DS 4400

Machine Learning and Data Mining I

Alina Oprea
Associate Professor, CCIS
Northeastern University

September 13 2018

Review

- Probability review
  - Random variables
  - Expectation, Variance, CDF, PDF
  - Example distributions
  - Independence and conditional independence
  - Bayes' Theorem
- Linear algebra review
  - Matrix, vectors
  - Inner products
  - Norms
  - Distance

Resources

Probability

- Review notes from Stanford's machine learning class
- Sam Roweis's probability review

Linear algebra

- Review notes from Stanford's machine learning class
- Sam Roweis's linear algebra review

Vectors and matrices

- Vector in R^n is an ordered set of n real numbers.
  - e.g. v = (1,6,3,4) is in R^4
  - A column vector: $\begin{pmatrix} 1 \\ 6 \\ 3 \\ 4 \end{pmatrix}$
  - A row vector: $\begin{pmatrix} 1 & 6 & 3 & 4 \end{pmatrix}$

- m-by-n matrix is an object in R^{mxn} with m rows and n columns, each entry filled with a (typically) real number:

$$\begin{pmatrix} 1 & 2 & 8 \\ 4 & 78 & 6 \\ 9 & 3 & 2 \end{pmatrix}$$

Matrix multiplication

We will use upper case letters for matrices. The elements are referred by $A_{i,j}$.

- Matrix product:

$$A \in \mathbb{R}^{m \times n} \qquad B \in \mathbb{R}^{n \times p}$$

$$C = AB \in \mathbb{R}^{m \times p}$$

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

e.g.

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

$$AB = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}$$

Matrix transpose

Transpose: You can think of it as
- "flipping" the rows and columns

OR
- "reflecting" vector/matrix on line

e.g. $\begin{pmatrix} a \\ b \end{pmatrix}^T = \begin{pmatrix} a & b \end{pmatrix}$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^T = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A+B)^T = A^T + B^T$

A is a symmetric matrix if $A = A^T$

Linear independence

- A set of vectors is linearly independent if none of them can be written as a linear combination of the others.
- Vectors $v_1,\ldots,v_k$ are linearly independent if $c_1v_1+\ldots+c_kv_k = 0$ implies $c_1=\ldots=c_k=0$
- Otherwise they are linearly dependent

$$\begin{pmatrix} | & | & | \\ v_1 & v_2 & v_3 \\ | & | & | \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

e.g.

$$\begin{pmatrix} 1 & 0 \\ 2 & 3 \\ 1 & 3 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

(u,v)=(0,0), i.e. the columns are linearly independent.

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix} \quad x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}$$

Linearly dependent

$$x_3 = -2x_1 + x_2$$

Inverse of a matrix

- Inverse of a square matrix A, denoted by A^{-1} is the *unique* matrix s.t.
  - $AA^{-1} = A^{-1}A = I$ (identity matrix)
- If A^{-1} and B^{-1} exist, then
  - $(AB)^{-1} = B^{-1}A^{-1}$,
  - $(A^T)^{-1} = (A^{-1})^T$
- For orthonormal matrices $\mathbf{A}^{-1} = \mathbf{A}^T$
- For diagonal matrices $\mathbf{D}^{-1} = \mathrm{diag}\{d_1^{-1}, \ldots, d_n^{-1}\}$

Rank of a Matrix

- rank(A) (the rank of a m-by-n matrix A) is

  The maximal number of linearly independent columns

  The maximal number of linearly independent rows

- If A is n by m, then
  - rank(A)<= min(m,n)

- Examples

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 & 3 \\ 0 & 5 & 2 \end{pmatrix}$$

System of linear equations

$$
\begin{aligned}
4x_1 &- 5x_2 = -13 \\
-2x_1 &+ 3x_2 = 9.
\end{aligned}
$$

Matrix formulation

$$Ax = b$$

$$A = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} -13 \\ 9 \end{bmatrix}.$$

If A has an inverse, solution is $x = A^{-1}b$

Linear regression

- One of the most widely used techniques
- Fundamental to many complex models
  - Generalized Linear Models
  - Logistic regression
  - Neural networks
  - Deep learning
- Easy to understand and interpret
- Efficient to solve in closed form
- Efficient practical algorithm (gradient descent)

Supervised Learning: Overview

Hypothesis space

Functions $\mathcal{F}$

$f : \mathcal{X} \rightarrow \mathcal{Y}$

Training data

$\{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}\}$

Training

LEARNING

find $\hat{f} \in \mathcal{F}$
s.t. $y_i \approx \hat{f}(x_i)$

Learning machine

New data

PREDICTION

$y = \hat{f}(x)$

x

Testing

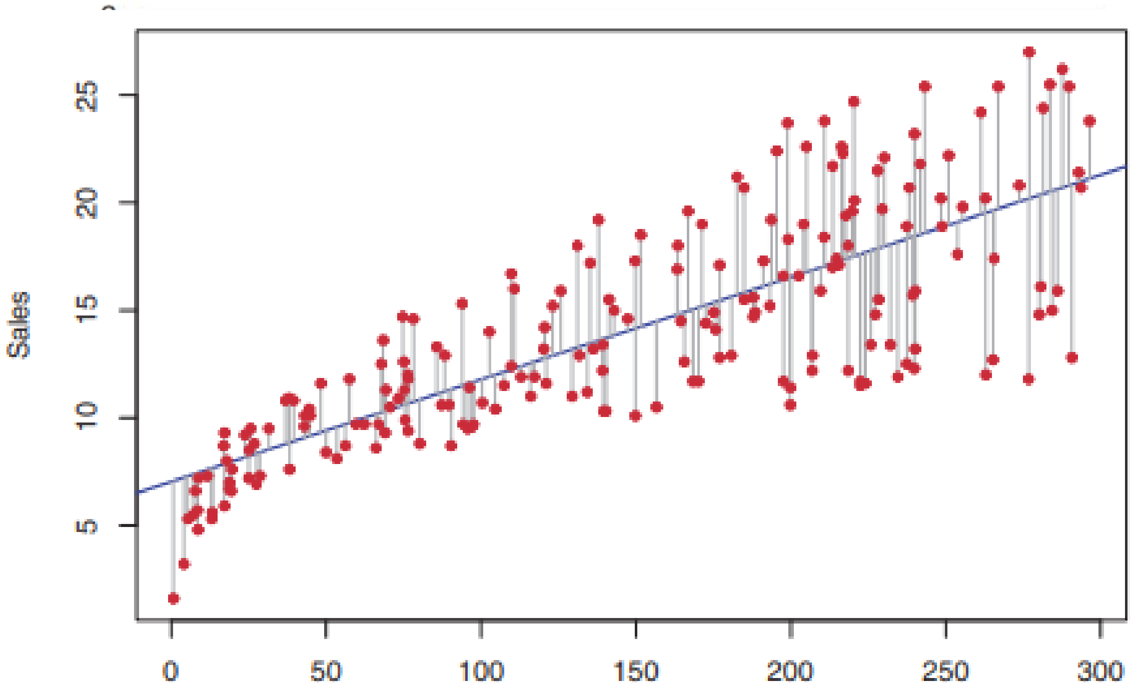
Linear regression

Given:

– Data $\boldsymbol{X} = \left\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\right\}$ where $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$ Features

