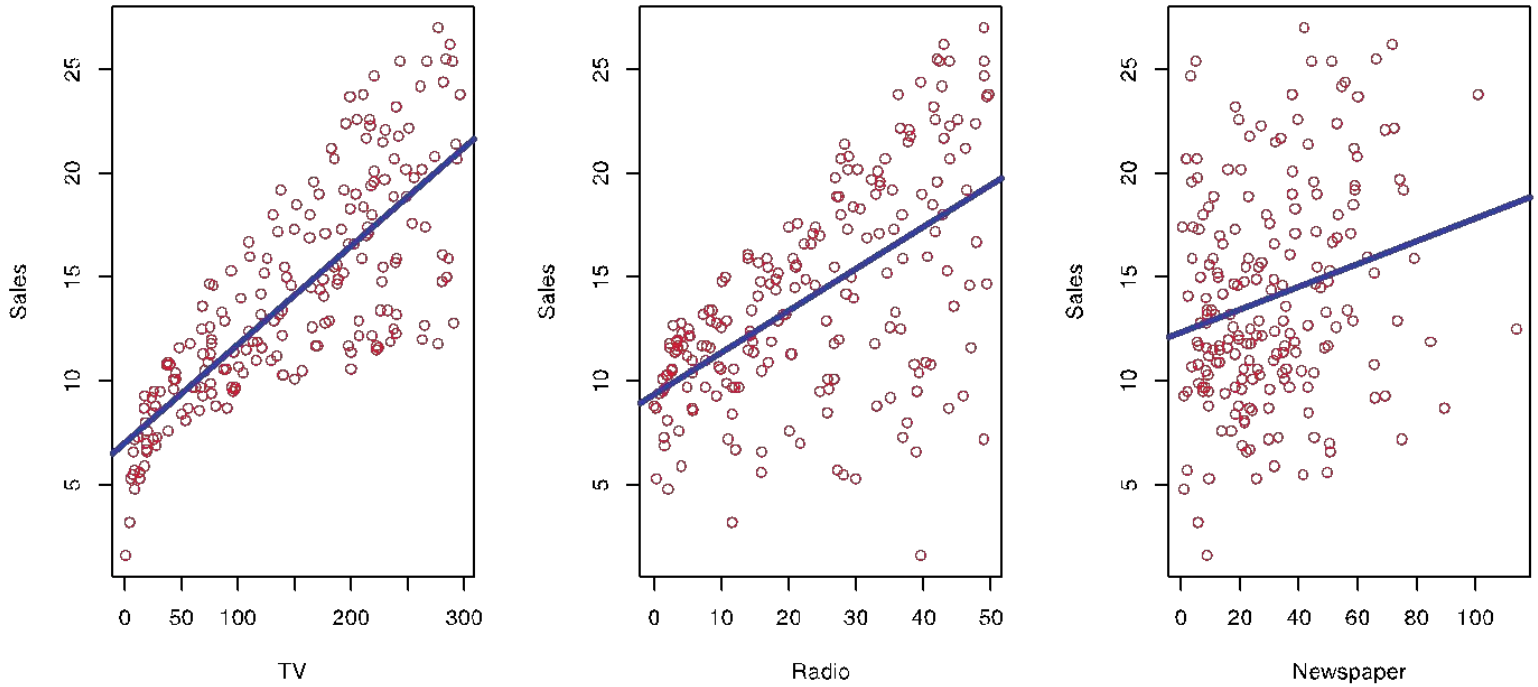
Cap. 3: Regressão Linear

Cristiano Leite de Castro - crislcastro@ufmg.br
André Paim Lemos – andrepaim@ufmg.br

Motivação

- Dada a base de dados *Advertising* referente a relação entre o valor gasto em milhares de dólares em propagandas (TV, rádio e jornais) sobre um determinado produto e sua venda (em milhares de unidades)



Advertising Data Set

```
[3]: advertising = pd.read_csv('Data/Advertising.csv', usecols=[1,2,3,4])
     advertising.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   TV         200 non-null    float64
 1   Radio      200 non-null    float64
 2   Newspaper  200 non-null    float64
 3   Sales      200 non-null    float64
dtypes: float64(4)
memory usage: 6.4 KB
```

```
[5]: advertising.head()
```

[5]:

	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

Motivação

- Existe alguma relação entre o orçamento de propaganda e as vendas?
- Quão forte é a relação entre o orçamento e as vendas?
- Qual mídia contribui mais significativamente para as vendas (TV, rádio ou jornal)?
- Com que acurácia podemos estimar o efeito de cada mídia nas vendas?
- Com que acurácia podemos prever vendas futuras?
- A relação é linear?
- Existe alguma sinergia entre as mídias?

Regressão Linear Simples

- Método simples para predizer uma variável de resposta quantitativa Y a partir de uma única variável preditiva X

- Assume-se que existe uma relação linear aproximada entre X e Y

$$Y \approx \beta_0 + \beta_1 X$$

- β_0 e β_1 são os *parâmetros* ou *coeficientes* do modelo

- Por exemplo, X pode ser definida como a valor em milhares de dólares gasto em propaganda de TV e Y como a quantidade de itens vendidos $(sales)$:

$$sales \approx \beta_0 + \beta_1 TV$$

Regressão Linear Simples

- Utiliza-se um *conjunto de dados de treinamento* para se estimar os parâmetros $\hat{\beta}_0$ e $\hat{\beta}_1$

- A partir do modelo resultante, podemos predizer valores futuros de vendas baseado em um determinado valor gasto com propaganda de TV através do modelo:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}x \tag{1}$$

- onde $\hat{y}$ é a previsão de Y dado o valor $X = x$

Estimação dos Coeficientes

- No caso da base *Advertising*, caso desejamos construir um modelo linear que relaciona o gasto com propaganda em TV com as vendas, utilizamos o conjunto de treinamento é composto por $n = 200$ pares de observações

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

- Esses pares são utilizados para estimar os parâmetros do modelo ($\hat{\beta}_0$, $\hat{\beta}_1$) de forma que a reta resultante seja mais próxima o possível das 200 observações.

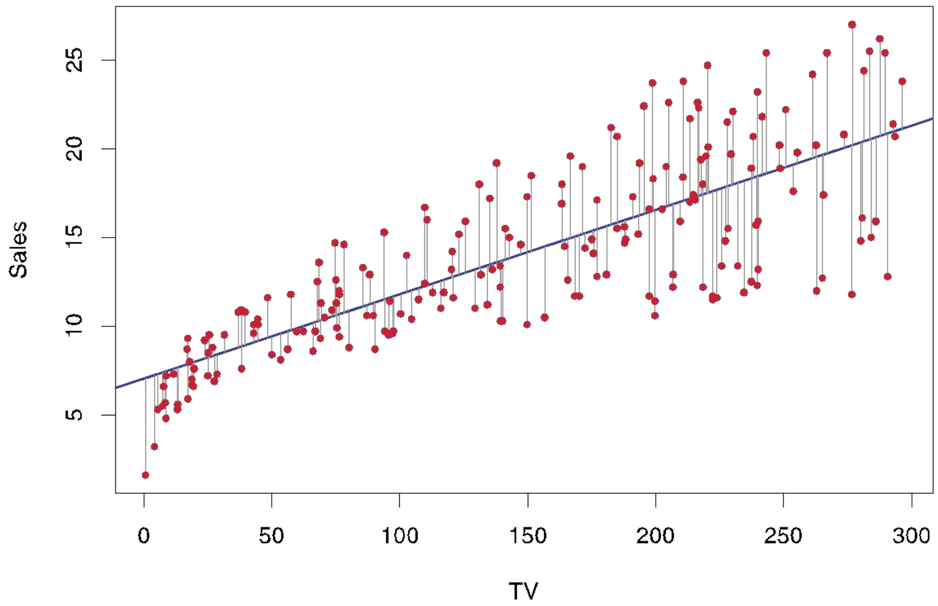
- Existem diversas formas de medir a *proximidade* entre o modelo resultante e as amostras de treinamento. A abordagem mais comum é o critério de *mínimos quadrados*.

