Regression Modeling with Actuarial and Financial Applications

Chapter 11: Categorical Dependent Variables

Chapter 13: Generalized Linear Models

Outline

Chapter 11: Categorical Dependent Variables

Example. MEPS Hospital Utilization

- Consider an extensive database from the Medical Expenditure Panel Survey (MEPS) on hospitalization utilization

Table: Hospitalization by Gender

		Male	Female
Not hospitalized	$y = 0$	902 (95.3%)	941 (89.3%)
Hospitalized	$y = 1$	44 (4.7%)	113 (10.7%)
Total		946	1,054

$$y_i = \begin{cases} 1 & i\text{th person was hospitalized during the sample period} \\ 0 & \text{otherwise} \end{cases} .$$

- Like linear regression techniques, we are interested in using characteristics of a person, such as their age, sex, education, income, prior health status and so forth, to help explain the dependent variable y.
- However, now the dependent variable is discrete and not even approximately normally distributed.

Linear Probability Model

- y_i has a Bernoulli distribution
  - The probability that the response equals 1 by $\pi_i = \Pr(y_i = 1)$.
  - The mean response is $\mathrm{E}\ y_i = 0 \times \Pr(y_i = 0) + 1 \times \Pr(y_i = 1) = \pi_i$.
  - Thus, the variance is related to the mean through the expression $\mathrm{Var}\ y_i = \pi_i(1 - \pi_i)$.

Linear Probability Model

- y_i has a Bernoulli distribution
  - The probability that the response equals 1 by $\pi_i = \Pr(y_i = 1)$.
  - The mean response is $\mathrm{E}\, y_i = 0 \times \Pr(y_i = 0) + 1 \times \Pr(y_i = 1) = \pi_i$.
  - Thus, the variance is related to the mean through the expression $\mathrm{Var}\, y_i = \pi_i(1 - \pi_i)$.
- The linear probability model is

$$y_i = \mathbf{x}_i' \beta + \varepsilon_i,$$

  - Assuming $\mathrm{E}\, \varepsilon_i = 0$, we have that $\mathrm{E}\, y_i = \mathbf{x}_i' \beta = \pi_i$
  - $\mathrm{Var}\, y_i = \mathbf{x}_i' \beta (1 - \mathbf{x}_i' \beta)$.

Drawbacks of the Linear Probability Model

- The expected response is a probability and thus must vary between 0 and 1. However, the linear combination, $\mathbf{x}_i'\beta$, can vary between negative and positive infinity. This mismatch implies, for example, that fitted values may be unreasonable.

Drawbacks of the Linear Probability Model

- The expected response is a probability and thus must vary between 0 and 1. However, the linear combination, $\mathbf{x}_i'\beta$, can vary between negative and positive infinity. This mismatch implies, for example, that fitted values may be unreasonable.
- Linear models assume homoscedasticity (constant variance) yet the variance of the response depends on the mean that varies over observations. The problem of varying variability is known as *heteroscedasticity*.

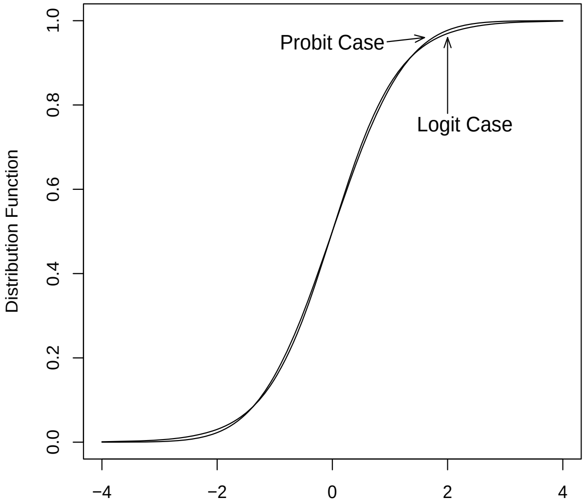
Drawbacks of the Linear Probability Model

- The expected response is a probability and thus must vary between 0 and 1. However, the linear combination, $\mathbf{x}_i'\beta$, can vary between negative and positive infinity. This mismatch implies, for example, that fitted values may be unreasonable.
- Linear models assume homoscedasticity (constant variance) yet the variance of the response depends on the mean that varies over observations. The problem of varying variability is known as *heteroscedasticity*.
- The response must be either a 0 or 1 although the regression models typically regards distribution of the error term as continuous. This mismatch implies, for example, that the usual residual analysis in regression modeling is meaningless.

Using nonlinear functions of explanatory variables

- The linear combination of explanatory variables, $\mathbf{x}_i'\beta$, is sometimes known as the "systematic component."
- We consider a function of explanatory variables, $\pi_i = \pi(\mathbf{x}_i'\beta) = \Pr(y_i = 1|\mathbf{x}_i)$.
- We focus on two special cases of the function $\pi(.)$:
  - $\pi(z) = \frac{1}{1+\exp(-z)} = \frac{e^z}{1+e^z}$, the logit case, and
  - $\pi(z) = \Phi(z)$, the probit case.
  - $\Phi(.)$ is the standard normal distribution function.
- Note that $\pi(z) = z$ yields the linear probability model.
- The inverse of the function, π^{-1}, is linear in the explanatory variables, that is, $\pi^{-1}(\pi_i) = \mathbf{x}_i'\beta$.
- The logit and probit are really close.

Comparison of Logit and Probit Distribution Functions



Threshold interpretation

- Both the logit and probit are special cases.
- To this end, suppose that there exists an underlying linear model, $y_i^* = \mathbf{x}_i'\beta + \varepsilon_i^*$.
  - We do not observe the response y_i^* yet interpret it to be the "propensity" to possess a characteristic.
  - For example, we might think about the speed of a horse as a measure of its propensity to win a race.

