

Reminder

Two variable samples

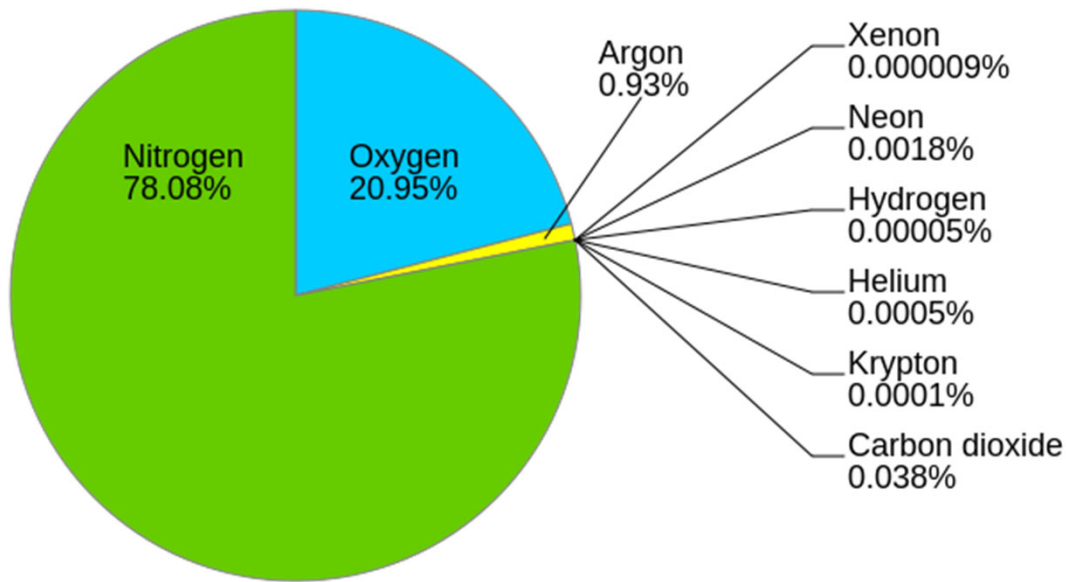


Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

- Oxygen can be distilled from the air
- Hydrocarbons need to be filtered out or the whole thing would go **kaboom!!!**
- When more hydrocarbons were removed, the remaining oxygen stays cleaner
- Except we don't know how dirty was the air to begin with

Linear regression

The **simple linear regression model** is given by

$$Y = \beta_0 + \beta_1 X + \varepsilon = \hat{Y} + \varepsilon$$

ε is the **random error term**

slope β_1 and intercept β_0 of the line are called **regression coefficients**

Note: Y , \hat{Y} , X and ε are random variables

The minimal assumption: $E(\varepsilon | x) = 0 \rightarrow$

$$E(Y | x) = \beta_0 + \beta_1 x + E(\varepsilon | x) = \beta_0 + \beta_1 x$$

$$Y = \beta_0 + \beta_1 X + \epsilon$$

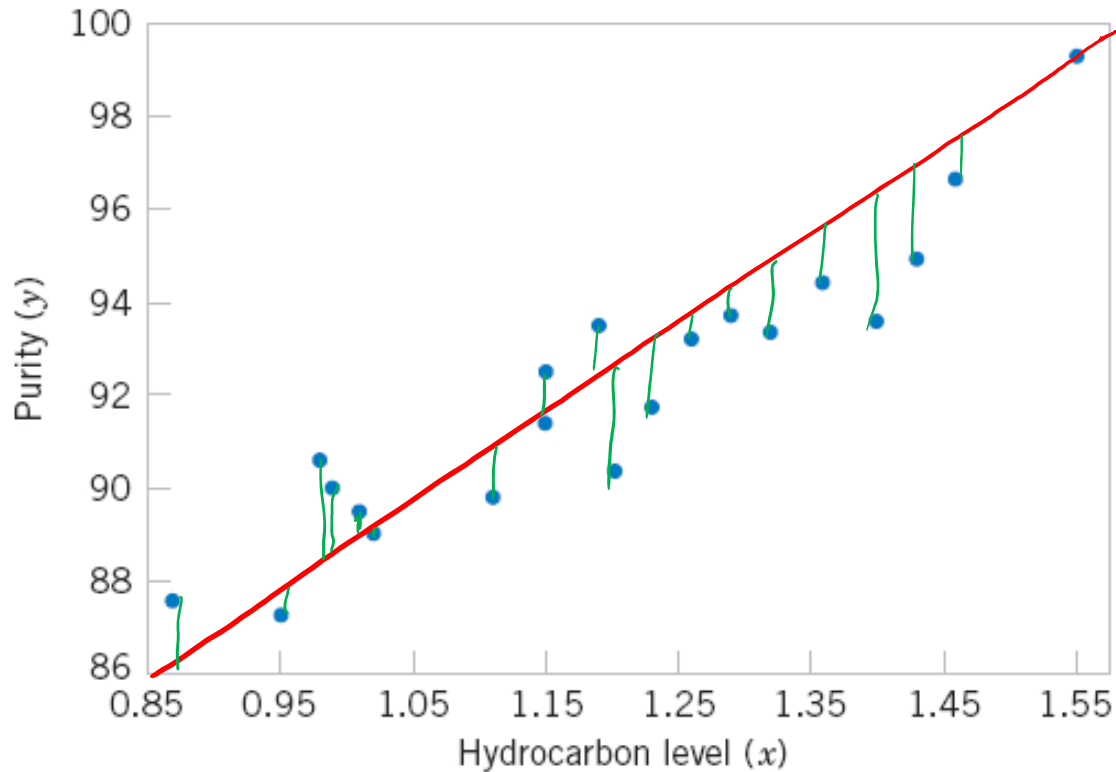


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

$$Y = 75 + 15 \cdot X + \epsilon$$

$$Y = \beta_0 + \beta_1 X + \epsilon ; E(\epsilon | x) = 0 \quad \forall x$$

How does one find β_0 & β_1 ?

$$\begin{aligned} \text{Cov}(Y, X) &= \text{Cov}(\beta_0 + \beta_1 X + \epsilon, X) = \\ &= \text{Cov}(\beta_0, X) + \beta_1 \text{Cov}(X, X) + \text{Cov}(\epsilon, X) \end{aligned}$$

$\text{Cov}(\beta_0, X) = 0$ since β_0 is constant

$$\text{Cov}(X, X) = E(X^2) - E(X)^2 = \text{Var}(X)$$

$$\text{Cov}(\epsilon, X) = E(\epsilon \cdot X) - E(\epsilon) \cdot E(X) =$$

$$= E(\epsilon \cdot X) = \sum_{\text{all } x} x \cdot E(\epsilon | x) = 0$$

Thus

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

$$\beta_0 = E(Y) - \beta_1 E(X)$$

Method of least squares

- The **method of least squares** is used to estimate the parameters, β_0 and β_1 by minimizing the sum of the squares of the vertical deviations in Figure 11-3.

Figure 11-3 Deviations of the data from the estimated regression model.

