HW 8 Key

SDS 291

April 20, 2020

#10.12 / 10.23

##a

```
Titanic$Sex<-relevel(Titanic$Sex, ref="male")
m1012a<-glm(Survived~Age+Sex, data=Titanic, family=binomial)
summary(m1012a)
```

```
##
## Call:
## glm(formula = Survived ~ Age + Sex, family = binomial, data = Titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7541  -0.6905  -0.6504   0.7576   1.8628
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.159839   0.219651  -5.280 1.29e-07 ***
## Age         -0.006352   0.006187  -1.027    0.305
## Sexfemale    2.465996   0.178455  13.819  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1025.57  on 755  degrees of freedom
## Residual deviance:  795.59  on 753  degrees of freedom
##   (557 observations deleted due to missingness)
## AIC: 801.59
##
## Number of Fisher Scoring iterations: 4
```

$log(odds)$ = -1.1598389 + -0.006352 age + 2.4659959 sexFemale

$\hat{\pi} = \frac{e^{-1.159839+(-.006352age)+(2.465996smoker)}}{1+e^{-1.159839+(-.006352age)+(2.465996smoker)}}$

- *check: if fit the model and wrote out both forms of the model correctly*
- *check-minus: if didn't fit correctly and only wrote one of the forms of the model*

##b

Age negatively associated with survival - as age increases, the log odds of being surviving decreases, adjusting for sex - and the relationship is no longer statistically significant.

Females have higher, and statistically significantly higher, log odds of survival than men, adjusted for age.

- *check-plus: they include correct and well-stated interpretations of odds ratios*
- *check: any resonable answer here that includes both direction and the statistical significance*
- *check-minus: if they don't talk about the direction* and *statistical significance (i.e., if they only mention statistical significance. . .)*

##c.

```
newdata = data.frame(Age=18, Sex="male")
prob_18male<-predict(m1012a,newdata, type="response")
odds_18male<-prob_18male/(1-prob_18male)
```

For a 18 year old man, the probability and odds of surviving the Titanic:

- Probability: 0.2185435
- Odds: 0.2796618
- *check-plus: if they (correctly) interpret the probability and odds in a sentence*
- *check: correctly calculated both*
- *check-minus: if they miscalculated* or *only included one, not both*

##d.

```
newdata = data.frame(Age=18, Sex="female")
prob_18female<-predict(m1012a,newdata, type="response")
odds_18female<-prob_18female/(1-prob_18female)

OR_18<-odds_18female/odds_18male
```

For a 18 year old female, the probability and odds of surviving the Titanic:

- Probability: 0.7670667
- Odds: 3.2930744

The Odds Ratio (OR) of the odds of death for a a 18 year old woman who smoked compared to a woman who did not smoke is:

OR: $\frac{Odds_{females}}{Odds_{males}} = 11.7752038$

- *check-plus: if they (correctly) interpret the probability and odds in a sentence* and *include the OR in a sentence*
- *check: correctly calculated all: odds, probability, and OR*
- *check-minus: if they miscalculated more than one thing, didn't include the OR, or only included odds but not probability*

##e.

```
newdata = data.frame(Age=50, Sex="male")
prob_50male<-predict(m1012a,newdata, type="response")
odds_50male<-prob_50male/(1-prob_50male)

newdata = data.frame(Age=50, Sex="female")
prob_50female<-predict(m1012a,newdata, type="response")
odds_50female<-prob_50female/(1-prob_50female)

OR_50<-odds_50female/odds_50male
```

Sex/Outcome	18yo female	18yo male	OR	50yo female	50yo male	OR
Probability	0.7670667	0.2185435		0.7288031	0.1858148	

Sex/Outcome	18yo female	18yo male	OR	50yo female	50yo male	OR
Odds	3.2930744	0.2796618	11.7752038	2.6873579	0.2282218	11.7752038

- *check-plus: if they (correctly) interpret the probability and odds in a sentence*
- *check: correctly calculated both odds and probability for both males and females*
- *check-minus: if they miscalculated* or *only included one, not both*

##f.

```
exp(coef(m1012a))
```

```
## (Intercept)         Age   Sexfemale
##   0.3135367   0.9936682  11.7752038
```

The odds ratio of survival for females compared to males is always the same for all ages - this is a reflection of "controlling for age".

Age here is the (linear) slope and females and males just have different intercepts; this is like the parallel slopes model we saw before in linear regression. The slope of the lines are still parallel, so the ratio of the odds for females compared to males is the same at all ages.

