BIOSTAT 651
Notes #1: Introduction to GLM

- Lecture Topics:
  - Class outline
  - Linear regression
  - Motivation: more general approaches
  - Generalized Linear Models (GLM)
  - Examples

Class outline

- Generalized Linear Model (GLM)
- First half: general framework
  - GLM Model: systematic and random components
  - Parameter estimation
  - Hypothesis test
- Second half: applications
  - Binary data
  - Multinomial data
  - Count data
  - Over-dispersion

Linear regression

- Linear regression: based on the assumption that error terms are *continuous* and *normally distributed*

- Linear regression model:

$$Y_i \;=\; \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + e_i$$

where

$$e_i \;\sim\; N(0, \sigma^2).$$

  - Relating p predictors for subject i to a response Y_i.
  - Assumptions can be summarized by:

$$Y_i \;\sim\; N(\mu_i, \sigma^2),$$
$$\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

where $\mathbf{x}_i^T \;=\; (1, X_{i1}, \ldots, X_{ip})$ and $\boldsymbol{\beta}^T \;=\; (\beta_0, \beta_1, \ldots, \beta_p)$.

Linear regression

- Assumptions:
  - Systematic component: predictor effect through linear regression on the mean (*linearity assumption)*

$$E[Y_i|\mathbf{x}_i] \;=\; \mu_i \;=\; \mathbf{x}_i^T\boldsymbol{\beta}$$

  - Random component: at each level of the predictor, variation in the response is characterized as

$$N(0, \sigma^2)$$

  - Independence (between subjects)

Generalizing the linear model

- In many applications, the distribution of a *continuous* response may be *non-normal*.
- In addition, the response may be *discrete*, e.g.,
  - binary ($Y_i = 1$, $Y_i = 0$)
  - unordered categorical or nominal ($Y_i \in \{1, \ldots, C\}$, with the ordering unimportant)
  - ordered categorical ($Y_i \in \{1, \ldots, C\}$, with the ordering of the index important)
  - count ($Y_i \in \{0, 1, \ldots, \infty\}$)
- In addition, a *non-linear* regression model relating the predictors to the mean may be needed.

Types of Responses

- Numeric response:
  - continuous

    eg., weight, blood pressure
  - discrete

    e.g., number of deaths, cancer cases, etc
- Categorical response:
  - nominal

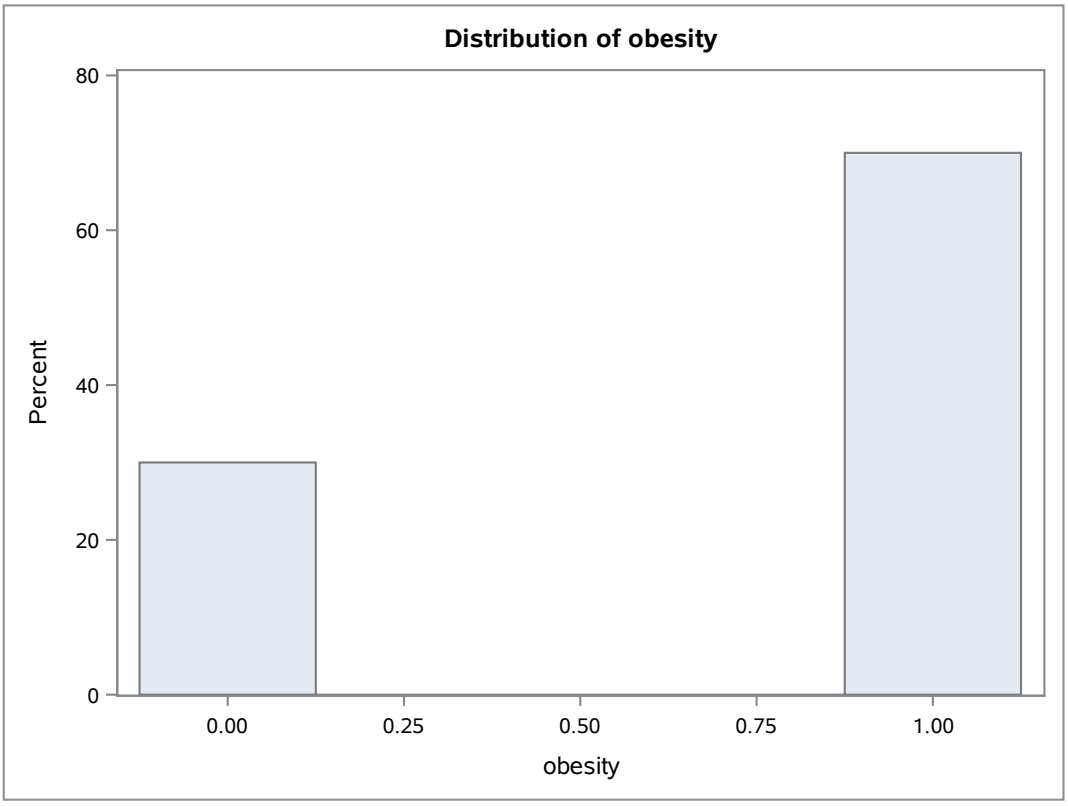
    e.g., blood type, gender, state
  - ordinal

