STAT3015/4030/7030:

Generalised Linear Modelling

Logistic regression for binary response variables

Semester 2 2017

Originally prepared by Bronwyn Loong

References

Ch 20, 21 - Ramsey and Schafer, Statistical Sleuth

Ch 2 - Faraway, Extending the linear model with R

Ch 5 - Gelman and Hill

Case Study I

A 1972-1981 health survey in The Hague, Netherlands discovered an association between keeping pet birds and increased risk of lung cancer. Researchers conducted a case-control study in 1985 at four hospitals in The Hague (population 450,000). They identified 49 cases of lung cancer among patients who were registered with a general practice, who were age 65 or younger, and who had resided in the city since 1965. They also selected 98 controls from a population of residents with the same general age structure. Data were gathered on sex, age, socio-economic status, years of smoking, average rate of smoking and a birdkeeping indicator.

Q: Age and smoking history are both known to be associated with lung cancer incidence. After age, socioeconomic status, and smoking have been controlled for, is there an additional risk associated with birdkeeping?

Case Study I

From our observed data, what is our response variable? What do we want to do with the response variable to answer the research question of interest?

Can we use classical linear regression to model our response? Why or why not?

What are the modelling alternatives to the classical linear regression model?

The logistic regression model

In the birdkeeping case study, the response variable (lung cancer) is binary, meaning it can take on values either 0 or 1. We also collected several other explanatory variables.

The mean of our binary variable is the probability of developing lung cancer. We want to model the probability of developing lung cancer as a function of the explanatory variables.

LOGISTIC REGRESSION enables us to do this.

Logistic regression as a generalised linear model

DEFINITION:

A generalised linear model (GLM) is a probability model in which the mean of a response variable is related to explanatory variables through a regression equation.

Let $\mu = \mu\{Y|X_1, ..., X_p\}$ denote the mean response.

For some specified **link** function $\mathbf{g}(\mu)$, we relate the mean response to a linear function of the covariates $X_1, ..., X_p$.

$$g(\mu) = \eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

for unknown parameters $(\beta_0, \beta_1, ..., \beta_p)$. η is called the linear predictor.

Q: what is the link function in classical linear regression?

Logistic regression as a generalised linear model

Q: What is an appropriate link function for binary response data? What considerations do we need to take into account when choosing the link function $g(\mu)$??

Logistic regression as a generalised linear model

Q: What is an appropriate link function for binary response data? What considerations do we need to take into account when choosing the link function $g(\mu)$??

Binary data:

$$Y = \begin{cases} 1 & \text{with probability} = p \\ 0 & \text{with probability} = (1-p) \end{cases}$$

Requirements for g

- g needs to be monotone (order of p's needs to be preserved)
- $0 \leq g^{-1}(\eta) \leq 1$ for any η - WHY??

Logistic regression as a generalised linear model

There are three common choices:

1. Logit: $\eta = g(\mu) = \log(p/(1-p))$
2. Probit $\eta = g(\mu) = \Phi^{-1}(p)$ where Φ is the standard normal cumulative distribution function
3. Complementary log-log: $\eta = g(\mu) = \log(-\log(1-p))$

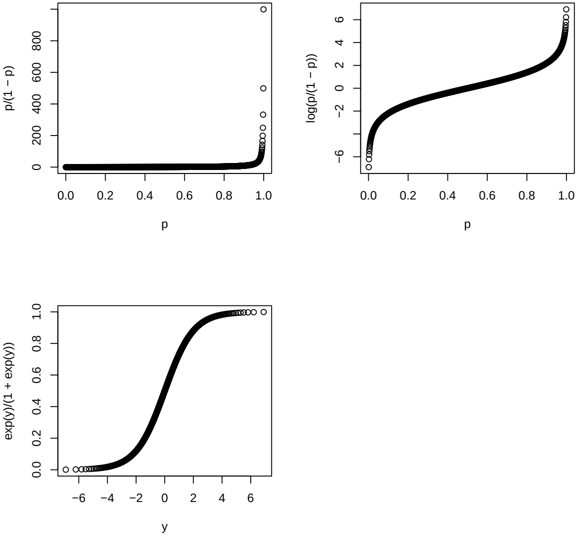
1. Logit link: $g(\mu) = logit(p) = \eta = \beta_0 + \beta_1 X_1 + ...\beta_p X_p$

The inverse of the logit function is called the **logistic function** (or inverse logit).

$$p = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

$$invlogit(\eta) = p$$

Logistic regression as a generalised linear model



Logistic regression as a generalised linear model

The inverse logit function transforms continuous values to the range (0,1) which is necessary since probabilities must be between 0 and 1.

