Ve406 Lecture 1

Jing Liu

UM-SJTU Joint Institute

May 14, 2018

- Course Description:

  This course provides an introduction to the process and procedures of statistical modelling. We will explore real data sets, examine various models for the data, assess the validity of their assumptions, and determine which conclusions we can make, if any. In this course you will learn how to program in R and how to use R for effective data analysis.

- Learning Outcomes:

  After successful completion of this course, you should be able to

  1. Explore data graphically.
  1. Select appropriate models.
  2. Implement those models in R.
  3. Examine those models critically.
  4. Interpret the results of those models to non-statisticians.

- Who should take this class?

  The prerequisite for this class is computer/programming knowledge at the level of Ve101 (or above), and statistics knowledge at the level of Ve401 (or above). Both undergraduates and graduate students are welcome to take the course.

- **Instructor:**

  Jing Liu

- **Lectures:**

Monday	(12:10pm – 1.50pm)	in E1-108
Wednesday	(12:10pm – 1.50pm)	in E1-108
Thursday (lab)	(06:20pm – 8.40pm)	in TBA Even Weeks Only

- **Office Hours:**

Monday	(08:30am – 09:40am)	in JI-Building 441A
Monday	(02:30pm – 03:40pm)	in E2-104
Thursday	(02:30pm – 03:40pm)	in E2-104 Even Weeks Only

- **Email:**

  stephen.liu@sjtu.edu.cn

- **Teaching Assistant/s:**

  See Canvas for his/her contact information

- **Assignment:**

  20% Assignments will be given in the form of problem sets, and may require extra reading and using of R.

- **Lab:**

  20% There will be fortnightly labs.

- **Project:**

  10%

  The project will be due one week before the final exam.

Programming/Analysis	5%
Interpretation/Presentation	5%

- **Exam:**

  50% There will be two exams:

Midterm	15%	
Final	35%	Not accumulative.

- For this course, the grade will be curved to achieve a median grade of "B".

- Honesty and trust are important. Students are responsible for familiarising themselves with what is considered as a violation of honour code.
- Assignments/labs/projects are to be solved by each student individually. You are encouraged to discuss problems with other students, but you are advised not to show your written work to others. Copying someone else's work is a very serious violation of the honour code.
- Students may read resources on the Internet, such as articles on Wikipedia, Wolfram MathWorld or any other forums, but you are not allowed to post the original assignment question online and ask for answers. It is regarded as a violation of the honour code.
- Only a single sheet, A4 size with your original handwritten notes on both sides, is allowed during the written exams.
- Since it is impossible to list all conceivable instance of honour code violations, the students has the responsibility to always act in a professional manner and to seek clarification from appropriate sources if their or another student's conduct is suspected to be in conflict with the intended spirit of the honour code.

- Fox and Weisberg (2010)

  An R Companion to Applied Regression

- Some Additional Material:
  - Grolemund and Wickham (2016)

    R for Data Science
  - Weisberg. (2013)

    Applied Linear Regression
  - Fox. (2015)

    Applied Regression Analysis and Generalized Linear Models
  - James et al. (2017)

    An introduction to Statistical learning: with Applications in R

- Teaching Schedule:

Week	Topics
1	Introduction Basic Probability and Statistics
2	Simple Linear Regression Least-Squares and Maximum Likelihood Lab 1
3	Diagnostics and Transformation Inference
4	Multiple Linear Regression Diagnostics and inference Lab 2
5	Polynomial and categorical predictors, and interactions Collinearity
6	Influential points and Outliers Model selection Lab 3

7	Heteroskedasticity **First Midterm Exam**
8	Correlated Noise Splines Lab 4
9	Additive Models Logistic Regression
10	Generalised Linear Models Generalised Additive Models Lab 5
11	Principal Component Analysis Factor Analysis
12	Mixture Models Survival Analysis (Optional) Lab 6
	Final Exam

Q: What is a statistical model?

