R Workshop: Linear & Logistic Regression

Brier Gallihugh, M.S.

2023-06-19

Table of contents

An Introduction to Regression

Linear Regression

```
library(tidyverse)
library(car)

data <- starwars %>% select(height,mass,sex) %>% na.omit()          ①
```

① The above code takes the `starwars` data set and uses the `select()` function to pull out the columns labelled height, mass and sex. Then using the `na.omit()` function, I've removed (using listwise deletion) any rows which contain missing values

Running a Linear Regression

Similar to correlation, regression (particularly multiple regression) is very common in the social sciences. As such, lets dive into an example again using the `starwars` data set.

```
data <- data %>%
  filter(sex == "male" | sex == "female") %>%                    ①
  mutate(sex = dplyr::recode(sex,                                ②
                             "male" = 1,
                             "female" = 0))

linear_regression <- lm(height ~ sex + mass, data = data)        ③

summary(linear_regression)                                       ④
```

① Here I want to filter out the data set for observations (rows) that meet the following conditions (in this case those with a sex "value" of either "male" or "female").
② I then want to recode the sex variables to a numeric value for the purposes of doing the linear regression using the `recode()` function. It takes the column as an input and takes the syntax "old value" to "new value".
③ This is a basic linear regression formula using the `lm()` function. It takes the form DV ~ IV + IV, df).
④ To see the results of the regression, we simply run the `summary()` function and specify the name we assigned the regression (i.e., linear_regression).

```
Call:
lm(formula = height ~ sex + mass, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-47.431  -3.892   2.007   8.108  48.526

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.8554     9.6656  11.779 4.91e-16 ***
sex         -18.0741     8.5390  -2.117   0.0393 *
mass          1.0144     0.1167   8.694 1.43e-11 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.78 on 50 degrees of freedom
Multiple R-squared:  0.6056,    Adjusted R-squared:  0.5898
F-statistic: 38.38 on 2 and 50 DF,  p-value: 7.938e-11
```

Assumptions of Linear Regression

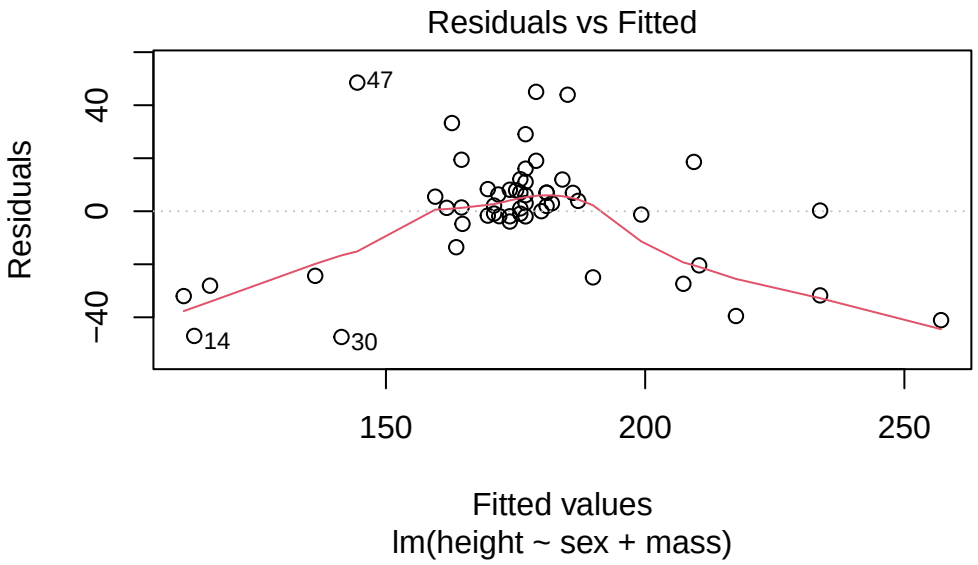
As some will attest to, we as social scientists don't always explicitly test our statistical assumptions (this is a problem). However, as a parametric test, regression consists of the following key assumptions: homogeneity of variance, residual normality, lack of multicollinearity, and the independence of errors. Further, we likely want to at least investigate the potentiality that there are outliers and influential cases. I'm going to show you how to test each of these assumptions.

Homogeneity of Variance

Graphical

```r
# Should Be a Straight Line
plot(linear_regression,1)                                    ①
```

① The `plot(object,1)` function when applied to a regression object will give you a plot assessing visually the homogeneity of variance assumption. The line should be roughly straight.



