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.

   Answer: This makes me suspect that the errors do not have equal variance. The data from year i is most likely dependent on the data from year i-1.

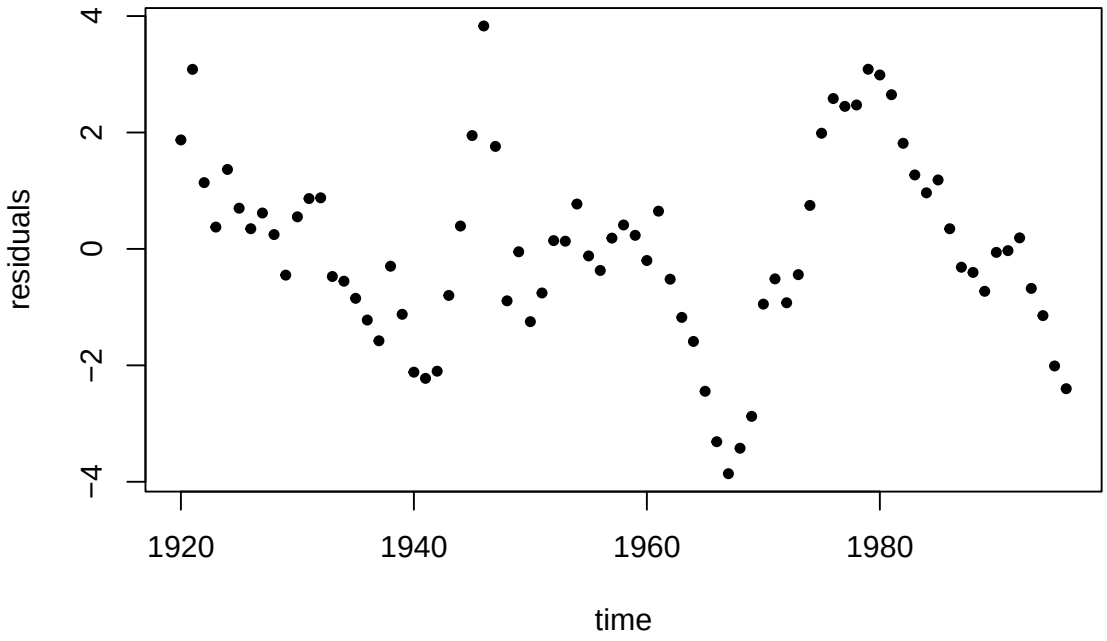
2. (6 points) Fit an OLS regression model with `divorce` as the response and all other variables except `year` as predictors. Check for serial correlation in the errors using a graphical method. Do you feel like the errors are serially correlated? Justify your answer. Include any plots in your response.

```r
library(faraway)

data(divusa, package='faraway')

model_ols <- lm(divorce~.-year, data=divusa)

plot(divusa$year, resid(model_ols), pch=20, xlab='time', ylab='residuals')
```



Answer: Yes, I feel like the errors are serially correlated. I plotted the residuals of the OLS model excluding 'year' vs. the attribute 'year' and there seems to be a clear correlation between the two.

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.

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.3.3
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
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
```

Answer: The test statistic is 0.2999 with a p-value of 2.2e-16. We reject the null hypothesis and concluded the errors follow an AR(1) process.

4. (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.

```
library(nlme)
```

```
## Warning: package 'nlme' was built under R version 4.3.3
```

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

ols_estimates <- coef(model_ols)
gls_estimates <- coef(model_gls)

data.frame(OLS = ols_estimates, GLS = gls_estimates)
```

```
##                      OLS         GLS
## (Intercept)  2.48784460 -7.05968163
## unemployed  -0.11125201  0.10764313
## femlab       0.38364928  0.31208493
## marriage     0.11867431  0.16432630
## birth       -0.12995915 -0.04990919
## military    -0.02673402  0.01794640
```

Answer: Above is a table with OLS and GLS estimates for the slope parameters.

