Outline

regression model

We have learned various regression model:

$$Y_i \overset{\mathcal{D}}{\sim} \mathcal{F}(\mu_i, \phi) \quad \text{independently},$$
$$g(\mu_i) = \beta_0 + \beta_1 X_{i,1} + \ldots + \beta_p X_{i,p}$$

where $\mathbb{E}(Y_i) = \mu_i$ and $\text{Var}(Y_i) = \phi V(\mu_i)$ and $V(\cdot)$ is the variance function.

Type	Response distribution	Variance function	link
Gaussian linear regression	$\mathcal{F} = N(\mu_i, \sigma^2)$	$V(\mu_i) = 1$	$g(\mu_i) = \mu_i$
Bernoulli logistic regression	$\mathcal{F} = \text{Ber}(\mu_i)$	$V(\mu_i) = \mu_i(1-\mu_i)$	$g(\mu_i) = \log(\frac{\mu_i}{1-\mu_i})$
Binomial regression	$\mathcal{F} = \text{Bin}(n_i, \mu_i)$	$V(\mu_i) = n_i\mu_i(1-\mu_i)$	$g(\mu_i) = \log(\frac{\mu_i}{1-\mu_i})$

- What if the independent assumption is violated?
- Example: repeated responses Y_{ij}, where i indexes the individual and j indexes repeations.
- Example: time-involving response, Y_{it}, where i indexes the individual and t indexes the time.

Correlated observations

- Example: repeated response, Y_{ij}, where i indexes the individual and j indexes repeated observations.
- Let $\boldsymbol{Y} = (Y_{1,1}, \ldots, Y_{1,m}, Y_{2,1}, \ldots, Y_{2,m}, \ldots, Y_{n,1}, \ldots, Y_{n,m})^T$, total sample size nm.
- Linear regression with correlated observations:

$$\boldsymbol{Y} \sim N(\mu_i, \sigma^2 \Sigma)$$
$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \ldots + \beta_p X_{i,p}$$

- $Cov(Y_{ij}, Y_{ij'} | \boldsymbol{X}) \neq Cov(Y_{ij}, Y_{i'j'} | \boldsymbol{X})$. Possible choice of $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho & \cdots & \rho & 0 & 0 & \cdots & 0 & 0 & \cdots \\ \rho & 1 & \cdots & \rho & 0 & 0 & \cdots & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 & \cdots \\ 0 & 0 & \cdots & 0 & 1 & \rho & \cdots & \rho & 0 & \cdots \\ 0 & 0 & \cdots & 0 & \rho & 1 & \cdots & \rho & 0 & \cdots \\ 0 & 0 & \cdots & 0 & \rho & \rho & \cdots & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Outline