Statistical

```
library(lmtest)
# Goldfield Quandt Test (Less than .05 BAD)
gqtest(linear_regression)                                    ①
```

① Statistically, we can also investigate homogeneity of variance using a Goldfield Quandt Test. This is one test we hope is not statistically significant. To do this, we use the `gqtest()` function in the `lmtest` package.

```
    Goldfeld-Quandt test

data:  linear_regression
GQ = 1.1014, df1 = 24, df2 = 23, p-value = 0.4096
alternative hypothesis: variance increases from segment 1 to 2
```

Residual Normality

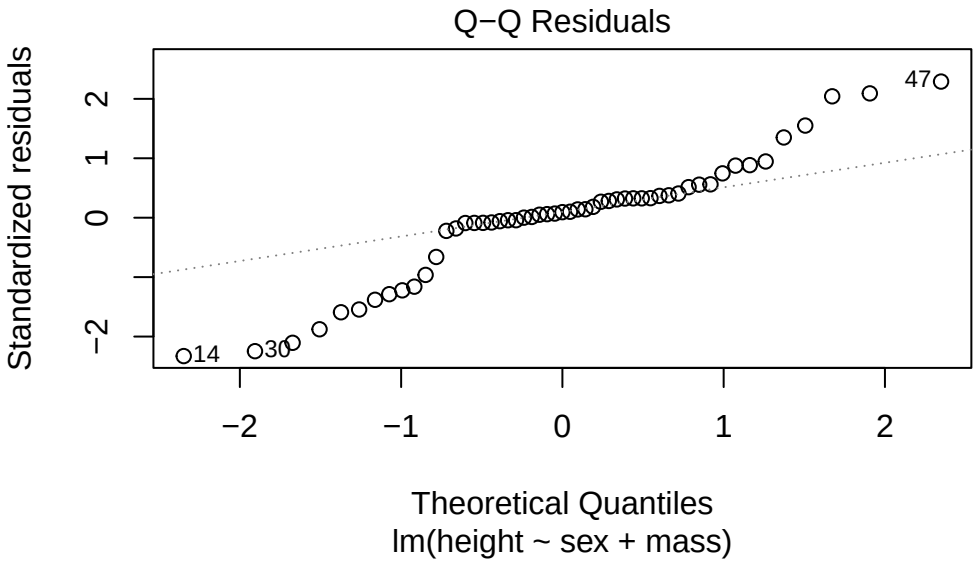
Contrary to popular belief, normality doesn't typically refer to the distribution of each individual variable in a regression. What matters is that the model residuals will be roughly normally

distributed. Below we will go through this. First, we'll look at the assumption graphically using what is known as a qq plot

Graphical

```
# Should Follow Straight Diagonal Line
plot(linear_regression,2)                                    ①
```

① Use the `plot(object,2)` with a regression will give you a standard qqplot with a reference line. This line should look linear.



Statistical

We can also assess the assumption statistically using a Shapiro Wilks test. Keep in mind this test can be heavily influenced by sample size but nonetheless it is a start.

```
shapiro.test(residuals(linear_regression))                   ①
```

① The `shapiro.test()` function will get us what we want. However, we need to make sure we get the residuals of the model so we need to use the `residuals()` function too.

```
    Shapiro-Wilk normality test

data:  residuals(linear_regression)
W = 0.93558, p-value = 0.006747
```

Multicollinearity

Multicollinearity (or too high of correlation between variables) can be an issue. One could look at the correlation matrix but that's highly subjective. A better idea might be to use what is known as variance inflation factors. Essentially higher values (i.e., north of 10) indicate an issue with said factor (i.e., variable). Below is the code for this using the `car` package.

Statistical (VIF)

```
# Individual Predictors (Less than 10 is OK)
car::vif(linear_regression)                                    ①

# Mean Across Predictors (Ideally Around 1)
mean(vif(linear_regression))                                   ②

# Tolerance (Greater than .20 Ideal)
1/vif(linear_regression)                                       ③
```

① Less than 10 is solid for this metric
② You want the mean value to be around 1.0
③ You want tolerance to be $>$ .20

```
   sex   mass
1.1485 1.1485
[1] 1.1485
      sex      mass
0.8707007 0.8707007
```

Independence of Errors

Long story short, errors shouldn't be correlated with each other. We can assess this statistically using the Durbin Watson test. Here we want the test to not be statistically significant. We can use the `durbinWatsonTest()` function from the `car` function for this. The code is shown below.

