NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Week 2 Assignment - Computational: OLS Linear Regression
MSDS 410

Modeling the US States Data

Data: The data for this assignment is the US State data set: USStates.CSV. It is a 12 variable dataset with n=50 records. The data, calculated from census data, consists of state-wide average or proportion scores for the non-demographic variables. As such, higher scores for the composite variables translate into having more of that quality. There is no other information available about this data.

Objective: Every dataset has a "story" to tell. It just doesn't have the voice to speak the story. In a sense, it is your job as the analyst to "tell" the story that the data has to offer. That is your objective here: To uncover the story this dataset has to tell.

Tasks: To achieve the objective please complete the following tasks enumerated below. You are to use R to obtain any graphs or statistics requested.

Data Description: https://cran.r-project.org/web/packages/Lock5Data/Lock5Data.pdf

Variable	Description
State	State name
Region	MW=Midwest, NE=Northeast, S=South, W=West
Population	Number of residents (in millions for 2014)
HouseholdIncome	Median household income (in $1,000's)
HighSchool	Percent of residents (ages 25-34) who are high school graduates
College	Percent of residents (ages 25-34) who are college graduates
Smokers	Percent of residents who smoke
PhysicalActivity	Percent who do 150+ minutes of aerobic physical activity per week
Obese	Percent obese residents (BMI 30+)
NonWhite	Percent nonwhite residents (in 2013)
HeavyDrinkers	Percent heavy drinkers (men: 3+ drinks/day, women 2+ drinks/day)
TwoParents	Percent of children living in two-parent households
Insured	Percent of adults (ages 18-64) who have any kind of health coverage

1. Given the variables in this dataset, which variables can be considered explanatory (X) and which considered response (Y)? Can any variables take on both roles? What is the population of interest for this problem (yes – this is a trick question!)?
    - Following variables are considered explanatory (X):
        - State
        - Region
        - Population
        - HighSchool
        - College
        - Smokers
        - PhysicalActivity
        - Obese
        - NonWhite
        - HeavyDrinkers
        - TwoParents
        - Insured
    - Following variable is considered response (Y):
        - HouseholdIncome
    - Can any variables take on both roles?
        - Yes, any variables can take on both roles, such as obese, college, insured, etc.
    - What is the population of interest for this problem (yes – this is a trick question!)?
        - Residents who earn median household income from 50 states

2. For the duration of this assignment, let's have HOUSEHOLDINCOME be the response variable (Y). Also, please consider the STATE, REGION and POPULATION variables to be demographic variables. Obtain basic summary statistics (i.e. n, mean, std dev.) for each variable. Report these in a table. Then, obtain all possible scatterplots relating the non-demographic explanatory variables to the response variable (Y).

