Regression and Linear Models (37252)

Lecture 11 – Logistic Regression II

Lecturer: Joanna Wang
Notes adopted from Scott Alexander

School of Mathematical and Physical Sciences, UTS

Autumn 2024

Acknowledgement

These notes are based, in part, on earlier versions prepared by Dr Ed Lidums and Prof. James Brown.

11. Lecture outline

Topics:

- multiple logistic regression
  - log-likelihood function
- model fit
  - introduction
- Wald test
  - upper-tail rejection
  - two-tail rejection
  - lower-tail rejection
- omnibus test
  - null hypothesis rejection
- partial omnibus test
- Hosmer-Lemeshow test
- pseudo-R^2 statistics

11. Lecture outline

Topics (continued):

- model fit
  - summary
- model selection
  - summary
- model fit example
  - testing association
  - initial model
  - predictions
  - pseudo R^2
  - omnibus test
  - Hosmer-Lemeshow test
  - outliers (Pearson's residuals)
  - final model

11. Lecture outline

Topics (continued):

- model selection example
  - predictors
  - fit
  - interaction
  - final model
  - parameter interpretation

A very good text, which has been used as a reference in compiling these notes, is Hosmer and Lemeshow (2000).

11. Multiple logistic regression

Recall the population model

$$Y|\boldsymbol{x} = p(\boldsymbol{x}) + \epsilon$$

where $\boldsymbol{x} \in \mathbb{R}^m$, Y is a Bernoulli RV defined on $\{1, 0\}$ and

$$p(\boldsymbol{x}) := \text{Prob}(Y = 1|\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{x}].$$

Last lecture we began modelling p using multiple logistic regression and saw that our fitted models could variously be described in the log-odds scale

$$\log\left(\frac{\hat{p}(\boldsymbol{x})}{1-\hat{p}(\boldsymbol{x})}\right) = \hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_j,$$

the odds scale

$$\frac{\hat{p}(\boldsymbol{x})}{1-\hat{p}(\boldsymbol{x})} = e^{\hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_j}$$

and the probability scale

$$\hat{p}(\boldsymbol{x}) = \frac{1}{1 + e^{-\hat{\beta}_0 - \sum_{j=1}^{m} \hat{\beta}_j x_j}}.$$

11. Multiple logistic regression – log-likelihood function

The most common method for estimating the parameters $\beta_0, \ldots, \beta_m$ is by **maximum likelihood**.

This involves the **likelihood function**

$$\mathcal{L}(\beta_0, \ldots, \beta_m | (\boldsymbol{X}_i, Y_i)) = \prod_{i=1}^{n} p(\boldsymbol{X}_i)^{Y_i} \big(1 - p(\boldsymbol{X}_i)\big)^{1-Y_i} \tag{1}$$

calculated on the sample data $(\boldsymbol{X}_i, Y_i)$.

The parameter estimates are those that maximise **log-likelihood** and are the solution of the optimisation problem

$$(\hat{\beta}_0, \ldots, \hat{\beta}_m) = \underset{(\beta_0, \ldots, \beta_m)}{\operatorname{argmax}} \log \mathcal{L}(\beta_0, \ldots, \beta_m | (\boldsymbol{X}_i, Y_i)). \tag{2}$$

(It is often easier to maximise log-likelihood than likelihood.)

Details can be found in the Chapter 1 of Hosmer and Lemeshow (2000).

11. Model fit – introduction

In this lecture we turn our attention to assessing model fit.

We started this process last lecture, describing a test on the coefficients not dissimilar to the T-tests used in linear regression.

We continue this process today and introduce two more tests:

1. **omnibus test** - a test of overall significance of the regression, fulfilling the role of the F-test in linear regression
2. **Hosmer-Lemeshow test** - a test of the adequacy of the regression, a type of test we did not consider in linear regression.

We also describe two R^2-type statistics that provide a quantitative measure of model fit.

We begin by revisiting the **Wald test** in a more general form than previously described.

