Qualifying Exam: CAS MA 576.
Boston University, Spring 2007

1. Let Y_i, $i = 1, \ldots, n$ be a r.v. giving us the number of success in a set of m_i trials with probability of success π_i, i.e. $Y_i \sim Binomial(m_i, \pi_i)$. Moreover, let the success probability π_i depend on the the p covariates of the ith covariate class $\boldsymbol{x_i}$. We seek to model the the above relation in a GLM framework.

   (a) What is the most important reason for not using the identity link in the above framework.

   (b) Name and write the expressions for two link functions by which we can provide a realistic model for π_i as a function of the covariates $\boldsymbol{x_i}$.

   (c) Write the likelihood of this model (for arbitrary link functions) and show that the residual deviance can be written as

$$D(y, \hat{\pi}) = 2 \sum_i \left(y_i \log(y_i / m_i \hat{\pi}_i) + (m_i - y_i) \log \left(\frac{m_i - y_i}{m_i(1 - \hat{\pi}_i)} \right) \right).$$

   (d) Wedderburn(1974) collected data on the incidence of *R. secalis*, commonly known as leaf blotch, on the leaves of ten varieties of barley grown at nine sites. The records on the proportion for each variety and sites are given below.

	variety									
site	1	2	3	4	5	6	7	8	9	10
1	0.0005	0.0000	0.0000	0.0010	0.0025	0.0005	0.0050	0.0130	0.0150	0.0150
2	0.0000	0.0005	0.0005	0.0030	0.0075	0.0030	0.0300	0.0750	0.0100	0.1270
3	0.0125	0.0125	0.0250	0.1660	0.0250	0.0250	0.0000	0.2000	0.3750	0.2625
4	0.0250	0.0050	0.0001	0.0300	0.0250	0.0001	0.2500	0.5500	0.0500	0.4000
5	0.0550	0.0100	0.0600	0.0110	0.0250	0.0800	0.1658	0.2950	0.2000	0.4350
6	0.0100	0.0500	0.0500	0.0500	0.0500	0.0500	0.1000	0.0500	0.5000	0.7500
7	0.0500	0.0010	0.0500	0.0500	0.5000	0.1000	0.5000	0.2500	0.5000	0.7500
8	0.0500	0.1000	0.0500	0.0500	0.2500	0.7500	0.5000	0.7500	0.7500	0.7500
9	0.1750	0.2500	0.4250	0.5000	0.3750	0.9500	0.6250	0.9500	0.9500	0.9500

   (e) Note that instead of modeling Y_i's, the number of leaf blotch, we need to model the proportion of success $Z_i = Y_i/m_i$, as we don't have data on m_i, the number of trials. Can we still use the form of the binomial regression? Justify your answer.

   (f) A suggested solution is to use quasibinomial regression, with the logit link and variance $V(\mu_i) = \sigma^2 \mu_i(1 - \mu_i)$, the scaling parameter would take care of the fact that Z_i proportion rather than number of success. Using the standard variance function for a quasibinomial we get the following R output

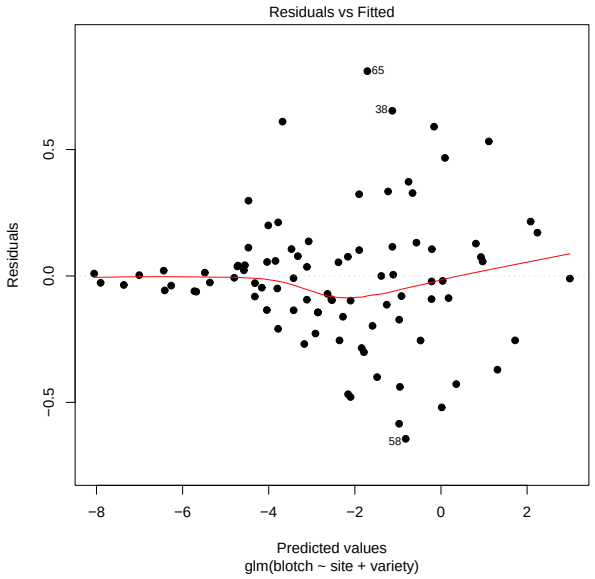
```
glm(formula = blotch ~ site + variety, family = quasibinomial,
    data = leafblotch)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.0546     1.4220  -5.664 2.84e-07 ***
site2         1.6391     1.4433   1.136 0.259880
site3         3.3265     1.3492   2.465 0.016068 *
site4         3.5822     1.3445   2.664 0.009512 **
site5         3.5838     1.3444   2.666 0.009479 **
site6         3.8932     1.3402   2.905 0.004876 **
site7         4.7299     1.3348   3.544 0.000698 ***
site8         5.5226     1.3346   4.138 9.39e-05 ***
```

```
site9          6.7945     1.3407    5.068  3.00e-06 ***
variety2       0.1501     0.7237    0.207  0.836293
variety3       0.6895     0.6724    1.025  0.308599
variety4       1.0481     0.6494    1.614  0.110919
variety5       1.6147     0.6257    2.581  0.011897 *
variety6       2.3711     0.6090    3.893  0.000219 ***
variety7       2.5712     0.6065    4.240  6.55e-05 ***
variety8       3.3419     0.6015    5.556  4.39e-07 ***
variety9       3.4999     0.6014    5.820  1.51e-07 ***
variety10      4.2529     0.6042    7.038  9.39e-10 ***
---

(Dispersion parameter for quasibinomial family taken to be 0.08878094)

    Null deviance: 40.8029  on 89  degrees of freedom
Residual deviance:  6.1264  on 72  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 8
```

(g) Is there a significant difference in the proportion of leaf blotch by site and by variety.

