STA 5207: Homework 8

Due: Friday, March 22 by 11:59 PM

Include your R code in an R chunks as part of your answer. In addition, your written answer to each exercise should be self-contained so that the grader can determine your solution without reading your code or deciphering its output.

Exercise 1 (The `divusa` Data Set) [50 points]

For this exercise, we will use the `divusa` data set from the `faraway` package. You can also find the data in `divusa.csv` on Canvas. The data set contains information on divorce rates in the USA from 1920 to 1996. The variables in the data set are

- `year`: the year from 1920-1996.
- `divorce`: divorce per 1000 women aged 15 or more.
- `unemployed`: unemployment rate.
- `femlab`: female participation in labor force aged 16+.
- `marriage`: marriages per 1000 unmarried women aged 16+.
- `birth`: births per 1000 women aged 15-44.
- `military`: military personnel per 1000 population.

In the following exercise, we will model the `divorce` variable in terms of `unemployed`, `femlab`, `marriage`, `birth`, and `military`.

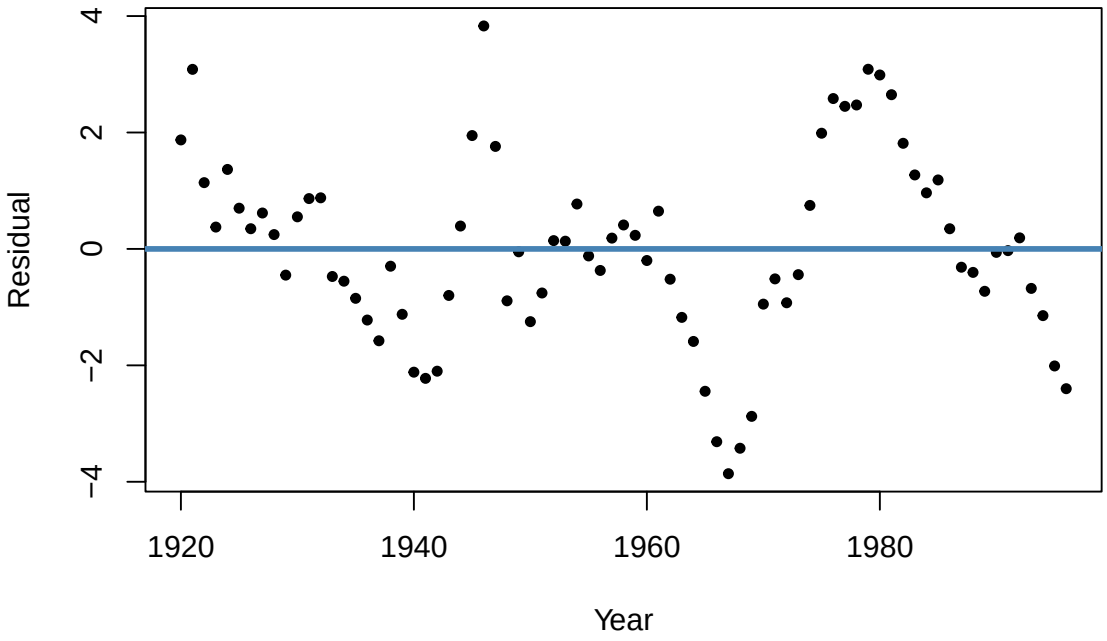
1. (2 points) The variable `year` is not being used in the model, but it shows that the measurements were taken across time. What does this make you suspect about the error term? No output need.

   There may be autocorrelation between the error terms.

2. (6 points) Fit an OLS regression model with `divorce` as the response and all other variables except `year` as predictors. Check for serial correlation in the errors using a graphical method. Do you feel like the errors are serially correlated? Justify your answer. Include any plots in your response.

```r
# fit the model using OLS
model_ols = lm(divorce ~ . - year, data = divusa)

# generate the fitted vs. year plot
plot(resid(model_ols) ~ year, data = divusa, pch = 20,
     xlab = 'Year', ylab = 'Residual')
abline(h=0, lwd=3, col='steelblue')
```



They are serially correlated since they are not scattered evenly along the line.

