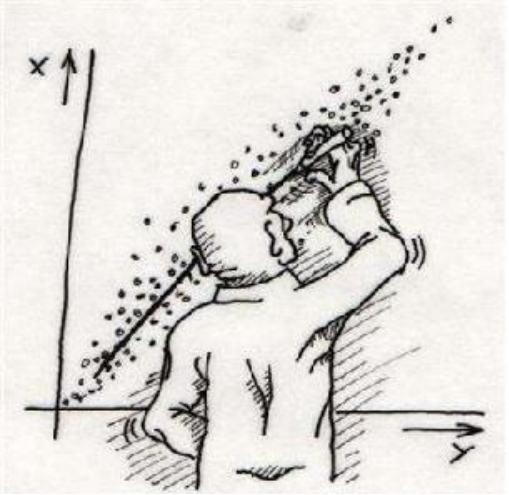
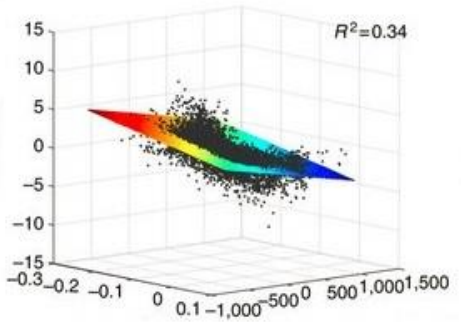
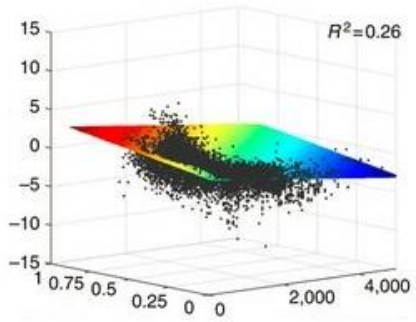


DATA 621 - Business Analytics and Data Mining





Meetup #3

Department of Data Analytics and IS
CUNY School of Professional Studies
The City University of New York

Dr. Nathan Bastian

Count Regression

Count Regression

- Positive Integer Response Variable: 0, 1, 2, 3, etc. (count outcomes)

- The response variable follows either:
  - "Poisson" Distribution
  - "Negative Binomial" Distribution

- The use of the linear regression model for count outcomes can result in inefficient, inconsistent, and biased estimates.

Further, linear regression is not usually a good choice because:

- The relationship between predictors and the response are not linear.

- Linear regression can yield a NEGATIVE prediction (counts must be >0).

- Error term distribution won't be random ("heteroskedastic").

Count Regression (cont.)

- The Poisson regression model is the most basic model.
- With this model the probability of a count is determined by a Poisson distribution, where the mean of the distribution is a function of the independent variables.
- This model has the defining characteristic that the conditional mean of the outcome is equal to the conditional variance.
- In practice, the conditional variance often exceeds the conditional mean.
- Dealing with this problem leads to the negative binomial regression model, which allows the variance to exceed the mean.

Poisson Distribution

POISSON PROBABILITY DENSITY FUNCTION:

$$\text{Prob}(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Where …

- λ is the mean value (expected number of times that an event has occurred)
- e is the natural exponent e = 2.71828
- k is any integer from 0, 1, …, Infinity

- Note that X is a random variable indicating the number of times that an event has occurred during an interval of time.

Poisson Distribution (cont.)

Properties:

$$\text{Prob}(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- Mean or Expected Value = λ
- Variance = λ

Therefore, if the **mean** and the **variance** are **the same** (known as *equidispersion*), it is a Poisson distribution (in practice, they won't be exactly the same, but if they are close then it is a Poisson distribution).

Poisson Distribution (cont.)

Example:

Assume a random variable follows a Poisson distribution with a mean value of λ=0.8, then what is the probability that X=3?

Prob(X=k) $= \frac{\lambda^k e^{-\lambda}}{k!}$

Prob(X=3) $= \frac{0.8^3 e^{-0.8}}{3!}$

Prob(X=3) $= \frac{0.512 * 0.4493}{6}$

Prob(X=3) = 0.0383

Poisson Distribution (cont.)

