Unlock the First Steps to Binary Data Analysis

SCSRU Workshop

Statistical Consulting and Survey Research Unit
University of Waterloo

2024-12-11

R and R packages

Throughout our discussion, we will be using R to perform the analysis. We will use the following packages in R:

- catdata,
- MASS,
- performance, and
- AER.

If you are conducting analysis using other software, please consult the respective handbook for detailed codes.

1. After collecting the data

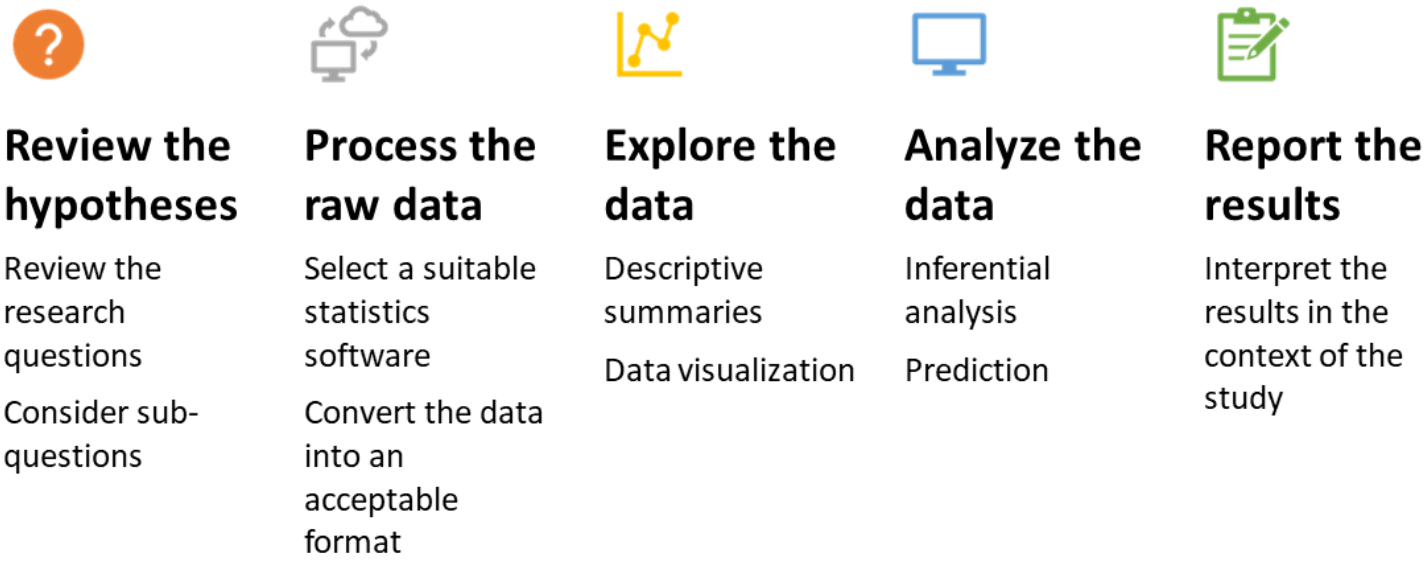


Figure 1: From raw data to results

A motivating example

- One of the data sets that we will use is the `heart` data set from the `catdata` package in R.
- This data set contains a retrospective sample of 462 males between ages 15 and 64 in South Africa where the risk of heart disease is considered high.

```
# install.packages("catdata") #Install the library if needed
library("catdata")
data("heart", package = "catdata") # load the data set

# convert to data frame
heart <- data.frame(heart)

# View(heart) #Spreadsheet view
```

This data set contains the variables:

- `y`: whether or not the subject has coronary heart disease,
- `sbp`: measurements of systolic blood pressure,
- `tobacco`: cumulative tobacco use,
- `ldl`: low density lipoprotein cholesterol,
- `adiposity`: adiposity,
- `famhist`: whether or not the subject has a family history of heart disease,
- `typea`: measures of type-A behavior,
- `obesity`: a measure of obesity,
- `alcohol`: current alcohol consumption, and
- `age`: the subject's age.

1.1. Explanatory and response variables

The most common goal in research is to understand the relationship between variables. These variables are typically categorized as:

- **Response variable** (or dependent variable): An outcome of the study or of interest.
- **Explanatory variable** (or independent variable): A measure in the study used to explain, predict or influence the response variable. Sometimes we also refer them as predictors or covariates.

Throughout this workshop, we want to investigate the factors that have an impact on coronary heart disease. Hence, $\mathtt{y}$ is the response variable, and the other variables are the explanatory variables.

2. Linear regression

We often encounter the linear regression (linear model) in data analysis because:

- the model assumptions are often found satisfactory among many data sets; and
- the interpretation of each parameter in the model is easy and clear.

When the assumptions of the linear regression model are satisfied, the model is powerful in terms of inference and interpretation.

Although linear models have the potential to answer many research questions, we may be interested in finding the association between an outcome and a set of covariates where the outcome is not necessarily continuous or normally distributed.

3. Generalized Linear Models

When the response variable is not continuous or normally distributed, we are most likely to use a family of model called the **generalized linear models**.

The generalized linear model is comprised of three components:

1. The Random Component: The distribution of the independently and identically distributed (i.i.d.) response variables are assumed to come from a parametric distribution that is a member of the exponential family.
2. The Systematic Component: The linear combination of explanatory variables and regression parameters.
3. The Link Function: The function that relates the mean of the distribution of the response to the linear predictor.

3.1 Link Functions

Recall that we are modelling the mean of the outcome through a link function as in

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}.$$

- The link function will essentially transform a non-linear outcome such that it can be linked to the covariates through a linear combination allowing us to fit a generalized linear model.
- For each distribution in the exponential family, there is a canonical link which is recommended to use as simplifies the process of finding maximum likelihood estimates in our model by ensuring that the mean of our outcome is mapped to $(-\infty, \infty)$ so we do not need to worry about constraints when optimizing.
- It also ensures $\boldsymbol{x}^T \boldsymbol{y}$ is a sufficient statistic for β.