11. Wald test

Wald test hypotheses

$$H_0\colon\ \beta_j = \beta_j^*,$$

where β_j^* is some hypothesised value of β_j we hope to exclude.

The alternative hypothesis can be any of

$$H_A\colon\ \beta_j > \beta_j^*$$
$$H_A\colon\ \beta_j \neq \beta_j^*$$
$$H_A\colon\ \beta_j < \beta_j^*.$$

Test statistic

The test statistic $z^*_{\hat{\beta}_j}$ is the value that the RV

$$Z_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}} \overset{\text{approx.}}{\sim} \mathrm{N}(0,1)$$

takes for the test.

11. Wald test – upper-tail rejection

Rejection of null hypothesis (upper-tail test)

H_0 is **rejected** in favour of H_A at **significance level** α if

$$z^*_{\hat{\beta}_j} > z_{1-\alpha}$$

where $z_{1-\alpha}$ is the **quantile** satisfying

$$\text{Prob}\big(Z_{\hat{\beta}_j} > z_{1-\alpha}\big) = \alpha.$$

Equivalently, H_0 is rejected if β^*_j falls outside the one-sided $100(1-\alpha)\%$ CI for β_j given by

$$\hat{\beta}_j - S_{\hat{\beta}_j} z_{1-\alpha} \leq \beta_j < \infty$$

or if the **p-value**

$$p = \text{Prob}(Z_{\hat{\beta}_j} > z^*_{\hat{\beta}_j}) < \alpha.$$

The null hypothesis H_0 is **retained** in any other case.

11. Wald test – two-tail rejection

Rejection of null hypothesis (two-tail test)

H_0 is **rejected** in favour of H_A at **significance level** $0 < \alpha < \frac{1}{2}$ if

$$|z^*_{\hat{\beta}_j}| > z_{1-\alpha/2}$$

where $z_{1-\alpha/2}$ is the **quantile** satisfying

$$\text{Prob}\big(Z_{\hat{\beta}_j} > z_{1-\alpha/2}\big) = \alpha/2.$$

Equivalently, H_0 is rejected if β^*_j falls outside the two-sided $100(1-\alpha)\%$ CI for β_j given by

$$\hat{\beta}_j - S_{\hat{\beta}_j} z_{1-\alpha/2} \leq \beta_j \leq \hat{\beta}_j + S_{\hat{\beta}_j} z_{1-\alpha/2}$$

or if the **p-value**

$$p = 2 \times \text{Prob}(Z_{\hat{\beta}_j} > |z^*_{\hat{\beta}_j}|) < \alpha.$$

The null hypothesis H_0 is **retained** in any other case.

11. Wald test – lower-tail rejection

Rejection of null hypothesis (lower-tail test)

H_0 is **rejected** in favour of H_A at **significance level** α if

$$z^*_{\hat{\beta}_j} < z_\alpha$$

where z_α is the **quantile** satisfying

$$\text{Prob}\big(Z_{\hat{\beta}_j} < z_\alpha\big) = \alpha.$$

Equivalently, H_0 is rejected if β^*_j falls outside the one-sided $100(1-\alpha)\%$ CI for β_j given by

$$-\infty < \beta_j \leq \hat{\beta}_j + S_{\hat{\beta}_j} z_\alpha$$

or if the **p-value**

$$p = \text{Prob}(Z_{\hat{\beta}_j} < z^*_{\hat{\beta}_j}) < \alpha.$$

The null hypothesis H_0 is **retained** in any other case.

11. Omnibus test

The **omnibus test** allows the overall statistical significance of the logistic regression model to be assessed, and should be seen as an analogue of the F-test in linear regression.

Omnibus-test hypotheses

$$H_0\text{: } \beta_1 = \cdots = \beta_m = 0$$
$$H_A\text{: } \beta_j \neq 0 \text{ for at least one } j \in \{1, \ldots, m\}.$$

Test statistic

The test statistic l_m^* is the value that the RV

$$L_m = -2 \log \frac{\mathcal{L}(\beta_0)}{\mathcal{L}(\beta_0, \ldots, \beta_m)} \sim \chi^2(m)$$

takes for this test, where $\mathcal{L}$ is the likelihood function in (1).