3. (6 points) Check for the presence of serial correlation in the errors using the Durbin-Watson test. Report the following:

   - The null and alternative hypotheses.
   - The value of the test statistic.
   - The p-value of the test.
   - A statistical decision at the $\alpha = 0.05$ significance level.

```
dwtest(model_ols, alternative = 'two.sided')
```

```
##
##  Durbin-Watson test
##
## data:  model_ols
## DW = 0.29988, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is not 0
```

Null Hypothesis: There is no serial correlation.

Alternative Hypothesis: There is a serial correlation in the errors.

p-value: $2.2e^{-16}$

We reject the null hypothesis, there is strong evidence that there is serial correlation between the errors.

1. (10 points) Model the serial correlation with an AR(1) process, meaning that $\Sigma_{ij} = \phi^{|i-j|}$. Use the ML method to estimate the parameters in the GLS fit. Create and report a table with the OLS estimates (model in part 2) and GLS estimates for the slope parameters.

```r
model_gls = gls(divorce ~ . - year,
                correlation = corAR1(form = ~ year),
                method = 'ML', data = divusa)
(summary(model_gls))
```

```
## Generalized least squares fit by maximum likelihood
##   Model: divorce ~ . - year
##   Data: divusa
##        AIC      BIC    logLik
##   179.9523 198.7027 -81.97613
##
## Correlation Structure: AR(1)
##  Formula: ~year
##  Parameter estimate(s):
##       Phi
## 0.9715486
##
## Coefficients:
##                 Value Std.Error   t-value p-value
## (Intercept) -7.059682  5.547193 -1.272658  0.2073
## unemployed   0.107643  0.045915  2.344395  0.0219
## femlab       0.312085  0.095151  3.279878  0.0016
## marriage     0.164326  0.022897  7.176766  0.0000
## birth       -0.049909  0.022012 -2.267345  0.0264
## military     0.017946  0.014271  1.257544  0.2127
##
##  Correlation:
##             (Intr) unmply femlab marrig birth
## unemployed -0.420
## femlab     -0.802  0.240
## marriage   -0.516  0.607  0.307
## birth      -0.379  0.041  0.066 -0.094
## military   -0.036  0.436 -0.311  0.530  0.128
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -1.4509327 -0.9760939 -0.6164694  1.1375377  2.1593261
##
## Residual standard error: 2.907665
## Degrees of freedom: 77 total; 71 residual
```

```r
(summary(model_ols))
```

```
##
## Call:
## lm(formula = divorce ~ . - year, data = divusa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8611 -0.8916 -0.0496  0.8650  3.8300
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.48784    3.39378   0.733   0.4659
```

```
## unemployed  -0.11125    0.05592  -1.989   0.0505 .
## femlab       0.38365    0.03059  12.543  < 2e-16 ***
## marriage     0.11867    0.02441   4.861 6.77e-06 ***
## birth       -0.12996    0.01560  -8.333 4.03e-12 ***
## military    -0.02673    0.01425  -1.876   0.0647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.65 on 71 degrees of freedom
## Multiple R-squared:  0.9208, Adjusted R-squared:  0.9152
## F-statistic: 165.1 on 5 and 71 DF,  p-value: < 2.2e-16
```

	Unemployed	Femlab	Marriage	Birth	Military
OLS	-0.113	.3836	.1187	-0.1300	-.0267
WLS	.108	.3121	.1643	-.0500	.0179

4. (10 points) Perform a t-test at the 5% significance level for each slope parameter for the OLS model in part 2 and the GLS model in part 4. Are there differences between which predictors are significant in the OLS model and which are significant in the GLS model? If so, state the changes.

   Unemployed becomes more significant in the GLS model

   Birth becomes more significant in the GLS model

   Military remains non-significant in both models.

   Femlab and marriage is highly significant in both models.

5. (5 points) For the GLS model in part 4, calculate and report the variance inflation factor (VIF) for each of the predictors using the `vif` function from the `car` package. Do any of these VIFs suggest we should be cautious about concluding a variable is "not significant" given the other predictors?

```
car::vif(model_gls)
```

```
## unemployed     femlab   marriage      birth   military
##   1.710203   1.905371   2.624558   1.148642   2.533990
```

   Unemployed: 1.71

   Femlab: 1.90

   Marriage: 2.62

   Birth: 1.15

   Military: 2.53

   None of the VIF's suggest that we should be cautious.