Distribution of Y	Canonical Link	Family parameter in R `glm` function
Normal (used for symmetric continuous data)	Identity: $g(\mu) = \mu$	`family = gaussian(link = "identity")`
Binomial (Binary data is a special case where n = 1)	Logistic: $g(\mu) = \log\left(\frac{\mu}{n-\mu}\right)$	`family = binomial(link = "logit")`
Poisson (used for discrete count data)	Log: $g(\mu) = \log(\mu)$	`family = poisson(link = "log")`
Gamma (used for continuous, positive, skewed or heteroskedastic data)	Reciprocal: $g(\mu) = \frac{1}{\mu}$	`family = Gamma(link = "inverse")`

Figure 2: Commonly used distributions and their canonical links

3.2 Assumptions

1. The outcome Y_i is independent between subjects and comes from a distribution that belongs to the exponential family,
2. There is a linear relationship between a transformation of the mean and the predictors through the link function, and
3. The errors are uncorrelated with constant variance, but not necessarily normally distributed.

We also assume that there is no multicollinearity among explanatory variables

Notes:

1. The methodology is particularly sensitive to these assumptions when sample sizes are small.
2. When collecting data, we also want to ensure that the sample is representative of the population of interest to answer the research question(s).

3.3 Model selection

When fitting GLMs, it is important to perform model selection procedures and assess the model fit before interpreting the results. We aim to find the simplest model that explains the relationship between the outcome and covariate(s) of interest.

3.3.1 Likelihood ratio test

- One of the most common ways to compare models against each other is through the likelihood ratio test (LRT).
- We can compare a full model to a nested model that contains a subset of variables that appear in the full model.
- LRTs tend to be the preferred method for building logistic regression models.

3.3.2 Information criteria

For nested or non-nested models, we can also perform model selection by

- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)

For both criteria, a smaller value represents a better fit. Although both AIC and BIC are similar, research has shown that each are appropriate for different tasks.

4. Logistic Regression

The logistic regression relates a binary outcome $\boldsymbol{Y_i}$ to a set of covariates $\boldsymbol{x_i}$ through the mean, as

$$g(\mu_i) = \boldsymbol{x}_i^T \beta$$

where Y_i is the binary outcome for individual i, $\boldsymbol{x}_i$ is a $p \times 1$ vector of covariates for individual i, and $g(\cdot)$ is a link function.

- Our primary interest is in estimating the coefficients β in our model.
- These coefficients can be interpreted as log odds ratio of the outcome for a one unit change in the corresponding covariate.

Interpretation of the coefficients β

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

Suppose x_1 is a binary covariate such as disease presence,

- β_1 is the estimated log odds ratio of the outcome Y for those with the disease versus without, controlling for other covariates in the model.

Suppose x_1 is a continuous covariate such as age,

- β_1 is the estimated log odds ratio associated with a one year increase in age, controlling for other covariates.

4.1 The `heart` data set

The researchers who collected the‘ data set are interested to see whether tobacco use has an effect on CHD diagnosis.

- The dependent variable is CHD diagnosis.
- Since the dependent variable is a binary variable, we cannot apply the linear model.
- The logistic regression model is a reasonable model for this scenario because we also wish to control for other variables/factors.

4.1.1 Data pre-processing

- Before building any kind of models in R, we need to pre-process or "clean" the data.
- The first thing we can do is ensure the covariates in our data set are the correct type.

```
str(heart)
```

```
## 'data.frame':    462 obs. of  10 variables:
##  $ y         : num  1 1 0 1 1 0 0 1 0 1 ...
##  $ sbp       : num  160 144 118 170 134 132 142 114 114 132 ...
##  $ tobacco   : num  12 0.01 0.08 7.5 13.6 6.2 4.05 4.08 0 0 ...
##  $ ldl       : num  5.73 4.41 3.48 6.41 3.5 6.47 3.38 4.59 3.83 5.8 ...
##  $ adiposity: num  23.1 28.6 32.3 38 27.8 ...
##  $ famhist   : num  1 0 1 1 1 1 0 1 1 1 ...
##  $ typea     : num  49 55 52 51 60 62 59 62 49 69 ...
##  $ obesity   : num  25.3 28.9 29.1 32 26 ...
##  $ alcohol   : num  97.2 2.06 3.81 24.26 57.34 ...
##  $ age       : num  52 63 46 58 49 45 38 58 29 53 ...
```

- The variables `sbp`, `tobacco`, `adiposity`, `obesity`, and `alcohol`, are continuous covariates. The output shows that these variables are recorded as `num`, i.e. numeric value. No further action is required.
- The variable `famhist` is a binary variable, but recorded as a numeric value. We need to convert it into a categorical variable.

```
# specify categorical variables as factors
heart$famhist_f <- as.factor(heart$famhist)
```

Data manipulation

Some data requires us to re-categorize, transform or manipulate some of the variables.

Suppose we want to convert age into a categorical variable to have a multi-level categorical variable in our analysis in the following manner:

- Group 1: 15 to 24
- Group 2: 25 to 34
- Group 3: 35 to 44
- Group 4: 45 to 54
- Group 5: 55 to 64

Data manipulation in R

```
#make a copy of age
heart$age_f <- heart$age

# overwrite it, making groups by age
heart$age_f[heart$age %in% 15:24] <- 1
heart$age_f[heart$age %in% 25:34] <- 2
heart$age_f[heart$age %in% 35:44] <- 3
heart$age_f[heart$age %in% 45:54] <- 4
heart$age_f[heart$age %in% 55:64] <- 5
#specify variable as factor
heart$age_f <- as.factor(heart$age_f)
```

We emphasize that this decision is just to demonstrate data manipulation tricks and later on, how to work with categorical factors in the model.

4.2 Fitting logisttic regression in R

To fit a logistic regression model in R, we fit a generalized linear model using the `glm()` function and specify a logistic link function by using the `family=binomial(link = "logit")"` argument.