– Corresponding labels $\boldsymbol{y} = \left\{y^{(1)}, \ldots, y^{(n)}\right\}$ where $y^{(i)} \in \mathbb{R}$

Response variables

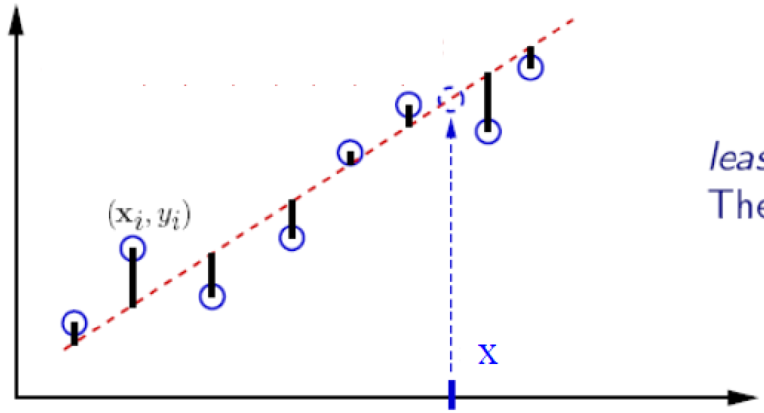


Hypothesis: linear model

- Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

Simple linear regression
Regression model is a line with 2 parameters: θ_0, θ_1

- Fit model by minimizing sum of squared errors



least squares (LSQ)
The fitted line is used as a predictor

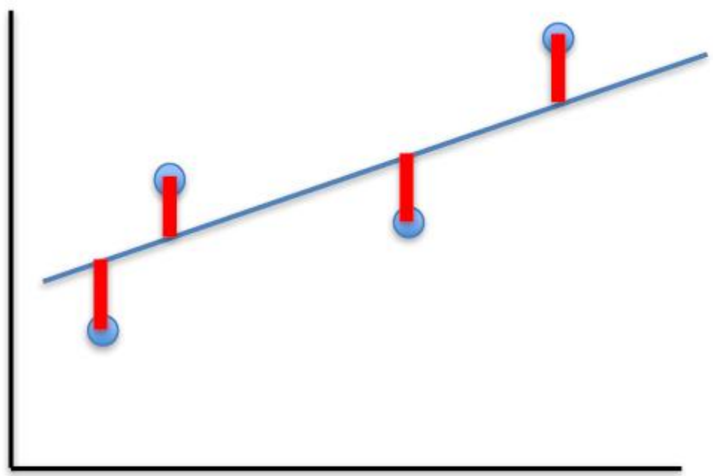
Least squares Linear Regression

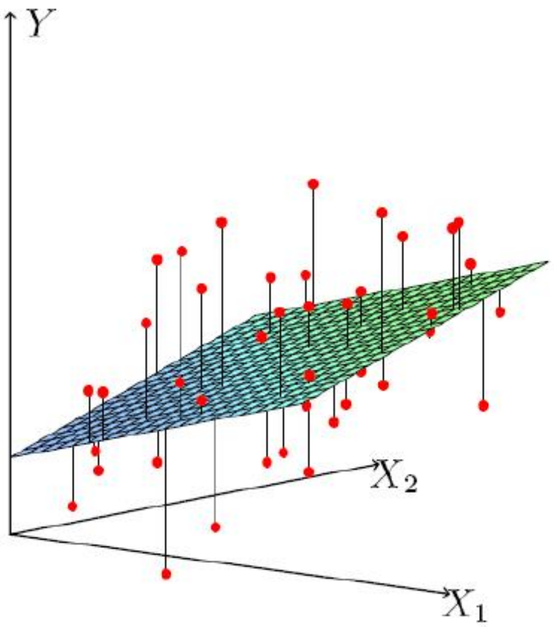
- Cost Function

$$J(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}} \left(\boldsymbol{x}^{(i)} \right) - y^{(i)} \right)^2$$

Mean Square Error (MSE)

