MAT 5030
Chapter 6:
Regression and correlation

Kazuhiko Shinki

Wayne State University

Reference

Our Chapters 6 & 11 are roughly equivalent to Chapters 2 & 3 of our second textbook: *Julian J. Faraway, "Practical Regression and Anova using R"*.

6.1 Simple linear regression

Independent and Dependent Variables:

When we assume that a variable $\boldsymbol{y}$ depends on a variable $\boldsymbol{x}$,

- $\boldsymbol{x}$ is called an **independent variable (or predictor)** .
- $\boldsymbol{y}$ is called an **dependent variable (or response)** .

Examples:

- $\boldsymbol{x}$: total calories intake of a mouse, $\boldsymbol{y}$: weight of the mouse.
- $\boldsymbol{x}$: interest rate this year, $\boldsymbol{y}$: inflation rate next year.
- $\boldsymbol{x}$: horse power of a car, $\boldsymbol{y}$: maximum speed of the car.

Note:
We use lower cases $\boldsymbol{x}$ and $\boldsymbol{y}$, because we reserve capital letters for matrix representation later in this chapter.

Simple linear regression:
Suppose data (x_i, y_i) $(i = 1, \cdots, n)$ are given.
Simple (=univariate) linear regression assumes the relationship:

$$y_i = a + bx_i + \epsilon_i \tag{1}$$

where ϵ_i is an **error** (or **residual**) and a and b are **parameters** to be estimated.

Our objective is to estimate a and b. ϵ_i will be obtained after getting a and b.

6.1 Simple linear regression

Probabilistic Interpretation:

- a, b: constants (unobserved)
- x: deterministic, i.e., constants (observed)
- ϵ: random variable (unobserved)
- y: random variable (observed)

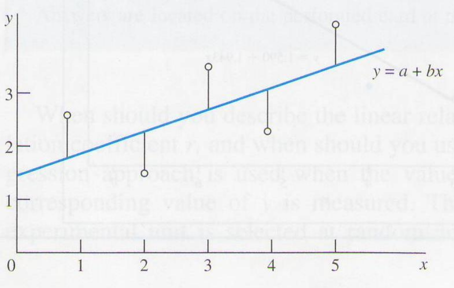
Additional assumptions such as ϵ_t is independent and identically distributed (IID) with a distribution $N(0, \sigma^2)$ (where σ is an unknown parameter) are often imposed.

6.1 Simple linear regression

Least square method:

In simple linear regression, we determine ***a*** and ***b*** to minimize the sum of squared errors (residuals):

$$SS_{Res} = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$



6.1 Simple linear regression

Theorem 1:

The following pair of $\hat{a}$ and $\hat{b}$ minimizes SS_{Res}.

$$\hat{b} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \quad (2)$$

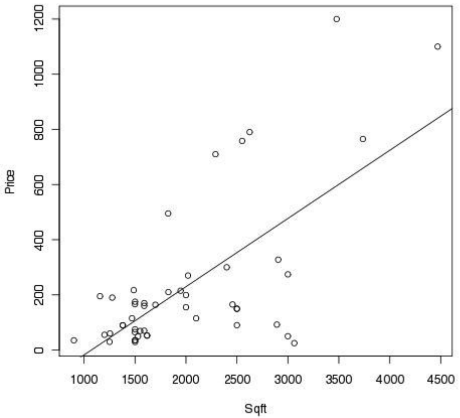
$$\hat{a} = \bar{y} - \hat{b}\bar{x} \quad (3)$$

$\hat{a}$ and $\hat{b}$ are called **estimates** for a and b. Estimates depend on x and y, so are random variables.

6.1 Simple linear regression

Example: 3 bedroom condo price in Detroit

We want to explain the condo price y (thousand dollars) by square footage x by a linear model: $y = a + bx + \epsilon$.



6.1 Simple linear regression

Sample Code:

```
> Data <- read.table("Condo.csv", sep=",", header=T)
> attach(Data)
> str(Data)
'data.frame': 48 obs. of  2 variables:
 $ Price: num  790 709.9 149.9 209.9 89.7 ...
 $ Sqft : int  2625 2290 2500 1827 1380 1380 1470 1276 1487 2000 ...

> LM1 <- lm(Price ~ Sqft) # Price = a + b*Sqft + epsilon
> LM1
Call:
lm(formula = Price ~ Sqft)

Coefficients:
(Intercept)         Sqft
  -265.0250       0.2473
```

The output says: $y = -265.0250 + 0.2473x + \epsilon$.