We begin by building the main-effects only logistic regression model considering all covariates previously described by:

```
#build the logistic model
heart_modelmaineffects <- glm(y ~ sbp + tobacco + ldl +
                                adiposity + famhist_f +
                                typea + obesity + alcohol +
                                age_f,
                              family=binomial(link = "logit"),
                              data=heart)
```

```r
#show the output
summary(heart_modelmaineffects)
```

```
##
## Call:
## glm(formula = y ~ sbp + tobacco + ldl + adiposity + famhist_f +
##     typea + obesity + alcohol + age_f, family = binomial(link = "logit"),
##     data = heart)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.166777   1.389558  -4.438 9.08e-06 ***
## sbp          0.007399   0.005745   1.288 0.197755
## tobacco      0.083147   0.026648   3.120 0.001807 **
## ldl          0.168207   0.059890   2.809 0.004976 **
## adiposity    0.027937   0.029331   0.952 0.340868
## famhist_f1   0.949461   0.229867   4.130 3.62e-05 ***
## typea        0.039052   0.012308   3.173 0.001510 **
## obesity     -0.076045   0.045265  -1.680 0.092959 .
## alcohol     -0.001241   0.004499  -0.276 0.782638
## age_f2       1.867250   0.792464   2.356 0.018460 *
## age_f3       1.899604   0.796106   2.386 0.017027 *
## age_f4       2.179759   0.809370   2.693 0.007078 **
## age_f5       2.710949   0.809071   3.351 0.000806 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 468.28  on 449  degrees of freedom
## AIC: 494.28
##
## Number of Fisher Scoring iterations: 6
```

4.2.1 Model selection with LRT

- The three covariates have small estimated coefficients, and p-values larger than 0.05.
- This is an indication that these variables may not be necessary in the model.
- We can see if these covariates are necessary in the model by testing the full model against one that does not contain the three covariates using the likelihood ratio test (LRT).

4.2.1 Model selection with LRT

```
heart_model2 <- glm(y ~ tobacco + ldl + famhist_f + typea +
                      obesity + age_f,
                    family=binomial(link = "logit"),
                    data=heart)

# perform the LRT
anova(heart_model2, heart_modelmaineffects, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ tobacco + ldl + famhist_f + typea + obesity + ag
## Model 2: y ~ sbp + tobacco + ldl + adiposity + famhist_f + ty
##     alcohol + age_f
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       452     471.07
## 2       449     468.28  3   2.7856   0.4259
```

Interpreting results from a LRT

- The p-value is larger than 0.05, indicating that we do NOT reject the null hypothesis.
- This means that the two models are not significantly different, statistically speaking.
- The test suggests that we can move forward with the simpler model.

```
##
## Call:
## glm(formula = y ~ tobacco + ldl + famhist_f + typea + obesity +
##     age_f, family = binomial(link = "logit"), data = heart)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.56962    1.16980  -4.761 1.92e-06 ***
## tobacco      0.08307    0.02597   3.199  0.00138 **
## ldl          0.18393    0.05839   3.150  0.00163 **
## famhist_f1   0.93218    0.22816   4.086 4.40e-05 ***
## typea        0.03717    0.01218   3.052  0.00228 **
## obesity     -0.03857    0.02976  -1.296  0.19491
## age_f2       1.88884    0.78761   2.398  0.01648 *
## age_f3       2.05322    0.78102   2.629  0.00857 **
## age_f4       2.42502    0.78318   3.096  0.00196 **
## age_f5       3.03253    0.77335   3.921 8.81e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 471.07  on 452  degrees of freedom
## AIC: 491.07
##
## Number of Fisher Scoring iterations: 6
```

We continue to remove obesity. Why?

```
# fit the nested model without obesity
heart_model3 <- glm(y ~ tobacco + ldl + famhist_f + typea +
                      age_f,
                    family=binomial(link = "logit"),
                    data=heart)

# perform the LRT
anova(heart_model3, heart_model2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ tobacco + ldl + famhist_f + typea + age_f
## Model 2: y ~ tobacco + ldl + famhist_f + typea + obesity + ag
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       453     472.79
## 2       452     471.07  1    1.716   0.1902
```

Interactions

- All of the remaining covariates in our model are statistically significant.
- We could stop the model selection procedure here.
- We could also consider interactions and higher-order terms in our model selection.

What are interaction terms?

- An interaction happens when the effect of an independent variable is affected by the value of another independent variable.
- There are two-factor interactions (2FIs), three-factor interactions (3FIs), etc.
- The higher order interactions are less likely to be significant. They are also harder to interpret.
- We recommend to not go beyond 2FIs unless the literature suggests that certain higher order interaction terms are meaningful.
- Experts' opinion can be helpful to identify meaningful higher-order terms.

Back to the heart data set

- Suppose the experts believe that those with a family history of heart disease may have different cholesterol levels than those who do not.
- Let's include the 2FI (`ldl*famhist_f`) into the model and check whether its presence is necessary.

```
# fit the nested model with an interaction term
heart_model4 <- glm(y ~ tobacco + ldl + famhist_f + typea +
                      age_f +ldl*famhist_f,
                    family=binomial(link = "logit"),
                    data=heart)
```

```r
# perform the LRT
anova(heart_model3, heart_model4, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ tobacco + ldl + famhist_f + typea + age_f
## Model 2: y ~ tobacco + ldl + famhist_f + typea + age_f + ldl
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       453     472.79
## 2       452     463.46  1   9.3244 0.002261 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

- The p-value is less than 0.05, indicating that we reject the null hypothesis that the simpler model fits as well as the larger model.
- That is, we should include the interaction term in our model.

```r
summary(heart_model4)
```

```
##
## Call:
## glm(formula = y ~ tobacco + ldl + famhist_f + typea + age_f +
##     ldl * famhist_f, family = binomial(link = "logit"), data = heart)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.61061    1.04614  -5.363 8.18e-08 ***
## tobacco         0.08901    0.02644   3.366 0.000762 ***
## ldl             0.01087    0.07413   0.147 0.883425
## famhist_f1     -0.81122    0.62828  -1.291 0.196643
## typea           0.03625    0.01243   2.917 0.003535 **
## age_f2          1.79383    0.78263   2.292 0.021902 *
## age_f3          1.95366    0.77616   2.517 0.011834 *
## age_f4          2.24868    0.77355   2.907 0.003650 **
## age_f5          2.96258    0.76746   3.860 0.000113 ***
## ldl:famhist_f1  0.34624    0.11725   2.953 0.003146 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 463.46  on 452  degrees of freedom
## AIC: 483.46
##
## Number of Fisher Scoring iterations: 6
```

Notes on model selection

- The main effects of `ldl` and `famhist_f` are now insignificant.
- The common practice is to keep both main effects in the model when the interaction term is in the model. This is to ease interpretation later on.
- For categorical variables, we do not test if individual levels of the covariate should be included.

4.2.2 Stepwise model selection procedure

- When there are many covariates, the built-in stepwise model selection procedure, `step()` may be a better option.
- The `step()` function can evaluate the model on the Akaike information criterion (AIC), or the Bayesian information criterion (BIC), where a smaller value represents a better fit.
- We can also specify which direction we want the function to search through: `forward`, `backward` or `both`.

The backward selection procedure