- Fit by solving $\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$

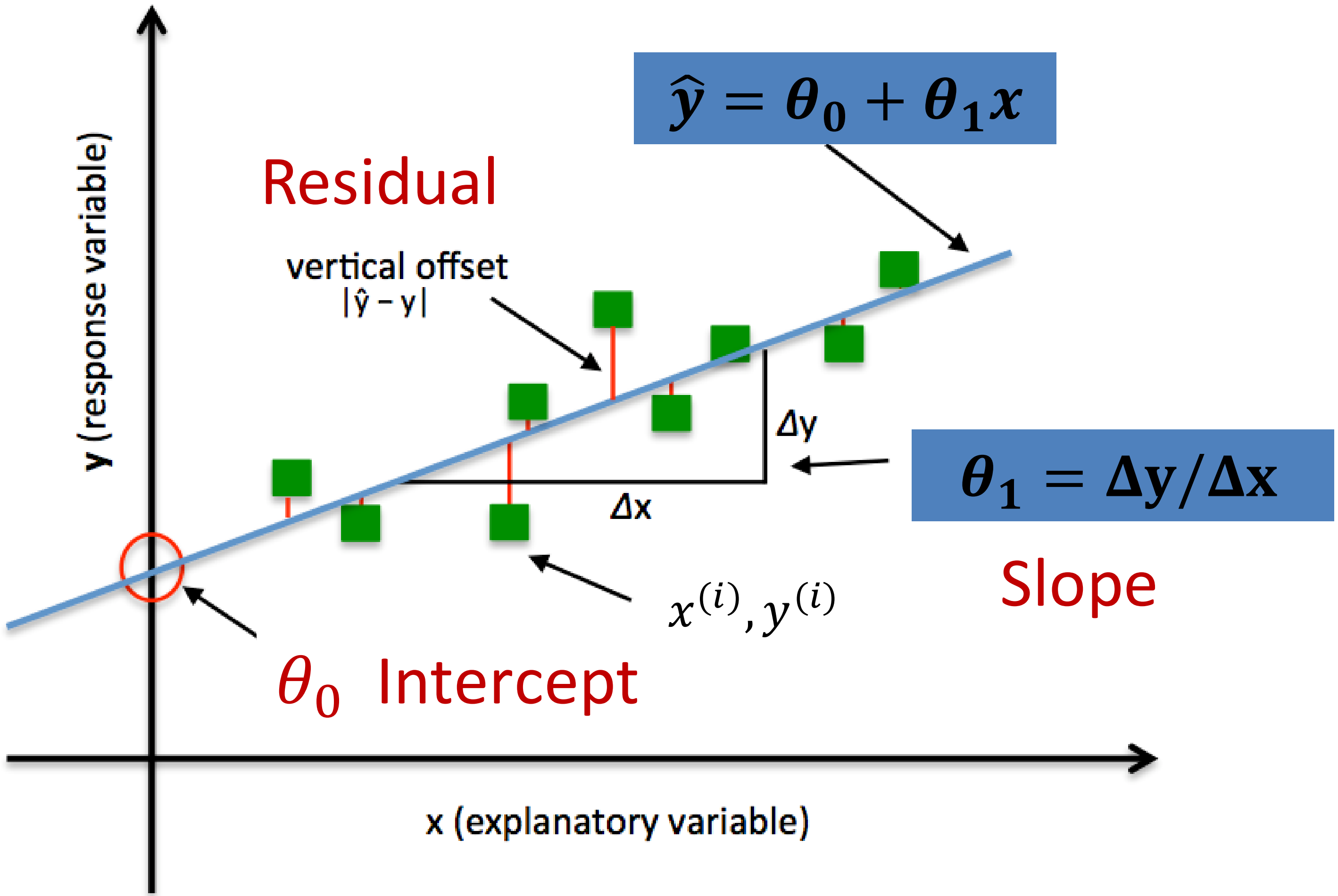




Terminology and Metrics

- Residuals
  - Difference between predicted values and actual values values
  - Predicted value for example i is: $\hat{y}^{(i)} = h_\theta\left(x^{(i)}\right)$
  - $R^{(i)} = \left|y^{(i)} - \hat{y}^{(i)}\right| = |y^{(i)} - (\theta_0 + \theta_1 x^{(i)})|$
- Residual Sum of Squares (RSS)
  - $RSS = \sum R^{(i)} = \sum \left[y^{(i)} - (\theta_0 + \theta_1 x^{(i)})\right]^2$
- Residual Standard Error (RSE)
  - $RSE = \sqrt{\frac{RSS}{n-2}}$
  - RSE^2 is a measure of variance of the model

Interpretation



$$h_\theta(x) = \theta_0 + \theta_1 x$$

MSE$= \frac{1}{n} \sum_{i=1}^{n} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$

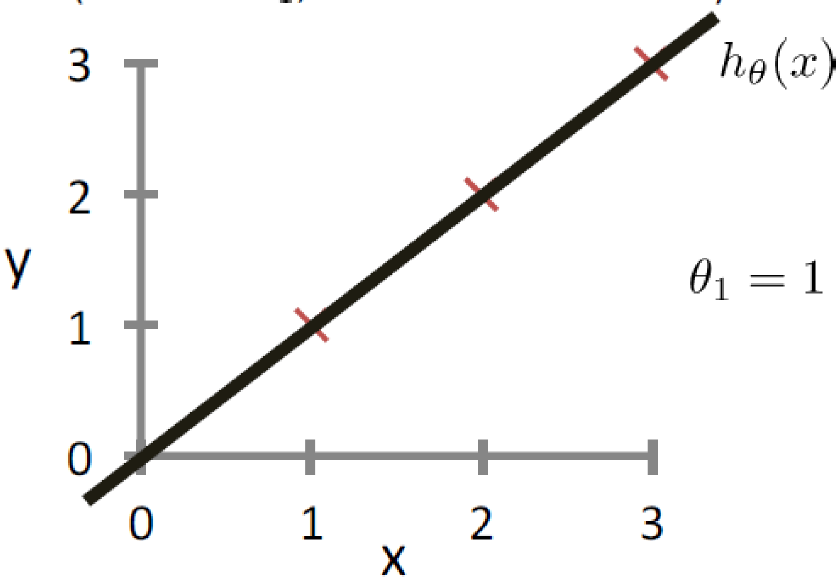
Intuition on MSE

$$J(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}} \left(\boldsymbol{x}^{(i)} \right) - y^{(i)} \right)^2$$

For insight on J(), let's assume $x \in \mathbb{R}$ so $\boldsymbol{\theta} = [\theta_0, \theta_1]$

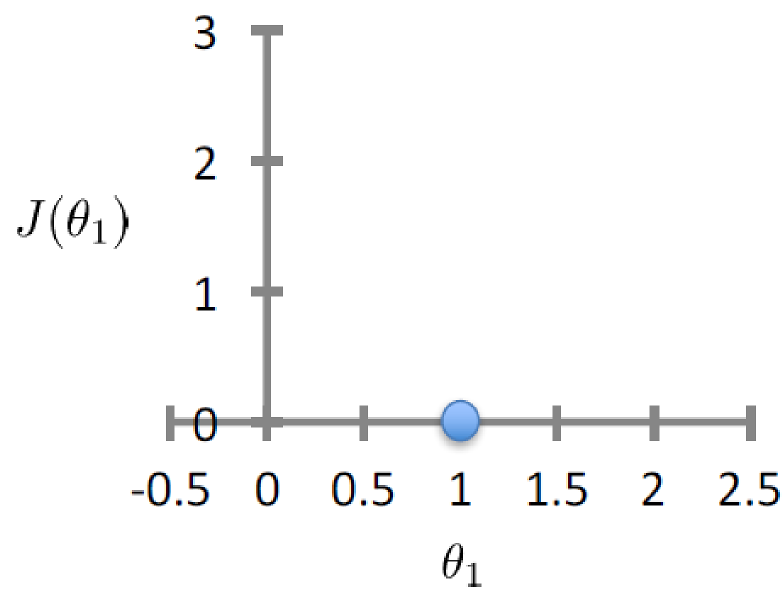
$h_\theta(x)$

(for fixed θ_1, this is a function of x)



$J(\theta_1)$

(function of the parameter θ_1)



Fix θ_0 =0

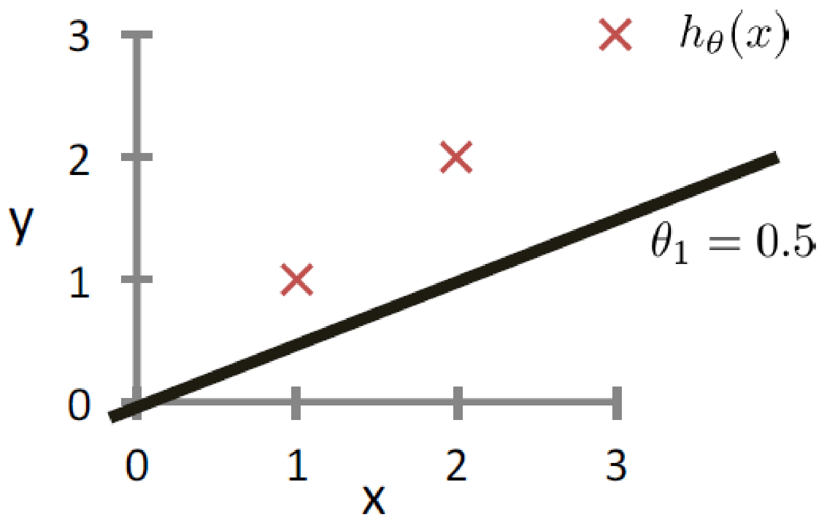
Intuition on cost function

$$J(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)} \right)^2$$

For insight on J(), let's assume $x \in \mathbb{R}$ so $\boldsymbol{\theta} = [\theta_0, \theta_1]$

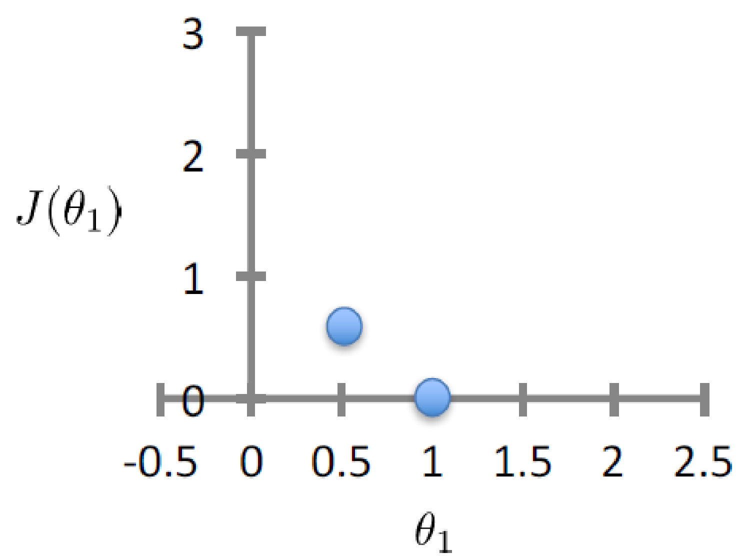
$h_\theta(x)$

(for fixed θ_1, this is a function of x)



