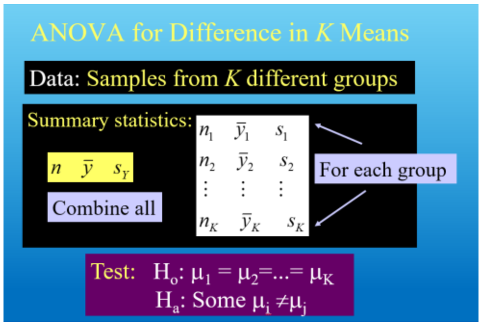
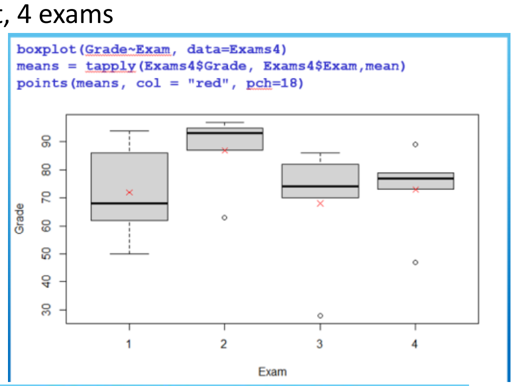
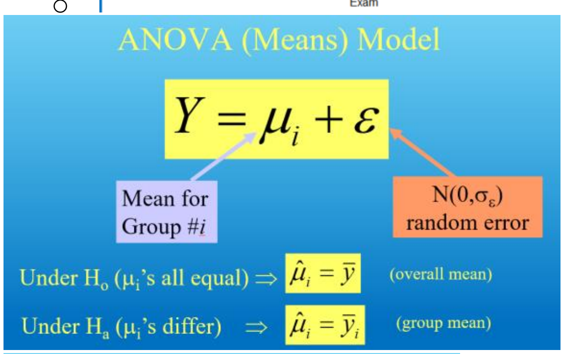
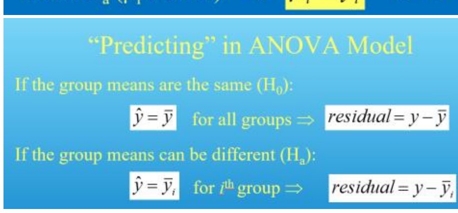
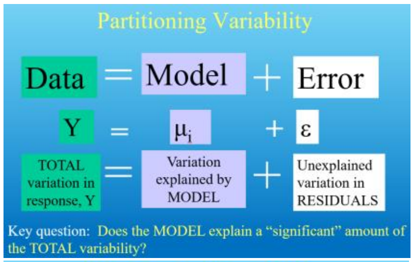
Class 17:

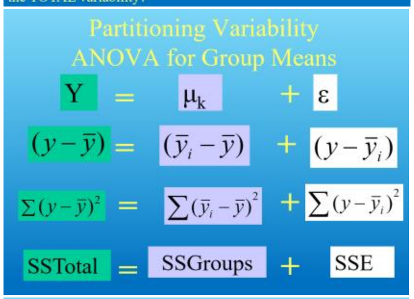
- Y= Quantitative, x= Categorical
    - "dummy" regression
- Difference in two means
    - Two-sample t-test
- Difference in more than two means
    - ANOVA
- Y= binary categorical, x= qualitative
    - Logistic regression
- 

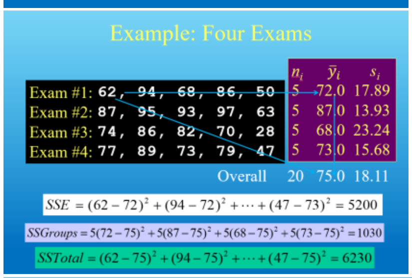
    - N = count
    - Y-bar = sample mean
    - S = standard deviation
- Boxplot, 4 exams
    - 