5. (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.

```
summary(model_ols)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  2.48784460 3.39377886  0.7330603 4.659351e-01
## unemployed  -0.11125201 0.05592466 -1.9893192 5.051940e-02
## femlab       0.38364928 0.03058675 12.5429880 1.106450e-19
## marriage     0.11867431 0.02441430  4.8608516 6.771809e-06
## birth       -0.12995915 0.01559500 -8.3333840 4.027096e-12
## military    -0.02673402 0.01424726 -1.8764329 6.470590e-02
```

```
coef(summary(model_gls))
```

```
##                   Value  Std.Error   t-value      p-value
## (Intercept) -7.05968163 5.54719305 -1.272658 2.072911e-01
## unemployed   0.10764313 0.04591511  2.344395 2.185974e-02
## femlab       0.31208493 0.09515139  3.279878 1.610994e-03
## marriage     0.16432630 0.02289698  7.176766 5.561148e-10
## birth       -0.04990919 0.02201218 -2.267345 2.641518e-02
## military     0.01794640 0.01427099  1.257544 2.126774e-01
```

Answer: The predictors that are significant in the OLS model (predictors with a p-value < 0.05) are femlab, marriage, and birth. The predictors that are significant in the GLS model are unemployed, femlab, marriage, and birth. The predictor "unemployed" is significant in the GLS model but not in the OLS model.

6. (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?

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.3
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following objects are masked from 'package:faraway':
##
##     logit, vif
```

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

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

Answer: We are looking for predictors that have VIFs larger than 5. Since none of these predictors have VIFs larger than 5, we do not need to be cautious about concluding a variable is "not significant" given the other predictors.

7. (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)
```

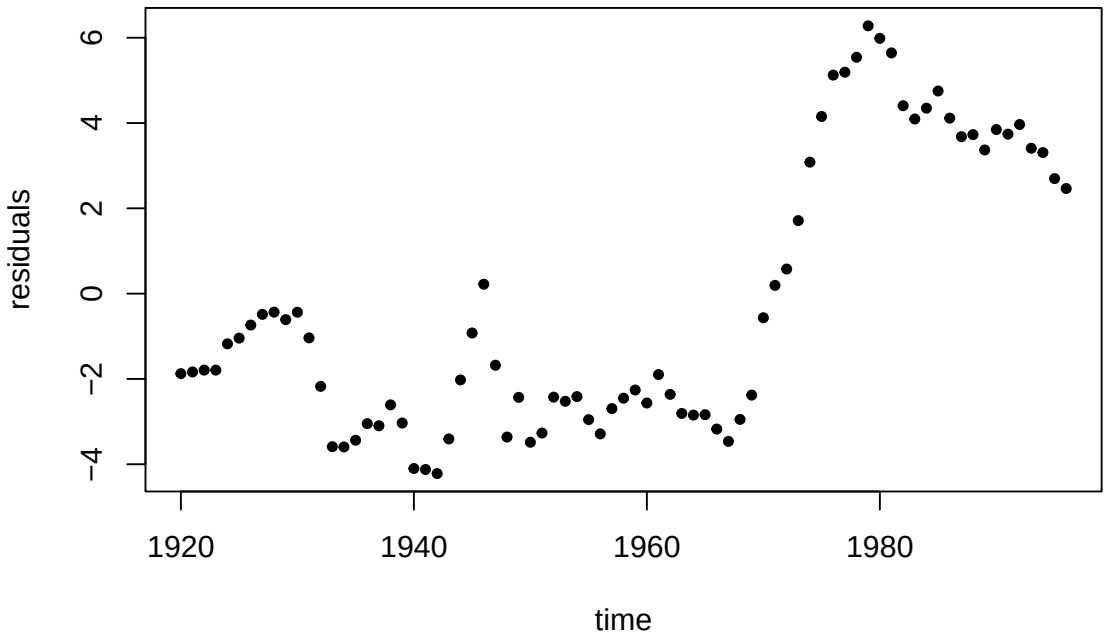
```
## Approximate 95% confidence intervals
##
##  Coefficients:
##                     lower        est.        upper
## (Intercept) -18.12047042 -7.05968163  4.001107160
```

```
## 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
```

Answer: We get a value of $\hat{\phi} = 0.972$ with a confidence interval of (0.653, 0.998). This interval does not cover zero, so we conclude that ϕ is significantly greater than zero, i.e., there is significant autocorrelation in the data.

8. (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.

```
plot(divusa$year, resid(model_gls), pch=20, xlab='time', ylab='residuals')
```



Answer: Yes, I feel like the errors are serially correlated. I plotted the residuals of the GLS model excluding 'year' vs. the attribute 'year' and there seems to be a clear correlation between the two.

