PSC 204B Homework 1

Due Date: January 19, 2024

Question 1 (2 points)

Fit a simple linear regression model using biological sex (*sex*; 0 = male; 1 = female) to predict depression (*CESD*). Write out the predicted regression equation using R markdown equation notation in the space below (see the lab), and interpret each parameter (see the lab for what is expected). Round all numbers to two decimal places.

```
## Code
mod1 <- lm(CESD ~ sex, data = hw1)
summary(mod1)
```

```
##
## Call:
## lm(formula = CESD ~ sex, data = hw1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.236  -3.236  -1.236   2.209  35.764
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.7910     0.1910   51.25   <2e-16 ***
## sex           2.4451     0.2262   10.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.433 on 3948 degrees of freedom
## Multiple R-squared:  0.02873,    Adjusted R-squared:  0.02849
## F-statistic: 116.8 on 1 and 3948 DF,  p-value: < 2.2e-16
```

- **Equation**: $\widehat{Depression}_i = 9.79 + 2.45 \times Sex_i$
- **(Intercept)**: The expected value of depression is 9.79 when sex is male.
- **Sex**: Females are predicted to have 2.45 more units of depression than males. This slope is significantly different from 0, indicating that Depression and sex are significantly associated.

Question 2 (6.5 points)

Part a) (2.5 points)

Fit a simple linear regression model using neuroticism (N) to predict stress (*Stress*). Write out the predicted regression equation and interpret each of the parameters found in the regression model.

```
## Code
mod2 <- lm(Stress ~ N, data = hw1)
summary(mod2)
```

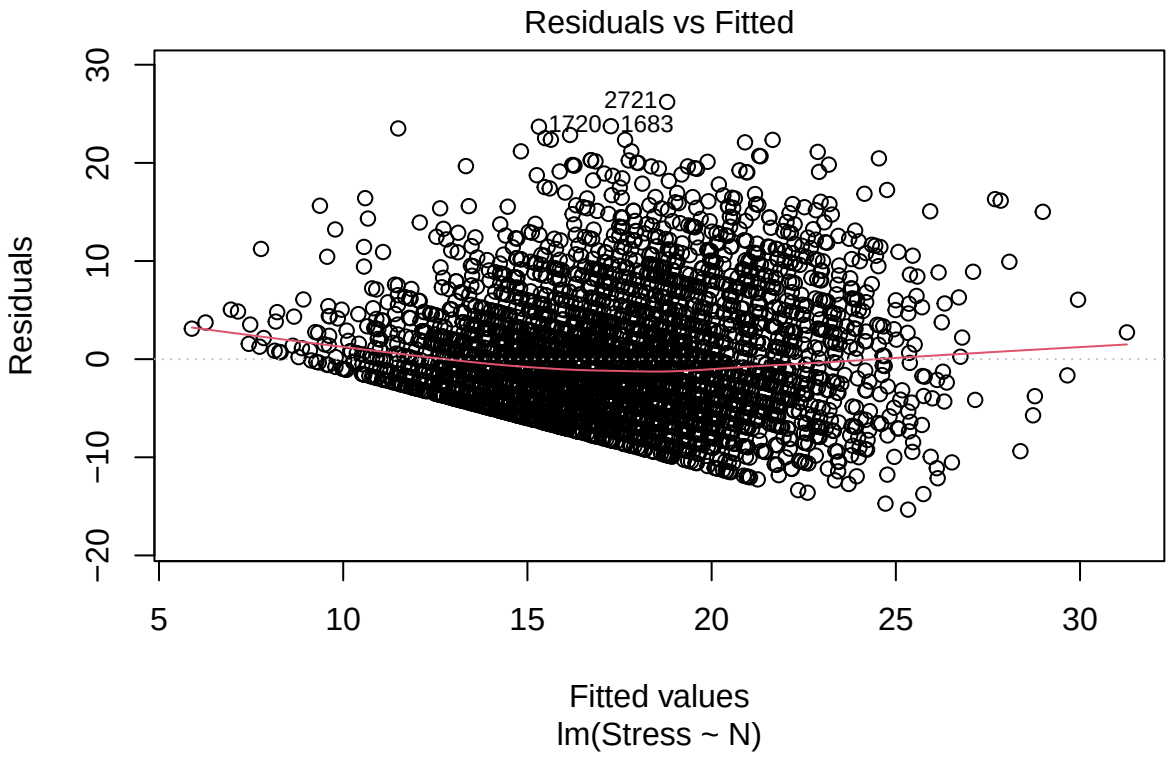
```
##
## Call:
## lm(formula = Stress ~ N, data = hw1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.333  -4.139  -1.027   2.943  26.207
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.26540    0.57417  -3.945  8.1e-05 ***
## N            0.35698    0.01026  34.791  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.97 on 3948 degrees of freedom
## Multiple R-squared:  0.2346, Adjusted R-squared:  0.2345
## F-statistic:  1210 on 1 and 3948 DF,  p-value: < 2.2e-16
```

