Name _____

POL 138 Quantitative Reasoning in Political Science
Practice · Linear regression

[Items 1 through 5] Below are coefficient estimates from a linear regression on data from residents in a hypothetical country. The linear regression used the number of years of education that a resident has (X) to predict that resident's support for the country's president on a scale from 0 to 100 (Y).

	Coefficient:
Constant/Intercept	90.00
Education	-3.00

1. Write an equation using X to predict Y.

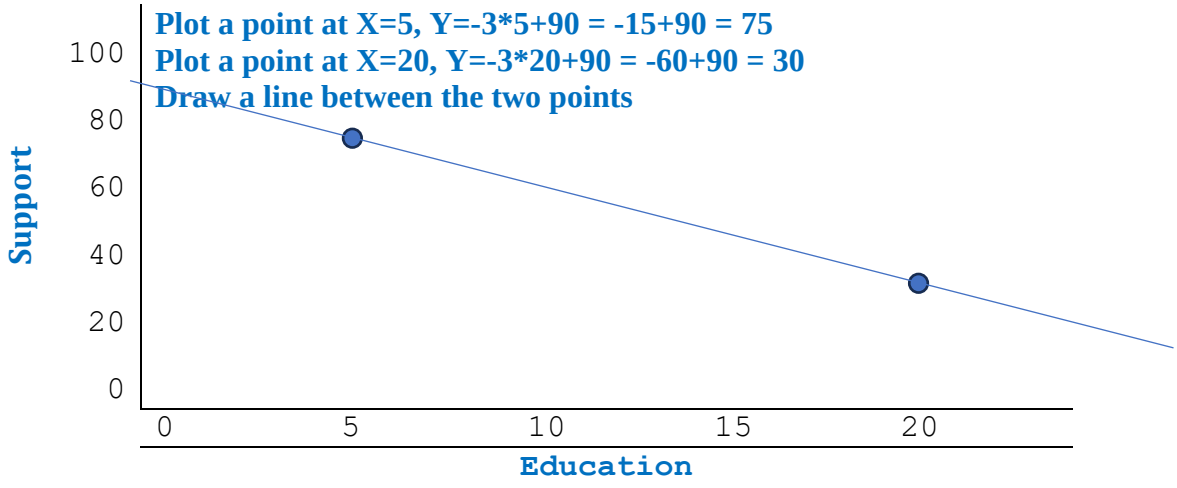
   Y = 90 + -3X
   Also acceptable:
   Y = -3X + 90
   SUPPORT = 90 + -3*Education
   SUPPORT = -3*Education + 90

2. In the plot below, label the Y-axis "Support for the President", and label the X-axis "Education".

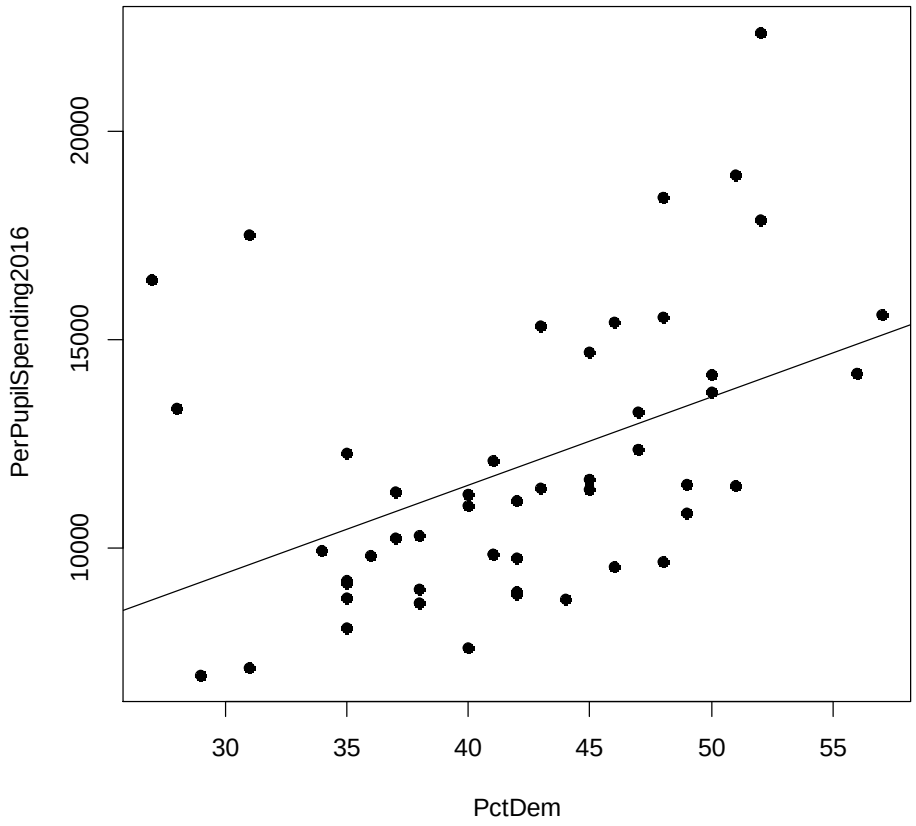
3. In the plot below, draw and label a point at the value of Y for which the X variable is 5 (the lowest observed level of education).

