Bayesian Methods: Review of Generalized Linear Models

RYAN BAKKER
University of Georgia

ICPSR Day 2

Likelihood and Maximum Likelihood Principles

- Likelihood theory is an important part of Bayesian inference: it is how the data enter the model.
- The basis is Fisher's principle: *what value of the unknown parameter is "most likely" to have generated the observed data.*
- Example: flip a coin 10 times, get 5 heads. MLE for p is 0.5.
- This is easily the most common and well-understood *general* estimation process.

- Starting details:
  - $\mathbf{Y}$ is a $n \times k$ design or observation matrix, $\boldsymbol{\theta}$ is a $k \times 1$ unknown coefficient vector to be estimated, we want $p(\boldsymbol{\theta}|\mathbf{Y})$ (joint sampling distribution or posterior) from $p(\mathbf{Y}|\boldsymbol{\theta})$ (joint probability function).
  - Define the likelihood function:
$$L(\boldsymbol{\theta}|\mathbf{Y}) = \prod_{i=1}^{n} p(Y_i|\boldsymbol{\theta})$$
which is no longer on the probability metric.
  - Our goal is the maximum likelihood value of $\boldsymbol{\theta}$:
$$\hat{\boldsymbol{\theta}} : \; L(\hat{\boldsymbol{\theta}}|\mathbf{Y}) \geq L(\boldsymbol{\theta}|\mathbf{Y}) \quad \forall \boldsymbol{\theta} \in \Theta$$
where Θ is the class of *admissable* values for $\boldsymbol{\theta}$.

Likelihood and Maximum Likelihood Principles (cont.)

- Its actually easier to work with the natural log of the likelihood function:

$$\ell(\boldsymbol{\theta}|\mathbf{Y}) = \log L(\boldsymbol{\theta}|\mathbf{Y})$$

- We also find it useful to work with the *score function*, the first derivative of the log likelihood function with respect to the parameters of interest:

$$\dot{\ell}(\boldsymbol{\theta}|\mathbf{Y}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}|\mathbf{Y})$$

- Setting $\dot{\ell}(\boldsymbol{\theta}|\mathbf{Y})$ equal to zero and solving gives the *MLE*: $\hat{\boldsymbol{\theta}}$, the "most likely" value of $\boldsymbol{\theta}$ from the parameter space Θ treating the observed data as given.

- The *Likelihood Principle* (Birnbaum 1962) states that once the data are observed, and therefore treated as given, all of the available evidence for estimating $\hat{\boldsymbol{\theta}}$ is contained in the (log) likelihood function, $\ell(\boldsymbol{\theta}|\mathbf{Y})$.
- Setting the score function from the joint PDF or PMF equal to zero and rearranging gives the likelihood equation:

$$\sum t(y_i) = n\frac{\partial}{\partial \boldsymbol{\theta}} E[\mathbf{y}]$$

where $\sum t(y_i)$ is the remaining function of the data,

Likelihood and Maximum Likelihood Principles (cont.)

- Nice properties of the MLE (very well known):
  - log likelihood unimodal for exponential family forms,
  - consistent,
  - asymptotically efficient (reaches the CRLB),
  - asymptotically normal: $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{P}} \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{\theta}})$,
  - $\sum t(y_i)$ is *sufficient* for $\boldsymbol{\theta}$.

- Example: Linear Regression.

  The likelihood equation for the residuals is:

$$\begin{aligned} L(\boldsymbol{\epsilon}) &= (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}\boldsymbol{\epsilon}'\boldsymbol{\epsilon}\right] \\ &= (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb})\right]. \end{aligned}$$

  $L(\boldsymbol{\epsilon})$ is concave for this equation, (not always guaranteed). The log of $L(\boldsymbol{\epsilon})$ is maximized at the same point as the function itself, so take the derivative with respect to b of the easier function, and solve for zero:

$$\begin{aligned} \log L(\boldsymbol{\epsilon}) &= -\frac{1}{n}\log(2\pi) - \frac{1}{n}\log(2\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb}) \\ \frac{\partial}{\partial b}\log L(\boldsymbol{\epsilon}) &= \frac{1}{\sigma^2}\mathbf{X}'(\mathbf{y}-\mathbf{Xb}) \equiv 0 \\ \hat{\mathbf{b}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \end{aligned}$$

Exponential Family Form

- The exponential family form of a PDF or PMF is:

$$\begin{aligned} f(z|\zeta) &= \exp\big[t(z)u(\zeta)\big]r(z)s(\zeta) \\ &= \exp\big[t(z)u(\zeta) + \log r(z) + \log s(\zeta)\big], \end{aligned}$$

  where: r and t are real-valued functions of z that do not depend on ζ, and s and u are real-valued functions of ζ that do not depend on z, and $r(z) > 0,\ s(\zeta) > 0\ \forall z, \zeta$.

- The canonical form obtained by transforming: $y = t(z)$, and $\theta = u(\zeta)$. Call θ the canonical parameter. This produces the final form:

$$f(y|\theta) = \exp\big[y\theta - b(\theta) + c(y)\big].$$

- The exponential family form is invariant to sampling:

$$f(\mathbf{y}|\theta) = \exp\big[\sum y_i\theta - nb(\theta) + \sum c(y_i)\big].$$

- And there often exists a *scale parameter*:

$$f(\mathbf{y}|\theta) = \exp\left[\frac{\sum y_i\theta - nb(\theta)}{\phi} + \sum c(y_i, \phi)\right].$$

Exponential Family Form (cont.)

- Example: normal PDF.

$$f(y|\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right]$$

$$= \exp\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y-\mu)^2\right]$$

$$= \exp\left[(\underbrace{y\mu}_{y\theta} - \underbrace{\frac{\mu^2}{2}}_{b(\theta)})/\underbrace{\sigma^2}_{\phi} + \underbrace{\frac{-1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)}_{c(y,\phi)}\right].$$

- Why is this handy
  - Consider the score function in this notation:

$$\dot{\ell}(\theta|\phi, y) = \frac{y - \frac{\partial}{\partial\theta}b(\theta)}{\phi}$$

  - Which actually has n data values:

$$\dot{\ell}(\theta|\phi, \mathbf{y}) = \frac{\sum t(y_i) - n\frac{\partial}{\partial\theta}b(\theta)}{\phi}$$

  - We then set this equal to zero and rearrange to get the *normal equation*:

$$\sum t(y_i) = n\frac{\partial}{\partial\theta}b(\theta)$$

  - Returning to the normal case:

$$b(\theta) = \frac{\theta^2}{2}, \text{ and } t(y) = y, \text{ so } \hat{\theta} = \frac{1}{n}\sum y_i.$$

Generalized Linear Model Theory

The Generalization

Start with the standard linear model meeting the Gauss-Markov conditions:

$$\underset{(n\times 1)}{\mathbf{V}} = \underset{(n\times p)(p\times 1)}{\boldsymbol{X}\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\epsilon}} \tag{1}$$

$$\underset{(n\times 1)}{E[\mathbf{V}]} = \underset{(n\times 1)}{\boldsymbol{\theta}} = \underset{(n\times p)(p\times 1)}{\boldsymbol{X}\boldsymbol{\beta}} \tag{2}$$

Generalize slightly with a new "linear predictor" based on the mean of the outcome variable:

$$\underset{(n\times 1)}{g(\boldsymbol{\mu})} = \underset{(n\times 1)}{\boldsymbol{\theta}} = \underset{(n\times p)(p\times 1)}{\boldsymbol{X}\boldsymbol{\beta}}$$

The generalization of the linear model has 4 components:

I. **Stochastic Component:** $\mathbf{Y}$ is the random or stochastic component which remains distributed i.i.d. according to a specific exponential family distribution with mean $\boldsymbol{\mu}$.