Threshold interpretation

- Both the logit and probit are special cases.
- To this end, suppose that there exists an underlying linear model, $y_i^* = \mathbf{x}_i'\beta + \varepsilon_i^*$.
  - We do not observe the response y_i^* yet interpret it to be the "propensity" to possess a characteristic.
  - For example, we might think about the speed of a horse as a measure of its propensity to win a race.
- Under the threshold interpretation, we do not observe the propensity but we do observe when the propensity crosses a threshold.
  - It is customary to assume that this threshold is 0, for simplicity.
  - We observe

$$y_i = \begin{cases} 0 & y_i^* \le 0 \\ 1 & y_i^* > 0 \end{cases} .$$

Threshold interpretation - Logit Case

- Assume a logit distribution function for the disturbances, so that

$$\Pr(\varepsilon_i^* \leq a) = \frac{1}{1+\exp(-a)}.$$

- Because the logit distribution is symmetric about zero, we have that $\Pr(\varepsilon_i^* \leq a) = \Pr(-\varepsilon_i^* \leq a)$.

$$\begin{aligned} \pi_i &= \Pr(y_i = 1|\mathbf{x}_i) = \Pr(y_i^* > 0) = \Pr(\varepsilon_i^* \leq \mathbf{x}_i'\beta) \\ &= \frac{1}{1+\exp(-\mathbf{x}_i'\beta)} = \pi(\mathbf{x}_i'\beta). \end{aligned}$$

- This establishes the threshold interpretation for the logit case.
- The development for the probit case is similar, and is omitted.

Random utility interpretation

- Think of an individual as selecting between two choices.
  - Preferences among choices are indexed by an unobserved utility function
  - Individuals select the choice that provides the greater utility.
- For the ith subject, we use the notation u_i for this utility.
  - Choice 1: $U_{i1} = u_i(V_{i1} + \varepsilon_{i1})$
  - Choice 2: $U_{i2} = u_i(V_{i2} + \varepsilon_{i2})$
  - Utility = function of an underlying value plus random noise.
- Choice $j = 1$ means
  - $U_{i1} > U_{i2}$
  - $y_i = 1$

Random utility interpretation

- Think of an individual as selecting between two choices.
  - Preferences among choices are indexed by an unobserved utility function
  - Individuals select the choice that provides the greater utility.
- For the ith subject, we use the notation u_i for this utility.
  - Choice 1: $U_{i1} = u_i(V_{i1} + \varepsilon_{i1})$
  - Choice 2: $U_{i2} = u_i(V_{i2} + \varepsilon_{i2})$
  - Utility = function of an underlying value plus random noise.
- Choice $j = 1$ means
  - $U_{i1} > U_{i2}$
  - $y_i = 1$
- Assuming that u_i is a strictly increasing function, we have

$$\begin{aligned} \Pr(y_i &= 1) = \Pr(U_{i2} < U_{i1}) = \Pr\left(u_i(V_{i2} + \varepsilon_{i2}) < u_i(V_{i1} + \varepsilon_{i1})\right) \\ &= \Pr(\varepsilon_{i2} - \varepsilon_{i1} < V_{i1} - V_{i2}). \end{aligned}$$

  - Assume that $V_{i2} = 0$ and $V_{i1} = \mathbf{x}_i'\beta$.
  - We may take the difference in the errors, $\varepsilon_{i2} - \varepsilon_{i1}$, to be normal or logistic, corresponding to the probit and logit cases, respectively.

Logistic regression

- Logit case - permits closed-form expressions, unlike the probit (normal distribution function).
  - *Logistic regression* is another phrase used to describe the logit case.
- Using $p = \pi(z)$, the inverse of π can be calculated as $z = \pi^{-1}(p) = ln(p/(1-p))$.
  - To simplify future presentations, we define

    $$\text{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$$

    to be the *logit function*.
- With logistic regression model, we represent the linear combination of explanatory variables as the logit of the probability, that is, $\mathbf{x}_i'\beta = \text{logit}(\pi_i)$.

Odds interpretation

- When the response y is binary, knowing only the probability p summarizes the distribution.
  - In some applications, a simple transformation of p has an important interpretation.
  - An important transformation: the *odds*, given by $p/(1-p)$.
  - For example, suppose that y indicates whether or not a horse wins a race, that is, $y=1$ if the horse wins and $y=0$ if the horse does not.
    - Interpret p to be the probability of the horse winning the race
    - As an example, suppose that $p=0.25$. Then, the odds of the horse winning the race is $0.25/(1-0.25)=0.3333$.

Odds interpretation

- When the response y is binary, knowing only the probability p summarizes the distribution.
    - In some applications, a simple transformation of p has an important interpretation.
    - An important transformation: the *odds*, given by $p/(1-p)$.
    - For example, suppose that y indicates whether or not a horse wins a race, that is, $y = 1$ if the horse wins and $y = 0$ if the horse does not.
        - Interpret p to be the probability of the horse winning the race
        - As an example, suppose that $p = 0.25$. Then, the odds of the horse winning the race is $0.25/(1-0.25) = 0.3333$.
- Odds have a useful interpretation from a betting standpoint.
    - Suppose that we are playing a fair game and that we place a bet of $1 with odds of one to three.
        - If the horse wins, then we get our $1 back plus winnings of $3.
        - If the horse loses, then we lose our bet of $1.
- The logit is the logarithmic odds function, also known as the *log odds* .

