Survey Package and Regression

Samantha-Jo Caetano

February 25, 2025

Let's work through the following R code:

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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
```

Loading in Data

The `dpylr` package has some data about Star Wars characters. Let's assume it is a representative sample of all characters seen in Episodes 1 to 9.

```
starwars<-starwars

glimpse(starwars)
```

```
## Rows: 87
## Columns: 14
## $ name       <chr> "Luke Skywalker", "C-3PO", "R2-D2", "Darth Vader", "Leia Or~
## $ height     <int> 172, 167, 96, 202, 150, 178, 165, 97, 183, 182, 188, 180, 2~
## $ mass       <dbl> 77.0, 75.0, 32.0, 136.0, 49.0, 120.0, 75.0, 32.0, 84.0, 77.~
## $ hair_color <chr> "blond", NA, NA, "none", "brown", "brown, grey", "brown", N~
## $ skin_color <chr> "fair", "gold", "white, blue", "white", "light", "light", "~
## $ eye_color  <chr> "blue", "yellow", "red", "yellow", "brown", "blue", "blue",~
## $ birth_year <dbl> 19.0, 112.0, 33.0, 41.9, 19.0, 52.0, 47.0, NA, 24.0, 57.0, ~
## $ sex        <chr> "male", "none", "none", "male", "female", "male", "female",~
## $ gender     <chr> "masculine", "masculine", "masculine", "masculine", "femini~
## $ homeworld  <chr> "Tatooine", "Tatooine", "Naboo", "Tatooine", "Alderaan", "T~
## $ species    <chr> "Human", "Droid", "Droid", "Human", "Human", "Human", "Huma~
## $ films      <list> <"A New Hope", "The Empire Strikes Back", "Return of the J~
## $ vehicles   <list> <"Snowspeeder", "Imperial Speeder Bike">, <>, <>, <>, "Imp~
## $ starships  <list> <"X-wing", "Imperial shuttle">, <>, <>, "TIE Advanced x1",~
```

```
head(starwars)
```

```
## # A tibble: 6 x 14
```

```
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 Luke Sky~    172    77 blond      fair       blue            19   male  mascu~
## 2 C-3PO        167    75 <NA>       gold       yellow         112   none  mascu~
## 3 R2-D2         96    32 <NA>       white, bl~ red             33   none  mascu~
## 4 Darth Va~    202   136 none       white      yellow          41.9 male  mascu~
## 5 Leia Org~    150    49 brown      light      brown           19   fema~ femin~
## 6 Owen Lars    178   120 brown, gr~ light      blue            52   male  mascu~
## # i 5 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>
```

We will be working with `mass`, and `height` throughout this class, so let's remove the NAs. Additionally, we will be working with `species`, let's categorize species to be `human`, or `other`.

```
summary(starwars$mass)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   15.00   55.60   79.00   97.31   84.50 1358.00      28
```

```
summary(starwars$height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    66.0   167.0   180.0   174.6   191.0   264.0       6
```

```
table(starwars$species)
```

```
##
##         Aleena       Besalisk         Cerean       Chagrian       Clawdite
##             1              1              1              1              1
##          Droid            Dug           Ewok      Geonosian         Gungan
##             6              1              1              1              3
##          Human           Hutt       Iktotchi        Kaleesh       Kaminoan
##             35             1              1              1              2
##        Kel Dor       Mirialan   Mon Calamari           Muun       Nautolan
##             1              2              1              1              1
##      Neimodian         Pau'an       Quermian         Rodian        Skakoan
##             1              1              1              1              1
##      Sullustan     Tholothian        Togruta          Toong      Toydarian
##             1              1              1              1              1
##     Trandoshan        Twi'lek     Vulptereen        Wookiee          Xexto
##             1              2              1              2              1
## Yoda's species         Zabrak
##             1              2
```

```
starwars_clean <- starwars %>%
  filter(!is.na(mass)) %>%
  filter(!is.na(height)) %>%
  mutate(species_clean = case_when(
    species == "Human" ~ "human",
    species == "Droid" ~ "droid",
    species != "Human" & species != "Droid" ~ "other")) %>%
  filter(!is.na(species_clean))

starwars_clean
```

```
## # A tibble: 56 x 15
##    name     height  mass hair_color skin_color eye_color birth_year sex   gender
```

