Introductory Statistics Lectures

Linear regression

How to mathematically model a linear relationship and make predictions.

Anthony Tanbakuchi
Department of Mathematics
Pima Community College

(Compile date: Tue May 19 14:51:27 2009)

Contents

1 Linear regression

1.1 Introduction

Motivation

Example 1. In 1929 Edwin Hubble published a paper[1] with data on 24 galaxies. He had recorded the distance and velocity of each galaxy relative to earth. The data is shown below with velocities in meters/year and distances in meters:

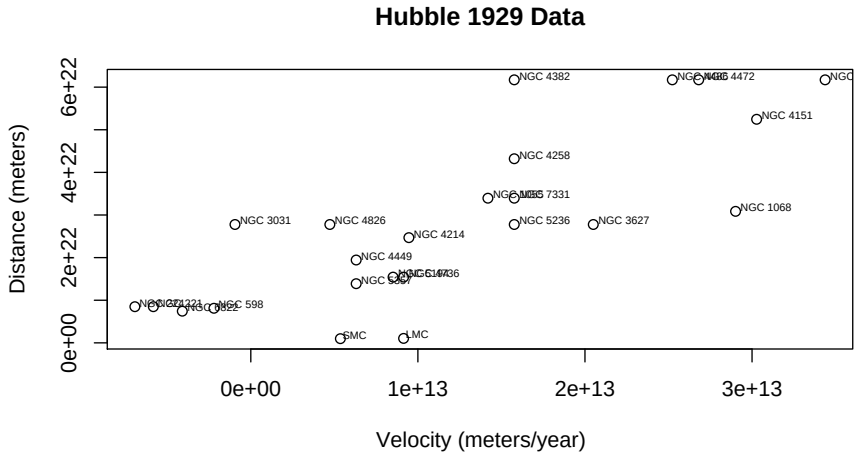
	Galaxy	Vel.m.year	Dist.m
1	SMC	5.36E+12	9.88E+20
2	LMC	9.15E+12	1.05E+21
3	NGC 6822	-4.10E+12	7.44E+21
4	NGC 598	-2.21E+12	8.12E+21
5	NGC 221	-5.83E+12	8.49E+21
6	NGC 224	-6.94E+12	8.49E+21
7	NGC 5357	6.31E+12	1.39E+22
8	NGC 4736	9.15E+12	1.54E+22
9	NGC 5194	8.51E+12	1.54E+22
10	NGC 4449	6.31E+12	1.94E+22
11	NGC 4214	9.46E+12	2.47E+22
12	NGC 3031	-9.46E+11	2.78E+22
13	NGC 3627	2.05E+13	2.78E+22
14	NGC 4826	4.73E+12	2.78E+22
15	NGC 5236	1.58E+13	2.78E+22
16	NGC 1068	2.90E+13	3.09E+22
17	NGC 1055	1.42E+13	3.39E+22
18	NGC 7331	1.58E+13	3.39E+22
19	NGC 4258	1.58E+13	4.32E+22
20	NGC 4151	3.03E+13	5.25E+22
21	NGC 4382	1.58E+13	6.17E+22
22	NGC 4472	2.68E+13	6.17E+22
23	NGC 4486	2.52E+13	6.17E+22
24	NGC 4649	3.44E+13	6.17E+22

Recall that:

value	name
1×10^3	thousand
1×10^6	million
1×10^9	billion
1×10^{12}	trillion
1×10^{15}	quadrillion
1×10^{18}	quintillion
1×10^{21}	sextillion

and 1 light year is 9.46×10^{15} m.

[1] You can read his original paper at http://www.pnas.org/misc/Hubble.pdf.



Question 1. Does there appear to be a relationship of a galaxy's distance and velocity?

Question 2. If we discovered a new galaxy moving 2.2×10^{13} m/year, could we predict it's distance from earth?

- Hubble inferred that the true relationship must be linear. If he knew the actual linear relationship, the **slope** of the line would yield the age of the universe!
- The true line is the **population parameter** that Hubble needed to estimate.
- The 24 galaxies were the **sample data** that he had collected.
- Thus, statistics should be able to **estimate** the true linear relationship.

DETERMINISTIC MODEL. DEFINITION 1.1

A model that can **exactly** predict the value of a variable. (algebra)

Example: The area of a circle can be determined exactly from it's radius: $A = \pi r^2$.