Logistic regression parameter interpretation

- Assume that jth explanatory variable, x_{ij}, is either 0 or 1.
- With the notation $\mathbf{x}_i = (x_{i1},...,x_{ij},\ldots,x_{iK})'$, we may interpret

$$\begin{aligned}\beta_j &= (x_{i1},...,1,\ldots,x_{iK})'\beta - (x_{i1},...,0,\ldots,x_{iK})'\beta \\ &= \ln\left(\frac{\Pr(y_i=1|x_{ij}=1)}{1-\Pr(y_i=1|x_{ij}=1)}\right) - \ln\left(\frac{\Pr(y_i=1|x_{ij}=0)}{1-\Pr(y_i=1|x_{ij}=0)}\right)\end{aligned}$$

- Exponentiating, we have the *odds ratio*

$$e^{\beta_j} = \frac{\Pr(y_i=1|x_{ij}=1)/\left(1-\Pr(y_i=1|x_{ij}=1)\right)}{\Pr(y_i=1|x_{ij}=0)/\left(1-\Pr(y_i=1|x_{ij}=0)\right)}.$$

  - The numerator of this expression is the odds when $x_{ij}=1$, whereas the denominator is the odds when $x_{ij}=0$.

Logistic regression parameter interpretation

- Assume that jth explanatory variable, x_{ij}, is either 0 or 1.
- With the notation $\mathbf{x}_i = (x_{i1}, ..., x_{ij}, \ldots, x_{iK})'$, we may interpret

$$\begin{aligned} \beta_j &= (x_{i1}, ..., 1, \ldots, x_{iK})'\beta - (x_{i1}, ..., 0, \ldots, x_{iK})'\beta \\ &= \ln\left(\frac{\Pr(y_i = 1 | x_{ij} = 1)}{1 - \Pr(y_i = 1 | x_{ij} = 1)}\right) - \ln\left(\frac{\Pr(y_i = 1 | x_{ij} = 0)}{1 - \Pr(y_i = 1 | x_{ij} = 0)}\right) \end{aligned}$$

- Exponentiating, we have the *odds ratio*

$$e^{\beta_j} = \frac{\Pr(y_i = 1 | x_{ij} = 1) / (1 - \Pr(y_i = 1 | x_{ij} = 1))}{\Pr(y_i = 1 | x_{ij} = 0) / (1 - \Pr(y_i = 1 | x_{ij} = 0))}.$$

  - The numerator of this expression is the odds when $x_{ij} = 1$, whereas the denominator is the odds when $x_{ij} = 0$.
- Thus, we can say that the odds when $x_{ij} = 1$ are $\exp(\beta_j)$ times as large as the odds when $x_{ij} = 0$.
  - To illustrate, if $\beta_j = 0.693$, then $\exp(\beta_j) = 2$.
  - From this, we say that the odds (for $y = 1$) are twice as great for $x_{ij} = 1$ as $x_{ij} = 0$.

Logistic regression parameter interpretation

- Similarly, assuming that *j*th explanatory variable is continuous (differentiable), we have

$$\begin{aligned}\beta_j &= \frac{\partial}{\partial x_{ij}} \mathbf{x}'\beta = \frac{\partial}{\partial x_{ij}} \ln\left(\frac{\Pr(y_i=1|x_{ij})}{1-\Pr(y_i=1|x_{ij})}\right) \\ &= \frac{\frac{\partial}{\partial x_{ij}}\Pr(y_i=1|x_{ij})/\left(1-\Pr(y_i=1|x_{ij})\right)}{\Pr(y_i=1|x_{ij})/\left(1-\Pr(y_i=1|x_{ij})\right)}.\end{aligned}$$

- Thus, we may interpret β_j as the proportional change in the odds, known as an *elasticity* in economics.

Likelihoods for maximum likelihood estimation

- The customary method of estimation is maximum likelihood.
- To provide intuition, we outline the ideas in the context of binary dependent variable regression models.
- The *likelihood* is the observed value of the density or mass function.
- For a single observation, the likelihood is

$$\begin{cases} 1-\pi_i & \text{if } y_i = 0 \\ \pi_i & \text{if } y_i = 1 \end{cases} .$$

Likelihoods for maximum likelihood estimation

- The customary method of estimation is maximum likelihood.
- To provide intuition, we outline the ideas in the context of binary dependent variable regression models.
- The *likelihood* is the observed value of the density or mass function.
- For a single observation, the likelihood is

$$\begin{cases} 1-\pi_i & \text{if } y_i = 0 \\ \pi_i & \text{if } y_i = 1 \end{cases} .$$

- The objective of maximum likelihood estimation is to find the parameter values that produce the largest likelihood.
  - Finding the maximum of the logarithmic function yields the same solution as finding the maximum of the corresponding function.
  - Because it is generally computationally simpler, we consider the logarithmic (log-) likelihood, written as

$$\begin{cases} \ln(1-\pi_i) & \text{if } y_i = 0 \\ \ln \pi_i & \text{if } y_i = 1 \end{cases} .$$

Log likelihood

- More compactly, the log-likelihood of a single observation is

$$y_i \ln \pi(\mathbf{x}_i'\beta) + (1 - y_i) \ln \left(1 - \pi(\mathbf{x}_i'\beta)\right),$$

  where $\pi_i = \pi(\mathbf{x}_i'\beta)$.
- Assuming independence, the log-likelihood of the data set is

$$L(\beta) = \sum_{i=1}^{n} \left\{ y_i \ln \pi(\mathbf{x}_i'\beta) + (1 - y_i) \ln \left(1 - \pi(\mathbf{x}_i'\beta)\right) \right\}.$$

  - The (log) likelihood is viewed as a function of the parameters, with the data held fixed.
  - In contrast, the joint probability mass (density) function is viewed as a function of the realized data, with the parameters held fixed.
- The method of maximum likelihood means finding the values of β that maximize the log-likelihood.

