Open Review Session

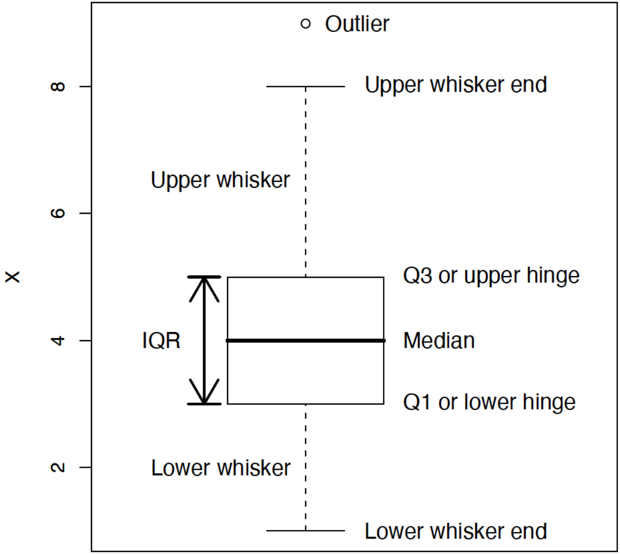
Roberta De Vito

BROWN Public Health

1

Asymmetry of a distribution

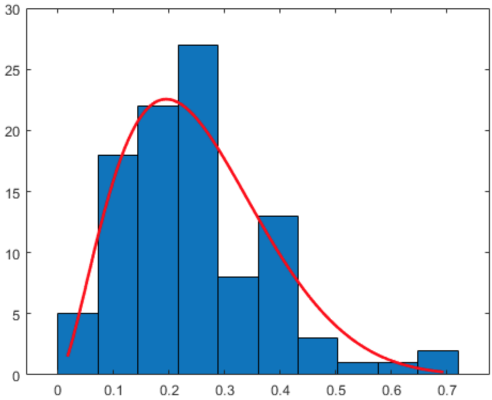
Boxplots



Outlier
Upper whisker end
Upper whisker
Q3 or upper hinge
IQR
Median
Q1 or lower hinge
Lower whisker
Lower whisker end
x
8
6
4
2

Mean Median and Mode

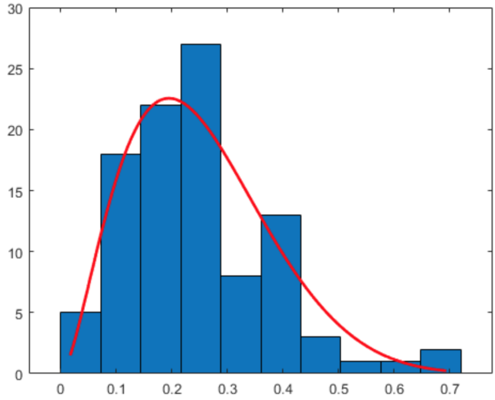
Questions



Do you think that this distribution is

1. Positive skew
2. Symmetric
3. Negative skew

Questions

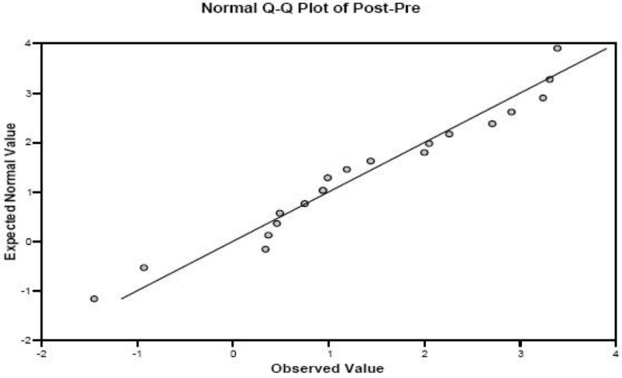


Can you write down the order of the mode, mean and median?

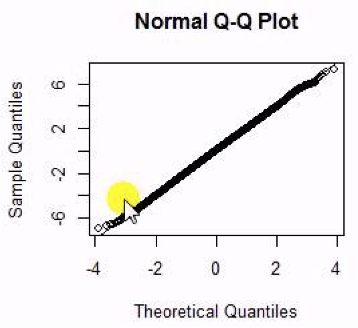
What is a QQ plot?

