

Class 17:

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

ANOVA for Difference in K Means

Data: Samples from K different groups

Summary statistics:

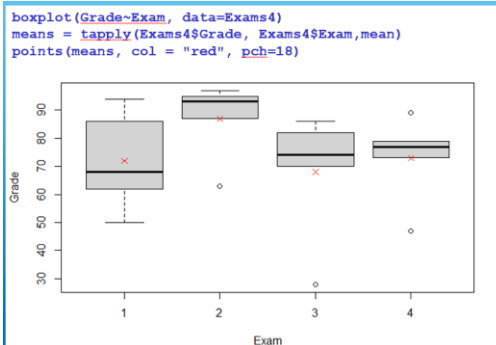
n	\bar{y}	s_y
n_1	\bar{y}_1	s_1
n_2	\bar{y}_2	s_2
\vdots	\vdots	\vdots
n_K	\bar{y}_K	s_K

Combine all

For each group

Test: $H_0: \mu_1 = \mu_2 = \dots = \mu_K$
 $H_a: \text{Some } \mu_i \neq \mu_j$

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



○

ANOVA (Means) Model

$$Y = \mu_i + \varepsilon$$

Mean for Group # i

$N(0, \sigma_\varepsilon)$ random error

Under H_0 (μ_i 's all equal) $\Rightarrow \hat{\mu}_i = \bar{y}$ (overall mean)

Under H_a (μ_i 's differ) $\Rightarrow \hat{\mu}_i = \bar{y}_i$ (group mean)

“Predicting” in ANOVA Model

If the group means are the same (H_0):

$$\hat{y} = \bar{y} \quad \text{for all groups} \Rightarrow \text{residual} = y - \bar{y}$$

If the group means can be different (H_a):

$$\hat{y} = \bar{y}_i \quad \text{for } i^{\text{th}} \text{ group} \Rightarrow \text{residual} = y - \bar{y}_i$$

Partitioning Variability

Data = **Model** + **Error**

Y = **μ_i** + **ε**

TOTAL = **Variation explained by MODEL** + **Unexplained variation in RESIDUALS**

Key question: Does the MODEL explain a "significant" amount of the TOTAL variability?

Partitioning Variability

ANOVA for Group Means

Y = **μ_k** + **ε**

$(y - \bar{y})$ = **$(\bar{y}_i - \bar{y})$** + **$(y - \bar{y}_i)$**

$\sum (y - \bar{y})^2$ = **$\sum (\bar{y}_i - \bar{y})^2$** + **$\sum (y - \bar{y}_i)^2$**

SSTotal = **SSGroups** + **SSE**

Example: Four Exams

	n_i	\bar{y}_i	s_i
Exam #1: 62, 94, 68, 86, 50	5	72.0	17.89
Exam #2: 87, 95, 93, 97, 63	5	87.0	13.93
Exam #3: 74, 86, 82, 70, 28	5	68.0	23.24
Exam #4: 77, 89, 73, 79, 47	5	73.0	15.68
Overall	20	75.0	18.11

$SSE = (62 - 72)^2 + (94 - 72)^2 + \dots + (47 - 73)^2 = 5200$

$SSGroups = 5(72 - 75)^2 + 5(87 - 75)^2 + 5(68 - 75)^2 + 5(73 - 75)^2 = 1030$

$SSTotal = (62 - 75)^2 + (94 - 75)^2 + \dots + (47 - 75)^2 = 6230$

ANOVA Table (for K Group Means)

$H_0: \mu_1 = \mu_2 = \dots = \mu_K$
 $H_a: \text{Some } \mu_i \neq \mu_j$

Source	d.f.	S.S.	M.S.	t.s.	P-value
Groups	$K - 1$	$SSGroups$	$\frac{SSGroups}{K - 1}$	$\frac{MSGroups}{MSE}$	$F_{K-1, n-K}$
Error	$n - K$	SSE	$\frac{SSE}{n - K}$		
Total	$n - 1$	$SSTotal$			

Small P-value \rightarrow Reject $H_0 \rightarrow$ There is a significant difference among the means of the K groups

Alternate Form

ANOVA Model for Means

$Y = \mu + \alpha_i + \varepsilon$

Grand Mean Effect for i^{th} group Random error

$(\mu_k = \mu + \alpha_k)$

$\hat{\mu} = \bar{y}$ $\hat{\alpha}_i = \bar{y}_i - \bar{y}$

$H_0: \mu_1 = \mu_2 = \dots = \mu_K$ \Leftrightarrow $H_0: \alpha_1 = \alpha_2 = \dots = \alpha_K = 0$
 $H_a: \text{Some } \mu_i \neq \mu_j$ $H_a: \text{Some } \alpha_i \neq 0$

Checking Conditions for ANOVA

$\varepsilon \sim N(0, \sigma_\varepsilon)$ Check with residuals

Zero mean: Always holds for sample residuals

Constant variance: Plot residuals vs. fits and/or compare std. dev.'s of groups (*Check if some group s_i is more than twice another*).

