Exercise_7

Ingrid Canelles

2026-02-18

```
# load libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.6
## v forcats   1.0.1     v stringr   1.6.0
## v ggplot2   4.0.1     v tibble    3.3.1
## v lubridate 1.9.4     v tidyr     1.3.2
## v purrr     1.2.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(openintro)
```

```
## Loading required package: airports
## Loading required package: cherryblossom
## Loading required package: usdata
```

```
library(modelr)
```

```
fit_int <- glm(spam ~ num_char * re_subj,
               data = email,
               family = binomial())

summary(fit_int)
```

Fit the logistic model with interaction

```
##
## Call:
## glm(formula = spam ~ num_char * re_subj, family = binomial(),
##     data = email)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.45231    0.07314 -19.857  < 2e-16 ***
## num_char          -0.06343    0.00784  -8.091 5.92e-16 ***
## re_subj1          -2.63983    0.49928  -5.287 1.24e-07 ***
## num_char:re_subj1 -0.07698    0.09559  -0.805    0.421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2437.2  on 3920  degrees of freedom
## Residual deviance: 2160.1  on 3917  degrees of freedom
## AIC: 2168.1
##
## Number of Fisher Scoring iterations: 9
```

Interpretation of the logistic model with interaction:

The number of characters ($\beta_1 = -0.063$, p < 0.001) has a negative and highly significant effect on spam probability, meaning that longer emails are less likely to be classified as spam.

The presence of "Re:" in the subject line ($\beta_2 = -2.640$, p < 0.001) is also strongly associated with a lower probability of spam.

The interaction term between the number of characters and the presence of "Re:" ($\beta_3 = -0.077$, p $= 0.421$) is not statistically significant. This indicates that there is no strong evidence that the effect of email length differs depending on whether the subject contains "Re:".

Overall, the main effects are important predictors of spam, but the interaction does not meaningfully improve the model.

```r
b <- coef(fit_int)

effect_re0 <- b["num_char"]
effect_re1 <- b["num_char"] + b["num_char:re_subj1"]

effect_re0
```

Derive the effect of num_char in each group

```
##    num_char
## -0.06343252
```

```r
effect_re1
```

```
##   num_char
## -0.1404077
```

Interpretation of the effect of num_char in each group:

When $re\_subj = 0$, the effect of num_char on the log-odds of spam is $\beta_1 = -0.063$.

When $re\_subj = 1$, the effect becomes $\beta_1 + \beta_3 = -0.140$.

Thus, the negative effect of the number of characters is stronger when "Re:" is present in the subject line. However, this difference is not statistically significant.

```
OR_re0 <- exp(10 * effect_re0)
OR_re1 <- exp(10 * effect_re1)

OR_re0
```

Odds ratio for +10 characters

```
##  num_char
## 0.5302932
```

```
OR_re1
```

```
##  num_char
## 0.2455936
```

Interpretation of the odds ratios for +10 characters:

For emails without “Re:” ($re\_subj = 0$), increasing the number of characters by 10 multiplies the odds of being spam by $\exp(10\beta_1) = 0.53$, corresponding to a reduction of about 47% in the odds of spam.

For emails with “Re:” ($re\_subj = 1$), a 10-character increase multiplies the odds by $\exp(10(\beta_1 + \beta_3)) = 0.25$, implying a reduction of approximately 75% in the odds of spam.

```
summary(fit_int)$coef["num_char:re_subj1", ]
```

Is the interaction statistically significant?

```
##    Estimate  Std. Error     z value    Pr(>|z|)
## -0.07697519  0.09558506 -0.80530573  0.42064328
```

```
# Or more formally via Likelihood Ratio Test:
fit_no <- glm(spam ~ num_char + re_subj,
              data = email,
              family = binomial())

anova(fit_no, fit_int, test = "Chisq")
```