11. Omnibus test – null hypothesis rejection

Rejection of null hypothesis

H_0 is **rejected** in favour of H_A at **significance level** α if

$$I_m^* > I_{1-\alpha}$$

where $I_{1-\alpha}$ is the **quantile** satisfying

$$\text{Prob}\big(L_m > I_{1-\alpha}\big) = \alpha.$$

Equivalently, H_0 is rejected if the **p-value**

$$p = \text{Prob}(L_m > I_m^*) < \alpha.$$

The null hypothesis H_0 is **retained** in any other case.

11. Partial omnibus test

As with the partial F-test in linear regression, there is a **partial omnibus test** in logistic regression.

Simple partition of the set parameters

Consider the partition of parameters $\{\beta_0, \ldots, \beta_q\}$ and $\{\beta_{q+1}, \ldots, \beta_m\}$.

By partition we mean

$$\{\beta_0, \ldots, \beta_q\} \bigcap \{\beta_{q+1}, \ldots, \beta_m\} = \emptyset$$

and

$$\{\beta_0, \ldots, \beta_q\} \bigcup \{\beta_{q+1}, \ldots, \beta_m\} = \{\beta_0, \ldots, \beta_m\}.$$

The partial omnibus test applied to such subsets of the slope parameters can be used to assess the statistical significance of the corresponding independent variables as a group.

11. Partial omnibus test

We describe the test for the second set of the partition.

Omnibus-test hypotheses

$$H_0\text{: } \beta_{q+1} = \cdots = \beta_m = 0$$
$$H_A\text{: } \beta_j \neq 0 \text{ for at least one } j \in \{q+1, \ldots, m\}.$$

Test statistic

The test statistic l^*_{m-q} is the value that the RV

$$L_{m-q} = -2 \log \frac{\mathcal{L}(\beta_0, \ldots, \beta_q)}{\mathcal{L}(\beta_0, \ldots, \beta_m)} \sim \chi^2(m-q) \tag{3}$$

takes for this test, where $\mathcal{L}$ is the likelihood function in (1).

The **rejection of the null hypothesis** occurs under the same conditions as in the full omnibus test.

11. Hosmer-Lemeshow test

The **Hosmer-Lemeshow test** is a statistical test for **goodness of fit** for logistic regression models.

Similar to a Chi-square goodness-of-fit test, it compares the observed values of the probabilities with the expected values and thereby determine if the model fits the data well.

The essential difference is the way observations are partitioned into the cells.

Whereas the Chi-square test allocates cell membership based on ranges of the independent variables, the Hosmer-Lemeshow test allocates cell membership based on ranges of the predicted probabilities.

11. Hosmer-Lemeshow test

Logstic regression models provide estimates of the probability of an outcome.

Based on these predicted probabilities, the sample is split into g groups (commonly $g = 10$).

For each group, the observed and expected number of successes and failure are obtained. They are then substituted into the Hosmer-Lemeshow test statistic.

Details of the test statistic can be found in Chapter 5 of Hosmer and Lemeshow (2000).

11. Hosmer-Lemeshow test

Hosmer-Lemeshow test hypotheses

H_0: the predicted probabilities match the observed probabilities

H_A: the predicted probabilities do not match the observed probabilities

We **reject the null hypothesis** using the $p < \alpha$ rule.

Interpretation

If the null is rejected then there is a statistically-significant difference in the predicted and observed probabilities as averaged over the observations in each subset.

From this we conclude the model has a poor fit with the data.

11. Pseudo R^2 statistics

In linear regression we use the statistic R^2 to quantify the proportion of total squared-variation in the data that is explained by the model.

The form of R^2 is a product of the decomposition of total sum of squares

$$SST = SSR + SSE$$

provided by ANOVA.