Normality: Histogram/normal plot of residuals

Independence: Pay attention to data collection

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 many pairwise comparisons

⇒ 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

$1-(1-0.05)^{20} \approx 64\%$!

Possible fixes:

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

Class 18:

Pairwise Comparisons AFTER ANOVA

Compute a CI for $\mu_i - \mu_j$

Pairwise t-tests for difference in means

$$H_0: \mu_i = \mu_j \quad \text{vs.} \quad H_a: \mu_i \neq \mu_j$$

Use the "usual" procedures except:

- (a) Estimate any σ with $\sqrt{MSE} = s_e$
- (b) Use the **error d.f.** for any t-values

Pairwise Inference After ANOVA

CI for $\mu_i - \mu_j$

$$(\bar{y}_i - \bar{y}_j) \pm t^* \sqrt{MSE} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

(use error d.f.)

Test

$$H_0: \mu_i = \mu_j$$

$$H_a: \mu_i \neq \mu_j$$

$$t = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{MSE} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

Confidence Intervals and Tests

A pair of means are considered significantly different at a 5% level \Leftrightarrow a 95% confidence interval for the difference *fails* to include zero.

Fisher's Least Significant Difference

Least Significant Difference

```
> tapply(Exams4$Grade, Exams4$Student, mean)
Barb Betsy Bill Bob Bud
75 91 79 83 47
> mod2=aov(Grade~Student, data=Exams4)
> summary(mod2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Student	4	4480	1120.00	9.6	0.0004676 ***
Residuals	15	1750	116.67		

Bud's mean grade is sig. different from the other four.

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

$$LSD = 2.132\sqrt{116.67} \sqrt{\frac{1}{4} + \frac{1}{4}} = 16.3$$

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

Possible fixes:

- (a) Do only a few pre-planned comparisons
- (b) Use a smaller α for each test.

Bonferroni adjustment: When doing m tests with a *overall* error rate of α^* , use $\alpha \approx \alpha^*/m$ for each test.

Bonferroni Significant Difference

```
> tapply(Exams4$Grade, Exams4$Student, mean)
Barb Betsy Bill Bob Bud
75 91 79 83 47
> mod2=aov(Grade~Student, data=Exams4)
> summary(mod2)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Student	4	4480	1120.00	9.6	0.0004676 ***
Residuals	15	1750	116.67		

Bud's mean grade is sig. different from the other four.

10 comparisons \Rightarrow use $\alpha = \frac{0.05}{10} = 0.005$ for each test

$$BSD = 3.286\sqrt{116.67} \sqrt{\frac{1}{4} + \frac{1}{4}} = 25.1$$

t* for 99.5%

Tukey's HSD (Honestly Significant Difference)

Replace t^* with a value q^* from the *studentized range distribution* (use R).

$$HSD = \frac{q^*}{\sqrt{2}} \sqrt{MSE} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

Depends on α , # groups- K , and error d.f.

R: use `qtukey(1- α , K , n- K)`

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

Assume: Factor A (Treatments) has K levels
Factor B (Blocks) has J levels
 $\Rightarrow n = K \cdot J$ data values

Two-way ANOVA: Main Effects Model

$$Y = \mu + \alpha_k + \beta_j + \varepsilon$$

Grand
Mean

Effect for k^{th}
treatment

Effect for j^{th}
block

Random
error

Randomize Block - Calculations

1. Find the mean for each treatment (row means), each block (column means), and grand mean.
2. Partition the SST_{Total} into three pieces:

$$SST_{\text{Total}} = SSA + SSB + SSE$$

SST_{Total}

Randomized Block ANOVA Table

Source	d.f.	S.S.	M.S.	t.s.	P-value
Factor A	$K-1$	SSA	$SSA/(K-1)$	MSA/MSE	
Factor B	$J-1$	SSB	$SSB/(J-1)$	MSB/MSE	
Error	$(K-1)(J-1)$	SSE	$SSE/(K-1)(J-1)$		
Total	$n-1$	SST_{Total}			

Testing TWO hypotheses:

$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_K = 0$
 $H_a: \text{Some } \alpha_k \neq 0$

$H_0: \beta_1 = \beta_2 = \dots = \beta_J = 0$
 $H_a: \text{Some } \beta_j \neq 0$

(Factor A: Difference in treatment means?)

