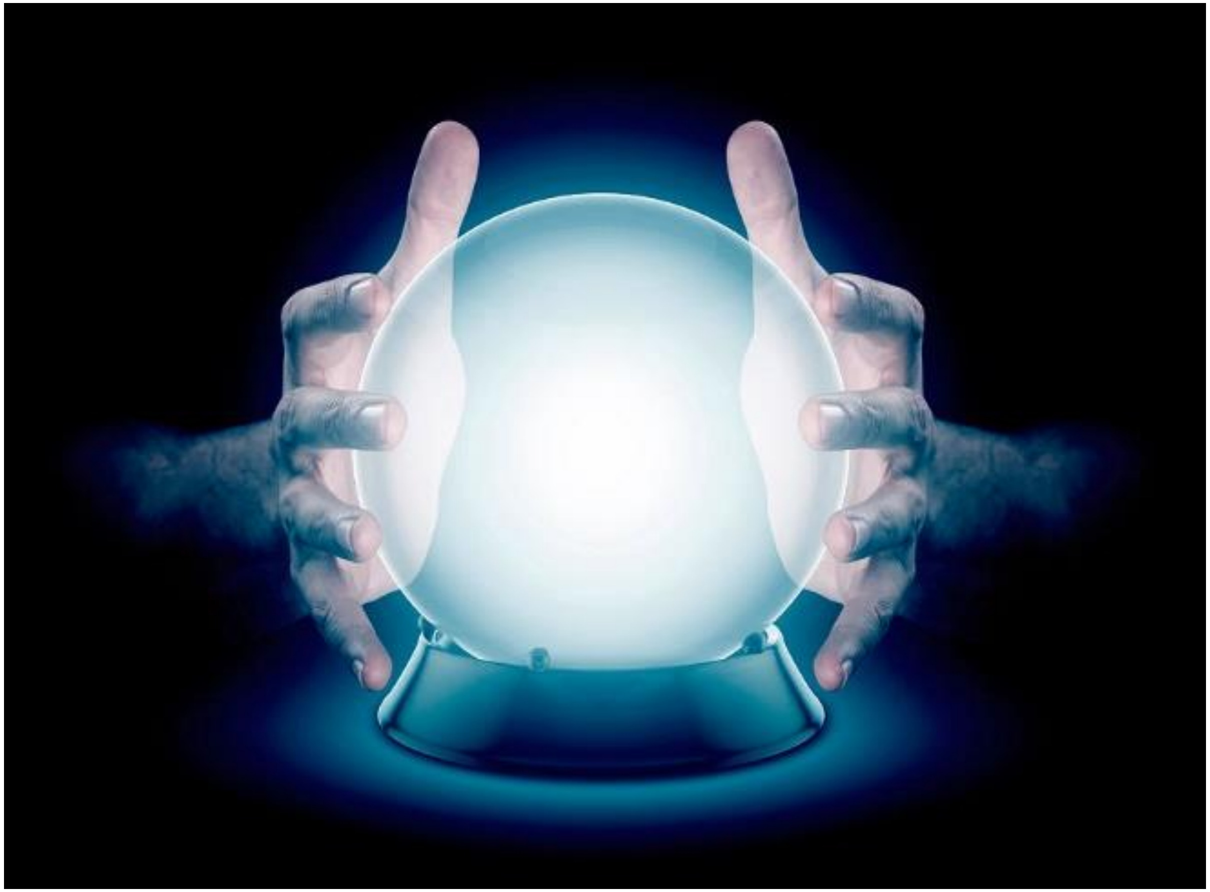
Predictive Modelling I



SML201: Introduction to Data Science, Spring 2019

Michael Guerzhoy

Linear regression

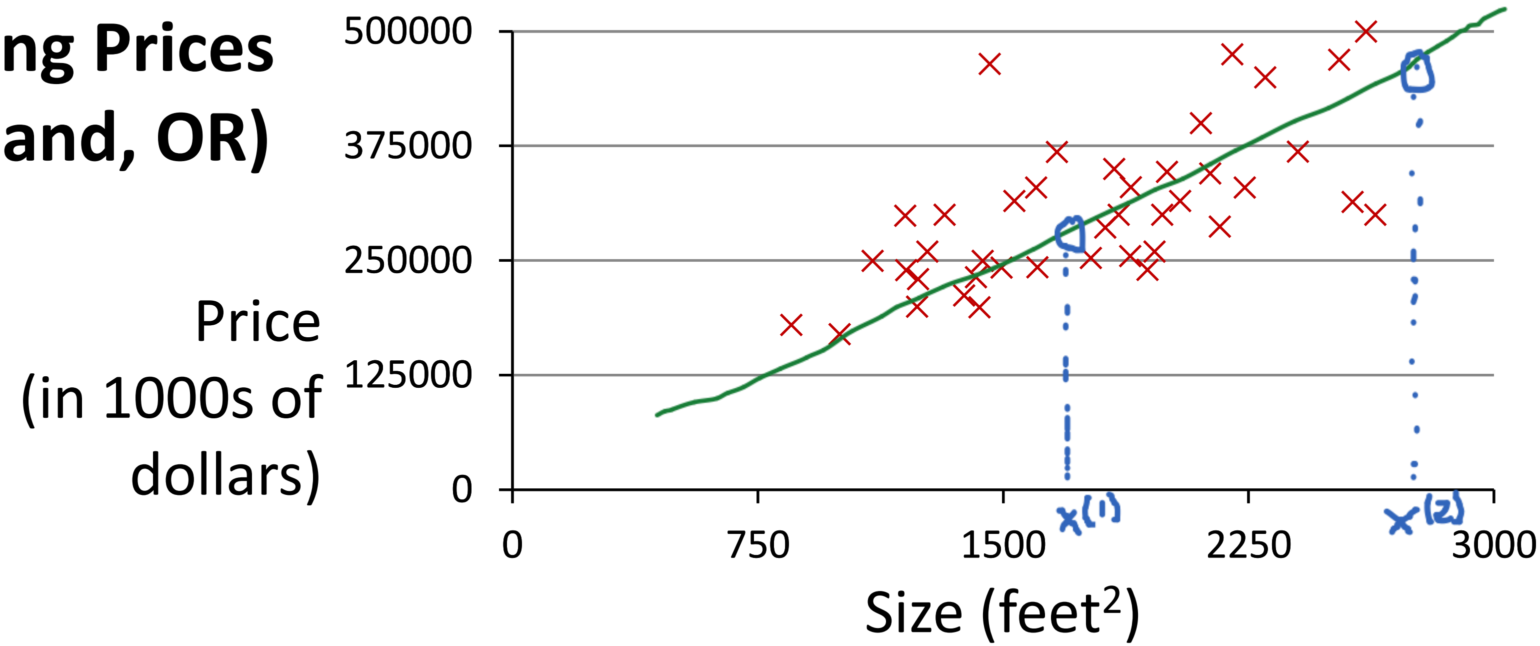
Training set of housing prices (Portland, OR)

Size in feet^2 (x)	**Price ($) in 1000's** (y)
$x^{(1)} = 2104$	$y^{(1)} = 460$
$x^{(2)} = 1416$	$y^{(2)} = 232$
$x^{(3)} = 1534$	$y^{(2)} = 315$
$x^{(4)} = 852$	$y^{(2)} = 178$
…	…

Notation:

m = Number of training examples
x's = "input" variable / features
y's = "output" variable / "target" variable

Housing Prices (Portland, OR)



- Equation for the "best" green line: price $= a_0 + a_1 size$
- Prediction for a new size x: $price(x) = a_0 + a_1 x$
- We take "best" to mean that on average, our predictions are not far off

