Statistical Thinking for Forensic Practitioners

Excel Lab on Part 7: Regression & Analysis of Variance

We will practice fitting regression models in Excel. Open the `strideLength.xlsx` file found on the course website. This file contains stride and leg length measurements taken from 50 males and 50 females. The variables collected from each cadaver were:

- `sex`: sex of the individual
- `legLength`: leg length of individual (in cm.)
- `strideLength:` stride length of the individual (in cm.)

Follow the steps in this article to include the Analysis ToolPak add-in in Excel. This will be needed to complete the exercises below.

1 Exploratory Data Analysis

It is good practice to visualize a data set prior to choosing models. We will be interested in predicting someone's stride length based on their leg length. As we have data on both males and females, we will want to indicate this information in our visualizations.

1. To color the points in a scatterplot based on sex, first title two new columns `male` and `female` to the right of the `strideLength` column.
2. In the first row of the `male` column, input the following formula: `=IF(A2="male",C2,NA())`. This will populate the cell with the value in cell `C2` (the stride length of the first individual) if they are male (which they are). Otherwise, it will populate the cell with the "null" value `NA`. Populate the rest of the cells in the `male` column in a similar manner (remember the trick of double-clicking the bottom right corner of a cell to vertically repeat a formula). Then populate the `female` cell in a similar manner after replacing `"male"` with `"female"` in the call to the `IF` function.
3. Highlight the `legLength` to `female` columns and create a scatterplot. Excel will try to use the `strideLength` data in this plot. To keep Excel from plotting this redundant information, hover over the plot and click on the filter button that appears on the right. Uncheck the `stridLength` box to remove this column's data from the plot. (Note: if these instructions don't pertain to your version of Excel, this guide has alternative methods.).
4. Based on the scatterplot, why might we want to fit a regression model with sex dummy variables?

   There's evidence that the linear relationship between stride length and leg length is different between males and females for these data. Using a sex dummy variable would give a regression model the flexibility to have a sex-dependent intercept or slope (or both).

2 Simple Linear Regression Model

Let's first naively fit a simple linear regression model without any additional structure. Let y_i be the stride length and x_i the leg length of the ith individual in the dataset, $i = 1, ..., 100$. This guide has a detailed discussion of fitting regression models using the Analysis ToolPak should you become lost at anytime during this exercise.

1. Define a simple linear regression model for the relationship between stride length and leg length for the ith individual using similar notation as the lecture notes (e.g., slide 24). Remember to specify assumptions made on the error term (like those discussed on slide 31).

   $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ **where** $\epsilon_i \sim N(0, \sigma_e^2)$ **(and the** ϵ_i**s are independent).**

2. Using the formulas for b_0, b_1 on slides 28 and 29, calculate the estimated regression coefficients for the simple linear regression model. (Hint: calculate b_1 first using the slide 29 formula and the `CORREL` and `STDEV.S` functions. Then calculate b_0 using the formula on slide 28 and the `AVERAGE` function.)

   The estimates are $b_0 = -25.206$ **and** $b_1 = 1.040$ **rounding to 3 decimal places.**

3. Ensure that the b_0, b_1 values you calculated above are correct by using the `INTERCEPT` and `SLOPE` functions to calculate b_0, b_1. (Note: these functions only work for calculating a SLR intercept/slope, so we won't be able to use them for more complicated models.)

   Again, $b_1 = 1.040$ **rounding to 3 decimal places.**

4. Now fit the SLR model using the Analysis TookPak. To do so...

   (a) Under the `Data` tab at the top of the spreadsheet, click `Data Analysis` on the right side.

   (b) Select `Regression` and click `OK`. Click the arrow to the right of `Input Y Range:` and highlight the data in the `strideLength` column (including the column title).

   (c) Do the same for `Input X Range:` highlighting the `legLength` column (a trick to quickly highlight multiple rows is to click on a row, hold down `Shift` and `Ctrl`, and press up/down on the arrow keys).

   (d) Make sure the boxes for `Labels` (which tells Excel that the first cells highlighted are the column titles), `Confidence Level` (set to 95%), `Residuals`, and `Line Fit Plots` are checked. Also make sure that `New Worksheet Ply` is selected (so that Excel prints the regression output into a new sheet).

   (e) Click `OK`.

5. Write the estimated response $\hat{y}_i$ given the values of the estimated coefficients (a la slide 51). Ensure that the estimated coefficients given by Excel match the ones calculated above.

   $\hat{y}_i = -25.206 + 1.039 x_i$

6. Suppose we obtained the leg length of a new individual where $x_{101} = 75$ cm. Based on the SLR model, what is this individual's estimated height?

