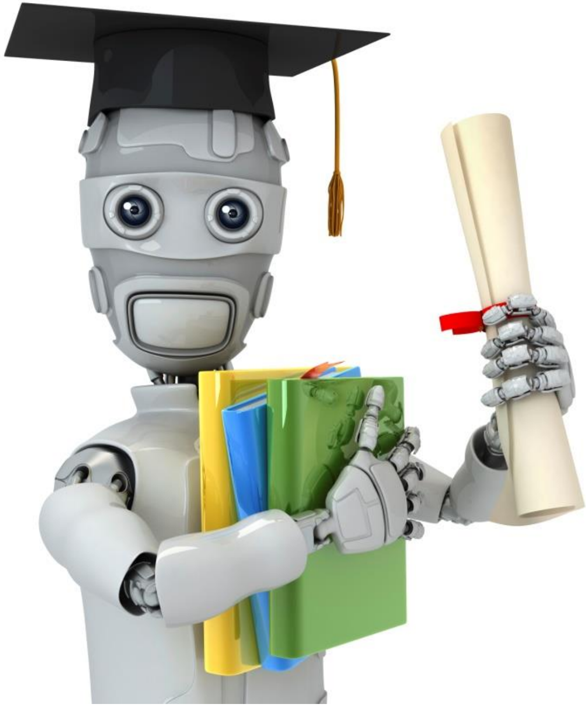
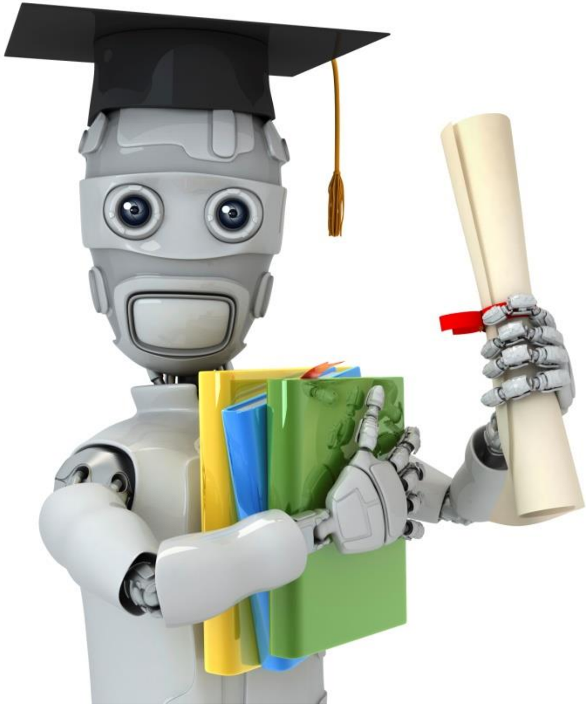


Linear regression

Fouad Hadj Selem

Machine Learning

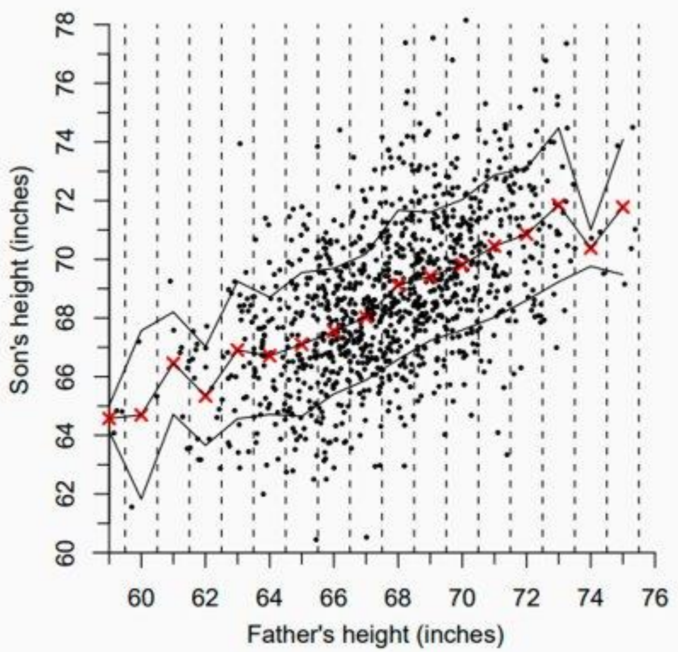


Machine Learning

Linear regression with one variable

Example 1: Pearson's Father-and-Son Data

Father-son pairs are grouped by father's height, to the nearest inch.

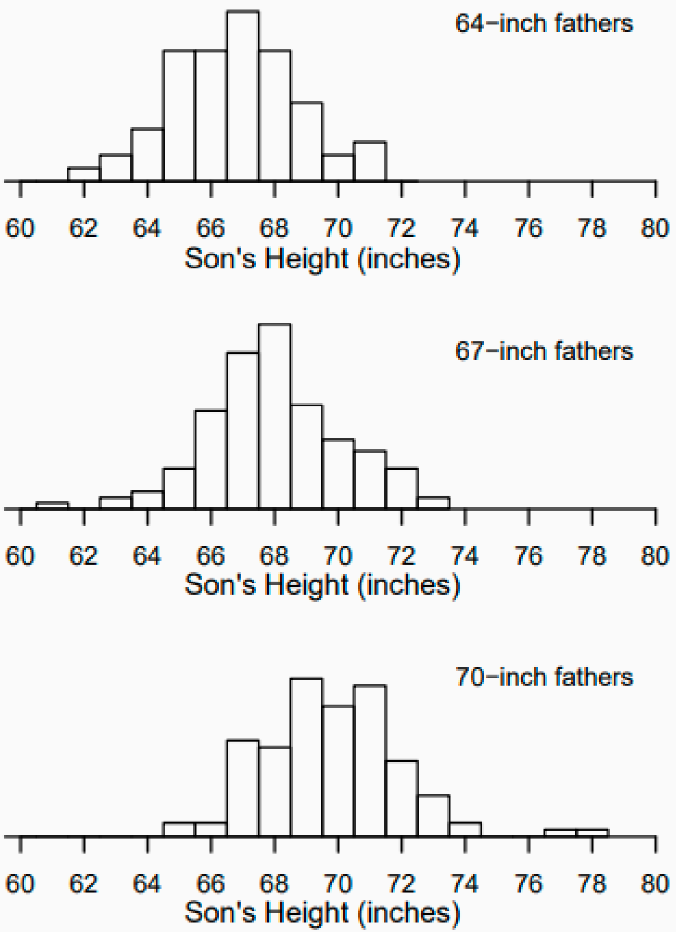


How do the

- mean of son's height (SH),
- SD of SH, and
- distribution of SH (histogram of SH)?

within each group change with father's height (FH)?