Mínimos Quadrados

- Define-se o *resíduo* como $e_i = y_i - \hat{y}_i$ e a *soma dos quadrados dos resíduos* como:

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2$$

- Escolhe-se os valores dos parâmetros que minimizem RSS.



Mínimos Quadrados

Derivando-se SSE em função dos parâmetros e igualando a 0, tem-se:

The Sum of Squared Errors (SSE) for the linear regression model:

$$SSE = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Derivative with respect to β_0:

$$\frac{\partial SSE}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1)$$

$$= -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)$$

Derivative with respect to β_1:

$$\frac{\partial SSE}{\partial \beta_1} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i)$$

$$= -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i)$$

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

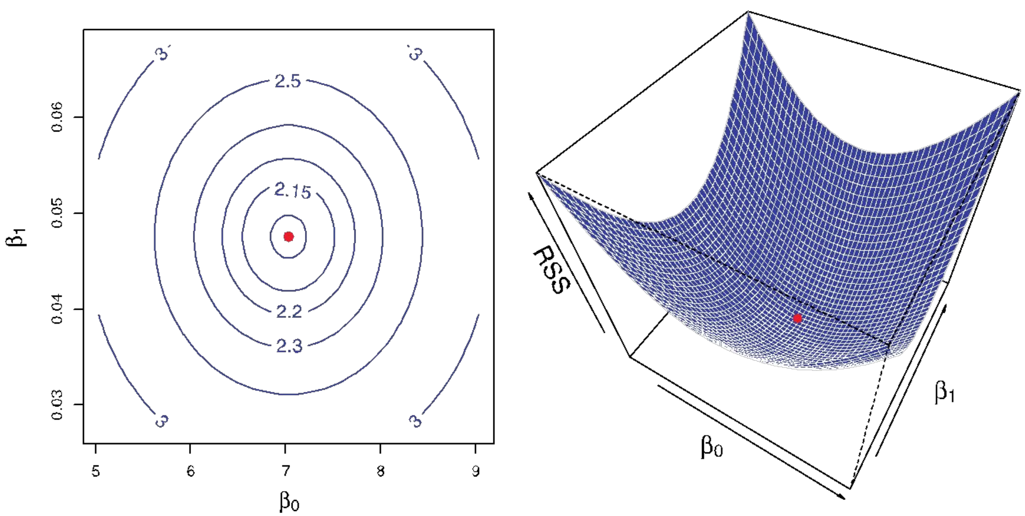
$$\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

Mínimos Quadrados

- Derivando-se RSS em função de cada uma os parâmetros e igualando-se a zero, encontra-se os seguintes valores para os parâmetros:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



Exercício

- Considere o seguinte modelo:

$$Y \approx \beta_0$$

- Dado um conjunto de n observações $y_1, y_2, \cdots, y_n$, encontre o estimador do parâmetro $\hat{\beta}_0$ utilizando utilizando o critério de mínimos quadrados

Solução do Exercício

$$SSE = \sum_{i=1}^{n}(y_i - \beta_0)^2.$$

Step 1: Compute the derivative of SSE with respect to β_0

$$\frac{d}{d\beta_0}SSE = \sum_{i=1}^{n} 2(y_i - \beta_0)(-1).$$

$$= -2\sum_{i=1}^{n}(y_i - \beta_0).$$

Step 2: Set the derivative to zero for minimization

$$-2\sum_{i=1}^{n}(y_i - \beta_0) = 0.$$

$$\sum_{i=1}^{n}(y_i - \beta_0) = 0.$$

Step 3: Solve for β_0

$$\sum_{i=1}^{n} y_i = n\beta_0.$$

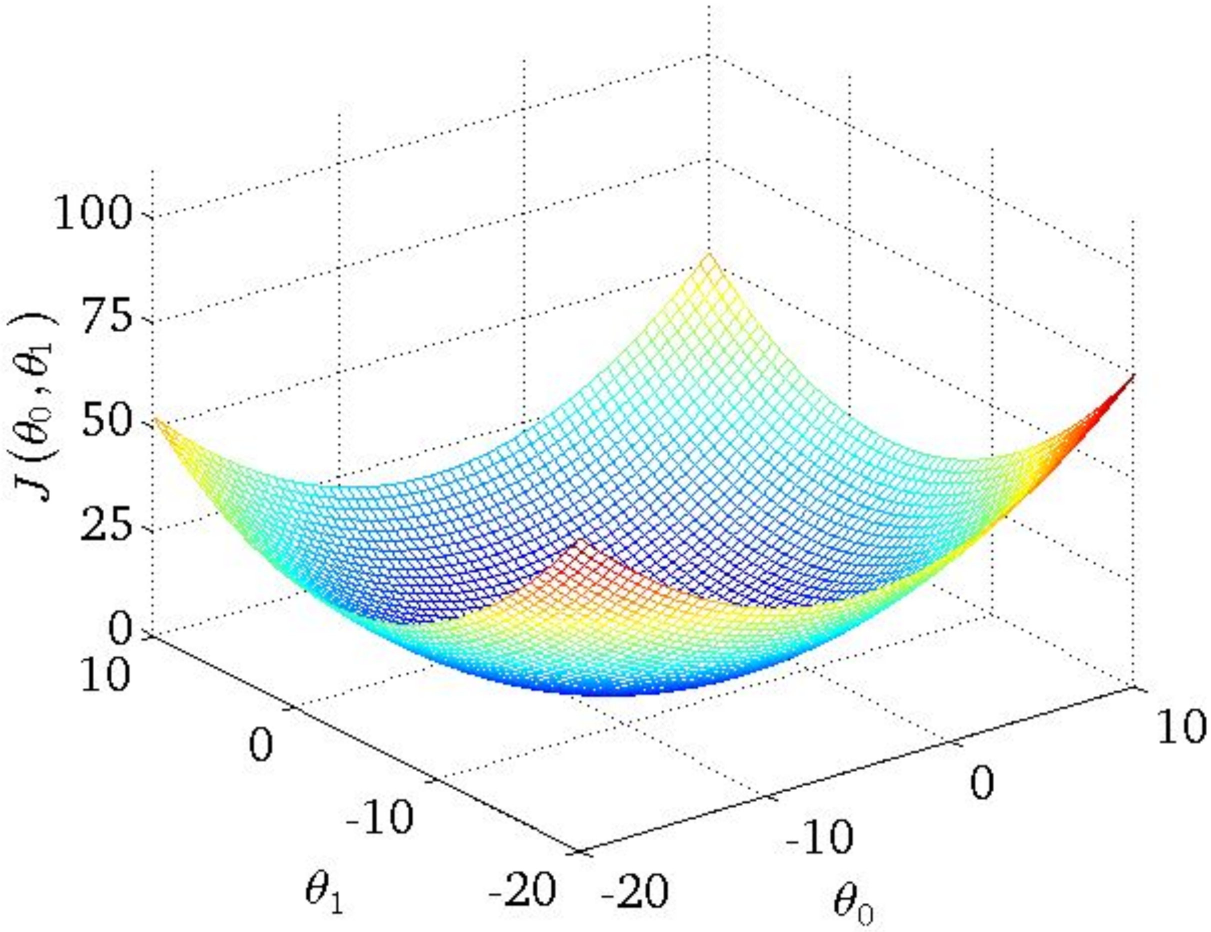
$$\beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Conclusion:

The value of β_0 that minimizes SSE is the **mean** of y_i:

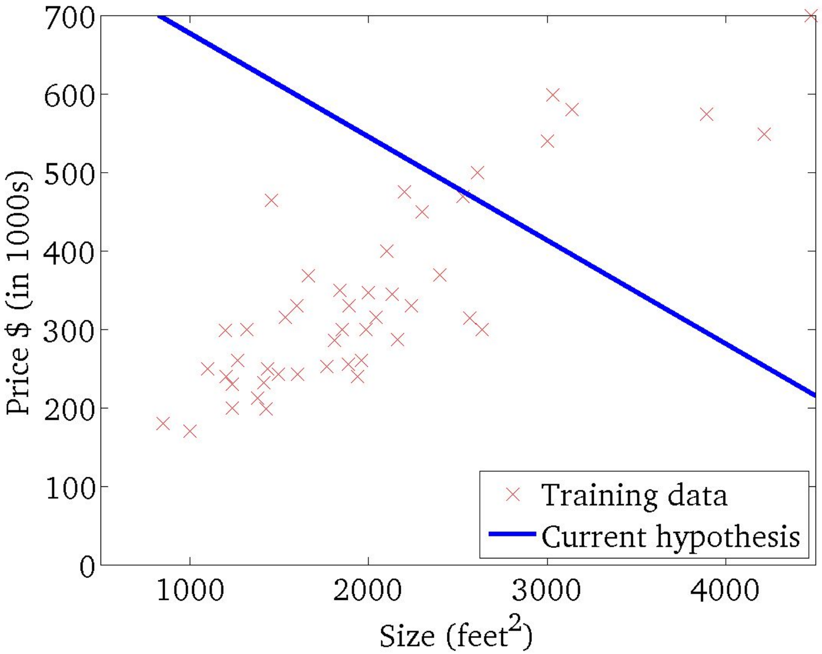
$$\beta_0 = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Gradient Descent



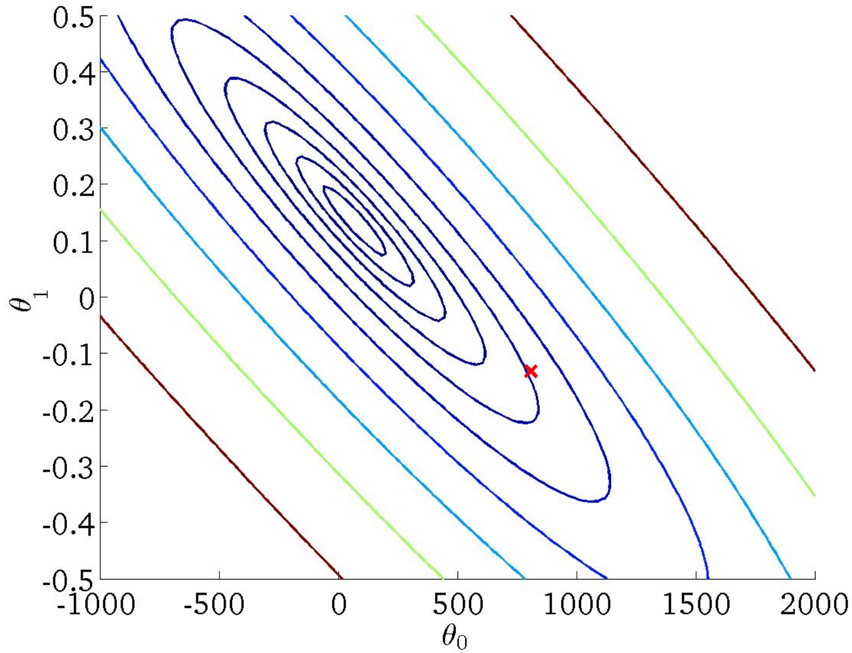
$$h_\theta(x)$$

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



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

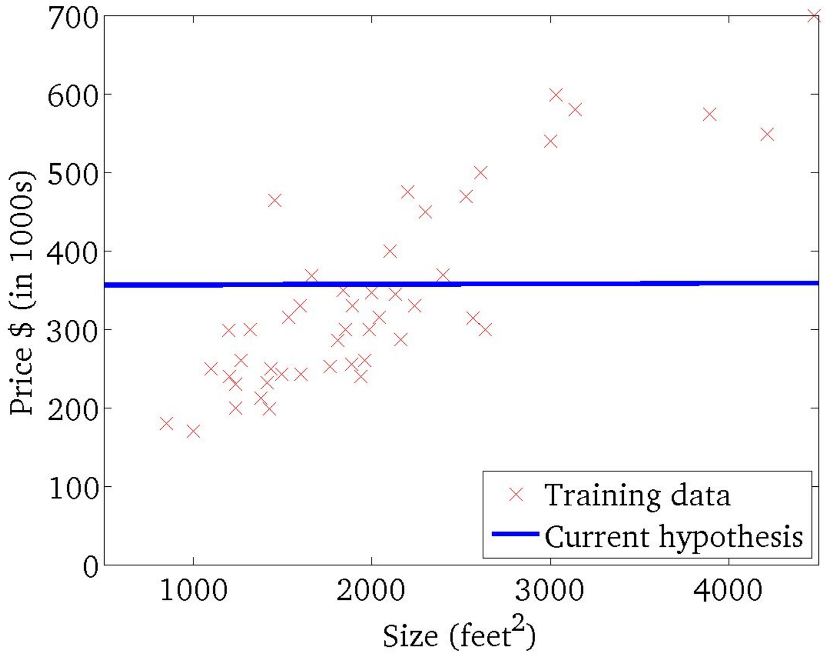
(function of the parameters θ_0, θ_1)