In logistic regression there is no such decomposition and hence no such equivalent statistic.

However, there are a variety of **pseudo R^2** statistics, including those developed by Cox and Snell and by Nagelkerke.

Calculation of these pseudo statistics uses the likelihood function of (1) in a way similar to its use in calculation of the test statistic (3) for the partial omnibus test.

We omit the details.

11. Model fit – summary

When assessing a model for fit there are two main questions.

1. Which factors are statistically significant?
   - use the omnibus test for overall significance of model
   - use Wald test for significance of individual predictors.

2. Given the factors, how "good" is the model?
   - use pseudo R^2 and classification table to assess fit with data
   - use Hosmer-Lemeshow test to judge whether further refinements may be necessary.

11. Model selection – summary

When selecting a model there are three main questions.

1. Is there a statistically significant association between response and predictor(s)?
   - visual inspection of plots
   - chi-square tests for categorical predictors, T-tests of correlation analysis for continuous predictors.
2. Are all "important" factors included in the model?
   - analysis of the modelling scenario – predictors should not be included solely for their statistical properties.
3. Does the model fit?

Although the statistical tools vary, the procedure of model selection in logistic regression is the same as the procedure in linear regression.

11. Model fit example

Last week we built a model for the probability of frequent GP visits with

$$GP = \begin{cases} 1, & \text{frequent (“success”)} \\ 0, & \text{infrequent (“failure”)} \end{cases}$$

as dummy response variable and

$$Gender = \begin{cases} 1, & \text{female} \\ 0, & \text{male} \end{cases}$$

as dummy predictor (data in *GPvisitExample.csv*, available on Canvas).

We defined $GP = 1$ (frequent GP visits) as “success” and $p = \text{Prob}(GP = 1)$.

11. Model fit example – testing association

Cross-tab data is reproduced below.

	datGP$GP		
datGP$Gender	0	1	Row Total
0	2146	764	2910
	1854.581	1055.419	
	45.792	80.466	
	0.737	0.263	0.468
	0.541	0.339	
	0.345	0.123	
1	1820	1493	3313
	2111.419	1201.581	
	40.222	70.678	
	0.549	0.451	0.532
	0.459	0.661	
	0.292	0.240	
Column Total	3966	2257	6223
	0.637	0.363	

11. Model fit example – testing association

With test statistic $\chi^2_1 = 237.157$, the null hypothesis of independence could be rejected.

```
Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  237.1566     d.f. =  1     p =  1.639466e-53

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  236.3435     d.f. =  1     p =  2.466067e-53
```

So a statistically-significant association between *Gender* and *GP* exists.

11. Model fit example – initial model

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.03279    0.04213  -24.52   <2e-16 ***
Gender       0.83474    0.05472   15.26   <2e-16 ***
```

which reports parameter estimates giving the fitted model

$$\hat{p}(Gender) = \frac{1}{1 + e^{1.033 - 0.835 \times Gender}}.$$

This week we examine the fit of this model and potential extensions with reference to R output.

11. Model fit example – predictions

The Classification table is a cross-tab of observed and predicted GP.

```
> class_table <- data.frame(observed = datGP$GP,
predicted = round(fitted(mod2),0))
> xtabs(~ observed + predicted, data = class_table)
        predicted
observed    0
       0 3966
       1 2257
```

The model predicts no frequent GP visits, obviously a major problem.

We can also see this from the fitted equation, with $\hat{p}(0) \approx 0.23$ and $\hat{p}(1) \approx 0.45$.

Note that

$$\hat{p} > 0.5 \Rightarrow \widehat{GP} = 1$$

and

$$\hat{p} < 0.5 \Rightarrow \widehat{GP} = 0.$$

11. Model fit example – pseudo R^2

We need a package to calculate both pseudo R^2 statistics mentioned earlier.

