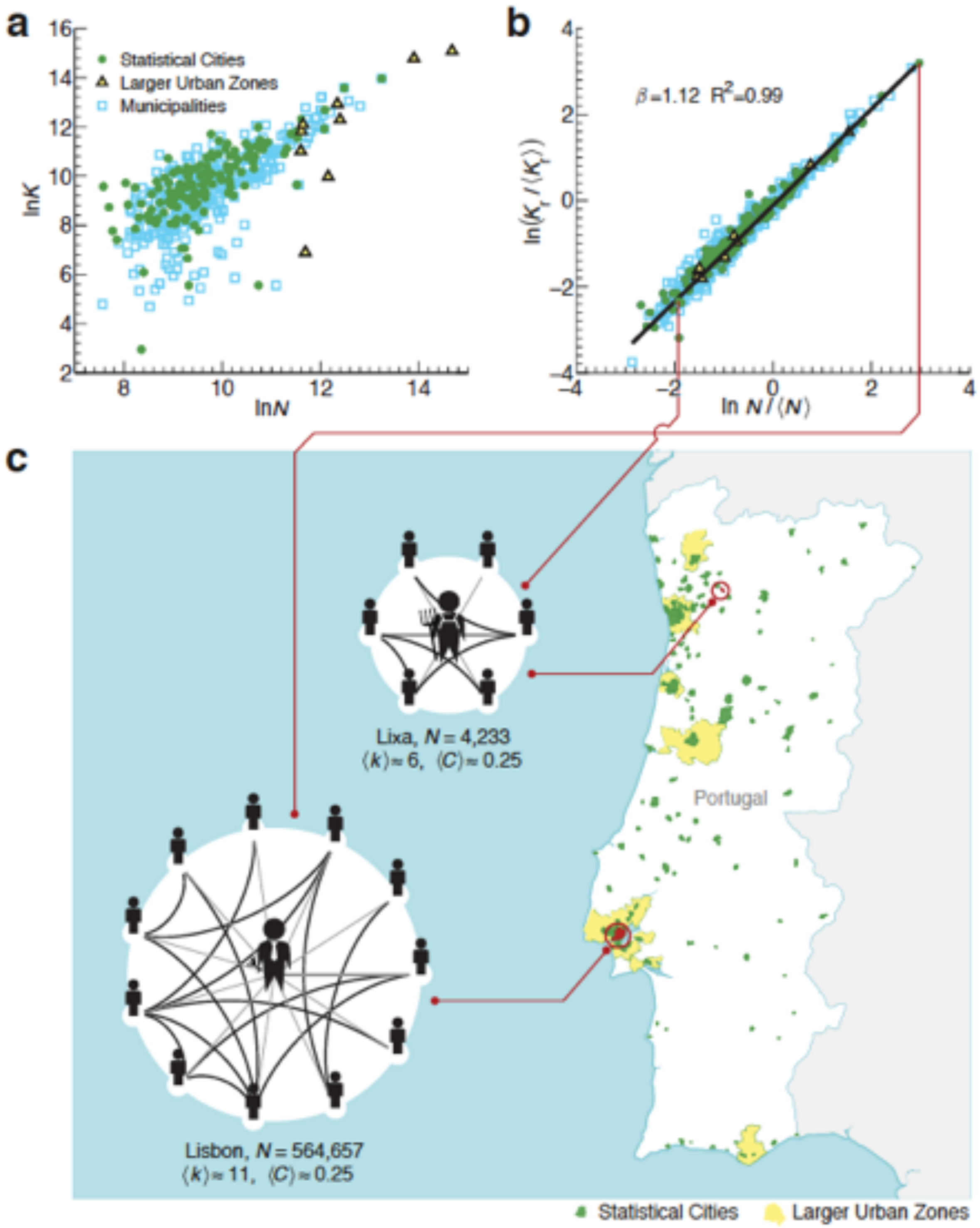


$$Wages \sim Population^{1.12}$$

$$Inventions \sim Population^{1.15}$$

Bettencourt, L. M., Lobo, J., Helbing, D., Kühnert, C., & West, G. B. (2007). Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the national academy of sciences*, *104*(17), 7301-7306.