Example 1: Pearson's Father-and-Son Data



Simple Linear Regression Model

Pearson's father-and-son data inspire the following assumptions for the simple linear regression (SLR) model:

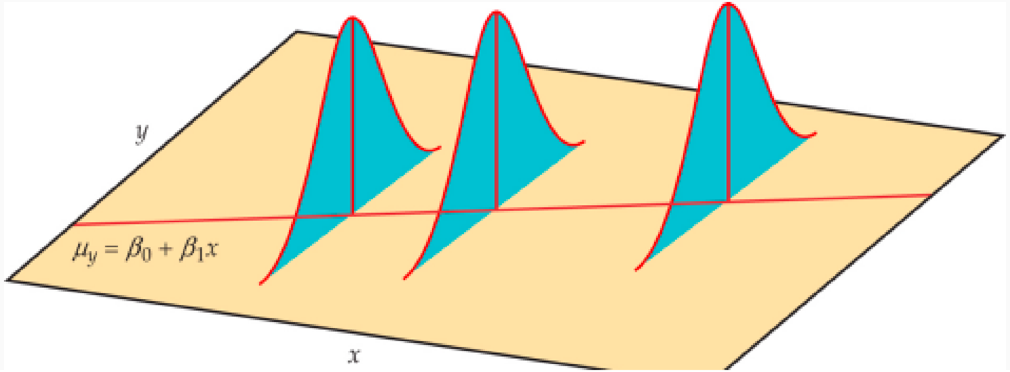
1. The means of Y is a linear function of X, i.e.,

$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

2. The SD of Y does not change with x, i.e.,

$$SD(Y \mid X = x) = \sigma \quad \text{for every } x$$

3. (Optional) Within each subpopulation, the distribution of Y is normal.



Simple Linear Regression Model

Equivalently, the SLR model asserts the values of X and Y for individuals in a population are related as follows

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

- the value of ε, called the **error** or the **noise**, varies from observation to observation, follows a normal distribution

$$\varepsilon \sim N(0, \sigma)$$

In the model, the line $y = \beta_0 + \beta_1 x$ is called the **population regression line.**

Data for a Simple Linear Regression Model

Suppose we have a SRS of n individuals from a population.

From individual i we observe the response y_i and the explanatory variable x_i:

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)$$

The SLR model states that

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Recall in the previous lecture, the least square line of the data above is

$$y = b_0 + b_1 x$$

in which

$$b_1 = r\frac{s_y}{s_x} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}, \quad b_0 = \overline{y} - b_1 \overline{x}$$

We can use b_1 to estimate β_1 and b_0 to estimate β_0.

Sample v.s. Population

Note the population regression line

$$y = \beta_0 + \beta_1 x$$

is different from the least square regression line

$$y = b_0 + b_1 x$$

- The latter is merely the least square line for a sample, while the former is the least square line for the entire population.
- The values of b_0 and b_1 will change from sample to sample.

$$b_1 = r\frac{s_y}{s_x} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}, \quad b_0 = \overline{y} - b_1\overline{x}$$

- We are interested in the population intercept β_0 and slope β_1, NOT the sample counterparts b_0 and b_1.

How Close Is b1 to β1?

Recall the slope of the least square line is

$$b_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

Under the SLR model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, replacing y_i in the formula above by $\beta_0 + \beta_1 x_i + \varepsilon_i$, we can show after some algebra that

$$b_1 = \beta_1 + \frac{\sum_i (x_i - \overline{x})\varepsilon_i}{\sum_i (x_i - \overline{x})^2}$$

From the above, one can get the mean, the SD, and the **sampling distribution** of b_1.

- $E(b_1) = \beta_1$ .... (b_1 is an **unbiased** estimate of β_1)
- $\text{SD}(b_1) = ?$................................(See the next slide)

Distribution of β1

The **sampling distribution** of b_1 is normal

$$b_1 \sim N\left(\beta_1, \frac{\sigma}{\sqrt{\sum(x_i - \overline{x})^2}}\right) \quad \Rightarrow \quad z = \frac{b_1 - \beta_1}{\sigma / \sqrt{\sum(x_i - \overline{x})^2}} \sim N(0,1)$$

This is (approx.) valid

- either if the errors ε_i are i.i.d. $N(0, \sigma)$
- or if the errors ε_i are independent and the sample size n is large

As σ is unknown, if replaced with s_e, the t-statistic below has a t-distribution with $n-2$ degrees of freedom

$$T = \frac{b_1 - \beta_1}{s_e / \sqrt{\sum(x_i - \overline{x})^2}} = \frac{b_1 - \beta_1}{SE(b_1)} \sim t_{n-2},$$

Confidence Intervals for β1

The $(1-\alpha)$ **confidence interval for** β_1 is given as

$$b_1 \pm t^* SE(b_1)$$

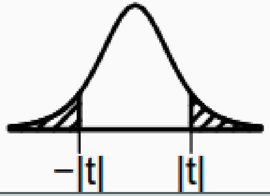
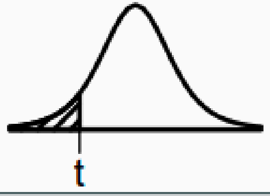
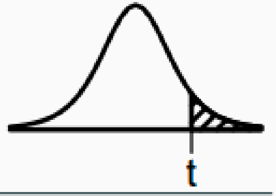
where t^* is the critical value for the $t_{(n-2)}$ distribution at confidence level $1-\alpha$.

Tests for β1

To **test the hypothesis** $H_0 : \beta_1 = a$, we use the t-statistic

$$t = \frac{b_1 - a}{SE(b_1)} \sim t_{n-2}$$

The p-value can be computed using the t-table based on the H_a:

H_a	$\beta_1 \neq a$	$\beta_1 < a$	$\beta_1 > a$
P-value			

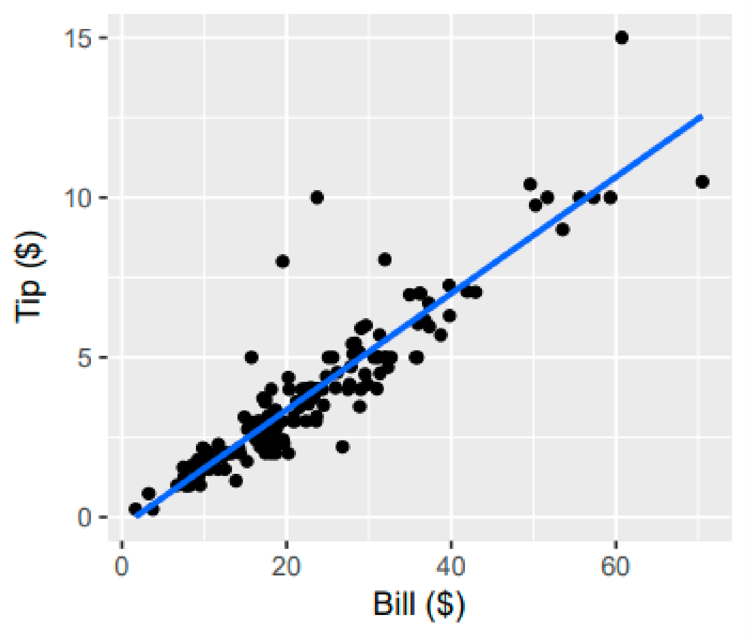
Observe that testing $H_0 : \beta_1 = 0$ is equivalent to testing whether x is useful in predicting y linearly.

- It is possible that r is small but β_1 is significantly different from 0.

Example: Restaurant Tips

The owner of a bistro called *First Crush* in Potsdam, NY, collected 157 restaurant bills over a 2-week period that he believes provide a good sample of his customers.