1. Let $\varepsilon_{(1)}, \ldots, \varepsilon_{(n)}$ be the ordered residuals with $\varepsilon_{(1)} \leq \varepsilon_{(2)} \leq \ldots \leq \varepsilon_{(n)}$.
2. Assume the ε are standardized by subtracting mean and dividing by standard error. This ensures that they have mean zero and variance one. Then, the distribution to compare to is a $\mathcal{N}(0, 1)$.
3. If the ε-s come from a $\mathcal{N}(0, 1)$ distribution, we expect $\varepsilon_{(k)}$ to be approximately equal to the $\frac{k}{n}$-th quantile of the $\mathcal{N}(0, 1)$.
4. A qq-plot plots the observed quantiles vs the theoretical quantiles. If points fall on a straight line, indication of the sample coming from a normal distribution.

QQ plot in practice



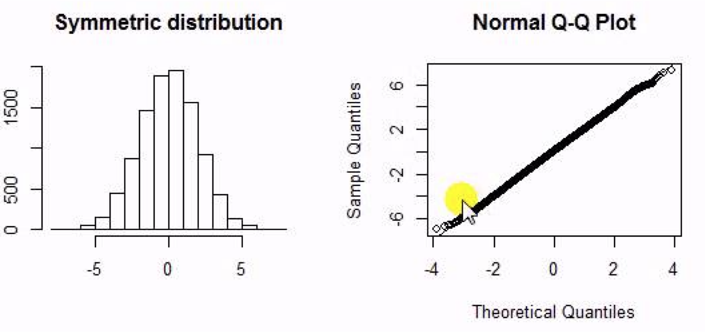
Question I QQplot



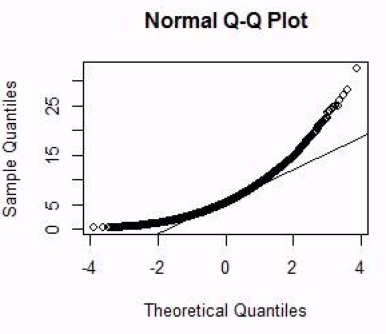
Do you think that this distribution is

1. Positive skew
2. Symmetric
3. Negative skew

Question I QQplot



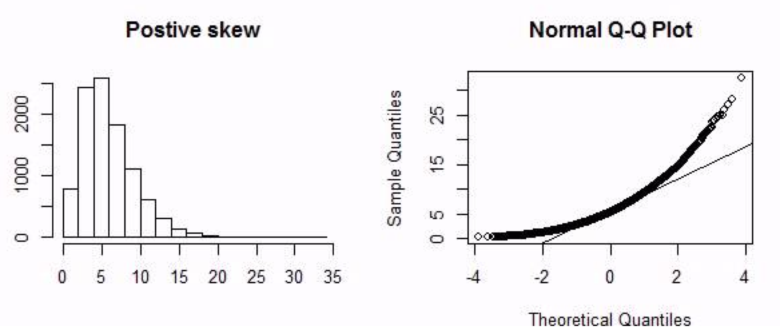
Question II QQplot



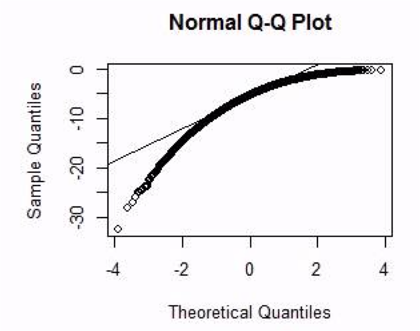
Do you think that this distribution is

1. Positive skew
2. Symmetric
3. Negative skew

Question II QQplot



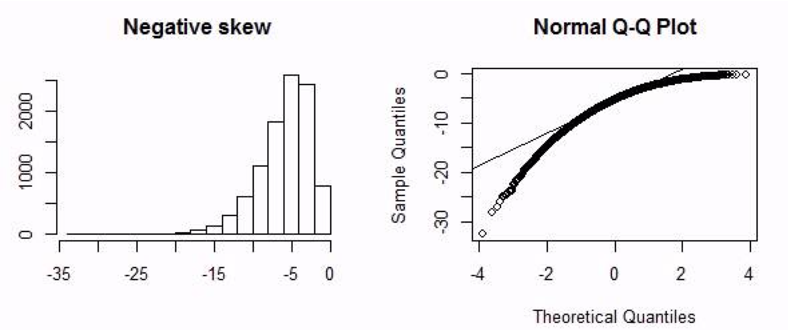
Question III QQplot



Do you think that this distribution is

1. Positive skew
2. Symmetric
3. Negative skew

Question III QQplot



2

Linear regression model

The linear regression

- $y_i = f(x_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$
- $f(x_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$

$$\Downarrow$$

Matrix form: $Y = X\beta + \epsilon$

Model Assumption

1. $E[Y_i|X_i] = \beta_0 + \beta_1 X_i$
2. $\epsilon \sim N(0, \sigma^2)$
