Multiple Linear Regression

Introduction

- In simple linear regression we studied the relationship between one explanatory variable and one response variable.
- Now, we look at situations where several explanatory variables works together to explain the response Y.

Example

- In a study of direct operating cost, Y, for 67 branch offices of consumer finance charge, four independent variables were considered:
  - X_1: Average size of loan outstanding during the year,
  - X_2 : Average number of loans outstanding,
  - X_3 : Total number of new loan applications processed, and
  - X_4 : Office salary scale index.
- The model for this example is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 X_3 + \beta_4 x_4 + \varepsilon$$

Multiple Linear Regression Model

- General regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

  - $\beta_0, \beta_1, \ldots, \beta_k$ are parameters
  - $X_1, X_2, \ldots, X_k$ are known constants
  - ε , the error terms are independent $N(o, \sigma^2)$

Estimating the parameters of the model

- The values of the regression parameters β_i are not known. We estimate them from data.
- As in the simple linear regression case, we use the least-squares method to fit a linear function

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

to the data.

Estimating the parameters of the model

- As in simple linear regression the least-squares method chooses the b's that minimize the sum of squares of the residuals.

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Closed form solution does exist (We will cover it later).

Estimating the parameters of the model

- The estimate of β_i is b_i and it indicates the change in the mean response per unit increase in X_i when the rest of the independent variables in the model are held constant.
- The parameters β_i are frequently called partial regression coefficients because they reflect the partial effect of one independent variable when the rest of independent variables are included in the model and are held constant

Estimating the parameters of the model

- The observed variability of the responses about this fitted model is measured by the variance

$$S^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

and the regression standard error

$$s = \sqrt{s^2}$$

Analysis of Variance Table

- The basic idea of the regression ANOVA table are the same in simple and multiple regression.
- The sum of squares decomposition and the associated degrees of freedom are:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$$SST = SSR + SSE$$

- *df*: $n - 1 = k + (n - k - 1)$

Analysis of Variance Table

Source	Sum of Squares	df	Mean Square	F-test
Regression	SSR	k	MSR= SSR/k	MSR/MSE
Error	SSE	n-k-1	MSE= SSE/n-k-1	
Total	SST	n-1		

F-test for the overall fit of the model

- To test the statistical significance of the regression relation between the response variable y and the set of variables x_1,…, x_k, i.e. to choose between the alternatives:

$$H_0: \ \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_a: \ \text{not all } \beta_i (i = 1, \cdots k) \text{ equal zero}$$

- We use the test statistic:

$$F = \frac{MSR}{MSE}$$

F-test for the overall fit of the model

- The decision rule at significance level α is:
  - Reject H_0 if $F > F(\alpha; k, n-k-1)$
  - Where the critical value $F(\alpha, k, n-k-1)$ can be found from an F-table.
- Note that when k=1, this test reduces to the F-test for testing in simple linear regression whether or not $\beta_1 = 0$

Interval estimation of βi

- For our regression model, we have:

$$\frac{b_i - \beta_i}{s(b_i)} \quad \text{has a t}-\text{distribution with n}-\text{k}-1 \text{ degrees of freedom}$$

- Therefore, an interval estimate for β_i with 1- α confidence coefficient is:

$$b_i \pm t(\frac{\alpha}{2}; n-k-1)\sqrt{\frac{MSE}{\sum(x-\bar{x})^2}}$$

T-test of βi

- To test:

$$H_0: \ \beta_i = 0$$
$$H_a: \beta_i \neq 0$$

- We may use the test statistic:

$$t = \frac{b_i}{s(b_i)}$$

- Reject H_0 if

$$t > t(\frac{\alpha}{2}; n-k-1) \quad or$$
$$t < -t(\frac{\alpha}{2}; n-k-1)$$

Multiple regression model Building

- Often, we have many explanatory variables, and our goal is to use these to explain the variation in the response variable.
- A model using just a few of the variables often predicts about as well as the model using all the explanatory variables.

Multiple regression model Building

- We may find that the reciprocal of a variable is a better choice than the variable itself, or that including the square of an explanatory variable improves prediction.
- We may find that the effect of one explanatory variable may depends upon the value of another explanatory variable. We account for this situation by including interaction terms.
- The simplest way to construct an interaction term is to multiply the two explanatory variables together.
- How can we find a good model?

Selecting the best Regression equation.

- When we are having a list of potentially useful independent variables, we can go over them and decide which variables can be screened out. We can look for variables that:
  - May not be fundamental to the problem
  - May be subject to large measurement error
  - May effectively duplicate another independent variable in the list.

Selecting the best Regression Equation.

- An automatic search procedure that develops sequentially the subset of explanatory variables to be included in the regression model is called *stepwise procedure*.
- It was developed to economize on computational efforts.
- It will end with the identification of a *single* regression model as “best”.

Example: Sales Forecasting

- Sales Forecasting
  - Multiple regression is a popular technique for predicting product sales with the help of other variables that are likely to have a bearing on sales.
- Example
  - The growth of cable television has created vast new potential in the home entertainment business. The following table gives the values of several variables measured in a random sample of 20 local television stations which offer their programming to cable subscribers. A TV industry analyst wants to build a statistical model for predicting the number of subscribers that a cable station can expect.

Example: Sales Forecasting

- Y = Number of cable subscribers (SUSCRIB)
- X_1 = Advertising rate which the station charges local advertisers for one minute of prim time space (ADRATE)
- X_2 = Kilowatt power of the station's non-cable signal (KILOWATT)
- X_3 = Number of families living in the station's area of dominant influence (ADI), a geographical division of radio and TV audiences (APIPOP)
- X_4 = Number of competing stations in the ADI (COMPETE)

Example: Sales Forecasting

- The sample data are fitted by a multiple regression model using Excel program.
- The marginal t-test provides a way of choosing the variables for inclusion in the equation.
- The fitted Model is

$$\text{SUBSCRIBE} = \beta_0 + \beta_1 \times \text{ADRATE} + \beta_2 \times \text{APIPOP} + \beta_3 \times COMPETE + \beta_4 \times SIGNAL$$

Example: Sales Forecasting

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.884267744					
R Square	0.781929444					
Adjusted R Square	0.723777295					
Standard Error	142.9354188					
Observations	20					
ANOVA						
	df	*SS*	*MS*	*F*	*ignificance F*	
Regression	4	1098857.84	274714.4601	13.44626923	7.52E-05	
Residual	15	306458.0092	20430.53395			
Total	19	1405315.85				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	51.42007002	98.97458277	0.51952803	0.610973806	-159.539	262.3795
AD_Rate	-0.267196347	0.081055107	-3.296477624	0.004894126	-0.43996	-0.09443
Signal	-0.020105139	0.045184758	-0.444954014	0.662706578	-0.11641	0.076204
APIPOP	0.440333955	0.135200486	3.256896248	0.005307766	0.152161	0.728507
Compete	16.230071	26.47854322	0.61295181	0.549089662	-40.2076	72.66778

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

Example: Sales Forecasting

- Do we need all the four variables in the model?
- Based on the partial t-test, the variables signal and compete are the least significant variables in our model.
- Let’s drop the least significant variables one at a time.

Example: Sales Forecasting

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.882638739					
R Square	0.779051144					
Adjusted R Square	0.737623233					
Standard Error	139.3069743					
Observations	20					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	1094812.92	364937.64	18.80498277	1.69966E-05	
Residual	16	310502.9296	19406.4331			
Total	19	1405315.85				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	51.31610447	96.4618242	0.531983558	0.602046756	-153.1737817	255.806
AD_Rate	-0.259538026	0.077195983	-3.36206646	0.003965102	-0.423186162	-0.09589
APIPOP	0.433505145	0.130916687	3.311305499	0.004412929	0.15597423	0.711036
Compete	13.92154404	25.30614013	0.550125146	0.589831583	-39.72506442	67.56815

Example: Sales Forecasting

- The variable Compete is the next variable to get rid of.

Example: Sales Forecasting

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.8802681					
R Square	0.774871928					
Adjusted R Square	0.748386273					
Standard Error	136.4197776					
Observations	20					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	1088939.802	544469.901	29.2562866	3.13078E-06	
Residual	17	316376.0474	18610.35573			
Total	19	1405315.85				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	96.28121395	50.16415506	1.919322948	0.07188916	-9.556049653	202.1184776
AD_Rate	-0.254280696	0.075014548	-3.389751739	0.003484198	-0.41254778	-0.096013612
APIPOP	0.495481252	0.065306012	7.587069489	7.45293E-07	0.357697418	0.633265086

Example: Sales Forecasting

- All the variables in the model are statistically significant, therefore our final model is:

$$\text{SUBSCRIBE} = 96.28 - 0.25 \times \text{ADRATE} + 0.495 \times \text{APIPOP}$$

Interpreting the Final Model

- What is the interpretation of the estimated parameters.
- Is the association positive or negative?
- Does this make sense intuitively, based on what the data represents?
- What other variables could be confounders?
- Are there other analysis that you might consider doing? New questions raised?

Multicollinearity

- In multiple regression analysis, one is often concerned with the nature and significance of the relations between the explanatory variables and the response variable.
- Questions that are frequently asked are:
  - What is the relative importance of the effects of the different independent variables?
  - What is the magnitude of the effect of a given independent variable on the dependent variable?

Multicollinearity

- Can any independent variable be dropped from the model because it has little or no effect on the dependent variable?
- Should any independent variables not yet included in the model be considered for possible inclusion?

- Simple answers can be given to these questions if
  - The independent variables in the model are uncorrelated among themselves.
  - They are uncorrelated with any other independent variables that are related to the dependent variable but omitted from the model.

Multicollinearity

- When the independent variables are correlated among themselves, *multicollinearity* or colinearity among them is said to exist.
- In many non-experimental situations in business, economics, and the social and biological sciences, the independent variables tend to be correlated among themselves.
- For example, in a regression of family food expenditures on the variables: family income, family savings, and the age of head of household, the explanatory variables will be correlated among themselves.

Multicollinearity

- Further, the explanatory variables will also be correlated with other socioeconomic variables not included in the model that do affect family food expenditures, such as family size.

Multicollinearity

- Some key problems that typically arise when the explanatory variables being considered for the regression model are highly correlated among themselves are:
  1. Adding or deleting an explanatory variable changes the regression coefficients.
  2. The estimated standard deviations of the regression coefficients become large when the explanatory variables in the regression model are highly correlated with each other.
  3. The estimated regression coefficients individually may not be statistically significant even though a definite statistical relation exists between the response variable and the set of explanatory variables.

Variance Inflation Factor (VIF)

- A formal method of detecting the presence of multicollinearity that is widely used is *Variance Inflation Factor*.
  - It measures how much the variances of the estimated regression coefficients are inflated as compared to when the independent variables are not linearly related.

$$VIF_j = \frac{1}{1 - R_j^2}, \quad j = 1,2, \cdots k$$

  - Is the coefficient of determination from the regression of the jth independent variable on the remaining k-1 independent variables. R_j^2

Variance Inflation Factor (VIF)

- A VIF near 1 suggests that multicollinearity is not a problem for the independent variables.
  - Its estimated coefficient and associated t value will not change much as the other independent variables are added or deleted from the regression equation.
- A VIF much greater than 1 indicates the presence of multicollinearity. A maximum VIF value in excess of 10 is often taken as an indication that the multicollinearity may be unduly influencing the least square estimates.
  - the estimated coefficient attached to the variable is unstable and its associated t statistic may change considerably as the other independent variables are added or deleted.

Multicollinearity Diagnostics

- The simple correlation coefficient between all pairs of explanatory variables (i.e., X_1, X_2, …, X_k) is helpful in selecting appropriate explanatory variables for a regression model and is also critical for examining multicollinearity.
- While it is true that a correlation very close to +1 or –1 does suggest multicollinearity, it is not true (unless there are only two explanatory variables) to infer multicollinearity does not exist when there are no high correlations between any pair of explanatory variables.

Example: Sales Forecasting

Pearson Correlation Coefficients, N = 20
Prob > |r| under H0: Rho=0

	SUBSCRIB	ADRATE	KILOWATT	APIPOP	COMPETE
SUBSCRIB	1.00000	-0.02848	0.44762	0.90447	0.79832
SUBSCRIB		0.9051	0.0478	<.0001	<.0001
ADRATE	-0.02848	1.00000	-0.01021	0.32512	0.34147
ADRATE	0.9051		0.9659	0.1619	0.1406
KILOWATT	0.44762	-0.01021	1.00000	0.45303	0.46895
KILOWATT	0.0478	0.9659		0.0449	0.0370
APIPOP	0.90447	0.32512	0.45303	1.00000	0.87592
APIPOP	<.0001	0.1619	0.0449		<.0001
COMPETE	0.79832	0.34147	0.46895	0.87592	1.00000
COMPETE	<.0001	0.1406	0.0370	<.0001	

Example : Sales Forecasting

$$\text{SUBSCRIBE} = 51.42 - 0.27 \times \text{ADRATE} - .02 \times \text{SIGNAL} + 0.44 \times \text{APIPOP} + 16.23 \times \text{COMPETE}$$

$$\text{SUBSCRIBE} = 51.32 - 0.26 \times \text{ADRATE} + 0.43 \times \text{APIPOP} + 13.92 \times \text{COMPETE}$$

$$\text{SUBSCRIBE} = 96.28 - 0.25 \times \text{ADRATE} + 0.495 \times \text{APIPOP}$$

Example:Sales Forecasting

$$\text{APIPOP} = \beta_0 + \beta_1 \times \text{SIGNAL} + \beta_2 \times \text{ADRATE} + \beta_3 \times COMPETE$$

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.878054					
R Square	0.770978					
Adjusted R Square	0.728036					
Standard Error	264.3027					
Observations	20					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	3762601	1254200	17.9541	2.25472E-05	
Residual	16	1117695	69855.92			
Total	19	4880295				
	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-472.685	139.7492	-3.38238	0.003799	-768.9402258	-176.43
Compete	159.8413	28.29157	5.649786	3.62E-05	99.86587622	219.8168
ADRATE	0.048173	0.149395	0.322455	0.751283	-0.268529713	0.364876
Signal	0.037937	0.083011	0.457012	0.653806	-0.138038952	0.213913

Example: Sales Forecasting

$$\text{Compete} = \beta_0 + \beta_1 \times \text{ADRATE} + \beta_2 \times \text{APIPOP} + \beta_3 \times SIGNAL$$

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.882936					
R Square	0.779575					
Adjusted R Square	0.738246					
Standard Error	1.34954					
Observations	20					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	103.0599	34.35329	18.86239	1.66815E-05	
Residual	16	29.14013	1.821258			
Total	19	132.2				
	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3.10416	0.520589	5.96278	1.99E-05	2.000559786	4.20776
ADRATE	0.000491	0.000755	0.649331	0.525337	-0.001110874	0.002092
Signal	0.000334	0.000418	0.799258	0.435846	-0.000552489	0.001221
APIPOP	0.004167	0.000738	5.649786	3.62E-05	0.002603667	0.005731

Example:Sales Forecasting

$$\text{Signal} = \beta_0 + \beta_1 \times \text{ADRATE} + \beta_2 \times \text{APIPOP} + \beta_3 \times COMPETE$$

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.512244					
R Square	0.262394					
Adjusted R Square	0.124092					
Standard Error	790.8387					
Observations	20					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	3559789	1186596	1.897261	0.170774675	
Residual	16	10006813	625425.8			
Total	19	13566602				
	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5.171093	547.6089	0.009443	0.992582	-1155.707711	1166.05
APIPOP	0.339655	0.743207	0.457012	0.653806	-1.235874129	1.915184
Compete	114.8227	143.6617	0.799258	0.435846	-189.7263711	419.3718
ADRATE	-0.38091	0.438238	-0.86919	0.397593	-1.309935875	0.548109

Example:Sales Forecasting

$$\text{ADRATE} = \beta_0 + \beta_1 \times \text{Signal} + \beta_2 \times \text{APIPOP} + \beta_3 \times COMPETE$$

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.399084					
R Square	0.159268					
Adjusted R Square	0.001631					
Standard Error	440.8588					
Observations	20					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	3	589101.7	196367.2	1.010346	0.413876018	
Residual	16	3109703	194356.5			
Total	19	3698805				
	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	253.7304	298.6063	0.849716	0.408018	-379.2865355	886.7474
Signal	-0.11837	0.136186	-0.86919	0.397593	-0.407073832	0.170329
APIPOP	0.134029	0.415653	0.322455	0.751283	-0.747116077	1.015175
Compete	52.3446	80.61309	0.649331	0.525337	-118.5474784	223.2367

Example: Sales Forecasting

- VIF calculation Results:

Variable	R- Squared	VIF
ADRATE	0.159268	1.19
COMPETE	0.779575	4.54
SIGNAL	0.262394	1.36
APIPOP	0.770978	4.36

- There is no significant multicollinearity.

Qualitative Independent Variables

- Many variables of interest in business, economics, and social and biological sciences are not quantitative but are qualitative.
- Examples of qualitative variables are gender (male, female), purchase status (purchase, no purchase), and type of firms.
- Qualitative variables can also be used in multiple regression.

Qualitative Independent Variables

- An economist wished to relate the speed with which a particular insurance innovation is adopted (y) to the size of the insurance firm (x_1) and the type of firm.
- The dependent variable is measured by the number of months elapsed between the time the first firm adopted the innovation and and the time the given firm adopted the innovation.
- The first independent variable, size of the firm, is quantitative, and measured by the amount of total assets of the firm.
- The second independent variable, type of firm, is qualitative and is composed of two classes-Stock companies and mutual companies.

Indicator variables

- Indicator, or dummy variables are used to determine the relationship between qualitative independent variables and a dependent variable.
- Indicator variables take on the values 0 and 1.
- For the insurance innovation example, where the qualitative variable has two classes, we might define the indicator variable x_2 as follows:

$$x_2 = \begin{matrix} 1 & \text{if stock company} \\ 0 & \text{otherwise} \end{matrix}$$

Indicator variables

- A qualitative variable with c classes will be represented by c-1 indicator variables.
- A regression function with an indicator variable with two levels (c = 2) will yield two estimated lines.

Interpretation of Regression Coefficients

- In our insurance innovation example, the regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- Where:
  - $x_1 =$ size of firm
  - $x_2 = \begin{cases} 1 & \text{if stock company} \\ 0 & \text{otherwise} \end{cases}$

Interpretation of Regression Coefficients

- To understand the meaning of the regression coefficients in this model, consider first the case of mutual firm. For such a firm, $x_2 = 0$ and we have:

$$\hat{y}_i = b_0 + b_1 x_1 + b_2(0) = b_0 + b_1 x_1 \quad Mutual \rightleftarrows \quad firms$$

- For a stock firm $x_2 = 1$ and the response function is:

$$\hat{y}_i = b_0 + b_1 x_1 + b_2(1) = (b_0 + b_2) + b_1 x_1 \quad Stock \quad firms$$

Interpretation of Regression Coefficients

- The response function for the mutual firms is a straight line, with y intercept β_0 and slope β_1.
- For stock firms, this also is a straight line, with the same slope β_1 but with y intercept $\beta_0+\beta_2$.
- With reference to the insurance innovation example, the mean time elapsed before the innovation is adopted is linear function of size of firm (x_1), with the same slope $\beta_1$for both types of firms.

Interpretation of Regression Coefficients

- β_2 indicates how much lower or higher the response function for stock firm is than the one for the mutual firm.
- β_2 measures the differential effect of type of firms.
- In general, β_2 shows how much higher (lower) the mean response line is for the class coded 1 than the line for the class coded 0, for any level of x_1.

Example: Insurance Innovation Adoption

- Here is the data set for the insurance innovation example:

Months Elapsed	Size	type of firm	Type
17	151	0	Mutual
26	92	0	Mutual
21	175	0	Mutual
30	31	0	Mutual
22	104	0	Mutual
0	277	0	Mutual
12	210	0	Mutual
19	120	0	Mutual
4	290	0	Mutual
16	238	1	Stock
28	164	1	Stock
15	272	1	Stock
11	295	1	Stock
38	68	1	Stock
31	85	1	Stock
21	224	1	Stock
20	166	1	Stock
13	305	1	Stock
30	124	1	Stock
14	246	1	Stock

Example: Insurance Innovation Adoption

- Fitting the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Where

- $x_1 =$ size of firm
- $$x_2 = \begin{matrix} 1 & \text{if stock company} \\ 0 & \text{otherwise} \end{matrix}$$

  - fitted response function is:

$$\hat{y} = 33.87 - .1061x_1 + 8.77x_2$$

Example: Insurance Innovation Adoption

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.95993655					
R Square	0.92147818					
Adjusted R Square	0.91224031					
Standard Error	2.78630562					
Observations	20					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	1548.820517	774.4103	99.75016	4.04966E-10	
Residual	17	131.979483	7.763499			
Total	19	1680.8				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	33.8698658	1.562588138	21.67549	8E-14	30.57308841	37.16664321
Size	-0.10608882	0.007799653	-13.6017	1.45E-10	-0.122544675	-0.089632969
type of firm	8.76797549	1.286421264	6.815789	3.01E-06	6.053860079	11.4820909

Example: Insurance Innovation Adoption

- The fitted response function is:

$$\hat{y} = 33.87 - .1061x_1 + 8.77x_2$$

- Stock firms response function is:

$$\hat{y} = (33.87 + 8.77) - .1061x_1$$

- Mutual firms response function is:

$$\hat{y} = 33.87 - .1061x_1$$

Moving Beyond Linearity

The linearity assumption is good in many machine learning problems.

However, there are other methods that offer a lot of flexibility, without losing the ease and interpretability of linear models we will cover few of them.