6.1 Simple linear regression

(Sample code for the figure:)

```
plot(Sqft, Price)  # scatter plot for (x,y)
abline(LM1)  # add regression line: y = a + bx
```

6.1 Simple linear regression

Matrix representation:
For programming and notational purposes, it is more convenient to use a matrix representation of the theorem above.

Let

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \beta = \begin{bmatrix} a \\ b \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Then (1) turns to $\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon$ and the above theorem becomes:

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{4}$$

($\boldsymbol{X}'$ is the transpose of $\boldsymbol{X}$).

(Proof of (4):)

Let

$$l(\beta) := SS_{Res} = (Y - X\beta)'(Y - X\beta).$$

$\frac{\partial l}{\partial \beta}$ has to be zero to minimize $l(\beta)$.

$$\frac{\partial l}{\partial \beta}(\hat{\beta}) = 2X'(Y - X\hat{\beta}) = 0$$
$$\Leftrightarrow \quad \hat{\beta} = (X'X)^{-1}X'Y. \qquad \blacksquare$$

6.1 Simple linear regression

Sample Code: 3 bedroom condo price

```
> Y <- Price
> X <- cbind(rep(1, times= 48), Sqft)

> solve(t(X) %*% X) %*% t(X) %*% Y  # beta hat = (X'X)^{-1} X'Y
             [,1]
     -265.0249880
Sqft    0.2472783
```

The estimated error $\hat{\epsilon}_i$ (scalar) and $\hat{\epsilon}$ (vector) are defined by:

$$\hat{\epsilon}_i = \mathbf{x}_i \hat{\beta} - y_i,$$

or in a matrix form,

$$\hat{\epsilon} = \mathbf{X} \hat{\beta} - \mathbf{Y}.$$

6.1 Simple linear regression

Variability of $\hat{\beta}$:

$\hat{\beta}$ is estimated by the data (in other words, $\hat{\beta}$ is random since it is a function of a random variable $\boldsymbol{Y}$), so it has some variability. To be concrete,

$$\hat{\beta} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}(\boldsymbol{X}\beta + \epsilon) = \beta + (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\epsilon \quad (5)$$

and ϵ causes some randomness.

6.1 Simple linear regression

Theorem 2:
Suppose ϵ_i's ($i = 1, \cdots, n$) are independent, $E[\epsilon_i] = 0$ and $E[\epsilon_i^2] = \sigma^2$. Then,

$$\mathrm{Var}(\hat{\beta}) = (X'X)^{-1}\sigma^2.$$

Note:

- $\mathrm{Var}(\hat{\beta})$ is a 2 by 2 covariance matrix:

$$\mathrm{Var}(\hat{\beta}) = \begin{bmatrix} \mathrm{Var}(\hat{a}) & \mathrm{Cov}(\hat{a}, \hat{b}) \\ \mathrm{Cov}(\hat{a}, \hat{b}) & \mathrm{Var}(\hat{b}) \end{bmatrix}$$

- Since σ^2 is not observed, we substitute it by $MSE := \frac{1}{n-2} \sum \hat{\epsilon}_i^2$.

6.1 Simple linear regression

(Proof of the Theorem 2:)

In general, for a random vector $\boldsymbol{e}$ with $\boldsymbol{E[e] = 0}$, the variance matrix $\mathrm{Var}(\boldsymbol{e}) = \boldsymbol{E[ee']}$. Hence, by (5),

$$\begin{aligned} \mathrm{Var}\hat{\beta} &= \mathrm{Var}((\boldsymbol{X'X})^{-1}\boldsymbol{X'}\epsilon) = \boldsymbol{E}\left[(\boldsymbol{X'X})^{-1}\boldsymbol{X'}\epsilon((\boldsymbol{X'X})^{-1}\boldsymbol{X'}\epsilon)'\right] \\ &= \boldsymbol{E}\left[(\boldsymbol{X'X})^{-1}\boldsymbol{X'}\epsilon\epsilon'\boldsymbol{X}((\boldsymbol{X'X})^{-1})\right] \\ &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'E}\left[\epsilon\epsilon'\right]\boldsymbol{X}((\boldsymbol{X'X})^{-1}) \qquad (6) \end{aligned}$$

