Homework 3

Multiple Regression

YOUR NAME

Due February 3rd, 2023

Contents

For this assignment, there are two datasets that you will use: hw2data.csv and motivation.Rdata

If you do the extra credit, you will also use the arh_hw3.Rdata dataset.

Question 1

Use the hw2 data to answer this question. The data come from the study below:

Kim, S. E., Kim, H. N., Cho, J., Kwon, M. J., Chang, Y., et al. (2016) Correction: Direct and indirect effects of five factor personality and gender on depressive symptoms mediated by perceived stress. *PLOS ONE, 11*: e0157204.

The `hw2` file contains the following variables:

- `Stress`: Total perceived stress score from self-reported stress questionnaire
- `CESD`: Total depression score for the Center for Epidemiological Studies Depression Scale
- `N`: Total score on neuroticism from the Revised NEO Personality Inventory
- `E`: Total score on extraversion from the Revised NEO Personality Inventory
- `O`: Total score on openness to Experience from the Revised NEO Personality Inventory
- `A`: Total score on agreeableness from the Revised NEO Personality Inventory
- `C`: Total score on conscientiousness from the Revised NEO Personality Inventory
- `sex`: Binary variable representing biological sex (0 = male; 1 = female)

Part a) [1 pt.]

Fit a linear model using Openness (`O`) and conscientiousness (`C`) to predict Depression (`CESD`), write the regression equation, and interpret each of the parameters found in the multiple regression model. Round all numbers to two decimal places.

```r
mod1a <- lm(CESD ~ O + C, data = hw2)
summary(mod1a)
```

```
##
## Call:
## lm(formula = CESD ~ O + C, data = hw2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.574  -3.543  -1.278   2.191  36.636
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.869830   0.630851  23.571  < 2e-16 ***
## O            0.037636   0.008959   4.201 2.72e-05 ***
## C           -0.120927   0.010991 -11.003  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.422 on 3947 degrees of freedom
## Multiple R-squared:  0.03225,    Adjusted R-squared:  0.03176
## F-statistic: 65.77 on 2 and 3947 DF,  p-value: < 2.2e-16
```

$$\widehat{\text{CESD}}_i = 14.87 + 0.038 \times Openness_i - 0.12 \times Conscientiousness_i$$

Interpretations

- **Intercept**: The predicted level of Total depression when Openness and Conscientiousness are both equal zero is 14.87.
- **Estimate of `O`**: There is a predicted increase of 0.038 points in depression scores for every one-unit increase in Openness, when holding Conscientiousness constant.
- **Estimate of `C`**: There is a predicted decrease of 0.12 points in depression scores for every one-unit increase in Conscientiousness, when holding Openness constant.

Part b) [1.5 pt.]

Repeat the multiple regression model from Part a, but with *standardized* predictors. Write the regression equation and interpret the slopes of the two predictors.

Based on this analysis is Openness (`O`) or Conscientiousness (`C`) a better predictor of Depression (`CESD`)? Explain your reasoning.

```r
hw2_std <- hw2 |>
  dplyr::mutate(
    dplyr::across(N:C, ~ scale(.x))
  )

mod1b <- lm(CESD ~ O + C, data = hw2_std)
summary(mod1b)
```

```
##
```

```
## Call:
## lm(formula = CESD ~ O + C, data = hw2_std)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.574  -3.543  -1.278   2.191  36.636
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.5342     0.1022 112.877  < 2e-16 ***
## O             0.4310     0.1026   4.201 2.72e-05 ***
## C            -1.1289     0.1026 -11.003  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.422 on 3947 degrees of freedom
## Multiple R-squared:  0.03225,    Adjusted R-squared:  0.03176
## F-statistic: 65.77 on 2 and 3947 DF,  p-value: < 2.2e-16
```

$$\widehat{\text{CESD}}_i = 11.53 + 0.43 \times Openness_i - 1.13 \times Conscientiousness_i$$

Interpretations

- **Intercept**: The predicted level of Total depression when Openness and Conscientiousness are both at their average is 11.53.
- **Estimate of `O`**: There is a predicted increase of 0.43 standard deviations in depression scores for every one-unit increase in Openness, when holding Conscientiousness constant.
- **Estimate of `C`**: There is a predicted decrease of 1.13 standard deviations in depression scores for every one-unit increase in Conscientiousness, when holding Openness constant.
- **Better predictor**: Consciousness, because its slope is higher in absolute value than the slope of Openness.

