Mock ALM midterm (Fall 2019)

Multiple choice questions

Multiple choice questions will have between 2-4 choice options and may cover anything from the course so far, including:

- Mean and median and when to use them
- Measures of spread for mean and median
- Right and left skew
- Interpreting frequencies and percents
- Use, interpretation, assumptions of chi-squared and one-sample t-test
- Identifying, using, interpreting histograms, boxplots, scatterplots
- When and why to use a correlation coefficient
- Interpreting correlation coefficients (strength, direction, significance)
- Assumptions of Pearson's r
- Assumptions of Spearman's rho
- Writing null and alternate hypotheses
- The NHST process
- When and why to use linear regression
- Assumptions for linear regression
- What outliers and influential values are
- The linear regression equation
- What residuals are
- The meaning of confidence intervals for correlation and regression
- Purpose of a Partial-F test and interpretation of partial-F test results

Short answer

The short answer part of the exam will provide you with R code and output for a correlation analysis and a linear regression analysis. Your job will be to write a thorough interpretation of the results. This may (or may not) entail assumption checking and diagnostics.

(1) What is the relationship between the distance to a syringe program and the percent of uninsured people in a county?

Codebook

The data are a sample of 500 counties in 2015 from the amFAR website. The variables have the following meanings:

- county: the county name
- STATEABBREVIATION: the two-letter abbreviation for the state the county is in
- dist_SSP: distance in miles to the nearest syringe services program
- HIVprevalence: people age 13 and older living with diagnosed HIV per 100,000
- opioid_RxRate: number of opioid prescriptions per 100 people
- pctunins: percentage of the civilian noninstitutionalized population with no health insurance coverage
- metro: county is non-metro, which includes open countryside, rural towns, or smaller cities with up to 49,999 people, or metro

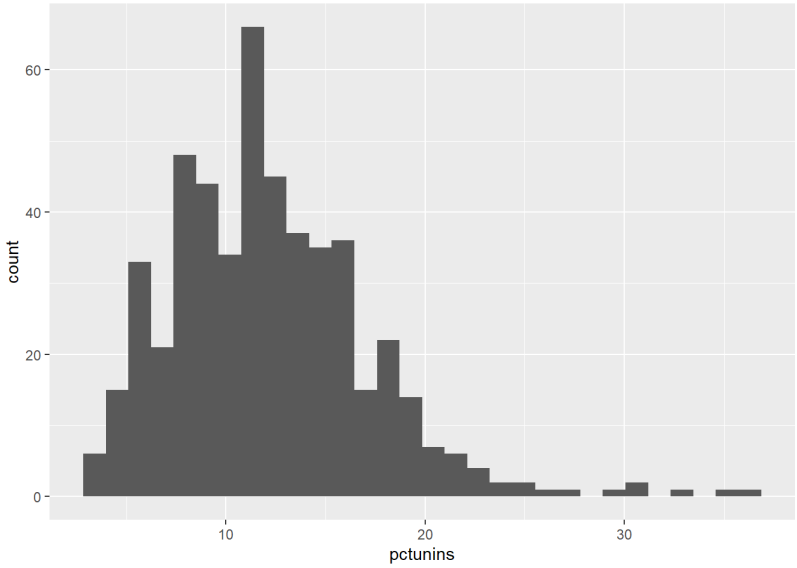
Importing data and descriptive stats

```
# distance to syringe program data
dist.ssp <- read.csv(file = "dist_ssp_amfar_ch9.csv")

# summary
summary(object = dist.ssp)
```