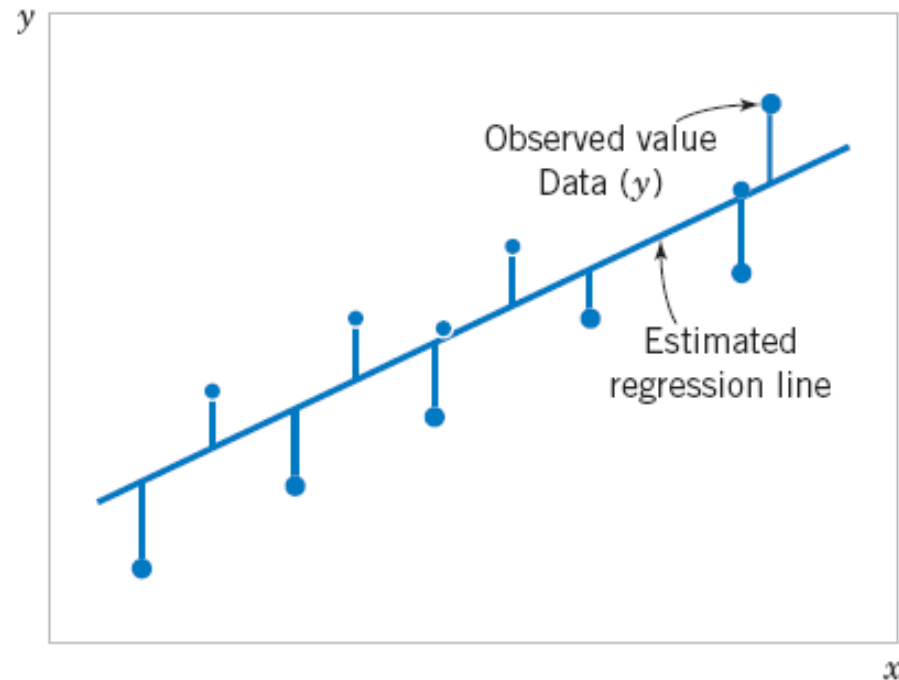


Figure 11-3 Deviations of the data from the estimated regression model.

Traditional notation

Definition

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (11-7)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{\left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}} \quad (11-8)$$

where $\bar{y} = (1/n) \sum_{i=1}^n y_i$ and $\bar{x} = (1/n) \sum_{i=1}^n x_i$.

Connection to Cov(X,Y)/Var(X) result

Definition

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (11-7)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{\left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad (11-8)$$

where $\bar{y} = (1/n) \sum_{i=1}^n y_i$ and $\bar{x} = (1/n) \sum_{i=1}^n x_i$.

Different types of y

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (11-7)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n \frac{y_i x_i}{n} - \frac{\left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right)}{n^2}}{\sum_{i=1}^n \frac{x_i^2}{n} - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n^2}} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad (11-8)$$

where $\bar{y} = (1/n) \sum_{i=1}^n y_i$ and $\bar{x} = (1/n) \sum_{i=1}^n x_i$.

$$\bar{y} = \sum y_i / n$$

$$\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$$

$$\varepsilon_i = y_i - \hat{y}_i$$

The analysis of variance identity is

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (11-24)$$

Symbolically,

$$SS_T = SS_R + SS_E \quad (11-25)$$

11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination (R^2) VERY COMMONLY USED

- The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

is called the **coefficient of determination** and is often used to judge the adequacy of a regression model.

- $0 \leq R^2 \leq 1$;
- We often refer (loosely) to R^2 as the amount of variability in the data explained or accounted for by the regression model.

11-2: Simple Linear Regression

Estimating σ_ε^2

An **unbiased estimator** of σ_ε^2 is

$$\hat{\sigma}_\varepsilon^2 = \frac{SS_E}{n - 2} \quad (11-13)$$

where SS_E can be easily computed using

$$SS_E = SS_T - \hat{\beta}_1 S_{xy} \quad (11-14)$$

Multiple Linear Regression

(Chapters 12-13 in
Montgomery, Runger)

12-1: Multiple Linear Regression Model

12-1.1 Introduction

- Many applications of regression analysis involve situations in which there are more than one regressor variable X_k used to predict Y .
- A regression model then is called a **multiple regression model**.

Multiple Linear Regression Model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + \varepsilon$$

One can also use powers and products of other variables or even non-linear functions like $\exp(x_i)$ or $\log(x_i)$

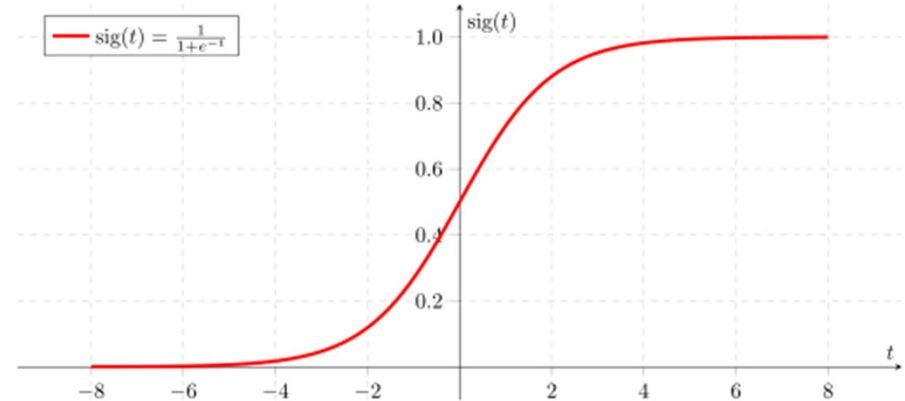
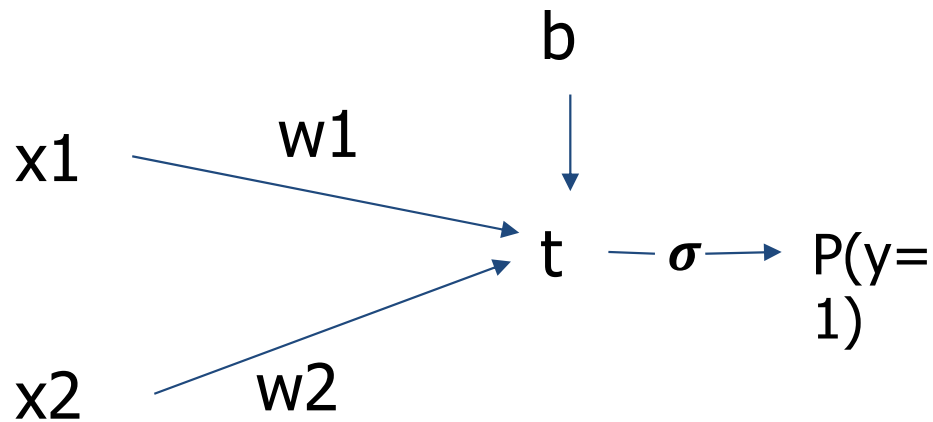
instead of x_3, \dots, x_k .