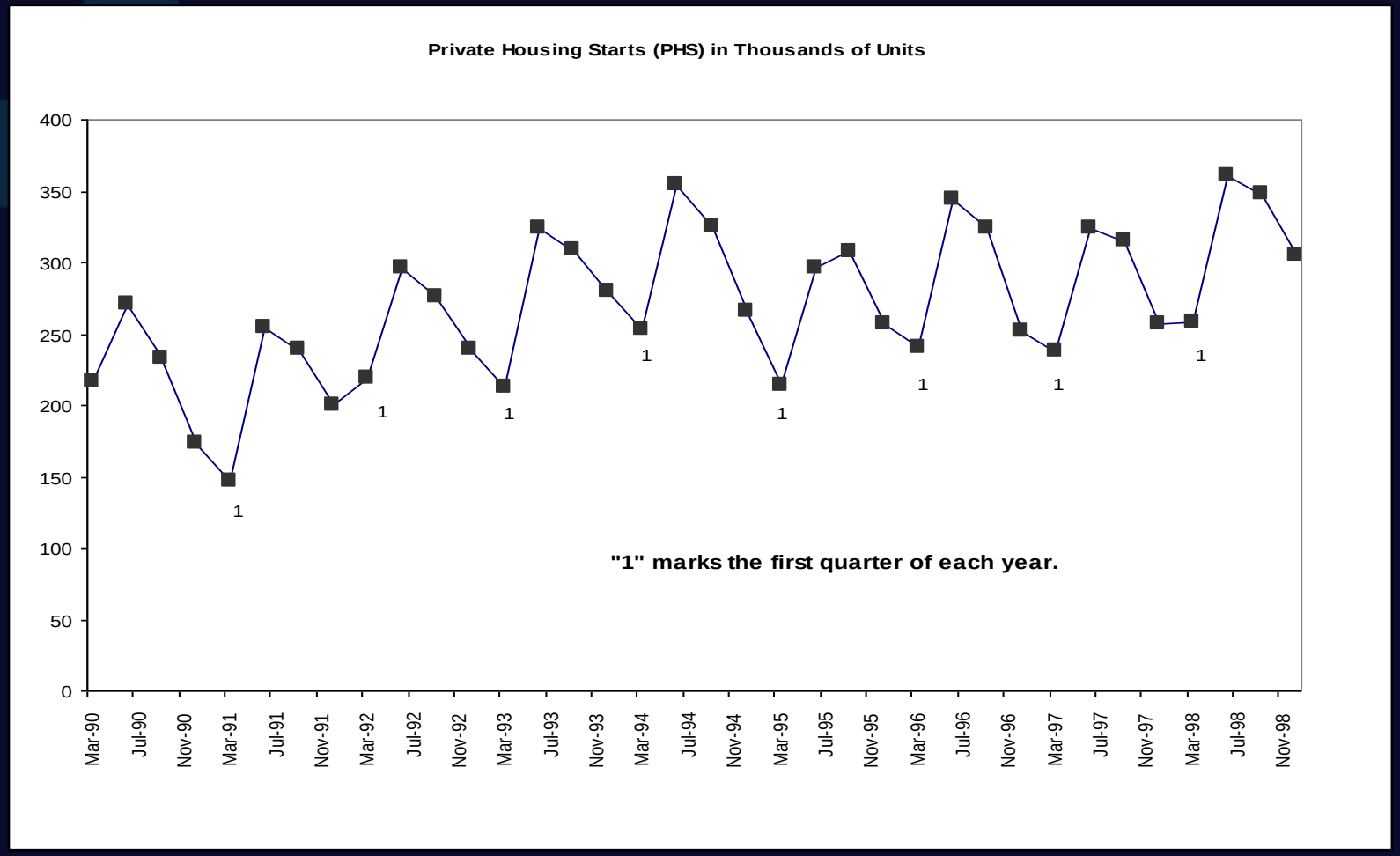
Accounting for Seasonality in a Multiple regression Model

- Seasonal Patterns are not easily accounted for by the typical causal variables that we use in regression analysis.
- An indicator variable can be used effectively to account for seasonality in our time series data.
- The number of seasonal indicator variables to use depends on the data.
- If we have p periods in our data series, we can not use more than P-1 seasonal indicator variables.

Example: Private Housing Starts (PHS)

- Housing starts in the United States measured in thousands of units. These data are plotted for 1990 Q1 through 1999 Q4.
- There are typically few housing starts during the first quarter of the year (January, February, March); there is usually a big increase in the second quarter of (April, May, June), followed by some decline in the third quarter (July, August, September), and further decline in the fourth quarter (October, November, December).

Example: Private Housing Starts (PHS)



Example: Private Housing Starts (PHS)

- To Account for and measure this seasonality in a regression model, we will use three dummy variables: Q2 for the second quarter, Q3 for the third quarter, and Q4 for the fourth quarter. These will be coded as follows:
  - Q2 = 1 for all second quarters and zero otherwise.
  - Q3 = 1 for all third quarters and zero otherwise
  - Q4 = 1 for all fourth quarters and zero otherwise.

Example: Private Housing Starts (PHS)

- Data for private housing starts (PHS), the mortgage rate (MR), and these seasonal indicator variables are shown in the following slide.
- Since we have assigned dummy variables for the second, third, and fourth quarters, the first quarter is the base quarter for our regression model.
- Note that any quarter could be used as the base, with indicator variables to adjust for differences in other quarters.

Example: Private Housing Starts (PHS)

PERIOD	PHS	MR	Q2	Q3	Q4
31-Mar-90	217	10.1202	0	0	0
30-Jun-90	271.3	10.3372	1	0	0
30-Sep-90	233	10.1033	0	1	0
31-Dec-90	173.6	9.9547	0	0	1
31-Mar-91	146.7	9.5008	0	0	0
30-Jun-91	254.1	9.5265	1	0	0
30-Sep-91	239.8	9.2755	0	1	0
31-Dec-91	199.8	8.6882	0	0	1
31-Mar-92	218.5	8.7098	0	0	0
30-Jun-92	296.4	8.6782	1	0	0
30-Sep-92	276.4	8.0085	0	1	0
31-Dec-92	238.8	8.2052	0	0	1
31-Mar-93	213.2	7.7332	0	0	0
30-Jun-93	323.7	7.4515	1	0	0
30-Sep-93	309.3	7.0778	0	1	0
31-Dec-93	279.4	7.0537	0	0	1
31-Mar-94	252.6	7.2958	0	0	0
30-Jun-94	354.2	8.4370	1	0	0
30-Sep-94	325.7	8.5882	0	1	0
31-Dec-94	265.9	9.0977	0	0	1
31-Mar-95	214.2	8.8123	0	0	0
30-Jun-95	296.7	7.9470	1	0	0
30-Sep-95	308.2	7.7012	0	1	0
31-Dec-95	257.2	7.3508	0	0	1
31-Mar-96	240	7.2430	0	0	0
30-Jun-96	344.5	8.1050	1	0	0
30-Sep-96	324	8.1590	0	1	0
31-Dec-96	252.4	7.7102	0	0	1
31-Mar-97	237.8	7.7905	0	0	0
30-Jun-97	324.5	7.9255	1	0	0
30-Sep-97	314.6	7.4692	0	1	0
31-Dec-97	256.8	7.1980	0	0	1
31-Mar-98	258.4	7.0547	0	0	0
30-Jun-98	360.4	7.0938	1	0	0
30-Sep-98	348	6.8657	0	1	0
31-Dec-98	304.6	6.7633	0	0	1
31-Mar-99	294.1	6.8805	0	0	0
30-Jun-99	377.1	7.2037	1	0	0
30-Sep-99	355.6	7.7990	0	1	0
31-Dec-99	308.1	7.8338	0	0	1

Example: Private Housing Starts (PHS)

- The regression model for private housing starts (PHS) is:

$$PHS = \beta_0 + \beta_1(MR) + \beta_2(Q2) + \beta_3(Q3) + \beta_4(Q4)$$

- In this model we expect b_1 to have a negative sign, and we would expect b_2, b_3, b_4 all to have positive signs.

Example: Private Housing Starts (PHS)

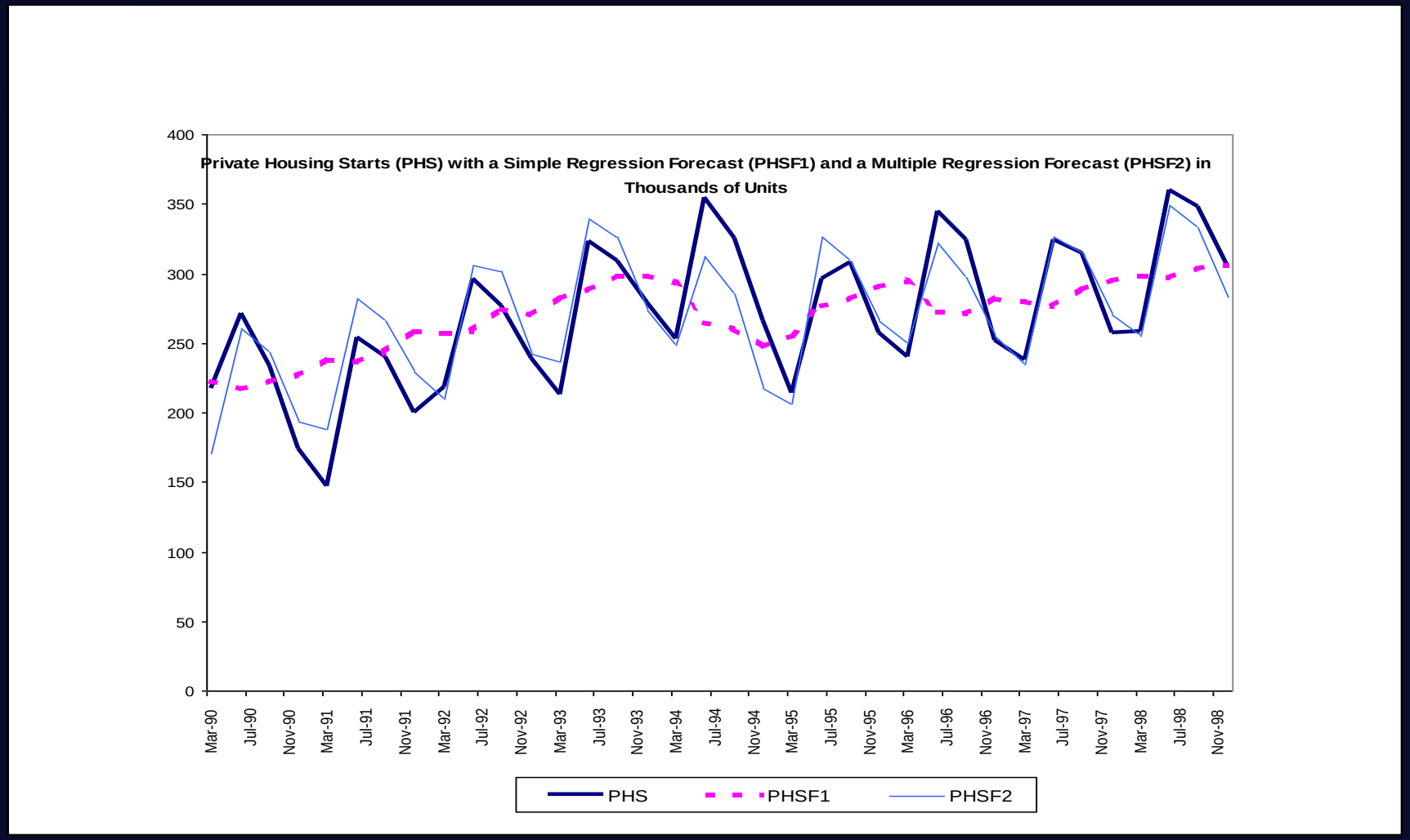
SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.885398221					
R Square	0.78393001					
Adjusted R Square	0.759236296					
Standard Error	26.4498851					
Observations	40					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	4	88837.93624	22209.48406	31.74613731	3.33637E-11	
Residual	35	24485.87476	699.5964217			
Total	39	113323.811				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	473.0650749	35.54169837	13.31014264	2.93931E-15	400.9115031	545.2186467
MR	-30.04838192	4.257226391	-7.058206249	3.21421E-08	-38.69102153	-21.40574231
Q2	95.74106935	11.84748487	8.081130334	1.6292E-09	71.689367	119.7927717
Q3	73.92904763	11.82881519	6.249911462	3.62313E-07	49.91524679	97.94284847
Q4	20.54778131	11.84139803	1.73524961	0.091495355	-3.491564078	44.5871267

Example: Private Housing Starts (PHS)

- Use the prediction equation to make a forecast for each of the fourth quarter of 1999.
- Prediction equation:

$$\hat{PHS} = 473.06 - 30.05(MR) + 95.74(Q2) + 73.93(Q3) + 20.55(Q4)$$

Example: Private Housing Starts (PHS)

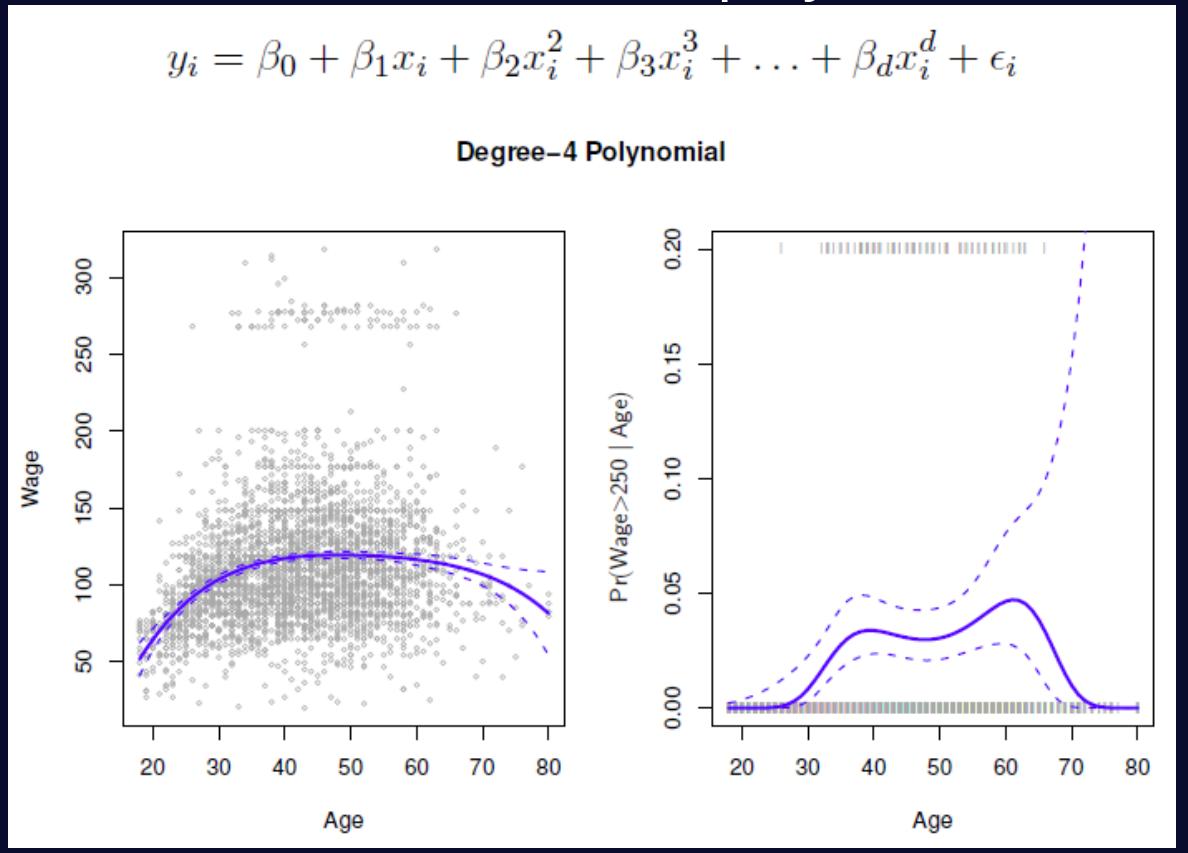


המרכז הבינלאומי
ללימודי הייטק וחדשנות

Polynomial Regression

- Replace the standard linear model with a polynomial function:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i$$



Polynomial Regression

- For large enough degree *d*, a polynomial regression allows us to produce an extremely non-linear curve.

- We do this by creating new variables $X_1 = X$, $X_2 = X^2$, etc. and then treat as multiple linear regression.

- In general, we are not really interested in the coefficients, but instead the fitted function values at any value x_0:

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4$$

Polynomial Regression

- Since $\hat{f}(x_0)$ is a linear function of the $\hat{\beta}_j$, we can get a simple expression for *pointwise variances* $\mathrm{Var}[\hat{f}(x_0)]$ at any value of x_0.

- In practice, It is common not to use *d* greater than 3 or 4.

Step Functions

- Using polynomial functions of the features as predictor in a linear model imposes a *global* structure on the non-linear function of X.
- To avoid imposing such a global structure, we can create transformations of a variable by cutting the variable into distinct regions.
- In particular, we use *step functions* to break the range of *X* into bins, where we fit a different constant in each bin.

Step Functions

- This amounts to converting a continuous variable into an *ordered categorical variable*.
- In greater detail, we create cut points (or knots) c_1, c_2,….,c_K in the range of X and then construct $K + 1$ new variables:

$$
\begin{aligned}
C_0(X) &= I(X < c_1), \\
C_1(X) &= I(c_1 \leq X < c_2), \\
C_2(X) &= I(c_2 \leq X < c_3), \\
&\vdots \\
C_{K-1}(X) &= I(c_{K-1} \leq X < c_K), \\
C_K(X) &= I(c_K \leq X),
\end{aligned}
$$

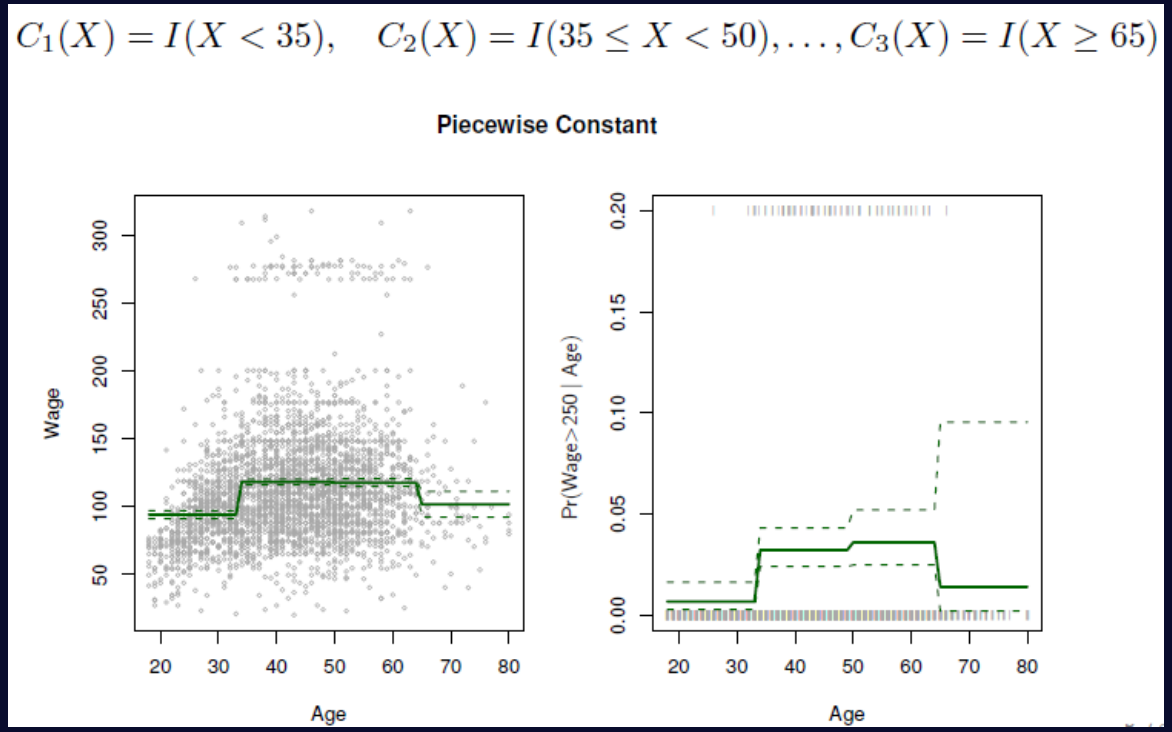
where $I(.)$ is an *indicator function* that returns a 1 if the condition is true and 0 otherwise.

Step Functions

- Note that for any value of X, $C_0(X) + C_1(X) + \ldots + C_K(X) = 1$, since X must be exactly in one of the $K + 1$ intervals.
- We then use OLS estimation to fit a linear model using these $K + 1$ new variables:
- For a given value of X, at most one of $C_1, C_2, \ldots, C_K$ can be non-zero.
- β_j represents the average increase in the response for X in $c_j \leq X \leq c_{j+1}$ relative to $X < c_1$.

Step Functions

$$C_1(X) = I(X < 35), \quad C_2(X) = I(35 \leq X < 50), \ldots, C_3(X) = I(X \geq 65)$$



- Unless there are natural breakpoints in the predictors, piece-wise constant functions can miss the action.

Basis Functions

- Polynomial and piece-wise constant regression models are special cases of a *basis function* approach.
- The idea is to have at hand a family of functions or transformations that can be applied to a variable X: $b_1(X),\ldots,b_K(X)$
- Instead of fitting a linear model in X, we fit the following model:

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \ldots + \beta_K b_K(x_i) + \epsilon_i$$

- Note that the basis functions $b_1(.),\ldots,b_K(.)$ are fixed and known.

Basis Functions

- For polynomial regression, the basis functions are $b_j(x_i) = x_i^j$
- For piece-wise constant functions, the basis functions are
$$b_j(x_i) = I(c_j \leq x_i \leq c_{j+1})$$
- Note that we can use OLS to estimate the unknown regression coefficients.
- Thus, all of the inference tools for linear models (standard errors, F-statistics, etc.) are available in this setting.

Regression Splines

- Regression splines are a flexible class of basis functions that extend upon the polynomial regressions and piece-wise constant regression approaches.
- They involve dividing the range of *X* into *K* distinct regions; within each region, a polynomial function is fit to the data.
- These polynomials are constrained so that they join *smoothly* at the region boundaries (or *knots*).
- Provided that the interval is divided into enough regions, this can produce an extremely flexible regression function.