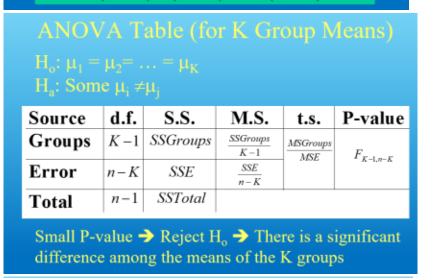
- 

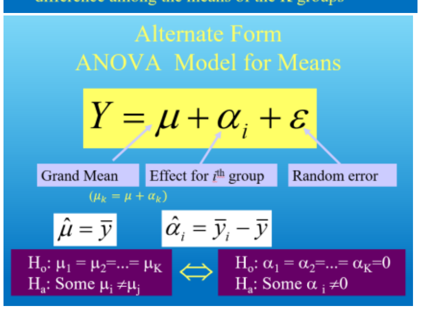
- 

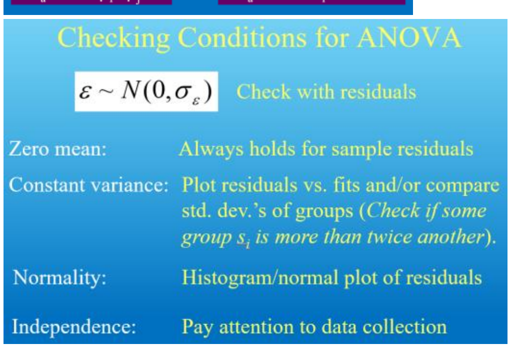
- 

- 

- 

- 

- 

- 

ANOVA conditions

ANOVA for Grades vs. Students

```
> tapply(Exams4$Grade,Exams4$Student,mean)
 Barb Betsy  Bill   Bob   Bud
   75    91    79    83    47

> round(tapply(Exams4$Grade,Exams4$Student,sd),2)
 Barb Betsy  Bill   Bob   Bud
10.30  4.24 10.98 11.40 14.45

> modS=aov(Grade~factor(Student),data=Exams4)
> summary(modS)

            Df Sum Sq Mean Sq F value    Pr(>F)
Student      4   4480  1120.0     9.6 0.000468 ***
Residuals   15   1750   116.7
```

There is a significant difference in mean exam score between the students.

-

Problem of Multiplicity

When doing <u>many</u> pairwise comparisons

$\Rightarrow$ likely to make a Type I error (find a false difference)

In the cartoon, even if there is no relationship between the color and acne, the chance of seeing at least one 0.05 significant test out of 20 independent tests is

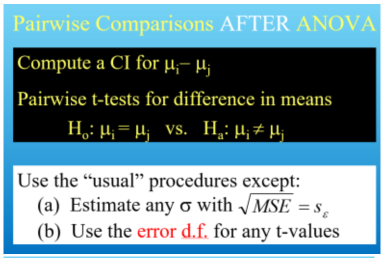
1-(1-0.05)^20 =64%!

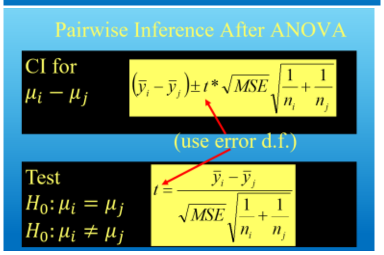
Possible fixes:

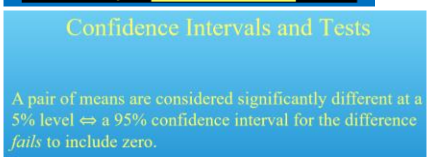
(a) Do only a few pre-planned comparisons
(b) Adjust the significance level used for each test.

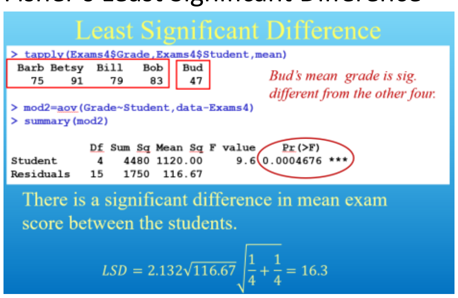
-

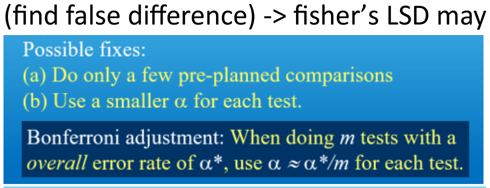
Class 18:

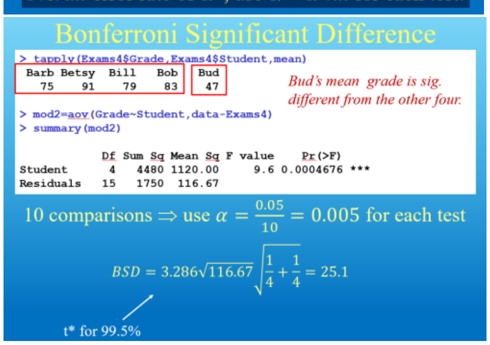
- 

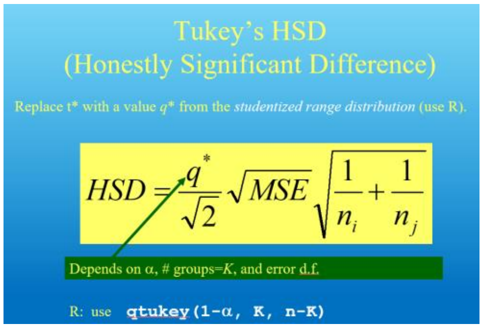
- 

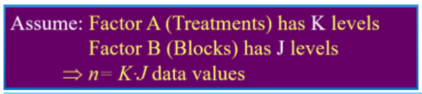
- 

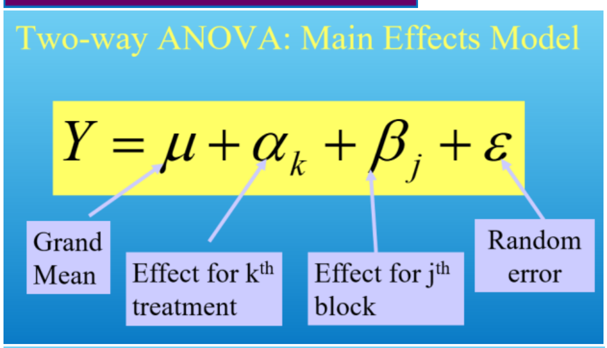
- Fisher's Least Significant Difference
- 

