Connor Johnson

Assignment 3

The first analysis will be a simple t-test on the means of the change in batting average that resulted from marriage, 1st child, and the 2nd child. This data was found by finding the difference in the averages of players in the MLB between the season before and after each of these three events occurred. The null hypothesis of the test will be that the mean is equal to zero. Since it was predicted earlier that the batting averages will decrease, the alternate hypothesis is the mean is less than zero.

Age will also be checked as a confounding variable on the response. This will be tested using a linear regression shown below.

Y = β0 + β1\*Age

* β0 – constant for the regression
* β1 – change in the response variable as age increases by one

The null hypothesis for this test is β1 = 0 and the alternate hypothesis is β1 ≠ 0. If the null hypothesis is false, then age will need to be considered as a confounding variable above in showing if the difference is significant. If age isn’t a significant predictor, then it won’t need to be included in the above model.

**ANOVA Regression**

The second analysis to be done is an ANOVA regression using the batting averages of all players from the sample in 2019. This will give more info on the long-term effects of the independent variables being studied. The analysis above only used data from the season before and after the players developed a new relationship. Because of this, the results only show what immediate effects their decisions may have. In this analysis, the players in the sample will be split into the following groups: Not married, married, no kids, one kid, two or more kids. Since these groups have multiple overlaps, they will be analyzed in pairs. The pairs of groups that will be tested are Married/Not Married, One Kid/No Kids, Two or More Kids/No Kids, Two or More Kids. The regression equation used for this is below.

Y = β0 + β1\*X1 + β2\*Age

* X1 – dummy variable that represents which category the player is in
* β0 – constant for the regression
* β1 – difference in the mean batting average between the two groups
* β2 – change in the response variable as the player’s age is increased by one

Since the goal of this analysis is to look for a difference between the groups being looked at, β1 is the variable that will be tested. The null hypothesis is β1 = 0 and the alternate hypothesis is β1 ≠ 0. A p-value that is less than 0.05 will show a statistically significant difference. Age acts as a control variable to avoid it confounding with the data. β2 shouldn’t be a very significant coefficient.

**­Results**

The goal of this study is to find if personal relationships in an athlete’s life have a negative effect on their performance, specifically in baseball. This is analyzed by testing if there is a significant change in a baseball player’s performance after he gets married and has children. The sample used for the analysis was taken from players that currently are in the starting lineup for a team in the MLB. Since the starting lineups can change often, the lineups were taken from lineups.com which shows the most recent starters for each team. Out of the 30 teams in the MLB, 103 players were found to be married or have children. Many of these players were removed from the sample because they didn’t get married while in the league or didn’t have enough at bats for their stats to be representative of their performance. The final sample has 54 players in the marriage sample, 48 players in the first child sample, and 32 players in the second child sample.

The first model is the regression that tests if age is a significant predictor in the change in batting average. All three regression models are shown below.

**Coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** | **VIF** |
| Constant | -0.0475 | 0.0612 | -0.78 | 0.441 |  |
| Age (Marriage) | 0.00167 | 0.00236 | 0.71 | 0.482 | 1.00 |
|  |  |  |  |  |  |

**Coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** | **VIF** |
| Constant | 0.0065 | 0.0400 | 0.16 | 0.871 |  |
| Age (1st Child) | -0.00061 | 0.00149 | -0.41 | 0.682 | 1.00 |
|  |  |  |  |  |  |

**Coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** | **VIF** |
| Constant | -0.0481 | 0.0488 | -0.99 | 0.332 |  |
| Age (2nd Child) | 0.00159 | 0.00169 | 0.94 | 0.356 | 1.00 |
|  |  |  |  |  |  |

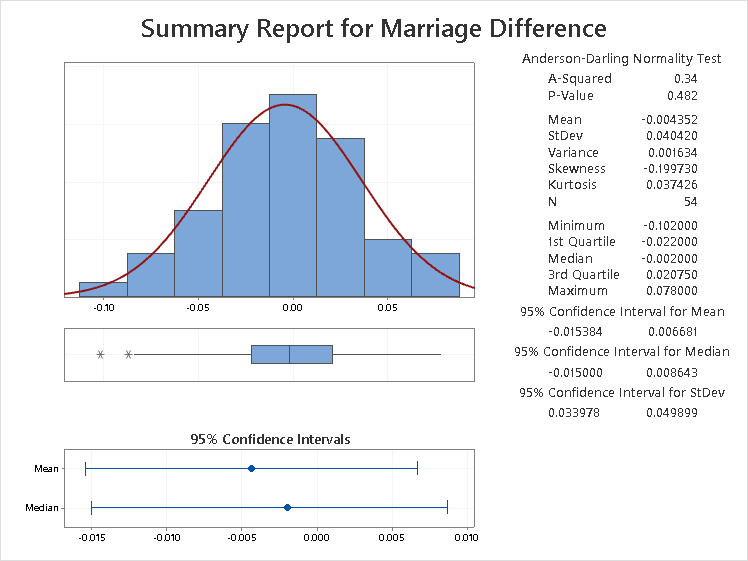
**Fig 1.** Linear Regressions of each response variable versus Age of the player.

