Regression Assignment

Robin Donatello

Last Updated 2017-11-06 14:28:25

Assignment Overview

You will perform 2 regression analyses in this assignment. For each you will interpret the regression coefficients, and test for a potential confounder Z.

1. Multiple Linear Regression: $Q \sim X + Z$
   - For this assignment choose a quantitative X and a binary confounder Z, where the bivariate $Q \sim Q$ relationship is significant.
2. Logistic Regression: $logit(B) \sim X + Z$
   - Your binary response variable Y must be coded as 1 (event) and 0 (non-event).
   - For this assignment, choose a binary X and a binary confounder Z.

You will then take one of the above analyses (or a new model) and

4. Add a third categorical (more than 2 levels) variable e.g.: $Q \sim Q + C$.

Submission Guidelines

- Use the template provided: [RMD]
- This is an individual assignment. I expect you to work with your team/partner on this, but each person must submit their own work, and own writing.
- Upload to the corresponding assignment on Blackboard by the due date.
- Share and learn from others by post a copy of this assignment to this Google Drive folder as you complete each model.

Instructions

1. Identify variables under consideration
   - Determine a third variable Z that you want to test as a potential confounder.
   - Consider the relationship and ask yourself: "*If I had to predict a future persons response based on the Predictor/Explanatory variable and some other variable(s), what would they be?*"
   - For the purposes of this assignment your potential confounder Z must be binary. In reality, confounders, covariates, mediators, and moderators can come in all data types.
2. Write out the null, alternative, and confounder Hypotheses statements.
   - **Null** - that there is no relationship between response and explanatory variables
   - **Alternative** - that there is a relationship between response and explanatory variables.
   - **Confounder** - that there is a relationship between response and explanatory variables after controlling for the confounding variable.
3. Fit the simple model
   - Model the response variable on the explanatory variable `y ~ x`
   - Determine if you are going to reject the null in favor of the alternative.
4. Fit the multivariable model.
   - Only do this if you rejected the null.
   - Model the response variable on the explanatory variable and the third variable. `y ~ x + z`
   - Determine if Z is a confounder by looking at the p-value for the explanatory variable.

  - If it is still significant, the third variable is not a confounding variable.
  - If it is no longer significant, the third variable is a confounding variable. This means that the third variable is explaining the relationship between the explanatory variable and the response variable.
5. Interpret all regression coefficients except the intercept.
6. Write a conclusion. A structured response template is provided.

```
library(knitr)
opts_chunk$set(warning=FALSE, message=FALSE)
library(dplyr)
library(ggplot2); library(gridExtra)
library(pander) # Used for printing nice linear model tables
panderOptions("digits", 3)
load("E:/MATH315/Project/addhealth_clean.Rdata")
```

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: Income (variable `income`).
- Quantitative predictor: Time you wake up in the morning (variable `wakeup`)
- Binary confounder: Gender (variable `female_c`)

2. State the research Hypothesis

- Null: There is no relationship between the time you wake up and your personal earnings
- Alternative: There is a relationship between the time you wake up and your personal earnings
- Confounder: There is still a relationship between the time you wake up and your personal earnings, after controlling for gender.

3. Fit the simple model

Is there a relationship between income and time a person wakes up?

```
lm.mod1 <- lm(income ~ wakeup, data=addhealth)
pander(summary(lm.mod1), digits=2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	43548	1126	39	3e-276
wakeup	-488	151	-3.2	0.0013

Table 2: Fitting linear model: income ~ wakeup

Observations	Residual Std. Error	R^2	Adjusted R^2
3814	24666	0.00272	0.00246

The estimate of the regression coefficient for `wakeup` is significant (b1=-488, p= 0.001). There is reason to believe that the time you wakeup is associated with your income.

4. Fit the multivariable model

Fit the same multiple linear regression model and include the potential confounding variable. Determine if the third variable is a confounder.