II. **Systematic Component:** $\boldsymbol{\theta} = \boldsymbol{X\beta}$ is the systematic component with an associated Gauss-Markov normal basis.

III. **Link Function:** the stochastic component and the systematic component are linked by a function of $\boldsymbol{\theta}$ which is *exactly the canonical link function*, summarized in the Table below. We can think of $g(\boldsymbol{\mu})$ as "tricking" the linear model into thinking that it is still acting upon normally distributed outcome variables.

IV. **Residuals:** Although the residuals can be expressed in the same manner as in the standard linear model, observed outcome variable value minus predicted outcome variable value, a more useful quantity is the deviance residual described in detail below.

Useful Distributions:

- Poisson Distribution
  - often used to model counts such as the number of arrivals, deaths, or failures, in a given time period.
  - assumes that for short time intervals, the probability of an arrival is fixed and proportional to the length of the interval.
  - indexed by "intensity parameter" equal to mean and variance.
- Binomial Distribution
  - summarizes the outcome of multiple binary outcome (Bernoulli) trials such as flipping a coin.
  - useful for modeling counts of success or failures given a number independent trials such as votes received given an electorate, international wars given country-dyads in a region, or bankruptcies given company starts.
  - parameterized by number of trials (n) and probability of success (p).

- Normal Distribution
  - outcome variable is interval measured and unbounded.
  - produces standard linear model.
- Gamma Distribution
  - useful for modeling terms that are required to be non-negative such as variances.
  - two important special cases: the χ^2 distribution is gamma$(\frac{\rho}{2}, \frac{1}{2})$ for ρ degrees of freedom, exponential distribution is gamma$(1, \beta)$.
- Negative Binomial Distribution
  - models the number of failures (y) before the r^{th} success.
  - parameterized by r and p.

Table 1: Natural Link Function Summary for Example Distributions

Distribution		Canonical Link: $\theta = g(\mu)$	Inverse Link: $\mu = g^{-1}(\theta)$
Poisson		$log(\mu)$	$\exp(\theta)$
Binomial	*logit link:*	$\log\left(\frac{\mu}{1-\mu}\right)$	$\frac{\exp(\theta)}{1+\exp(\theta)}$
	probit link:	$\Phi^{-1}(\mu)$	$\Phi(\theta)$
	cloglog link:	$\log(-\log(1-\mu))$	$1-\exp(-\exp(\theta))$
Normal		μ	θ
Gamma		$-\frac{1}{\mu}$	$-\frac{1}{\theta}$
Negative Binomial		$\log(1-\mu)$	$1-\exp(\theta)$

Poisson GLM of Capital Punishment Data

The model is developed from the Poisson link function, $\boldsymbol{\theta} = \log(\boldsymbol{\mu})$, with the objective of finding the best $\boldsymbol{\beta}$ vector in:

$$\begin{aligned}
\underbrace{g^{-1}(\boldsymbol{\theta})}_{17\times 1} &= g^{-1}(\boldsymbol{X}\boldsymbol{\beta}) \\
&= \exp\left[\boldsymbol{X}\boldsymbol{\beta}\right] \\
&= \exp\left[\mathbf{1}\beta_0 + \mathbf{INC}\beta_1 + \mathbf{POV}\beta_2 + \mathbf{BLK}\beta_3 + \mathbf{CRI}\beta_4 + \mathbf{SOU}\beta_5 + \mathbf{DEG}\beta_6\right] \\
&= E[\mathbf{Y}] = E[\mathbf{EXE}].
\end{aligned}$$

Table 2: Capital Punishment in the United States – 1997

State	Executions	Median Income	Percent Poverty	Percent Black	Violent Crime/100K	South	Proportion w/Degrees
Texas	37	34453	16.7	12.2	644	1	0.16
Virginia	9	41534	12.5	20.0	351	1	0.27
Missouri	6	35802	10.6	11.2	591	0	0.21
Arkansas	4	26954	18.4	16.1	524	1	0.16
Alabama	3	31468	14.8	25.9	565	1	0.19
Arizona	2	32552	18.8	3.5	632	0	0.25
Illinois	2	40873	11.6	15.3	886	0	0.25
South Carolina	2	34861	13.1	30.1	997	1	0.21
Colorado	1	42562	9.4	4.3	405	0	0.31
Florida	1	31900	14.3	15.4	1051	1	0.24
Indiana	1	37421	8.2	8.2	537	0	0.19
Kentucky	1	33305	16.4	7.2	321	0	0.16
Louisiana	1	32108	18.4	32.1	929	1	0.18
Maryland	1	45844	9.3	27.4	931	0	0.29
Nebraska	1	34743	10.0	4.0	435	0	0.24
Oklahoma	1	29709	15.2	7.7	597	0	0.21
Oregon	1	36777	11.7	1.8	463	0	0.25
	EXE	**INC**	**POV**	**BLK**	**CRI**	**SOU**	**DEG**

Source: United States Census Bureau, United States Department of Justice.

```
dp.97 <- read.table("http://web.clas.ufl.edu/~jgill/GLM.Data/cpunish.dat",header=T)
attach(dp.97)
dp.out <- glm(EXECUTIONS ~ INCOME + PERPOVERTY + PERBLACK + log(VC100k96) + SOUTH
                        + PROPDEGREE, family=poisson)

glm.summary <- function (in.object, alpha = 0.05)
{
      lo <- in.object$coefficient - qnorm(1-alpha/2) * sqrt(diag(summary(in.object)$cov.unscaled))
      hi <- in.object$coefficient + qnorm(1-alpha/2) * sqrt(diag(summary(in.object)$cov.unscaled))
      out.mat <- round(cbind(in.object$coefficient, sqrt(diag(glm.vc(in.object))), lo, hi),5)
      dimnames(out.mat)[[2]] <- c("Coefficient","Std. Error",
            paste(1-alpha,"CI Lower"),paste(1-alpha,"CI Upper"))
      out.mat
}

glm.summary(dp.out)
```