    e.g., low/medium/high; age group; calendar period

Example: Binary Response

- Childhood obesity data:
  - Response: obesity ($Y_i = 1$ if obese; $Y_i = 0$ otherwise)
  - Predictors: Age (in years) and Smoking Status
- Fit a linear regression model with normality assumption
  - Imagine a histogram of Y
  - What assumptions of the linear model are clearly violated for binary responses?

The UNIVARIATE Procedure



Logistic and Linear Regression

- Recall our assumptions in linear regression:

  $Y_i \sim$ Normal

  $V(Y_i) = \sigma^2$, constant variance

- Suppose Y_i takes on one of only two possible values (0 or 1), as in our example
  - e.g., Y_i=0 (alive) or 1 (dead)
  - e.g., Y_i=0 (no lung cancer) or 1 (lung cancer present)

- Clearly, Y_i does not follow a normal distribution
- In fact, $Y_i \sim$ Bernoulli(π_i), where $\pi_i = \pi(\mathbf{x}_i) \equiv P(Y_i = 1|\mathbf{x}_i)$

Bernoulli Distribution

- we've set $\pi_i = P(Y_i = 1|\mathbf{x}_i)$
- recall that for binary random variables:

  $E[Y_i] = \pi_i$

  $V[Y_i] = \pi_i(1 - \pi_i)$

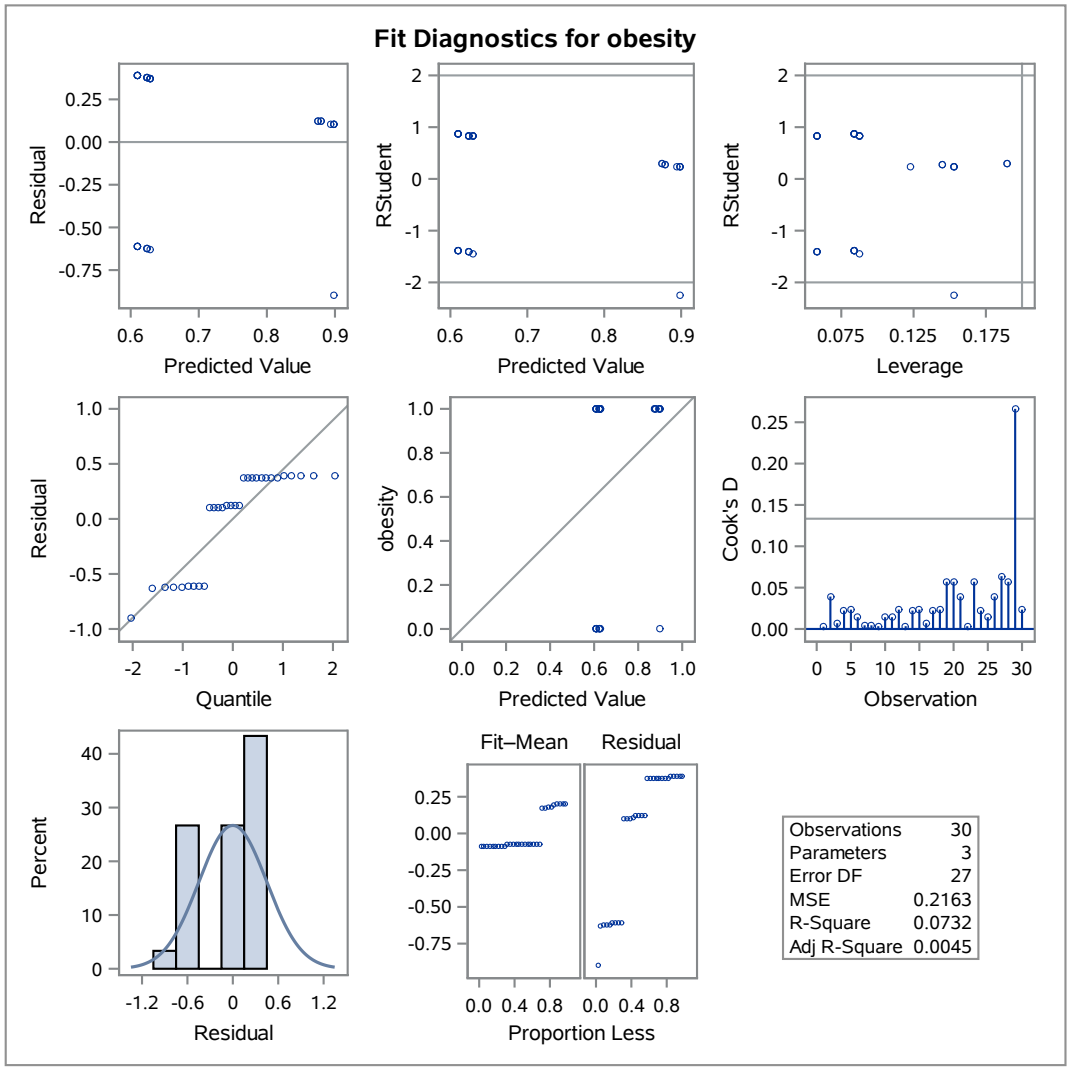
  note: $\pi_i = \pi(\mathbf{x}_i)$
- hence, constant variance assumption is inherently violated, as variance is a function of the mean
- therefore, linear regression is invalid: normality and constant variance assumptions blatantly violated