(h) Is there a natural ranking of the sites and varieties based on the proportion of leaf blotch? If so what are these rankings?

(i) Comment on the following residual vs fitted plot using the above analysis.



2. Let $Y_1, \ldots, Y_n$ be independent r.v. measuring proportions. Suppose we are interested in modeling the mean response (hear proportion) $\mu_i = E(Y_i)$ as a function of a single covariate x. We will use the link function

$$logit(\mu_i) = \beta_0 + \beta_1(x_i - \bar{x}) \text{ i.e. } \mu_i = \frac{\exp[\beta_0 + \beta_1(x_i - \bar{x})]}{1 + \exp[\beta_0 + \beta_1(x_i - \bar{x})]}$$

to model this relationship.

(a) First show that $\dfrac{\partial \mu_i}{\partial \beta_0}$ can be expressed as $\mu_i(1-\mu_i)$ and $\dfrac{\partial \mu_i}{\partial \beta_1}$ can be expressed as $\mu_i(1-\mu_i)(x_i - \bar{x})$.

(b) Now consider the variance function $V(\mu_i) = \sigma^2 \mu_i^2 (1-\mu_i)^2$. Recall the result

$$cov(\hat{\beta}) = \sigma^2 (D^T V^{-1} D)^{-1}, \text{ where } D = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \text{ and } V = \text{ Diagonal matrix of the variance relationship}$$

Using the above result or otherwise show that

$$var(\beta_0) = \sigma^2/n \text{ and } var(\beta_1) = \sigma^2 / \sum_i (x_i - \bar{x})^2$$

and β_0 and β_1 are uncorrelated.

(c) Would you get the uncorrelatedness if you started with the link

$$logit(\mu_i) = \beta_0 + \beta_1(x_i)?$$

Why? [Restrict your answer for this part to at most 3 sentences]

(d) Note that if σ^2 is known, the variance covariance of β_0 and β_1 are independent of $\boldsymbol{\mu}$ in (c). Is this true in general for any link and variance functions? If your answer is yes, prove it and if your answer is no give a counter example.

(e) We will now use this model to analyze the leaf blotch example from Wedderburn(1974). The data appears in question 1. Wedderburn proposed to use the variance function $V(\mu_i) = \sigma^2 \mu_i^2 (1-\mu_i)^2$, which does not give rise to any standard pmf. [Note that a quasi-binomial likelihood can be obtained using the standard variance for the logit link, $V(\mu_i) = \sigma^2 \mu_i (1-\mu_i)$ which we have used in Question 1]
Using Wedderburn's variance function we get the following R output

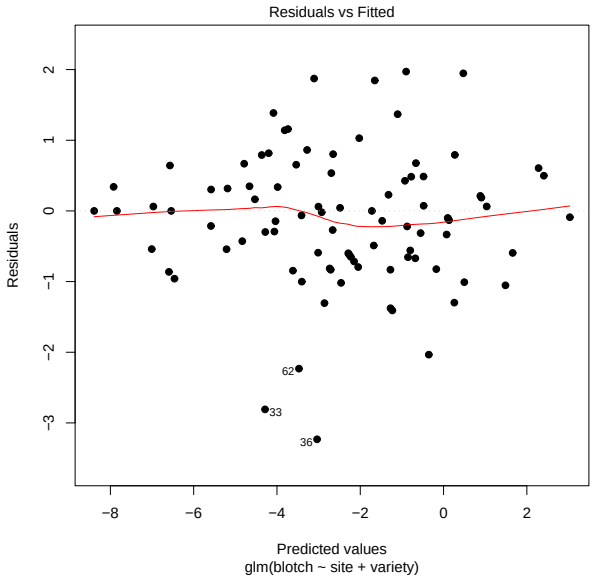
```
Call:
glm(formula = blotch ~ site + variety, family = quasi(link = "logit",
    variance = "mu^2(1-mu)^2"), data = leafblotch)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.92253    0.44463 -17.818  < 2e-16 ***
site2        1.38308    0.44463   3.111  0.00268 **
site3        3.86013    0.44463   8.682 8.18e-13 ***
site4        3.55697    0.44463   8.000 1.53e-11 ***
site5        4.10841    0.44463   9.240 7.48e-14 ***
site6        4.30541    0.44463   9.683 1.13e-14 ***
site7        4.91811    0.44463  11.061  < 2e-16 ***
site8        5.69492    0.44463  12.808  < 2e-16 ***
site9        7.06762    0.44463  15.896  < 2e-16 ***
variety2    -0.46728    0.46868  -0.997  0.32210
variety3     0.07877    0.46868   0.168  0.86699
variety4     0.95418    0.46868   2.036  0.04544 *
```

```
variety5      1.35276     0.46868    2.886   0.00514 **
variety6      1.32859     0.46868    2.835   0.00595 **
variety7      2.34066     0.46868    4.994 3.99e-06 ***
variety8      3.26268     0.46868    6.961 1.30e-09 ***
variety9      3.13556     0.46868    6.690 4.10e-09 ***
variety10     3.88736     0.46868    8.294 4.33e-12 ***
---

(Dispersion parameter for quasi family taken to be 0.9884758)

    Null deviance: 370.523  on 89  degrees of freedom
Residual deviance:  66.267  on 72  degrees of freedom
```

(f) Is there a significant difference according to site and and variety.

(g) Is there a natural ranking of the sites and varieties? If so what are these rankings?

(h) Comment on the shape of the residual plot.



(i) If you have answered Question 1, compare the two residual plots.