CatData HW6

Matthew Vanaman

05-16-2019

All work and code are shown in the appendix.

4.3

(a): In the linear probability model, the coefficient is treated the same way as it is in linear regression. That is, it represents a slope such that for every unit increase x (decade, in this case), there is a corresponding unit change in y (the probability of pitching a complete game, in this case, the change of which = -0.0694).

(b):

$$\hat{\pi} = 0.7578 - 0.0694(12),$$
$$= -0.075.$$

Percentages can't be negative! So not possible.

(c):

$$\hat{\pi} = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}.$$
$$\hat{\pi} = \frac{e^{1.148-0.315(12)}}{1+e^{1.148-0.315(12)}},$$
$$= 0.0671071.$$

This is much more plausible.

4.5

(a):

The model comes out to be 15.04 + (-0.23)x. When temperature is 0, the log odds are 15.04. The log odds decrease by 0.23 with every unit increase in temperature.

(b):

$$\hat{\pi} = \frac{e^{15.04-0.23(31)}}{1+e^{15.04-0.23(31)}}.$$
$$= 0.9996331.$$

(c):

You need to solve for x:

$$0.50 = \frac{e^{15.04-0.23(x)}}{1+e^{15.04-0.23(x)}}.$$

The temperature comes out to be 64.8. To get the linear approximation (i.e. rate of change), take the derivative of the logistic function above (with respect to x) to get its probability density function:

$$f'(x) = \frac{\beta e^{\alpha+\beta x}}{(1+e^{\alpha-\beta(x)})^2}$$

$$= \frac{15.04(e^{15.04-0.23(64.8)})}{(1+e^{15.04-0.23(64.8)})^2}$$

$$= -0.0580407.$$

At 64.8 degrees, for every unit increase in temperature, there is a 0.058 decrease in probability.

(d):

Since the coefficient is the log odds ratio, you exponentiate it to get the odds. Exp(-0.2321627) = 0.7928171. This seems kind of difficult to interpret though, so it might be easier to take $1/OR = 1/0.79 = 1.27$ and say that for every unit increase in temperature, the odds of the o-rings *working correctly* increase by 1.27 - that is, every unit increase in temperature means there is 27% greater odds of the orings working the way they're supposed to (or 27% greater probability if we think events are rare enough to warrent using the odds ratio as an approximation of the risk ratio).

(e):

The Wald test is:

$$z^2 = \left(\frac{\beta}{SE}\right)^2,$$

$$= \left(\frac{-0.2321627}{0.1082}\right)^2,$$

$$= 4.6039476.$$

z^2 approximates the chi-square; the p-value is 0.0318984 with one degree of freedom.

The likelihood ratio test is:

$$-2(\ell_0 - \ell_1),$$

$$-2(-14.1335764 - 10.1575963),$$

$$7.95196.$$

χ^2 p-value at 1 degree of freedom is 0.0048035. Temperature is probably important here.

A4.15

(a)

The CMH test is:

$$\frac{[\sum_k (n_{11k} - \mu_{11k})]^2}{\sum_k \text{Var}(n_{11k})},$$

where:

$$\mu_{11k} = \frac{n_{1+k}n_{+1k}}{n_{++k}},$$

and:

$$\text{Var}_{11k} = \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n_{++k}^2(n_{++k}-1)}.$$

The CMH test statistic comes out to be 7.8149199 with 1 df, and a p-value of 0.0051817. We reject the null hypothesis that merit pay decisions are independent of race.

(b)

You could use the wald test to test the difference between the race parameter and the intercept while excluding distrit from the model. This would mean the intercept is the effect of race when race is Black, and the parameter is the effect when race is White. The parameter for White is zero according to the nuul. The wald test is $\left(\frac{\beta}{SE}\right)^2$, which comes out the be 8.4655366 with a p-value of 0.0036194. We reject the null of no difference between race in merit decision pay.

(c)

Unlike with the CMH, the model tells you the log-odds ratio which you can exponentiate to get the odds ratio. This tells us the magnitude of difference in decision pay rates between whites and blacks. Exponentiating the parameter for Whites, you get 2.2371917, meaning Whites have 123% greater odds of getting a pay increase.

