Supervised Learning

Chapter IV: Linear Regression

Johannes Jurgovsky

Outline

Linear Regression

1. Motivation
2. Simple Linear Regression
3. Multiple Linear Regression
4. Polynomial Linear Regression

1. Motivation

Motivation:

What is Regression Analysis?

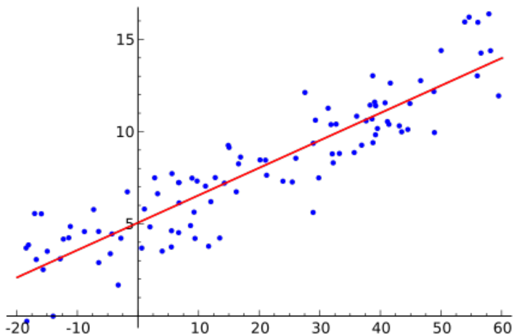
Regression Analysis: Statistical process for estimating the relationships among variables

- ❑ relationship between a dependent variable & one or more independent variables ('predictors')
- ❑ how does the value of the dependent variable change when one of the independent variables is varied (while the other independent variables are held fixed)

Regression Function: Function of the independent variables.

Usage:

- ❑ Prediction / Forecasting of dependent vairable given the independent variables
- ❑ Quantify strength of the relationship between the dependent variable and the independent variables



Source: `wikipedia.org`

Motivation:
Regression Models

Regression models relate Y to a function of X and β: $Y \approx f(X, \beta)$ with

- dependent (endogenous)[1] variable Y
- independent (exogenous) variables X
- unknown *parameters* β, a vector of size k (also called *regression coefficients*).

Form of the function f must be specified:

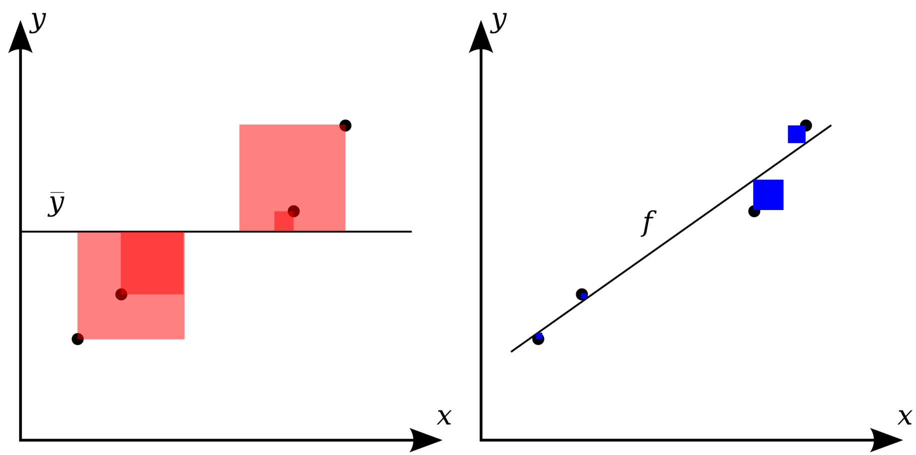
- Linear Regression
- Polynomial Regression

[1] Naming convention in econometrics.

Motivation:
Regression Models

Typical setup: n data points of the form (Y, X) observed; $n >> k$.

- enough information to estimate a unique value for β that best fits the data in some sense
- regression model can be viewed as an overdetermined system in β
- regression analysis finds a solution for unknown parameters β that will e.g. minimize the distance between the measured and the predicted values of the dependent variable Y ($=$ method of least squares).