$J(\theta_1)$

(function of the parameter θ_1)



Based on example by Andrew Ng

$$J([0, 0.5]) = \frac{1}{2 \times 3} \left[(0.5 - 1)^2 + (1 - 2)^2 + (1.5 - 3)^2 \right] \approx 0.58$$

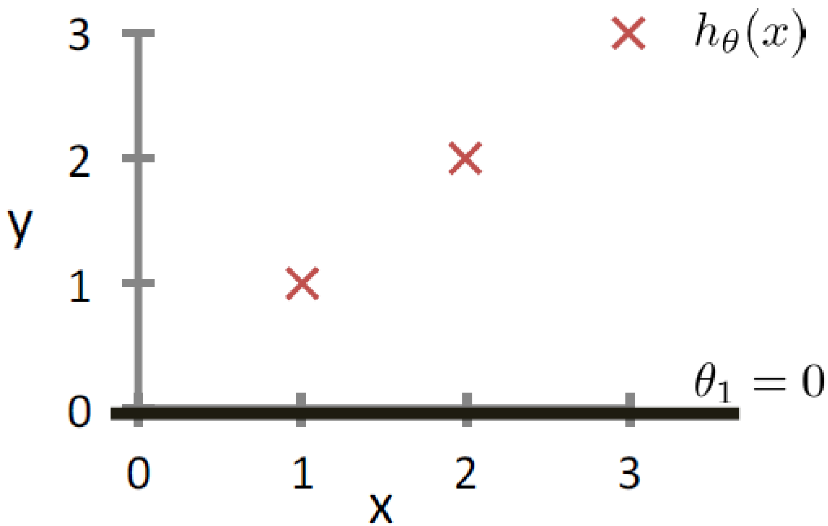
Intuition on cost function

$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2$$

For insight on J(), let's assume $x \in \mathbb{R}$ so $\boldsymbol{\theta} = [\theta_0, \theta_1]$

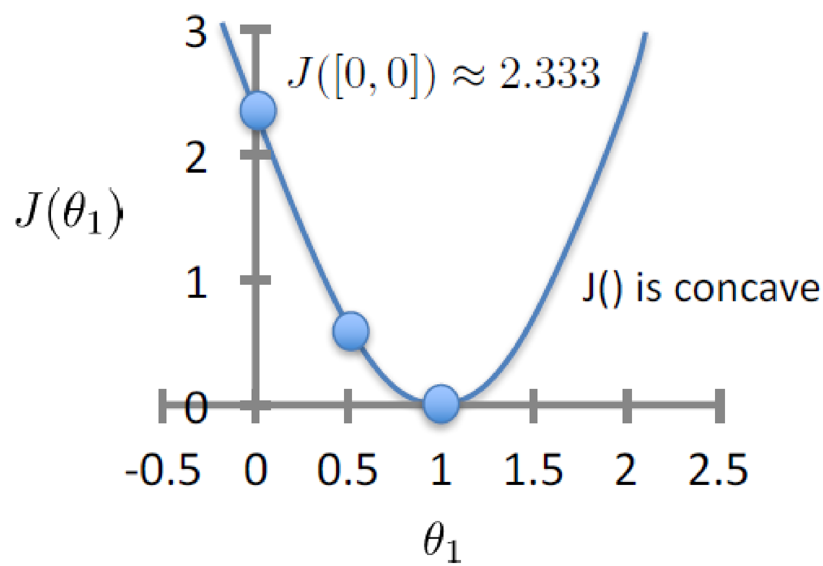
$h_\theta(x)$

(for fixed θ_1, this is a function of x)



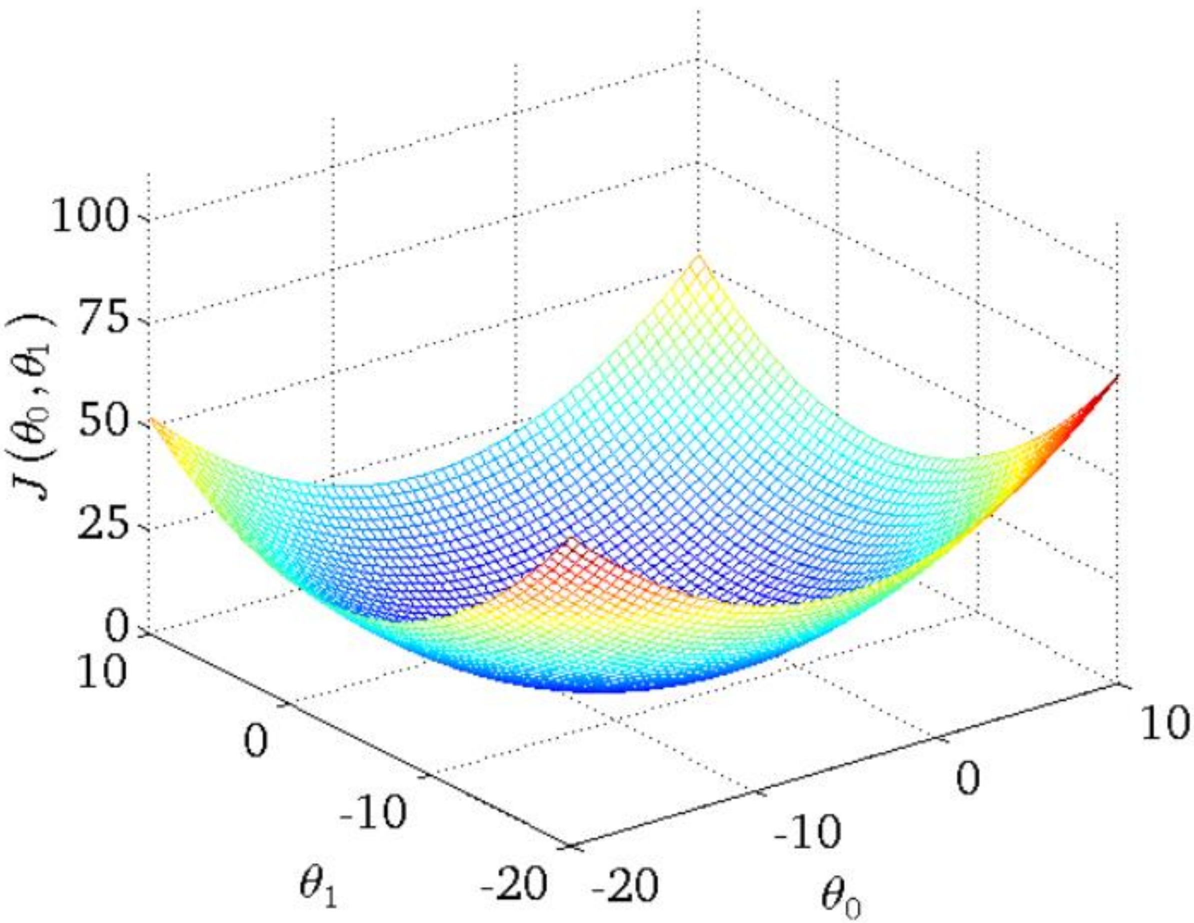
$J(\theta_1)$

(function of the parameter θ_1)