Since ϵ_i's are independent, $\boldsymbol{E}[\epsilon_i\epsilon_j] = \boldsymbol{0}$ if $\boldsymbol{i} \neq \boldsymbol{j}$ and $\boldsymbol{E}[\epsilon_i\epsilon_j] = \sigma^2$ if $\boldsymbol{i = j}$. Hence, $\boldsymbol{E}[\epsilon\epsilon'] = \sigma^2\boldsymbol{I_n}$ where $\boldsymbol{I_n}$ is a $\boldsymbol{n}$ by $\boldsymbol{n}$ identity matrix. Therefore,

$$\begin{aligned} (6) &= (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\sigma^2\boldsymbol{I_2X}((\boldsymbol{X'X})^{-1}) \\ &= \sigma^2(\boldsymbol{X'X})^{-1}. \qquad \blacksquare \end{aligned}$$

Theorem 3:
Suppose $\epsilon_i \sim N(0, \sigma^2)$ are IID, then

$$T_{\hat{a}} := \frac{\hat{a} - a}{SE(\hat{a})} \sim t(n-2), \quad T_{\hat{b}} := \frac{\hat{b} - b}{SE(\hat{b})} \sim t(n-2). \tag{7}$$

where $SE(\hat{a})$ is the square root of $\mathrm{Var}(\hat{a})$ (same for $\hat{b}$), and $\mathrm{Var}(\hat{a})$ is the (1,1) entry of $\mathrm{Var}(\hat{\beta})$.

We can test $H_0 : a = 0$ (or in general $H_0 : a = a_0$ for any constant a_0) with this theorem (same for $\hat{b}$).

6.1 Simple linear regression

Sample code: 3 bedroom condo price

```
> summary(LM1)

Call:
lm(formula = Price ~ Sqft)

Residuals:
    Min      1Q  Median      3Q     Max
-467.49  -78.24   -8.05   62.47  604.99

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -265.0250    85.2354  -3.109  0.00321 **
Sqft           0.2473     0.0399   6.198 1.46e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 205.8 on 46 degrees of freedom
Multiple R-squared: 0.4551,Adjusted R-squared: 0.4432
F-statistic: 38.41 on 1 and 46 DF,  p-value: 1.461e-07
```

6.1 Simple linear regression

The above result shows:
$SE(\hat{a}) = \sqrt{\mathrm{Var}(\hat{a})} = 85.2354$, $SE(\hat{b}) = 0.0399$,
and $\hat{a}/SE(\hat{a})$ and $\hat{b}/SE(\hat{b})$ follow t- distribution with $df = 46$.

$\hat{a}$ and $\hat{b}$ are significantly different from zero with p-values 0.00321 and 1.46×10^{-7}.

Coefficients of Determination:
The '`lm`' function output "$R^2 = 0.4551$" and "adjusted $R^2 = 0.4432$". These values are calculated as:

$$\begin{aligned} R^2 &= 1 - \frac{SS_{Err}}{SS_{Tot}} \\ adj.R^2 &= 1 - \frac{SS_{Err}/(n-2)}{SS_{Tot}/(n-1)} \end{aligned}$$

where

$$\begin{aligned} SS_{Err} &:= (Y - X\hat{\beta})'(Y - X\hat{\beta}) = \sum (y_i - x_i\hat{\beta})^2 \\ SS_{Tot} &:= (Y - \bar{Y})'(Y - \bar{Y}) = \sum (y_i - \bar{y})^2. \end{aligned}$$

and $\bar{Y}$ is an n by 1 vector whose components are all $\bar{y}$.

6.1 Simple linear regression

Meaning of R^2:
$\boldsymbol{SS_{Err}}$ is the sum of squared errors (ϵ_i^2), and $\boldsymbol{SS_{Tot}}$ is the sum of squared errors when we do not use $\boldsymbol{x}$. Hence, $\boldsymbol{R^2}$ represents how much variation of $\boldsymbol{y}$ (in terms of deviation square) is explained by $\boldsymbol{x}$.

When we have $\boldsymbol{p}$ predictors $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots, \boldsymbol{x}^{(p)}$, the adjusted $\boldsymbol{R^2}$ is defined by

$$adj.R^2 \;=\; 1 - \frac{SS_{Err}/(n-p-1)}{SS_{Tot}/(n-1)}$$

It adjusts a higher $\boldsymbol{R^2}$ when we predict a small number of observations with many predictors. To compare two linear regression models with different numbers of predictors, the adjusted $\boldsymbol{R^2}$ should be used.