- *check-plus: if they give an thorough answer of how and why*
- *check: if they say "yes" it will be the same at all ages without much explanation*
- *check-minus: if they reach the wrong conclusion*

RECOMMENDED

No need to grade these #10.13 / 10.24

##a

```
m1013a<-glm(Survived~Age*Sex, data=Titanic, family=binomial)
summary(m1013a)
```

```
##
## Call:
## glm(formula = Survived ~ Age * Sex, family = binomial, data = Titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1262  -0.7348  -0.5194   0.7699   2.2632
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.298750   0.277699  -1.076    0.282
## Age           -0.036367   0.009263  -3.926 8.63e-05 ***
## Sexfemale      0.599858   0.408050   1.470    0.142
## Age:Sexfemale  0.065718   0.013686   4.802 1.57e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1025.57  on 755  degrees of freedom
## Residual deviance:  770.56  on 752  degrees of freedom
##   (557 observations deleted due to missingness)
## AIC: 778.56
##
## Number of Fisher Scoring iterations: 4
```

The addition of the interaction term changes the models for Females and Males – the slope for age now varies by sex.

$log(odds)_{females}$ = -0.2987495 + -0.0363669 age + 0.5998582 smoker + 0.0657179 smoker:age

$log(odds)_{females}$ = (-0.2987495 + 0.5998582) + (-0.0363669 + 0.0657179) $\cdot age$

$log(odds)_{females}$ = 0.3011086 + 0.0293511 $\cdot age$

$log(odds)_{males}$ = -0.2987495 + -0.0363669 $\cdot age$

We now have a model with different intercepts *and* different slopes, so we wouldn't expect the odds ratio between smokers and non-smokers to be constant – they now vary by age. We can see from the equations above that the slope for age among smokers is now smaller that it is for non-smokers. What we see below is that as a woman who smokes gets older, the ratio of the odds of death between her and a similar aged nonsmoker gets smaller.

Smoking/Outcome	18yo female	18yo male	OR	50yo female	50yo male	OR
Probability	0.6962339	0.278211		0.8542911	0.1074466	
Odds	2.2920065	0.3854465	8.2383736	5.8629996	0.1203812	54.566625

```r
exp(coef(m1013a))
```

```
##   (Intercept)           Age     Sexfemale Age:Sexfemale
##     0.7417452     0.9642865     1.8218604     1.0679254
```

-0.5 if don't illustrate multiple models -0.5 if no explanation of how the interaction term changed the OR

##b

```r
newtitanic<-Titanic %>% filter(!is.na(Age) & !is.na(Sex))
mSex<-glm(Survived~Sex, data=newtitanic, family=binomial)
anova(mSex,m1013a, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Survived ~ Sex
## Model 2: Survived ~ Age * Sex
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       754     796.64
## 2       752     770.56  2   26.088 2.163e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can reject the null hypothesis that $\hat{\beta}_2 = \hat{\beta}_4 = 0$, since the χ^2 statistic is large (and above the critical value of 6) and the p-value is small (p<0.001) and conclude that adjusting for age and letting the relationship between age and survival differ between males and females does a better job of estimating the log(odds) of death than the simple regression model with just sex alone.

-0.5 if misestimated (it's fine if the order of models are switched and the test statistic is negative) -0.5 if compared to the model from 10.12 with sex and age instead of a different model with just sex

#10.19 / 10.31 ##a

```r
newtitanic2<-Titanic %>% filter(PClass!="*") %>%
  mutate(Surv_fct=as.factor(Survived),
    PClass2=as.factor(PClass))
newtitanic2$Surv_fct<-relevel(newtitanic2$Surv_fct, ref="1")
gmodels::CrossTable(newtitanic2$PClass, newtitanic2$Surv_fct,
  prop.r=TRUE, prop.c=FALSE, prop.chisq = FALSE, prop.t = FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |-------------------------|
##
##
## Total Observations in Table:  1312
##
##
##                    | newtitanic2$Surv_fct
## newtitanic2$PClass |         1 |         0 | Row Total |
## -------------------|-----------|-----------|-----------|
##                1st |       193 |       129 |       322 |
##                    |     0.599 |     0.401 |     0.245 |
## -------------------|-----------|-----------|-----------|
```

```
##                2nd |       119 |       160 |       279 |
##                    |     0.427 |     0.573 |     0.213 |
## -------------------|-----------|-----------|-----------|
##                3rd |       138 |       573 |       711 |
##                    |     0.194 |     0.806 |     0.542 |
## -------------------|-----------|-----------|-----------|
##       Column Total |       450 |       862 |      1312 |
## -------------------|-----------|-----------|-----------|
##
##
```

We see that the proportion of survival is highest among 1st Class passengers. 2nd and 3rd class passengers have lower odds of survival, likely based on their further distance from the lifeboats.

-0.5 if probabilities are miscalculated -0.5 if no explanation of the results in the context of the problem

##b