Example: the general two-variable quadratic regression has 6 constants:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1)^2 + \beta_4 (x_2)^2 + \beta_5 (x_1 x_2) + \varepsilon$$

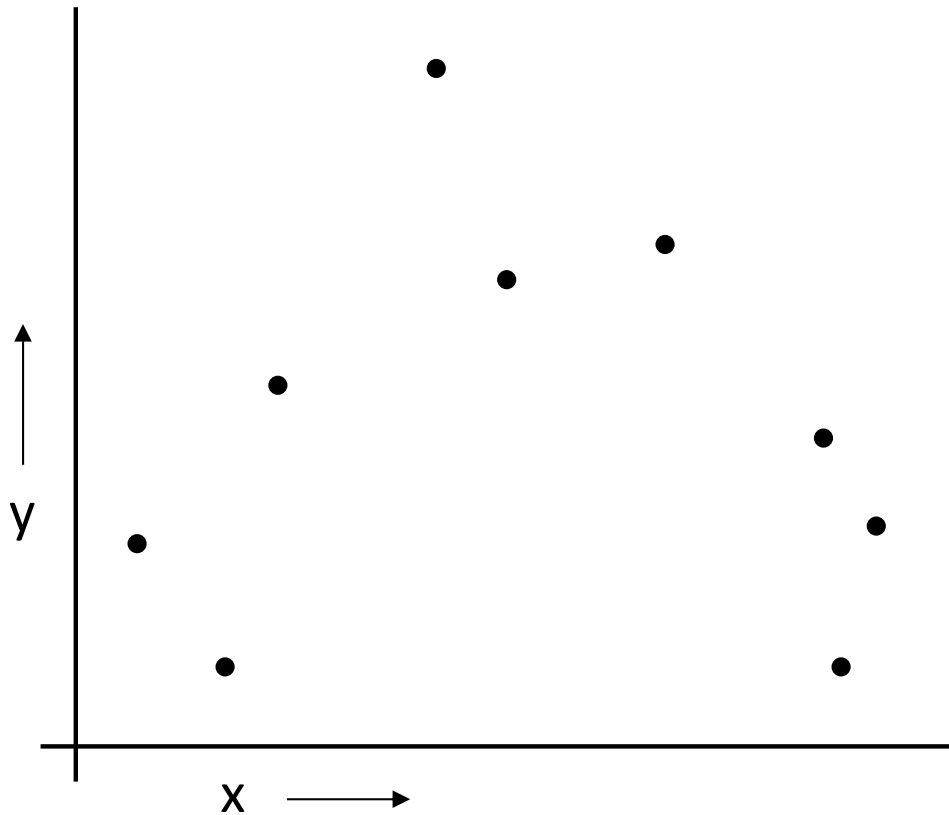
Logistic Regression

$$P(y=1) = \sigma(x_1 * w_1 + x_2 * w_2 + b)$$



How to know where to stop
adding new variables or
powers of old variables?

A Regression Problem

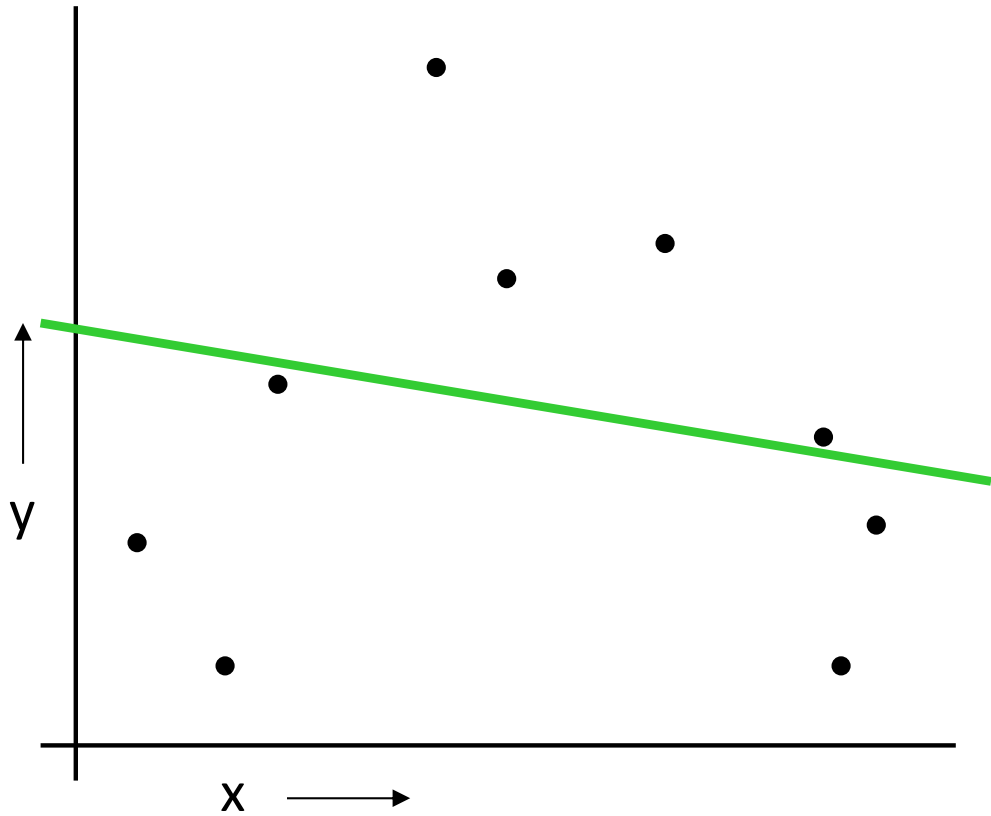


$$y = f(x) + \text{noise}$$

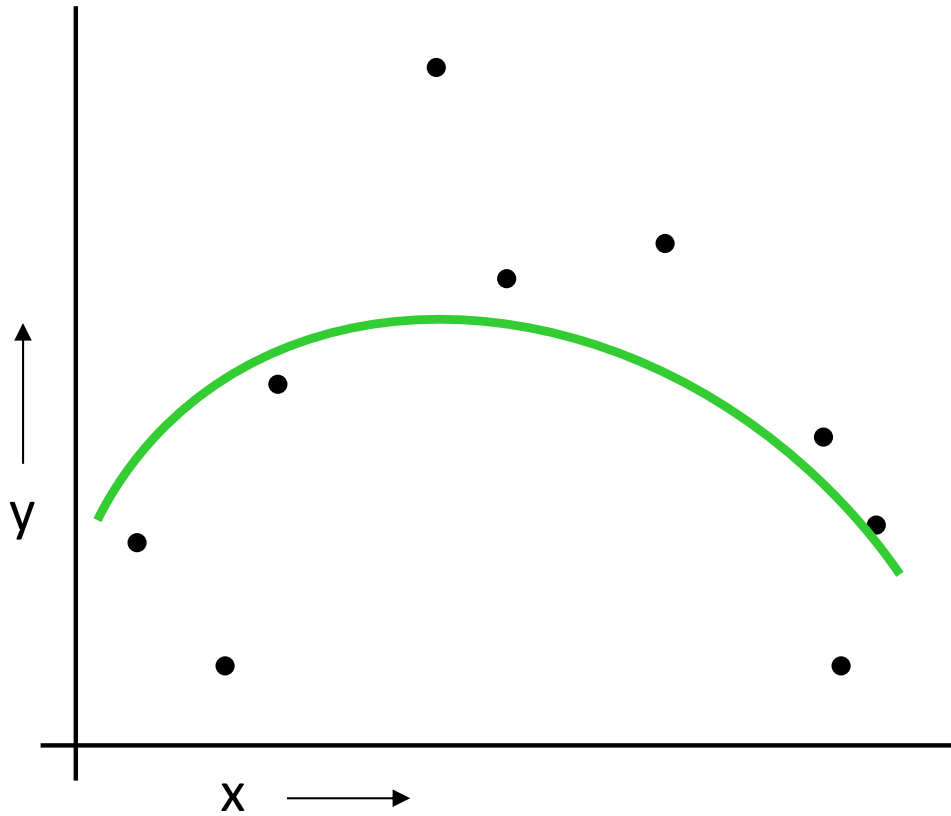
Can we learn f from this data?