4. In the plot below, draw and label a point at the value of Y for which the X variable is 20 (the highest observed level of education).

5. In the plot below, draw a straight line between the above two points.



[Items 6 through 12] The plot below indicates, for each state in the United States, the per pupil spending in the state in dollars in 2016 (PerPupilSpending2016) and the percentage of state residents who were Democrats in 2016 (PctDem). The line in the plot is the linear regression line of best fit. Below the plot is the statistical output about this line of best fit.



Coefficients:

	Est.	p-value
Constant/Intercept	3052	0.243
PctDem	212	0.001

6. Below, write the equation for the line of best fit.

   Y = 3052 + 212*PctDem

7. Based on the line of best fit and to the nearest dollar, what is the predicted per pupil spending in 2016 for a state in which 50 percent of state residents were Democrat? For this, enter 50 percent into the equation as "50", not "0.50".

   Y = 3052 + 212*PctDem = 3052 + 212*50 = 3052 + 10600 = 13,652

Let's use the same output, for a linear regression that predicts per pupil spending in a state in the United States in 2016, using the percentage of state residents who were Democrats (PctDem).

Coefficients:

	Est.	p-value
Constant/Intercept	3052	0.243
PctDem	212	0.001

8. Does the output contain enough evidence to conclude at the conventional level in political science that the percentage Democrat in a state associated with per pupil spending in that state, at least on average?
   - ☐ Yes
   - ☐ No

9. Does the output contain enough evidence to conclude at the conventional level in political science that a higher percentage Democrat in a state caused a higher per pupil spending in that state, at least on average?
   - ☐ Yes
   - ☐ No

10. What does the 3052 intercept coefficient mean?
    - ☐ The average per pupil spending in a state was $3052.
    - ☐ If percentage Democrat in a state were zero, the predicted per pupil spending in that state would be $3052.
    - ☐ If per pupil spending in a state were zero, the predicted percentage Democrat in the state would be 3052.
    - ☐ Compared to a state with a certain percentage Democrat, a state with a one-unit higher percentage Democrat was predicted to have a per pupil spending that is $3052 higher.
    - ☐ Compared to a state with a certain per pupil spending, a state with a one-unit higher per pupil spending was predicted to have a percentage Democrat that is 3052 higher.

11. What does the 212 PctDem coefficient mean?
    - ☐ The average per pupil spending in a state was $212.
    - ☐ If percentage Democrat in a state were zero, the predicted per pupil spending in that state would be $212.
    - ☐ If per pupil spending in a state were zero, the predicted percentage Democrat in the state would be 212.
    - ☐ Compared to a state with a certain percentage Democrat, a state with a one-unit higher percentage Democrat was predicted to have a per pupil spending that is $212 higher.
    - ☐ Compared to a state with a certain per pupil spending, a state with a one-unit higher per pupil spending was predicted to have a percentage Democrat that is 212 higher.