6. (5 points) Report the estimated value of the autocorrelation parameter ϕ and its associated 95% confidence interval. Does the interval indicate that ϕ is significantly different from zero at the 5% significance level?

```
intervals(model_gls)
```

```
## Approximate 95% confidence intervals
##
##  Coefficients:
```

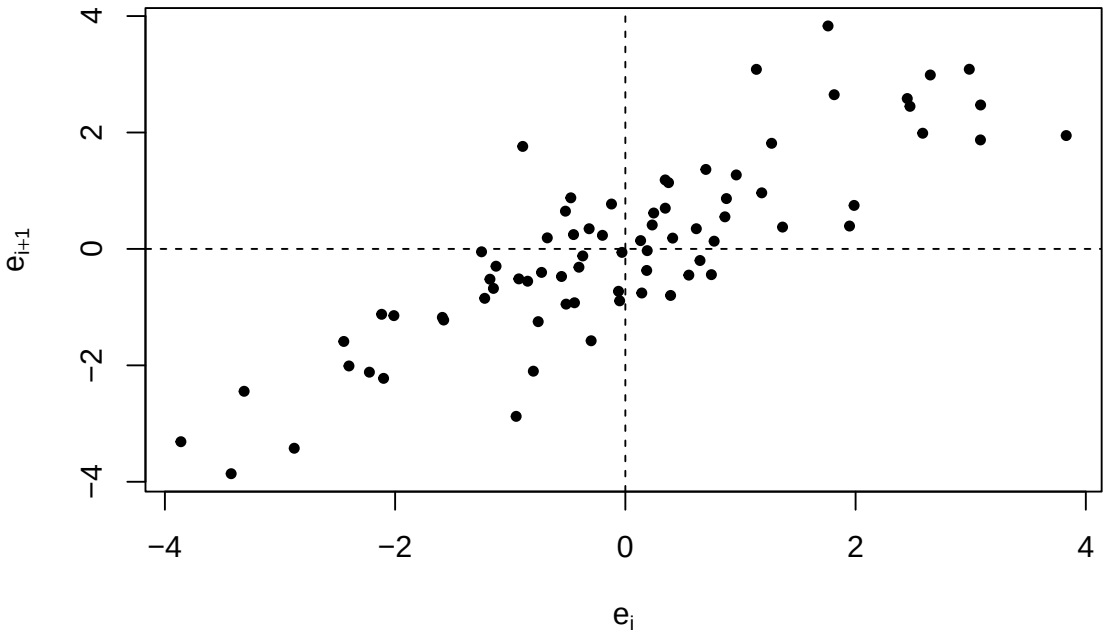
```
##                    lower        est.        upper
## (Intercept) -18.12047042 -7.05968163  4.001107160
## unemployed    0.01609101  0.10764313  0.199195251
## femlab        0.12235846  0.31208493  0.501811412
## marriage      0.11867101  0.16432630  0.209981587
## birth        -0.09380023 -0.04990919 -0.006018159
## military     -0.01050915  0.01794640  0.046401944
##
##  Correlation structure:
##         lower      est.     upper
## Phi 0.6527537 0.9715486 0.9980196
##
##  Residual standard error:
##     lower      est.     upper
##  0.797364  2.907665 10.603078
```

The confidence interval is (.653, .998), which does not include 0 so it is significantly different from zero at the 5% significance level.

7. (6 points) Check for serial correlation in the normalized errors of the GLS model in part 4 using a graphical method. Do you feel like the normalized errors are serially correlated? Justify your answer. Include any plots in your response.

```r
# plot of e_i vs. e_{i+1}
n = length(resid(model_ols))
plot(tail(resid(model_ols), n-1), head(resid(model_ols), n-1), pch = 20,
     xlab=expression(e[i]), ylab=expression(e[i+1]))

# lines at the x and y axes
abline(h=0, v=0, lty='dashed')
```



The points seem to follow a line, suggesting that they are serially correlated.