	Coefficient	Std. Error	0.95 CI Lower	0.95 CI Upper
(Intercept)	-6.30665	4.17678	-14.49299	1.87969
INCOME	0.00027	0.00005	0.00017	0.00037
PERPOVERTY	0.06897	0.07979	-0.08741	0.22534
PERBLACK	-0.09500	0.02284	-0.13978	-0.05023
log(VC100k96)	0.22124	0.44243	-0.64591	1.08838
SOUTH	2.30988	0.42875	1.46955	3.15022
PROPDEGREE	-19.70241	4.46366	-28.45102	-10.95380

Table 3: Modeling Capital Punishment in the United States: 1997

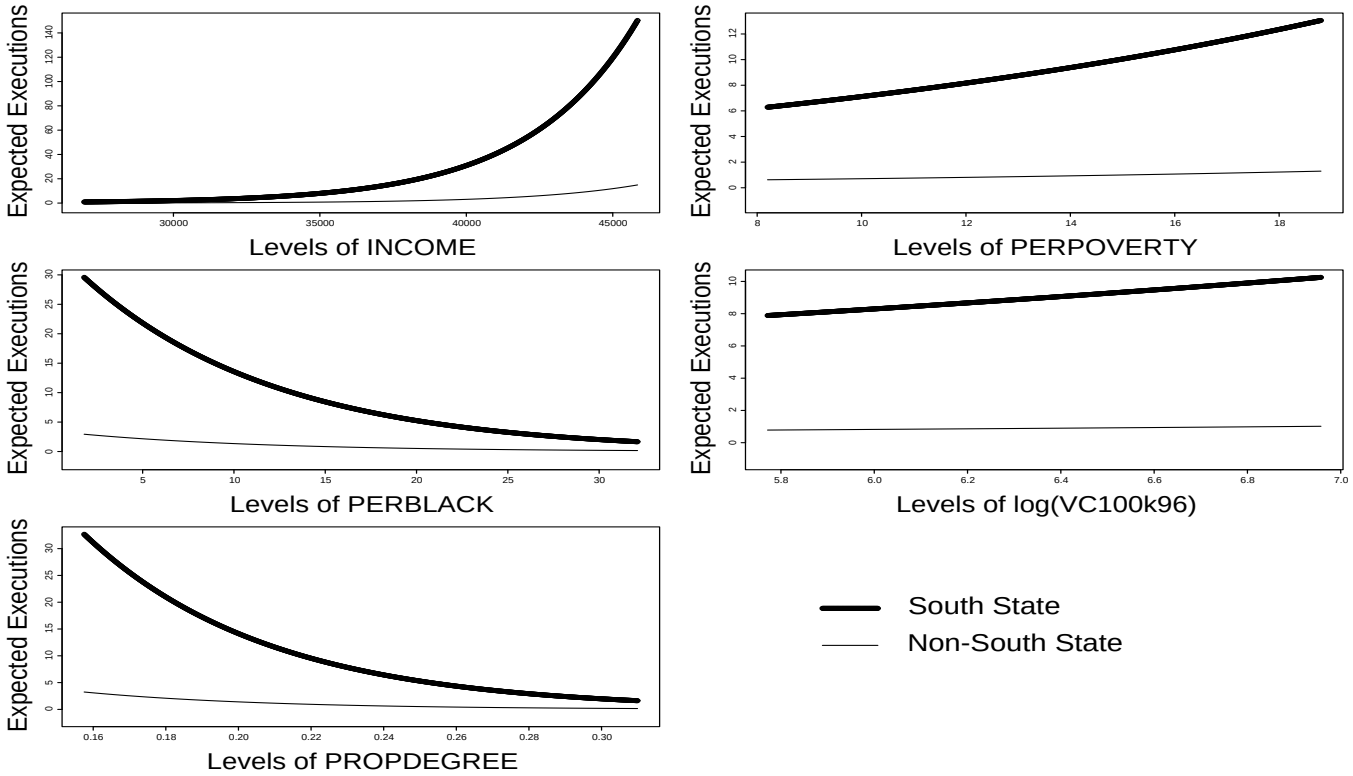
	Coefficient	Standard Error	95% Confidence Interval
(Intercept)	-6.30665	4.17678	[-14.49299: 1.87969]
Median Income	0.00027	0.00005	[0.00017: 0.00037]
Percent Poverty	0.06897	0.07979	[-0.08741: 0.22534]
Percent Black	-0.09500	0.02284	[-0.13978: -0.05023]
log(Violent Crime)	0.22124	0.44243	[-0.64591: 1.08838]
South	2.30988	0.42875	[1.46955: 3.15022]
Degree Proportion	-19.70241	4.46366	[-28.45102:-10.95380]

Null deviance: 136.573, $df = 16$ Maximized $\ell()$: -31.7375

Summed deviance: 18.212, $df = 11$ AIC: 77.475

$\mathbf{VC} = (-\mathbf{A})^{-1} =$

Int	**INC**	**POV**	**BLK**	log(**CRI**)	**SOU**	**DEG**
17.445501654	−0.000131052	−0.198325558	0.017689695	−1.484011921	0.368916884	−4.651658695
−0.000131052	0.000000003	0.000001862	0.000000113	0.000004171	−0.000006245	−0.000094858
−0.198325558	0.000001862	0.006365688	0.000158039	0.003911954	−0.017825119	0.121451892
0.017689695	0.000000113	0.000158039	0.000521871	−0.003283494	−0.005090192	−0.033679253
−1.484011921	0.000004171	0.003911954	−0.003283494	0.195742167	−0.001384018	0.397439934
0.368916884	−0.000006245	−0.017825119	−0.005090192	−0.001384018	0.183825030	0.298730196
−4.651658695	−0.000094858	0.121451892	−0.033679253	0.397439934	0.298730196	19.924250374


Expected Executions
Levels of INCOME
Levels of PERPOVERTY
Levels of PERBLACK
Levels of log(VC100k96)
Levels of PROPDEGREE
South State
Non-South State

Gamma GLM of Electoral Politics in Scotland

- On September 11, 1997 Scottish voters overwhelming (74.3%) approved the establishment of the first Scottish national parliament in nearly three hundred years.
- On the same ballot, the voters gave strong support (63.5%) to granting this parliament taxation powers.
- Data: 32 *Unitary Authorities* (also called council districts), U.K. government sources, includes 40 potential explanatory variables

The model for these data using the gamma link function is produced by:

$$\begin{aligned}
\underbrace{g^{-1}(\boldsymbol{\theta})}_{32\times 1} &= g^{-1}(\boldsymbol{X\beta}) \\
&= -\frac{1}{\boldsymbol{X\beta}} \\
&= -\left[\mathbf{1}\beta_0 + \mathbf{COU}\beta_1 + \mathbf{UNM}\beta_2 + \mathbf{MOR}\beta_3 + \mathbf{ACT}\beta_4 + \mathbf{AGE}\beta_5\right]^{-1} \\
&= E[\mathbf{Y}] = E[\mathbf{YES}].
\end{aligned}$$

The systematic component here is $\boldsymbol{X\beta}$, the stochastic component is $\mathbf{Y} = \mathbf{YES}$, and the link function is $\boldsymbol{\theta} = -\frac{1}{\boldsymbol{\mu}}$.