```
lm.mod2 <- lm(income ~ wakeup + female_c, data=addhealth)
pander(summary(lm.mod2), digits=2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	48669	1207	40	2.1e-296
wakeup	-611	149	-4.1	4.4e-05
female_cFemale	-8527	789	-11	8.1e-27

Table 4: Fitting linear model: income ~ wakeup + female_c

Observations	Residual Std. Error	R^2	Adjusted R^2
3813	24297	0.0324	0.0318

The relationship between income and wake up time is still significant after controlling for gender. Gender is not a confounder.

5. Interpret the regression coefficients.

- b_1: Holding gender constant, for every hour later a person wakes up, their predicted average income drops by $611.
- b_2: Controlling for the time someone wakes up in the morning, the predicted average income for females is $8,527 lower than for males.

6. Conclusion

- Use the numerical results from the multivariable model to fill in the values in the conclusion below.
- Look at the Adjusted R^2 to see how much of the variance in the response you are accounting for with the predictor.
- To get the confidence intervals for the regression coefficients you will use the function `confint()`

```
kable(confint(lm.mod2), digits=2)
```

	2.5 %	97.5 %
(Intercept)	46303.12	51035.72
wakeup	-904.24	-318.41
female_cFemale	-10074.58	-6979.58

Replace the **bold** words with your variables, the `highlighted` words with data from your analysis, and choose between *conclusion options*.

After adjusting for the potential confounding factor of **third variable**, **explanatory variable** (b1 = `parameter estimate, CI confidence interval range, p = significance value`) was *__significantly/not significantly__ and **positively/negatively** associated with **response variable**. Approximately `R-Square*100` of the variance of **response** can be accounted for by **explanatory** after controlling for **third variable**. Based on these analyses, **third variable** *is not/is* a confounding factor because the association between **explanatory** and **response** *is still/is no longer* significant after accounting for **third variable**.

So the conclusion for this analysis reads:

> After adjusting for the potential confounding factor of **gender**, **wake up time** (b_1 = `-611, 95% CI: (-904, -318), p<.0001`) was **significantly** and **negatively** associated with **income**. Approximately `3.2%` of the variance of **income** can be accounted for by **wake up time** after controlling for **gender**. Based on these analyses, **gender** *is not* a confounding factor because the association between **wake up time** and **income** *is still* significant after accounting for **gender**.

Additional example interpretation

> After adjusting for the potential confounding factor of gender, an adolescent's weight (Beta = 1.34, 95% CI -0.53, 3.21, p = .1558) was not significantly associated with the number of cigarettes smoked in the past 30 days. Approximately 0.78% of the variance in cigarettes smoked can be accounted for by weight after controlling for gender. Based on these analyses, gender is a confounding factor because the association between weight and cigarettes smoked is no longer significant after accounting for gender.

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: Poverty (variable `poverty`). This is an indicator if reported personal income is below $10,210.
- Binary predictor: Ever smoked a cigarette (variable `eversmoke_c`)
- Binary confounder: Gender (variable `female_c`)

2. State hypotheses

- Null hypothesis: There is no relationship between the probability of living below the poverty level ever being a smoker.
- Alternative hypothesis: There is a relationship between the probability of living below the poverty level and ever being a smoker.
- Confounding hypothesis: There *still is* a relationship between the probability of living below the poverty level and ever being a smoker, after controlling for gender.

3. Fit the simple model

Fit the logistic regression model (a.k.a generalized linear model) of the explanatory variable on the response variable. Decide to reject the null hypothesis in favor of the alternative.

```r
log.mod.1 <- glm(poverty~eversmoke_c, data=addhealth, family='binomial')
summary(log.mod.1)
```

```
##
## Call:
## glm(formula = poverty ~ eversmoke_c, family = "binomial", data = addhealth)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7064  -0.7064  -0.7064  -0.6210   1.8659
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.54800    0.06444 -24.021  < 2e-16 ***
## eversmoke_cSmoker  0.28711    0.07737   3.711 0.000207 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4907.4  on 4834  degrees of freedom
## Residual deviance: 4893.3  on 4833  degrees of freedom
##   (1669 observations deleted due to missingness)
## AIC: 4897.3
##
## Number of Fisher Scoring iterations: 4
```

The p-value for the b1 estimate of the regression coefficient for `eversmoke_c` is significant at 0.0002. There is reason to believe that smoking status is associated with the probability of living below the poverty level.