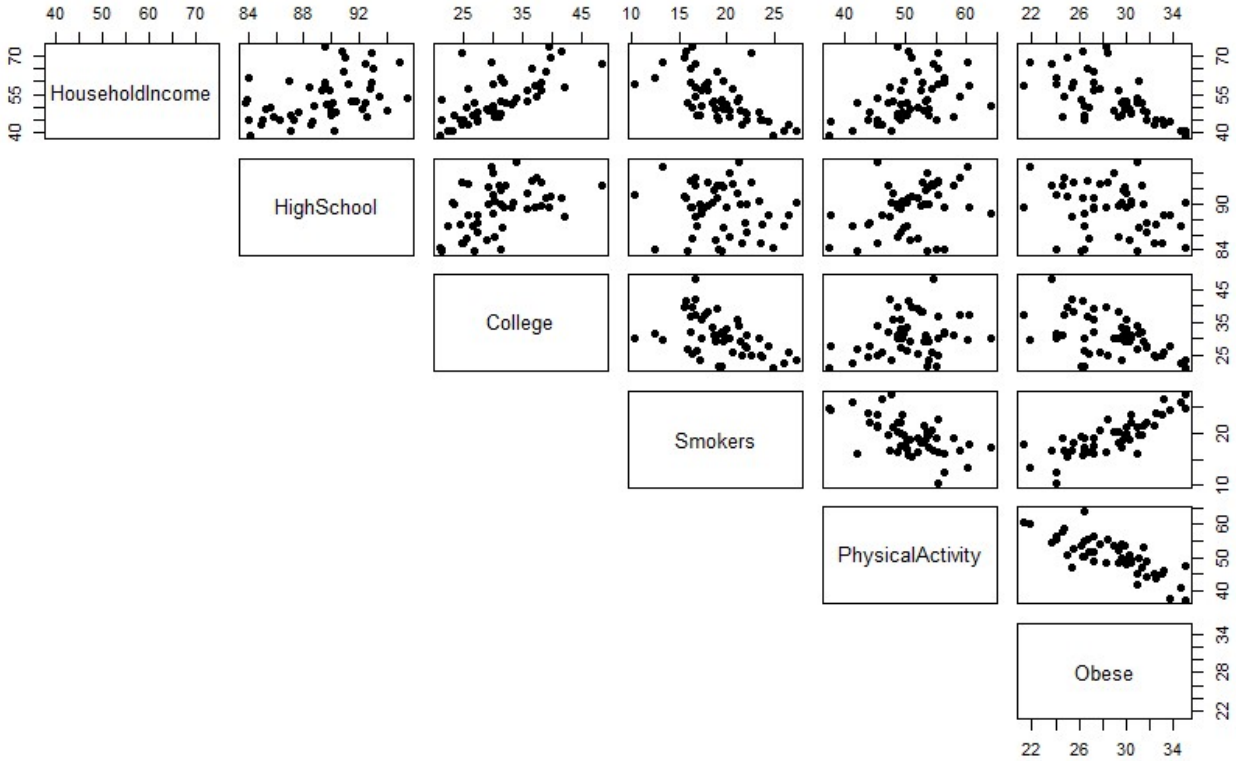
Summary Statistics of Demographics

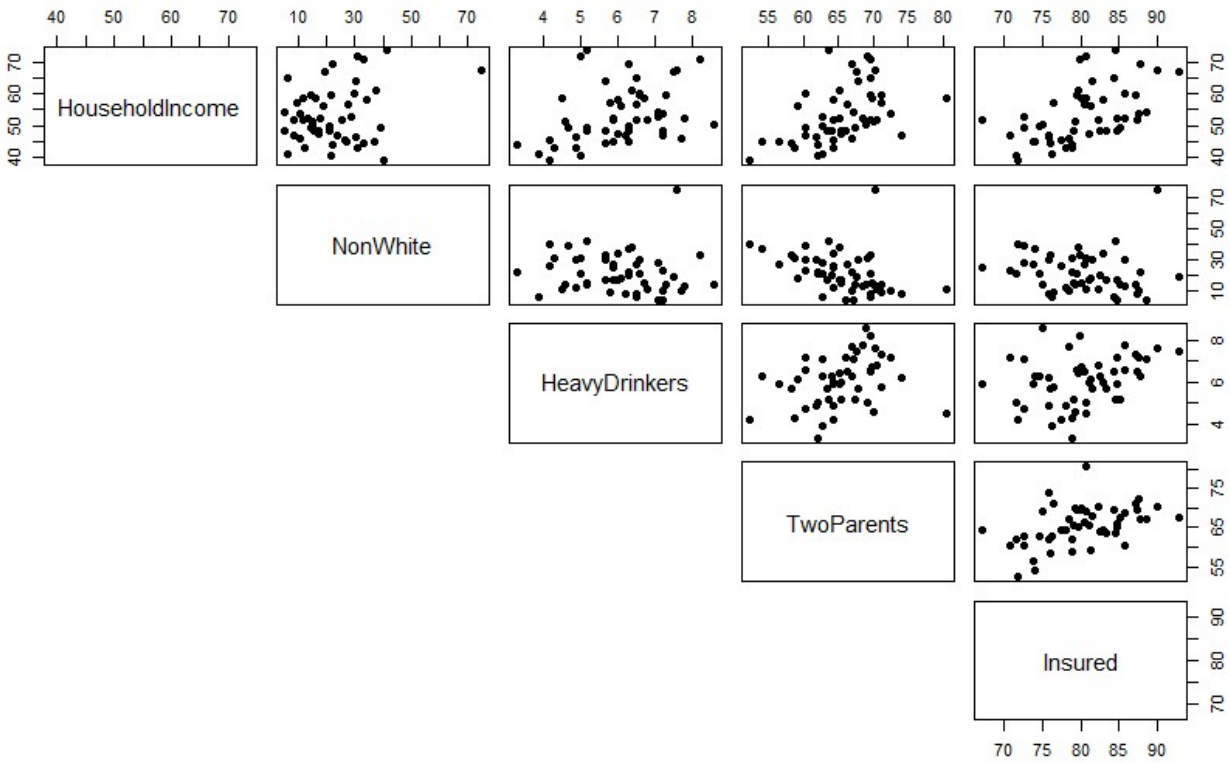
State			Region			Population		
Alabama	:	1	MW	:	13	Min.	:	0.584
Alaska	:	1	NE	:	11	1st Qu.	:	1.858
Arizona	:	1	S	:	13	Median	:	4.532
Arkansas	:	1	W	:	13	Mean	:	6.364
California	:	1				3rd Qu.	:	6.983
Colorado	:	1				Max.	:	38.803
(Other)	:	44						