Table 4: Taxation Powers Vote for the Scottish Parliament – 1997

	Proportion Voting Yes	Council Tax	% Female Unemploy.	Standardized Mortality	% Active Economically	% Aged 5–15
Aberdeen City	0.603	712	21.0	105	82.4	12.3
Aberdeenshire	0.523	643	26.5	97	80.2	15.3
Angus	0.534	679	28.3	113	86.3	13.9
Argyll & Bute	0.570	801	27.1	109	80.4	13.6
Clackmannanshire	0.687	753	22.0	115	64.7	14.6
Dumfries & Galloway	0.488	714	24.3	107	79.0	13.8
Dundee City	0.655	920	21.2	118	72.2	13.3
East Ayrshire	0.705	779	20.5	114	75.2	14.5
East Dunbartonshire	0.591	771	23.2	102	81.1	14.2
East Lothian	0.627	724	20.5	112	80.3	13.7
East Renfrewshire	0.516	682	23.8	96	83.0	14.6
Edinburgh City	0.620	837	22.1	111	74.5	11.6
Western Isles	0.684	599	19.9	117	83.8	15.1
Falkirk	0.692	680	21.5	121	77.6	13.7
Fife	0.647	747	22.5	109	77.9	14.4
Glasgow City	0.750	982	19.4	137	65.3	13.3
Highland	0.621	719	25.9	109	80.9	14.9
Inverclyde	0.672	831	18.5	138	80.2	14.6
Midlothian	0.677	858	19.4	119	84.8	14.3
Moray	0.527	652	27.2	108	86.4	14.6
North Ayrshire	0.657	718	23.7	115	73.5	15.0
North Lanarkshir	0.722	787	20.8	126	74.7	14.9
Orkney Islands	0.474	515	26.8	106	87.8	15.3
Perth and Kinross	0.513	732	23.0	103	86.6	13.8
Renfrewshire	0.636	783	20.5	125	78.5	14.1
Scottish Borders	0.507	612	23.7	100	80.6	13.3
Shetland Islands	0.516	486	23.2	117	84.8	15.9
South Ayrshire	0.562	765	23.6	105	79.2	13.7
South Lanarkshire	0.676	793	21.7	125	78.4	14.5
Stirling	0.589	776	23.0	110	77.2	13.6
West Dunbartonshire	0.747	978	19.3	130	71.5	15.3
West Lothian	0.673	792	21.2	126	82.2	15.1
	YES	**COU**	**UNM**	**MOR**	**ACT**	**AGE**

Source: U.K. Office for National Statistics, the General Register Office for Scotland, the Scottish Office.

```
scotland.df <- read.table("http://web.clas.ufl.edu/~jgill/GLM.Data/scotvote.dat",header=T)
scottish.vote.glm <- glm((PerYesTax/100) ~ CouncilTax * PerClaimantFemale + StdMortalityRatio
                              + Active + GDP + Percentage5to15,
                    family=Gamma, data=scotland.df)
```

Table 5: Modeling the Vote for Parliamentary Taxation: 1997

	Coefficient	Standard Error	95% Confidence Interval
(Intercept)	-1.77653	1.14789	[-4.14566: 0.59261]
Council Tax	0.00496	0.00162	[0.00162: 0.00831]
Female Unemployment	0.20344	0.05321	[0.09363: 0.31326]
Standardized Mortality	-0.00718	0.00271	[-0.01278:-0.00159]
Economically Active	0.01119	0.00406	[0.00281: 0.01956]
GDP	-0.00001	0.00001	[-0.00004: 0.00001]
Percent Aged 5–15	-0.05187	0.02403	[-0.10145:-0.00228]
Council Tax:Female Un.	-0.00024	0.00007	[-0.00040:-0.00009]

Null deviance: 0.536072, $df = 31$ Maximized $\ell()$: 63.89

Summed deviance: 0.087389, $df = 24$ AIC: -111.78

RESIDUALS AND MODEL FIT

- General Notation: $D = \sum_{i=1}^{n} d(\boldsymbol{\theta}, y_i)$
- Linear Model Residual Vector: $\mathbf{R}_{standard} = \mathbf{Y} - \mathbf{X}\beta$
- Response Residual Vector: $\mathbf{R}_{Response} = \mathbf{Y} - g^{-1}(\boldsymbol{X\beta})$
- Pearson Residual Vector: $\mathbf{R}_{Pearson} = \frac{\mathbf{Y}-\boldsymbol{\mu}}{\sqrt{VAR[\boldsymbol{\mu}]}}$ (the sum of the Pearson residuals for a Poisson generalized linear model is the Pearson χ^2 goodness-of-fit measure)
- $\mathbf{R}_{Working} = (\mathbf{y} - \boldsymbol{\mu})\frac{\partial}{\partial\theta}\boldsymbol{\mu}$ (from the last step of Iteratively Reweighted Least Squares algorithm).

Individual Deviance Function:

$$R_{Deviance} = \frac{(y_i - \mu_i)}{|y_i - \mu_i|}\sqrt{|d(\boldsymbol{\theta}, y_i)|} \qquad \text{where:} \qquad d(\boldsymbol{\theta}, y_i) = -2\left[\ell(\hat{\boldsymbol{\theta}}, \psi|y_i) - \ell(\tilde{\boldsymbol{\theta}}, \psi|y_i)\right].$$

