Advanced Statistics  
Assignment 1  
Claude M. Schrader

**Research Question:**

How well do the percentage of population that are college graduates, the percentage of households with broadband availability in January 2004, and population density explain the variation in percentage of Kentucky households subscribing to broadband internet in 2005?

**Measures:**

adopt05 is the percentage of Kentucky households subscribing to broadband from a 2005 survey. This is the dependant variable.

college tracks the percent of college graduates in the population in 2004.

hhpct\_0401 is the percentage of Kentucky households with broadband available to them.

popden2000 is the Population Density as of the 2000 census.

adopt05 college hhpct\_0401 popden2000

nbr.val 120.000000000 120.000000000 120.00000000 120.000000

nbr.null 0.000000000 0.000000000 0.00000000 0.000000

nbr.na 0.000000000 0.000000000 0.00000000 0.000000

min 0.071000000 0.041000000 0.00400000 16.100000

max 0.548000000 0.472000000 1.00000000 1336.700000

range 0.477000000 0.431000000 0.99600000 1320.600000

sum 28.654000000 24.981000000 67.18200000 9964.300000

median 0.214500000 0.196000000 0.60050000 45.300000

mean 0.238783333 0.208175000 0.55985000 83.035833

SE.mean 0.009236575 0.007408363 0.02594906 13.992603

CI.mean.0.95 0.018289340 0.014669299 0.05138173 27.706752

var 0.010237717 0.006586062 0.08080245 23495.153075

std.dev 0.101181606 0.081154554 0.28425771 153.281287

coef.var 0.423738141 0.389838134 0.50773906 1.845966

**Methods:**

The Null hypothesis: college, hhpct\_0401, and popden2000 do not have a statistically significant effect on adopt05

The Alternative hypothesis: these three variables will have a statistically significant effect on adopt05.

The regression test performed will try to fit a line to a scatterplot of each of the dependent variables with adopt05, and generate data that will enable us to quantify the accuracy and significance of this test.

For the test statistic, R will perform a T test, which uses the t distribution to give you the t score for a particular value. In this case, it will perform a T test on the slope of each fitted line and of the intercept. If this T value is greater than 1.96 or less than -1.96, the value in question is statistically significant at the 95% confidence level.

The line fitted to the dataset can also be useful for drawing inferences about other samples than the data. This line isn’t aware of things like the fact that a percentage cannot be greater than 100%, so some careful attention to this inferred data is necessary.

**Results:**

After trying a few models, I settled on model 6, which has adopt05 as the dependent variable, and college, hhpct\_0401, and popden2000 as the three independent variables. With the highest R2 and adjusted R2, this model explains the largest amount of variance in adopt05. With t scores over 1.96, and p values for the t test less than 0.01, the fit lines for all three independent variables are statistically significant. The p value of the F statistic is also less than .01, which indicates that the fit of the line is statistically significant.

Model 1:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.0673175 0.0312489 2.154 0.0333 \*

medhhinc 0.0046882 0.0008243 5.687 0.0000000953 \*\*\*

Residual standard error: 0.09002 on 118 degrees of freedom

Multiple R-squared: 0.2151, Adjusted R-squared: 0.2085

F-statistic: 32.35 on 1 and 118 DF, p-value: 0.00000009531

Model 3:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.402968 0.168809 2.387 0.018597 \*

medhhinc 0.001405 0.001071 1.312 0.192023

college 0.503936 0.136904 3.681 0.000354 \*\*\*

medage -0.007055 0.003502 -2.015 0.046249 \*

Residual standard error: 0.08291 on 116 degrees of freedom

Multiple R-squared: 0.3455, Adjusted R-squared: 0.3286

F-statistic: 20.41 on 3 and 116 DF, p-value: 0.0000000001087

Model 6:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.08623234 0.02123268 4.061 0.0000890 \*\*\*

college 0.42934669 0.10008186 4.290 0.0000372 \*\*\*

hhpct\_0401 0.08673592 0.02800457 3.097 0.002450 \*\*

popden2000 0.00017598 0.00005173 3.402 0.000918 \*\*\*

Residual standard error: 0.07628 on 116 degrees of freedom

Multiple R-squared: 0.446, Adjusted R-squared: 0.4316

F-statistic: 31.12 on 3 and 116 DF, p-value: 0.000000000000007787

**Conclusion:**

The results of model 6 seem reasonable to me, at first analysis. College grads, broadband availability, and population density all have a statistically significant effect on broadband adoption. This seems to indicate that more urban areas had higher broadband adoption in 2005, This all seems to make sense when you consider the concentration of white collar jobs in urban areas, and the difficulty in maintaining affordable and reliable broadband in rural areas.

To refine this analysis, I would want to repeat it for multiple years, looking for trends and changes over time. It would also be useful to gather data on education and marketing campaigns if possible, to help determine whether any effort and money spent on outreach was allocated properly.