Q: What is a regression model?

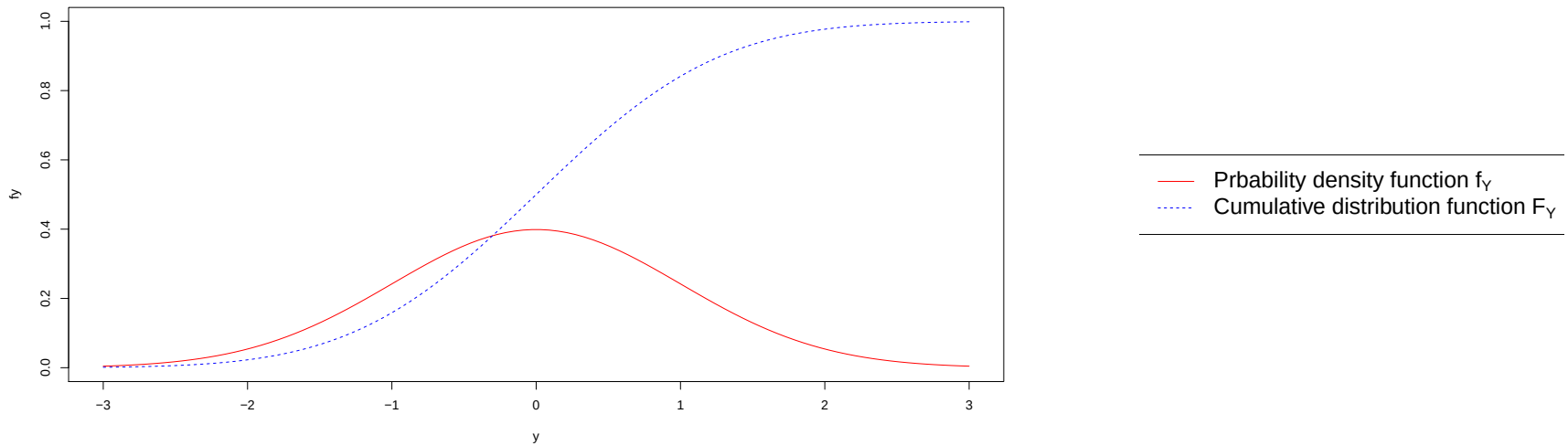
- Regression analysis is about statistical models that involve only one

  dependent variable

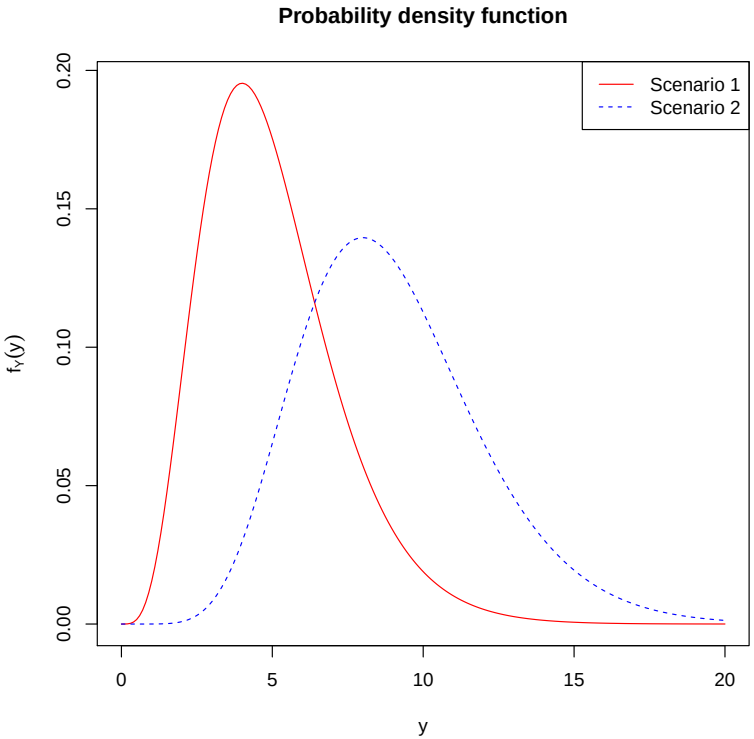
  and one or more independent variables.
- It focuses on modelling the distribution of the dependent variable

