Lecture Outline

- Regression models
  - Linear vs logistic vs Poisson vs GLM variants

- Adjustment for confounders
  - Linear continuous
  - Transformed continuous
  - Polynomial continuous
  - Dummy variables
    - Quantile categorization vs scientific categorization
    - Piecewise linear
    - Splines

Common Regression Models

- According to the parameter and contrast across groups
  - Mean (differences) → Linear regression (a GLM)
  - Mean (ratios) → GLM with log link (a GLM)
  - Geom Means (ratios) → Linear regression on logs (a GLM)
  - Odds (ratios) → Logistic regression (a GLM)
  - Rates (ratios) → Poisson regression (a GLM)
  - Hazards (ratios) → Proportional Hazards regr
  - Quantiles (ratios) → Parametric (AFT) survival regr

- A special case for this course
  - Cumulative incidence → Exponential regression (a GLM)
General Regression

- General notation for variables and parameter
  \[ Y_i \text{ Response measured on the } i\text{th subject} \]
  \[ X_i \text{ Value of the POI for the } i\text{th subject} \]
  \[ W_{i1}, W_{i2}, \ldots \text{ Value of adjustment variables for the } i\text{th subject} \]
  \[ \theta_i \text{ Parameter (summary measure) of distribution of } Y_i \]

- The parameter might be the mean, geometric mean, odds, rate, instantaneous risk of an event (hazard), etc.

- (Sometimes we will use multiple regression predictors to model the scientific POI
  - E.g., linear and quadratic terms, dummy variables, etc.)

Multiple Regression

- General notation for multiple regression model
  \[ g(\theta) = \beta_0 + \beta_1 \times X_1 + \beta_2 \times W_{i1} + \beta_3 \times W_{i2} + \ldots \]
  \[ g(\theta) \text{ "link" function used for modeling} \]
  \[ \beta_0 \text{ "Intercept"} \]
  \[ \beta_i \text{ "Slope for Pred of Interest } X_i \text{"} \]
  \[ \beta_j \text{ "Slope for covariate } W_{i,j-1} \text{"} \]

- The link function can usually be viewed as either none (means) or log (geometric mean, odds, hazard) but some exceptions
  - log odds viewed by many as logit mean
  - complementary log log link for proportions measuring cumulative incidence, which relate to log link on exponential hazards
    - \[ \log(-\log \rho) = \log \lambda + \log \Delta t \]

Generalized Linear Model (GLM)

- A special subset of my general regression model is the "Generalized Linear Model"
  - \( \theta \) is the mean of response \( Y \)
  - \( Y \) has a 1-parameter exponential family distribution
    - Includes normal (Gaussian), binomial, Poisson, exponential, gamma, negative binomial, log-normal...

- Estimating equations derived from maximum likelihood theory in these "regular" models (i.e., distributions that satisfy certain technical assumptions)
  - MLEs are "consistent": arbitrarily close to the truth with large \( n \)
  - MLEs are "asymptotically efficient": have greatest precision possible in general (meet Cramer-Rao lower bound)

GLM Estimating Equations

- Use maximum likelihood estimation assuming a parametric distribution for the data
  \[ Y_i \mid \tilde{X}_i \sim \tilde{x}_i \sim (\mu_i, \sigma_i^2 = V(\mu_i)) \]
  \[ \tilde{x}_i \hat{\beta} = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots \]

"Score function estimating equation":
\[ U(\hat{\beta}) = \sum_{i=1}^{n} \frac{Y_i - \mu_i}{\sigma_i^2} \frac{\partial \mu_i}{\partial \beta} = \sum_{i=1}^{n} \frac{Y_i - \mu_i}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta} \]

Find regression parameter estimates \( \hat{\beta} \) to satisfy:
\[ U(\hat{\beta}) = \sum_{i=1}^{n} \frac{Y_i - \hat{\mu}_i}{V(\hat{\mu}_i)} \frac{\partial \mu_i}{\partial \beta} \bigg|_{\hat{\mu}_i, \beta} = 0 \]

