THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Linear Model

10/1/2019

Correlation and Regression

- Models and Data Reduction, what is a model?

Description of how the world works and how the data where generated - makes expectation about relationships

Example: Normal Model

Correlation and Simple Linear Regression

What is Correlation?

Correlation analysis is used to measure strength of the association (linear relationship) between two variables

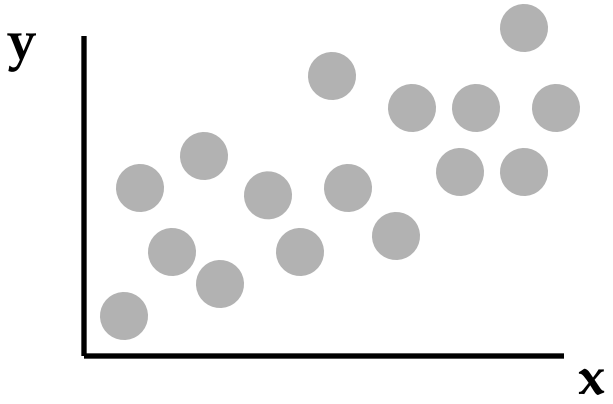
- ❖ Only concerned with strength of the relationship
- ❖ No causal effect is implied

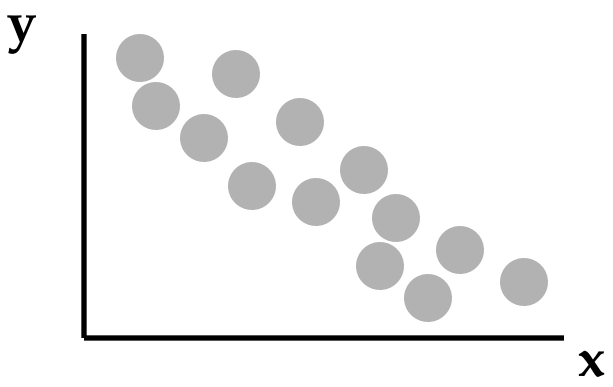
What is Scatter Plot?

