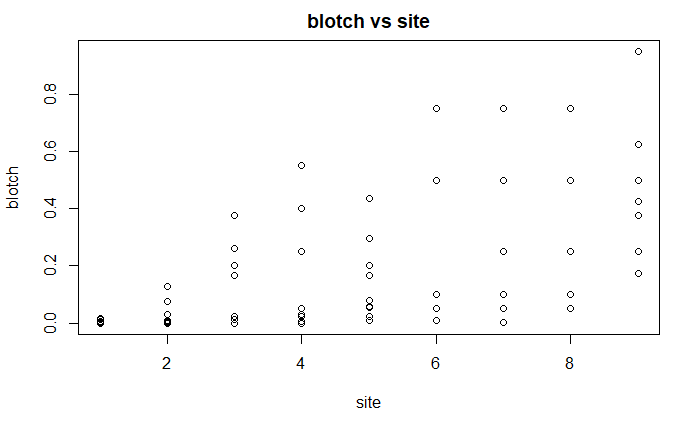
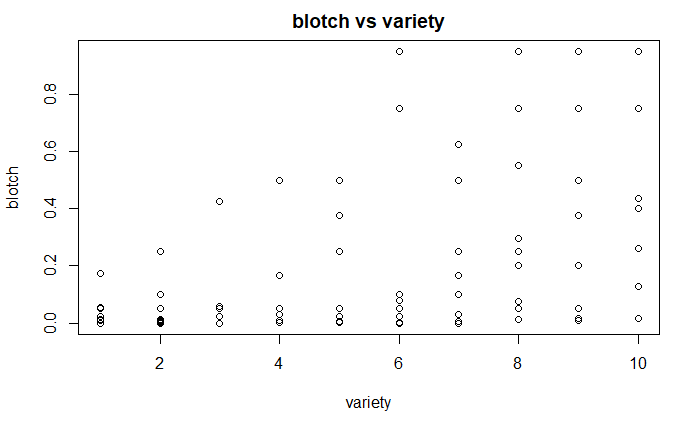
**STAT34700-HW4**

**Problem 1**

**(a)**





**Comments:**

From the plot of blotch vs site, we can see that there is variance for responses of each site. The variance of responses for different sites are not the same. As the site number increasing, the variance of response becomes greater.

From the plot of blotch vs variety, we can see that there is variance for responses of each variety. The variance of responses for different varieties are not the same. As the variety number increasing, the variance of response becomes greater.

**(b)**

We fit a binomial GLM with *blotch (the percentage leaf area affected by leaf blotch)* as the binomial response, *site* and *variety* as categorical predictors. In this generalized linear model, we use logit link function and binomial distribution family.

> fitbin = glm(blotch~site+variety, family=binomial, data=leafblotch)

**The GLM formula:**

**Output:**

> sumary(fitbin)

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.05460 4.77233 -1.6878 0.09146

site2 1.63906 4.84395 0.3384 0.73508

site3 3.32648 4.52822 0.7346 0.46258

site4 3.58219 4.51217 0.7939 0.42726

site5 3.58384 4.51208 0.7943 0.42703

site6 3.89320 4.49798 0.8655 0.38674

site7 4.72994 4.47976 1.0558 0.29104

site8 5.52259 4.47923 1.2329 0.21760

site9 6.79450 4.49961 1.5100 0.13104

variety10 4.25287 2.02789 2.0972 0.03598

variety2 0.15008 2.42879 0.0618 0.95073

variety3 0.68946 2.25656 0.3055 0.75996

variety4 1.04814 2.17961 0.4809 0.63060

variety5 1.61468 2.09983 0.7690 0.44192

variety6 2.37110 2.04396 1.1600 0.24603

variety7 2.57117 2.03542 1.2632 0.20651

variety8 3.34193 2.01886 1.6554 0.09785

variety9 3.49994 2.01822 1.7342 0.08289

n = 90 p = 18

Deviance = 6.12644 Null Deviance = 40.80291 (Difference = 34.67647)

The deviance is 6.12644.

**Test Good-of-fit:**

Provided that the response is binomial distributed, the deviance is approximately Chi-squared distributed with 72 degrees of freedom if the model is correct. Thus, we can use the deviance to test whether the model is an adequate fit.

**Output:**

> pchisq(deviance(fitbin),df.residual(fitbin),lower=FALSE)

[1] 1

**Comments:**

The null hypothesis is that the model has a good fit on the data. Because the p-value is greater than significance level 0.05, then we cannot reject the null hypothesis and conclude that the model is an adequate fit.

