

# 719HW6

## Question 1

### A

The saturated model has 6 parameters. 1 from the intercept, 1 from the treatment, 2 from the disease, and 2 from the interaction of disease and treatment.

### B

Model 1:  $\text{logit}(p) = \beta_0 + \beta_1 * T$

Model 2:  $\text{logit}(p) = \beta_0 + \beta_1 * D$

Model 3:  $\text{logit}(p) = \beta_0 + (\beta_1 * T) + (\beta_2 * D) + (\beta_3 * T * D)$

### C

Model 1 is suitable to answer question 1. Model 2 is suitable to answer question 2. Model 3 is suitable to answer question 3.

### D

#### Build df

```
y <- c(51, 67, 42, 143, 97, 40)
n <- c(200, 158, 89, 315, 263, 75)
severity <- rep(c(0,1,2), 2)
treatment <- c(rep(0, 3), rep(1, 3))

df <- cbind.data.frame(y, n, severity, treatment)
df
```

	y	n	severity	treatment
1	51	200	0	0
2	67	158	1	0
3	42	89	2	0
4	143	315	0	1
5	97	263	1	1
6	40	75	2	1

#### Fit models

```
mod_1 <- glm(cbind(y, n-y) ~ treatment, data = df, family = binomial(link = "logit"))
mod_2 <- glm(cbind(y, n-y) ~ severity, data = df, family = binomial(link = "logit"))
mod_3 <- glm(cbind(y, n-y) ~ treatment*severity, data = df, family = binomial(link = "logit"))
```

### E

DF of the null deviance: 5

DF of the residual deviance for Model 1: 4

DF of the residual deviance for Model 2: 4

DF of the residual deviance for Model 3: 2

## F

```
mod1_sum <- summary(mod_1)
```

The p-value for the model's treatment coefficient is 0.0186188, which is statistically significant. This means the coefficient is significantly different from 0 and we can use the estimator to estimate the estimand.

The odds of hospitalization increases by a factor of 1.347, or 35%, when an individual is administered Drug B compared to Drug A. Drug A is better for preventing hospitalization.

## G

```
mod2_sum <- summary(mod_2)
```

The p-value for the model's severity coefficient is 0.0151081, which is statistically significant. This means the coefficient is significantly different from 0 and we can use it to estimate our parameter. It also means that severity is statistically significantly associated with hospitalization.

A 1 unit increase in severity results in a 1.231 odds increase of hospitalization when comparing Drug A to Drug B, which is 23.1% increase.

## H

```
mod3_sum <- summary(mod_3)
```

$$= \frac{\exp(\hat{\beta}_0 + \hat{\beta}_2)}{\exp(\hat{\beta}_0 + \hat{\beta}_1 + 2\hat{\beta}_2 + 2\hat{\beta}_3)} = \exp(-\hat{\beta}_1 - \hat{\beta}_2 - 2\hat{\beta}_3) = \exp(-0.7032 - .511 - 2 * (-.5122)) = \exp(-.1898) = 0.827$$

95% CI:

```
L <- cbind(c(0,-1, -1, -2))
Lb <- t(L)%*%mod_3$coef
varLb <- t(L)%*%mod3_sum$cov.unscaled%*%L
CILb <- Lb + c(-1, 1)*qnorm(1-.05/2)*sqrt(varLb)
```

Warning in c(-1, 1) \* qnorm(1 - 0.05/2) \* sqrt(varLb): Recycling array of length 1 in vector-array arithmetic is deprecated.

Use c() or as.vector() instead.

Warning in Lb + c(-1, 1) \* qnorm(1 - 0.05/2) \* sqrt(varLb): Recycling array of length 1 in array-vector arithmetic is deprecated.

Use c() or as.vector() instead.

```
exp(CILb)
```

```
[1] 0.5536296 1.2382083
```

## I

I can use model 3 to answer question 3. The interaction term is significant, which means the treatment effect changes for different levels of severity, and vice versa.

To show this, I will compare the effect of a change in severity in two different situations, one where treatment == 0 and one where treatment == 1.

```
mod3_b <- mod3_sum$coefficients
```

$$\frac{\exp(\beta_0 + \beta_1 * 0 + \beta_2 * 1 + \beta_3 * 0 * 1)}{\exp(\beta_0 + \beta_1 * 0 + \beta_2 * 2 + \beta_3 * 0 * 2)} = \frac{0.6205}{1.0343} = 0.5999$$

$$\frac{\exp(\beta_0 + \beta_1 * 1 + \beta_2 * 1 + \beta_3 * 1 * 1)}{\exp(\beta_0 + \beta_1 * 1 + \beta_2 * 2 + \beta_3 * 1 * 2)} = \frac{0.7503}{0.7494} = 1.0012$$

This shows that the odds ratio between levels of severity changes differently for different levels of disease. This shows that disease severity is an effect modifier.