Conversely the logit function maps the range (0,1) to the range $(-\infty, \infty)$.

The slope of the inverse-logit function is curved → the expected difference in p corresponding to a fixed difference in x is not constant.

- Where is the curve steepest? What is the slope of the curve at this point?
- Numerical example: we can calculate logit(0.5)=0 and logit (0.6)=0.405. We can also calculate logit(0.9)=2.2 and logit(0.93)=2.6. Interpret and compare these two pairs of results.

Logistic regression as a generalised linear model

The inverse logit function transforms continuous values to the range (0,1) which is necessary since probabilities must be between 0 and 1.

Conversely the logit function maps the range (0,1) to the range $(-\infty, \infty)$.

The slope of the inverse-logit function is curved $\rightarrow$ the expected difference in p corresponding to a fixed difference in x is not constant.

- Where is the curve steepest? What is the slope of the curve at this point?
- Numerical example: we can calculate logit(0.5)=0 and logit (0.6)=0.405. We can also calculate logit(0.9)=2.2 and logit(0.93)=2.6. Interpret and compare these two pairs of results.

Changes on the logit scale are compressed at the ends of the probability scale - WHY??

Logistic regression as a generalised linear model

Non-constant variance
What did we assume about the variance of error terms in classical linear regression??

Logistic regression as a generalised linear model

Non-constant variance
What did we assume about the variance of error terms in classical linear regression?? $Var(Y|X_1, .., X_p) = \sigma^2$.

For logistic regression, what is our variance function?

Logistic regression as a generalised linear model

Non-constant variance

What did we assume about the variance of error terms in classical linear regression?? $Var(Y|X_1, .., X_p) = \sigma^2$.

For logistic regression, what is our variance function?

$\mathbf{Var(Y|X_1, .., X_p) = p(1 - p)}$

$\rightarrow$ the variance function is a function of the mean, there are no additional parameters like σ^2

Logistic regression as a generalised linear model

Recap: Logistic regression model
Binary data:

$$Y = \begin{cases} 1 & \text{with probability} = p \\ 0 & \text{with probability} = (1-p) \end{cases}$$

$$\mu(Y|X_1, .., X_p) = p$$

$$g(p) = logit(p) = \log(p/(1-p)) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p = X^T\beta$$

$$Var(Y|X_1, .., X_p) = p(1-p)$$

GLM notes

(we will come back to these points later)

- The link function is one of the central ideas of GLMs. Its purpose is to link the mean of the response to a **linear** function of the predictors. The function g(μ) is linear in the unknown parameters $\beta_0, \beta_1, ..., \beta_p$.
- The mean response $\mu\{Y|X_1, ..., X_p\} = g^{-1}(\eta)$ is *non-linear* in the unknown parameters.
- Other types of response can be modelled, not just binary. And the linear predictor can accommodate qualitative and quantitative predictors, transformations of these predictors and combinations of these predictors. Hence the term *generalised* linear models.

Interpretation of logistic regression coefficients

The logit is the log odds function. If $\text{Odds} = p/(1-p) = 1$, what does this imply?

Interpretation of logistic regression coefficients

The logit is the log odds function. If Odds $= p/(1-p) = 1$, what does this imply?

How do we relate the β's to our binary response??

$$\text{Odds of Y=1} : \exp(\beta_0 + \beta_1 X_1 + ...\beta_p X_p)$$

What about when our predictor values change?

Interpretation of logistic regression coefficients

The logit is the log odds function. If Odds $= p/(1-p) = 1$, what does this imply?