PROBABILISTIC MODEL. DEFINITION 1.2

A model where one variable an be used to **approximate** the value of another variable. More specifically, one variable is not completely determined by the other variable.

Example: Forearm length of an individual can be used to estimate the approximate height of an individual but not an exact height.

LINEAR MODELS

Equation of a line: algebra
Recall

$$y = mx + b \tag{1}$$

- x is the **independent** variable.
- y is the **dependent variable**. (Since y depends on x.)
- b is the y-intercept.
- m is the slope

Equation of a line: statistics
We will estimate the linear probabilistic model as

$$\boxed{y = \beta_0 + \beta_1 x + \epsilon} \tag{2}$$

using the best fit line

$$\hat{y} = b_0 + b_1 x$$

- x is the **predictor variable**.
- $\hat{y}$ is the **response variable**.
- b_0 is the y-intercept
- b_1 is the slope

b_0 and b_1 are sample statistics that we use to estimate the population parameters β_0 and β_1. The term ϵ represents the random error in the model.

Assumptions about random error ϵ

1. ϵ are independent in the probabilistic sense.
2. $\mu_\epsilon = 0$, common variance σ_ϵ^2.
3. ϵ are normally distributed.

DEFINITION 1.3 RESIDUAL e.

The residual is the "error" in the regression equation prediction for the **sample** data (versus ϵ). For each (x_i, y_i) observed sample data, we can plug x_i into the regression equation and estimate $\hat{y}_i$. The residual is the difference of the **observed** y_i from the **predicted** $\hat{y}_i$.

$$e_i = y_i - \hat{y}_i \tag{3}$$
$$= (\text{observed } y) - (\text{predicted } y) \tag{4}$$

1.2 Simple Linear regression

USE

Often used to help answer:

1. Can I use X to predict Y?
2. What is the equation that models X and Y?
3. What is the best predicted value for Y if the value of $X = X'$?

COMPUTATION

STEPS FOR REGRESSION

Use the following steps to model a linear relationship between two quantitative variables:

1. Determine which variable is the predictor variable (x) and which variable is the response variable (y).
2. Make a scatter plot of the data to determine if the relationship is linear.
3. Check for influential points.
4. Determine if the linear correlation coefficient is significant.
5. Write the model and determine the coefficients (b_0 and b_1).
6. Plot the regression line on the data.
7. Check the residuals for any patterns.

What is a best fit line?

LEAST-SQUARES PROPERTY. DEFINITION 1.4

We will define the "best" fit line to be the line that minimizes the squared residuals. Thus, the best line results in the **smallest possible** sum of squared error (SSE):

$$\text{SSE} = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 \tag{5}$$

The linear regression equation

$$\boxed{\hat{y} = b_0 + b_1 x} \tag{6}$$

Where b_0 and b_1 satisfying the least-squares property are:

$$\boxed{b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}} \tag{7}$$

$$\boxed{b_0 = \bar{y} - b_1 \bar{x}} \tag{8}$$

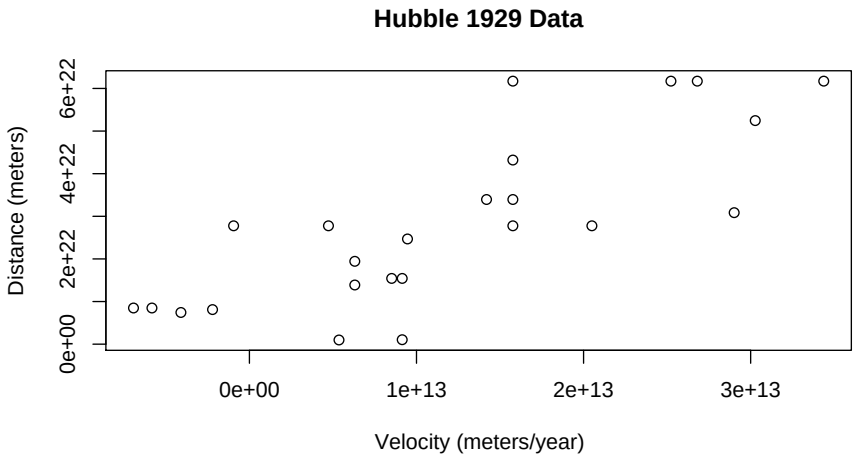
A COMPLETE EXAMPLE

Finding the regression equation for our example

We will attempt to find a suitable regression equation for the 1929 Hubble data. (1) The velocity is the **predictor** variable (x) and the distance is the

response variable (y). (2) Plot the data to check for a linear relationship & (3) check for influential points:

```
R: x = Vel.m.year
R: y = Dist.m
R: plot(x, y, xlab = "Velocity (meters/year)", ylab = "Distance (↩
    meters)",
+     main = "Hubble 1929 Data")
```



