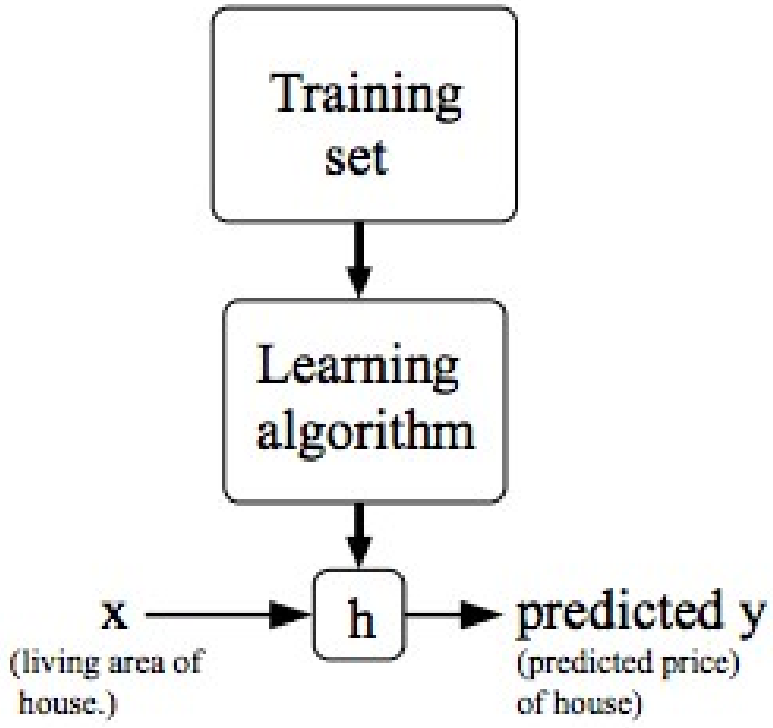
Représentation du modèle

Afin d'établir une notation pour une utilisation future, nous utiliserons $x^{(i)}$ pour désigner les variables "d'entrée" (la surface habitable dans cet exemple), également appelées caractéristiques d'entrée, et $y^{(i)}$ pour désigner la variable "de sortie" ou cible que nous essayons de prédire (le prix). Une paire $(x^{(i)}, y^{(i)})$ est appelée un exemple d'apprentissage, et l'ensemble de données que nous utiliserons pour apprendre - une liste de m exemples d'apprentissage $(x^{(i)}, y^{(i)})$; $i=1,\ldots,m$ - *est* appelé un ensemble d'apprentissage. Notez que l'exposant "(i)" dans la notation est simplement un index dans l'ensemble d'apprentissage, et n'a rien à voir avec l'exponentiation. Nous utiliserons également $\mathcal{X}$ pour désigner l'espace des valeurs d'entrée et $\mathcal{Y}$ pour désigner l'espace des valeurs de sortie. Dans cet exemple, $\mathcal{X} = \mathcal{Y} = \mathbb{R}$.

Pour décrire le problème de l'apprentissage supervisé de manière un peu plus formelle, notre objectif est, étant donné un ensemble d'apprentissage, d'apprendre une fonction $h : \mathcal{X} \to \mathcal{Y}$ afin que $h(x)$ soit un "bon" prédicteur de la valeur correspondante de y. Pour des raisons historiques, cette fonction h est appelée une hypothèse. Vu de manière imagée, le processus se présente donc comme suit :



Lorsque la variable cible que nous essayons de prédire est continue, comme dans notre exemple de logement, nous appelons le problème d'apprentissage un problème de régression. Lorsque y ne peut prendre qu'un petit nombre de valeurs discrètes (comme si, étant donné la surface habitable, nous voulions prédire si un logement est une maison ou un appartement, par exemple), nous parlons de problème de classification.