He wanted to study the payment and tipping patterns of its patrons.



Regression in R

Regression in R is as simple as `lm(y ~ x)`, in which "`lm`" stands for "`l`inear `m`odel"

```
> tips = read.table("RestaurantTips.txt",h=T)
> lm(Tip ~ Bill, data=tips)

Call:
lm(formula = Tip ~ Bill, data = tips)

Coefficients:
(Intercept)         Bill
    -0.2923       0.1822
```

It is better to save the model as an object,

```
lmtips = lm(Tip ~ Bill, data=tips)
```

and then we can get a more detailed output by viewing the `summary()` of the model object. The output is shown in the next slide

Regression in R

```
> summary(lmtips)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.292267   0.166160  -1.759   0.0806 .
Bill         0.182215   0.006451  28.247   <2e-16 ***
---
Residual standard error: 0.9795 on 155 degrees of freedom
Multiple R-squared:  0.8373,Adjusted R-squared:  0.8363
F-statistic: 797.9 on 1 and 155 DF,  p-value: < 2.2e-16
```

- The column "`Estimate`" gives the LS estimate for the intercept $b_0 = -0.292267$ and the slope $b_1 = 0.182215$
- The column "`Std. Error`" gives $\mathrm{SE}(b_0)$ and $\mathrm{SE}(b_1)$:

$$SE(b_0) = 0.166160, \quad SE(b_1) = 0.006451$$

Example: Test for the Slope β1

A general rule for waiters is to tip 15 to 20% of the pre-tax bill. That is, $\beta_0 = 0$ and β_1 is between 0.15 to 0.20.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.292267	0.166160	-1.759	0.0806
Bill	0.182215	0.006451	28.247	<2e-16

- R tests $\beta_0 = 0$ for us: t-statistic $= -1.759$, 2-sided p-value $= 0.0806$
- To test $H_0 : \beta_1 = 0.2$ v.s. $H_A : \beta_1 < 0.2$. The t-statistic is

$$t = \frac{b_1 - 0.2}{SE(b_1)} = \frac{0.182215 - 0.2}{0.006451} = -2.757$$

with df $= 155$, the one-sided p-value is < 0.005.

one tail	0.1	0.05	0.025	0.01	0.005
two tails	0.2	0.10	0.050	0.02	0.010
df 150	1.29	1.66	1.98	2.35	2.61
200	1.29	1.65	1.97	2.35	2.60

Conclusion: Customers of this restaurant gave less than 20% the bill as tips on average.

How to Read R Outputs for Regression?

```
Residual standard error: 0.9795 on 155 degrees of freedom
Multiple R-squared:  0.8373,Adjusted R-squared:  0.8363
F-statistic: 797.9 on 1 and 155 DF,  p-value: < 2.2e-16
```

- `Residual standard error: 0.9795 on 155 degrees of freedom`
  This gives the estimate s_e of σ, which is 0.9795.
  $\text{df} = n - 2 = 157 - 2 = 155$
- `Multiple R-squared: 0.8373` gives $r^2 = 0.8373$, Bill size explained 83.73% of the variation in tipping amount.
  The correlation between bill size and tips is
  $r = \sqrt{r^2} = \sqrt{0.8373} = 0.915$.
- `Adjusted R-squared`: Ignore this.
- `F-statistic: 797.9 on 1 and 155 DF,  p-value: < 2.2e-16` Skip.

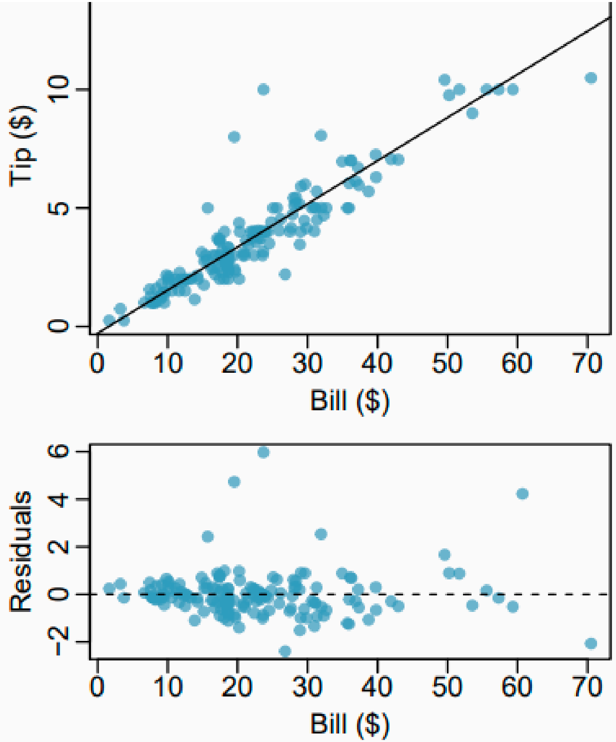
Checking Conditions for Simple Linear Regression Model?

1. Linearity
2. Constant variability
3. (Optional) Nearly normal residuals

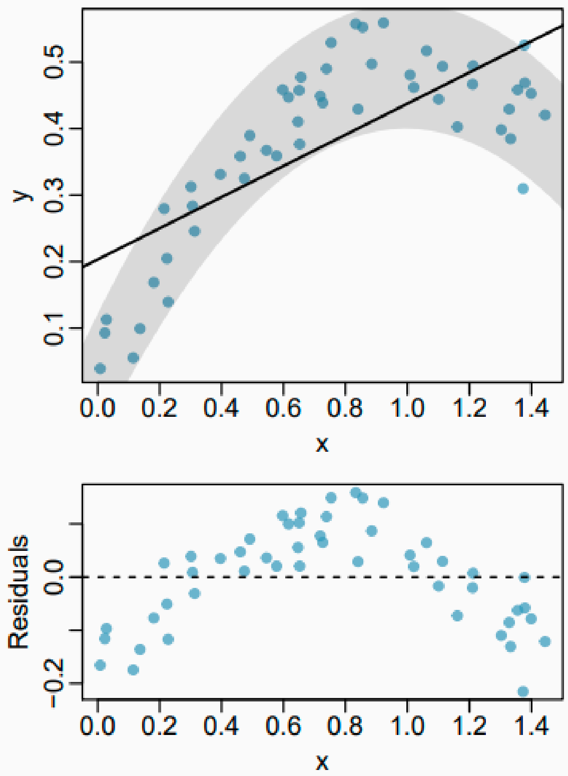
Tools for checking conditions:

- Residual plot

If conditions are satisfied, points should scatter evenly around the zero line in the residual plot.



Example



What condition is this linear model obviously violating?