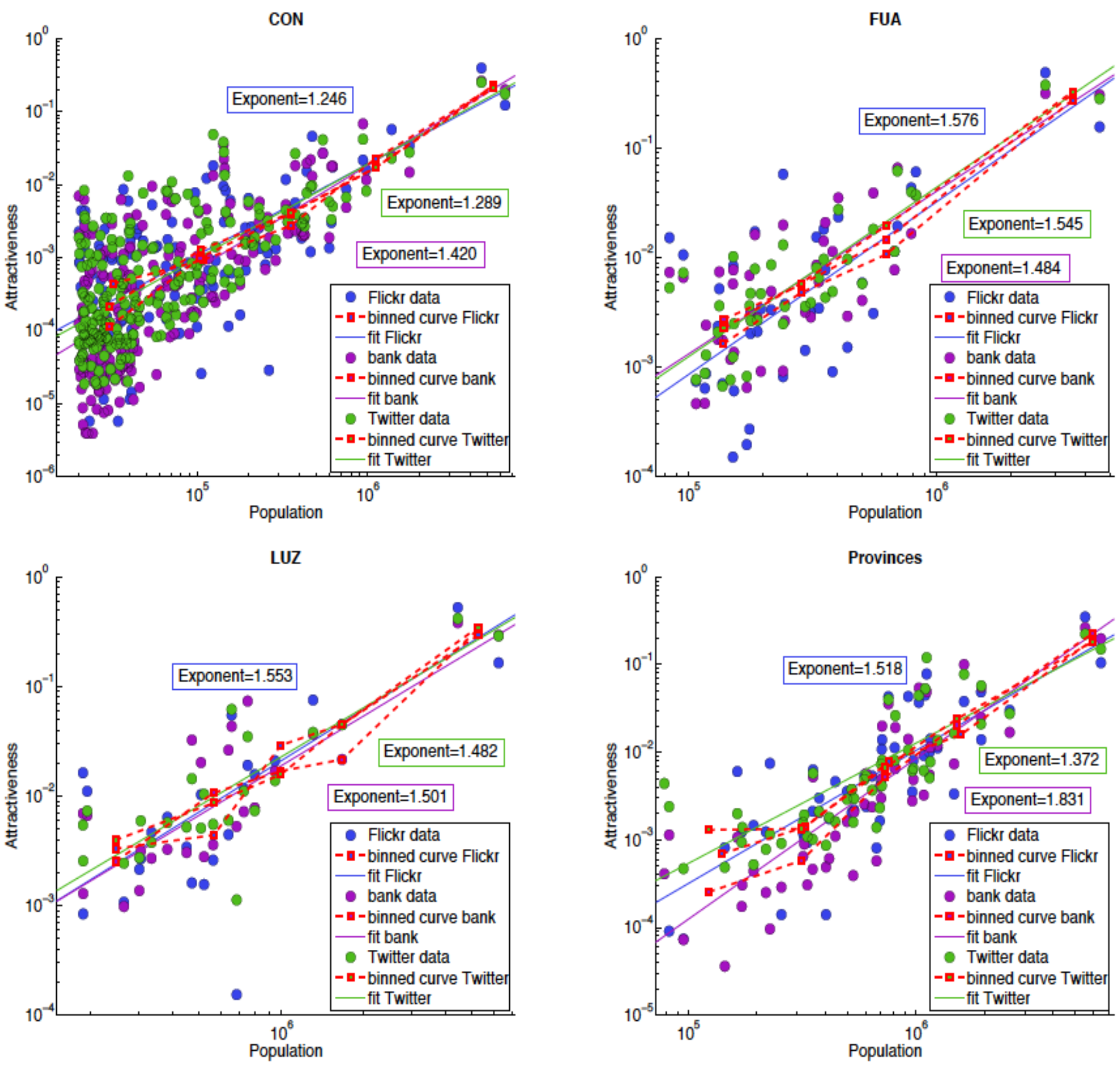
Power law scaling



$$Communication \sim Population^{1.15}$$

Schläpfer, M., Bettencourt, L. M., Grauwin, S., Raschke, M., Claxton, R., Smoreda, Z., ... & Ratti, C. (2014). The scaling of human interactions with city size. *Journal of The Royal Society Interface*, *11*(98), 20130789.

Power law scaling



$$Visitors \sim Population^{1.5}$$

Sobolevsky, S., Bojic, I., Belyi, A., Sitko, I., Hawelka, B., Arias, J. M., & Ratti, C. (2015). Scaling of city attractiveness for foreign visitors through big data of human economical and social media activity. *arXiv preprint arXiv:1504.06003*.

