**WEEK I:**

* Why do we Fit models to data
* To estimate distribution properties like means, variance etc.
* Summarize the relationship between variables and predict them.
* Example: Test performance and age: **1.** Estimating marginal mean of performance across all ages **2.** Estimating the mean conditional on age.
* **Quantiles:** Divide the data into equal groups. 50% quantile divides the data into 2 equal groups, 75% quantile divides data into 4 equal groups 3 below and 1 above.
* **Q-Q plot:** Plots the quantile of given quantity to quantiles of any distribution and then we visualise the correlation using a straight line. If points lie on the straight line it means they are distributed as per the distribution considered.

**Types of variables:**

* Categorical variable
* Continuos variable
* Dependent and independent variables

**Different study design generate different data:**

* Simple random samples
* Clustered samples
* Longitudinal data

**Objectives of model fitting:**

* Inferences about relationship between variables
* Forecasting future outcomes.
* Plotting predictions and predicting uncertainty
* *model = sm.OLS.from\_formula("BPXSY1 ~ RIDAGEYR + RIAGENDRx", data=da)*
* *res = model.fit()*
* *print(res.summary())*
* *The above code is creating a multiple linear regression where the target variable is BPXSY1 and the two predictor variables are RIDAGEYR and RIAGENDRx.*

**Ordinary least squares, Generalized linear model, Generalize estimate equations, Multilevel models** are curve fitting methods used for creating models

* ***Ordinary least squares***: Plot of slope of regression line v/s sum of squared distance is plotted and point with zero slope is found

**WEEK II:**

**Linear regression:**

* Develop a model to predict mean cartwheel distance for a population: Will height and if they completed the cartwheel affect this mean. Linear regression of a scattered data gives an estimate of mean, in our case mean cartwheel distance for a particular height.
* We get a test statistic, p value for the slope of regression line and when we plot 95% confidence interval for the regression line, the min interval is at specific value considered

**Causation and correlation:**

* Causation indicates occurrence of one variable due to another, Correlation indicates size and direction of relationship between two variables.

**Prediction interval** gives the range where future observations might fall, **confidence interval** shows the range associated with population mean.

**Logistic regression:**

* To check if the probability of completing the cart wheel depends on age, instead of linear regression, we use logit function. Linear regression results in probability above 1 which is not feasible.
* Hence, instead of y=ax+b
  + - * + logit(y) = ax + b
* logit(y) = y/(1-y)
* p/(1-p) is the odds
* **Loss function**
* **Hyperparameters**
* Residual plots can be used to check correctness of the model built but, in logistic regression our values are only 0 and 1 so it might not be helpful.
* Python’s confidence interval is for 95% confidence

**WEEK III:**

**Models for dependent data:**

* Multilevel model: Regression coefficients vary across clusters. We can infer the variability of coefficients in larger populations as compared to samples.
* The coefficients depend on clusters by including a term which considers the effect of these clusters.
* For multilevel models, data should be organised in clusters, clusters should be randomly sampled, gender, caste are not randomly sampled & we should model correlation between clusters, we should have interest in variance in between cluster.
* Advantages of multilevel model: It is efficient because one parameter measures the variance of given coeff in diff cluster than the coeff for each cluster

**Multilevel regression models:**

* We need to define distributions for random variables.
* Multilevel because level is the regression model between dependent and independent variable and level 2 is only intercept regression where intercept is the fixed parameter.
* After making a model, we can test the significance of parameters using hypothesis testing one method is likelihood ratio testing.
* **Likelihood testing:** Does the probability of observed data change if we remove a given parameter
* We assume that random effects have normal distribution.