Source: Wikipedia.org

2. Simple Linear Regression

Simple Linear Regression:

Simple Linear Regression Model

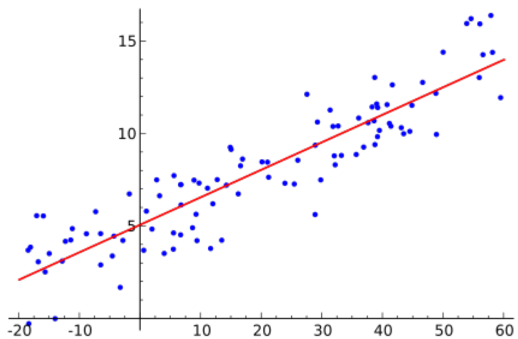
Simple: Y and X are scalar values $\rightarrow n$ data points of the form (x_i, y_i) are observed

Linear: linear relationship between dependent and independent variables:

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

- Stochastic **error term** ϵ_i: Captures all factors influencing the dependent variable y_i other than x_i (also called *disturbance term* or *noise*).

- Graphically: draw a line through the data points with
  - $f(x) = \beta_0 + \beta_1 x$
  - $\beta_0 =$ y-intercept
  - $\beta_1 =$ slope
- Assumption (commonly): ϵ_i follow gaussian distribution with mean 0 and identical variance $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$



Source: wikipedia.org

Simple Linear Regression:

Estimating regression coefficients (parameter vector)

Ordinary least squares (OLS)

- Simplest (and most common) estimator
- Minimizes the sum of squared errors

$$\underset{\beta_0,\beta_1}{\text{argmin}} \sum_{i=1}^{n} \epsilon_i^2 = \underset{\beta_0,\beta_1}{\text{argmin}} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

- Calculating the partial derivatives and setting them to 0 leads to

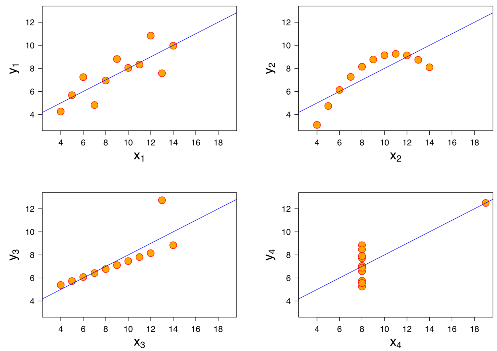
$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \textit{and} \qquad b_0 = \bar{y} - b_1\bar{x}$$

  , where b_0 and b_1 are estimates for β_0 and β_1 from our sample.
  Notice: $b_1 = \frac{Cov(X,Y)}{Var(X)}$
- Many other estimators exist (especially important if the assumptions regarding the error are not met).

Simple Linear Regression:

Coefficient of Determination r^2 - Goodness of fit (1)

- Different data sets can lead to the same linear regression model, e.g. Anscombe's quartet
- Measure for the *Goodness of fit* needed
- **Coefficient of determination** (R^2 or r^2): Indicates the proportion of the variance in the dependent variable that is predictable from the independent variable(s)



Source: wikipedia.org
Anscombe's quartet: Four datasets that have nearly identical simple descriptive statistics.

- Caveats: r^2 does not indicate whether
  - the independent variables are a cause of the changes in the dependent variable
  - the correct regression was used
  - the most appropriate set of independent variables has been chosen
  - the model might be improved by using transformed versions of the existing set of independent variables
  - there are enough data points to make a solid conclusion

Simple Linear Regression:

Coefficient of Determination r^2 - Goodness of fit (2)

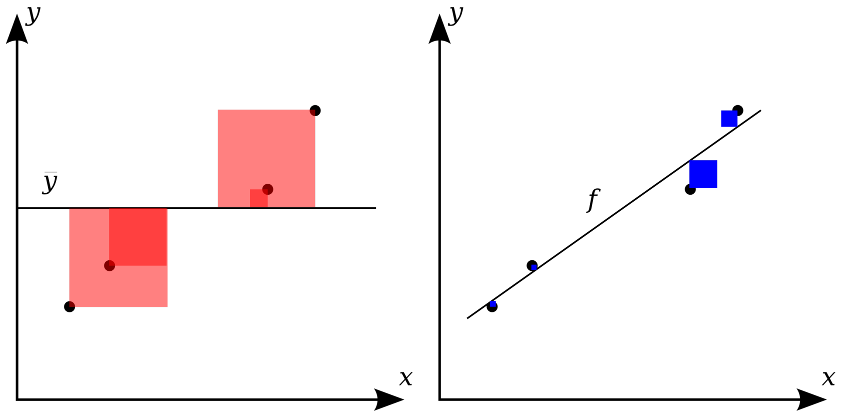
- **Total sum of squares** (proportional to the variance of the data)
  $SS_{tot} = \sum_i (y_i - \bar{y})^2$
- **Explained sum of squares**
  $SS_{reg} = \sum_i (f(x_i) - \bar{y})^2$
- **Sum of squares of residuals** / residual sum of squares
  $SS_{res} = \sum_i (y_i - f(x_i))^2$

