Pset 4: Multiple regression

Alejandro Sarria

Part 1

1.

a. We find the error by plugging in the values for X and Y in observation A (2 and 9, respectively) and solving for the error term in the data generating formula.

9 = 2 + 3*2 + error

9 - 2 - 6 = error

1 = error

b. Solving the regression equation for the error results in the residual.

9 = 1.9 + 3.1*2 + error

9 - 1.9 - 6.2 = error

0.9 = error

Then, the residual of observation A is 0.9

2.

- Top left: Y ~ X + D
- Top right: Y ~ X
- Bottom left: Y ~ X
- Bottom right: Y ~ X

3.

a. When controlling for children under 5, the model estimates 76.18 more hours of work each year per unit increase in years of education.

b. When controlling for years of education, the standard error for children under 5 is 19.28

c. In the third model, the predicted work hours for a woman with 0 years of education and 0 children under 5 is 306.55

d. In the second model, a woman with 2 children and 10 years of education would be estimate to work 778.965 hours per year.

e. To calculate the "statistical significance" of children under 5 we need to calculate the t-statistic (coefficient / std. error of coefficient).

```
model2_t <- -238.853 / 19.693
model3_t <- -251.181 / 19.28

print(model2_t)
```

```
[1] -12.12883
```

```
print(model3_t)
```

```
[1] -13.02806
```

In both models, the t-statistic for children under 5 is well below -1.96, meaning that they are both statistically significant.

4.

a. A one year increase in education the number of hours worked in a year by the derivative of the model, that is: $110.230 - 3.162(Edu)$.

b. $110.230 - 3.162(16) = 59.638$

c. It gets less positive as the number of years of education increases as $\beta_2 Edu^2$ will eventually be higher than β_1 (110.230).

d. As we add more polynomials to the model, we risk overfitting to the sample data, loosing validity when comparing to the population.

5.

a. The coefficient suggests that homeowners work 50.174 more hours per year than non-homeowners. The t-statistic for Model 1 is 1.52, leaving it below the threshold for statistical significance (1.96).

```
50.174/32.923
```

```
[1] 1.52398
```

b. People with 4 children under the age of 5 work 150.592 less than people with 3.

c. The table alone is not enough to tell. One way to test it would be to run a new model including both having 2 and having 3 children under 5 and looking at the significance of each variable.

6.

a. In the model, a unit increase in education predicts an increase of $110.073 - 53.994 Homeowner$

b. The interaction term shows that the effect of education on hours worked is 110.073 for non-homeowners and 56.079 for homeowners.

c. A one unit increase in Education predicts a 6.93% increase in hours worked.

d. A 1% increase in education predicts a $832.347 * Log(1.01) = 8.28$ increase in hours worked.

e. Applying a log transformation requires the value for Hours Worked to be higher than 0 as log0 is undefined. The reduction in the number of observations for model 2 may be a result of removing cases in which there were no hours worked.

7.

b. When error terms are correlated across time, such that knowing the error term in one period gives us some information about the error term in the next period.

8.

b. Creating a plot with Y on the y-axis and X on the x-axis, and a line reflecting the predicted values of the regression, and seeing if the spread of the values around the predicted values change over the range of X.

c.

d.

9.

The researchers could apply weights the observations in the sample so that the underrepresented sectors of the population have a higher weight in the analysis.

10.

a. You're doing research on unusual sexual practices. You ask people whether they've ever engaged in these practices, which many people might prefer to keep secret, even if they've actually done them.

Part 2.

1.

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
d <- read.csv("dengue.csv")
model1 <- lm(NoYes ~ humid + temp,
             data = d)
summary(model1)
```

```
Call:
lm(formula = NoYes ~ humid + temp, data = d)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.1055 -0.1442 -0.0062  0.1767  1.3624 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.407272   0.019033 -21.398   <2e-16 ***
humid        0.052851   0.001955  27.034   <2e-16 ***
temp        -0.003286   0.001778  -1.848   0.0648 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.331 on 1995 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.5492,    Adjusted R-squared:  0.5488 
F-statistic:  1215 on 2 and 1995 DF,  p-value: < 2.2e-16
```

2.

