

Name Jenny

Student Number _____

STA 312f2022 Quiz 8

1. Hens (female chickens) are fed with one of three different feed types: A, B or C. The response variable is the mean weight of the eggs they lay, based on 100 eggs from each chicken. Hen's age is a covariate. Assume that the relationship between age and expected egg weight can be approximated by a straight line over the range of the data, and that the *slope of the line might not be the same* for the different feed types.

- (a) (1 point) ^{huvv} Write a regression equation with age as the first explanatory variable. Call it x . You ~~do not~~ need ~~not~~ say how your dummy variables are defined. You will do that in the next part. Needless to say, there is no ϵ in your answer.

$$E(Y|x) = \beta_0 + \beta_1 x + \beta_2 d_1 + \beta_3 d_2 + \beta_4 x d_1 + \beta_5 x d_2$$

Names of the dummy variables do not matter.

- (b) (1 point) Make a table with three rows, one for each feed type. Feed type C should be the reference category. Make columns showing how your dummy variables are defined. Add another, wider column showing expected egg weight for each feed type. The *symbols* for the dummy variables for feed type will not appear in this last column, because they are either zero or one.

	d_1	d_2	$E(Y x)$
A	1	0	$\beta_0 + \beta_1 x$
B	0	1	$\beta_0 + \beta_2 + (\beta_1 + \beta_4) x$
C	0	0	$\beta_0 + \beta_3 + (\beta_1 + \beta_5) x$

ORDER AND GROUPING of terms does not matter, though of course my way makes the rest easier.

- (c) (1 point) You want to test whether the three regression lines have equal slopes. Give the null hypothesis in symbols.

$$H_0: \beta_4 = \beta_5 = 0 \quad (\text{Must have } = 0 \text{ or one full point off})$$

- (d) (1 point) Write that last null hypothesis in matrix form as $L\beta = 0$.

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

- (e) (1 point) You want to test whether the effect of feed type depends on the chicken's age. Give the null hypothesis in symbols.

$$H_0: \beta_4 = \beta_5 = 0 \quad (\text{Again, need } = 0 \text{ for any marks})$$

- (f) (1 point) You want to test whether the slope of the regression line relating age to expected egg weight is different for feed types A and B. Give the null hypothesis in symbols.

$$H_0: \beta_4 = \beta_5 \quad (\text{One mark off if } = 0)$$

2. This question is ^{uses} based on your printout from the birth weight data. It is based on the model with these explanatory variables: Age, mother's weight, race, smoking status during pregnancy, and an indicator for any first-trimester visits to the doctor (1 = one or more, 0 = none). If you do not have a printout, don't answer Question 2.

(a) (2 points) ^{holding} Controlling for all other variables, are the chances of having a low birth weight baby different for White and Other mothers? ^{constant}

- i. Based on the 0.05 significance level, answer Yes or No.

Yes

- ii. Give the test statistic (either z or χ^2). On your printout, circle the number and write "Question 2" beside it.

$$z = 2.037$$

- iii. Give the p -value. This is a number.

$$p = 0.03259$$

- iv. In plain, non-statistical language, what do you conclude?

Controlling for all other variables, mothers of race Other are more likely to have a low birth weight baby than white mothers

- (b) (2 points) Estimate the probability of a low birth weight baby for a 20-year-old, White, non-smoking mother weighing 140 pounds, who made no first trimester visits to the doctor. The answer is a number. Circle your answer.

$$x'\hat{\beta} = 0.35174 - 20 \times 0.019288 - 140 \times 0.012507 + 0 + 0 + 0 + 0 = -1.785, \text{ and}$$

$$\hat{\pi} = \frac{e^{x'\hat{\beta}}}{1 + e^{x'\hat{\beta}}} = 0.144$$

Attach your printout for Question 2 (Homework Question 2). Make sure your name is written on the printout.

R version 4.2.0 (2022-04-22) -- "Vigorous Calisthenics"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.78 (8075) x86_64-apple-darwin17.0]