(4) Ensure that linear correlation is significant!

```
R: cor.test(x, y)
        Pearson's product-moment correlation

data:  x and y
t = 6.0173, df = 22, p-value = 4.681e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.56509 0.90436
sample estimates:
   cor
0.7887
```

The linear correlation is significant (p-value ≈ 0), thus a linear model is appropriate. The velocity can predict $r^2 = 62.2\%$ of the variation in distance.

(5) The Model. We believe a **linear** model for predicting distance based on velocity is appropriate:

$$\text{distance} = b_0 + b_1 \cdot \text{velocity}$$
$$\hat{y} = b_0 + b_1 \cdot x$$

Finding the regression equation:

(a) Define needed variables:

```
R: x.bar = mean(x)
R: x.bar
```

```
[1] 1.1767e+13
R: y.bar = mean(y)
R: y.bar
[1] 2.816e+22
```

(b) Find the slope using equation 7:

```
R: b1 = sum((x - x.bar) * (y - y.bar))/sum((x - x.bar)^2)
R: b1
[1] 1339313079
```

(c) Find the y-intercept using equation 8:

```
R: b0 = y.bar - b1 * x.bar
R: b0
[1] 1.2400e+22
```

Thus our linear model for this relationship is:

$$\hat{y} = (1.24e + 22) + (1.34e + 09) \cdot x$$

Question 3. Since the **slope** should indicate the age of the universe, what is the universe's age?

MAKING PREDICTIONS

To predict the **average** value of the response variable for a given x: plug in the specific value of x that you wish to make a prediction for in the regression equation.

Cautions when making predictions

- Stay within the scope of the data. Don't predict outside the range of sample x values.
- Ensure your model is applicable for what you wish to predict. Is it the same population? Is the data current?

When the linear correlation is NOT significant

If the linear correlation is not significant, you should not use the regression model for predictions.

In this case, **the best point estimate for y is $\bar{y}$.**

Question 4. For our Hubble data model, what is the maximum velocity that we can use to make a prediction?

Question 5. A new galaxy has been discovered. We need to estimate it's distance from earth, however, it's much easier to estimate it's velocity. The velocity is approximately 2.2×10^{13} m/year. What is the best point estimate for it's distance?

Question 6. If the linear correlation coefficient had not been significant for the Hubble data, what would be the best point estimate of a galaxy's distance if it's velocity is 2.2×10^{13} m/year?

1.3 Regression using R

Rather than typing in the equations for b_0 and b_1 each time, R can calculate them for us and make the necessary diagnostic plots:

R Command

SIMPLE LINEAR REGRESSION:

```
results=lm(y ~ x)
results
plot(x, y)
abline(results)
plot(x, results$resid)
qqnorm(results$resid)
```

For simple linear regression use a model: `y ~ x` to indicate that y is linearly related to x. Both x and y are **ordered vectors** of data. Output shows regression coefficients, plots the data with the regression line, and plot residuals.

Using R to find the regression model:

```
R: x = Vel.m.year
R: y = Dist.m
R: results = lm(y ~ x)
R: results
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
```

```
    1.24e+22        1.34e+09
```

Question 7. What is the regression model?

Note that for the Hubble data, it physically makes sense for the intercept to be 0.[2] Let's modify the model:

```
R: results = lm(y ~ x + 0)
R: results
Call:
lm(formula = y ~ x + 0)

Coefficients:
        x
1.88e+09
```

Question 8. What is our new regression model?