Parameter estimation

- The customary method of finding the maximum is taking partial derivatives with respect to the parameters of interest and finding roots of the these equations.
- In this case, taking partial derivatives with respect to β yields the *score equations*

$$\frac{\partial}{\partial \beta} L(\beta) = \sum_{i=1}^{n} \mathbf{x}_i \left(y_i - \pi(\mathbf{x}_i'\beta) \right) \frac{\pi'(\mathbf{x}_i'\beta)}{\pi(\mathbf{x}_i'\beta)(1-\pi(\mathbf{x}_i'\beta))} = \mathbf{0}.$$

- The solution of these equations, say $\mathbf{b}_{MLE}$, is the maximum likelihood estimator.

Parameter estimation

- The customary method of finding the maximum is taking partial derivatives with respect to the parameters of interest and finding roots of the these equations.
- In this case, taking partial derivatives with respect to β yields the *score equations*

$$\frac{\partial}{\partial \beta} L(\beta) = \sum_{i=1}^{n} \mathbf{x}_i \left(y_i - \pi(\mathbf{x}_i'\beta) \right) \frac{\pi'(\mathbf{x}_i'\beta)}{\pi(\mathbf{x}_i'\beta)(1-\pi(\mathbf{x}_i'\beta))} = \mathbf{0}.$$

- The solution of these equations, say $\mathbf{b}_{MLE}$, is the maximum likelihood estimator.
- To illustrate, for the logit case, the score equations reduce to

$$\frac{\partial}{\partial \beta} L(\beta) = \sum_{i=1}^{n} \mathbf{x}_i \left(y_i - \pi(\mathbf{x}_i'\beta) \right) = \mathbf{0}.$$

  where $\pi(z) = 1/(1+\exp(-z))$.
- When the model contains an intercept term, we can write the first row of this expression as $\sum_{i=1}^{n} \left(y_i - \pi(\mathbf{x}_i'\mathbf{b}_{MLE}) \right) = 0$, so the sum of observed values equals the sum of fitted values.

Inference – Regression coefficient standard errors

- An estimator of the asymptotic variance of β may be calculated taking partial derivatives of the score equations.

$$\mathbf{I}(\beta) = \frac{\partial^2}{\partial \beta \partial \beta'} L(\beta)$$

  is the *information matrix*.
- To illustrate, using the logit function, straightforward calculations show that the information matrix is

$$\mathbf{I}(\beta) = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \pi(\mathbf{x}_i'\beta)(1 - \pi(\mathbf{x}_i'\beta)).$$

- The square root of the $(j+1)st$ diagonal element of this matrix evaluated at $\beta = \mathbf{b}_{MLE}$ yields the standard error for $b_{j,MLE}$, denoted as $se(b_{j,MLE})$.

Inference – Model fit

- To assess the overall model fit, it is customary to cite *likelihood ratio test statistics* in nonlinear regression models.
- For example, to test the overall model adequacy $H_0 : \beta = \mathbf{0}$, we use the statistic

$$LRT = 2 \times (L(\mathbf{b}_{MLE}) - L_0),$$

where L_0 is the maximized log-likelihood with only an intercept term.

  - Under the null hypothesis H_0, this statistic has a chi-square distribution with K degrees of freedom.

- Another measure of model fit is the so-called $max - scaled\ R^2$, defined as $R^2_{ms} = R^2 / R^2_{\max}$, where

$$R^2 = 1 - \left(\frac{\exp(L_0/N)}{\exp(L(\mathbf{b}_{MLE})/N)} \right)$$

and $R^2_{\max} = 1 - \exp(L_0/N)^2$. Here, L_0/N represents the average value of this log-likelihood.

Data

- Data from the Medical Expenditure Panel Survey (MEPS), conducted by the U.S. Agency of Health Research and Quality (AHRQ).
  - A probability survey that provides nationally representative estimates of health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian population.
  - Collects detailed information on individuals of each medical care episode by type of services including
    - physician office visits,
    - hospital emergency room visits,
    - hospital outpatient visits,
    - hospital inpatient stays,
    - all other medical provider visits, and
    - use of prescribed medicines.
  - This detailed information allows one to develop models of health care utilization to predict future expenditures.
  - We consider MEPS data from the first panel of 2003 and take a random sample of $n = 2,000$ individuals between ages 18 and 65.

Dependent Variable

- Our dependent variable is an indicator of positive expenditures for inpatient admissions.
- For MEPS, inpatient admissions include persons who were admitted to a hospital and stayed overnight.
- In contrast, outpatient events include hospital outpatient department visits, office-based provider visits and emergency room visits excluding dental services.
  - Hospital stays with the same date of admission and discharge, known as "zero-night stays," were included in outpatient counts and expenditures.
  - Payments associated with emergency room visits that immediately preceded an inpatient stay were included in the inpatient expenditures.
  - Prescribed medicines that can be linked to hospital admissions were included in inpatient expenditures, not in outpatient utilization.