Exercise 2 (The `gala` Data Set) [40 points]

For this exercise, we will use the `gala` data set from the `faraway` package. You can also find the data set in `gala.csv` on Canvas. The data set contains the following variables:

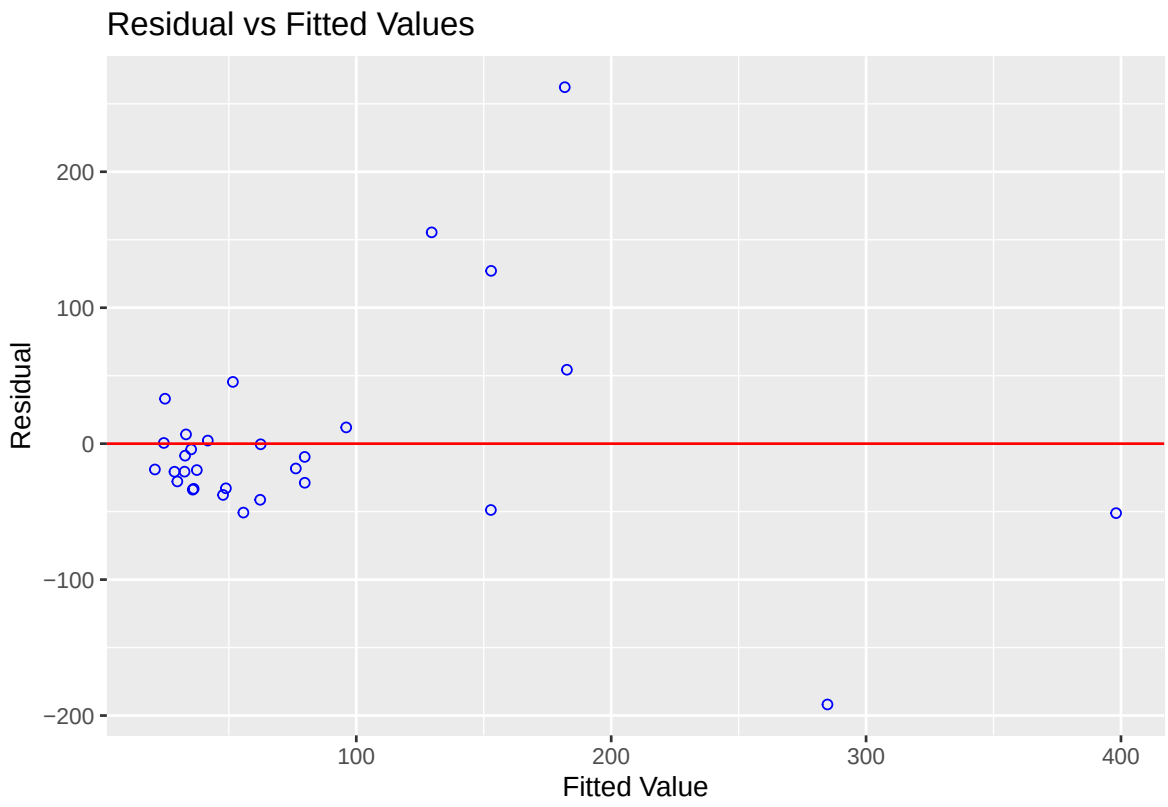
- `Species`: The number of plant species found on the island.
- `Area`: The area of the island (km^2).
- `Elevation`: The highest elevation of the island (m).
- `Nearest`: The distance from the nearest island (km).
- `Scruz`: The distance from Santa Cruz island (km).
- `Adjacent`: The area of the adjacent island (km^2).

In the following exercise, we will model `Species` in terms of `Area`, `Elevation`, and `Nearest`.

1. (5 points) Perform OLS regression with `Species` as the response and `Area`, `Elevation`, and `Nearest` as the predictors. Check the constant variance assumption for this model using a graphical method and a hypothesis test at the $\alpha = 0.05$ significance level. Do you feel it has been violated? Justify your answer. Include any plots in your response.

```r
# fit the model using OLS
model_ols = lm(Species ~ Area + Elevation + Nearest, data = gala)

# fitted-vs-residuals plot
ols_plot_resid_fit(model_ols)
```




```r
bptest(model_ols)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_ols
## BP = 11.184, df = 3, p-value = 0.01077
```

They seem to be clustered in the graph, and the BP test suggests that there is a violation of the constant variation assumption. I would conclude that there is a violation of the constant variance assumption.