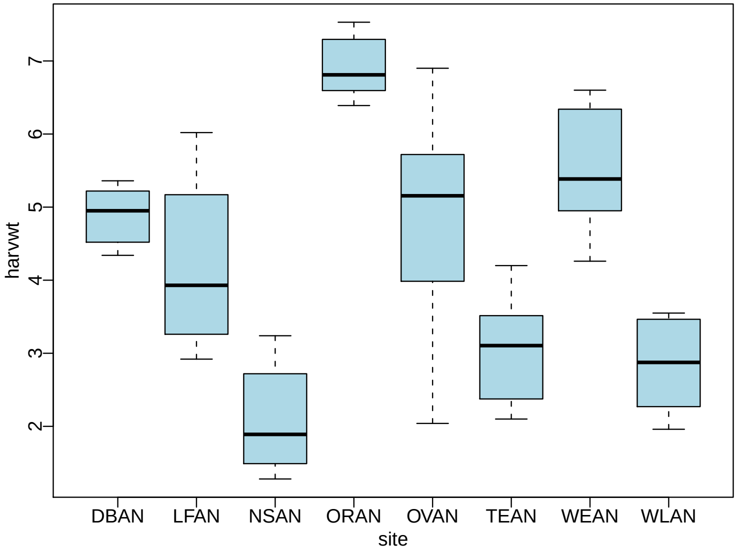
Corn example

- Subset of a larger data set on corn grown on the island Antigua.
- Response: harvest weight (`harvwt`) per plot, 64 plots in total.
- 8 sites, with 8 separate plots within each site where the corn is grown under the same treatment conditions.
- Does the site have an effect on the harvest weight?

```
> corn = read.table("corn.txt", header = T)
> summary(corn)
      site                    harvwt
 DBAN    : 8       Min.   :1.280
 LFAN    : 8       1st Qu.:2.935
 NSAN    : 8       Median :4.300
 ORAN    : 8       Mean   :4.292
 OVAN    : 8       3rd Qu.:5.442
 TEAN    : 8       Max.   :7.530
 WEAN    : 8
 WLAN    : 8
```

Corn example

```
> plot(harvwt ~ site,data=corn,pch=16,col="lightblue")
```



Outline

First analysis: Standard one-way ANOVA

Standard regression: we could use one-way ANOVA with 8 fixed parameters for the site mean weights and a single plot-level source of error.

Model:

$$
\begin{aligned}
y_i &= \beta_1 + \beta_2 \cdot \mathbb{1}_{\text{site 2}} + \cdots + \beta_8 \cdot \mathbb{1}_{\text{site 8}} + e_i \\
&= \mu + \alpha_1 \cdot \mathbb{1}_{\text{site 1}} + \alpha_2 \cdot \mathbb{1}_{\text{site 2}} \cdots + \alpha_8 \cdot \mathbb{1}_{\text{site 8}} + e_i
\end{aligned}
$$

where $i = 1, \ldots, 64$ indexes observations and $e_i \sim \text{iid}\mathcal{N}(0, \sigma^2)$.

Fixed, unknown parameters:

- $\beta_j,\ j = 1, \ldots, 8$:
  intercept $\beta_1 =$ mean corn yield for site 1, and adjustments $\beta_2, \ldots, \beta_8$ for other sites,
- Or equivalently, μ and $\alpha_j,\ j = 1, \ldots, 8$:
  intercept $\mu =$ grand mean over all sites, and adjustements $\alpha_1, \ldots, \alpha_8$ around this overall mean for each site, with the constraint $\sum_{j=1}^{8} \alpha_j = 0$.
- σ^2

Standard regression model

By default, `R` uses the β formula, with the first site as the reference level.

```
> corn.lm = lm(harvwt ~ site, data = corn)
> summary(corn.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.090      0.309    6.76  8.6e-09 ***
siteWLAN       0.751      0.437    1.72    0.091 .
siteTEAN       0.946      0.437    2.16    0.035 *
siteLFAN       2.118      0.437    4.84  1.0e-05 ***
siteOVAN       2.743      0.437    6.27  5.5e-08 ***
siteDBAN       2.795      0.437    6.39  3.5e-08 ***
siteWEAN       3.436      0.437    7.86  1.3e-10 ***
siteORAN       4.825      0.437   11.03  1.1e-15 ***

Residual standard error: 0.88 on 56 degrees of freedom
Multiple R-squared: 0.767,      Adjusted R-squared: 0.737
F-statistic: 26.3 on 7 and 56 DF,  p-value: 1.55e-15
```

Standard regression model

We may request the formula with an overall mean and α adjustments that sum up to 0:

```
> options(contrasts=c("contr.sum", "contr.poly"))
> corn.lm = lm(harvwt ~ site, data = corn)
> summary(corn.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.2917     0.1093   39.26  < 2e-16 ***
site1        -2.2017     0.2892   -7.61  3.4e-10 ***
site2        -1.4505     0.2892   -5.02  5.7e-06 ***
site3        -1.2555     0.2892   -4.34  6.0e-05 ***
site4        -0.0842     0.2892   -0.29    0.772
site5         0.5408     0.2892    1.87    0.067 .
site6         0.5933     0.2892    2.05    0.045 *
site7         1.2345     0.2892    4.27  7.7e-05 ***

Residual standard error: 0.875 on 56 degrees of freedom
Multiple R-squared: 0.767,      Adjusted R-squared: 0.737
F-statistic: 26.3 on 7 and 56 DF,  p-value: 1.55e-15
```