6.2 Residuals and fitted values

Fitted values:

The 'fitted', 'fitted.values' and '[5]' return fitted values: $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\beta}$. This is useful if you want to use fitted values for some purposes (e.g., fair market value for condo).

Sample code:

```
> LM1$fitted.values
        1         2         3         4         5         6
384.08044 301.24222 353.17066 186.75239  76.21901  76.21901
.....
       43        44        45        46        47        48
658.80658 105.89240 450.10373 476.80979 595.00879 840.30882

> LM1[5]    # same as LM1$fitted.values

> fitted(LM1) # same as LM1$fitted.values
```

6.2 Residuals and fitted values

Residuals:

The '`resid`', '`residuals`' and '`[2]`' return residuals: $\hat{\epsilon} = \boldsymbol{y} - \boldsymbol{x}\hat{\beta}$. This is useful to check the IID assumption on ϵ. If it has some patterns, the simple linear regression may not be appropriate and we may have to develop another model.

Sample code:

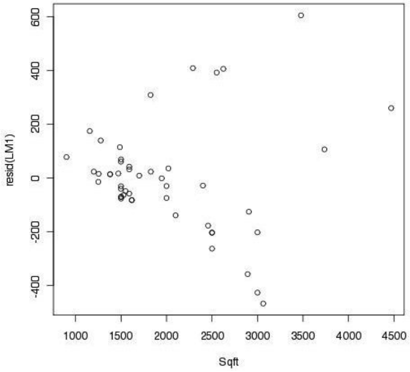
```
> LM1$residuals
        1          2          3          4          5          6
405.91956  408.65778 -203.27066   23.14761   13.45099   13.45099
.....
       43         44         45         46         47         48
106.19342   69.10760 -358.10373 -426.80979  604.99121  259.69118

> resid(LM1)   # same as LM1$residuals
> LM1[2]   # same as LM1$residuals
```

6.2 Residuals and fitted values

The following residual plot suggests:

1. A larger square footage makes its residual larger (in absolute value).
2. The observations around 3000 sqft have a negative bias.



6.2 Residuals and fitted values

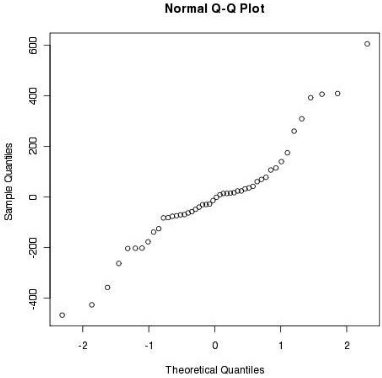
Possible remedies for the above phenomenon:

- Transform both variables (e.g., by **log**) before regression.
- Look into the observations with around 3000 sqft, and search for other variables negatively affecting these properties.

6.2 Residuals and fitted values

Q-Q plot for residuals:

The normal Q-Q plot for residuals indicates that the residuals have heavier tails than normal. This is partially due to (1) in the previous slide.



6.2 Residuals and fitted values

Exercise 1:
Let $(x_i, y_i) =$ $(0,0),(0.5,0.25),(1,1),(1.5,2.25),(2,4),(2.5,6.25),(3,9)$. Regress y on x, make a scatter plot with the regression line, and make a residual plot. Do you see any problems with the residual plot?

Prediction by a regression line:

Suppose we obtained $\hat{\beta}$ based on $(x_1, y_1), \cdots, (x_n, y_n)$. When we get a new observation x_0, our best estimate for y_0 is:

$$\hat{y}_0 = \boldsymbol{x}_0'\hat{\beta} \quad (= \hat{a} + \hat{b}x_0)$$

where we use the notation $\boldsymbol{x}_0 := (1, x_0)'$ (2 by 1 column vector). $\boldsymbol{x}_0$ denotes this red $\boldsymbol{x}_0$ hereafter.

6.3 Prediction and confidence bands

Confidence interval for y_0:

$\hat{y}_0$ has variability since $\hat{\beta}$ is an estimated quantity (i.e., a random variable).