Percent of Positive Expenditures by Explanatory Variable

Category	Variable	Description	Percent of data	Percent Positive Expend
Demography	AGE	Age in years between 18 to 65 (mean: 39.0)		
	GENDER	1 if female	52.7	10.7
		0 if male	47.3	4.7
Ethnicity	ASIAN	1 if Asian	4.3	4.7
	BLACK	1 if Black	14.8	10.5
	NATIVE	1 if Native	1.1	13.6
	WHITE	Reference level	79.9	7.5
Region	NORTHEAST	1 if Northeast	14.3	10.1
	MIDWEST	1 if Midwest	19.7	8.7
	SOUTH	1 if South	38.2	8.4
	WEST	Reference level	27.9	5.4
Education	COLLEGE	1 if college or higher degree	27.2	6.8
	HIGHSCHOOL	1 if high school degree	43.3	7.9
	Reference level is lower than high school degree		29.5	8.8
Self-rated physical health	POOR	1 if poor	3.8	36.0
	FAIR	1 if fair	9.9	8.1
	GOOD	1 if good	29.9	8.2
	VGOOD	1 if very good	31.1	6.3
	Reference level is excellent health		25.4	5.1
Self-rated mental health	MNHPOOR	1 if poor or fair	7.5	16.8
		0 if good to excellent mental health	92.6	7.1
Any activity limitation	ANYLIMIT	1 if any functional/activity limitation	22.3	14.6
		0 if otherwise	77.7	5.9
Income compared to poverty line	HINCOME	1 if high income	31.6	5.4
	MINCOME	1 if middle income	29.9	7.0
	LINCOME	1 if low income	15.8	8.3
	NPOOR	1 if near poor	5.8	9.5
	Reference level is poor/negative		17.0	13.0
Insurance coverage	INSURE	1 if covered by public/private health insurance in any month of 2003	77.8	9.2
		0 if have no health insurance in 2003	22.3	3.1
Total			100.0	7.9

Comparison of Binary Regression Models

	Logistic				Probit	
	Full Model		Reduced Model		Reduced Model	
Effect	Parameter Estimate	*t*-ratio	Parameter Estimate	*t*-ratio	Parameter Estimate	*t*-ratio
Intercept	-4.239	-8.982	-4.278	-10.094	-2.281	-11.432
AGE	-0.001	-0.180				
GENDER	0.733	3.812	0.732	3.806	0.395	4.178
ASIAN	-0.219	-0.411	-0.219	-0.412	-0.108	-0.427
BLACK	-0.001	-0.003	0.004	0.019	0.009	0.073
NATIVE	0.610	0.926	0.612	0.930	0.285	0.780
NORTHEAST	0.609	2.112	0.604	2.098	0.281	1.950
MIDWEST	0.524	1.904	0.517	1.883	0.237	1.754
SOUTH	0.339	1.376	0.328	1.342	0.130	1.085
COLLEGE	0.068	0.255	0.070	0.263	0.049	0.362
HIGHSCHOOL	0.004	0.017	0.009	0.041	0.003	0.030
POOR	1.712	4.385	1.652	4.575	0.939	4.805
FAIR	0.136	0.375	0.109	0.306	0.079	0.450
GOOD	0.376	1.429	0.368	1.405	0.182	1.412
VGOOD	0.178	0.667	0.174	0.655	0.094	0.728
MNHPOOR	-0.113	-0.369				
ANYLIMIT	0.564	2.680	0.545	2.704	0.311	3.022
HINCOME	-0.921	-3.101	-0.919	-3.162	-0.470	-3.224
MINCOME	-0.609	-2.315	-0.604	-2.317	-0.314	-2.345
LINCOME	-0.411	-1.453	-0.408	-1.449	-0.241	-1.633
NPOOR	-0.201	-0.528	-0.204	-0.534	-0.146	-0.721
INSURE	1.234	4.047	1.227	4.031	0.579	4.147
Log-Likelihood	-488.69		-488.78		-486.98	
AIC	1,021.38		1,017.56		1,013.96	

Comparison of Binary Regression Models

- From the t-values of the Full Model, one might consider a more parsimonious model by removing statistically insignificant variables.
  - The table shows a "Reduced Model," where age and mental health status variables have been removed.
  - However, twice the change in the log likelihood was only $2 \times (-488.78 - (-488.69)) = 0.36$.
  - Comparing this to a chi-square distribution with $df = 2$ degrees of freedom results in a p-value$= 0.835$, indicating that the drop is not statistically significant.
- The table also provides probit model fits.
  - The results are similar to the logit model fits.
  - Examine the sign of the coefficients and their significance.

Chapter 13: Generalized Linear Models

GLM Ingredients

- This extension of the linear model is so widely used that it is known as *the* "generalized linear model," or as the acronym GLM.
- GLM generalizes the linear model in three ways
- 1. Mean as a function of linear predictors
  - Call the linear combination of explanatory variables the *systematic component*, denoted as $\eta_i = \mathbf{x}_i'\beta$
  - The *link* function relates the mean to the systematic component

$$\eta_i = \mathbf{x}_i'\beta = \mathrm{g}\left(\mu_i\right).$$

  - g(.) a smooth, invertible function. The inverse $\mu_i = \mathrm{g}^{-1}(\mathbf{x}_i'\beta)$, is the mean function.
  - Some examples we have seen:
    - $\mathbf{x}_i'\beta = \mu_i$, for (normal) linear regression,
    - $\mathbf{x}_i'\beta = \exp(\mu_i)/(1+\exp(\mu_i))$, for logistic regression and
    - $\mathbf{x}_i'\beta = \ln(\mu_i)$, for Poisson regression.

GLM Ingredients II

- 2. The GLM extends linear modeling through the use of the *linear exponential family of distributions*
  - *Not* the exponential distribution - it is a generalization.
  - This family includes the normal, Bernoulli and Poisson distributions as special cases.

GLM Ingredients II

- 2. The GLM extends linear modeling through the use of the *linear exponential family of distributions*
  - *Not* the exponential distribution - it is a generalization.
  - This family includes the normal, Bernoulli and Poisson distributions as special cases.
- 3. GLM modeling is robust to the choice of distributions.
  - The linear model sampling assumptions focused on:
    - the form of the mean function (assumption F1),
    - non-stochastic or exogenous explanatory variables (F2),
    - constant variance (F3) and
    - independence among observations (F4).
  - GLM models maintain assumptions F2 and F4
  - GLM models extend F1 through the link function.
  - To extend F3, the variance depends on the choice of distributions

Variance as a Function of the Mean

Table: Variance Functions for Selected Distributions

Distribution	Variance Function $v(\mu)$
Normal	1
Bernoulli	$\mu(1-\mu)$
Poisson	μ
Gamma	μ^2
Inverse Gaussian	μ^3

- The choice of the variance function drives many inference properties, not the choice of the distribution.