Site means

```
> means      = with(corn, tapply(harvwt, site, mean))
> means
  NSAN   WLAN   TEAN   LFAN   OVAN   DBAN   WEAN   ORAN
2.0900 2.8412 3.0362 4.2075 4.8325 4.8850 5.5262 6.9150

> mean(means)
[1] 4.291719
```

The adjustment α_8 for the last site was not given, but it has to be $-(-2.2017 - 1.4505 + \cdots + 1.2345)$.

Outline

Second analysis: random effect model

or one-way ANOVA with random effects.

We can consider the 8 sites as part of a larger population of sites across the island, and consider their mean harvest weights as random from some normal distribution.

Second analysis: random effect model

Model:

$$y_i = \mu + \alpha_{j[i]} + e_i$$

where $j[i] = 1, \ldots, 8$ indicates which of the 8 sites contains the ith observation.

Two levels of variation:

- at the plot level: $e_i \sim$ iid $\mathcal{N}(0, \sigma^2)$.
- at the site level: $\alpha_j \sim$ iid $\mathcal{N}(0, \sigma^2_\alpha)$ are random effects for the sites, $j = 1, \ldots, 8$.

Fixed, unknown parameters:

- μ_α: overall mean corn yield over the entire population of plots and sites.
- σ^2_α: variance of the mean sites' corn yield over the population of sites.
- and σ^2: variance of the plot corn yield over the population of plots within the same site.

- Random effect model:

$$y_i = \mu + \alpha_{j[i]} + e_i$$

where $\alpha_j \sim$ iid $N(0, \sigma_\alpha^2)$ and $e_i \sim$ iid $N(0, \sigma_e^2)$, $i = 1, \ldots, 64$, $j[i] = 1, \ldots, 8$ indexes the number of sites.

- Equivalent interpretation:

$$\mathbb{E}(\boldsymbol{Y}) = \mu 1,$$
$$\text{Var}(\boldsymbol{Y}) = (\sigma_e^2 + \sigma_a^2)\boldsymbol{\Sigma},$$

where $\rho = \frac{\sigma_\alpha^2}{\sigma_e^2 + \sigma_a^2}$ and $\boldsymbol{\Sigma}$ equals

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho & \cdots & \rho & 0 & 0 & \cdots & 0 & 0 & \cdots \\ \rho & 1 & \cdots & \rho & 0 & 0 & \cdots & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 & \cdots \\ 0 & 0 & \cdots & 0 & 1 & \rho & \cdots & \rho & 0 & \cdots \\ 0 & 0 & \cdots & 0 & \rho & 1 & \cdots & \rho & 0 & \cdots \\ 0 & 0 & \cdots & 0 & \rho & \rho & \cdots & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Using `lmer` in `R`

`lme4`: most recently developed `R` package for fitting linear models with random effects. To install it:

```
>install.packages("lme4")
```

`lmer`: function to use instead of `lm` (Linear Mixed Effects in R). To use it, first load the package, once per session:

```
>library(lme4)
```

The model formula with a random effect in `lmer` differs from `lm`: need to include a term of the form `(pred | group)` where

- `group` is the variable defining the groups within which the random effect applies.
- `pred` defines a model matrix (often just an intercept 1): each group is to have its own coefficients for this regression model.
- `lme4` can also be used for fitting random-effect non-Gaussian models (e.g. Bernoulli logistic regression, or Poisson regression).