$$f(x) = -0.15x + 800$$

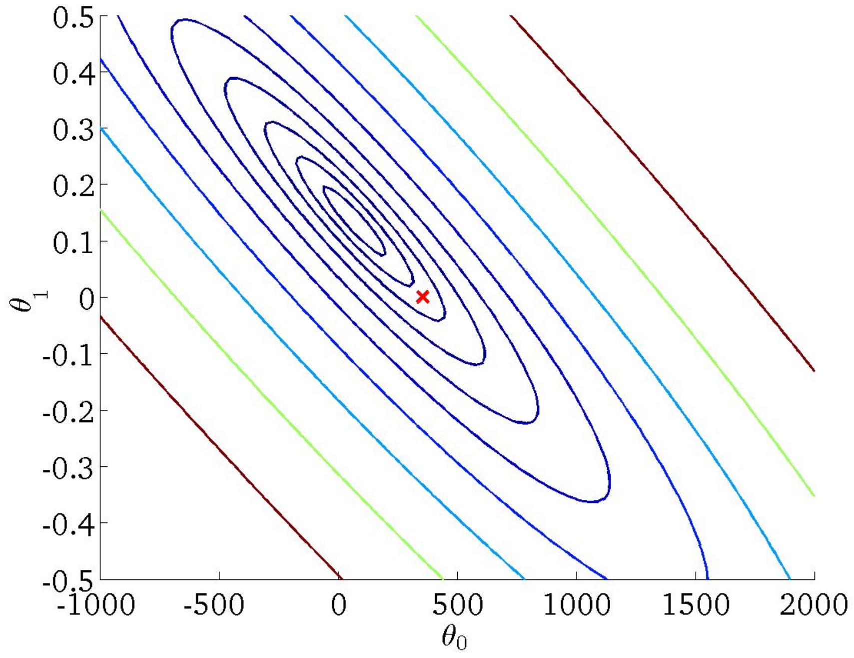
$$h_\theta(x)$$

(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)

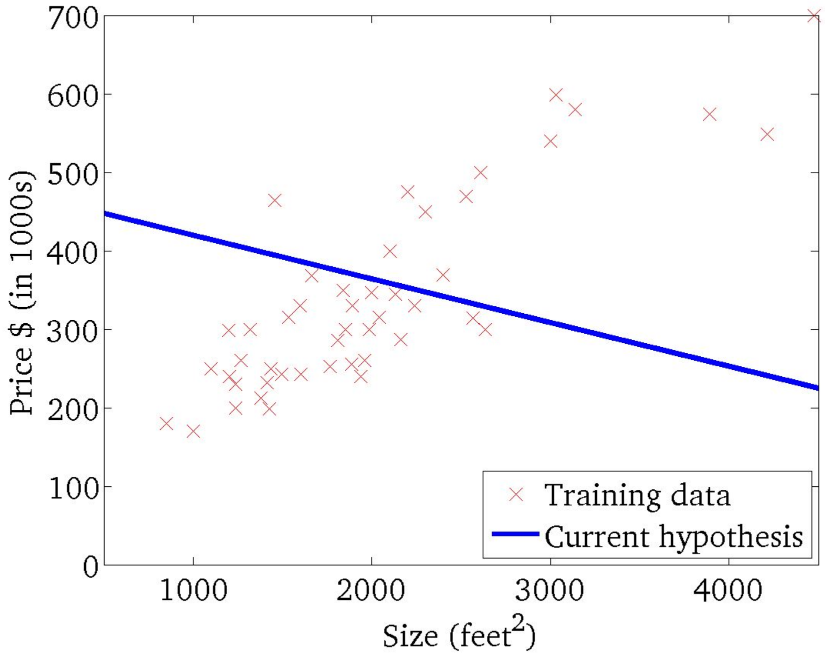


$$f(x) = 0x + 360$$

Andrew Ng

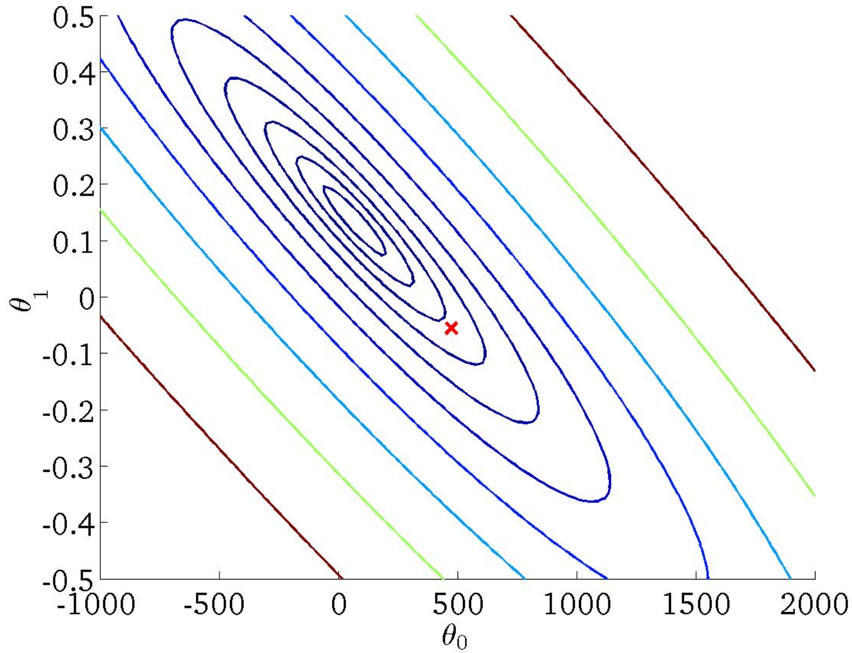
$$h_\theta(x)$$

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



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

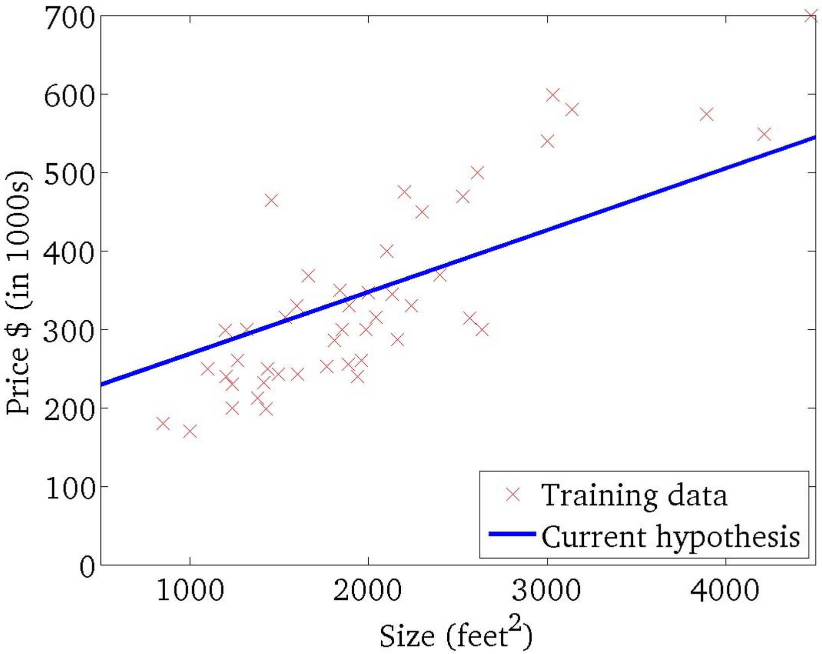
(function of the parameters θ_0, θ_1)