12. Which of the following is the best interpretation of the 0.243 p-value for the intercept coefficient, if using the conventional p-value threshold in political science?
    - ☐ The output contains sufficient evidence to infer that the intercept coefficient was zero.
    - ☐ The output contains sufficient evidence to infer that the intercept coefficient was not zero.
    - ☐ The output does not contain sufficient evidence to infer that the intercept coefficient was zero.
    - ☐ The output does not contain sufficient evidence to infer that the intercept coefficient was not zero.

[Items 13 through 17] The linear regression output below is for the four points in the plot below, in which an outcome (OUTCOME) was predicted based on a respondent's highest level of education (EDUCATION), coded 1 for a high school degree, 2 for some college, 3 for an undergraduate degree, and 4 for a graduate degree.

	Coeff.	p-value
EDUCATION	11	0.168
Constant/Intercept	30	0.592

13. Based on the linear regression output above, write an equation to predict OUTCOME using a predictor for EDUCATION.

    OUTCOME = 30 + 11*EDUCATION

14. Use the equation to predict the level of OUTCOME for a respondent with a graduate degree.

    OUTCOME = 30 + 11*EDUCATION = 30 + 11*4 = 30 + 44 = 74

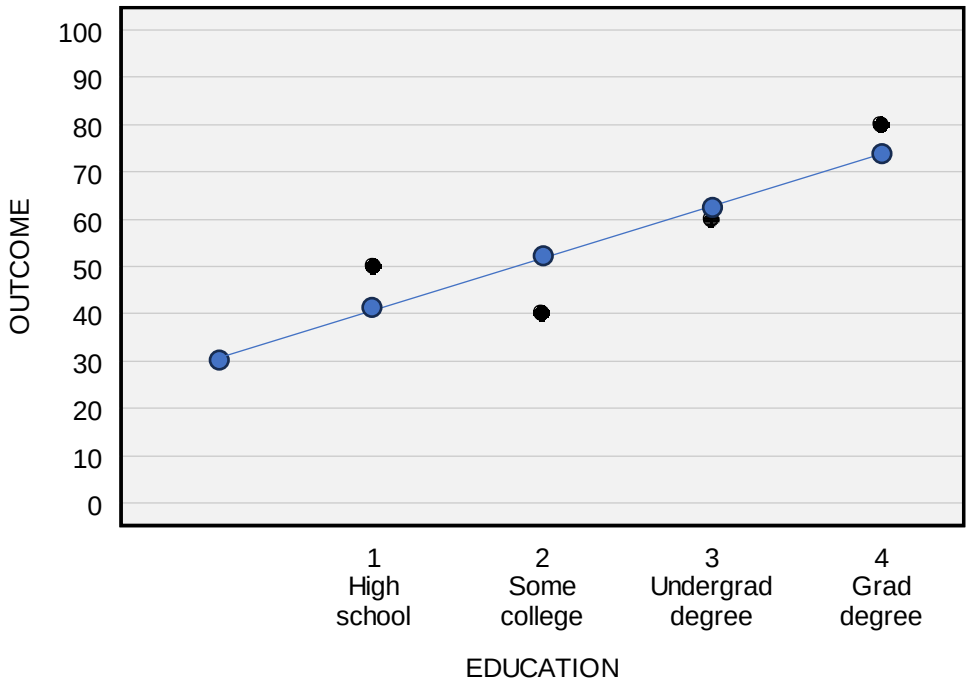
15. Interpret the coefficient for the constant/intercept.

    The 30 constant/intercept indicates the predicted level of the OUTCOME when all predictors are zero. In this case, the predicted OUTCOME is 30 when EDUCATION is zero. (The variable EDUCATION does not have a zero level, but a linear regression draws a straight line that goes on forever, so the linear regression line will hit points that do not exist or are not possible).

16. Interpret the coefficient and p-value for the EDUCATION predictor.

    The 11 coefficient for EDUCATION indicates that, compared to having a particular level of EDUCATION, having a one-unit higher level of EDUCATION is predicted to associate with a level of the outcome that is 11 units higher. (This is not necessarily a one *year* increase in education; instead, this is a one-unit increase in how the EDUCATION predictor has been coded). The p=0.168 p-value for EDUCATION indicates that the analysis did not provide sufficient evidence to conclude at the conventional level in political science that EDUCATION associated with OUTCOME any stronger than would be expected in random data.

