Logistic Regression

$$Odds = \frac{\pi}{1-\pi}$$

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_i x_{ji}$$

Estimate α and β maximum likelihood

$$\eta_i = \alpha + \beta x_i$$

$$\frac{\hat{\pi}_i}{1-\hat{\pi}_i} = \exp(\eta_i)$$

$$\hat{\pi}_i = \frac{\exp(\eta_i)}{1-\exp(\eta_i)} = (1+\exp(-\eta_i))^{-1}$$

Generalized Linear Models

$$g(\mu_i) = \eta_i = \boldsymbol{\beta}^\top \boldsymbol{x_i}$$

- *Linear component*: $\eta_i = \boldsymbol{\beta}^\top \boldsymbol{x_i}$

- **Link function:** let univariate function g be monotonic (strictly decreases or increases) and differentiable

- *Random component*: y_i are independent and from an exponential family, which impose the variance of y_i depends on μ_i through a variance function $var(y_i) = \phi V(\mu_i)$ where ϕ is called the *dispersion parameter.*

Classic linear regression assumes normal dist.

Logistic regression assumes $g(\mu) = \log[\frac{\mu}{(1-\mu)}]$ and y_i is a Bernoulli trial($\mathbb{V}(y) = \mu(1-\mu)$). Other possible links for binary responses include

- Probit:

$g(\mu) = \Phi^{-1}(\mu)$, where Φ is the cumulative standard normal distribution

- Complementary log-log:

$$g(\mu) = \log(-\log(1-\mu))$$

Maximum Likelihood Estimation of Proportions

Suppose we draw a random sample of size $n = 5$ from some population, send them an offer, and $x = 2$ respond. What is our best guess of the response probability?

$$L(\pi) = \binom{5}{2}\pi^2(1-\pi)^3$$

- Let $l(\pi) = \log(L(\pi))$, *log-likelihood*

$$l(\pi) = \log(10) + 2\log(\pi) + 3\log(1-\pi)$$

$$\frac{dl(\pi)}{d\pi} = \frac{2}{\pi} - \frac{3}{1-\pi} = 0 \Rightarrow \hat{\pi} = \frac{2}{5}$$

Maximum Likelihood Estimation of Means

Suppose we draw a random sample of size $n = 3$ from a normal population with unknown mean μ and known variance $\sigma^2 = 5$. The observed values are $x_1 = 4, x_2 = 5, x_3 = 6$. What is the best guess of μ?

Pick μ so that the probability of observing the three values given μ is maximized:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$$

$$L(\mu) = \Pi_{i=1}^3(\frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{1}{2\sigma^2}(x_i-\mu)^2]$$

$$l(\mu) = \Sigma_{i=1}^3[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_i-\mu)^2]$$

$$= -k_1 - k_2\Sigma_{i=1}^3(x_i-\mu)^2$$

$$\frac{dl(\mu)}{d\mu} = 2k_2\Sigma_{i=1}^3(x_i-\mu) = 0 \Rightarrow \mu = \frac{1}{3}\Sigma_{i=1}^3 x_i$$

Note: $\frac{d^2 l(\mu)}{d\mu^2} = -6k_2 < 0$ thus maximum

Likelihood Function for Logistic Regression

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i$$

Probability distribution: $f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

$\Rightarrow f(y_1, \ldots, y_n) = \Pi_{i=1}^n \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

$\log(f) = l = \log(L) =$ Log-Likelihood

$$\Sigma_{i=1}^n y_i(\alpha + \beta x_i) - \Sigma_{i=1}^n \log[1+\exp(\alpha+\beta x_i)]$$

Maximize this with respect to α and β

Residual versus null deviance

Residual deviance: deviance for full model

Null deviance: deviance for the intercept-only model (think of as SST)

Difference between them measures how much variation is explained by the model and plays the role of extra sums of squares. Can test overall significance of the model b/c it has a **Chi-squared distribution**.

```
> fit$null.deviance - fit$deviance
[1] 1095.99
```

Likelihood-Ratio Test

We can test $H_0: \beta_{p+1} = \cdots = \beta_{p+q} = 0$ with the test statistic $D_1 - D_2$, which has a chi square distribution with q degrees of freedom (the D's represent deviance)