Fonction de coût

Nous pouvons mesurer la précision de notre fonction d'hypothèse en utilisant une **fonction de coût**. Celle-ci prend une différence moyenne (en fait une version plus sophistiquée d'une moyenne) de tous les résultats de l'hypothèse avec des entrées de x et la sortie réelle de y.

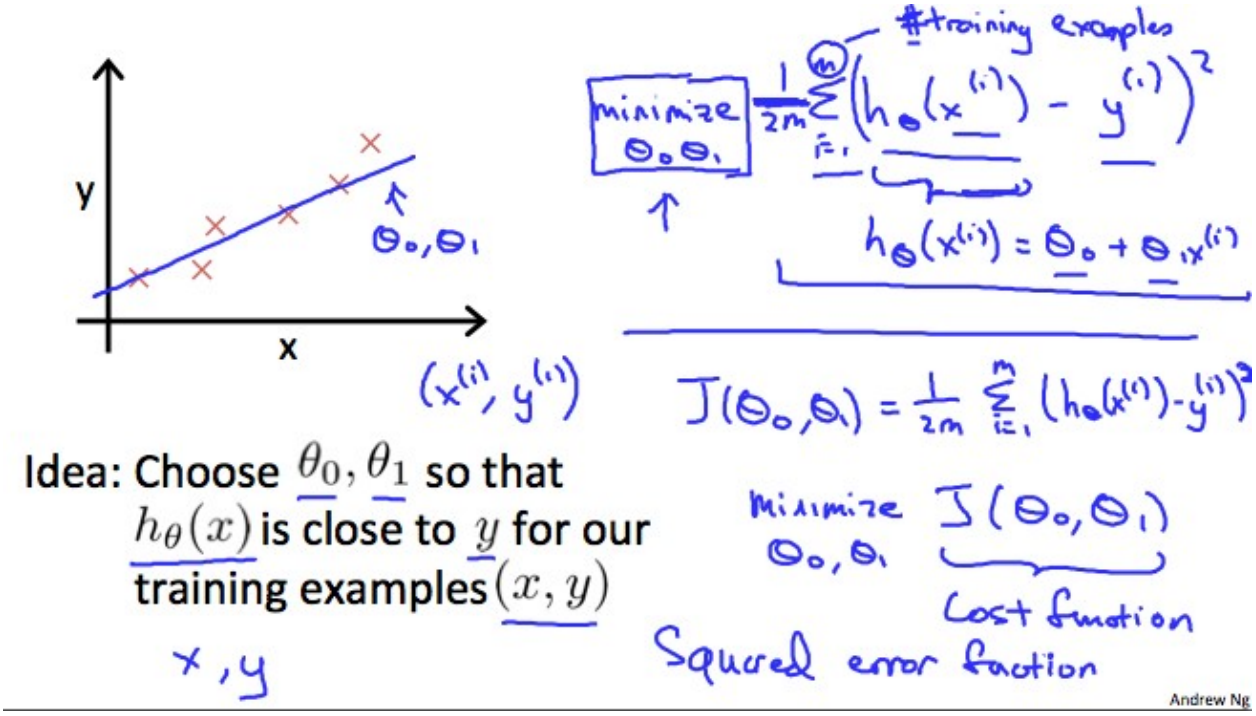
$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x_i) - y_i)^2$$

Pour le décomposer, c'est $\frac{1}{2}\ \bar{x}$ où $\bar{x}$ est la moyenne des carrés de , $h_\theta(x_i) - y_i$

ou la différence entre la valeur prédite et la valeur réelle.

