Problem Set 2

Due date: 25 September

Please upload your completed assignment to the ELMs course site (under the assignments menu). Remember to include an annotated script file for all work with R and show your math for all other problems (if applicable, or necessary). Please also upload your completed assignment to the Github repository that you have shared with us. *We should be able to run your script with no errors.*

Total points: 30

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.3     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- 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 becor
```

```
library(states)
```

```
Attaching package: 'states'

The following object is masked from 'package:readr':

    parse_date
```

```
library(poliscidata)
```

```
Registered S3 method overwritten by 'gdata':
  method         from
  reorder.factor gplots
```

```
library(ggplot2)
library(wbstats)
library(countrycode)
library(broom)
library(janitor)
```

```
Attaching package: 'janitor'

The following object is masked from 'package:poliscidata':

    crosstab

The following objects are masked from 'package:stats':

    chisq.test, fisher.test
```

```
library(ggridges)
library(modelsummary)
```

Question 1

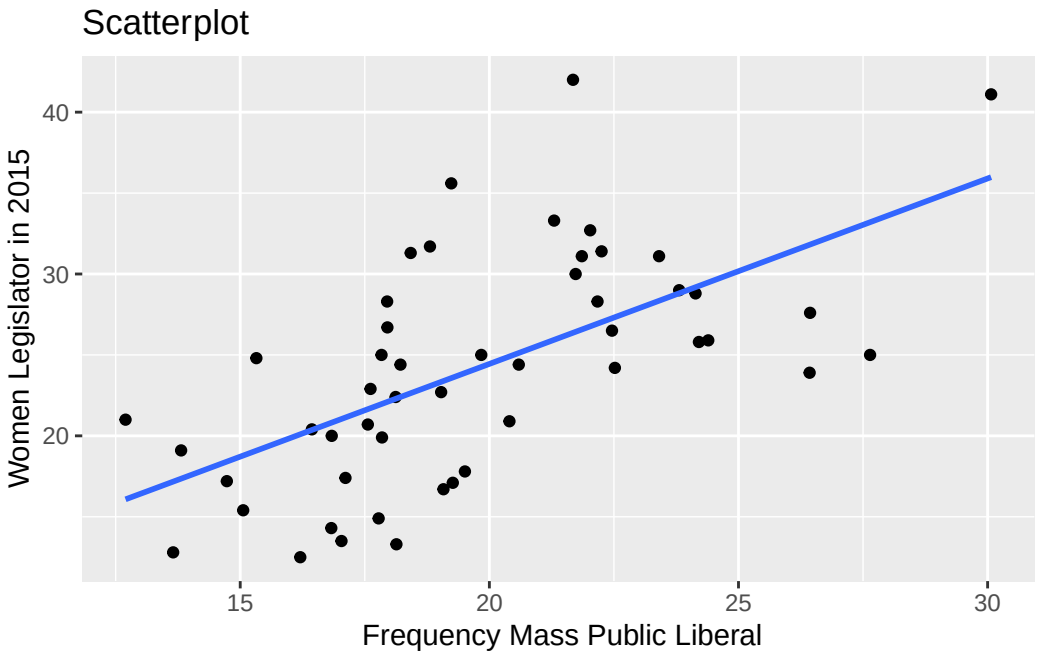
Points: 5

Using the `states` data, produce a scatterplot of the variables `womleg_2015` and `libpct_m` (with `womleg_2015` as the dependent variable on the y-axis). Describe the scatterplot and include a copy of it. Note any suspected outliers, if any (a visual inspection will suffice for this question). Lastly, give the general equation for the correlation between `womleg_2015` and `libpct_m` (include as much information as possible), but do not solve it.

> **Note**
>
> The `states` data set can be found in `poliscidata::states`. Take a look at `?states` to see what these variables measure.

```
# Scatterplot
ggplot(data = states, aes(x = libpct_m, y = womleg_2015)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Scatterplot",
    x = " Frequency Mass Public Liberal",
    y = "Women Legislator in 2015"
  )
```

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




```
cor(states$womleg_2015, states$libpct_m)
```

```
[1] 0.6088832
```

ANSWER: Correlation equation:

$$Y = \beta_0 + \beta_1(x)$$

Y = womleg_2015 + 0.6088(libpct_m)

Scatter plot description: There are some outliers that we cannot explain on the scatter plot

Question 2

Points: 5

Regress `womleg_2015` (as the dependent variable) on `libpct_m` and report the results in a professionally formatted table. Write the model equation with the estimated coefficients and interpret them. What does the value of R^2 tell us about this model?

Model equation: Y = womleg_2015 + (libpct_m)(x)

$$Y = 1.53 + 1.15(libpct_m)$$