Table 6: Deviance Functions

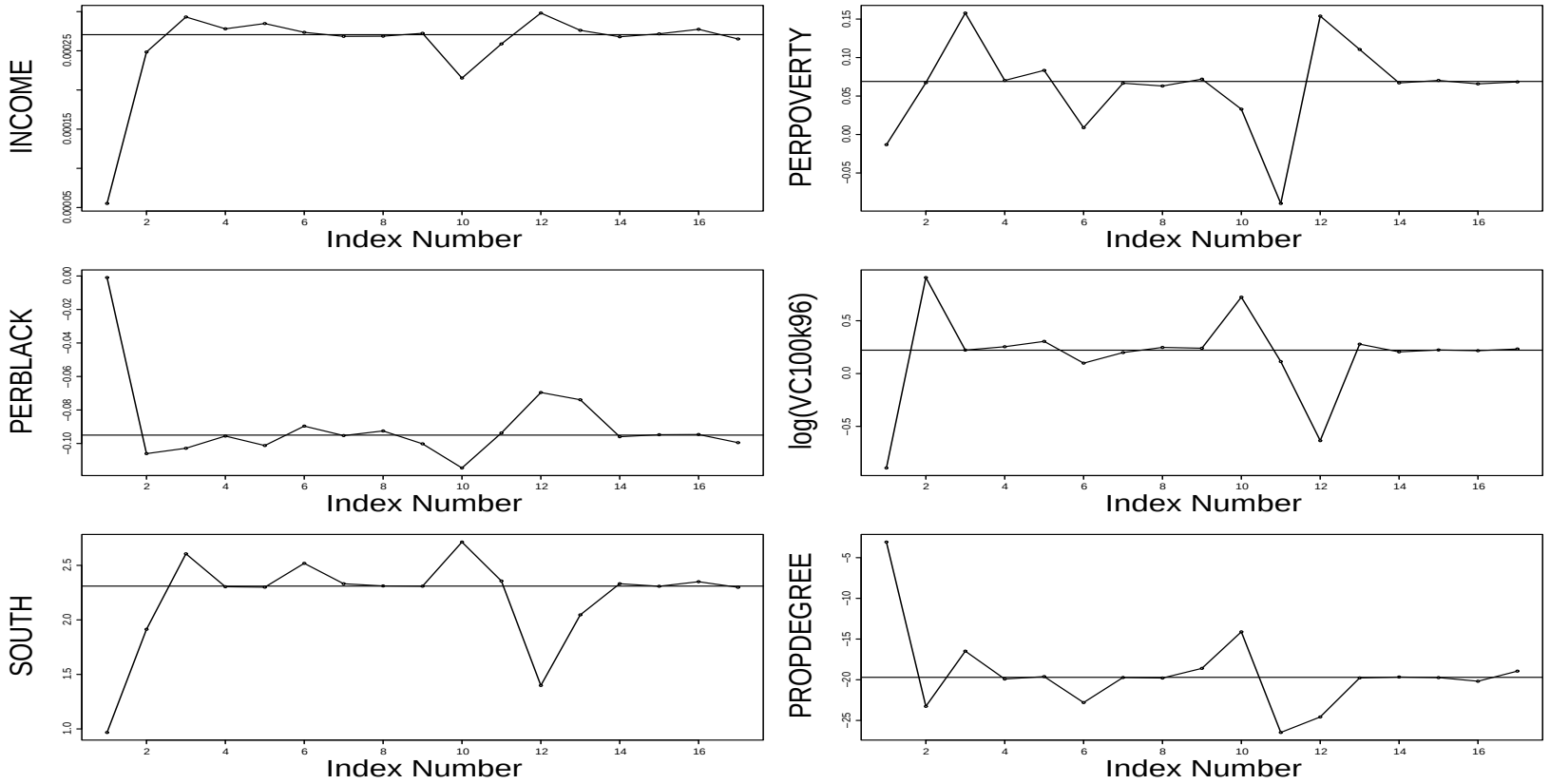
Distribution	Canonical Parameter	Deviance Function
Poisson(μ)	$\theta = log(\mu)$	$2\sum\left[y_i \log\left(\frac{y_i}{\mu_i}\right) - y_i + \mu_i\right]$
Binomial(m,p)	$\theta = log\left(\frac{\mu}{1-\mu}\right)$	$2\sum\left[y_i \log\left(\frac{y_i}{\mu_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \mu_i}\right)\right]$
Normal(μ, σ)	$\theta = \mu$	$\sum\left[y_i - \mu_i\right]^2$
Gamma(μ, δ)	$\theta = -\frac{1}{\mu}$	$2\sum\left[-\log\left(\frac{y_i}{\mu_i}\right)\frac{y_i - \mu_i}{\mu_i}\right]$
Negative Binom(μ, p)	$\theta = \log(1-\mu)$	$2\sum\left[y_i \log\left(\frac{y_i}{\mu_i}\right) + (1+y_i)\log\left(\frac{1+\mu_i}{1+y_i}\right)\right]$

Poisson GLM of Capital Punishment, Continued

Table 7: Residuals From Poisson Model of Capital Punishment

	Response	Pearson	Working	Deviance	Anscombe
Texas	1.70755431	0.28741478	0.04837752	0.28515874	0.28292493
Virginia	0.87407687	0.30671010	0.10762321	0.30136452	0.29629097
Missouri	4.59530299	3.86395636	3.24898061	2.86925916	2.27854829
Arkansas	0.26481208	0.13694108	0.07081505	0.13544624	0.13391171
Alabama	0.95958171	0.67097152	0.46916278	0.62736060	0.58874967
Arizona	0.95395198	0.93375106	0.91397549	0.82741022	0.74425671
Illinois	0.13924315	0.10197129	0.07467388	0.10084230	0.09963912
South Carolina	-0.38227185	-0.24752186	-0.16027167	-0.25478237	-0.26235519
Colorado	-0.95901329	-0.68428704	-0.48826435	-0.75706323	-0.84845827
Florida	-1.82216650	-1.08543456	-0.64657649	-1.25272634	-1.49557143
Indiana	-2.17726883	-1.21566195	-0.67880001	-1.42915840	-1.74185735
Kentucky	-2.31839936	-1.26926054	-0.69489994	-1.49593905	-1.83715998
Louisiana	-1.60160305	-0.99359914	-0.61640776	-1.13620002	-1.33738726
Maryland	0.10161119	0.10709684	0.11287657	0.10527242	0.10341466
Nebraska	0.07022962	0.07261924	0.07506941	0.07194451	0.07107841
Oklahoma	0.49917358	0.70406163	0.99304011	0.62019695	0.55401828
Oregon	-0.90510552	-0.65451282	-0.47330769	-0.72189767	-0.80517526

Figure 1: Jackknife Index Plot: Capital Punishment Model



New and Old Ways to Look at Model Fit

- Approximation to Pearson's Statistic.

$$X^2 = \sum_{i=1}^{n} \mathbf{R}^2_{Pearson} = \sum_{i=1}^{n} \left[\frac{\mathbf{Y} - \boldsymbol{\mu}}{\sqrt{VAR[\boldsymbol{\mu}]}} \right]^2 . \tag{3}$$

  If the sample size is sufficiently large, then $\frac{X^2}{a(\phi)} \sim \chi^2_{n-p}$ where n is the sample size and p is the number of explanatory variables including the constant.

- Summed Deviance.

  Given sufficient sample size, it is also true that $D(\boldsymbol{\theta}, \mathbf{y})/a(\psi) \sim \chi^2_{n-p}$. It is also common to contrast this with the *null deviance*: the deviance function calculated for a model with no covariates (mean function only).

- Akaike Information Criterion.
  minimizes the negative likelihood penalized by the number of parameters:

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\theta}}|\mathbf{y}) + 2p \tag{4}$$

  where $\ell(\hat{\boldsymbol{\theta}}|\mathbf{y})$ is the maximized model log likelihood value and p is the number of explanatory variables in the model (including the constant). (AIC has a bias towards models that overfit with extra parameters since the penalty component is obviously linear with increases in the number of explanatory variables, and the log likelihood often increases more rapidly.)

- Schwartz Criterion/Bayesian Information Criterion (BIC).

$$\text{BIC} = -2\ell(\hat{\boldsymbol{\theta}}|\mathbf{y}) + p\log(n) \tag{5}$$

  where n is the sample size.

- Graphical Techniques

Negative Binomial GLM, Congressional Activity: 1995

- Negative binomial distribution has the same sample space (i.e. on the counting measure) as the Poisson, but contains an additional parameter which can be thought of as gamma distributed and therefore used to model a variance function.
- Used by many to fit a count model with overdispersion.
- compare the number of bills assigned to committee in the first 100 days of the 103^{rd} and 104^{th} Houses as a function of the number of members on the committee, the number of subcommittees, the number of staff assigned to the committee, and a dummy variable indicating whether or not it is a high prestige committee.
- Model is developed with the link function: $\theta = log(1 - \mu)$.