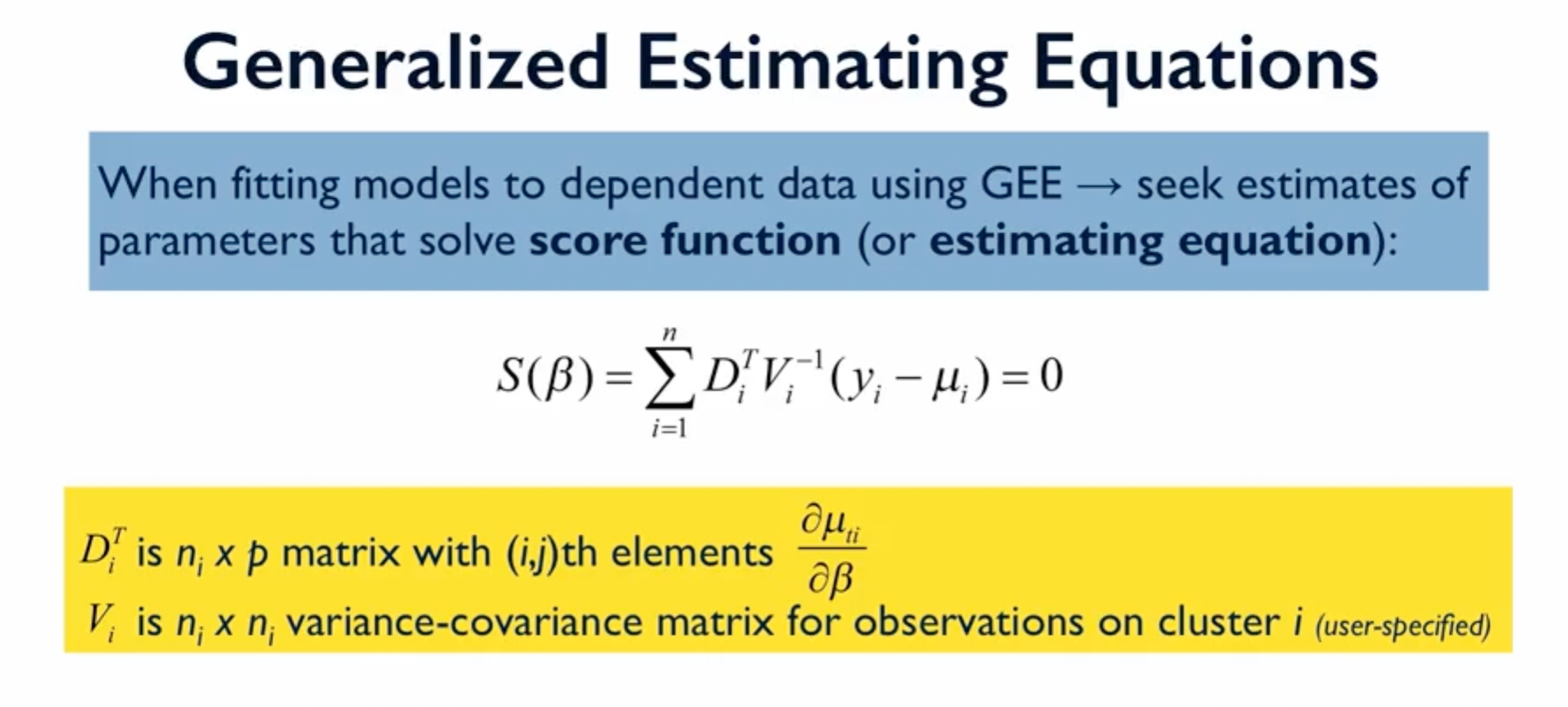
For logistic regression, multilevel modeling is difficult because of non normal distribution outcomes, it is difficult to find the likelihood functions.