Based on example
by Andrew Ng

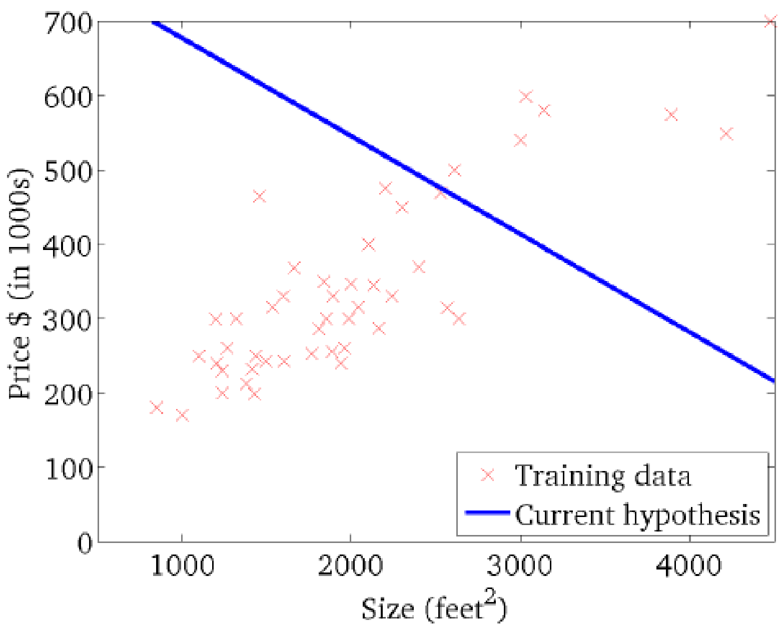
Cost function



Relation between h and J

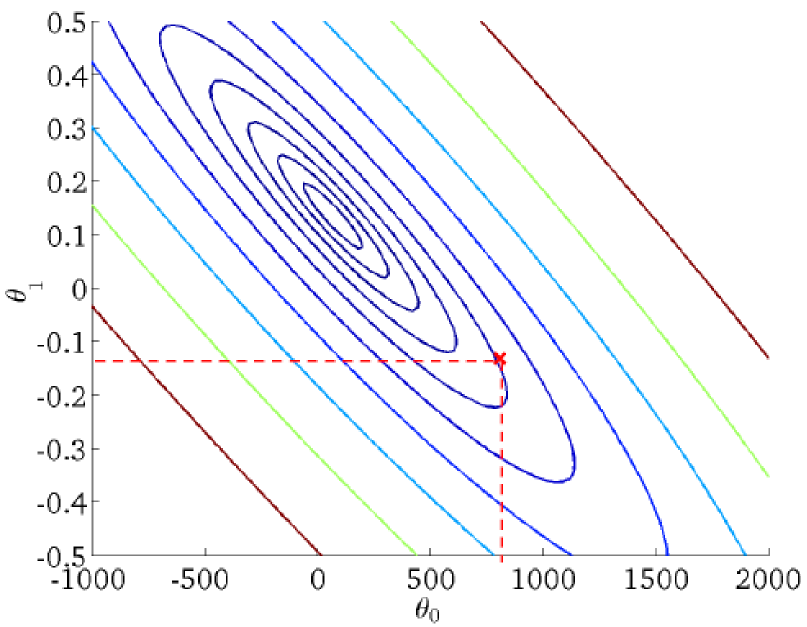
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

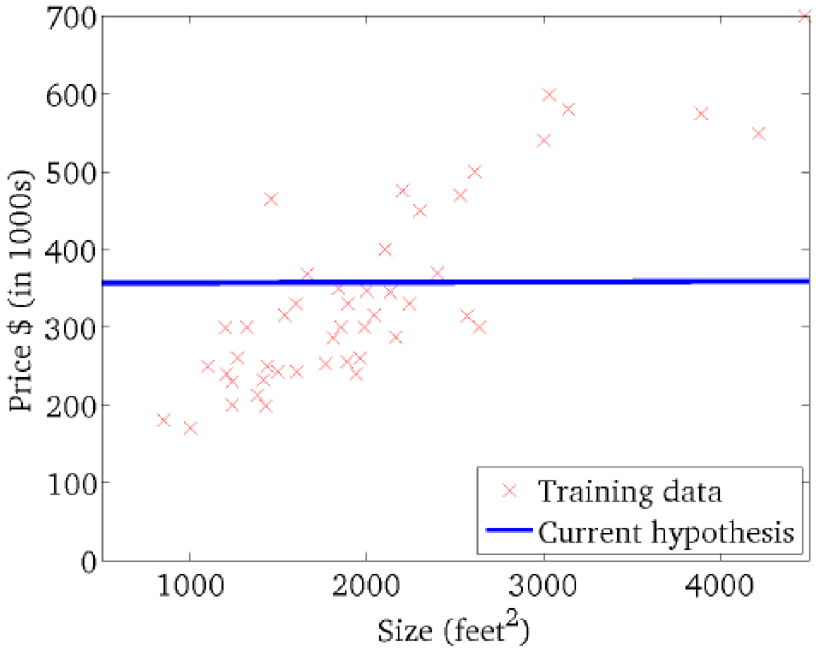
(function of the parameters θ_0, θ_1)



Relation between h and J

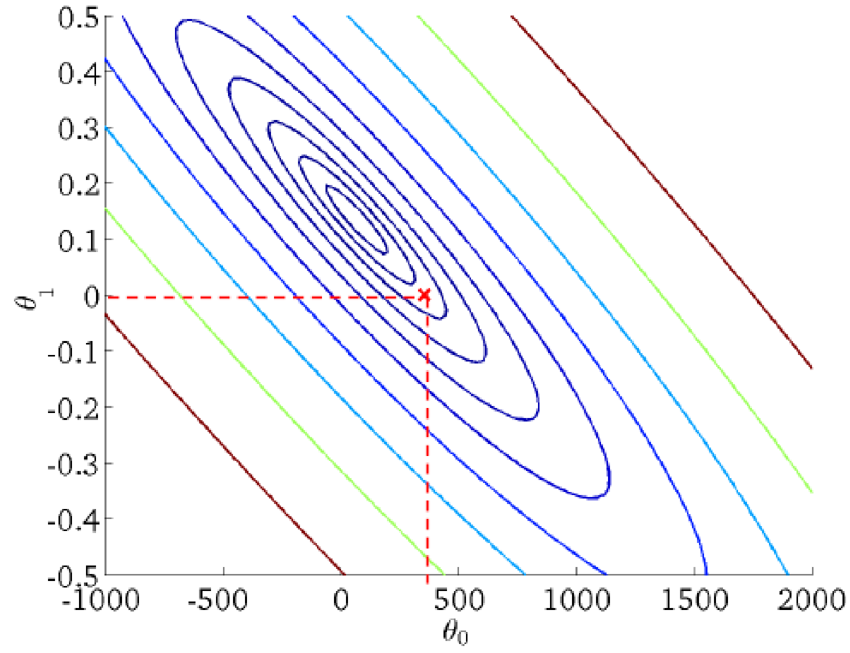
$$h_\theta(x)$$

(for fixed θ_0, θ_1, this is a function of x)