Let's consider three methods...

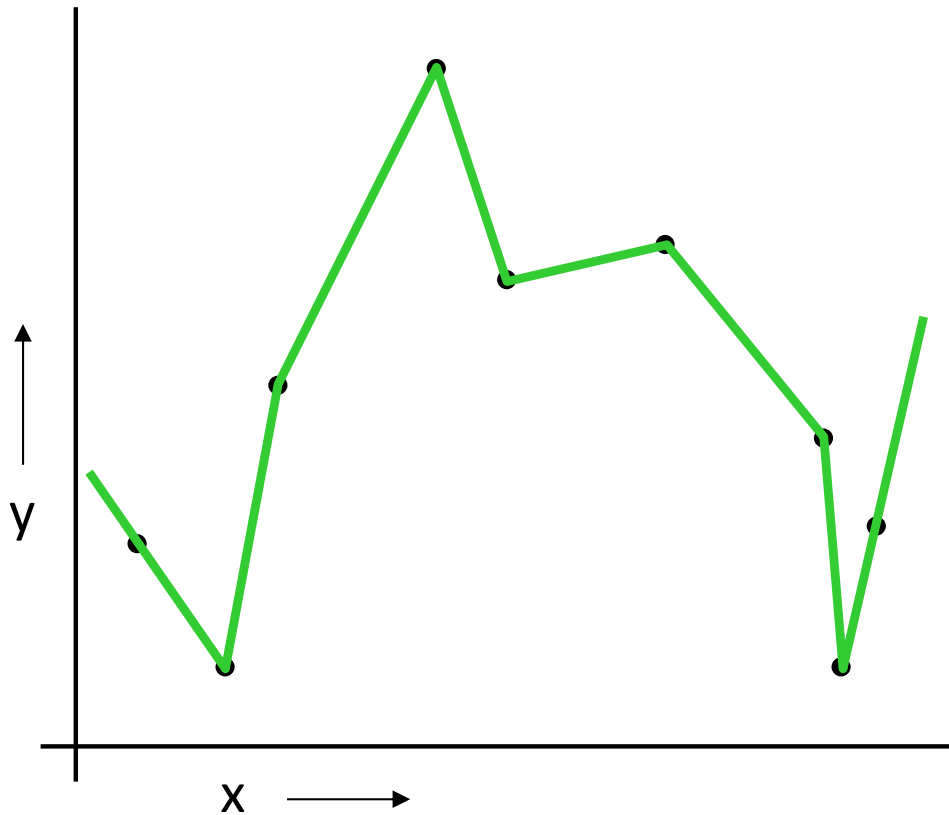
Linear Regression



Quadratic Regression

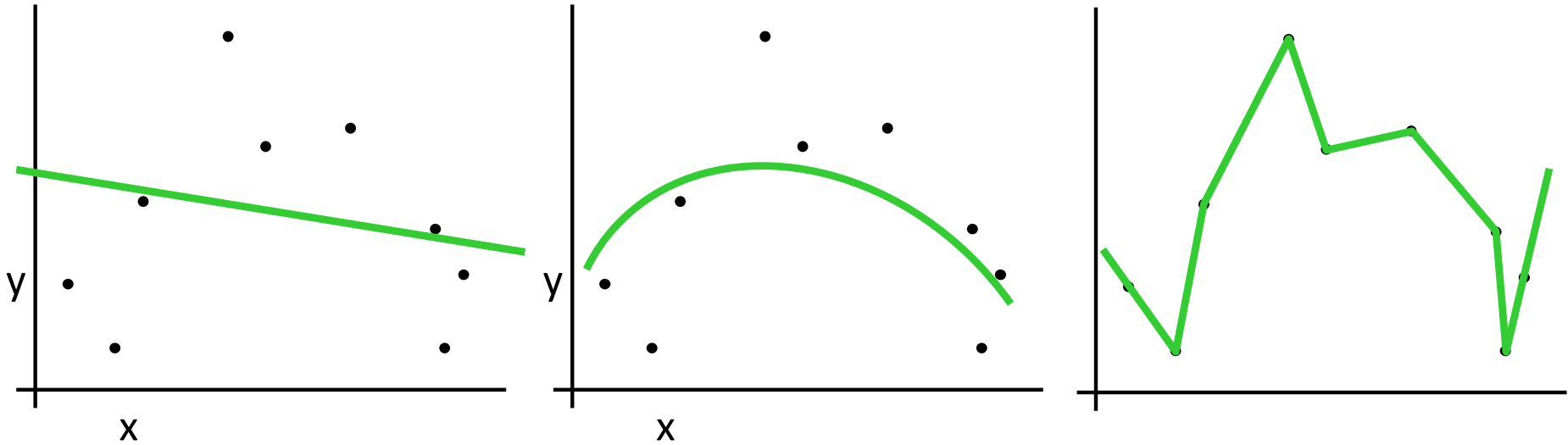


Join-the-dots



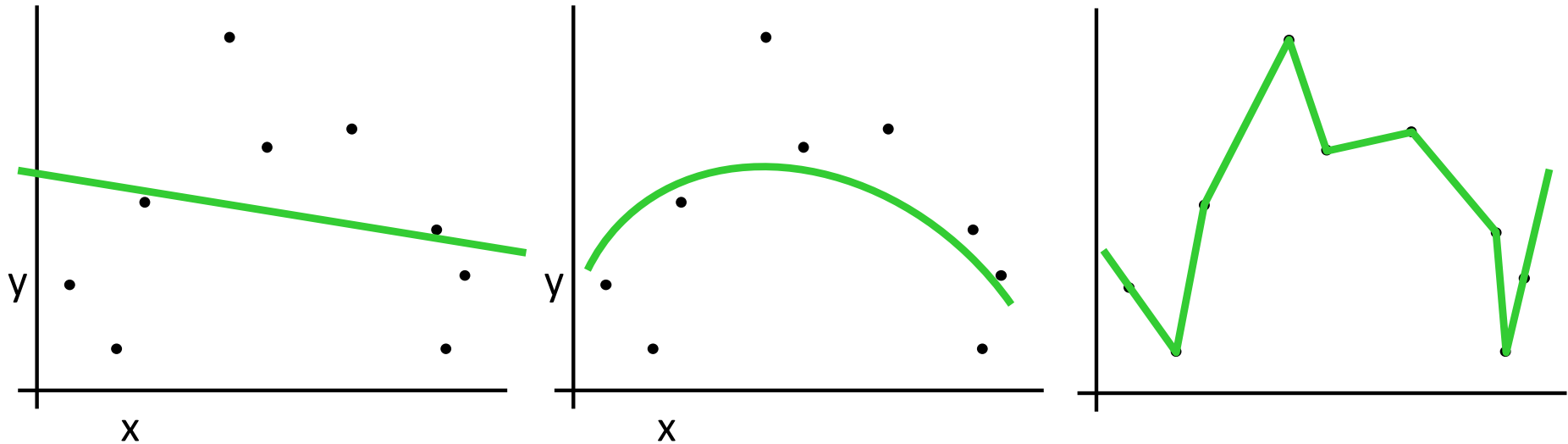
Also known as **piecewise linear nonparametric regression** if that makes you feel better

Which is best?