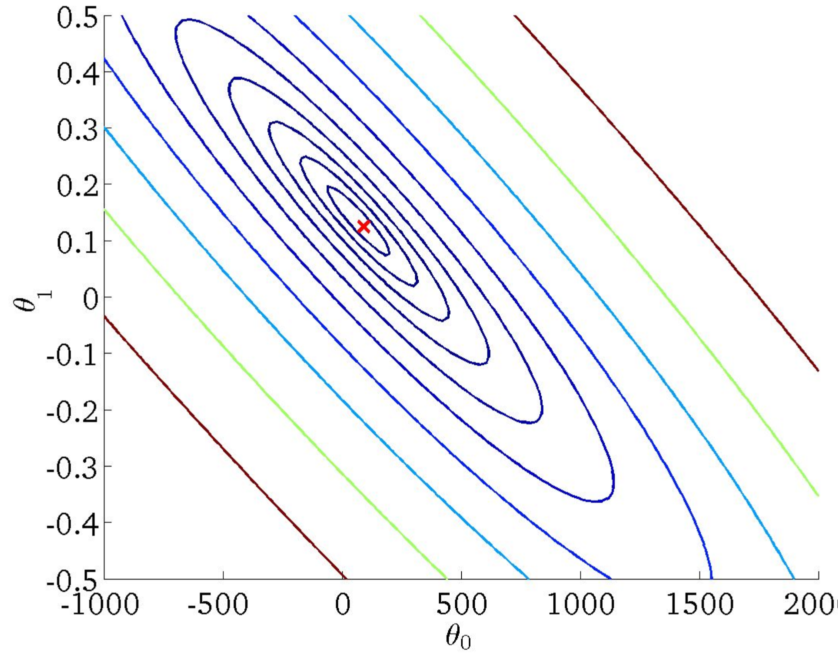


$$f(x) = -0.025x + 460$$

$h_\theta(x)$ $J(\theta_0, \theta_1)$

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





$$f(x) = 0.13x + 210$$

Widrow-Hoff Learning Rule

Gradient descent algorithm

$$\text{repeat until convergence } \{$$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$

$$(\text{for } j = 1 \text{ and } j = 0)$$

$$\}$$

Linear Regression Model

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

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Widrow-Hoff Learning Rule

- assuming we have only **one training example** (x, y), so that we can neglect the sum in the definition of J.

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} \left(h_\theta(x) - y\right)^2 \\
&= 2 \cdot \frac{1}{2} \left(h_\theta(x) - y\right) \cdot \frac{\partial}{\partial \theta_j} \left(h_\theta(x) - y\right) \\
&= \left(h_\theta(x) - y\right) \cdot \frac{\partial}{\partial \theta_j} \left(\sum_{i=0}^{n} \theta_i x_i - y \right) \\
&= \left(h_\theta(x) - y\right) x_j
\end{aligned}
$$

Widrow-Hoff Learning Rule

For a single training example, this gives the update rule:[1]

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)}.$$

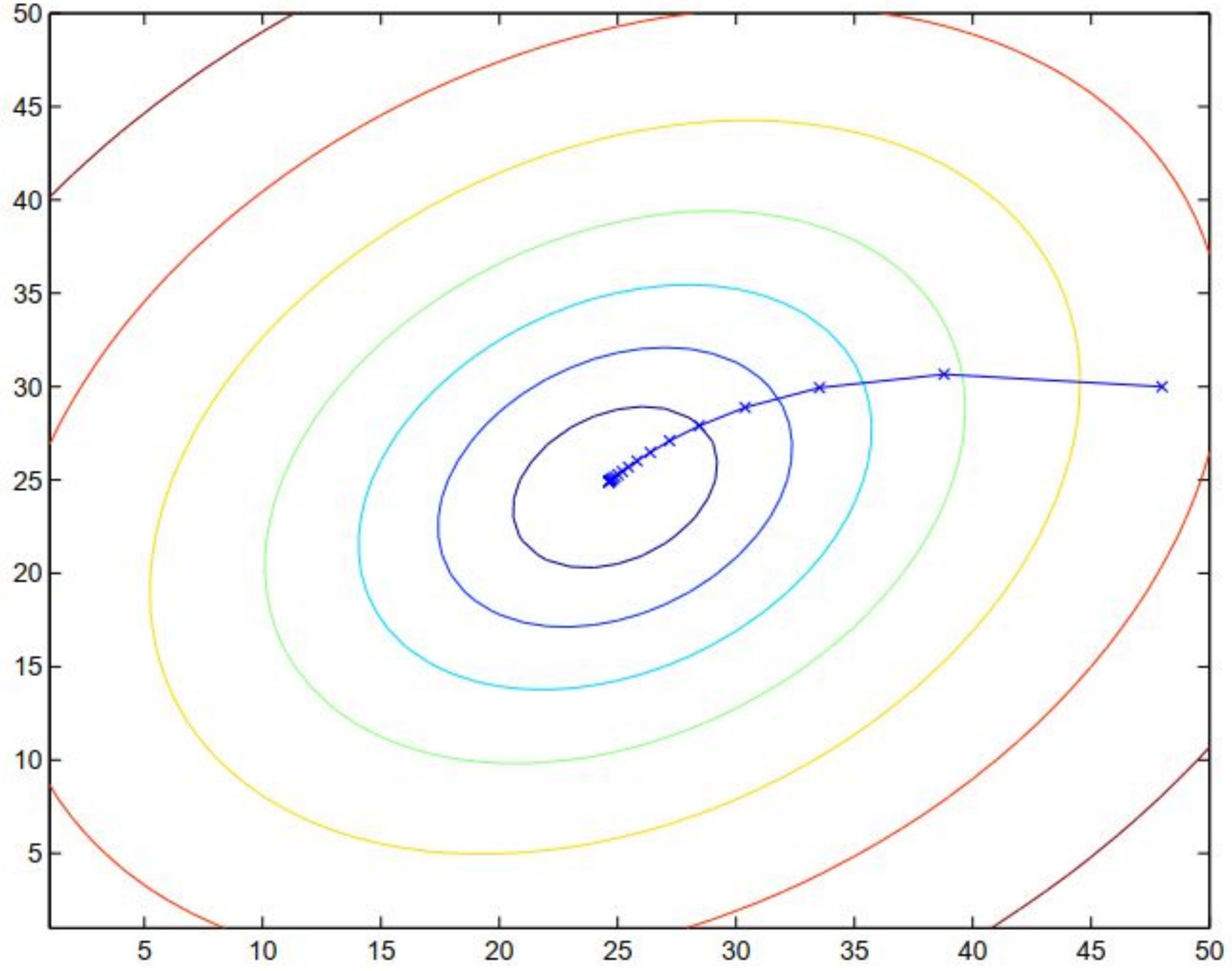
- Gradient Descent (Widrow-Hoff) Rule considerando m training examples:

Repeat until convergence {

$$\theta_j := \theta_j + \alpha \sum_{i=1}^{m} \left(y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)} \qquad \text{(for every } j\text{)}.$$

}

Widrow-Hoff Learning Rule



Reta de Regressão Populacional

- O modelo aproximado pode ser escrito como

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

- β_0 é o coeficiente linear (qual o valor esperado de Y quando $X = 0$) e β_1 o coeficiente angular (qual a variação de Y dado um incremento de X em uma unidade).