Example:

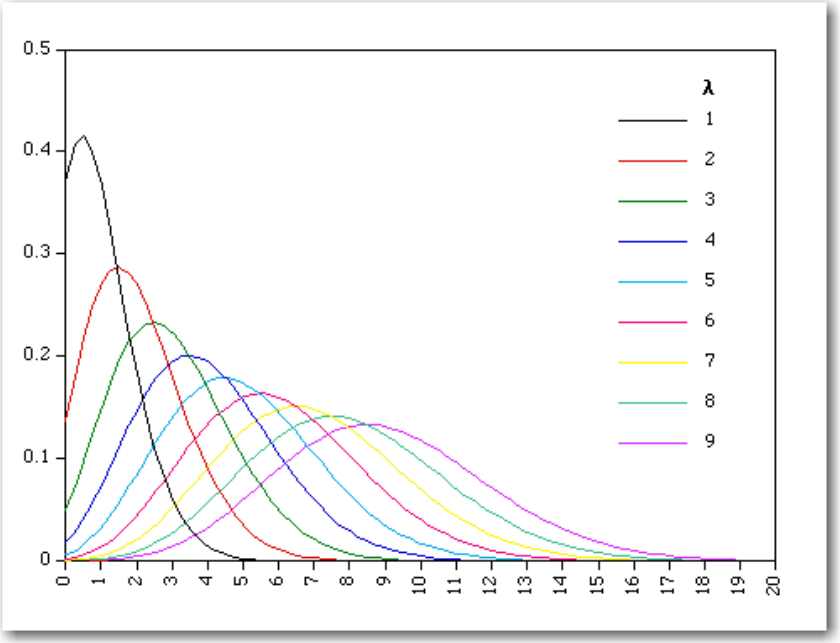
Assume a random variable follows a Poisson distribution with a mean value of λ=0.8, what are the probabilities of X=0, 1,…,6?

K	Prob(x=K)
0	0.4493
1	0.3595
2	0.1438
3	0.0383
4	0.0077
5	0.0012
6	0.0002

Poisson Distribution (cont.)

Graph:

- This is a graph of the Poisson distribution for various values of λ.
- Notice that as λ gets larger, the **Poisson** distribution looks like a **Normal** distribution.
- Also, as λ increases, the probability of 0's decreases!



Poisson Regression

Poisson Transform:

Recall that the response variable can be transformed prior to regression:

- LINEAR REGRESSION: $G(y) = y$ *(Identity or "Do Nothing")*
- LOGISTIC REGRESSION: $G(y) = \ln(y/(1-y))$ *(LOGIT Transform)*
- POISSON REGRESSION: $G(Y) = \ln(Y)$ (Log Transform)

Note: *Some commercially available software will handle the possibility that Y=0 for Poisson Regression and Y=0 or 1 for Logistic Regression. For Poisson, for example, Y can be replaced by a small number greater than 0 or else add 1 to the target and subtract 1 after the regression. There are other techniques, but they are beyond the scope of this course. Suffice it to say that you don't need to worry about this.*

Poisson Regression (cont.)

The likelihood function for *n* independent Poisson observations is a product of probabilities:

$$L(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X}) = \prod_{i=1}^{N} \frac{\lambda_i{}^{y_i} e^{-\lambda_i}}{y_i!}$$

where $\lambda = \exp(\boldsymbol{x\beta})$

- After the **G(Y) = ln(Y)** transformation, the regression is conducted using Maximum Likelihood Estimation. Since the likelihood function is globally concave, if a maximum is found it will be unique.

- The result of the regression will be the NATURAL LOG of the Count, so the Count value must be determined by exponentiating the output.

Interpretation of parameters: for a one unit change in x_k, the expected count changes by a factor of $\exp(\beta_k)$, holding all other variables constant.

Negative Binomial Distribution

NEGATIVE BINOMIAL PROBABILITY DENSITY FUNCTION:

$$\text{Prob}(X=n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}$$

Where:

- n is the number of attempts
- r is the number of successes
- p is the probability of success

Negative Binomial Distribution (cont.)

NEGATIVE BINOMIAL PROBABILITY DENSITY FUNCTION:

$$\text{Prob}(X=n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}$$

… SIMPLIFIED FORMULA:

$$\text{Prob}(X=n) = \frac{(n-1)!}{(r-1)!(n-r)!} p^r (1-p)^{n-r}$$

Negative Binomial Distribution (cont.)

Properties:

$$\text{Prob}(X=n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}$$

- Mean or Expected Value = r(1-p)/p = λ
- Variance = $r(1-p)/p^2$

Therefore, if the **variance** is **greater than** the **mean** (known as overdispersion), then it is a negative binomial distribution.

Negative Binomial Distribution (cont.)

Example:

Assume a random variable follows a Negative Binomial distribution. Assume that there is a probability of success of 0.7. What is the probability that it will take 5 tries (n=5) to get 3 successes (r=3)?

Prob(n=5) $= \frac{(n-1)!}{(r-1)!(n-r)!} p^r (1-p)^{n-r}$

Prob(n=5) $= \frac{(5-1)!}{(3-1)!(5-3)!} (0.7)^3 (1-0.7)^{5-3}$

Prob(n=5) $= \frac{24}{(2)(2)} (0.343)(0.3)^2$

Prob(n=5) $= (6)(0.343)(0.09)$

Prob(n=5) = 0.18522

Negative Binomial Distribution (cont.)