- **Coefficient of determination:** $r^2 = 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{reg}}{SS_{tot}}$

$\rightarrow$ Ratio of the explained variance (which is SS_{reg}/n) to the total variance (which is SS_{tot}/n)

$\rightarrow$ Ranges from 0 to 1

$\rightarrow$ $r^2 = 1$: The regression line fits the data perfectly.



Source: Wikipedia.org

Exercise

The Chocolate Bar Company

Source: Pixabay.com

3. Multiple Linear Regression

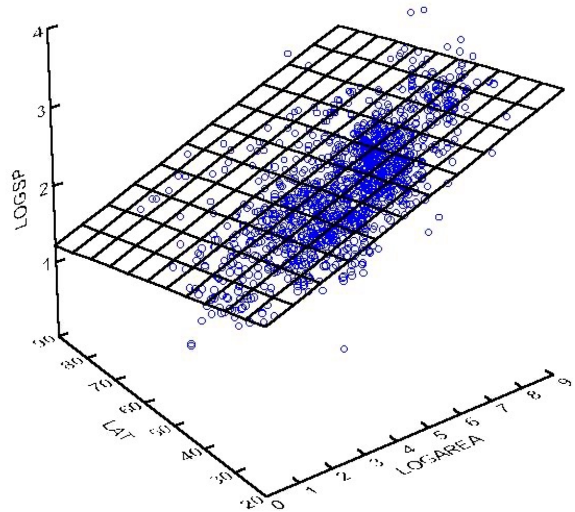
Multiple Linear Regression:

Multiple: The scalar variable y is dependent on multiple values $x_{i1}, x_{i2}, \ldots, x_{ik}$:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

Leading to the following system of equations (for n data points):

$$\begin{aligned} y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \epsilon_1 \\ y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \epsilon_2 \\ &\vdots \\ y_n &= \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \epsilon_n \end{aligned}$$



Source: http://ordination.okstate.edu/MULTIPLE.htm

Which can be written more concisely as $\mathbf{y} = \mathbf{X}\beta + \epsilon$ with

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \quad \textit{and} \quad \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

Multiple Linear Regression:

Estimating regression coefficients for Multiple Linear Regression

- **OLS** can be used in basically the same way to minimize $||\mathbf{y} - \mathbf{X}\beta||^2$
- Setting the partial derivatives to 0 and solving for the regression coefficients $\mathbf{b}$ as an estimation for β leads to

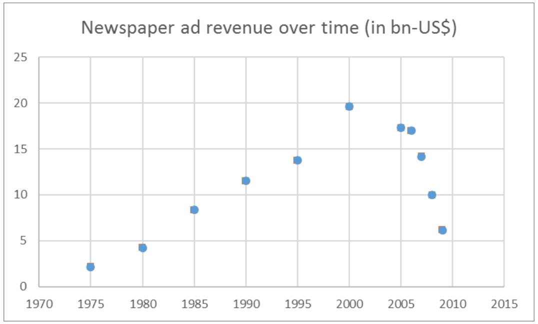
$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{pmatrix} \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} & \cdots & \sum_i x_{i1}x_{ik} \\ \sum_i x_{i2}x_{i1} & \sum_i x_{i2}^2 & \cdots & \sum_i x_{i2}x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_i x_{ik}x_{i1} & \sum_i x_{ik}x_{i2} & \cdots & \sum_i x_{ik}^2 \end{pmatrix}^{-1} \cdot \begin{pmatrix} \sum_i x_{i1}y_i \\ \sum_i x_{i2}y_i \\ \vdots \\ \sum_i x_{ik}y_i \end{pmatrix}$$