Predictive Accuracy of Classifiers

- GLMs minimize deviance

- $AIC = deviance + 2p$ common penalized measure

- **Classification rate**: $\frac{TN+TP}{TN+FP+FN+TP}$

- **Recall:** $\frac{TP}{FN+TP}$

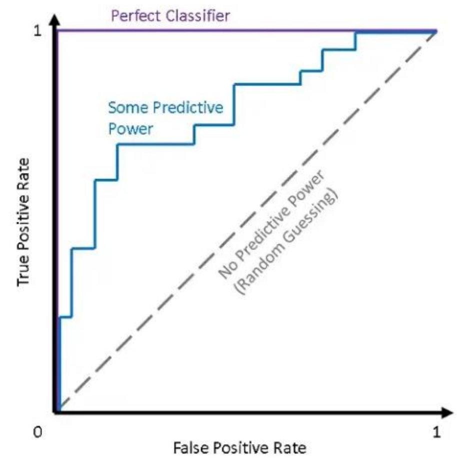
- **TPR (True Positive Rate) (Sensitivity):** $\frac{TP}{TP+FN}$

- **FPR (False Positive Rate) (1-Specificity):** $\frac{FP}{TN+FP}$

- **Precision:** $\frac{TP}{FP+TP}$

- **F1:** $\frac{2}{\frac{1}{Recall}+\frac{1}{Precision}} = 2\frac{P*R}{P+R}$

- **AUC** (Area under curve)**:** ROC curve is produced by calculating and plotting the true positive rate against the false positive rate for a single classifier at a variety of thresholds



Scoring Model:

- ***Proxy behavior***: behavior that has been observed in the past that is similar to future behavior you would like to predict, e.g., response to similar offer sent yesterday

Warning: your model may not work because this is only a proxy behavior. Seasonality, the state of the economy, what competition is doing, etc. usually all affect response

- ***Target period***: time period when proxy offer was active

- ***Base period***: a period of time prior to the target period. Information from the base period will be used to predict proxy behavior

Performance usually assessed with a *gains table*

1. Find quantiles of predicted values $\hat{y}$

2. Compute number of responders and revenue by quantile, also averages

3. Compute cumulative counts and revenues by quantile, also averages and lifts

A	B	C	D	E	F	G	H	I	J	K	L	M
Quantile of $\hat{y}$		Amount by Quantile				Cumulative					Lift	
%	n	Num Resp	Rev Amt	Resp Rate	Avg Amt	Num Cont	Num Resp	Rev Amt	Resp Rate	Avg Amt	Resp Rate	Rev Amt
1	10569	1681	159501	0.159	15.1	10569	1681	159501	0.159	15.1	2.98	3.18
2	10569	477	40241	0.0451	3.81	21138	2158	199742	0.102	9.45	1.91	1.99
3	10568	307	23716	0.0290	2.24	31706	2465	223458	0.0777	7.05	1.45	1.49
4	10569	201	15484	0.0190	1.46	42275	2666	238942	0.0631	5.65	1.18	1.19
5	10569	159	11608	0.0150	1.10	52844	2825	250549	0.0535	4.74	1	1

- **Columns A and B**: Quantiles of $\hat{y}$

- **Columns C and D**: number of responders and total revenue by quantile

- **Columns E and F**: response rate and average revenue in quantile, e.g., $\frac{1681}{10569} = 15.9\%$ and $\frac{96728}{10569} = \$15.1$ per contact

- **Column G**: *Depth* of contacts or cumulative counts, e.g., $10{,}569 + 10{,}569 = 21{,}138$

- **Columns H and I**: cumulative responders and revenue by quantile:

Row 2: $1{,}681 + 444 = 2{,}158$ responders and $159{,}501 + 40{,}241 = 199{,}742$ revenue at 40%

Last row: 2,825 total responders and 250,549 total revenue

- **Columns J and K**: cumulative response rate and average revenue in quantile:

Row 2: $\frac{2{,}158}{21{,}138} = 9.45\%$ and $\frac{199{,}742}{21{,}138} = \$9.45$ per contact at 40%

Last row: **contacting at random** gives 5.35% respond rate, \$4.74 per contact

- **Columns L and M**: lift of model over random guessing, e.g. $\frac{15.1\%}{5.35\%} = 2.98$ indicates the response rate from using model to pick best 20% of the names is improved by 57% over picking names at random. Revenue more than tripled (3.93)!

