STAC67: Regression Analysis

Lecture 1

Sohee Kang

Jan. 13, 2021

Regression: What is it?

- Simply: The most widely used statistical tool for understanding relationships among variables.
- A conceptaully simple method for investigating relationships between one or more factors of outcome of interest.
- The relationship is expressed in the form of an equation or a model connecting the outcome to the factors.
- Examples of application:
  - **Epidemiology**: what are the social factors to contribute the death rate of Covid-19 in Canada?
  - **Business**: determing price and marketing strategy:
    - Estimate the effect of price and advertisement on sales
    - Decide what is optimal price and campaign?
  - Straight prediction questions:
    - What will be the interest rates be next month?
    - What will be the length of hospital stay of a surgical patient?
  - Explanation and understanding:
  - Much more . . . .

Origin of Regression (History)

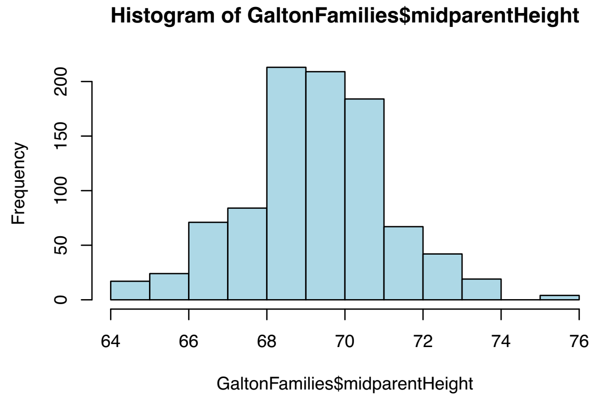
- In 1886, Sir Francis Galton (1822-1911) invented the term, **"regression"** after he analyzed the data on the heights of parents and their children.

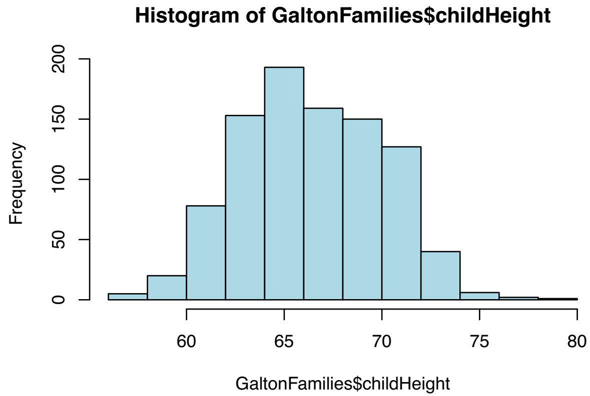
"Regression towards mediocrity in hereditary stature". The Journal of the Anthropological Institute of Great Britain and Ireland, Vol 15, pages 246-263 (or Wikipedia!)