4. Fit the multivariable model

Fit the same logistic regression model and include the potential confounding variable. **This is only done if there is a significant relationship between the explanatory and response variable.** Determine if the third variable is a confounder.

```r
log.mod.2 <- glm(poverty~eversmoke_c + female_c, data=addhealth, family='binomial')
summary(log.mod.2)
```

```
##
## Call:
## glm(formula = poverty ~ eversmoke_c + female_c, family = "binomial",
##     data = addhealth)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8335  -0.7048  -0.5652  -0.4716   2.1221
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.14045    0.08662  -24.71  < 2e-16 ***
## eversmoke_cSmoker  0.38725    0.07886    4.91 9.09e-07 ***
## female_cFemale     0.87450    0.07690   11.37  < 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: 4906.9  on 4833  degrees of freedom
## Residual deviance: 4755.4  on 4831  degrees of freedom
##   (1670 observations deleted due to missingness)
## AIC: 4761.4
##
## Number of Fisher Scoring iterations: 4
```

The p-value for the regression coefficient estimate of `eversmoke_c` is still significant at <.0001 after controlling for gender. Thus gender is **not** a confounder.

5. 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. You will see one of three things:

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

	OR	LCL	UCL
(Intercept)	0.12	0.10	0.14
eversmoke_cSmoker	1.47	1.26	1.72
female_cFemale	2.40	2.06	2.79

```
# Try to figure out what the functions coef() and confint() do?
# Type coef(log.mod.2) and conf(log.mod.2) into the console and make sure you understand what output it
```

- **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).

  For this example, the OR for the explanatory variable of eversmoker is 1.47 and the OR for gender is 2.40.

Interpreting confidence intervals for Odds Ratios

- Confidence intervals are a range for the population's predicted odds ratio based on the sample data. We are 95% confident that any given population's odds ratio would range between those two values.
- When the confidence intervals for the explanatory variables and third variables do not overlap, the variable with the higher values we can interpret as being more strongly associated with our response variable.
- For both the odds ratio and confidence interval interpretation, when you add in third variables it is explained in the same way except that you are controlling for the third variable or explanatory variable.
- When you add a third variable to the logistic regression, if you determine one is a confound then you do not interpret the variable that becomes non significant in the odds ratio.

  After controlling for gender, smokers have 1.47 times the odds of reporting making below the federal poverty level compared to non smokers. After controlling for smoking status, females have 2.4 time the odds of reporting annual earned wages below the federal poverty level compared to males. Gender is a stronger predictor of earning a lower wage than smoking status.

6. Conclusion

Replace the **bold** words with your variables, the `highlighted` words with data from your analysis, and choose between *conclusion options*.

After adjusting for the potential confounding factor of **third variable**, **explanatory variable** (`OR odds ratio estimate, CI confidence interval range, p = significance value`) was *significantly/not significantly* and *positively/negatively* associated with the likelihood of **response variable**. In this analysis, the odds ratio tells us that those who are [**describe what dummy code 1 of your explanatory variable means here**] are `0.05` times *more (if OR greater than 1)/less (if OR less than 1)* likely to [**describe what dummy code 1 of your response variable means here**]. Based on these analyses, **third variable** *is not/is* a confounding factor because the association between **explanatory** and **response** *is still/is no longer* significant after accounting for **third variable**.

So the conclusion for this analysis reads:

After adjusting for the potential confounding factor of **gender**, **smoking status** (`1.47, CI 1.26-1.72, p <.0001`) was *significantly* and *positively* associated with the likelihood of **earning under the poverty level**. In this analysis, the odds ratio tells us that those who **have ever smoked** are `1.47` times *more* likely to **earn income below the federal poverty level**. Based on these analyses, **gender** *is not* a confounding factor because the association between **smoking** and **poverty status** *is still* significant after accounting for **gender**.