   $\hat{y}_{101} = 52.78$ **cm rounding to 3 decimal places.**

7. Consider the `RESIDUAL OUTPUT` given by Excel. The `Residuals` column contains the values of $e_i = y_i - \hat{y}_i$ for each $i = 1, ..., 100$. Use the values in this column and the formulas given on slide 32 to calculate the Mean Squared Error (MSE).

   MSE = 40.802 rounding to 3 decimal places.

8. Using the MSE calculated above and the formula given on slide 39, calculate the estimated variance for b_1. Use this value to check that the standard error reported by Excel for b_1 is correct. (Note: you will need to click back to the `strideLength` tab to calculate the denominator of the estimated variance of b_1. You may also need to create a new column to calculate $(x_i - \bar{x})^2$ for each x_i.)

   The estimated variance is $S^2_{b_1} = .011$ **meaning SE**$(b_1) = .103$ **which agrees with the Excel output, rounding to 3 decimal places.**

9. Using the standard error for b_1 calculated above, verify the t-statistic value given by Excel assuming the hypothesis test performed is $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.

   The t-statistic is $t = \frac{b_1 - 0}{SE(b_1)} = \frac{1.040}{.103} = 10.1$ rounding to 1 decimal place, which approximately agrees with the t statistic computed by Excel.

10. Using the t statistic calculated above and the `T.DIST.2T` function, verify the p-value computed by Excel. (Note: the degrees of freedom for a regression model are $(n-1)-(p-1) = n-p$ where p is the number of regression parameters. For SLR, we assume β_0, β_1 are the regression parameters.)

    Depending on the decimal precision used above, the p-value will be approximately $p = 7.2 \times 10^{-17}$ (basically 0). The degrees of freedom for a SLR hypothesis test are $n - 2 = 100 - 2 = 98$.

11. Using the value and standard error of b_1 and the `T.INV` function, very the confidence interval computed by Excel for b_1. (Note: recall from Part 6 that for a $(1-\alpha)\%$ confidence interval, we use $\frac{\alpha}{2}$ to calculate the t critical value.)

    The confidence interval is $b_1 \pm t^*_{98,.975} SE(b_1) = 1.040 \pm .103 = [.836, 1.244]$ rounding to 3 decimal places.

12. Using the values in the `RESIDUAL OUTPUT`, create a scatterplot with `Predicted strideLength` on the horizontal axis and `Residuals` on the vertical axis. Based on this residual plot, what assumption(s) made about the error terms appear to be violated? (Note: use language as on slide 46 to describe the violations.)

    There does appear to be some sort of pattern to the residuals, but it's difficult to identify. It could roughly be characterized as a downward trend (for the record, *any* identifiable pattern in the residuals is bad, so being specific about the pattern is not always necessary). Additionally, the variance may not be constant due to the somewhat conical shape.

3 SLR with Sex Dummy Variables and Interactions

Let d_i represent a dummy variable that equals 1 if subject i is female and 0 if they are male (note: this is reversed from the example given in the lecture notes). Let z_i represent the interaction between d_i and x_i (i.e., $z_i = x_i$ for females and $z_i = 0$ for males similar to slide 64.)

13. Define a model that extends the previous SLR model with the inclusion of d_i and z_i.

    $y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 z_i + \epsilon_i$ **where $\epsilon_i \sim N(0, \sigma^2)$ independently.**

14. Write the two regression models for y_i if the ith individual is male vs. if they are female (similar to slides 65 and 66). How is this model more "flexible" than the SLR model?

    For males, we assume $y_i|\text{male} = \beta_0 + \beta_1 x_i + \beta_2(0) + \beta_3(0) + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$. For females, we assume $y_i|\text{female} = \beta_1 + \beta_1 x_i + \beta_2(1) + \beta_3(x_i) + \epsilon_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_i + \epsilon_i$. This model is more flexible in that it allows the fitted model to have a different intercept and slope for the male and female groups.

15. In Excel we need to create new columns to represent these additional variables. You may want to copy the contents of the `legLength` column into a new column to the right of the columns you created previously to make the next few steps easier.

    (a) Create a `sexIndic` column using the `IF` function where a cell contains a 1 if `sex="female"` and 0 otherwise (note: this will be similar to the creation of the `male` and `female` columns above). This will represent the d_i variables.

    (b) Create a column for z_i by multiplying the contents of the `sexIndic` and `legLength` columns together. You may call this column simply `sexIndic*legLength`.

    (c) Click again on the `Data Analysis` button under the `Data` tab and select `Regression`. You should not have to change the `Input Y Range:` selection from before, but you will now want to select the data in the `legLength`, `secIndic`, and `sexIndic*legLength` columns for `Input X Range:`.

(d) Make sure the boxes for `Labels` (which tells Excel that the first cells highlighted are the column titles), `Confidence Level` (set to 95%), `Residuals`, and `Line Fit Plots` are checked. Click `OK`.