```
gmodels::CrossTable(newtitanic2$PClass, newtitanic2$Surv_fct,
  prop.r=TRUE, prop.c=FALSE, prop.chisq = FALSE, prop.t = FALSE,
  chisq = TRUE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |-------------------------|
##
##
## Total Observations in Table:  1312
##
##
##                     | newtitanic2$Surv_fct
## newtitanic2$PClass |         1 |         0 | Row Total |
## -------------------|-----------|-----------|-----------|
##                1st |       193 |       129 |       322 |
##                    |     0.599 |     0.401 |     0.245 |
## -------------------|-----------|-----------|-----------|
##                2nd |       119 |       160 |       279 |
##                    |     0.427 |     0.573 |     0.213 |
## -------------------|-----------|-----------|-----------|
##                3rd |       138 |       573 |       711 |
##                    |     0.194 |     0.806 |     0.542 |
## -------------------|-----------|-----------|-----------|
##       Column Total |       450 |       862 |      1312 |
## -------------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  172.5191     d.f. =  2     p =  3.45104e-38
```

```
##
##
##
```

H_0: There is no relationship between PClass and Survival. H_A: There is a relationship between PClass and Survival.

Since the χ^2 test statistic large (172.519) and the p-value is <0.001, we can reject the null hypothesis and conclude that there is a relationship between PClass and Survival.

-0.5 if no hypotheses are mentioned

##c

```
m1019c<-glm(Survived~PClass2, data=newtitanic2, family=binomial)
summary(m1019c)
```

```
##
## Call:
## glm(formula = Survived ~ PClass2, family = binomial, data = newtitanic2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3526  -0.6569  -0.6569   1.0118   1.8108
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.4029     0.1137   3.543 0.000396 ***
## PClass22nd   -0.6989     0.1661  -4.208 2.58e-05 ***
## PClass23rd   -1.8265     0.1481 -12.335  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1687.2  on 1311  degrees of freedom
## Residual deviance: 1514.1  on 1309  degrees of freedom
## AIC: 1520.1
##
## Number of Fisher Scoring iterations: 4
```

```
exp(coef(m1019c))
```

```
## (Intercept)  PClass22nd  PClass23rd
##   1.4961240   0.4971179   0.1609744
```

The odds of surviving for a passenger in 1st class is 1.5 to 1. These odds are significantly greater than 0.

The odds of surviving for a passenger in 2nd class is 0.5 times (i.e., lower) than for 1st Class passengers. This difference was statistically significant, as evidenced by the z statistic>1.96 and the p-value<0.05.

The odds of surviving for a passenger in 3rd class is 0.16 times (i.e., lower) than for 1st Class passengers.

-0.5 if miscalculated (if PClass is the reference, and all three PClass variables are in the model) -0.5 if no interpretation*

##d

```
m1019c_1st<-coef(m1019c)[1]
m1019c_2nd<-coef(m1019c)[2]
```

```
m1019c_3rd<-coef(m1019c)[3]


newdata_1st = data.frame(PClass2="1st")
prob_1st<-predict(m1019c,newdata_1st, type="response")
odds_1st<-prob_1st/(1-prob_1st)

newdata_2nd = data.frame(PClass2="2nd")
prob_2nd<-predict(m1019c,newdata_2nd, type="response")
odds_2nd<-prob_2nd/(1-prob_2nd)

newdata_3rd = data.frame(PClass2="3rd")
prob_3rd<-predict(m1019c,newdata_3rd, type="response")
odds_3rd<-prob_3rd/(1-prob_3rd)

OR_2nd<-odds_2nd/odds_1st

OR_3rd<-odds_3rd/odds_1st
```

Smoking/Outcome	1st Class	2nd Class	OR	3rd Class	OR
Probability	0.5993789	0.4265233		0.1940928	
Odds	1.496124	0.74375	0.4971179	0.2408377	0.1609744

We get the same fitted probabilities from the model than we do from the table in #9.

```
gmodels::CrossTable(newtitanic2$PClass, newtitanic2$Surv_fct,
  prop.r=TRUE, prop.c=FALSE, prop.chisq = FALSE, prop.t = FALSE,
  chisq = TRUE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |-------------------------|
##
##
## Total Observations in Table:  1312
##
##
##                     | newtitanic2$Surv_fct
## newtitanic2$PClass |         1 |         0 | Row Total |
## -------------------|-----------|-----------|-----------|
##                1st |       193 |       129 |       322 |
##                    |     0.599 |     0.401 |     0.245 |
## -------------------|-----------|-----------|-----------|
##                2nd |       119 |       160 |       279 |
##                    |     0.427 |     0.573 |     0.213 |
## -------------------|-----------|-----------|-----------|
##                3rd |       138 |       573 |       711 |
##                    |     0.194 |     0.806 |     0.542 |
## -------------------|-----------|-----------|-----------|
```

```
##       Column Total |       450 |       862 |      1312 |
## -------------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  172.5191     d.f. =  2     p =  3.45104e-38
##
##
##
```

-0.5 if miscalculated if they don't match

##e

```
anova(m1019c, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     1311     1687.2
## PClass2  2   173.14      1309     1514.1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes, it does match the answer from above.

-0.5 if they don't calculate this correctly