Data Visualization

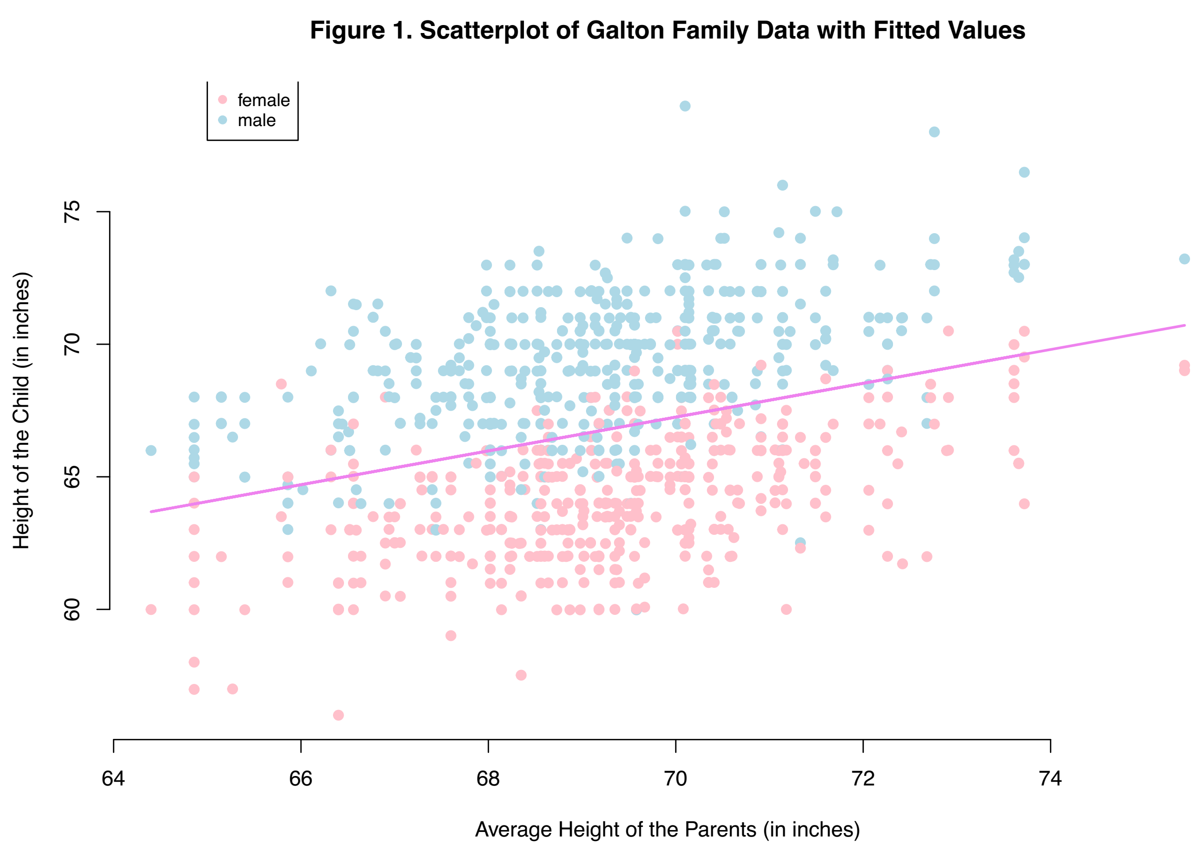
Example: Galton family data

- The GaltonFamilies(HistData) dataset lists the individual observations for 934 adult children born to 205 fathers and mothers He wrote that, “the average regression of the offspring is a constant fraction of their respective mid-parental deviations.” For height, Galton estimated this regression coefficient to be about two-thirds (2/3).





Data Visualization



Review

- Covariance and Correlation Coefficient

Suppose we have observations on n subjects consisting of a **dependent** or **response variable** Y and an **independent** or **explanatory variable** X.

- Measure both direction and strength of the relationship between Y and X.

Obs	Y	X
1	y_1	x_1
2	y_2	x_2
$\vdots$	$\vdots$	$\vdots$
n	y_n	x_n

Covariance and Correlation

calculate this $\iint (x-\mu_x)(y-\mu_y) f_{xy}\,dx\,dy$

Def. $Cov(X, Y) = E((X - \mu_x)(Y - \mu_y))$, where $\mu_x = E(X)$, $\mu_y = E(Y)$

$Z_x = \frac{X-\mu_x}{\sqrt{Var(X)}}$, $Z_y = \frac{Y-\mu_y}{\sqrt{Var(Y)}}$

standardized X   standardized Y

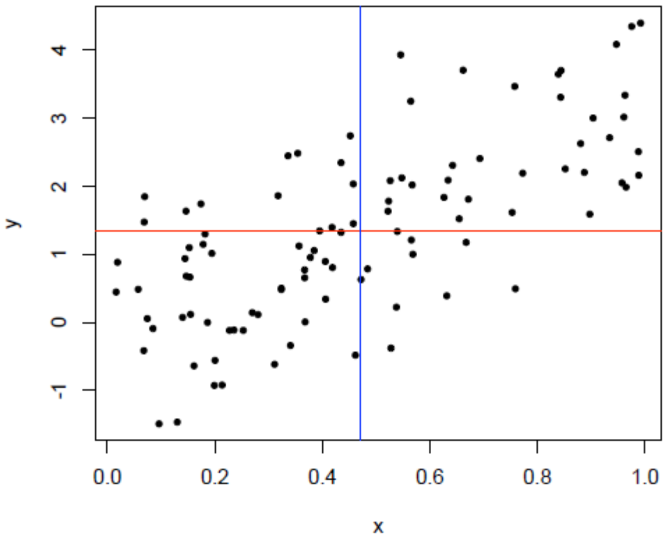
$$Cov(Z_x, Z_y) = \rho_{xy} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = Corr(X, Y)$$