16. Write the estimated response $\hat{y}_i$ given the values of the estimated coefficients (a la slide 51).

    $\hat{y}_i = -82.234 + 1.734x_i + 37.266d_i - .400z_i$ **rounding to 3 decimal places.**

17. Suppose we obtain the leg length of a new female where $x_{101} = 71$ cm. Based on the fitted model, what is this individual's estimated height?

    $\hat{y}_{101} = -82.234 + 1.734(71) + 37.266(1) - .400(71) = 49.7$ **cm. approximately.**

18. Create a residual plot based on the `RESIDUAL OUTPUT` similar to the one created above. Do you notice any improvement over the previous residual plot? What might be done to improve upon this model?

    Not really an improvement. This plot again seems to exhibit the same violation(s) pointed out for the previous plot. Perhaps we need to use a higher-order model (e.g., quadratic regression) to fit the trend more effectively.

4 Quadratic Regression with Sex Dummy Variables and Interactions

Quadratic regression models can be used to explain a curved relationship between a response and predictor. In this example, we may want to model a curved relationship between stride length, y_i, and leg length, x_i. We will be combining the indicator/interaction variables from the previous model with the quadratic regression models discussed on slides 71-78. We will let w_i represent the interaction term between d_i, the sex dummy/indicator variable, and x_i^2, the quadratic leg length term (so $w_i = x_i^2$ if the ith individual is female and 0 otherwise).

19. Define a model that extends the SLR with d_i and z_i model used previously by including a quadratic leg length term, x_i^2, and a quadratic interaction w_i defined above. (Note: your model should have regression coefficients β_0 through β_5.)

    $y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 z_i + \beta_4 x_i^2 + \beta_5 w_i + \epsilon_i$ **where** $\epsilon_i \sim N(0, \sigma^2)$ **independently.**

20. Similar to SLR + dummy and interaction model, we will need to create 2 new columns representing our new variables.

    (a) Create a `legLengthSquared` column by squaring the contents in the `legLength` column.

    (b) Create a `sexIndic*legLengthSquared` column by multiplying the contents of the `sexInidic` and `legLengthSquared` columns.

    (c) Click again on the `Data Analysis` button under the `Data` tab and select `Regression`. You should not have to change the `Input Y Range:` selection from before, but you will now want to select the data in the `legLength`, `secIndic`, `sexIndic*legLength`, `legLengthSquared`, and `sexIndic*legLengthSquared` columns for `Input X Range:`.

    (d) Make sure the boxes for `Labels` (which tells Excel that the first cells highlighted are the column titles), `Confidence Level` (set to 95%), `Residuals`, and `Line Fit Plots` are checked. Click `OK`.

21. Write the estimated response $\hat{y}_i$ given the values of the estimated coefficients.

    $\hat{y}_i = 320.802 - 8.23x_i - 2278.109d_i + 64.27z_i + .061x_i^2 - .452w_i$**.**

22. Sift through the fitted line plots printed by Excel until you find the one titled `legLengthSquared Line Fit Plot`. It may be helpful to make this plot larger to see the fitted values. Based on this visualization, it still doesn't appear as that the female quadratic fit follows the trend we see in the female data (the female predicted values are just a tad higher than what we might feel they should be). Based on the scatterplot, explain why this might be expected. (Hint: think about a property of the female data that might "throw off" a least squares estimator.)

    The female data seem to have some "extreme" stride length observations causing the estimated quadratic curve to be "pulled" upwards.

23. Create a residual plot based on the `RESIDUAL OUTPUT` similar to the ones created above. Of the two assumptions we can check with this residual plot (that the error terms have (1) mean 0 and (2) constant variance), one of these assumptions appears to be more effectively met by this quadratic regression model than by the linear models considered previously. However, it is clear that the other assumption is violated. Identify which assumption is met and which assumption is violated. How could we *further* extend this model to fix this violation? (Hint: think of an assumption we could make on the error terms. Your answer to question 22 may be useful to consider.)

The residuals appear to be more balanced above and below the horizontal 0 line for this model, indicating that the 0 expectation assumption is more effectively met than it was for the 2 previous models. However, it is clearer from this residual plot that the constant variance assumption is violated (the conical shape is more pronounced). In question 22, it was noted that the female data appear to have more "extreme" values than the male data set, indicating that the variance may be different for the two sexes. We could assume different variances for the two sexes in our model. Mathematically, this could be expressed by assuming for the first 50 observations (the males in the dataset) that $\epsilon_i \sim N(0, \sigma_m^2)$, $i = 1, ..., 50$, and for the latter 50 observations (the females) that $\epsilon_i \sim N(0, \sigma_f^2)$, $i = 51, ..., 100$. At this point, with so many qualifications made to the model on the basis of sex, we could consider this as fitting two separate models: one for the males and one for the females. Such is the way that model exploration often proceeds.