Piecewise Polynomials

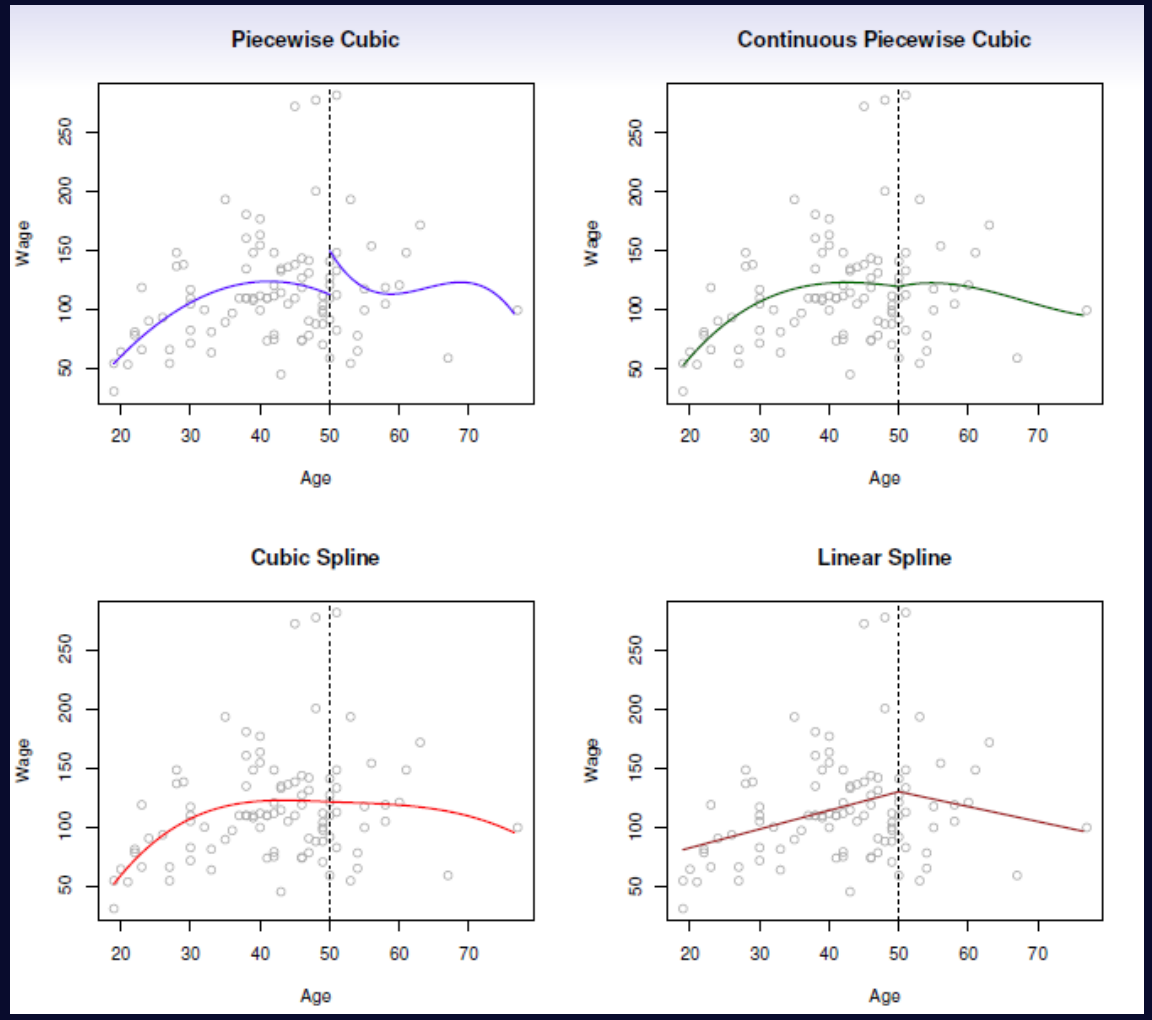
- Instead of fitting a high-degree polynomial over the entire range of *X*, *piece-wise polynomial regression* involves fitting separate low-degree polynomials over different regions of *X*.
- Here, the beta coefficients differ in different parts of the range of *X*; the points where the coefficients change are called *knots*.
- Example: A piecewise cubic polynomial with a single knot at a point *c* takes the following form:

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \epsilon_i & \text{if } x_i < c; \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \epsilon_i & \text{if } x_i \geq c. \end{cases}$$

Piecewise Polynomials

- Each of the polynomial functions can be fit using OLS applied to simple functions of the original predictor.
- Using more knots leads to a more flexible piecewise polynomial.
- If general, if we place *K* different knots through the range of *X*, then we end up fitting *K* + 1 different polynomials.
- It is better to add *constraints* to the polynomials (e.g. continuity).
- *Splines* have the maximum amount of continuity

Piecewise Polynomials



Piecewise Polynomials

- Each constraint that we impose effectively frees up one degree of freedom, by reducing the complexity of the resulting piecewise polynomial fit.

- The general definition of a degree-*d* spline is that it is a piecewise degree-*d* polynomial, with continuity in derivatives up to degree *d* – 1 at each knot.

- Thus, a linear spline is obtained by fitting a line in each region of the predictor space defined by the knots, requiring continuity at each knot.

Linear Splines

A linear spline with knots at $\xi_k,\ k = 1, \ldots, K$ *is a piecewise linear polynomial continuous at each knot.*

We can represent this model as

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i,$$

where the b_k are *basis functions*.

$$\begin{aligned} b_1(x_i) &= x_i \\ b_{k+1}(x_i) &= (x_i - \xi_k)_+, \quad k = 1, \ldots, K \end{aligned}$$

Here the $()_+$ means *positive part*; i.e.

$$(x_i - \xi_k)_+ = \begin{cases} x_i - \xi_k & \text{if } x_i > \xi_k \\ 0 & \text{otherwise} \end{cases}$$

Cubic Splines

A cubic spline with knots at ξ_k, $k = 1, \ldots, K$ is a piecewise cubic polynomial with continuous derivatives up to order 2 at each knot.

Again we can represent this model with truncated power basis functions

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i,$$

$$\begin{aligned} b_1(x_i) &= x_i \\ b_2(x_i) &= x_i^2 \\ b_3(x_i) &= x_i^3 \\ b_{k+3}(x_i) &= (x_i - \xi_k)_+^3, \quad k = 1, \ldots, K \end{aligned}$$

where

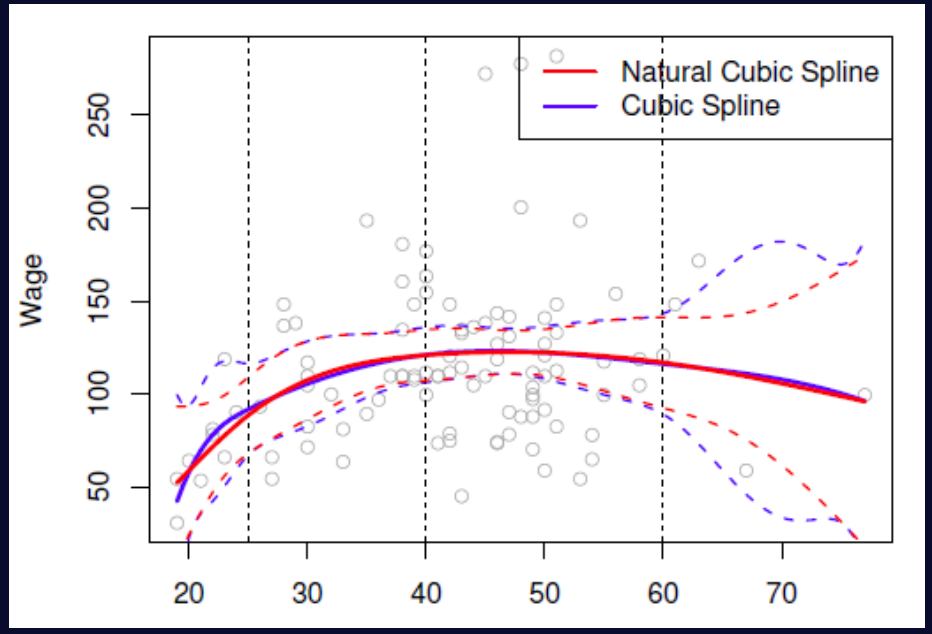
$$(x_i - \xi_k)_+^3 = \begin{cases} (x_i - \xi_k)^3 & \text{if } x_i > \xi_k \\ 0 & \text{otherwise} \end{cases}$$

Natural Splines

- A *natural spline* is a regression spline with additional boundary constraints.
- The function is required to be linear at the boundary (in the region where X is smaller than the smallest knot, or larger than the largest knot).
- This additional constraint means that natural splines generally produce more stable estimates at the boundaries.

Natural Splines

- A natural cubic spline extrapolates linearly beyond the boundary knots. This adds 4 = 2 x 2 extra constraints, and allows us to put more internal knots for the same degrees of freedom as a regular cubic spline.



Choosing the Location of Knots

- The regression spline is most flexible in regions that contain a lot of knots, because in those regions the polynomial coefficients can change rapidly.
- One option is to place more knots in places where we feel the function might vary most rapidly, and to place fewer knots where it seems more stable.
- In practice, it is common to place knots in a uniform fashion. For example, one strategy is to decide K, the number of knots, and then place them at appropriate quantiles of the observed X.

Choosing the Number of Knots

- One option is to try out different numbers of knots and see which produces the best looking curve.
- However, a more objective approach is to use cross-validation.
- The procedure is repeated for different number of knots *K*; then the value of *K* giving the smallest RSS is chosen.
- Splines allow us to place more knots, and hence flexibility, over regions where the function *f* seems to be changing rapidly, and fewer knots where *f* appears more stable.

Smoothing Splines

- We create regression splines by specifying a set of knots, producing a sequence of basis functions, and then use OLS to estimate the spline coefficients.

- What we really want is a function *g* that makes RSS small and *smooth*. Thus, consider the following criterion for fitting a smooth function *g*(*x*) to some data (known as a *smoothing spline*):

$$\underset{g \in \mathcal{S}}{\text{minimize}} \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

where λ is a nonnegative tuning parameter.

Smoothing Splines

- The first term is a loss function (RSS), which tries to make $g(x)$ match the data at each x_i.
- The second term is a *roughness penalty* and controls how wiggly $g(x)$ is; this is modulated by the tuning parameter λ.
- The larger the value of λ, the smoother g will be. The smaller the value of λ, the more wiggly the function.
- As $\lambda \rightarrow \infty$, the function $g(x)$ becomes linear.

Smoothing Splines

- It turns out that the solution is a natural cubic spline, with a knot at every unique value of x_i.
- The tuning parameter λ controls the level of roughness (i.e. the effective degrees of freedom).
- Smoothing splines avoid the knot-selection issue, leaving a single λ to be chosen.
- The vector of *n* fitted values (for a particular choice of λ) can be written as $\hat{\boldsymbol{g}}_\lambda = \boldsymbol{S}_\lambda \boldsymbol{y}$, where $\boldsymbol{S}_\lambda$ is the *n* x *n* smoother matrix.

Smoothing Splines

- The *effective degrees of freedom* are trace($\boldsymbol{S}_\lambda$), which equals:

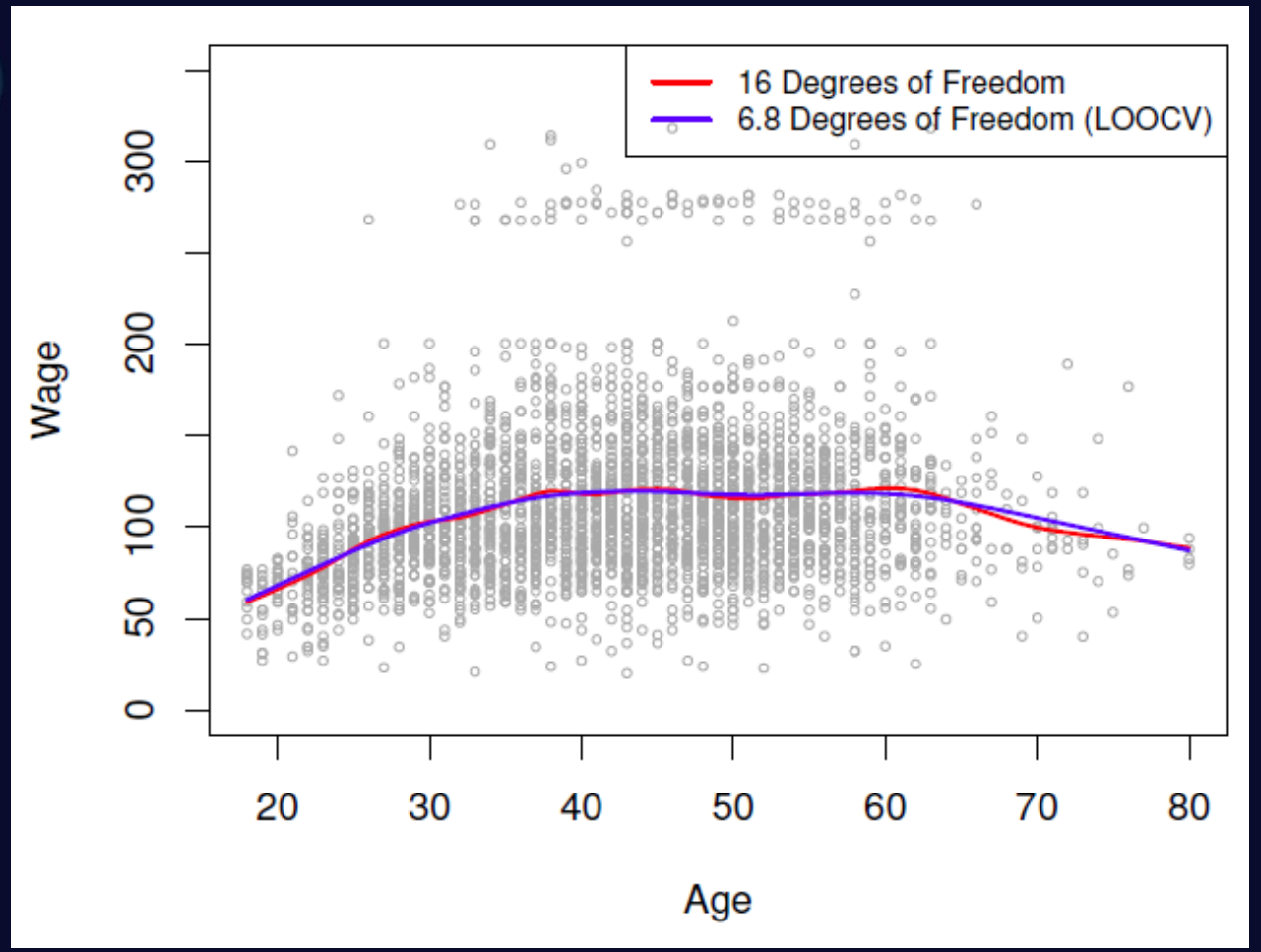
$$df_\lambda = \sum_{i=1}^{n} \{\mathbf{S}_\lambda\}_{ii}$$

which is the sum of the diagonal elements of the matrix $\boldsymbol{S}_\lambda$.

- We can find λ. The error is given as:

$$\mathrm{RSS}_{cv}(\lambda) = \sum_{i=1}^{n} (y_i - \hat{g}_\lambda^{(-i)}(x_i))^2 = \sum_{i=1}^{n} \left[\frac{y_i - \hat{g}_\lambda(x_i)}{1 - \{\mathbf{S}_\lambda\}_{ii}} \right]^2$$

Smoothing Splines



Local Regression

- *Local regression* is a different approach for fitting flexible non-linear functions, which involves computing the fit at a target point x_0 using only the nearby training observations.

- It is a *memory-based* procedure because we need all the training data each time we wish to compute a prediction.

- The *span* plays a role like that of the tuning parameter λ in smoothing splines; it controls the flexibility of the non-linear fit.

Local Regression

- The smaller the value of the span *s*, the more *local* and wiggly will be our fit.
- A very large value of *s* will lead to a global fit to the data using all of the training observations.
- We can use cross-validation to choose *s* or specify it directly.
- Another choice to be made includes how to define the weighting function *K*, and whether to fit a linear, constant, or quadratic regression.

Local Regression

Algorithm 7.1 *Local Regression At* $X = x_0$

1. Gather the fraction $s = k/n$ of training points whose x_i are closest to x_0.

2. Assign a weight $K_{i0} = K(x_i, x_0)$ to each point in this neighborhood, so that the point furthest from x_0 has weight zero, and the closest has the highest weight. All but these k nearest neighbors get weight zero.

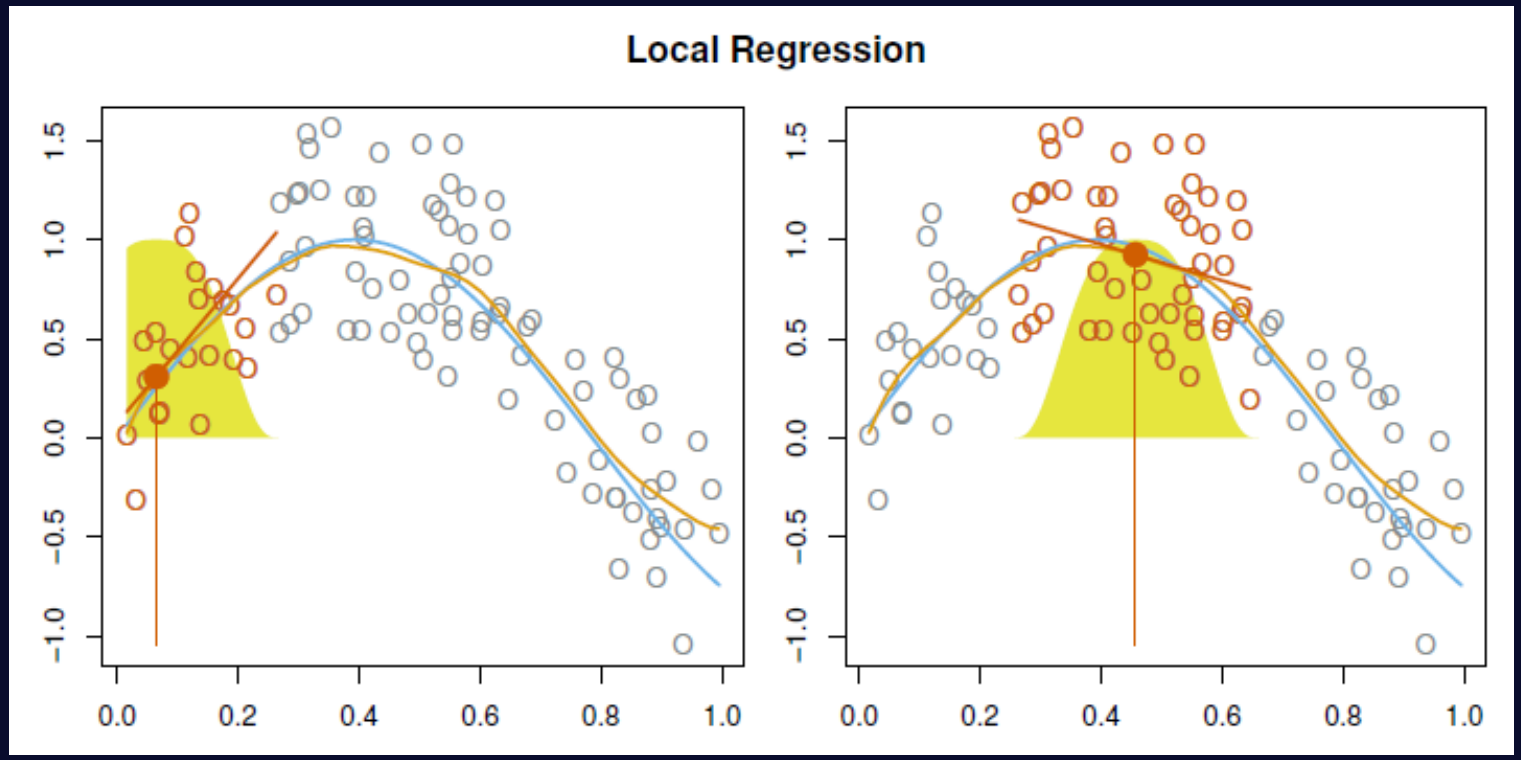
3. Fit a *weighted least squares regression* of the y_i on the x_i using the aforementioned weights, by finding $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

$$\sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2. \tag{7.14}$$

4. The fitted value at x_0 is given by $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Local Regression

- With a sliding weight function, we fit separate linear fits over the range of *X* by weighted least squares.



Generalized Additive Models

- *Generalized additive models* (GAMs) allow for flexible nonlinearities in several variables, but retains the additive structure of linear models.
- GAMs can be applied with both quantitative and qualitative responses.
- In particular, we replace each linear component of the multiple linear regression model with a (smooth) non-linear function.

Generalized Additive Models

$$
\begin{aligned}
y_i &= \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \\
&= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i.
\end{aligned}
$$

- It is called an *additive* model because we calculate a separate f_j for each X_j, and then add together all of their contributions.

- We can use any of the previously discussed methods (smoothing splines, local regression, polynomial regression, etc.) as building blocks for fitting an additive model.

Generalized Additive Models

- GAMs allow us to fit a non-linear f_j to each X_j, so that we can automatically model non-linear relationships that standard linear regression will miss.

- This means that we do not need to manually try out many different transformations on each variable individually.

- The non-linear fits can potentially make more accurate predictions for the response Y.

Generalized Additive Models

- Because the model is additive, we can still examine the effect of each X_j on Y individually while holding all of the other variables fixed.
- Thus, if we are interested in inference, GAMs provide a useful representation.
- The smoothness of the function f_j for the variable X_j can be summarized via degrees of freedom.

Generalized Additive Models

- The main limitation of GAMs is that the model is restricted to be additive.
- With many variables, important interactions can be missed. However, we can manually add interaction terms to the GAM model by including additional predictors of the form $X_j \times X_k$.
- Although we have not yet covered classification problems, note that GAMs can also be used in situations where Y is qualitative.

Influential Points

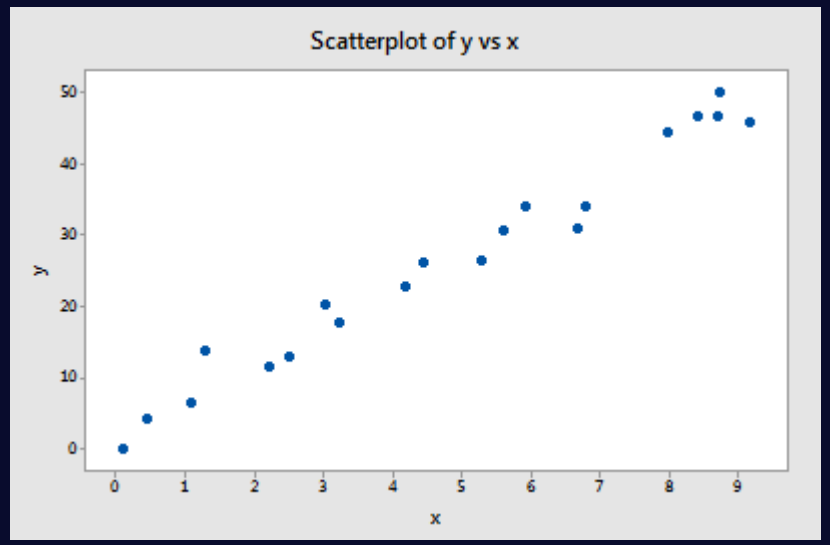
- Data observations can potentially be influential in different ways.
- If an observation has a response value that is very different from the predicted value based on a model, then that observation is called an outlier.
- If an observation has a particularly unusual combination of predictor values (e.g., one predictor has a very different value for that observation compared with all the other data observations), then that observation is said to have high leverage.

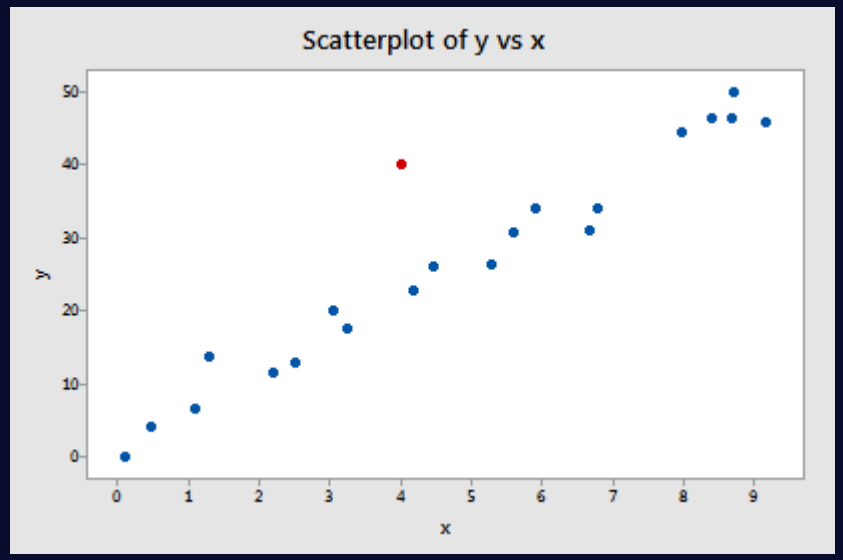
Influential Points

- An observation can be both an outlier and have high leverage.
- Due to the potential in affecting the significance and the performance of the regression model, it is important to know how to detect outliers and high leverage data points.
- After detecting such points, we need to determine whether or not the points actually have an undue influence on our model.