Example:

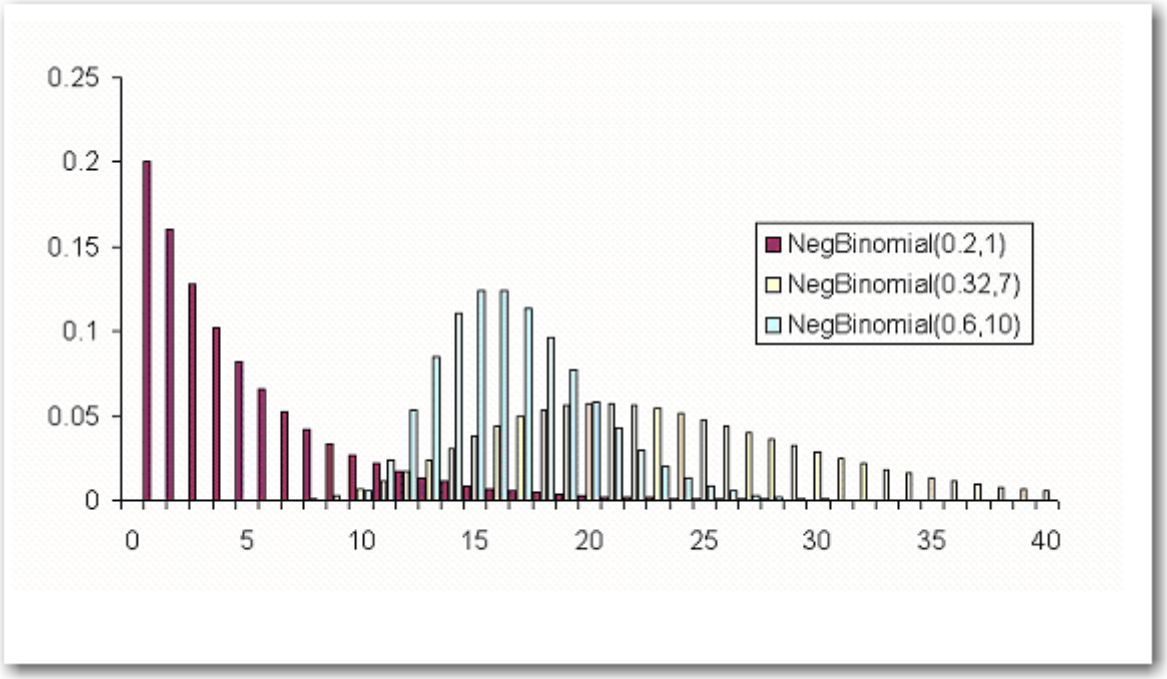
Assume a random variable follows a Negative Binomial distribution with a probability of success of 0.7. What is the probability that there will be “N” trials to achieve r=3 successes?

N	Prob(x=N)
3	0.343
4	0.309
5	0.185
6	0.093
7	0.042
8	0.018
9	0.007

Negative Binomial Distribution (cont.)

Graph:

- This is a graph of the Negative Binomial distribution of various values of p and r.



Negative Binomial Regression

Negative Binomial Transform:

The Negative Binomial Transform is similar to Poisson Regression in that it uses the natural log transformation.

- LINEAR REGRESSION: *G(y) = y (Identity or "Do Nothing")*
- LOGISTIC REGRESSION: *G(y) = ln(y/(1-y)) (LOGIT Transform)*
- POISSON REGRESSION: G(Y) = ln(Y) (Log Transform)
- **NEGATIVE BINOMIAL REGRESSION:** G(Y) = ln(Y) (Log Transform)

Note: *Poisson Regression is actually a special case of Negative Binomial Regression where the mean and the variance are equal. Many times, if the variance and the mean are similar, both Poisson and Negative Binomial Regressions will converge to the same result.*

Negative Binomial Regression (cont.)

- If there is overdispersion, then the estimates from the Poisson regression model are consistent but inefficient. The standard errors will be biased downward, resulting in spuriously large z-values.

- In Negative Binomial regression, the Poisson regression model is extended by capturing observed heterogeneity in the mean and by adding a parameter that allows the conditional variance of *y* to exceed the conditional mean.

- After the **G(Y) = ln(Y)** transformation, the negative binomial regression model is also estimated using Maximum Likelihood Estimation.

- Again, the result of the regression will be the NATURAL LOG of the Count, so the Count value must be determined by exponentiating the output.

Interpretation of parameters: for a one unit change in x_k, the expected count changes by a factor of $\exp(\beta_k)$, holding all other variables constant.