[Workspace restored from /Users/brunner/.RData]
[History restored from /Users/brunner/.Rapp.history]

```
> # Birth weight data for Assignment 8
> rm(list=ls()); options(scipen=999)
> library(MASS); # help(birthwt)
> head(birthwt)
  low age lwt race smoke ptl ht ui ftv bwt
85   0  19 182   2    0  0  0  1  0 2523
86   0  33 155   3    0  0  0  0  3 2551
87   0  20 105   1    1  0  0  0  1 2557
88   0  21 108   1    1  0  0  1  2 2594
89   0  18 107   1    1  0  0  1  0 2600
91   0  21 124   3    0  0  0  0  0 2622
>
> # Explore a bit first
> # with(birthwt, table(ftv))
> # with(birthwt, table(race)) # 1 = white, 2 = black, 3 = other
> # summary(birthwt) # All numeric
>
> bweight = within(birthwt, {
+   race = factor(race, labels = c('White', 'Black', 'Other')) # Order maintained
+   anyftv = ftv; anyftv[anyftv>1] = 1
+ })
> with(bweight, table(race)) # Now what does it look like?
race
White Black Other
  96   26   67
> contrasts(bweight$race) # White still reference category.
      Black Other
White    0     0
Black    1     0
Other    0     1
>
>
> # Full model
> model1 = glm(low ~ age + lwt + race + smoke + anyftv, data=bweight, family=binomial)
> summary(model1)
```

```
Call:
glm(formula = low ~ age + lwt + race + smoke + anyftv, family = binomial,
    data = bweight)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.5289	-0.9075	-0.5797	1.3007	2.0550

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.351747	1.110641	0.317	0.75147
age	-0.019288	0.034977	-0.551	0.58133
lwt	-0.012507	0.006387	-1.958	0.05021 .
raceBlack	1.224274	0.517471	2.366	0.01799 *
raceOther	0.908265	0.425003	2.137	0.03259 *
smoke	1.024100	0.387048	2.646	0.00815 **
anyftv	-0.149082	0.352620	-0.423	0.67245

Question 2

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 234.67 on 188 degrees of freedom
 Residual deviance: 214.40 on 182 degrees of freedom
 AIC: 228.4

Number of Fisher Scoring iterations: 4

```
>
> # 12a
> se = sqrt(diag(vcov(model1))); se
(Intercept)      age      lwt  raceBlack  raceOther      smoke      anyftv
1.110641332 0.034976846 0.006386949 0.517471090 0.425003136 0.387048253 0.352619776
>
> # For (b) and (c), see summary
>
> betahat = coefficients(model1); betahat
(Intercept)      age      lwt  raceBlack  raceOther      smoke      anyftv
0.35174720 -0.01928760 -0.01250698 1.22427400 0.90826519 1.02409982 -0.14908197
>
> # 12d Correcting for all the other variables, the odds of a low birth weight baby are an
estimated \underline{\hspace{10mm}} times as great for a mother who smokes during pregnancy.
> exp(betahat[6]) # 2.78
smoke
2.784588
>
> # 12e (i) LR test for race
> norace = glm(low ~ age + lwt + smoke + anyftv, data=bweight, family=binomial)
> anova(norace,model1, test="Chisq")
Analysis of Deviance Table
```

Model 1: low ~ age + lwt + smoke + anyftv
 Model 2: low ~ age + lwt + race + smoke + anyftv

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	184	222.19			
2	182	214.40	2	7.7919	0.02032 *

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

```

>
> # 12e (ii) Wald test for race
> Vhat = vcov(model1)
> # For Wald tests: Wtest = function(L,Tn,Vn,h=0) # H0: L theta = h
> source("http://www.utstat.utoronto.ca/brunner/Rfunctions/Wtest.txt")
> LL = rbind(c(0,0,0,1,0,0,0),
+           c(0,0,0,0,1,0,0))
> Wtest(LL,betahat,Vhat)
              W              df      p-value
7.40790875 2.00000000 0.02462595
>
> # (12f) See summary
>
> # 12g Correcting for all the other variables, the odds of a low birth weight baby are an
estimated \underline{\hspace{10mm}} times as great for a Black mother, compared to a White mother.
> exp(betahat[4])
raceBlack
3.401696
>
> # 12h Give a 95% confidence interval for that last number.
> # Indicator for Black is 4th element
> a = betahat[4] - 1.96*se[4]; b = betahat[4] + 1.96*se[4]
> c(exp(a),exp(b))
raceBlack raceBlack
1.233716 9.379414
>
> # 12i) For a 26-year-old, 130 pound, Black non-smoking mother with no first trimester visits to
the doctor, estimate the probability of a low birth weight baby and give a 95% confidence interval
for the probability.
>
> # Estimate first without using predict.
> betahat
(Intercept)      age      lwt  raceBlack  raceOther      smoke  anyftv
0.35174720 -0.01928760 -0.01250698 1.22427400 0.90826519 1.02409982 -0.14908197
> x = c(1,26,130,1,0,0,0); xb = sum(x*betahat)
> pihat = exp(xb)/(1+exp(xb)); pihat # 0.3655479
[1] 0.3655479
> # Now try predict
> head(bweight)
  low age lwt  race smoke ptl ht ui ftv  bwt anyftv
85   0  19 182 Black    0   0  0  1   0 2523     0
86   0  33 155 Other    0   0  0  0   3 2551     1
87   0  20 105 White    1   0  0  0   1 2557     1
88   0  21 108 White    1   0  0  1   2 2594     1
89   0  18 107 White    1   0  0  1   0 2600     0
91   0  21 124 Other    0   0  0  0   0 2622     0
> onemom = data.frame(age=26, lwt=130, race='Black', smoke=0, anyftv=0)
> pred1 = predict(model1, newdata=onemom, type = "response", se.fit = TRUE); pred1
$fit
      1
0.3655479