```
##    <chr>     <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
##  1 Luke Sk~    172    77 blond      fair       blue            19   male  mascu~
##  2 C-3PO       167    75 <NA>       gold       yellow         112   none  mascu~
##  3 R2-D2        96    32 <NA>       white, bl~ red             33   none  mascu~
##  4 Darth V~    202   136 none       white      yellow          41.9 male  mascu~
##  5 Leia Or~    150    49 brown      light      brown           19   fema~ femin~
##  6 Owen La~    178   120 brown, gr~ light      blue            52   male  mascu~
##  7 Beru Wh~    165    75 brown      light      blue            47   fema~ femin~
##  8 R5-D4        97    32 <NA>       white, red red             NA   none  mascu~
##  9 Biggs D~    183    84 black      light      brown           24   male  mascu~
## 10 Obi-Wan~    182    77 auburn, w~ fair       blue-gray       57   male  mascu~
## # i 46 more rows
## # i 6 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>, species_clean <chr>
```

```
n=nrow(starwars_clean)
```

Task 1: Create a Simple Linear Model to predict mass

Use the `svyglm()` function in the `survey` library to run a linear regression estimation of mass given height. Assume it was a Simple Random Sample and use the finite population correction with $N = 224$.

```
library(survey)
```

```
## Loading required package: grid
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
## Loading required package: survival
##
## Attaching package: 'survey'
## The following object is masked from 'package:graphics':
##
##     dotchart
```

```
N=224
n=nrow(starwars_clean)

fpc.srs = rep(N, n)

starwars.design <- svydesign(id=~1, data=starwars_clean, fpc=fpc.srs)

mysvylm <- svyglm(mass ~ height, starwars.design)
summary(mysvylm)
```

```
##
## Call:
## svyglm(formula = mass ~ height, design = starwars.design)
##
## Survey design:
```

```
## svydesign(id = ~1, data = starwars_clean, fpc = fpc.srs)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.24762   21.15216  -0.532    0.597
## height        0.62893    0.07741   8.125 6.21e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 29705.79)
##
## Number of Fisher Scoring iterations: 2
```

What is the model in this example?

$$y_{mass} = \beta_0 + \beta_1 x_{height} + \epsilon$$

This is a "simple linear regression" model (because there is only one numeric x variable).

What is the estimate of the model?

$$\hat{y}_{mass} = \hat{\beta}_0 + \hat{\beta}_1 x_{height}$$
$$\hat{y}_{mass} = -11.25 + 0.63 x_{height}$$

Task 2: Create a Linear Model to predict mass

Use the `lm()` function in the `survey` library to run a linear regression estimation of mass given height.

```
#install.packages("survey")
library(survey)

## Using the Survey Library
summary(lm(mass ~ height, data=starwars_clean))
```

```
##
## Call:
## lm(formula = mass ~ height, data = starwars_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -62.14  -30.42  -22.47  -18.75 1259.19
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.2476   114.3454  -0.098    0.922
## height        0.6289     0.6434   0.977    0.333
##
## Residual standard error: 173.9 on 54 degrees of freedom
## Multiple R-squared:  0.01739,    Adjusted R-squared:  -0.0008114
## F-statistic: 0.9554 on 1 and 54 DF,  p-value: 0.3327
```

What is the model in this example?

$$y_{mass} = \beta_0 + \beta_1 x_{height} + \epsilon$$

This is a "simple linear regression" model (because there is only one numeric x variable).

What is the estimate of the model?

$$\hat{y}_{mass} = \hat{\beta}_0 + \hat{\beta}_1 x_{height}$$
$$\hat{y}_{mass} = -11.25 + 0.63 x_{height}$$

What is the different between the output here and the outcome in Task 1? What is similar?

Standard errors are different, but estimates are the same.

What happens if you change the N in Task 1? Try setting $N = 87, 224, 1000, 10000$

Task 3: Create a Linear Model to predict mass

Run a linear regression estimation of mass given height and species.

```
## Using the Survey Library
fpc.srs = rep(N, n)

starwars.design <- svydesign(id=~1, data=starwars_clean, fpc=fpc.srs)

mysvylm <- svyglm(mass ~ height + species_clean, starwars.design)
summary(mysvylm)
```

```
##
## Call:
## svyglm(formula = mass ~ height + species_clean, design = starwars.design)
##
## Survey design:
## svydesign(id = ~1, data = starwars_clean, fpc = fpc.srs)
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -21.3802    16.3062  -1.311    0.196
## height               0.6509     0.1122   5.799 3.99e-07 ***
## species_cleanhuman -14.6399    10.1665  -1.440    0.156
## species_cleanother  20.1821    35.3932   0.570    0.571
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 29431.98)
##
## Number of Fisher Scoring iterations: 2
```

```
## Using lm
mymodel<-lm(mass ~ height + species_clean, data=starwars_clean)
summary(mymodel)
```

```
##
## Call:
## lm(formula = mass ~ height + species_clean, data = starwars_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -76.43  -39.53  -19.94   -2.05 1245.29
```