```
#fit a null model, including the intercept
heart_empty <- glm(y ~ 1,
                   family=binomial(link = "logit"),
                   data=heart)
#fit a full model, including the interaction
heart_full <- glm(y ~ sbp + tobacco + ldl + adiposity +
                    famhist_f + typea + obesity +
                    alcohol + age_f + ldl*famhist_f,
                  family=binomial(link = "logit"),
                  data=heart)
heart_step <- step(heart_full, scope = list(upper=heart_empty),
                   direction = c("backward"),
                   k=2, trace=0)
```

```r
summary(heart_step)
```

```
##
## Call:
## glm(formula = y ~ tobacco + ldl + famhist_f + typea + age_f +
##     ldl:famhist_f, family = binomial(link = "logit"), data = heart)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.61061    1.04614  -5.363 8.18e-08 ***
## tobacco         0.08901    0.02644   3.366 0.000762 ***
## ldl             0.01087    0.07413   0.147 0.883425
## famhist_f1     -0.81122    0.62828  -1.291 0.196643
## typea           0.03625    0.01243   2.917 0.003535 **
## age_f2          1.79383    0.78263   2.292 0.021902 *
## age_f3          1.95366    0.77616   2.517 0.011834 *
## age_f4          2.24868    0.77355   2.907 0.003650 **
## age_f5          2.96258    0.76746   3.860 0.000113 ***
## ldl:famhist_f1  0.34624    0.11725   2.953 0.003146 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 463.46  on 452  degrees of freedom
## AIC: 483.46
##
## Number of Fisher Scoring iterations: 6
```

- The model chosen by the `step()` function in this case is exactly the same as the one we obtained by the LRT.
- However, depending on the order we perform the LRTs, or the direction of the stepwise algorithm, we can obtain different model results.
- LRTs tend to be preferred for building logistic regression models where we want to draw claims and perform hypothesis tests.
- AIC based algorithms tend to be preferred for forecasting problems.

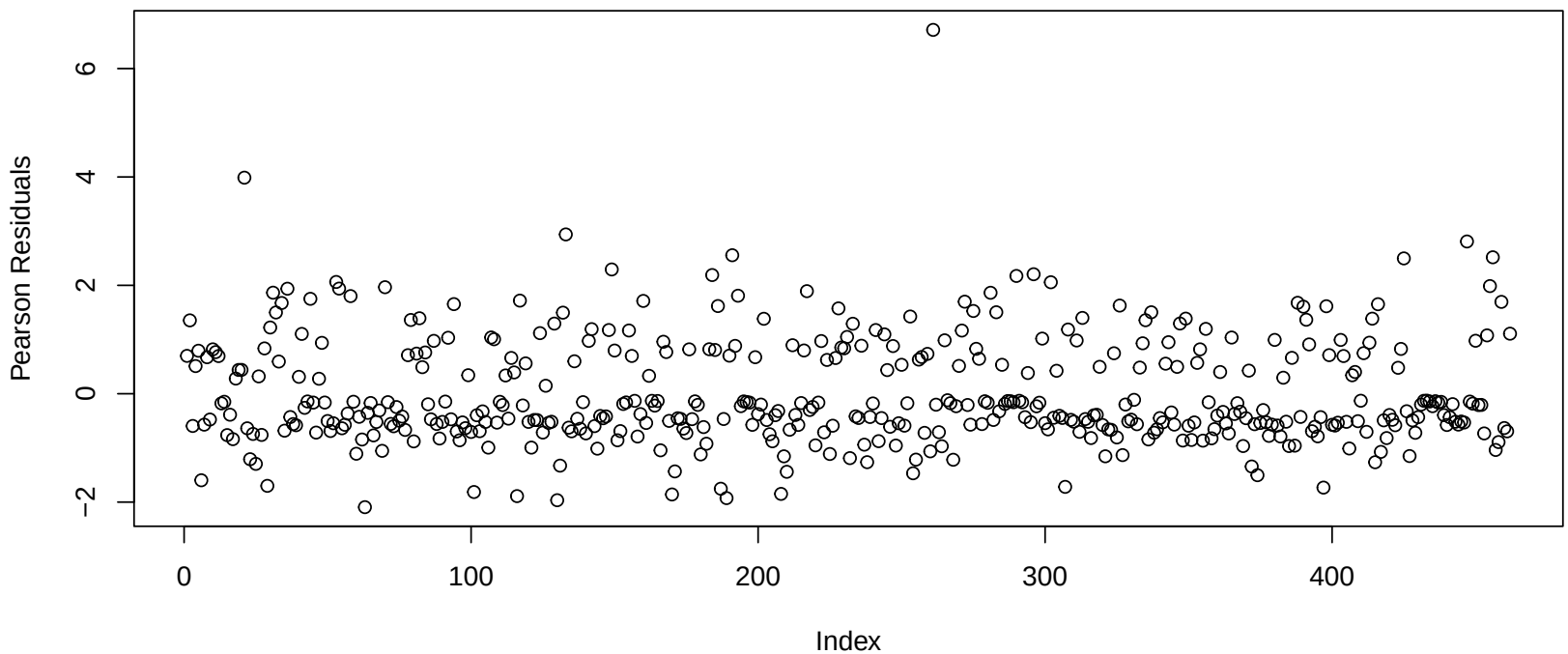
4.3 Model diagnostics

- The goal of the researcher is to see if there is a relationship between CHD diagnosis and tobacco use.
- Recall that the LRT is preferable for logistic regression.
- As such, we will continue our analysis with the model chosen by our LRTs.
- Before interpreting the chosen model, we must assess the model fit.

4.3.1 Residual analysis

We begin by plotting the Pearson residuals to give an idea of the model fit.