$$J(\theta_0, \theta_1)$$

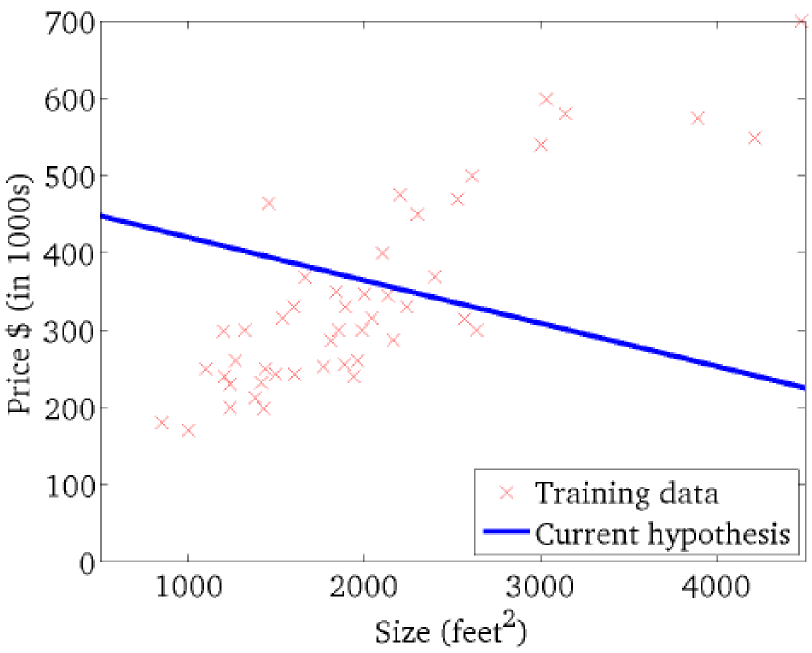
(function of the parameters θ_0, θ_1)



Relation between h and J

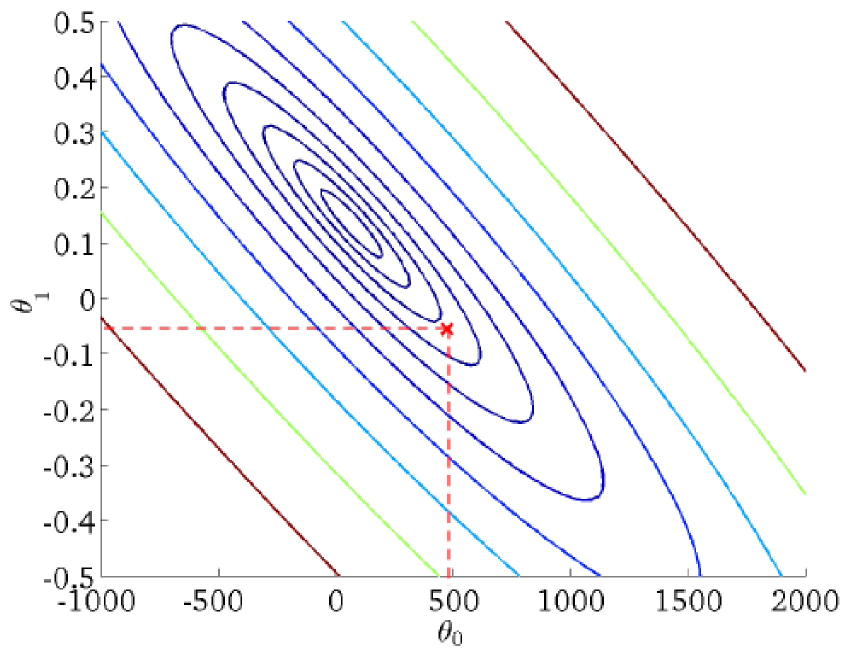
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

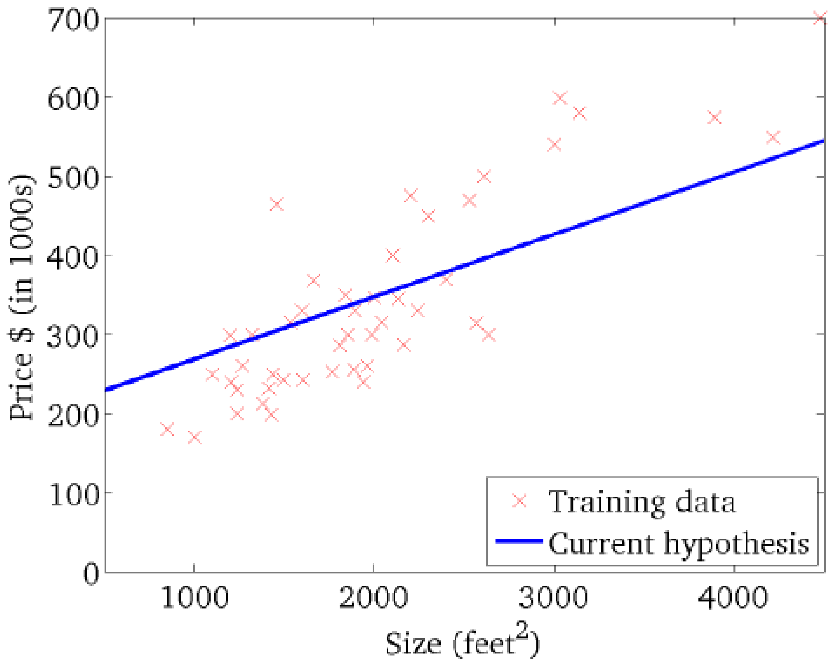
(function of the parameters θ_0, θ_1)



Relation between h and J

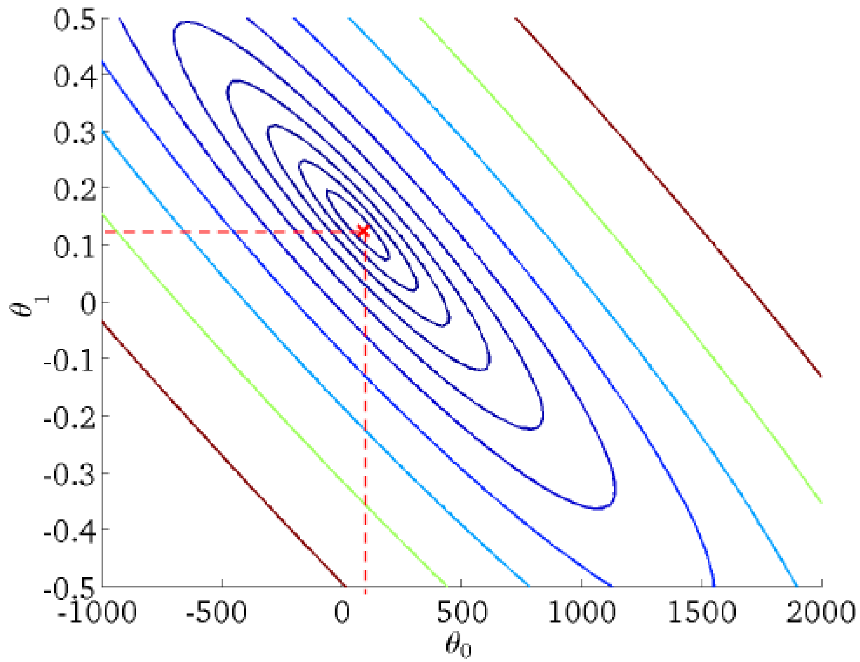
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



How to find optimal model parameters θ to minimize cost J?