```
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -21.3802   129.6105  -0.165    0.870
## height               0.6509     0.6782   0.960    0.342
## species_cleanhuman -14.6399   100.4204  -0.146    0.885
## species_cleanother  20.1821    96.4256   0.209    0.835
##
## Residual standard error: 176.4 on 52 degrees of freedom
## Multiple R-squared:  0.02644,    Adjusted R-squared:  -0.02972
## F-statistic: 0.4708 on 3 and 52 DF,  p-value: 0.7039
```

What is the model in this example?

$$y_{mass} = \beta_0 + \beta_1 x_{height} + \beta_2 x_{human} + \beta_3 x_{other} + \epsilon$$

This is a "simple linear regression" model (because there is only one numeric x variable).

What is the estimate of the model?

$$\hat{y}_{mass} = \hat{\beta}_0 + \hat{\beta}_1 x_{height} + \hat{\beta}_2 x_{human} + \hat{\beta}_3 x_{other}$$
$$\hat{y}_{mass} = -21.38 + 0.65 x_{height} - 14.64 x_{human} + 20.18 x_{other}$$

Note, the variables x_{human} and x_{other} are "dummy" variables. They are coded such that they indicate whether or not the character is in that species category.

Based on the output, what is the estimated mass of a human character who is 175 units tall?

```
-21.3802493 + 0.6509304 *175 - 14.6399467*1 +20.1821107*0
```

```
## [1] 77.89262
```

```
predict(mysvylm, tibble(height=175, species_clean="human"))
```

```
##     link     SE
## 1 77.893 3.1774
```

```
predict(mymodel, tibble(height=175, species_clean="human"))
```

```
##        1
## 77.89262
```

Based on the output, what is the estimated mass of a droid character who is 175 units tall?

```
-21.3802493 + 0.6509304 *175 - 14.6399467*0 +20.1821107*0
```

```
## [1] 92.53257
```

```
predict(mysvylm, tibble(height=175, species_clean="droid"))
```

```
##     link     SE
## 1 92.533 9.4044
```

```
predict(mymodel, tibble(height=175, species_clean="droid"))
```

```
##        1
## 92.53256
```

What is the expected difference in the mass of a human vs a droid character of the same height?

```
77.89262-92.53256 = -14.63994
```

We expect human characters to have a mass of 14.64 units lower than droid characters of the same height.

Task 4: Create a Logistic Regression Model to predict if mass exceeds 100lb.

Run a logistic regression model mass being over 100lbs given height and species.

```
## Create a new variable called `mass_over100`.

starwars_clean <- starwars_clean %>%  mutate(
  mass_over100 = case_when(
    mass > 100 ~ 1, ## 1 = Yes, over 100lbs
    mass <= 100 ~ 0))


## Using lm
my_glm <- glm(mass_over100 ~ height + species_clean,
              data=starwars_clean,
              family = "binomial")
summary(my_glm)
```

```
##
## Call:
## glm(formula = mass_over100 ~ height + species_clean, family = "binomial",
##     data = starwars_clean)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -11.45548    4.69328  -2.441   0.0147 *
## height               0.06231    0.02430   2.565   0.0103 *
## species_cleanhuman  -2.14452    1.76197  -1.217   0.2236
## species_cleanother  -2.04237    1.74018  -1.174   0.2405
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 49.375  on 55  degrees of freedom
## Residual deviance: 36.612  on 52  degrees of freedom
## AIC: 44.612
##
## Number of Fisher Scoring iterations: 6
```

What is the model in this example?

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{height} + \beta_2 x_{human} + \beta_3 x_{other}$$

What is the estimate of the model?

$$log\Big(\frac{\hat{p}}{1-\hat{p}}\Big) = \hat{\beta}_0 + \hat{\beta}_1 x_{height} + \hat{\beta}_2 x_{human} + \hat{\beta}_3 x_{other}$$

$$log\Big(\frac{\hat{p}}{1-\hat{p}}\Big) = -11.45 + 0.06 x_{height} - 2.14 x_{human} - 2.04 x_{other}$$

Based off the above output, what is the predicted probability of a human who is 175cm tall being over 100lbs?

Answer: 0.05729392

```
pred_odds <- -11.45547800 + 0.06230662*175 -2.14452294*1 -2.04236670*0
exp(pred_odds)/(1+exp(pred_odds))
```

```
## [1] 0.06318953
```

```
pred_odds2 <- predict(my_glm, tibble(species_clean="human", height=175))
exp(pred_odds2)/(1+exp(pred_odds2))
```

```
##          1
## 0.06318953
```

```
pred_prob <- predict(my_glm, tibble(species_clean="human", height=175), type = "response")
pred_prob
```

