Robin Singh

NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

Assignment #1: OLS Linear Regression
MSHA 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.

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!)?

Explanatory: Region, Population, Household Income, Highschool, College, and Two Parents

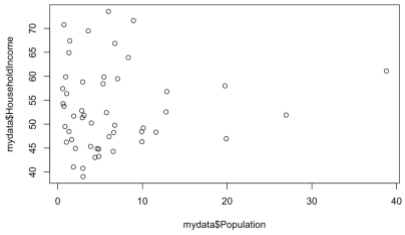
Response: Insured, Obese, Heavy Drinkers, Smokers, Physical Activity

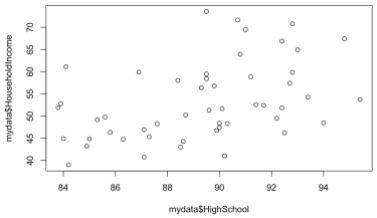
Both: Highschool, Household Income, College, and Physical Activity

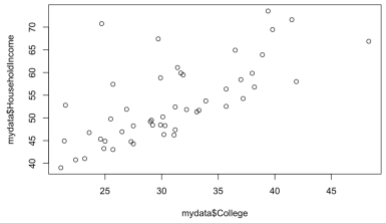
Population of Interest: US population

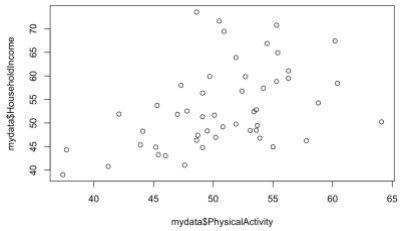
2. For the duration of this assignment, let's have HOUSEHOLDINCOME be the response variable (Y). Also, please consider the STATE, REGION and POPULATION variables to be demographic variables. Obtain basic summary statistics (i.e. n, mean, std dev.) for each variable. Report these in a table. Then, obtain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

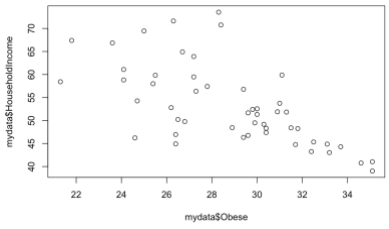
```
   State              Region            Population     HouseholdIncome   HighSchool       College         Smokers
Length:50          Length:50          Min.   : 0.584   Min.   :39.03   Min.   :83.80   Min.   :21.10   Min.   :10.30
Class :character   Class :character   1st Qu.: 1.858   1st Qu.:46.81   1st Qu.:87.10   1st Qu.:25.90   1st Qu.:16.65
Mode  :character   Mode  :character   Median : 4.532   Median :51.76   Median :89.70   Median :30.15   Median :19.05
                                      Mean   : 6.364   Mean   :53.28   Mean   :89.32   Mean   :30.83   Mean   :19.32
                                      3rd Qu.: 6.983   3rd Qu.:58.72   3rd Qu.:91.62   3rd Qu.:35.25   3rd Qu.:21.48
                                      Max.   :38.803   Max.   :73.54   Max.   :95.40   Max.   :48.30   Max.   :27.30
PhysicalActivity     Obese          NonWhite     HeavyDrinkers     TwoParents       Insured
Min.   :37.40    Min.   :21.30   Min.   : 4.80   Min.   :3.300   Min.   :52.30   Min.   :67.30
1st Qu.:47.65    1st Qu.:26.40   1st Qu.:13.35   1st Qu.:5.200   1st Qu.:62.70   1st Qu.:76.15
Median :50.65    Median :29.40   Median :20.75   Median :6.150   Median :65.45   Median :79.90
Mean   :50.73    Mean   :28.77   Mean   :22.16   Mean   :6.046   Mean   :65.52   Mean   :80.15
3rd Qu.:54.12    3rd Qu.:31.07   3rd Qu.:30.23   3rd Qu.:6.775   3rd Qu.:69.50   3rd Qu.:84.47
Max.   :64.10    Max.   :35.10   Max.   :75.00   Max.   :8.600   Max.   :80.60   Max.   :92.80
```