HAI Example

- We evaluated the effects of pay-for-performance (P4P) financial incentive program participation on existing levels of care quality and whether there was a decline in healthcare associated infections (HAIs) for participating hospitals.

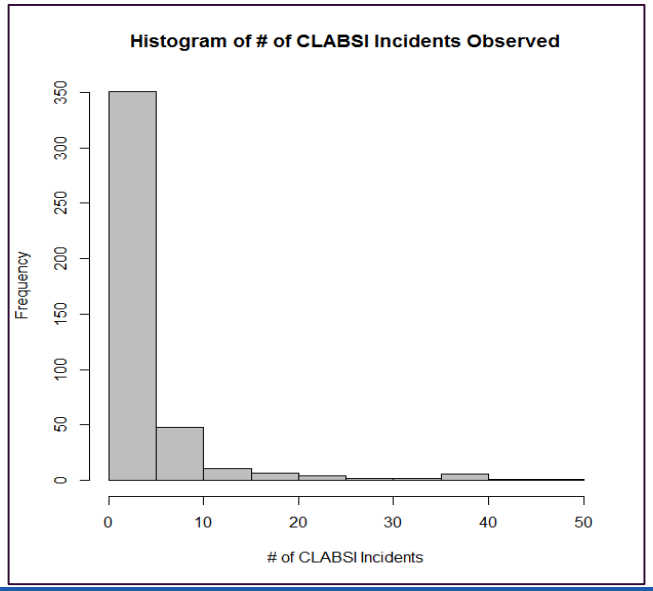
- In particular, we measured the impact of Highmark's Quality Blue (*QB*) P4P financial incentive program on reducing central line-associated blood stream infections (CLABSI).

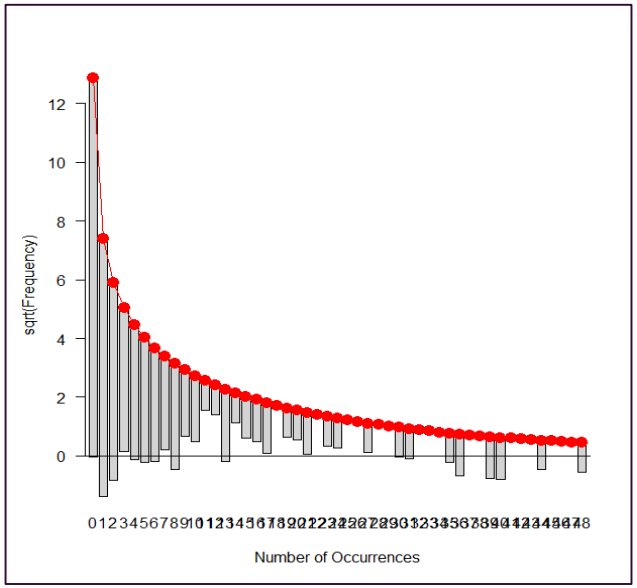
Did hospitals that participated in Highmark's QB program have a lower expected number of CLABSI compared to hospitals that did not participate in QB?

HAI Example (cont.)

- **Data:** Hospital-level CLABSI data for 149 hospitals in Pennsylvania from 2008 through 2013
  - 52 hospitals participating in QB and 97 not participating

- **Response variable:** # of CLABSI observed (CLABSI_Obs)
  - Best fit by **negative binomial distribution** (Chi-square = 44.92, df = 42, p-value = 0.35, $\alpha = 0.05$)





HAI Example (cont.)

- **Independent variables**:
  - QB (1 = Participation in QB, 0 otherwise)
  - Beds (number of hospital beds)
  - PatD (number of patient days)
  - Discharges (total number of discharges)
  - CLD (number of central line days).

- In order to avoid overdispersion that often occurs with Poisson generalized linear models, we performed **negative binomial regression**.

- This results in the following count regression model, where the parameters are estimated using maximum likelihood estimation:

$$\ln(CLABSI\_Obs) = \beta_0 + \beta_1 QB + \beta_2 Beds + \beta_3 PatD + \beta_4 Discharges + \beta_5 CLD + \varepsilon$$

HAI Example (cont.)

- Approximately 82% of the observations had 0-5 CLABSI per year and approximately 10% had 5-10 CLABSI per year.

- Output of the negative binomial regression model:

	Coefficient	IRR (95% CI)	Std. Error	Z	p
Intercept	−3.360e-01	0.714 [0.553, 0.919]	9.95e-02	−2.713	0.007
QB	−3.183e-01	0.727 [0.546, 0.966]	1.15e-01	−2.203	0.027
Patient days	3.360e-05	1.000 [1.000, 1.000]	9.259e-06	3.910	<0.001
Discharges	−5.982e-05	0.999 [0.999, 1.000]	3.759e-05	−1.592	0.111
CLD	−2.362e-05	0.999 [0.999, 1.000]	2.566e-05	−0.920	0.357

Note. CI = confidence interval; IRR = incidence-risk ratio; QB = Quality Blue.