How do we relate the β's to our binary response??

$$\text{Odds of Y=1} : \exp(\beta_0 + \beta_1 X_1 + ...\beta_p X_p)$$

What about when our predictor values change? The ratio of the odds at $X_1 = B$ relative to the odds at $X_1 = A$ for fixed values of the other X's is

$$\exp(\beta_1(B - A))$$

In particular, if X_1 increases by one unit, the odds of Y=1 will change by a **multiplicative** factor of $\exp(\beta_1)$, other variables being the same.

Interpretation of logistic regression coefficients

Example: Suppose $\beta_1 = 0.2$, then there is a multiplicative change of $\exp(0.2) = 1.22$ in the odds of $Y = 1$ per unit increase in X_1.

Suppose the odds of $Y = 1$ change from 1 to 1.22. What is the equivalent change on the probability scale?

Case Study I

Model: $logit(p) = \beta_0 + \beta_1 AG + \beta_2 BK$

Variable	Coef.	Std. error
Constant	-2.32950	1.53189
AG	0.01615	0.02569
BK	1.40271	0.38004

What is the effect of birdkeeping on the odds of lung cancer?
Construct a 95% confidence interval for this estimate?

Estimation of logistic regression coefficients

In classical linear regression, we used the method of least squares to obtain parameter estimates. That is we found $\hat{\beta}$ to minimise the function:

$$d(\hat{\beta}) = \sum_{i=1}^{n} (Y_i - X_i^T \hat{\beta})^2$$

Linear regression assumes linearity of the model, independence of the errors, and normally distributed and additive error terms.

That is, $(Y_i - X_i^T \hat{\beta})$ is the error, and we want to minimise the sum of squared errors (acoss all data points)

For the logistic regression, do we have linearity? Do we have independence of the error terms? Are our error terms normally distributed? Are they additive?

Estimation of logistic regression coefficients

In classical linear regression, we used the method of least squares to obtain parameter estimates. That is we found $\hat{\beta}$ to minimise the function:

$$d(\hat{\beta}) = \sum_{i=1}^{n} (Y_i - X_i^T \hat{\beta})^2$$

Linear regression assumes linearity of the model, independence of the errors, and normally distributed and additive error terms.

That is, $(Y_i - X_i^T \hat{\beta})$ is the error, and we want to minimise the sum of squared errors (acoss all data points)

For the logistic regression, do we have linearity? Do we have independence of the error terms? Are our error terms normally distributed? Are they additive?
$\rightarrow$ need to seek alternative estimation methods.

Maximum likelihood estimation

Recall, the likelihood is a function of the parameter(s) given the data. The *maximum likelihood estimate* (MLE) is the value of the parameter(s) that gives the largest probability to the observed data, that is, maximises the likelihood function. For mathematical ease, we generally maximise the log of the likelihood function.

Maximum likelihood estimation

Recall, the likelihood is a function of the parameter(s) given the data. The *maximum likelihood estimate* (MLE) is the value of the parameter(s) that gives the largest probability to the observed data, that is, maximises the likelihood function. For mathematical ease, we generally maximise the log of the likelihood function.

Binary data: for a single binary response Y, the probability model is:

$$Pr(Y = y) = p^y (1 - p)^{1-y}$$

For n such responses, and if the responses are independent, then the probability model is:

$$Pr(Y_1 = y_1, ..., Y_n = y_n) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

Maximum likelihood estimation

So for logistic regression (that is, binary response and log-link), the likelihood function is:

$$L(\beta|y,X) = \prod_{i=1}^{n} (\mathrm{invlogit}(X_i\beta))^{y_i}(1-\mathrm{invlogit}(X_i\beta))^{1-y_i}$$

The log-likelihood function is:

$$\begin{aligned}\log L(\beta|y,X) &= \sum_{i=1}^{n} y_i \log((\mathrm{invlogit}(X_i\beta))) + (1-y_i)\log((1-\mathrm{invlogit}(X_i\beta))) \\ &= \sum_{i=1}^{n} y_i\eta_i - \log(1+\exp(\eta_i))\end{aligned}$$

Maximum likelihood estimation

To find $\hat{\beta}_{MLE}$, compute the derivative, $\frac{\partial \log L(\beta|y,X)}{\partial\beta}$, equate to zero, and solve for β.