```
> -2*logLik(mod2)
'log Lik.' 7911.039 (df=2)
> library("DescTools")
> round(PseudoR2(mod2, c("CoxSnell", "Nagelkerke")),3)
  CoxSnell Nagelkerke
     0.038      0.052
```

Both R^2 figures show there is little explanatory power in the model.

The statistic **-2 Log likelihood** is calculation of $-2 \log \mathcal{L}(\beta_0, \beta_1)$, where $\log \mathcal{L}$ is the the quantity maximised when estimating model parameters (see 1-(2)).

Note that as model fit improves, $-2 \log \mathcal{L}$ decreases towards zero and increases otherwise.

11. Model fit example – omnibus test

Now the omnibus test of the hypotheses

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0.$$

```
> mod_null <-glm(GP ~ 1, family = "binomial", data = datGP)
> anova(mod_null, mod2, test = "LRT")
Analysis of Deviance Table

Model 1: GP ~ 1
Model 2: GP ~ Gender
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      6222     8151.5
2      6221     7911.0  1   240.45 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

With p-value close to zero we reject the null hypothesis that $\beta_1 = 0$, a result also obtainable using the Wald test on β_1.

11. Model fit example – omnibus test

The current model has a significant predictor in *Gender*, but the model is a poor fit given the pseudo R^2 statistics and its failure to predict $\hat{p}(\textit{Gender}) > 0.5$, associated with fitted value $\widehat{GP} = 1$.

We now add the independent variable *LogIncBin* (treated as continuous) representing the decile of log income.

```
> mod3 <- glm(GP ~ Gender + LogIncBin, family = "binomial", data = datGP)
> summary(mod3)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.549463   0.066616  -8.248   <2e-16 ***
Gender       0.806607   0.055113  14.635   <2e-16 ***
LogIncBin   -0.086710   0.009502  -9.126   <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: 8151.5  on 6222  degrees of freedom
Residual deviance: 7826.7  on 6220  degrees of freedom
AIC: 7832.7
```

11. Model fit example – omnibus test

```
> -2*logLik(mod3)
'log Lik.' 7826.672 (df=3)
> round(PseudoR2(mod3, c("CoxSnell", "Nagelkerke")),3)
  CoxSnell Nagelkerke
     0.051      0.070
> anova(mod_null, mod3, test = "LRT")
Analysis of Deviance Table

Model 1: GP ~ 1
Model 2: GP ~ Gender + LogIncBin
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      6222     8151.5
2      6220     7826.7  2   324.82 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
```

Pseudo R^2 has gone up slightly and both predictors are significant.

11. Model fit example – Hosmer-Lemeshow test

The biggest problem of the first model was its failure to predict $\hat{p} > 0.5$, a situation of poor model fit the Hosmer-Lemeshow test is deigned to detect.

```
> class_table3 <- data.frame(observed = datGP$GP, predicted
= round(fitted(mod3),0))
> xtabs(~ observed + predicted, data = class_table3)
        predicted
observed    0    1
       0 3598  368
       1 1884  373
```

The revised model now provides predictions where $\hat{p} > 0.5$, as can be inferred from the classification table above.

Is it possible to test if the model needs further improvement?

11. Model fit example – Hosmer-Lemeshow test

```
> library("ResourceSelection")
> h1 <- hoslem.test(datGP$GP, fitted(mod3), g=10)
> h1

Hosmer and Lemeshow goodness of fit (GOF) test

data:  datGP$GP, fitted(mod3)
X-squared = 24.053, df = 8, p-value = 0.002245
```

We reject the null hypothesis that the fitted probabilities adequately fit the data.