Used to show the relationship between two variables

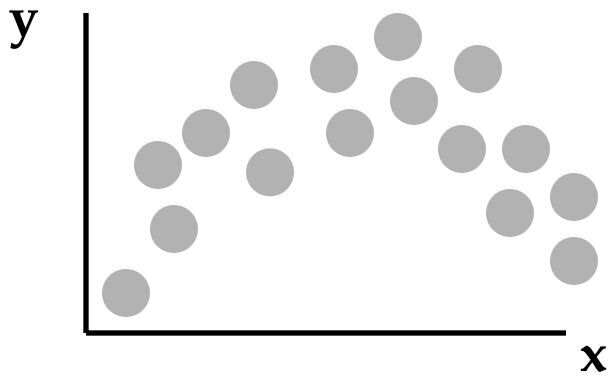
Types of Relationships

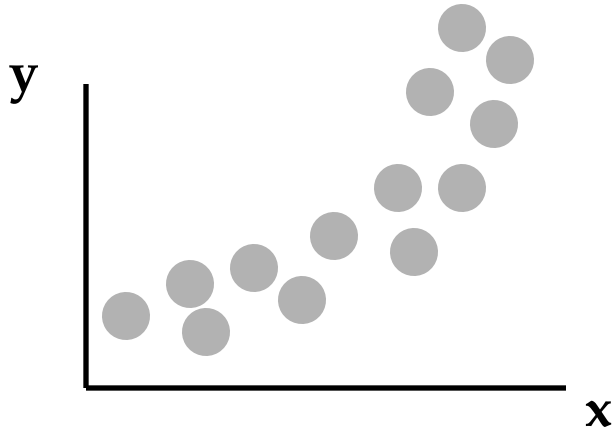
Linear relationships



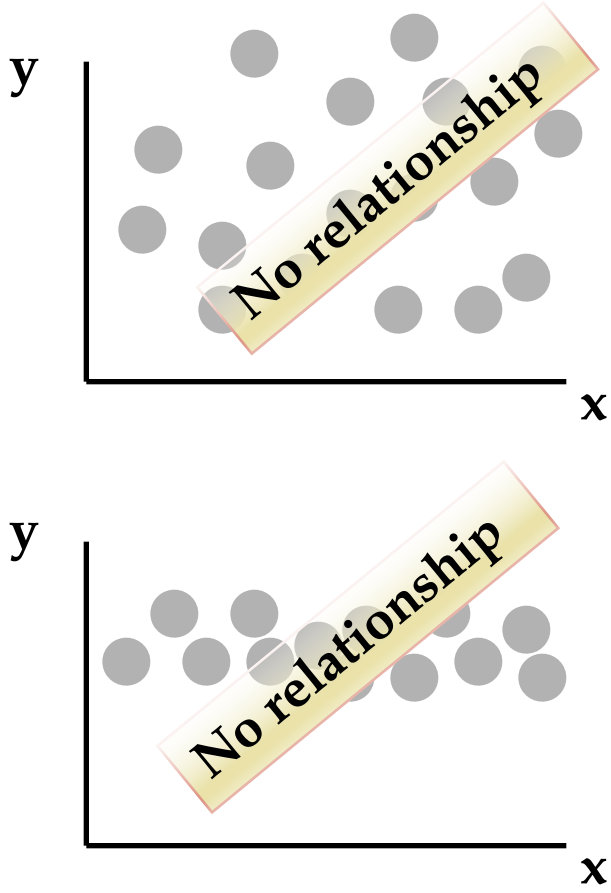


Curvilinear relationships





What's the Relationship?



Correlation Coefficients

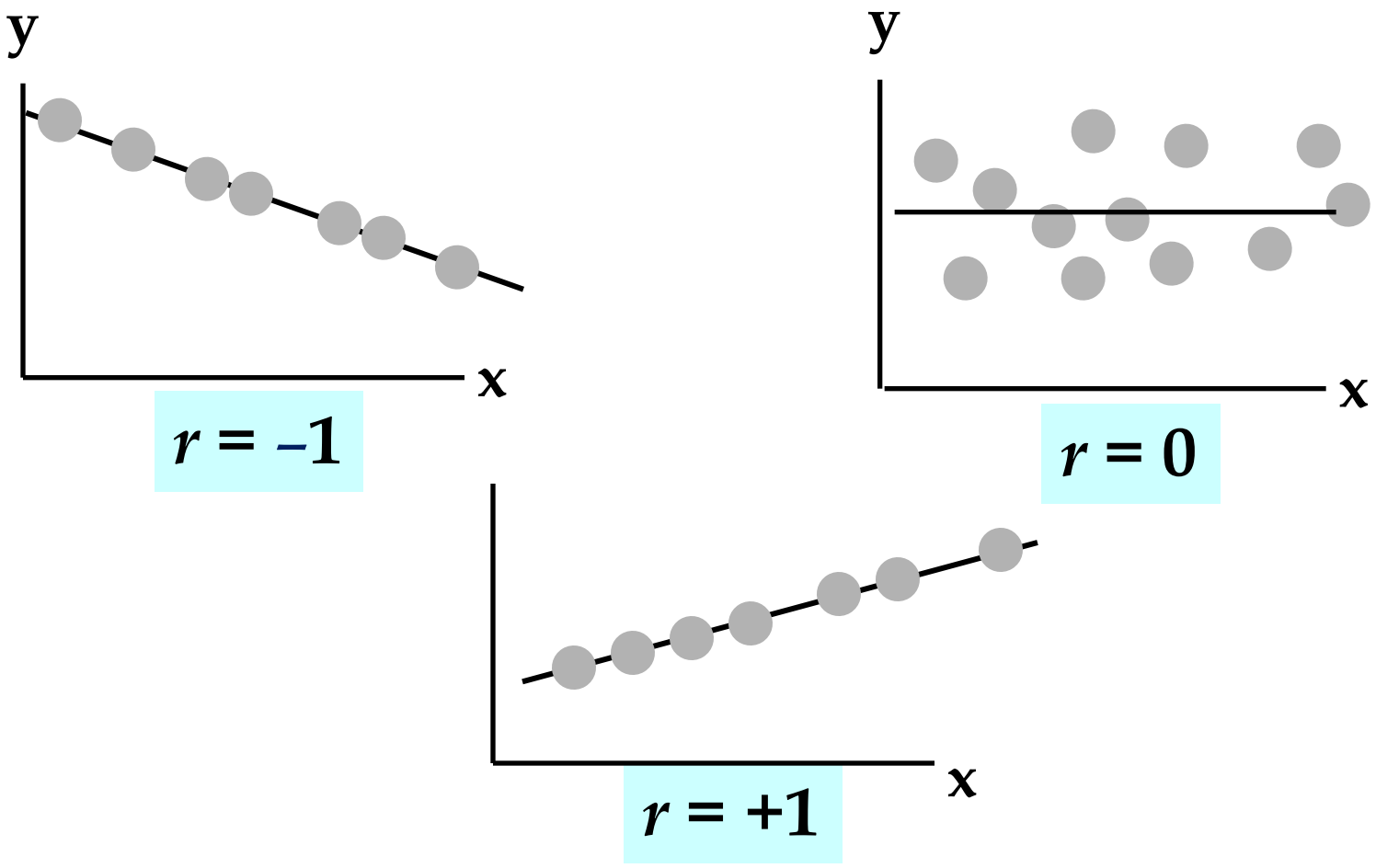
The population correlation coefficient ρ (rho) measures the strength of the association between the variables