We can also quantify the uncertainty in our MLEs via the Fisher information matrix:

$$\mathscr{I}(\beta) = -E\left[\frac{\partial \log L^2(\beta|y,X)}{\partial\beta\partial\beta^T}\right]$$

and

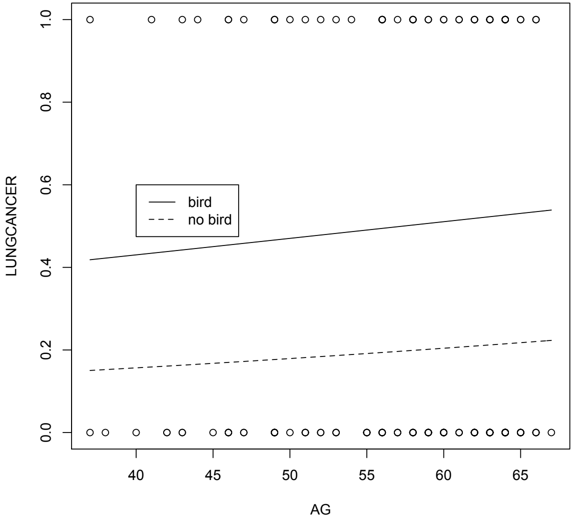
$$\widehat{Var}(\hat{\beta}) = \mathscr{I}^{-1}(\hat{\beta})$$

Case Study I

(R code)

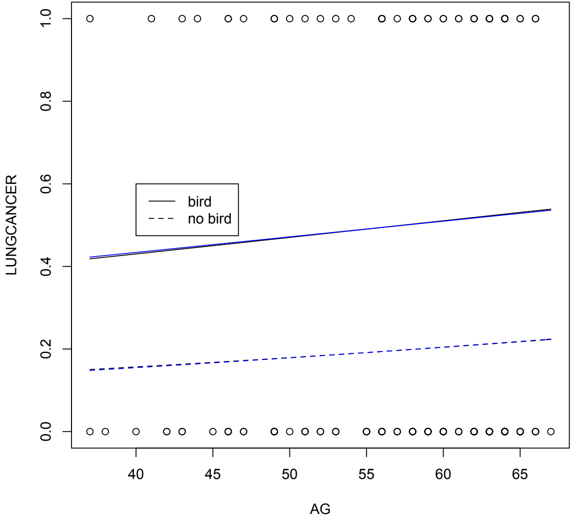
```
data_bk<-read.table("bird.csv",header=T,sep=",")
attach(data_bk)
names(data_bk)
y<-ifelse(LC=="LUNGCANCER",1,0)
bk<-ifelse(BK=="BIRD",1,0)
m1<-glm(y~AG+bk, family=binomial,data_bk)
summary(m1)
plot(AG,y,ylab="LUNGCANCER")
curve(invlogit(cbind(1,x,1)%*%coef(m1)),add=TRUE,lty=1)
curve(invlogit(cbind(1,x,0)%*%coef(m1)),add=TRUE,lty=2)
legend(40,0.6,c("bird","no bird"),lty=c(1,2))
```

Case Study I



Case Study I

(blue lines - probit link)



Inference

Is birdkeeping statistically significant? How many predictors should I include in my model? How can I assess the adequacy of my model fit??

Tests and Confidence Intervals for Single Coefficients

By MLE theory

$$\hat{\beta}_j \underset{\cdot}{\sim} (\beta_j, SE(\hat{\beta}_j))$$

A test based on this approximate normality of maximum likelihood estimates is referred to as **Wald's test**.

Do we need to use t-quantiles? No, t-theory applies to normally distributed response variables.

Tests and Confidence Intervals for Single Coefficients

Model: $logit(p) = \beta_0 + \beta_1 AG + \beta_2 BK$ Example:

Variable	Coef.	Std. error
Constant	-2.32950	1.53189
AG	0.01615	0.02569
BK	1.40271	0.38004

Is birdkeeping statistically significant?

Tests and Confidence Intervals for Single Coefficients

Model: $logit(p) = \beta_0 + \beta_1 AG + \beta_2 BK$ Example:

Variable	Coef.	Std. error
Constant	-2.32950	1.53189
AG	0.01615	0.02569
BK	1.40271	0.38004

Is birdkeeping statistically significant?