```
# Plot the residuals
plot(residuals(heart_model4, type = "pearson"),
     ylab = "Pearson Residuals")
```



From the residual plot

- There are two points that deviate far away from the rest.
- To identify the two subjects with high residual values, we can use:

```
# sort residuals largest to smallest and select the first
sort(residuals(heart_model4, type = "pearson"),
     decreasing = T)[1:2]
```

```
##       261        21
## 6.714401 3.987004
```

- To see if these are influential observations, we can refit the logistic regression model without these observations.
- If the estimates of our model change greatly, then we should remove these two observations as they may affect inference and predictions made with the logistic regression model.

Let's make a second data set without the 261st observation and see if the results of the model change.

```
heart2 <- heart[-261,] # removing the 261st observation

#fit the model using heart2
heart_model4_2 <- glm(y ~ tobacco + ldl + famhist_f + typea +
                          age_f + ldl*famhist_f,
                      family=binomial(link = "logit"),
                      data=heart2)
```

Do you see that the estimated coefficients and standard errors change after removing the observation?

- The estimated regression coefficients and standard errors of the `age_f` variable changed greatly.
- This shows that observation 261 is an influential observation, and should be removed.

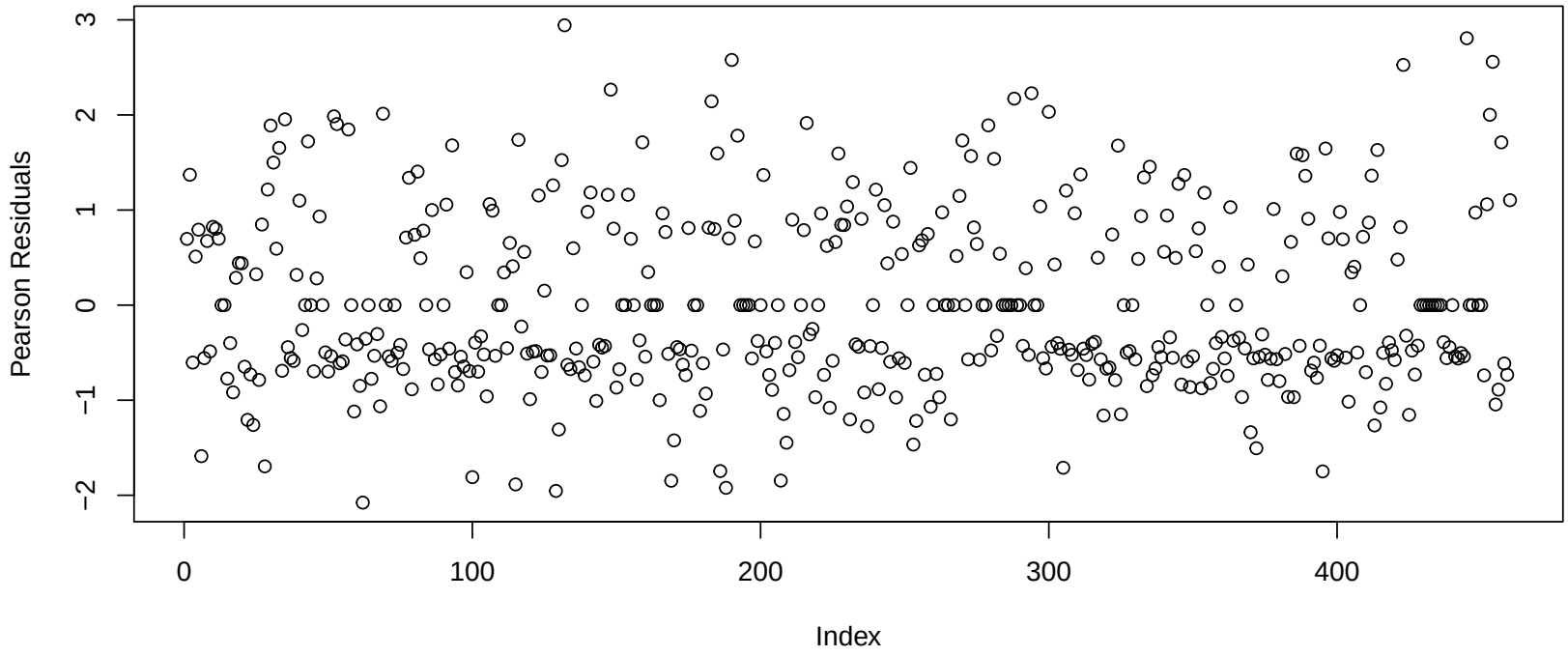
Another observation

The Pearson residual of observation 21 is large too.

```
heart3 <- heart[-c(21, 261),]

#fit the model using heart3
heart_model4_3 <- glm(y ~ tobacco + ldl + famhist_f + typea +
                          age_f + ldl*famhist_f,
                      family=binomial(link = "logit"),
                      data=heart3)
```

Before moving forward with this new model, we need to check the residual plot again.



Most residual values fall between $(-2, 2)$ (with no values beyond ± 3), which indicates a proper model fit.

What do you notice?

```
summary(heart_model4_3)
```

```
##
## Call:
## glm(formula = y ~ tobacco + ldl + famhist_f + typea + age_f +
##     ldl * famhist_f, family = binomial(link = "logit"), data = heart3)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -20.65184  793.20342  -0.026 0.979229
## tobacco          0.08771    0.02635   3.328 0.000874 ***
## ldl              0.03073    0.07479   0.411 0.681168
## famhist_f1      -0.65602    0.63603  -1.031 0.302337
## typea            0.03471    0.01254   2.768 0.005642 **
## age_f2          16.82365  793.20311   0.021 0.983078
## age_f3          16.97712  793.20310   0.021 0.982924
## age_f4          17.26763  793.20310   0.022 0.982632
## age_f5          17.97553  793.20309   0.023 0.981920
## ldl:famhist_f1   0.32012    0.11794   2.714 0.006643 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 591.85  on 459  degrees of freedom
## Residual deviance: 446.13  on 450  degrees of freedom
## AIC: 466.13
##
## Number of Fisher Scoring iterations: 17
```

The model summary shows that some covariates have large estimated standard errors and are no longer significant, indicating that after removing the influential observations we should re-do our model fitting.