- Remark: $(\mathbf{X}^T\mathbf{X})^{-1}$ can be computed efficiently with Gauss-Jordan Algorithm
- Remark: We dropped the explicit notation of b_0 and instead add an additional column of 1s to $\mathbf{X}$ instead.

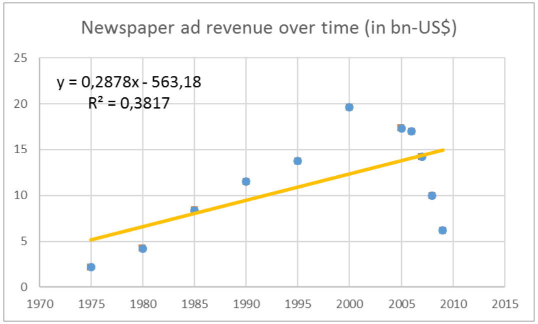
4. Polynomial Linear Regression

Polynomial Linear Regression:

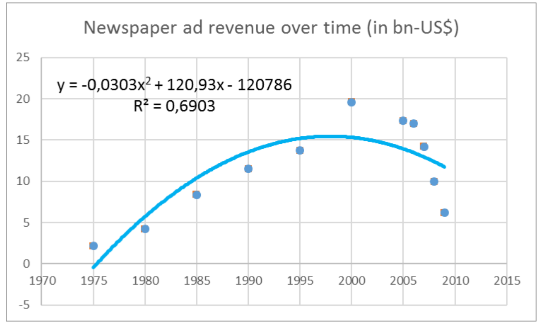
Not every dataset can be fitted with a linear regressor



Original dataset



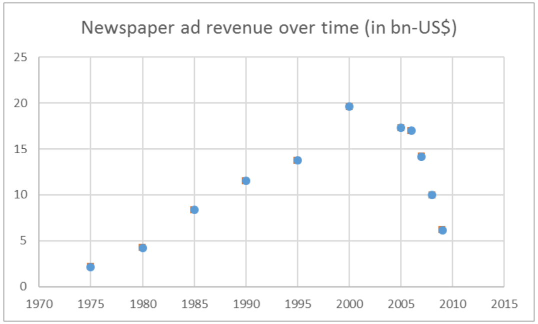
Linear Regression



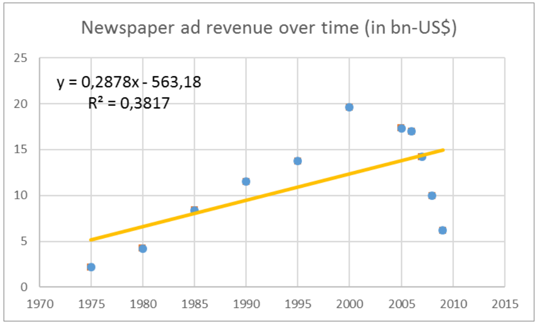
Quadratic Regression (x, x^2)

Polynomial Linear Regression:

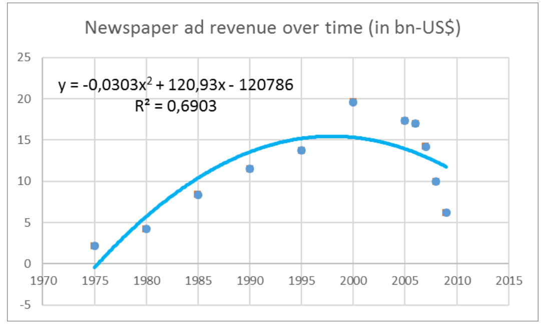
Not every dataset can be fitted with a linear regressor



