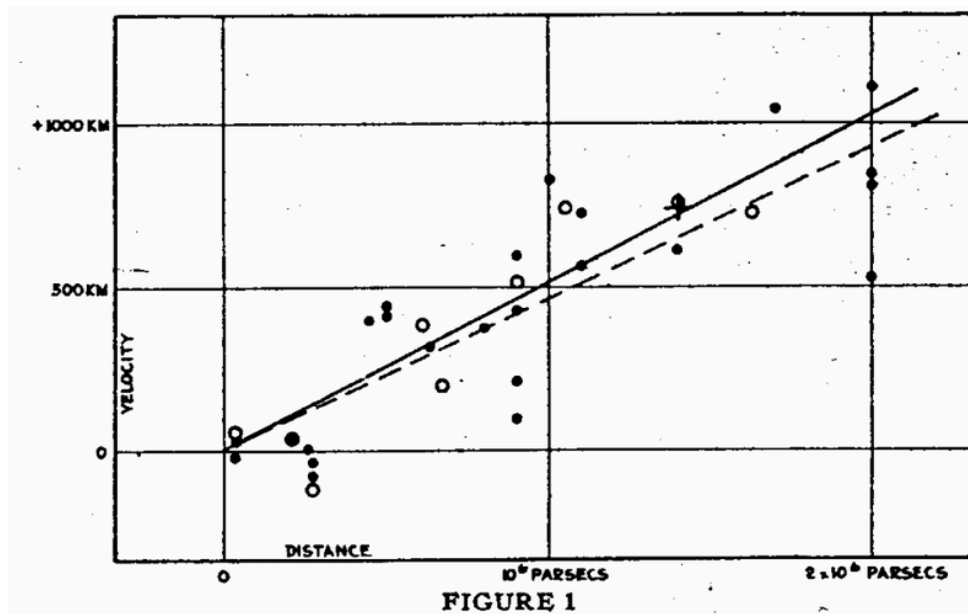


COMP47460 Tutorial

Regression and Gradient Descent

Q1.

In 1929 Edwin Hubble published a paper which revolutionised astronomy. It is the basis of Hubble's law which was the first observational evidence for the expansion of the universe. Hubble was able to measure the recession velocity (km/sec) (how fast it is moving away from us) and the distance from the earth in megaparsecs of various galaxies (1 megaparsec is about 3×10^{22} metres, a long way!)



- Just looking at the graph from Hubble's original paper, would you have confidence in his conclusion that there is a linear relationship between the speed of galaxy and its distance from earth?
- Using OLS find the best fit linear model of the data (β_0 , β_1). (do this by hand)
- Compute the correlation coefficient. (by hand)
- Using a two-tailed t-test, determine if the relationship between distance and velocity is significant for a p-value of 0.05.
- The Andromeda Galaxy is the closest spiral galaxy to us at 0.613 megaparsecs. What is its recession velocity?
- The slope β_1 is known as the Hubble constant H_0 . The latest measurements of the Hubble Space Telescope determine it to be

73.00±1.75 (km/s/megaparsec). How close was Hubble with his original data?

- g) Using Weka, run the LinearRegression model and confirm your results.

hubble_constant.csv

```
cov_xy: 189.231820652
sd(x), sd(y): 0.645495752352 371.254666198
<x>, <y>: 0.911375 373.125
beta_1 454.158440923
beta_0 -40.7836490959
corr(x,y) 0.789639487935
r_sq 0.623530520907
```

OLS Regression Results

Dep. Variable:	recession_velocity (km/s)		R-squared:	0.624		
Model:	OLS		Adj. R-squared:	0.606		
Method:	Least Squares		F-statistic:	36.44		
Date:	Thu, 02 Nov 2017		Prob (F-statistic):	4.48E-06		
Time:	07:20:36		Log-Likelihood:	-163.83		
No. Observations:	24		AIC:	331.7		
Df Residuals:	22		BIC:	334.0		
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t 	[0.025	0.975]
const	-40.7836	83.439	-0.489	0.630	-213.825	132.258
distance (megaparsec)	454.1584	75.237	6.036	0.000	298.126	610.191

Omnibus:	0.126	Durbin-Watson:	2.089
Prob(Omnibus):	0.939	Jarque-Bera (JB):	0.293
Skew:	0.138	Prob(JB):	0.864
Kurtosis:	2.535	Cond. No.	3.22

Linear Regression with Weka:

The image displays two screenshots of the Weka Explorer interface, illustrating the process of running a Linear Regression model.