Cost functions

- If the correct price is $y^{(i)}$ and we predicted $\left(a_0 + a_1 x^{(i)}\right)$, we are off by

$$|y^{(i)} - \left(a_o + a_1 x^{(i)}\right)| \qquad \text{("error")}$$

- If we square this quantity, we still have a measure of how far off we are

$$\left(y^{(i)} - \left(a_o + a_1 x^{(i)}\right)\right)^2 \text{ ("squared error")}$$

- Overall, we will be off (in terms of squared errors) by

$$\sum_{i=1}^{m} \left(y^{(i)} - \left(a_o + a_1 x^{(i)}\right)\right)^2$$

Simple Linear Regression

- Find the a_0 and a_1 such that

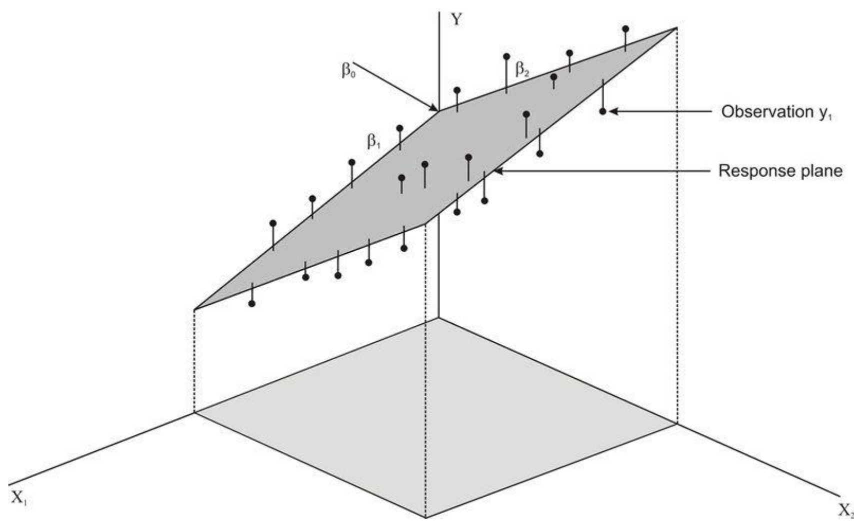
$\sum_{i=1}^{m}\left(y^{(i)} - \left(a_o + a_1 x^{(i)}\right)\right)^2$ is as small as possible

- Graphically, roughly corresponds to "draw the best line through the scatterplot"