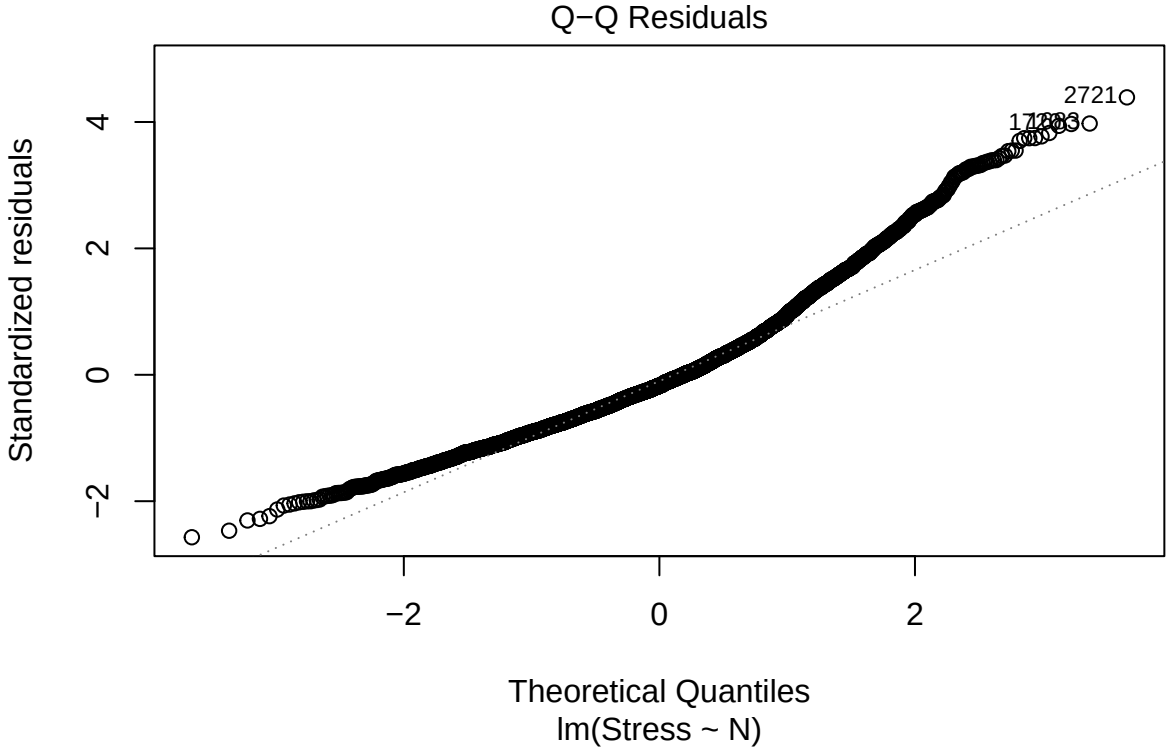
- **Equation**: $\widehat{Stress}_i = -2.66 + 0.36 \times Neuroticism_i$
- **(Intercept)**: The expected value of stress is -2.27 when Neuroticism is equal 0.
- **N**: An one-unit increase in Neuroticism is associated with a 0.36 increase in Stress. This slope is significantly different from 0, indicating that Stress and Neuroticism are significantly associated.