(Factor B: Difference in block means?)

ANOVA Table

Class 19:

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

Example: $Y = \text{GPA}$

Factor A = Year in School (FY, So, Jr, Se)

Factor B = Major (Psych, Bio, Math)

FY is hard? $\Rightarrow \alpha_1 < 0$ (main effect)

Bio is easy? $\Rightarrow \beta_2 > 0$ (main effect)

Jr in Math is hard? $\Rightarrow \gamma_{33} < 0$ (interaction effect)

Factorial Design

Assume: Factor A has K levels, Factor B has J levels.

To estimate an interaction effect, we need *more than one* data value in each combination of factors.

Let n_{kj} = sample size in $(k,j)^{\text{th}}$ cell

Def: For a balanced design, n_{kj} is constant for all cells.

$$n_{kj} = c$$

$c = 1 \Rightarrow$ randomized block design

$c > 1 \Rightarrow$ balanced factorial design

Factorial Design

Example: Glue Strength

Factor A: Thickness (*thin, moderate, heavy*)

Factor B: Glue Type (*plastic, wood*)

Response: Force required to separate parts (newtons)

Data:	Plastic	Wood	K = 3 J = 2 c = 2 n = 12
Thin	52 64	72 60	
Moderate	67 55	78 68	
Heavy	86 72	43 51	

Two-way ANOVA Table (with interaction)

Source	d.f.	S.S.	M.S.	t.s.	p-value
Factor A	K-1	SSA	SSA/(K-1)	MSA/MSE	
Factor B	J-1	SSB	SSB/(J-1)	MSB/MSE	
A x B	(K-1)(J-1)	SSAB	SSAB/df	MSAB/MSE	
Error	JK(c-1)	SSE	SSE/df		
Total	n-1	SSY			

$H_o: \text{All } \alpha_k = 0$

$H_A: \text{Some } \alpha_k \neq 0$

$H_o: \text{All } \beta_j = 0$

$H_A: \text{Some } \beta_j \neq 0$

$H_o: \text{All } \gamma_{kj} = 0$

$H_A: \text{Some } \gamma_{kj} \neq 0$

Two-way ANOVA Table

Levene's Test for Grades versus Students

$$H_0: \sigma_{Barb}^2 = \sigma_{Betsy}^2 = \sigma_{Bill}^2 = \sigma_{Bob}^2 = \sigma_{Bud}^2$$

$$H_0: \sigma_i^2 \neq \sigma_j^2 \text{ For at least one pair of students (i, j)}$$

Levene

Class 20:

Logistic Regression

In all of our regression models (so far) the response variable, Y , has been quantitative.

What if we want to model a categorical response?

Categorical Response Variables

Examples:

Whether or not a person smokes

Binary Response

$$Y = \begin{cases} \text{Non-smoker} \\ \text{Smoker} \end{cases}$$

Success of a medical treatment

$$Y = \begin{cases} \text{Survives} \\ \text{Dies} \end{cases}$$

Opinion poll responses

Ordinal Response

$$Y = \begin{cases} \text{Agree} \\ \text{Neutral} \\ \text{Disagree} \end{cases}$$

Categorical Response Variables

Examples:

Political preference

Nominal response

$$Y = \begin{cases} \text{Democrat} \\ \text{Republican} \\ \text{Independent} \end{cases}$$

Three “flavors” of logistic regression:

binary

ordinal

nominal

Binary Logistic Regression

Response variable (Y) is *categorical* with just two categories (yes/no or success/failure or 0/1 ...).

One approach: Code the response Y as a (0,1) dummy (indicator) variable.

Assume we have a single quantitative predictor X .

Binary Logistic Regression

Binary Logistic Regression Model

Y = Binary response

X = Quantitative predictor

π = proportion of 1's (yes, success,...) at any x

Modeling the probability of success instead!

$$\pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Properties of Logistic Modeling

$$\pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

- π is between 0 and 1
 - Allowing representation of a probability.
- π is monotone increasing in the exponent $\beta_0 + \beta_1 x$
 - Allowing predicting the exponent through a linear model.
- Instead of predicting a response of 0 or 1, the logistic regression predicts π , the chance of observing a response of 1.