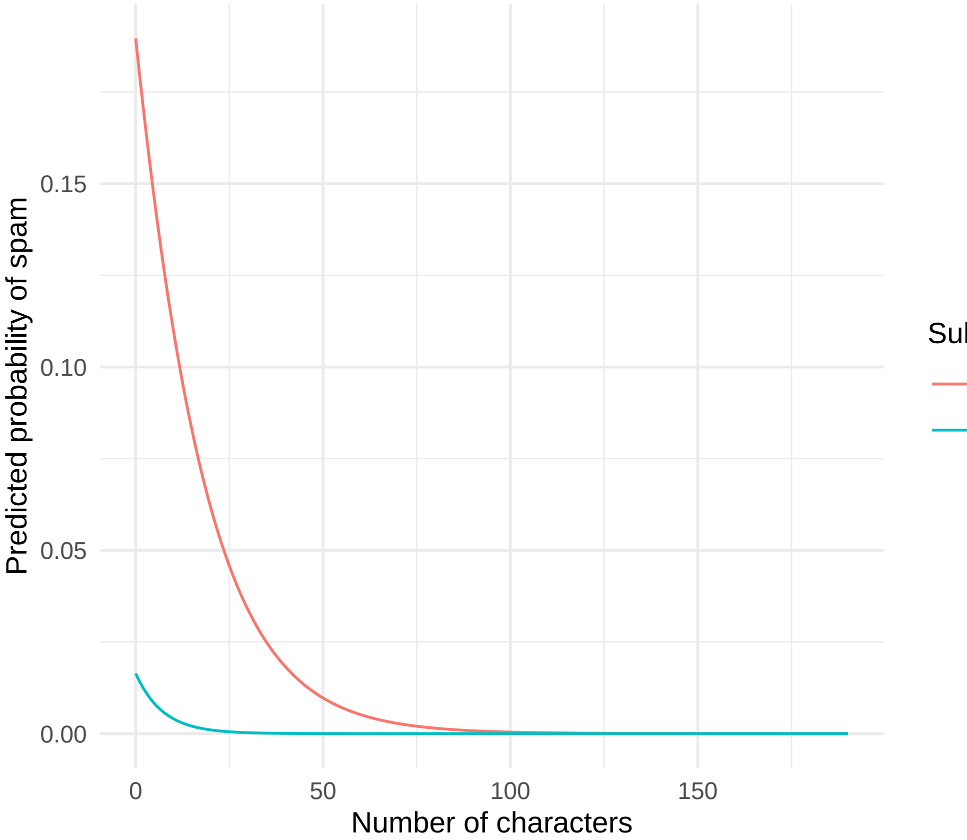
```
## Analysis of Deviance Table
##
## Model 1: spam ~ num_char + re_subj
## Model 2: spam ~ num_char * re_subj
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3918     2161.0
## 2      3917     2160.1  1  0.84088   0.3591
```

Interpretation of the interaction significance:

The interaction term between num_char and re_subj is not statistically significant (p = 0.421). This indicates that there is no strong evidence that the effect of the number of characters on spam probability differs depending on whether the subject contains “Re:”. The Likelihood Ratio Test confirms this result. Comparing the model with and without interaction yields a p-value of 0.359, which is well above 0.05. Therefore, adding the interaction does not significantly improve model fit. Overall, the simpler model without interaction is sufficient.

```r
# Create a proper prediction grid
grid <- data_grid(email,
                  num_char = seq(min(email$num_char),
                                 max(email$num_char),
                                 length.out = 200),
                  re_subj = levels(email$re_subj))

# Obtain predicted probabilities
grid$pred_int <- predict(fit_int,
                         newdata = grid,
                         type = "response")
# Plot predicted curves
ggplot(grid, aes(x = num_char, y = pred_int, color = re_subj)) +
  geom_line() +
  labs(x = "Number of characters",
       y = "Predicted probability of spam",
       color = "Subject contains 'Re:'") +
  theme_minimal()
```



Plot predicted spam probabilities

Interpretation of the predicted probability curves:

The predicted probability of spam decreases sharply as the number of characters increases, confirming the strong negative effect of email length. Emails without “Re:” in the subject (re_subj = 0) consistently have a higher predicted probability of spam compared to those with “Re:”. Although the slope appears slightly steeper when re_subj = 1, the curves are fairly similar overall, which aligns with the previous statistical results indicating that the interaction term is not significant. Therefore, while email length strongly reduces spam probability and “Re:” substantially lowers baseline spam risk, the difference in slopes between the two groups is not meaningful.