Cette fonction est également appelée "fonction d'erreur quadratique", ou "erreur quadratique moyenne". La moyenne est divisée par deux $(1/2)$ par commodité pour le calcul de la descente du gradient, car le terme dérivé de la fonction carrée annulera le terme $(1/2)$

L'image suivante résume ce que fait la fonction de coût :



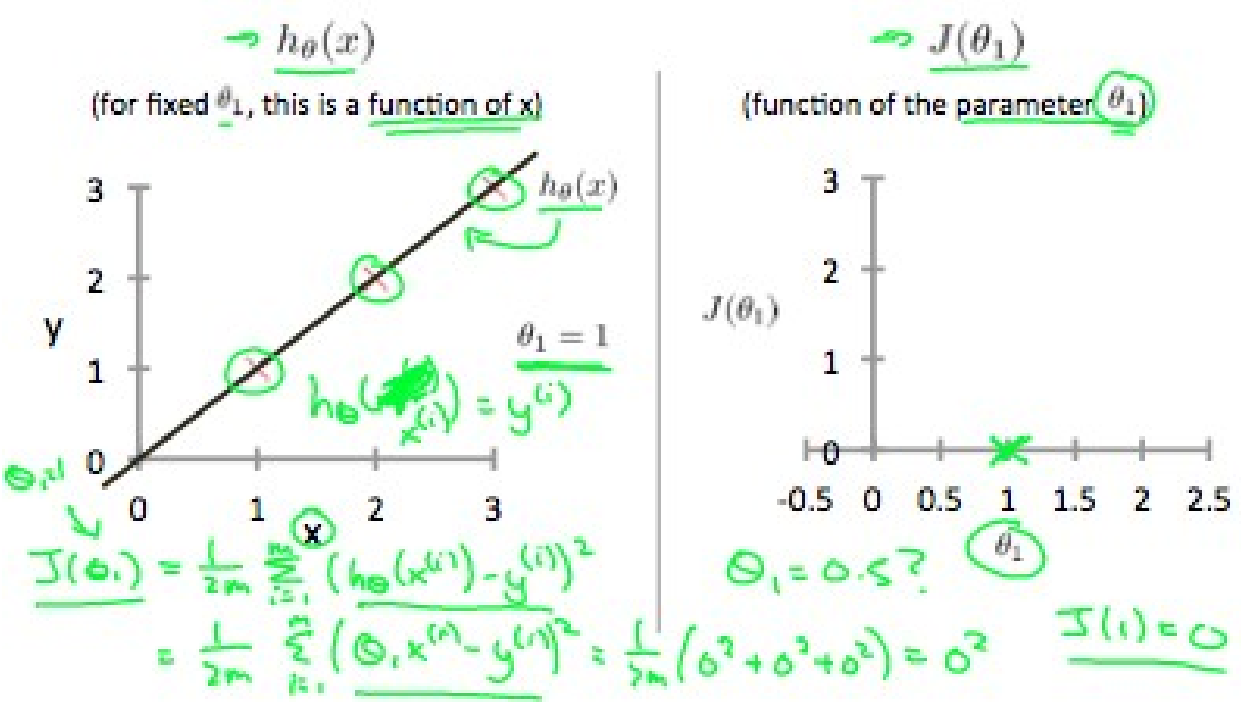
Fonction de coût - Intuition I

Si nous essayons d'y penser en termes visuels, notre ensemble de données d'apprentissage est éparpillé sur le plan x-y. Nous essayons de créer une ligne droite (définie par $h_\theta(x)$) qui passe par ces points de données dispersés.

Notre objectif est d'obtenir la meilleure ligne possible. La meilleure ligne possible sera telle que les distances verticales moyennes au carré des points dispersés par rapport à la ligne seront les plus faibles. Idéalement, la ligne devrait passer par tous les points de notre ensemble de données d'apprentissage.