- Quasi-likelihood and generalized estimating equations (GEE) use same general equation, but do not assume parametric distribution
  - Only model mean and mean-variance relationship
“Family”: Mean Variance Relationship

- If we trust our linear model and data distribution completely, then using the correct mean-variance relationship will be more efficient
  - Weights observations less if the variance is greater
  - Works because every weighted average will correctly estimate the straight line relationship

\[
\begin{align*}
\text{Gaussian} & \quad U(\beta) = \sum_{i=1}^{n} \frac{Y_i - \mu_i}{\sigma} \frac{\partial \mu_i}{\partial \beta} \\
\text{Bernoulli} & \quad U(\beta) = \sum_{i=1}^{n} \frac{Y_i - \mu_i}{\mu_i(1 - \mu_i)} \frac{\partial \mu_i}{\partial \beta} \\
\text{Poisson} & \quad U(\beta) = \sum_{i=1}^{n} \frac{Y_i - \mu_i}{\mu_i} \frac{\partial \mu_i}{\partial \beta} \\
\text{Exponential} & \quad U(\beta) = \sum_{i=1}^{n} \frac{Y_i - \mu_i}{\mu_i^2} \frac{\partial \mu_i}{\partial \beta}
\end{align*}
\]

Role of Link Function

- Conceptually, we can use any link function with any distribution family

  - There are often advantages of using the “canonical” link for each family: These tend to be the default value in statistical software
    - Gaussian distribution \( \Rightarrow \) Identity link
    - Binomial distribution \( \Rightarrow \) Logit link (on means)
    - Poisson distribution \( \Rightarrow \) Log link
    - Exponential distribution \( \Rightarrow \) Inverse link: \( g(\mu) = 1/\mu \)

Binary Data: Role of Link Function

- With Bernoulli data, we might consider three different link functions

\[
\begin{align*}
\text{Identity} & \quad U(\beta) = \sum_{i=1}^{n} \frac{Y_i - X_i \beta}{X_i \beta(1 - X_i \beta)} X_i \Rightarrow RD \\
\text{Log} & \quad U(\beta) = \sum_{i=1}^{n} \frac{Y_i - e^{X_i \beta}}{1 - e^{X_i \beta}} X_i \Rightarrow RR \\
\text{Logit} & \quad U(\beta) = \sum_{i=1}^{n} \left( Y_i - \frac{e^{X_i \beta}}{1 + e^{X_i \beta}} \right) X_i \Rightarrow OR
\end{align*}
\]

Uses of Regression

- Modeling questions about associations
  - “Defining contrasts” for associations between response and POI
  - “Defining contrasts” for detecting effect modification

- Enabling comparisons across POI groups that are “otherwise similar” with respect to other variables
  - Adjusting for confounding
  - Adjusting to gain precision
Borrowing Information

- Use other groups to make estimates in groups with sparse data
- Intuitively: 67 and 69 year olds would provide some relevant information about 68 year olds
- Assuming straight line relationship in modeled covariates tells us how to adjust data from other (even more distant) age groups
  - If we do not know about the exact functional relationship, we might want to borrow information only close to each group

Defining “Contrasts”

- Define a comparison across groups to use when answering scientific question
- If straight line relationship in parameter, slope for POI compares parameter between groups differing by 1 unit in X when all other covariates in model are equal
- If nonlinear relationship in parameter, slope is average comparison of parameter between groups differing by 1 unit in X “holding covariates constant”
  - Statistical jargon: a “contrast” across the groups
- If multiple regression predictors model the POI, interpretation of the contrast is more difficult

Comparison of Models

- The major difference between regression models is interpretation of the parameters
  - Summary: Mean, geometric mean, odds, hazards
  - Comparison of groups: Difference, ratio
- Issues related to inclusion of covariates remain the same
  - Address the scientific question
    - Predictor of interest; Effect modifiers
  - Address confounding
  - Increase precision