$= E((Z_x - \mu_{z_x})(Z_y - \mu_{z_y}))$

$= E(Z_x Z_y) = \frac{1}{\sigma_x \sigma_y} E((x-\mu_x)(y-\mu_y))$

- $-1 \leq \rho_{xy} \leq 1$
- When the relationship is perfectly linear then $|\rho| = 1$.
- if two variables are independent then $\rho = 0$. (Note: the inverse does not hold)

Sample Covariance and Correlation



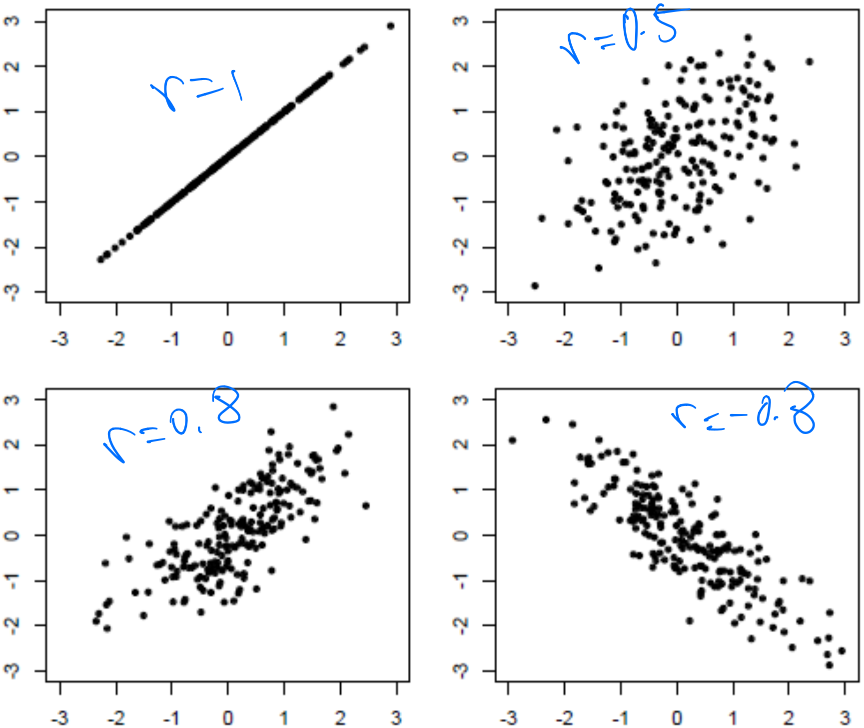
$Cov(X, Y) = \dfrac{\sum_i (Y_i - \bar{Y})(X_i - \bar{X})}{n-1}$

= Sample Covariance

Let $Z_Y = \dfrac{Y - \bar{Y}}{\sigma_Y}$ Let $Z_X = \dfrac{X - \bar{X}}{\sigma_X}$

$$Cor(Y, X) = Cov(Z_y, Z_x) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{y_i - \bar{y}}{s_y} \right) \left(\frac{x_i - \bar{x}}{s_x} \right)$$ = Sample Correlation Coefficient

Correlation



$r_{xy} = r_{yx}$

Question: what are main differences between correlation and regression model? Corr : measures linear relation between x,y

Test for Population correlation

population correlation

When $\rho = 0$, and the joint distribution of (X, Y) is bivariate normal, and it can be shown that:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

has a student's t distiruibution with $n-2$ digress of freedom

$$H_0 : \rho = 0 \qquad vs \qquad H_1 : \rho \neq 0$$

- Testing Procedure

- Calculating the observed value of t (call this t_{obs})
- Compute the p-value for the test

Simulation

```
par(mfrow=c(2,2))
library(mvtnorm)
```

— multivariate normal package

```
## Warning: package 'mvtnorm' was built under R version 3.5.3
```

```
sigma.1 = matrix(c(1, 0, 0, 1), ncol=2)
sigma.2 = matrix(c(1, 0.9, 0.9, 1), ncol=2)
```

