EDUC 641 Lab: Applied Statistics in Education and Human Services I

Lab 9: 11/20 and 11/21

## Helpsheet for Assignment 4

**Helpsheet**

**Don’t forget to load packages: library(**tidyverse**)** and **library(**modelsummary**).**

**Use cont.csv dataset for assignment 4.**

* 1. **Use the datasummary\_skim() function to create continuous and categorical summary tables.**

1. **Select the variables you’re only looking at in your analysis**.

desc\_data <- your\_data %>%

**select(**variable\_1, variable\_2, variable\_3**)**

1. **For continuous variable:**

**datasummary\_skim(**desc\_data**,**

**type =** "numeric",

**histogram =** FALSE,

**title =** "Descriptive Statistics of Continuous Variables",

**output =** “table/descriptive\_cont.docx”**)**

1. **For categorical variable:**

**datasummary\_skim(**desc\_data**,**

**type =** "categorical",

**title =** "Descriptive Statistics of Categorical Variables",

**output =** “table/descriptive\_cat.docx”**)**

Then, edit the table as needed in the docx format to make it more readable.

**1.2 Using the geom\_point() and geom\_smooth() functions**

* 1. **ggplot(**your\_data, **aes(**x = continuous\_variable1, y = continuous\_variale2**))** + **geom\_point()** +

**geom\_smooth(** method = **lm,** se = **FALSE )** + # line of best fit

**labs(title =** “ Your Title for Your Graph ” ,

**x =** “Continuous Variable 1”,

**y =** “Continuous Variable 2” **)**

**1.3 Remember to discuss the direction, strength, and magnitude of the relationship.**

**2.2 Use Ordinary Least Squares estimation strategy to fit a model.**

1. **OLS fit**

fit <- **lm(**outcome **~** predictor, **data =** your\_data**)**

1. **Formatted table**

**modelsummary(**fit, **stars**=T,

**gof\_omit** = "Adj.|AIC|BIC|RMSE|Log",

**coef\_rename** = c("name\_of\_predictor" = "polished\_predictor\_name"),

**output**="tables/model\_fit.docx"**)**

**2.4. Test statistical assumptions of OLS.**

1. **First store predicted and residual values in your data frame.**

your\_data$predict<- **predict(**fit**)**

your\_data$resid <- **residuals(**fit**)**

1. **To view the distribution of the residuals and determine whether they’re normally distributed, use:**

**Summary statistics -**

**summary(**your\_data$resid**)**

**sd(**your\_data$resid**)**

**Distribution -**

**ggplot(data=** your\_data**,**

**aes(resid)) +**

**geom\_density() +**

**xlab("Raw residuals") + ylab("Proportion")**

1. **Plotting the residuals versus fitted values, to see if the residuals are symmetrically distributed, use:**

**ggplot(data=** your\_data**,**

**aes(x =** predict**, y =** resid**)) +**

**geom\_point() +**

**geom\_hline(yintercept =** 0, **color =** "red", **linetype="**dashed"**) +**

**ylab("Residuals") + xlab("Fitted values")**