Statistical (Durbin Watson Test)

```
library(car)
car::durbinWatsonTest(linear_regression)
```

```
 lag Autocorrelation D-W Statistic p-value
   1      0.09534174      1.807892   0.498
 Alternative hypothesis: rho != 0
```

Finding Outliers & Influential Cases

Residuals

Typically, residuals outside of the range of -2.5 to 2.5 are a potential problem. However, this doesn't mean we should discard the data necessarily. We have to feel they are distinctly not in the population we're hoping to investigate. However, typically as a rule of thumb, having more than 5% of your cases being outside of this range is not ideal.

```
# Greater than 5% of Data Potentially Problematic
data$residuals <- resid(linear_regression)
summary(data$residuals)

problem_residuals <- data %>% filter(residuals > 2.5 | residuals < -2.5)①

nrow(problem_residuals)/nrow(data)                                    ②
```

① This will filter the data set into observations which meet the criteria
② This tells me what percent of the cases are potentially problematic

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-47.431   -3.892    2.007    0.000    8.108   48.526
[1] 0.7358491
```

Influential Cases

Maybe a more useful investigation is that involving influential cases (i.e., cases which have undue influence on the results). There are several metrics one can you. I'm just going to focus on Cooks distance. Essentially values over 1 are potentially problematic. Below is the code for this

```r
data$cooks_dist <- round(cooks.distance(linear_regression),5)          ①
# Greater Than 1 is Problematic
summary(data$cooks_dist)                                               ②
```

① This rounds the values to 5 decimal places for ease of reading in addition to calculating the Cooks distance for each observation.

② `summary()` function says the max value is .36 so there are no individual cases here that are having undue influence on the results.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.00016 0.00205 0.03116 0.02715 0.36309
```

Logistic Regression

So logistic regression is something I had to learn in order to do this workshop. Essentially however, logistic regression is used to give odds ratios of whether an observation belongs in one of two categories based on a set of inputs. In industry, this is actually considered a form of machine learning (so is most other regression). You can actually go above and beyond two categories but we're not going to talk about polynomial regression in this workshop. Below is how we run a logistic regression in R.

Running a Logistic Regression

```r
# Create Logistic Model w/ Multiple Predictors
log_regression <- glm(sex ~ height + mass, family = binomial(link = "logit"), data = data)
summary(log_regression)                                                ②

# Determine Chi Square Diff
modelChi <- log_regression$null.deviance - log_regression$deviance     ③
modeldf <- log_regression$df.null - log_regression$df.residual
chisq_prob <- 1 - pchisq(modelChi,modeldf)
print(chisq_prob)

# Odds Ratios
exp(log_regression$coefficients)                                       ④

# CI
exp(confint(log_regression))                                           ⑤

# Effect Size
```

```
effect_size <- modelChi/log_regression$null.deviance                    ⑥
print(effect_size)
```

① This is your basic logistic regression formula. The `family = binomial(link = "logit")` tells R that we want this model to be a logistic regression.
② Summarizes the results of the logistic regression
③ Calculates a Chi Square Test to determine if the model is better than chance (less than .05)
④ Exponentiation the coefficients will give you the odds ratios
⑤ We can also get the confidence intervals of the odds ratios using the `confint()` function
⑥ We can calculate an effect size using this formula

```
Call:
glm(formula = sex ~ height + mass, family = binomial(link = "logit"),
    data = data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.27693    2.11194   1.552  0.12075
height      -0.06519    0.02458  -2.652  0.00799 **
mass         0.14554    0.04719   3.084  0.00204 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 48.292  on 52  degrees of freedom
Residual deviance: 30.705  on 50  degrees of freedom
AIC: 36.705

Number of Fisher Scoring iterations: 6

[1] 0.0001517089
(Intercept)      height        mass
 26.4943310   0.9368878   1.1566654
                2.5 %       97.5 %
(Intercept) 0.5524946 3056.5379217
height      0.8856929    0.9779379
mass        1.0685581    1.2942792
[1] 0.3641807
```

Assumptions of Logistic Regression

Like regular regression, logistic regression also has assumptions that must be met. They are the linearity of each predictor with the **log** of the outcome, independence of errors, and multicollinearity. I will show you how to test each of these below.