Binary Logistic Regression Model

Y = Binary response

X = Quantitative predictor

π = proportion of 1's (yes, success,...) at any x

Probability form

$$\pi = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Logit form

$$\log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 x$$

Linear Model

binary logistic regression model

Predicting Proportion of "Success"

In regression the model predicts the *mean* Y for any combination of predictors.

What's the "mean" of a 0/1 indicator variable?

$$\bar{y} = \frac{\sum y_i}{n} = \frac{\text{\# of 1's}}{\text{\# of trials}} = \text{Proportion of "success"}$$

Goal for this model: Predict the "true" proportion of success, π , at *any* value of the predictor.

Golf Putts Probabilities

$$\hat{p} = \frac{\text{\# made}}{\text{\# trials}}$$

Length	3	4	5	6	7
\hat{p}	0.832	0.739	0.565	0.488	0.328
$\hat{\pi}$	0.826	0.730	0.605	0.465	0.330

$$\hat{\pi} = \frac{e^{3.257 - 0.5661 \text{Length}}}{1 + e^{3.257 - 0.5661 \text{Length}}}$$

Probability form

Odds

The *odds* against a certain horse winning a race are 4 to 1. What does that mean?

4 losses for every 1 win

$$P(\text{Win}) = 1/5$$

$$P(\text{Loss}) = 4/5$$

$$\text{Odds} = \frac{P(\text{Win})}{P(\text{Loss})} = \frac{1/5}{4/5} = \frac{1}{4}$$

Odds

Odds

If π = proportion of “yes” (success, 1, ...)
the odds of yes are(is) $\frac{P(yes)}{P(no)} = \frac{\pi}{1-\pi}$

With a little bit of algebra...

$$odds = \frac{\pi}{1-\pi} \Leftrightarrow \pi = \frac{odds}{1+odds}$$

There is a one-one monotone correspondence between odds and π .

Odds and Logistic Regression

Logit form: $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$

⇒ The logistic model assumes a linear relationship between the predictor and $\log(odds)$.

$$\log(odds) = \beta_0 + \beta_1 X$$

- Interpretation of β_1 : Change in log odds per unit change in X.
- Interpretation of β_0 (when within range of X): log odds when X=0.

odds and logistic regression, logit, b1, b0

Back to Putting Data

Since we have lots of putts, we can estimate \hat{p} (proportion of putts made) at each length

$$\hat{p} = \frac{\# \text{ made}}{\# \text{ trials}}$$

and the odds $\widehat{odds} = \frac{\# \text{ made}}{\# \text{ missed}} = \frac{\hat{p}}{1-\hat{p}}$

and find $\log(\widehat{odds})$ at each length.

Odds Ratio

A common way to compare two groups is to look at the *ratio* of their odds

$$\text{Odds Ratio} = \text{OR} = \frac{\text{Odds}_1}{\text{Odds}_2}$$

Odds ratio

Odds Ratios for Putts

From fitted logistic:

Length	3	4	5	6	7
$\hat{\pi}$	0.826	0.730	0.605	0.465	0.331
\widehat{odds}	4.75	2.70	1.53	0.87	0.49

$$e^{\hat{\beta}_1} = e^{-0.566} = 0.57$$

	4 to 3 feet	5 to 4 feet	6 to 5 feet	7 to 6 feet
Odds Ratio	0.57	0.57	0.57	0.57

In a logistic model, the odds ratio when changing the predictor by *one* is *constant*.

Interpreting “Slope” using Odds Ratio

$$\log(odds) = \beta_0 + \beta_1 x \Rightarrow odds = e^{\beta_0 + \beta_1 x}$$

What happens when we increase x by one?

$$e^{\beta_0 + \beta_1(x+1)} = e^{\beta_0 + \beta_1 x} \cdot e^{\beta_1}$$

When we increase x by one, the odds increase/decrease by a multiplicative factor of e^{β_1} (odds ratio).

In the putts example: The odds of making a putt decrease by a factor of 0.57 ($e^{-0.566}$) for every extra foot of length.

Putting Data

Odds using data from 4 feet = 2.84

Odds using data from 3 feet = 4.94

$$\rightarrow \text{Odds ratio (4 ft to 3 ft)} = \frac{2.84}{4.94} = 0.57$$

The odds of making a putt from 4 feet are 57% of the odds of making from 3 feet.