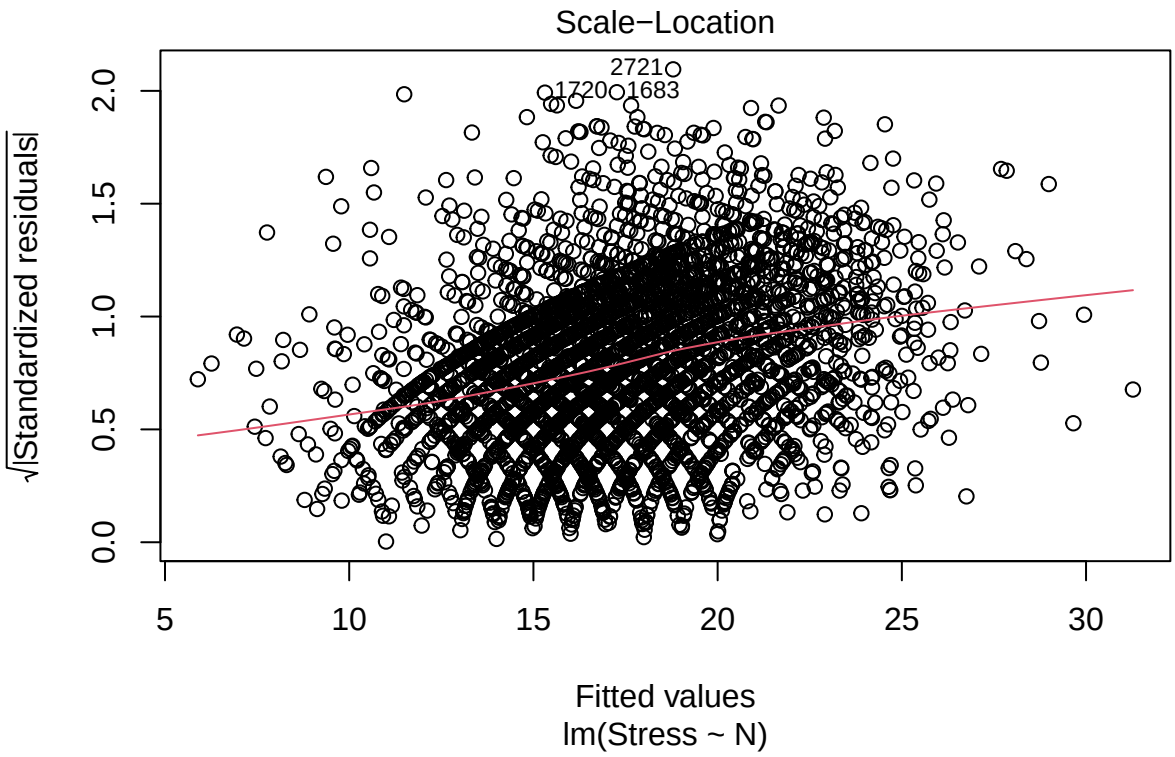
Part b) (1 point)

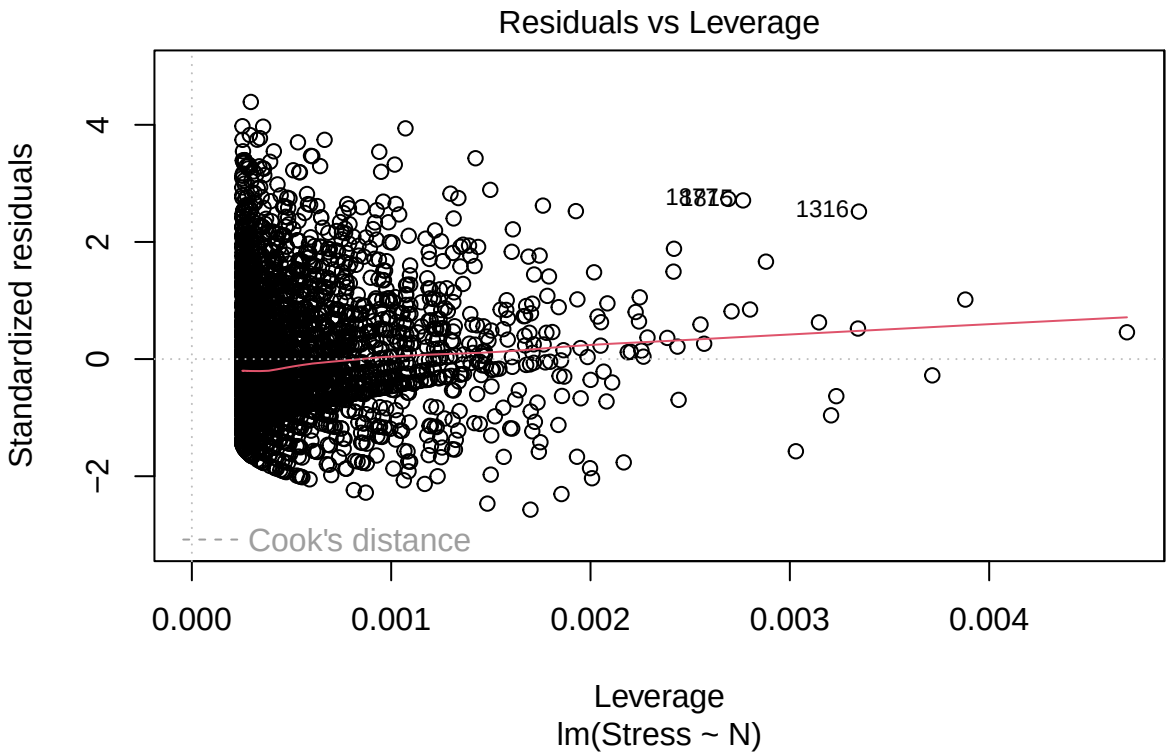
Plot the model diagnostic plots, and assess whether the assumptions of linearity, homogeneity of variances, and normality of residuals are met. Explain your reasoning.

```
plot(mod2)
```