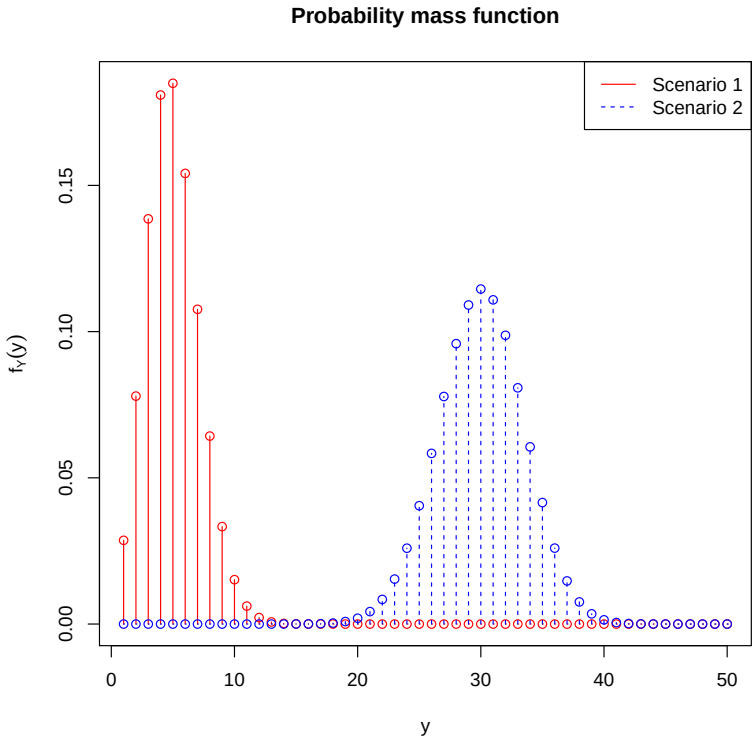
  $$f_Y(y)$$

  using the independent variables X_j.



- Depending on the values that the independent variables take,



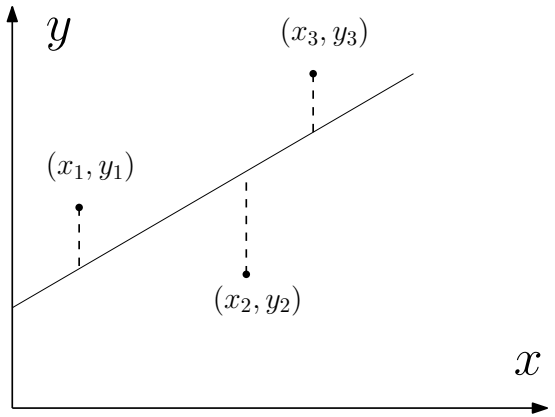


Y may follow a different distribution. e.g.

the distribution of height depends on gender

- Roughly speaking, regression is about finding a rule of picking distributions for Y from a space of infinitely many distributions that agrees with the data.

- Notice this view of regression models is very different from OLS.



$$\underset{\alpha,\beta}{arg\,min} \sum_{i=1}^{3} \varepsilon_i^2$$

where

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- You will see that the two views are equivalent, but one is easier to generalise.
- Regression models usually attempt to address one of two questions:
  - Explaining the dependencies between

$$\underbrace{\text{the explanatory variables}}_{\text{the independent variables}} \quad \text{and} \quad \underbrace{\text{the response}}_{\text{the dependent variable}}$$