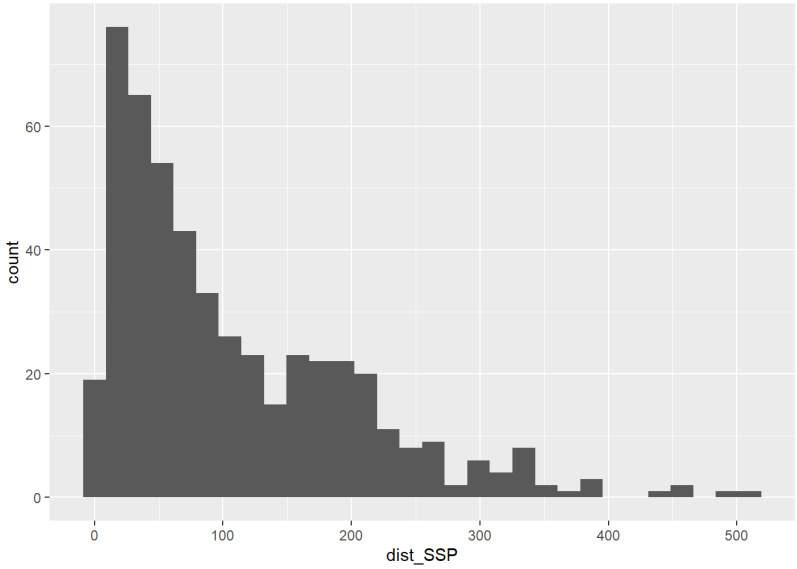
```
##               county    STATEABBREVIATION    dist_SSP
##  jackson county   :  5   TX     : 50       Min.   :  0.00
##  jefferson county :  5   GA     : 30       1st Qu.: 35.12
##  lincoln county   :  5   KS     : 21       Median : 75.94
##  washington county:  5   NC     : 21       Mean   :107.74
##  benton county    :  4   TN     : 21       3rd Qu.:163.83
##  decatur county   :  4   KY     : 19       Max.   :510.00
##  (Other)          :472   (Other):338
##  HIVprevalence     opioid_RxRate       pctunins           metro
##  Min.   :  -1.00   Min.   :  0.20   Min.   : 3.00   metro    :226
##  1st Qu.:  52.98   1st Qu.: 45.12   1st Qu.: 8.60   non-metro:274
##  Median : 101.15   Median : 62.40   Median :11.70
##  Mean   : 165.75   Mean   : 68.33   Mean   :12.18
##  3rd Qu.: 210.35   3rd Qu.: 89.95   3rd Qu.:15.00
##  Max.   :2150.70   Max.   :345.10   Max.   :35.90
##
```

Distribution of the two variables

```
# open tidyverse
library(package = "tidyverse")

# distributions of variables
dist.ssp %>%
  ggplot(aes(x = pctunins)) +
  geom_histogram()
```

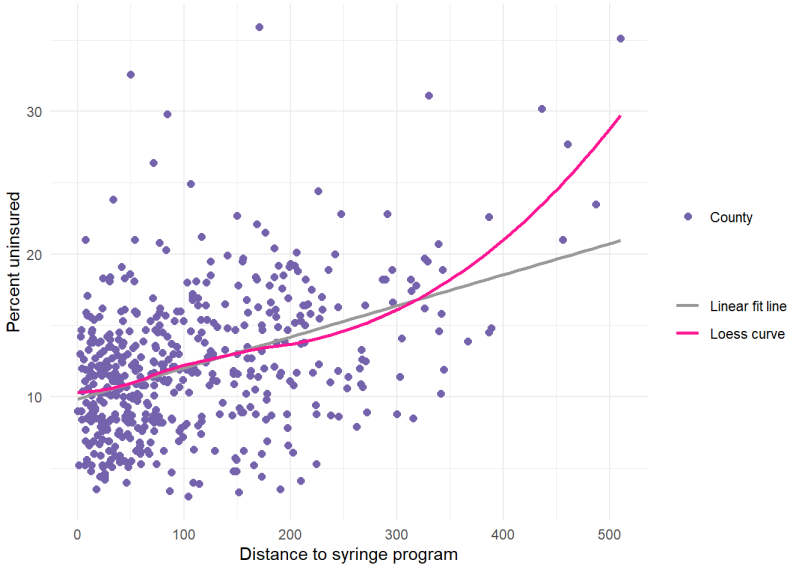



```
dist.ssp %>%
  ggplot(aes(x = dist_SSP)) +
  geom_histogram()
```



Relationship between the two variables

```
# relationship between uninsured percent and syringe program
dist.ssp %>%
  ggplot(aes(y = pctunins, x = dist_SSP)) +
  geom_point(aes(size = "County"), color = "#7463AC") +
  geom_smooth(aes(color = "Linear fit line"), method = "lm", se = FALSE) +
  geom_smooth(aes(color = "Loess curve"), se = FALSE) +
  theme_minimal() +
  labs(y = "Percent uninsured",
       x = "Distance to syringe program") +
  scale_color_manual(values = c("gray60", "deeppink"), name= "") +
  scale_size_manual(values = 2, name = "")
```