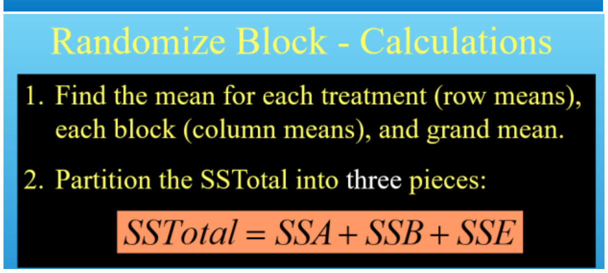
- Problem of Multiplicity: when doing many pairwise comparisons -> likely to make Type I error (find false difference) -> fisher's LSD may be too lenient
- 

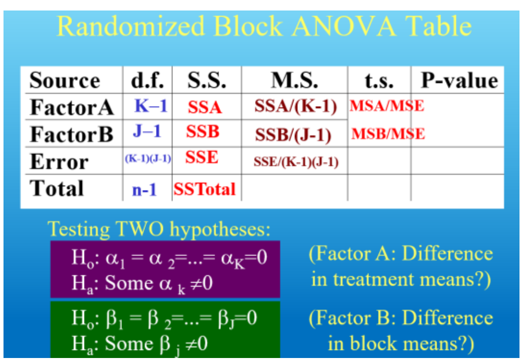
- 

- 

- A simple block design has two factors with exactly one data value in each combination of the factors
- 

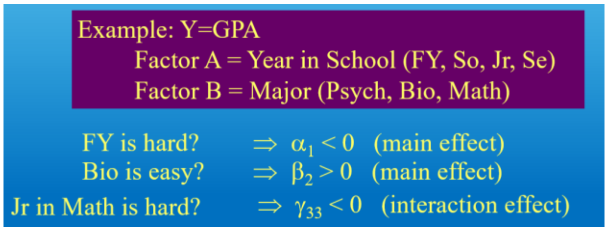
- 

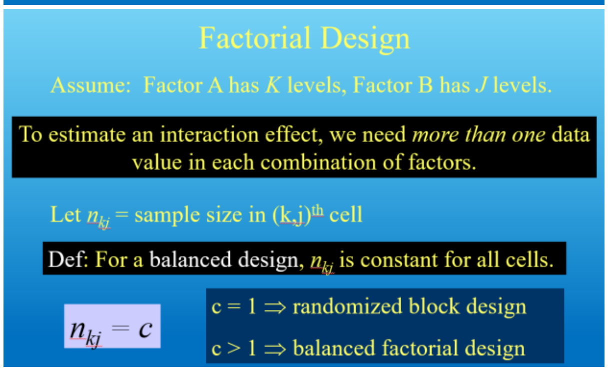
- 

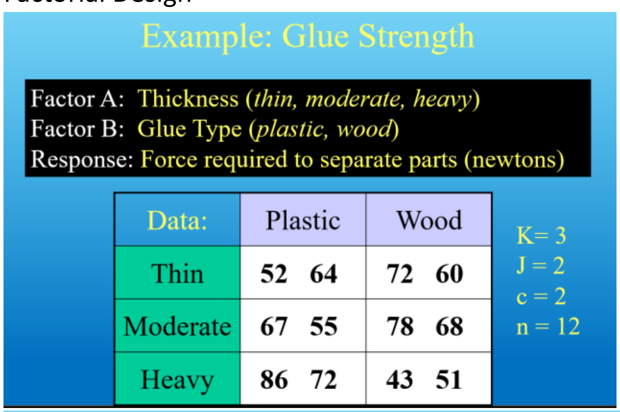
SSTotal
- 

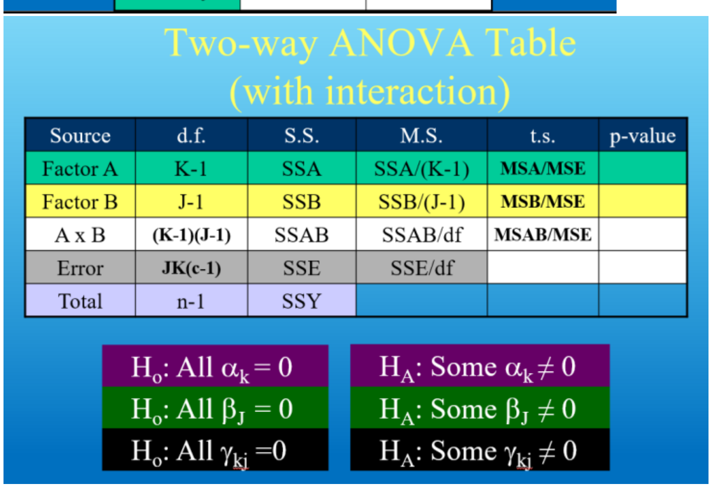
ANOVA Table

Class 19:

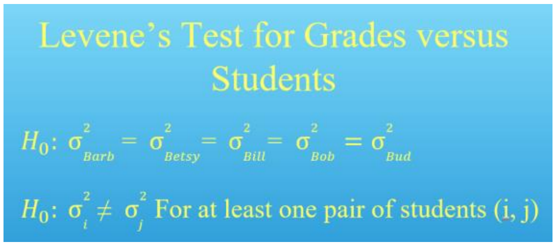
- An *interaction effect* occurs when a significant difference is present at a specific *combination* of factors.
- 