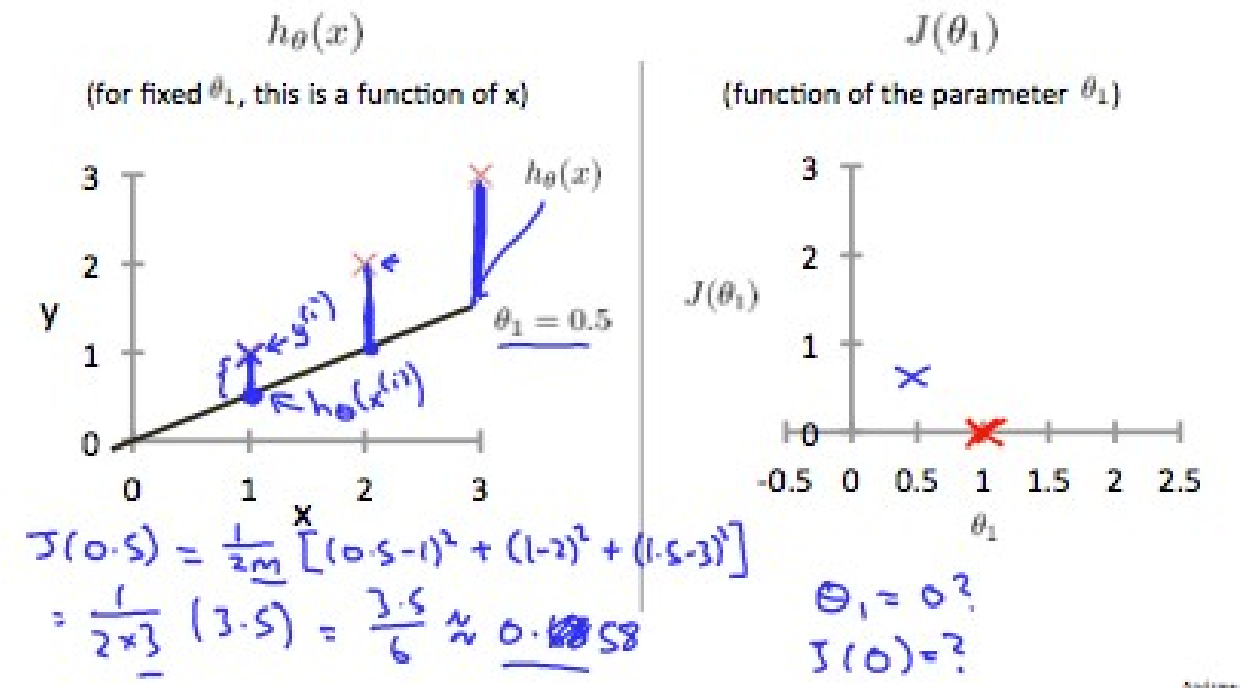
Dans un tel cas, la valeur de $J(\theta_0,\theta_1)$ sera 0.

L'exemple suivant montre la situation idéale où nous avons une fonction de coût de 0.