Correlation analysis

```
# Pearson and Spearman correlations
cor.test(x = dist.ssp$pctunins,
         y = dist.ssp$dist_SSP)
```

```
##
##  Pearson's product-moment correlation
##
## data:  dist.ssp$pctunins and dist.ssp$dist_SSP
## t = 10.11, df = 498, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3371858 0.4828903
## sample estimates:
##       cor
## 0.4126744
```

```
cor.test(x = dist.ssp$pctunins,
         y = dist.ssp$dist_SSP,
         method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  dist.ssp$pctunins and dist.ssp$dist_SSP
## S = 13642588, p-value = 1.958e-15
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3451532
```

Provide a complete interpretation of the results in order to answer the research question. Check the assumptions and use your assumption checking results to select the most appropriate statistical results to interpret.

(2) Use linear regression to determine whether the percent of people living a dollar per day or less and the access to basic water can help to predict the number of school-aged people in school?

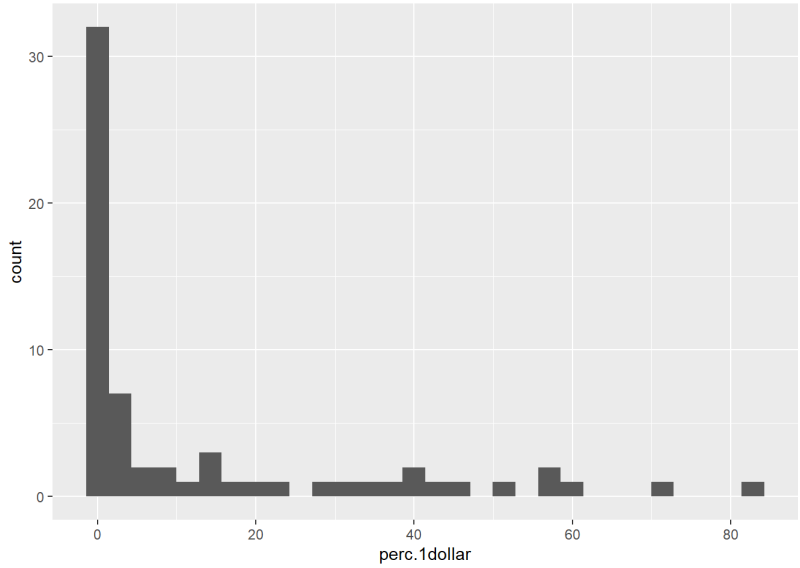
Codebook:

- country: the name of the country
- med.age: the median age of the citizens in the country
- perc.1dollar: percentage of citizens living on $1 per day or less
- perc.basic2015sani: percentage of citizens with basic sanitation access
- perc.safe2015sani: percentage of citizens with safe sanitation access
- perc.basic2015water: percentage of citizens with basic water access
- perc.safe2015water: percentage of citizens with safe water access
- perc.in.school: percentage of school-age people in primary and secondary school
- female.in.school: percentage of female school-age people in primary and secondary school
- male.in.school: percentage of male school-age people in primary and secondary school