Random effect model with `lmer`

```
> corn.lmer = lmer(harvwt ~ (1 | site), data = corn)
> summary(corn.lmer)

Linear mixed model fit by REML

Formula: harvwt ~ (1 | site)
   Data: corn
 AIC BIC logLik deviance REMLdev
 195 201  -94.5      190     189

Random effects:
 Groups   Name        Variance Std.Dev.
 site     (Intercept) 2.417    1.555
 Residual             0.765    0.875
Number of obs: 64, groups: site, 8

Fixed effects:
            Estimate Std. Error t value
(Intercept)     4.29       0.56    7.66
```

Outline

Comparing models

- Same residual $\hat{\sigma} = 0.87$ ($\hat{\sigma}^2 = 0.765$): plot-level variation
- Same overall mean value: 4.29

```
> means      = with(corn, tapply(harvwt, site, mean))
> means
  NSAN   WLAN   TEAN   LFAN   OVAN   DBAN   WEAN   ORAN
2.0900 2.8412 3.0362 4.2075 4.8325 4.8850 5.5262 6.9150
> mean(means)
[1] 4.291719
```

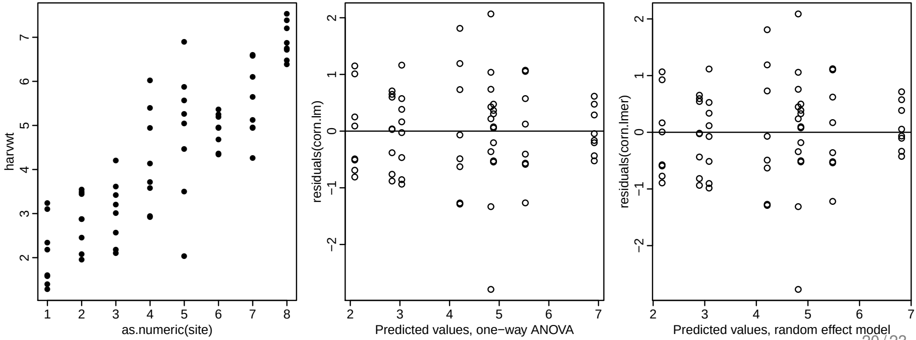
- Estimated variance of observations: $1.55^2 + 0.87^2$
  Covariance of observations at the same site: 1.55^2
  Square correlation of observations at the same site:
  $1.55^2/(1.55^2 + 0.87^2) = 0.76$
- This is close to $R^2 = 0.77$ in standard one-way ANOVA, which corresponds to how much variance is explained by the site-level variability.

Residual plot

```
layout(matrix(1:3, 1,3))
plot(harvwt ~ as.numeric(site),data=corn,pch=16) # plot data first

plot(residuals(corn.lm) ~ fitted(corn.lm),        #residual plot from
     xlab="Predicted values, one-way ANOVA")      #standard 1-way ANOVA
abline(h=0)

plot(residuals(corn.lmer) ~ fitted(corn.lmer),   #from random effect
    xlab="Predicted values, random effect model")            # model
abline(h=0)
```



Motivations for random and mixed effect models

Random effects are used to include sources of variation at more than one level. Examples:

- Repeated measures — when a single individual is measured multiple times, it is often appropriate to model two levels of variation, one for individuals and one for measurements.
- Split-plot designs — in agricultural or ecological studies, it is often the case that sites are broken into plots and possibly subplots. Variables can be measured at the site, plot, subplot, or individual measurement level.
- Also appropriate for non-nested variables. For example, measurements could be clustered by year and by site if a single site is measured over multiple years.

Motivations for random effect models

- We have discussed random effect for Gaussian response; random effect model also exists for non-Gaussian responses (e.g. Bernoulli, Poisson regression with correlated responses.)
- Accounting for both individual and group level variation in estimating group-level effects.
- Modeling individual level regression coefficients.
- Estimation of effects for subgroups.

We can have covariates and error associated with the plot level and separate covariates and error associated with the site level.

When is it worth fitting mixed effect models?

- If there are few groups, (or group size is small?) there may not be much data to estimate random effects and there is little to gain.
- The complexity of mixed-effect models is greater than classical fixed-effect models. The added complexity is often worthwhile, but perhaps not when there are only a small number (say less than five) individuals in a group.