- 

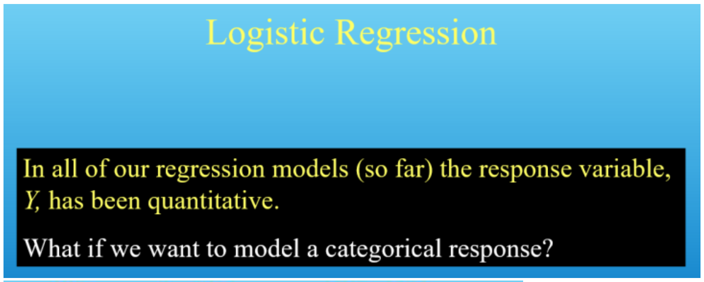
- Factorial Design
- 

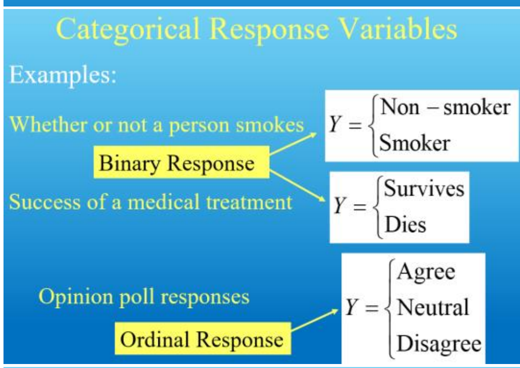
- 

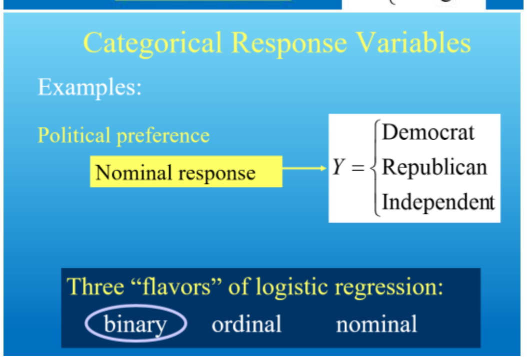
Two-way ANOVA Table

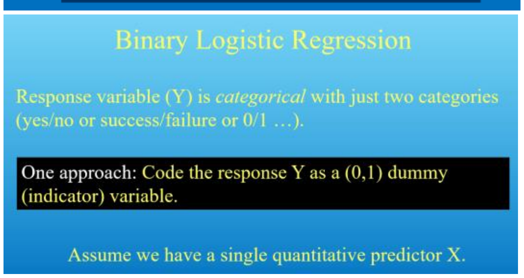
- 
Levene's Test for Grades versus Students
H_0: $\sigma^2_{Barb} = \sigma^2_{Betsy} = \sigma^2_{Bill} = \sigma^2_{Bob} = \sigma^2_{Bud}$
H_0: $\sigma^2_i \neq \sigma^2_j$ For at least one pair of students (i, j)
Levene

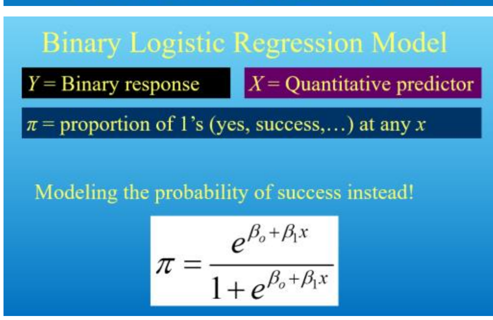
Class 20:

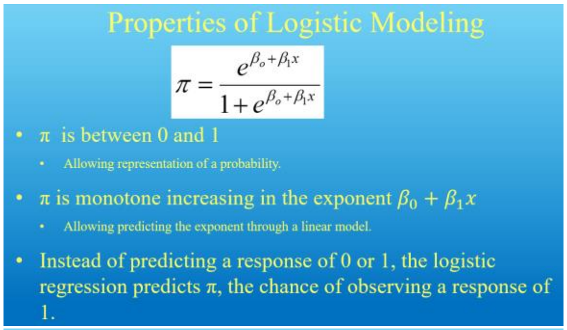
- 

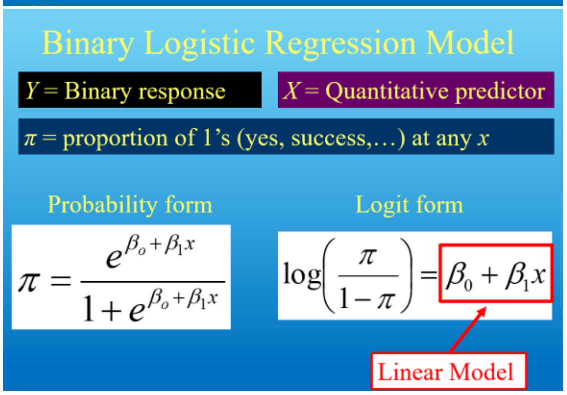
- 

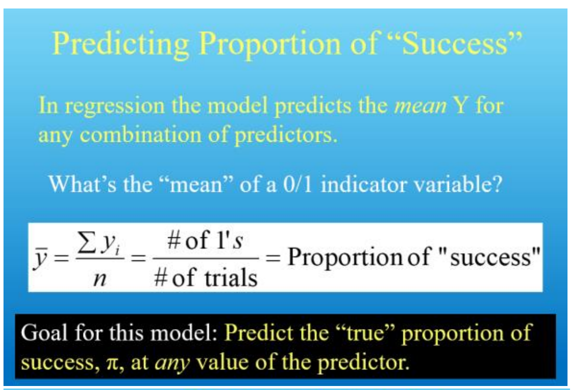
- 

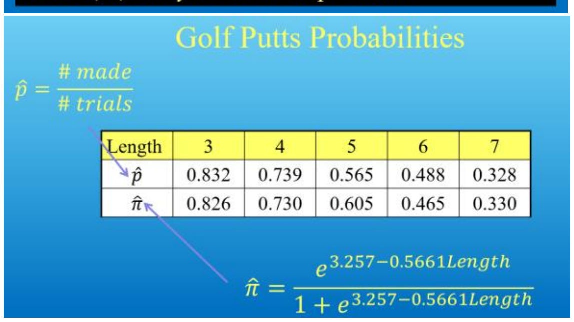
-  Binary Logistic Regression

- 

- 

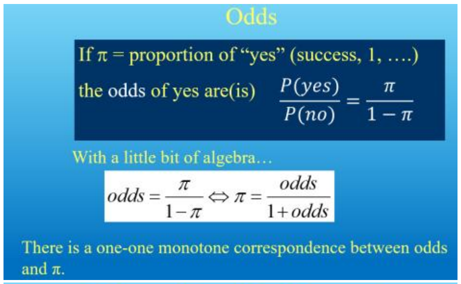
- 

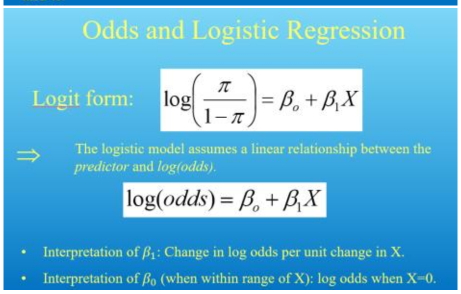
binary logistic regression model
- 

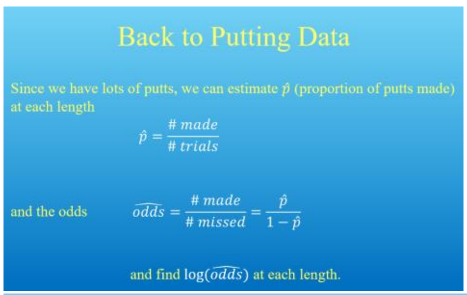
- 

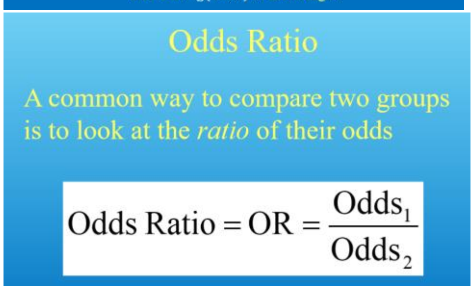
Probability form
- 

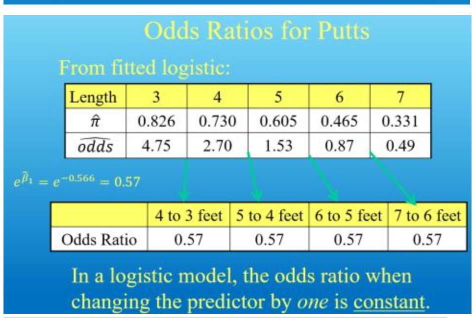
Odds

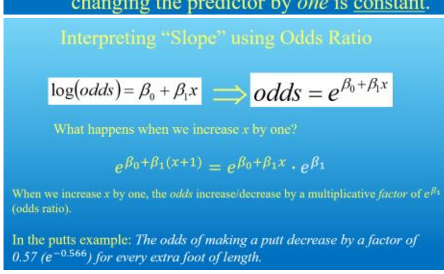
- 

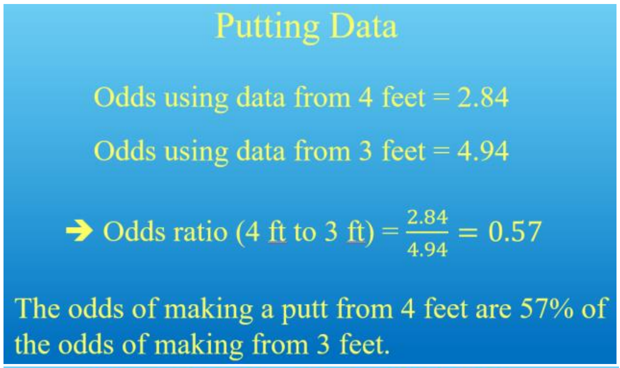
-  odds and logistic regression, logit, b1, b0