Additional example interpretation

- After adjusting for the potential confounding factor of gender, being overweight (OR 0.920, CI 0.822 – 1.028, p = .1420) was not significantly associated with the likelihood of participating in an active sport. In this analysis, the odds ratio tells us that those adolescents who are overweight are 0.920 times less likely to participate in an active sport. Based on these analyses, gender is a confounding factor because the association between being overweight and active sport participation is no longer significant after accounting for gender.
- After adjusting for the potential confounding factor of gender, being overweight (OR 3.65, CI 1.573 – 4.891, p = .0001) was significantly and positively associated with the likelihood of participating in an active sport. In this analysis, the odds ratio tells us that those adolescents who are overweight are 3.65 times more likely to participate in an active sport. Based on these analyses, gender is not a confounding factor because the association between being overweight and active sport participation is still significant after accounting for gender.

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. * If the confounder was significant, use the multivariable model including confounder. * If the confounder was not significant, use the bivariate model without the confounder. * Interpret the regression coefficients for at least two levels of the categorical variable.

1. Identify variables and their data type

- Outcome: BMI (variable `BMI`). This is a quantitiative measure.
- Predictor: Income (variable `income`). This is a quantitative measure.
- Predictor: general health (variable `genhealth`). This is a categorical measure.

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 $x_2 = 1$ when `genhealth='Very good'`, and 0 otherwise,
- let $x_3 = 1$ when `genhealth='Good'`, and 0 otherwise,
- let $x_4 = 1$ when `genhealth='Fair'`, and 0 otherwise,
- let $x_5 = 1$ when `genhealth='Poor'`, and 0 otherwise.

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

The mathematical model would look like:

$$Y \sim \beta_0 + \beta_1 * x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

3. State hypothesis in words and symbols

- Null: General health is not associated with BMI after controlling for income. All levels of general health have the same relationship with income.
- Alternative: General health is associated with BMI after controlling for income - at least one level of General health has a different association with income.