- On average, those hospitals that participated in the *QB* program had 0.727 times the CLABSI as those hospitals that did not participate in the program, holding all other variables constant (a 27% reduction in expected CLABSI).

Multinomial Logistic Regression

Multinomial Logistic Regression

- We sometimes wish to classify a response variable that has more than two classes.
- The two-class (i.e., binary) logistic regression model has a multiple class (i.e., multinomial) extension.
- Here, the response variable, *Y*, can take on any of *m* qualitative values, which we number 1, 2, … ,*m.*
- Let π_{ij} denote the probability that the *i*th observation falls in the *j*th category of the response variable; that is:
  - $\pi_{ij} \equiv \Pr(Y_i = j), \text{for } j = 1, \ldots, m$

Multinomial Logistic Regression (cont.)

- We have *k* predictor variables of interest, $X_1, \ldots, X_k$, on which the π_{ij} depend.

- This dependence between the response variable and predictors is modeled using a multinomial logistic distribution:

$$\pi_{ij} = \frac{\exp(\gamma_{0j} + \gamma_{1j}X_{i1} + \cdots + \gamma_{kj}X_{ik})}{1 + \sum_{l=1}^{m-1}\exp(\gamma_{0l} + \gamma_{1l}X_{i1} + \cdots + \gamma_{kl}X_{ik})} \text{ for } j = 1, \ldots, m-1$$

$$\pi_{im} = 1 - \sum_{j=1}^{m-1} \pi_{ij} \text{ (for category } m)$$

Multinomial Logistic Regression (cont.)

- In this multinomial logit model, there is one set of parameters, $\gamma_{0j}, \gamma_{1j}, \ldots, \gamma_{kj}$, for each response category but the baseline.

- The use of a baseline category is one way to avoid redundant parameters because of the restriction that the response category probability for each observations must sum to one.

- Upon some algebraic manipulation, we get the following model:

$$\ln \frac{\pi_{ij}}{\pi_{im}} = \gamma_{0j} + \gamma_{1j} X_{i1} + \cdots + \gamma_{kj} X_{ik} \text{ for } j = 1, \ldots, m-1$$

Multinomial Logistic Regression (cont.)

- The regression coefficients represent effects on the log-odds of membership in category *j* versus the baseline category *m*.
- These regression coefficients are estimated using the method of maximum likelihood.
- Note that it is convenient to impose the restriction $\sum_{j=1}^{m} \pi_{ij} = 1$ by setting $\gamma_m = 0$ (making category *m* the baseline).
- This allows us to interpret γ_{kj} as the effect of X_k on the logit of category *j* relative to category *m* (baseline).

Multinomial Logistic Regression (cont.)

- In addition, we can form the log-odds of membership in any pair of category *j* and *j'* (other than category *m*), where the regression coefficients for the logit between any pair of categories are the differences between corresponding coefficients for the two categories.

- The following equation allows us to interpret $(\gamma_{kj} - \gamma_{kj'})$ as follows: for a unit change in X_k, the logit of category *j* versus category *j'* is expected to change by $(\gamma_{kj} - \gamma_{kj'})$ units, holding all other variables constant.

$$\ln\frac{\pi_{ij}}{\pi_{ij'}} = \ln\frac{\pi_{ij}/\pi_{im}}{\pi_{ij'}/\pi_{im}} = \ln\frac{\pi_{ij}}{\pi_{im}} - \ln\frac{\pi_{ij'}}{\pi_{im}} =$$
$$= (\gamma_{0j} - \gamma_{0j'}) + (\gamma_{1j} - \gamma_{1j'})X_{i1} + \cdots + (\gamma_{kj} - \gamma_{kj'})X_{ik}$$

Generalized Linear Models

Generalized Linear Models

- Generalized linear models (GLMs) extend ordinary linear regression to non-normal response distributions.

- GLMs allow us to analyze the linear relationship between predictor variables and the mean of the response variable when it is not reasonable to assume the data is distributed normally.

- The response distribution must come from the Exponential Family of Distributions, including:
  - Normal, Bernoulli, Binomial, Poisson, Gamma, etc.

Generalized Linear Models (cont.)

- GLMs are a general class of linear models that are made up of three components:
  1. **Random Component:** Specifies the conditional distribution of the response variable given the value of the predictor variables in the model.
  2. **Systematic Component:** Predictor variables in a linear predictor function (***Xβ***) (i.e., linear function of regressors).
  3. **Link Function:** A smooth and invertible function (g(.)) that transforms the expectation of the response variable to the systematic component.

$$g(\mu_i) = \sum_j \beta_j x_{ij} \qquad \mu_i = g^{-1}\left(\sum_j \beta_j x_{ij}\right)$$

Generalized Linear Models (cont.)