Simple linear regression

- Dataset $x^{(i)} \in R, y^{(i)} \in R, h_\theta(x) = \theta_0 + \theta_1 x$
- $J(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)^2$ loss

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{2}{n}\sum_{i=1}^{n}\left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right) = 0$$

$$\frac{\partial J(\theta)}{\partial \theta_1} = \frac{2}{n}\sum_{i=1}^{n} x^{(i)}\left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right) = 0$$

- Solution of min loss

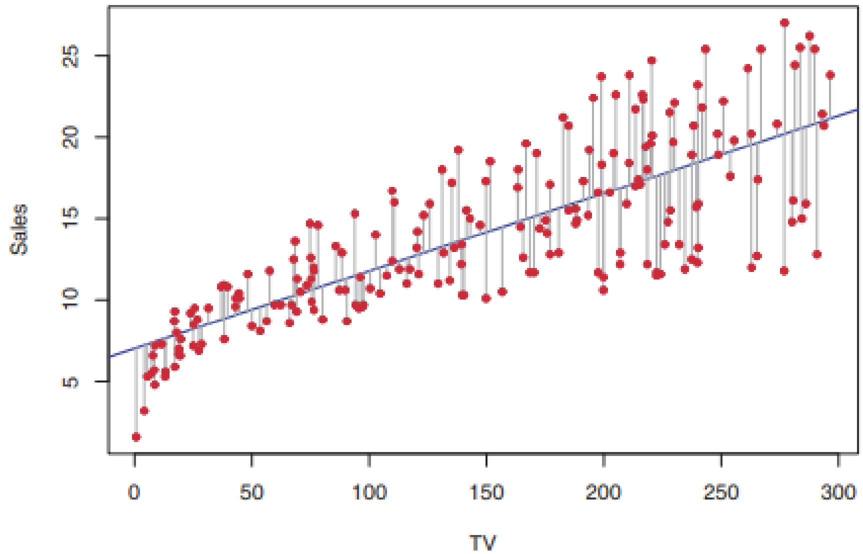
$$-\ \theta_0 = \bar{y} - \theta_1 \bar{x}$$

$$-\ \theta_1 = \frac{\sum (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum\left(x^{(i)} - \bar{x}\right)^2}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x^{(i)}}{n}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y^{(i)}}{n}$$

Hypothesis Testing

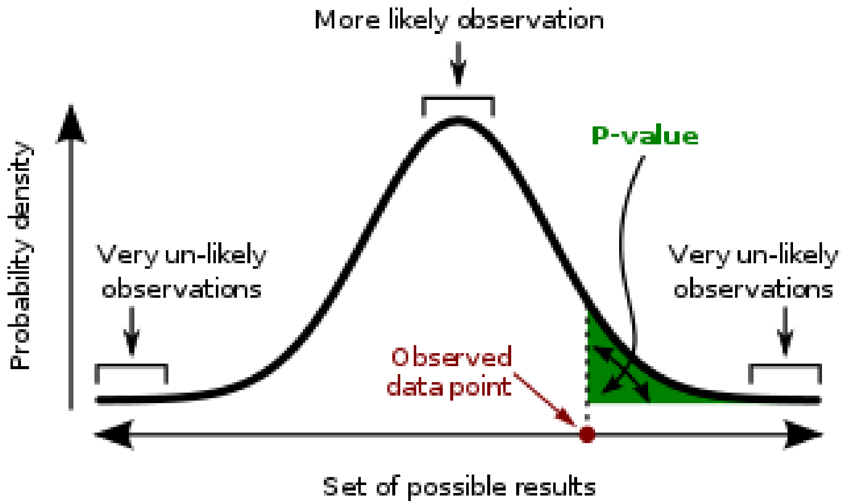


- Is there some relationship between X and Y?
- Null Hypothesis: No relationship between X and Y
    - Equivalent to $\theta_1 = 0$
- Alternative Hypothesis: There is relationship between X and Y
    - Equivalent to $\theta_1 \neq 0$

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Reject Null Hypothesis

Hypothesis Testing



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

- If the p value is very small, the null hypothesis can be rejected!
- If the p value is large, we cannot say anything about the null hypothesis (whether it's true or not)

How Well Does the Model Fit?

- Residual Sum of Squares
  - $RSS = \sum R^{(i)} = \sum\left[y^{(i)} - (\theta_0 + \theta_1 x^{(i)})\right]^2$
- Total Sum of Squares
  - $TSS = \sum\left[y^{(i)} - \bar{y}\right]^2$
  - Total variance of the response
- Proportion of variability in Y that can be explained using X
  - $R^2 = 1 - \frac{RSS}{TSS} \in [0,1]$
- Correlation between feature and response

$$\mathrm{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

For simple regression R^2 is equal to Cor(X,Y)!

Lab example

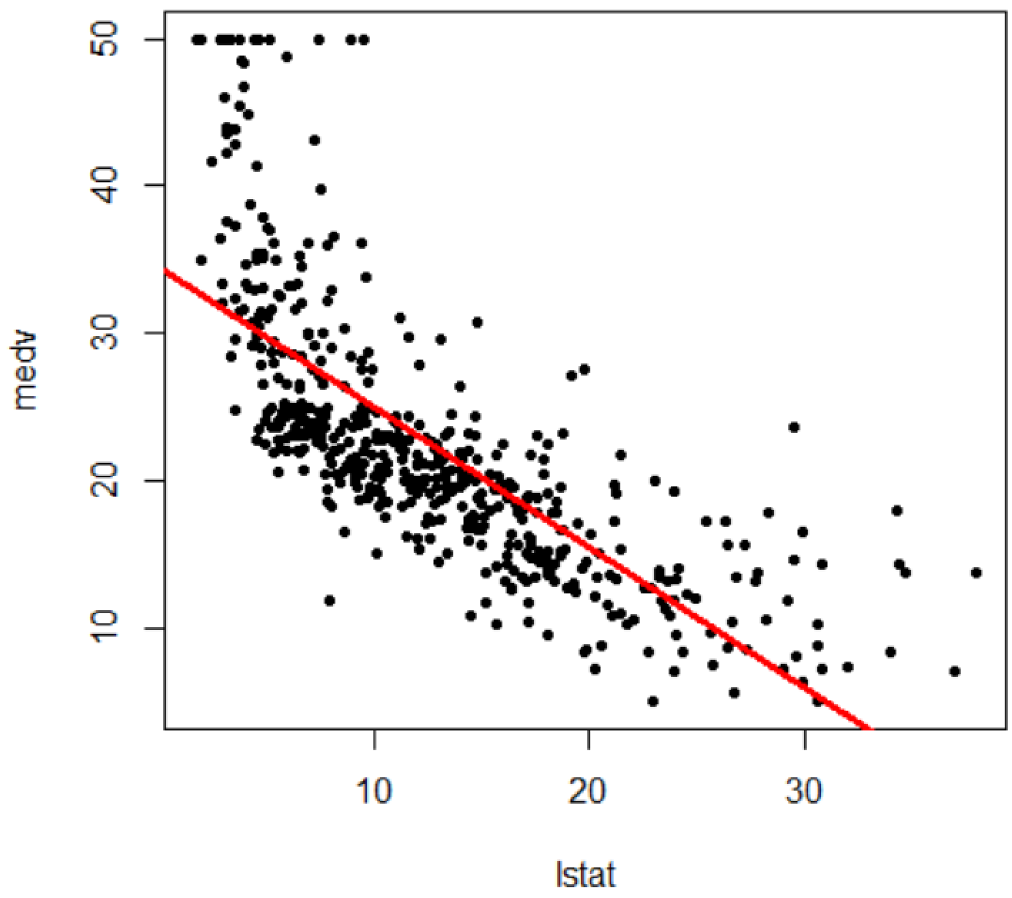
```
>
> library(MASS)
> fix(Boston)
>
```