Part c) [0.5 pt.]

Add `N` as another standardized predictor to the model created in Part b. Write the regression equation, and identify what the best predictor of depression (`CESD`) is in the model.

```r
mod1c <- lm(CESD ~ O + C + N, data = hw2_std)
summary(mod1c)
```

```
##
## Call:
## lm(formula = CESD ~ O + C + N, data = hw2_std)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.2243  -3.4145  -0.5357   2.5046  30.4171
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.53418    0.09459 121.937  < 2e-16 ***
## O            0.28269    0.09515   2.971  0.00299 **
## C            0.26280    0.10934   2.404  0.01628 *
## N            2.79808    0.10891  25.691  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.945 on 3946 degrees of freedom
## Multiple R-squared:  0.1709, Adjusted R-squared:  0.1703
## F-statistic: 271.2 on 3 and 3946 DF,  p-value: < 2.2e-16
```

$$\widehat{\text{CESD}}_i = 11.53 + 0.28 \times Openness_i + 0.26 \times Conscientiousness_i + 2.80 \times Neuroticism_i$$

The best predictor of CESD is:

- Neuroticism

Part d) [1 pt.]

Create a table summarizing the results of your models. The table does not have to be perfectly compliant with APA formatting, but it should be presentable (see the tables in Lab for expectations). The table should have all numbers rounded to 2 decimal places, names for the models, and should include confidence intervals.

```r
stargazer(mod1a, mod1b, mod1c, type = "latex", header = FALSE, title = "Multiple regression model predi
          column.labels = c("Raw predictors", "Standardized predictors", "Standardized predictors"))
```

Table 1: Multiple regression model predicting Depression scores from Openness, Conscientiousness (1 and 2), and Neuroticism (3).

	Dependent variable:		
	CESD		
	Raw predictors	Standardized predictors	Standardized predictors
	(1)	(2)	(3)
O	0.04^{***}	0.43^{***}	0.28^{***}
	(0.02, 0.06)	(0.23, 0.63)	(0.10, 0.47)
C	-0.12^{***}	-1.13^{***}	0.26^{**}
	(−0.14, −0.10)	(−1.33, −0.93)	(0.05, 0.48)
N			2.80^{***}
			(2.58, 3.01)
Constant	14.87^{***}	11.53^{***}	11.53^{***}
	(13.63, 16.11)	(11.33, 11.73)	(11.35, 11.72)
Observations	3,950	3,950	3,950
R^2	0.03	0.03	0.17
Adjusted R^2	0.03	0.03	0.17
Residual Std. Error	6.42 (df = 3947)	6.42 (df = 3947)	5.94 (df = 3946)
F Statistic	65.77^{***} (df = 2; 3947)	65.77^{***} (df = 2; 3947)	271.18^{***} (df = 3; 3946)
Note:			*p<0.1; **p<0.05; ***p<0.01

Question 2

Part a) [1 pt.]