```
#fit a null model
heart_empty2 <- glm(y ~ 1,
                    family=binomial(link = "logit"),
                    data=heart3)
#fit a full model, including the interaction
heart_full2 <- glm(y ~ sbp + tobacco + ldl + adiposity +
                     famhist_f + typea + obesity +
                     alcohol + age_f + ldl*famhist_f,
                    family=binomial(link = "logit"),
                    data=heart3)

heart_step2 <- step(heart_full2, scope = list(upper=heart_empty2
                    direction = c("both"),
                    k=2,
                    trace = 0) #don't print every step
```

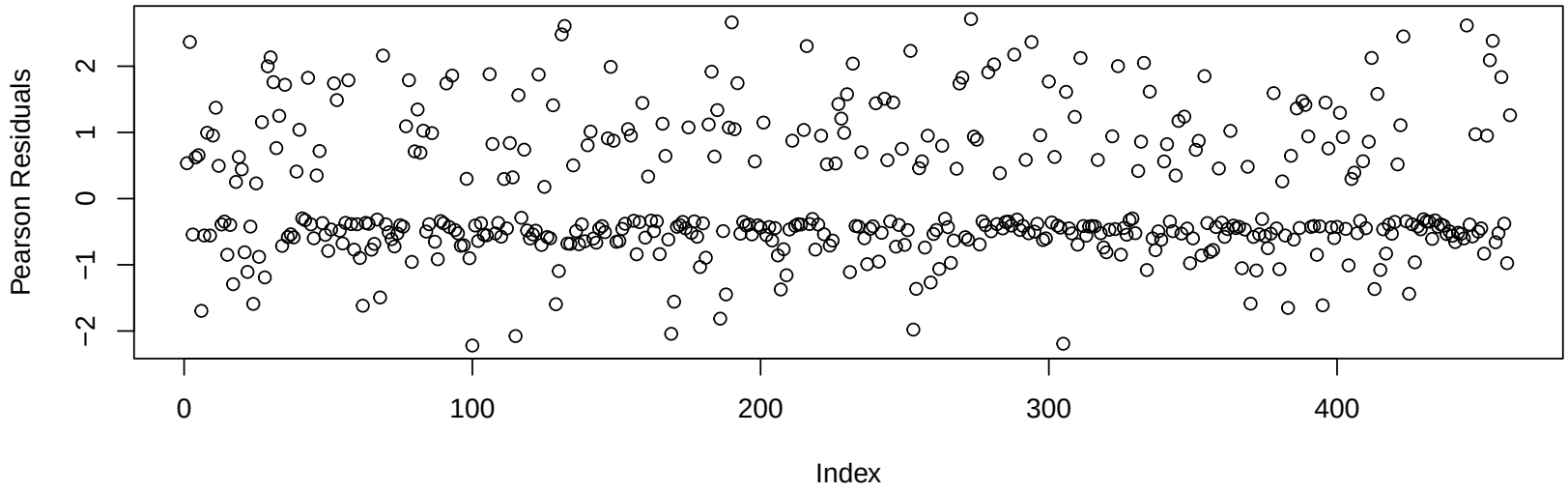
- The estimates of age_f and the corresponding standard error are large in comparison to other covariates.
- The stepwise procedure (and an LRT) will not suggest us to remove age_f.
- The variable age_f is not adding any value to our model.
- Let's try to refit this model without age_f:

```
# fit a new model
heart_model7 <- glm(y ~ tobacco + ldl + famhist_f +
                        typea + ldl*famhist_f ,
                    family=binomial(link = "logit"),
                    data=heart3)
```

How does this look?

```
summary(heart_model7)
```

```
##
## Call:
## glm(formula = y ~ tobacco + ldl + famhist_f + typea + ldl * famhist_f,
##     family = binomial(link = "logit"), data = heart3)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.47029    0.74548  -4.655 3.24e-06 ***
## tobacco         0.13821    0.02539   5.444 5.22e-08 ***
## ldl             0.09899    0.07189   1.377  0.16850
## famhist_f1     -0.39576    0.61446  -0.644  0.51953
## typea           0.02383    0.01182   2.016  0.04385 *
## ldl:famhist_f1  0.30096    0.11609   2.592  0.00953 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 591.85  on 459  degrees of freedom
## Residual deviance: 486.83  on 454  degrees of freedom
## AIC: 498.83
##
## Number of Fisher Scoring iterations: 4
```



The residuals are behaving as expected with no extreme values.

4.3.2 Multicollinearity

- We should also check for multicollinearity in our model.
- To do so, we can use the `vif()` function from the `car` package.
- We typically are concerned about multicollinearity when VIF values are above 10.