The sample correlation coefficient r is an estimate of ρ and is used to measure the strength of the linear relationship in the sample observations

Characteristics of Coefficient

- Range between –1 and 1 (inclusive)
- The closer to –1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker the linear relationship

Examples Correlations



Calculating Correlation

Sample correlation coefficient:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2][\sum (y - \bar{y})^2]}}$$

or the algebraic equivalent:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

where:

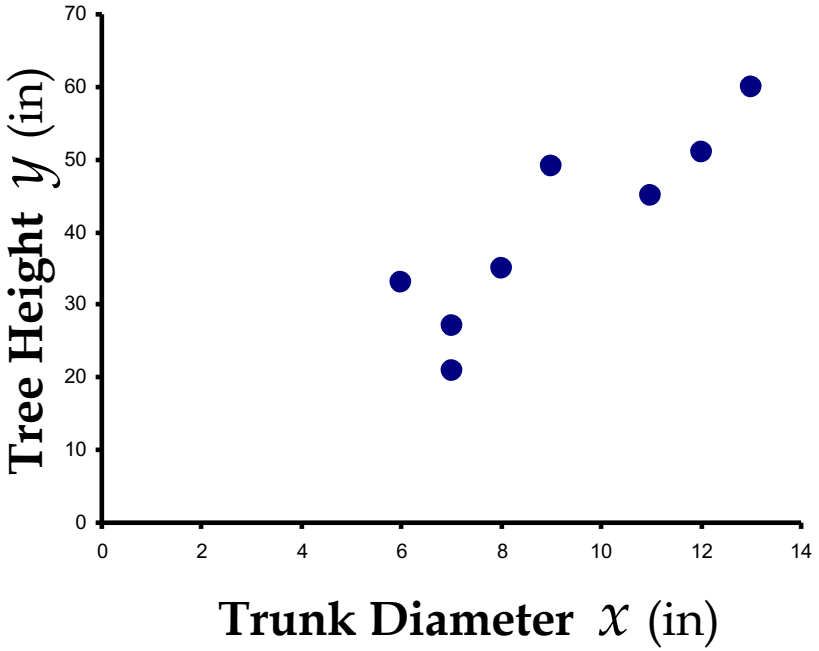
r = Sample correlation coefficient
n = Sample size
x = Value of the independent variable
y = Value of the dependent variable

Example

Tree Height	Trunk Diameter			
y	x	xy	y^2	x^2
35	8	280	1225	64
49	9	441	2401	81
27	7	189	729	49
33	6	198	1089	36
60	13	780	3600	169
21	7	147	441	49
45	11	495	2025	121
51	12	612	2601	144
Σ=321	**Σ=73**	**Σ=3142**	**Σ=14111**	**Σ=713**

