# Problem Set 8

**Due date: 20 November**

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**

## Question 1

*Points: 5*

For the following regression equation, $\hat{Y} = 8.5 + 6x + \epsilon$, the standard error for $\beta_0$ is 2.5, the standard error for $\beta_1$ is 3.5, and the sample size is 2000. Find the t-statistic, 95% confidence interval, and p-value (using a two-tailed test) for $\beta_1$.

Is $\beta_1$ statistically significant at the 0.05-level with a two-tailed test? Why or why not?

**Ans 1:**

1. We know that: The t-statistic is calculated using the formula:

$$t = (estimateoftheparameter - hypotheziedvalue)/Stanarderroroftheestimate$$

2. For Beta1, the hypothesized value is usually 0 (to test if Beta1 is significantly different from zero), and its estimate right here is 6.
3. We know that the confidence interval can be given as:
    1. $$CI = estimate + / - t_c * StandardError$$

Where the degrees of freedom is 2000- 2 = 1998 and we estimate it for a 95% confidence interval

4. **Determine Statistical Significance:** If the p-value is less than the significance level (0.05 in this case), then Beta1 is considered statistically significant.

The calculated values are as follows:

1. **t-statistic for** Beta1: The t-statistic is approximately 1.71.
2. **95% Confidence Interval for** beta1: The confidence interval ranges from approximately -0.86 to 12.86.
3. **p-value for the two-tailed test:** The p-value is approximately 0.087.

To determine if Beta1 is statistically significant at the 0.05 level for a two-tailed test, we compare the p-value with the significance level (0.05). Since the p-value (0.087) is greater than 0.05, we fail to reject the null hypothesis. Therefore, Beta1 is not statistically significant at the 0.05 level.

This means that, based on our sample data, we do not have enough evidence to conclude that Beta1 is different from zero in the population.

## Question 2

*Points: 5*

Suppose you estimate an OLS regression and retrieve a $R^2$ value of 0.45. If the Total Sum of Squares (TSS) from that regression equals 4,700, what is the value for the Residual Sum of Squares (RSS)?

**Ans 2:**

The Residual Sum of Squares (RSS) can be calculated using the RSquare value from an Ordinary Least Squares (OLS) regression and the Total Sum of Squares (TSS). The R Square value represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It is calculated as:

1. $$R^2 = 1 - RSS/TSS$$

Now, let's calculate the RSS.

The value of the Residual Sum of Squares (RSS) for this regression is 2,585.

## Question 3

*Points: 5*

Suppose you estimate a bivariate regression with a sample size of 102 and obtain a regression coefficient ($\beta_1$) of 5.0. What is the largest standard error that $\beta_1$ could have and still be statistically significant (i.e., reject the null hypothesis of no relationship) at the 0.05 level with a one-tailed test?

**Ans 3:**

To determine the largest standard error that Beta 1 could have and still be statistically significant at the 0.05 level with a one-tailed test, we need to consider the critical value from the t-distribution for our given sample size and significance level, and then use the formula for the t-statistic:

1.
$$t = \beta_1/SE(\beta_1)$$

Where:

- 1 1 is the regression coefficient, in this case, 5.0.
- SE of the standard error of 1 1.
- t is the t-statistic.

For a one-tailed test at the 0.05 significance level and with a sample size of 102 (which gives us 101 degrees of freedom since df = n - 1 for bivariate regression), we will find the critical t-value. The largest standard error SE that still results in a statistically significant result is when the calculated t-statistic equals this critical t-value.

The critical t-value for a one-tailed test at the 0.05 significance level with 101 degrees of freedom is approximately 1.66. Given this, the largest standard error that the regression coefficient could have and still be statistically significant at this level is approximately 3.01.

This means if the standard error of beta 1 is less than or equal to 3.01, the result will be significant at the 0.05 level for a one-tailed test.

## Question 4

*Points: 5*

Using the `states` dataset from the `poliscidata` package, produce a scatterplot of the variables `romney2012` and `hispanic10` (with `romney2012` as the dependent variable on the y-axis). Fit a regression line to the scatterplot. Describe the scatterplot and include a copy of it. Note any suspected outliers, if any (a visual inspection will suffice for this question).

> Note
>
> The variable `romney2012` measures the percentage of the state's vote that Mitt Romney received in the 2012 presidential election, and `hispanic10` indicates the percentage of the state's population that identified as Hispanic in 2010.