$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

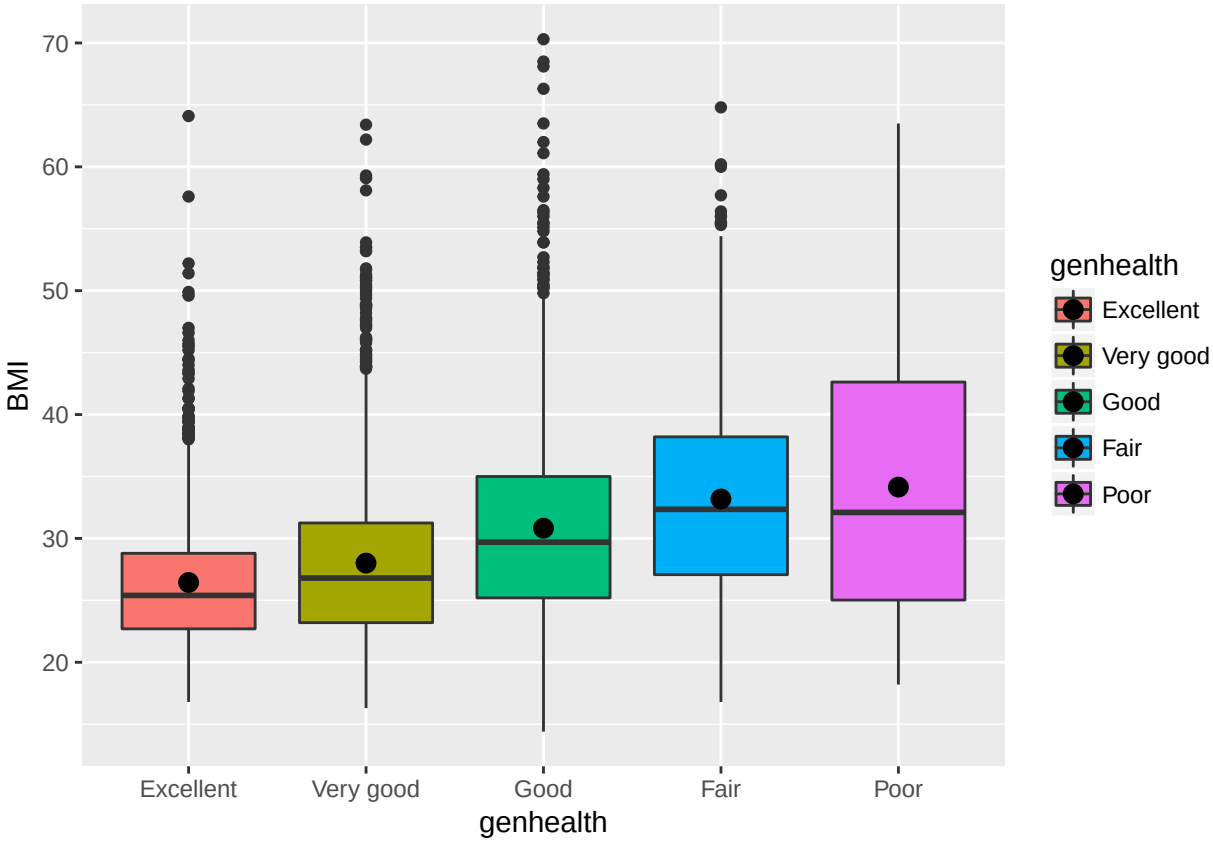
H_A : at least one β_j is not 0

3. Visualize this relationship

Create a graphic to explore the relationship between the categorical variable and the response variable.

- If your outcome is quantitative, this will be a side by side boxplot
- If your outcome is binary, this will be side by side barcharts

```
toplot <- addhealth %>% select(BMI, genhealth) %>% na.omit()
ggplot(toplot, aes(y=BMI, x=genhealth, fill=genhealth)) + geom_boxplot() +
    stat_summary(fun.y="mean", geom="point", size=3, position=position_dodge(width=0.75))
```



I would expect that the average BMI differs significantly across general health category.

4. Fit the multivariable model with both predictors.

Print out the coefficients and 95% CI's.

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

```
##
## Call:
## lm(formula = BMI ~ income + genhealth, data = addhealth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.837  -4.802  -1.091   3.441  39.132
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.652e+01  3.298e-01  80.409  < 2e-16 ***
## income             -4.735e-06  4.686e-06  -1.010    0.312
## genhealthVery good  1.602e+00  3.100e-01   5.166 2.52e-07 ***
## genhealthGood       4.758e+00  3.245e-01  14.664  < 2e-16 ***
## genhealthFair       6.917e+00  5.039e-01  13.726  < 2e-16 ***
## genhealthPoor       9.350e+00  1.392e+00   6.717 2.13e-11 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.967 on 3771 degrees of freedom
##   (2727 observations deleted due to missingness)
## Multiple R-squared:  0.09217,    Adjusted R-squared:  0.09096
## F-statistic: 76.57 on 5 and 3771 DF,  p-value: < 2.2e-16
```

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

```
##                     2.5 % 97.5 %
## (Intercept)          25.9   27.2
## income                0.0    0.0
## genhealthVery good    1.0    2.2
## genhealthGood         4.1    5.4
## genhealthFair         5.9    7.9
## genhealthPoor         6.6   12.1
```

5. Interpret the regression coefficients.

- b_1: After controlling for general health, 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: Those reporting very good health have 1.6 (0.99, 2.2, p<.0001) higher BMI compared to those reporting excellent health.
- b_3:Those reporting good health have 4.8 (4.1, 5.4, p<.0001) higher BMI compared to those reporting excellent health.
- b_4: Those reporting fair health have 6.9 (5.9, 7.9, p<.0001) higher BMI compared to those reporting excellent health.
- b_5: Those reporting poor health have 9.4 (6.6, 12.1, p<.0001) higher BMI compared to those reporting excellent health.

6. Conclusion

After controlling for general health, income is not significantly associated with BMI. General health is significantly associated with BMI, the average BMI increases as reported general health decreases.