Residuals vs Fitted
2721
1720
1683
Residuals
30
20
10
0
−10
−20
5
10
15
20
25
30
Fitted values
lm(Stress ~ N)


Q−Q Residuals
Standardized residuals
−2
0
2
4
2721
−2
0
2
Theoretical Quantiles
lm(Stress ~ N)


Scale−Location
2721
1720
1683
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
5
10
15
20
25
30
Fitted values
lm(Stress ~ N)



- **Linearity**: The predicted line of the Residuals vs. Fitted plot is expected to be flat to indicate zero correlation between the predicted and residual values. The assumption of linearity in this model is well met.
- **Homogeneity of Variances**: We can also use the Residuals vs. Fitted plot to investigate the homogeneity of variances. Different from the random pattern expected when the variance is homogeneous, we observe a peculiar pattern, indicating heterocedasticity.
- **Normality of Residuals**: The normality of residuals can be assessed with the QQ Plot. All of the points are expected to fall on the diagonal line. In this model, we observe an expressive departure from normality of the residuals since the standardized residuals greater than zero show a positive slope, deviating from the reference dotted line.

Part c) (2.5 points)

Repeat the regression analysis above in Part a, but using mean-centered Neuroticism (N) to predict *Stress*. Call the mean-centered Neuroticism variable *N_c*. Write out the predicted regression equation and interpret each of the parameters found in the regression model. **In each of your interpretations, comment on whether the parameter changed from the original analysis, and why or why not.**

```
## Code
hw1$N_c <- scale(hw1$N, scale = FALSE)

mod3 <- lm(Stress ~ N_c, data = hw1)
summary(mod3)
```

```
##
## Call:
```

```
## lm(formula = Stress ~ N_c, data = hw1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.333  -4.139  -1.027   2.943  26.207
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.43544    0.09499  183.54   <2e-16 ***
## N_c          0.35698    0.01026   34.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.97 on 3948 degrees of freedom
## Multiple R-squared:  0.2346, Adjusted R-squared:  0.2345
## F-statistic:  1210 on 1 and 3948 DF,  p-value: < 2.2e-16
```

- **Equation**: $\widehat{Stress}_i = 17.44 + 0.36 \times Neuroticism_{c_i}$
- **(Intercept)**: The predicted value of stress is 17.44 when Neuroticism at its mean level. By centering the predictor variable the value of the intercept also changed because now we've changed the reference point for the Neuroticism.
- **N**: The predicted value of N_c does not differ from the predicted value of N because centering the preditor does not alter its linear relationship with the outcome. It still means that an one-unit increase in Neuroticism is associated with a 0.36 increase in Stress. The statistical significance of the slope also did not change.

Part d) (0.5 points)

Using the model from part (b), what is the predicted score for someone who scores 1.5 points *above the mean*?

```
# Code

17.44 + 0.36 * 1.5
```

```
## [1] 17.98
```

Question 3 (1.5 points)

Create two scatter plots to illustrate the relation between Stress and Neuroticism in the two analyses performed in Question 2. Arrange the graphs in two rows. Stress should be on the y axis in both graphs. In the first row, show the relationship between Stress and uncentered Neuroticism (N). In the second row, show the relationship between Stress and mean-centered Neuroticism (N_c). On each graph, add a line of best. Be sure to label your axes, and make sure that all graphs have the same x-axis and y-axis range.