**Ans 4:**

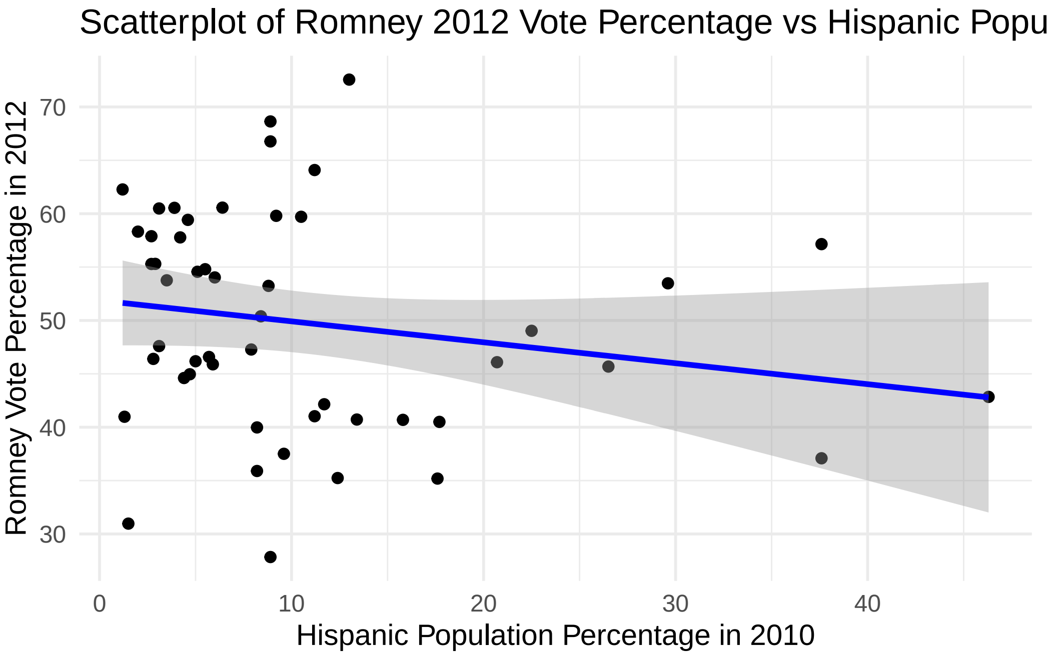
```r
#Loading the data:
data <- poliscidata::states
```

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

```r
library(ggplot2)

# Creating a scatterplot with a regression line
ggplot(data, aes(x = hispanic10, y = romney2012)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "Scatterplot of Romney 2012 Vote Percentage vs Hispanic Population Percenta
       x = "Hispanic Population Percentage in 2010",
       y = "Romney Vote Percentage in 2012") +
  theme_minimal()
```

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



As shown above we have the regression equation along with it's confidence interval.

Based on the regression equation we can see that there seems to be a fairly strong negative correlation between Romney's vote percentage and the Hispanic population.

Seems based on the size of the confidence interval and the number of points not in the attached line , we see that the majority of points can be considered outliers. There is a cluster of points near the 0% level, which indicates the strongest portion of the data points.

### Question 5

*Points: 10*

Estimate a bivariate regression with `romney2012` as the dependent variable and `hispanic10` as the independent variable and report the results in a professionally formatted table. In as much detail as possible, describe (and interpret) the regression results.

**Ans 5:**

```
library(broom)
library(dplyr)
```

```
Attaching package: 'dplyr'

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

    filter, lag

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

    intersect, setdiff, setequal, union
```

```
library(knitr)

states_df <- data.frame(data$romney2012, data$hispanic10)
names(states_df) <- c("romney2012", "hispanic10")

# Performing the regression analysis
reg_model <- lm(romney2012 ~ hispanic10, data = states_df)

# Tidying the regression results
tidy_results <- tidy(reg_model)

# Displaying the regression results in a formatted table
kable(tidy_results, caption = "Regression Results: Romney Vote Percentage vs Hispanic Popu
```

Table 1: Regression Results: Romney Vote Percentage vs Hispanic Population Percentage

| term | estimate | std.error | statistic | p.value |
|:---|---:|---:|---:|---:|
| (Intercept) | 51.8766881 | 2.0996679 | 24.70709 | 0.0000000 |
| hispanic10 | -0.1961252 | 0.1448262 | -1.35421 | 0.1820106 |

**Interpretation of the Regression Results:**

1. **Intercept**:

    - This is the expected value of `romney2012` when `hispanic10` is zero. In other words, it's the estimated percentage of votes for Romney in 2012 in a state with no Hispanic population, according to this model. In this. case we can see that the intercept value is 51.87 and the p_valvue is significant at this level.

2. **Coefficient of `hispanic10`** :

- This coefficient represents the change in the `romney2012` vote percentage for each one percentage point increase in the `hispanic10` population. A negative value would indicate that as the percentage of the Hispanic population increases, Romney's vote share tends to decrease, and vice versa. In this case we have a negative coefficient and hence there is a negative relationship between the variables. However the associated p_value is not very significant at a 95% interval and hence we cannot say that the relationship holds at a significant degree.

3. **Standard Error (SE):**

   - The standard errors for Beta0 and Beta1 provide a measure of the variability or precision of the respective estimates. Smaller values indicate more precise estimates. Standard error in this case is alright but it needs to be looked at in relationship with other values in a linear regression.

4. **t values and P-values:**

   - The t values test the null hypothesis that each coefficient is equal to zero (no effect). A large absolute t value and a small p-value (typically <0.05) suggest that the effect of that variable is statistically significant..

   - P_values should generally not be used very heavily , any deviation of the assumptions of linear regression assumptions will result in a lower P_value. For example if there is any non-linearity between the variables even for a few data pints , you would see a lower in P-value.