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 cesd: “Center for Epidemiological Studies-Depression (CESD) total score - Baseline”
    - PREDICTOR variable indtot: ““Inventory of Drug Use Consequences (InDue) total score - Baseline””
    - decide if you want to transform either variable cesd or indtot and if so, what transformation you applied and why - you can also decide not to transform (i.e. tradeoffs between model fit and interpretability of your results) - discuss your reasoning
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 cesd: “Center for Epidemiological Studies-Depression (CESD) total score - Baseline”
    - GROUP variable racegrp: “Racial Group of Respondent”
    - 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 or equivalent).
    - 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?

Variables in HELP dataset to be used for Homework 04

Use these variables from HELP dataset for Homework 04

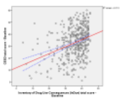
	Variable Label
cesd	CESD total score - Baseline
racegrp	Racial Group of Respondent
indtot	Inventory of Drug Use Consequences (InDue) total score - Baseline

Simple Linear Regression of Center for Epidemiological Studies-Depression (cesd) by Inventory of Drug Use Consequences Total Score (indtot)

A linear regression model was performed for the Center for Epidemiological Studies-Depression (CESD) by Inventory of Drug Use Consequences Total Score (INDTOT). The fitted model was significant (F(1,451)=57.251, p<.001), with INDTOT explaining 11.3% of the variability in INDTOT (adjusted R2=0.113). The association between INDTOT and CESD was moderate with a standardized slope term = 0.336. In terms of the original units (unstandardized coefficient), for every 10 points higher someone scored on their INDTOT, their CESD depressive symptoms score increased by 5.87 points. The intercept estimated an average CESD score of 11.866 for an INDTOT=0 (which is unlikely, especially in this population).

	Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
	B	SE_B	β	t	p-value	Lower Bound	Upper Bound
Intercept	11.866	2.828		4.196	<.001	6.309	17.423
SF36 Mental Composite Score	0.587	0.078	0.336	7.566	<.001	0.435	0.740

Figure: Plot of the fitted line for CESD by INDTOT with 95% Confidence Intervals overlaid

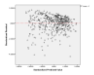


Regression Diagnostics

The residuals look very good – nice normal distribution and almost perfectly linear normal probability plot (P-P plot on right).

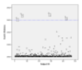
Homoscedasticity

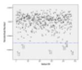
The variability of the residuals appears to be pretty constant across the predicted values. The variability (scatter about the horizontal line) is slightly less at higher predicted values, but this is minor and there are no obvious trends.



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 (for ID’s 1, 2, 3, 4, 5, 6). These cases also had higher “Cook’s D” distances indicating more leverage on the model fit. But these levels are not too 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 – Center for Epidemiological Studies-Depression (cesd) versus Race (racegrp)

Black subjects had the lowest CESD scores averaging 30.18 +/- 12.99, followed by Hispanic with average CESD scores 34.36 +/- 10.72, and then White subjects with average CESD scores of 35.30 +/- 11.85 with Other-race subjects having the highest average CESD scores 35.92 +/- 12.10. There were significant differences between the 4 races (F(3, 449)=6.304, p<.001). The homogeneity of variance test (Levene's test) was not statistically significant (F(3,449)=0.944, p=.419), so no additional adjustment of the F-test was done.

Descriptive Statistics

Dependent Variable: cesd CESD total score - Baseline

racenum Racial Group of Respondent	Mean	Std. Deviation	N
1 black	30.18	12.993	211
2 hispanic	34.36	10.717	50
3 other	35.92	12.099	26
4 white	35.30	11.854	166
Total	32.85	12.514	453

Tests of Between-Subjects Effects

Dependent Variable: cesd CESD total score - Baseline

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	2861.028[a]	3	953.676	6.304	.000	.040
Intercept	266261.807	1	266261.807	1759.988	.000	.797
racenum	2861.028	3	953.676	6.304	.000	.040
Error	67927.462	449	151.286			
Total	559562.000	453				
Corrected Total	70788.490	452				

a. R Squared = .040 (Adjusted R Squared = .034)

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 any other pairs of race groups.

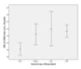
Multiple Comparisons

Dependent Variable: cesd CESD total score - Baseline

Sidak

		Mean Difference			95% Confidence Interval	
Racial Group 1	Racial Group 2	(Group 1 vs 2)	Std. Error	Sig.	Lower Bound	Upper Bound
1 black	2 hispanic	-4.18	1.935	.173	-9.29	.93
1 black	3 other	-5.74	2.556	.142	-12.50	1.01
1 black	4 white	-5.12*	1.276	.000	-8.49	-1.75
2 hispanic	3 other	-1.56	2.974	.996	-9.42	6.30
2 hispanic	4 white	-.94	1.984	.998	-6.18	4.30
3 other	4 white	.62	2.594	1.000	-6.23	7.48

<u>Figure: Errorbar plot of CESD Means and 95% Confidence Intervals By Race</u>



NOTE: The widths of the confidence intervals shown here are influenced by the sample sizes – black and whites has the highest numbers of subjects so their confidence intervals are the smallest whereas Hispanic and Other have smaller numbers and large confidence intervals. One way to visualize variance between the groups regardless of sample size is to use the standard deviation which is not dependent on sample size. Here is an alternative plot showing the means +/- 1 standard deviation versus means +/- the 95% confidence intervals.