Regression Assignment

Assignment Overview

You will perform 4 regression analyses in this assignment. For each analysis you will

- Fit the simple, and multivariable model.
- Test for a potential confounder Z.
- Interpret ALL regression coefficients (including the intercept) of the multivariable model.

Remember, to test if Z confounds the relationship between X and Y, the relationship between X and Y must be significant

1. Multiple Linear Regression: $Q \sim Q + Z$
   - One quantitative and a potential binary confounder.
2. Logistic Regression: $logit(B) \sim Q + Z$
   - One quantitative and a potential binary confounder.
   - Your binary response variable Y must be coded as 1 (event) and 0 (non-event).
3. Log-Linear Regression: $log(Q) \sim X + Z$
   - X and Z can be of any data type.
4. Any of the above analyses (or a new model) with a categorical (more than 2 levels) predictor variable e.g.: $Q \sim X + C$.
   - You do NOT need to test a for a potential confounder here.
   - Review detailed instructions in the example below. You have to write out the mathematical model.

Instructions

1. Identify variables under consideration
   - Determine a third variable Z that you want to test as a potential confounder.
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 bivariate model
   - Model the response variable on the explanatory variable `y ~ x`
   - Write a simple sentence on whether or not there is a relationship.
   - For models where you must test for a confounder, this relationship must be significant.
4. Fit the multivariable model.
   - 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.
   - Assess model fit by examining the residuals.
5. Interpret all regression coefficients.
6. Write a conclusion.

Multiple Linear Regression

1. Identify variables

If you have a likert "Strongly Agree" to "Strongly Disagree" variable that has at least 5 levels you can treat it as a pseudo-Quantitative Variable for this assignment.

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

2. State hypotheses

- 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)
summary(lm.mod1)
```

```
##
## Call:
## lm(formula = income ~ wakeup, data = addhealth)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31297 -15622  -5622   9245 209865
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  43548.4     1126.2  38.667  < 2e-16 ***
## wakeup        -487.7      151.2  -3.225  0.00127 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24670 on 3812 degrees of freedom
##   (2690 observations deleted due to missingness)
## Multiple R-squared:  0.002721,   Adjusted R-squared:  0.002459
## F-statistic:  10.4 on 1 and 3812 DF,  p-value: 0.001271
```

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.

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

```
##
## 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
```

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

> 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/

```r
library(stargazer)
stargazer(lm.mod1, lm.mod2,  type='latex', ci=TRUE, single.row=TRUE, digits=0,
          omit.stat="rsq", column.labels=c("SLR", "MLR"), model.numbers=FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Nov 11, 2019 - 3:26:31 PM

Table 1:

	Dependent variable:	
	income	
	SLR	MLR
wakeup	−488*** (−784, −191)	−611*** (−904, −319)
female_cFemale		−8,527*** (−10,074, −6,980)
Constant	43,548*** (41,341, 45,756)	48,669*** (46,304, 51,035)
Observations	3,814	3,813
Adjusted R^2	0	0
Residual Std. Error	24,666 (df = 3812)	24,297 (df = 3810)
F Statistic	10*** (df = 1; 3812)	64*** (df = 2; 3810)
Note:		*p<0.1; **p<0.05; ***p<0.01

5. Interpret the regression coefficients of the multivariable model.

- b_0: For a male (gender=0) who wakes up at midnight (wakeup=0), their predicted average income is \$48,669.4 (95% CI \$46,303.9, \$51,035)
- b_1: Holding gender constant, for every hour later a person wakes up, their predicted average income drops by \$611 (95% CI \$319, \$904).
- b_2: Controlling for the time someone wakes up in the morning, the predicted average income for females is \$8,527 (95% CI \$6,980, \$10,074) lower than for males.

6. Template 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.

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**.

After wordsmithing/editing for sentence flow,

> 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 and the time you wake up in the morning.
- Alternative hypothesis: There is a relationship between the probability of living below the poverty level and the time you wake up in the morning.
- Confounding hypothesis: There *still is* a relationship between the probability of living below the poverty level and the time you wake up in the morning 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.

```
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_cSmoked at least once  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.

