Week ON11

Elmer V Villanueva

04 May 2020

SET YOUR WORKING DIRECTORY!

```
setwd("C:/Users/1/Dropbox/00 - Working Folder/Teaching/DPH112/2019-2020/Week ON11")
```

Announcements

- The two coursework assessments and the final paper have been released. The deadlines are

Assessment	Due Date	Days Till Deadline
Coursework 1	16 May	12
coursework 2	30 May	26
Final Paper	17 June	44

- All the assessments are to be completed INDIVIDUALLY. No collusion is allowed. This means that you cannot discuss this assignment with other students, nor can you share your approach or code.
- The running count for students forwarding errors is as follows:

Student	Items Identified
Yijia Jiang	6
Jing Wang	3
Xinwen Hu	1
Yuxuan Wu	1

Reading

Read and understand Vittinghoff et al., Chapter 5.

Introduction

We have learned much about basic regression using the simple, then the multiple, linear regression models. We have learned to define models, specify their assumtions, estimate the model, test the assumptions, draw inferences and make predictions. We can apply these steps when we learn new models classes. This is what we are going to do. We are going to extend the linear regression model to see if it can be useful for different types of problems.

The types of problems we have been considering these past 10 weeks have been limited to a single kind. This might not have been obvious to you. However, when you think about it, all our dependent variables (i.e.,

Y) have been *continuous* variables. Linear regression model are appropriate when the dependent variable is continuously scaled. They are NOT appropriate when the dependent variable is a different scale. For these kinds of problems, you will need different kinds of regression models.

IMPORTANT: The dependent variable determines the type of regression model that you will apply. This is a good rule of thumb, but it does not always work.

What happens when the dependent variable is a *binary* one? That is, Y takes on only two states. This is a common question in health and biology. For example, we want to predict whether a drug cures or does not cure a disease. We want to know whether the drug dosage is related to the occurrence or non-occurrence of allergies. We want to know whether the distance you stand apart from other people results in you catching or not catching SARS-CoV-2.

In these problems, all our dependent variables are categorical variables. Specifically, they are *binary* categorical variables.

What happens when we try to fit a linear regression on a binary dependent variable? Let's try that now.

Example 1A

Following a heart attack or heart surgery, it is useful to have patients undergo a period of cardiac rehabilitation. However, not all people who are offered this service choose to take it. Gallagher and colleagues wanted to ask the question whether the age of a patient was related to the patient's eventual participation in a cardiac rehabilitation program. More formally, we want to know if we are able to predict whether a patient of a certain age would or would not participate in the program. Why would we want to learn this? Well, if we know that most young women have a low chance of participating in the program, perhaps we can redisng the program for them to better suit their needs (perhaps by holding it at home, for example).

The table gives the ages of women discharged from a hospital in Australia following a heart attack or an invasive procedure on their hearts. We also have information about whether they did (ATT=1) or did not (ATT=0) attend a cardiac rehabilitation program.

Table 1. Ages of women participating (ATT=1) and not participating (ATT=0) in a cardiac rehabilitation program.

ATT = 0				ATT = 1	
50	73	46	74	74	62
59	75	57	59	50	74
42	71	53	81	55	61
50	69	40	74	66	69
34	78	73	77	49	76
49	69	68	59	55	71
67	74	72	75	73	61
44	86	59	68	41	46
53	49	64	81	64	69
45	63	78	74	46	66
79	63	68	65	65	57
46	72	67	81	50	60
62	64	55	62	61	63
58	72	71	85	64	63
70	79	80	84	59	56
60	75	75	39	73	70
67	70	69	52	73	70
64	73	80	67	65	63
62	66	79	82	67	63
50	75	71	84	60	65
61	73	69	79	69	67

ATT = 0				ATT = 1	
69	71	78	81	61	68
74	72	75	74	79	84
65	69	71	85	66	69
80	76	69	92	68	78
69	60	77	69	61	69
77	79	81	83	63	79
61	78	78	82	70	83
72	62	76	85	68	67
67	73	84	82	59	47
80				57	64
				66	