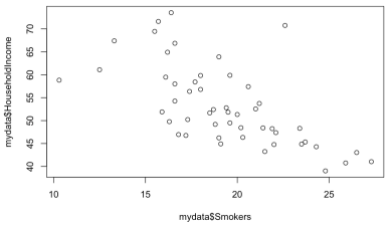


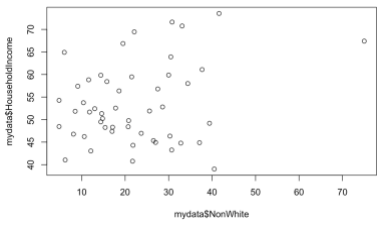


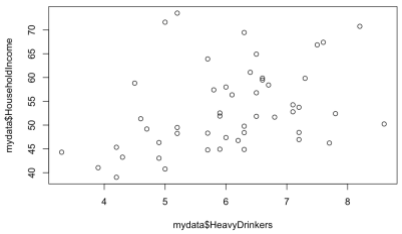


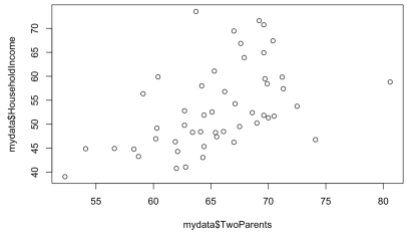


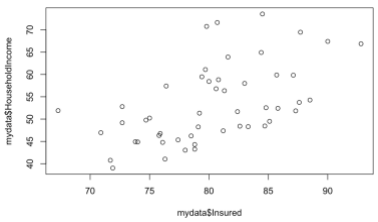












3. Obtain all possible pairwise Pearson Product Moment correlations of the non-demographic variables with Y and report the correlations in a table. Given the scatterplots from step 2) and the correlation coefficients, is simple linear regression an appropriate analytical method for this data? Why or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

```
[1] "Population: 0.0737382749503407"
[1] "HighSchool: 0.430844783768895"
[1] "College: 0.685590939825705"
[1] "PhysicalActivity: 0.440416649426189"
[1] "Obese: -0.649111608835359"
[1] "NonWhite: 0.252941779441709"
[1] "HeavyDrinkers: 0.373014275725829"
[1] "TwoParents: 0.477644344118505"
[1] "Insured: 0.549678617344339"
[1] "Smokers: -0.637522473059737"
```

The variables College, Obese, Insured, and Smokers all have some linear relationships to House Income, as their absolute values are greater than 0.5, and hence would be good to measure together towards their combined impact to the Y variable.

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.

```
Call:
lm(formula = mydata$HouseholdIncome ~ mydata$College)

Residuals:
   Min     1Q Median     3Q    Max
-7.319 -4.245 -2.203  2.652 23.484

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     23.0664     4.7187   4.888 1.18e-05 ***
mydata$College   0.9801     0.1502   6.525 3.94e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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: 3.941e-08
```

```
Anova Table (Type II tests)

Response: mydata$HouseholdIncome
                Sum Sq Df F value    Pr(>F)
mydata$College 1739.4  1  42.572 3.941e-08 ***
Residuals      1961.1 48
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] 0.4700349
```

You want to start with the College variable because it is the variable with the highest Pearson Correlation value to Household Income.

```
Call:
lm(formula = mydata$HouseholdIncome ~ mydata$College)

Coefficients:
   (Intercept)  mydata$College
       23.0664          0.9801
```

Equation: y = 23.0664 + 0.9801 * x

23.0664 is the initial amount for y when the x variable equals zero.

0.9801 is the additional amount added for every unit change in the x variable.

In this case, the x variable represents the percentage of people who have gone to college, and y represents the household income of that person, or at least an estimate.