Summary Statistics of the Rest of the Data

	HouseholdIncome	HighSchool	College	Smokers	PhysicalActivity
N	50	50	50	50	50
Min	39.03	83.80	21.10	10.30	37.40
1st Quartile	46.81	87.10	25.90	16.65	47.65
Median	51.76	89.70	30.15	19.05	50.65
Mean	53.28	89.32	30.83	19.32	50.73
10% Trimmed Mean	52.61	89.38	30.52	19.24	50.83
Standard Deviation	8.69	3.11	6.08	3.52	5.51
3rd Quartile	58.72	91.62	35.25	21.48	54.13
Max	73.54	95.40	48.30	27.30	64.10
Range	34.51	11.60	27.20	17.00	26.70
Interquartile Range	11.91	4.52	9.35	4.83	6.48
Skew	0.65	-0.19	0.56	0.08	-0.19
Kurtosis	-0.33	-0.81	0.06	0.20	0.36

	Obese	NonWhite	HeavyDrinkers	TwoParents	Insured
N	50	50	50	50	50
Min	21.30	4.80	3.30	52.30	67.30
1st Quartile	26.40	13.35	5.20	62.70	76.15
Median	29.40	20.75	6.15	65.45	79.90
Mean	28.77	22.16	6.05	65.52	80.15
10% Trimmed Mean	28.79	21.09	6.06	65.66	80.16
Standard Deviation	3.37	12.69	1.18	5.17	5.49
3rd Quartile	31.08	30.22	6.78	69.50	84.48
Max	35.10	75.00	8.60	80.60	92.80
Range	13.80	70.20	5.30	28.30	25.50
Interquartile Range	4.68	16.87	1.58	6.80	8.33
Skew	-0.12	1.51	-0.15	-0.08	-0.01
Kurtosis	-0.54	4.68	-0.37	0.96	-0.35





3. Obtain all possible pairwise Pearson Product Moment correlations of the non-demographic variables with Y and report the correlations in a table. Given the scatterplots from step 2) and the correlation coefficients, is simple linear regression an appropriate analytical method for this data? Why or why not?