(a) Constant variability

(b) Linear relationship

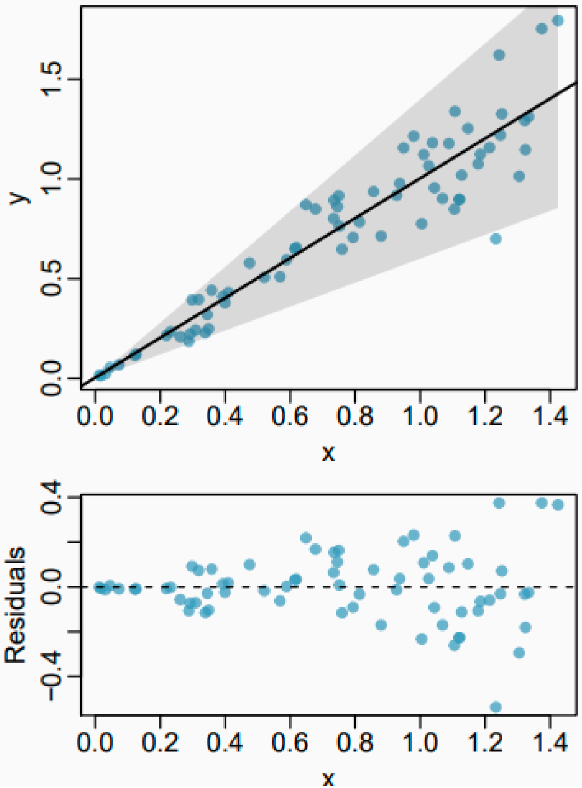
(c) *Linear relationship*

(d) Normal residuals

(e) No extreme outliers

Note the correlation between the residuals and x remains zero, but zero correlation $\neq$ no association. It can be a non-linear association

Checking Conditions : Constant Variability?

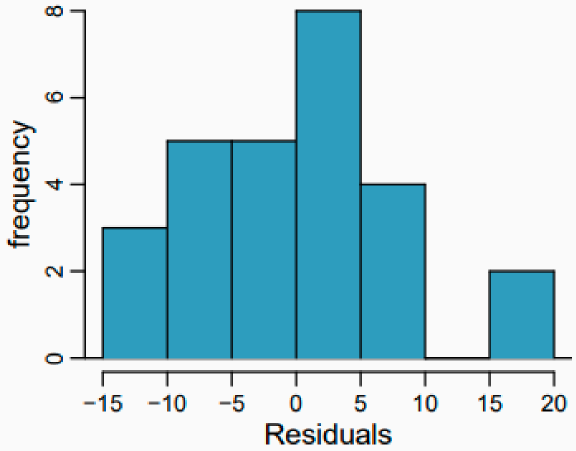


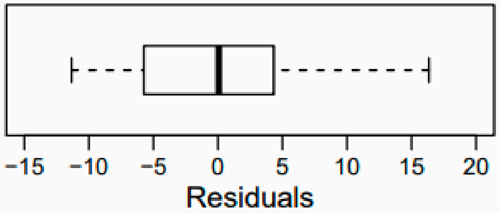
The variability of points around the least-squares line should be roughly constant, implying the variability of residuals around the 0 line should be roughly constant as well, called *homoscedasticity.*

If not, called *heterocedasticity*, predictions made in areas of larger variability will be worse. May try weighted least-square method or transforming the response.

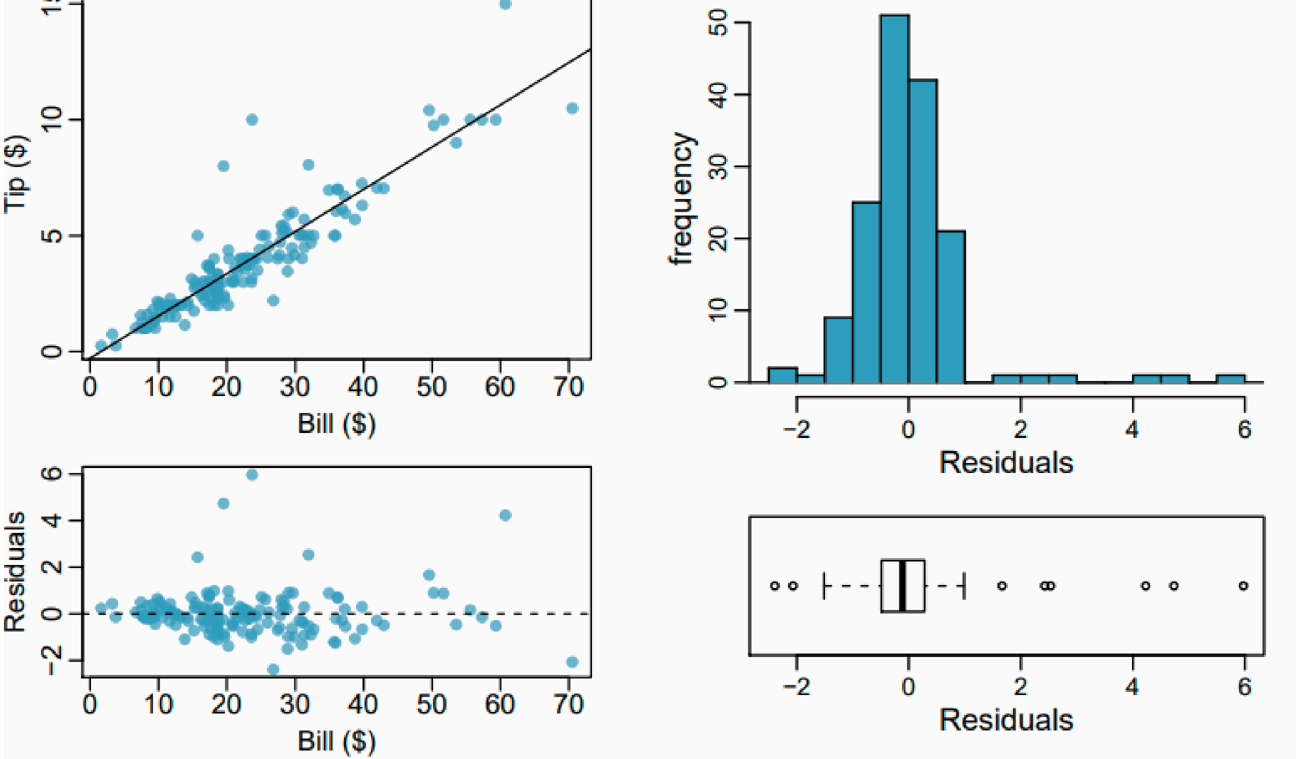
Conditions: Nearly Normal Residuals?

- Less relevant than the first two conditions
- Diagnosis: Check the histogram or boxplot of residuals
- If the linearity or constant variability condition is clearly violated, there is no need to check the normality of residuals.

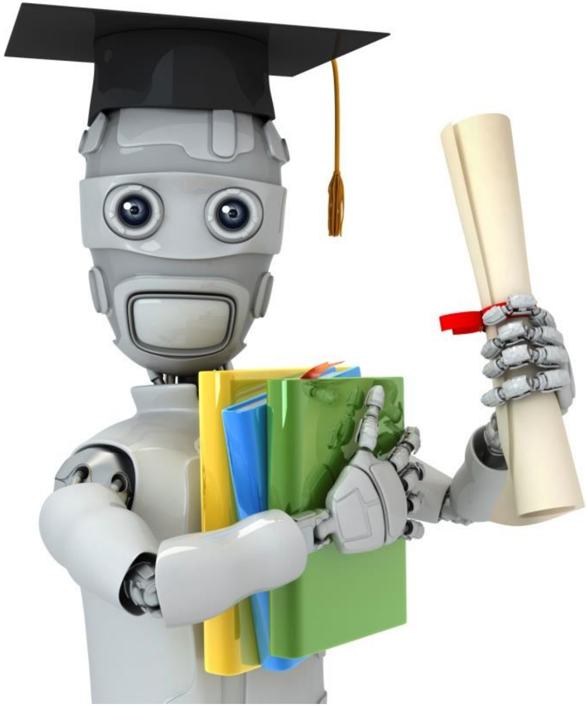




Checking Conditions for the Restaurant Tip Data



The constant variability condition seems to be violated.
The size of residual seems to increase with Bill.