```r
library(marginaleffects)
model2 <- glm(NoYes ~ humid + temp,
             data = d,
             family = binomial(link = "logit"))
summary(model2)
```

```
Call:
glm(formula = NoYes ~ humid + temp, family = binomial(link = "logit"), 
    data = d)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.59219    0.30290 -21.764   <2e-16 ***
humid        0.30474    0.01985  15.350   <2e-16 ***
temp         0.03987    0.01933   2.063   0.0391 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2711.7  on 1997  degrees of freedom
Residual deviance: 1357.0  on 1995  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 1363

Number of Fisher Scoring iterations: 6
```

```r
avg_slopes(model2)
```

```
  Term Estimate Std. Error     z Pr(>|z|)     S    2.5 % 97.5 %
 humid  0.03173    0.00155 20.47   <0.001 307.0 0.028692 0.0348
 temp   0.00415    0.00201  2.06   0.0393   4.7 0.000204 0.0081

Type: response
Comparison: dY/dX
```

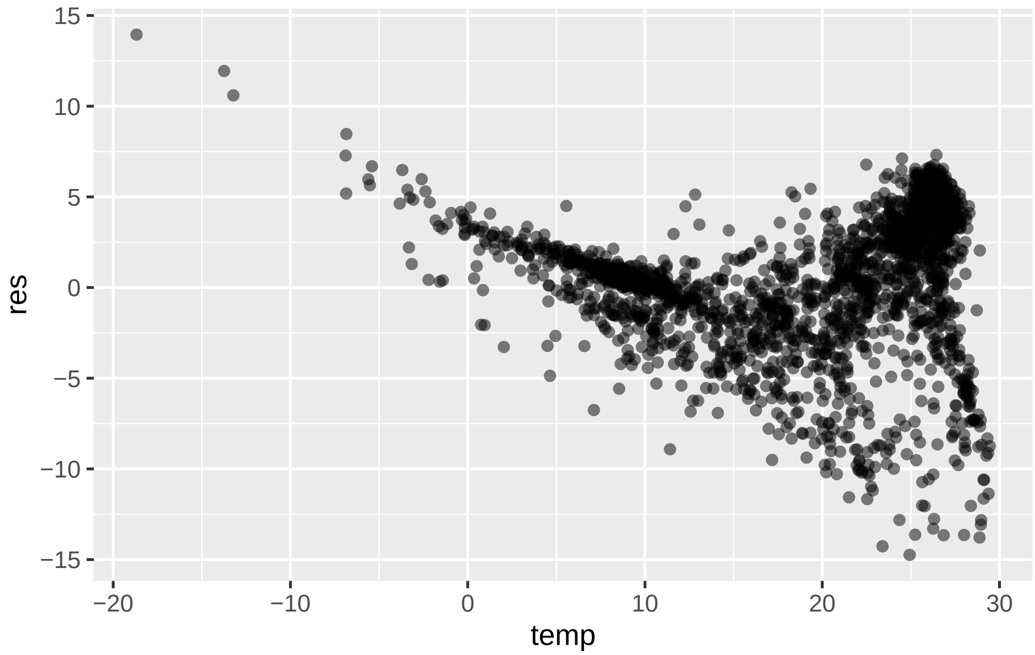
In the logit model, a unit increase in humid increases the probability of observing dengue by 3.173% while a unit increase in temperature increases the probability by 0.415%.

3.

```r
d <- d |>
  filter(!is.na(humid))

model3 <- lm(humid ~ temp,
             data = d)
summary(model3)
```

```
Call:
lm(formula = humid ~ temp, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-14.7416  -1.8751   0.4252   2.5564  13.9461

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.36152    0.21141   11.17   <2e-16 ***
temp         0.77885    0.01052   74.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.79 on 1996 degrees of freedom
Multiple R-squared:  0.7331,    Adjusted R-squared:  0.733
F-statistic:  5483 on 1 and 1996 DF,  p-value: < 2.2e-16
```

```r
d <- d |>
  mutate(res = model3$residuals)

ggplot(d, aes(x=temp,y=res)) +
  geom_point(alpha=.5)
```



There seems to be heteroskedasticity, the variance of the residuals increases as temperature rises.

```
model4 <- fixest::feols(humid ~ temp,
              data = d,
              se = "hetero")
summary(model4)
```

```
OLS estimation, Dep. Var.: humid
Observations: 1,998
Standard-errors: Heteroskedasticity-robust
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 2.361523   0.195914 12.0539 < 2.2e-16 ***
temp        0.778849   0.011529 67.5563 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 3.78837   Adj. R2: 0.732993
```

There was little change in the standard errors of the model using heteroskedasticity-robust standard errors so there might have been much issue to begin with.

4.

```r
model5 <- fixest::feols(log(humid) ~ temp,
             data = d,
             se = "hetero")
summary(model5)
```

```
OLS estimation, Dep. Var.: log(humid)
Observations: 1,998
Standard-errors: Heteroskedasticity-robust
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 1.659049   0.018220 91.0589 < 2.2e-16 ***
temp        0.056482   0.000886 63.7548 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.248727   Adj. R2: 0.770082
```

In this model, a unit increase in temperature is associated with a $100 * (e^{0.056} - 1) = 5.76$ percent increase in humidity.

5.

The relationship between humidity and temperature is exponential (this can be seen in the last graph). Applying a logarithmic transformation to humidity “reverses” the exponential relationship, making it linear.