Titanic Regression

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
rawdata <- read.csv("/cloud/project/train.csv", header = TRUE)
mean_age <- mean(rawdata$Age, na.rm = TRUE)
```

Impute missing values for Age

```r
logistic_df <- rawdata %>%
  select(Survived,Pclass, Sex, Age) %>%
  mutate(Sex = as.factor(Sex), Pclass = as.factor(Pclass), Survived = as.factor(Survived)) %>%
  mutate(Age2 = ifelse(is.na(Age), mean_age, Age))
```

Train a logistic model

```r
model <- glm(Survived ~ Pclass + Sex + Age2, data = logistic_df, family = binomial)
```

Use the model to make predictions pos = Survived then add 1) predicted probabilites 2) calculate predicted survival count 3) logit log(p/(1-p))

```r
Prob = predict(model, type = "response")
logistic_df <- cbind(logistic_df,Prob)
logistic_df <- logistic_df %>% mutate(Predict = ifelse(Prob > 0.5, 1, 0))
logistic_df <- logistic_df %>% mutate(logit = log(Prob/(1-Prob)))
```

Predicted

```r
table(logistic_df$Predict)
```

```
##
##   0   1
## 568 323
```

Actual

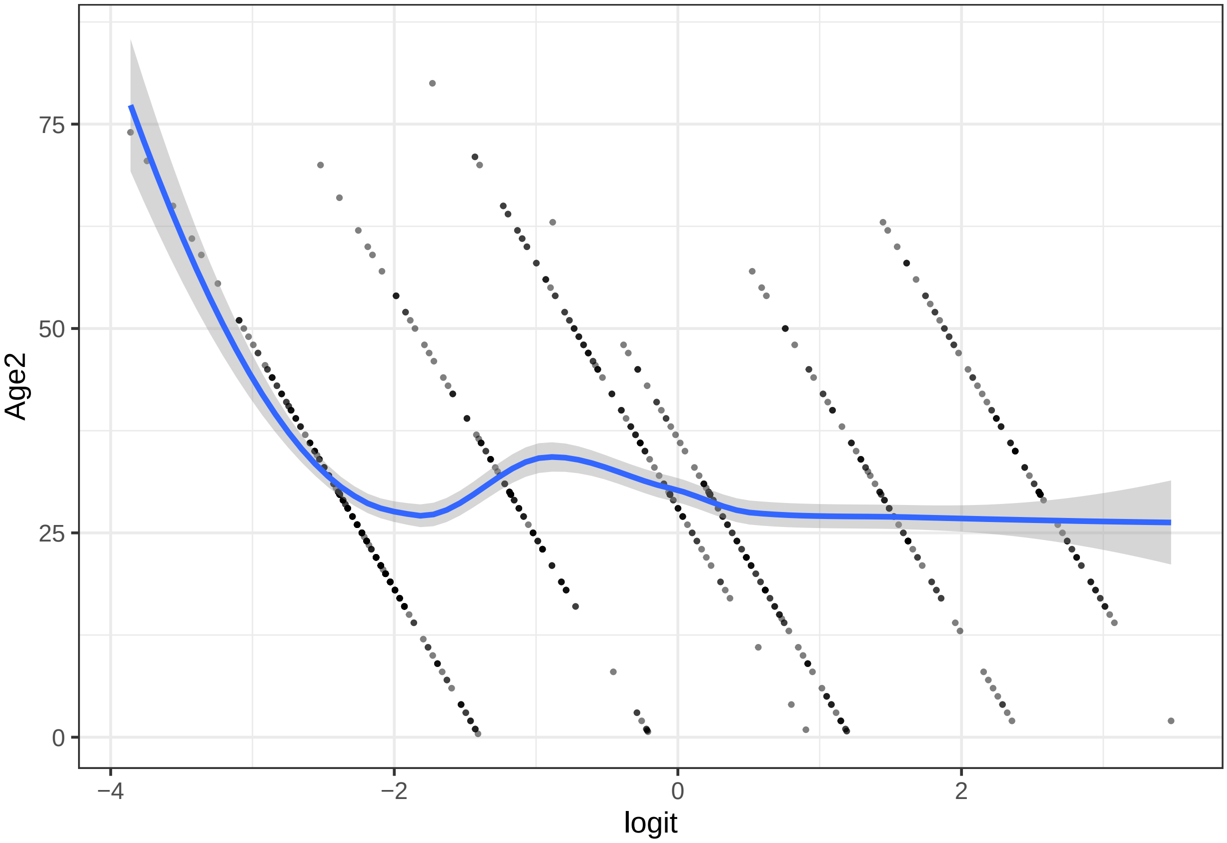
```r
table(logistic_df$Survived)
```

```
##
##   0   1
## 549 342
```

Plot the functional relationship between Age2 and the logit

```
ggplot(logistic_df, aes(logit, Age2))+
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```


Age2
logit