Why not choose the method with the best fit to the data?

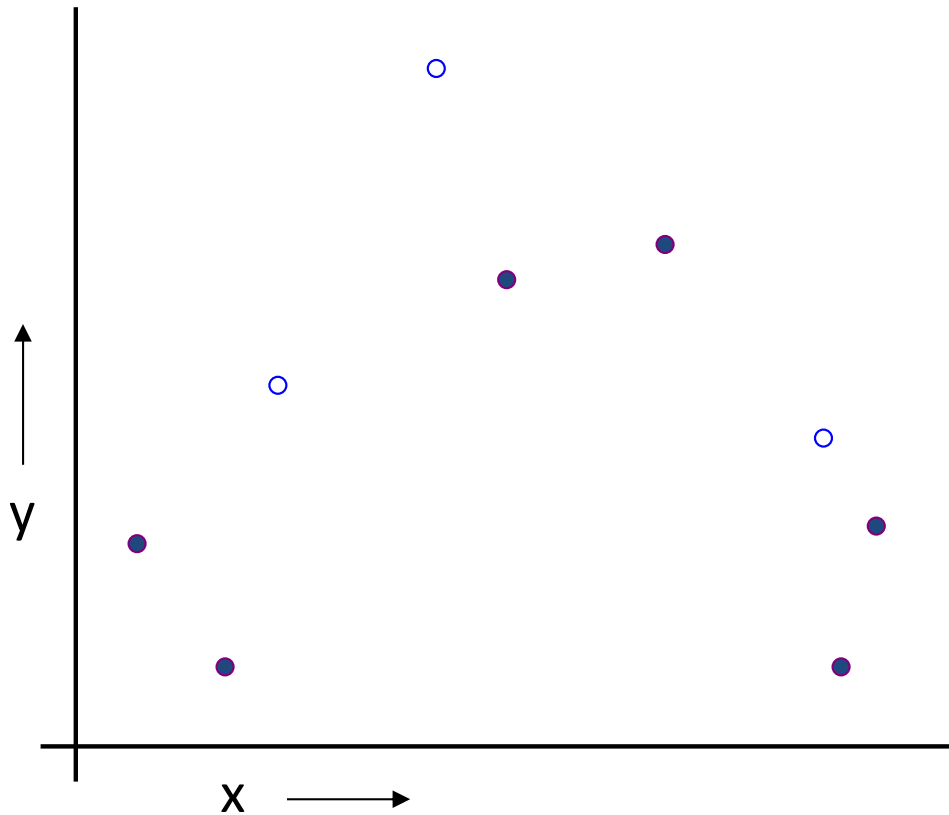
What do we really want?



Why not choose the method with the best fit to the data?

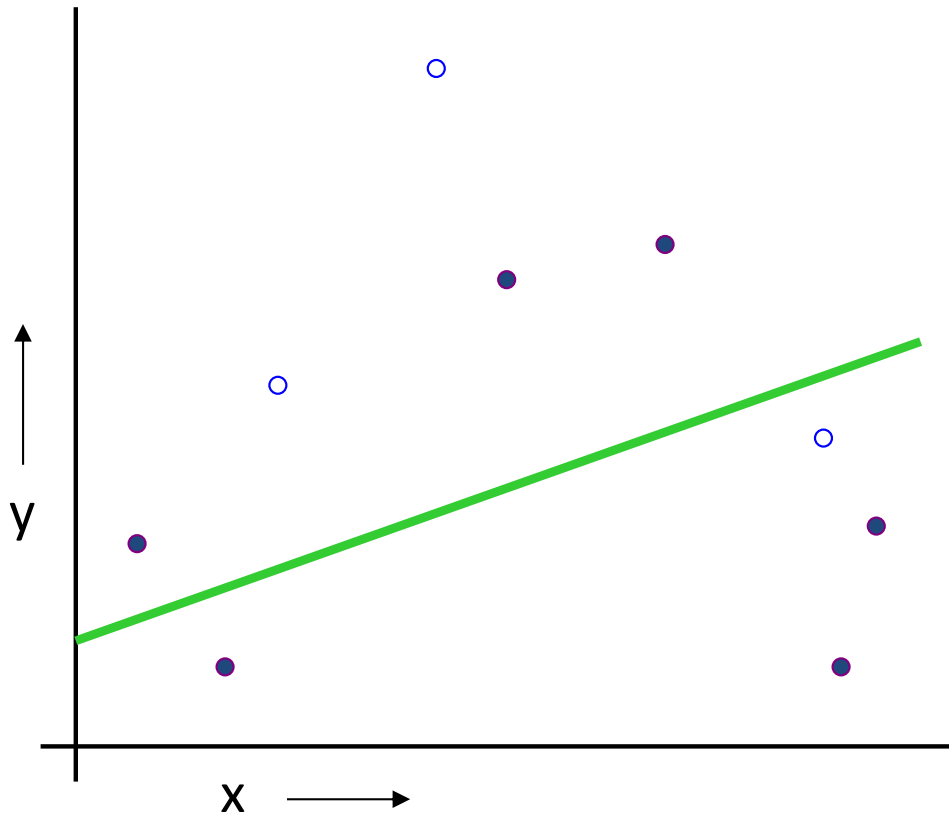
“How well are you going to predict future data drawn from the same distribution?”

The test set method



1. Randomly choose 30% of the data to be in a **test set**
2. The remainder is a **training set**

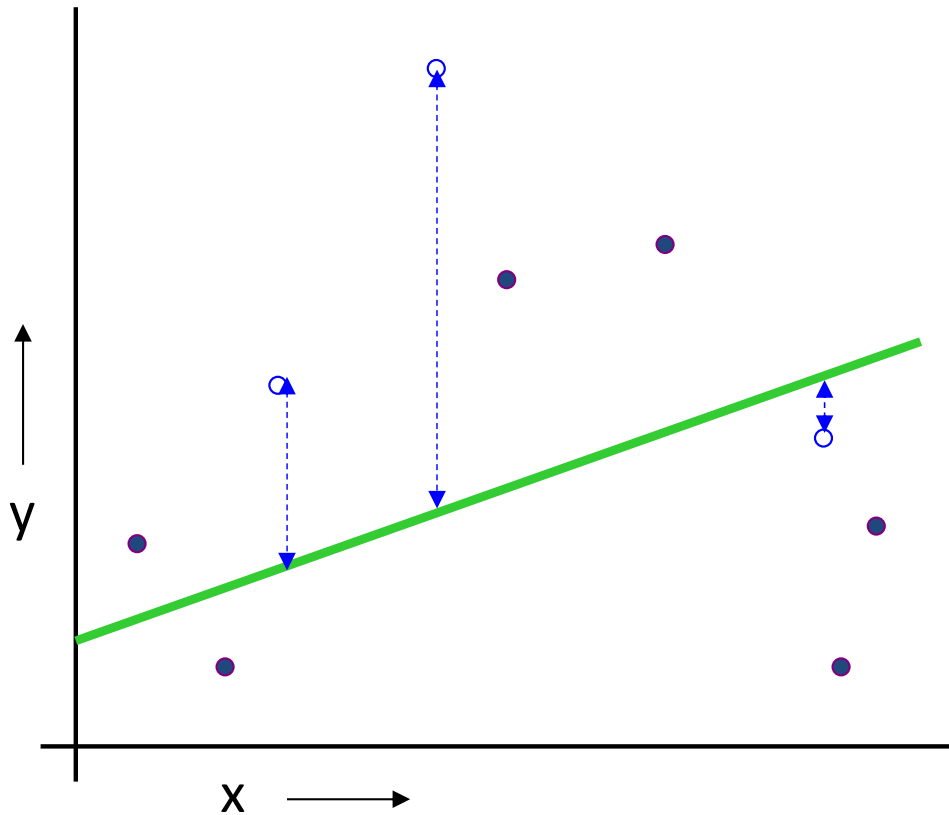
The test set method



(Linear regression example)