The p-value for all three variables is much greater than 0.2. Because of this, we can assume that that Age does not have a significant effect on the response variable and does not need to be considered in the model.

Now, the difference in batting average for all three tests can be analyzed for its significance. The figures below show the distribution and descriptive statistics of the three.

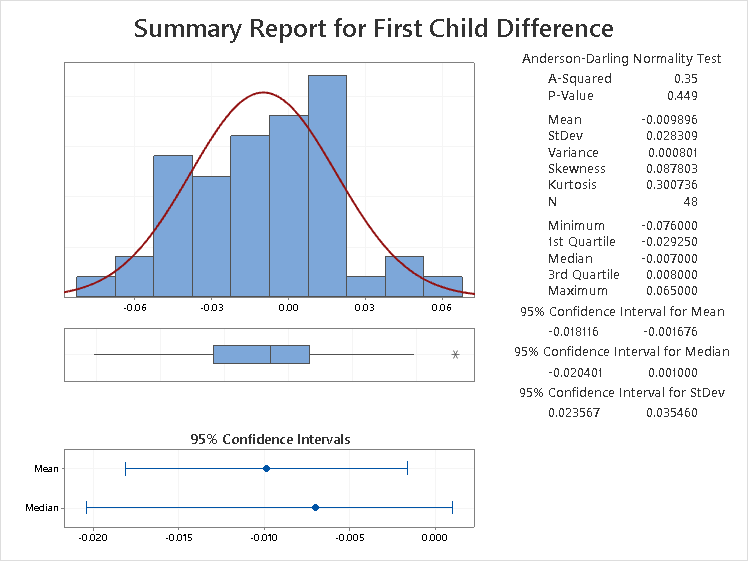
**Table 1**

*Histogram and descriptive statistics of the Marriage Difference response variable*



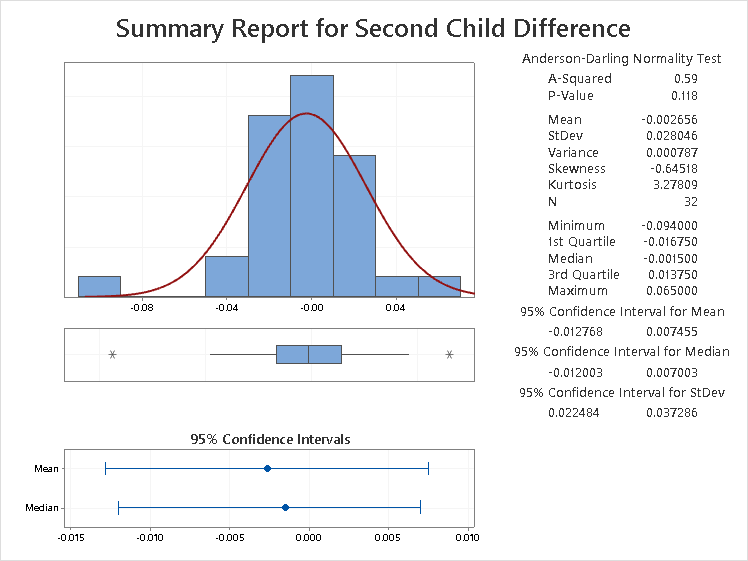
**Table 2**

*Histogram and descriptive statistics of the First Child Difference response variable*



**Table 3**

*Histogram and descriptive statistics of the First Child Difference response variable*



From figures 2, 3, and 4, it can be seen that the distributions of all three variables are relatively normal. So, the t-test statistic can be used to show if they are significantly less than zero. These tests are shown below.

**Figure 5**

*Test Statistics and p-values for all response variables.*

**Marriage Dif. 1st Child Dif. 2nd Child Dif.**

|  |  |  |  |
| --- | --- | --- | --- |
| Null hypothesis | | | H₀: μ = 0 |
| Alternative hypothesis | | | H₁: μ < 0 |
| **T-Value** | **P-Value** |
| -0.54 | 0.298 |

|  |  |  |  |
| --- | --- | --- | --- |
| Null hypothesis | | | H₀: μ = 0 |
| Alternative hypothesis | | | H₁: μ < 0 |
| **T-Value** | **P-Value** |
| -0.79 | 0.216 |

|  |  |  |  |
| --- | --- | --- | --- |
| Null hypothesis | | | H₀: μ = 0 |
| Alternative hypothesis | | | H₁: μ < 0 |
| **T-Value** | **P-Value** |
| -2.42 | 0.010 |

Since we’re looking for 95% confidence in the test, the p-value must be less than 0.05 for the mean of the response variable to be significantly less than zero. As seen in Figure 5, the First Child Difference is the only variable to have a p-value less than 0.05 with p = 0.010. Thus, the null hypothesis is rejected and the mean is presumed to be less than zero. This tells us that the first child of a baseball player has an immediate effect on their performance. This effect is small as shown by the data, but still statistically significant. For the marriage response variable, the T-value is -0.79 with p = 0.216. For the second child variable, the T-value is -0.54 and p = 0.298. Because the values are both much greater than 0.05, the null hypothesis can’t be rejected meaning there isn’t enough evidence to show the difference is less than zero.