Interpreting "Slope" using Odds Ratio

$$\log(odds) = \beta_0 + \beta_1 x \Rightarrow odds = e^{\beta_0 + \beta_1 x}$$

What happens when we increase x by one?

$$e^{\beta_0 + \beta_1(x+1)} = e^{\beta_0 + \beta_1 x} \cdot e^{\beta_1}$$

When we increase x by one, the *odds* increase/decrease by a multiplicative factor of e^{β_1} (odds ratio).

In the putts example: The odds of making a putt decrease by a multiplicative factor of 0.57 ($e^{-0.566}$) for every extra foot of length.

slope odds ratio

CI for Slope and Odds Ratio

Using the SE for the slope, find a CI for β_1 with

$$\hat{\beta}_1 \pm z^* \cdot SE$$

To get CI for the odds ratio (e^{β_1}) exponentiate the CI for β_1

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.25684	0.36893	8.828	<2e-16 ***
Length	-0.56614	0.06747	-8.391	<2e-16 ***

CI for slope: $-0.566 \pm 1.96(0.06747) = (-0.698, -0.434)$

CI for OR: $(e^{-0.698}, e^{-0.434}) = (0.497, 0.648)$

confidence interval

Test for Individual Coefficients

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$

$$t.s. = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Supplied by R

Interpret as with individual *t*-tests in ordinary regression

$$P\text{-value} = 2P(Z > |t.s.|)$$

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.25684	0.36893	8.828	<2e-16 ***
Length	-0.56614	0.06747	-8.391	<2e-16 ***

test for individual coefficients

Class 22

Similar tests/measures for logistic regression?

Recall: "Ordinary" Regression

```
lm(formula = Active ~ Rest)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	8.75340	5.60773	1.561	0.12
Rest	1.18387	0.08214	14.413	<2e-16 ***

Residual standard error: 14.39 on 310 degrees of freedom
 Multiple R-squared: 0.4012, Adjusted R-squared: 0.3993
 F-statistic: 207.7 on 1 and 310 DF, p-value: < 2.2e-16

Tests for individual coefficients

Compare models

Test for overall fit

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

$-2 \ln(L)$ for Constant (H_0) Model

For a constant model: $L_0 = \hat{\pi}^{\#yes}(1 - \hat{\pi})^{n-\#yes}$
 $\log(L_0) = \#yes \cdot \log(\hat{\pi}) + \#no \cdot \log(1 - \hat{\pi})$

Combining all putts: 338 made out of 587

$\hat{\pi} = \frac{338}{587} = 0.5758$ $L_0 = 0.5758^{338} 0.4242^{249}$

$\log(L_0) = 338 \log(0.576) + 249 \log(0.424) = -400.1$
 $-2\log(L_0) = 800.2$ ← Null deviance

-
- Example: Golf Putts

$L = \prod \hat{\pi}_i^{y_i} (1 - \hat{\pi}_i)^{1-y_i}$

Length	3	4	5	6	7
Made	84	88	61	61	44
Missed	17	31	47	64	90
	0.826	0.730	0.605	0.465	0.330

$L = 0.826^{84} 0.174^{17} 0.730^{88} 0.270^{31} \dots 0.330^{44} 0.670^{90}$

$\log(L) = 84 \log(0.826) + 17 \log(0.174) + \dots + 44 \log(0.330) + 90 \log(0.670) = -359.9$

Coefficients are chosen to get $\log(L)$ as big as possible

$-2\log(L) = 719.9$ ← Minimize residual deviance

- G statistic slide 19

Evaluating Overall Fit

Test for overall fit
 (Similar to regression ANOVA)

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

Compare to χ^2 with k d.f. (chi-square)

predictors

Null deviance: 800.21 on 586 degrees of freedom
 Residual deviance: 719.89 on 585 degrees of freedom

$G = 800.2 - 719.9 = 80.3$

`1-pchisq(80.3, 1)`
 [1] 0

p-value ≈ 0 , Reject H_0

- evaluating overall fit

Categorical Predictors with Multiple Categories in Logistic Regression

Two approaches:

Method #1: Logistic regression for *Survive* with *AgeGroup* as a quantitative predictor.

Method #2: Use dummy (indicator) variables for the age categories as predictors in a logistic regression model for *Survive*.

-

Method #1: AgeGroup as Quantitative Pred

```
> ICUmod = glm(Survive~AgeGroup, data=ICU, family=binomial)
> summary(ICUmod)