11. Model fit example – Hosmer-Lemeshow test

The contingency data used as the basis of the chi-square statistic is also reported.

```
> cbind(h1$observed,h1$expected)
                y0  y1    yhat0    yhat1
[0.195,0.209] 521 121 512.2068 129.7932
(0.209,0.239] 489 137 481.1095 144.8905
(0.239,0.29]  652 210 627.3068 234.6932
(0.29,0.327]  331 194 358.6355 166.3645
(0.327,0.352] 340 214 360.4729 193.5271
(0.352,0.393] 356 250 374.3084 231.6916
(0.393,0.435] 356 281 366.8522 270.1478
(0.435,0.478] 392 292 364.8157 319.1843
(0.478,0.521] 348 370 351.4767 366.5233
(0.521,0.543] 181 188 168.8154 200.1846
```

11. Model fit example – Hosmer-Lemeshow test

The Hosmer-Lemeshow test tells us the model needs improving if it is to have any utility. We add the interaction term

$$Gender\_LogIncBin = Gender \times LogIncBin$$

and the dummy variable

$$LogIncBin5 = \begin{cases} 1 & LogIncBin = 5 \\ 0 & LogIncBin \neq 5 \end{cases} .$$

Hosmer and Lemeshow goodness of fit (GOF) test

```
data:  datGP$GP, fitted(mod4)
X-squared = 5.8731, df = 8, p-value = 0.6614
```

Now we can retain the null hypothesis that the predicted probabilities adequately fit the data.

11. Model fit example – outliers (Pearson's residuals)

We can use the dependent variable data and the fitted probabilities to calculate **Pearson's Residuals**, which in terms of our current example is

$$r_i = \frac{GP_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}.$$

(To introduce this in a general sense would have required us to consider multiplicities of the covariate patterns, the same reason we did not described the Hosmer-Lemeshow test statistic.)

We can use these statistics to assess model fit, determining any observations where $|r_i| > 2$ as potential outliers.

As we know from OLS/GLS, outliers have the potential to have undue influence over the parameter estimates.

Many outliers is a sign of a poorly fitting model which, hopefully, we will also have been alerted to by other statistical tests (time does not permit detailed outlier analysis for logistic regression.)

11. Model fit example – final model

The parameter estimates for the final model are shown below.

```
> mod4 <- glm(GP ~ Gender + LogIncBin + Gender_LogIncBin + LogIncBin5, family
= "binomial", data = datGP)
> summary(mod4)
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.29915    0.09044  -3.308 0.000940 ***
Gender            0.44233    0.11589   3.817 0.000135 ***
LogIncBin        -0.12900    0.01503  -8.582  < 2e-16 ***
Gender_LogIncBin  0.06936    0.01938   3.580 0.000344 ***
LogIncBin5       -0.26346    0.09172  -2.872 0.004073 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8151.5  on 6222  degrees of freedom
Residual deviance: 7805.5  on 6218  degrees of freedom
AIC: 7815.5
```

All independent variables and the interaction term are significant.

11. Model fit example – final model

...and quantitative assessment of the model's fit

```
> -2*logLik(mod4)
'log Lik.' 7805.462 (df=5)
> round(PseudoR2(mod4, c("CoxSnell", "Nagelkerke")),3)
  CoxSnell Nagelkerke
     0.054      0.074
```

11. Model selection example

Now we look at model selection using the **forward selection** method.

Recall we have already used this technique with OLS/GLS, and the rules remain the same. However, there are flavours of the technique when it comes to binary logistic regression, and these relate to the statistic used to judge inclusion/exclusion.

We use "Forward: AIC", which is forward selection method involving the Akaike Information Criterion (AIC). Other varieties using backward or stepwise selection can also be used.

11. Model selection example – predictors

Using the same data set, we start with *Gender*, *LogIncBin* and *Health* as potential predictors, where $Health \in \{1, \ldots, 5\}$, $Health = 1$ representing the highest state of self-reported health and $Health = 5$ taken as reference category.

```
> datGP$Gender <- as.factor(datGP$Gender)
> datGP$Health <- as.factor(datGP$Health)
> datGP$Health <- relevel(datGP$Health, ref = "5")
> datGP$Gender_Health <- as.factor(datGP$Gender_Health)
> min.model = glm(GP ~ 1, family = "binomial", data = datGP)
> max.model <- formula(glm(GP ~ Health + Gender+ LogIncBin ,
family = "binomial", data = datGP))
```