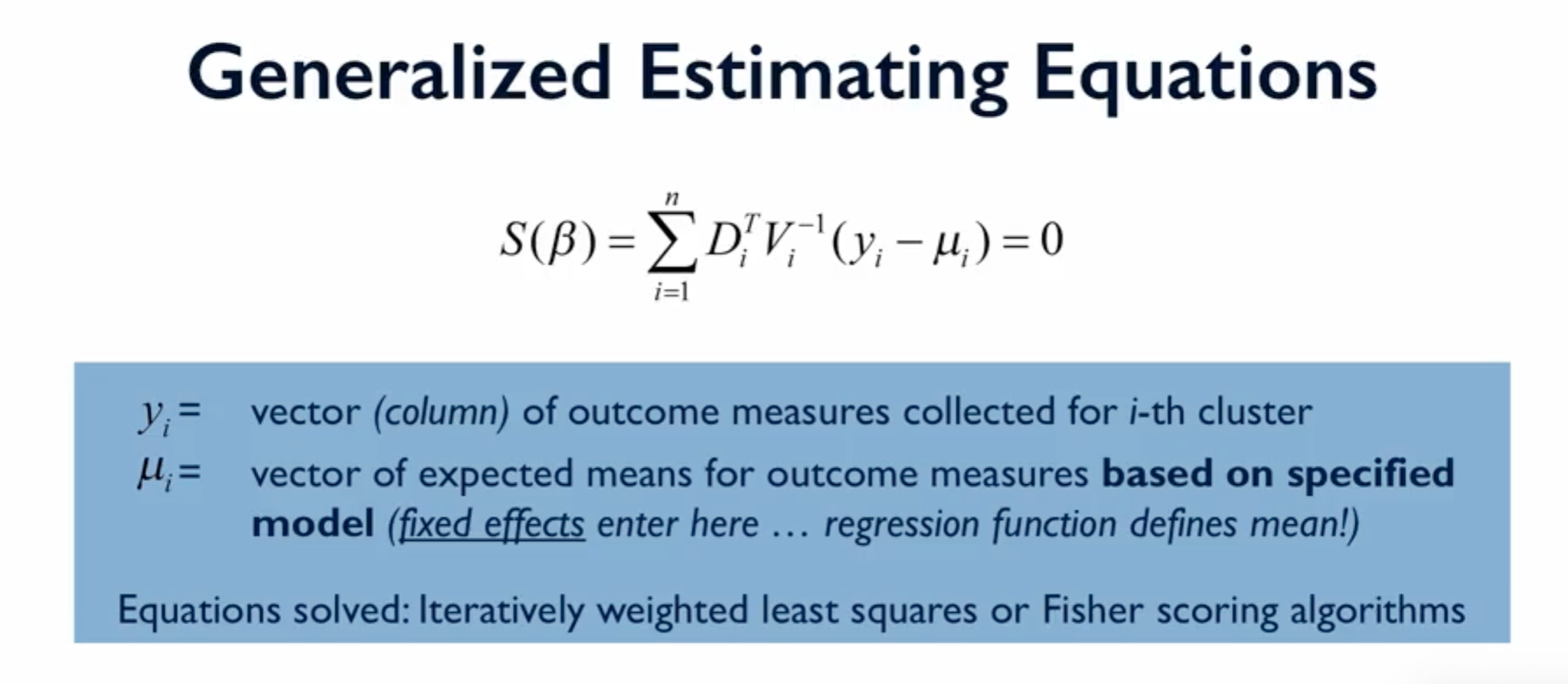
**Marginal model:**

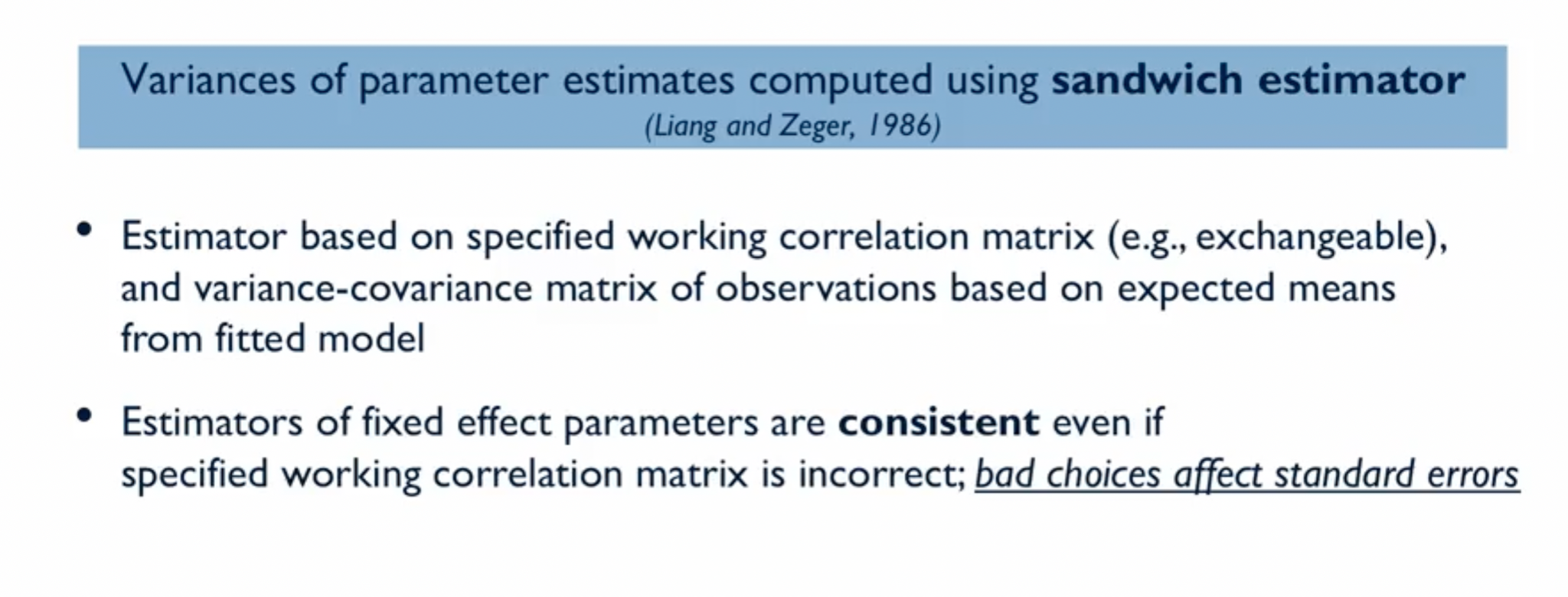
Modelling overall average relationship between dependent variable and independent variable and not include random effects.