Interpretation of Parameters

- Intercept
  - Corresponds to a population with all modeled covariates equal to zero
  - Most often outside range of data; quite often impossible; very rarely of interest by itself
- Slope
  - A comparison between groups differing by 1 unit in corresponding covariate, but agreeing on all other modeled covariates
    - Identity link: a difference in the summary measure
    - Log link: a ratio in the summary measure (difference in logs)
  - Sometimes impossible to use this definition when modeling interactions or complex curves
    - I most often resort to looking at predicted summary measures
Regression with Identity Link

- E.g., modeling mean of response $Y$ on predictors $X, W_1, W_2, \ldots$

Model

\[ \theta_i = \beta_0 + \beta_1 x_i + \beta_2 x_i + \cdots \]

\[ X_i = 0 \quad W_{ji} = 0 \quad \theta = \beta_0 \]

\[ X_i = x \quad W_{ji} = w_{ji} \quad \theta = \beta_0 + \beta_1 x + \beta_2 w_{ji} + \cdots \]

\[ X_i = x + 1 \quad W_{ji} = w_{ji} \quad \theta = \beta_0 + \beta_1 x + \beta_1 + \beta_2 w_{ji} + \cdots \]

Difference of $\theta$: \[ \beta_i \]

---

Linear Regression: Ordinary, Weighted

- Use least squares estimation
  - Ordinary least squares: assume equal variance
  - Weighted least squares: use actual variances

\[ Y_i | \bar{X}_i = \bar{x} \sim (\bar{x}, \tilde{\beta}, \sigma^2) \]

\[ \bar{x} \tilde{\beta} = \beta_0 + \beta_1 x_i + \beta_2 + \cdots \]

"Least squares": Minimize \[ \sum_{i=1}^{n} \frac{(Y_i - \bar{x} \tilde{\beta})^2}{\sigma_i^2} \]

"Score function estimating equation": \[ U(\beta) = \sum_{i=1}^{n} \frac{Y_i - \bar{x} \tilde{\beta}}{\sigma_i} \]

Find regression parameter estimates $\hat{\beta}$ to satisfy:

\[ U(\hat{\beta}) = \sum_{i=1}^{n} \frac{Y_i - \bar{x} \hat{\beta}}{\sigma_i} = 0 \]

---

Regression with Identity Link

- Most common example: Linear regression for difference in means
  - Ordinary (unweighted) least squares
    - Optimal (in some sense): independent, homoscedastic response
  - Weighted least squares: weighting by inverse variance
    - Optimal (in some sense): independent, heteroscedastic response
  - General least squares: weighting by covariance matrix
    - Optimal (in some sense) for correlated response data

- Optimality: For any distribution having a variance, then
  - "Best linear unbiased estimate": greatest precision among all unbiased estimators that are a linear combination of the data
  - Gauss-Markov theorem
    - If response is normal within groups and linear model correct, then
      - Unbiased and efficient (greatest possible precision)
      - Also MLE

---

Properties of Least Squares Estimates

- In unweighted regression with no covariates
  - Estimated intercept is the sample mean
- In unweighted regression with a single binary estimator $X$
  - Estimated intercept is the sample mean for group with $X=0$
  - Estimated slope is the difference in sample means
    - Group with $X=1$ minus group with $X=0$
- In unweighted regression with as many parameters as distinct combinations of predictors
  - The fitted value for every group will correspond exactly to the sample mean
- In weighted regressions the last two may not hold, unless all subjects with the same covariate values have the same weights
Linear Regression: Binary Response

• A mean-variance relationship: mean $\mu = p \rightarrow$ variance $p(1-p)$
• Optimal (BLUE) estimates would use weighted least squares
  – We have to estimate the weights in an iterative search