Example: Linear regression

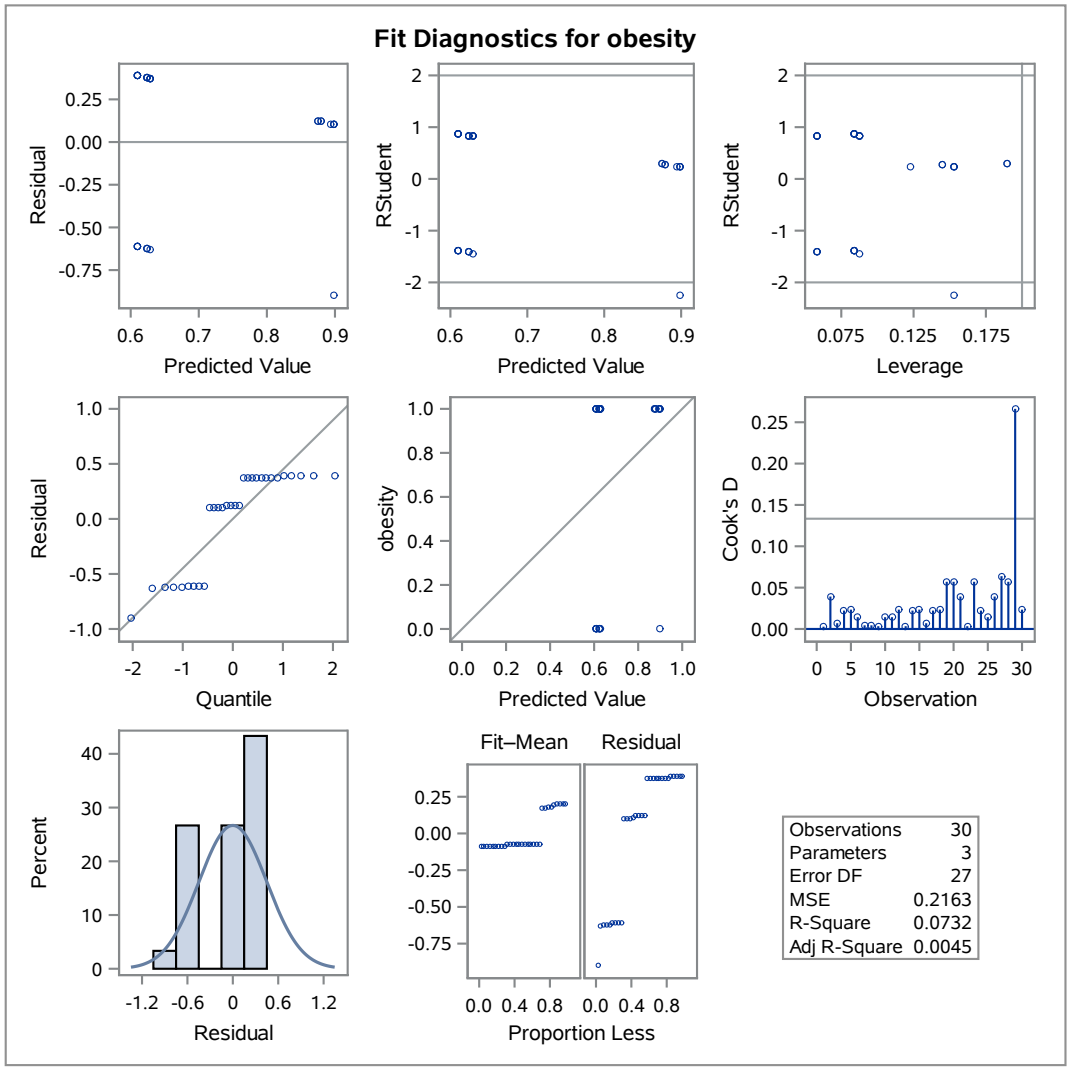
```
DATA Weights;
INPUT id wt age smoke obesity;
datalines;
1 22509.41 7 0 1
2 33452.27 7 1  0
3 13380.91 3 0  1
4 24947.45 8 1  1
5 15875.65 4 1 1
. . .


proc reg;
model obesity = age smoke;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: obesity



The REG Procedure
Model: MODEL1
Dependent Variable: obesity



Linear Regression, Binary Data

- suppose we ignore the binary nature of Y_i, and fit the following linear regression model:

$$E[Y_i] \quad = \quad \mathbf{x}_i^T \boldsymbol{\beta}$$

- we fit the linear model, obtaining $\widehat{\boldsymbol{\beta}}$ and the estimated means:

$$\widehat{Y}_i \quad = \quad \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}$$