	HouseholdIncome
HighSchool	0.4308448
College	0.6855909
Smokers	-0.6375225
PhysicalActivity	0.4404166
Obese	-0.6491116
NonWhite	0.2529418
HeavyDrinkers	0.3730143
TwoParents	0.4776443
Insured	0.5496786

- Given the scatterplots from step two and the correlation coefficients, simple linear regression is an appropriate analytical method for this data because we observed stronger positive and negative correlation variables between predictors and response variables. Highly positive correlated variables are college, insured, etc. and highly negatively correlated variables are a smoker and obese.

4. Fit a simple linear regression model to predict Y using the COLLEGE explanatory variable. Use the base STAT lm(Y~X) function. Why would you want to start with this explanatory variable? Call this Model 1. Report the results of Model 1 in equation form and interpret each coefficient of the model in the context of this problem. Report the ANOVA table and model fit statistic, R-squared. Use the summary statistics from steps 2) and 3) to verify, by hand computation, the estimates for the slope and intercept.
    - `model1 <- lm(HouseholdIncome ~ College, data = mydata1)`

- I want to start with College explanatory variable because it is the highest correlated variable among other explanatory variables.
- Y_hat = 23.0664 + 0.9801 * College + error
- Interpretation: If HouseholdIncome is median household income in $1,000's and college is a percent of residents from ages 25 to 34 who graduated college, then College (or betta 1) variable measures the change in median household income given another college graduate. For every additional college graduate, on average, the median household income goes up by 0.98 units or percentage. The error term in here is a catch-all for what we miss with this simple model: the true relationship is probably not linear, there may be other variables that cause variation in Y (HouseholdIncome), and there may be measurement error.
- ANOVA - Analysis of Variance Table:

```
Response: HouseholdIncome
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
College	1	1739.4	1739.36	42.572	3.94E-08
Residual	48	1961.1	40.86		

- Model fit statistic, R-squared

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	23.0664	4.7187	4.888	1.18E-05
College	0.9801	0.1502	6.525	3.94E-08

```
Residual standard error: 6.392 on 48 degrees of freedom
Multiple R-squared:   0.47,     Adjusted R-squared:  0.459
F-statistic: 42.57 on 1 and 48 DF,  p-value: 0.00000003941
```

- Slope (betta 1) = r (sd_y / sd_x) = 0.6855909 * (8.69 / 6.08) = 0.9801
- Intercept (betta 0) = mean (y) – slope * mean (x) = 53.28 – 0.9801 * 30.83 = 23.066

5. Write R-code to calculate and create a variable of predicted values based on Model 1. Use the predicted values and the original response variable Y to calculate and create a variable of residuals (i.e. residual = Y – Y_hat = observed minus predicted) for Model 1. Using the original Y variable, the predicted, and/or residual variables, write R-code to:

```
Y_hat <- predict(model1)
residual <- mydata1$HouseholdIncome - Y_hat
residual
```

- Square each of the residuals and then add them up. This is called sum of squared residuals, or sums of squared errors. (SSE)
  - `sum(residual^2) = 1961.13`
- Deviate the mean of the Y's from the value of Y for each record (i.e. Y – Y_bar). Square each of the deviations and then add them up.)This is called sum of squares total. (SST)
  - `sum((mydata1$HouseholdIncome - mean(mydata1$HouseholdIncome))^2) = 3700.488`
- Deviate the mean of the Y's from the value of predicted (Y_hat) for each record (i.e. Y_hat – Y_bar). Square each of these deviations and then add them up. This is called the sum of squares due to regression. (SSR)
  - `sum((Y_hat - mean(mydata1$HouseholdIncome))^2) = 1739.359`
- Calculate a statistic that is: (Sum of Squares due to Regression) / (Sum of squares Total). (R Squared)

- 1739.359 / 3700.488 = 0.470035

Verify and note the accuracy of the ANOVA table and R-squared values from the regression printout from part 4), relative to your computations here.

- SSE is verified: 1961.1
- SSR is verified: 1739.4
- SST is verified: SST = SSE + SSR = 1961.1 + 1739.4 = 3700.5
- R squared is verified: 0.47
  - we can also verify it with correlation coefficient: R^2 = r^2 = 0.6856^2 = 0.47

6. Fit a multiple linear regression model to predict Y using COLLEGE and INSURED as the explanatory variables. Use the base lm(Y~X) function. Call this Model 2. Report the results of Model 2 in equation form, interpret each coefficient of the model in the context of this problem, and report the model fit statistic, R-squared. How have the coefficients and their interpretations changed? Calculate the change in R-squared from Model 1 to Model 2 and interpret this value. For this specific problem, is it OK to use the hypothesis testing results to determine if the additional explanatory variable should be retained or not? Think statistically using first principals. Discuss. NOTE: The topic of hypothesis testing in regression is the focus of Module 2 – you should NOT need to read anything about hypothesis testing to answer this.
   - `model2 <- lm(HouseholdIncome ~ College + Insured, data = mydata1)`
   - Y_hat = 9.6728 + 0.8411* College + 0.2206*Insured + error
   - Interpretation: for every resident who graduated from college, we are getting an increase in about 0.84% change in median household income, holding all other factors fixed. For every resident who obtained health insurance, we are getting a raise in about 0.22% change in median household income, holding all other factors fixed.
   - When I keep the insured residents constant, if the residents who graduated college increase by a percentage (or unit), then on average, the median household income increases by 0.84%. When I keep the college graduate residents constant, if the health care insured residents increase by a percentage (or unit), then on average, the median household income increases by 0.22%.