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

Results Correlation Coefficient



$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

$$= \frac{8(3142) - (73)(321)}{\sqrt{[8(713) - (73)^2][8(14111) - (321)^2]}}$$

$$= 0.886$$

Measure of Correlation

- ~~0 < |r| < .3 - weak correlation~~
- ~~3 < |r| < .7 - moderate correlation~~
- ~~|r| > 0.7 - strong correlation~~

TABLE 10-6 Critical Values for Correlation Coefficient r

n	$\alpha = 0.05$	$\alpha = 0.01$	n	$\alpha = 0.05$	$\alpha = 0.01$	n	$\alpha = 0.05$	$\alpha = 0.01$
3	1.00	1.00	13	0.53	0.68	23	0.41	0.53
4	0.95	0.99	14	0.53	0.66	24	0.40	0.52
5	0.88	0.96	15	0.51	0.64	25	0.40	0.51
6	0.81	0.92	16	0.50	0.61	26	0.39	0.50
7	0.75	0.87	17	0.48	0.61	27	0.38	0.49
8	0.71	0.83	18	0.47	0.59	28	0.37	0.48
9	0.67	0.80	19	0.46	0.58	29	0.37	0.47
10	0.63	0.76	20	0.44	0.56	30	0.36	0.46
11	0.60	0.73	21	0.43	0.55			
12	0.58	0.71	22	0.42	0.54			

Correlation Versus Prediction

What's the difference and how can they work together?

Correlation is a measure of how closely one variable (usually Y) changes in accordance with other variables (typically X_i)

Prediction is the process of estimating scores on one variable (Y), on the basis of other variables/predictors (X_i)

Regression

- Best with quantitative data to increase the inferential power
- Try to model or explain relationship between Y and X_i
- Y - *response, output, dependent* variable
- X_i - *predictor, input, independent, explanatory, regessor* variables
- Y must be continuous variable, X categorical or quantitative
- Centering/scaling (z-score) variables is often recommended
- Intercept is then interpreted as **the expected Y value when the predictor are set to their means**. Otherwise, the intercept is the expected Y value when the predictors are at 0, which may not be realistic or interpretable (e.g. predictors are height/weight)

Scaling into z-scores

It is assuring to know the following fact of z-scores:

Chebyshev's theorem / inequality

For any number k greater than 1, at least $(1 - 1/k^2)$ of the data values lie k standard deviations of the mean.

Std Dev	Normal Distribution	Any Distribution
1	68.27%	≥ 0.00%
2	95.45%	≥ 75.00%
3	99.73%	≥ 88.89%
4	99.99%	≥ 93.75%

What are the most commonly used Regression techniques?

1. Bivariate Regression
2. Multiple Regression
3. Logistic Regression

Three Main Types of Multiple Regression?

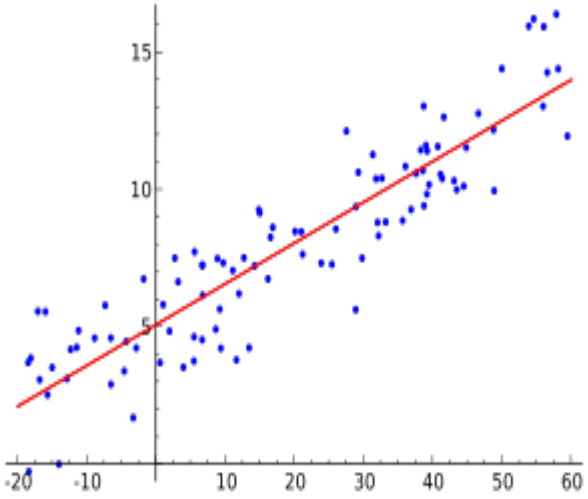
1. Simultaneous – standard issue

2. Step-wise – variables added to the regression equation based on their correlation value with the dependent variable

3. Hierarchical – Order is up to the researcher but typically items that are wanted to be controlled are entered into the equation first