- note: $0 \leq \widehat{Y}_i \leq 1$ need not hold, which is a major limitation since $E[Y_i]$ is a probability

- thus, we need to model some function of $E[Y_i]$, as opposed to $E[Y_i]$ itself

Logistic Function

- we need to find a transformation of $E[Y_i]$ to model as a linear function of covariates
- define the inverse-logit function:

$$\frac{e^x}{1+e^x}$$

- clearly,

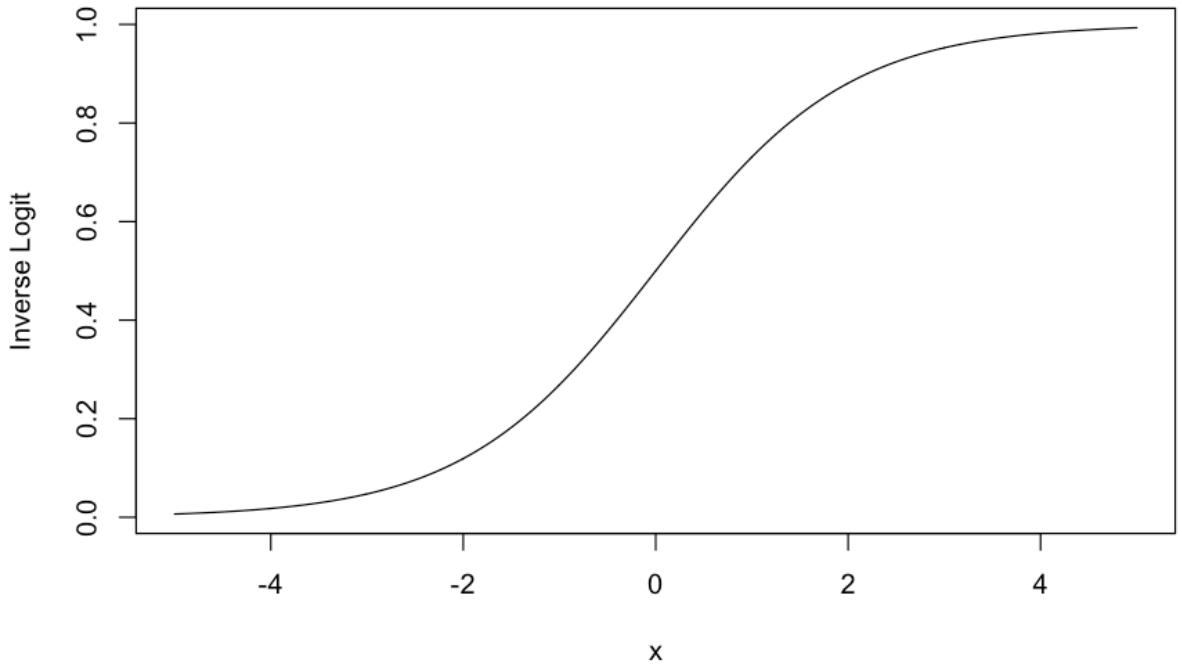
$$0 \leq \frac{e^x}{1+e^x} \leq 1, \text{for all } x$$

- this motivates the model:

$$E[Y_i] = \mu_i = \frac{e^{\mathbf{x}_i^T \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T \boldsymbol{\beta}}}, \text{ or}$$

$$log \frac{\mu_i}{1-\mu_i} = \mathbf{x}_i^T \boldsymbol{\beta}$$

- Inverse-logit function:



Example: Logistic regression