<u>Generalized Logistic Regression Model</u>

Let $Y \; in \; \{1,2,\dots,K\}$ be a multinomial r.v. for the outcome with K values. The probability that observation i comes from class k is $\pi_{ik} = P(Y_i = k)$, where $\pi_{i1} + \cdots + \pi_{iK} = 1$

Approaches:

- *One vs all*: fit K separate logistic regression models

- *Generalized logit*: fit $K-1$ models. Pick class 1 as the base category, but, as with the binary logit, this choice is arbitrary (models equivalent with a different base category)

Generalized logit model math:

- Model ($k = 2,\dots,K$):

$$\log\left(\frac{\pi_k}{\pi_1}\right) = \boldsymbol{\beta}_k^\top \mathbf{x} \qquad (k = 2,\dots,K)$$

- Solving for π_k:

$$\pi_k = \pi_1 \exp(\boldsymbol{\beta}_k^\top \mathbf{x}) \qquad (k = 2,\dots,K)$$

- The probabilities must sum to 1:

$$\begin{aligned} 1 = \Sigma_{j=1}^{K} \pi_j &= \pi_1 + \Sigma_{(j=2)}^{K} \pi_1 \exp\left(\boldsymbol{\beta}_j^\top \mathbf{x}\right) \\ &= \pi_1 [1 + \Sigma_{j=2}^{K} \exp(\boldsymbol{\beta}_j^\top \mathbf{x})] \end{aligned}$$

- Solving for π_1:

$$\pi_1 = \frac{1}{1 + \Sigma_{j=2}^{K} \exp(\boldsymbol{\beta}_j^\top \mathbf{x})}$$

- Substituting back into formula for π_k:

$$\pi_k = \frac{\exp(\boldsymbol{\beta}_k^\top \mathbf{x})}{1 + \Sigma_{j=2}^{K} \exp(\boldsymbol{\beta}_j^\top \mathbf{x})} \qquad (k = 2,\dots,K)$$

Generalized logit estimation:

- Likelihood function:

$$L(\mathbf{B}) = \Pi_{i=1}^{n} \Pi_{k=1}^{K} \pi_{ik}^{z_{ik}}$$

Where $z_{ik} = 1$ when $y_i = k$ and $z_{ik} = 0$ otherwise

- The log-likelihood is the cost function

$$\log\left(L(\mathbf{B})\right) = \mathrm{l}(\mathbf{B}) = \Sigma_{\mathrm{i}=1}^{\mathrm{n}} \Sigma_{k=1}^{\mathrm{K}} z_{ik} \log(\pi_{ik})$$

- Denote optimal values of $\boldsymbol{\beta}_k$ by $\boldsymbol{b}_k$. Compute estimated probabilities that each observation is in class k as:

$p_{i1} = \frac{1}{1+\Sigma_{j=2}^{K} \exp(\mathbf{b}_j^\top \mathbf{x}_i)}$ and $p_{ik} = \exp(\mathbf{b}_k \mathbf{x}^\top)\, p_{i1}$

- The *maximum likelihood classifier* assigns i to the class with the largest p_{ik}

Base category is $y = 1$

$$\log\left(\frac{\pi_2}{\pi_1}\right) = -8.69 + 1.42x_1 + 0.99x_2$$

$$\log\left(\frac{\pi_3}{\pi_1}\right) = -7.64 + 0.66x_1 + 1.52x_2$$

All four models give similar results

- There are three possible base categories and all give equivalent models

```
Call: multinom(formula = y ~ x1 + x2, data = train)

Coefficients:
  (Intercept)        x1        x2
2   -8.693704 1.4200155 0.9893357
3   -7.644341 0.6568461 1.5195375
Residual Deviance: 68.61044

multinom(formula = factor(y, levels = c(2, 1, 3)) ~ x1 + x2, data = train)
Coefficients:
  (Intercept)         x1         x2
1    8.691775 -1.4197191 -0.9891887
3    1.047763 -0.7629366  0.5302861
Residual Deviance: 68.61043

multinom(formula = factor(y, levels = 3:1) ~ x1 + x2, data = train)
Coefficients:
  (Intercept)         x1         x2
2   -1.047771  0.7629246 -0.5302758
1    7.643941 -0.6567638 -1.5194762
Residual Deviance: 68.61043
```