We want to test $H_0 : \beta_2 = 0$ vs $H_A : \beta_2 \neq 0$. The z-statistic $=$ $\frac{1.40271}{0.38004}$=3.691. The two-sided p-value =$2Pr(Z > 3.691) = 0.000223$.

Tests and Confidence Intervals for Single Coefficients

Construct a 95% confidence interval for the odds of lung cancer for birdkeepers relative to non-bird keepers.

A 95% confidence interval estimate for the coefficient of BK in the logistic regression model is

$$1.40271 \pm 1.96 \times 0.38004 = 0.6578 \text{ to } 2.1476$$

Taking antilogarithms the 95% confidence interval estimate for the odds of lung cancer for birdkeepers relative to the odds of lung cancer for non birdkeepers is (exp(0.6578) , exp(2.1476)) $=$ $(1.93, 8.56)$

Tests and Confidence Intervals for Single Coefficients

Using the birdkeeping model fit above, what are the odds of lung cancer for a 60 year old relative to a 45 year old?

Tests and Confidence Intervals for Single Coefficients

Using the birdkeeping model fit above, what are the odds of lung cancer for a 60 year old relative to a 45 year old?

The log-odds change by 0.01615 for each extra year of age. A 95% confidence interval estimate is (-0.034,0.067)

Therefore, for 15 years extra age, the log-odds change by $15 \times 0.01615 = 0.24225$. A 95% confidence interval for the change in log-odds of lung cancer for an additional 15 years of age is $(15 \times -0.034, 15 \times 0.067) = (-0.51, 1.00)$.

Taking anti-logarithms: It is estimated that the odds of lung cancer for a 60 year old are 1.2741 times greater than the odds for a 45 year old (95% CI estimate: (0.60,2.71))

Comparing full and reduced models

With linear regression we used the extra-sums-of squares F-test and calculated R^2 to assess model fit.

Can we do the same here??

Comparing full and reduced models

With linear regression we used the extra-sums-of squares F-test and calculated R^2 to assess model fit.

Can we do the same here??

Our error terms are not additive as in linear regression - squared error is not the mathematically optimal measure of model error. (same reason we did not use least squares to estimate parameters in the logistic regression).

Comparing full and reduced models

The Likelihood Ratio Test

(similar to extra-sum-of squares F-test)

$$LRT = 2(\log(LMAX_{full}) - \log(LMAX_{reduced}))$$

where LMAX is the likelihood of the parameters evaluated at its maximum for the relevant model.

Properties:

When the reduced model (and therefore the null hypothesis) is correct, the LRT has approximately a chi-square distribution with degrees of freedom = difference in no. of parameters between full and reduced model.

When would you reject the null (reduced model) in favour of the alternative (full model)??

Comparing full and reduced models

The Drop in Deviance Test

In GLMS, we can calculate a quantity known as *deviance* (analagous to residual standard deviation in regression)

We will learn more about deviance later, for now, note that deviance is measure of error, lower deviance means a better fit to the data.

$$\text{deviance} = \text{Constant} - 2\log(\text{LMAX})$$

It follows that:

$$\text{LRT} = \text{deviance}_{\text{reduced}} - \text{deviance}_{\text{full}}$$

Comparing full and reduced models

The Drop in Deviance Test

- Also note, if a predictor is added that is simply random noise, we expect deviance to decrease by 1 on average.
- When an informative predictor is added, we expect deviance to decrease by more than 1.
- When d predictors are added to a model, we expect deviance to decrease by more than d.

Suppose the full model has d extra parameters. We compare the LRT to a χ^2_d distribution (which has mean d and standard deviation $\sqrt{2d}$). So if the drop-in-deviance is far in excess of d, we can reject the null hypothesis.

Case Study I

A drop-in-deviance test for the association of the odds of lung cancer with birdkeeping, after accounting for other factors.

Full model:

$$logit(p) = \beta_0 + \beta_1 FM + \beta_2 AG + \beta_3 SS + \beta_4 YR + \beta_5 BK$$

```
> m3<-glm(y∼SEX+AG+SS+YR+BK, family=binomial,data_bk)
> summary(m3)
..........
Null deviance: 187.14 on 146 degrees of freedom
Residual deviance: 155.24 on 141 degrees of freedom
```

Residual deviance is the deviance for the current model

Null deviance is the deviance for a model with no predictors and just an intercept term.