Linear Exponential Family of Distributions

- *Definition.* The distribution of the *linear exponential family* is

$$\mathrm{f}(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + S(y,\phi)\right).$$

  - y is a dependent variable and θ is the parameter of interest.
  - ϕ is a scale parameter, often assumed known.
  - $b(\theta)$ depends only on the parameter θ, not the dependent variable.
  - $S(y,\phi)$ is a function of y and the scale parameter, not the parameter θ.

Linear Exponential Family of Distributions

- *Definition.* The distribution of the *linear exponential family* is

$$\mathrm{f}(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + S(y,\phi)\right).$$

  - y is a dependent variable and θ is the parameter of interest.
  - ϕ is a scale parameter, often assumed known.
  - $b(\theta)$ depends only on the parameter θ, not the dependent variable.
  - $S(y,\phi)$ is a function of y and the scale parameter, not the parameter θ.

- Example: Normal distribution - use $\theta = \mu$ and $\phi = \sigma^2$,

$$\begin{aligned}\mathrm{f}(y;\mu,\sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\ &= \exp\left(\frac{(y\mu-\mu^2/2)}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\ln\left(2\pi\sigma^2\right)\right).\end{aligned}$$

  - Also $b(\theta) = \theta^2/2$ and $S(y,\phi) = -y^2/(2\phi) - \ln\left(2\pi\sigma^2\right)/2$.

Table of Linear Exponential Family of Distributions

Table: Selected Distributions of the One-Parameter Exponential Family

Distribution	Parameters	Density or Mass Function	Components	E y	Var y
General	$\theta,\ \phi$	$\exp\left(\frac{y\theta-b(\theta)}{\phi}+S(y,\phi)\right)$	$\theta,\ \phi, b(\theta), S(y,\phi)$	$b'(\theta)$	$b''(\theta)\phi$
Normal	μ,σ^2	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$	$\mu,\sigma^2, \frac{\theta^2}{2}, -\left(\frac{y^2}{2\phi}+\frac{\ln(2\pi\phi)}{2}\right)$	$\theta=\mu$	$\phi=\sigma^2$
Binomal	π	$\binom{n}{y}\pi^y(1-\pi)^{n-y}$	$\ln\left(\frac{\pi}{1-\pi}\right), 1, n\ln(1+e^\theta),$ $\ln\binom{n}{y}$	$n\frac{e^\theta}{1+e^\theta}$ $=n\pi$	$n\frac{e^\theta}{(1+e^\theta)^2}$ $=n\pi(1-\pi)$
Poisson	λ	$\frac{\lambda^y}{y!}\exp(-\lambda)$	$\ln\lambda, 1, e^\theta, -\ln(y!)$	$e^\theta=\lambda$	$e^\theta=\lambda$
Gamma	α,β	$\frac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}\exp(-y\beta)$	$-\frac{\beta}{\alpha}, \frac{1}{\alpha}, -\ln(-\theta), -\phi^{-1}\ln\phi$ $-\ln\left(\Gamma(\phi^{-1})\right)+(\phi^{-1}-1)\ln y$	$-\frac{1}{\theta}=\frac{\alpha}{\beta}$	$\frac{\phi}{\theta^2}=\frac{\alpha}{\beta^2}$
Inverse Gaussian	μ,λ	$\sqrt{\frac{\lambda}{2\pi y^3}}\exp\left(-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right)$	$-\mu^2/2, 1/\lambda, -\sqrt{-2\theta},$ $\theta/(\phi y)-0.5\ln(\phi 2\pi y^3)$	$(-2\theta)^{-1/2}$ $=\mu$	$\phi(-2\theta)^{-3/2}$ $=\frac{\mu^3}{\lambda}$

Link Functions

- In the linear exponential family, we can show that
$$\mu_i = \mathrm{E}\, y_i = b'(\theta_i) \quad \text{and} \quad \mathrm{Var}\, y_i = \phi_i b''(\theta_i).$$
- Both θ and ϕ may vary by subject i
  - Because the θ determines the mean, we think of it as the mean, or location, parameter
  - Thus, think of ϕ as the scale, or dispersion, parameter
  - Typically, when the scale parameter varies by i, it is according to $\phi_i = \phi / w_i$, that is, a constant divided by a known weight w_i.
- Recall the link function
$$\eta_i = \mathbf{x}_i'\beta = \mathrm{g}(\mu_i) = \mathrm{g}(b'(\theta_i)).$$
  - The link function allows us to introduce explanatory variable to determine the mean.
- The model parameters are β and ϕ.

Choosing the Link Function

- The systematic component, $\eta_i = \mathbf{x}_i'\beta$, ranges over $(-\infty, \infty)$.
- Would like the range for $\mathrm{g}(\mu)$ to be comparable
  - Example, use the log-link, $\mathbf{x}_i'\beta = \ln(\mu_i)$, for Poisson regression.

Choosing the Link Function

- The systematic component, $\eta_i = \mathbf{x}_i'\beta$, ranges over $(-\infty, \infty)$.
- Would like the range for $\mathrm{g}(\mu)$ to be comparable
  - Example, use the log-link, $\mathbf{x}_i'\beta = \ln(\mu_i)$, for Poisson regression.
- Bernoulli distribution examples
  - Logit: $g(\mu) = \mathrm{logit}(\mu) = \ln(\mu/(1-\mu))$.
  - Probit: $g(\mu) = \Phi^{-1}(\mu)$.
  - Complementary log-log: $g(\mu) = \ln(-\ln(1-\mu))$.