4.31

The odds ratio of marijuana use for alcohol users vs non-users, ignoring whether or not they smoke, is 61.87. The odds ratio of marijuana use for smokers vs non-smokers, ignoring whether or not they drink, is 25.14. These are pretty large, and may be deceptive because they are not conditioned on the other variable. On theoretical grounds, there may be reason to suspect that the use of cigarettes in addition to alcohol (or vice versa) has a compounding effect on the odds of marijuana use, so it makes sense to run a logistic regression with main effects for cigarette use and alcohol use as well as a cigarette x alchohol interaction. Using the likelihood ratio test $(-2(\ell_0 - \ell_1))$, we should first test to see whether we gain anything from adding cigarettes to an intercept-only model. We get a LRT statistic 751.81, $p < 0.001$ with 1 degree of freedom, strongly suggesting that the cigarette model is an improvement over the intercept-only model. Big surprise. But what about adding alcohol to the model? The LRT is 91.64 $p < 0.001$ with 1 degree of freedom. It would thus be wise to include alcohol in the model. Lastly, what about adding an interaction term? Adding the interaction, the LRT is 0.37 $p = 0.54$ with 1 degree of freedom. The model does not seem to improve, and that the interaction term does not come out significant in this model serves as further evidence in favor of its exclusion. However, for substantive reasons, I think it is still informative to have in the model for reasons discussed below.

Examining the main-effects plus interaction model, with alcohol, cigarette, and alcohol+cigarette use predicting the use of marijuana, we find that the conditional odds ratio for marijuana use is 9.73, $p = 0.014$ for smokers vs non-smokers, such that smokers have 9.73 greater odds of marijuana use compared to non-sokers. This is large, but not nearly as large as the unconditional odds ratio. For alcohol, the conditional odds ratio is 13.46, $p < 0.001$ which is again quite sizable but not as large as the unconditional odds ratio. Importantly though, the odds ratio for the interaction is 1.8, $p = 0.53$. Because we fail to reject the null here, we continue to assume the null is true, thus that the odds ratio difference here is due to sampling variability. Despite the lack of statistical significance, this is still informative to have in the model because there may be practical reasons why you'd want to point out that adding smoking on top of drinking (or vice versa) doesn't really affect the odds of smoking marijuana. For example, an intervention that aims to prevent alcohol users from picking up smoking might not be worth the money, and that that money might be better spent on trying to prevent alcohol use in people who don't smoke and cigarette use in people who don't drink.

4.16

(a)

MBTI Dataset

	TF	JP	EI	SN	AlcYes	AlcNo
1	t	j	e	s	10	67
2	t	p	e	s	8	34
3	f	j	e	s	5	101
4	f	p	e	s	7	72
5	t	j	e	n	3	20
6	t	p	e	n	2	16
7	f	j	e	n	4	27
8	f	p	e	n	15	65
9	t	j	i	s	17	123
10	t	p	i	s	3	49
11	f	j	i	s	6	132
12	f	p	i	s	4	102
13	t	j	i	n	1	12
14	t	p	i	n	5	30
15	f	j	i	n	1	30
16	f	p	i	n	6	73

MBTI Model Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.1140	0.2715	-7.79	0.0000
TFt	0.6873	0.2206	3.12	0.0018
JPp	0.2022	0.2266	0.89	0.3721
EIi	-0.5550	0.2170	-2.56	0.0105
SNs	-0.4292	0.2340	-1.83	0.0666

The model formula is $\text{logit}(\hat{\pi}) = -2.11 - 0.56(E/I) - 0.43(S/N) + 0.69(T/F) + 0.2(J/P)$. The intercept is the ENFJ type; that is, when EI = E, SN = N, TF = F, and JP = J.

(b)

For an ESTJ,

$$\begin{aligned}\hat{\pi} &= \frac{e^{-2.11-0.2-0.43}}{(1+e^{-2.11-0.2-0.43})},\\ &= 0.13.\end{aligned}$$

(c)

The formula for ENTP would come out to be $\text{logit}(\hat{\pi}) = -2.11 + 0.69 + 0.2$. To get the probability,

$$\begin{aligned}\hat{\pi} &= \frac{e^{-2.11+0.69+0.2}}{(1+e^{-2.11+0.69+0.2})},\\ &= 0.23.\end{aligned}$$

4.17

(a)

In table 4.14, the intercept captures Introverted Feelers. To get $\hat{\pi}$ for introverted feelers::

$$\begin{aligned}\hat{\pi} &= \frac{e^{-2.8291}}{(1 + e^{-2.8291)}}, \\ &= 0.06.\end{aligned}$$

(b)