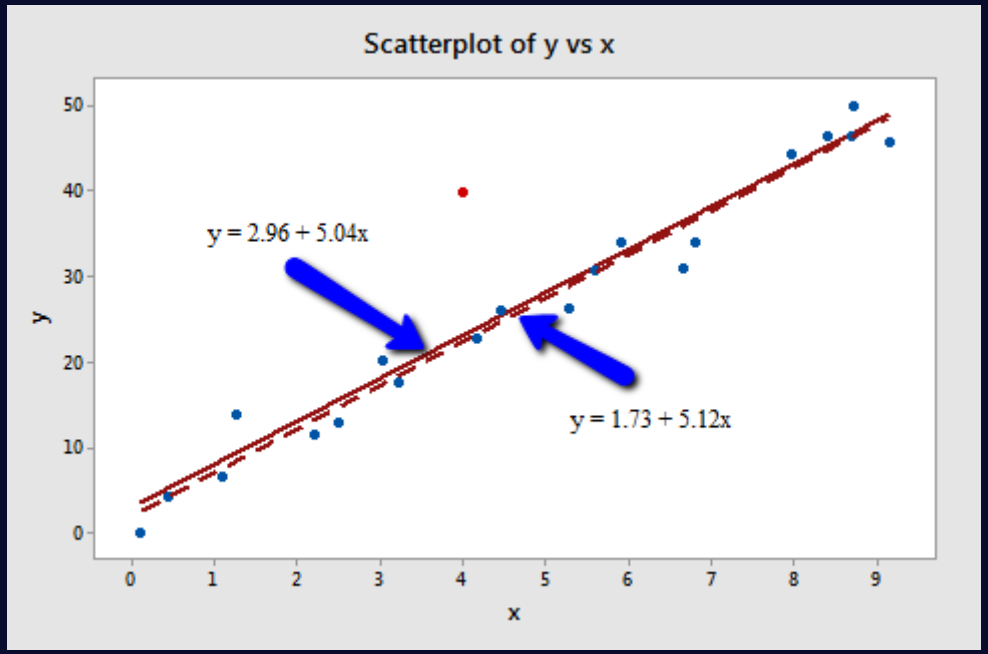
Influential Points - Outliers

An **outlier** is a data point whose response *y* does not follow the general trend of the rest of the data





Influential Points - Outliers



Influential Points - Outliers

When red data point is included

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
4.71075	91.01%	90.53%	89.61%

Model Summary

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.96	2.01	1.47	0.157	
X	5.037	0.363	13.86	0.000	1.00

Regression Equation

$y = 2.96 + 5.037 x$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

Influential Points - Outliers

When red data point is not included

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.59199	97.32%	97.17%	96.63%

Model Summary

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.73	1.12	1.55	0.140	
X	5.117	0.200	25.55	0.000	1.00

Regression Equation

$y = 1.73 + 5.117 x$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

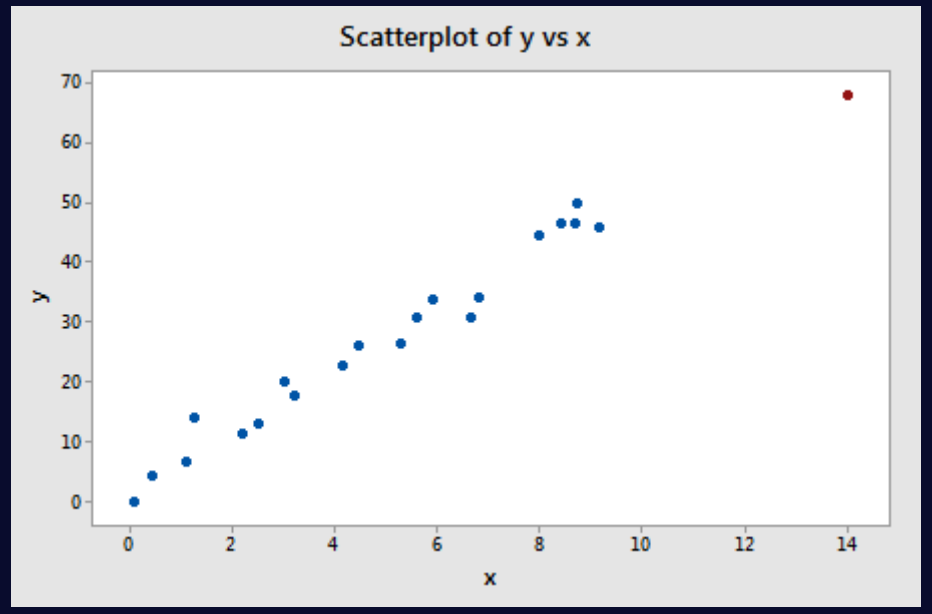
Influential Points - Outliers

- The r^2 value has decreased slightly, but the relationship between y and x would still be deemed strong.
- The standard error, which is used in calculating our confidence interval for , is larger when the red data point is included, thereby increasing the width of our confidence interval. increases, not because the data point is influential in any way.
- In each case, the P-value for testing is less than 0.001. In either case, we can conclude that there is sufficient evidence at the 0.05 level to conclude that, in the population, x is related to y.

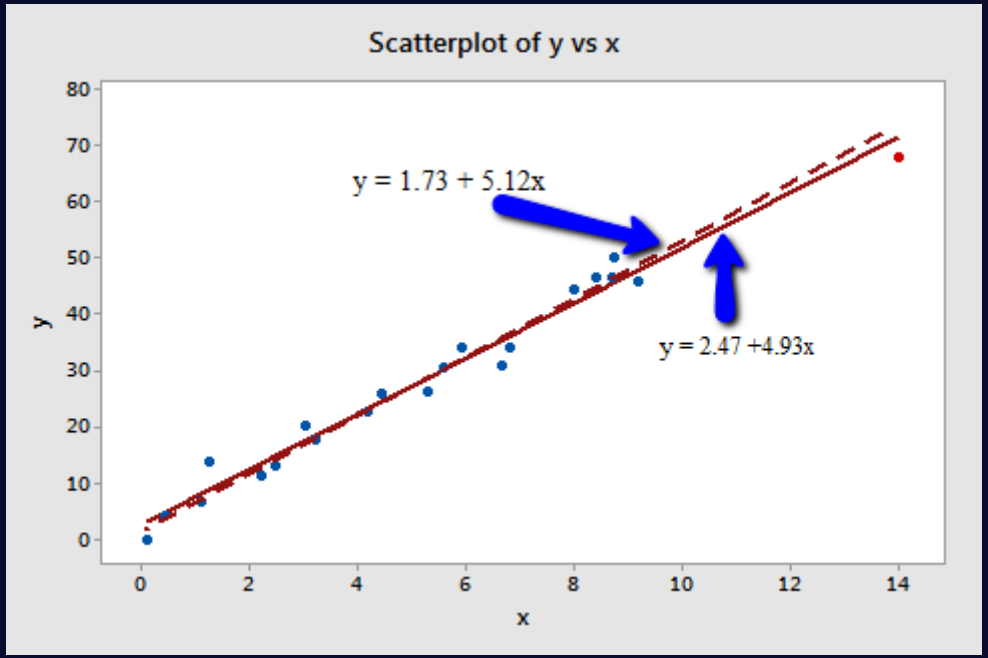
Influential Points – High Leverage

- A data point has high leverage if it has "extreme" predictor x values. With a single predictor, an extreme x value is simply one that is particularly high or low.

Influential Points – High Leverage



Influential Points – High Leverage



Influential Points – High Leverage

When red data point is included

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.70911	97.74%	97.62%	97.04%

Model Summary

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.47	1.08	2.29	0.033	
x	4.927	0.172	28.66	0.000	1.00

Regression Equation

$y = 2.47 + 4927\ x$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

Influential Points – High Leverage

When red data point is not included

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.59199	97.32%	97.17%	96.63%

Model Summary

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.73	1.12	1.55	0.140	
x	5.117	0.200	25.55	0.000	1.00

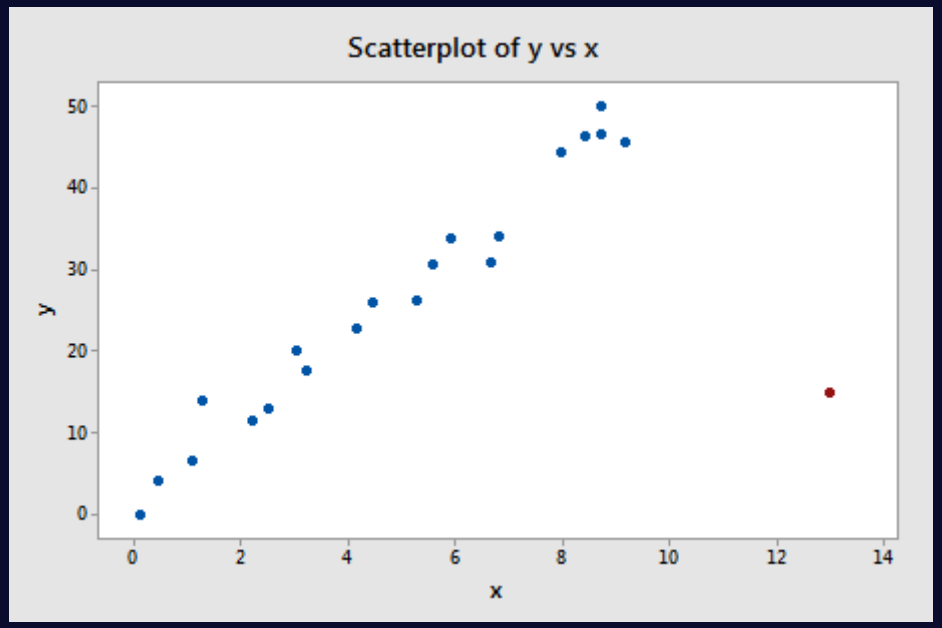
Regression Equation

$y = 1.73 + 5.117\ x$

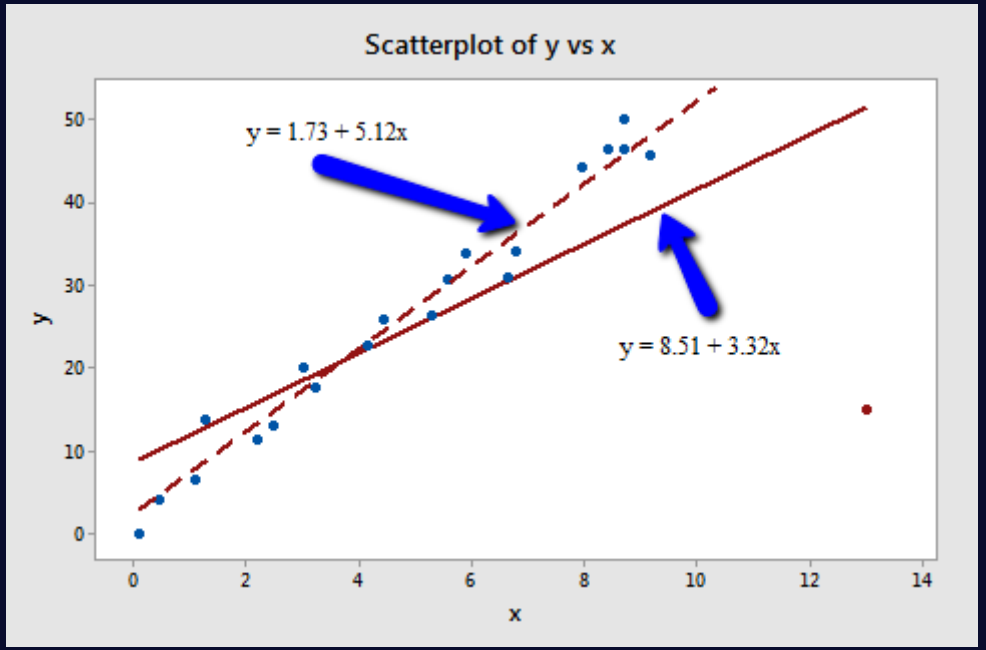
Influential Points – High Leverage

- The value has hardly changed at all, increasing only slightly from 97.3% to 97.7%. In either case, the relationship between y and x is deemed strong.
- The standard error of is about the same in each case — 0.172 when the red data point is included, and 0.200 when the red data point is excluded. Therefore, the width of the confidence intervals for
- would largely remain unaffected by the existence of the red data point. You might take note that this is because the data point is not an outlier heavily impacting MSE.
- In each case, the P-value for testing is less than 0.001. In either case, we can conclude that there is sufficient evidence at the 0.05 level to conclude that, in the population, x is related to y.

Example



Example



Example

- When red data point is included

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
10.4459	55.19%	52.84%	19.11%

Model Summary

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	8.50	4.22	2.01	0.058	
x	3.320	0.686	4.484	0.000	1.00

Regression Equation

$y = 8.50 + 3.320\ x$

Example

- When red data point is not included

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.59199	97.32%	97.17%	96.63%

Model Summary

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.73	1.12	1.55	0.140	
x	5.117	0.200	25.55	0.000	1.00

Regression Equation

$y = 1.73 + 5.117\ x$

Example

- The r^2 value has decreased substantially from 97.32% to 55.19%. If we include the red data point, we conclude that the relationship between y and x is only moderately strong, whereas if we exclude the red data point, we conclude that the relationship between y and x is very strong.
- The standard error is almost 3.5 times larger when the red data point is included — increasing from 0.200 to 0.686. This increase would have a substantial effect on the width of our confidence interval for
- In each case, the P-value for testing the slope is less than 0.001. In both cases, we can conclude that there is sufficient evidence at the 0.05 level to conclude that, in the population, x is related to y. Note, however, that the t-statistic decreases dramatically from 25.55 to 4.84 upon inclusion of the red data point.

- Before finding a way to detect such errors let’s derive the closed form solution for linear regression.

Closed form solution

- Consider the case where we have n data points where each data point consists of d features.

Closed form solution

- Recall that the linear regression model can be written as follows:

$$h(\boldsymbol{x}) = \sum_{j=0}^{d} \theta_j x_j$$

- Or in vectorized format:

$$h(\boldsymbol{x}) = \boldsymbol{\theta}^\intercal \boldsymbol{x}$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} \qquad \boldsymbol{x}^\intercal = \begin{bmatrix} 1 & x_1 & \dots & x_d \end{bmatrix}$$

Closed form solution

- We define:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_1^{(1)} & \dots & x_d^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(i)} & \dots & x_d^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(n)} & \dots & x_d^{(n)} \end{bmatrix}$$

$\mathbb{R}^{(d+1)\times 1}$ $\mathbb{R}^{n\times(d+1)}$

$$h\left(\boldsymbol{x}^{(i)}\right) = \sum_{j=0}^{d} \theta_j x_j^{(i)}$$

$$\boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

Closed form solution

- We define:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_1^{(1)} & \dots & x_d^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(i)} & \dots & x_d^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(n)} & \dots & x_d^{(n)} \end{bmatrix}$$

$\mathbb{R}^{(d+1)\times 1}$ $\mathbb{R}^{n\times(d+1)}$

$$h\left(\boldsymbol{x}^{(i)}\right) = \sum_{j=0}^{d} \theta_j x_j^{(i)}$$

$$\boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

Closed form solution

- For the regression loss function:

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

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

$$= \frac{1}{2n}\underbrace{(\boldsymbol{X\theta} - \boldsymbol{y})^{\intercal}}_{\mathbb{R}^{1\times n}}\underbrace{(\boldsymbol{X\theta} - \boldsymbol{y})}_{\mathbb{R}^{n\times 1}}$$

$\boldsymbol{X}: \mathbb{R}^{n\times(d+1)}$, $\boldsymbol{\theta}: \mathbb{R}^{(d+1)\times 1}$

המרכז הבינלאומי ללימודי הייטק וחדשנות

Closed form solution

- In order to derive the minimum we can calculate

$$\frac{\partial}{\partial \boldsymbol{\theta}} J(\boldsymbol{\theta}) = 0$$

Closed form solution

$$\mathcal{J}(\theta) = \frac{1}{2n}(X\theta - y)^\intercal (X\theta - y)$$

1 x 1

$$\propto \theta^\intercal X^\intercal X\theta - y^\intercal X\theta - \theta^\intercal X^\intercal y + y^\intercal y$$

$$\propto \theta^\intercal X^\intercal X\theta - 2\theta^\intercal X^\intercal y + y^\intercal y$$

$$\frac{\partial}{\partial\theta}(\theta^\intercal X^\intercal X\theta - 2\theta^\intercal X^\intercal y + \cancel{y^\intercal y}) = 0$$

$$(X^\intercal X)\theta - X^\intercal y = 0$$

$$(X^\intercal X)\theta = X^\intercal y$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

Closed form solution

- From previous slide we get the optimal solution

$$\boldsymbol{\theta} = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{y}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1^{(1)} & \dots & x_d^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(i)} & \dots & x_d^{(i)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(n)} & \dots & x_d^{(n)} \end{bmatrix} \qquad \boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

Closed form solution

- What if the matrix is not invertible?
  - Use pseudo-inverse
  - Remove redundant features
  - Make sure n>>d

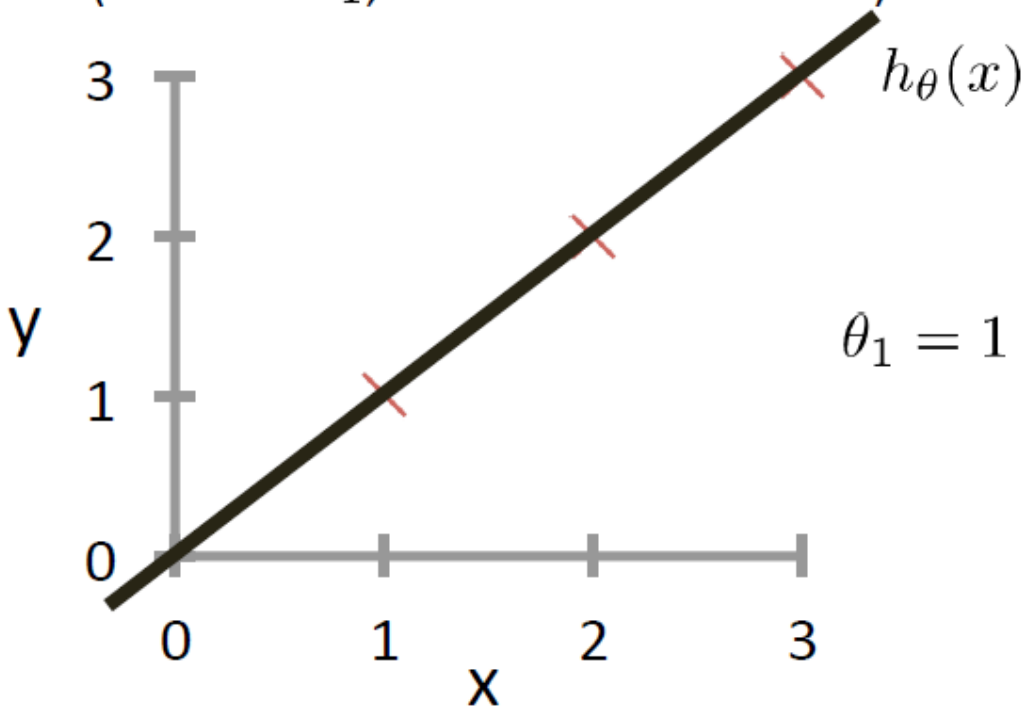
$$\theta = (X^\intercal X)^{-1} X^\intercal y$$

Problems with closed form solution

- If n is large the computation is slow ~O(n^3)
- Problematic for Big-Data scenario

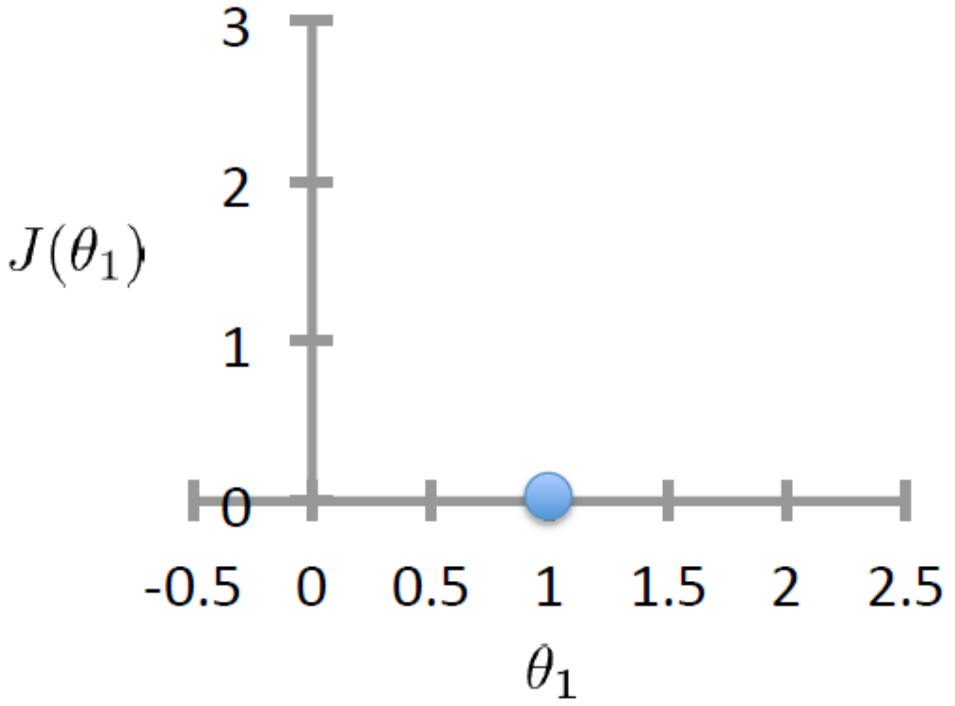
$h_\theta(x)$

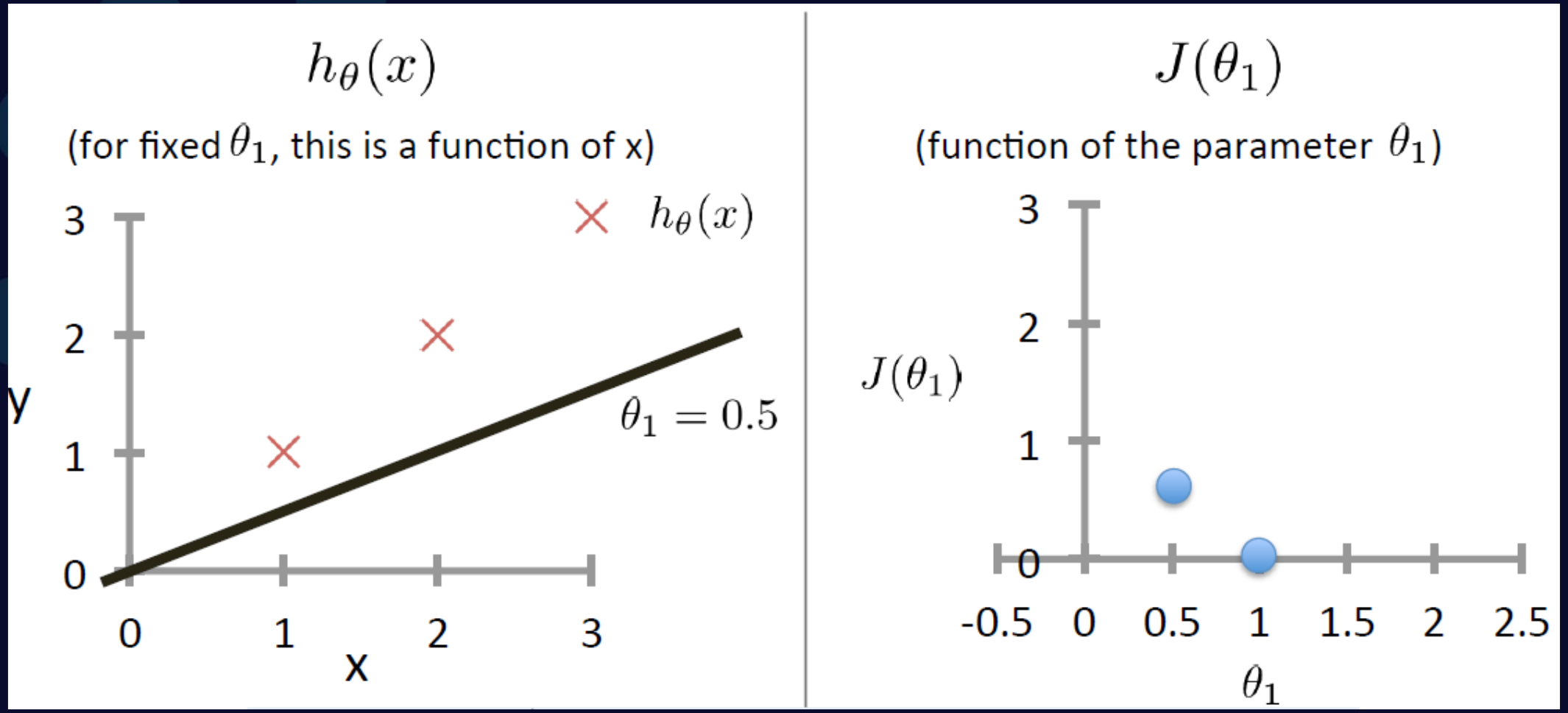
(for fixed θ_1, this is a function of x)