"Score function estimating equation":

$$U(\beta_j) = \sum_{i=1}^{n} \frac{Y_i - \hat{x}_j \beta_j}{\sigma_i^2} x_{ji} = \sum_{i=1}^{n} \frac{Y_i - \hat{x}_j \beta_j}{\hat{x}_j \beta_j (1 - \hat{x}_j \beta_j)} x_{ji} = 0$$

Find regression parameter estimates $\hat{\beta}_j$ to satisfy:

Stata: Binary Response with Identity Link

• Weighted least squares using iteratively estimated weights:
  - glm $y \ x1 \ldots$,family(binomial) link(identity)
  - binreg $y \ x1 \ldots$,rd
  - cs $y \ x1$

• Ordinary (unweighted) least squares:
  - glm $y \ x1 \ldots$,family(gaussian) link(identity)
  - regress $y \ x1 \ldots$

• Robust standard errors can be obtained with either
  – Specify option robust with glm or regress
  – Specify option vce(robust) with binreg

Examples: From Homework #2

• Modeling females’ risk of dying within 4 years as a function of
  – Estrogen use (0 or 1)
  – Prior history of cardiovascular disease (0 or 1)
  – Age (65 – 100 years)

• Derived variable $estr\_prev = estrogen \ast prevdis$

• Comparisons
  – No covariates
  – Binary covariate
  – Saturated model with two binary covariates
  – Continuous covariate

Relative Advantages

• IF the linear model is correct, weighting by the estimated mean-
  variance relationship is most efficient
• HOWEVER, if the linear model is not correct, we can end up with
  negative variance estimates
  – Not a good thing to use when weighting
• The unweighted analysis
  – is unbiased if the linear model is correct,
  – interpretable as a linear trend when the linear model is not exactly
    correct, and
  – can be adjusted for the heteroscedasticity when making inference
Regression with Log Link

- E.g., modeling geometric mean, odds, rate, hazard of response \( Y \) on predictors \( X, W_1, W_2, \ldots \)

Model:

\[
\log(\theta) = \beta_0 + \beta_1 \times X_i + \beta_2 \times W_i + \cdots
\]

\[
X_i = 0 \quad W_{ji} = 0 \quad \log(\theta) = \beta_0
\]

\[
X_i = x \quad W_{ji} = w_{ji} \quad \log(\theta) = \beta_0 + \beta_1 \times x + \beta_2 \times w_{ji} + \cdots
\]

\[
X_i = x + 1 \quad W_{ji} = w_{ji} \quad \log(\theta) = \beta_0 + \beta_1 \times x + \beta_1 + \beta_2 \times w_{ji} + \cdots
\]

Difference of \( \log(\theta) \):

\[
\beta_i
\]

Regression with Log Link: Back Transform

- E.g., modeling geometric mean, odds, rate, hazard of response \( Y \) on predictors \( X, W_1, W_2, \ldots \)

Model:

\[
\log(\theta) = \beta_0 + \beta_1 \times X_i + \beta_2 \times W_i + \cdots
\]

\[
X_i = 0 \quad W_{ji} = 0 \quad \theta = e^{\beta_0}
\]

\[
X_i = x \quad W_{ji} = w_{ji} \quad \theta = e^{\beta_0 + \beta_1 \times x + \beta_2 \times w_{ji} + \cdots}
\]

\[
X_i = x + 1 \quad W_{ji} = w_{ji} \quad \theta = e^{\beta_0 + \beta_1 \times x + \beta_1 + \beta_2 \times w_{ji} + \cdots}
\]

Ratio of \( \theta \):

\[
e^{\beta_i}
\]

Logistic Regression

\[
Y_i | \bar{X}_i = \bar{x}_i \sim B \left( 1, \frac{e^{x_i \beta}}{1 + e^{x_i \beta}} \right)
\]

"Score function estimating equation":

\[
U(\beta) = \sum_{m} \left( Y - \frac{e^{x_i \beta}}{1 + e^{x_i \beta}} \right) X_i
\]

Find regression parameter estimates \( \hat{\beta} \) to satisfy:

\[
U(\hat{\beta}) = 0
\]

Stata: Logistic regression

- Weighted least squares using iteratively estimated weights:
  - logistic y x1 ..., or logit y x1 ...
  - glm y x1 ..., family(binomial) link(logit)
  - binreg y x1 ...