Let us enter the data with the dependent variable `ATT` and the indepdent variable `AGE`.

```
AGE <- c(50, 59, 42, 50, 34, 49, 67, 44, 53, 45,
         79, 46, 62, 58, 70, 60, 67, 64, 62, 50,
         61, 69, 74, 65, 80, 69, 77, 61, 72, 67,
         80, 73, 75, 71, 69, 78, 69, 74, 86, 49,
         63, 63, 72, 64, 72, 79, 75, 70, 73, 66,
         75, 73, 71, 72, 69, 76, 60, 79, 78, 62,
         73, 46, 57, 53, 40, 73, 68, 72, 59, 64,
         78, 68, 67, 55, 71, 80, 75, 69, 80, 79,
         71, 69, 78, 75, 71, 69, 77, 81, 78, 76,
         84, 74, 59, 81, 74, 77, 59, 75, 68, 81,
         74, 65, 81, 62, 85, 84, 39, 52, 67, 82,
         84, 79, 81, 74, 85, 92, 69, 83, 82, 85,
         82,

         74, 50, 55, 66, 49, 55, 73, 41, 64, 46,
         65, 50, 61, 64, 59, 73, 73, 65, 67, 60,
         69, 61, 79, 66, 68, 61, 63, 70, 68, 59,
         57, 66, 62, 74, 61, 69, 76, 71, 61, 46,
         69, 66, 57, 60, 63, 63, 56, 70, 70, 63,
         63, 65, 67, 68, 84, 69, 78, 69, 79, 83,
         67, 47, 64)
ATT <- c(rep(0, times = 121), rep(1, times = 63))
CARDIAC <- data.frame(AGE, ATT)
str(CARDIAC)
```

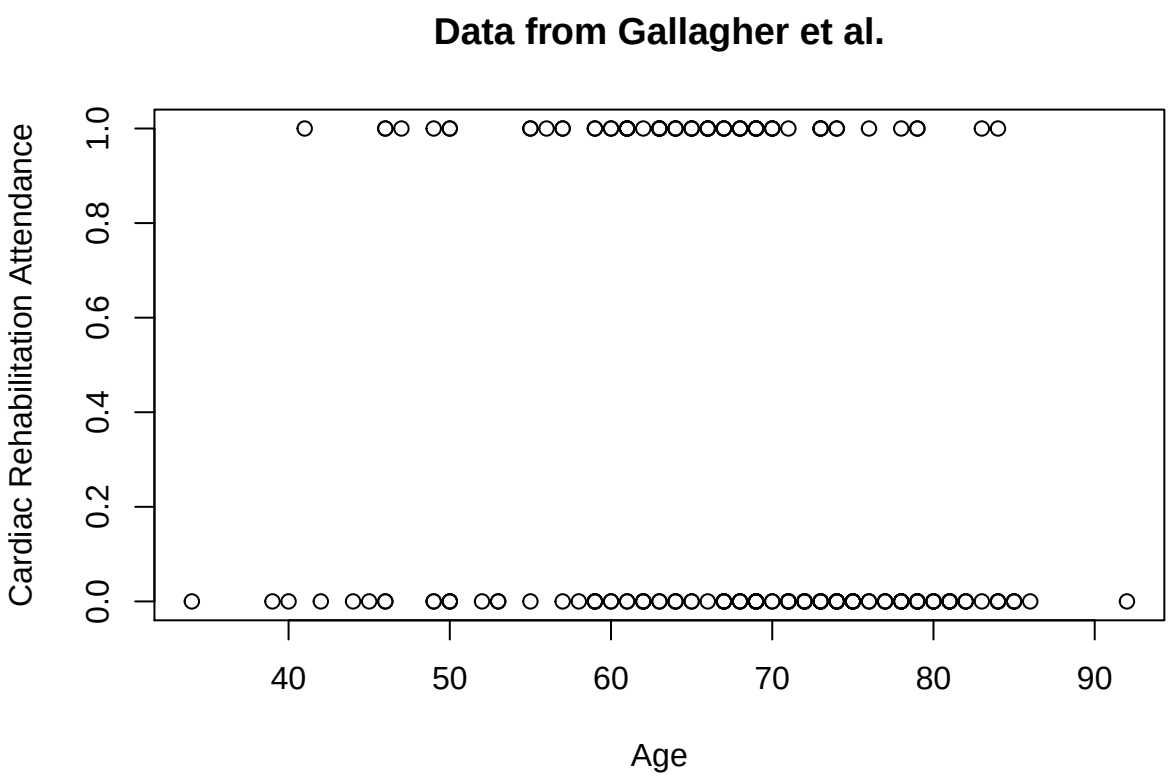
```
## 'data.frame':    184 obs. of  2 variables:
##  $ AGE: num  50 59 42 50 34 49 67 44 53 45 ...
##  $ ATT: num  0 0 0 0 0 0 0 0 0 0 ...
```

```
head(CARDIAC)
```

```
##   AGE ATT
## 1  50   0
## 2  59   0
## 3  42   0
## 4  50   0
## 5  34   0
## 6  49   0
```

Let's draw a picture.

```r
plot(ATT ~ AGE,
     ylab = "Cardiac Rehabilitation Attendance",
     xlab = "Age",
     main = "Data from Gallagher et al.")
```



There is nothing wrong with this graph. R has done what you have asked for. Note that the dependent variable is classified as a `numeric` variable. This is wrong, because 0 and 1 are labels and have no numeric information in the typical sense. Thus, we need to convert it to a `factor` variable.

```r
ATT.F <- factor(ATT, levels = c(0, 1),
                labels = c("No Rehab", "Rehab"))
CARDIAC <- data.frame(AGE, ATT, ATT.F)
str(CARDIAC)
```