Linearity w/ Log of Outcome (Each Predictor)

Statistical

```
data$massINT <- data$mass*log(data$mass)                                    ①
data$heightINT <- data$height*log(data$height)

linearity_assumption <- glm(sex ~ height + mass + massINT + heightINT, family = binomial(l

summary(linearity_assumption)
```

① You want the interaction terms of each predictor with the log of the outcome. These need to go into the logistic regression.
② You can see them added here. You want none of the interaction terms to be statistically significant

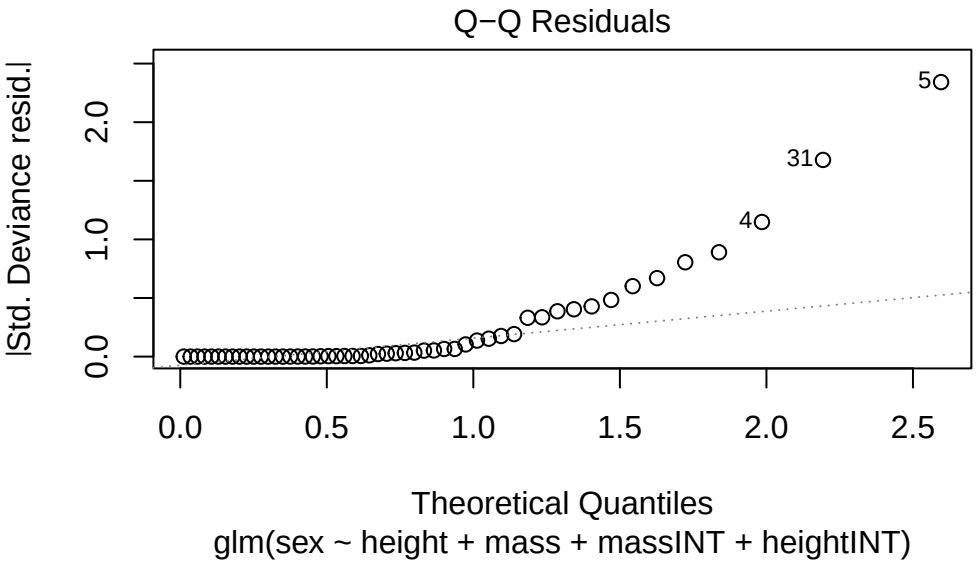
```
Call:
glm(formula = sex ~ height + mass + massINT + heightINT, family = binomial(link = "logit"),
    data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 1042.3027   703.0356   1.483   0.1382
height       -41.2936    26.8629  -1.537   0.1242
mass           5.3989     3.2018   1.686   0.0918 .
massINT       -0.9727     0.5926  -1.641   0.1007
heightINT      6.7461     4.3829   1.539   0.1238
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 48.2922  on 52  degrees of freedom
Residual deviance:  8.4973  on 48  degrees of freedom
AIC: 18.497
```

```
Number of Fisher Scoring iterations: 13
```

Graphical

You can also just plot the logistic regression with the interaction terms and see if you get a roughly straight line.

```
plot(linearity_assumption,2)
```



Independence of Errors

Like regular regression, we can test this assumption using a Durbin Watson Test. The code here is below.

Statistical

```
durbinWatsonTest(log_regression)
```

```
 lag Autocorrelation D-W Statistic p-value
   1       0.0998344      1.776247   0.462
 Alternative hypothesis: rho != 0
```

Multicollinearity

Multicollinearity can also be assessed just like regression by using the `vif()` function as well as the `1/vif()` function. We want these to be less than 10 and greater than .20 respectively.

Statistical

```
vif(log_regression)
```

```
  height     mass
4.812415 4.812415
```

```
1/vif(log_regression)
```

```
   height      mass
0.2077959 0.2077959
```

Finding Outliers & Influential Cases

Residuals

Residuals can be assessed the same was as it is in regular linear regression. The code is below as a refresher.

```
# Greater than 5% of Data Potentially Problematic
data$residuals_log <- resid(log_regression)
summary(data$residuals_log)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.6335  0.0324  0.2731  0.1136  0.3751  2.1968
```

```
log_prob <- data %>% filter(residuals_log > 2 | residuals_log < -2)
```

```
round(nrow(log_prob)/nrow(data),3)
```

```
[1] 0.038
```

Influential Cases

As with residuals and linear regression compared to logistic regression, the same is true for influential cases. You can see the code for this below.

```
data$cooks_dist_log <- cooks.distance(log_regression)
# Greater Than 1 is Problematic
summary(data$cooks_dist_log)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
0.0000000 0.0002653 0.0006915 0.0267272 0.0199929 0.5227436 
```