```
##          1
## 0.06318953
```

```
starwars.design <- svydesign(id=~1, data=starwars_clean, fpc=fpc.srs)

mysvyglm <- svyglm(mass_over100 ~ height + species_clean,
              family = "binomial", starwars.design)

summary(mysvyglm)
```

```
##
## Call:
## svyglm(formula = mass_over100 ~ height + species_clean, design = starwars.design,
##     family = "binomial")
##
## Survey design:
## svydesign(id = ~1, data = starwars_clean, fpc = fpc.srs)
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -11.45548    3.97846  -2.879  0.00577 **
## height               0.06231    0.02150   2.897  0.00550 **
## species_cleanhuman  -2.14452    1.03816  -2.066  0.04386 *
## species_cleanother  -2.04237    1.05966  -1.927  0.05940 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 0.8722449)
##
## Number of Fisher Scoring iterations: 6
```

Task 5: Create a logistic regression for Stratified Random Sampling

For the starwars data, let's assume the study instead collected the data by stratifying by species, where in the population there 16 droids, 80 humans and 128 other species. Use the finite population correction with N=224 to estimate.

Run a logistic regression model mass being over 100lbs given height and species.

```
starwars_clean <- starwars_clean %>%
  mutate(pop_fpc = case_when(species_clean=="droid" ~ 16,
                             species_clean=="human" ~ 80,
                             species_clean=="other" ~ 128)) %>%
```

```r
  mutate(samp_wt = case_when(species_clean=="droid" ~ 16/4,
                             species_clean=="human" ~ 80/20,
                             species_clean=="other" ~ 128/32))


strata.design <- svydesign(id=~1,
                       strata=~species_clean,
                       weights = ~samp_wt,
                       fpc=~pop_fpc,
                       data=starwars_clean)

mysvyglm <- svyglm(mass_over100 ~ height + species_clean,
              family = "binomial", strata.design)

summary(mysvyglm)
```

```
##
## Call:
## svyglm(formula = mass_over100 ~ height + species_clean, design = strata.design,
##     family = "binomial")
##
## Survey design:
## svydesign(id = ~1, strata = ~species_clean, weights = ~samp_wt,
##     fpc = ~pop_fpc, data = starwars_clean)
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -11.45548    3.99707  -2.866  0.00607 **
## height               0.06231    0.02149   2.899  0.00554 **
## species_cleanhuman  -2.14452    1.14044  -1.880  0.06588 .
## species_cleanother  -2.04237    1.15723  -1.765  0.08369 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 0.8722449)
##
## Number of Fisher Scoring iterations: 6
```

Logistic Regression UCLA analysis of SRS with population of N=6000 code

```
mydata <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")

## Standard Logistic Regression
mylogit<-glm(admit ~ gre + gpa +
               as.factor(rank), data=mydata, family="binomial")
summary(mylogit)
```

```
##
## Call:
## glm(formula = admit ~ gre + gpa + as.factor(rank), family = "binomial",
##     data = mydata)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -3.989979   1.139951  -3.500 0.000465 ***
## gre               0.002264   0.001094   2.070 0.038465 *
## gpa               0.804038   0.331819   2.423 0.015388 *
## as.factor(rank)2 -0.675443   0.316490  -2.134 0.032829 *
## as.factor(rank)3 -1.340204   0.345306  -3.881 0.000104 ***
## as.factor(rank)4 -1.551464   0.417832  -3.713 0.000205 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 499.98  on 399  degrees of freedom
## Residual deviance: 458.52  on 394  degrees of freedom
## AIC: 470.52
##
## Number of Fisher Scoring iterations: 4
```

```
## log(p/(1-p)) = -3.99 + 0.002*gre + 0.804*gpa-0.675*x3
##                           -1.34*x4 - 1.55*x5


## Survey Estimation for Logistic Regression
n=length(mydata$admit)
N=6000

#install.packages("survey")
library(survey)
## Using the Survey Library
fpc.srs = rep(N, n)

ucla.design <- svydesign(id=~1, data=mydata, fpc=fpc.srs)

mysvyglm <- svyglm(admit ~ gre + gpa + as.factor(rank),
                   ucla.design, family="binomial")
summary(mysvyglm)
```

```
##
```

```
## Call:
## svyglm(formula = admit ~ gre + gpa + as.factor(rank), design = ucla.design,
##     family = "binomial")
##
## Survey design:
## svydesign(id = ~1, data = mydata, fpc = fpc.srs)
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -3.989979   1.099498  -3.629 0.000322 ***
## gre               0.002264   0.001065   2.126 0.034156 *
## gpa               0.804038   0.333433   2.411 0.016348 *
## as.factor(rank)2 -0.675443   0.303806  -2.223 0.026764 *
## as.factor(rank)3 -1.340204   0.332843  -4.027 6.79e-05 ***
## as.factor(rank)4 -1.551464   0.401947  -3.860 0.000133 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 0.9962149)
##
## Number of Fisher Scoring iterations: 4
```