  - Predicting the range of response values given a set of values for

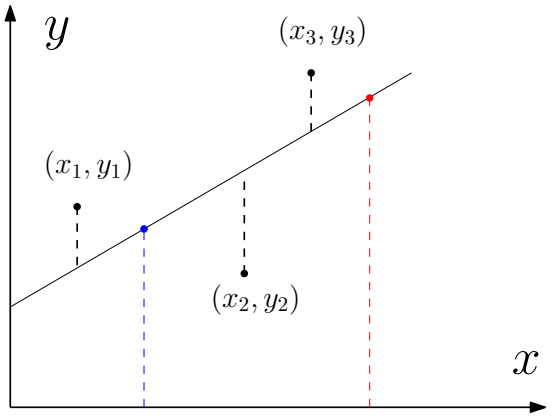
$$\underbrace{\text{the predictors}}_{\text{the independent variables}}$$

- In terms of prediction, the linear model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

  is very nature when there is only one predictor.

- It is particularly easy to understand and implement using OLS.



$$\underset{\alpha,\beta}{arg\,min} \sum_{i=1}^{3} \varepsilon_i^2$$

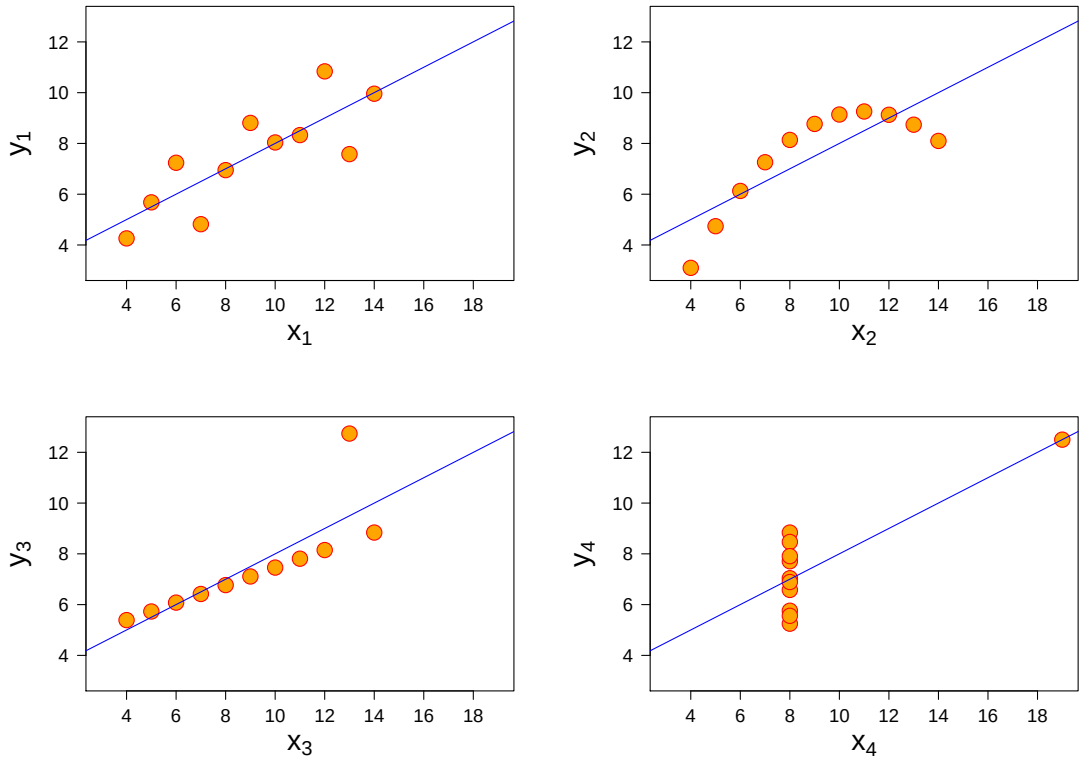
where

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

- It is even not hard to imagine how one would proceed with finding

$$f_Y$$

Q: But why the linear model? What is exactly complicated with nonlinearity?



