Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.

Professor

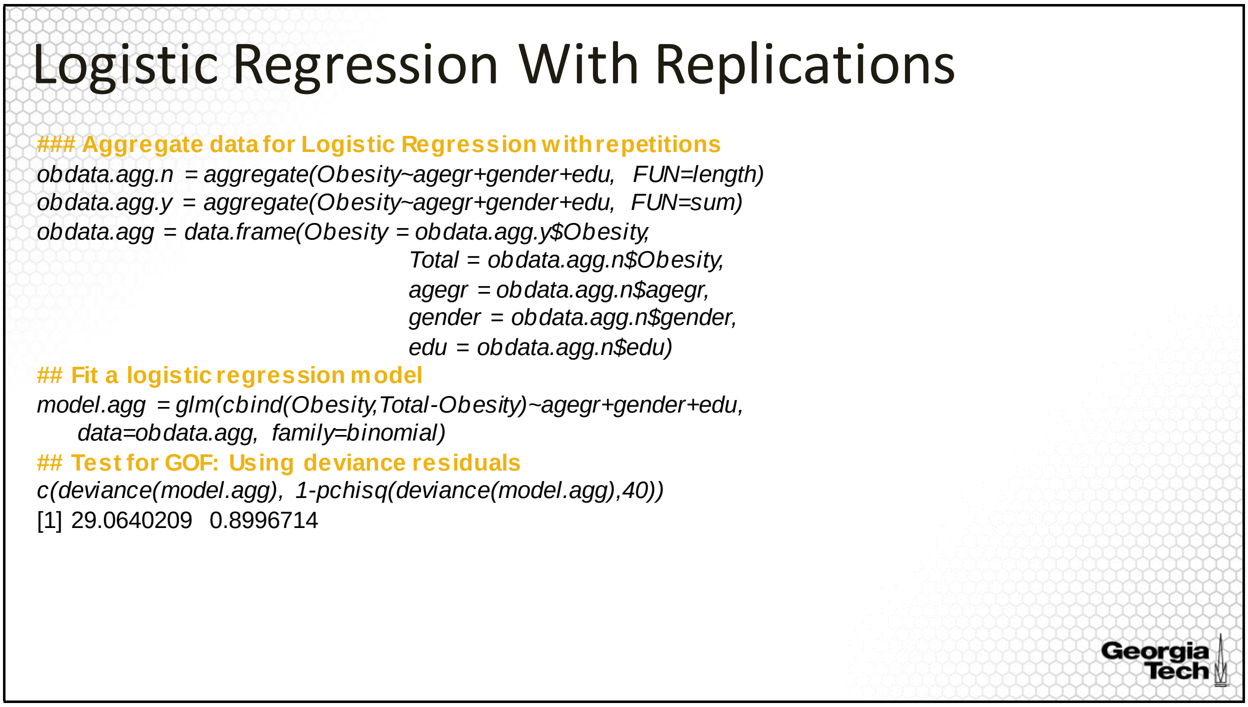
Stewart School of Industrial and Systems Engineering