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.

```
library(olsrr)
```

```
## Warning: package 'olsrr' was built under R version 4.3.3
```

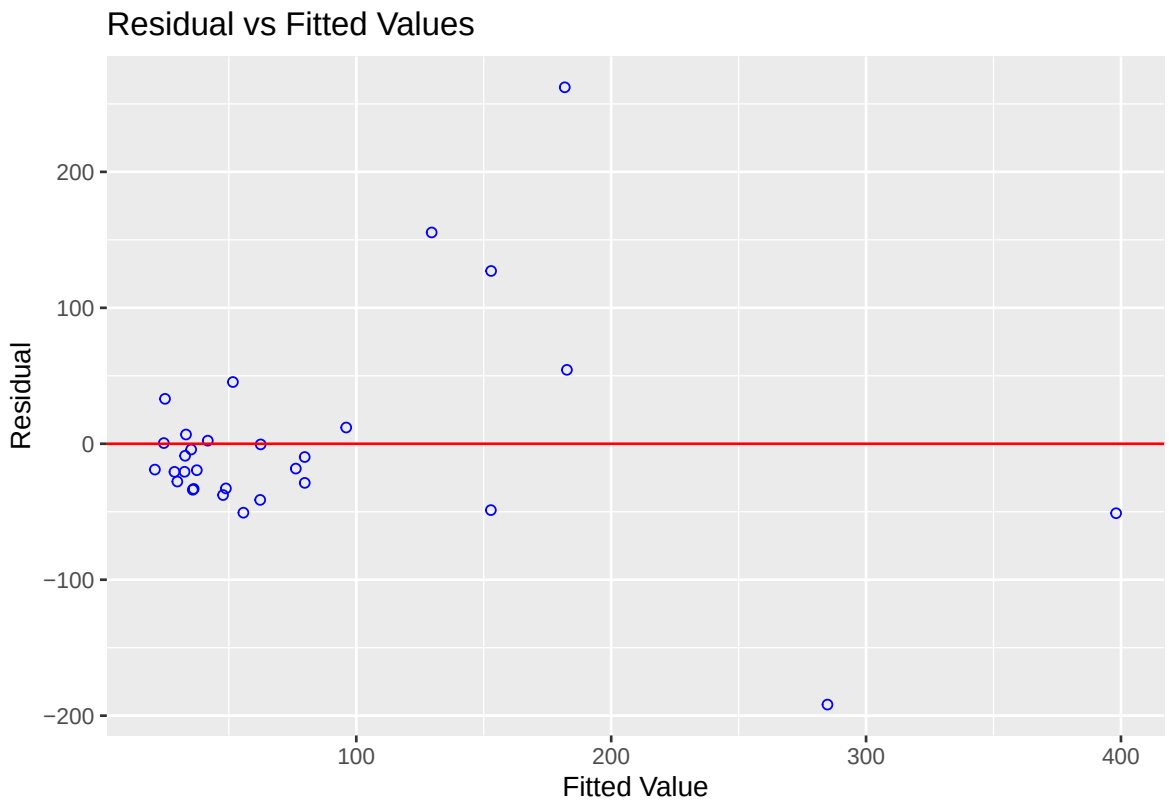
```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:faraway':
##
##     hsb
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
library(lmtest)

data(gala, package='faraway')
model_ols <- lm(Species ~ Area + Elevation + Nearest, data=gala)
ols_plot_resid_fit(model_ols)
```




```
bptest(model_ols)
```

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

Answer: Looking at the graph, the errors seem to be heteroscedastic. Upon performing a Breush-Pagan test, we get a p-value of 0.011. At a the $\alpha = 0.05$ significance level, we reject the null hypothesis and conclude that the errors are heteroscedastic and thus do not have constant variance.

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.

```
model_wts <- lm(abs(resid(model_ols))~Area+Elevation+Nearest,data=gala)

coef(model_wts)
```

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

Answer: The estimated regression equation using all 3 predictors is $|e_i| = 5.868 - 0.036 area_i + 0.143 elevation_i - 0.256 nearest_i$

3. (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.

```
weights = 1 / fitted(model_wts)^2

model_wls = lm(Species~Area+Elevation+Nearest, data=gala, weights=weights)

ols_estimates <- coef(model_ols)
wls_estimates <- coef(model_wls)

data.frame(OLS = ols_estimates, WLS = wls_estimates)
```

```
##                       OLS        WLS
## (Intercept) 16.46471112 5.65938759
## Area         0.01908464 0.02237259
## Elevation    0.17133627 0.17395271
## Nearest      0.07122724 0.40384843
```

Answer: Above is a table with the OLS and WLS estimates for the slope parameters.