Linear regression with multiple variables

Machine Learning

Multiple regression

- Simple linear regression: Bivariate - two variables: y and x
- Multiple linear regression: Multiple variables: y and $x_1, x_2, \cdots$

- Suppose we have n independent observations

$$(\boldsymbol{x}_1^\top, Y_1), \ldots, (\boldsymbol{x}_n^\top, Y_n),$$

where $\boldsymbol{x}_i$ is a $(p-1)$-vector of known (explanatory) values.

- A natural extension of simple linear regression is to consider the model with more than one predictor variables

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{i,p-1} + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\epsilon_i \sim NID(0, \sigma^2)$ and p is the number of regression parameters to estimate.

- Equivalently, we can write in matrix notation

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & \boldsymbol{x}_1^\top \\ 1 & \boldsymbol{x}_2^\top \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_n^\top \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Regression parameter estimation?

Still minimise residual sum of squares to get:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{Y}$$

if $(\mathbf{X}^\top \mathbf{X})^{-1}$ exists.

Assumptions

- Assumptions of the linear model

$$E(Y \mid X) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$
$$Var(Y \mid X) = \sigma^2$$

1. Errors $\varepsilon_i \sim N(0, \sigma^2)$.
2. Error variances are equal.
3. Errors are independent.
4. Y has a linear dependence on X.

T-test

Hypothesis concerning one of the terms

$$\boxed{\begin{aligned} H_0 &: \beta_i = b_i \\ H_1 &: \beta_i \neq b_i \end{aligned}}$$

t-test statistic: $t = \dfrac{\hat{\beta}_i - b_i}{\sqrt{c_{ii} MSE}} \sim t(n-p-1)$

If H0 is true, $t \sim t(n-p-1)$,
so we reject H0 at level α if $|t| > 2t_{\alpha/2}(n-p-1)$

The confidence interval for β_i is $b_i \pm t_{\alpha/2}\sqrt{c_{ii} MSE}$

```
> summary(lm(Fuel~.,data=fueldata))
Call:
lm(formula = Fuel ~ ., data = fueldata)

Residuals:
     Min       1Q   Median       3Q      Max
-163.145  -33.039    5.895   31.989  183.499

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  154.1928   194.9062    0.791  0.432938
Tax           -4.2280     2.0301   -2.083  0.042873 *
Dlic           0.4719     0.1285    3.672  0.000626 ***
Income        -6.1353     2.1936   -2.797  0.007508 **
logMiles      18.5453     6.4722    2.865  0.006259 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-test

Hypothesis test for the reduced model

$$\boxed{\begin{aligned} &H_0 : \beta_{r+1} = \cdots \beta_p = 0 \\ &H_1 : not\ H_0 \end{aligned}}$$

F test statistic: $F = (\dfrac{SSE_R - SSE_F}{p-r})/(\dfrac{SSE_F}{n-p-1})$

If H0 is true, $F \sim F(p-r, n-p-1)$
so we reject H0 at level α if $F > F_\alpha(p-r, n-p-1)$

From this test, we conclude that the hypotheses are plausible or not. And we say that which model is adequate.

```
> summary(lm(Fuel~.,data=fueldata))
Call:
lm(formula = Fuel ~ ., data = fueldata)

Residuals:
    Min       1Q  Median      3Q     Max
-163.145  -33.039   5.895  31.989  183.499