3. Error term is independent of (uncorrelated with) covariate(s)

$$Corr(X, \epsilon) = 0$$

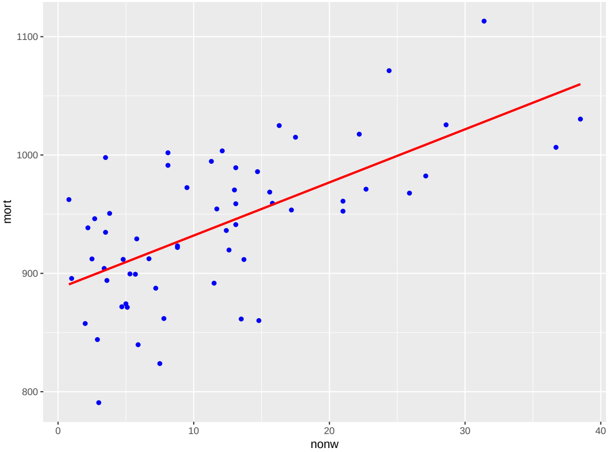
4. Variance of error term is same, regardless of value of x (homoscedasticity)

$$Var(\epsilon) = \sigma^2$$

Example: Pollution data set

`ID`	code for the identification of the sample
`OVR65`	% of 1960 SMSA population aged 65 or older
`EDUC`	Median school years completed by those over 22
`HOUS`	% of housing units with all facilities
`DENS`	Population per sq. mile in urbanized areas, 1960
`NONW`	% non-white population in urbanized areas, 1960
`WWDRK`	% employed in white collar occupations
`POOR`	% of families with income < 3000
`HC`	Relative hydrocarbon pollution potential
`NOX`	Same for nitric oxides
`SO2`	Same for sulphur dioxide
`HUMID`	Annual average % relative humidity at 1pm
`MORT`	Total age-adjusted mortality rate per 100,000
`PREC`	Average annual precipitation in inches

How do we find regression line that fits best?



Example: Pollution data set

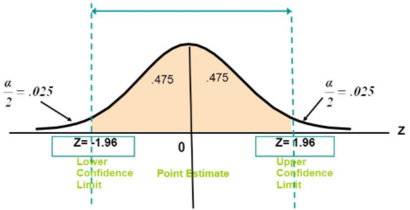
ID	code for the identification of the sample
OVR65	% of 1960 SMSA population aged 65 or older
EDUC	Median school years completed by those over 22
HOUS	% of housing units with all facilities
DENS	Population per sq. mile in urbanized areas, 1960
NONW	% non-white population in urbanized areas, 1960
WWDRK	% employed in white collar occupations
POOR	% of families with income < 3000
HC	Relative hydrocarbon pollution potential
NOX	Same for nitric oxides
SO2	Same for sulphur dioxide
HUMID	Annual average % relative humidity at 1pm
MORT	Total age-adjusted mortality rate per 100,000
PREC	Average annual precipitation in inches