- Note the least squares principle can also be used for

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \qquad \text{or} \qquad y_i = \frac{\beta_1 x_i}{\beta_2 + x_i} + \varepsilon_i$$

- Consider a random variable Y that follows some *known* distribution.

$$f_Y$$

Q: Suppose we would like to predict the value of Y. What is the "best" guess?

- Let us denote our guess by

$$m$$

Q: What is a sensible and common way to measure the how good m is ?

- If we don't care about positive more than negative errors, then

$$\mathrm{MSE}(m) = \mathbb{E}\left[(Y-m)^2\right]$$

is the traditional choice, which can be decomposed into

$$\mathrm{MSE}(m) = \mathrm{Var}[Y] + (\mathbb{E}[Y-m])^2$$

Q: Can you recall why the above is true and what does it mean?

Q: What is bias in statistics?

$$\text{Bias}\left(\hat{\theta}, \theta\right) = \mathbb{E}\left[\hat{\theta} - \theta\right]$$

where $\hat{\theta}$ is often an estimator, and θ is a population parameter, thus fixed.

- So we have the simplest form of the bias-variance decomposition,

$$\text{MSE}(m) = \text{Var}[Y] + (\mathbb{E}[Y - m])^2 = \text{Var}[Y] + (\mathbb{E}[Y] - m)^2$$

which confirms the expected value is the best single number for predicting Y

$$m = \mathbb{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y)\, dy$$

- Note changing our prediction, m, does nothing to the true distribution of Y.

- Now imagine we have two random variables, say

$$X \quad \text{and} \quad Y$$

- Suppose we know X in the sense we also know

$$f_X,$$

and we would like to use this knowledge to improve our prediction of Y.

Q: What is the "best" guess m now?

- It is clear that our guess should be a function of x,

$$m(x)$$

- Again we would like to make MSE as small as possible

$$\mathbb{E}\left[\big(Y - m(X)\big)^2\right]$$

- We use conditional expectations to reduce

$$\mathbb{E}\left[\left(Y - m(X)\right)^2\right]$$

back to the problem we already solved

$$\mathbb{E}\left[\left(Y - m(X)\right)^2\right] = \mathbb{E}_X\left[\mathbb{E}_Y\left[\left(Y - m(X)\right)^2 \mid X\right]\right]$$

- For each possible value x, the best value is just the conditional mean

$$m(x) = \mu(x) = \mathbb{E}\left[Y \mid X = x\right]$$

which is known as the true regression function.

- Now we have reached the part that requires us to use a linear approximation, it is one of two things, that are too complicated, and need to be modelled.

- We restrict the regression function to have the linear form

$$m(x) = \beta_0 + \beta_1 x$$

and find the choice of β_0 and β_1 so that it is the best under this restriction.

- This is achieved by simply minimise MSE with respect to β_0 and β_1

$$\mathrm{MSE}(\beta_0, \beta_1) = \mathbb{E}\left[\left(Y - (\beta_0 + \beta_1 X)\right)^2\right]$$