- All three models have the same deviance (cost function value)

- Let $\boldsymbol{\beta}_{ij}$ be slopes for $\log\left(\frac{\pi_i}{\pi_j}\right) = \boldsymbol{\beta}_{ij}^\top \mathbf{x}$. Note that $\frac{\pi_i}{\pi_j} = e^{\boldsymbol{\beta}_{ij}^\top \mathbf{x}}$

- Clearly, $\boldsymbol{\beta}_{ij} = -\boldsymbol{\beta}_{ij}$, e.g.,

$$\log\left(\frac{\pi_3}{\pi_1}\right) = \beta_{31}^\top \mathbf{x} \Rightarrow \frac{\pi_3}{\pi_1} = e^{(\beta_{31}^\top \mathbf{x})} \Rightarrow \frac{\pi_1}{\pi_3} = e^{-\beta_{31}^\top \mathbf{x}}$$

- $\boldsymbol{\beta}_{23} = \boldsymbol{\beta}_{21} - \boldsymbol{\beta}_{31}$, e.g.,

$$0.7629246 = 1.4200155 \; - \; 0.6568461$$

$$\begin{aligned} \frac{\pi_2}{\pi_3} = \frac{\pi_2}{\pi_3} * \frac{\pi_1}{\pi_1} = \frac{\pi_2}{\pi_1} * \frac{\pi_1}{\pi_3} &= e^{\beta_{21}^\top \mathbf{x}} * e^{-\beta_{31}^\top \mathbf{x}} \\ &= e^{(\beta_{21} - \beta_{31})^\top \mathbf{x}} \end{aligned}$$

<u>Making predictions</u>

- **Binary logit**: individual chooses between two options and selects the one that provides greater utility

- **Multinomial logit**: individual chooses among more than two alternatives and selects the one that provides the greatest utility

- **Ordered logit**: individual reveals the strength of his or her preferences with respect to a single outcome

- **Conditional logit**: allows variables that <u>vary across alternatives</u> and possibly across the individuals as well, e.g., choice of mode of transportation (e.g., train, bus, car). Characteristics or attributes of these include time waiting, how long it takes to get to work, and cost.

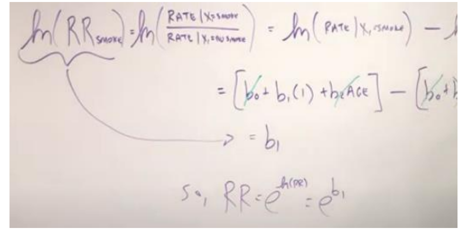
- **Log-linear analysis**: class of models that subsumes (includes or absorbs) the logit models and more.

<u>Poisson</u>

- The PMF of a Poisson distribution:

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, \mu > 0, y = 0,1,2,\dots$$

- Suppose we observe a random sample of ordered pairs $(x_i, y_i), i = 1,\dots,n$ from where y_i has a Poisson distribution with mean μ_i. More generally, x_i could be a vector of predictors. Assume (log link function) (log(mu) -> rate ratio)



$$\log(\mu_i) = \eta_i = \alpha + \beta x_i \Longleftrightarrow \mu_i = e^{\eta_i}$$

- The likelihood is:

$$L = \Pi_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}$$

- The log-likelihood of a Poisson model is:

$$l = \log(L) = \Sigma_{i=1}^{n} [y_i \log(\mu_i) - \mu_i - \log(y_i!)]$$

- The log-likelihood of a saturated Poisson model is:

$$\hat{\mu}_i \equiv y_i \Rightarrow l_s = \Sigma_{i=1}^{n} [y_i \log(y_i) - y_i - \log(y_i!)]$$