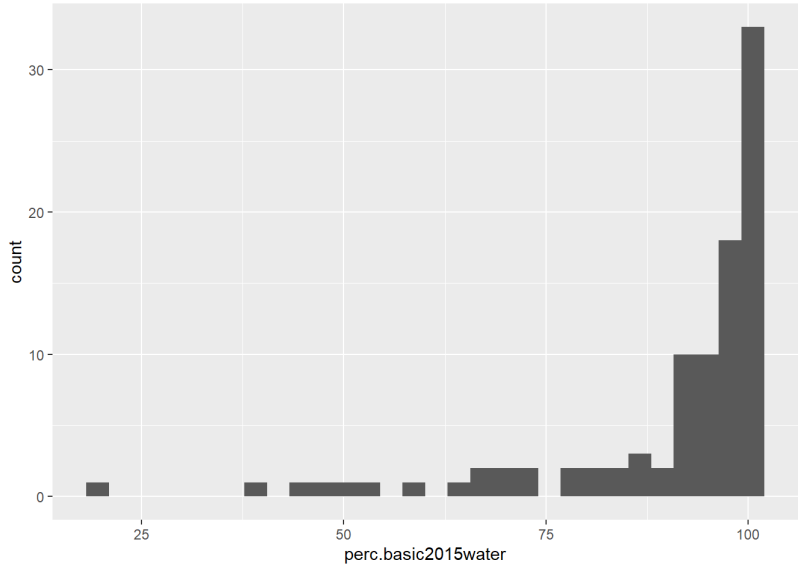
```
# import water and education data
water.educ <- read.csv("water_educ_2015_who_unesco_ch8.csv")

# make smaller data set with complete data
water.educ.small <- water.educ %>%
  select(perc.1dollar, perc.basic2015water, perc.in.school)

# summarize
summary(object = water.educ.small)
```

```
##   perc.1dollar   perc.basic2015water perc.in.school
##  Min.   : 1.00   Min.   : 19.00      Min.   :33.32
##  1st Qu.: 1.00   1st Qu.: 88.75      1st Qu.:83.24
##  Median : 1.65   Median : 97.00      Median :92.02
##  Mean   :13.63   Mean   : 90.16      Mean   :87.02
##  3rd Qu.:17.12   3rd Qu.:100.00      3rd Qu.:95.81
##  Max.   :83.80   Max.   :100.00      Max.   :99.44
##  NA's   :33      NA's   :1
```

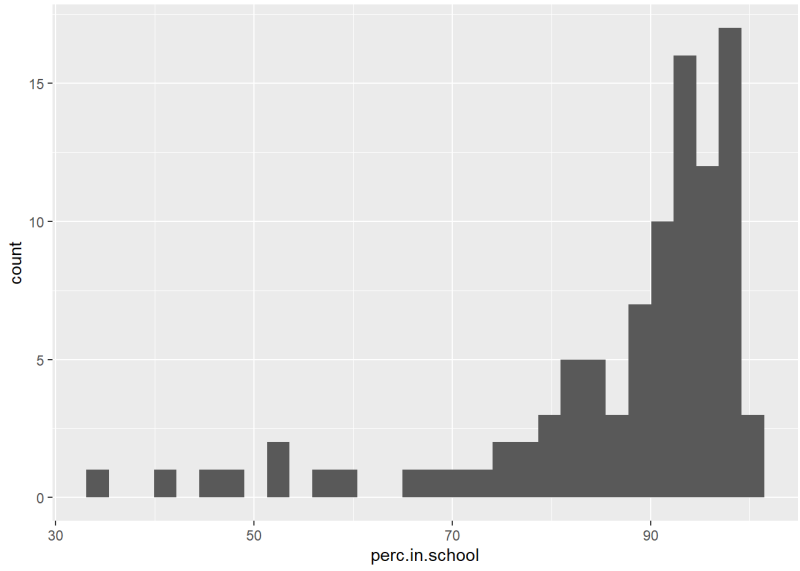
```
# variable distributions
water.educ.small %>%
  ggplot(aes(x = perc.1dollar)) +
  geom_histogram()
```



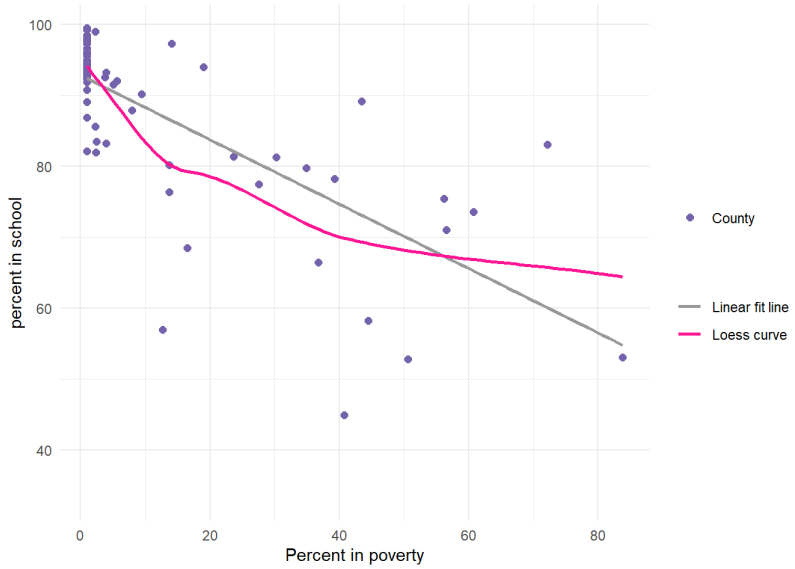

```
water.educ.small %>%
  ggplot(aes(x = perc.basic2015water)) +
  geom_histogram()
```