- Robust standard errors can be obtained with either
  - Specify option robust with glm or logistic / logit
  - Specify option vce(robust) with binreg
Poisson Regression

\[ Y_i | \lambda_i = \lambda_i \sim P(\lambda_i) \quad \log(\lambda_i) = \bar{x}_i \beta = \beta_0 + x_{i1} \beta_1 + x_{i2} \beta_2 + \ldots \]

"Score function estimating equation":

Logit link

\[ U(\beta) = \sum_{i=1}^{n} (Y_i - t_i e^{\beta})^2 \]

Find regression parameter estimates \( \hat{\beta} \) to satisfy:

\[ U(\hat{\beta}) = 0 \]

Note that we formulate regression with "exposure" having a known coefficient of 1 for \( \log(t) \):

\[ \log(\lambda_{it}) = \log(t) + \beta_0 + x_{i1} \beta_1 + x_{i2} \beta_2 + \ldots \]

Stata: Poisson regression

- Weighted least squares using iteratively estimated weights:
  - `poisson y x1 ..., exposure(t)`
  - `glm y x1 ..., family(poisson) link(log)`

- Robust standard errors can be obtained with either
  - Specify option `robust` with `glm` or `poisson`
  - Specify option `vce(robust)` with `binreg`

Properties of MLEs with Canonical Link

- In regression with no covariates
  - Estimated intercept is the sample mean
- In regression with a single binary estimator \( X \)
  - Estimated intercept is the sample mean for group with \( X=0 \)
  - Estimated slope is the difference in sample means
    - Group with \( X=1 \) minus group with \( X=0 \)
- In regression with as many parameters as distinct combinations of predictors
  - The fitted value for every group will correspond exactly to the sample mean

Stata: Regression Commands

- Linear regression (means, proportions)
  - `regress Y X W1 W2 W3, robust`
- Linear regression (geometric means)
  - `regress logY X W1 W2 W3, robust`
- Logistic regression (odds)
  - `logistic Y X W1 W2 W3, [robust]`
  - `logit Y X W1 W2 W3, [robust]`
- Poisson regression (rates)
  - `poisson Y W1 W2 W3, robust`
- Proportional hazards regression (hazards)
  - `stcox Y X W1 W2 W3, robust`
### Stata: glm Regression Commands

- **Difference of means, proportions (linear regression)**
  - `glm Y X W1, robust family(gaussian) link(identity)`

- **Difference of log means, log proportions**
  - `glm Y X W2, robust link(log)`

- **Difference of log geometric means (linear regression on logs)**
  - `glm logY X W1 W2, robust`

- **Difference of log odds (logistic regression)**
  - `glm Y X W1, [robust] family(binomial) link(logit)`

- **Difference of log rates (Poisson regression)**
  - `glm Y X W1 W2, robust family(poisson) link(log)`

- **Back transforming regression parameters with log link**
  - `glm Y X W1 W2, robust family(...) link(...) eform`

### Stratification

- **When we are uninterested in making inference about associations between some (categorical) variables and the response, we can consider stratified analyses**
  - Perform an analysis within each stratum defined by a unique combination of the stratifying variables
  - Compute a (weighted) average of the results from those analyses

- **But stratification adjusts for covariates and all interactions among those covariates**
  - E.g., sex, race, and the sex-race interaction

- **Our habit in regression is to just adjust for the covariates (the “main effect”), and consider interactions less often**