Theorem 4:

$$\mathrm{Var}(\hat{y}_0) = \mathrm{Var}(x_0'\hat{\beta}) = (x_0' \left(X'X\right)^{-1} x_0)\sigma^2.$$

Using this variance, the **$(1-\alpha)$ confidence interval for Ey_0** $(0 < \alpha < 1)$ is:

$$\hat{y}_0 \pm t_{1-\alpha/2}\sqrt{x_0' \left(X'X\right)^{-1} x_0}S. \tag{8}$$

where $S^2 := \frac{SS_{Err}}{n-2}$ is an estimated σ^2 and the degrees of freedom for $t_{1-\alpha/2}$ is $(n-2)$. Ey_0 is within the interval (8) with probability $(1-\alpha)$.

6.3 Prediction and confidence bands

Prediction interval for y_0:

We often want to know the variability of $y_0 = x_0'\beta + \epsilon_0$, not $E[y_0]$. Given $(x_1, y_1), \cdots, (x_n, y_n), x_0$, it is obtained by the following.

Theorem 5:

$$\mathrm{Var}(x_0'\hat{\beta} + \epsilon_0) = (1 + x_0' \left(X'X\right)^{-1} x_0)\sigma^2.$$

Note:
An additional '1' appears because of ϵ_0. $x_0'\hat{\beta}$ and ϵ_0 are independent.

Then the **prediction interval** for y_0 is given by:

$$\hat{y}_0 \pm t_{1-\alpha/2}\sqrt{1 + x_0' \left(X'X\right)^{-1} x_0}S. \tag{9}$$

y_0 is within (9) with $1 - \alpha$ probability.

Note 2:

The proof of Theorems 4 and 5 are similar to Theorem 2.

Which one should we use, confidence interval or prediction interval?

The prediction interval shows the variability of $\boldsymbol{y_0}$ in usual sense.

The confidence interval is for the expectation of $\boldsymbol{y_0}$. If we have many new observations whose $\boldsymbol{x}$ values are $\boldsymbol{x_0}$, the sample mean of their $\boldsymbol{y}$ value is likely in the confidence interval.

6.3 Prediction and confidence bands

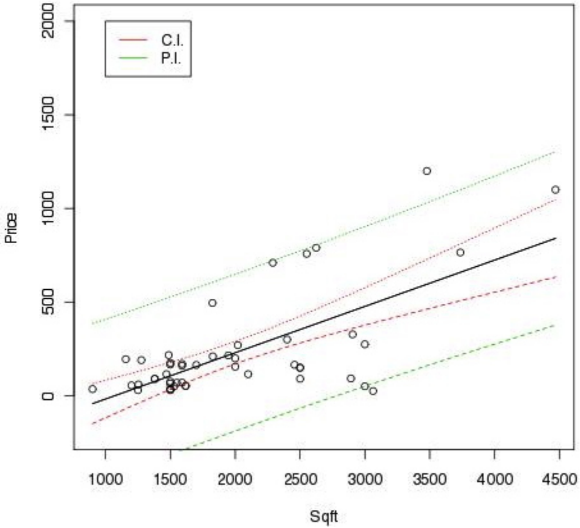
Confidence bands and prediction bands:

Confidence intervals for all possible x_0 is a band. It is called a **confidence band** . A confidence band is the band for the regression line.

Prediction intervals for all possible x_0 is a band. It is called a **prediction band** . A prediction band is the band for observations.

Both bands get wider as x gets farther away from $\bar{x}$.

6.3 Prediction and confidence bands



6.3 Prediction and confidence bands

Sample Code:

```
Ylim <- c(-200, 2000)
Xlim <- range(Sqft)

CI <- predict(LM1, int="c")
PI <- predict(LM1, int="p")

CI <- CI[order(Sqft),]    # CI sorted by Sqft
PI <- PI[order(Sqft),]    # PI sorted by Sqft

jpeg("Condo-CIPI") # output a jpeg file
plot(Sqft, Price, xlim=Xlim, ylim=Ylim)
matlines(sort(Sqft), CI, xlim=Xlim, ylim=Ylim, col=c(1,2,2))
# line plot for data frame of CI
matlines(sort(Sqft), PI, xlim=Xlim, ylim=Ylim, col=c(1,3,3))
legend(1000, 2000, c("C.I.","P.I."), lty=1, col=2:3)
dev.off() # jpeg() ended
```

6.3 Prediction and confidence bands

Exercise 2:

Using "iris" data,

1. regress "Sepal.Length" on "Sepal.Width" for the first 50 observations (i.e., setosa species) by simple linear regression.
2. create a scatter plot with the regression line, confidence and prediction bands.

Exercise 3:

Calculate the confidence and prediction bands for the regression in Exercise 2 by using the formulae (8) and (9). Compare the results to Exercise 2 (they should be the same).