- O termo ϵ é definido como o *erro* do modelo e representa tudo o que o modelo não pode representar:

  - a relação não ser realmente linear
  - o fato de existirem outras variáveis que causem variações em Y
  - erro de medição das variáveis.

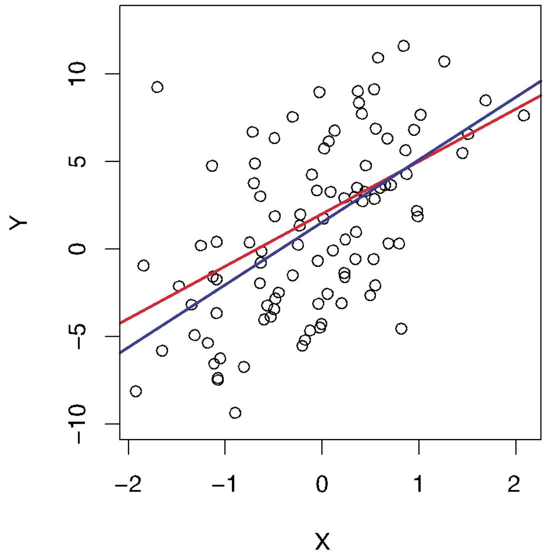
Reta de Regressão Populacional

- *Reta de Regressão Populacional* representa a melhor aproximação *linear* entre X e Y.

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

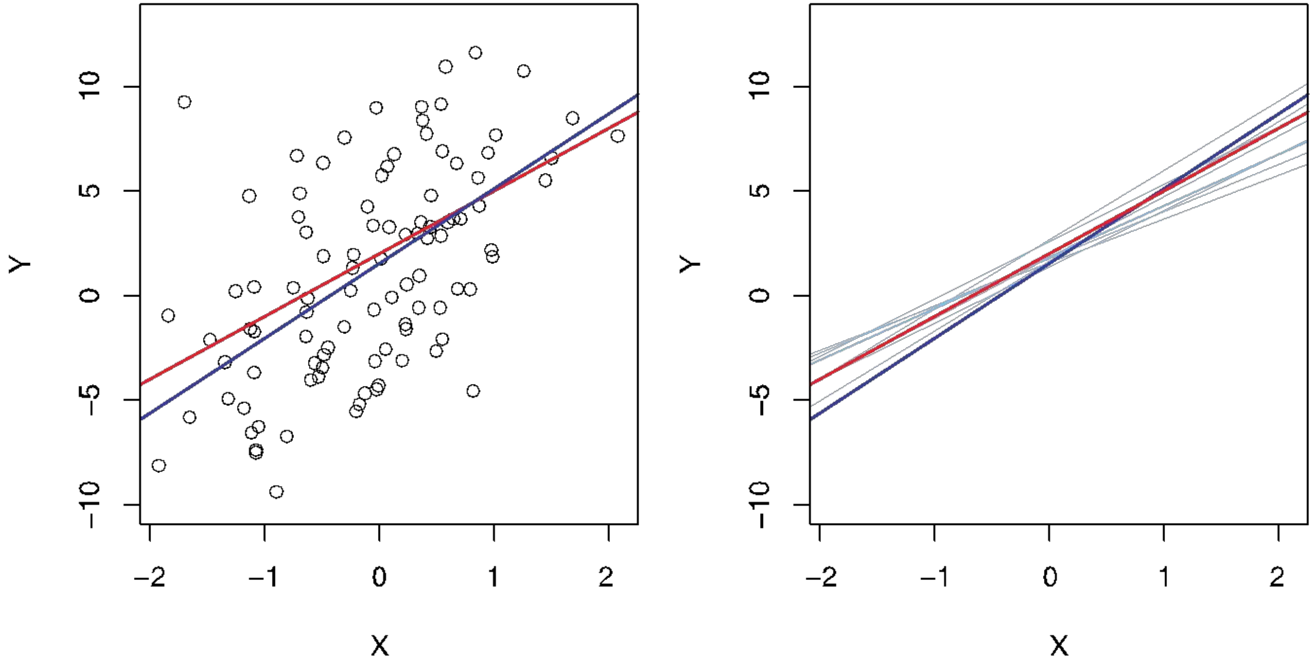
- Os coeficientes estimados pelo critério de mínimos quadrados representam a *Reta de Regressão de Mínimos Quadrados*

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



Reta de Regressão Populacional

- A *Reta de Regressão Populacional* é, geralmente, desconhecida
- A *Reta de Regressão de Mínimos Quadrados* pode ser estimada a partir de observações
  - É dependente das observações utilizadas na estimativa



Parâmetros Populacionais

- β_0 e β_1 são os *parâmetros populacionais* que desejamos conhecer
- Para isso, utiliza-se uma estimativa, ou seja, $\hat{\beta}_0$ e $\hat{\beta}_1$ são os estimadores amostrais desses parâmetros.
- Exemplo, deseja-se saber o peso médio de uma população
  - A *média populacional* μ da variável aleatória peso Y é desconhecida
  - Caso tenhamos acesso a um conjunto de n observações de Y, $y_1, y_2, \cdots, y_n$; podemos *estimar* μ
  - Uma boa estimativa seria $\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, ou seja, a *média amostral*
  - A média amostral e populacional são diferentes, porém, em geral, a média amostral pode ser uma boa estimativa da média populacional
  - Da mesma forma os coeficientes populacionais β_0 e β_1 são desconhecidos, porém podem ser estimados por $\hat{\beta}_0$ e $\hat{\beta}_1$

Viés do Estimador

- Considere o caso anterior da estimativa da média populacional através de $\hat{\mu}$
- O estimador $\hat{\mu} = \bar{y}$ é dito ser *não viciado* (não viesado)
  - A estimativa a partir de um determinado conjunto de observações pode resultar em um valor *sobrestimado* para μ e um outro conjunto de observações pode gerar um valor *subestimado*
  - Porém, a média de um grande número de estimativas é exatamente μ
- *Estimador não viciado* não sobrestima ou subestima o parâmetro populacional *sistematicamente*
- Os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ baseados no critério de mínimos quadrados são não viesados.

Erro Padrão

- *Dado que o estimador não é viesado, como medir sua acurácia para estimar os parâmetros populacionais?*
- Utiliza-se o *erro padrão* do estimador
  - No caso da média amostral, temos que:

$$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

  - onde σ^2 é o desvio padrão das realizações y_i da variável Y (dado que estas sejam independentes)
- O erro padrão é uma medida média de quanto a estimativa difere do parâmetro populacional.