### Combining Across Subgroups

- **Based on the properties of independent, normally distributed estimates**

  For independent \( \hat{\theta}_1 \sim N(\theta_1, se_1^2) \) \( \hat{\theta}_2 \sim N(\theta_2, se_2^2) \)

  \[
  a \hat{\theta}_1 + b \hat{\theta}_2 \sim N(a\theta_1 + b\theta_2, a^2se_1^2 + b^2se_2^2)
  \]

  \[
  \hat{\theta}_1 - \hat{\theta}_2 \sim N(\theta_1 - \theta_2, se_1^2 + se_2^2)
  \]

  \[
  \frac{\hat{\theta}_1}{\hat{\theta}_2} \sim N\left(\frac{\theta_1}{\theta_2}, \frac{1}{\theta_2^2}\left(se_1^2 + \frac{\theta_1^2}{\theta_2^2}se_2^2\right)\right)
  \]

### Stratification vs Regression

- **Generally, any stratified analysis could be performed as a regression model**

- **Stratification adjusts for covariates and all interactions among those covariates**
  - E.g., sex, race, and the sex-race interaction

- **Any covariates modeled in each stratum’s analysis would have to be modeled as interactions**
  - E.g., sex stratified analyses of response adjusted for age could be modeled in an unstratified analysis with sex-age interaction

- **Our habit in regression is to just adjust for the covariates (the “main effect”), and consider interactions less often**
Example

- We are interested in exploring the incidence of colorectal cancer by birthplace among whites in the US
  - Cases identified through the SEER registry 1973-1987
  - Available data: birthplace (US, 25 non-US, unknown), age in 5 year groups, sex
  - Denominator data from US census data

- Compare
  - Directly standardized rates using Stata `ir`
  - Poisson regression using Stata `poisson`

Adjustment for Covariates

- We “adjust” for other covariates

- Define groups according to
  - Predictor of interest, and
  - Other covariates

- Compare the distribution of response across groups which
  - differ with respect to the Predictor of Interest, but
  - are the same with respect to the other covariates
  - “holding other variables constant”

Adjustment for Confounders

- We “adjust” for other covariates

- Define groups according to
  - Predictor of interest, and
  - Other covariates

- Compare the distribution of response across groups which
  - differ with respect to the Predictor of Interest, but
  - are the same with respect to the other covariates
  - “holding other variables constant”
Adjusted Analyses

• We include covariates in a regression model for three reasons
  – Modeling our primary question about associations
    • Predictor of interest
    • Effect modification
  – Adjusting for confounders
  – Adding precision to our analysis

• As a rule, it is most important to accurately model the true relationship between the response and confounders
  – We answer questions about general trends in our POI and effect modifiers
  – We tend to gain most precision (if any) from rough adjustment for precision variables

Linear Predictors

• The most commonly used regression models use “linear predictors”
  – “Linear” refers to linear in the parameters
  – The modeled predictors can be transformations of the scientific measurements
    • Examples

\[
g[\theta | X_i, W_i] = \beta_0 + \beta_{\log X} \times \log(X_i)
\]
\[
g[\theta | X_i, W_i] = \beta_0 + \beta_X \times X_i + \beta_X^2 \times X_i^2
\]

Transformations of Predictors

• We transform predictors to provide more flexible description of complex associations between the response and some scientific measure
  • Threshold effects
  • Exponentially increasing effects
  • U-shaped functions
  • S-shaped functions
  • etc.

Ex: Cubic Relationship

• Linear regression

FEV vs Height in Children
Ex: Threshold Effect of Dose?

- Linear regression

Plasma Beta-carotene at 3 months by Dose

Plasma Beta-carotene at 9 months by Dose

Ex: U-shaped Trend?