```r
# there are multiple vif functions so car:: specifies
# we are using the vif function from the car package
car::vif(heart_model7)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##       tobacco           ldl     famhist_f         typea ldl:f
##      1.034016      1.649904       7.596317      1.013287
```

We see that we only have low-moderate variance inflation factors (VIFs), indicating that multicollinearity is not an issue in this model.

4.4 Interpretation

Now, we are ready to discuss how to interpret the results. There are three ways to discuss the results:

- odd ratios,
- confidence intervals, and
- probabilities.

4.4.1 Odds

The odds of an event is the probability of success (π) divided by its probability of failure ($1 - \pi$):

$$Odds = \frac{\pi}{1 - \pi}.$$

If we have a fair coin, what is the odds of getting a head? The probability of getting a head is 0.5 and the probability of not getting a head is 0.5. The odds of getting a head is

$$Odds = \frac{0.5}{1 - 0.5} = 1$$

Odd ratios

The odd ratio is a comparison of two odds,

$$OR = \frac{Odds_1}{Odds_2},$$

where $Odds_1$ is the odds of Event 1 and $Odds_2$ is the odds of Event 2.

The main interest is to determine whether an odd ratio (OR)

- is equal to 1,
- larger than 1, or
- smaller than 1.

- When $OR = 1$, this implies that the probabilities of the two events happening are the same.
- When $OR > 1$, this implies that Event 1 is more likely to happen than Event 2.
- When $OR < 1$, then we say that Event 1 is less likely to happen than Event 2.

Revisiting the model

In mathematical notation, the model we chose from the selection procedure is written as

$$\log\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_2 * x_3.$$

- π_i is the probability of having CHD,
- x_1 is the measurement of cumulative tobacco use,
- x_2 is the ldl cholesterol measurement,
- x_3 is an indicator for family history, and
- x_4 is the measure of type-A behavior.
- We interpret each coefficients (β) as the log odds ratio of the outcome for a one unit change in the corresponding covariate, controlling for the other covariates in the model.

- The `summary()` output provided the estimates and standard errors of the log odds ratios.
- To obtain estimates of the odds ratio, we take the exponential of the coefficient.
- The estimates of the standard error for the coefficients (log-odds) will be useful for hypothesis testing and constructing confidence intervals.

Example 4.1

- The `tobacco` covariate's estimated coefficient ($\widehat{\beta_1}$) is 0.138.
- We estimate that a one unit increase in `tobacco` is associated with a **log odds ratio** of chronic heart disease equal to 0.138, controlling for the other factors in the model.
- Alternatively, we can say that a one unit increase in `tobacco` is associated with an **odds ratio** of chronic heart disease equal to exp(0.138) $= 1.148$, controlling for other factors.
- A one unit increase in `tobacco` is estimated to be 114.8% more likely to be diagnosed with CHD.

4.4.2 Confidence interval

- Confidence intervals (CIs) are useful in communicating the uncertainty in our estimates and are typically presented along with our estimate.
- The CIs take the form of

$$\widehat{\beta} \pm t \times \widehat{se}(\widehat{\beta}),$$

- The value of t depends on the level of confidence of the CI.
- There are 3 common level of confidence: 90%, 95% and 99%.
- Higher level of confidence corresponds to larger value of t.
- The 99% CI of an estimate is the widest among them, and the 90% CI is the most narrow.
- Although we would like to create 100% CIs, 100% CIs are meaningless because they are (∞, ∞).

The 95% confidence intervals

- We will continue our discussion with the most common CI – the 95% CI.
- A 95% CI for a log odds ratio is:

$$\widehat{\beta} \pm 1.96 \times \widehat{se}(\widehat{\beta}),$$

where $\widehat{se}(\widehat{\beta})$ is the estimated standard error of the regression coefficient.

Example

The 95% confidence interval for the log odds ratio of `tobacco` is

$$0.138 \pm 1.960 \times 0.025 = (0.089, 0.187).$$

- The lower bound of this confidence interval is 0.089.
- The upper bound of the confidence interval is 0.187.

To find the 95% CI for the odds ratio, we exponentiate both sides of the CI as:

$$(\exp(0.089), \exp(0.187)) = (1.093, 1.206)$$

- The odds ratio is estimated to be 1.148 (95% CI: (1.093, 1.206)) controlling for the other factors in the model.
- This 95% CI does not contain the value of OR $= 1$ in it.
- We say that we are 95% confident that higher tobacco use is associated with an increased odds of developing CHD.

We can obtain the 95% confidence interval of all the individual covariates in the model using the function `confint.default()`:

```
logORs <- cbind(coef(heart_model7),
                confint.default(heart_model7))
colnames(logORs) <- c("logOR", "Lower", "Upper")
round(logORs,4) #round to 4 decimal places
```

```
##                    logOR   Lower   Upper
## (Intercept)     -3.4703 -4.9314 -2.0092
## tobacco          0.1382  0.0884  0.1880
## ldl              0.0990 -0.0419  0.2399
## famhist_f1      -0.3958 -1.6001  0.8086
## typea            0.0238  0.0007  0.0470
## ldl:famhist_f1   0.3010  0.0734  0.5285
```

Similarly, we exponentiate `logORs` to obtain the 95% CIs of the ORs:

```
ORs <- exp(logORs)
colnames(ORs) <- c("Odds Ratio", "Lower", "Upper")
round(ORs,4) #round to 4 decimal places
```

```
##                  Odds Ratio  Lower  Upper
## (Intercept)          0.0311 0.0072 0.1341
## tobacco              1.1482 1.0925 1.2068
## ldl                  1.1041 0.9590 1.2711
## famhist_f1           0.6732 0.2019 2.2447
## typea                1.0241 1.0007 1.0481
## ldl:famhist_f1       1.3512 1.0762 1.6964
```

- So far, we have discussed obtaining individual estimates and the corresponding CIs.
- For more complex estimates where we may be interested in combinations of covariates, it can be useful to create a table to determine what regression coefficients we want to use.

Example

Let's look at estimating the odds ratio of CHD for a one unit increase in `ldl` among those with a family history of CHD, controlling for the other factors.

Recall the mathematical model from our selection procedure

$$\log\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0+\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_2 x_3.$$

- x_2 represents `ldl`, and
- x_3 represents `famhist_f`.

- We wish to estimate the odds ratio of CHD for a one unit increase in `ldl`, which is the same as looking at $x_2 = 1$ versus $x_2 = 0$.
- We also are interested in only those with a family history of CHD, represented by $x_3 = 1$.
- All of the other covariates are held constant.
- Combining the above, we are comparing

$$\beta_0+\beta_1 x_1 + \beta_2(1) + \beta_3(1) + \beta_4 x_4 + \beta_5(1)(1)$$

to

$$\beta_0+\beta_1 x_1 + \beta_2(0) + \beta_3(1) + \beta_4 x_4 + \beta_5(0)(1)$$

If we look at the difference of these equations, we have

$$\begin{array}{l} \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3(1) + \beta_4 x_4 + \beta_5(1)(1) \\ \underline{-(\beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(1) + \beta_4 x_4 + \beta_5(0)(1))} \\ \beta_2 \qquad\qquad\qquad\qquad\qquad\qquad + \beta_5 \end{array}$$

- We need to estimate and interpret $\beta_2 + \beta_5$.
- The estimate of the log odds ratio is $\widehat{\beta_2} + \widehat{\beta_5} = 0.099 + 0.301 = 0.400$.
- The estimated odds ratio is $\exp(0.400) = 1.492$.
- So, we estimate that a one unit increase in low density lipoprotein cholesterol is associated with an odds ratio of CHD equal to 1.492, controlling for other factors.

To estimate the confidence interval, we must

1. find the estimated standard error of $\widehat{\beta_2} + \widehat{\beta_5}$ manually,
2. construct a confidence interval for this quantity (the logOR), and then
3. exponentiate each bound of the confidence interval.