1. Randomly choose 30% of the data to be in a **test set**
2. The remainder is a **training set**
3. Perform your regression on the **training set**

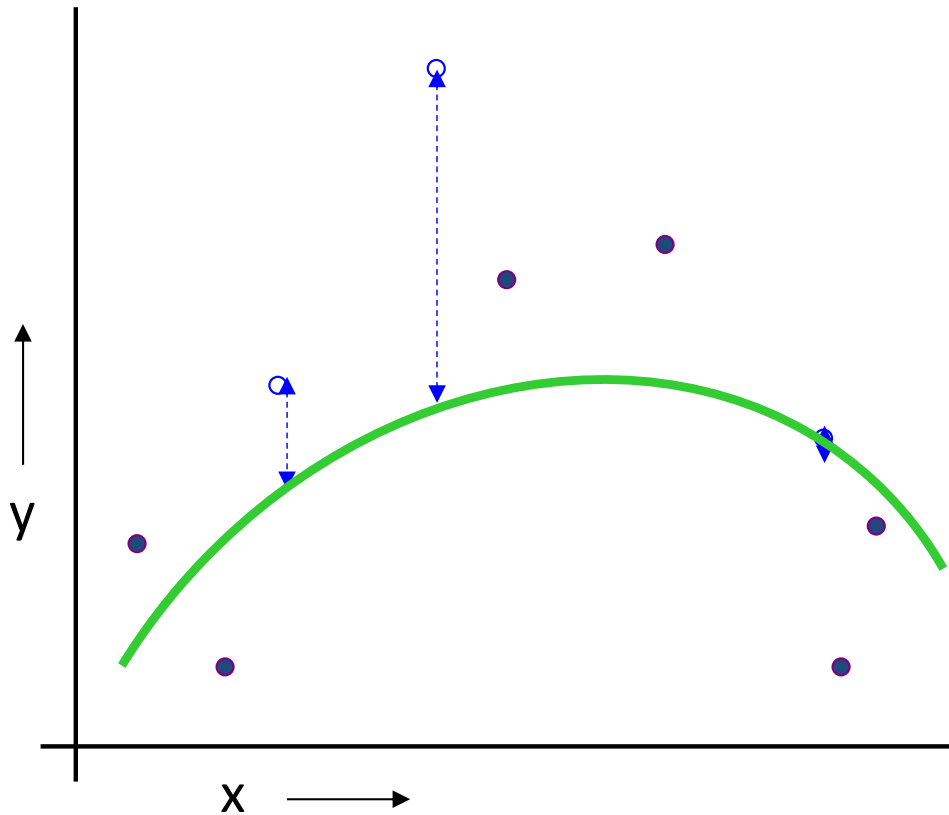
The test set method



(Linear regression example)
Mean Squared Error = 2.4

1. Randomly choose 30% of the data to be in a **test set**
2. The remainder is a **training set**
3. Perform your regression on the training set
4. Estimate your future performance with the test set

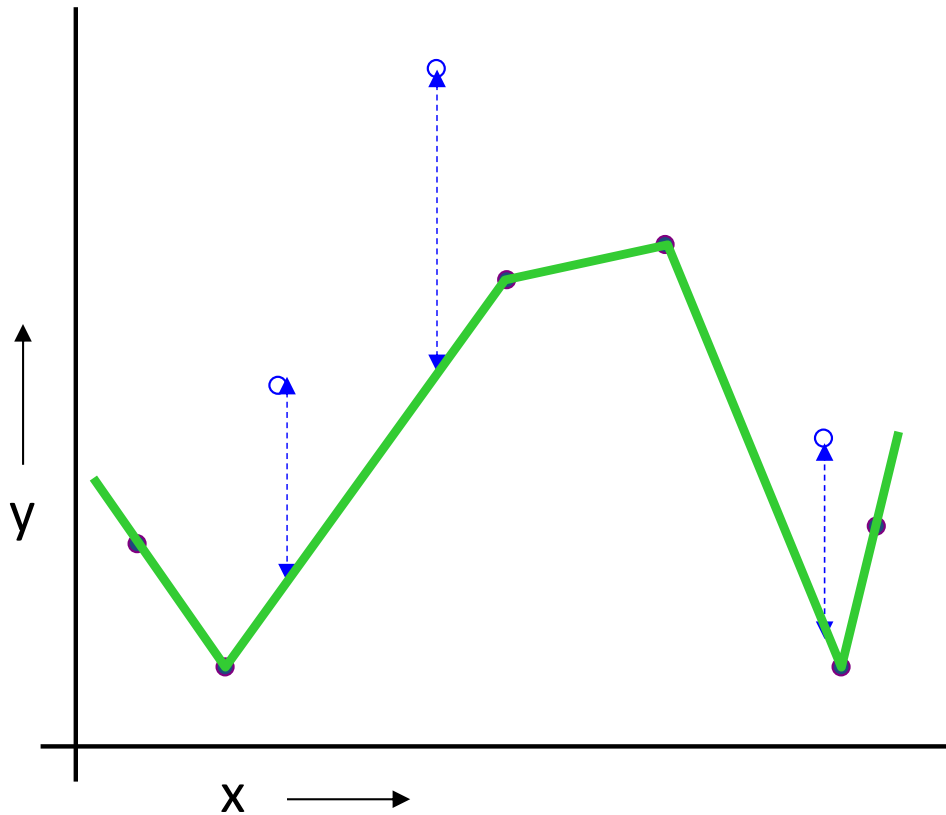
The test set method



(Quadratic regression example)
Mean Squared Error = 0.9

1. Randomly choose 30% of the data to be in a **test set**
2. The remainder is a **training set**
3. Perform your regression on the training set
4. Estimate your future performance with the test set