Top Screenshot: Preprocessing and Attribute Selection

- Left Panel (Classifiers):** Shows the 'functions' folder expanded, with 'LinearRegression' selected.
- Main Panel:**
 - Current relation:** 'hubble_constant', Attributes: 2, Sum of weights: 24, Instances: 24.
 - Attributes:** A list of attributes is shown, with 'distance (megaparsec)' and 'recession_velocity (km/s)' selected.
 - Selected attribute:** 'distance (megaparsec)', Type: Numeric, Missing: 0 (0%), Distinct: 15, Unique: 10 (42%).
 - Statistics:** Minimum: 0.032, Maximum: 2, Mean: 0.911, StdDev: 0.645.
 - Class:** 'recession_velocity (km/s)...'
 - Visualize:** A bar chart is displayed, showing the distribution of the selected attribute.

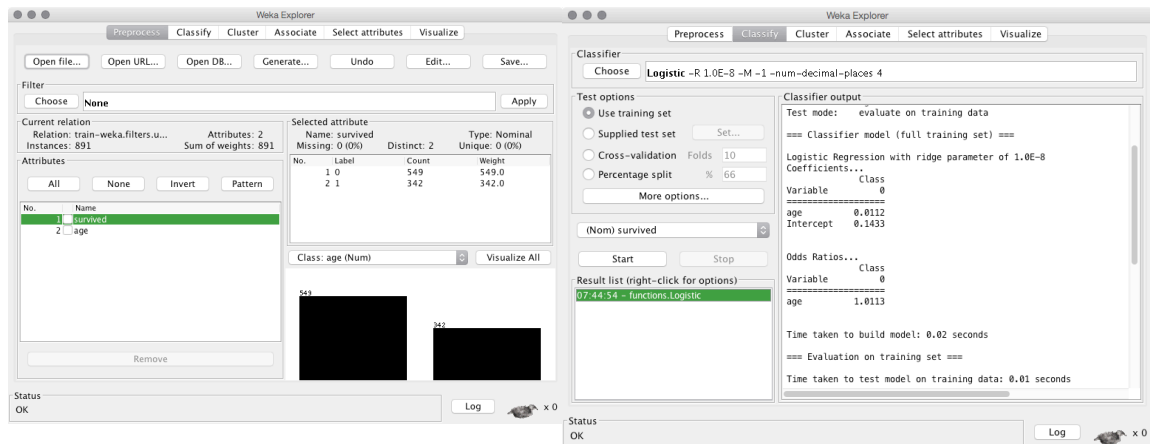
Bottom Screenshot: Classification Results

- Classifier:** 'LinearRegression -S 1 -C -R 1.0E-8 -additional-stats -num-decimal-places 4'.
- Test options:** 'Use training set' is selected. Other options include 'Supplied test set', 'Cross-validation' (Folds: 10), and 'Percentage split' (%: 66).
- Classifier output:**
 - Variable Coefficient SE of Coef t-Stat:**

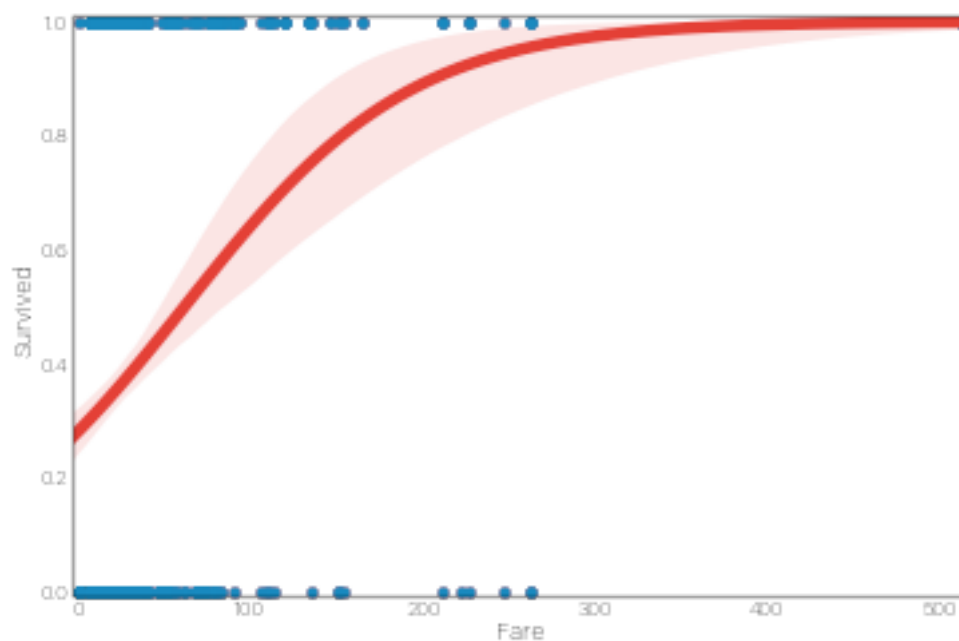
Variable	Coefficient	SE of Coef	t-Stat
distance (megaparsec)	454.1584	75.2371	6.0364
const	-40.7836	83.4389	-0.4888
 - Degrees of freedom = 22**
 - R² value = 0.6235**
 - Adjusted R² = 0.60642**
 - F-statistic = 36.4377**
 - Time taken to build model: 0.06 seconds**
 - === Evaluation on training set ===**
 - Time taken to test model on training data: 0 seconds**
 - === Summary ===**
 - Correlation coefficient: 0.7896
 - Mean absolute error: 179.0426
 - Root mean squared error: 222.995
 - Relative absolute error: 59.423 %
 - Root relative squared error: 61.3571 %
 - Total Number of Instances: 24