$J(\theta_1)$

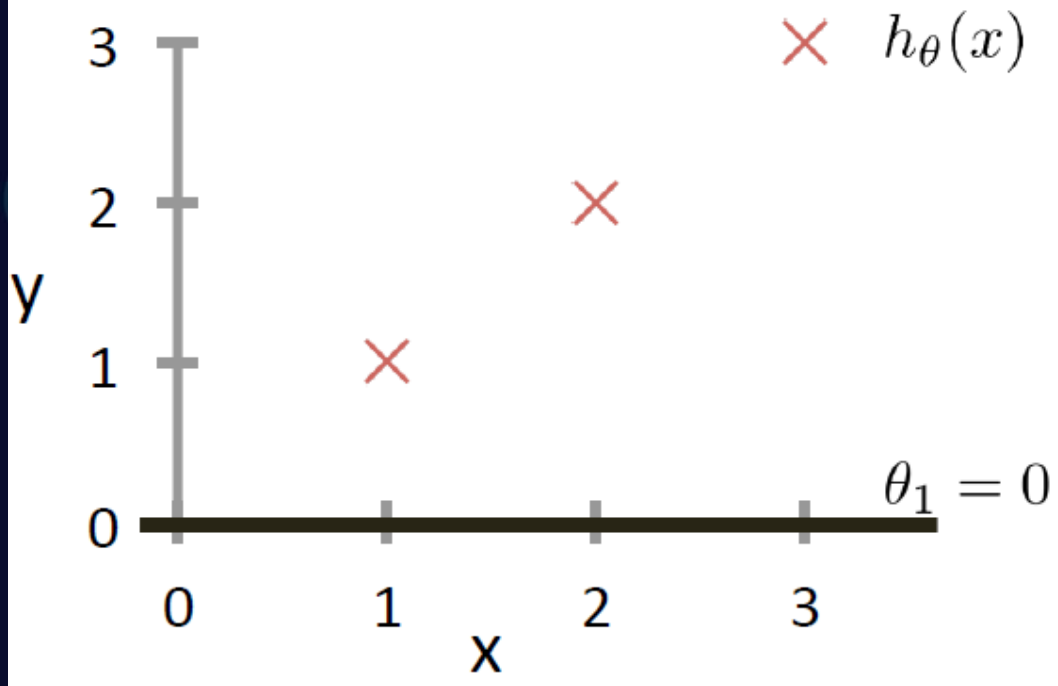
(function of the parameter θ_1)





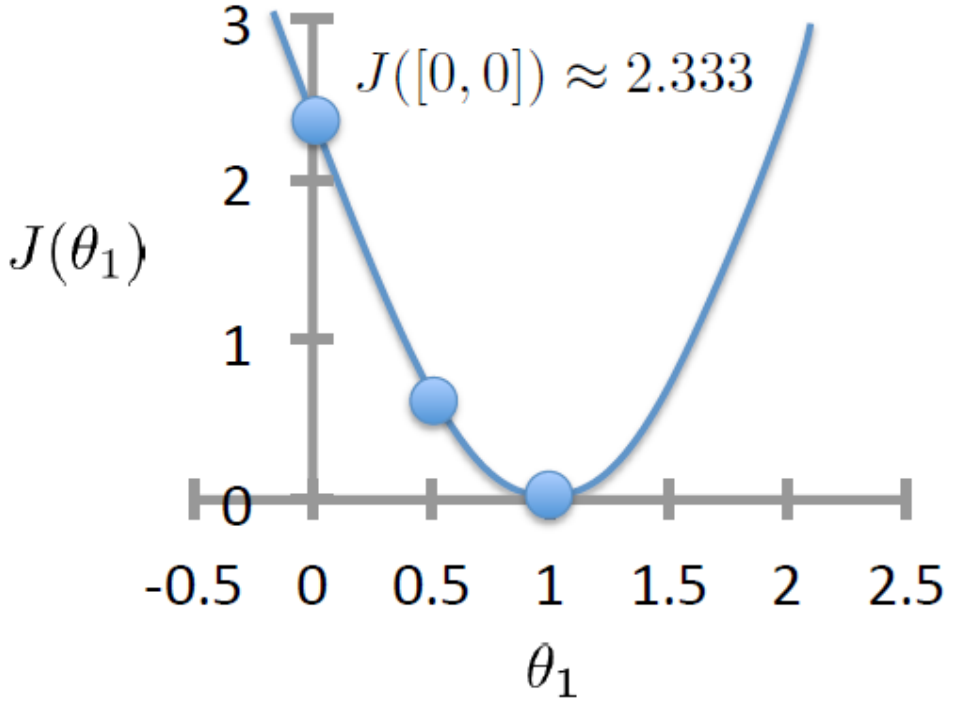
$h_\theta(x)$

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



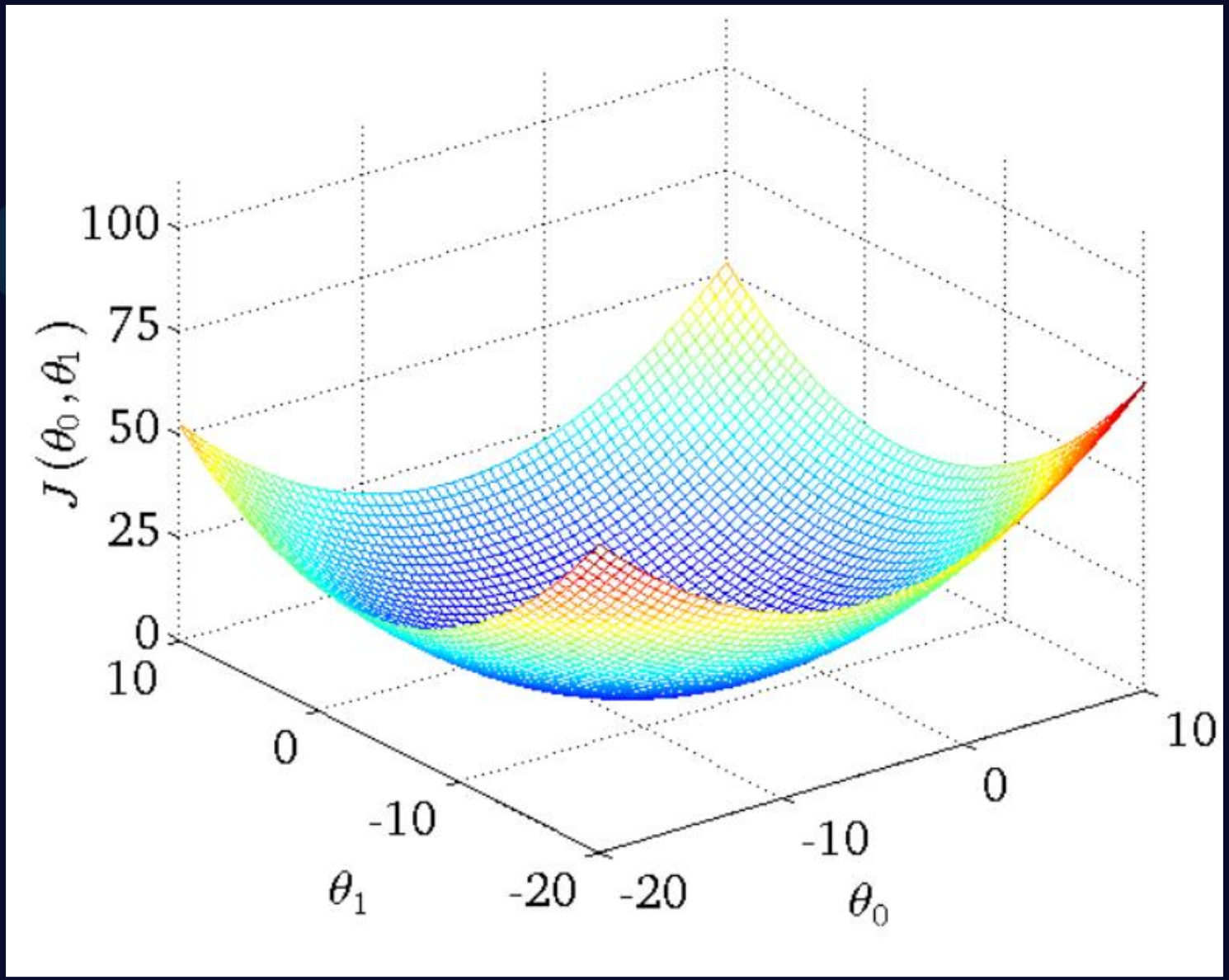
$J(\theta_1)$

(function of the parameter θ_1)



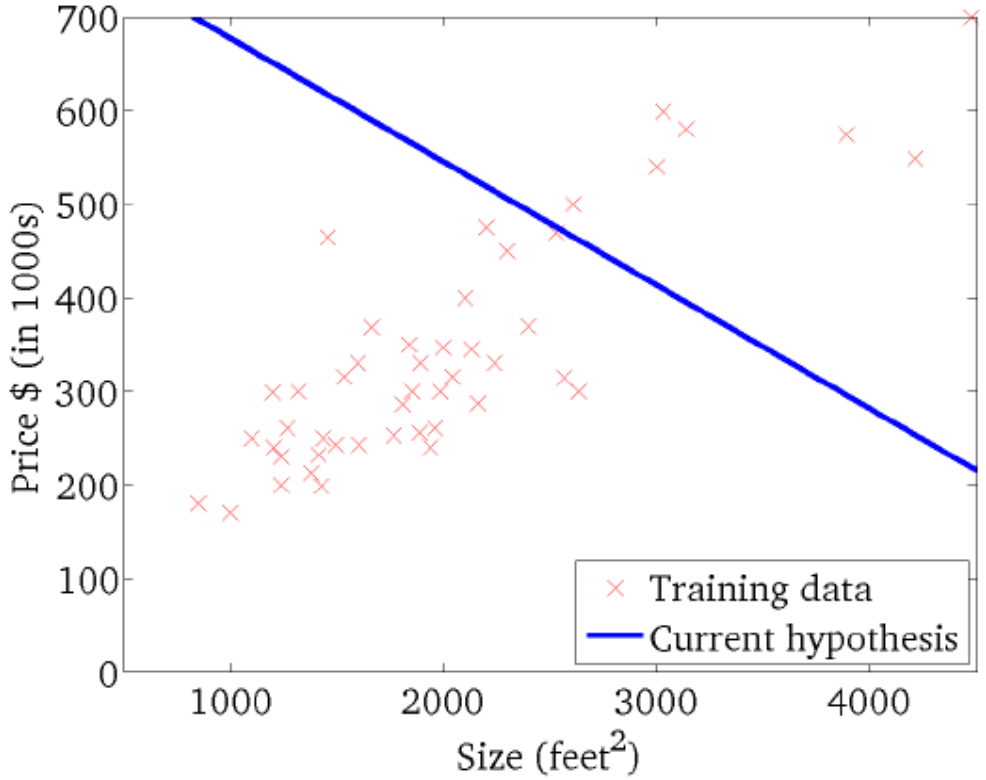
INT

המרכז הבינלאומי
ללימודי הייטק וחדשנות



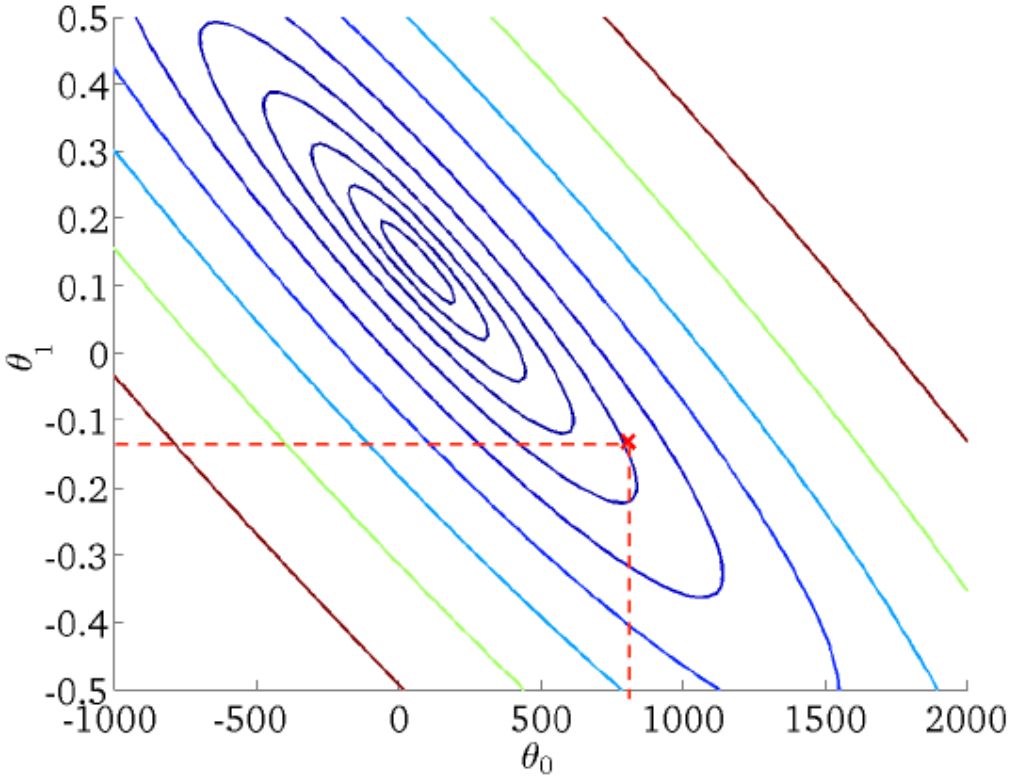
$h_\theta(x)$

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



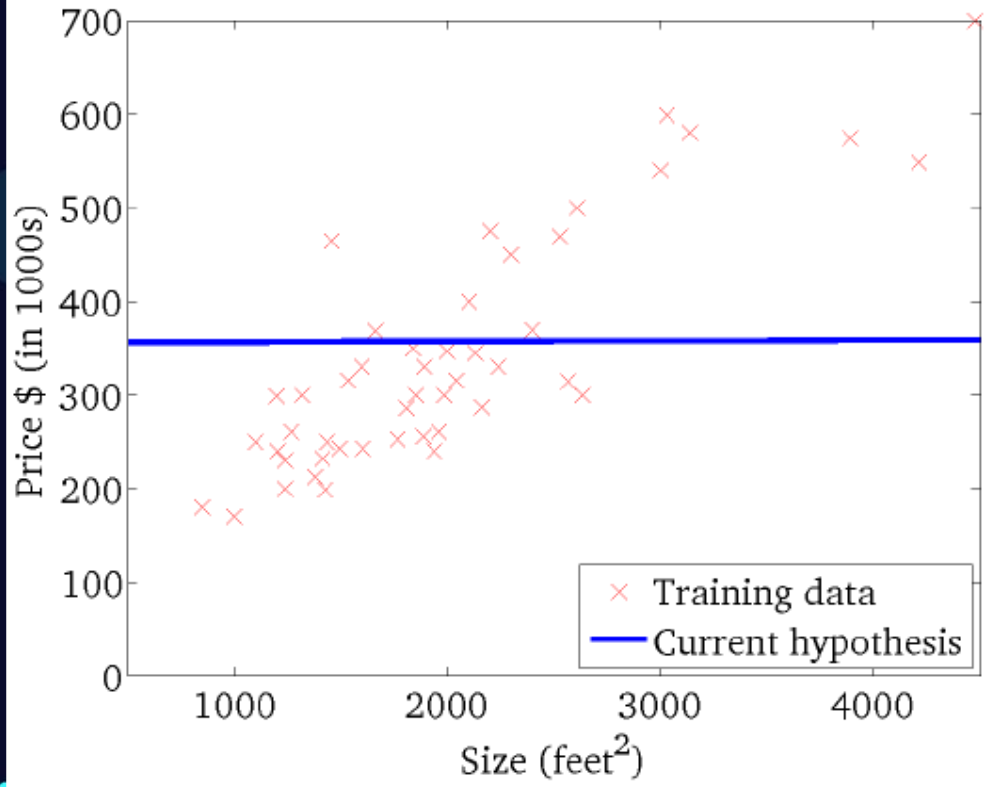
$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



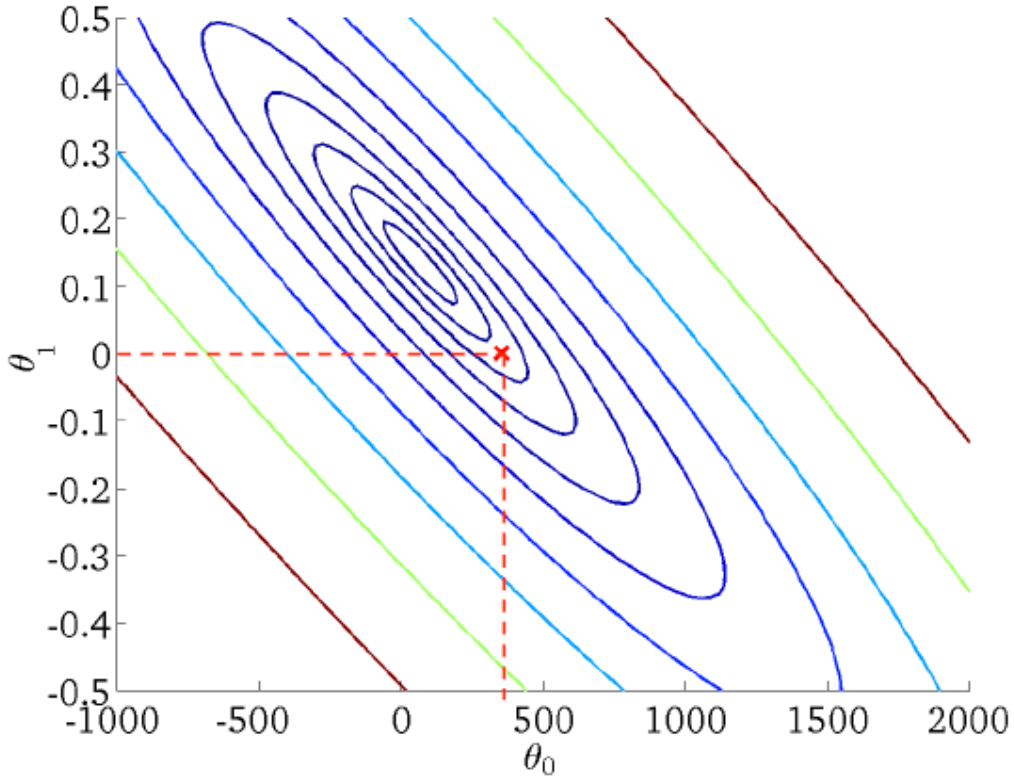
$h_\theta(x)$

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



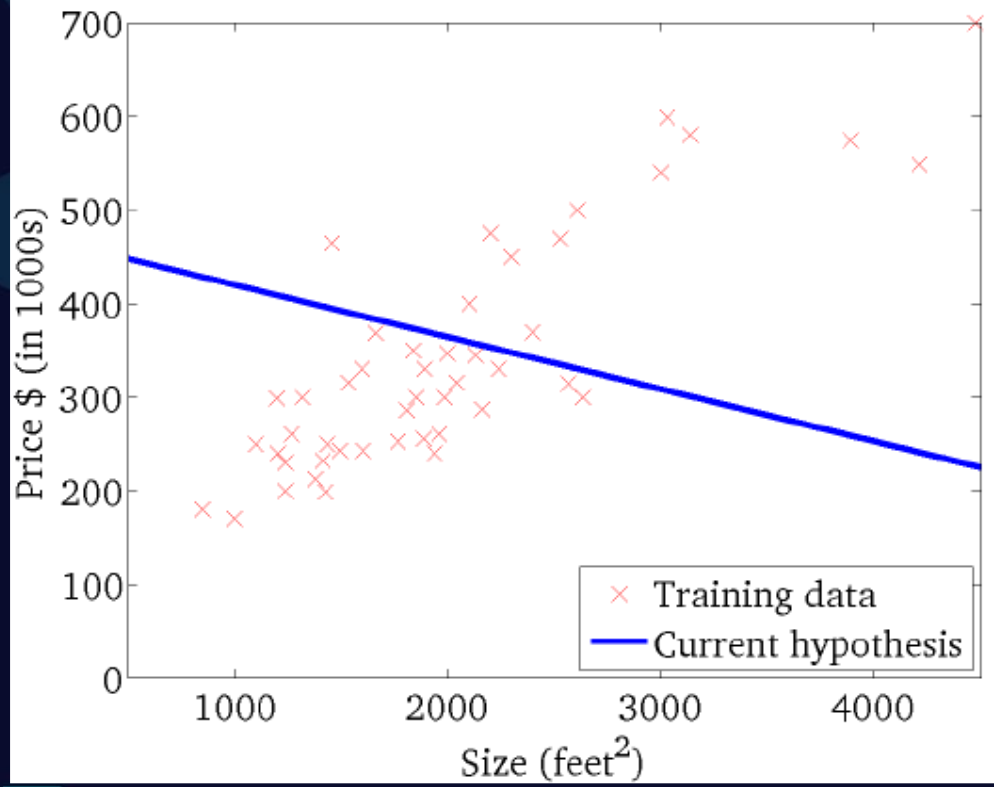
$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