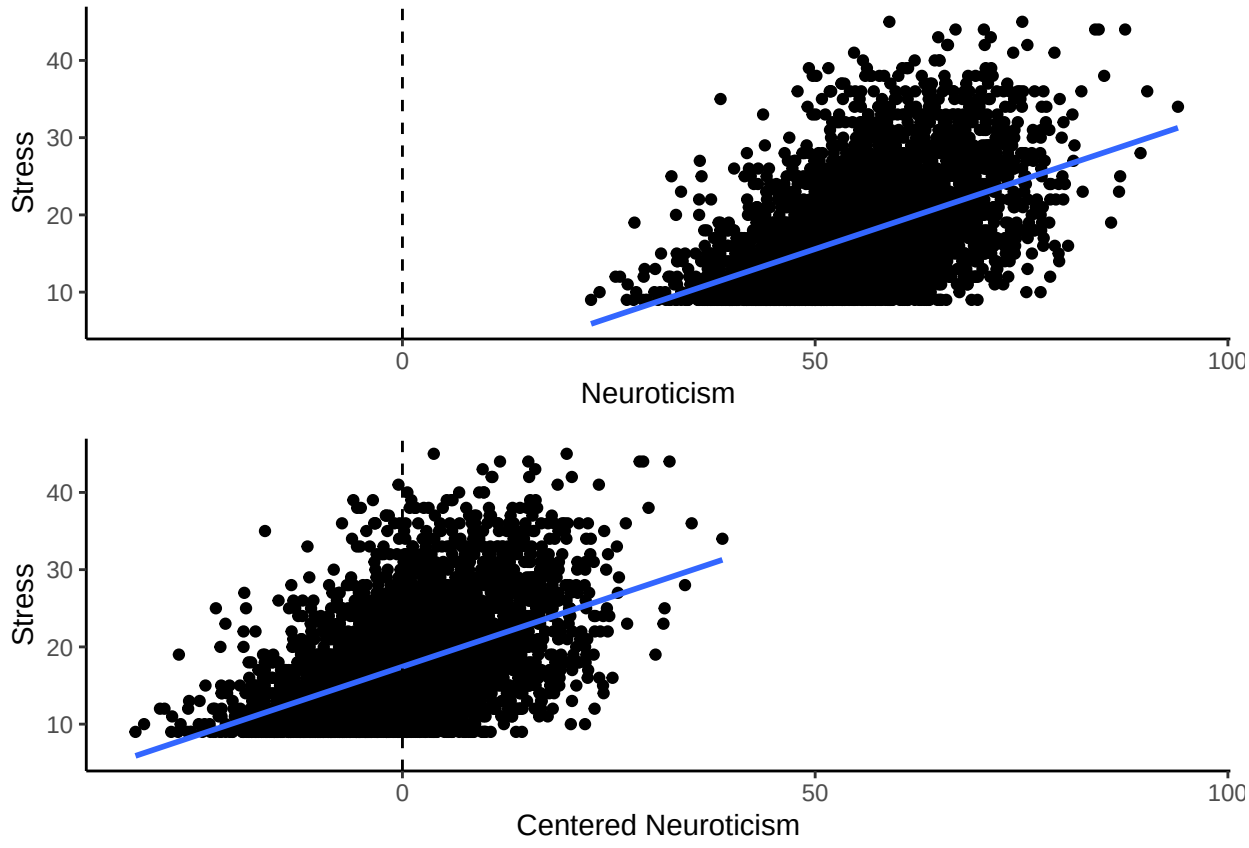
```
## Code

# Original, uncentered Data
g1 <- ggplot(data = hw1, aes(y = Stress, x = N)) +
  geom_point() +
  theme_classic() +
  xlab('Neuroticism') +
  ylab('Stress') +
  geom_smooth(method = 'lm', se = F) +
  geom_vline(xintercept = 0, linetype = "dashed") +
```

```
  coord_cartesian(xlim = c(-32, 94))

# Mean-centered FNE
g2 <- ggplot(data = hw1, aes(y = Stress, x = N_c)) +
  geom_point() +
  theme_classic() +
  xlab('Centered Neuroticism') +
  ylab('Stress') +
  geom_smooth(method = 'lm', se = F) +
  geom_vline(xintercept = 0, linetype = "dashed") +
  coord_cartesian(xlim = c(-32, 94))

ggarrange(g1, g2, nrow = 2)
```



Extra Credit (1 point)

Conduct a regression to assess whether min-centered Neuroticism (*N_min*) predicts **median-centered** *Stress* (*Stress_med*). In the space below, report and interpret the intercept. Keep in mind that the value of the intercept is now relative to something since it has been centered.

```
## Code

hw1$N_min <- hw1$N - min(hw1$N)
hw1$Stress_med <- hw1$Stress - median(hw1$Stress)
```