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  154.1928  194.9062    0.791  0.432938
Tax           -4.2280    2.0301   -2.083  0.042873 *
Dlic           0.4719    0.1285    3.672  0.000626 ***
Income        -6.1353    2.1936   -2.797  0.007508 **
logMiles      18.5453    6.4722    2.865  0.006259 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 64.89 on 46 degrees of freedom
Multiple R-squared: 0.5105,    Adjusted R-squared: 0.4679
F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.33e-07
```

checking model conditions using graphs

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

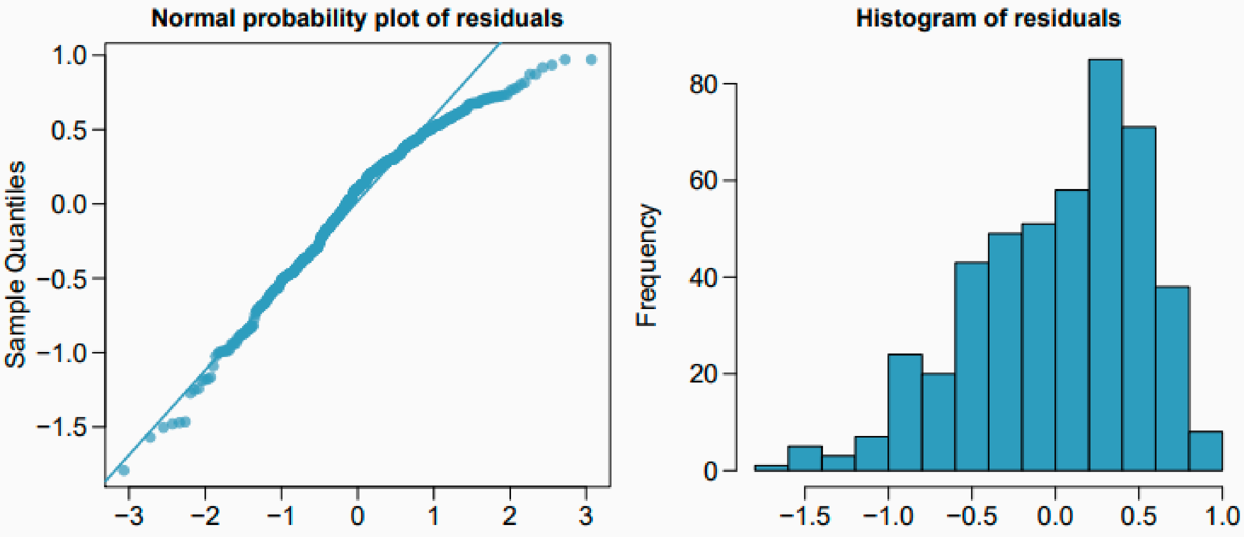
The model depends on the following conditions

1. residuals are nearly normal (primary concern relates to residuals that are outliers)
2. residuals have constant variability
3. residuals are independent
4. each variable is linearly related to the outcome

We often use graphical methods to check the validity of these conditions, which we will go through in detail in the following slides.

Nearly Normal Residuals ?

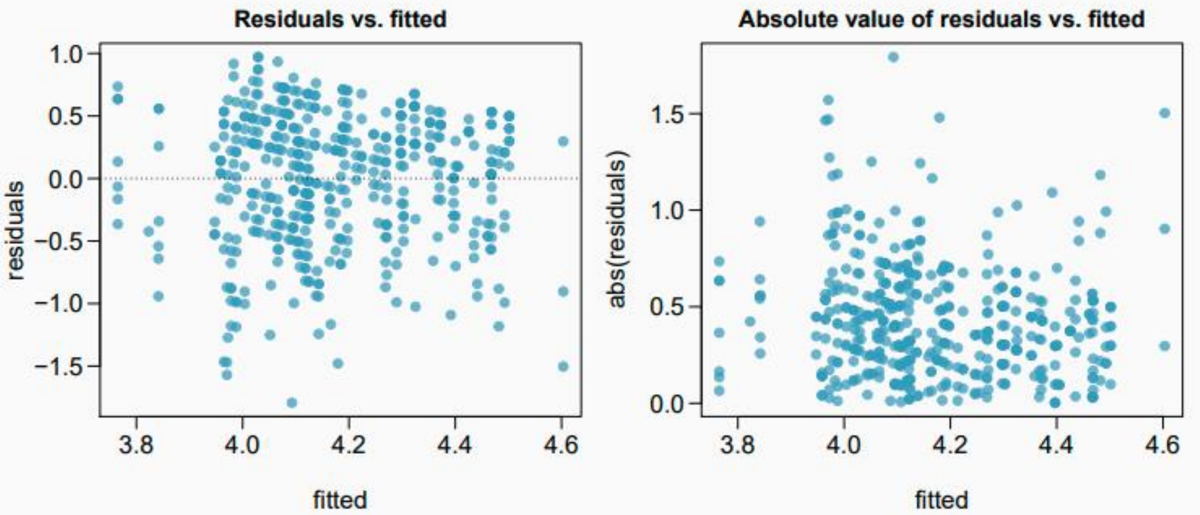
normal probability plot and/or histogram of residuals:



Does this condition appear to be satisfied?

Checking Constant Variance

scatterplot of residuals and/or absolute value of residuals vs. fitted (predicted):



Checking Constant Variance

- When we did simple linear regression (one explanatory variable) we checked the constant variance condition using a plot of *residuals vs. x*.
- With multiple linear regression (2+ explanatory variables) we checked the constant variance condition using a plot of *residuals vs. fitted*.

Why are we using different plots?

In multiple linear regression there are many explanatory variables, so a plot of residuals vs. one of them wouldn't give us the complete picture.

Residuals are Independent?

- Checking for independent residuals allows us to indirectly check for independent observations.
- If observations and residuals are independent, we would not expect to see an increasing or decreasing trend in the scatterplot of residuals vs. order of data collection.
- This condition is often violated when we have time series data. Such data require more advanced time series regression techniques for proper analysis.

scatterplot of residuals vs. order of data collection:

Coefficient of Determination

- Fact : $\hat{\sigma}^2 = \dfrac{SSE}{n-(p+1)} (= MSE)$ is an unbiased estimator of σ^2.
- If e is normally distributed, $\dfrac{SSE}{\sigma^2} = \dfrac{(n-(p+1))\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-(p+1))$
- Define SSreg=SYY-SSE (SYY= the sum of squares of Y)

  As with the simple regression, the coefficient of determination is

  $$R^2 = \frac{SS_{reg}}{SYY} = 1 - \frac{SSE}{SYY}$$

  It is also called the multiple correlation coefficient because it is the maximum of the correlation between Y and any linear combination of the terms in the mean function.

Coefficient of Determination

R^2 can be calculated in three ways:

1. square the correlation coefficient of x and y (how we have been calculating it)
2. square the correlation coefficient of y and $\hat{y}$
3. based on definition:

$$R^2 = \frac{\text{explained variability in } y}{\text{total variability in } y}$$

For single-predictor linear regression, having three ways to calculate the same value may seem like overkill.