Using the `hw2` data, predict `Stress` from the following independent variables, in a series of regression models:

- Model 1: `Stress` predicted by Openness (`O`)
- Model 2: `Stress` predicted by `sex`
- Model 3: `Stress` predicted by Openness (`O`) plus `sex`
- Model 4: `Stress` predicted by Openness (`O`), `sex`, and the interaction between Openness (`O`) and `sex`

Standardize all the appropriate variables in the analyses and present the results of the analyses in a table.

```
mod2a <- lm(Stress ~ O, data = hw2_std)
mod2b <- lm(Stress ~ sex, data = hw2_std)
mod2c <- lm(Stress ~ O + sex, data = hw2_std)
mod2d <- lm(Stress ~ O*sex, data = hw2_std)
```

```
stargazer(mod2a, mod2b, mod2c, mod2d, type = "latex", header = FALSE, title = "Multiple regression model
          column.labels = c("Model 1", "Model 2", "Model 3", "Model 4"))
```

Table 2: Multiple regression model predicting Stress scores from Openness and Sex.

	Dependent variable:			
	Stress			
	Model 1	Model 2	Model 3	Model 4
	(1)	(2)	(3)	(4)
O	0.17		0.02	−0.43*
	(−0.05, 0.38)		(−0.19, 0.23)	(−0.91, 0.05)
sex1		1.99***	1.99***	2.09***
		(1.53, 2.46)	(1.52, 2.46)	(1.61, 2.57)
O:sex1				0.56**
				(0.03, 1.10)
Constant	17.44***	16.01***	16.02***	15.90***
	(17.22, 17.65)	(15.62, 16.41)	(15.62, 16.42)	(15.49, 16.32)
Observations	3,950	3,950	3,950	3,950
R^2	0.001	0.02	0.02	0.02
Adjusted R^2	0.0003	0.02	0.02	0.02
Residual Std. Error	6.82 (df = 3948)	6.76 (df = 3948)	6.77 (df = 3947)	6.76 (df = 3946)
F Statistic	2.32 (df = 1; 3948)	70.31*** (df = 1; 3948)	35.17*** (df = 2; 3947)	24.88*** (df = 3; 3946)
Note:				*p<0.1; **p<0.05; ***p<0.01

Part b) [1 pt.]

Write the estimated equations from the regression in Model 4 for when sex = 0 (males) and when sex = 1 (females). Make sure to simplify the equations.

You may find equations 1-3 in this paper useful.

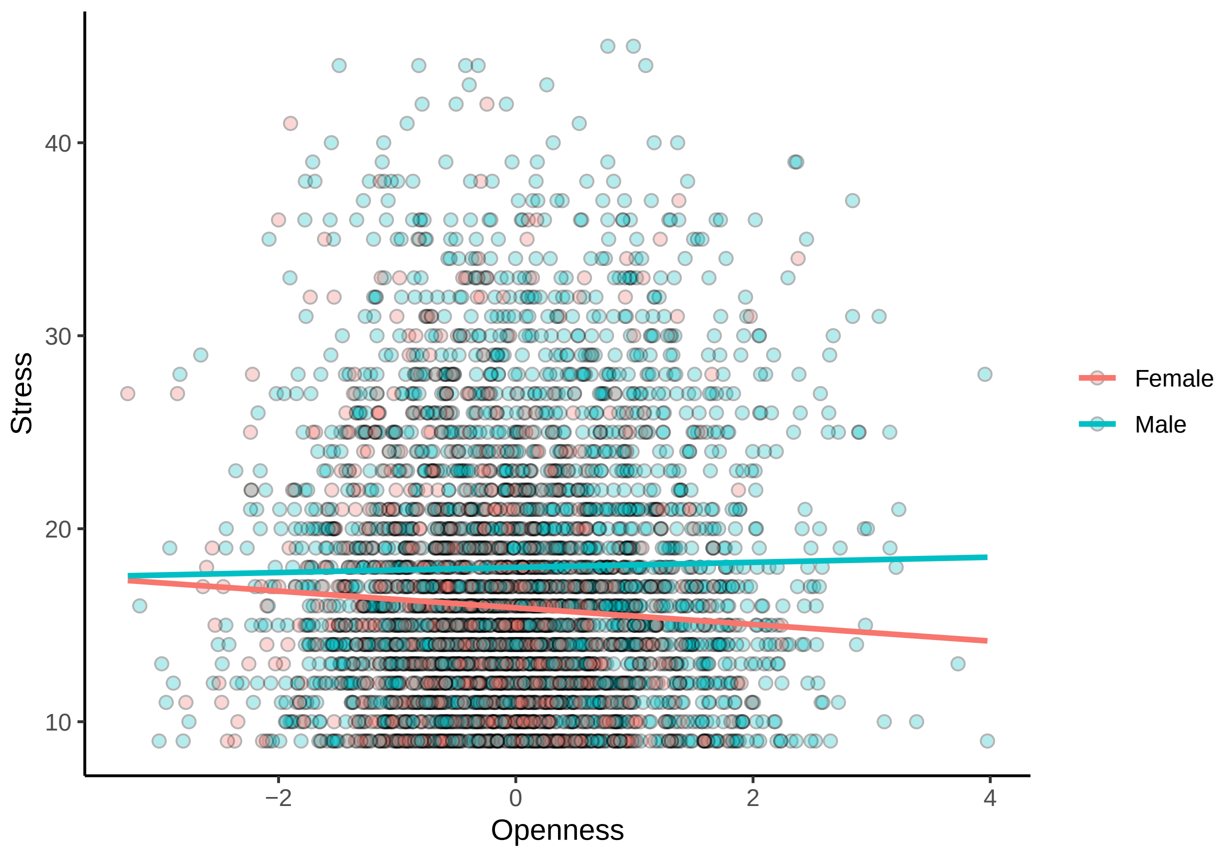
$$\widehat{\text{Stress}}_i = 15.90 - 0.43 \times Openness_i, \quad \text{when sex} = 0$$