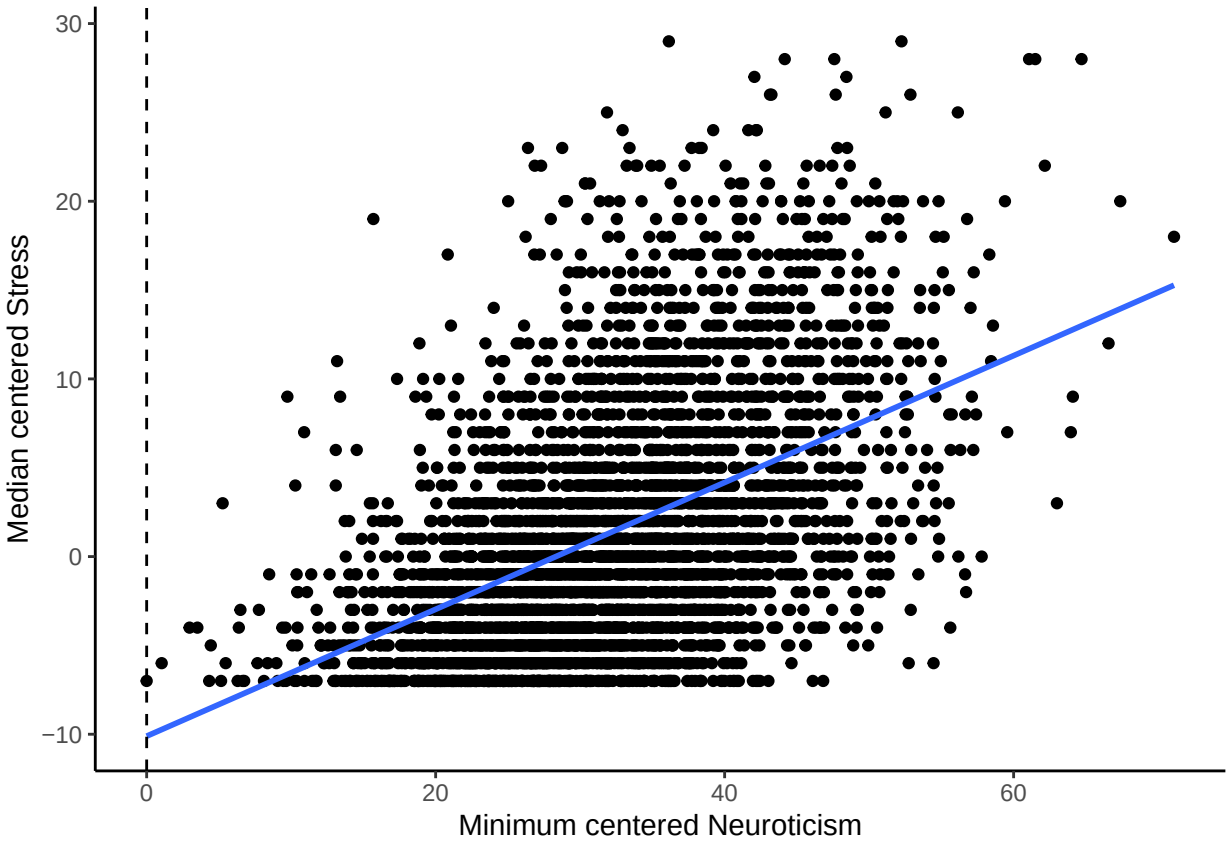
```r
mod4 <- lm(Stress_med ~ N_min, data = hw1)
summary(mod4)
```

```
##
## Call:
## lm(formula = Stress_med ~ N_min, data = hw1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.333  -4.139  -1.027   2.943  26.207
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.10836    0.34513  -29.29   <2e-16 ***
## N_min         0.35698    0.01026   34.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.97 on 3948 degrees of freedom
## Multiple R-squared:  0.2346, Adjusted R-squared:  0.2345
## F-statistic:  1210 on 1 and 3948 DF,  p-value: < 2.2e-16
```

- **Report the value of the intercept**: -10.11
- **Interpret the intercept**: The intercept now represents the predicted level of median-centered Stress when Neuroticism is at its minimum value.

```r
ggplot(data = hw1, aes(y = Stress_med, x = N_min)) +
  geom_point() +
  theme_classic() +
  xlab('Minimum centered Neuroticism') +
  ylab('Median centered Stress') +
  geom_smooth(method = 'lm', se = F) +
  geom_vline(xintercept = 0, linetype = "dashed")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```


Median centered Stress
30
20
10
0
−10
0
20
40
60
Minimum centered Neuroticism