The Demographics of Obesity:
Goodness of Fit

1

About This Lesson

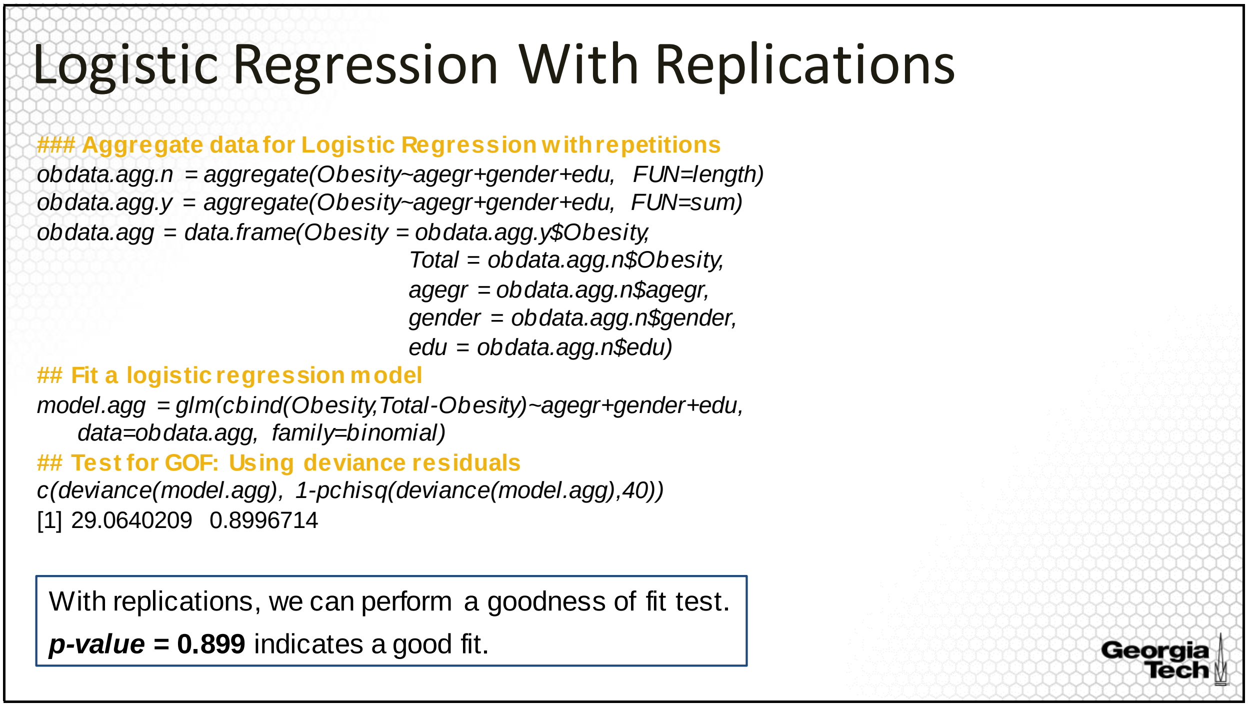
2



Logistic Regression With Replications

```
### Aggregate data for Logistic Regression with repetitions
obdata.agg.n = aggregate(Obesity~agegr+gender+edu, FUN=length)
obdata.agg.y = aggregate(Obesity~agegr+gender+edu, FUN=sum)
obdata.agg = data.frame(Obesity = obdata.agg.y$Obesity,
                        Total = obdata.agg.n$Obesity,
                        agegr = obdata.agg.n$agegr,
                        gender = obdata.agg.n$gender,
                        edu = obdata.agg.n$edu)
## Fit a logistic regression model
model.agg = glm(cbind(Obesity,Total-Obesity)~agegr+gender+edu,
    data=obdata.agg, family=binomial)
## Test for GOF: Using deviance residuals
c(deviance(model.agg), 1-pchisq(deviance(model.agg),40))