The estimated conditional odds for the EI variable is

$$e^{0.58} = 1.79.$$

(c)

To get CIs for the conditional odds ratio:

$$(e^{0.1589}, e^{1.0080}) = (1.17,\ 2.74).$$

(d)

We just need the inverse of part (b), so

$$\begin{aligned}\frac{1}{1.79} &= 0.56, \\ \text{CI} &= (\frac{1}{2.74}, \frac{1}{1.17}), \\ &= (0.36,\ 0.85).\end{aligned}$$

(e)

You could calculate the likelihood ratio test by taking $-2(\ell_0 - \ell_1)$, where ℓ is the model sans EI and ℓ_1 is the model with EI. This approximates the chi-squared, and luckily the output has already done this procedure for us. The LR statistic is 7.28, which has a p-value of 0.007. We conclude that EI has effect on the response above and beyond TF.

5.4

(a)

The deviance approximates the chi-square, so with a reported deviance of 11.1491 and 11 degrees of freedom, the p-value comes out to 0.43. Failure to reject the null indicates a lack of evidence for poor fit.

(b)

The JP would have to go because it's Chi-Squared value is the smallest of all predictors. It is also not significant, $p = 0.37$.

(c)

We can conduct a deviance test. Deviance of model minus deviance of model 2 with 6 degrees of freedom is $11.1491 - 3.74 = 7.41$, $p = 0.28$. Probably want the more parsimonious model without the interaction.

Appendix

4.3

(a): NA.

(b):

`0.7578 - 0.0694*(12).`

(c):

`exp(1)^(1.148 - 0.315 * 12) / (1 + exp(1)^(1.148 - 0.315 * 12)).`

4.5

```
shuttle <- as.data.frame(matrix(c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12,
                    13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23,

                    66, 70, 69, 68, 67, 72, 73, 70, 57, 63,
                    70, 78, 67, 53, 67, 75, 70, 81, 76, 79, 75, 76, 58,

                    0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0,
                    1 ,0 ,0 ,0, 0, 0, 0, 1, 0, 1),

                  ncol = 3, nrow=23))
colnames(shuttle) <- c("Ft", "Temp", "TD")
shuttle_mod <- glm(TD ~ Temp, family = binomial, data = shuttle)
```

(a):

The model comes out to be:

`round(coef(shuttle_mod)["(Intercept)"], 2)` + (-0.23)x.

(b):

`exp(1)^(15.04 - 0.23 * 31) / (1 + exp(1)^(15.04 - 0.23 * 31))`

(c): probability density function:

`(-0.2321627 * exp(1)^(15.0429016 - 0.2321627 * 64.79465)) / (1 + exp(1)^(15.0429016 - 0.2321627 * 64.79465))^2`

(d): NA

(e):

The Wald test is: `(-0.2321627/0.1082)^2.`

the p-value is `pchisq(4.6039476, df=1, lower.tail=FALSE)` w

The likelihood ratio test is:

`-2 * (logLik(glm(TD ~ 1, family = binomial, data = shuttle)) - logLik(shuttle_mod)).`

χ^2 p-value at 1 degree of freedom is:

`pchisq(7.95196, df=1, lower.tail=FALSE).`

A4.15

(a)

```r
merit <- array(c(24,9,47,12,
                 10,3,45,8,
                 5,4,57,9,
                 16,7,54,10,
                 7,4,59,12),
               dim = c(2,2,5),
               dimnames = list(
                 meritPay = c("yes", "no"),
                 race = c("black", "white"),
                 district = c("NC", "NE", "NW", "SE", "SW")))

cmh <- mantelhaen.test(merit, z=district, correct = FALSE, exact = FALSE)
```

The CMH test statistic comes out to be:

`cmh$statistic` with 1 df, and a p-value of `cmh$p.value`.

(b)

```r
#logistic regression model
merit_mod <- glm((yes/(yes + no)) ~ race ,
                 family = binomial,
                 weights = yes + no,
                 data = extension)
```

The wald test is $\left(\frac{\beta}{SE}\right)^2$, which comes out the be:

`((coef(merit_mod)[("raceWhites")] / (summary(merit_mod)$coefficients[2, 2]))^2)` with a p-value of `pchisq(8.4655366, df=1, lower.tail=FALSE)`.