```
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_cSmoked at least once  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

Below I create a table containing the odds ratio estimates and 95% CI for those estimates using the multivariable model.

```
# 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_cSmoked at least once	1.47	1.26	1.72
female_cFemale	2.40	2.06	2.79

- After controlling for gender, smokers have 1.5 (1.3, 1.7) times the odds of reporting making below the federal poverty level compared to non smokers.
- After controlling for smoking status, females have 2.4 (2.1, 2.8) time the odds of reporting annual earned wages below the federal poverty level compared to males.

6. Template 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**.

Log Transformed Response

1. Identify variables

- Quantitative outcome that has been log transformed: Income (variable `logincome`)
- Binary predictor: Ever smoked a cigarette (variable `eversmoke_c`)
- Binary confounder: Gender (variable `female_c`)

2. State hypothesis

- Null hypothesis: There is no relationship between the time you wake up in the morning and your income level
- Alternative hypothesis: There is a relationship between the time you wake up in the morning and your income level.
- Confounding hypothesis: There *still is* a relationship between the time you wake up in the morning and your income, after controlling for gender.

3. Fit the simple model

$$ln(Y) \sim \beta_0 + \beta_1 x_1$$

```
ln.mod.1 <- lm(logincome~wakeup, data=addhealth)
summary(ln.mod.1)
```

```
##
## Call:
## lm(formula = logincome ~ wakeup, data = addhealth)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.23205 -0.34407 -0.00154  0.35211  1.97669
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.537321   0.024294 433.738  < 2e-16 ***
## wakeup      -0.012113   0.003262  -3.713 0.000207 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5321 on 3812 degrees of freedom
##   (2690 observations deleted due to missingness)
## Multiple R-squared:  0.003604,   Adjusted R-squared:  0.003343
## F-statistic: 13.79 on 1 and 3812 DF,  p-value: 0.0002075
```

There is a significant relationship between the time you wake up and the natural log of your income.

4. Fit the multivariable model

$$ln(Y) \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

```
ln.mod.2 <- lm(logincome~wakeup + female_c, data=addhealth)
summary(ln.mod.2)
```

```
##
## Call:
## lm(formula = logincome ~ wakeup + female_c, data = addhealth)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32215 -0.33473 -0.00461  0.34058  2.01559
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.653062   0.025995 409.805  < 2e-16 ***
## wakeup         -0.014907   0.003218  -4.633 3.73e-06 ***
## female_cFemale -0.192710   0.017000 -11.336  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5233 on 3810 degrees of freedom
##   (2691 observations deleted due to missingness)
## Multiple R-squared:  0.03611,    Adjusted R-squared:  0.0356
## F-statistic: 71.36 on 2 and 3810 DF,  p-value: < 2.2e-16
```

There is still significant relationship between the time someone wakes up and the natural log of their income.

5. Interpret the regression coefficients.

- For every hour later one wakes up in the morning, one can expect to earn `1-exp(-0.015)` $= 1.4\%$ less income than someone who wakes up one hour earlier. This is after controlling for gender.
- Females have on average `1-exp(-0.19)` $= 17\%$ percent lower income than males, after controlling for the wake up time.

6. Conclusion

Don't forget to add the confidence intervals and p-values into your conclusion.

```
1-exp(confint(ln.mod.2)[-1,])
```

```
##                      2.5 %      97.5 %
## wakeup          0.02099299 0.008561652
## female_cFemale 0.20231394 0.147326777
```

Both gender and time one wakes up are significantly associated with the amount of personal earnings one makes. Waking up later in the morning is associated with 1.4% (95% CI 0.8%-2%, p<.0001) percent lower income than someone who wakes up one hour earlier. Females have 17% (95% CI 15%-20%, p<.0001) percent lower income than males.

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.

1. Identify variables

- 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 with levels: Excellent (reference), Very good, Good, Fair and Poor.

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. Fit the multivariable model.

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) %>% kable()
```

	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

4. Interpret the regression coefficients.

- b_0: The predicted BMI for individuals with no income and excellent health is 26.5 (25.9, 27.2).
- 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.

5. 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.