2. (8 points) Perform a regression of the absolute value of the residuals from the model in part 1 against the predictors `Area`, `Elevation`, and `Nearest` using OLS. Report the estimated regression equation using all 3 predictors.

```r
model_wls = lm(abs(resid(model_ols)) ~ Area + Elevation + Nearest, data = gala)

# extract the coefficient estimates.
coef(model_wls)
```

```
## (Intercept)        Area   Elevation     Nearest
##  5.86799406 -0.03612868  0.14338356 -0.25577502
```

$$|e_i| = 5.867 - .0361 Area_i + .1434 Elevation_i - .2558 Nearest_i$$

1. (8 points) Perform WLS using the inverse of the squared fitted values from the model in part 2 as weights, i.e, `weights` $= 1/(\text{fitted values})^2$. Create and report a table with the OLS estimates (model in part 1) and WLS estimates for the slope parameters.

```
# calculate the weights as 1 / (fitted values)^2
weights = 1 / fitted(model_wls)^2

# run WLS
model_wls = lm(Species ~ Area + Elevation + Nearest, data = gala, weights = weights)
summary(model_ols)
```

```
##
## Call:
## lm(formula = Species ~ Area + Elevation + Nearest, data = gala)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -191.856  -33.111  -18.626    5.673  262.209
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.46471   23.38884   0.704  0.48772
## Area         0.01908    0.02676   0.713  0.48216
## Elevation    0.17134    0.05452   3.143  0.00415 **
## Nearest      0.07123    1.06481   0.067  0.94718
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 80.84 on 26 degrees of freedom
## Multiple R-squared:  0.5541, Adjusted R-squared:  0.5027
## F-statistic: 10.77 on 3 and 26 DF,  p-value: 8.817e-05
```

```
summary(model_wls)
```

```
##
## Call:
## lm(formula = Species ~ Area + Elevation + Nearest, data = gala,
##     weights = weights)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8301 -0.9225 -0.3112  0.4494  3.4435
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.65939    7.79303   0.726   0.4742
## Area         0.02237    0.03283   0.682   0.5015
## Elevation    0.17395    0.06066   2.868   0.0081 **
```

```
## Nearest      0.40385    0.17093   2.363   0.0259 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.31 on 26 degrees of freedom
## Multiple R-squared:  0.5339, Adjusted R-squared:  0.4801
## F-statistic: 9.927 on 3 and 26 DF,  p-value: 0.0001544
```

	Col1	Col2	Col3
OLS	.01908	.17134	.07123
WLS	.02237	.17395	.40385

3. (8 points) Perform a t-test at the 5% significance level for each slope parameter for the OLS model in part 1 and the WLS model in part 3. Are there differences between which predictors are significant in the OLS model and which are significant in the WLS model? If so, state the changes.

   Elevation remains significant in both models.

   Nearest is only significant in WLS

   Area remains non-significant for both

4. (5 points) For the WLS model in part 3, calculate and report the variance inflation factor (VIF) for each of the predictors using the `vif` function from the `car` package. Do any of these VIFs suggest we should be cautious about concluding a variable is "not significant" given the other predictors?

```r
vif(model_wls)
```

