Regression Assignment

Robin Donatello

Last Updated 2018-11-17 12:10:57

Assignment Overview

You will perform 3 regression analyses in this assignment. The variable types for the coefficients are pre-specified so that you can practice interpretations of different types of variables. **You must use the variable types listed here**

1. Multiple Linear Regression: $Q \sim Q + B$ (one quantitative and one binary predictor)
2. Logistic Regression: $logit(B) \sim Q + B$
3. Either of the two above analyses (or a new model) add a third categorical (more than 2 levels) variable e.g.: $Q \sim Q + B + C$. (one quantitative, one binary, one categorical)

Instructions

0. Use the template provided: [RMD] for R users, and [Word] for SPSS users.
1. Identify variables under consideration.
2. Write the mathematical model being fit.

   SPSS users use the Equation editor in Word to create these. R users write the equation directly below in the Rmarkdown file using LaTeX script (example below).

```
$$y_{i} \sim \beta_{0} + \beta_{1}x_{1i} + \beta_{2}x_{2i} + \epsilon_{i}$$
```

3. Fit the model in your software program of choice.
    - Include confidence intervals for the coefficients.
4. Interpret all regression coefficients except the intercept.
    - For logistic regression, calculate and interpret the Odds Ratios

---

Multiple Linear Regression

1. Identify variables

If you have a "Strongly Agree" to "Strongly Disagree" variable that you have kept all 5 levels, you can treat it as a Quantitative Variable.

- Quantitative outcome (y): Income (variable `income`).
- Quantitative predictor (x_1): Time you wake up in the morning (variable `wakeup`)
- Binary predictor (x_2): Gender of individual as an indicator of being female (variable `gender`, 0=male, 1=female)

2. Write the mathematical model

$$y_i \sim \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

3. Fit the multivariable model

Don't forget to calculate the confidence interval for each coefficient to use in your conclusion.

```
lm.mod <- lm(income ~ wakeup + female_c, data=addhealth)
summary(lm.mod)
```

```
##
## Call:
## lm(formula = income ~ wakeup + female_c, data = addhealth)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36047 -15141  -5252   8678 205610
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     48669.4     1206.9  40.325  < 2e-16 ***
## wakeup           -611.3      149.4  -4.092 4.37e-05 ***
## female_cFemale  -8527.1      789.3 -10.803  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24300 on 3810 degrees of freedom
##   (2691 observations deleted due to missingness)
## Multiple R-squared:  0.03236,    Adjusted R-squared:  0.03185
## F-statistic:  63.7 on 2 and 3810 DF,  p-value: < 2.2e-16
```

```
confint(lm.mod)
```

```
##                      2.5 %     97.5 %
## (Intercept)     46303.1191 51035.7225
## wakeup           -904.2448  -318.4088
## female_cFemale -10074.5798 -6979.5823
```

> Optional new package `stargazer` for printing regression models as columns. Great for comparisons, looks like journal articles. Your R code chunk header must look like this: ` ```{r, results='asis'} ` and be sure to use the correct output format: `type='html'` or `type='latex'`. Vignette found at: https://www.jakeruss.com/cheatsheets/stargazer/

```
library(stargazer)
stargazer(lm.mod, type='html', ci=TRUE, single.row=TRUE, digits=1, omit.stat="rsq")
```

Dependent variable:

income

wakeup

-611.3*** (-904.2, -318.5)

female_cFemale

-8,527.1*** (-10,074.1, -6,980.1)

Constant

48,669.4*** (46,303.9, 51,035.0)

Observations

3,813

Adjusted R2

0.03

Residual Std. Error

24,297.2 (df = 3810)

F Statistic

63.7*** (df = 2; 3810)

Note:

p<0.1; ***p<0.05;*** p<0.01

4. Interpret the regression coefficients.

- b_1: Holding gender constant, for every hour later a person wakes up, their predicted average income drops by 611 (318, 904) dollars. This is a significant association ($p < .01$).
- b_2: Controlling for the time someone wakes up in the morning, the predicted average income for females is 8,527 (6980, 10,074) dollars lower than for males. This is a significant association ($p < .01$).

Logistic Regression

Your outcome variable must be coded as 1 (event) and 0 (non-event). Recoding this way ensures you are predicting the presence of your categorical variable and not the absence of it.

1. Identify variables

- Binary outcome (y): Poverty (variable `poverty`). This is an indicator if reported personal income is below $10,210.
- Quantitative predictor (x_1): Time you wake up in the morning (variable `wakeup`)
- Binary predictor (x_2): Gender (variable `female_c`)

2. Write the mathematical model

$$logit(y_i) \sim \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

3. Fit the multivariable model

```r
log.mod <- glm(poverty~wakeup + female_c, data=addhealth, family='binomial')
summary(log.mod)
```

```
##
## Call:
## glm(formula = poverty ~ wakeup + female_c, family = "binomial",
##     data = addhealth)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0597  -0.7703  -0.5423  -0.5141   2.1124
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.18642    0.11857 -18.439  < 2e-16 ***
## wakeup          0.04587    0.01351   3.396 0.000683 ***
## female_cFemale  0.84822    0.07660  11.074  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4909.2  on 4832  degrees of freedom
## Residual deviance: 4772.7  on 4830  degrees of freedom
##   (1671 observations deleted due to missingness)
## AIC: 4778.7
##
## Number of Fisher Scoring iterations: 4
```