- **Linear Regression:** Continuous outcome that is conditionally *Normally* distributed with constant standard deviation.

- **Logistic Regression / Probit Regression:** Binary outcome (success or failure), where the random component has *Binomial* distribution.

- **Poisson Regression:** Count outcome (number of events in a length of time), where the random component has *Poisson* distribution.

- **Negative Binomial Regression:** When Count data have V(Y) > E(Y).

- **Gamma Regression:** Continuous outcome with skewed distribution and variation that increases with the mean can be modeled with a *Gamma* distribution.

Generalized Linear Models (cont.)

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Identity	μ_i	η_i
Log	$\log_e \mu_i$	e^{η_i}
Inverse	μ_i^{-1}	η_i^{-1}
Inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
Square-root	$\sqrt{\mu_i}$	η_i^2
Logit	$\log_e \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
Complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

NOTE: μ_i is the expected value of the response; η_i is the linear predictor; and $\Phi(\cdot)$ is the cumulative distribution function of the standard-normal distribution.

Generalized Linear Models (cont.)

- The general desire is to select a link function that renders the regression of *Y* on the *X*s linear.

- A promising link should behave reasonably in relation to the range of the expected response.

- Examples:
  - For binomial data, we modeled the probability of "success", represented by μ_i. As a probability, μ_i is confined to the unit interval [0, 1]. The logit, probit, log-log, and complementary log-log links map this interval to the entire real line.
  - For count data, we modeled the expected count. The log link maps μ_i to the whole real line.

Generalized Linear Models (cont.)

- GLMs are fit to data by the method of maximum likelihood (ML), providing not only estimates of the regression coefficients but also estimated asymptotic standard errors of the coefficients.

- Typically, these ML estimating equations can be solved by iterative reweighted least squares (IRWLS).

- Applying the maximum-likelihood equations and obtaining estimates by IRWLS is known as quasi-likelihood estimation.

- **Testing GLMs:** The ANOVA for linear models has an analog in the analysis of deviance for GLMs.

Panel Regression

Panel Regression

- A **panel data** set has both a cross-sectional and a time series dimension, where the *same* individuals, families, firms, cities, states, etc. are followed across time (i.e., we observe repeated cross-sections of the same individuals over time).

- Examples:
  - Annual unemployment rates of each state over several years
  - Quarterly sales of individual stores over several quarters
  - Wages for the same worker, working at several different jobs

Panel Regression (cont.)

- For proper regression analysis of panel data, we cannot assume that the observations are independently distributed across time.

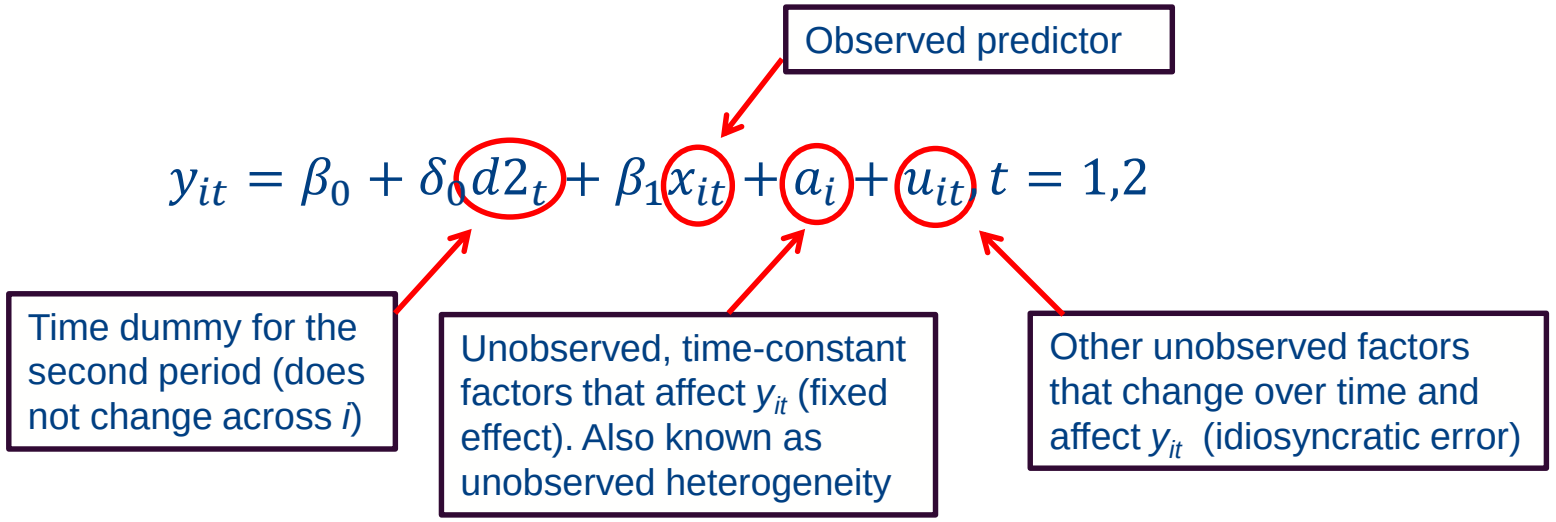
- For example, unobserved factors (such as ability) that affect someone's wage in 2016 will also affect that person's wage in 2017. This unobserved heterogeneity is a form of omitted variable bias.