```
#get the variance covariance matrix
varcov <- vcov(heart_model7)
#represents \beta_2 + \beta_5 (first place is beta_0)
L <- c(0, 0, 1, 0, 0, 1)

var_est <- L%*%varcov %*% L
#vector of coefficients from model
beta_est <- L%*%coef(heart_model7)

CI <- c(beta_est - 1.96*sqrt(var_est),
        beta_est + 1.96*sqrt(var_est))

exp(CI) #exponentiate to get CI for OR
```

```
## [1] 1.247937 1.783199
```

4.4.3 Probabilities

- In some scenarios, we are more interested in estimating or predicting the probability of the outcome instead of the odds ratio for given covariates.
- We can obtain a prediction using the `predict()` function.

Example: Suppose we are interested in the probability that a 25 year old with `spb` $= 150$, `tobacco` $= 0$, `ldl` $= 6$, `adiposity` $= 24$, no family history of CHD, `typea` $= 60$, `obesity` $= 30$, and `alcohol` $= 10$.

```r
# make a vector of the new information as it would appear in the
# dataframe (excluding y). Use colnames(heart) to see the order
newsubject <- data.frame(sbp = 150,
                         tobacco = 0,
                         ldl = 6,
                         adiposity = 24,
                         famhist = 0,
                         typea = 60,
                         obesity = 30,
                         alcohol = 10,
                         age = 25,
                         age_f = 2,
                         famhist_f = 1)

newsubject$age_f <- as.factor(newsubject$age_f)
newsubject$famhist_f <- as.factor(newsubject$famhist_f)

predict(heart_model7, newdata = newsubject)
```

```
##           1
## -0.03674586
```

The output is the estimate of $\log \frac{\pi}{1-\pi} = -0.037$.

To estimate the probability, we take the expit() of this estimate, or obtain

$$\frac{\exp(-0.037)}{1+\exp(-0.037)} = 0.491$$

We estimate this hypothetical individual to have a 49.1% probability of having CHD given their covariates.

5. Review of applying GLMs

1. Choose the appropriate link function based on the response variable.
2. Fit the model using the `glm()` function.
3. Choose the set of covariates and interaction terms using LRTs or the automatic stepwise selection procedure.
4. Check the model assumptions such as residuals and multicollinearity.
5. Repeat Step 3 and 4 as needed.
6. Interpret the results in the context of the study.

6. Final thoughts

- The logistic regression models belong to the family of generalized linear model.
- There are other generalized linear models such as the Poisson and Gamma regression model.
- The procedure to fit these models is similar to our demonstration in this workshop.
- We encourage you to consider the various generalized linear models when performing data analysis.

The MIDI steps of data analysis

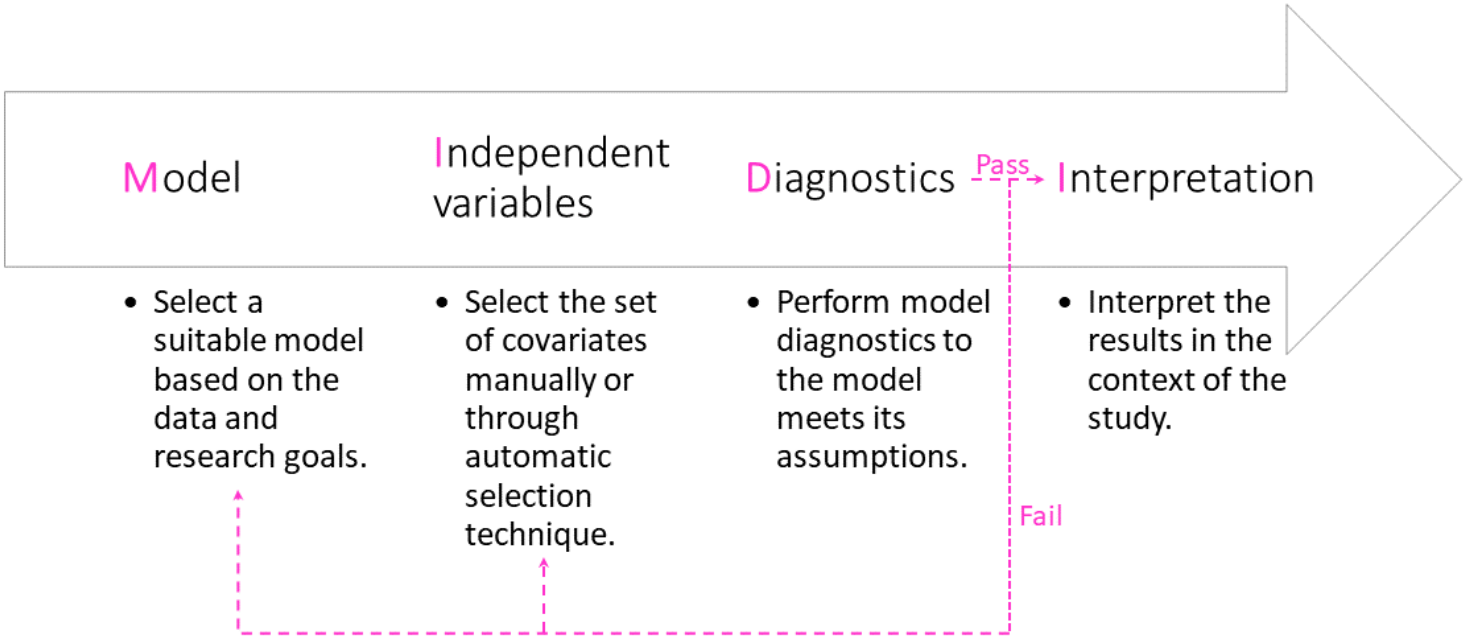


Figure 3: Recommended steps to data analysis

6.1. Statistical vs practical significance

The p-values are commonly used as an indicator of significance/importance. However, we want to remind readers that:

- Statistical inference techniques test for statistical significance.
- Statistical significance means that the effect observed in a sample is very unlikely to occur if the null hypothesis is true.
- Whether this observed effect has practical importance is an entirely different question. The experts in the field of interest determine whether these results have any practical importance.

Some notes about p-values

The ASA's Statement on p-values:

- P-values can indicate how incompatible the data are with a specified statistical model.
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- Proper inference requires full reporting and transparency
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Beyond this workshop

For those who are interested to learn more, the SCSRU hosts statistics seminars and workshops focusing on topics commonly encountered by researchers on campus. Please check our website for future events.

Thank you!

The Statistical Consulting and Survey Research Unit (SCSRU) is the unit through which the Department of Statistics and Actuarial Science provides statistical advice to those working on research problems.