The lm function in R: what are we looking?

```
Call:
lm(formula = mort ~ nonw + so2 + educ + nonw)

Residuals:
    Min      1Q  Median      3Q     Max 
-94.201 -19.410   1.294  16.537  92.986 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1156.06487   71.68018  16.128  < 2e-16 ***
nonw           3.70485    0.58615   6.321 4.55e-08 ***
so2            0.25699    0.08298   3.097 0.003054 ** 
educ         -24.92413    6.28208  -3.967 0.000209 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39.02 on 56 degrees of freedom
Multiple R-squared:  0.6266,	Adjusted R-squared:  0.6066 
F-statistic: 31.33 on 3 and 56 DF,  p-value: 5.063e-12
```

Inference

- $H_0 : \beta_1 = 0$
- 95% confidence intervals



- R^2
- F-statistics: Does the model fit better than a model with only an intercept?

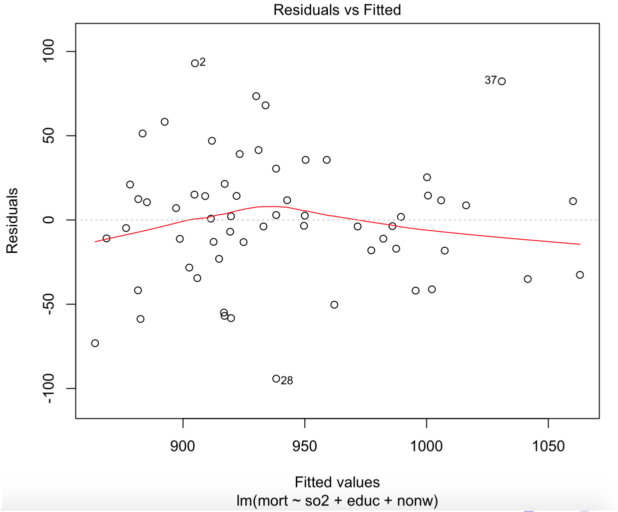
Diagnostics and Assumption Checking

1. is the linear relationship a good assumption?
2. is the error term variance constant?
3. are the error term normally distributed?
4. are there any outliers?
5. do we repeat some information?

1. Non-linearity of the data

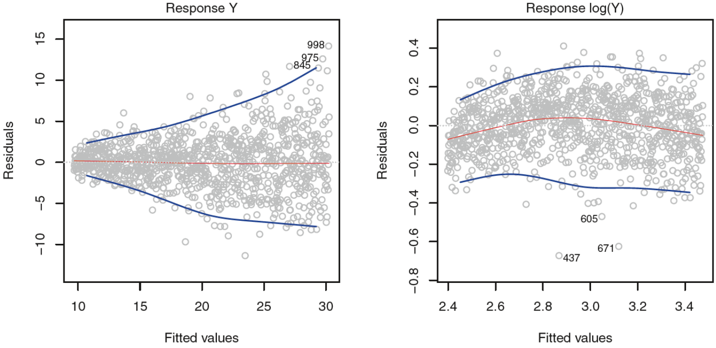
Residual plot of fitted values vs. residuals should

- have no discernible pattern
- be scattered evenly around 0

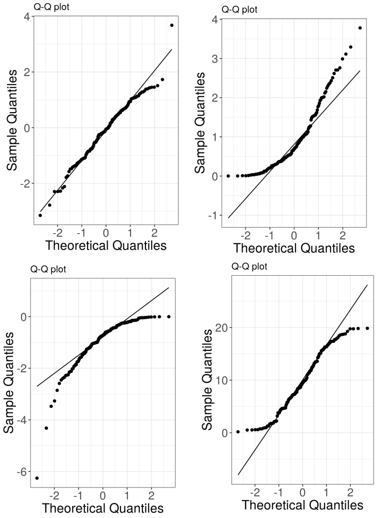


2. Non-constant Variance of Error Terms: Heteroscedasticity

- Patterns might indicate wrong form of model variable
- Funnel shape in the residual plot: transform Y