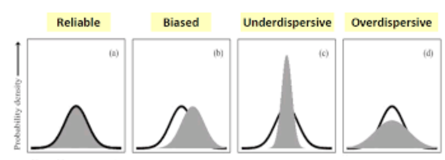
- Let $\mu_i \equiv \exp\left(\hat{\alpha} + \hat{\beta} x_i\right)$ be the MLEs. The deviance is:

$$\begin{aligned} D^2 = -2(l - l_s) = -2 \Bigg[&\Sigma_{i=1}^{n} y_i \log\left(\frac{\hat{\mu}_i}{y_i}\right) \\ &+ \Sigma_{i=1}^{n} (y_i - \hat{\mu}_i) \Bigg] \end{aligned}$$

Where $y_i \log\left(\frac{\hat{\mu}_i}{y_i}\right) = 0$ for $y_i = 0$

<u>Beyond Poisson</u>

- Recall that for Poisson Y, $\mathbb{E}(Y) = \mathbb{V}(Y) = \mu$, but in practice we may find $\mathbb{E}(Y) < \mathbb{V}(Y)$, called the problem of **overdispersion** (or **underdispersion** $\mathbb{E}(Y) > \mathbb{V}(Y)$)



- **Negative binomial distribution** (NBD) **regression** models allow for overdispersion

- Beyond overdispersion, we often observe too many zero values for either Poisson or NBD. What to do?

Zero-inflated Poisson (ZIP) assumes a mixture distribution

$$P(Y = 0) = \pi + (1 - \pi) e^{-\mu},$$

$$P(Y = y) = (1 - \pi) \frac{e^{-\mu} \mu^y}{y!}, \mu > 0, y = 1,2,\dots$$

Where $\pi \in [0,1]$ is the probability of extra (structural) zeros

Zero-inflated negative binomial (ZINP)

Hurdle models

- GLMs can accommodate other distributions including exponential and gamma

Gamma regression

- The PMF of a *gamma distribution* is:

$$f(y) = \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-\frac{y}{\beta}}, \alpha > 0, \beta > 0, y > 0$$

- Where α is the *shape* parameter and β is the *scale* parameter. The mean of a gamma distribution is $\mu = \alpha\beta$, so $\beta = \frac{\mu}{\alpha}$. The pdf can be written in terms of μ:

$$f(y) = \left(\frac{\alpha}{\mu}\right)^\alpha \Gamma(\alpha)^{-1} y^{\alpha-1} e^{-\frac{y\alpha}{\mu}}$$

- The log pdf is given by:

$$\log(f(y)) = \alpha\log(\alpha) - \alpha\log(\mu) - \log[\Gamma(\alpha)]$$
$$+(\alpha-1)\log(y) - \frac{y\alpha}{\mu} = \alpha\left(-\frac{y}{\mu} - \log(\mu)\right)$$
$$-\log(\Gamma(\alpha)) + \alpha\log(\alpha y) - \log(y)$$

- The log-likelihood is:

$$l = \Sigma_{i=1}^n \left[\alpha\left(-\frac{y_i}{\mu_i} - \log(\mu_i)\right) - \log[\Gamma(\alpha)] + \alpha\log(\alpha y_i) - \log(y_i)\right]$$

- The log-likelihood of the saturate model is:

$$l_s = \Sigma_{i=1}^n[\alpha(-1 - \log(y_i)) - \log[\Gamma(\alpha)] + \alpha\log(\alpha y_i) - log\,(y_i)]$$

- The deviance is $-2(l - l_s)$, but notice how the last three terms of l and l_s are identical and thus cancel out. Thus the deviance is:

$$-2(l - l_s) = -2\alpha\Sigma_{i-1}^n\left(-\frac{y_i}{\mu} - \log(\mu_i) + 1 + \log(y_i)\right)$$
$$= -2\alpha\Sigma_{i-1}^n\left(\log\left(-\frac{y_i}{\mu_i}\right) - \frac{y_i - \mu_i}{\mu_i}\right)$$

Regression Terms and Symbols:

Term	ACT	JWHT	Other
Sum of squared errors	**SSE**	**RSS**	
Total sum of squares	**SST**	**TSS**	
Mean squared error	**MSE**		S_e^2
Residual Standard Error		**RSE**	