17. Draw the line of best fit on the plot below, for the above regression output.



For the plot above, figure out where 0 would be on the X-axis. Then plot a point at X=1 and Y=30, to indicate the Y-intercept. Then we need to plot another point to be able to draw a straight line of best fit. For this, I'll pick a level of EDUCATION of 4, which we already calculated in Item 14. So we can plot a point at which X=4 and Y=74. Then we can draw a line between the two points.

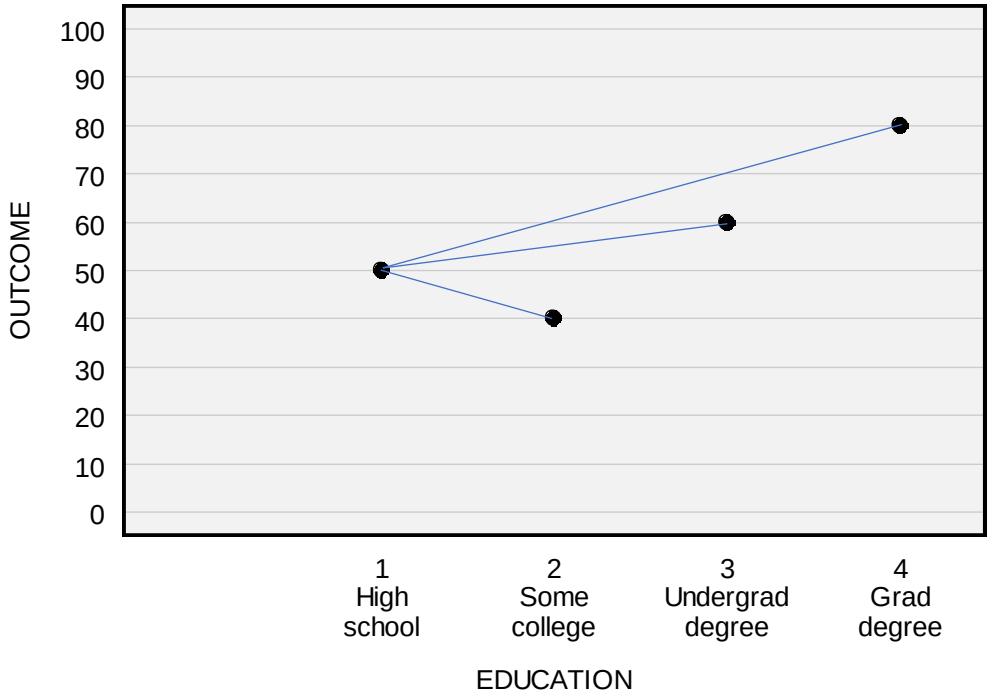
[Items 18 through 20] The linear regression output below is for the four points in the plot below:

	Coeff.
Constant/Intercept	50
EDUCATION: Some college	-10
EDUCATION: Undergrad degree	10
EDUCATION: Grad degree	30

18. Based on the linear regression output above, write an equation to predict OUTCOME using the included categories for EDUCATION. Use abbreviations of SC (Some college), UG (Undergrad degree), and GR (Grad degree).

OUTCOME = 50 + -10*(Some college) + 10*(Undergrad degree) + 30*(Grad degree)

19. Draw the lines of best fit on the plot below, for the above regression output.



For the plot above, the constant/intercept indicates the predicted OUTCOME when all predictors are zero. So, in this case, that is when "Some college" is zero, when "Undergrad degree" is zero, and when "Grad degree" is zero; so that's a person who has a high school education. So let's plot a point at X is "High school" and Y is 50.

For this regression, EDUCATION has been entered as a categorical predictor, so the plot will have multiple lines, each starting from the omitted reference category, which in this case is High school.