$$\widehat{\text{Stress}}_i = (15.90 + 2.09) + (-0.43 + 0.56) \times Openness_i, \quad \text{when sex} = 1$$

Part c) [1 pt.]

Create a scatter plot depicting the interaction analysis above. Make sure that data points and regression line for males and females is clearly labeled, and that the differences between males and females are apparent. You made need to try different combinations of colors and/or panels to make a nice graph.

```r
ggplot(data = hw2_std, aes(x = O, y = Stress, fill = sex)) +
  geom_point(shape = 21, size = 2, alpha = .3) +
  geom_smooth(aes(color = sex), method = "lm", se = F, fullrange = T) +
  xlab("Openness") +
  scale_colour_discrete(labels = c("Female", "Male")) +
  scale_fill_discrete(labels = c("Female", "Male")) +
  theme_classic() +
  theme(legend.title = element_blank())
```



Part d) [1 pt.]

Answer the following questions:

i) On its own, which variable had a bigger effect on `Stress`: Openness (`O`) or `sex`?
ii) In which group was there a stronger association between Openness (`O`) and `Stress`: males or females?

Answers:

- Sex
- Females

Question 3

Part a) [1 pt.]

Using the motivation dataset, run a regression with `motivation` predicted by `difficulty` and write the regression equation. Then, create a quadratic model with `motivation` predicted by `difficulty` and write the regression equation. Do not standardize `difficulty`, but make sure it is mean-centered. You can do this using the `scale()` function and setting `scale = FALSE`, e.g.,

```
# example of mean-centering the variable `x`
x <- 1:10
mean_centered_x <- scale(x, center = TRUE, scale = FALSE)
mean(x)
```

```
## [1] 5.5
```

```
mean(mean_centered_x)
```

```
## [1] 0
```

```
motivation$difficulty_c <- motivation$difficulty - mean(motivation$difficulty, na.rm = TRUE)
```

```
# first model
mod3a <- lm(motivation ~ difficulty_c, data=motivation)
summary(mod3a)
```

```
##
## Call:
## lm(formula = motivation ~ difficulty_c, data = motivation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -178.54  -54.84   17.30   73.03   99.32
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  156.6948     5.8566  26.755   <2e-16 ***
## difficulty_c  -0.4170     0.5137  -0.812    0.418
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82.82 on 198 degrees of freedom
## Multiple R-squared:  0.003317,   Adjusted R-squared:  -0.001717
## F-statistic: 0.659 on 1 and 198 DF,  p-value: 0.4179
```