Table 8: Bills Assigned to Committed, First 100 Days

Committee	Size	Subcommittees	Staff	Prestige	Bills–103rd	Bills–104th
Appropriations	58	13	109	1	9	6
Budget	42	0	39	1	101	23
Rules	13	2	25	1	54	44
Ways and Means	39	5	23	1	542	355
Banking	51	5	61	0	101	125
Economic/Educ. Opportunities	43	5	69	0	158	131
Commerce	49	4	79	0	196	271
International Relations	44	3	68	0	40	63
Government Reform	51	7	99	0	72	149
Judiciary	35	5	56	0	168	253
Agriculture	49	5	46	0	60	81
National Security	55	7	48	0	75	89
Resources	44	5	58	0	98	142
Transport./Infrastructure	61	6	74	0	69	155
Science	50	4	58	0	25	27
Small Business	43	4	29	0	9	8
Veterans Affairs	33	3	36	0	41	28
House Oversight	12	0	24	0	233	68
Standards of Conduct	10	0	9	0	0	1
Intelligence	16	2	24	0	2	4

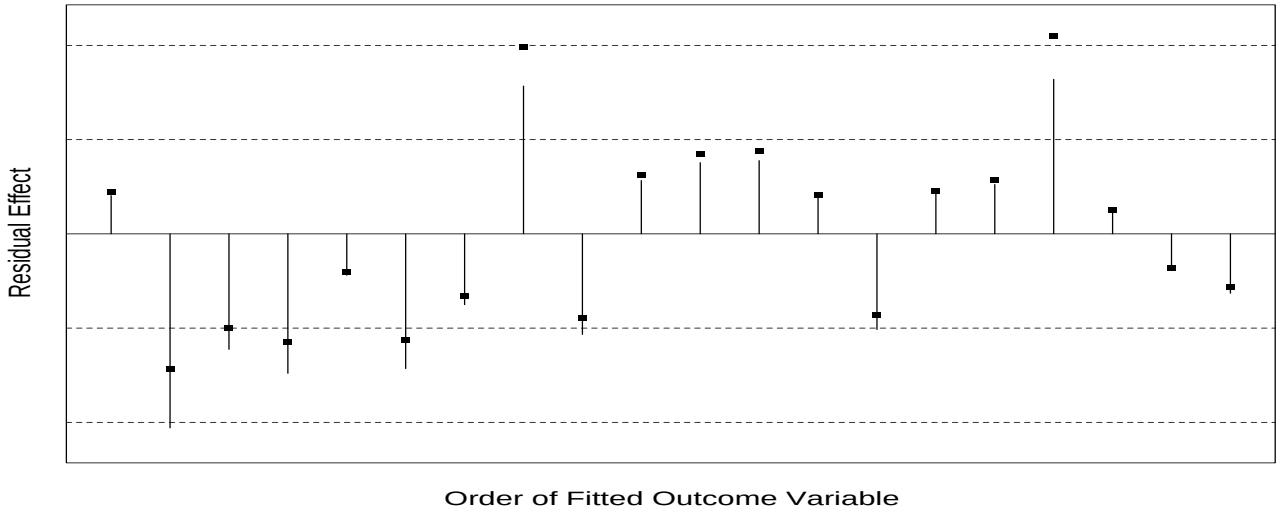
```
committee.dat <- read.table("/export/home/jgill/Book.GLM/Example.Committee/committee.dat",
      header=T,col.names=1)
attach(committee.dat)
committee.out <- glm.nb(BILLS104 ~ SIZE + SUBS ttee.out <- glm.nb(BILLS104 ~ SIZE +
      SUBS * (log(STAFF)) + PRESTIGE + BILLS103)

resp <- resid(committee.out,type="response")
pears <- resid(committee.out,type="pearson")
working <- resid(committee.out,type="working")
devs <- resid(committee.out,type="deviance")
```

Table 9: Modeling Bill Assignment – 104th House, First 100 Days

	Coefficient	Standard Error	95% Confidence Interval
(Intercept)	-6.80543	2.54651	[-12.30683:-1.30402]
Size	-0.02825	0.02093	[-0.07345: 0.01696]
Subcommittees	1.30159	0.54370	[0.12701: 2.47619]
log(Staff)	3.00971	0.79450	[1.29329: 4.72613]
Prestige	-0.32367	0.44102	[-1.27644: 0.62911]
Bills in $103^{\mathbf{rd}}$	0.00656	0.00139	[0.00355: 0.00957]
Subcommittees:log(STAFF)	-0.32364	0.12489	[-0.59345:-0.05384]
Null deviance: 107.314, $df = 19$			Maximized $\ell()$: 10559
Summed deviance: 20.948, $df = 13$			AIC: 121130

Figure 2: Residual Diagnostics: Bill Assignment Model



(points=pearson, lines=deviance)

Two-Stage GLM, the World Copper Market: 1951–1975

- common managerial economic problem: the estimation of a model of supply and demand functions for a certain good given data.
- central problem, endogeneity: price affects demand and demand affects price.
- classic solution: implement a two-stage process in which the endogenous variable for price is regressed onto some exogenous variables to create a predicted price vector, then this predicted price vector is used as one of a set of explanatory variables to regress quantity.
- the model is fully identified if the first stage of the model has one or more explanatory variables not included in the second stage.
- if the regression technique used in this process is the standard linear model, then this is called two-stage least squares (2SLS).
- consider a model for the world demand for copper over the years 1951–1975.
- 2SLS model using: world consumption of copper in 1,000 metric tons (**QTY**), the constant dollar adjusted price of copper (**PRI**), and aluminum (**ALM**, which is a substitute in many industrial settings),

an index of real per capita income base 1970 (INC), and an annual a measure of manufacturer inventory change (INV). As an attempt to capture technological improvements in manufacturing over this period, the authors use a simple integer time index $1 - 25$ (TME) over the years.

- 2SLS:

$$\textbf{Stage 1: } \mathit{Predicted}(\mathbf{PRI}) = \mathbf{1}\beta_{10} + \mathbf{INC}\beta_{11} + \mathbf{ALM}\beta_{12} + \mathbf{INV}\beta_{13} + \mathbf{TME}\beta_{14}$$
$$\textbf{Stage 2: } E[\mathbf{QTY}] = \mathbf{1}\beta_{20} + \mathit{Predicted}(\mathbf{PRI})\beta_{21} + \mathbf{INC}\beta_{22} + \mathbf{ALM}\beta_{23}.$$

- issue: there is evidence that technological improvement is not a linear change over these years, and in particular that most innovations occurred early in the time period. Using an integer scale as the 2SLS model has done, imposes a strict linearity condition here.
- A histogram of the outcome variable indicates a strongly right-skewed distribution, suggesting that the linear model might not be the best choice. In addition, there is a slight downturn for the last production value, indicating a discontinuation of the linear trend.