Erro Padrão

- O erro padrão associado aos estivadores $\hat{\beta}_0$ e $\hat{\beta}_1$ são:

$$
\begin{aligned}
SE(\hat{\beta}_0)^2 &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] \\
SE(\hat{\beta}_1)^2 &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}
$$

- em que $\sigma^2 = Var(\epsilon)$, assumindo que os erros ϵ_i são descorrelacionados (o que nem sempre é verdade, mas pode ser uma boa aproximação)

- Geralmente, σ^2 não é conhecido, mas pode ser estimado como: $RSE = \sqrt{RSS/(n-2)}$

* o slide 7 apresenta um exemplo em que os erros não são descorrelacionados.

Intervalo de Confiança

- O erro padrão pode ser utilizado para computar um *intervalo de confiança*
- Um intervalo de confiança de 95% é definido como o intervalo de valores de forma que este contenha o valor desconhecido do parâmetro estimado com 95% de probabilidade
  - Ou seja, caso sejam realizadas várias estimativas do parâmetro desconhecido e o intervalo de confiança seja estimado para cada uma das estimativas, o parâmetro desconhecido estará dentro do intervalo 95% dadas vezes)
- Para a regressão linear, o intervalo de confiança de 95% da estimativa $\hat{\beta}_1$ pode ser aproximado por:

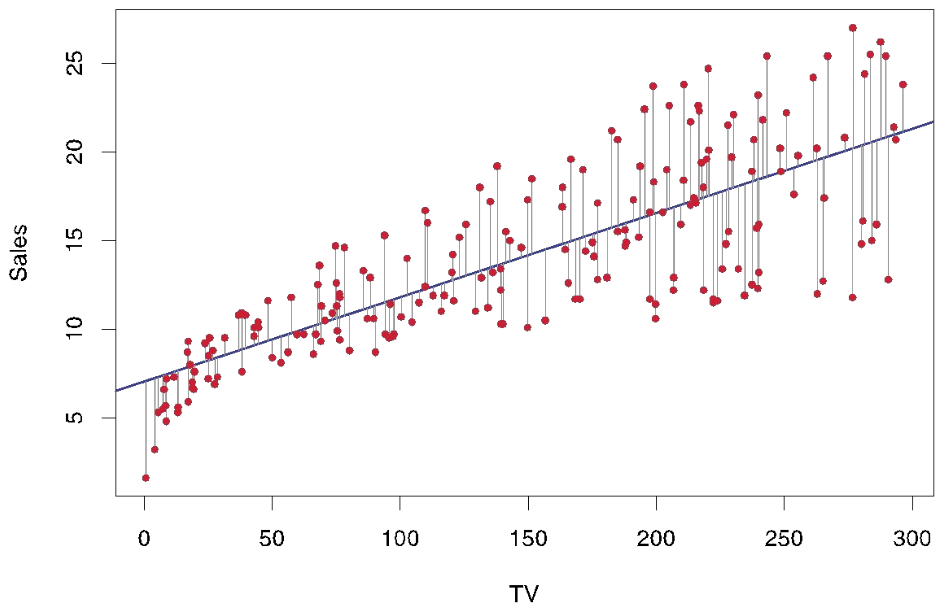
$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

Intervalo de Confiança

- Para o caso da base de dados *Advertising* ($Y = sales$, $X = TV$), o intervalo de confiança de 95% para β_0 é $[6.130, 7.935]$ e para β_1 é $[0.042, 0.053]$.

- Pode-se concluir que, na ausência de propaganda, o total de vendas será, na média, algum valor entre 6130 e 7940 unidades

- Além disso, cada aumento de \$1000 em propaganda de TV acarretará um aumento médio nas vendas entre 42 e 53 unidades.



Teste de Hipótese

- O erro padrão também pode ser utilizado para realizar *testes de hipótese* nos coeficientes

- O teste mais comum define a seguinte *hipótese nula*

$$H_0 : \beta_1 = 0$$

versus a *hipótese alternativa*

$$H_a : \beta_1 \neq 0$$

- Caso $\beta_1 = 0$, o modelo se resume a $Y = \beta_0 + \epsilon$ e X não é associado a Y, ou seja, não existe correlação entre X e Y

Teste de Hipótese

- Para testar a hipótese nula, é necessário determinar se a estimativa $\hat{\beta}_1$ é suficientemente distante de zero.

- *Como definir qual a distância mínima necessária?*

- Depende da acurácia da estimativa, ou seja, depende de $SE(\hat{\beta}_1)$
  - Caso $SE(\hat{\beta}_1)$ seja pequeno, mesmo valores pequenos de $\hat{\beta}_1$ podem prover forte evidência de que $\beta_1 \neq 0$
  - Caso contrário, $\hat{\beta}_1$ deve possuir um valor absoluto alto para que possamos rejeitar a hipótese nula

- Na prática, computa-se a *estatística t*:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

que terá distribuição *T de Student* com $n-2$ graus de liberdade, assumindo-se que $\beta_1 = 0$

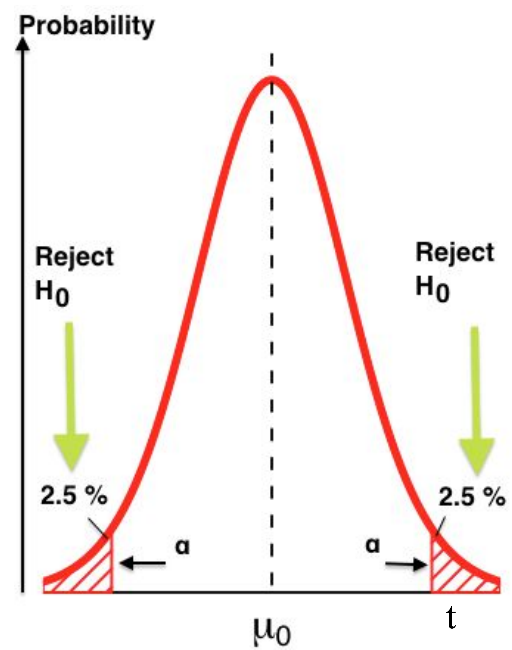
Teste de Hipótese

```
[11]: est = smf.ols('Sales ~ TV', advertising).fit()
      est.summary().tables[1]
```

[11]:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.0326	0.458	15.360	0.000	6.130	7.935
TV	0.0475	0.003	17.668	0.000	0.042	0.053

distribuição t para n>=30 tem o formato similar a uma distribuição normal.



p-value:

probabilidade de se observar valores iguais ou maiores que *t*, assumindo que $\boldsymbol{\beta}_1 = 0$.

se essa probabilidade for muito pequena (< 2.5% ou 1%), então a hipótese nula pode ser rejeitada.

Teste de Hipótese

	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

- os *coeficientes* são muito grandes comparados aos seus *erros-padrão*, o que resulta em valores das *estatisticas t* muito grandes.
- as probabilidades se observar tais valores de *t* quando H_0 é igual a 0 é muita pequena, virtualmente 0.
- conclui-se assim que a H_0 pode ser rejeitada.