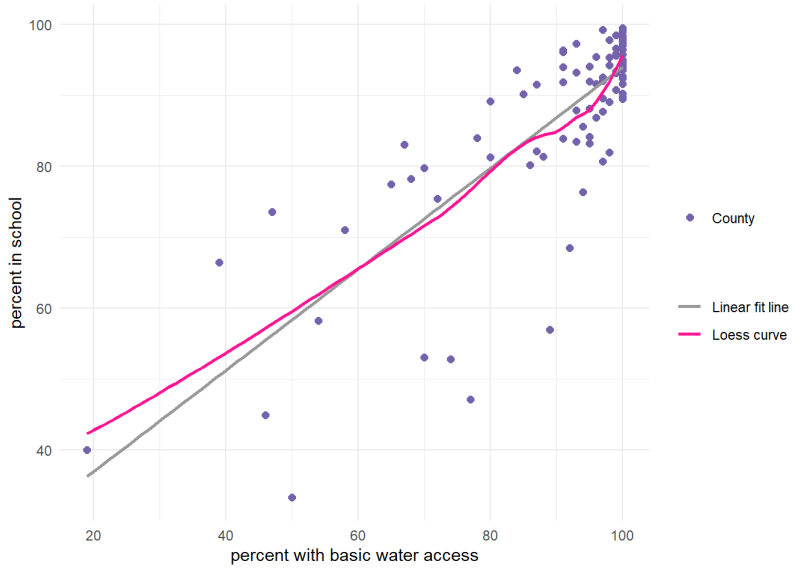

```
water.educ.small %>%
  ggplot(aes(x = perc.in.school)) +
  geom_histogram()
```



```
# relationship between
water.educ.small %>%
  ggplot(aes(y = perc.in.school, x = perc.1dollar)) +
  geom_point(aes(size = "County"), color = "#7463AC") +
  geom_smooth(aes(color = "Linear fit line"), method = "lm", se = FALSE) +
  geom_smooth(aes(color = "Loess curve"), se = FALSE) +
  theme_minimal() +
  labs(y = "percent in school",
       x = "Percent in poverty") +
  scale_color_manual(values = c("gray60", "deeppink"), name= "") +
  scale_size_manual(values = 2, name = "")
```




```
water.educ.small %>%
  ggplot(aes(y = perc.in.school, x = perc.basic2015water)) +
  geom_point(aes(size = "County"), color = "#7463AC") +
  geom_smooth(aes(color = "Linear fit line"), method = "lm", se = FALSE) +
  geom_smooth(aes(color = "Loess curve"), se = FALSE) +
  theme_minimal() +
  labs(y = "percent in school",
       x = "percent with basic water access") +
  scale_color_manual(values = c("gray60", "deeppink"), name= "") +
  scale_size_manual(values = 2, name = "")
```




```
# linear regression
school.by.pov.water <- lm(formula = perc.in.school ~ perc.1dollar +
                        perc.basic2015water, data = water.educ.small,
                        na.action = na.exclude)
summary(object = school.by.pov.water)
```

```
##
## Call:
## lm(formula = perc.in.school ~ perc.1dollar + perc.basic2015water,
##     data = water.educ.small, na.action = na.exclude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.899  -2.703   1.718   4.649  16.803
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         51.48993   11.85655   4.343  5.4e-05 ***
## perc.1dollar        -0.19011    0.09035  -2.104 0.039502 *
## perc.basic2015water  0.42455    0.12116   3.504 0.000864 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.161 on 61 degrees of freedom
##   (33 observations deleted due to missingness)
## Multiple R-squared:  0.6069, Adjusted R-squared:  0.5941
## F-statistic:  47.1 on 2 and 61 DF,  p-value: 4.275e-13
```

```
# confidence interval for regression parameters
ci.dist.by.unins.met <- confint(school.by.pov.water)
ci.dist.by.unins.met
```

```
##                         2.5 %       97.5 %
## (Intercept)         27.7812925 75.198576127
## perc.1dollar        -0.3707807 -0.009436747
## perc.basic2015water  0.1822857  0.666819031
```