Let's use the equation in Item 18 to plot the remaining points. Let's start with Some college. Let's plug in 1 for each category that applies and 0 for each category that does not apply:

```
OUTCOME = 50 + -10*(Some college) + 10*(Undergrad degree) + 30*(Grad degree)
OUTCOME = 50 + -10*(1)            + 10*(0)                + 30*(0)
OUTCOME = 50 + -10                + 0                     + 0
OUTCOME = 40
```

So we can draw a line from (X=High school, Y=50) to (X=Some college, Y=40).
Let's use the equation in Item 18 to get a prediction for Undergrad degree:

```
OUTCOME = 50 + -10*(Some college) + 10*(Undergrad degree) + 30*(Grad degree)
OUTCOME = 50 + -10*(0)            + 10*(1)                + 30*(0)
OUTCOME = 50 + 0                  + 10                    + 0
OUTCOME = 60
```

So we can draw a line from (X=High school, Y=50) to (X=Undergrad degree, Y=60).
Let's use the equation in Item 18 to get a prediction for Grad degree:

```
OUTCOME = 50 + -10*(Some college) + 10*(Undergrad degree) + 30*(Grad degree)
OUTCOME = 50 + -10*(0)            + 10*(0)                + 30*(1)
OUTCOME = 80
```

So we can draw a line from (X=High school, Y=50) to (X=Grad degree, Y=80).

20. Suppose that the omitted category of EDUCATION was "Undergrad degree". Indicate what each coefficient would be in the output below:

If Undergrad degree is the omitted category, then the constant/intercept is the predicted outcome for Undergrade degree. Based on the prior item, this is 60, so let's enter that below.

The coefficient for High school is how much the predicted outcome for High school differs from the omitted category; from the prior item, the predicted outcome for High school is 50 and the predicted outcome for the omitted category of undergrad degree is 60, so the difference is -10.

The coefficient for Some college is how much the predicted outcome for Some college differs from the omitted category; from the prior item, the predicted outcome for Some college is 40 and the predicted outcome for the omitted category of undergrad degree is 60, so the difference is -20.

The coefficient for Grad degree is how much the predicted outcome for Grad degree differs from the omitted category; from the prior item, the predicted outcome for Grad degree is 80 and the predicted outcome for the omitted category of undergrad degree is 60, so the difference is +20.

	Coeff.
Constant/Intercept	60
EDUCATION: High school	-10
EDUCATION: Some college	-20
EDUCATION: Grad degree	20

[Items 21 through 25] Below is linear regression output using data from the ANES 2020 Time Series Study. The outcome is a respondent's feeling thermometer rating about rural Americans (FTRURAL), coded from 0 for very cold to 100 for very warm. The predictor is a categorical measure of the respondent's political party identification (PARTY), coded into three categories, with Democrat as the omitted category.

```
. reg FTRURAL i.PARTY
                                        Number of obs   =     6,227
------------------------------------------------------------------------------
     FTRURAL |      Coef.    p-value     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       PARTY |
Independent  |          7     <0.001            6           9
 Republican  |         20     <0.001           19          21
             |
   _constant |         62     <0.001           61          62
------------------------------------------------------------------------------
```

21. Write an equation to predict the outcome. If needed, use the abbreviations D for Democrat, I for Independent, and R for Republican.

    FTRURAL = 62
    + 7*Independent
    + 20*Republican

22. What does the 62 coefficient for the constant/intercept indicate?
    - ☐ Across the sample, the mean rating about rural Americans was 62.
    - ☐ Across Democrats, the mean rating about rural Americans was 62.
    - ☐ Across non-Democrats, the mean rating about rural Americans was 62.

23. The 95% confidence interval for Independent is [6, 9], which is 3 units wide. The 99% confidence interval for Independent will be ___
    - ☐ less than 3 units wide
    - ☐ more than 3 units wide

24. What does the 20 coefficient for Republican indicate?
    - ☐ Across Republicans, the mean rating about rural Americans was 20.
    - ☐ The mean rating about rural Americans was 20 units higher among Republicans than among Democrats.
    - ☐ The mean rating about rural Americans was 20 units higher among Republicans than among Independents.
    - ☐ The mean rating about rural Americans was 20 units higher among Republicans than among non-Republicans.

25. Suppose that the linear regression above had omitted the category for Independent and had included the categories for Democrat and Republican. In that regression, which of the following would be the coefficient for Republican?
    - ☐ -20
    - ☐ -7
    - ☐ 13
    - ☐ 20
    - ☐ 27
    - ☐ None of the above