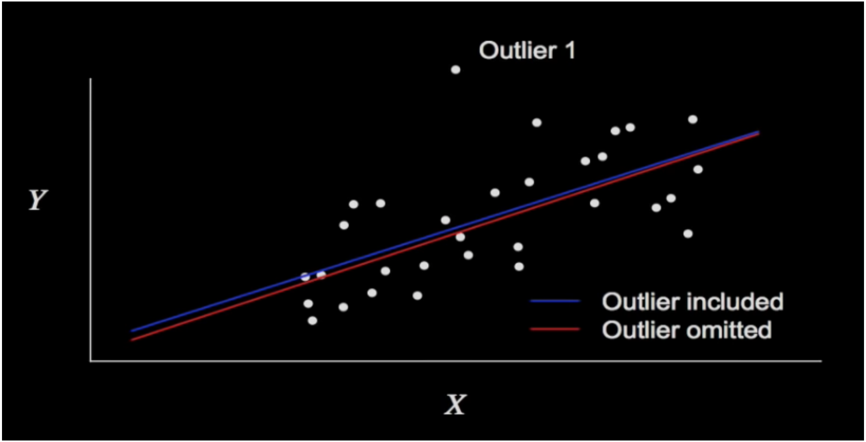


3. Normal distribution



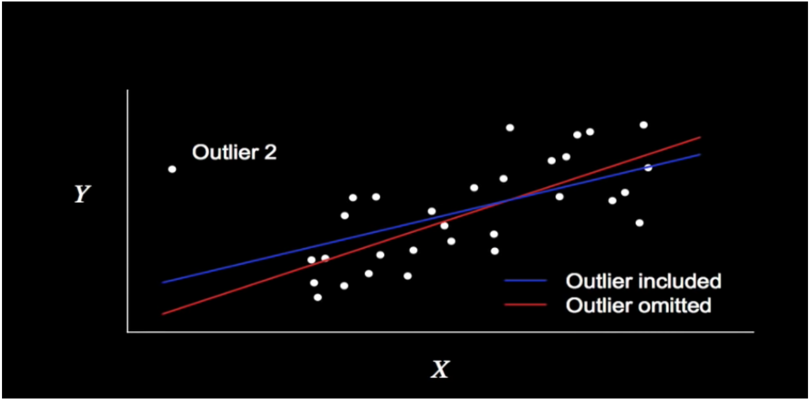
4. Outliers

Outliers for Y



4. Outliers

Outliers for X (High Leverage)



5. Collinearity

- Collinearity refers to when the predictors are highly correlated.
- Repetition of information
- Leads to increased standard errors of the regression coefficients $\rightarrow$ fail to reject $H_0 : \beta_j = 0$
- take a look at the correlation of two covaiates

3 Logistic regression model

The logistic function

Mathematical model

- the outcome

$$y_i = \begin{cases} 1 & \text{if household i switched to a new well} \\ 0 & \text{if household i continued using its own well} \end{cases}$$

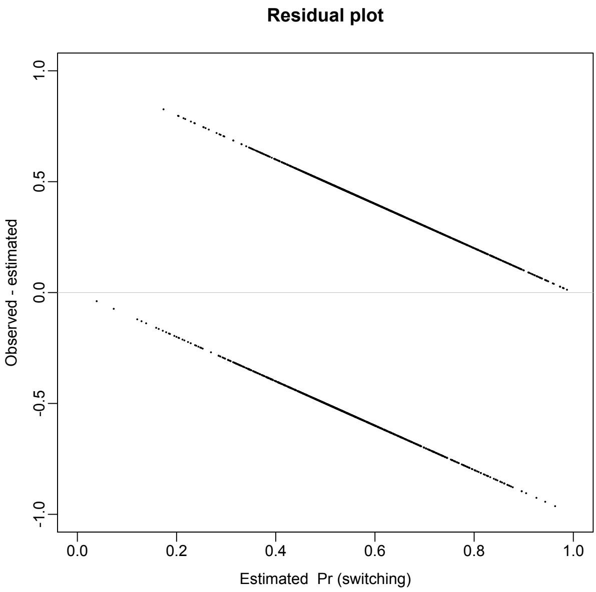
- The distance (in meters) to the closest known safe well
- The arsenic level of respondent's well
- Whether any members of the household are active in community organizations
- The education level of the head of household.

