APSTA-GE: 2003 Intermediate Quantitative Methods

Fall 2020

Class Exercise Questions

Created on: 10/06/2020

Instructions:

To submit, please make a copy of this file, work on solutions, and **paste the sharing link** to the F column of the group assigning list.

Load `lung_capacity0.csv`, and **add a new column** labelling *the row ID*.

1. How does lung capacity change as people get older?

   (in one sentence)

   a. Make a scatter plot of lung capacity (on y-axis) and age (on x-axis).

      Eyeball the pattern. What do you see?

      (in one or two sentences)

   b. Report the sample size.
2. Conduct a regression analysis to answer the following questions:

   - Which variable is the dependent variable (D.V.)?

   - Which variable is the independent variable (I.V.)?
3. Answer below questions based on the regression result:

   a. What's the regression coefficient of `age` on `lung capacity`?

      Report the appropriate regression coefficient and its standard error (S.E.).

   b. Is this coefficient statistically significant? What test did you use?

      Write down the null and alternative hypotheses.

      Report test statistic and p-value.

   c. Report a 95% confidence interval for the regression coefficient of `age` on `lung capacity` based on the results from the model.
4. Based on the model, calculate the fitted values and residuals for all observations in the model. Answer the following questions based on your results:

a. What are the mean values of the fitted values and residuals?
b. Calculate or report the total sum of squares (TSS), model sum of squares (MSS) and residual sum of squares (RSS).
   Verify that: TSS = MSS + RSS
c. What's the R-square of this model?
d. What's the correlation between `lung capacity` and age? Verify that the R-square is the correlation squared.

5. Based on this model, how much difference do we expect to see between 20 years old and 25 year old? And between 60 years old and 65 years old?

   Hint: By the nature of linear relationships, it doesn't matter at what age ranges.
   This Model assumes that 1 year age difference leads to the same change in `lung capacity`.

   Another interesting question:

   Can you prescribe a 95% confidence interval for the average difference in `lung capacity` associated with a 5-year age difference?

   Hint: `beta1` (β_1) is the difference due to 1-year difference in age.

   `5 * beta1` ($5 \times \beta_1$) is the difference due to a 5-years difference in age.

   Now the standard error in `beta 1 hat` ($\hat{\beta_1}$) is ___?

   What's the standard error in $5 \times \beta_1$?

   It is $5 \times S.E.$.

   $Var(5 \times \beta_1) = 25 \times Var(\beta_1) = 25 \times S.E.(\beta_1)^2$

   Confidence Interval (C.I.) $= (\beta_1 - 2 \times S.E.,\ \beta_1 + 2 \times S.E.)$

   Now $5 \times \beta_1$, $C.I. = 5 \times C.I.(\beta_1)$

   $Var(\beta) = [S.E.(\beta)]^2$

   $Var(5 \times \beta) = 25 \times Var(\beta)$

   $S.E.(5 \times \beta) = \sqrt{}(Var(5 \times \beta)$

6. Estimate standardized regression coefficients using two different approaches:

a. Generate a new variable based on `lung capacity` by dividing its original value by its standard deviation.

   Generate a new variable based on `age` by dividing its original value by its standard deviation.

   Run a regression using these two new variables. Report the regression coefficients (intercept and slope)

b. Use `lm.beta` to estimate the standardized regression coefficients.

c. Compare standardized slope with correlation.

d. What is the meaning of the intercept now?

7. Conduct regression diagnostics.

   R: stat package (more traditional)

   influence.measures function

   R: car (fancier plots)

   Regression Diagnostics

   a. Plot a studentized residual plot against $\hat{Y}$ (fitted values).

      - How is figure the same or different from:
        - Residual versus $\hat{Y}$?
        - Residual (or studentized residual) versus `age`?
        - Difference between standardized and studentized residuals?
      - Any outliers? Use *-2* and *2* as fences.
      - Based on eyeballing, does linear relationship hold?
        - Test linearity by including a quadratic term.
      - Based on eyeballing, does equal variance hold?
        - Test for equal variance.
      - Any observations have leverage, and high influence?
        - Eyeball by overlaying the fitted line - outliers along x and y axes?
        - Use a fancy plot.
        - Use Cook's D, Deffits.
      - Remove outlier data points, what happens? Things to check:

    - p-value of β_1, difference in values of new and old β_1 estimates.
    - SE values of new and old β_1 estimates.
    - R squares of the new and old models.
- Remove influential data points, what happens?
- For curious minds, if we remove data points that have high leverage but not an outlier or influential, what will happen?

  Can you fabricate some points to test the ideas?

  Hint: you can add a new observation with a pair of `lung capacity` and `age` values of your choice and attach this new row to the existing data.
- Lastly, think about when should we remove point(s) and when should we not remove point(s) from regression? What are the pros and cons?