6.4 Correlation

(Pearson's) Correlation coefficient r:

The (Pearson's population) correlation coefficient r between two variables x and y quantifies the strength of linear relationship between x and y.

$$r := \frac{\mathrm{Cov}(x, y)}{\sqrt{\mathrm{Var}(x)}\sqrt{\mathrm{Var}(y)}} = \frac{E\left[(x - Ex)(y - Ey)\right]}{\sqrt{E\left[(x - Ex)^2\right]}\sqrt{E\left[(y - Ey)^2\right]}}$$

6.4 Correlation

Correlation coefficient for sample:
For n observations $(x_1, y_1), \cdots, (x_n, y_n)$, the best estimate for $\hat{r}$ is:

$$\hat{r} = \frac{s_{xy}}{s_x s_y} \tag{10}$$

where

$$\begin{aligned} s_{xy} &:= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1} \\ s_x &:= \text{the sample standard deviation of } x_i\text{'s} \\ &= \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{1}{n-1}\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)} \\ s_y &:= \text{the sample standard deviation of } y_i\text{'s} \end{aligned} \tag{11}$$

s_{xy} is called the **sample covariance** of x and y.

6.4 Correlation

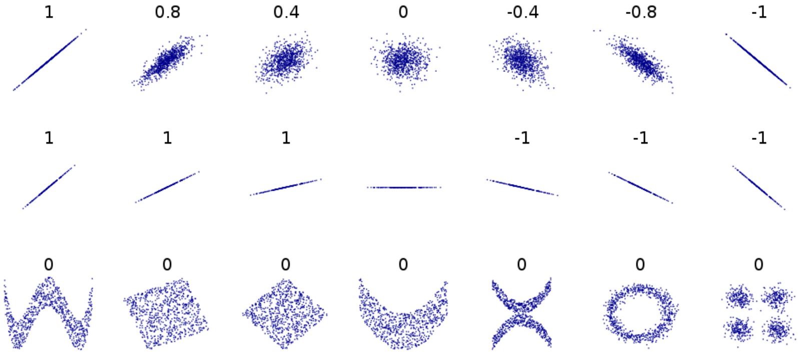
Other expressions:

$$\hat{r} = \frac{\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}\sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \quad (12)$$

$$s_{xy} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{n-1} \quad (13)$$

6.4 Correlation

Graphical Examples:



Source: Wikipedia.

6.4 Correlation

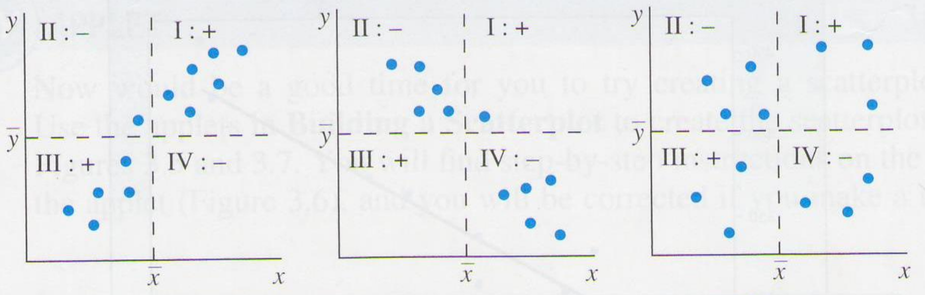
Properties of r (and $\hat{r}$):

- r is always between -1 and 1.
- If $r = 1$, x_i and y_i have 'perfect' positive linear relationship. I.e., all (x_i, y_i)'s are on a straight line with a positive slope.
  - E.g., $(x_1, y_1) = (1, 2), (x_2, y_2) = (-3, -6)$ and $(x_3, y_3) = (2, 4)$.
- If $r > 0$, x_i and y_i have positive linear relationship.
- If $r = 0$, there is no linear relationship.
  - **There may be some non-linear relationship.**
- If $r < 0$, x_i and y_i have negative linear relationship.
- If $r = -1$, x_i and y_i have 'perfect' negative linear relationship. I.e., all (x_i, y_i)'s are on a straight line with a negative slope.
  - E.g., $(x_1, y_1) = (1, -2), (x_2, y_2) = (-3, 6)$ and $(x_3, y_3) = (2, -4)$.