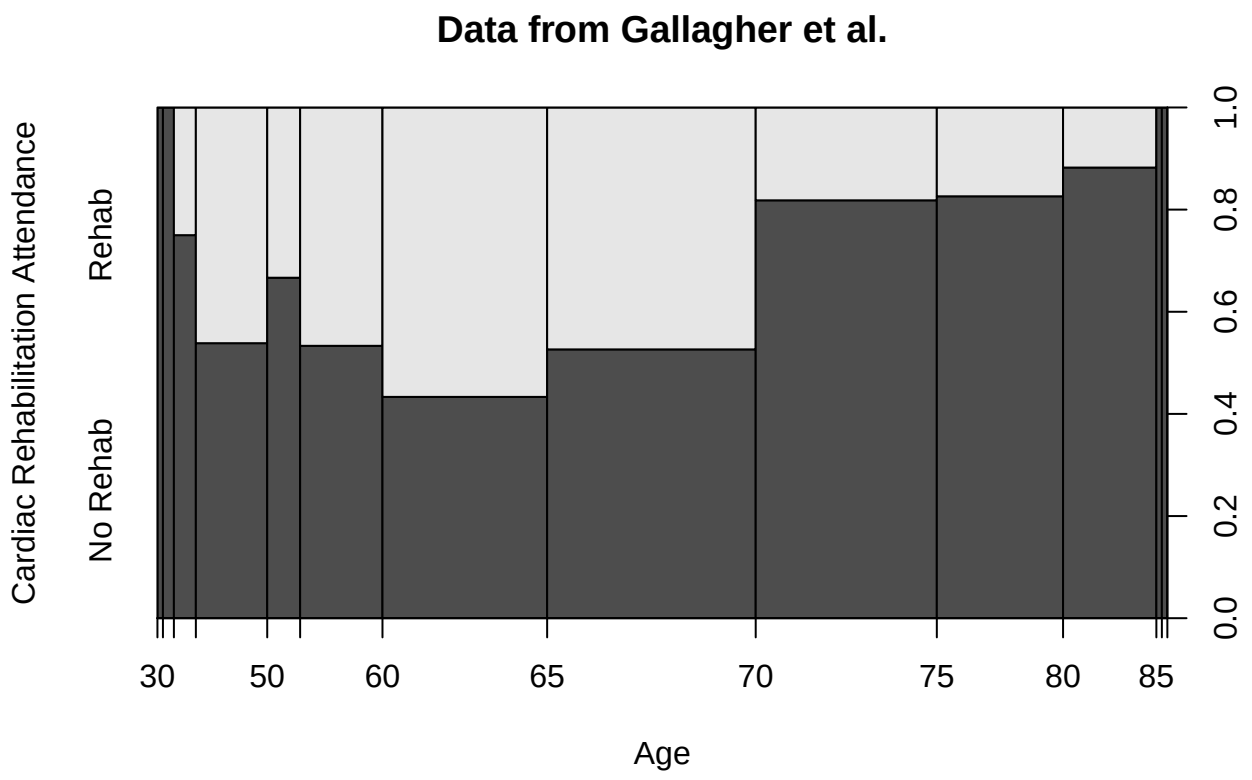
```
## 'data.frame':    184 obs. of  3 variables:
##  $ AGE  : num  50 59 42 50 34 49 67 44 53 45 ...
##  $ ATT  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ATT.F: Factor w/ 2 levels "No Rehab","Rehab": 1 1 1 1 1 1 1 1 1 1 ...
```

```r
head(CARDIAC)
```

```
##   AGE ATT    ATT.F
## 1  50   0 No Rehab
## 2  59   0 No Rehab
## 3  42   0 No Rehab
## 4  50   0 No Rehab
## 5  34   0 No Rehab
## 6  49   0 No Rehab
```

Now, let's see if the graph makes more sense.

```
plot(ATT.F ~ AGE,
     ylab = "Cardiac Rehabilitation Attendance",
     xlab = "Age",
     main = "Data from Gallagher et al.")
```



R, recognising that one of the variables is a `factor` variable, produced a mosaic plot.