Regression Review

Linear regression finds the straight line, called the **least squares regression line** or LSRL, that best represents observations in a data set. Suppose Y is a dependent variable, and X is an independent variable. Then, the equation for the regression line would be: $\hat{y} = b_0 + b_1 x$



Linear regression equation

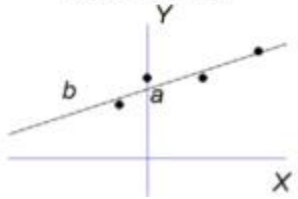
(without error)

$$\hat{Y} = bX + a$$

predicted values of Y

slope = rate of increase/decrease of Y hat for each unit increase in X

Y-intercept = level of Y when X is *0*.



Assumptions

- Normally shaped data – Histogram or other tests
- Linearity : dependent vs independent – Scatter Plot
- Multicollinearity : occurs when independent variables are not independent – checked with VIF
  - vif = 1 : not correlated
  - 1 < vif < 5 : moderately correlated
  - vif > 10 : highly correlated
- Auto-correlation : Occurs when residuals are not independent from each other – stock prices example- Durbin-Watson d tests
- Homoscedasticity : Error term along the regression line are equal – Scatter Plot – can convert the dependent variable.

R – formula input format

- the basic formula is Y ~ X , e.g. fit ← glm(Y ~ X)
- forcing intercept to be zero: fit ← glm(Y ~ X – 1)
- use all available variables: fit ← glm(Y ~ . , data = 'my_df')
- exclude some vars: fit ← glm(Z ~. – X – Y , data = 'my_df')
- include more predictors/vars: Z ~ X + Y
- include interaction term X:Y (i.e. X times Y)
- include the vars and their interaction X*Y (i.e. X+Y+X:Y)
- as-is, multiplied I(X*Y)
  - X*X will only include X, while I(X*X) will include X:X
- conditional term X|Y (i.e. X given Y)
- include the vars up to 3-way interaction term (X+Z+W)^3
  - (X+Z+W)^3 = X+Z+W + (X:Z +Z:W+W:X) +X:Z:W

Coefficient of Determination – r^2

This statistic quantifies the proportion of variance of y explained by x (statistical-wise, not causal-effect).

– example in R below: 26.97% variability in trunk Height is explained by Girth

```
Call:
lm(formula = Height ~ Girth, data = trees)

Residuals:
     Min       1Q   Median       3Q      Max
-12.5816  -2.7686   0.3163   2.4728   9.9456

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.0313     4.3833  14.152 1.49e-14 ***
Girth         1.0544     0.3222   3.272  0.00276 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.538 on 29 degrees of freedom
Multiple R-squared:  0.2697,    Adjusted R-squared:  0.2445
F-statistic: 10.71 on 1 and 29 DF,  p-value: 0.002758
```

R Regression Output

- Residuals? Essentially the difference between the actual observed response values and the response values that the model predicted. Want to see how symmetrical around zero the residuals are.

R Regression Output: Coefficients

- Formula Call? Regression equation
- Estimate? Coefficients for the regression equation
- Standard Error?

  Measures the average amount that the coefficient estimates vary from the actual average value of our response variable.

- t value? No. of Stand Dev. coefficient is from zero. Estimate/Stand Error – larger is better
- P value?

  A small p-value indicates that it is unlikely we will observe a relationship between the predictor and response variables due to chance, a p-value of 5% or less is a good cut-off point

R Regression Output

- Residual Standard Error

  The average amount that the response will deviate from the **true regression line.**

- Multiple R-squared?

  Provides a measure of how well the model is fitting the actual data. We get is **.2697** or roughly 27% of the variance found in the response variable can be explained by the predictor variable.

R Regression Output

- F-statistic

  Used to determine statistical significance of the model. Further from 1 the better. In our example the critical value at alpha level .05 for 1 and 22 is 4.30

  Essentially, a small p-value for F-statistics concludes to reject the null model, which is the constant model that all the coefficients are zero.

Practice

- Using the mtcars dataset create a model that predicts mpg based on cyl, disp, hp and wt
- How many variables help to predict mpg?
- How well does the model fit the data?
- How accurate is the prediction?
- What are the VIF values?