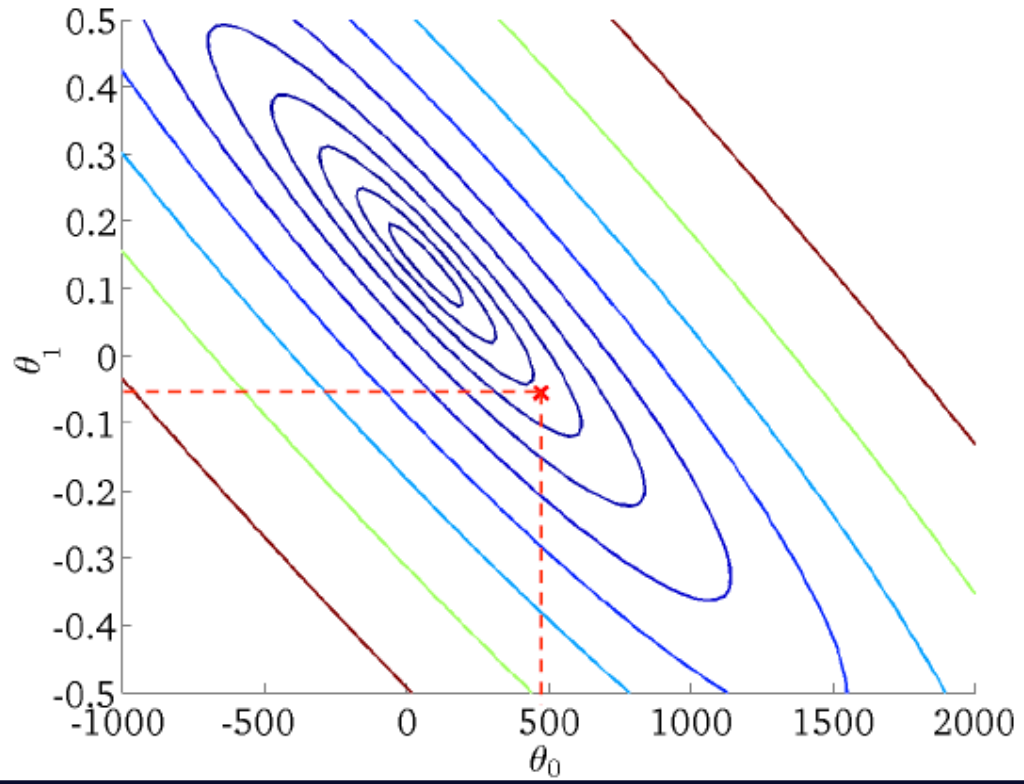
$h_\theta(x)$

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



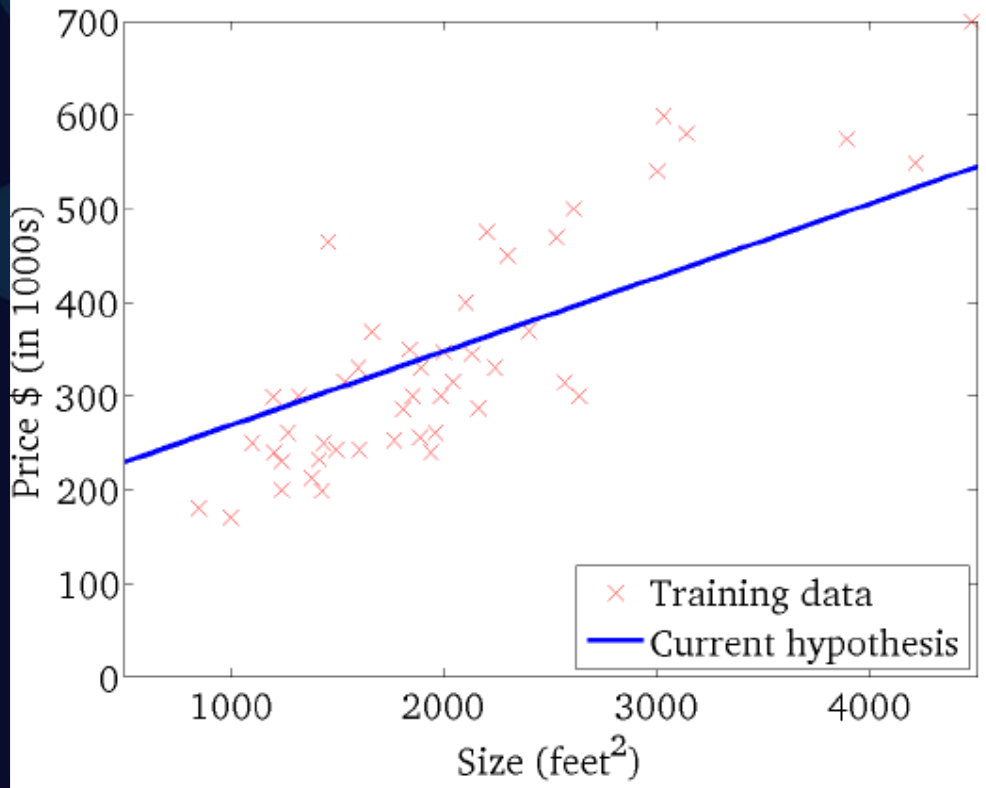
$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



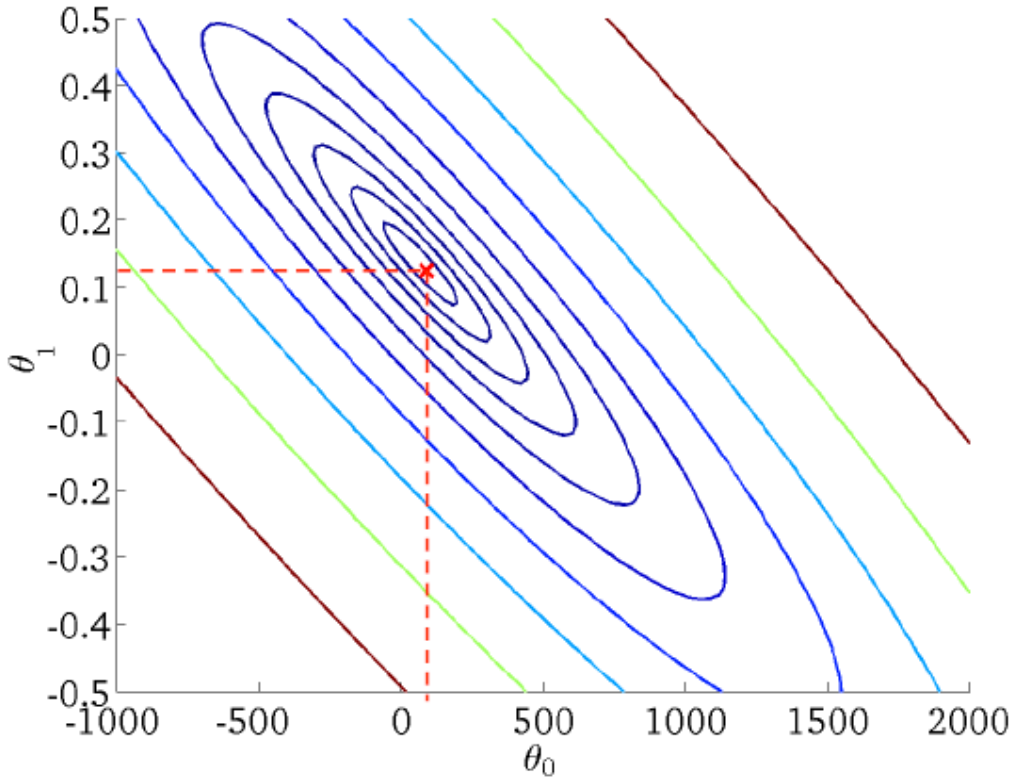
$h_\theta(x)$

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



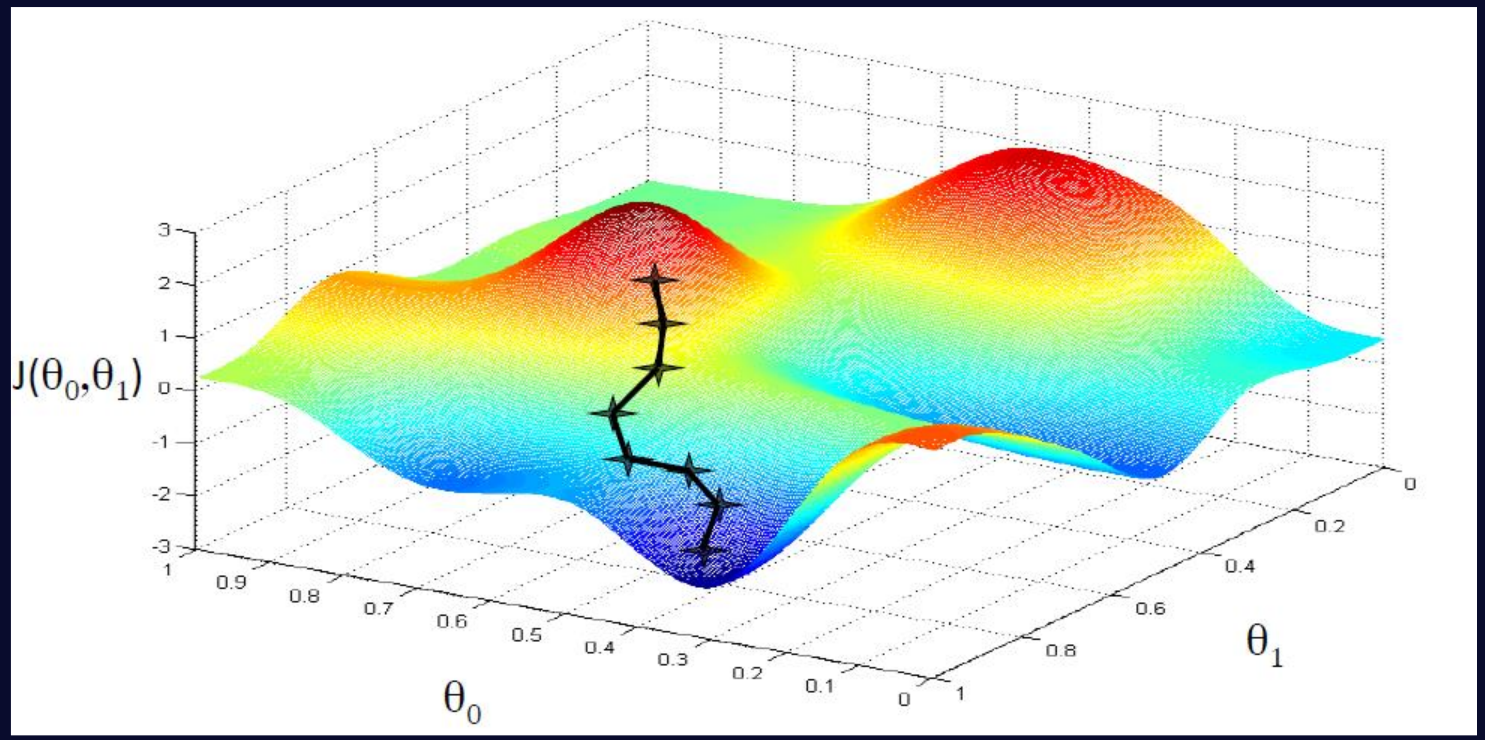
$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)

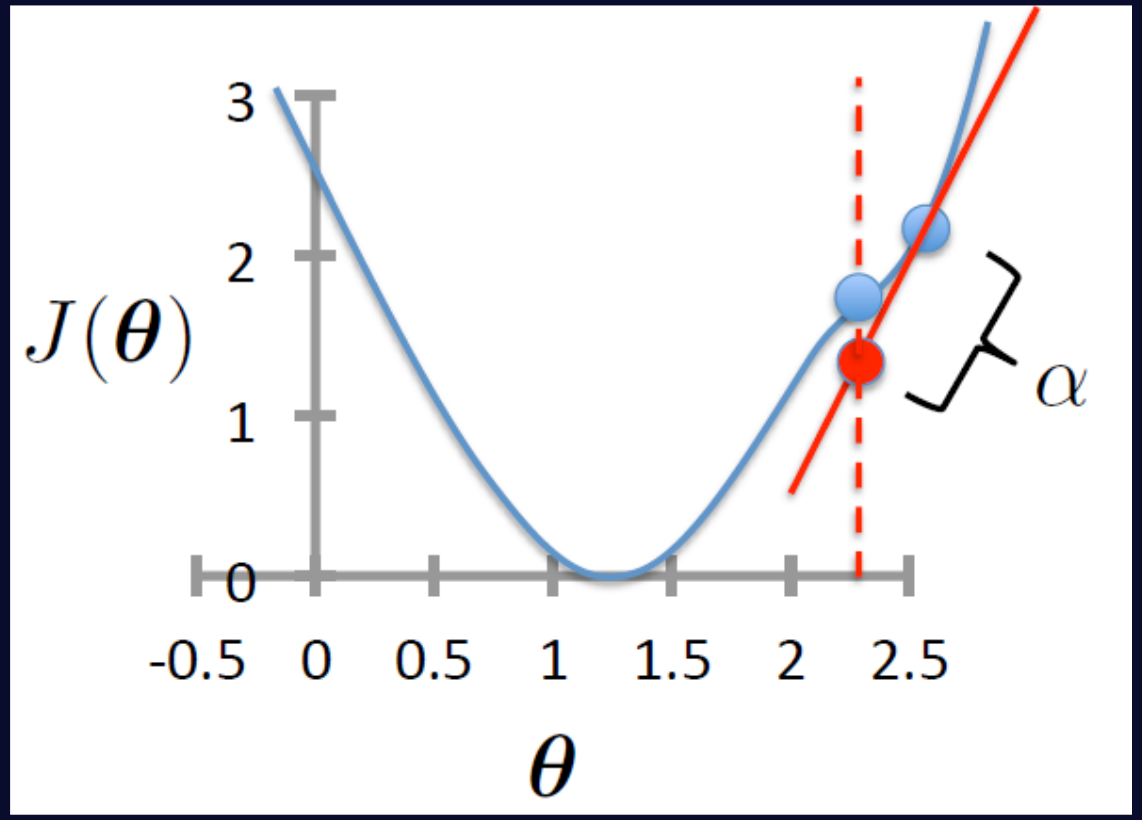


Search Procedure

- Start with an initial guess
- Until we reach minimum: find a new value and proceed



Gradient Decent



Gradient Decent

- Initialize $\boldsymbol{\theta}$
- Repeat until convergence

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\boldsymbol{\theta})$$

simultaneous update
for j = 0 ... d

learning rate (small)
e.g., α = 0.05

GD for Linear Regression

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} J(\boldsymbol{\theta}) &= \frac{\partial}{\partial \theta_j} \frac{1}{2n} \sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}} \left(\boldsymbol{x}^{(i)} \right) - y^{(i)} \right)^2 \\
&= \frac{\partial}{\partial \theta_j} \frac{1}{2n} \sum_{i=1}^{n} \left(\sum_{k=0}^{d} \theta_k x_k^{(i)} - y^{(i)} \right)^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=0}^{d} \theta_k x_k^{(i)} - y^{(i)} \right) \times \frac{\partial}{\partial \theta_j} \left(\sum_{k=0}^{d} \theta_k x_k^{(i)} - y^{(i)} \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{k=0}^{d} \theta_k x_k^{(i)} - y^{(i)} \right) x_j^{(i)}
\end{aligned}
$$

INT
המרכז הבינלאומי
ללימודי הייטק וחדשנות

GD for Linear Regression

- Initialize $\boldsymbol{\theta}$
- Repeat until convergence

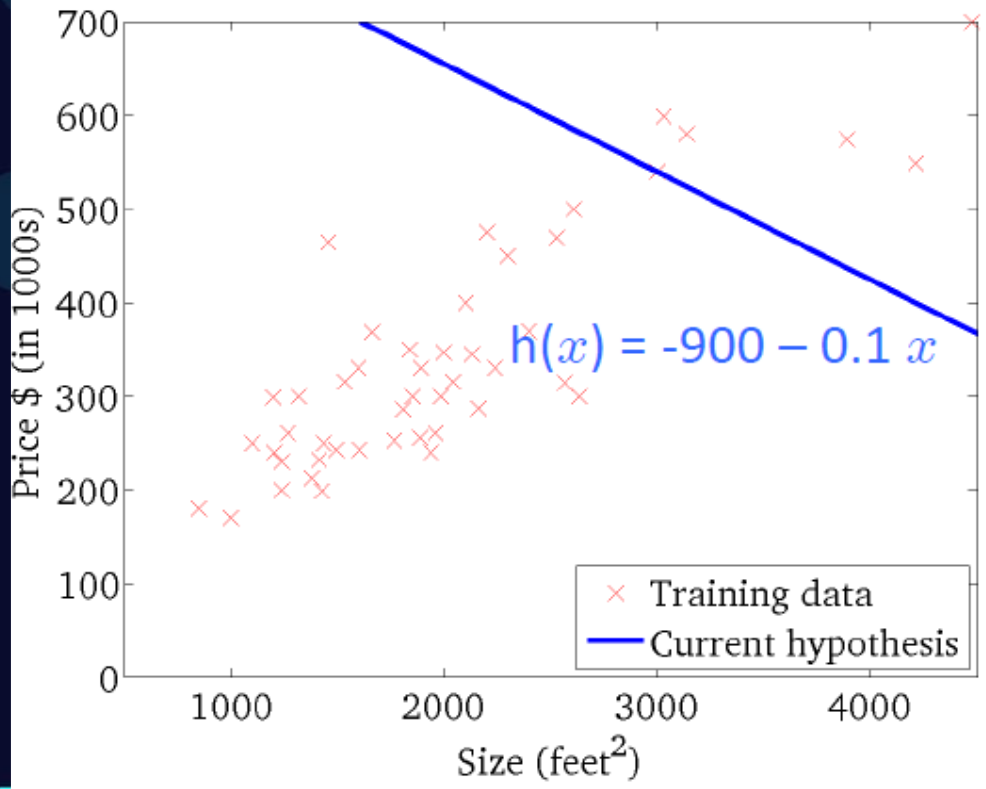
$$\theta_j \leftarrow \theta_j - \alpha \frac{1}{n} \sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}} \left(\boldsymbol{x}^{(i)} \right) - y^{(i)} \right) x_j^{(i)}$$

simultaneous update for $j = 0 \ldots d$

- Assume convergence when the updates are small then a predefined threshold.

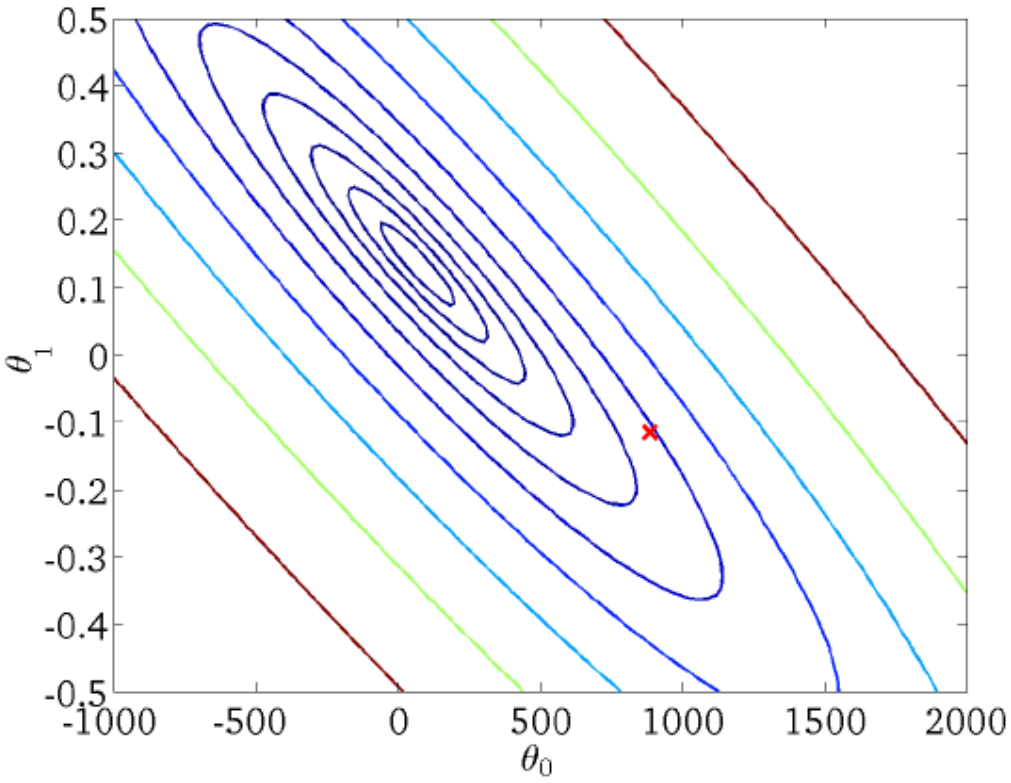
$h_\theta(x)$

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



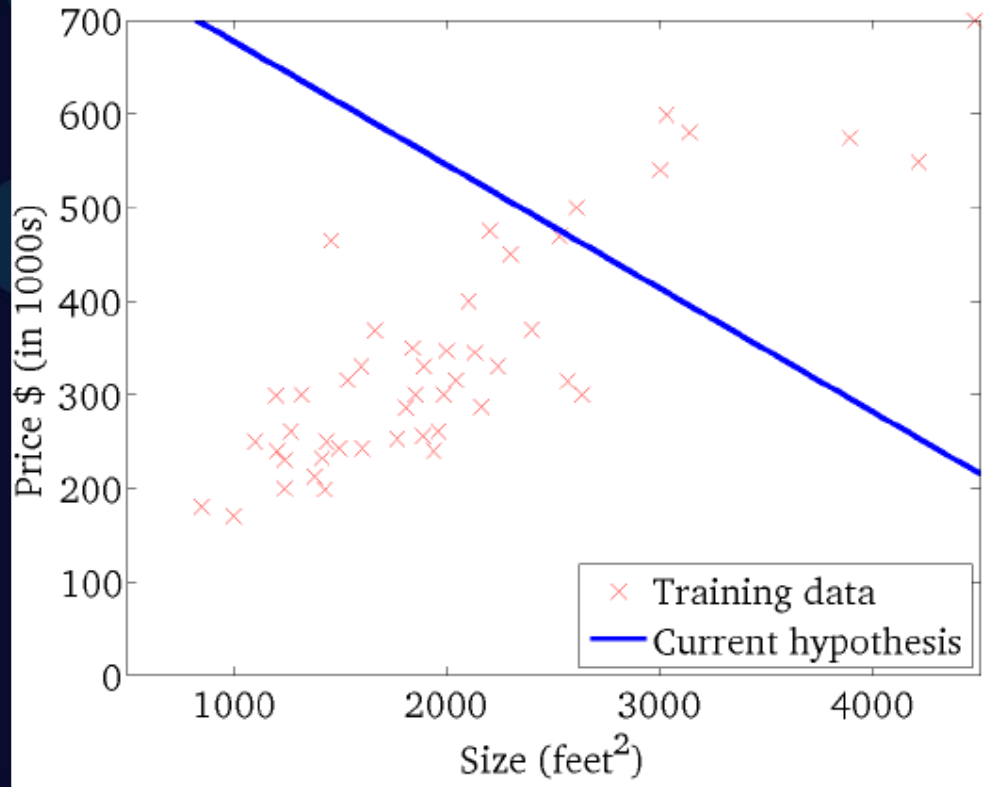
$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