Now, let's try to fit a *SIMPLE LINEAR REGRESSION MODEL* onto these data. (We will use the variable `ATT` to do this. Try to use the factor form of `ATT` and see what happens. Simply remove the hashtags in the code snippet below.) , to be clear, we want R to fit the model $Y = \beta_0 + \beta_1 AGE$.

```
CARDIAC.LM <- lm(ATT ~ AGE, data = CARDIAC)
summary(CARDIAC.LM)
```

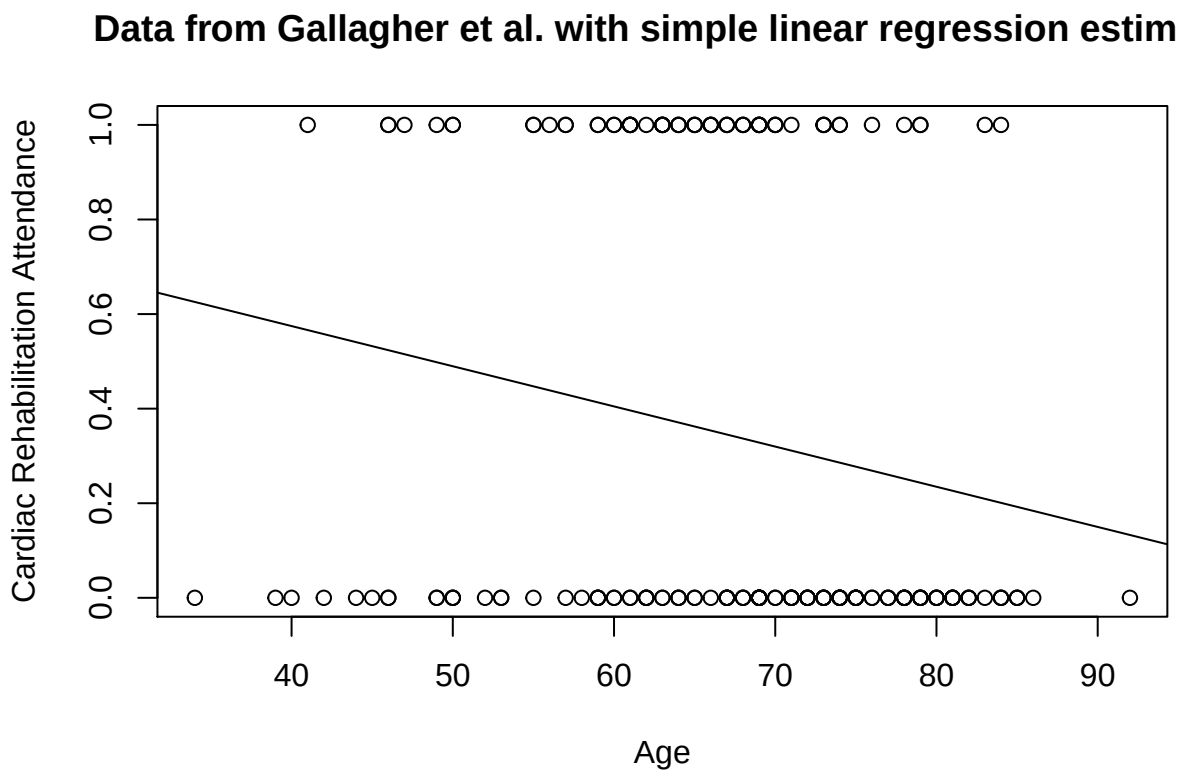
```
##
## Call:
## lm(formula = ATT ~ AGE, data = CARDIAC)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6256 -0.3368 -0.2518  0.6038  0.7992
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.914505   0.216184   4.230 3.69e-05 ***
## AGE         -0.008496   0.003169  -2.681  0.00802 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.468 on 182 degrees of freedom
## Multiple R-squared:  0.03799,    Adjusted R-squared:  0.0327
## F-statistic: 7.187 on 1 and 182 DF,  p-value: 0.00802
```

```
#CARDIAC.LM2 <- lm(ATT.F ~ AGE, data = CARDIAC)
#summary(CARDIAC.LM2)
```

Let's try to visualise the line derived by R.

```
plot(ATT ~ AGE,
     ylab = "Cardiac Rehabilitation Attendance",
     xlab = "Age",
     main = "Data from Gallagher et al. with simple linear regression estimate")
abline(CARDIAC.LM)
```



The resulting formula is $ATT = 0.9145 - 0.0085AGE$. Note that R had no trouble producing a formula. However, how are we to interpret this formula? ATT is a binary state with `ATT = 1` being attendance and `ATT = 0` being non-attendance. What, then, does an ATT value between 0 and 1 mean? Thus, we can see that the resulting estimate is quite useless to us because it gives us meaningless results.

The problem here is the `form` of the dependent variable. We must change it so that the resulting estimates produce meaningful interpretations. In what ways can we change the dependent variable?

