**Homework 04 – Answer Key**

For Homework 04, you will be using the HELP dataset, learn more at:

* <https://melindahiggins2000.github.io/N736Fall2017_HELPdataset/> &
* <https://github.com/melindahiggins2000/N736Fall2017_HELPdataset>

Complete the following:

1. Perform a Simple Linear Regression for:
   * OUTCOME variable indtot: “Inventory of Drug Use Consequences (InDue) total score - Baseline”
   * PREDICTOR variable mcs: “SF36 Mental Composite Score - Baseline”
   * decide if you want to transform either variable indtot or mcs and if so, what transformation you applied and why
2. Perform regression diagnostics:
   * check the normality of the residuals (histogram and Q-Q plots)
   * check for linearity - is there any systematic relationship between the residuals and the predicted (or fitted) values?
   * homoscedasticity - plot of standardized residuals versus fitted values - this is known as a “Scale-Location” graph.
   * check for outliers and data points with high leverage or influence: outliers are often identified with standardized residuals > 3 (or <-3) and influential observations are often identified using Cook’s D
3. Provide a summary of the regression results.
   * provide a **FIGURE** of the model, in this case a scatterplot with the fitted line overlaid and 95% confidence intervals if you can
   * Make a **TABLE** presenting the fitted regression model (coefficients and tests of significance for those coefficients)
   * describe the variance explained by the model (based on r2)
   * describe the model itself based on the y-intercept and slope terms
   * note any limitations or issues with the model fit or interpretation of the model
4. Perform a One-way ANOVA for:
   * OUTCOME variable indtot: “Inventory of Drug Use Consequences (InDue) total score - Baseline”
   * GROUP variable racegrp: “Racial Group of Respondent”
   * I would suggest merging “other” and “hispanic” together and create a 3-group variable for race, since the “other” category is only about 6% of the sample.
   * options - you can use either an ANOVA or GLM modeling approach
   * if the GROUP variable is significant, also perform *post hoc* tests - use some kind of pairwise error rate adjustment (i.e. bonferroni, sidak, Tukey’s HSD, etc) - be sure to report which one you used and why
5. Perform model diagnostics:
   * homoscedasticity - look at a test for equal variance (Levene’s test or Bartlett’s test).
   * if this test of equal variances fails, you may want to report a modified F-test (e.g. Welch’s test)
6. Present a summary of the ANOVA results.
   * Make a **FIGURE** of the group mean differences - either an error-bar plot or a series of boxplots one for each group to show the group differences in the outcome
   * Make a **TABLE** presenting the ANOVA results
   * describe the model results - was the GROUP (racegrp) significant?
   * If GROUP is significant, what did the post hoc tests reveal?

**Simple Linear Regression of Inventory of Drug Use Consequences Total Score (indtot) against SF36 Mental Composite Score (mcs)**

A linear regression model was performed for the Inventory of Drug Use Consequences Total Score (INDTOT) against the SF36 Mental Composite Score (MCS). The fitted model was significant (F(1,451)=76.407, p<.001), with MCS explaining 14.3% of the variability in INDTOT (adjusted R2=0.143). The association between mcs and indtot was moderate-to-strong with a standardized slope term = -0.381. In terms of the original units, for every 10 points higher someone scored on their MCS, their drug use consequences decreased by 2.12 points. The intercept estimated an average INDTOT score of 42.445 for an MCS=0 (which is unlikely).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Unstandardized  Coefficients | | Standardized  Coefficients |  |  | 95.0% Confidence  Interval for B | |
|  | B | SEB | β | t | p-value | Lower Bound | Upper Bound |
| Intercept | 42.445 | .829 |  | 51.203 | <.001 | 40.816 | 44.074 |
| SF36 Mental Composite Score | -.212 | .024 | -.381 | -8.741 | <.001 | -.260 | -.164 |

**Figure: Plot of the fitted line for INDTOT against MCS with 95% Confidence Intervals overlaid**



**Regression Diagnostics**

The residuals are slightly skewed to the left, but the skewness is mild and the sample size here is large (n=453). These are OK.