6.4 Correlation

The signs of $(x_i - \bar{x})(y_i - \bar{y})$:

By (12), the sum of $(x_i - \bar{x})(y_i - \bar{y})$ (over all i's) determines the sign of r. One can see that positive $(x_i - \bar{x})(y_i - \bar{y})$'s correspond to a positive relationship.



(a) Positive pattern (b) Negative pattern (c) No pattern

6.4 Correlation

Example:

(x_i, y_i)'s are $(1, 20), (3, 8), (6, 10), (8, 0)$ and $(2, 12)$ ($n = 5$).

Table: Calculation for $\hat{r}$

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1	20	1	400	20
2	3	8	9	64	24
3	6	10	36	100	60
4	8	0	64	0	0
5	2	12	4	144	24
sum (Σ)	20	50	114	708	128
mean	4	10	-	-	-

$$s_x^2 = \frac{114 - 20^2/5}{5-1} = 8.5, \quad s_y^2 = \frac{708 - 50^2/5}{5-1} = 52$$

$$s_{xy} = \frac{128 - 20 \times 50/5}{5-1} = -18, \quad \hat{r} = \frac{-18}{\sqrt{8.5}\sqrt{52}} \approx -0.856$$

Exercise 4:

Recalculate the correlation coefficient $\hat{r}$ above with R in two ways:

1. by "cor" function, and
2. by the expressions (12),

and make a scatter plot for (x_i, y_i) $(i = 1, \cdots, 5)$.

6.4 Correlation

Testing on r:

Using the following theorem, one can test $H_0 : r = r_0$ (usually $r_0 = 0$) against $r \neq r_0$.

Theorem 6 (Fisher's transformation):

Suppose (x_i, y_i) $(i = 1, \cdots, n)$ are IID bivariate normal (*1). When n is large, approximately

$$\frac{1}{2}\log\frac{1+\hat{r}}{1-\hat{r}} \sim N\left(\frac{1}{2}\log\frac{1+r_0}{1-r_0}, \frac{1}{n-3}\right)$$

where **log** is natural logarithm.

(*1) $ux + vy$ is normal for any constants u and v.

6.4 Correlation

Sample code: 3 bedroom condo price in Detroit

```
> cor(Sqft, Price)
[1] 0.6745883
> cor.test(Sqft, Price)

Pearson's product-moment correlation

data:  Sqft and Price
t = 6.1979, df = 46, p-value = 1.461e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4830398 0.8045179
sample estimates:
      cor
0.6745883
```

The r is significantly different from 0 since p-value $< \mathbf{0.05}$.

6.4 Correlation

Exercise 5:

1. Test $\boldsymbol{r = 0}$ between "Sepal.Length" and "Sepal.Width" for the first 50 observations (i.e., setosa species) of the iris data by "cor.test".
2. Do the same test by Fisher transformation.

6.4 Correlation

$\hat{r}$ and regression line:

$\hat{r}$ and $\hat{b}$ (in the regression line $y = \hat{a} + \hat{b}x$) have a correspondence:
$\hat{b} = r \cdot \frac{S_y}{S_x}$.

In other words, $\hat{b} = \hat{r}$ if $S_x = S_y = 1$. If not, $\hat{b}$ is "$\hat{r}$ multiplied by the scale of y (S_y) and divided by the scale of x (S_x)".

6.4 Correlation

Example: 3 bedroom condo price in Detroit

```
> lm(Price ~ Sqft)

Call:
lm(formula = Price ~ Sqft)

Coefficients:
(Intercept)         Sqft
  -265.0250       0.2473

> cor(Sqft, Price) * sd(Price) / sd(Sqft)
[1] 0.2472783
```

6.4 Correlation

Definition:

When there are random variables $X_1, \cdots, X_n$, $X_{(i)}$ $(i = 1, \cdots, n)$ denotes the i-th smallest observation of all, and is called the **i-th order statistic** .

Example:

If $x_1 = 11.5,\ x_2 = 7.9,\ x_3 = 5.4,\ x_4 = 10.1$, then
$x_{(1)} = 5.4,\ x_{(2)} = 7.9,\ x_{(3)} = 10.1,\ x_{(4)} = 11.5$.

Spearman's ρ:

Spearman's $\boldsymbol{\rho}$ is the Pearson's correlation coefficient for the order statistics, which are invariant by monotone transformation.
Properties of $\boldsymbol{\rho}$ is similar to $\boldsymbol{r}$, for instance, $-\mathbf{1} \leq \boldsymbol{\rho} \leq \mathbf{1}$, while there is no simple exact relationship between Pearson's $\boldsymbol{r}$ and Spearman's $\boldsymbol{\rho}$.