J

NO YES YES

Only model 1 vs Model 3 & Model 2 vs Model 3 can be compared using the LRT.

```
anova(mod_1, mod_3, test = "LRT")
```

Analysis of Deviance Table

Model 1: cbind(y, n - y) ~ treatment

Model 2: cbind(y, n - y) ~ treatment \* severity

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	4	25.5436			
2	2	9.8977	2	15.646	0.0004004 ***

---

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

```
anova(mod_2, mod_3, test = "LRT")
```

Analysis of Deviance Table

Model 1: cbind(y, n - y) ~ severity

Model 2: cbind(y, n - y) ~ treatment \* severity

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	4	25.2144			
2	2	9.8977	2	15.317	0.0004721 ***

---

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

The larger model is preferred in both cases since  $H_0$  = Smaller model is preferred,  $H_a$  : Larger model is preferred and  $p < 0.05$ , so the interaction term should be included.

## Question 2

A

```
y <- c(55, 52, 57, 55, 50, 50)
n <- c(102, 99, 108, 76, 81, 90)
force <- rep(c(40, 150, 350), 2)
newstor <- c(rep("Control", 3), rep("Treatment", 3))

df2 <- cbind.data.frame(y, n, force, newstor)
```

```
fit2.1 <- glm(cbind(y, n-y) ~ newstor, family = binomial(link = "logit"), data = df2)
fit2.1
```

Call: glm(formula = cbind(y, n - y) ~ newstor, family = binomial(link = "logit"), data = df2)

Coefficients:

(Intercept)	newstorTreatment
0.1231	0.3985

Degrees of Freedom: 5 Total (i.e. Null); 4 Residual

Null Deviance: 10.45

Residual Deviance: 5.173 AIC: 38.74

```
fit2.2 <- glm(cbind(y, n-y) ~ newstor*log(force), family = binomial(link = "logit"), data = df2)
fit2.2
```

Call: glm(formula = cbind(y, n - y) ~ newstor \* log(force), family = binomial(link = "logit"), data = df2)

Coefficients:

(Intercept)	newstorTreatment
0.23389	1.97711
log(force)	newstorTreatment:log(force)
-0.02274	-0.31862

Degrees of Freedom: 5 Total (i.e. Null); 2 Residual

Null Deviance: 10.45

Residual Deviance: 0.02773 AIC: 37.6

## B

```
df2_ungrouped_left <- as.data.frame(rbind(
  cbind(force=rep(df2[, "force"], df2[, "y"]), n=1, y=1),
  cbind(force=rep(df2[, "force"], df2[, "n"]-df2[, "y"]), n=1, y=0)
))

df2_ungrouped_right <- as.data.frame(rbind(
  cbind(newstor=rep(df2[, "newstor"], df2[, "y"]),
  cbind(newstor=rep(df2[, "newstor"], df2[, "n"]-df2[, "y"]))
))

df2_ungrouped <- cbind(df2_ungrouped_left, df2_ungrouped_right)

df2_ungrouped$y <- as.numeric(df2_ungrouped$y)
df2_ungrouped$newstor <- as.factor(df2_ungrouped$newstor)
```

```
fit2.1.2 <- glm(y ~ newstor, family = binomial(link = "logit"), data = df2_ungrouped)
fit2.1.2
```

```
Call: glm(formula = y ~ newstor, family = binomial(link = "logit"),
  data = df2_ungrouped)
```

Coefficients:

```
(Intercept)  newstorTreatment
      0.1231         0.3985
```

Degrees of Freedom: 555 Total (i.e. Null); 554 Residual

Null Deviance: 758.6

Residual Deviance: 753.4 AIC: 757.4

```
fit2.2.2 <- glm(y ~ newstor * log(force), family = binomial(link = "logit"), data = df2_ungrouped)
fit2.2.2
```

```
Call: glm(formula = y ~ newstor * log(force), family = binomial(link = "logit"),
  data = df2_ungrouped)
```

Coefficients:

```
(Intercept)          newstorTreatment
      0.23389          1.97711
log(force) newstorTreatment:log(force)
     -0.02274          -0.31862
```

Degrees of Freedom: 555 Total (i.e. Null); 552 Residual

Null Deviance: 758.6

Residual Deviance: 748.2 AIC: 756.2

The models give the same conclusions using grouped and ungrouped data, though some calculated values are different such as the deviances.

## C

```
an_1 <- anova(fit2.1, fit2.2, test = "LRT")
an_1
```

Analysis of Deviance Table

Model 1: cbind(y, n - y) ~ newstor

Model 2: cbind(y, n - y) ~ newstor \* log(force)

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	4	5.1730			
2	2	0.0277	2	5.1452	0.07634 .
---					

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

```
pval_1 <- an_1$`Pr(>Chi)`[2]
```

```
an_2 <- anova(fit2.1.2, fit2.2.2, test = "LRT")
an_2
```

#### Analysis of Deviance Table

Model 1: y ~ newstor

Model 2: y ~ newstor \* log(force)

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	554	753.36			
2	552	748.22	2	5.1452	0.07634 .

---

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

```
pval_2 <- an_2$`Pr(>Chi)`[2]
```

They do give the same result.

## D

$H_0$  : Model 1 better than Model 2

$H_a$  : Model 2 better than Model 1

The p-value is not significant for either test (0.0763 & 0.0763), meaning we fail to reject the null and conclude that the simpler model is better, i.e. we do not need to include log(force) at all in the model. It is better to use just the storage condition to predict the outcome than a model that includes storage and log(force).