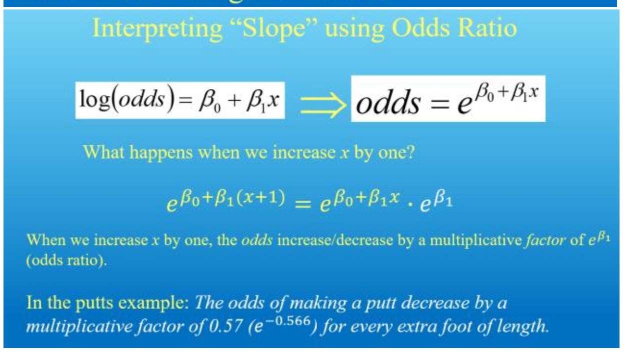
- 

-  Odds ratio

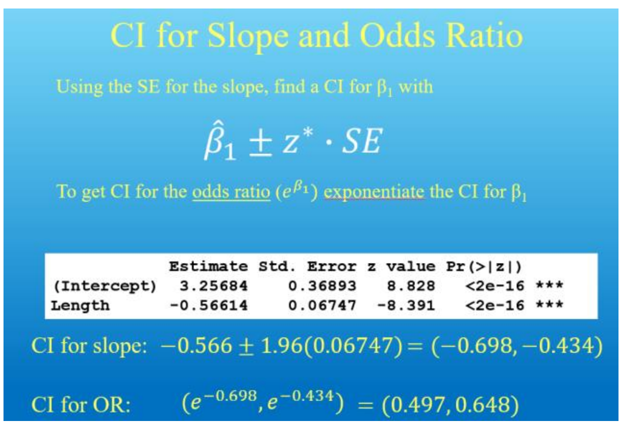
- 

- 

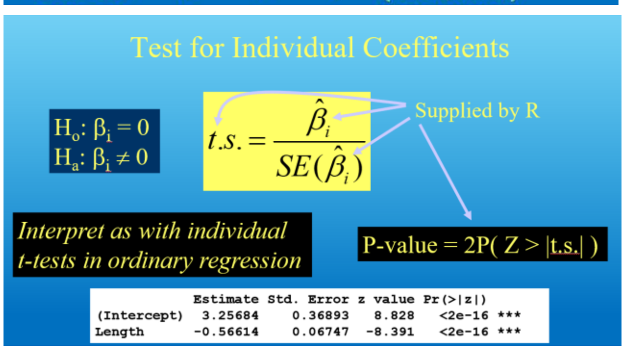
- 

- 

slope odds ratio

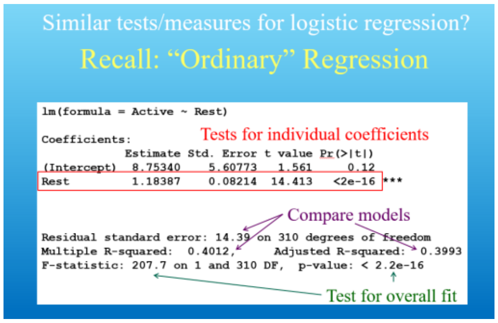
- 

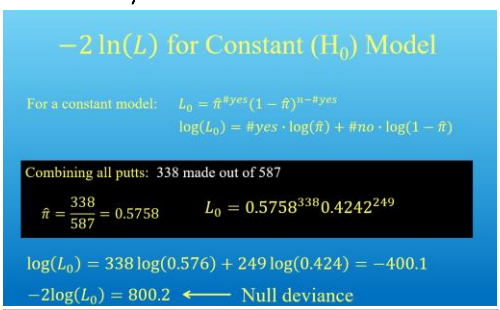
confidence interval

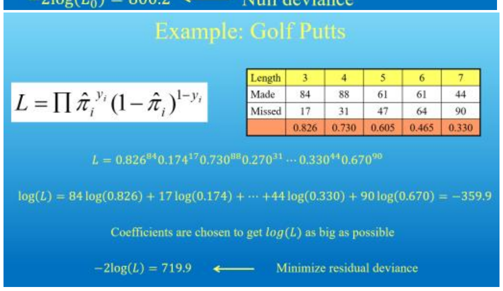
- 

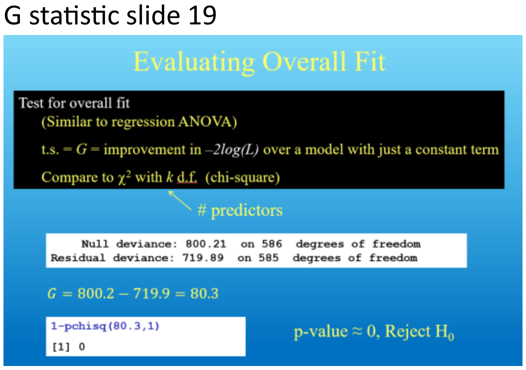
test for individual coefficients

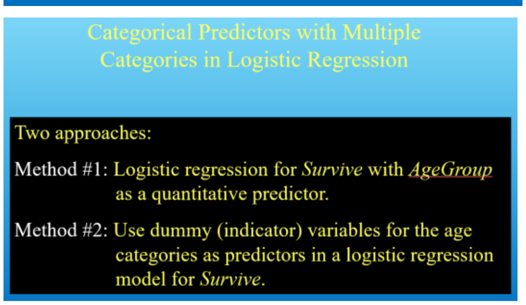
Class 22

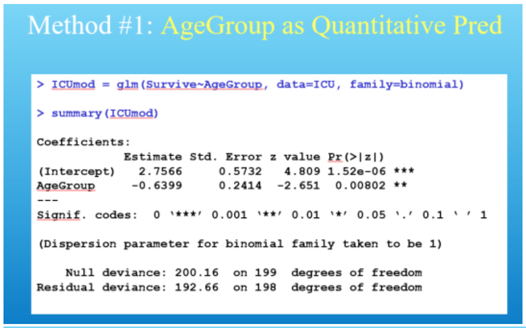
- 

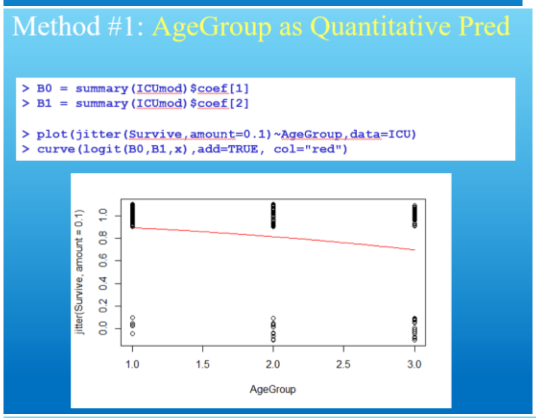
ordinary regression
- Parameters are chosen to *maximize* the *likelihood* of the observed sample. (Maximum Likelihood Estimation)
- 

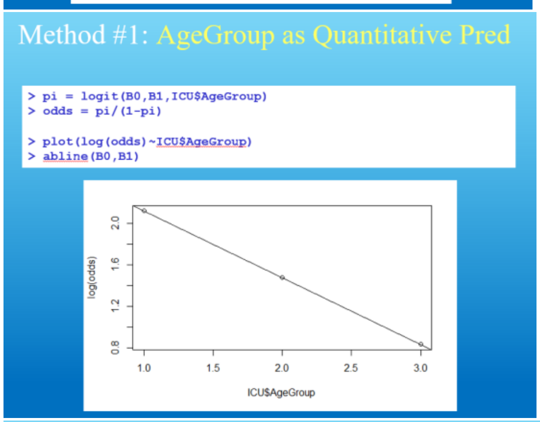
- 

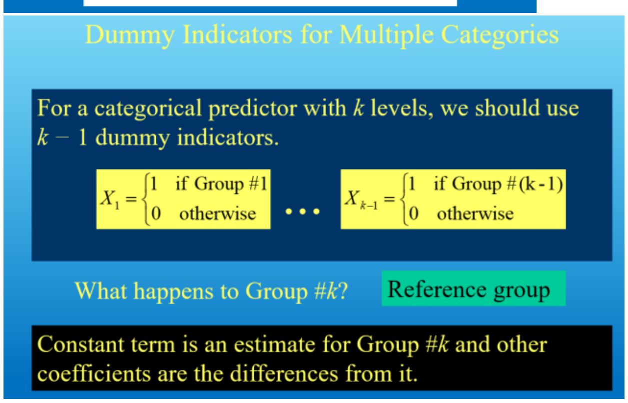
- G statistic slide 19
- 

evaluating overall fit
- 

- 

- 

- 

- 

- Binary logistic regression model slide 29

- **Interpreting Individual Tests**

  Similar issues to ordinary regression:
  - Is the predictor helpful, *given the other predictors* are already in the model?
  - Beware of problems due to multicollinearity.
  - Try to keep the model simple.