By hand:

```{r}
slope=cor(mydata$College,mydata$HouseholdIncome) *(sd(mydata$HouseholdIncome)/sd(mydata$College))
print(paste("slope:",slope))
intercept=mean(mydata$HouseholdIncome)-slope*mean(mydata$College)
print(paste("intercept:",intercept))
```

```
[1] "slope: 0.980144089210707"
[1] "intercept: 23.0664377296339"
```

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:

```
#mydata$predictedY = intercept + (slope * mydata$College)
mydata$Y_hat=predict(model1)
mydata$residual=mydata$HouseholdIncome-mydata$Y_hat
```

- Square each of the residuals and then add them up. This is called sum of squared residuals, or sums of squared errors.

```{r}
sse=sum((mydata$residual)^2)
print(paste('sse:',sse))
```

```
[1] "sse: 1961.12951169001"
```

- 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.

```{r}
sst=sum((mydata$HouseholdIncome-mean(mydata$HouseholdIncome))^2)
print(paste('sst:',sst))
```

```
[1] "sst: 3700.48829208"
```

- 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.

```{r}
ssr=sum((mydata$Y_hat-mean(mydata$HouseholdIncome))^2)
print(paste('ssr:',ssr))
```

```
[1] "ssr: 1739.35878038999"
```

- Calculate a statistic that is: (Sum of Squares due to Regression) / (Sum of squares Total)

```{r}
ssr/sst
```

```
[1] 0.4700349
```

- 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.

```
Anova Table (Type II tests)

Response: mydata$HouseholdIncome
               Sum Sq Df F value    Pr(>F)
mydata$College 1739.4  1  42.572 3.941e-08 ***
Residuals      1961.1 48
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
[1] 0.4700349
```

```
Call:
lm(formula = mydata$HouseholdIncome ~ mydata$College)

Residuals:
   Min     1Q Median     3Q    Max
-7.319 -4.245 -2.203  2.652 23.484

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     23.0664     4.7187   4.888 1.18e-05 ***
mydata$College   0.9801     0.1502   6.525 3.94e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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: 3.941e-08
```

R^2, Sum of squared residuals, and sum of squared errors match

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:

```
model2=lm(mydata$HouseholdIncome~mydata$College + mydata$Insured)
model2
```

```
Call:
lm(formula = mydata$HouseholdIncome ~ mydata$College + mydata$Insured)

Coefficients:
   (Intercept)  mydata$College  mydata$Insured
        9.6728          0.8411          0.2206
```

Equation: y = 9.6728 + 0.8411*x1 + 0.2206 * x2

9.6728 is the initial amount for y when the x variables equal zero; y-intercept

0.8411 is the additional amount added for every unit change in the x1 variable (College).

0.2206 is the additional amount added for every unit change in the x2 variable (Insured).

In this case, the x1 variable represents the percentage of people who have gone to college, the x2 variable represents the percentage of people who have been insured and y would represent the household income of that person, or at least an estimate.

Summary:

```
Call:
lm(formula = mydata$HouseholdIncome ~ mydata$College + mydata$Insured)

Residuals:
   Min     1Q Median     3Q    Max
-6.918 -4.545 -2.125  4.357 22.709

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.6728    14.8628   0.651 0.518339
mydata$College   0.8411     0.2098   4.010 0.000216 ***
mydata$Insured   0.2206     0.2321   0.950 0.346759
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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: 2.116e-07
```

R^2=0.48

The coefficient for College has decreased from 0.9801 to 0.8411. This is due to the impact the other variable, whether or not the person is insured, now has an impact on the household income value as well. The y-intercept value has significantly decreased from 23.0664 to 9.6728. This is the value if both variables x1 and x2 were zero.

Difference in R^2 from model 1 to model 2:

```
```{r}
summary(model1)$r.squared-summary(model2)$r.squared
```

[1] -0.009993449
```

For this problem, it may not be so beneficial to use hypothesis testing to determine whether or not the additional explanatory value should be retained or not. This is because the difference in the r^2 values is very little, however, the question being asked is the combined impact of variables on the household income, and we need to find the impact of the variables collectively to see how the various x variables impact the Y.