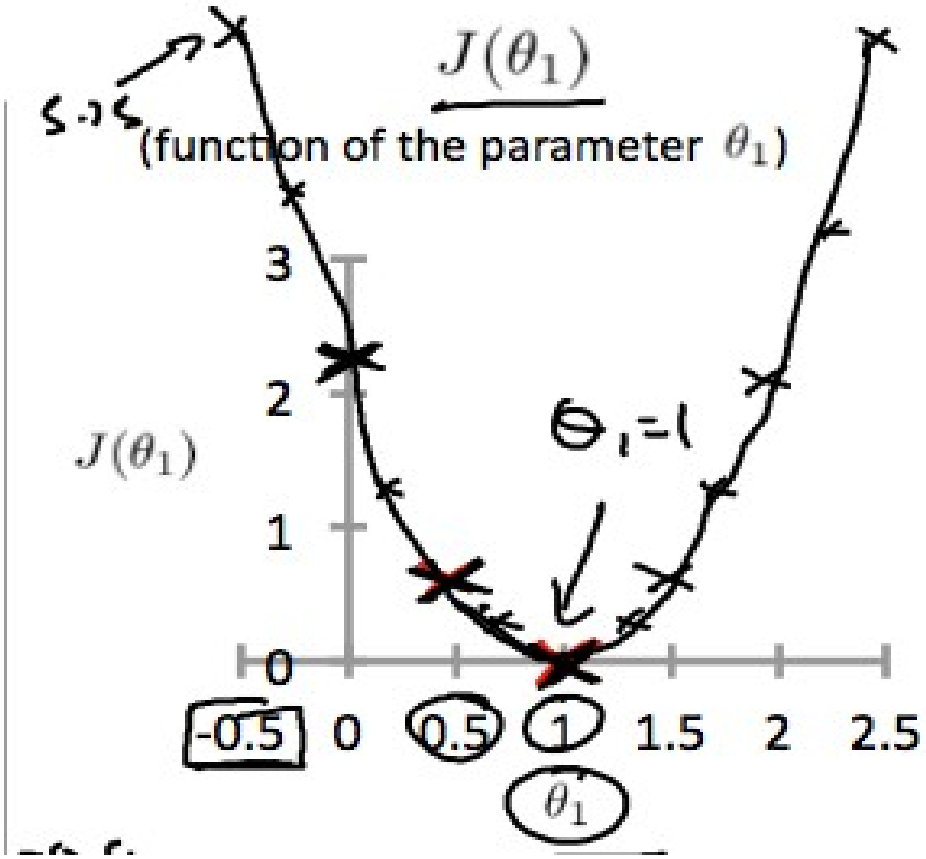


Lorsque $\theta_1=1$, nous obtenons une pente de 1 qui passe par chaque point de données de notre modèle.

À l'inverse, lorsque $\theta_1=0,5$, nous voyons la distance verticale entre notre ajustement et les points de données augmenter.



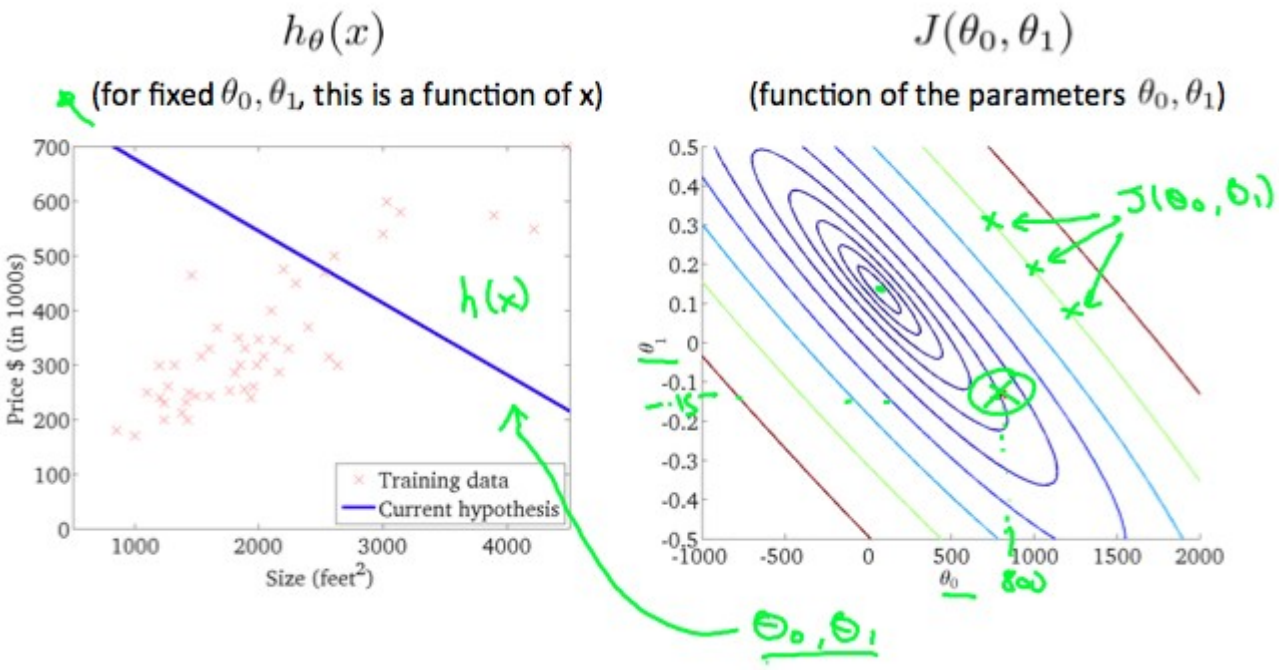
Cela porte notre fonction de coût à ***0,58***. En traçant plusieurs autres points, on obtient le graphique suivant :