```
Coefficients:
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept)    9.6728   14.8628    0.651  0.518339
College        0.8411    0.2098     4.01  0.000216
Insured        0.2206    0.2321     0.95  0.346759

Residual standard error: 6.398 on 47 degrees of freedom
Multiple R-squared:   0.48,     Adjusted R-squared:  0.4579
F-statistic: 21.69 on 2 and 47 DF,  p-value: 0.0000002116
```

   - Getting the health insurance did not add much because the corresponding p-value is no longer significant, with a value of around 0.35. Model 1 R-squared value was 0.47, and Model 2 R-squared value was 0.48. Adding the insured variable moved the R-squared amount by 0.01. We can also see that there is a high correlation (about 70%) between college and insured variables. We can drop the insured variable and explore other variables.
   - The p-values indicate that college variable is related to medium household income, but that there is no evidence that the insured variable is associated with medium household income, in the presence of college variable. If there is a small p-value, then we can infer that there is an association between the predictor and the response. We reject the null hypothesis—that is, we declare a relationship to exist between X and Y —if the p-value is small enough. So, there is an association between the college graduates and the household income but not insured residents and household income.

7. In a sequential fashion, continue to add in the non-demographic variables into the prediction model, one variable at a time. Make a table summarizing the change in R-squared that is associated with each variable added. Based on this information, what variables should be retained for a "best" predictive model? What criteria seems appropriate to you?
During this problem, practice interpreting coefficients for each model. Do any of the interpretations become counter intuitive as you fit more and more complex models? What does, or would, this mean for the model being developed? You do not need to report all of the coefficient interpretations, but this is a general question to contemplate and skill to use in model determination. Please write a short summary of your conclusions here.

	R-Squared Value	Change Points
College	0.47	
Smokers	0.59	0.12
HighSchool	0.62	0.02
PhysicalActivity	0.62	0.00
Obese	0.62	0.01
NonWhite	0.71	0.09
HeavyDrinkers	0.71	0.00
TwoParents	0.74	0.02
Insured	0.74	0.00

```
Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -8.56579   16.38635  -0.523  0.603720
College      0.70795    0.13138   5.389 0.0000025
NonWhite     0.27469    0.06447   4.261  0.000103
Smokers     -0.38538    0.26233  -1.469  0.148766
TwoParents   0.63155    0.17826   3.543  0.000935