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?

```
print(paste("College r^2:",summary(model1)$r.squared))
print(paste("College + Insured r^2:",summary(model2)$r.squared))
model3=lm(mydata$HouseholdIncome~mydata$College + mydata$Insured+mydata$HighSchool)
print(paste("College + Insured + HighSchool r^2:",summary(model3)$r.squared))
model4=lm(mydata$HouseholdIncome~mydata$College + mydata$Insured+mydata$HighSchool + mydata$Smokers)
print(paste("College + Insured + HighSchool + Smokers r^2:",summary(model4)$r.squared))
model5=lm(mydata$HouseholdIncome~mydata$College + mydata$Insured+mydata$HighSchool+mydata$Smokers +mydata$PhysicalActivity)
print(paste("College + Insured + HighSchool + Smokers + Physical Activity r^2:",summary(model5)$r.squared))
model6=lm(mydata$HouseholdIncome~mydata$College + mydata$Insured+mydata$HighSchool+mydata$Smokers +mydata$PhysicalActivity + mydata$Obese)
print(paste("College + Insured + HighSchool + Smokers + Physical Activity + Obese r^2:",summary(model6)$r.squared))
model7=lm(mydata$HouseholdIncome~mydata$College + mydata$Insured+mydata$HighSchool+mydata$Smokers +mydata$PhysicalActivity + mydata$Obese + mydata$NonWhite)
print(paste("College + Insured + HighSchool + Smokers + Physical Activity + Obese + NonWhite r^2:",summary(model7)$r.squared))
model8=lm(mydata$HouseholdIncome~mydata$College + mydata$Insured+mydata$HighSchool+ mydata$Smokers +mydata$PhysicalActivity + mydata$Obese + mydata$NonWhite +mydata$HeavyDrinkers)
print(paste("College + Insured + HighSchool + Smokers + Physical Activity + Obese + NonWhite + HeavyDrinkers r^2:",summary(model8)$r.squared))
model9=lm(mydata$HouseholdIncome~mydata$College + mydata$Insured+mydata$HighSchool+ mydata$Smokers +mydata$PhysicalActivity + mydata$Obese + mydata$NonWhite +mydata$HeavyDrinkers + mydata$TwoParents)
print(paste("College + Insured + HighSchool + Smokers + Physical Activity + Obese + NonWhite + HeavyDrinkers + TwoParents r^2:",summary(model9)$r.squared))
```

```
[1] "College r^2: 0.470034936771093"
[1] "College + Insured r^2: 0.480028386255677"
[1] "College + Insured + HighSchool r^2: 0.484353917327217"
[1] "College + Insured + HighSchool + Smokers r^2: 0.61753877304019"
[1] "College + Insured + HighSchool + Smokers + Physical Activity r^2: 0.618367977147063"
[1] "College + Insured + HighSchool + Smokers + Physical Activity + Obese r^2: 0.627974181667562"
[1] "College + Insured + HighSchool + Smokers + Physical Activity + Obese + NonWhite r^2: 0.711209961087337"
[1] "College + Insured + HighSchool + Smokers + Physical Activity + Obese + NonWhite + HeavyDrinkers r^2: 0.711452250776483"
[1] "College + Insured + HighSchool + Smokers + Physical Activity + Obese + NonWhite + HeavyDrinkers + TwoParents r^2: 0.73547778661442"
```