(c)

Exponentiating the parameter for Whites, you get:

`exp((coef(merit_mod)[("raceWhites")]))`,

4.31

```r
maryj <- data.frame(c("Yes", "Yes", "No", "No"),
                              c("Yes", "No", "Yes", "No"),
                              c(as.integer(c(911, 44, 3, 2))),
                              c(as.integer(c(538, 456, 43, 279))))
colnames(maryj) <- c("AlcUse", "CigUse", "MaryjYes", "MaryjNo")

maryUse <- maryj$MaryjYes/(maryj$MaryjYes+maryj$MaryjNo)


mary_intercept <- glm(maryUse ~ 1,
                      weights = MaryjYes + MaryjNo,
                      family = binomial,
                      data = maryj)

mary_cig <- glm(MaryjYes/(MaryjYes+MaryjNo) ~ CigUse,
                weights = MaryjYes + MaryjNo,
                family = binomial,
                data = maryj)

mary_cig_alc <- glm(MaryjYes/(MaryjYes+MaryjNo) ~ CigUse + AlcUse,
                    weights = MaryjYes + MaryjNo,
                    family = binomial,
                    data = maryj)

mary_full <- glm(MaryjYes/(MaryjYes+MaryjNo) ~ CigUse * AlcUse,
                 weights = MaryjYes + MaryjNo,
                 family = binomial,
                 data = maryj)

# print good looking p-values
pvalr <- function(pvals, sig.limit = .001, digits = 3, html = FALSE) {

  roundr <- function(x, digits = 1) {
    res <- sprintf(paste0('%.', digits, 'f'), x)
    zzz <- paste0('0.', paste(rep('0', digits), collapse = ''))
    res[res == paste0('-', zzz)] <- zzz
    res
  }

  sapply(pvals, function(x, sig.limit) {
    if (x < sig.limit)
      if (html)
        return(sprintf('&lt; %s', format(sig.limit))) else
          return(sprintf('< %s', format(sig.limit)))
    if (x > .1)
      return(roundr(x, digits = 2)) else
        return(roundr(x, digits = digits))
  }, sig.limit = sig.limit)
}
```

\begin{doublespace}

The odds ratio of marijuana use for alcohol users vs non-users, ignoring whether or not they smoke, is: `round(((911+44)*(43+279)) / ((2+3) * (538 + 456)),2)`.

The odds ratio of marijuana use for smokers vs non-smokers, ignoring whether or not they drink, is: `round(((911+3)*(456+279)) / ((44+2) * (538 + 43)),2)`.

We should first test to see whether we gain anything from adding cigarettes to an intercept-only model. We get a LRT statistic: `round(-2 * (logLik(mary_intercept) - logLik(mary_cig)),2)`, *p* `pvalr(pchisq((-2 * (logLik(mary_intercept) - logLik(mary_cig))), df=1, lower.tail=FALSE))` with 1 degree of freedom.

But what about adding alcohol to the model? The LRT is:

`round(-2 * (logLik(mary_cig) - logLik(mary_cig_alc)),2)` p `pvalr(pchisq((-2 * (logLik(mary_cig) - logLik(mary_cig_alc))), df=1, lower.tail=FALSE))` with 1 degree of freedom.

Adding the interaction, the LRT is:

`round(-2 * (logLik(mary_cig_alc) - logLik(mary_full)),2)` $p =$ `pvalr(pchisq((-2 * (logLik(mary_cig_alc) - logLik(mary_full))), df=1, lower.tail=FALSE))` with 1 degree of freedom.

The conditional odds ratio for marijuana use is:

`round(exp((coef(mary_full)[("CigUseYes")])),2)`, $p =$ `pvalr(coef(summary(mary_full))[2,4], digits = 3)` for smokers vs non-smokers.

For alcohol, the conditional odds ratio is:

`round(exp((coef(mary_full)[("AlcUseYes")])),2)`, p `pvalr(coef(summary(mary_full))[3,4])`.

the odds ratio for the interaction is:

`round(exp((coef(mary_full)[("CigUseYes:AlcUseYes")])),2)`, $p =$ `pvalr(coef(summary(mary_full)`
`digits = 3)`.