1. Select structure for the mean using regression coefficient and predictor variables
2. Select variances and covariances for observations from same cluster, not explained by means
3. Compare different fits of these variances and choose the best one.

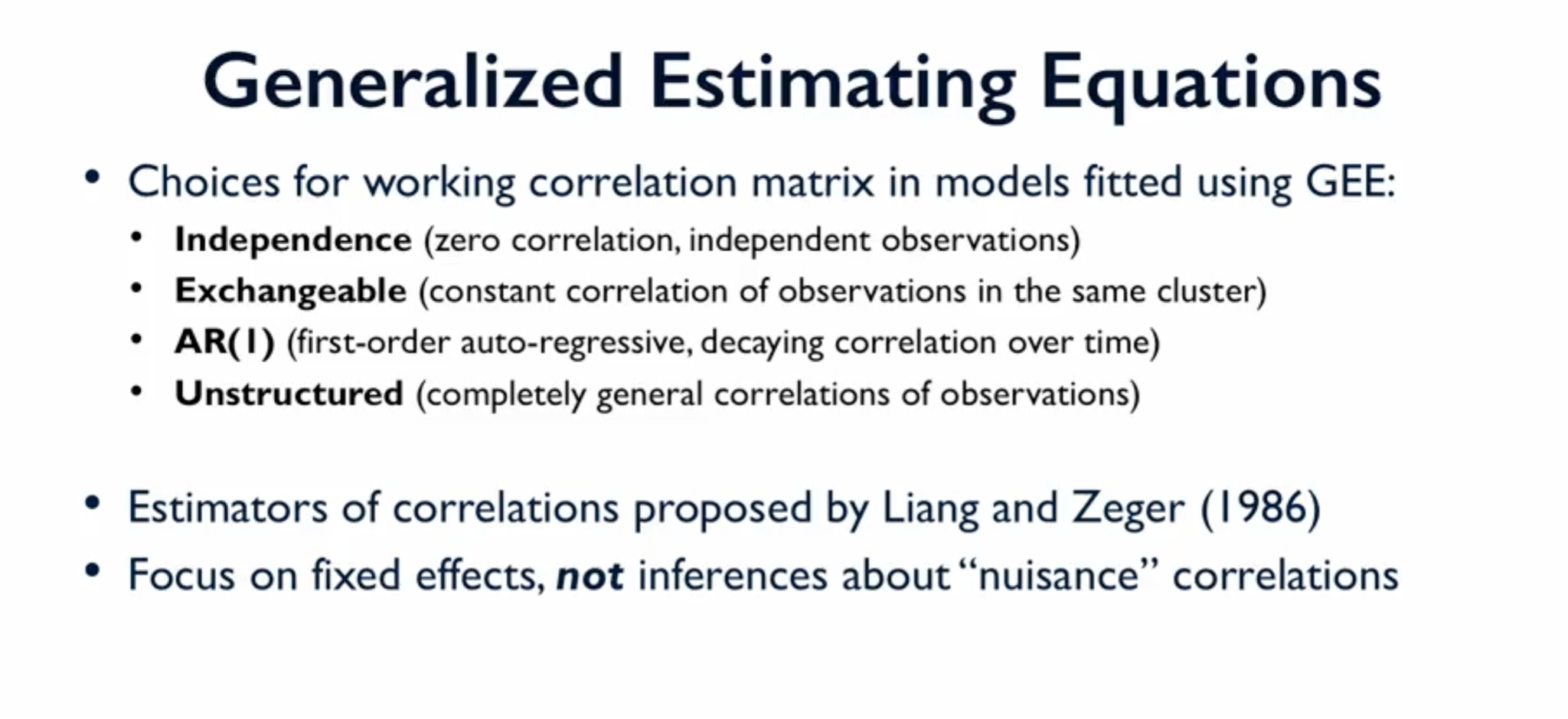
These are faster as compared to Multilevel, accommodates non normal models appropriately.

**GEE Generalised es**

****

****

The fixed effect parameters converge to true values and thus we make a guess for correlations and not covariances becoz for non normal model covariances are related to means. Here, even if our guess is incorrect, variances vary but our mean estimates are consistent



QIC is used to check the fit of model.

**timating equations:**

**Covariance:** Measure of how variability between 2 variables

**WEEK IV:**

**Bayesian statistics:**

More the number of observations, the distribution has lower standard deviation. Initial belief is **prior** and new updated belief is posterior belief.

Adding cost function can help us get a better understanding of world

Initially we need to select distributions for every parameter using our belief.

STAN: program to fit bayesian models