Acurácia do Modelo

- Uma vez que a hipótese nula referente a ausência de correlação entre X e Y é rejeitada, uma pergunta natural seria quantificar a capacidade do modelo estimado de descrever os dados

- Uma medida para medir a qualidade do modelo é o *erro padrão do resíduo*

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

- Uma medida alternativa é a estatística R^2

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

proportion of variability in Y that can be explained using X

em que $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$

- R^2 corresponde ao coeficiente de correlação linear, para o modelo de regressão linear simples

Regressão Linear Múltipla

- Dado o base de dados *Advertising*, como estender a análise para as outras duas mídias?

- Solução Inicial: estimar um modelo linear simples para cada um das mídias

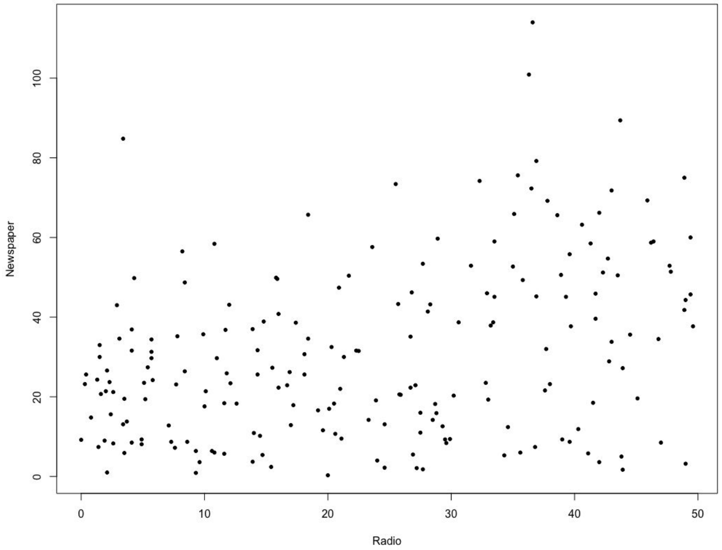
	Coefficient	Std. error	t-statistic	p-value
`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	< 0.0001

	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

Regressão Linear Múltipla

- Solução inicial não é satisfatória
- Como fazer um previsão de vendas dados os valores das três mídias?
- Ignora as correlações existentes entre as mídias



	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Regressão Linear Múltipla

- *Regressão Linear Múltipla*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

  em que p corresponde ao número de variáveis preditoras

- Interpreta-se β_i como o efeito médio em Y de um aumento de uma unidade em X_i, *caso todas as outras variáveis predadoras sejam fixas*

- Para a base *Advertising* temos

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper + \epsilon$$

Regressão Linear Múltipla

- De forma análoga a regressão linear simples, uma vez estimados os coeficientes $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$, pode-se fazer predições através da equação

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

- Os coeficientes também podem ser estimados utilizando o critério de mínimos quadrados

Ordinary Least Squares:

$$\theta = (X^T X)^{-1} X^T \vec{y}.$$

Regressão Linear Múltipla

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	−0.001	0.0059	−0.18	0.8599

TABLE 3.4. *For the* `Advertising` *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

Regressão Linear Múltipla

- *Faz sentido a variável preditiva Newspaper ser significativa na regressão linear simples e não significativa na regressão múltipla?*
- Sim! *Newspaper* tem uma correlação de 0.35 com *Radio*
- Tendência de se gastar mais dinheiro com propaganda em jornal em mercados em que mais dinheiro é gasto com propaganda em rádio
- Regressão linear simples, somente baseada na variável *Newspaper*, podemos observar que valores altos para *Newspaper* estejam associados a valores altos de *sales*, mesmo que propaganda em jornal não afete as vendas
- Outro exemplo, uma regressão entre ataques de tubarão e o consumo de sorvete em praias pode apresentar uma relação positiva significativa apesar do aumento do consumo de sorvete não causar aumento nas ocorrências de ataques de tubarão

a variável newspaper “pega carona” no efeito que a variável radio provoca em vendas

Regressão Linear Múltipla

- Ao menos uma das variáveis preditoras $X_1, X_2, \cdots, X_p$ são relevantes para a previsão?
- Todas as variáveis contribuem para prever Y ou apenas um subconjunto?
- Qual a qualidade do modelo?
- Dado um conjunto de valores das variáveis preditoras, qual o valor deve ser previsto e qual a acurácia da previsão?

Presença de Relação entre X e Y

- Para responder a primeira pergunta, testamos a seguinte hipótese nula:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

versus a hipótese alternativa

$$H_a : \text{ao menos um } \beta_j \text{ é não nulo}$$

- Essa hipótese é testada utilizando a *estatística F*

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$$

- Caso a hipótese nula seja verdadeira, a estatística F segue uma distribuição F com p e $n - p - 1$ graus de liberdade (calcula-se o p-valor e é possível determinar se a hipótese nula pode ser rejeitada)

Presença de Relação entre X e Y

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

TABLE 3.6. *More information about the least squares model for the regression of number of units sold on TV, newspaper, and radio advertising budgets in the* `Advertising` *data. Other information about this model was displayed in Table 3.4.*

- when there is no relationship between the response and predictors, one would expect the F -statistic to take on a value close to 1.

Escolhendo as Variáveis Importantes

- A abordagem de usar a estatística F para testar qualquer associação entre os preditores e a resposta funciona bem quando p é relativamente pequeno.
- Quando o número de preditores é tão grande quanto, ou maior que o número de observações, então deve-se usar uma abordagem para *Seleção de Preditores*
- Abordagem ingênua para *seleção de preditores* envolve testar todos os modelos possíveis
  - $p = 2$: 4 modelos candidatos
  - $p = 30$: $2^{30} = 1073741824$ modelos candidatos!
  - Só aplicável, se p é muito pequeno!

Escolhendo as Variáveis Importantes

- Uma solução seria utilizar uma heurística
- *Forward Selection*:
  1. Incia-se com um modelo contendo apenas β_0
  2. Testa-se p modelos de regressão linear simples, cada um incluindo uma das p variáveis
  3. Escolhe-se o modelo associado ao menor RSS
  4. Repete-se os passos 2 e 3, testando as variáveis remanescentes até que um critério de parada seja aingido
- Exemplo de critério de parada: a adição de qualquer uma das variáveis remanescentes tenha um p-valor maior que um limiar

Escolhendo as Variáveis Importantes

- Uma solução seria utilizar uma heurística
- *Backward Selection*:
  1. Incia-se com todas as variáveis no modelo
  2. Remove-se a variável associada ao maior p-valor (variável menos significante)
  3. Recalcula-se o novo modelo com $p - 1$ variáveis
  4. Repete-se os passos 2 e 3 até que um critério de parada seja atingido
- Exemplo de critério de parada: todas as variáveis remanescentes no modelo tenham um p-valor menor que um limiar

Previsões

- Uma vez estimado o modelo, pode-se estimar valores de Y, baseado em valores observados de X

- Porém, deve-se considerar que:

  - Os coeficientes $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ são estimativas de $\beta_0, \beta_1, \cdots, \beta_p$
  - *O plano de mínimos quadrados*

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

  é uma estimativa do *verdadeiro hiperplano populacional*

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Referências

- Capítulo 3 do livro James, Gareth, et al. *An Introduction to Statistical Learning*. Vol. 112. New York: Springer, 2013 – Section 10.3