**(c)**

We fit a quasi-binomial model with the same response and predictors as above.

> fitqbin = glm(blotch~site+variety, family=quasibinomial, data=leafblotch)

**Output:**

> sumary(fitqbin)

Estimate Std. Error t value Pr(>|t|)

(Intercept) -8.05460 1.42197 -5.6644 2.838e-07

site2 1.63906 1.44331 1.1356 0.2598797

site3 3.32648 1.34924 2.4655 0.0160684

site4 3.58219 1.34445 2.6644 0.0095118

site5 3.58384 1.34443 2.6657 0.0094789

site6 3.89320 1.34022 2.9049 0.0048764

site7 4.72994 1.33480 3.5436 0.0006976

site8 5.52259 1.33464 4.1379 9.387e-05

site9 6.79450 1.34071 5.0678 3.000e-06

variety10 4.25287 0.60423 7.0385 9.388e-10

variety2 0.15008 0.72369 0.2074 0.8362926

variety3 0.68946 0.67237 1.0254 0.3085988

variety4 1.04814 0.64944 1.6139 0.1109195

variety5 1.61468 0.62567 2.5807 0.0118974

variety6 2.37110 0.60902 3.8933 0.0002188

variety7 2.57117 0.60648 4.2395 6.551e-05

variety8 3.34193 0.60154 5.5556 4.393e-07

variety9 3.49994 0.60135 5.8201 1.511e-07

Dispersion parameter = 0.08878

n = 90 p = 18

Deviance = 6.12644 Null Deviance = 40.80291 (Difference = 34.67647)

The dispersion parameter is 0.089.

This dispersion estimate can be derived from the binomial GLM. The dispersion parameter equals to the Pearson Chi-square statistic divided by the degree of freedom.

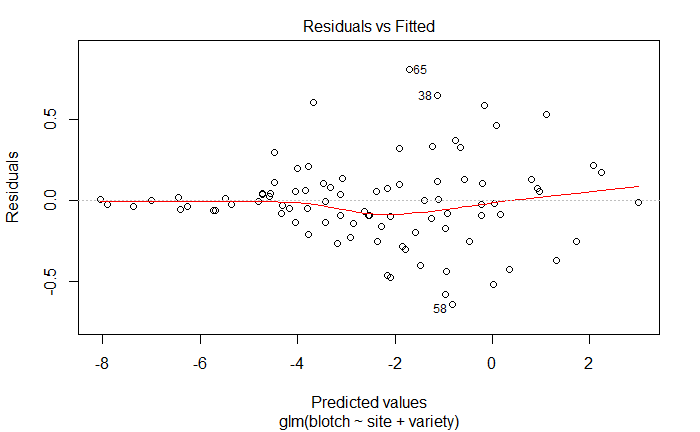
**Output:**

> dispersion = sum(residuals(fitbin,type = "pearson")^2)/df.residual(fitbin)

> dispersion

[1] 0.08878094

**(d)**



**Comments:**

From the residual plot we can see that from predicted values -8 to -5, the variance of residual is very small and close to zero. From predicted values -5 to -1, the variance of residual is increasing. From the fitted value -1, the variance of residual decreases as the predicted value increasing. As a result, the variance of residual is not constant in this model.

**(e)**

In the original binomial model, the variance function . To make the variance function we should define a weights function . In this problem, we can use the fitted in part (c).

Then we fit the binomial model with this weight function.

> fit\_1e = glm(blotch~site+variety,family=quasibinomial,data=leafblotch,weights = miu)

**Output:**

> sumary(fit\_1e)

Estimate Std. Error t value Pr(>|t|)