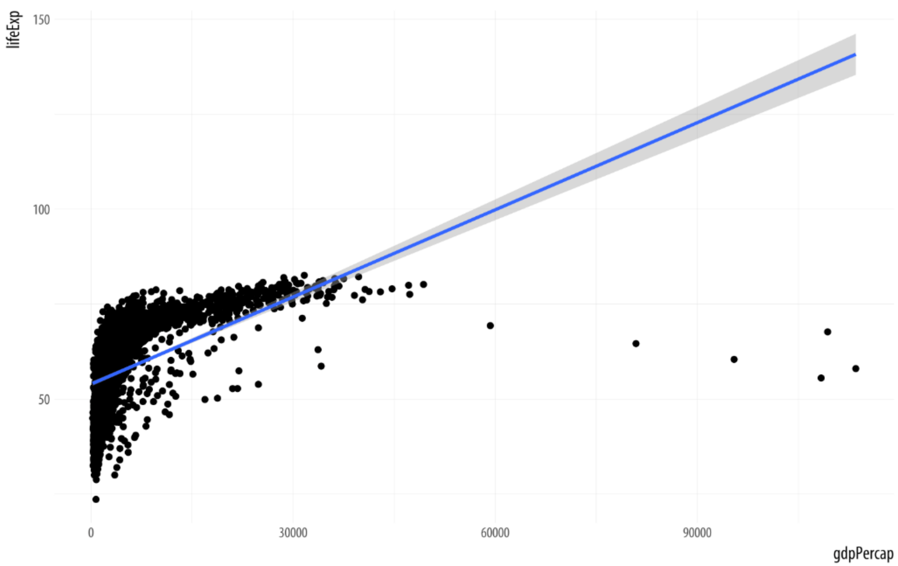
Multiple Linear Regression

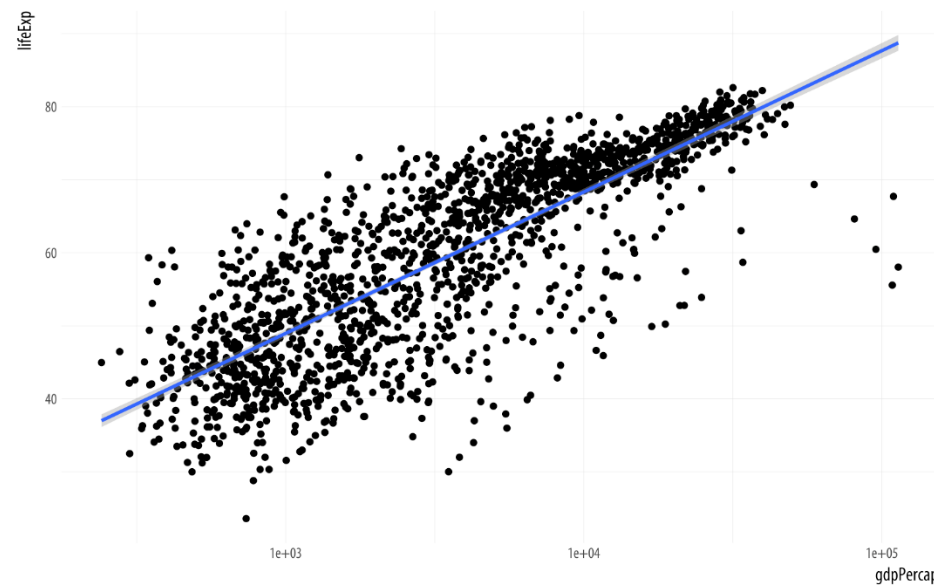
- *n* quantities per case
- Find $a_0, a_1, \dots, a_n$ such that

$$\sum_{i=1}^{m}\left(y^{(i)} - \left(a_o + a_1 x_1^{(i)} + a_2 x_2^{(i)} + \cdots + a_n x_n^{(i)}\right)\right)^2$$ is small



Transforming inputs





Seems better to predict lifeExp using a new variable, $\log(gdpPercap)$

Content from Kieran Healy, *Data Visualization*

(switch to R)

Categorical variables

- Continuous variables are numbers
  - Number, size, or weight of something
  - If x = 19 and x = 21 make sense, like x = 20 would make sense too
- Categorical variables indicate categories
  - Country, Continent, …
  - Cannot put continents on a single scale

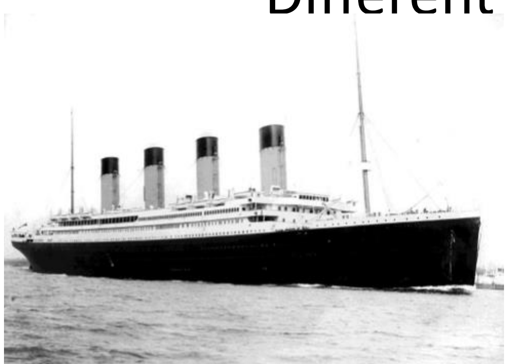
Predicting with categorical variables

- Suppose we are trying to predict y using one categorical variable (e.g., continent) which has k possible categories (e.g., Cont1, Cont2, …, Cont5)
- $y^{(i)} \approx a_{1,0} + a_{1,1}I_{i,1} + a_{1,2}I_{i,2} + \cdots + a_{1,k-1}I_{i,k-1}$
  - $I_{i,1} = \begin{cases} 1, if\ the\ i-th\ row\ contains\ Cont1 \\ 0, otherwise \end{cases}$
  - $I_{i,2} = \begin{cases} 1, if\ the\ i-th\ row\ contains\ Cont2 \\ 0, otherwise \end{cases}$
  - …
- Note: if the i-th point is Contk, then the prediction is a_0
  - Potentially different predictions for each continent, despite the fact that we didn't include $I_{i,k}$