Case Study I

Reduced model:

$$logit(p) = \beta_0 + \beta_1 FM + \beta_2 AG + \beta_3 SS + \beta_4 YR$$

```
> m4<-glm(y~SEX+AG+SS+YR, family=binomial,data_bk)
> summary(m4)
Null deviance: 187.14 on 146 degrees of freedom
Residual deviance: 166.53 on 142 degrees of freedom

> deviance(m4)-deviance(m3)
11.29165
> df.residual(m4)-df.residual(m3)
1
> pchisq(deviance(m4)-deviance(m3),
df.residual(m4)-df.residual(m3),lower=FALSE)
0.0007785639
```

Conclusion??

Drop in Deviance test

We can also test the significance of a single term using the drop in deviance test.

NOTE: This is not the same as Wald's test for a single coefficient (we can show they are not mathematically equivalent). The drop-in-deviance test is generally more reliable.

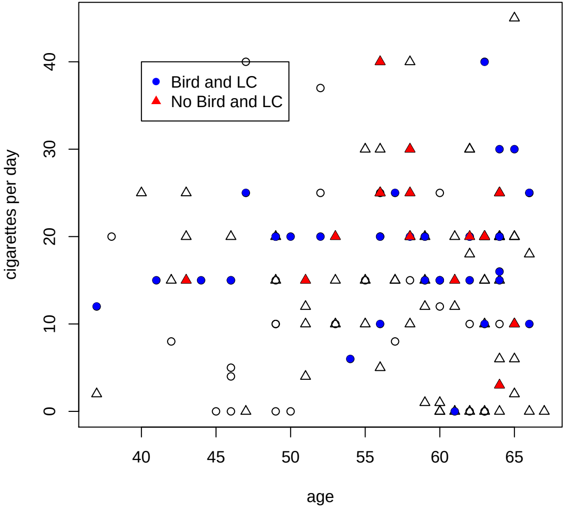
Case Study I - Step by Step Analysis

Goal: examine odds of lung cancer for birdkeepers relative to persons with similar demographic and smoking characteristics who do not keep pet birds.

1. Find an adequate model

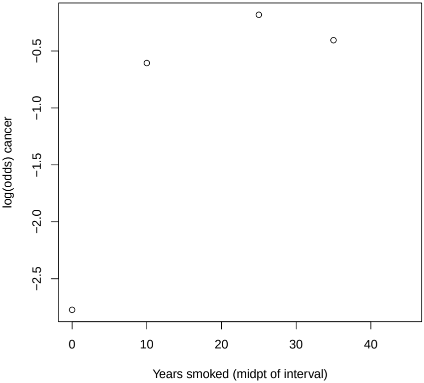
Graphical methods: graph of binary response versus predictor not particularly useful. Can try a scatterplot of one of the explanatory variables versus another, with codes to indicate whether the response is 0 or 1.

Case Study I - Step by Step Analysis



Case Study I - Step by Step Analysis

Grouping: group observations with similar values of explanatory variables. Calculate sample proportions of binary response of 1 in the group, apply logit transformation, draw scatterplot of sample logit versus midpoint of grouped version of explanatory variable.



Case Study I - Step by Step Analysis

Informal testing of extra model terms - test for significance of polynomial and interaction terms (use drop in deviance tests).

Case Study I - Step by Step Analysis

2. Final Model
Full model:

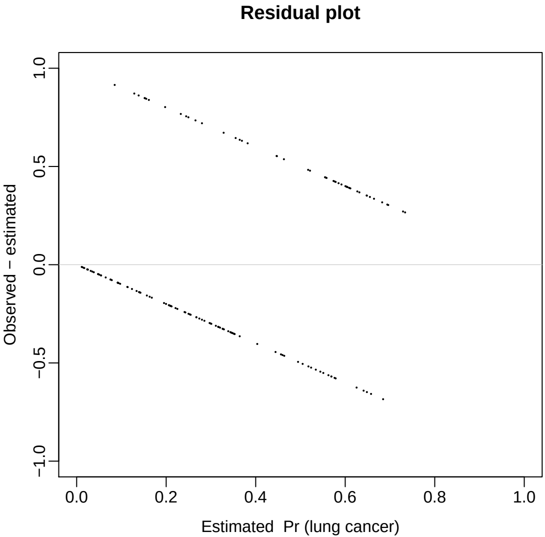
$$logit(p) = \beta_0 + \beta_1 FM + \beta_2 AG + \beta_3 SS + \beta_4 YR + \beta_5 BK$$

The coefficient estimate $\hat{\beta}_5 = 1.3349$ with $SE(\hat{\beta}_5) = 0.4091$.