However, in multiple linear regression, we can't calculate R^2 as the square of the correlation between x and y because we have multiple xs.
And next we'll learn another measure of explained variability, adjusted R^2, that requires the use of the third approach, ratio of explained and unexplained variability.

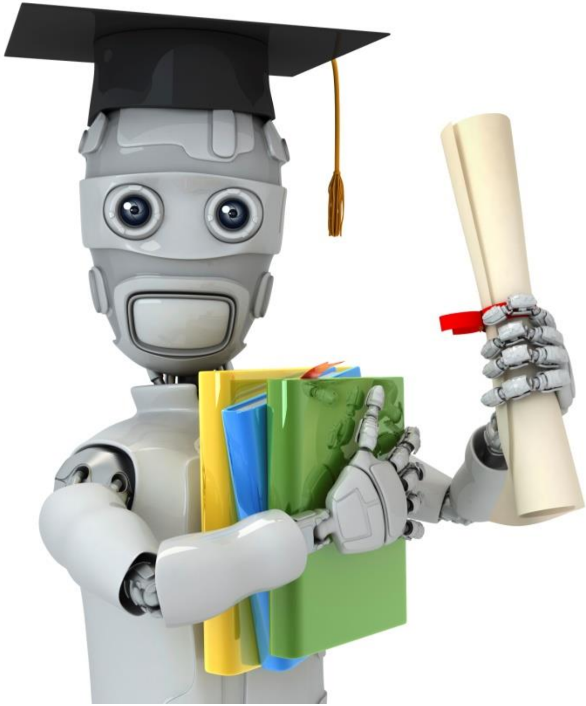
Adjusted Coefficient

$$R^2_{adj} = 1 - \left(\frac{SS_{Error}}{SS_{Total}} \times \frac{n-1}{n-p-1} \right)$$

where n is the number of cases and p is the number of predictors (explanatory variables) in the model.

- Because p is never negative, R^2_{adj} will always be smaller than R^2.
- R^2_{adj} applies a penalty for the number of predictors included in the model.
- Therefore, we choose models with higher R^2_{adj} over others.

- When any variable is added to the model R^2 increases.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adjusted R^2 does not increase.



Machine Learning

Model selection

Backward-Elimination

1. R^2_{adj} approach:
   - Start with the full model
   - Drop one variable at a time and record R^2_{adj} of each smaller model
   - Pick the model with the highest increase in R^2_{adj}
   - Repeat until none of the models yield an increase in R^2_{adj}
2. p-value approach:
   - Start with the full model
   - Drop the variable with the highest p-value and refit a smaller model
   - Repeat until all variables left in the model are significant

Example Backward-Elimination: R2 Approach

Step	Variables included & p-value								
Full	beauty 0.00	gender male 0.00	age 0.01	formal yes 0.04	lower yes 0.29	native non english 0.06	minority yes *0.35*	students 0.30	tenure tenure track 0.02	tenur tenure 0.0
Step 1	beauty 0.00	gender male 0.00	age 0.01	formal yes 0.04	lower yes *0.38*	native non english 0.03		students 0.34	tenure tenure track 0.02	tenur tenure 0.0
Step 2	beauty 0.00	gender male 0.00	age 0.01	formal yes 0.05		native non english 0.02		students *0.44*	tenure tenure track 0.01	tenur tenure 0.0
Step 3	beauty 0.00	gender male 0.00	age 0.01	formal yes *0.06*		native non english 0.02			tenure tenure track 0.01	tenur tenure 0.0
Step 4	beauty 0.00	gender male 0.00	age 0.01			native non english *0.06*			tenure tenure track 0.01	tenur tenure 0.0
Step 5	beauty 0.00	gender male 0.00	age 0.01						tenure tenure track 0.01	tenur tenure 0.0

Best model: beauty + gender + age + tenure

Example Backward-Elimination: p_value Approach

Step	Variables included	R^2_{adj}
Full	beauty + gender + age + formal + lower + native + minority + students + tenure	*0.0839*
Step 1	gender + age + formal + lower + native + minority + students + tenure	0.0642
	beauty + age + formal + lower + native + minority + students + tenure	0.0557
	beauty + gender + formal + lower + native + minority + students + tenure	0.0706
	beauty + gender + age + lower + native + minority + students + tenure	0.0777
	beauty + gender + age + formal + native + minority + students + tenure	0.0837
	beauty + gender + age + formal + lower + minority + students + tenure	0.0788
	beauty + gender + age + formal + lower + native + students + tenure	*0.0842*
	beauty + gender + age + formal + lower + native + minority + tenure	0.0838
	beauty + gender + age + formal + lower + native + minority + students	0.0733
Step 2	gender + age + formal + lower + native + students + tenure	0.0647
	beauty + age + formal + lower + native + students + tenure	0.0543
	beauty + gender + formal + lower + native + students + tenure	0.0708