Data Editor

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
2	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
3	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
6	0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
7	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
8	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1
9	0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5
10	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9
11	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15
12	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9
13	0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7
14	0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4
15	0.63796	0	8.14	0	0.538	6.096	84.5	4.4619	4	307	21	380.02	10.26	18.2
16	0.62739	0	8.14	0	0.538	5.834	56.5	4.4986	4	307	21	395.62	8.47	19.9

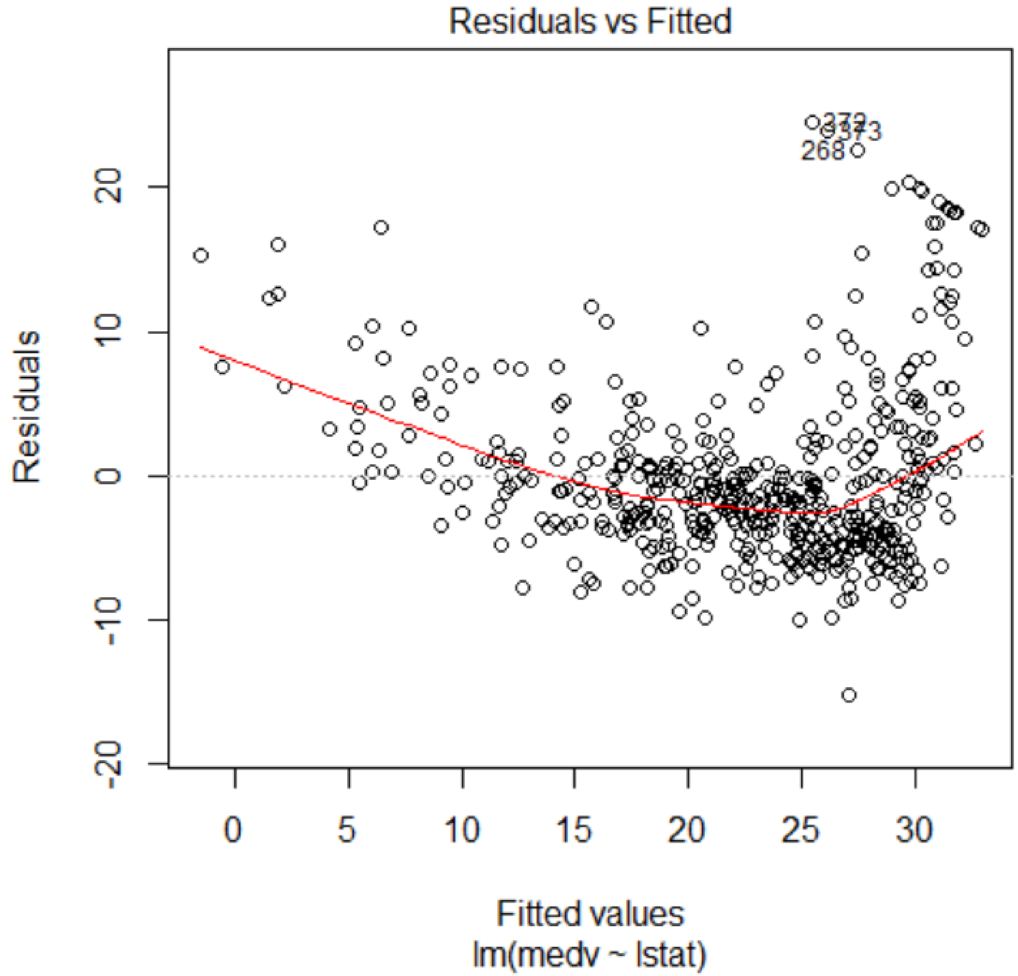
Simple LR

```
>
> lm.fit=lm(medv~lstat,data=Boston)
> plot(lstat,medv,pch=20)
> abline(lm.fit,lwd=3,col="red")
>
```



Residual plot

```
> plot(predict(lm.fit), residuals(lm.fit))
>
> plot(lm.fit, which=1)
>
```



Estimated responses

Simple LR

```
>
> lm.fit=lm(medv~lstat,data=Boston)
> summary(lm.fit)

Call:
lm(formula = medv ~ lstat, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-15.168  -3.990  -1.318   2.034  24.500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.216 on 504 degrees of freedom
Multiple R-squared:  0.5441,    Adjusted R-squared:  0.5432
F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

Coef not zero!

$$RSE = \sqrt{MSE}$$

R^2 measures linear relationship between X and Y

Multiple LR

```
> lm.fit=lm(medv~nox+rm+lstat+ptratio+rad+dis,data=Boston)
> summary(lm.fit)

Call:
lm(formula = medv ~ nox + rm + lstat + ptratio + rad + dis, d$

Residuals:
     Min       1Q   Median       3Q      Max
-12.8663  -3.1525  -0.5509   1.9870  27.1748

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  40.61722    5.07480   8.004 8.53e-15 ***
nox         -20.16431    3.57710  -5.637 2.90e-08 ***
rm            4.04507    0.41938   9.645  < 2e-16 ***
lstat        -0.59197    0.04846 -12.217  < 2e-16 ***
ptratio      -1.12748    0.12634  -8.924  < 2e-16 ***
rad           0.05399    0.03682   1.466    0.143
dis          -1.19580    0.16840  -7.101 4.29e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.988 on 499 degrees of freedom
Multiple R-squared:  0.7093,    Adjusted R-squared:  0.7058
F-statistic:   203 on 6 and 499 DF,  p-value: < 2.2e-16
```

Review linear regression

- Simple linear regression: one dimension
- Multiple linear regression: multiple dimensions
- Minimize cost (loss) function
  - MSE: average of squared residuals
- Can derive closed-form solution

$$- \theta_0 = \bar{y} - \theta_1 \bar{x}$$

$$- \theta_1 = \frac{\sum (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum (x^{(i)} - \bar{x})^2}$$

Acknowledgements

- Slides made using resources from:
  - Andrew Ng
  - Eric Eaton
  - David Sontag
- Thanks!