Residual standard error: 4.798 on 45 degrees of freedom
Multiple R-squared:   0.72,     Adjusted R-squared:  0.6951
F-statistic: 28.93 on 4 and 45 DF,  p-value: 0.00000000000625
```

- College, nonwhite, smokers, and two-parents variables gave R-squared value 0.72.
- For every resident who graduated from college, we are getting an increase in about 0.71% change in median household income, holding all other factors fixed. For every nonwhite resident, we are getting an increase in about 0.28% change in median household income, holding all other factors fixed. For every resident who smokes, we are getting a decrease in about 0.39% change in median household income, holding all other factors fixed. For every child who lives in two-parent households, we are getting an increase in about 0.63% change in median household income, holding all other factors fixed.
- However, a large p-value indicates that there is no evidence that the smokers' variable is associated with medium household income. When we drop the smokers variable, we lose about 0.01 R-squared value. Thus, college, nonwhite, and two-parents variables should be retained for a "best" predictive model, and they give R-squared value 0.71.
- Adding non-demographic variables into the model one variable at a time helped us to utilize forward selection criteria. Looking R squared, and p-values seem appropriate to retain for a "best" predictive model.
- Conclusion: the percent of residents who graduated from college, nonwhite, and children living in two-parent households are good predictors of medium household income.

8. Now that you have a sense of which explanatory variables contribute to explaining HOUSEHOLDINCOME, refit a model using only the set of variables you consider to be appropriate

to model Y. Report this model, interpret the coefficients, and interpret R-squared in the context of this problem. Discuss why is it necessary to refit this model.

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-27.6406	10.12129	-2.731	0.00892
College	0.78053	0.12326	6.332	9.17E-08
NonWhite	0.3136	0.05951	5.27	3.53E-06
TwoParents	0.76175	0.15661	4.864	1.38E-05

```
Residual standard error: 4.858 on 46 degrees of freedom
Multiple R-squared:  0.7066,    Adjusted R-squared:  0.6875
F-statistic: 36.93 on 3 and 46 DF,  p-value: 0.000000000002633
```

- `model3 <- lm(HouseholdIncome ~ College + TwoParents + NonWhite, data = mydata1)`
- A tiny increase in R-squared value provides additional evidence that smokers, high school graduates, physical activity lovers, obese, heavy drinkers, and health care insured variables can be dropped from the model. Essentially, they provide no real improvement in the model fit to the training samples, and their inclusion will likely lead to poor results on independent test samples due to overfitting. In contrast, the model containing only College as a predictor had an R-squared value of 0.47. Adding nonwhite and two-parents residents to the model leads to a substantial improvement (from 0.47 to 0.71) in R-squared value. This implies that a model that uses college graduate, nonwhite, two parent residents to predict medium household income is substantially better than one that uses only college graduates.
- For every resident who graduated from college, we are getting an increase in about 0.78% change in median household income, holding all other factors fixed. For every nonwhite resident, we are getting an increase in about 0.31% change in median household income, holding all other factors fixed. For every child who lives in two-parent households, we are getting an increase in about 0.76% change in median household income, holding all other factors fixed.

9. You are welcome to conduct any other analyses you wish to embellish your understanding of this dataset.
    - It will be better if we can go through this assignment in class. I can learn more by looking at how other people interpreted the model.

10. Given what you've learned from this modeling endeavor, what overall conclusions do you draw? What is the "Story" contained in this data? What have you learned? What are your Prescriptive Recommendations for action based on this evidence? Finally, feel free to reflect on what you've learned from a modeling perspective.
    - I learned how each predictor influence the response varaible. I also learned there is a relationship between residents who were surveyed from 50 states with various backgrounds and medium household income. By running some analysis, I learned how to identify the strength of the relationship and how to predict the household income with a high level of accuracy. We can predict median household income with three variables: the percent of residents who graduated from college, nonwhite, and children living in two-parent households.

Assignment Document:

Results should be presented and discussed in the numerical order of the questions given. The report should not contain unnecessary results or information. Tables are highly effective for summarizing

data across multiple models. The document MUST be submitted in pdf format. Please use the naming convention: Assign1_YourLastName.pdf.