Logistic regression model with arsenic

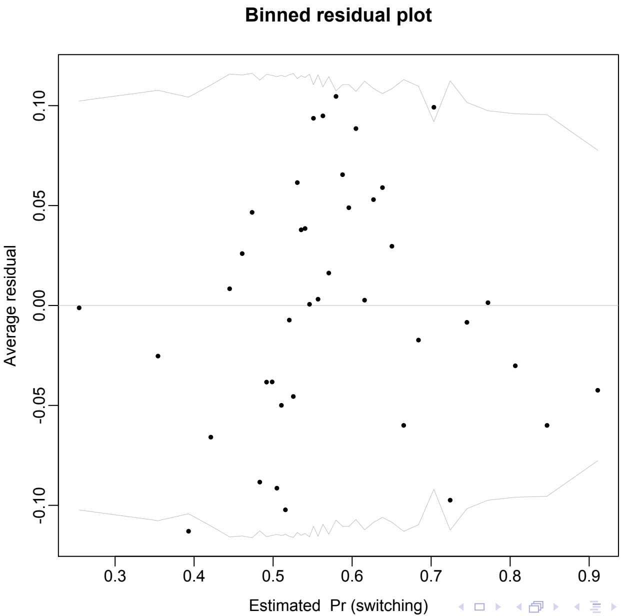
$$\text{Percent Change in the Odds} = (e^{\beta_1} - 1) \times 100$$

```
> summary(fit.3)

Call:
glm(formula = switch ~ dist100 + arsenic, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6351  -1.2139   0.7786   1.0702   1.7085

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.002749   0.079448   0.035    0.972
dist100     -0.896644   0.104347  -8.593   <2e-16 ***
arsenic      0.460775   0.041385  11.134   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4118.1  on 3019  degrees of freedom
Residual deviance: 3930.7  on 3017  degrees of freedom
AIC: 3936.7

Number of Fisher Scoring iterations: 4
```

Evaluating, Checking: Residuals



Evaluating, Checking: Binned Residuals



The Deviance in our data set

Model	Null Deviance	Residual Deviance
dist	4118.1	4076.2
dist100	4118.1	4076.2
arsenic	4118.1	3930.7
interaction	4118.1	3927.6
center	4118.1	3927.6
social	4118.1	3905.4
educ	4118.1	3907.9
log	4118.1	3863.1

4

Generalized Linear Model

A generalized linear model

1. $y = (y_1, \ldots, y_n)$
2. X and coefficients β
3. a link function g so that $\hat{y} = g^{-1}(X\beta)$
4. a data distribution
5. other parameters, for example?

Link Distribution

Family	Default Link Function
binomial	(link = "logit")
gaussian	(link = "identity")
Gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

5 Select the best model

Method for selecting the best model

- AIC
- BIC
- R^2

Forward Stepwise Selection

Aim: choose $K \leq P$

$$y_i = \beta_0 + \epsilon_i$$

Forward Stepwise Selection

Aim: choose $K \leq P$

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

Forward Stepwise Selection

Aim: choose $K \leq P$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

Forward Stepwise Selection

Aim: choose $K \leq P$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$$

Backward Stepwise Selection

Aim: choose $K \leq P$

$$y_i = \beta_0 + \cdots + \beta_p x_{ip} + \epsilon_i$$

Backward Stepwise Selection

Aim: choose $K \leq P$

$$y_i = \beta_0 + \cdots + \beta_{p-1} x_{i(p-1)} \epsilon_i$$

Backward Stepwise Selection

Aim: choose $K \leq P$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$$

6 Multi-level Model

Multi-level Model 1

$$y_i \sim N(\alpha_{j[i]} + \beta x_i, \sigma_y^2)$$

$$\alpha_j \sim N(\mu_0, \sigma_\alpha^2)$$

```
lmer(formula = y ~ x + (1 | county))
            coef.est coef.se
(Intercept)  1.46     0.05
x           -0.69     0.07
Error terms:
 Groups   Name        Std.Dev.
 county   (Intercept) 0.33
 Residual             0.76
# of obs: 919, groups: county, 85
deviance = 2163.7
```

```
      (Intercept)
  1        -0.27
  2        -0.53
  3         0.02
  . . .
  85       -0.08
```