(Switch to R)

Titanic Survival Case Study

- The RMS *Titanic*
  - British passenger liner
  - Collided with an iceberg during her maiden voyage
  - 2224 people aboard, 710 survived
- People on board
  - 1st class, 2nd class, 3rd class passengers (price of ticker + social class played a role)
  - Different ages, genders



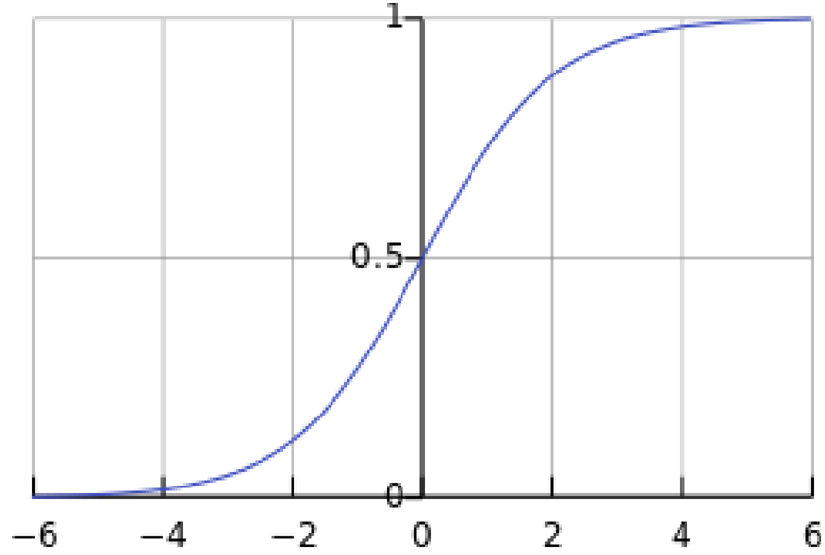
Predicting Survival

- Trying to predict a categorical variable (died/ survived)
- Convert (arbitrarily) "died" to 0 and "survived" to 1
- But $a_o + a_1 x_1^{(i)} + a_2 x_2^{(i)} + \cdots + a_n x_n^{(i)}$ could be any real number
- Solution: compute

$$p^{(i)} = \sigma(a_o + a_1 x_1^{(i)} + a_2 x_2^{(i)} + \cdots + a_n x_n^{(i)})$$

Logistic function

- $\sigma(y) = \frac{1}{1+\exp(-y)}$



Inputs can be in $(-\infty, \infty)$, outputs will always be in $(0, 1)$

Logistic regression: prediction

$$p^{(i)} = \sigma(a_o + a_1 x_1^{(i)} + a_2 x_2^{(i)} + \cdots + a_n x_n^{(i)})$$

- $0 < p^{(i)} < 1$
- Interpret $p^{(i)}$ as the probability that the variable that we are predicting (i.e., $y^{(i)}$) is 1
  - For now, think of a probability is a number between 0 and 1 where 0 indicates that the event will not happen and 1 indicates that the event will happen.

(Switch to R)

Logistic regression: cost function

- For linear regression, we had
$$\sum_{i=1}^{m}\left(y^{(i)}-\left(a_o+a_1x_1^{(i)}+a_2x_2^{(i)}+\cdots+a_nx_n^{(i)}\right)\right)^2=\sum_{i=1}^{m}\left(y^{(i)}-pred(x^{(i)})\right)^2$$
- The cost is small if the predictions are close to the actual y's

- For logistic regression:
$$-\sum_{i}(y^{(i)}\log p^{(i)}+\left(1-y^{(i)}\right)\log(1-p^{(i)}))$$
- Idea: the cost is small if the p's are close to the y's
- Won't go into detail here