[1] 29.0640209  0.8996714
```

3



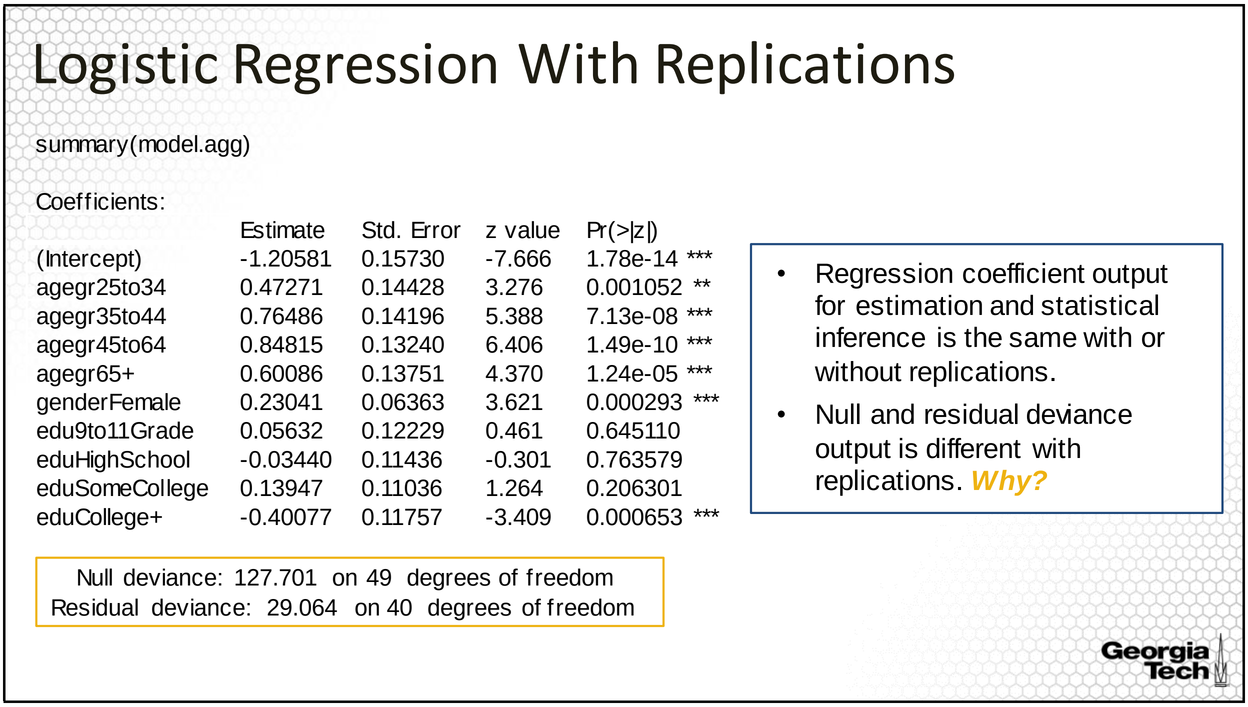
Logistic Regression With Replications

```
### Aggregate data for Logistic Regression with repetitions
obdata.agg.n = aggregate(Obesity~agegr+gender+edu, FUN=length)
obdata.agg.y = aggregate(Obesity~agegr+gender+edu, FUN=sum)
obdata.agg = data.frame(Obesity = obdata.agg.y$Obesity,
                        Total = obdata.agg.n$Obesity,
                        agegr = obdata.agg.n$agegr,
                        gender = obdata.agg.n$gender,
                        edu = obdata.agg.n$edu)
## Fit a logistic regression model
model.agg = glm(cbind(Obesity,Total-Obesity)~agegr+gender+edu,
    data=obdata.agg, family=binomial)
## Test for GOF: Using deviance residuals
c(deviance(model.agg), 1-pchisq(deviance(model.agg),40))
[1] 29.0640209  0.8996714
```