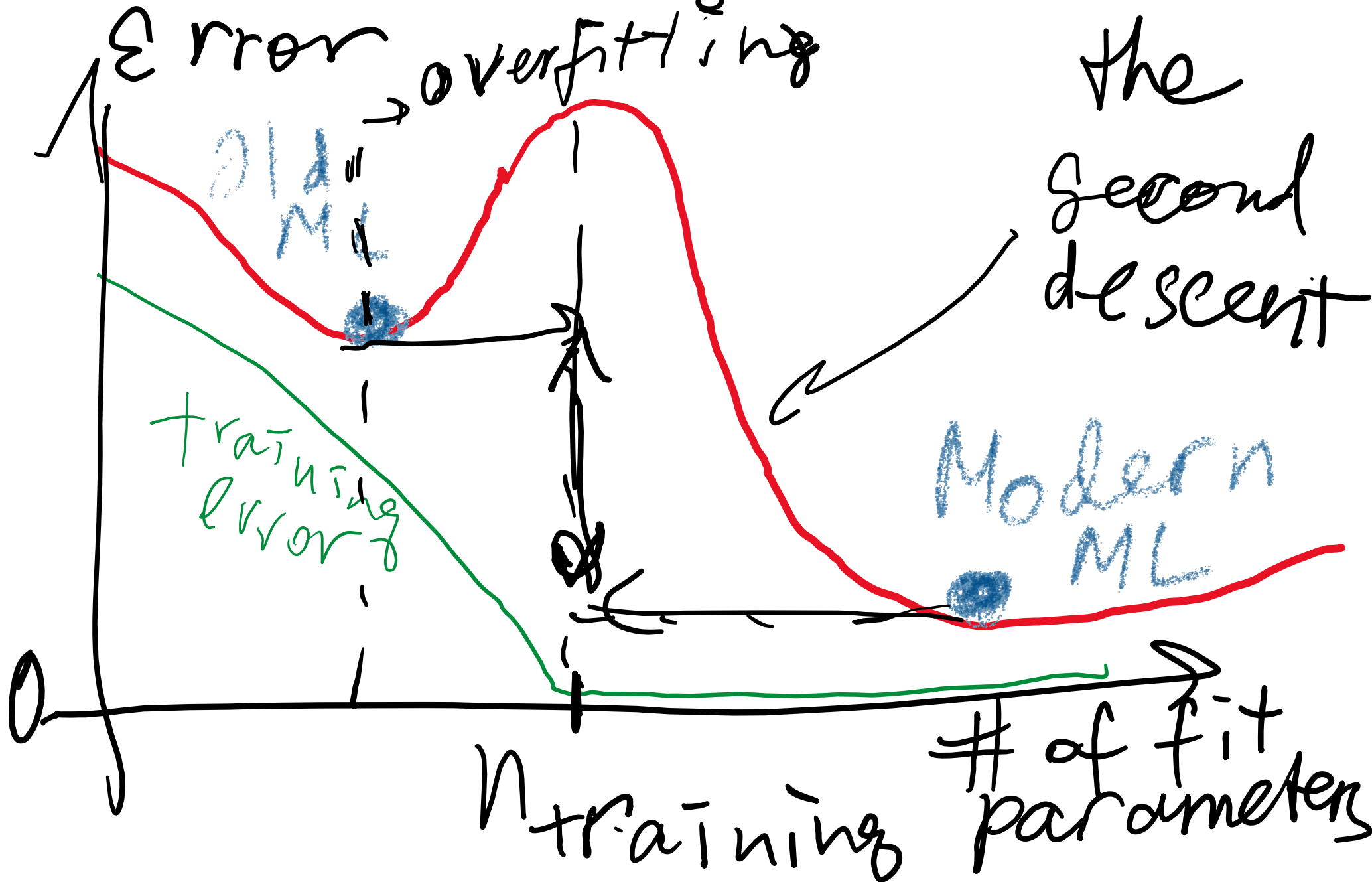
The test set method



(Join the dots example)
Mean Squared Error = 2.2

1. Randomly choose 30% of the data to be in a **test set**
2. The remainder is a **training set**
3. Perform your regression on the training set
4. Estimate your future performance with the test set

Double descend- the main reason modern Machine Learning works so well



12-1: Multiple Linear Regression Model

12-1.3 Matrix Approach to Multiple Linear Regression

Suppose the model relating the regressors to the response is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \quad i = 1, 2, \dots, n$$

In matrix notation this model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (12-6)$$

12-1: Multiple Linear Regression Model

12-1.3 Matrix Approach to Multiple Linear Regression

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

12-1.3 Matrix Approach to Multiple Linear Regression

We wish to find the vector $\hat{\beta}$ that minimizes the sum of squares of error terms:

$$L = \sum_{i=1}^n \varepsilon_i^2 = \varepsilon' \varepsilon = (\mathbf{y} - \mathbf{X}\beta)' (\mathbf{y} - \mathbf{X}\beta)$$

$$0 = \frac{\partial L}{2\partial \beta} = -\mathbf{X}' (\mathbf{y} - \mathbf{X}\beta) = -\mathbf{X}' \mathbf{y} + (\mathbf{X}' \mathbf{X}) \beta$$

The resulting least squares estimate is

$$\hat{\beta} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} \quad (12-7)$$

Analog of $\frac{1}{\text{Var}(x)}$

Analog of $\text{Cov}(x, y)$

Multiple Linear Regression Model

$$\hat{\beta} = (X'X)^{-1} X'y$$

H is an idempotent matrix

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y,$$

$$\hat{y} = Hy, \quad \text{and} \quad e = (I - H)y.$$



$$H = H^2; \quad H^2 = X \underbrace{(X'X)^{-1} X' X (X'X)^{-1}}_I X = X(X'X)^{-1} X' = H$$

Vectors \hat{y} & e are orthogonal since

$$\hat{y}'e = y'H(I-H)y = 0 \quad \text{since}$$

$$H(I-H) = H - H^2 = H - H = 0.$$

12-1: Multiple Linear Regression Models

12-1.4 Properties of the Least Squares Estimators

Unbiased estimators:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}] \\ &= \boldsymbol{\beta} \end{aligned}$$

Covariance Matrix of Estimators:

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{bmatrix}$$

12-1: Multiple Linear Regression Models

12-1.4 Properties of the Least Squares Estimators

Individual variances and covariances:

$$V(\hat{\beta}_j) = \sigma^2 C_{jj}, \quad j = 0, 1, 2$$
$$\text{cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 C_{ij}, \quad i \neq j$$

In general,

$$\text{cov}(\hat{\beta}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \mathbf{C}$$

12-1: Multiple Linear Regression Models

Estimating error variance σ_ε^2

An unbiased estimator of error variance σ_ε^2 is

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^n e_i^2}{n - p} = \frac{SS_E}{n - p} \quad (12-16)$$

Here $p=k+1$ for k -variable multiple linear regression

R² and Adjusted R²

The **coefficient of multiple determination R²**

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

The **adjusted R²** is

$$R_{\text{adj}}^2 = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} \quad (12-23)$$

Handwritten red annotations: A red arrow points to the numerator of the fraction, and another red arrow points to the denominator. The fraction is written as $\frac{\sum \epsilon^2}{\sum y^2}$ in red.

- The adjusted R² statistic penalizes **adding terms** to the MLR model.
- It can help guard against **overfitting** (including regressors that are not really useful)

How to know where to stop adding variables?

- Adding new variables x_i to MLR
watch the adjusted R^2
- Once the adjusted R^2
no longer increases = stop.
Now you did the best you can.