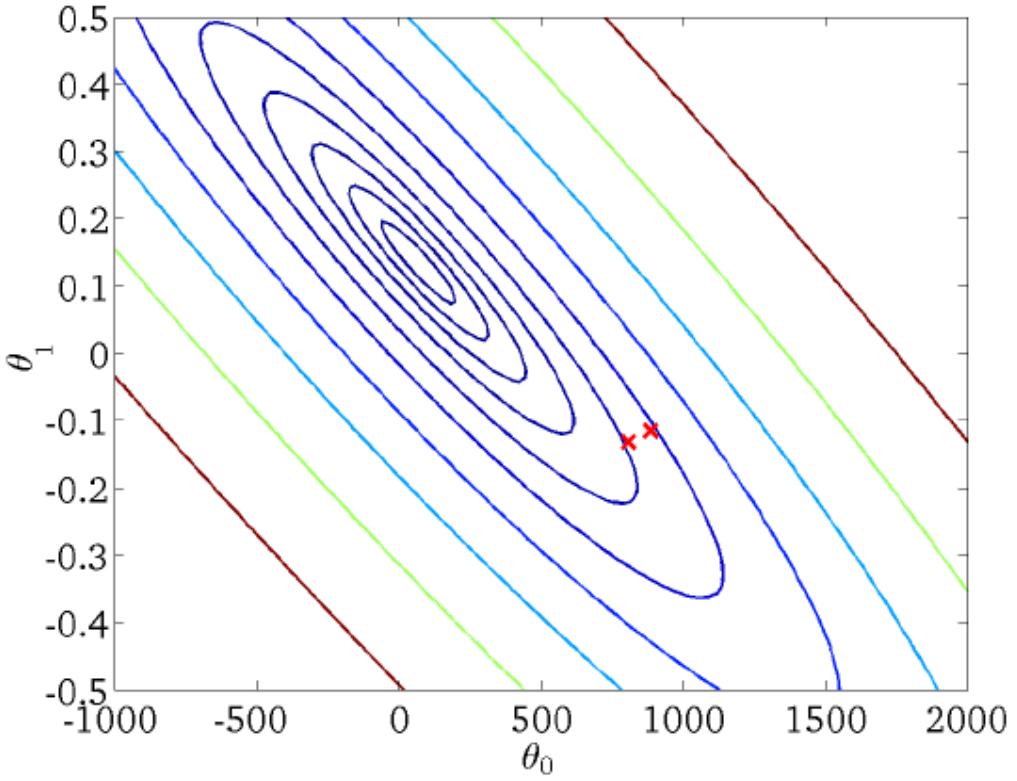
$h_\theta(x)$

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



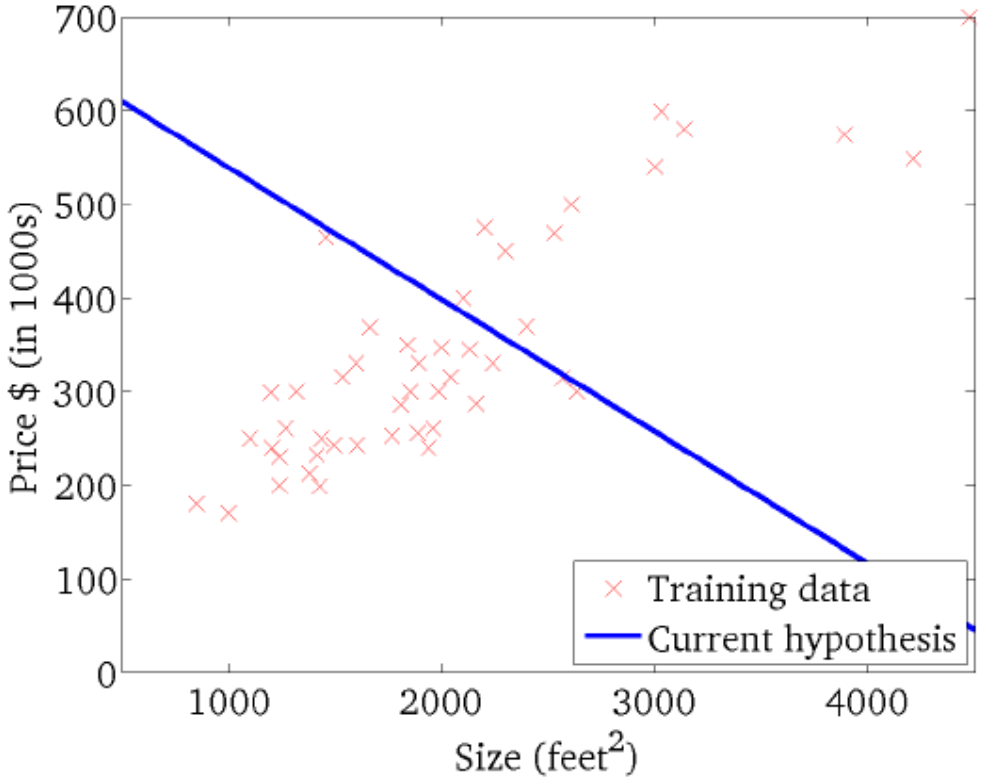
$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



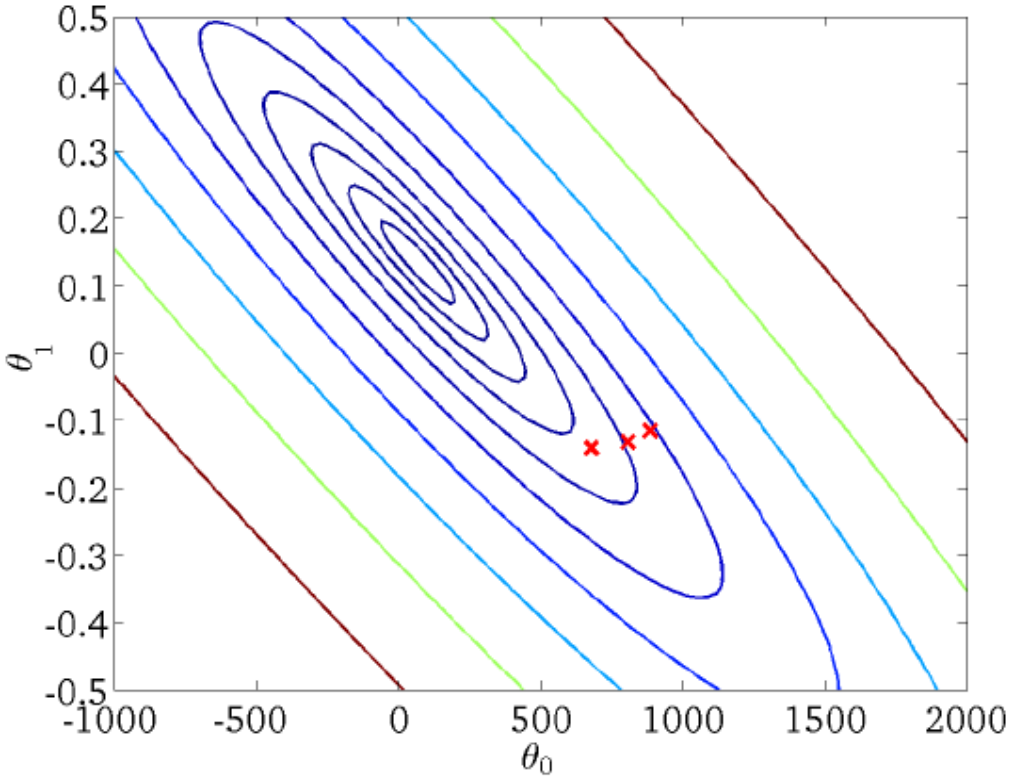
$h_\theta(x)$

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



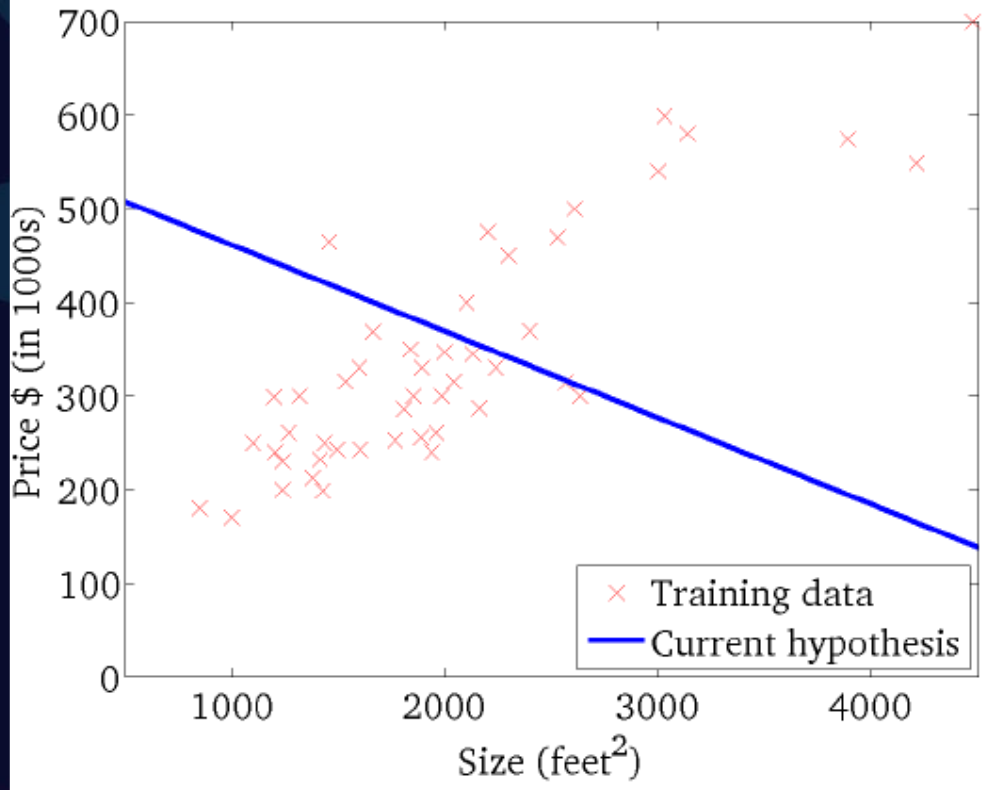
$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



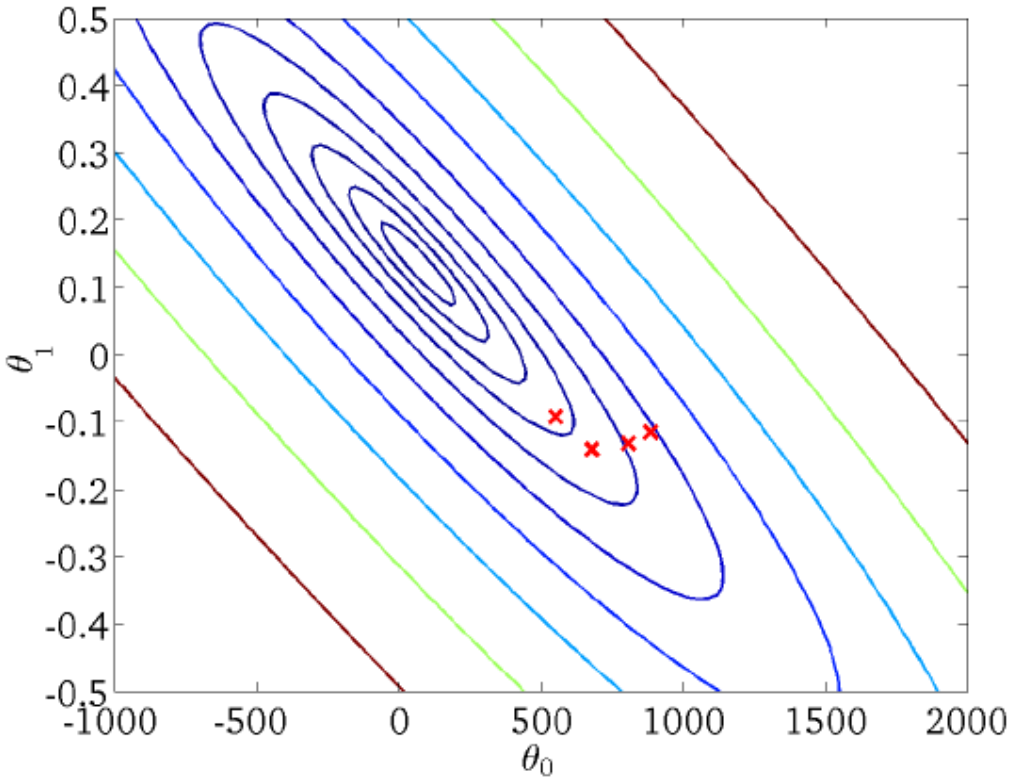
$h_\theta(x)$

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



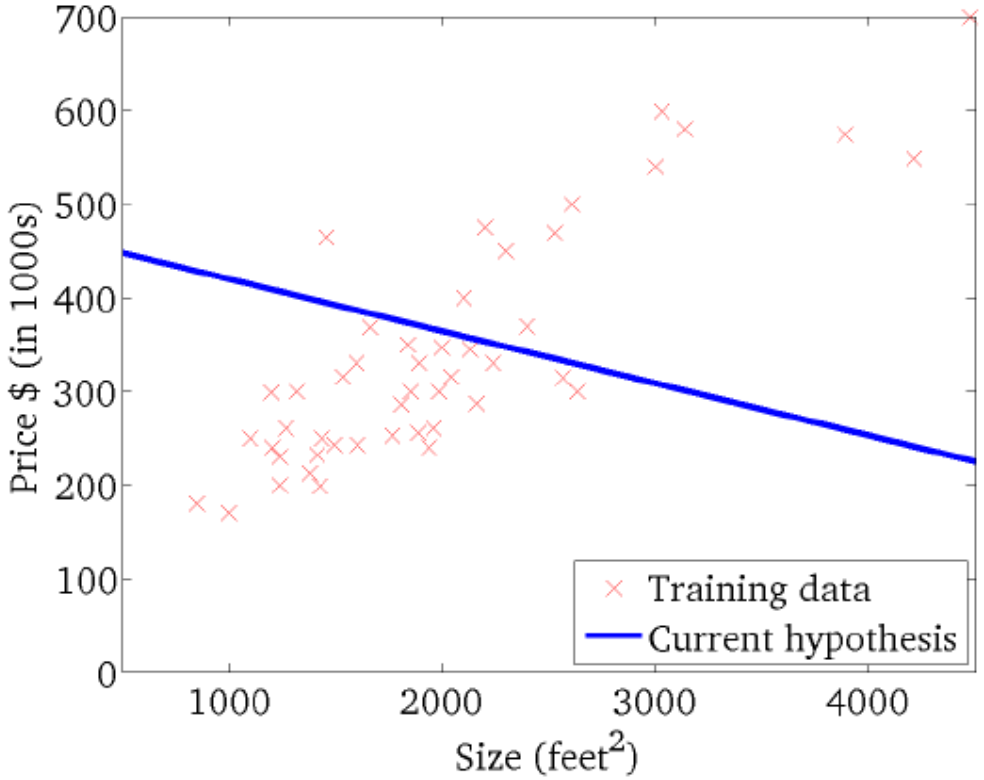
$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



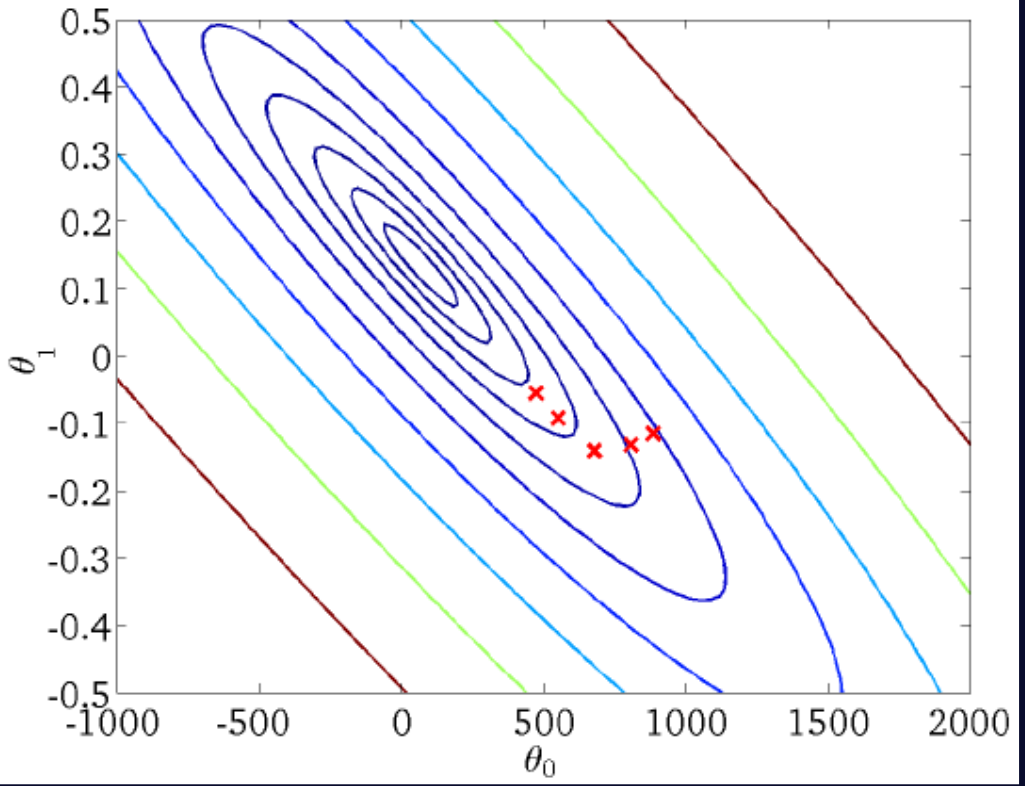
$h_\theta(x)$

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



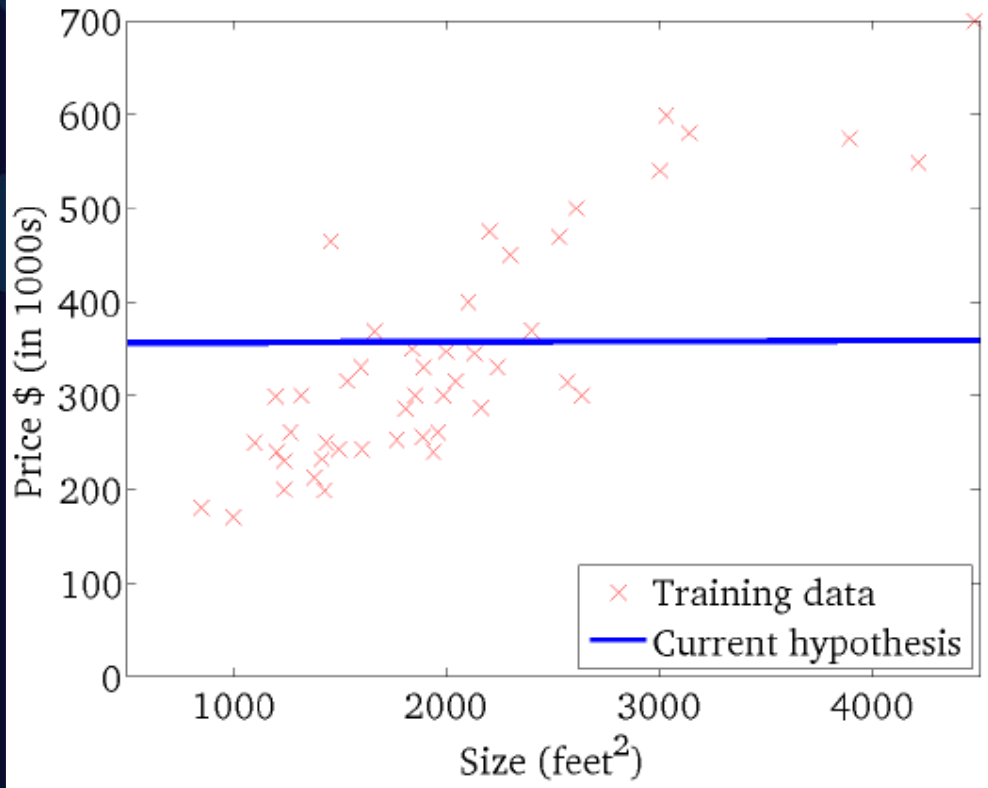
$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