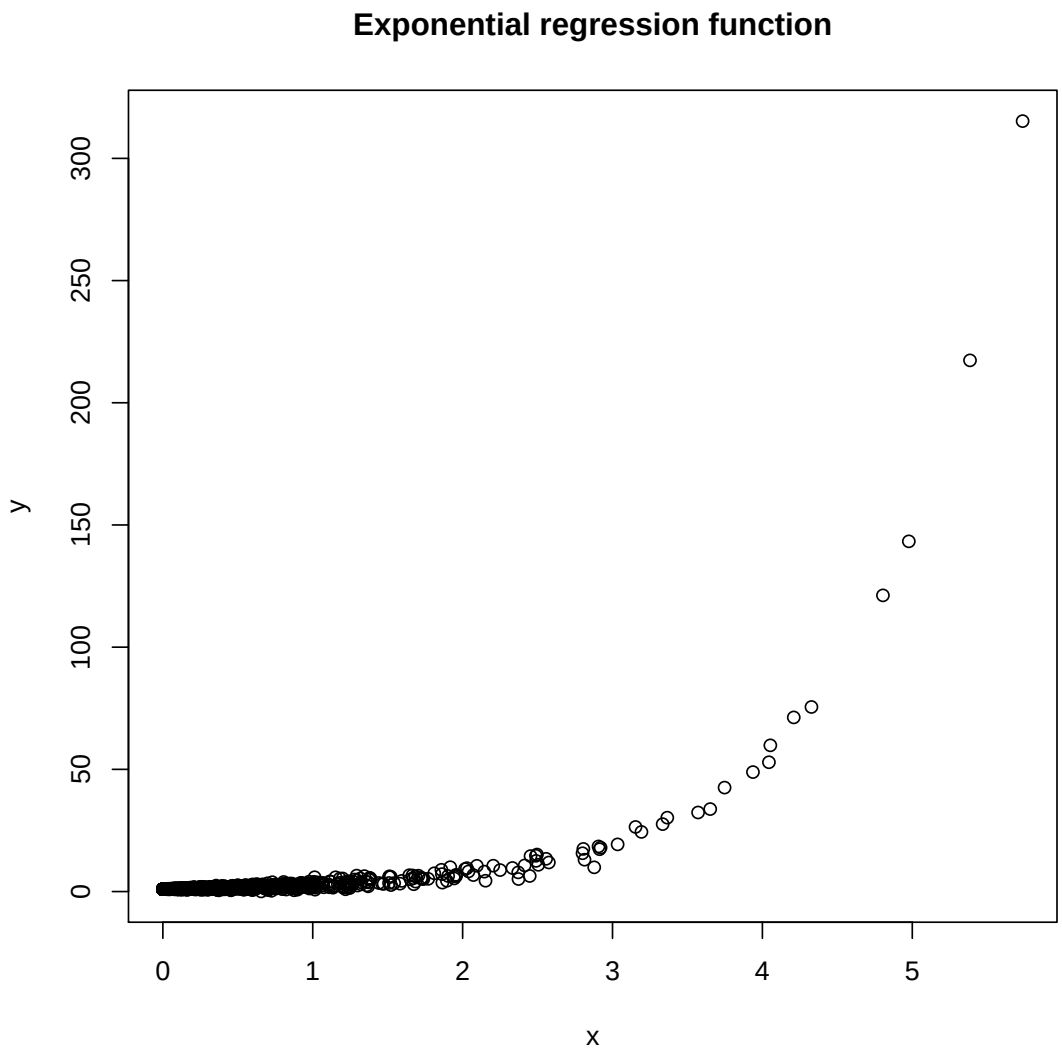
Q: Can you see why the following choice is the best?

$$\beta_0 = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X] \qquad \text{where} \quad \beta_1 = \frac{\mathrm{Cov}[X, Y]}{\mathrm{Var}[X]}$$

- Notice means play no role in β_1, only the variance and covariance matter.