\end{doublespace}

4.16

(a)

```
# dataset
myers <- data.frame(c("t", "t", "f", "f", "t", "t", "f", "f",
                      "t", "t", "f", "f", "t", "t", "f", "f"),

                    c("j", "p", "j", "p", "j", "p", "j", "p",
                      "j", "p", "j", "p", "j", "p", "j", "p"),

                    c("e", "e", "e", "e", "e", "e", "e", "e",
                      "i", "i", "i", "i", "i", "i", "i", "i"),

                    c("s", "s", "s", "s", "n", "n", "n", "n",
                      "s", "s", "s", "s", "n", "n", "n", "n"),

                    c(as.integer(c(10, 8, 5, 7, 3, 2, 4, 15,
                                   17, 3, 6, 4, 1, 5, 1, 6))),

                    c(as.integer(c(67, 34, 101, 72, 20, 16, 27,
                                   65, 123, 49, 132, 102, 12, 30, 30, 73))))
colnames(myers) <- c("TF", "JP", "EI", "SN", "AlcYes", "AlcNo")

myers_table <- print(xtable::xtable(myers,
                                    caption="MBTI Dataset",
                                    align = "ccccccc"),
                     caption.placement ="top")
```

MBTI Dataset

	TF	JP	EI	SN	AlcYes	AlcNo
1	t	j	e	s	10	67
2	t	p	e	s	8	34
3	f	j	e	s	5	101
4	f	p	e	s	7	72
5	t	j	e	n	3	20
6	t	p	e	n	2	16
7	f	j	e	n	4	27
8	f	p	e	n	15	65
9	t	j	i	s	17	123
10	t	p	i	s	3	49
11	f	j	i	s	6	132
12	f	p	i	s	4	102
13	t	j	i	n	1	12
14	t	p	i	n	5	30
15	f	j	i	n	1	30
16	f	p	i	n	6	73

```
# model
AlcUse <- myers$AlcYes/(myers$AlcYes + myers$AlcNo)
mbti_mod <- summary(glm(AlcUse ~ TF + JP + EI + SN,
                         weights = AlcYes + AlcNo,
                         family = binomial,
                         data = myers))

mbti_mod_table <- print(xtable::xtable(mbti_mod,
                                        caption="MBTI Model Output",
                                        align="lcccc"),
                         caption.placement ="top")
```

MBTI Model Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.1140	0.2715	-7.79	0.0000
TFt	0.6873	0.2206	3.12	0.0018
JPp	0.2022	0.2266	0.89	0.3721
EIi	-0.5550	0.2170	-2.56	0.0105
SNs	-0.4292	0.2340	-1.83	0.0666

```
# coefficients
intercept <- coef(mbti_mod)[("(Intercept)"),1]
EI <- coef(mbti_mod)[("EIi"),1]
SN <- coef(mbti_mod)[("SNs"),1]
TF <- coef(mbti_mod)[("TFt"),1]
JP <- coef(mbti_mod)[("JPp"),1]
```

The model formula is:
$\text{logit}(\hat{\pi}) = \text{‘}round(intercept, 2)\text{“}round(EI, 2)\text{‘}(E/I) - 0.43(S/N) + \text{‘}round(TF, 2)\text{‘}(T/F) + \text{‘}round((JP), 2)\text{‘}(J/P)$.

(b)

For an ESTJ,

$$\hat{\pi} = \frac{e^{`round(intercept,2)`-`round(JP,2)``round(SN,2)`}}{(1+e^{`round(intercept,2)`-`round(JP,2)``round(SN,2)`})},$$

$$= `round(exp(1)^{(}intercept - JP - SN)/(1 + round(exp(1)^{(}intercept - JP - SN), 2)), 2)`.$$

(c)

```
x <- round(intercept + TF + JP,2)
```

The formula for ENTP would come out to be:

$\text{logit}(\hat{\pi}) = `round((intercept), 2)` + `round(TF, 2)` + `round((JP), 2)`.$

To get the probability:

$$\hat{\pi} = \frac{e^{`round(intercept,2)`+0.69+`round(JP,2)`}}{(1+e^{`round(intercept,2)`+`rround(TF,2)`+`round(JP,2)`})},$$

$$= `round(exp(x)/(1 + exp(x)), 2)`.$$

4.17

(a) NA

(b) NA

(c) NA

(d) NA

(e)

`pvalr(pchisq(7.28, df=1, lower.tail=FALSE))`

5.4

(a)

`pvalr(pchisq(11.1491, df=11, lower.tail=FALSE))`

(b)

$p =$ `pvalr(pchisq(0.80, df=1, lower.tail=FALSE)).`

(c)

$11.1491 - 3.74 = `round(11.1491 - 3.74, 2)`$, $p =$ `pvalr(pchisq(7.4091, df=6, lower.tail=FALSE)).`