	beauty + gender + age + lower + native + students + tenure	0.0776
	beauty + gender + age + formal + native + students + tenure	*0.0846*
	beauty + gender + age + formal + lower + native + tenure	0.0844
	beauty + gender + age + formal + lower + native + students	0.0725
Step 3	gender + age + formal + native + students + tenure	0.0653
	beauty + age + formal + native + students + tenure	0.0534
	beauty + gender + formal + native + students + tenure	0.0707
	beauty + gender + age + native + students + tenure	0.0786
	beauty + gender + age + formal + students + tenure	0.0756
	beauty + gender + age + formal + native + tenure	*0.0855*
	beauty + gender + age + formal + native + students	0.0713
Step 4	gender + age + formal + native + tenure	0.0667
	beauty + age + formal + native + tenure	0.0553
	beauty + gender + formal + native + tenure	0.0723
	beauty + gender + age + native + tenure	0.0806

Best model: beauty + gender + age + formal + native + tenure

Forward-selection

1. R^2_{adj} approach:
   - Start with regressions of response vs. each explanatory variable
   - Pick the model with the highest R^2_{adj}
   - Add the remaining variables one at a time to the existing model, and once again pick the model with the highest R^2_{adj}
   - Repeat until the addition of any of the remanning variables does not result in a higher R^2_{adj}
2. $p - value$ approach:
   - Start with regressions of response vs. each explanatory variable
   - Pick the variable with the lowest significant p-value
   - Add the remaining variables one at a time to the existing model, and pick the variable with the lowest significant p-value
   - Repeat until any of the remaining variables does not have a significant p-value

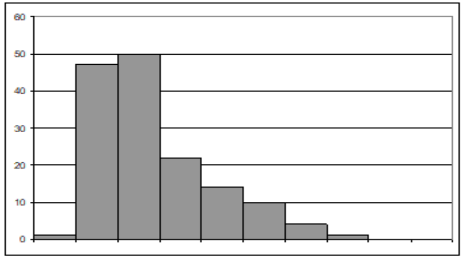
Moving Beyond Linearity (in the features)

Common intrinsically linear models and required transformations

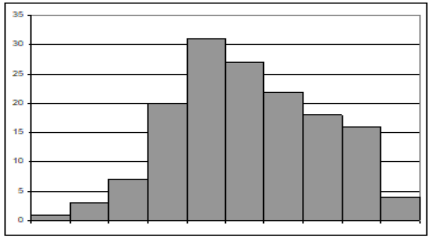
True relationship	Transformation	Linearized model
$y = \beta_0 x^{\beta_1}$	$y' = \log y, x' = \log x$	$y' = \log \beta_0 + \beta_1 x'$
$y = \beta_0 e^{\beta_1 x}$	$y' = \log y$	$y' = \log \beta_0 + \beta_1 x$
$y = \beta_0 + \beta_1 \log x$	$x' = \log x$	$y = \log \beta_0 + \beta_1 x'$
$y = \frac{x}{\beta_0 x - \beta_1}$	$y' = \frac{1}{y}, x' = \frac{1}{x}$	$y' = \beta_0 - \beta_1 x'$

The effect of the ln transformation

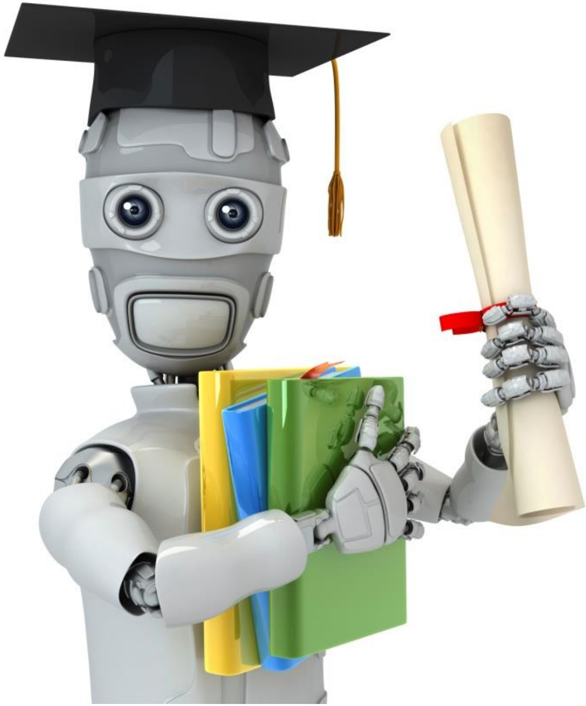
- It spreads out values that are close to zero
- Compacts values that are large



x



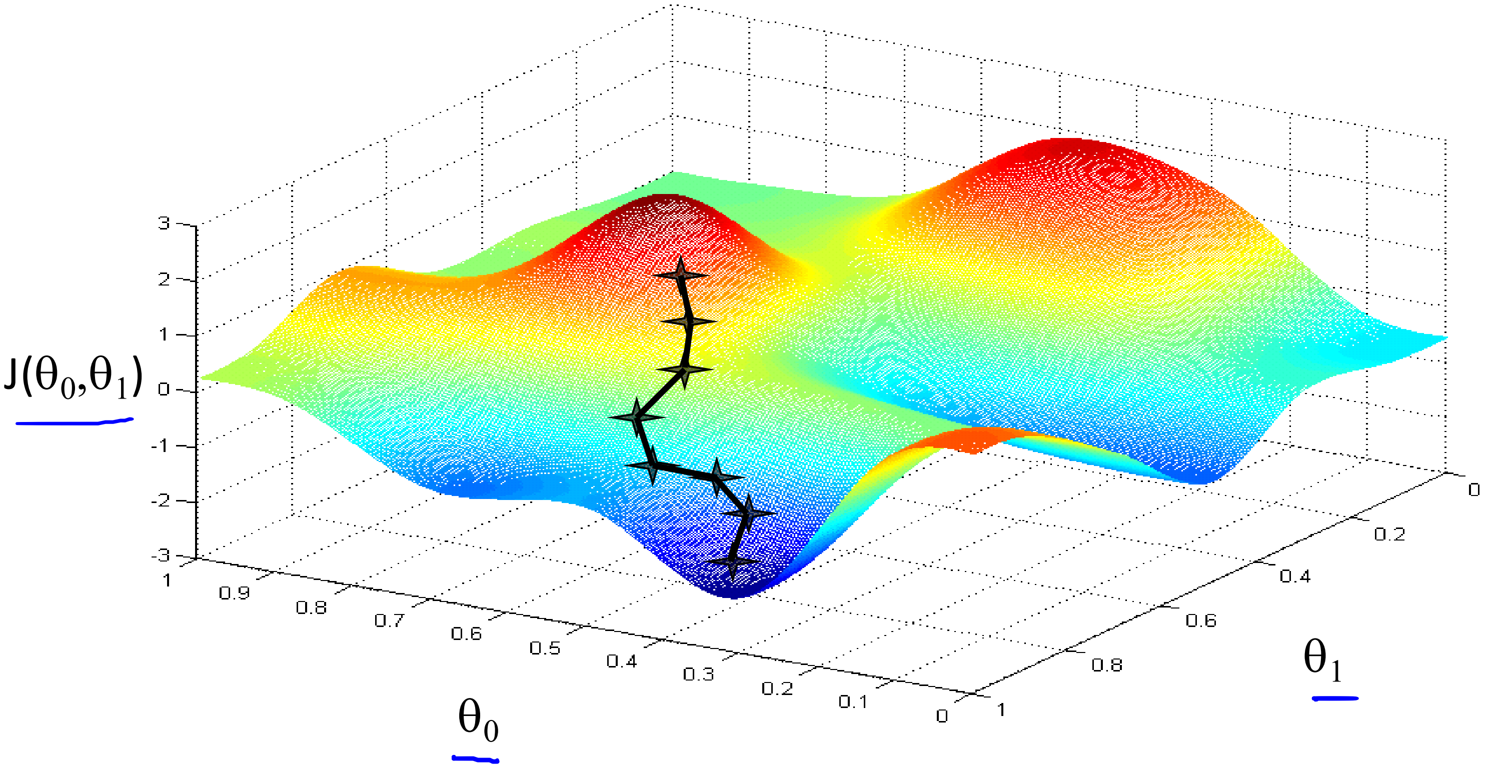
$x^{(new)} = \ln(x)$



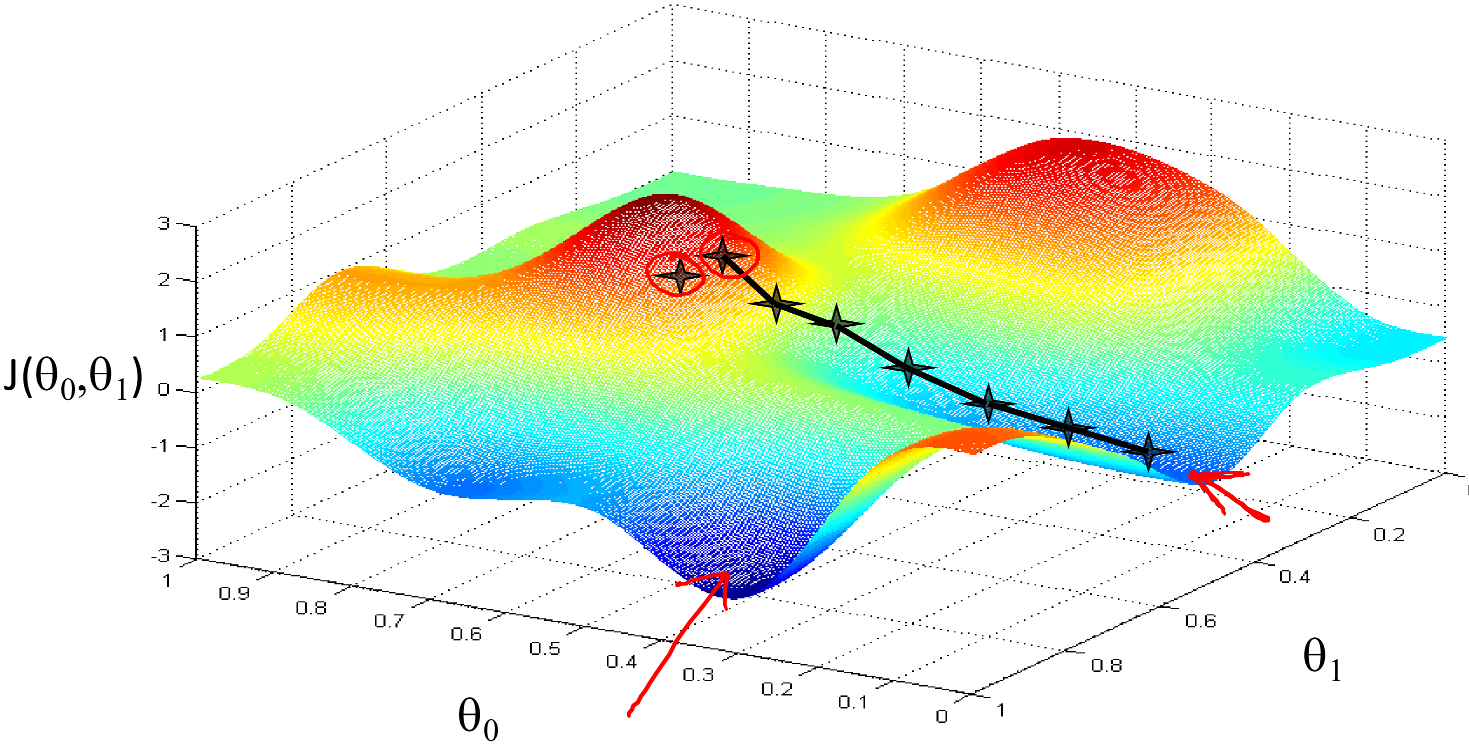
Machine Learning

See you next chapter

Thanks :)



$J(\theta_0,\theta_1)$
3
2
1
0
-1
-2
-3
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
θ_0
0
0.2
0.4
0.6
0.8
1
θ_1



$J(\theta_0,\theta_1)$
3
2
1
0
-1
-2
-3
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
θ_0
0
0.2
0.4
0.6
0.8
1
θ_1

Gradient descent algorithm

$$\text{repeat until convergence } \{$$
$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \qquad (\text{for } j = 0 \text{ and } j = 1)$$
$$\}$$

Correct: Simultaneous update

$$\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$
$$\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$
$$\theta_0 := \text{temp0}$$
$$\theta_1 := \text{temp1}$$

Incorrect:

$$\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$
$$\theta_0 := \text{temp0}$$
$$\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$
$$\theta_1 := \text{temp1}$$