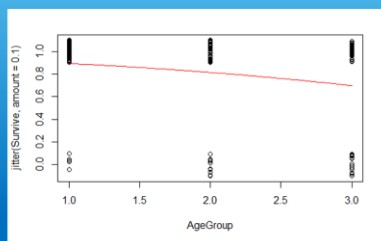
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.7566     0.5732   4.809 1.52e-06 ***
AgeGroup      -0.6399     0.2414  -2.651 0.00802 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 192.66  on 198  degrees of freedom
```

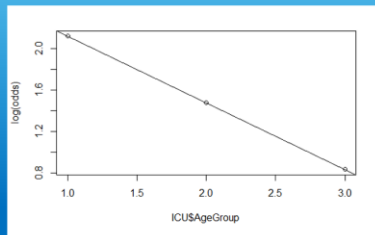
Method #1: AgeGroup as Quantitative Pred

```
> B0 = summary(ICUmod)$coef[1]
> B1 = summary(ICUmod)$coef[2]
> plot(jitter(Survive,amount=0.1)~AgeGroup,data=ICU)
> curve(logit(B0,B1,x),add=TRUE, col="red")
```



Method #1: AgeGroup as Quantitative Pred

```
> pi = logit(B0,B1,ICU$AgeGroup)
> odds = pi/(1-pi)
> plot(log(odds)~ICU$AgeGroup)
> abline(B0,B1)
```



Dummy Indicators for Multiple Categories

For a categorical predictor with k levels, we should use $k - 1$ dummy indicators.

$$X_1 = \begin{cases} 1 & \text{if Group \#1} \\ 0 & \text{otherwise} \end{cases} \quad \dots \quad X_{k-1} = \begin{cases} 1 & \text{if Group \#(k-1)} \\ 0 & \text{otherwise} \end{cases}$$

What happens to Group $\#k$? **Reference group**

Constant term is an estimate for Group $\#k$ and other coefficients are the differences from it.

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 ANOVA)

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

t.s. = G = *improvement* in $-2\log(L)$ 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

$1 - \text{pchisq}(8.57, 2)$

[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

	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 $i \geq 2$

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

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

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

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

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

nested lrt

Comparing Full to Reduced Models

```
ICUmod.4 = glm(Survive~factor(AgeGroup)+Emergency, data=ICU,  
family=binomial)
```

Null deviance: 200.16 on 199 degrees of freedom
Residual deviance: 171.16 on 196 degrees of freedom

FULL

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

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

Reduced

$$H_0: \beta_3 = 0 \quad \text{vs.} \quad H_a: \beta_3 \neq 0$$

```
1 - pchisq(summary(ICUmod.2)$deviance - summary(ICUmod.4)$deviance, 1)
```

Reject H_0 (p-value= 6.187652e-06). The Emergency term significantly improves the model.

This is also often called a “Drop-in-Deviance” test.

drop in deviance test

bestglm for Model Selection

Requirements to use `bestglm()`

1. Only the response and possible predictor variables should be within the dataframe

```
MedGPA.1 = within(MedGPA, {Accept = NULL})
```

2. The response variable must be the last column in the dataframe.

```
MedGPA.2 = MedGPA.1[,c(2:10,1)]
```

bestglm for model selection

bestglm for Model Selection

```
bestglm(MedGPA.2, family=binomial)
```

Bayesian Information Criteria

Morgan-Tatar search since family is non-gaussian.

BIC
BICq equivalent for q in (0.407407122288894, 0.830512766582046)
Best Model:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-39.4708940	12.2144951	-3.231480	0.001231510
SexM	-2.8403423	1.1580871	-2.452616	0.014182182
GPA	5.3344003	2.4807386	2.150327	0.031529326
PS	1.0247592	0.4722984	2.169728	0.030027451
WS	-0.7177605	0.3496614	-2.052730	0.040098780
BS	1.7914617	0.6434984	2.783941	0.005370279

Bayesian Information Criteria

$$k \log(n) - 2 \log(L(\theta))$$

n : sample size

k : number of predictors

θ : set of all parameters.

$L(\theta)$: probability of obtaining the data which you have, supposing the model being tested was a given.

- Selection criteria, similar to Mallows's Cp
- Smaller values indicate preferred models

BIC – Bayesian information criteria

Comparing Models by BIC

Δ BIC	Evidence against higher BIC
0 to 2	Little
2 to 6	Positive
6 to 10	Strong
>10	Very strong

```
MedGPA.2.bestglm = bestglm(MedGPA.2, family=binomial)
MedGPA.2.bestglm$BestModels
```

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