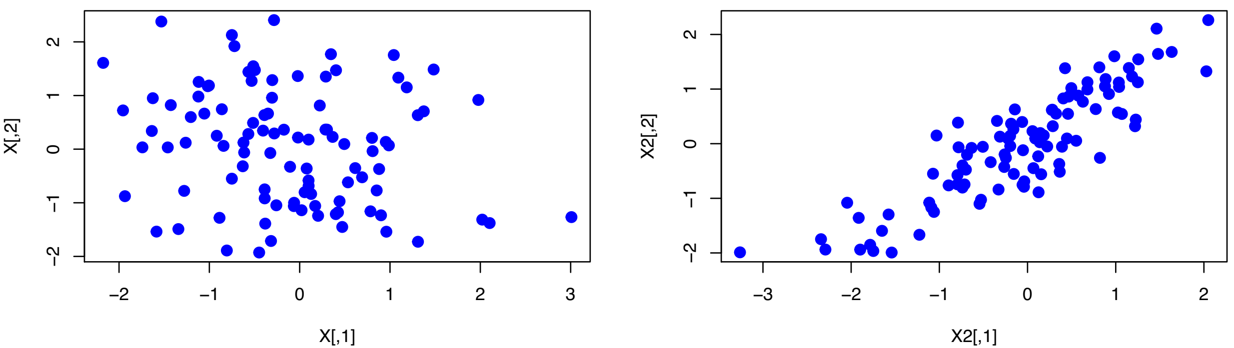
r N r μ r Σ

```
X = rmvnorm(100, mean=c(0,0), sigma.1)
plot(X, pch=20, cex=2, col="blue")
```

r generate sample

```
X2 = x = rmvnorm(100, mean=c(0,0), sigma.2)

plot(X2, pch=20, cex=2, col="blue")
```



Simulation

```
x = X2[,1]
y = X2[, 2]
cor.test(x, y)
```

For $H_0: \rho = 0$ $H_1: \rho \neq 0$

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 18.084, df = 98, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.822453 0.915802
## sample estimates:
##       cor
## 0.8771691
```

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(98)$$
$$= 20.73$$

$\alpha = 5\%$

since p-value $< \alpha$

Thus we reject the null hypothesis.

```
r= cor(x, y)
t = r*sqrt(98)/sqrt(1-r^2)
t
```

```
## [1] 18.08385
```

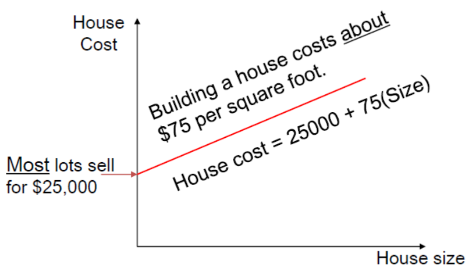
Relationship between variables

What factor or variable affects the price of house?

- Relation of the form

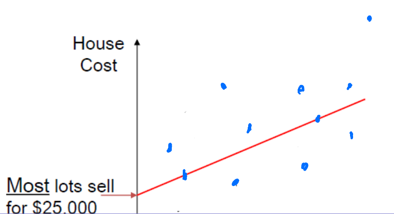
1. Mathematical Relation: $Y = f(X)$

where, X, Y are variables and f is a function



2. Statistical Relation:

$$Y = f(X) + \epsilon$$



Data Collection for regression analysis

- **Observational study**
  - Investigator has no control over the explanatory variables (X)
  - Limitation: not adequate for cause-and-effect A strong association does not necessarily means a cause-and-effect relationship
- **Experiment**
  - Investigator exercises control over the explanatory variables (X) through random assignment
  - Random assignment balances out effect of other variables that might affect Y
  - Gold standard for cause-and-effect conclusions

- Always understand how data is collected

The Regression Process

1. The researcher must clearly define the question(s) of interest in the study
2. The response variable Y must be decided on, based on the question of interest.
3. A set of potentially relevant covariates, which can be measured, needs to be defined.
4. Data is collected.
5. Model Specification.
6. Decide on a method for fitting the specified model
7. Fit the model - typically using software such as R
8. Examine the fitted model for violations of assumptions.
9. Conduct hypothesis testing for questions of interest.
10. Report the results from statistical inference.