With replications, we can perform a goodness of fit test.

***p-value* = 0.899** indicates a good fit.

4

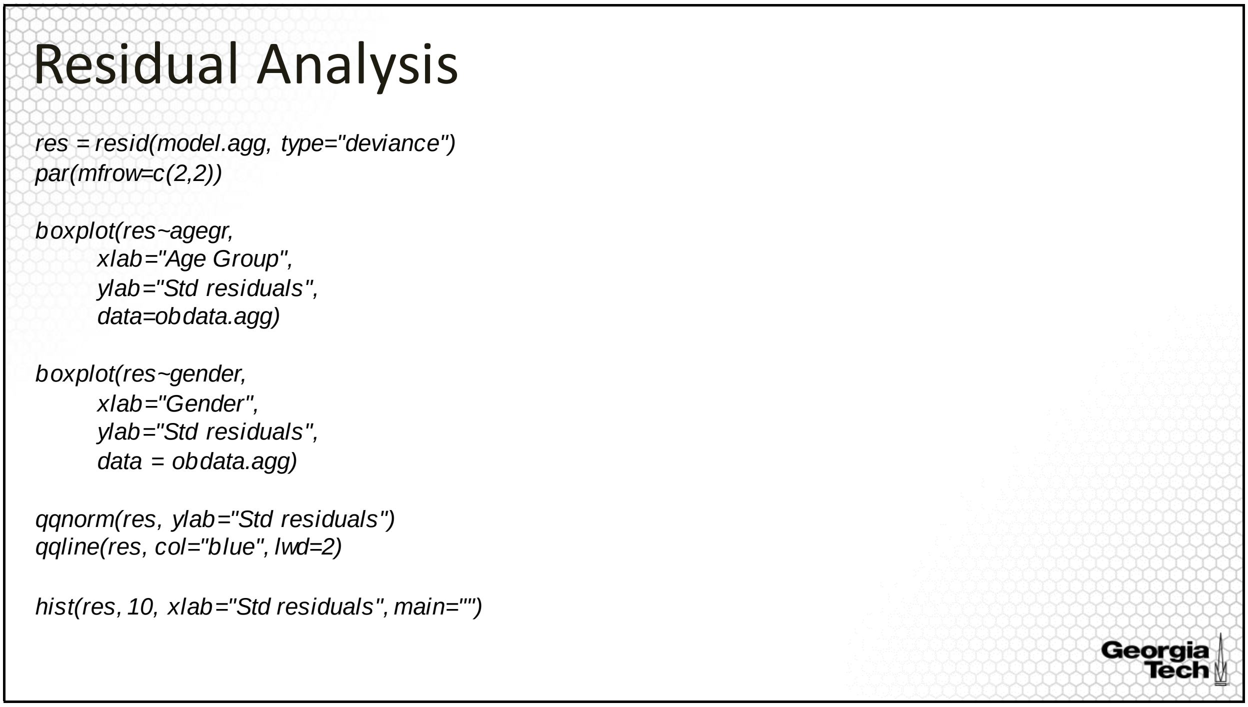


Logistic Regression With Replications

```
summary(model.agg)

Coefficients:
                Estimate   Std. Error  z value   Pr(>|z|)
(Intercept)     -1.20581   0.15730     -7.666    1.78e-14 ***
agegr25to34     0.47271    0.14428     3.276     0.001052  **
agegr35to44     0.76486    0.14196     5.388     7.13e-08 ***
agegr45to64     0.84815    0.13240     6.406     1.49e-10 ***
agegr65+        0.60086    0.13751     4.370     1.24e-05 ***
genderFemale    0.23041    0.06363     3.621     0.000293  ***
edu9to11Grade   0.05632    0.12229     0.461     0.645110
eduHighSchool   -0.03440   0.11436     -0.301    0.763579
eduSomeCollege  0.13947    0.11036     1.264     0.206301
eduCollege+     -0.40077   0.11757     -3.409    0.000653  ***

   Null deviance: 127.701  on 49  degrees of freedom
 Residual deviance:  29.064  on 40  degrees of freedom
```

- Regression coefficient output for estimation and statistical inference is the same with or without replications.
- Null and residual deviance output is different with replications. ***Why?***

5

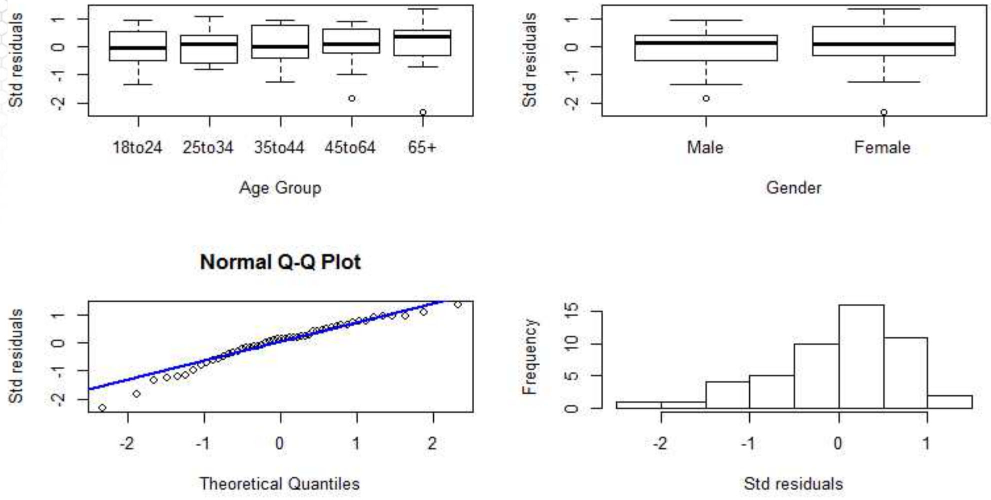


Residual Analysis

```
res = resid(model.agg, type="deviance")
par(mfrow=c(2,2))

boxplot(res~agegr,
        xlab="Age Group",
        ylab="Std residuals",
        data=obdata.agg)

boxplot(res~gender,
        xlab="Gender",
        ylab="Std residuals",
        data = obdata.agg)

qqnorm(res, ylab="Std residuals")
qqline(res, col="blue", lwd=2)

hist(res, 10, xlab="Std residuals", main="")
```

6

Residual Analysis



7

Prediction of Adult Obesity: Results

- Both gender and age group factors are statistically significant factors in explaining the variability in the classification of adults by obesity.
  - **But the fitted model with education, gender, and age group factors does not improve prediction.**
- After factor aggregation, goodness of fit can be performed.
- The *p-value* of the deviance test for goodness of fit is high, indicating good fit.
  - **But residual analysis suggests that there may be some departures from normality and thus from goodness of fit.**
- Models with different link functions or including interaction terms have not shown improvement. *(Results not shown in this lecture.)*
- The sample size is large enough for reliable statistical inference.

8

Prediction of Adult Obesity: Results

- Both gender and age group factors are statistically significant factors in explaining the variability in the classification of adults by obesity.
  - **But the fitted model with education, gender, and age group factors does not improve prediction.**
- After factor aggregation, goodness of fit can be performed.
- The *p-value* of the deviance test for goodness of fit is high, indicating good fit.
  - **But residual analysis suggests that there may be some departures from normality and thus from goodness of fit.**
- Models with different link functions or including interaction terms have not shown improvement. *(Results not shown in this lecture.)*
- The sample size is large enough for reliable statistical inference.

What can be done to improve the model fit and the predictive power?

- Include other factors in the model, such as income level, unemployment, race, and ethnicity, among others.
- Consider interaction terms between age, education, and gender groups and other factors.

Summary