The odds of lung cancer for birdkeepers are estimated to be 3.80 times as large as the odds of lung cancer for non bird keepers, after the other explanatory variables are accounted for. (95% CI: (1.70,8.47))

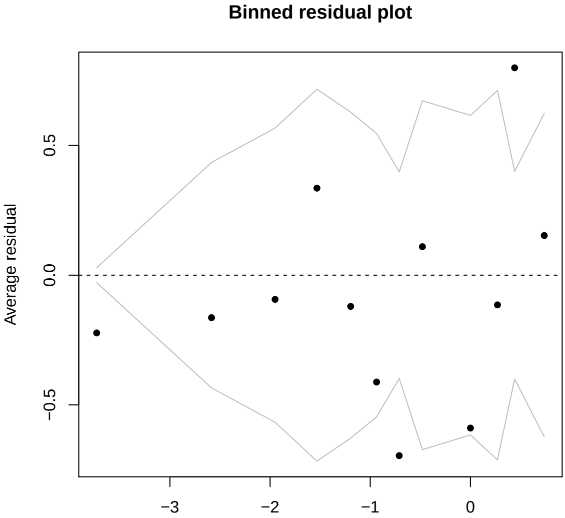
Checking residuals

$residual_i = y_i - logit^{-1}(X_i\beta)$. Data are discrete so plots of raw residuals are generally not useful



Checking residuals

Plot binned residuals: divide data into categories (bins) based on fitted values, then plot average residuals versus average fitted values for each bin.



Identifiability and separation

Nonidentifiable - parameters cannot be estimated. Reasons in logistic regression

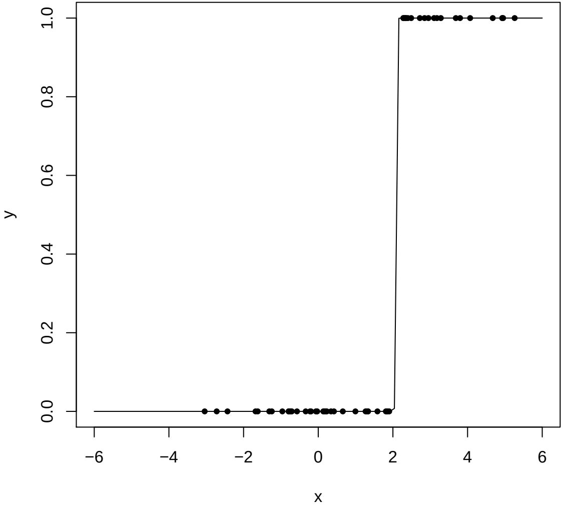
1. Predictors are collinear
2. Separation -

   • Predictor x_j completely aligned with outcome, so that $y = 1$ for all the cases where x_j exceeds some threshold T, and y=0 for all cases where $x_j < T$, then the best estimate of the coefficient β_j is ∞. (or $SE(\hat{\beta}_j)$ is very large)

   • Conversely, if y=1 for all cases where $x_j < T$, and y=0 for all cases where $x_j > T$. then β_j will be $-\infty$. (or $SE(\hat{\beta}_j)$ is very large)

   • More generally, separation problem if any linear combination of predictors is perfectly aligned with outcome.

Identifiability and separation



Identifiability and separation

```
> x<-sample(seq(0,100),100,replace=TRUE)
> y<-rep(0,100)
> y[x>50]<-1
>
> model<-glm(y~x,family=binomial)
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurre
> summary(model)


Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -612.60  129131.92  -0.005    0.996
x               12.13    2554.88   0.005    0.996
```