One way to do this is to transform the dependent variable. Instead of Y, let's take the probability of Y. That is, let's fit the new model $Y' = Pr(Y) = Y/N = \beta_0 + \beta_1 AGE$. Note that we are only changing the left-hand side of the equation. The right-hand side stays the same.

```r
ATT.1PRIME <- ATT/length(ATT)
CARDIAC2 <- data.frame(AGE, ATT.1PRIME)
plot(ATT.1PRIME ~ AGE,
     ylab = "Probability of Cardiac Rehabilitation Attendance",
     xlab = "Age",
     main = "First Transformation of Y")
```


```r
CARDIAC2.LM1 <- lm(ATT.1PRIME ~ AGE, data = CARDIAC2)
summary(CARDIAC2.LM1)
```

```
##
## Call:
## lm(formula = ATT.1PRIME ~ AGE, data = CARDIAC2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.003400 -0.001830 -0.001368  0.003281  0.004343
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.970e-03  1.175e-03   4.230 3.69e-05 ***
## AGE         -4.618e-05  1.722e-05  -2.681  0.00802 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002543 on 182 degrees of freedom
```

```
## Multiple R-squared:  0.03799,    Adjusted R-squared:  0.0327
## F-statistic: 7.187 on 1 and 182 DF,  p-value: 0.00802
```

```
plot(ATT.1PRIME ~ AGE,
     ylab = "Probability of Cardiac Rehabilitation Attendance",
     xlab = "Age",
     main = "First Transformation of Y")
abline(CARDIAC2.LM1)
```

This seems to make more sense. The resulting equation is $Y' = Pr(Y) = 0.0050 - 4.618 \times 10^{-5} AGE$. This formula suggests that the probabilty of attendance in cardiac rehabilitation decreases as the woman ages. For every 10 years of life, the probability of attendance drops by about 0.56%.

There is a problem with this model. While the transformation is fine, the model is not. This is because the model will be able to produce estimates greater than 1 and less than 0. Since the dependent variable is a probability, then values less than 0 and greater than 1 are impossible. Thus, we are back to where we started.

Let us try another transformation. This time, instead of the probability, we can use the odds. Recall from DPH101 that the odds is the ratio of the occurrence to the non-occurrence of the event (i.e., attendance versus non-attendance). The great thing about the odds is that is has no upper limit. It is bound by zero and $+\infty$. Thus, we don't need to worry about the model producing estimates greater than 1.

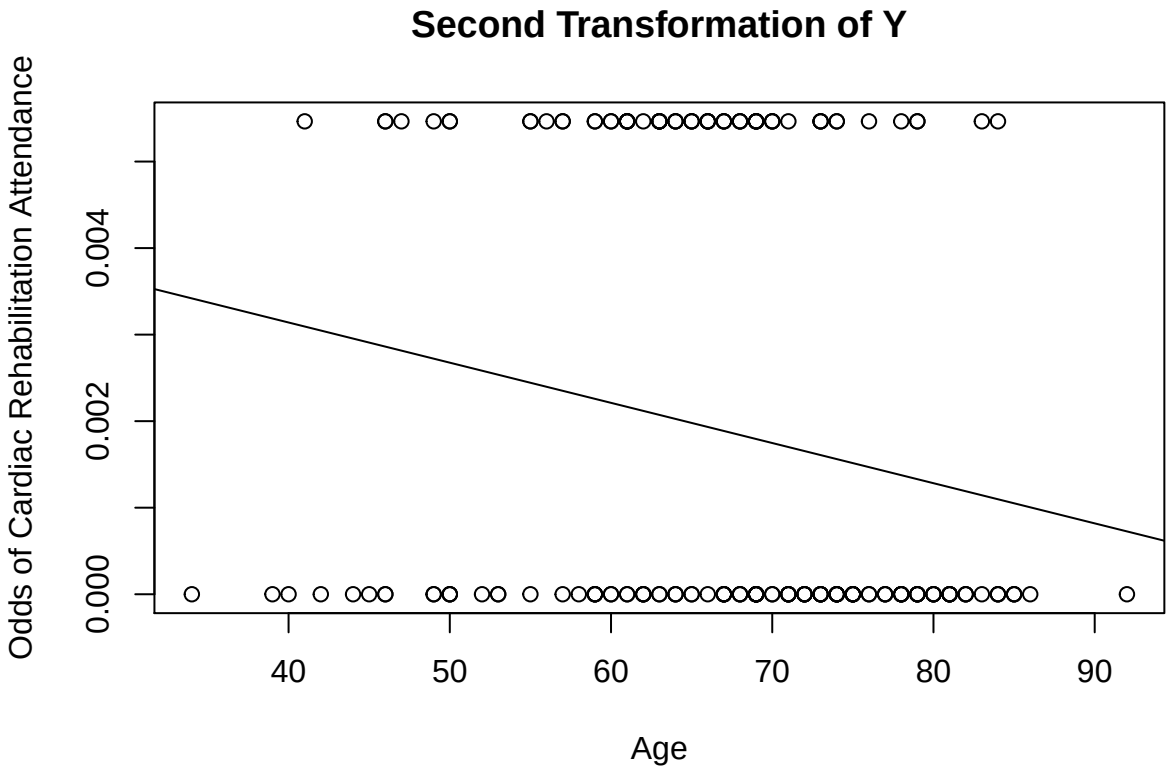
Let's try that transformation. We fit the model $Y'' = O(Y) = \dfrac{Y'}{1 - Y'} = \beta_0 + \beta_1 AGE$.

```
ATT.2PRIME <- ATT.1PRIME/(1-ATT.1PRIME)
CARDIAC2 <- data.frame(AGE, ATT.1PRIME, ATT.2PRIME)
plot(ATT.2PRIME ~ AGE,
     ylab = "Odds of Cardiac Rehabilitation Attendance",
```

```
     xlab = "Age",
     main = "Second Transformation of Y")
CARDIAC2.LM2 <- lm(ATT.2PRIME ~ AGE, data = CARDIAC2)
summary(CARDIAC2.LM2)
```

```
##
## Call:
## lm(formula = ATT.2PRIME ~ AGE, data = CARDIAC2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.003419 -0.001840 -0.001376  0.003299  0.004367
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.997e-03  1.181e-03   4.230 3.69e-05 ***
## AGE         -4.643e-05  1.732e-05  -2.681  0.00802 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.002557 on 182 degrees of freedom
## Multiple R-squared:  0.03799,    Adjusted R-squared:  0.0327
## F-statistic: 7.187 on 1 and 182 DF,  p-value: 0.00802
```

```
plot(ATT.2PRIME ~ AGE,
     ylab = "Odds of Cardiac Rehabilitation Attendance",
     xlab = "Age",
     main = "Second Transformation of Y")
abline(CARDIAC2.LM2)
```