```
regress_m_1 <- lm(womleg_2015 ~ libpct_m, states)
regress_m_1
```

```
Call:
lm(formula = womleg_2015 ~ libpct_m, data = states)

Coefficients:
(Intercept)     libpct_m
      1.524        1.146
```

```
modelsummary(
  regress_m_1,
  statistic = NULL,
  coef_rename = c("libpct_m" = "Public_liberal"),
  gof_map = "nobs"
)
```

```
summary(regress_m_1)
```

	(1)
(Intercept)	1.524
Public_liberal	1.146
Num.Obs.	50

```
Call:
lm(formula = womleg_2015 ~ libpct_m, data = states)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0061 -3.9376 -0.5102  4.1746 15.6324

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5240     4.3293   0.352    0.726
libpct_m      1.1460     0.2155   5.318 2.71e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.616 on 48 degrees of freedom
Multiple R-squared:  0.3707,    Adjusted R-squared:  0.3576
F-statistic: 28.28 on 1 and 48 DF,  p-value: 2.709e-06
```

```
glance(regress_m_1)
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic    p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>      <dbl> <dbl>  <dbl> <dbl> <dbl>
1     0.371         0.358  5.62      28.3 0.00000271     1  -156.  318.  324.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

ANSWER: R squared is around .3707, which shows us a relatively weak relationship with the regression model and won't be able describe our dependent variable with confidence.

Question 3

Points: 5

Based on this regression, find the predicted value, the observed value, and compute the residual for the state of Colorado and then the state of Georgia. Lastly, compute the total aggregate error from those two select observations combined (i.e., Colorado and Georgia).

> Tip
>
> Think RSS.

```
regress_m_1
```

```
Call:
lm(formula = womleg_2015 ~ libpct_m, data = states)

Coefficients:
(Intercept)     libpct_m
      1.524        1.146
```

```
tidy(regress_m_1)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic    p.value
  <chr>          <dbl>     <dbl>     <dbl>      <dbl>
1 (Intercept)     1.52     4.33      0.352 0.726
2 libpct_m        1.15     0.215     5.32  0.00000271
```

```
library(poliscidata)
library(states)
regress_m_1_cg <- tidy(regress_m_1)

augment(regress_m_1)
```

```
# A tibble: 50 x 8
   womleg_2015 libpct_m .fitted .resid   .hat .sigma  .cooksd .std.resid
         <dbl>    <dbl>   <dbl>  <dbl>  <dbl>  <dbl>    <dbl>      <dbl>
 1        28.3     17.9    22.1  6.21  0.0248   5.60 0.0159        1.12
 2        14.3     16.8    20.8 -6.51  0.0326   5.59 0.0234       -1.18
 3        20       16.8    20.8 -0.819 0.0325   5.67 0.000369     -0.148
 4        35.6     19.2    23.6 12.0   0.0204   5.39 0.0488        2.16
```

```
 5        25.8     24.2    29.3 -3.46  0.0493   5.65 0.0104       -0.633
 6        42       21.7    26.4 15.6   0.0255   5.18 0.104         2.82
 7        28.3     22.2    26.9  1.37  0.0286   5.67 0.000901      0.247
 8        24.2     22.5    27.3 -3.13  0.0313   5.66 0.00518      -0.566
 9        24.4     20.6    25.1 -0.721 0.0210   5.67 0.000181     -0.130
10        22.9     17.6    21.7  1.19  0.0267   5.67 0.000632      0.215
# i 40 more rows
```

```
beta_0 <- regress_m_1_cg |>
  filter(term == "(Intercept)") |>
  pull(estimate)

beta_0
```

```
[1] 1.52399
```

```
beta_1 <- regress_m_1_cg |>
  filter(term == "libpct_m") |>
  pull(estimate)
beta_1
```

```
[1] 1.145986
```

```
?state
states |>
  filter(stateid%in%c("CO    ", "GA    ")) |>
  select(stateid, libpct_m)
```

```
  stateid libpct_m
1  CO     21.67878
2  GA     17.61538
```

```
#predicted value CO

beta_0 + beta_1 * 21.67878
```

```
[1] 26.36756
```

```
#predicted value GA

beta_0 + beta_1 * 17.61538
```

```
[1] 21.71096
```

```
#observed

#GA=17.61538 CO=21.67878

#residual 4.0634

 21.67878-17.61538
```

```
[1] 4.0634
```

ANSWER:

Predicted values

GA: 21.71096

CO: 26.36756

Residual: 4.0634

Question 4

Points: 5

Using the `states` dataset, assess the relationship between the following two variables: `obama_win12` and `gun_rank3`. Construct a cross-tab and describe the nature of the relationship (if any) in detail.