- For this reason, special models and methods have been developed to analyze panel data.

- We first turn to the analysis of the simplest kind of panel data: for a cross-section of individuals, schools, firms, cities, etc., where we have two years of data; call these $t = 1$ and $t = 2$.

Panel Regression (cont.)

- **Fixed Effects (Unobserved Effects) Model**
  - One way to mitigate the omitted variable bias problem with panel data is to view the unobserved factors affecting the dependent variable as consisting of two types: those that are constant and those that vary over time.
  - Letting *i* denote the cross-sectional unit and *t* the time period, we can write a model with a single observed predictor variable as:



Panel Regression (cont.)

- How should we estimate the parameter of interest, β_1, given two years of panel data?
- One possibility is just to pool the two years and use OLS.
- Drawback:
  - In order for pooled OLS to produce a consistent estimator of β_1, we would have to assume that the unobserved effect, a_i, is uncorrelated with the predictor variable, x_{it}.
  - In the model, we can substitute $v_{it} = a_i + u_{it}$ (called composite error). From OLS, we must assume that v_{it} is uncorrelated with x_{it} in order to estimate β_1 consistently (or else there is an endogeneity issue).
  - Thus, even if we assume that the idiosyncratic error u_{it} is uncorrelated with x_{it}, pooled OLS is biased and inconsistent if a_i and x_{it} are correlated.
  - This resulting bias in pooled OLS is called **heterogeneity bias**.

Panel Regression (cont.)

- In most applications, the main reason for collecting panel data is to allow for the unobserved effect, a_i, to be correlated with the predictor variable(s).

- Because a_i is constant over time, we can difference the data across the two years. More precisely, for a cross-sectional observation *i*, write the two years as:

$$y_{i2} = (\beta_0 + \delta_0) + \beta_1 x_{i2} + a_i + u_{i2} \quad (t = 2)$$

$$y_{i1} = \beta_0 + \beta_1 x_{i1} + a_i + u_{i1} \quad (t = 1)$$

- If we subtract the second equation from the first, we get:

$$(y_{i2} - y_{i1}) = \delta_0 + \beta_1 (x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$$

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

Panel Regression (cont.)

- This is known as the **first-differenced equation**.
- When we obtain the OLS estimator of β_1 from this equation, we call the resulting estimator the **first-differenced panel estimator.**
  - This a way to consistently estimate causal effects in the presence of time-invariant endogeneity.
  - For consistency, strict exogeneity has to hold in the original equation (i.e., the idiosyncractic error at each time *t,* u_{it}, is uncorrelated with the predictor variable in *both* time periods).
  - Note that we allow x_{it} to be correlated with unobservables that are constant over time.
  - First-differenced estimates will be imprecise if predictor variables does not change over time for any cross-sectional observation, or if it changes by the same amount for every observation.

Panel Regression (cont.)

- It should be noted that we can also use differencing with more than two time periods.

- When using more than two time periods, we must assume that Δu_{it} is uncorrelated over time for the usual standard errors and test statistics to valid.

- Note that we can correct for the presence of AR(1) serial correlation in Δu_{it} by using FGLS.

- If there is no serial correlation in the errors, the usual methods for dealing with heteroscedasticity are valid.

Panel Regression (cont.)

- In addition to first differencing, we can estimate unobserved effects panel data models using two other common methods.

- **Fixed effects estimator** (which uses a transformation to remove the unobserved effect prior to estimation)

- **Random effects estimator** (which is attractive when we think the unobserved effect is uncorrelated with all the predictor variables)

Panel Regression (cont.)