The resulting model is $Y'' = O(Y) = 0.0050 - 4.643 \times 10^{-5} AGE$. Note that the intercept and sloe estimates are quite similar to the previous model. However, wew gain the advantage of being immune to impossible values above 1. However, this model is still problematic because it may produce estimates less than zero. Like probabilities, negative odds are impossibilities. Thus, we need to do one final transformation.

Note that our model range is now $0 < Y'' < +\infty$. We need a transformation such that our range becomes $-\infty < Y''' < +\infty$.

One simple transformation is to take the natural logarithm of Y''. That is, if $0 < Y'' < +\infty$ and we take the natural logarithm of all terms, $ln(0) < ln(Y'') < ln(+\infty)$ we arrive at the range we need: $-\infty < Y''' < +\infty$.

In short, the final transformation of Y is $Y''' = ln(O(Y)) = ln\left(\frac{Y'}{1-Y'}\right) = \beta_0 + \beta_1 AGE$. This final transformation – the natural logarithm of the odds – has a special name. We call this the *logit* transformation.

$Y''' = ln(O(Y)) = logit(Y) = \beta_0 + \beta_1 AGE$

The resulting regression technique is, therefore, called *logistic regression.*

Note all the transformation has occured in the left-hand side of the equation; the right hand side stays the same. Thus, the way we can use what we learned in the linear regression model to interpret the logistic regression model. This is why the logistic model is part of a group of models that involve transformations of the left-hand side of the equation while preserving the right-hand side. These model extensions are called *general linear models* or GLM.

In `R`, logistic regression uses the `glm()` function. It has a similar structure as the `lm()` function we used for linear regression. An important difference is that you should specify the `family` and `link` options under `glm()`. (Note that `glm` also allows you to use `factor` variables.)

```
CARDIAC.LM3 <- glm(ATT.F ~ AGE, data = CARDIAC,
                  family = binomial(link = 'logit'))
summary(CARDIAC.LM3)
```

```
##
## Call:
## glm(formula = ATT.F ~ AGE, family = binomial(link = "logit"),
##     data = CARDIAC)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4345  -0.8975  -0.7646   1.3674   1.7589
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.87467    0.98088   1.911  0.05598 .
## AGE         -0.03788    0.01462  -2.590  0.00959 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 236.48  on 183  degrees of freedom
## Residual deviance: 229.52  on 182  degrees of freedom
## AIC: 233.52
##
## Number of Fisher Scoring iterations: 4
```

The resulting equation is $logit(ATT) = 1.8747 - 0.0378AGE$.

Let us interpret each of the estimated model parameters starting with the intercept. In this case, $b_0 = 1.8747$. This means that this is the value of the $logit(ATT)$ when the woman at the time of her birth. This is out of the scope of our data, so it quite meaningless.

The slope coefficient is $b_1 = -0.0378$. For every year of life that a woman ages, the natural logarithm of the odds of attendance in a cardiac rehabilitation program decreases by 3.78%.