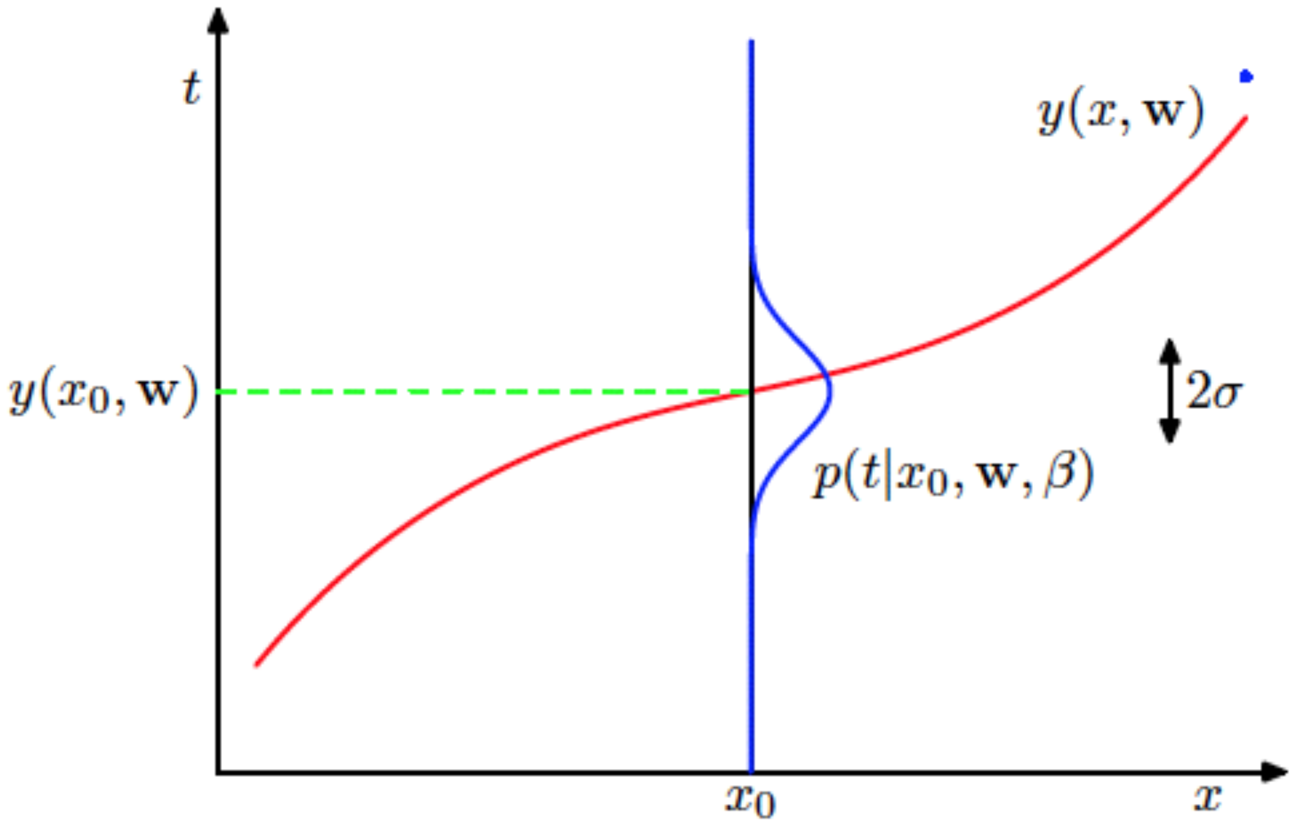
Why sum of squares or residuals?

Linear Model - probabilistic approach

$$p(y|x, w) = \mathcal{N}(y|w_1 x + w_0, \sigma^2)$$

$$y = w_1 x + w_0 + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$



Bishop, Christopher M. *Pattern recognition and machine learning*. springer, 2006.
These materials are included under the fair use exemption and are restricted from further use

Linear Model - max-likelihood

$$\prod_i p(y_i|x_i, w, \sigma) \to \max$$

$$\mathcal{N}(y|w_1 x + w_0, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - w_1 x - w_0)^2}{2\sigma^2}}$$

$$\log\left(\prod_j p(y_j|x_j, w, \sigma)\right) = \sum_j \log\left(\mathcal{N}(y|w_1 x + w_0, \sigma^2)\right) =$$

$$= -\sum_j \frac{(y_j - w_1 x_j - w_0)^2}{2\sigma^2} - N\log(\sigma) - N\log(\sqrt{2\pi}) \to \max$$

Linear Model - sigma estimation

$$\frac{RSS(\hat{w})}{2\sigma^2} + N\log(\sigma) \to \min$$

$$\frac{\partial \frac{RSS(\hat{w})}{2\sigma^2} + N\log(\sigma)}{\partial\sigma} = 0,$$

$$-\frac{RSS(\hat{w})}{\sigma^3} + \frac{N}{\sigma} = 0,$$

$$\sigma^2 = \frac{RSS(\hat{w})}{N}.$$

$$\sigma^2 = \frac{RSS(\hat{w})}{N-2}$$