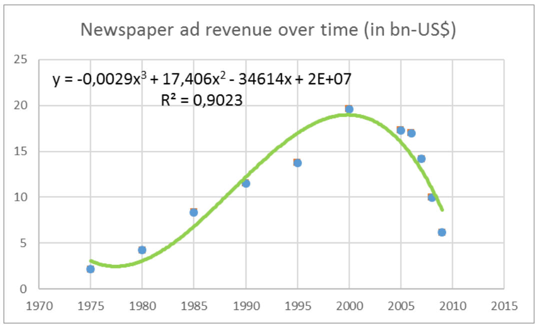
Original dataset



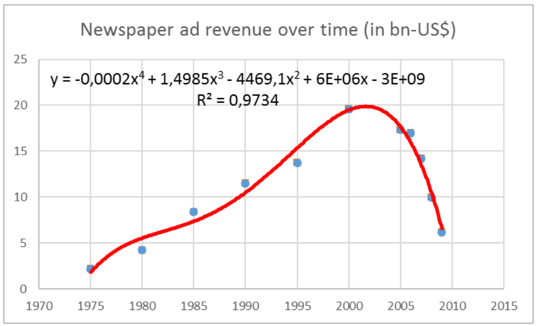
Linear Regression



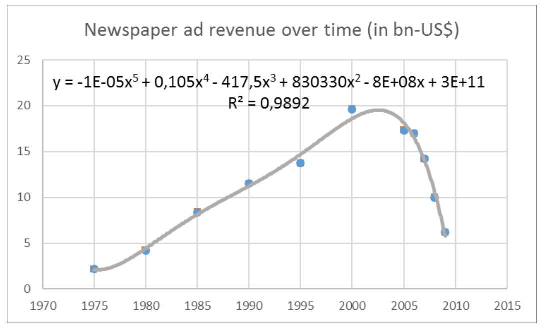
Quadratic Regression (x, x^2)



Polynomial regression of degree 3



Polynomial regression of degree 4



Polynomial regression of degree 5

Polynomial Linear Regression:

Polynomial Regression with one independent variable

- Instead of the function $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ used in the Simple Linear Regression, use a **polynomial function of degree d**:

  Degree 2: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ (quadratic model)
  Degree 3: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i$ (cubic model)
  ...
  Degree d: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_d x_i^d + \epsilon_i$

- This can be transformed into Multiple Linear Regression: Simply treat $x, x^2, \ldots$ as distinct variables, e.g. for degree 3, replace x_i by x_{i1}, x_i^2 by x_{i2} and x_i^3 by x_{i3}, leading to

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

Polynomial Linear Regression:

Polynomial Regression with multiple independent variables

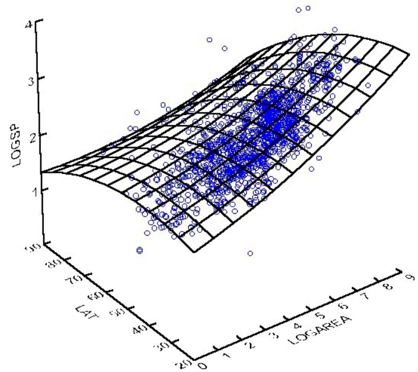
- Polynomial Regression can easily be extended to multiple independent variables, by using all combinations of features and polynomial terms.

- **Example:** y is dependent on two features x_1 and x_2. Fit a (standard) polynomial of degree 2:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2 + \beta_5 x_{i1} x_{i2} + \epsilon_i$$

  $\rightarrow$ Multiple Linear Regression in 6 parameters.