Matlab exercise

- Every group works with
g0=2907; g1=1527; g2=2629; g3=2881;
g4=1144; g5=1066;
- Compute **Multiple Linear Regression (MLR)**:
where
y=exp_t (g0); x1= exp_t (g1); x2= exp_t (g2);
- **How much better** the MLR did compared to the
Single Linear Regression (SLR)?
- **Continue increasing** the number of genes in x
until **R_adj** starts to decrease

How I did it

- `g0=2907; g1=1527; g2=2629; g3=2881;g4=1144; g5=1066;`
- `y=exp_t(g0,:)' ;`
- `%% first use one x to predict y`
- `x=exp_t(g1,:)' ;`
- `figure; plot(x,y,'ko')`
- `lm=fitlm(x,y)`
- `y_fit=lm.Fitted;`
- `hold on;`
- `plot(x,lm.Fitted,'r-');`
- `%% now use 2 x's to predict y`
- `x=[exp_t(g1,:)', exp_t(g2,:)]';`
- `lm2=fitlm(x,y)`
- `y_fit=lm2.Fitted;`
- `hold on; plot(x(:,1),y_fit,'gd');`
- `%% now use m x's to predict y`
- `corr_matrix=corr(exp_t');`
- `g0=2907;`
- `[u v]=sort(corr_matrix(g0,:), 'descend');`
- `x=[exp_t(v(2:m+1),:)]';`
- `lm3=fitlm(x,y)`
- `y_fit=lm3.Fitted;`
- `plot(x(:,1),y_fit,'s');`

Credit: XKCD
comics

WHY ARE THERE SLAVES IN THE BIBLE

WHY DO TWINS HAVE DIFFERENT FINGERPRINTS
WHY ARE AMERICANS AFRAID OF DRAGONS

WHY IS HTTPS CROSSED OUT IN RED
WHY IS THERE A LINE THROUGH HTTPS
WHY IS THERE A RED LINE THROUGH HTTPS ON FACEBOOK
WHY IS HTTPS IMPORTANT

QUESTIONS FOUND IN GOOGLE AUTOCOMPLETE



WHY ARE THERE WEEKS
WHY DO I FEEL DIZZY

WHY DO WHALES JUMP
WHY ARE WITCHES GREEN
WHY ARE THERE MIRRORS ABOVE BEDS
WHY DO I SAY UH
WHY IS SEA SALT BETTER
WHY ARE THERE TREES IN THE MIDDLE OF FIELDS
WHY IS THERE NOT A POKEMON MMO
WHY IS THERE LAUGHING IN TV SHOWS
WHY ARE THERE DOORS ON THE FREEWAY
WHY ARE THERE SO MANY SVCHOST.EXE RUNNING
WHY AREN'T THERE ANY COUNTRIES IN ANTARCTICA
WHY ARE THERE SCARY SOUNDS IN MINECRAFT
WHY IS THERE KICKING IN MY STOMACH
WHY ARE THERE TWO SLASHES AFTER HTTP
WHY ARE THERE CELEBRITIES
WHY DO SNAKES EXIST
WHY DO OYSTERS HAVE PEARLS
WHY ARE DUCKS CALLED DUCKS
WHY DO THEY CALL IT THE CLAP
WHY ARE KYLE AND CARTMAN FRIENDS
WHY IS THERE AN ARROW ON AANG'S HEAD
WHY ARE TEXT MESSAGES BLUE
WHY ARE THERE MUSTACHES ON CLOTHES
WHY ARE THERE MUSTACHES ON CARS
WHY ARE THERE MUSTACHES EVERYWHERE
WHY ARE THERE SO MANY BIRDS IN OHIO
WHY IS THERE SO MUCH RAIN IN OHIO
WHY IS OHIO WEATHER SO WEIRD

WHY DO IGUANAS DIE
WHY AREN'T THERE DINOSAUR GHOSTS

WHY AREN'T ECONOMISTS RICH
WHY DO AMERICANS CALL IT SOCCER
WHY ARE MY EARS RINGING
WHY ARE THERE SO MANY AVENGERS
WHY ARE THE AVENGERS FIGHTING THE X MEN
WHY IS WOLVERINE NOT IN THE AVENGERS

WHY ARE THERE SWARMS OF GNATS
WHY IS THERE PHLEGM
WHY ARE THERE SO MANY CROWS IN ROCHESTER, MN
WHY IS PSYCHIC WEAK TO BUG
WHY DO CHILDREN GET CANCER
WHY IS POSEIDON ANGRY WITH ODYSSEUS
WHY IS THERE ICE IN SPACE

WHY ARE THERE ANTS IN MY LAPTOP

WHY ARE THERE BRIDESMAIDS
WHY DO DYING PEOPLE REACH UP
WHY AREN'T THERE VARICOSE ARTERIES
WHY ARE OLD KUNGONS DIFFERENT



WHY ARE THERE TINY SPIDERS IN MY HOUSE
WHY DO SPIDERS COME INSIDE
WHY ARE THERE HUGE SPIDERS IN MY HOUSE
WHY ARE THERE LOTS OF SPIDERS IN MY HOUSE
WHY ARE THERE SPIDERS IN MY ROOM
WHY ARE THERE SO MANY SPIDERS IN MY ROOM
WHY DO SPIDER BITES ITCH
WHY IS DYING SO SCARY



WHY IS THERE AN OWL IN MY BACKYARD
WHY IS THERE AN OWL OUTSIDE MY WINDOW
WHY IS THERE AN OWL ON THE DOLLAR BILL
WHY DO OWLS ATTACK PEOPLE
WHY ARE AK 47s SO EXPENSIVE
WHY ARE THERE HELICOPTERS CIRCLING MY HOUSE
WHY ARE THERE GODS
WHY ARE THERE TWO SPOCKS

WHY ARE DOGS AFRAID OF FIREWORKS
WHY IS THERE NO KING IN ENGLAND

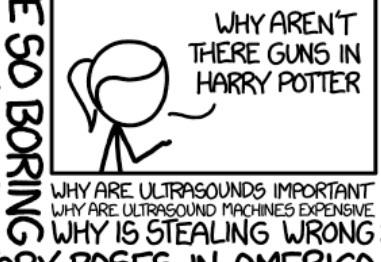
WHY IS PROGRAMMING SO HARD
WHY IS THERE A 0 OHM RESISTOR
WHY DO AMERICANS HATE SOCCER
WHY DO RHYMES SOUND GOOD
WHY DO TREES DIE
WHY IS THERE NO SOUND ON CNN
WHY AREN'T POKEMON REAL
WHY AREN'T BULLETS SHARP
WHY DO DREAMS SEEM SO REAL

WHY IS THERE NO GPS IN LAPTOPS
WHY DO KNEES CLICK
WHY AREN'T THERE E GRADES
WHY IS ISOLATION BAD
WHY DO BOYS LIKE ME
WHY DON'T BOYS LIKE ME
WHY IS THERE ALWAYS A JAVA UPDATE
WHY ARE THERE RED DOTS ON MY THIGHS
WHY IS LYING GOOD



WHY IS MT VESUVIUS THERE
WHY DO THEY SAY T MINUS
WHY ARE THERE OBELISKS
WHY ARE WRESTLERS ALWAYS WET
WHY ARE OCEANS BECOMING MORE ACIDIC
WHY IS ARWEN DYING
WHY AREN'T MY QUAIL LAYING EGGS
WHY AREN'T MY QUAIL EGGS HATCHING
WHY AREN'T THERE ANY FOREIGN MILITARY BASES IN AMERICA

WHY ARE CIGARETTES LEGAL
WHY ARE THERE DUCKS IN MY POOL
WHY IS JESUS WHITE
WHY IS THERE LIQUID IN MY EAR
WHY DO Q TIPS FEEL GOOD
WHY DO GOOD PEOPLE DIE



WHY IS LIFE SO BORING
WHY ARE ULTRASOUNDS IMPORTANT
WHY ARE ULTRASOUND MACHINES EXPENSIVE
WHY IS STEALING WRONG