4. Interpret the Odds Ratio estimates

The regression coefficients b_p from a logistic regression must be *exponentiated* before interpretation. This is done by raising the constant e to the value of the coefficient. So, $OR = e^b$. Below I create a table containing the odds ratio estimates and 95% CI for those estimates using the confounding model.

```r
# For your assignment - replace the saved model object `log.mod` with whatever YOU named this model.
data.frame(
  OR  = exp(coef(log.mod)),
  LCL = exp(confint(log.mod))[,1],
  UCL = exp(confint(log.mod))[,2]
  ) %>%
  kable(digits=2, align = 'ccc')
```

	OR	LCL	UCL
(Intercept)	0.11	0.09	0.14
wakeup	1.05	1.02	1.07
female_cFemale	2.34	2.01	2.72

You will see one of three things:

- **OR = 1** = equal chance of response variable being YES given any explanatory variable value. You are not able to predict participants' responses by knowing their explanatory variable value. This would be a non significant model when looking at the p-value for the explanatory variable in the parameter estimate table.
- **OR > 1** = as the explanatory variable value increases, the presence of a YES response is more likely. We can say that when a participant's response to the explanatory variable is YES (1), they are more likely to have a response that is a YES (1).
- **OR <1** = as the explanatory variable value increases, the presence of a YES response is less likely. We can say that when a participant's response to the explanatory variable is YES (1) they are less likely to have a response that is a YES (1).

- After controlling for gender, those that wake up one hour later have 1.05 (1.02, 1.07) times the odds of reporting annual earned wages below the federal poverty level compared to someone waking up one hour earlier. This is a significant association ($p < .001$), but the magnitude of the increase is very small.
- After controlling for the time someone wakes up, females have 2.34 (2.01, 2.72) times the odds of reporting annual earned wages below the federal poverty level compared to males. This is a significant association ($p < .001$)

Categorical predictors

For any of the regression models above, or a new model if you choose, add a categorical variable with more than 2 levels as a *third* predictor. Be sure to define EACH indicator variable for this categorical variable and state what the reference group is.

1. Identify variables and their data type

- Response (y): BMI (variable `BMI`). This is a quantitative measure.
- Predictor(x_1): Income (variable `income`). This is a quantitative measure.
- Predictor(x_2): Smoking status (variable `eversmoke_c`). This is a binary measure.
- Predictor: general health (variable `genhealth`). This is a categorical measure with 5 levels.

2. Write the mathematical model.

Define what each x is, and write the mathematical model. State what group is the reference group.

- Let x_1 be `income`
- Let x_2 be `eversmoke_c`, an indicator of ever smoking.
- Let $x_3 = 1$ when `genhealth='Very good'`, and 0 otherwise,
- let $x_4 = 1$ when `genhealth='Good'`, and 0 otherwise,
- let $x_5 = 1$ when `genhealth='Fair'`, and 0 otherwise,
- let $x_6 = 1$ when `genhealth='Poor'`, and 0 otherwise.

The reference group for `genhealth` is `Excellent`.

The mathematical model would look like:

$$y_i \sim \beta_0 + \beta_1 * x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + \epsilon_i$$

3. Fit the multivariable model.

```r
gh.model <- lm(BMI~income + eversmoke_c + genhealth, data=addhealth)
summary(gh.model)
```

```
##
## Call:
## lm(formula = BMI ~ income + eversmoke_c + genhealth, data = addhealth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.620  -4.767  -0.988   3.508  39.448
##
```

```
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                  2.712e+01  3.583e-01  75.681  < 2e-16 ***
## income                      -5.061e-06  4.678e-06  -1.082    0.279    
## eversmoke_cSmoked at least once -1.016e+00  2.369e-01  -4.289 1.84e-05 ***
## genhealthVery good           1.659e+00  3.099e-01   5.354 9.13e-08 ***
## genhealthGood                4.864e+00  3.249e-01  14.968  < 2e-16 ***
## genhealthFair                7.041e+00  5.047e-01  13.950  < 2e-16 ***
## genhealthPoor                9.461e+00  1.389e+00   6.810 1.13e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.952 on 3763 degrees of freedom
##   (2734 observations deleted due to missingness)
## Multiple R-squared:  0.09693,    Adjusted R-squared:  0.09549 
## F-statistic: 67.32 on 6 and 3763 DF,  p-value: < 2.2e-16
```

```r
round(confint(gh.model),1)
```

```
##                                 2.5 % 97.5 %
## (Intercept)                      26.4   27.8
## income                            0.0    0.0
## eversmoke_cSmoked at least once  -1.5   -0.6
## genhealthVery good                1.1    2.3
## genhealthGood                     4.2    5.5
## genhealthFair                     6.1    8.0
## genhealthPoor                     6.7   12.2
```

4. Interpret the regression coefficients.

- b_1: After controlling for general health and smoking status, for every additional \$1 a person makes annually, their BMI decreases .0000047. This is not a significant relationship. A more meaningful interpretation would be to look at a \$1000 increase in annual income. For every additional \$1,000,000 in income a person makes annually, their BMI decreases by 4.7.
- b_2: After controlling for income level and general health, those who have smoked at least once have on average 1.02 (0.6, 1.5, p<.0001) lower BMI compared to those who have never smoked.
- b_3: After controlling for income level and smoking status, those reporting very good health have 1.7 (1.1, 2.3, p<.0001) higher BMI compared to those reporting excellent health.
- b_4: After controlling for income level and smoking status, those reporting good health have 4.9 (4.2, 5.5, p<.0001) higher BMI compared to those reporting excellent health.
- b_5: After controlling for income level and smoking status, those reporting fair health have 7.0 (6.1, 8.0 p<.0001) higher BMI compared to those reporting excellent health.
- b_6: After controlling for income level and smoking status, those reporting poor health have 9.5 (6.7, 12.2, p<.0001) higher BMI compared to those reporting excellent health.