- Inflammatory marker vs cholesterol
  - Linear regression

Ex: S-shaped trend

- In vitro cytotoxic effect of Doxorubicin with chemosensitizers
  - Poisson regression

“1:1 Transformations”

- Sometimes we transform 1 scientific measurement into 1 modeled predictor
  - Ex: log transformation will sometimes address apparent “threshold effects”
  - Ex: cubing height produces more linear association with FEV
Log Transformations

- Linear regression

```
<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>log X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>2</td>
</tr>
<tr>
<td>0.5</td>
<td>3</td>
</tr>
<tr>
<td>1.0</td>
<td>4</td>
</tr>
<tr>
<td>1.5</td>
<td>5</td>
</tr>
<tr>
<td>2.0</td>
<td>6</td>
</tr>
</tbody>
</table>
```

"1:Many Transformations"

- Sometimes we transform 1 scientific measurement into several modeled predictor
  - Ex: "polynomial regression"
  - Ex: "dummy variables" ("factored variables")
  - Ex: "piecewise linear"
  - Ex: "splines"

Polynomial Regression

- Fit linear term plus higher order terms (squared, cubic, …)
- Can fit arbitrarily complex functions
  - An n-th order polynomial can fit n+1 points exactly
- Generally very difficult to interpret parameters
  - I usually graph function when I want an interpretation
  - Use Stata’s `predict` command
- Special uses
  - 2nd order (quadratic) model to look for U-shaped trend
  - Test for linearity by testing that all higher order terms have parameters equal to zero

Stata: “Predicted Values”

- After computing a regression model, Stata will provide “predicted values” for each case
  - Covariates times regression parameter estimates for each case
  - “predict varname”
Dummy Variables

- Indicator variables for all but one group
  - This is the only appropriate way to model nominal (unordered) variables
    - E.g., for marital status
      - Indicator variables for
        » married (married = 1, everything else = 0)
        » widowed (widowed = 1, everything else = 0)
        » divorced (divorced = 1, everything else = 0)
        » (single would then be the intercept)
  - Often used for other settings as well
  - Equivalent to “Analysis of Variance (ANOVA)”

Ex: Interpretation of Slopes

- Based on coding used
  - Intercept corresponds to “reference” group
  - Slope for other terms compare each group to the “reference”

- Tests for association must test all predictors together

- Need to be very careful in overinterpreting “statistical significance” among groups
  - (More later)

Stata: Dummy Variables

- Stata has a facility to automatically create dummy variables
  - Old way: Prefix regression commands with “xi: ”
  - Now just prefix variables to be modeled as dummy variables with “i.varname”
  - (Stata will drop the lowest category by default)

Continuous Variables

- We can also use dummy variables to represent continuous variables
  - Continuous variables measured at discrete levels
    - E.g., dose in an interventional experiment
  - Continuous variables divided into categories
Relative Advantages

• Dummy variables fit groups exactly
  – If no other predictors in the model, parameter estimates correspond exactly with descriptive statistics
• With continuous variables, dummy variables assume a “step function” is true
• Modeling with dummy variables ignores order of predictor of interest

Flexible Methods

Flexible Modeling of Predictors

• We do have methods that can fit a wide variety of curve shapes
  – Dummy variables
    • A step function with tiny steps
  – Polynomials
    • If high degree: allows many patterns of curvature
  – Piecewise linear or piecewise polynomial
  – Splines
    • Piecewise linear or piecewise polynomial joined at knots
  – Fractional polynomial

Stata: Linear Splines

• Stata will make variable that will fit piecewise linear curves
  – Joined at “knots”
  – Lines in between
• mkspline newvar0 #k1 newvar1 #k2 newvar2 ... #kp varp= oldvar
  – Regression on newvar0 ... newvarp
    • Straight lines between min and k1, k1 and k2, etc.
Stata: fracpoly

- Stata will make variables modeling “fractional polynomials”
  - Can fit many different shapes depending on degree of the fractional polynomials
  - Can ask Stata to find “best” degree of the fractional polynomials: “fracpoly”
  - Can ask Stata to make new variables to model fractional polynomial of desired degree: “fracgen”

fracpoly

- Command
  fracpoly regression command yvar xvar, degree(#)

Example
  fracpoly regression logslry yrdeg, degree(3)