**Figure 2**

*Regression and Analysis of the Married/Not Married groups of players*

**Regression Equation**

|  |  |  |
| --- | --- | --- |
| Average | = | 0.2762- 0.00077 Married - 0.00056 Age |

**Analysis of Variance**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Age | 1 | 0.000345 | 0.000345 | 0.30 | 0.586 |
| Married | 1 | 0.000012 | 0.000012 | 0.01 | 0.919 |
| Error | 113 | 0.130467 | 0.001155 |  |  |
| Lack-of-Fit | 24 | 0.019788 | 0.000824 | 0.66 | 0.874 |
| Pure Error | 89 | 0.110680 | 0.001244 |  |  |
| Total | 115 | 0.131011 |  |  |  |

**Figure 3**

*Regression and Analysis of the One Kid/No Kids groups of players.*

**Regression Equation**

|  |  |  |
| --- | --- | --- |
| Average | = | 0.2679 - 0.00559 One Kid\_1 - 0.00020 Age |

**Analysis of Variance**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Age | 1 | 0.000039 | 0.000039 | 0.03 | 0.854 |
| One Kid | 1 | 0.000647 | 0.000647 | 0.56 | 0.455 |
| Error | 113 | 0.129833 | 0.001149 |  |  |
| Lack-of-Fit | 24 | 0.034348 | 0.001431 | 1.33 | 0.167 |
| Pure Error | 89 | 0.095484 | 0.001073 |  |  |
| Total | 115 | 0.131011 |  |  |  |
|  |  |  |  |  |  |

**Figure 4**

*Regression and Analysis of the Two or More Kids/No Kids groups of players.*

|  |  |  |
| --- | --- | --- |
| Average | = | 0.2660 - 0.00728 Two or More Kids - 0.00015 Age |

**Regression Equation**

**Analysis of Variance**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Age | 1 | 0.000024 | 0.000024 | 0.02 | 0.884 |
| Two or More Kids | 1 | 0.000825 | 0.000825 | 0.72 | 0.398 |
| Error | 113 | 0.129654 | 0.001147 |  |  |
| Lack-of-Fit | 23 | 0.026445 | 0.001150 | 1.00 | 0.471 |
| Pure Error | 90 | 0.103209 | 0.001147 |  |  |
| Total | 115 | 0.131011 |  |  |  |

**Figure 5**

*Regression and Analysis of the Two or More Kids/No Kids groups of players.*

|  |  |  |
| --- | --- | --- |
| Average\_1 | = | 0.3157 - 0.0006 Two or More Kids - 0.00197 Age |

**Regression Equation**

**Analysis of Variance**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Age\_1 | 1 | 0.001748 | 0.001748 | 1.23 | 0.273 |
| Two or More Kids\_1 | 1 | 0.000004 | 0.000004 | 0.00 | 0.959 |
| Error | 48 | 0.068253 | 0.001422 |  |  |
| Lack-of-Fit | 20 | 0.035158 | 0.001758 | 1.49 | 0.164 |
| Pure Error | 28 | 0.033095 | 0.001182 |  |  |
| Total | 50 | 0.070381 |  |  |  |

*Note:* This regression doesn’t include players with no kids so N = 51 instead of N=117.

***Married/Not Married***

The regression shows a drop in average for players that are married. Age has a p-value p = 0.586 so it isn’t significant enough to be a major confounding variable. The p-value of the marriage coefficient is p = 0.919. A p-value this high shows the null hypothesis cannot be rejected. There isn’t enough evidence to show a difference in batting average between the groups.

***One Kid/No Kids***

The regression shows players with one kid having a lower batting average. The F-value for the age coefficient is F = 0.03 with a p-value of p = 0.854. So, age doesn’t need to be considered as a confounding variable. The F-value for the One Child coefficient is F = 0.56 with a p-value p = 0.455. The data for these groups is more significant than the marriage groups, but not significant enough to reject the null hypothesis. It can’t be concluded that the difference between the means of the two groups is significantly less than zero.

***Two or More Kids/No Kids***

This regression also shows a negative difference between the two groups. The age coefficient has an F-value F = 0.02 and p-value p = 0.884. So, age isn’t significant enough to be considered. The F-value of the Kids coefficient is F=0.72 with p-value p=0.398. This is more significant than the last two analyses but most of this likely comes from the One Kid factor. However, this p-value is too small to suggest that the difference is significant.

***Two or More Kids/One Kid***

Finally, this last regression also shows a negative difference between the groups. Age has F-value F = 1.23 and p-value p = 0.273. This is more significant than the other regressions but still too weak to be considered. The coefficient for the difference between the groups has F-value F = 0.00 and p-value .959. This is an extremely high p-value which shows almost no difference between the groups. The null hypothesis can’t be rejected and the means of the groups can’t be assumed to be different.