4. (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.

```
summary(model_ols)$coefficients
```

```
##                Estimate  Std. Error    t value    Pr(>|t|)
## (Intercept) 16.46471112 23.38884127 0.70395583 0.487717672
## Area         0.01908464  0.02676432 0.71306297 0.482158270
## Elevation    0.17133627  0.05451917 3.14267925 0.004151162
## Nearest      0.07122724  1.06480583 0.06689223 0.947179238
```

```
coef(summary(model_wls))
```

```
##               Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) 5.65938759 7.79302547 0.7262119 0.474196308
## Area        0.02237259 0.03282529 0.6815657 0.501540058
## Elevation   0.17395271 0.06066342 2.8675057 0.008098258
## Nearest     0.40384843 0.17093444 2.3625926 0.025911489
```

Answer: The predictors that are significant in the OLS model (predictors with a p-value < 0.05) is only Elevation. The predictors that are significant in the WLS model are Elevation and Nearest. The predictor "Nearest" is significant in the WLS model but not in the OLS model.

5. (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?

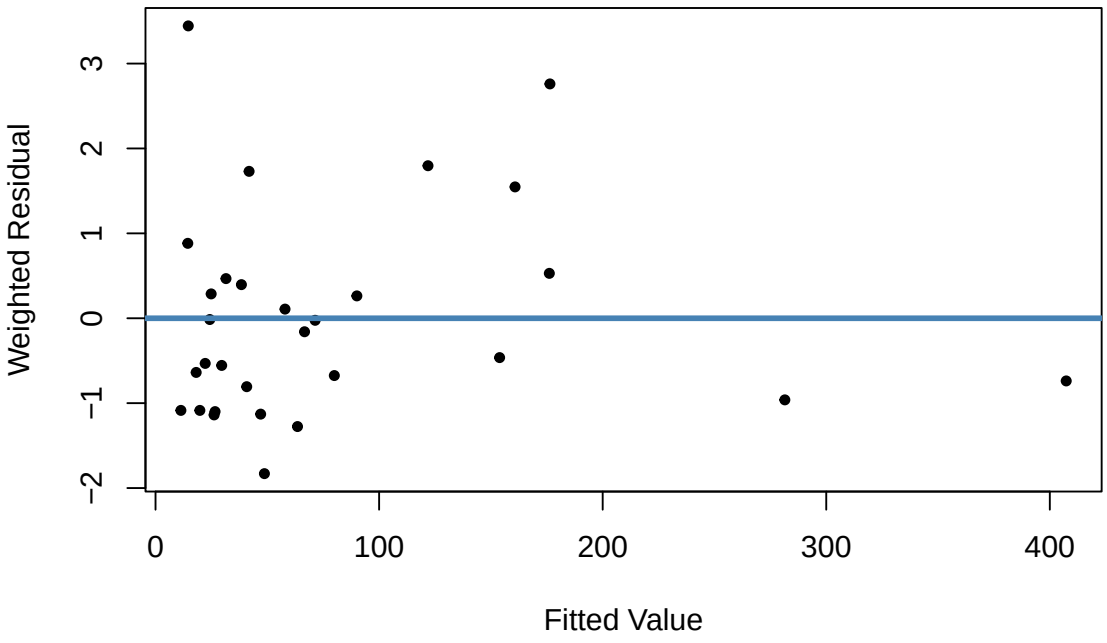
```
car::vif(model_wls)
```

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

Answer: We are looking for predictors that have VIFs larger than 5. Since none of these predictors have VIFs larger than 5, we do not need to be cautious about concluding a variable is "not significant" given the other predictors.

6. (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.

```
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
```

Answer: The errors look to have constant variance and the p-value of 1 supports this due to it being greater than 0.05. Therefore we do not reject the null and assume

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

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.

   Answer: We should use n_i as our weights.

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.

```
data <- read.csv('chibus.csv')

wls_model <- lm(perceived_time ~ computed_time, data = data, weights = n)
summary(wls_model)
```

```
##
## Call:
## lm(formula = perceived_time ~ computed_time, data = data, weights = n)
##
## 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
```

   Answer: Our estimated regression equation for this model is $y = 2.293 + 1.132 computedtime$.