Choosing the Link Function

- The systematic component, $\eta_i = \mathbf{x}_i'\beta$, ranges over $(-\infty,\infty)$.
- Would like the range for $\mathrm{g}(\mu)$ to be comparable
  - Example, use the log-link, $\mathbf{x}_i'\beta = \ln(\mu_i)$, for Poisson regression.
- Bernoulli distribution examples
  - Logit: $g(\mu) = \mathrm{logit}(\mu) = \ln(\mu/(1-\mu))$.
  - Probit: $g(\mu) = \Phi^{-1}(\mu)$.
  - Complementary log-log: $g(\mu) = \ln(-\ln(1-\mu))$.
- Another choice principle: The *canonical* link
  - The choice of g that is the inverse of $b'(\theta)$ is called the canonical link.
  - The systematic component equals the parameter of interest ($\eta = \theta$).

Table: Mean Functions and Canonical Links for Selected Distributions

Distribution	Mean function $b'(\theta)$	Canonical link $g(\mu)$
Normal	θ	μ
Bernoulli	$e^\theta/(1+e^\theta)$	$\mathrm{logit}(\mu)$
Poisson	e^θ	$\ln\mu$
Gamma	$-1/\theta$	$1/\mu$
Inverse Gaussian	$(-2\theta)^{-1/2}$	$1/\mu^2$

Maximum Likelihood Estimation

- The usual method of parameter estimation is maximum likelihood.
- For example, the log-likelihood is

$$\ln \mathrm{f}(\mathbf{y}) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi_i} + S(y_i, \phi_i) \right\}.$$

- With a canonical link, $\theta_i = \eta_i = \mathbf{x}_i' \beta$.
- See the text for more information on this topic ...

Overdispersion

- For some distributions, such as the normal and gamma distributions, we estimate ϕ after the estimation of β, using maximum likelihood.
- For others, such as the binomial and Poisson, the scale parameter ϕ is known.
  - Although the scale parameter is theoretically known, the data suggest a different value.
  - We introduce an extra parameter that can be estimated from the data
  - This is known as “quasi-binomial” or “quasi-Poisson”.
  - The variance is of the form $\mathrm{Var}\; y_i = \sigma^2 \phi b''(\theta_i)/w_i$.
  - Can estimate the additional scale parameter σ^2 as a Pearson’s chi-square statistic divided by the error degrees of freedom. That is,

$$\widehat{\sigma}^2 = \frac{1}{N-K} \sum_{i=1}^{n} w_i \frac{\left(y_i - b'(\mathbf{x}_i'\mathbf{b}_{MLE})\right)^2}{\phi b''(\mathbf{x}_i'\mathbf{b}_{MLE})}.$$

Goodness of Fit Statistics

- R^2 is not a useful statistic in nonlinear models, in part because of the analysis of variance decomposition is no longer valid.
  - The Sum of Cross-Products is not zero in non-linear models.

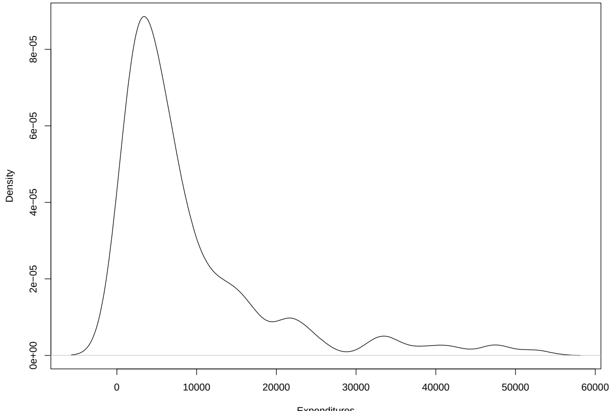
$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \widehat{y}_i)^2 + \sum_i (\widehat{y}_i - \bar{y})^2 + 2 \times \sum_i (y_i - \widehat{y}_i)(\widehat{y}_i - \bar{y}).$$

Total SS = Error SS + Regression SS + 2 × Sum of Cross-Products.

- For discrete data, consider reporting Pearson's chi-square statistic (either grouped or individual)
- General information criteria, including AIC and BIC, (defined in Section 11.9) are regularly cited in GLM studies.

MEPS Data

- There are 157 people with positive inpatient expenditures
- Smooth Empirical Histogram of Positive Inpatient Expenditures. The largest expenditure is omitted.
- The skewed histogram suggests using a gamma distribution.



Median Expenditures by Explanatory Variable - $n = 157$ with Positive Expends

Category	Variable	Description	Percent of data	Median Expend
	COUNTIP	Number of expenditures (median: 1.0)		
Demography	AGE	Age in years between 18 to 65 (median: 41.0)		
	SEX	1 if female	72.0	5,546
		0 if male	28.0	7,313
Ethnicity	ASIAN	1 if Asian	2.6	4,003
	BLACK	1 if Black	19.8	6,100
	NATIVE	1 if Native	1.9	2,310
	WHITE	Base category	75.6	5,695
Region	NORTHEAST	1 if Northeast	18.5	5,833
	MIDWEST	1 if Midwest	21.7	7,999
	SOUTH	1 if South	40.8	5,595
	WEST	Base category	19.1	4,297
Education	COLLEGE	1 if college or higher degree	23.6	5,611
	HIGHSCHOOL	1 if high school degree	43.3	5,907
	Base category is lower than high school degree		33.1	5,338
Self-rated physical health	POOR	1 if poor	17.2	10,447
	FAIR	1 if fair	10.2	5,228
	GOOD	1 if good	31.2	5,032
	VGOOD	1 if very good	24.8	5,546
	Base category is excellent health		16.6	5,277
Self-rated mental health	MPOOR	1 if poor or fair	15.9	6,583
		0 if good to excellent mental health	84.1	5,599
Any activity limitation	ANYLIMIT	1 if any functional/activity limitation	41.4	7,826
		0 if otherwise	58.6	4,746
Income compared to poverty line	Base category is high income		21.7	7,271
	MINCOME	1 if middle income	26.8	5,851
	LINCOME	1 if low income	16.6	6,909
	NPOOR	1 if near poor	7.0	5,546
	POORNEG	if poor/negative income	28.0	4,097
Insurance coverage	INSURE	1 if covered by public/private health insurance in any month of 2003	91.1	5,943
		0 if have no health insurance in 2003	8.9	2,668
Total			100.0	5,695