$se.fit
      1
0.1240606

$residual.scale
[1] 1

```

```

> # CI using the delta method
> lower95 = pred1$fit - 1.96*pred1$se
> upper95 = pred1$fit + 1.96*pred1$se
> c(lower95,upper95)
      1      1
0.1223890 0.6087068
> # CI starting from difference in log odds. Quoting,
> # The second way is to get an interval for the log odds, and transform it.
> # It works because the probability is an increasing function of the log odds.
> # The default for predict is x'betahat
> pred2 = predict(model1,newdata=onemom, se.fit = TRUE); pred2
$fit
      1
-0.5513644

$se.fit
[1] 0.5349226

$residual.scale
[1] 1

> a = pred2$fit - 1.96*pred2$se; b = pred2$fit + 1.96*pred2$se
> Lower95 = exp(a)/(1+exp(a)); Upper95 = exp(b)/(1+exp(b))
> c(Lower95,Upper95)
      1      1
0.1680078 0.6217738
>
> # 12j Comparing Black and Other mothers
> # (i) Wald test
> LblackVSother = rbind(c(0,0,0,1,-1,0,0))
> Wtest(LblackVSother,betahat,Vhat) # W = 0.353, p = 0.5523908
      W      df    p-value
0.3530515 1.0000000 0.5523908
> # (ii) Odds ratio
> orB0 = exp(betahat[4]-betahat[5]); orB0 # 1.37 (?)
raceBlack
1.371642
> # (iii) CI for odds ratio
> seB0 = sqrt(as.numeric(LblackVSother %*% Vhat %*% t(LblackVSother))); seB0
[1] 0.5318389
> a = betahat[4]-betahat[5] - 1.96*seB0; b = betahat[4]-betahat[5] + 1.96*seB0
> c(exp(a),exp(b)) # 0.4836492 3.8900154
raceBlack raceBlack
0.4836492 3.8900154
>
> # Do it with a different reference category as a cross check
> # Remember the original coding of race is 123
> bw2 = within(bweight,
+ {contrasts(race) = contr.treatment(3,base=3) # Reference is Other
+ })
> with(bw2,contrasts(race))
      1 2
White 1 0
Black 0 1
Other 0 0
> model2 = glm(low ~ age + lwt + race + smoke + anyftv, family = binomial, data = bw2)
> summary(model2)

```

Call:

```
glm(formula = low ~ age + lwt + race + smoke + anyftv, family = binomial,
    data = bw2)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.5289	-0.9075	-0.5797	1.3007	2.0550

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.260012	1.017375	1.238	0.21553
age	-0.019288	0.034977	-0.551	0.58133
lwt	-0.012507	0.006387	-1.958	0.05021 .
race1	-0.908265	0.425003	-2.137	0.03259 *
race2	0.316009	0.531839	0.594	0.55239
smoke	1.024100	0.387048	2.646	0.00815 **
anyftv	-0.149082	0.352620	-0.423	0.67245

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 234.67 on 188 degrees of freedom
 Residual deviance: 214.40 on 182 degrees of freedom
 AIC: 228.4

Number of Fisher Scoring iterations: 4

```
> # Notice how the p-value for race2 = 0.55239, same as for the earlier Wald test
> # Now odds ratio Black/Other = 1.371642
> betahat2 = coefficients(model2); betahat2
(Intercept)      age      lwt      race1      race2      smoke      anyftv
 1.26001240 -0.01928760 -0.01250698 -0.90826519  0.31600880  1.02409982 -0.14908197
> orB0
raceBlack
 1.371642
> exp(betahat2[5])
      race2
 1.371642
>
> # Maybe I should always name the columns after I set the contrasts, like this
> with(bw2, contrasts(race))
      1 2
White 1 0
Black 0 1
Other 0 0
> bw2 = within(bw2,{
+ colnames(contrasts(race)) = c('White','Black')
+ })
> with(bw2, contrasts(race))
      White Black
White    1     0
Black    0     1
Other    0     0
> model2 = glm(low ~ age + lwt + race + smoke + anyftv, family = binomial, data = bw2)
> summary(model2)
```

Call:

```
glm(formula = low ~ age + lwt + race + smoke + anyftv, family = binomial,
    data = bw2)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.5289	-0.9075	-0.5797	1.3007	2.0550

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.260012	1.017375	1.238	0.21553
age	-0.019288	0.034977	-0.551	0.58133
lwt	-0.012507	0.006387	-1.958	0.05021 .
raceWhite	-0.908265	0.425003	-2.137	0.03259 *
raceBlack	0.316009	0.531839	0.594	0.55239
smoke	1.024100	0.387048	2.646	0.00815 **
anyftv	-0.149082	0.352620	-0.423	0.67245

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 234.67 on 188 degrees of freedom
Residual deviance: 214.40 on 182 degrees of freedom
AIC: 228.4

Number of Fisher Scoring iterations: 4

>
>
>
>
>
>
>
>
>