Ainsi, comme objectif, nous devrions essayer de minimiser la fonction de coût. Dans ce cas, ***θ1=1*** est notre minimum global.

Fonction de coût - Intuition II

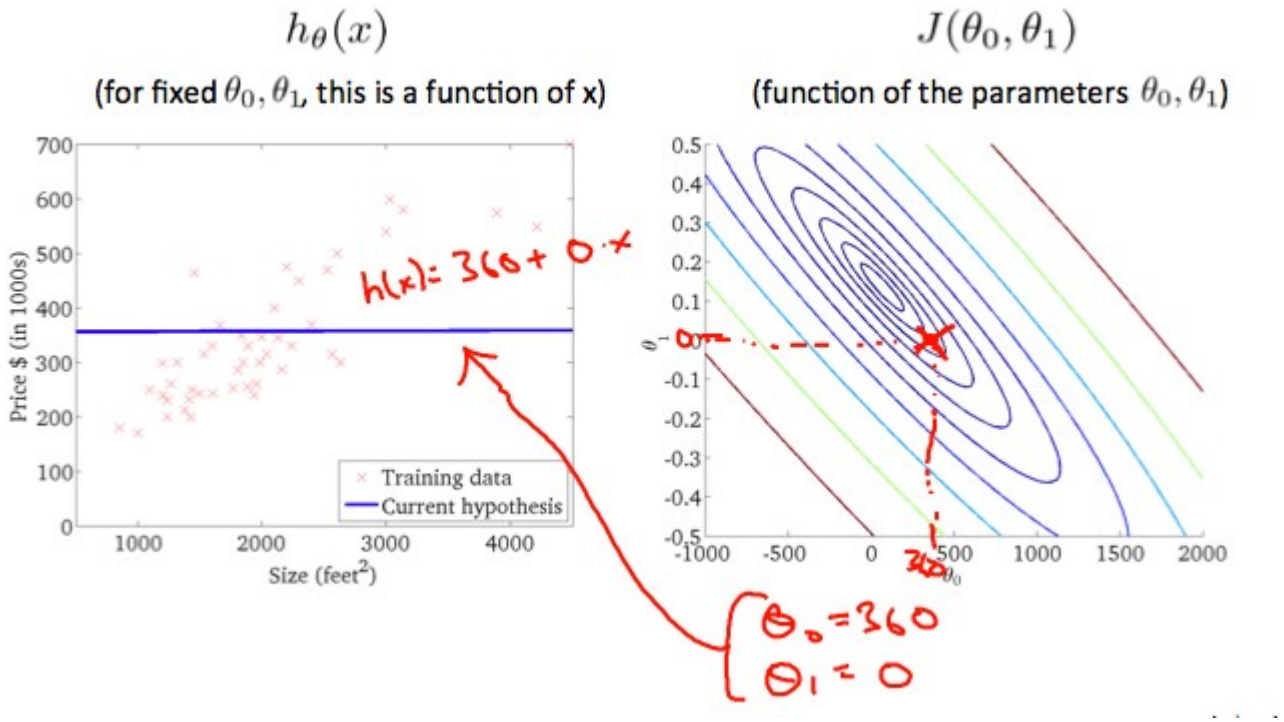
Un graphique de contour est un graphique qui contient de nombreuses lignes de contour. Une courbe de niveau d'une fonction à deux variables a une valeur constante à tous les points de la même ligne. Un exemple d'un tel graphique est celui qui figure à droite ci-dessous.



