University of Kansas **Dr. William Duncan**
Institute for Policy and Social Research **Summer REU Workshop**
Sponsored by ARISE **Regression and its Applications**

1 Overview

This worksheet covers topics related to hypothesis tests, and confidence intervals for the normal distribution.

2 Questions

Question 1:
The RIOPA dataset has information on the concentration of various air pollutant measurements and temperature measurements. Pollutants are known to trap heat and increase temperature. One pollutant in the RIOPA dataset is called acrolein.

(a) In R, run a regression where you use the variable **acrolein** to predict the variable **temp**.

(b) What is the slope of the regression line? Is the slope positive or negative?

(c) What is the correlation?

(d) What is the formula for and the value of the test statistic for testing Ho: $\rho = 0$ vs. Ha: $\rho \neq 0$

  (i) What are the degrees of freedom?
  (ii) What is the p-value for testing the correlation?
  (iii) Give the conclusion of the test in context of testing at the 5% significance level.

(e) What is the regression line?

  (i) Find the predicted value and residual for temp if the acrolein concentration is 1.616 ppm.

(f) Interpret the slope of the line in context.

(g) What is the standard error of the slope?

(h) What are the null and alternative hypotheses for testing the slope?

  (i) What is the p-value for a test of the slope?
  (ii) Give the conclusion of the test in context if testing at a 5% significance level.

(i) Compare the two p-values, from the test for correlation and the test for slope.

(j) How many measurements were taken and included in the dataset?

(k) Use the sum of squares values to compute R2, and compare the result with the value given in the output

  (i) Interpret R2 in context.

(l) What are the F-statistic and p-value of the ANOVA test?

  (i) How does this p-value compare to the two found in parts (d) and (g)?

Question 2:
Another air pollutant in the RIOPA dataset is hexaldehyde. In this question we will look at the relation of hexaldehyde and temperature measurements in the dataset. You can find the output for the regression below. Use the output to answer the questions below.

Source	SS	df	MS	Number of obs	=	851
				F(1, 849)	=	63.66
Model	2393.9384	1	2393.9384	Prob > F	=	0.0000
Residual	31926.9943	849	37.6054114	R-squared	=	0.0698
				Adj R-squared	=	0.0687
Total	34320.9327	850	40.3775679	Root MSE	=	6.1323

temp	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
hexaldehyde	.5427493	.0680249	7.98	0.000	.4092326	.676266
_cons	19.51741	.3116969	62.62	0.000	18.90563	20.1292

(a) What is the least squares regression equation?

(b) Is the hexaldehyde concentration a good predictor of the number of field goals made? Justify your answer using a statistical test and specific values in the output.

(c) Interpret the value of R2 for this model.

(d) Use the regression equation to find the predicted number of field goals made if the hexaldehyde concentration is 4.407 ppm.

Question 3:
It is often useful to be able to run two different models with the same outcome and compare those models.

Model 1: In R, run a regression that predicts temp with the following three explanatory variables: **formaldehyde**, **benzaldehyde**, and **methylglyoxal**.

(a) What is the regression equation?

(b) Use this equation to predict the temperature in a location with a **formaldehyde** concentration of 58.9509, **benzaldehyde** concentration of 5.2213, and a **methylglyoxal** concentration of 3.5969.

(c) Test to see if this model is useful for predicting **temp** with the following steps:

  (i) State the null and alternative hypotheses.
  (ii) Give the value of the test statistic.
  (iii) Give the p-value.
  (iv) At the 5% significance level, what is the generic conclusion?
  (v) Is the model useful for predicting **temp**?

(d) Test to see if **benzaldehyde** is useful for predicting **temp** if **formaldehyde** and **methylglyoxal** are already in the model with the following steps:

  (i) State the null and alternative hypotheses.
  (ii) Give the value of the test statistic.
  (iii) Give the p-value.
  (iv) At the 5% significance level, what is the generic conclusion?
  (v) Is **benzaldehyde** useful for predicting **temp**?

(e) State and interpret R2 for this model.

(f) Which explanatory variable is least significant in predicting temp? How do you know?

Model 2: Rerun the regression without the least significant explanatory variable.

(g) What is the regression equation?

(h) Compare the values for the adjusted R2 for each model below.

  (i) What is the value of the adjusted R2 for Model 1?
  (ii) What is the value of the adjusted R2 for Model 2?
  (iii) Does this change indicate that Model 2 is better than Model 1 for predicting temp? Explain.

3 Appendix

variable name	storage type	variable label
linkid	str8	Observation identification number
homeid	str5	Home identification number
state	str2	State
type	str24	Air type
sampleid	long	Sample identification number
temp	double	Temperature, measured in Celsius
formaldehyde	float	Concentration of formaldehyde in the air, parts per million (ppm)
acetaldehyde	float	Concentration of acetaldehyde in the air, parts per million (ppm)
acetone	float	Concentration of acetone in the air, parts per million (ppm)
acrolein	double	Concentration of acrolein in the air, parts per million (ppm)
crotonaldehyde	double	Concentration of crotonaldehyde in the air, parts per million (ppm)
benzaldehyde	double	Concentration of benzaldehyde in the air, parts per million (ppm)
methylglyoxal	double	Concentration of methylglyoxal in the air, parts per million (ppm)
hexaldehyde	double	Concentration of hexaldehyde in the air, parts per million (ppm)