It is very hard for a lay person to understand the meaning of a logit, so we can back transform this value. Recall, $logit(ATT) = ln(O(ATT))$. We can re-interpret the logit as odds by $O(ATT) = e^{logit(ATT)} = e^{-0.0378} = 0.9629$. Thus, we can say that for every year of life that a woman ages, the odds of her attendance in a cardiac rehabilitation program is 96.29%. Another way to express this is that for every year of life, the odds of a woman participating in a cardiac rehabilitation program changes by $(O(ATT) - 1) \times 100\% = (0.9629 - 1) = -3.71\%$.

In statistics and epidemiology, $e^{logit(Y)}$ has a special name that you might have encountered before. It is the *odds ratio*. Thus, yet another way of expressing the findings above is to say that $OR = 0.9629$ or that the odds of participation decrease by 3.71% for every year of life lived.

IMPORTANT: For the most part, we will NOT use logit estimates, but convert to odds ratios.

Visualising the Model

What does the model look like if we were to plot it in the probability scale?

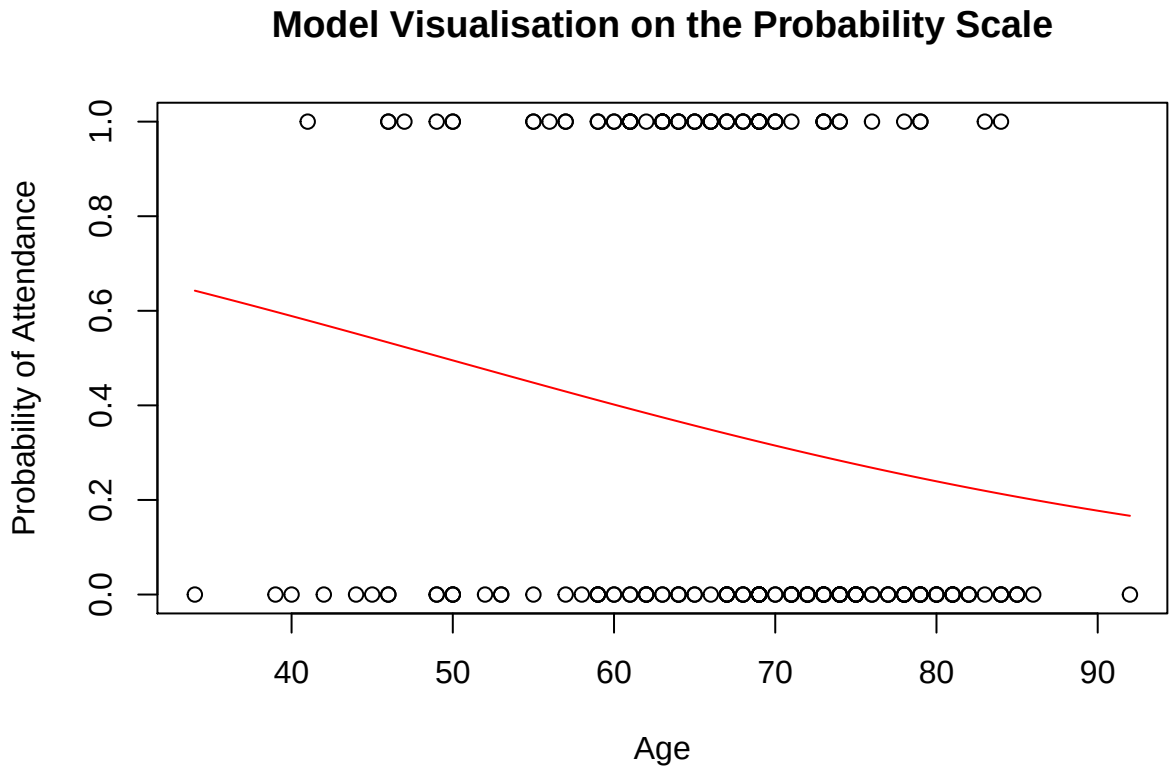
First, we must derive the model estimates for the range of the independent variable.

```
range(AGE)
```

```
## [1] 34 92
```

We use this information to produce a set of very fine points.

```
PRED.X <- seq(34, 92, 0.01)
PRED.Y <- predict(CARDIAC.LM3, list(AGE = PRED.X), type = "response")
plot(ATT ~ AGE,
     ylab = "Probability of Attendance",
     xlab = "Age",
     main = "Model Visualisation on the Probability Scale")
lines(PRED.X, PRED.Y, col = "red")
```