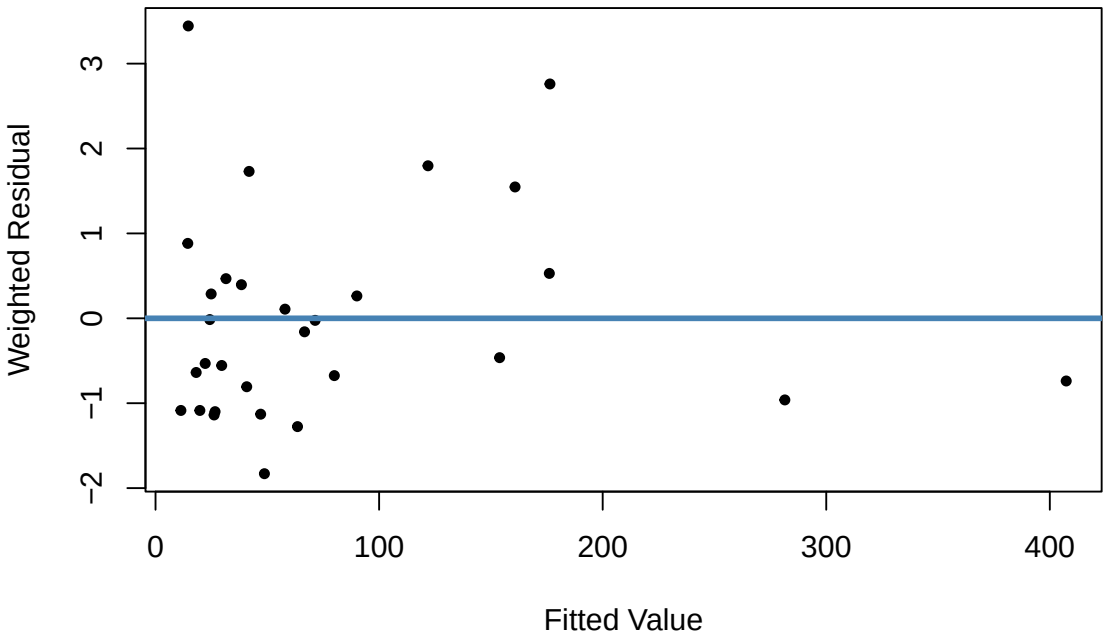
```
##      Area Elevation   Nearest
##  2.154149  2.156878  1.002607
```

   They are all below 5, suggesting that multicolinearity is not a problem.

5. (6 points) Check the constant variance assumption on the weighted residuals of the WLS model using a a graphical method and a hypothesis test at the $\alpha = 0.05$ significance level. Do you feel that it has been violated? Justify your answer. Include any plots in your response.

```r
plot(fitted(model_wls), weighted.residuals(model_wls),
     pch = 20, xlab = 'Fitted Value', ylab = 'Weighted Residual')

abline(h=0, lwd=3, col='steelblue')
```




```
bptest(model_wls)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_wls
## BP = 0.000812, df = 3, p-value = 1
```

Both the hypothesis test and the graphical test support that there is no violation of the constant variance assumption.

Exercise 3 (WLS for Survey Data) [10 points]

```
data = read.csv("chibus.csv")
```

For this exercise, we will use the the `chibus` data set, which can be found in `chibus.csv` on Canvas. Each observation in this data set represents a pair of zones in the city of Chicago. The variables in the data set are

- `computed_time`: travel times, computed from bus timetables augmented by walk times from zone centers to bus-stops (assuming a walking speed of 3 mph) and expected waiting times for the bus (= half of the time between successive buses).
- `perceived_time`: average travel times as reported to the U.S. Census Bureau by n travelers.

- `n`: number of travelers per observations for each case.

In the following exercise, we will model `perceived_time` in terms of `computed_time`.

1. (5 points) The variable `n` is not being used in the model, but it shows that the response is recorded as an average over different groups of size n_i. Based on this observation, what would make for a good choice of weights? No output is needed.

   $w_i = n_i$

2. (5 points) Perform WLS with `perceived_time` as the response and `computed_time` as the predictor using the weights you chose in part 1. Report the estimated regression equation for this model.

```
weights = data$n  # Using the number of travelers as weights

# Fit the WLS model
model_wls = lm(perceived_time ~ computed_time, data = data, weights = weights)

# Get the summary of the model to report the regression equation
summary(model_wls)
```

```
##
## Call:
## lm(formula = perceived_time ~ computed_time, data = data, weights = weights)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -20.278  -7.661  -0.680   4.543  33.219
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.2932     4.5903   0.500    0.621
## computed_time   1.1319     0.1475   7.676 1.46e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.01 on 30 degrees of freedom
## Multiple R-squared:  0.6626, Adjusted R-squared:  0.6514
## F-statistic: 58.93 on 1 and 30 DF,  p-value: 1.458e-08
```

$$perceived\_time = 2.29 + 1.132 computed\_time$$