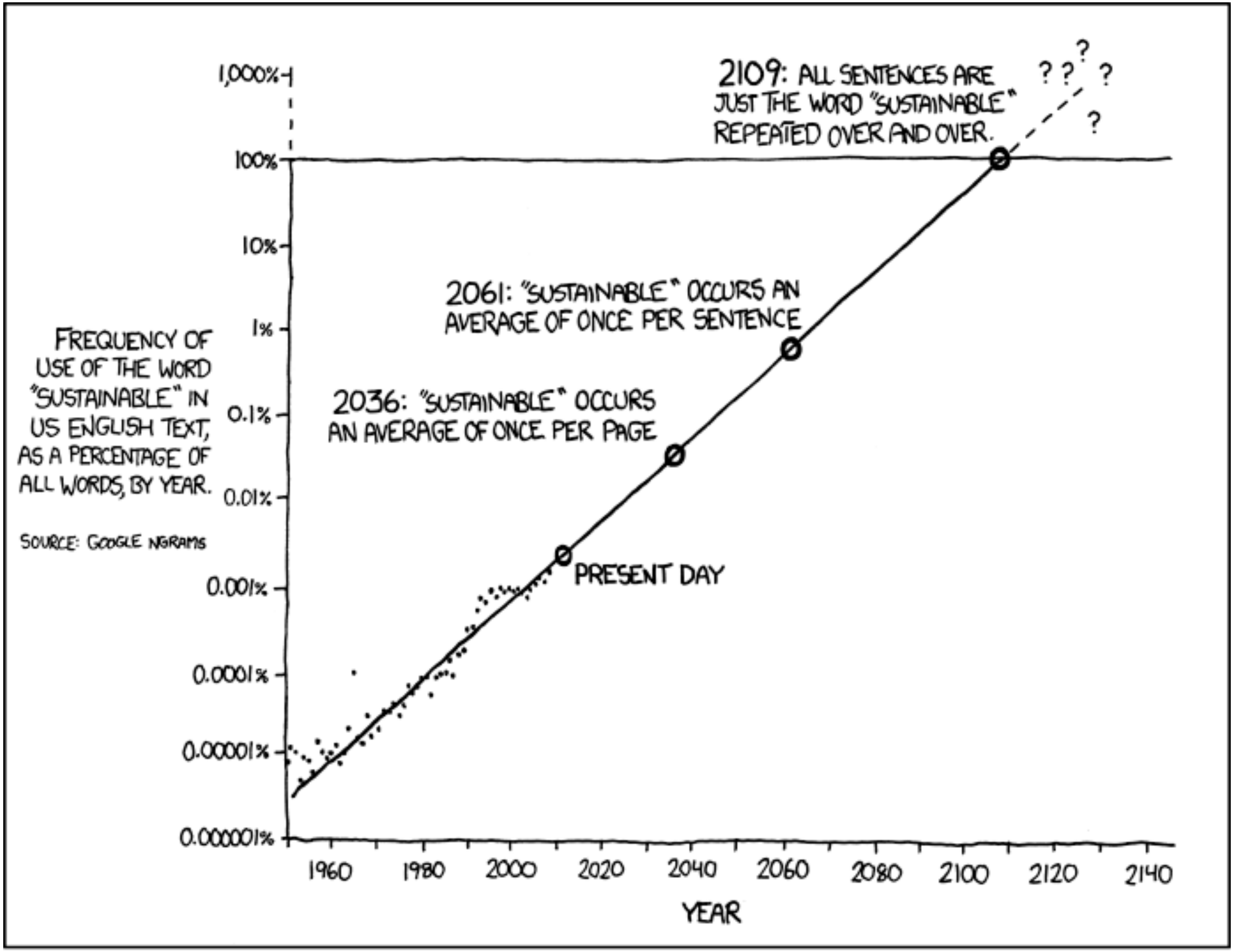
- **Beware:** The number of parameters grows
  - exponentially in the number of independent features and
  - exponentially in the number of the degree of the polynomial



Quadratic function in 2 Variables (Source)

Polynomial Linear Regression:

The dangers of using Linear Regression for extrapolation



THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.

(Source)

Further Reading

- https://en.wikipedia.org/wiki/Linear_regression
- http://scikit-learn.org/stable/modules/linear_model.html
- http://onlinestatbook.com/2/regression/regression.html