```
proc logistic;
model obesity (event='1') = age smoke;
run;
```

Logistic regression for binary responses

The LOGISTIC Procedure

Model Information	
Data Set	WORK.WEIGHTS
Response Variable	obesity
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	30
Number of Observations Used	30

Response Profile		
Ordered Value	**obesity**	**Total Frequency**
1	0	9
2	1	21

Probability modeled is obesity=1.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	38.652	40.176
SC	40.053	44.379
-2 Log L	36.652	34.176

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	2.4762	2	0.2899
Score	2.1960	2	0.3335
Wald	1.9249	2	0.3819

Logistic regression for binary responses

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	1.9242	1.6964	1.2867	0.2567
age	1	0.0266	0.2286	0.0135	0.9074
smoke	1	-1.5967	1.1526	1.9190	0.1660

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
age	1.027	0.656	1.607
smoke	0.203	0.021	1.939

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	57.1	**Somers' D**	0.349
Percent Discordant	22.2	**Gamma**	0.440
Percent Tied	20.6	**Tau-a**	0.152
Pairs	189	**c**	0.675

Example: Ordered Response

- Example: The University of Regensburg conducted an investigation on senior psychology students regarding future job prospects. One of the key questions was whether they expected to find adequate employment after obtaining their degree.

- Response: Ordered categorical 1-3:
  1. Don't expect to find adequate employment
  2. Not sure
  3. Will obtain adequate employment immediately
     - Predictor: Age in years
- Scale of response: ordered categorical

Example: Ordered Response (continued)

Age	Response		
in years	1	2	3
19	1	2	0
20	5	18	2
21	6	19	2
22	1	6	3
23	2	7	3
24	1	7	5
25	0	0	3
26	0	1	0
27	0	2	1
29	1	0	0
30	0	0	2
31	0	1	0
34	0	1	0

- Multinomial logistic regression:
  - Systematic component: logit function
  - Random component: multinomial distribution

Example: Count Response

- Cellular Differentiation (Piegorsch, Weinberg & Margolin, 1988):
  - Interest in the effect of two agents of immuno-activating ability that may introduce cell differentiation.
  - Do the agents TNF (tumor necrosis factor) and IFN (interferon) simulate cell differentiation independently or is there a synergetic effect?
- Response: number of cells that exhibited markers after exposure was recorded.
- Covariates:

  TNF

  IFN

Number of cells differentiating	Dose of TNF (U/ml)	Dose of IFN (U/ml)
11	0	0
18	0	4
20	0	20
39	0	100
22	1	0
38	1	4
52	1	20
69	1	100
31	10	0
68	10	4
69	10	20
128	10	100
102	100	0
171	100	4
180	100	20
193	100	100

Example: Count Response (continued)

- Response: count
  - Poisson distribution often used for modeling counts
  - regression version of Poisson model:

    $E[Y_i] = \mathbf{x}_i^T \boldsymbol{\beta}$

- Q: What issues arise if linear regression is used?

  consider properties of Poisson variate ...

Notation: Counts, Rates, Person-Time

- Before deciding on a regression method, we set up some notation:
  - Y_i = event count, cell i
  - Y_i can take values $0,\ 1,\ 2,\ 3\ldots$
  - λ_i = event rate, cell i

$$\lambda_i = E\left[Y_i\right] \tag{1}$$

  - covariates: ${\mathbf{x}_i}^T = (1, X_{i1}, \ldots, X_{ip})$

Linear Regression for Count Data

- We return to our question of an appropriate modeling strategy
- Q: Could we model Y_i using linear regression?
  - Y_i is discrete and non-negative; violation of Normality assumption
  - $\lambda_i = E[Y_i]$ is an event rate; should be positive.
  - usually, when Y_i is a count, $V[Y_i]$ is related to $E[Y_i]$, in violation of the constant variance assumption

Poisson Regression: Deriving the Model

- We now set up our model equation...
- Begin by writing the rates as a linear function:

  $$\lambda_i = \mathbf{x}_i{}^T \boldsymbol{\beta}$$

- If we were to fit this model, there is no guarantee that $\mathbf{x}_i{}^T \widehat{\boldsymbol{\beta}} > 0$
- How did we handle a parallel issue when deriving our model for binary data?

Poisson Regression Model

- Solution: since $e^x \geq 0$ for all x,

$$\begin{aligned} \lambda_i &= e^{\mathbf{x}_i^T \boldsymbol{\beta}} \\ \log(\lambda_i) &= \mathbf{x}_i^T \boldsymbol{\beta} \end{aligned} \tag{2}$$

- Distribution assumption

$$Y_i \sim Poisson(\lambda_i)$$

- We fit this model through the method of maximum likelihood

Summary

- Linear regression is inappropriate for each of these examples.
  - need a more general regression framework accounting for response data having a variety of measurement scales.
  - methods for model fitting and inference under this framework.
- Ideally, some elements of linear regression should carry over.
- Generalizations to more complex settings (correlated data, censored observations, etc) will be necessary in many applications (BIOSTAT 653, BIOSTAT 675)

Comment

All models are wrong, but some are useful.