 

**Homoscedasticity**

The variability of the residuals appears to be pretty constant across the predicted values.



**Outliers**

In reviewing the Cook’s D distances and the Standardized Residuals, there appears to be a few cases where the residuals were <-3 (more than 3 standard deviations from the mean) indicating influential outliers. These were for cases (IDs) 12, 50, 280, 379, 396, 418, 466. However, given the slightly skewness, these outliers are not extreme. We could rerun the regression model using a bootstrapping approach to obtain a more robust slope estimate to minimize the influence of these few outliers.

 

**ONEWAY ANOVA – Inventory of Drug Use Consequences total score (INDTOT) versus Race (racegrp)**

White subjects had the highest INDTOT scores averaging 37.58 +/- 6.04, followed by Hispanic/Other with average scores 35.74 +/- 6.91 with Black subjects having the lowest scores averaging 34.27 +/- 7.72. There were significant differences between the 3 races (F(2, 450)=10.350, p<.001). While the homogeneity of variance test (Levene’s test was statistically significant (F(2,450)=3.495, p=.031), the ratio of the variances for the race with the highest variance to the smallest variance = (7.716)2/(6.039)2 = 1.63 which is < 3 and < 4 both common rules of thumb when reviewing assumptions of equal variance.

|  |  |  |  |
| --- | --- | --- | --- |
| **Descriptive Statistics** | | | |
| Dependent Variable: indtot Inventory of Drug Use Consequences (InDue) total score - Baseline | | | |
| race3 | Mean | Std. Deviation | N |
| 1 black | 34.27 | 7.716 | 211 |
| 2 hispanic/other | 35.74 | 6.914 | 76 |
| 4 white | 37.58 | 6.039 | 166 |
| Total | 35.73 | 7.152 | 453 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Tests of Between-Subjects Effects** | | | | | | |
| Dependent Variable: indtot Inventory of Drug Use Consequences (InDue) total score - Baseline | | | | | | |
| Source | Type III Sum of Squares | df | Mean Square | F | Sig. | Partial Eta Squared |
| Corrected Model | 1016.782a | 2 | 508.391 | 10.350 | .000 | .044 |
| Intercept | 483860.948 | 1 | 483860.948 | 9850.224 | .000 | .956 |
| race3 | 1016.782 | 2 | 508.391 | 10.350 | .000 | .044 |
| Error | 22104.821 | 450 | 49.122 |  |  |  |
| Total | 601387.000 | 453 |  |  |  |  |
| Corrected Total | 23121.603 | 452 |  |  |  |  |
| a. R Squared = .044 (Adjusted R Squared = .040) | | | | | | |

**POST HOC – Using Sidak multiple pairwise comparison tests (you may have used a different adjustment)**

The significant difference in races was between Blacks and Whites (p<.001). No significant differences were seen between Hispanic/Other and Whites nor between Blacks and Hispanic/Other.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Multiple Comparisons** | | | | | | |
| Dependent Variable: indtot Inventory of Drug Use Consequences (InDue) total score - Baseline | | | | | | |
| Sidak | | | | | | |
| (I) race3 | (J) race3 | Mean  Difference (I-J) | Std.  Error | Sig. | 95% Confidence Interval | |
| Lower Bound | Upper Bound |
| 1 black | 2 hispanic/other | -1.47 | .938 | .315 | -3.71 | .78 |
| 4 white | -3.31\* | .727 | .000 | -5.05 | -1.57 |
| 2 hispanic/other | 1 black | 1.47 | .938 | .315 | -.78 | 3.71 |
| 4 white | -1.84 | .971 | .165 | -4.17 | .48 |
| 4 white | 1 black | 3.31\* | .727 | .000 | 1.57 | 5.05 |
| 2 hispanic/other | 1.84 | .971 | .165 | -.48 | 4.17 |
| Based on observed means.  The error term is Mean Square(Error) = 49.122. | | | | | | |
| \*. The mean difference is significant at the .05 level. | | | | | | |

**Figure: Errorbar plot of INDTOT Means and 95% Confidence Intervals By Race**