Table 10: The World Copper Market: 1951–1975

Year	World Copper Consumption	Copper Price	Aluminum Price	Income Index	Inventory Change
1951	3173.00	26.56	19.76	0.70	0.97679
1952	3281.10	27.31	20.78	0.71	1.03937
1953	3135.70	32.95	22.55	0.72	1.05153
1954	3359.10	33.90	23.06	0.70	0.97312
1955	3755.10	42.70	24.93	0.74	1.02349
1956	3875.90	46.11	26.50	0.74	1.04135
1957	3905.70	31.70	27.24	0.74	0.97686
1958	3957.60	27.23	26.21	0.72	0.98069
1959	4279.10	32.89	26.09	0.75	1.02888
1960	4627.90	33.78	27.40	0.77	1.03392
1961	4910.20	31.66	26.94	0.76	0.97922
1962	4908.40	32.28	25.18	0.79	0.99679
1963	5327.90	32.38	23.94	0.83	0.96630
1964	5878.40	33.75	25.07	0.85	1.02915
1965	6075.20	36.25	25.37	0.89	1.07950
1966	6312.70	36.24	24.55	0.93	1.05073
1967	6056.80	38.23	24.98	0.95	1.02788
1968	6375.90	40.83	24.96	0.99	1.02799
1969	6974.30	44.62	25.52	1.00	0.99151
1970	7101.60	52.27	26.01	1.00	1.00191
1971	7071.70	45.16	25.46	1.02	0.95644
1972	7754.80	42.50	22.17	1.07	0.96947
1973	8480.30	43.70	18.56	1.12	0.98220
1974	8105.20	47.88	21.32	1.10	1.00793
1975	7157.20	36.33	22.75	1.07	0.93810
YEAR	**QTY**	**PRI**	**INC**	**ALM**	**INV**

- Instead of the two-stage least squares linear model, a two-stage gamma GLM with $\boldsymbol{\theta} = -\frac{1}{\boldsymbol{\mu}}$ is built with the following specification:

$$\textbf{Stage 1: } Predicted(\mathbf{PRI}) = g^{-1}\big[\mathbf{1}\beta_{10} + \mathbf{INC}\beta_{11} + \mathbf{ALM}\beta_{12} + \mathbf{INV}\beta_{13} + log(\mathbf{TME})\beta_{14}\big]$$

$$\textbf{Stage 2: } E[\mathbf{QTY}] = g^{-1}\big[\mathbf{1}\beta_{20} + Predicted(\mathbf{PRI})\beta_{21} + \mathbf{INC}\beta_{22} + \mathbf{ALM}\beta_{23}\big]$$

where the $g^{-1}(\boldsymbol{X\beta})$ is the gamma link function.

```
copper.data <- as.matrix(read.table("/export/home/jgill/Book.GLM/Example.Copper/copper.dat",
          header=T,row.names=1))
copper.factors <- data.frame(copper.data)
attach(copper.factors)

copper.stage1.linear <- glm(COPPERPRICE ~ INCOMEINDEX + ALUMPRICE + INVENTORYINDEX + TIME,
     family=gaussian)
copper.stage2.linear <- glm(WORLDCONSUMPTION ~
     copper.stage1.linear$fitted.values + INCOMEINDEX + ALUMPRICE,
     family=gaussian)

copper.stage1 <- glm(COPPERPRICE ~ INCOMEINDEX + ALUMPRICE + INVENTORYINDEX + log(TIME),
     family=Gamma)
copper.stage2 <- glm(WORLDCONSUMPTION ~
     copper.stage1$fitted.values + INCOMEINDEX + ALUMPRICE,
     family=Gamma)
```

Table 11: Modeling the World Copper Market: 1951–1975

	Coefficient	Standard Error	95% Confidence Interval
(Intercept)	0.00080558	0.00006566	[0.00066904: 0.00094212]
Predicted(PRI)	0.00000449	0.00000162	[0.00000111: 0.00000786]
INC	-0.00058689	0.00006905	[-0.00073049:-0.00044329]
ALM	-0.00001082	0.00000234	[-0.00001568:-0.00000596]
Null deviance: 2.36735, $df = 24$			Maximized $\ell()$: -185.755
Summed deviance: 0.14290, $df = 21$			AIC: 379.51

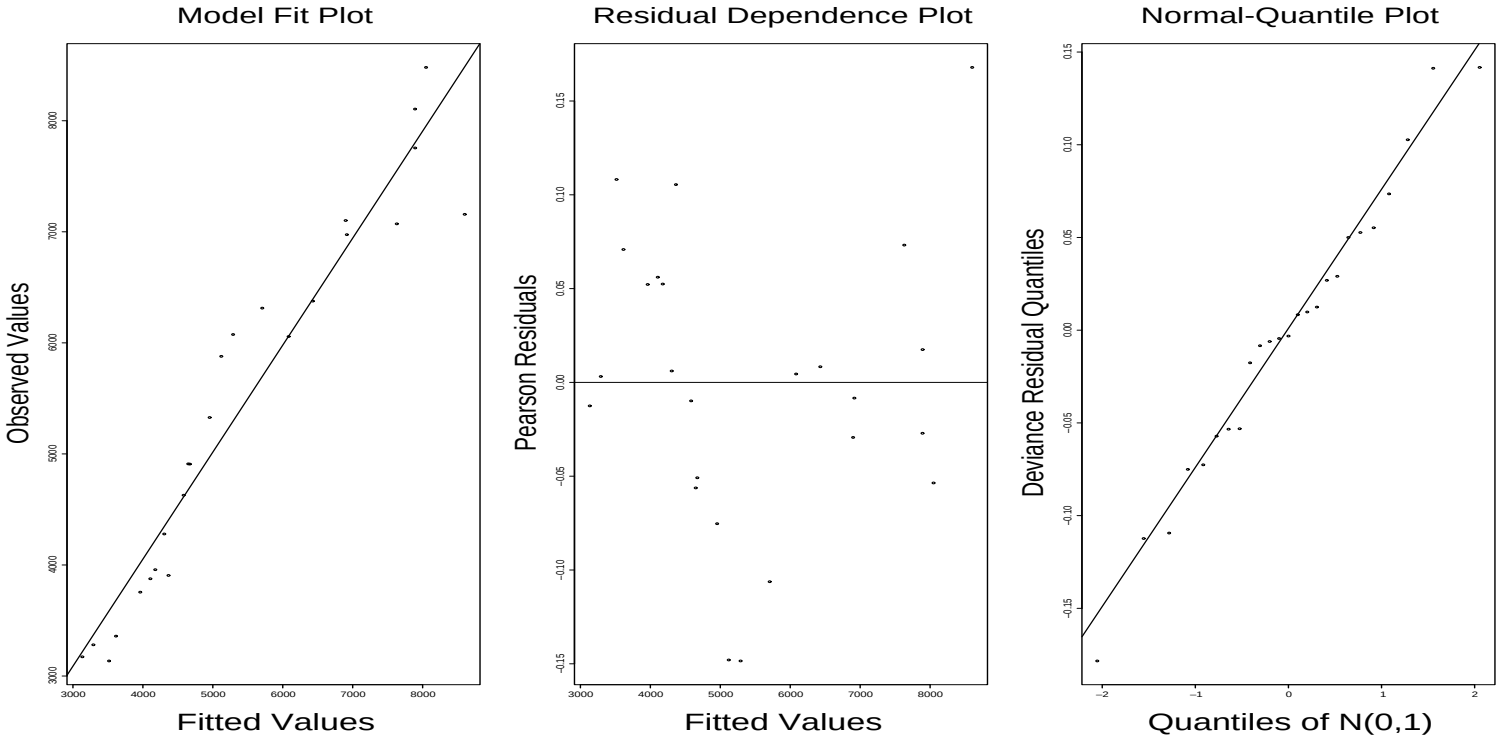
- Initially the sign on the coefficient for price is surprising since a positive value implies that higher prices are associated with greater demand which contradicts basic theory for a normal good (the 2SLS model had a negative sign). However, once we recall that the link function is necessarily acting on the linear predictor, this makes sense.