11. Model selection example – predictors

```
> library("stats")
> fwd.model <- step(min.model, direction='forward', scope=max.model)
Start:  AIC=8153.49
GP ~ 1

            Df Deviance    AIC
+ Health     4   6915.8 6925.8
+ Gender     1   7911.0 7915.0
+ LogIncBin  1   8047.2 8051.2
<none>           8151.5 8153.5

Step:  AIC=6925.79
GP ~ Health

            Df Deviance    AIC
+ Gender     1   6707.3 6719.3
+ LogIncBin  1   6895.2 6907.2
<none>           6915.8 6925.8
```

11. Model selection example – predictors

```
Step:  AIC=6719.3
GP ~ Health + Gender

             Df Deviance    AIC
+ LogIncBin   1   6694.0 6708.0
<none>            6707.3 6719.3

Step:  AIC=6708.05
GP ~ Health + Gender + LogIncBin
```

11. Model selection example – predictors

```
> summary(fwd.model)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.20174    0.43094   5.109 3.24e-07 ***
Health1     -4.40431    0.43482 -10.129  < 2e-16 ***
Health2     -3.26076    0.42954  -7.591 3.17e-14 ***
Health3     -2.05787    0.43142  -4.770 1.84e-06 ***
Health4     -0.77235    0.45418  -1.701 0.089034 .
Gender1      0.85205    0.06112  13.941  < 2e-16 ***
LogIncBin   -0.03858    0.01060  -3.640 0.000272 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8151.5  on 6222  degrees of freedom
Residual deviance: 6694.0  on 6216  degrees of freedom
AIC: 6708
```

11. Model selection example – fit

Having added all significant variables, we move on to analysis of model fit.

```
> -2*logLik(fwd.model)
'log Lik.' 6694.046 (df=7)
> round(PseudoR2(fwd.model, c("CoxSnell", "Nagelkerke")),3)
  CoxSnell Nagelkerke
     0.209      0.286
> hoslem.test(datGP$GP, fitted(fwd.model), g=10)
Hosmer and Lemeshow goodness of fit (GOF) test

data:  datGP$GP, fitted(fwd.model)
X-squared = 13.356, df = 8, p-value = 0.1002
```

P-value of the Hosmer-Lemeshow test does not reject the null hypothesis.

The reported statistics provide a guide for the modelling process sequence:

1. build main effects model
2. check Hosmer-Lemeshow test results
3. add interaction terms if required by result in 2.

11. Model selection example – interaction

We consider adding some interaction terms may be useful we add these to the model.

```
> mod5 <- glm(GP ~ Health + Gender+ LogIncBin + Gender_LogIncBin + Gender_Healt
family = "binomial", data = datGP)
> summary(mod5)
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.22403    0.54400   4.088 4.34e-05 ***
Health1          -4.20219    0.55529  -7.568 3.80e-14 ***
Health2          -2.97367    0.54289  -5.477 4.32e-08 ***
Health3          -1.93283    0.54562  -3.542 0.000396 ***
Health4          -0.70766    0.57864  -1.223 0.221341
Gender1           0.94513    0.90623   1.043 0.296983
LogIncBin        -0.08247    0.01653  -4.989 6.05e-07 ***
Gender_LogIncBin  0.07482    0.02157   3.468 0.000524 ***
Gender_Health1   -0.47506    0.91627  -0.518 0.604130
Gender_Health2   -0.61138    0.90505  -0.676 0.499342
Gender_Health3   -0.31726    0.90887  -0.349 0.727037
Gender_Health4   -0.18988    0.95345  -0.199 0.842144
Gender_Health5         NA         NA      NA       NA
---
    Null deviance: 8151.5  on 6222  degrees of freedom
Residual deviance: 6677.8  on 6211  degrees of freedom
AIC: 6701.8
```