MEPS Data

- Percent of Data
  - The Table shows that the sample is 72% female, almost 76% white and over 91% insured.
  - There are relatively few expenditures by Asians, Native Americans and the uninsured in our sample.
- Median Expenditures by categorical variable
- Potentially important determinants of the amount of medical expenditures
  - gender,
  - a poor self-rating of physical health and
  - income that is poor or negative.

Gamma and Inverse Gaussian Regression Models

	Gamma				Inverse Gaussian	
	Full Model		Reduced Model		Reduced Model	
Effect	Parameter Estimate	*t*-value	Parameter Estimate	*t*-value	Parameter Estimate	*t*-value
Intercept	6.891	13.080	7.859	17.951	6.544	3.024
COUNTIP	0.681	6.155	0.672	5.965	1.263	0.989
AGE	0.021	3.024	0.015	2.439	0.018	0.727
GENDER	-0.228	-1.263	-0.118	-0.648	0.363	0.482
ASIAN	-0.506	-1.029				
BLACK	-0.331	-1.656	-0.258	-1.287	-0.321	-0.577
NATIVE	-1.220	-2.217				
NORTHEAST	-0.372	-1.548	-0.214	-0.890	0.109	0.165
MIDWEST	0.255	1.062	0.448	1.888	0.399	0.654
SOUTH	0.010	0.047	0.108	0.516	0.164	0.319
COLLEGE	-0.413	-1.723	-0.469	-2.108	-0.367	-0.606
HIGHSCHOOL	-0.155	-0.827	-0.210	-1.138	-0.039	-0.078
POOR	-0.003	-0.010	0.167	0.706	0.167	0.258
FAIR	-0.194	-0.641				
GOOD	0.041	0.183				
VGOOD	0.000	0.000				
MNHPOOR	-0.396	-1.634	-0.314	-1.337	-0.378	-0.642
ANYLIMIT	0.010	0.053	0.052	0.266	0.218	0.287
MINCOME	0.114	0.522				
LINCOME	0.536	2.148				
NPOOR	0.453	1.243				
POORNEG	-0.078	-0.308	-0.406	-2.129	-0.356	-0.595
INSURE	0.794	3.068				
Scale	1.409	9.779	1.280	9.854	0.026	17.720
Log-Likelihood	-1,558.67		-1,567.93		-1,669.02	
AIC	3,163.34		3,163.86		3,366.04	

Gamma and Inverse Gaussian Regression Models

- A gamma regression model using a logarithmic link was fit to inpatient expenditures using all explanatory variables.
    - Many variables are not statistically significant.
    - Common in expenditure analysis, where variables help predict the frequency although are not as useful in explaining severity.
    - Collinearity - too many variables in a fitted model can lead to statistical insignificance of important variables and even cause signs to be reversed.
- Reduced Model
    - Removed the Asian, Native American and the uninsured variables - they account for a small subset of our sample.
    - Used only the POOR variable for self-reported health status and only POORNEG for income, essentially reducing these categorical variables to binary variables.
    - AIC is about the same as the full model - a reasonable alternative.
    - The variables COUNTIP (inpatient count), AGE, COLLEGE and POORNEG, are statistically significant variables.
- Also fit of an inverse gaussian model with a log link.
    - AIC - does not fit nearly as well as the gamma regression model.
    - All variables are statistically insignificant - difficult to interpret.

Tweedie Distribution

- The Tweedie distribution is defined as a Poisson sum of gamma random variables
  - Suppose that N has a Poisson distribution with mean λ, representing the number of claims.
  - Let y_j be i.i.d., independent of N
  - Each y_j has a gamma distribution with parameters α and β, representing the amount of a claim.
  - $S_N = y_1 + \ldots + y_N$ is Poisson sum of gammas.
- The Tweedie distribution
  - Discrete component - the probability of zero claims is

$$\Pr(S_N = 0) = \Pr(N = 0) = e^{-\lambda}.$$

  - Continuous component - for $y > 0$, the density is

$$\mathrm{f}_S(y) = \sum_{n=1}^{\infty} e^{-\lambda} \frac{\lambda^n}{n!} \frac{\beta^{n\alpha}}{\Gamma(n\alpha)} y^{n\alpha-1} e^{-y\beta}.$$

Tweedie Distribution and GLM

- We can define three parameters μ, ϕ, p through the relations

$$\lambda = \frac{\mu^{2-p}}{\phi(2-p)}, \qquad \alpha = \frac{2-p}{p-1} \qquad \text{and} \qquad \frac{1}{\beta} = \phi(p-1)\mu^{p-1}.$$

- With this new parameterization, it can be readily shown that the Tweedie distribution is a member of the linear exponential family.
- Easy calculations show that

$$\mathrm{E}\ S_N = \mu \qquad \text{and} \qquad \mathrm{Var}\ S_N = \phi \mu^p,$$

where $1 < p < 2$.

  - Thus, the Tweedie can be used in a GLM with μ as a function of the systematic component η.
  - Examining variances, the Tweedie distribution can also be viewed as a choice that is intermediate between the Poisson ($p = 1$) and the gamma ($p = 2$) distributions.