**37252 Regression and Linear Models**

**Lab 9: Simple Logistic Regression**

This lab is marked out of 26.

Please save your file in PDF format with name

**37252\_Lab9\_Surname\_FirstName**

**Due: 12 noon Wednesday 15 May 2024**

In this week’s lab we continue our example from last week. The data are available in **37252\_Lab9\_data.csv** which can be downloaded from Canvas.

The variables we consider are summarised in the table below.

|  |  |  |
| --- | --- | --- |
| **Name** | **Role** | **Description** |
| $$good$$ | response | successful field goal attempt: 1 (yes), 0 (no) |
| $$qtr$$ | predictor | game time quarter (1, 2, 3, 4) |
| $$distance$$ | predictor | kicking distance |

**Simple logistic regression with categorical variable**

Recall from Lab 8 that we found there was no statistically-significant relationship between $good$ and $qtr$. Let’s ignore this for the moment and fit a logistic regression model anyway.

As $qtr$ is a four-state categorical variable, we require three binary dummy variables. We will code them as

$$\left(qtr\_{1},qtr\_{2},qtr\_{3}\right)=\left\{\begin{array}{c}\left(1,0,0\right), qtr=1\\\left(0,1,0\right), qtr=2\\\begin{matrix}\left(0,0,1\right), qtr=3\\\left(0,0,0\right), qtr=4\end{matrix}\end{array}\right.$$

We need to specify this variable as a “factor” before fitting the model.

> read.csv("~/2024\_37252/Labs/Lab9/37252\_Lab9\_data.csv")

> NFLdat$qtr <- as.factor(NFLdat$qtr)

> NFLdat$qtr <- relevel(NFLdat$qtr, ref = "4")

> mod1 <- glm(good ~ qtr, family = "binomial", data = NFLdat)

> summary(mod1)

R output is displayed below.

Call:

glm(formula = good ~ qtr, family = "binomial", data = NFLdat)

Deviance Residuals:

 Min 1Q Median 3Q Max

-2.1330 0.4658 0.4914 0.5851 0.5851

Coefficients:

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

(Intercept) 2.0532 0.1972 10.410 <2e-16 \*\*\*

qtr1 0.1132 0.2993 0.378 0.705

qtr2 -0.3750 0.2429 -1.544 0.123

qtr3 -0.3485 0.2848 -1.224 0.221

---

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

(Dispersion parameter for binomial family taken to be 1)

 Null deviance: 810.25 on 1025 degrees of freedom

Residual deviance: 805.10 on 1022 degrees of freedom

AIC: 813.1

Number of Fisher Scoring iterations: 4

1. Write down the fitted logistic regression model in log-odds scale, odds scale and probability scale **[3 marks]**.

Recall in Lab 8 we used two-way table analysis to calculate

$$odds\left(qtr=1\right)=8.727$$

$$odds\left(qtr=4\right)=7.793$$

$$oddsRatio\left(qtr=1 v. qtr=4\right)=1.120$$

1. Using the regression model, carry-out the calculations to show that the odds of successful field goal in quarter 1 and in quarter 4 almost match those calculated in Lab 8 **[2 marks]**. Using the regression model (or R output) show that the odds ratio almost matches that from Lab 8 **[1 mark]**.
2. Determined if the model predicts a successful field goal for kicks taken in quarter1 and in quarter 4, i.e. determine if

$$\hat{good}\left(qtr=1\right)=1$$

and

$$\hat{good}\left(qtr=4\right)=1.$$

**[3 marks]**

1. Provide interpretations of $\hat{β}\_{qtr\_{1}}=0.113$ on the log-odds scale and of $e^{\hat{β}\_{qtr\_{1}}}=1.120$ on the odds scale **[2 marks]**.

**Simple logistic regression with continuous predictor**

Recall from Lab 8 that we found there was a statistically-significant relationship between $good$ and $distanceBin$, where $distanceBin$ represented kick distance quartile. Let’s infer from this the existence of a statistically-significant relationship between $good$ and $distance$ and build a logistic regression model.

> mod2 <- glm(good ~ distance, family = "binomial", data = NFLdat)

> summary(mod2)

R output is displayed below.

Call:

glm(formula = good ~ distance, family = "binomial", data = NFLdat)

Deviance Residuals:

 Min 1Q Median 3Q Max

-2.9500 0.2047 0.3491 0.5846 1.2341

Coefficients:

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

(Intercept) 6.75473 0.54691 12.351 <2e-16 \*\*\*

distance -0.12083 0.01234 -9.788 <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: 810.25 on 1025 degrees of freedom

Residual deviance: 680.53 on 1024 degrees of freedom

AIC: 684.53

Number of Fisher Scoring iterations: 6

1. Write down the fitted logistic regression model in log-odds scale, odds scale and probability scale **[3 marks]**.
2. Interpret the impact of $distance$ on the log-odds scale and the odds scale **[2 marks]**.
3. Perform a hypothesis test to determine if the regression is significant at the 0.05 level. Write down the hypotheses **[1 mark]**, the test statistic and p-value **[1 mark]**, the result of the test **[1 mark]** and a conclusion in non-mathematical language **[1 mark]**.
4. Use a scatterplot to plot $\hat{p}$ against $distance$ and describe the relationship **[2 marks]**.
5. Calculate the “median effective level” $distance\_{mel}$ by solving

$$\hat{p}\left(distance\right)={1}/{2}$$

for $distance$ **[2 marks]**. Describe the relationship between $distance\_{mel}$ and $\hat{good}$ **[2 marks]**.