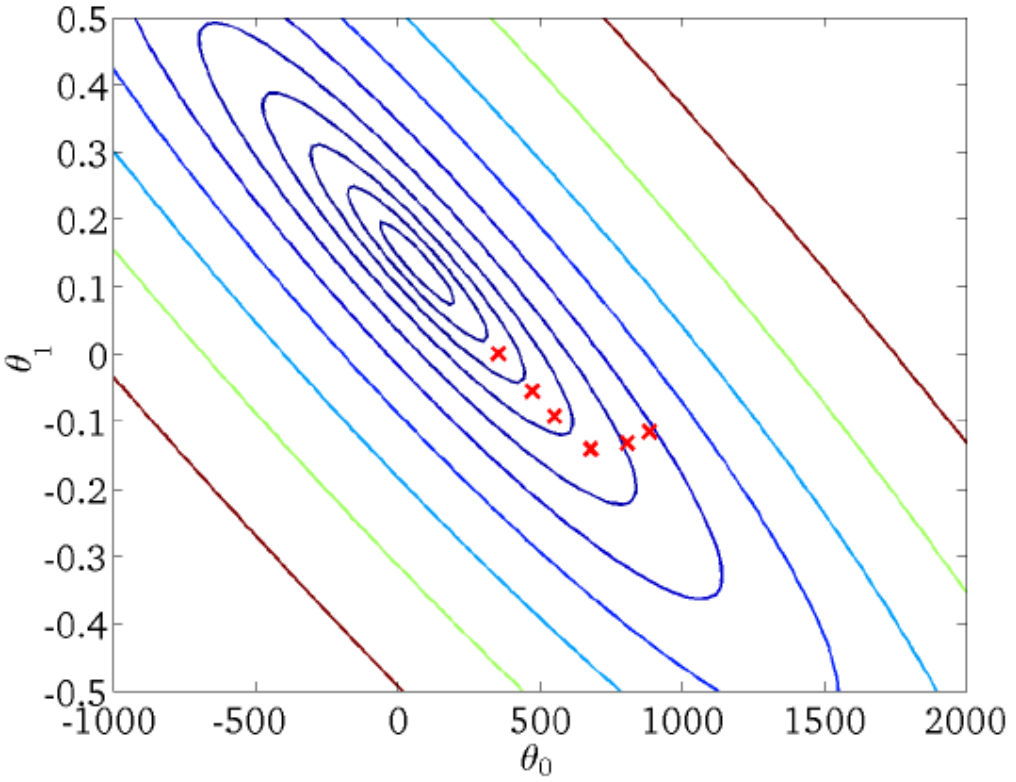
$h_\theta(x)$

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



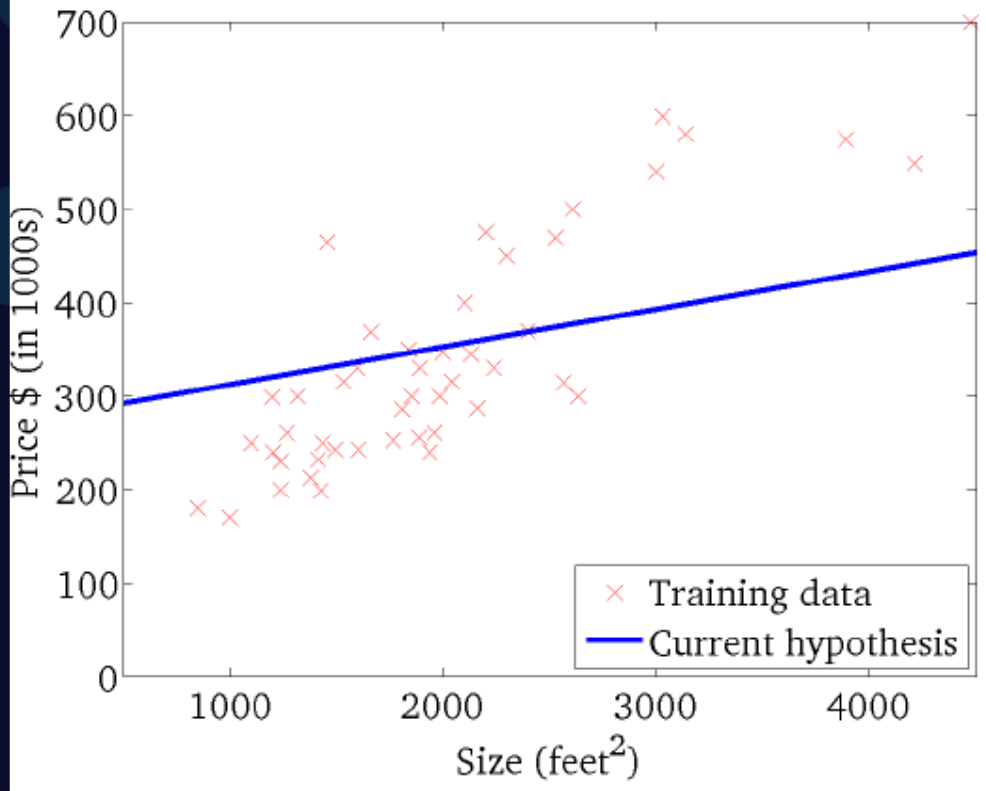
$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



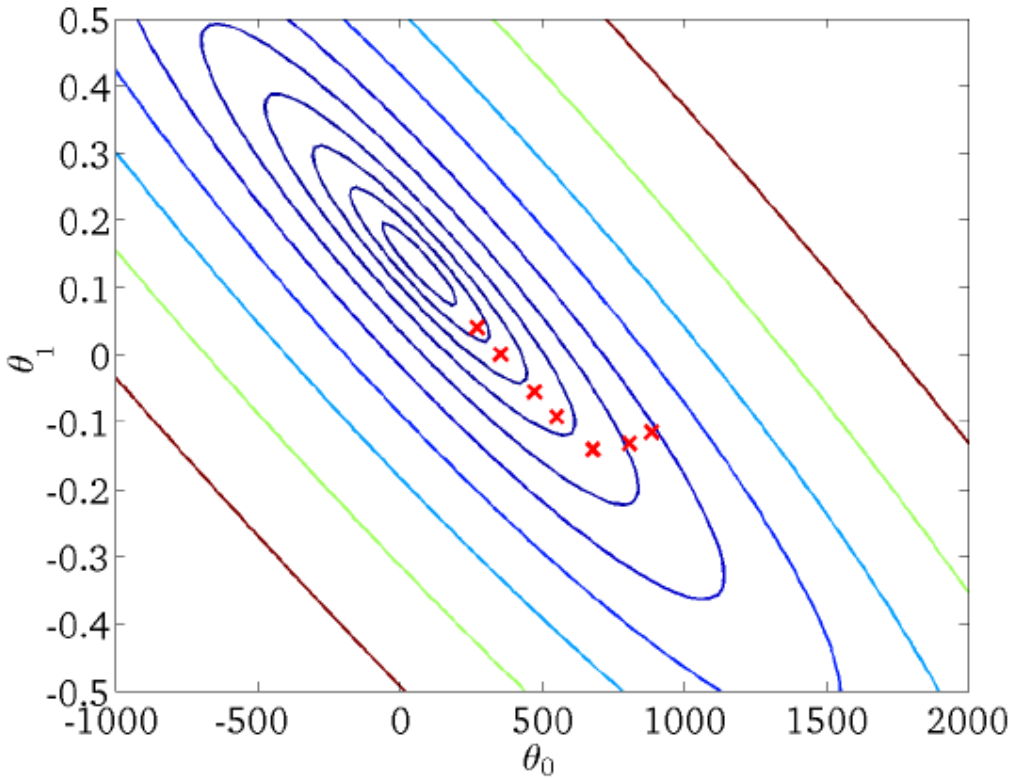
$h_\theta(x)$

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



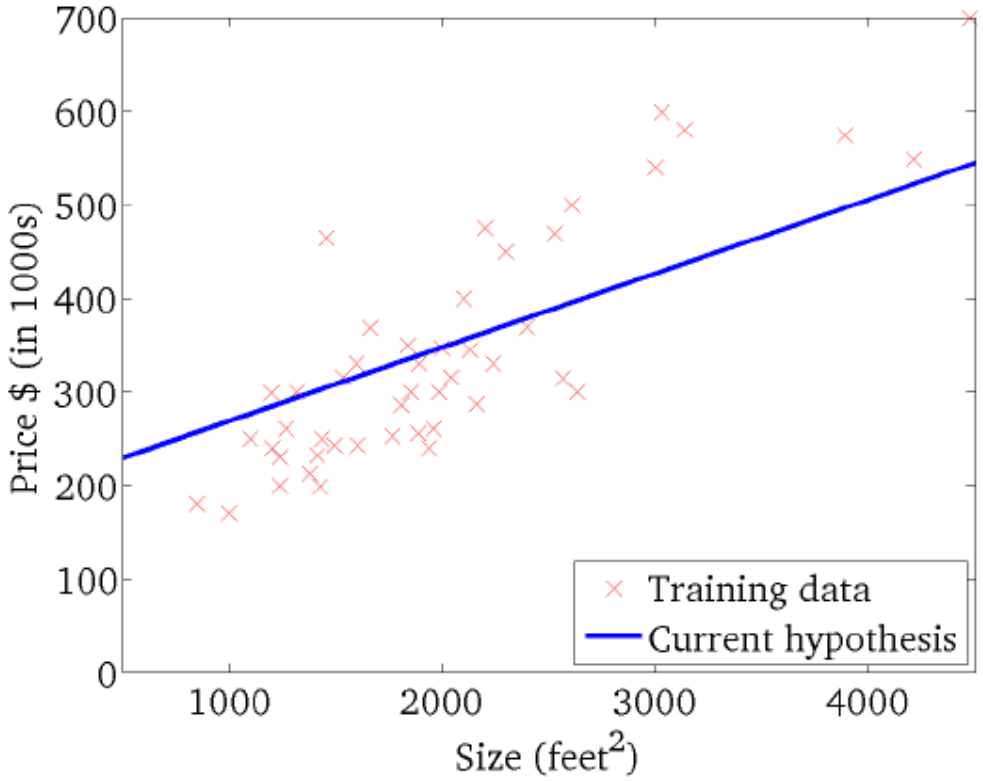
$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



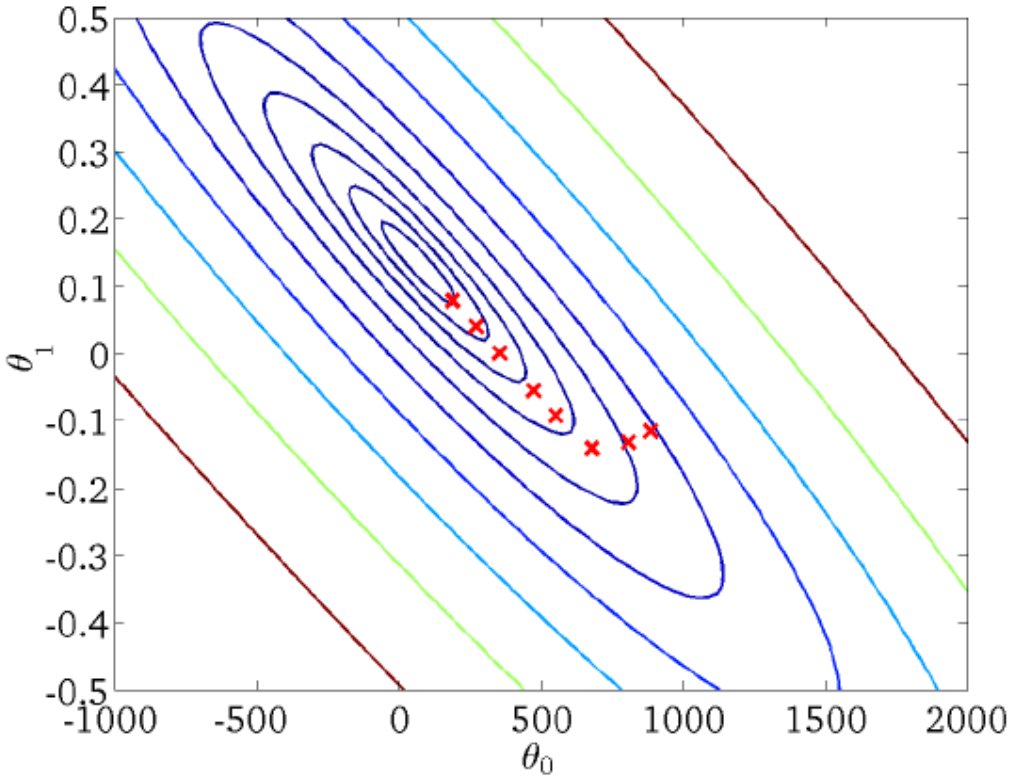
$h_\theta(x)$

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



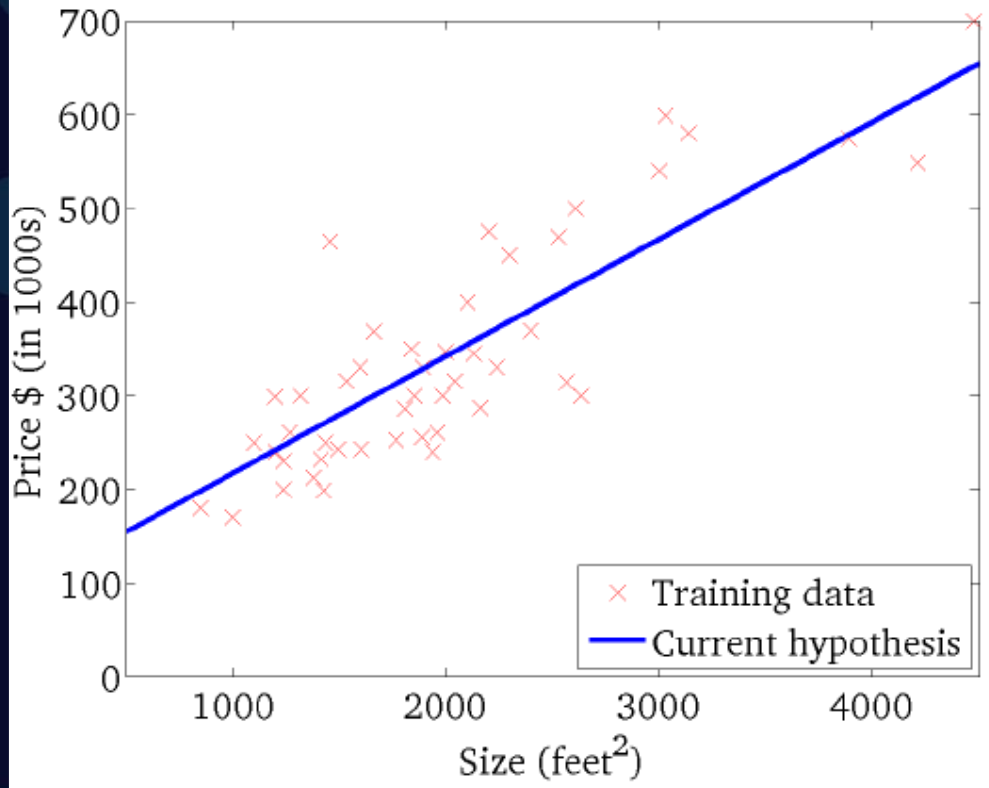
$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



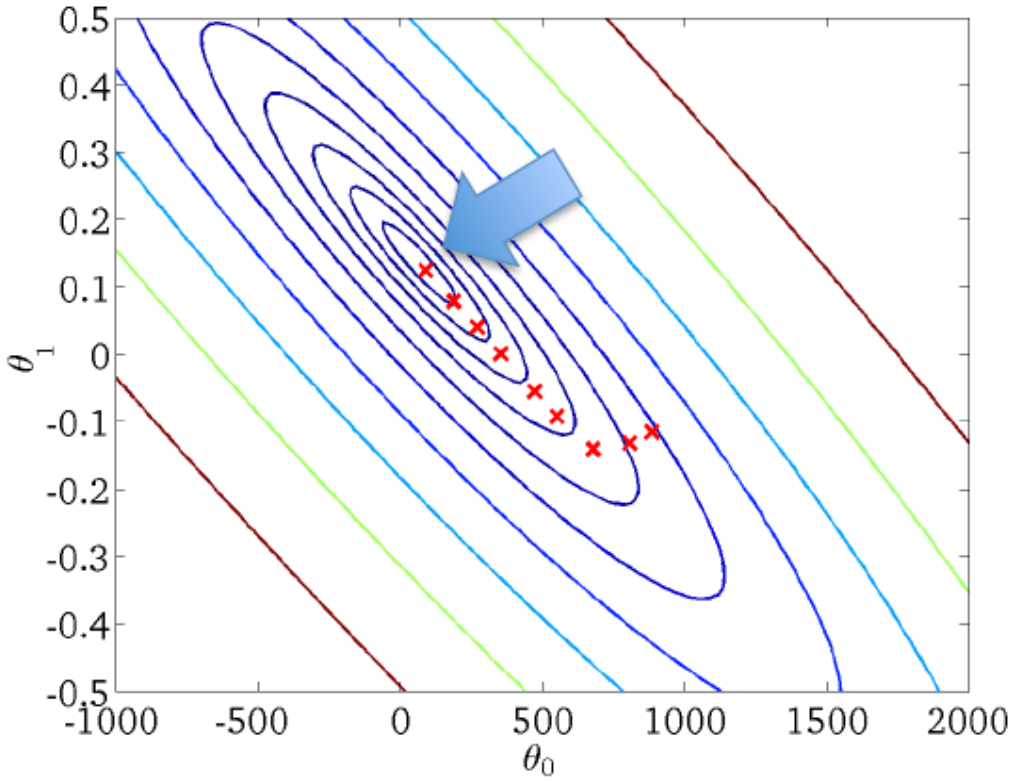
$h_\theta(x)$

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

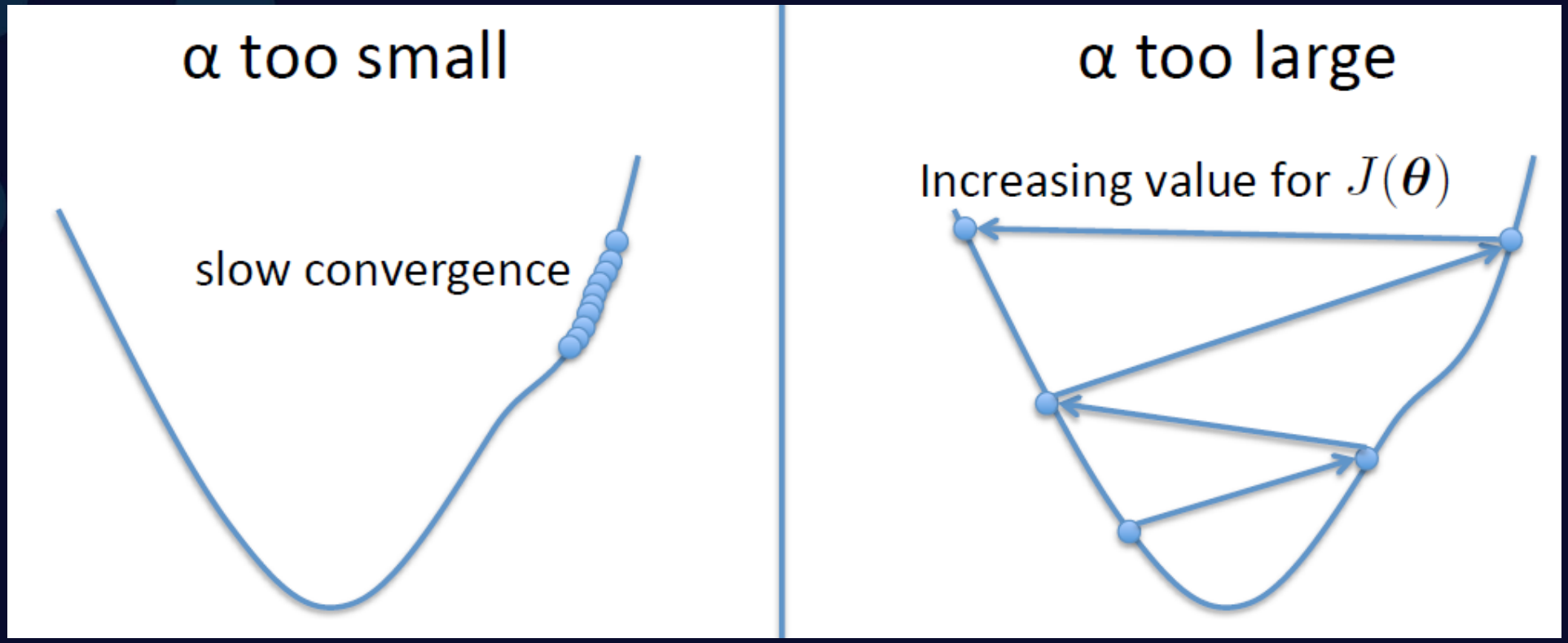


$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



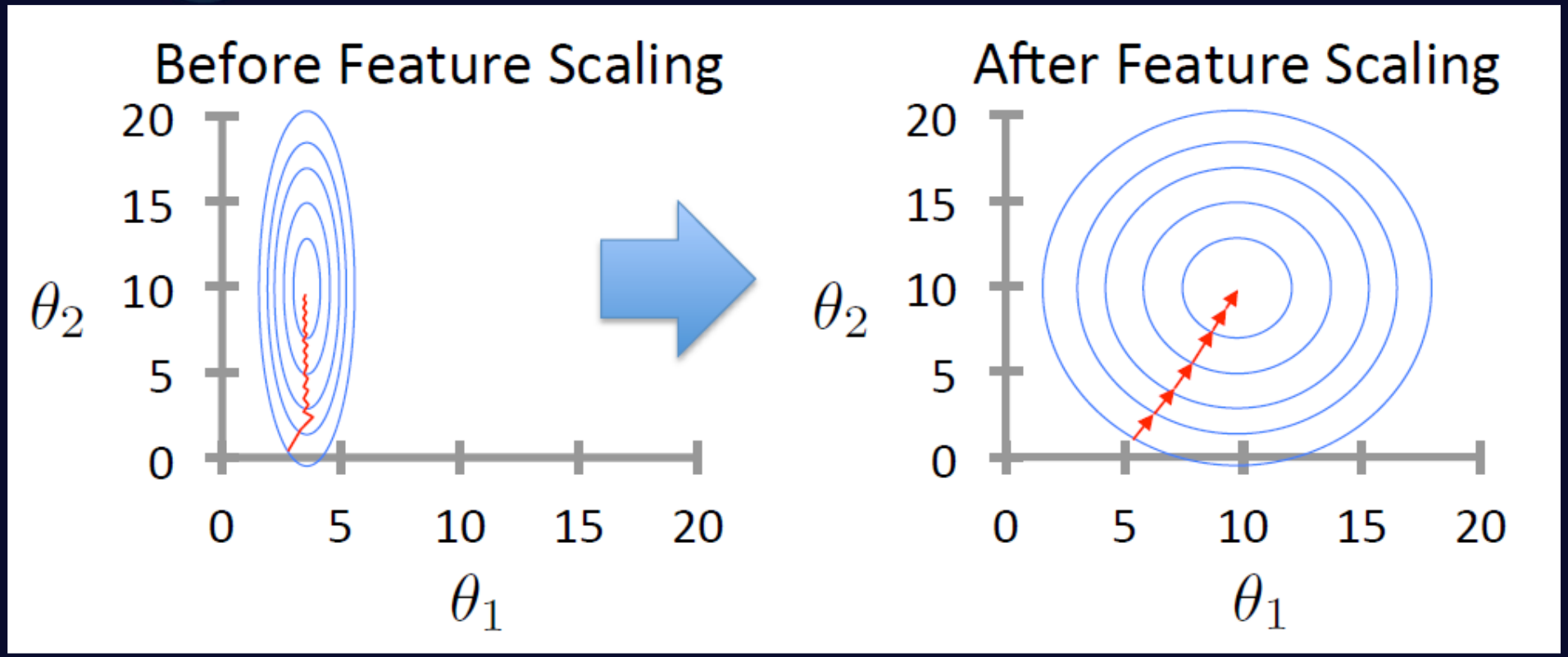
Choosing the learning rate



Gradient Descent	Closed Form Solution
• Requires multiple iterations • Need to choose α • Works well when n is large • Can support incremental learning	• Non-iterative • No need for α • Slow if n is large – Computing $(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X})^{-1}$ is roughly O(n^3)

Improving GD Convergence

- Make sure that the features are scaled



Improving GD Convergence

- Feature scaling:

$$x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_j}{s_j}$$

for j = 1...d
(not x_0!)