11. Model selection example – interaction

```
> mod5_Gender_LogIncBin <- glm(GP ~ Health + Gender + LogIncBin + Gender_Health, family
= "binomial", data = datGP)
> anova(mod5_Gender_LogIncBin, mod5, test = "LRT")
Analysis of Deviance Table

Model 1: GP ~ Health + Gender + LogIncBin + Gender_Health
Model 2: GP ~ Health + Gender + LogIncBin + Gender_LogIncBin + Gender_Health
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      6212     6689.9
2      6211     6677.8  1   12.061 0.0005148 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> mod5_Gender_Health <- glm(GP ~ Health + Gender + LogIncBin + Gender_LogIncBin, family
= "binomial", data = datGP)
> anova(mod5_Gender_Health, mod5, test = "LRT")
Analysis of Deviance Table

Model 1: GP ~ Health + Gender + LogIncBin + Gender_LogIncBin
Model 2: GP ~ Health + Gender + LogIncBin + Gender_LogIncBin + Gender_Health
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      6215     6683.3
2      6211     6677.8  4    5.507   0.2391
```

While the interaction term for *Gender* $\times$ *LogIncBin* is significant, that for *Gender* $\times$ *Health* is not and can be removed from the model.

11. Model selection example – final model

Doing so and re-running the model generates the following output.

```
> mod6 <- glm(GP ~ Health + Gender+ LogIncBin + Gender_LogIncBin, family =
"binomial", data = datGP)
> summary(mod6)
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       2.41342    0.43609   5.534 3.13e-08 ***
Health1          -4.39041    0.43484 -10.097  < 2e-16 ***
Health2          -3.24659    0.42954  -7.558 4.09e-14 ***
Health3          -2.04247    0.43146  -4.734 2.20e-06 ***
Health4          -0.76238    0.45423  -1.678 0.093266 .
Gender1           0.47575    0.12978   3.666 0.000246 ***
LogIncBin        -0.08029    0.01665  -4.821 1.43e-06 ***
Gender_LogIncBin  0.07019    0.02149   3.266 0.001090 **
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8151.5  on 6222  degrees of freedom
Residual deviance: 6683.3  on 6215  degrees of freedom
AIC: 6699.3
```

We see that all terms are now significant.

11. Model selection example – final model

```
> -2*logLik(mod6)
'log Lik.' 6683.339 (df=8)
> round(PseudoR2(mod6, c("CoxSnell", "Nagelkerke")),3)
  CoxSnell Nagelkerke
     0.210      0.288
> hoslem.test(datGP$GP, fitted(mod6), g=10)

Hosmer and Lemeshow goodness of fit (GOF) test

data:  datGP$GP, fitted(mod6)
X-squared = 9.2781, df = 8, p-value = 0.3194
```

We also see that the pseudo R^2 statistics have improved as a result.

11. Model selection example – final model

The model we have just fitted on the log-odds scale is

$$\begin{aligned}\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 2.413 &+ 0.476\,Gender - 0.08\,LogIncBin \\ &-4.39h_1 - 3.247h_2 - 2.042h_3 - 0.762h_4 \\ &+0.07\,Gender \times LogIncBin\end{aligned}$$

where h_i is dummy variable for $Health = i$, $i \in \{1, \ldots, 4\}$ ($Health = 5$ is reference category).

By taking the exponential of the last we have the fitted model on the odds scale.

Solving the log-odds (or odds) equation for $\hat{p}$ gives the fitted model on the probability scale.

11. Model selection example – parameter interpretation

As detailed in Lecture 10, we can provide physical interpretations for the parameters.

For example, -4.39 is the predicted difference in log-odds and $e^{-4.39} = 0.012$ the predicted multiple of the odds (odds ratio) of frequent GP visits for those with $Health = 1$ compared to those with $Heath = 5$.

For variables appearing in interaction terms, a more nuanced interpretation is required.

On the log-odds scale, 0.07 is the predicted difference in the change in log-odds for a unit increase in *LogIncBin* of females compared to males.

References I

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. 2nd edition.