Multi-level Model 2

$$y_i \sim N(\alpha_{j[i]} + \beta x_i, \sigma_y^2)$$

$$\alpha_j \sim N(\gamma_0 + \gamma_1 u_j, \sigma_\alpha^2)$$

```
lmer(formula = y ~ x + u.full + (1 | county))
            coef.est coef.se
(Intercept)  1.47     0.04
x           -0.67     0.07
u.full       0.72     0.09
Error terms:
 Groups   Name        Std.Dev.
 county   (Intercept) 0.16
 Residual             0.76
# of obs: 919, groups: county, 85
deviance = 2122.9
```

Varying intercepts and slopes

$$y_i = N(\alpha_{j[i]} + \beta_{j[i]} x_i, \sigma_y^2)$$

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right)$$

Varying intercepts and slopes: R output

```
M3 <- lmer (y ~ x + (1 + x | county))
display (M3)
```

which yields

```
lmer(formula = y ~ x + (1 + x | county))
             coef.est coef.se
(Intercept)   1.46     0.05
x            -0.68     0.09
Error terms:
 Groups    Name        Std.Dev. Corr
 county    (Intercept) 0.35
           x           0.34     -0.34

    Residual             0.75
   # of obs: 919, groups: county, 85
   deviance = 2161.1
```

Varying intercepts and slopes: R output

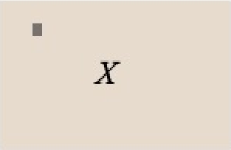
```
fixef (M3)

(Intercept)            x
       1.46        -0.68
```

Varying intercepts and slopes: R output

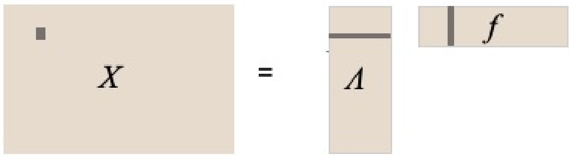
```
ranef (M3)

   (Intercept)      x
1        -0.32   0.14
2        -0.53  -0.09
3         0.01   0.01
. . .
85       -0.08   0.03
```

7 Unsupervised Learning

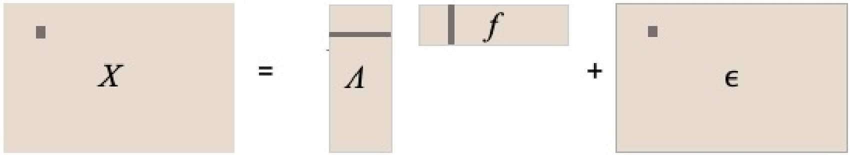
The matrix form



The matrix form



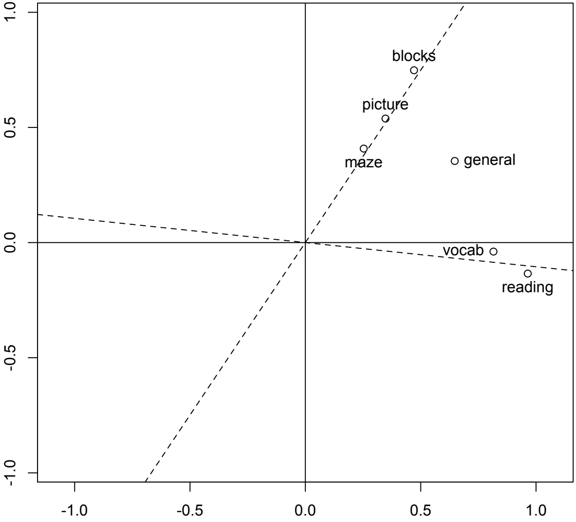
The matrix form



Factor analysis

Property of the model: Identifiability

Plotting the Factor Model



Plotting the Factor Model with Varimax Rotation