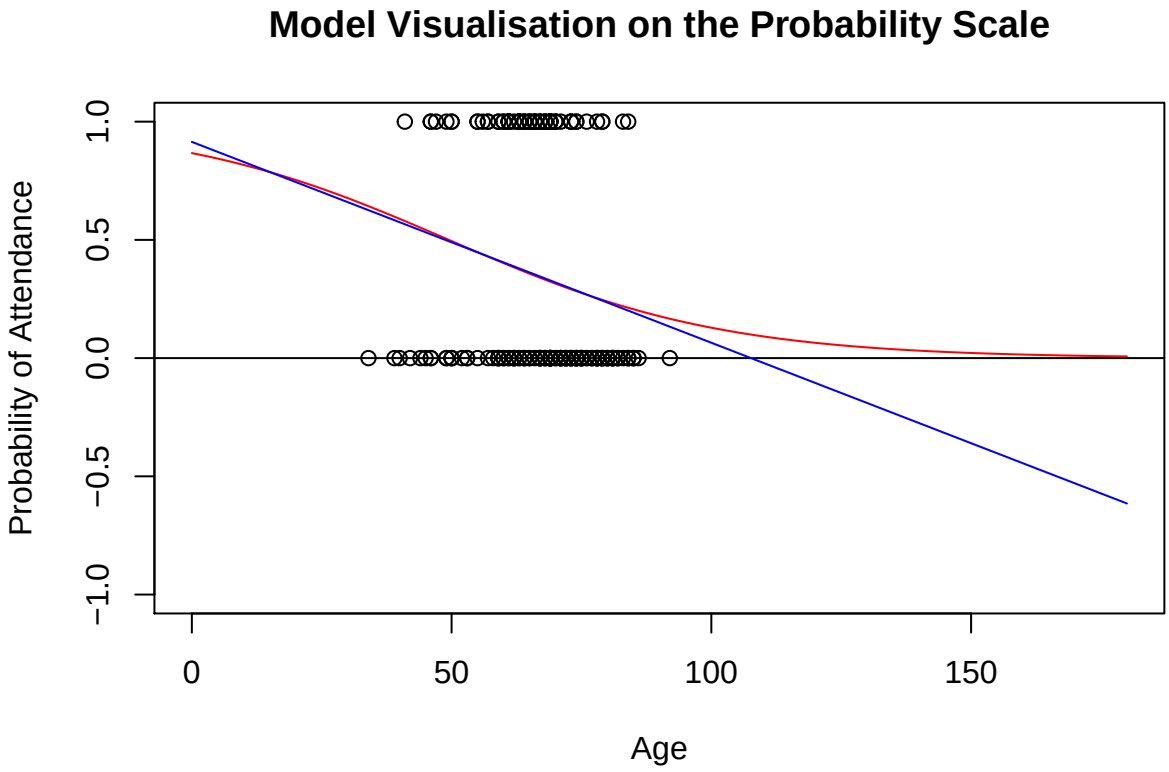


Note that the predicted line is actually probabilities estimated by the model are actually curved. This is because probabilities are bound by zero and one.

Look at what happens when I extend the horizontal axis.

```
PRED.X <- seq(0, 180, 0.01)
PRED.Y <- predict(CARDIAC.LM3, list(AGE = PRED.X), type = "response")
PRED.Y2 <- predict(CARDIAC.LM, list(AGE = PRED.X), type = "response")
plot(ATT ~ AGE,
     ylab = "Probability of Attendance",
     xlab = "Age",
     main = "Model Visualisation on the Probability Scale",
     xlim = range(0:180),
     ylim = range(-1:1))
abline(h = 0)
lines(PRED.X, PRED.Y, col = "red")
```

```
lines(PRED.X, PRED.Y2, col = "blue")
```



The shape of the logit function on the unit scale is S-shaped or *sigmoidal*. This is another important characteristic to note. In comparison, I have shown in blue the estimated probabilities derived from the simple linear regression model. Note how the estimated probabilies go below zero.

Confidence Intervals for the Parameter Estimates

Deriving confidence intervals for parameter estimates of the logistic regression model proceeds similarly as in linear regression.

Example 1B.

Produce 90% confidence intervals for the intercept and slope estimates.

```
confint(CARDIAC.LM3, level = 0.90)
```

```
## Waiting for profiling to be done...
```

```
##                    5 %        95 %
## (Intercept)  0.2761812  3.51716830
## AGE         -0.0624566 -0.01415749
```

The 90% confidence intervals for the intercept and slope are $(0.2762, 3.5172)$ and $(-0.0625, -0.0141)$, respectively. Remember, these values are in log odds or logit units.

```
exp(confint(CARDIAC.LM3, level = 0.90))
```

```
## Waiting for profiling to be done...
```

```
##                    5 %       95 %
## (Intercept) 1.3180867 33.6888963
## AGE         0.9394538  0.9859423
```

The 90% confidence interval for the odds ratio for age is (0.9394, 0.9859).

Producing predictions

The process of producing predictions based on the model proceeds in the same way as for linear regression models. It was demonstrated brielfy above. I will leave it to you to review the notes and apply the same techniques here.

Qualitative Independent Variables

What we learned about the interpretation of qualitative independent variables in previous lectures also holds here. Again, I leave you to review the notes and apply the techniques to logistic regression.

THE END