Question 9. What is our new point estimate for the age of the universe?[3]

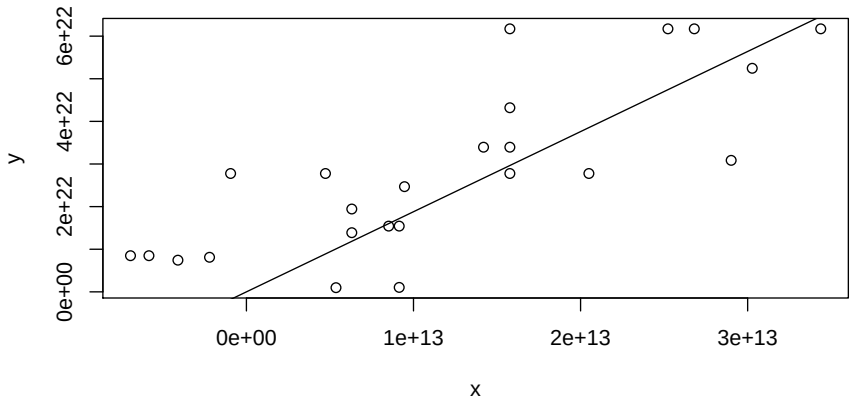
Question 10. Rather than a point estimate for the age of the universe, what statistical tool would provide more information?

Plotting the regression equation on the data to check model
Use the following commands:

```
R: plot(x, y)
R: abline(results)
```

[2] If the distance of the galaxy is 0, we are in our galaxy which has a 0 velocity relative to us.

[3] Modern estimates for the age of the universe from 12-14 billion years. However, it's not quite that simplistic, read more at http://map.gsfc.nasa.gov/universe/uni_age.html

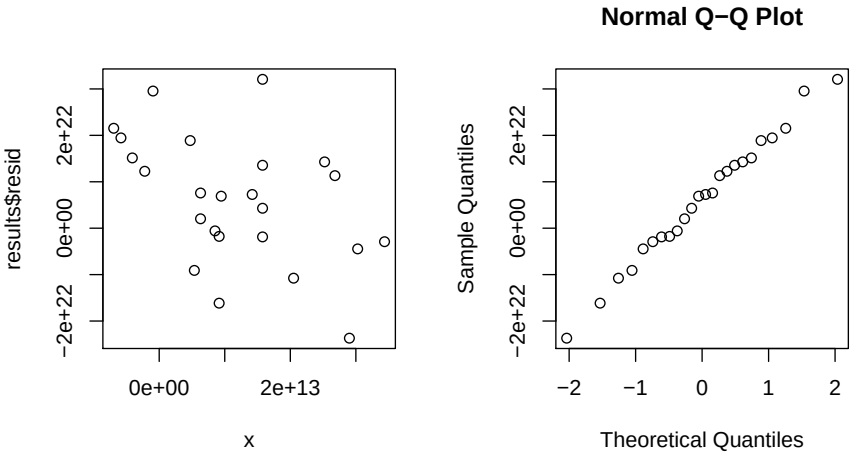


RESIDUAL PLOTS

Checking our assumptions:

1. Linear model: no obvious pattern in residuals, should look random.
2. $\mu_\epsilon = 0$: mean should be zero.
3. σ_ϵ^2 constant: variance along x should be constant.
4. ϵ are normally distributed. (Can check with QQ plot of residuals.)

```
R: par(mfrow = c(1, 2))
R: plot(x, results$resid)
R: qqnorm(results$resid)
```



1.4 Summary

Linear Regression and Predictions
Requirements: (1) linear relationship (2) residuals are random (independent), have constant variance across x and are normally distributed.

1. Determine **predictor variable** (x) and **response variable** (y).
2. Check for linear relationship: `plot(x,y)` (otherwise stop!)
3. Check for influential points.
4. Check for statistically significant correlation: `cor.test(x,y)`
   If a significant relation **does not exist**, the best prediction for **any** x is $\bar{y}$.
5. Find the regression equation: `results=lm(y ~ x)` .
6. Plot data & line: `plot(x, y); abline(results)`
7. Plot residuals: `plot(x, results$resid); qqnorm(results$resid)`
   Don't predict outside of sample data x values!

1.5 Additional Examples

Use Data Set 1 in Appendix B:

Example 2. Find a linear model to predict the leg length (cm) of men based on their height (in). Then predict the leg length of a 68 in male. Also, how much variation does the model explain?

Example 3. Find a linear model to predict the cholesterol level (mg) of men based on their weight (lbs) . Then predict the cholesterol of a man weighing 200 lbs.