$$\widehat{\text{motiv}}_i = 156.69 - 0.42 \times difficulty_i$$

```
# second model
mod3b <- lm(motivation ~ difficulty_c + I(difficulty_c^2), data=motivation)
summary(mod3b)
```

```
##
## Call:
## lm(formula = motivation ~ difficulty_c + I(difficulty_c^2), data = motivation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0023 -1.9198  0.1666  2.1002  6.7772
```

```
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        247.962609   0.309808   800.4   <2e-16 ***
## difficulty_c        -2.544384   0.018978  -134.1   <2e-16 ***
## I(difficulty_c^2)   -0.702136   0.001769  -396.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.935 on 197 degrees of freedom
## Multiple R-squared:  0.9988, Adjusted R-squared:  0.9987
## F-statistic: 7.899e+04 on 2 and 197 DF,  p-value: < 2.2e-16
```

$$\widehat{\text{motiv}}_i = 274.96 - 2.54 \times difficulty_i - 0.70 \times difficulty_i^2$$

Which model accounts for more variance in motivation (i.e., which model has a higher R^2)?

Answer: The second model with the quadratic term.

Part b) [1 pt.]

For the quadratic model, provide how you would interpret the following:

- **Intercept**: The predicted level of motivation when difficulty is at its mean level is 247.96.
- **Estimate of `difficulty` (not the square of difficulty)**: For every unit change in Difficulty, Motivation is predicted to change by -2.54 units.

Extra Credit [3 pts.]

Using two of the three variables in `arh_hw3`, create a regression model in which suppression occurs. The variables should be standardized in the regression model. You may need to test out different combinations of independent variables, dependent variables, and covariates.

Remember to examine the necessary R^2 values to confirm that suppression has occurred. Once you have identified a model that leads to suppression, report the regression equation below, and identify: the type of suppression that is occurring, the suppressor variable, and the supressed variable. You may include a brief explanation if you would like.

```r
mod4a <- lm(FearDeath ~ Depression2004, data = arh_hw3)
summary(mod4a)
```

```
##
## Call:
## lm(formula = FearDeath ~ Depression2004, data = arh_hw3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1725 -1.6679 -0.5195  2.4137 23.5102
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    12.25238    0.31236  39.225   <2e-16 ***
## Depression2004  0.02968    0.02347   1.265    0.206
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.854 on 836 degrees of freedom
## Multiple R-squared:  0.00191,    Adjusted R-squared:  0.0007162
## F-statistic:   1.6 on 1 and 836 DF,  p-value: 0.2063
```

```r
mod4b <- lm(FearDeath ~ Depression2004 + SelfWorth2004
            , data = arh_hw3)
summary(mod4b)
```

```
##
## Call:
## lm(formula = FearDeath ~ Depression2004 + SelfWorth2004, data = arh_hw3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8973  -1.9887  -0.2209   1.7843  17.3279
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.19248    0.46773  11.101  < 2e-16 ***
## Depression2004 -0.07747    0.02069  -3.745 0.000193 ***
## SelfWorth2004   1.34514    0.07355  18.288  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 835 degrees of freedom
## Multiple R-squared:  0.2874, Adjusted R-squared:  0.2856
## F-statistic: 168.3 on 2 and 835 DF,  p-value: < 2.2e-16
```

Regression Model

$$\widehat{\text{fdeath}}_i = 5.19 - 0.07 \times Depression_i + 1.34 \times SelfWorth_i$$

Type of Suppression Occurring * Classical suppression

The Suppressor Variable * Depression

The Suppressed Variable * Self Worth

Explanation: * The depression suppresses the proportion of variance in self worth that is irrelevant for fear of death.