- **G-Test for Overall Fit**
  (Similar to regression ANOVA455)

  $H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$ vs. H_a: Some $\beta_i \neq 0$

  t.s. = G = *improvement* in $-2log(L)$ over a model with just a constant term

  Compare to χ^2 with k d.f.

  \# predictors

  ```
      Null deviance: 200.16  on 199  degrees of freedom
  Residual deviance: 191.59  on 197  degrees of freedom
  ```

  $G = 200.16 - 191.59 = 8.57$

  Reject H_0

  ```
  1-pchisq(8.57,2)
  [1] 0.01377362
  ```

  g test for overall fit

- **Method #2: Survive ~ Middle + Old**

  ```
  ICUmod.2 = glm(Survive~factor(AgeGroup), data=ICU, family=binomial)
  summary(ICUmod.2)
  ```

  log(odds) for Young

  ```
  Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
  (Intercept)           2.3795     0.4675   5.090 3.57e-07 ***
  factor(AgeGroup)2    -1.1184     0.5422  -2.063  0.03915 *
  factor(AgeGroup)3    -1.4413     0.5439  -2.650  0.00805 **
  ```

  Change in log(odds) for Middle and Old—compared to Young

  ```
  1-pchisq(ICUmod.2$null.deviance - ICUmod.2$deviance,
  ICUmod.2$df.null - ICUmod.2$df.residual)
  [1] 0.01375896
  ```