- The first difference for price using its first and third quartile (thus bracketing the interquartile range), keeping the other two variables constant at their mean:

$$\begin{aligned} E[\mathbf{QTY}_{Q1}] &= 5566.772 \\ E[\mathbf{QTY}_{Q3}] &= 4527.485 \\ \text{first difference: } &- 1039.287 \end{aligned} \tag{6}$$

  So as price moves from the 25^{th} percentile to the 75^{th} percentile, the expected drop in world for demand is a little over one million (1,039,287) metric tons.

Figure 3: Diagnostics: World Copper Market Model



Running GLMs in R

Running generalized linear models in R is remarkably simple and is setup to resemble the process for linear models to the greatest extent possible. Instead of using the `lm` function, the user calls the `glm` function. The basic syntax of the command is:

- CALL: `glm(formula, family = gaussian, data,`
  `weights = NULL, subset = NULL, na.action,`
  `model = TRUE, X = FALSE, y = TRUE,`
  `control = glm.control(epsilon=0.0001, maxit=10,`
  `trace=FALSE), contrasts = NULL)`
- FORMULA: a symbolic description of the model to be fit. For instance, the specification `Y ~ X1 + X2 + X1:X2` states that the outcome variable `Y` is modelled by two explanatory variables, `X1` and `X2`, and their interaction effect. For details about specifying formulas, see Chambers and Hastie (1993), Section 2.3.
- FAMILY: a description of the GLM link function and subsequent error distribution in the model fit. The `glm` function admits the following link function specifications:

Binomial	`binomial(link=‘‘logit’’)`
Normal	`gaussian(link=‘‘identity’’)`
Gamma	`Gamma(link=‘‘inverse’’)`
Inverse Gamma	`inverse.gaussian(link =‘‘1/mu2̂’’)`
Poisson	`poisson(link =‘‘log’’)`
Negative Binomial†	`negative.binomial(a=1,link=‘‘log’’)`
Quasi-Likelihood	`quasi(link=‘‘identity",variance=‘‘constant’’)`
Quasi-Likelihood/Binomial	`quasibinomial(link=‘‘logit’’)`
Quasi-Likelihood/Poisson	`quasipoisson(link=‘‘log’’)`

†Requires the Venables and Ripley `MASS` library extension.

- DATA: if the variables are to be taken from a different environment from which the `glm` call is made, then a data frame can be specified with this parameter.
- WEIGHTS: an optional vector of user-specified regression weights to be used in the fitting process.
- SUBSET: an optional vector specifying a subset of observations to be used in the fitting process. Use the `S` language subset rules for conditioning.
- NA.ACTION: a function that tells `glm` how to handle “NA”s. The default is the environmentally set “options$na.action”, which can be changed by the user or overridden in the function call. Common actions include: “na.fail” and “na.omit.”
- CONTROL: adjustment of numerical parameters that are used by the IWLS algorithm when mode-finding. `glm.control$epsilon` is the convergence threshold value for zero,

`glm.control$maxit` is the maximum number of IWLS iterations, and `glm.control$trace` is a logical value that turns on or off printed iteration information (i.e. like the `Gauss` software "max-like" procedure does as a default).

- MODEL: a logical value indicating whether the designated model frame should be included as part of the returned object.
- X,Y: logical values indicating whether the output variable vector and explanatory variable matrix used should be returned as components of the returned object.
- CONTRASTS: an optional list of factor contrasts. See Chambers and Hastie (1993), Section 2.3, and Venables and Ripley (1999), Section 6.2.

```
> options()$contrasts
   unordered     ordered
"contr.treatment"   "contr.poly"

> N <- factor(Nlevs <- c(1,4))
> N
[1] 1 4
Levels: 1 4
```

```
> contr.sum(N)
 [,1]
1  1
2 -1

> contr.treatment(N)
 2
1 0
2 1

> contr.helmert(N)
 [,1]
1 -1
2  1
```

```
> contr.poly(N)
      .L
[1,] -0.7071068
[2,]  0.7071068

> N <- factor(Nlevs <- c(1,4,8))

> contr.sum(N)
 [,1] [,2]
1   1   0
2   0   1
3  -1  -1

> contr.treatment(N)
 2 3
1 0 0
2 1 0
3 0 1
```

```
> contr.helmert(N)
 [,1] [,2]
1  -1  -1
2   1  -1
3   0   2

> contr.poly(N)
               .L         .Q
[1,] -7.071068e-01  0.4082483
[2,] -7.850462e-17 -0.8164966
[3,]  7.071068e-01  0.4082483
```

Multiple Imputation Using MICE

```
library(nnet)
library(mice)

anes.2000.mat <- read.table("http://www.clas.ufl.edu/~jgill/Turnout/nes2000.dat")
imp.2000.anes <- mice(anes.2000.mat,m=5)

anes.2000.imp.mat.1 <- complete(imp.2000.anes,1)
anes.2000.imp.mat.2 <- complete(imp.2000.anes,2)
anes.2000.imp.mat.3 <- complete(imp.2000.anes,3)
anes.2000.imp.mat.4 <- complete(imp.2000.anes,4)
anes.2000.imp.mat.5 <- complete(imp.2000.anes,5)

# 5 MODELS RUN HERE.
```

```
coef.mat <- cbind(out.mat2[,1],out.mat2[,1],out.mat3[,1],out.mat4[,1],out.mat5[,1])
var.mat <- cbind(out.mat2[,2],out.mat2[,2],out.mat3[,2],out.mat4[,2],out.mat5[,2])^2
impute.coef.vec <- apply(coef.mat,1,mean)
between.var <- apply(coef.mat,1,var)
within.var <- apply(var.mat,1,mean)
m <- 5
impute.se.vec <- sqrt(within.var + ((m+1)/m)*between.var)
```