The variables that should be retained for the best predictive model are College, Non-White, and perhaps Two Parents, because they have very low P values. Although, the p value for Two Parents is about 0.06, so could be omitted, but it is still quite close to 0.05. In my opinion it should be left in the model to see how it impacts Household Income.

```
Call:
lm(formula = mydata$HouseholdIncome ~ mydata$College + mydata$Insured +
    mydata$HighSchool + mydata$Smokers + mydata$PhysicalActivity +
    mydata$Obese + mydata$NonWhite + mydata$HeavyDrinkers + mydata$TwoParents)

Residuals:
   Min     1Q Median     3Q    Max
-7.541 -2.543 -1.260  1.515 15.204

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -15.52228   33.84632  -0.459 0.648996
mydata$College         0.61379    0.19794   3.101 0.003528 **
mydata$Insured         0.02526    0.25319   0.100 0.921014
mydata$HighSchool      0.22500    0.52624   0.428 0.671257
mydata$Smokers        -0.26301    0.42024  -0.626 0.534959
mydata$PhysicalActivity  -0.02829    0.25515  -0.111 0.912257
mydata$Obese           -0.27036    0.51896  -0.521 0.605257
mydata$NonWhite         0.27281    0.06866   3.973 0.000288 ***
mydata$HeavyDrinkers    0.52234    0.84689   0.617 0.540883
mydata$TwoParents       0.50137    0.26304   1.906 0.063847 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.947 on 40 degrees of freedom
Multiple R-squared:  0.7355,    Adjusted R-squared:  0.676
F-statistic: 12.36 on 9 and 40 DF,  p-value: 4.541e-09
```

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.

Two variables that stood out to me as being counter intuitive are the variables Obese and the variable Physical Activity. If one is physically active, they are less likely to be obese. I believe one should be omitted as it is redundant in the model. One could also argue that High School and College are somewhat redundant, as to go to college one must have had completed High School.

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.

Model after omitting the Obese and High School Variables:

```
Call:
lm(formula = mydata$HouseholdIncome ~ mydata$College + mydata$Smokers +
    mydata$PhysicalActivity + mydata$NonWhite + mydata$HeavyDrinkers +
    mydata$TwoParents)

Residuals:
   Min     1Q Median     3Q    Max
-6.554 -2.464 -1.102  1.276 15.335

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -16.32814   18.49351  -0.883  0.38219
mydata$College         0.70356    0.13359   5.266 4.21e-06 ***
mydata$Smokers        -0.25876    0.29472  -0.878  0.38483
mydata$PhysicalActivity  0.06416    0.20012   0.321  0.75005
mydata$NonWhite        0.28142    0.06483   4.341 8.47e-05 ***
mydata$HeavyDrinkers   0.69099    0.79124   0.873  0.38735
mydata$TwoParents      0.59904    0.18448   3.247  0.00226 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.81 on 43 degrees of freedom
Multiple R-squared:  0.7312,    Adjusted R-squared:  0.6937
F-statistic:  19.5 on 6 and 43 DF,  p-value: 7.888e-11
```

I believe the variables Obese and High School are already accounted for in the variables Physical Activity and College respectively. Adding those variables created some redundancy in the data. Omitting those variables could allow for a better analysis. The R^2 value has also barely changed, resulting in 0.7312. This is a very small change compared to having all of the variables within, which had an R^2 value of 0.735. If this was a bigger dataset, adding redundant variables could result in bias in a model. Hence, removing the variables even though it decreases the accuracy of this model slightly, may be beneficial in other models which may cause the model to overestimate some features.

9. 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.

   Throughout this assignment, I learned how various variable impact household income within this dataset and analyzed some correlations within the data. However, correlation does not mean causation. Some variable impacted the model more than others, as indicated through the P-values, found within the summaries of the models. This assignment allowed me to become more comfortable working with linear regression models within R Studio. Although I had some experience in the past through MSDS 401, this assignment game me a refresh on how to use the different functions within the software.

   This assignment also briefed me on the importance of preprocessing your data in order to create a more accurate model. I felt that having both obesity and physical activity as variables in the model and both College and High School as variables in the model, may cause some overestimation within the linear regression model. With the little change in the R^2 value, it shows that if such variables existed in bigger datasets, it could cause some form of over estimation of some variables, leading to less accurate models.

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.