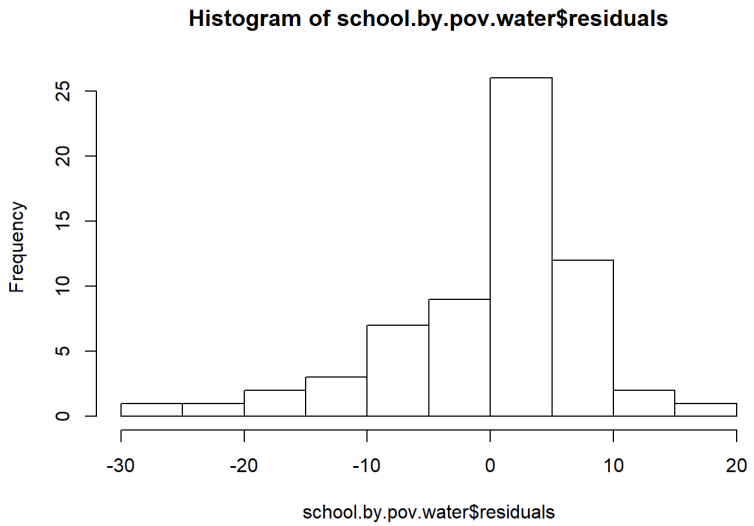
```
# Breusch-Pagan test
library(package = "lmtest")
bptest(formula = school.by.pov.water)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  school.by.pov.water
## BP = 7.9079, df = 2, p-value = 0.01918
```

```
# Durbin-Watson test
dwtest(formula = school.by.pov.water)
```

```
##
##  Durbin-Watson test
##
## data:  school.by.pov.water
## DW = 1.7445, p-value = 0.1499
## alternative hypothesis: true autocorrelation is greater than 0
```

```
# Plot residuals
hist(x = school.by.pov.water$residuals)
```




```
# multicollinearity
car::vif(school.by.pov.water)
```

```
##        perc.1dollar perc.basic2015water
##            3.250689            3.250689
```

```
# Outliers and influential
water.educ.small.diag <- water.educ.small %>%
  mutate(standardres = rstandard(model = school.by.pov.water)) %>%
  mutate(lever = hatvalues(model = school.by.pov.water)) %>%
  mutate(cooks.dist = cooks.distance(model = school.by.pov.water))%>%
  mutate(predict.distance = predict(object = school.by.pov.water))

 water.educ.small.diag %>%
  filter(cooks.dist > 4/n() & abs(x = lever) > 2*(3+1)/n() |
           cooks.dist > 4/n() & abs(x = standardres) > 1.96 |
           abs(x = standardres) > 1.96 & abs(x = lever) > 2*(3+1)/n() )
```

```
##   perc.1dollar perc.basic2015water perc.in.school standardres      lever
## 1         36.8                  39       66.40913   0.7924158 0.31361895
## 2         83.8                  70       53.11508  -1.8280405 0.33535562
## 3         72.2                  67       83.01217   2.2892836 0.19108948
## 4         50.6                  74       52.80126  -2.6192393 0.08148165
## 5         60.7                  47       73.52982   1.7946848 0.13452989
## 6         40.8                  46       44.95451  -2.4967750 0.19264657
## 7         12.7                  89       56.96193  -3.6927635 0.01570351
##   cooks.dist predict.distance
## 1 0.09563613         61.05147
## 2 0.56203920         65.27749
## 3 0.41268077         66.20909
## 4 0.20286218         73.28731
## 5 0.16688675         59.90429
## 6 0.49583270         63.26291
## 7 0.07251913         86.86071
```

Provide a complete interpretation of the results in order to answer the research question. Include statements about assumptions and diagnostics. Even if you do not meet the assumptions, interpret the linear regression model as if you did meet assumptions.