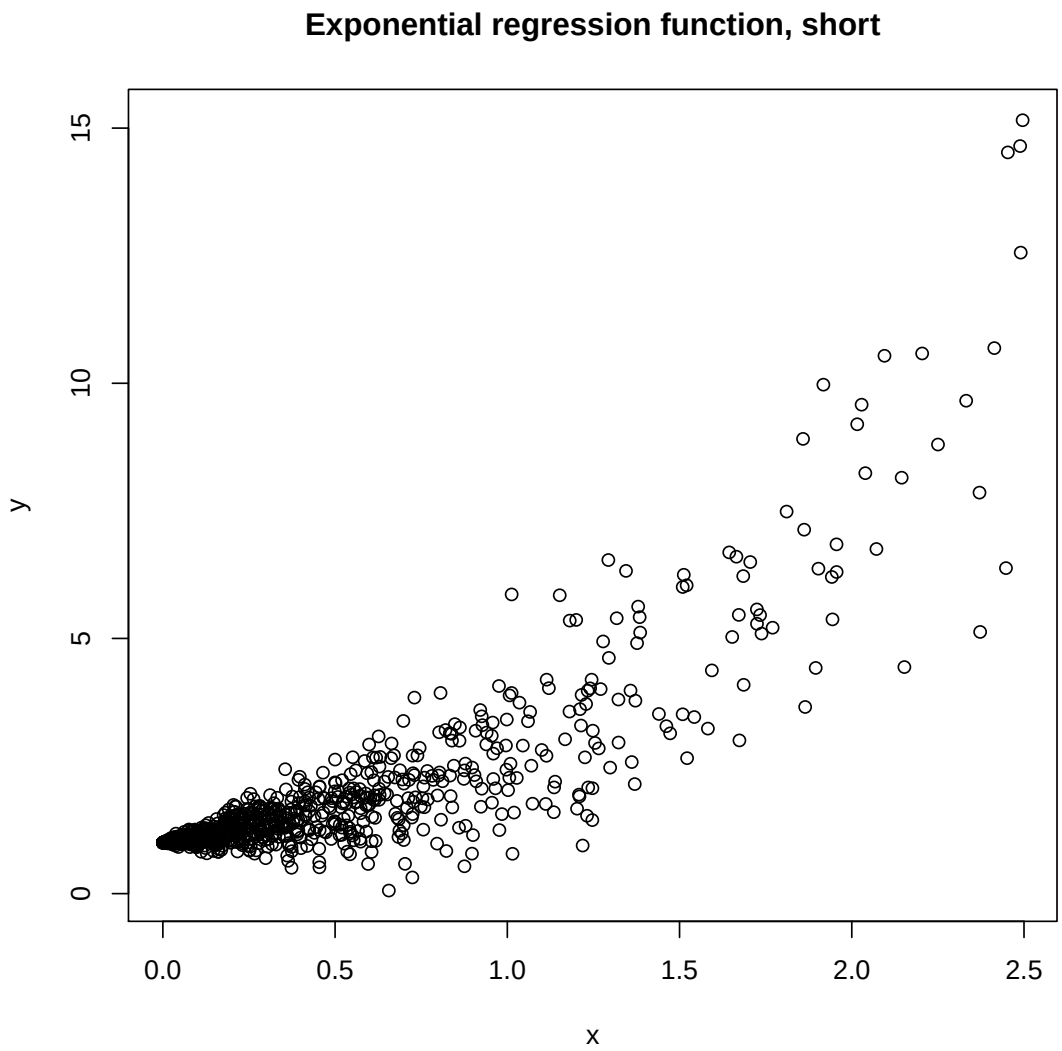
- It is the best amongst linear predictors, but it can be far from the truth. e.g.

$$\mathbb{E}\left[Y \mid X\right] = e^x$$



- We did not assume the relationship between X and Y is linear, but

$$\mu(x) = \mathbb{E}[Y \mid X] = \mu(x_0) + (x - x_0)\mu'(x_0) + \frac{1}{2} + (x - x_0)^2\mu''(x_0) + \cdots$$



- Recall regression is about finding a rule of picking distributions for Y from a space of infinitely many distributions that agrees with the data.
- If we know everything about X and Y, then we have argued "best" guess is

$$m = \beta_0 + \beta_1 x$$

where

$$\beta_0 = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X] \qquad \text{and} \qquad \beta_1 = \frac{\operatorname{Cov}[X, Y]}{\operatorname{Var}[X]}$$

Q: What do we do when we don't have f_X and f_Y?

- It is natural to use sample values to create estimators

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{and} \qquad \hat{\beta}_1 = \frac{c_{xy}}{s_x^2}$$

Q: What do you notice?