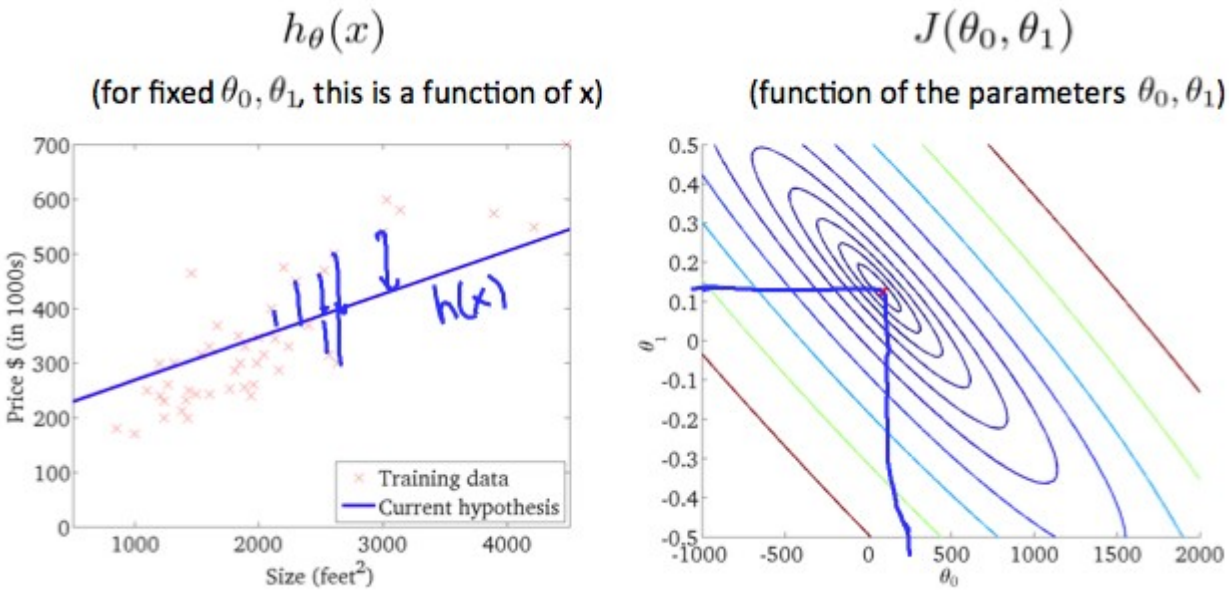
En prenant n'importe quelle couleur et en suivant le " cercle ", on s'attendrait à obtenir la même valeur de la fonction de coût.

Par exemple, les trois points verts trouvés sur la ligne verte ci-dessus ont la même valeur pour $J(\theta_0,\theta_1)$ et par conséquent, ils se trouvent le long de la même ligne. Le x encerclé affiche la valeur de la fonction de coût pour le graphique de gauche lorsque $\theta_0 = 800$ et $\theta_1 = -0{,}15$.

En prenant un autre $h(x)$ et en traçant son contour, on obtient les graphiques suivants :



Lorsque $\theta 0 = 360$ et $\theta 1 = 0$, la valeur de $J(\theta 0,\theta 1)$ dans le tracé du contour se rapproche du centre, ce qui réduit l'erreur de la fonction de coût. En donnant à notre fonction d'hypothèse une pente légèrement positive, on obtient un meilleur ajustement des données.



Le graphique ci-dessus minimise la fonction de coût autant que possible et, par conséquent, les résultats de $\theta 1$ et $\theta 0$ tendent à être autour de $0{,}12$ et 250 respectivement. Le tracé de ces valeurs sur notre graphique de droite semble placer notre point au centre du "cercle" le plus intérieur.

Nous commençons à parler d'un algorithme pour trouver automatiquement cette valeur de theta zéro et theta un qui **minimise la fonction de coût J**.