6.4 Correlation

Example:

Let (x_i, y_i)'s be $(1, 20), (3, 8), (6, 10), (8, 0)$ and $(2, 12)$ ($n = 5$), Rx_i (Ry_i) be the rank of x_i (y_i, respectively).

Table: Calculation for $\hat{\rho}$

i	x_i	y_i	Rx_i	Ry_i	Rx_i^2	Ry_i^2	$Rx_i Ry_i$
1	1	20	1	5	1	25	5
2	3	8	3	2	9	4	6
3	6	10	4	3	16	9	12
4	8	0	5	1	25	1	5
5	2	12	2	4	4	16	8
sum (Σ)	20	50	15	15	55	55	36
mean	4	10	3	3	-	-	-

$$s_{Rx}^2 = \frac{55 - 15^2/5}{5-1} = 2.5, \quad s_{Ry}^2 = \frac{55 - 15^2/5}{5-1} = 2.5,$$

$$s_{RxRy} = \frac{36 - 15 \times 15/5}{5-1} = -2.25, \quad \hat{\rho} = \frac{-2.25}{\sqrt{2.5}\sqrt{2.5}} \approx -0.9.$$

6.4 Correlation

ρ is invariant by monotone transformation:

```
> X <- c(1,3,6,8,2)
> Y <- c(20,8,10,0,12)
>
> cor(X,Y) # Pearson's r
[1] -0.8561727
> cor(X,Y, method="spearman") # Spearman's rho
[1] -0.9

> cor(sqrt(X),Y^0.4) # r for transformed X and Y
[1] -0.8268442
> cor(sqrt(X),Y^0.4,method="spearman")  # rho for transformed X and Y
[1] -0.9
```

6.4 Correlation

Testing:

Testing a hypothesis $H_0 : \rho = \rho_0$ is done by simulating the distribution of $\hat{\rho}$ or relying on a similar asymptotic result as Pearson's correlation.

```
> cor.test(X, Y, method="spearman")

Spearman's rank correlation rho

data:  X and Y
S = 38, p-value = 0.08333
alternative hypothesis: true rho is not equal to 0
sample estimates:
 rho
-0.9
```

$H_0 : \rho = 0$ is not rejected with $\alpha = 0.05$ since p-value > 0.05. In general, it is difficult to reject H_0 when the sample size is very small.

6.4 Correlation

Kendall's τ:

Kendall's τ is a different version of correlation coefficient which is invariant by monotone transformation.

$$\tau := \frac{\#\left[(x_i - x_j)(y_i - y_j) > 0\right] - \#\left[(x_i - x_j)(y_i - y_j) < 0\right]}{n(n-1)/2}$$

where $\#[\bullet]$ means the number of observations which satisfy $\bullet$ among all $1 \leq i < j \leq n$.

$-1 \leq \tau \leq 1$ is always satisfied, similarly to r and ρ.

6.4 Correlation

Example:

Let (x_i, y_i)'s be $(1, 20), (2, 12), (3, 8), (6, 10), (8, 0)$ $(n = 5)$.

$$\begin{aligned}
\#\left[(x_i - x_j)(y_i - y_j) > 0\right] &= 1 \quad (3, 8) \text{ vs } (6, 10) \text{ only} \\
\#\left[(x_i - x_j)(y_i - y_j) < 0\right] &= 9 \\
\frac{n(n-1)}{2} &= 10 \\
\hat{\tau} &= \frac{1-9}{10} = -0.8.
\end{aligned}$$

6.4 Correlation

Sample Code:

```
> cor(X,Y, method="kendall")
[1] -0.8
> cor(sqrt(X),Y, method="kendall") # invariant by monotone transform
[1] -0.8
```

Testing on τ:

Test on τ is done either by simulating the distribution of $\hat{\tau}$ under the null, or by using normal approximation of the distribution.

6.4 Correlation

Exercise 6:

1. Calculate Spearman's ρ and Kendall's τ between "Sepal.Length" and "Sepal.Width" for the first 50 observations (i.e., setosa species) of the iris data by 'cor' function.
2. Recalculate Kendall's τ from the definition for the same data.
3. Test $\rho = 0$ and $\tau = 0$ with "cor.test" function.