(Intercept) -7.78288 0.41679 -18.6734 < 2.2e-16

site2 1.32998 0.50434 2.6370 0.010240

site3 3.83452 0.42511 9.0202 1.916e-13

site4 3.59325 0.46226 7.7731 4.063e-11

site5 4.02142 0.43099 9.3307 5.077e-14

site6 4.09415 0.44773 9.1443 1.126e-13

site7 4.84618 0.45396 10.6753 < 2.2e-16

site8 5.50358 0.45758 12.0277 < 2.2e-16

site9 6.96006 0.45632 15.2525 < 2.2e-16

variety10 3.78624 0.43688 8.6666 8.731e-13

variety2 -0.43093 0.45211 -0.9532 0.343702

variety3 0.18377 0.44203 0.4157 0.678831

variety4 0.99995 0.39809 2.5119 0.014252

variety5 1.21714 0.43046 2.8276 0.006070

variety6 1.54080 0.46592 3.3070 0.001473

variety7 2.18037 0.43131 5.0552 3.151e-06

variety8 3.12639 0.42793 7.3059 3.005e-10

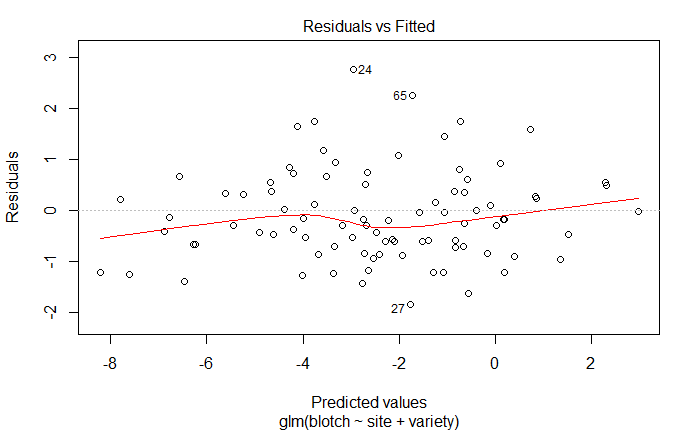
variety9 3.10408 0.43431 7.1472 5.912e-10

Dispersion parameter = 1.00027

n = 90 p = 18

Deviance = 71.15858 Null Deviance = 1106.42620 (Difference = 1035.26761)

Plot the residual against predicted values:



**Comments:**

From the residual plot we can see that the variance of residual is still not constant but he residual points are spread out, compared to the residual points in part (d). Hence, the variance of residual of this model seems more constant than that from the quasi-binomial model.

**(f)**

We use the F-test to test the significance of predictors in the full model.

**Output:**

> drop1(fit\_1e, test="F")

Single term deletions

Model:

blotch ~ site + variety

Df Deviance F value Pr(>F)

<none> 71.16

site 8 624.40 69.973 < 2.2e-16 \*\*\*

variety 9 256.14 20.797 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ?\*\*?0.001 ?\*?0.01 ??0.05 ??0.1 ??1

**Test the significance of predictor *site:***

The null hypothesis : the predictor *site* is not significant in the full model.

The corresponding p-value is smaller than the significance level , so we can reject the null hypothesis. Then we conclude that the predictor *site* is significant in the full model.

**Test the significance of predictor *variety:***

The null hypothesis : the predictor *variety* is not significant in the full model.

The corresponding p-value is smaller than the significance level , so we can reject the null hypothesis. Then we conclude that the predictor *variety* is significant in the full model.

**(g)**

We can find the interaction term by looking at the residual plot in part (e). From the residual plot, we can see that the data point with the largest residual value is no.24. For this data point, the *site* level is 3 and *variety* level is 4. The large residual value means that there should be some correlation between these two levels. As a result, the combination of site 3 and variety 4 shows the greatest indication of an interaction.

**R Code**

#a

library(faraway)

library(ggplot2)

data(leafblotch)

leafblotch$site=as.numeric(as.character(leafblotch$site))

leafblotch$variety=as.numeric(as.character(leafblotch$variety))

plot(leafblotch$site,leafblotch$blotch,xlab="site", ylab="blotch", main="blotch vs site")

plot(leafblotch$variety,leafblotch$blotch,xlab="variety", ylab="blotch", main="blotch vs variety")

#b

leafblotch$site=as.factor(as.character(leafblotch$site))

leafblotch$variety=as.factor(as.character(leafblotch$variety))

fitbin = glm(blotch~site+variety, family=binomial, data=leafblotch)

summary(fitbin)

#good-of-fit test

pchisq(deviance(fitbin),df.residual(fitbin),lower=FALSE)

#c

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

sumary(fitqbin)

#calculate dispersion parameter

dispersion = sum(residuals(fitbin,type = "pearson")^2)/df.residual(fitbin)

dispersion

#d

plot(fitqbin)

#e

miu = 1/(fitted(fitqbin)\*(1-fitted(fitqbin)))

fit\_1e = glm(blotch~site+variety,family=quasibinomial,data=leafblotch,weights = miu)

plot(fit\_1e)

sumary(fit\_1e)

#f

drop1(fit\_1e, test="F")