> **Note**
>
> The variable `Obama_win12` is a dichotomous indicator of whether Obama won the state in 2012 (Obama won; Obama lost). The variable `gun_rank3` represents the general (ordinal) extent of gun restrictions in each state (more restrictions; middle restrictions; less restrictions).

obama_win12		More restr	Mid	Less restr	All
No	N	1	5	18	24
	% row	4.2	20.8	75.0	100.0
Yes	N	14	9	3	26
	% row	53.8	34.6	11.5	100.0
All	N	15	14	21	50
	% row	30.0	28.0	42.0	100.0

> **Caution**
>
> Please note that you would customarily want a greater number of observations within each cell before conducting such an analysis.

```
library(poliscidata)
library(modelsummary)

datasummary_crosstab(obama_win12 ~ gun_rank3, data = states)
```

ANSWER: The relationship between gun restriction and Obama winning states in 2012, when there were more restrictions on guns, Obama won(restriction 14). When Obama had less restriction (1) on guns, he lost the state.

Question 5

Points: 5

I hypothesize that religious identifiers in the mass public are less likely to support federal government support of scientific research. I use data from the General Social Survey to evaluate this hypothesis. In particular, I use a three-category indicator of religious attendance to measure religious identification (low attendance; moderate attendance; high attendance) and a three-category indicator of perceptions toward the federal government's support for scientific research (federal government provides "too little" support; "about right"; federal government provides "too much" support). Complete the cross-tab below so that you may properly evaluate my hypothesis.

> **Note**
>
> Table entries are raw counts of observations within each cell.

Relig. Attendance	Supporting Scientific Research: Too Little	About Right	Too much	Total
Low	342	356	106	
Moderate	190	213	51	
High	182	287	91	
Total				

ANSWER: The column totals

Too little Column: 714, 39.2739%

About right: 856, 47.0847%

Too much: 248, 13.6414%

The row totals:

Low: 804, 44.2244%

Moderate: 454, 24.9725%

High: 560, 30.8030

total row/column: 1,818

Evaluate: The hypothesis is moderately weak. When there is high religious attendance, we should see a high number in perception that the government provides "too much" government support for scientific research; but we don't, its 5.0055% compared to 10.0110% (too little) and 15.7866% (about right).

But if we go the other way, if they have low religious attendance, based on the hypothesis we should expect that their perception is either "too little" or "about right". This is what we see in the cross tab (low 18.8119%, about right 19.5819% and too much 5.8310%)

Question 6

Points: 5

Say I wish to explore the relationship between the relative advantage of Democrats (`dem_advantage`) in a state and abortion policy (`abort_rank3`). The `dem_advantage` variable is a continuous indicator where higher values represent a greater Democratic advantage among the mass public; `abort_rank3` is an ordinal indicator for the extent of abortion restrictions in each state (fewer restrictions; middle restrictions; more restrictions). To explore this relationship, complete the following:

Part A

Create a new variable (i.e., `dem_adv`) based on the `dem_advantage` variable. Calculate the summary statistics of `dem_advantage` and assign the following values to our new variable: if `dem_advantage` is less than the first quartile, set `dem_adv` to `Low`; if the value for `dem_advantage` is greater than the first quartile and less than the third quartile, set the value to `Mid`; and if the value of `dem_advantage` is greater than the third quartile, set the value to `High`.

```
quartiles <- quantile(states$dem_advantage, probs = c(0.25, 0.75))
quartiles
```

```
   25%    75%
-4.875 10.925
```

```
states <- states <- states %>%
  mutate(dem_adv = case_when(
    dem_advantage < quartiles[1] ~ "Low",
    dem_advantage >= quartiles[1] & dem_advantage < quartiles[2] ~ "Mid",
    dem_advantage > quartiles[2] ~ "High"
  ))

summary(states$dem_advantage)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-36.700  -4.875   0.950   0.948  10.925  24.000
```

ANSWER: High: 13 Mid: 24 Low: 13

abort_rank3		High	Low	Mid	All
More restr	N	0	6	11	17
	% row	0.0	35.3	64.7	100.0
Mid	N	4	5	8	17
	% row	23.5	29.4	47.1	100.0
Less restr	N	9	2	5	16
	% row	56.2	12.5	31.2	100.0
All	N	13	13	24	50
	% row	26.0	26.0	48.0	100.0

Part B

Create a crosstab using R; include your results in a professionally formatted table.

```
datasummary_crosstab(abort_rank3 ~ dem_adv, data = states)
```

Part C

What relationship (if any) is there between the relative advantage of Democrats is a given state and the restrictiveness of Abortion policy?

We see a relationship between the relative advantage of Democrats and restrictiveness of abortion policy;

- When a state has lower abortion restrictions, the Democrat advantage in the state is high 56.2%,
- When the state as more restrictive abortion policy, we see the democrat advantage is not high (0.00%) at all, it is low (35.3%) to mid (64.7%)