- **Recall: Nested F-test**

  Purpose: Test a subset of predictors

  Ex: $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$

  $H_0: \beta_3 = \beta_4 = \beta_5 = 0$ vs. $H_a: Some\ \beta_i \neq 0$ for i>2

  Basic idea: Is the improvement (reduction in SSE) "significant" for the number of extra predictors?

  i.e. Compare "full" model to "reduced" model

  t.s.= F-ratio (interpret similar to ANOVA)

  nested f-test

- **Nested LRT for Logistic Regression**
  (Likelihood Ratio Test)

  Purpose: Test a subset of predictors

  Ex: $\log(odds) = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$

  $H_0: \beta_3 = \beta_4 = \beta_5 = 0$ vs. $H_a: Some\ \beta_i \neq 0$ for i>2

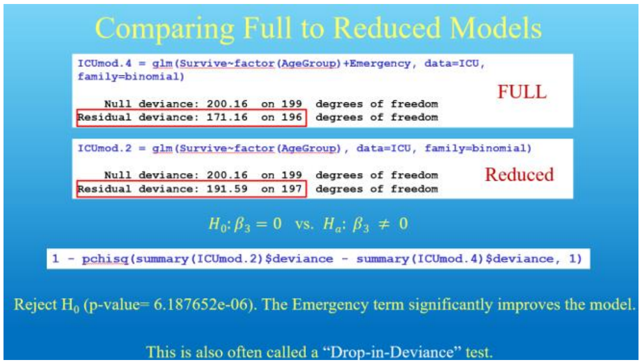
  Basic idea: Is the improvement, change in $-2\log(L)$, "significant" for the number of extra predictors?

  i.e. Compare "reduced" model to "full" model

  $\chi^2 = -2\log(L_{Reduced}) - (-2\log(L_{Full}))$

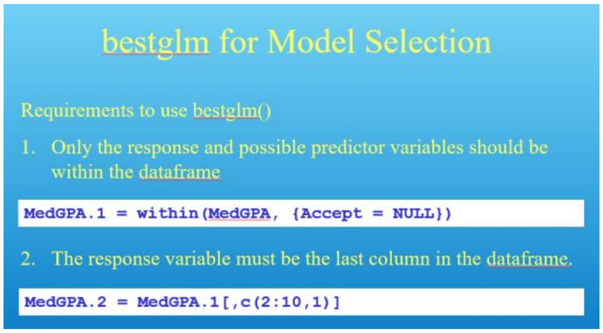
  Chi-square d.f.=#extra predictors tested

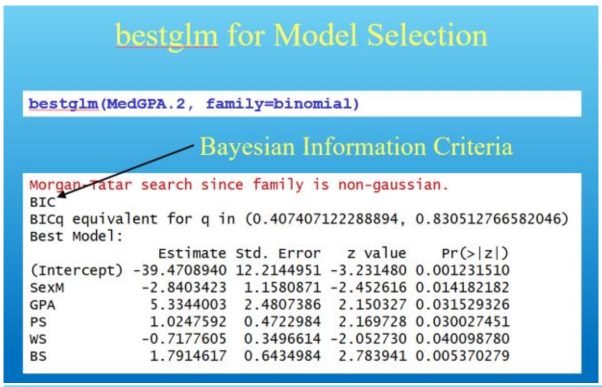
  nested lrt

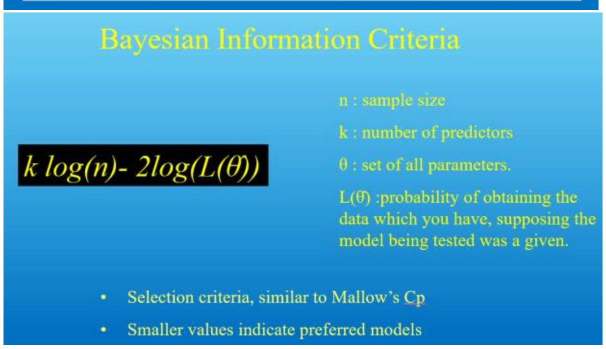
- 

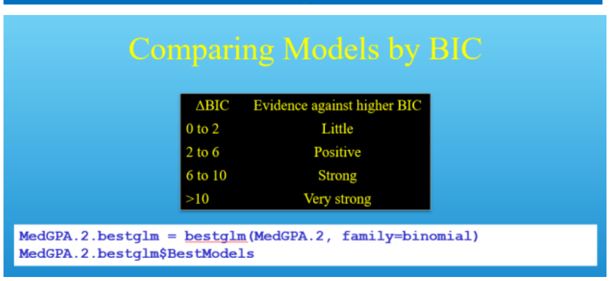
drop in deviance test

Class 23

- 

bestglm for model selection
- 

- 

BIC – Bayesian information criteria
- 

compare bic models
-

Code:

18: 21, 22, 23, 24, 31

19: 13, 20, 21, 22, 26, 27, 28, 30, 31, 32, 33

20: 9, 14, 16, 19, 20, 22, 23, 24, 25, 31, 32

21: 14, 15, 22

22: 11, 17, 19, 21, 27, 30, 33, 36

23: 9, 13, 15, 16, 17, 18