- **Fixed effects estimation**

Fixed effect, potentially correlated with predictor variables

$$y_{it} = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it},\ i = 1, \ldots, N, t = 1, \ldots, T$$

Form time-averages for each individual

$$\bar{y}_i = \beta_1 \bar{x}_{i1} + \cdots + \beta_k \bar{x}_{ik} + \bar{a}_i + \bar{u}_i$$

$$\Rightarrow [y_{it} - \bar{y}_i] = \beta_1 [x_{it1} - \bar{x}_{i1}] + \cdots + \beta_k [x_{itk} - \bar{x}_{ik}] + [u_{it} - \bar{u}_i]$$

Because $a_i - \bar{a}_i = 0$ (the fixed effect is removed)

- **Estimate time-demeaned equation by OLS**
  - Uses time variation in *y* and *x within* each cross-sectional observation (known as the within-estimator).

Panel Regression (cont.)

- Under a strict exogeneity assumption on the predictor variables, the fixed effects estimator is unbiased.

- The R-squared of the time-demeaned equation is inappropriate (as it is interpreted as the amount of time variation in the y_{it} that is explained by the time variation in the predictor variables).

- Although the effect of time-invariant variables cannot be estimated in a fixed effects model, the effect of *interactions* between time-variant variables (e.g., year dummy) and time-invariant variables *can* be estimated.

- If a full set of time dummies are included, the effect of variables whose change over time is constant cannot be estimated.

Panel Regression (cont.)

- A traditional view of the fixed effects approach is to assume that the unobserved effect, a_i, is a parameter to be estimated for each *i*.
- The fixed effects estimator is equivalent to introducing a dummy variable for each cross-sectional observation in the original regression, along with the predictor variables, and using pooled OLS (**least squares dummy variable regression**):

$$y_{it} = a_1 ind1_{it} + a_2 ind2_{it} + \ldots + a_N indN_{it}$$
$$+ \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + u_{it}$$

For example, =1 if the observation stems from individual N, =0 otherwise

  - After fixed effects estimation, the fixed effects can be estimated as:

$$\hat{a}_i = \bar{y}_i - \hat{\beta}_1 \bar{x}_{i1} - \cdots - \hat{\beta}_k \bar{x}_{ik}, \; i = 1, \ldots, N$$

Estimated individual effect for individual i

Panel Regression (cont.)

- **Fixed effects or first differencing?**
  - Remember that first differencing can also be used if T > 2.
  - In the case T = 2, fixed effects and first differencing are identical.
  - For T > 2, fixed effects is more efficient if classical assumptions hold.
  - First differencing may be better in the case of severe serial correlation in the errors, for example if the errors follow a random walk.
  - If T is very large (and N not so large), the panel has a pronounced time series character and problems such as strong dependence arise.
  - In these cases, it is probably better to use first differencing.
  - Otherwise, it is a good idea to compute both and check robustness.

Panel Regression (cont.)

- **Random effects model**

The unobserved effect is assumed to be uncorrelated with each predictor variable in all time periods

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it},\ i = 1, \ldots, N, t = 1, \ldots, T$$

Random effects assumption: $Cov(x_{itj}, a_i) = 0$

The composite error $a_i + u_{it}$ is uncorrelated with the explanatory variables but it is serially correlated for observations coming from the same *i*:

$$Cov(a_i + u_{it}, a_i + u_{is}) = Cov(a_i, a_i) = \sigma_a^2$$

Under the assumption that idiosyncratic errors are serially uncorrelated

Panel Regression (cont.)

- **Estimation in the random effects model**
  - Under the random effects assumptions, predictor variables are exogenous so that pooled OLS provides consistent estimates.
  - If OLS is used, standard errors have to be adjusted for the fact that the composite errors are serially correlated across time for given *i*.
  - But, because of the serial correlation, OLS is not efficient. However, GLS can be used to solve the serial correlation problem.
  - One can transform the model so that it satisfies the necessary assumptions:

$$[y_{it} - \lambda \bar{y}_i] = \beta_1 [x_{it1} - \lambda \bar{x}_{i1}] + \cdots + \beta_k [x_{itk} - \lambda \bar{x}_{ik}]$$

Quasi-demeaned data

$$+ [a_i - \lambda \bar{a}_i + u_{it} - \lambda \bar{u}_i]$$

Error can be shown to satisfy assumptions

Panel Regression (cont.)

- **Estimation in the random effects model (cont.)**

with $\lambda = 1 - \left[\sigma_u^2/(\sigma_u^2 + T\sigma_a^2)\right]^{1/2}$, $0 \leq \lambda \leq 1$

  - The quasi-demeaning parameter is unknown but it can be estimated.
  - FGLS using the estimated λ is called **random effects estimation.**
  - If the random effect is relatively unimportant compared to the idosyncratic error, FGLS will be close to pooled OLS (because $\lambda \to 0$).
  - If the random effect is relatively important compared to the idiosyncratic term, FGLS will be similar to fixed effects (because $\lambda \to 1$).
  - Random effects estimation works for time-invariant variables.
  - Note: unobserved effects are seldomly uncorrelated with predictor variables, making fixed effects models more convincing in practice.