SSE = $\sum_{i=1}^n [y_i - f(X_i)]^2$

MSE = $\frac{SSE}{n}$

$$\boldsymbol{R^2} = 1 - \frac{SSE}{SST} = 1 - \frac{RSS}{(n-1)var(y)}$$

Penalized Estimates

$\boldsymbol{R^2} = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}}$ or AIC = deviance + 2p

Multicollinearity

Pipe: e.g., $x_1 \to x_2 \to y$. If you are studying $x_1 \to y$ then do not control for x_2

Latent construct: predictors manifestations of common, underlying **latent construct**, e.g., $w \to x_1$ *and* $w \to x_2$. Often, estimate w and use it instead of x_1 and x_2

Back door confound (**fork**): include control to block back-door path, e.g., if $w \to y$ and $w \to x \to y$ then control for w to study $x \to y$

Collider: usually do not control for colliders, e.g., if $x \to w$ and $y \to w$, then do not control for collider w when studying $x \to y$

Extras

F = $\frac{\frac{SST-SSE}{p}}{\frac{SSE}{n-p-1}} = \frac{\left(\frac{\Delta SSE}{\Delta df}\right)}{S_e^2}$, bottom from full model

1. (9 points) The Poisson distribution has the following PMF:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad (\lambda > 0; \quad x = 0, 1, 2, \ldots).$$

Suppose that we have a random sample of $n = 2$ observations, x_1 and x_2, from a Poisson distribution with parameter λ, where the observations are independent. This problem will derive the maximum likelihood estimate of λ.

(a) (2 points) Find $L(\lambda) = P(X_1 = x_1 \cap X_2 = x_2)$ *Answer:*

$$L(\lambda) = \frac{\lambda^{x_1}e^{-\lambda}}{x_1!} \cdot \frac{\lambda^{x_2}e^{-\lambda}}{x_2!} = \frac{\lambda^{x_1+x_2}e^{-2\lambda}}{x_1!x_2!}$$

(b) (2 points) Find $l(\lambda) = \log[L(\lambda)]$. *Answer:* $l(\lambda) = (x_1+x_2)\log\lambda - 2\lambda - \log(x_1!x_2!)$

(c) (3 points) Find the value of λ that maximizes $l(\lambda)$. *Answer:*

$$\frac{dl(\lambda)}{d\lambda} = \frac{x_1 + x_2}{\lambda} - 2 = 0$$

so we find $\lambda = (x_1 + x_2)/2$, *the sample mean.*

(d) (2 points) How do you know that your value of λ is a maximum? *Answer: The second derivative is*

$$\frac{d^2l(\lambda)}{d\lambda^2} = -\frac{x_1 + x_2}{\lambda^2} < 0,$$

since both x_j *and* λ *are not negative.*

2. (30 points) Data from 37 patients receiving a non-depleted allogeneic bone marrow transplant were examined to see which variables were associated with the development of acute graft-versus-host disease (GvHD), which is a binary response variable. The predictor variables are the age of the recipient (`Rage`), the age of the donor (`Dage`), whether or not the donor had been pregnant (`preg`), an index of mixed epidermal cell-lymphocyte reactions (`index`) and the `type` of leukemia (there are 3 types, 1= acute myeloid leukemia, 2=acute lymphocytic leukemia and 3=chronic myeloid leukemia).

(a) (3 points) I first estimated a full logistic regression model, with the output below:

```
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -5.343720   2.624293  -2.036   0.0417
Rage               0.003882   0.084106   0.046   0.9632
Dage               0.112187   0.081436   1.378   0.1683


preg               1.705127   1.199851   1.421   0.1553
log(index)         1.835935   0.892520   2.057   0.0397
as.factor(type)2 -0.405541   1.256857  -0.323   0.7470
as.factor(type)3  1.676694   1.297599   1.292   0.1963

    Null deviance: 51.049  on 36  degrees of freedom
Residual deviance: 26.288  on 30  degrees of freedom
```

Test whether the overall model is significant at the 5% level. State the null and alternative and your decision. Hint: the 95th percentile of a chi-square distribution with 6 degrees of freedom is 12.59. *Answer:* $H_0 : \beta_1 = \cdots = \beta_6 = 0$ *versus* H_1 *: at least one* $\beta_j \neq 0$. *The test statistic is* $51.049 - 26.288 = 24.761 > 12.59$ *so we reject* H_0.