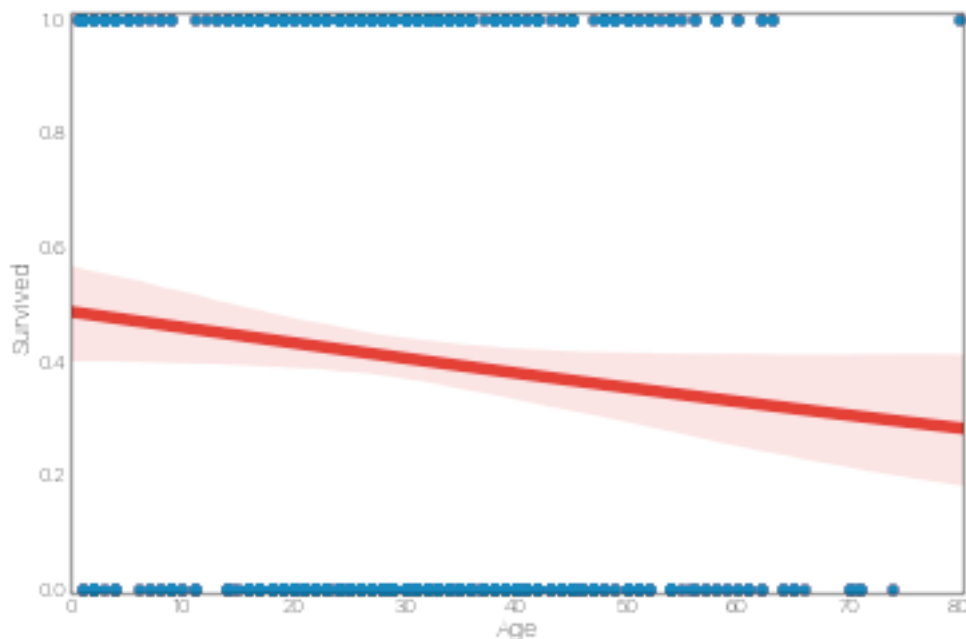
Q2: Titanic:

This is a Weka exercise to build and interpret a logistic regression model:



There doesn't appear to be a strong relationship between survived and age, or fare.





There is not a strong relationship either between Age and Survival.

The odds ratio for *Age* is: 1.0175 which implies a positive relationship between age and survived.

The odds ratio for *Fare*: 0.9839 which implies a negative relationship between Fare and Survived.

Q3.

The OLS model is:

$$\text{recession_velocity (km/s)} = 454.1584 \text{ distance (megaparsec)} - 40.7836$$

After 100000 iterations with $\alpha=0.01$ the model is:

$$\text{recession_velocity (km/s)} = 456.8329 \text{ distance (megaparsec)} - 34.9857$$

which is quite far from the OLS model, and doesn't seem to improve much with more iterations.

Setting $\alpha=1e-4$ and $\text{epochs}=1e6$ we find:

$$\text{recession_velocity (km/s)} = 454.178 \text{ distance (megaparsec)} - 40.7225$$

Could implement a table to show the cost function as we increase the number of iterations.