(b) (3 points) Test whether the type of leukemia has an effect, i.e., do we need the two leukemia dummies? A model without the two leukemia dummies has a residual deviance of 29.27. Hint: the 95% percentile of a chi-square distribution with 2 df is 5.992. *Answer:* $H_0 : \beta_5 = \beta_6 = 0$ *versus* H_1 *: at least one of* β_5 *and* β_6 *is nonzero. The test statistics is* $29.27 - 26.288 = 2.984 < 5.992$ *so we cannot reject* H_0.

(c) (2 points) I created three dummies for the three types of leukemia (they equal 1 for that type and 0 otherwise): `aml`, `all` and `cml`). I entered all variables into a backward selection logistic regression model giving the following model:

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.5464     0.9485  -2.685  0.00726
log(index)    1.4877     0.7197   2.067  0.03872
cml           2.2506     1.1060   2.035  0.04187
preg          2.4955     1.1012   2.266  0.02344

    Null deviance: 51.049  on 36  degrees of freedom
Residual deviance: 28.848  on 33  degrees of freedom
```

State the estimated regression equation. *Answer:*

$$\log\left(\frac{\pi}{1-\pi}\right) = -2.5465 + 1.4877\log(\texttt{index}) + 2.2506\texttt{cml} + 2.4955\texttt{preg}$$

(d) (3 points) Use this model to estimate the *probability* if GvHD for someone with an index of 1.10, acute lymphocytic leukemia, and whose donor had been pregnant. *Answer:* $\hat{\eta} = -2.5464 + 1.4877 * \log(1.1) + 2.4955 = 0.0909$ *and* $\hat{\pi} = e^{0.0909} = 0.52$.

(e) (2 points) Using the model from the previous part, how many *times* greater are the *odds* (not probability or log odds) of GvHD for someone with chronic myeloid leukemia compared with one of the other two types of leukemia? *Answer:* $e^{2.2506} =$ 9.49 *times more likely.*

3. (8 points) Complete the following table giving the Kaplan-Meier estimates of the survival function and retention rate.

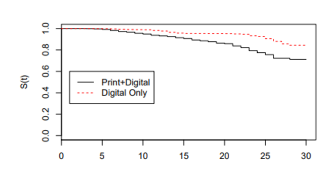
t	Number Cancels	Number Censored	Number at Risk	Retention Rate	Survival Function
1	0	5	1000	$1 - 0/1000 = 1$	**1**
2	5	10	**995**	$1 - 5/995 = .995$	**.995**
3	7	8	**980**	$1 - 7/980 = .993$	$.995(.993) = .988$

4. (6 points) Multiple logistic regression was used to construct a prognostic index to predict significant coronary artery disease from data on 348 patients with valvular heart disease who had undergone routine coronary arteriography before value replacement. Forward stepwise selection was used, with a significance level for entry into the model of 0.05. The prognostic index obtained was based on a model with seven variables.

(a) (2 points) The regression coefficient for a family history of ischaemic heart disease (coded 0=no and 1=yes) was 1.167. What is the estimated "odds ratio" for having significant coronary artery disease associated with a positive family history? Hint: e^β. *Answer: exp(1.167) = 3.21.*

(b) (4 points) One of the variables in the model was the estimated total number of cigarettes ever smoked, calculated as the average number smoked annually $\times$ the number of years smoking. The regression coefficient was 0.0106 per 1000 cigarettes. What total number of cigarettes ever smoked carries the same risk as a family history of ischaemic heart disease? Convert this figure into years of smoking 20 cigarettes per day. (Assume there are 365 days in a year.) *Answer: 1.167/(.0106*20*365/1000) = 15.08 years.*

5. The plot below is from a KM model of the newspaper data, stratifying by the publication type (print+digital versus digital only). What does the plot tell you?



Answer: Digital only survive longer than print+digital.

6. Expect a question interpreting a Poisson regression similar to the news desert homework 7, but not as complicated

7. Expect something interpreting a multinomial logit