**Dataset Description**

The data NorthValleyRealtor.xlsx contain information on homes sold by the North Valley Real Estate group within the last year. Within this file you will find the following fields:

* Record - Property identification number
* Agent – Name of the real estate agent assigned to the property
* Price – Market price in US dollars
* Size – Livable square feet of the property
* Bedrooms – Number of bedrooms
* Baths – Number of baths, which takes numbers 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5
* Pool – Does this home have a pool? (1 = yes, 0 = no)
* Garage – Does the home have a detached garage? (1 = yes, 0 = no)
* Days – Number of days the property was on the market
* Township – Area where the property is located
* Mortgage type – Fixed or adjustable. The fixed mortgage is a 30 year, fixed interest rate loan. The adjustable rate loan begins with an introductory rate of 3% for the first five years, then the interest rate is based on the current interest rates plus 1% (i.e., the interest rate AND the payment is likely to change each year after the 5th year.).
* Years – The number of years that the mortgage loan has been paid
* FICO – the credit score of the mortgage loan holder. The highest score is 850; an average score is 680; a low score is below 680. The score reflects a person’s ability to pay their debts.
* Default – Is the mortgage loan in default? (1 = yes, 0 = no)

**Correlation and Simple Linear Regression (35pts)**

1. If we perform a test to investigate the correlation between Price and Size. State the null hypothesis and alternate hypothesis for this research question.

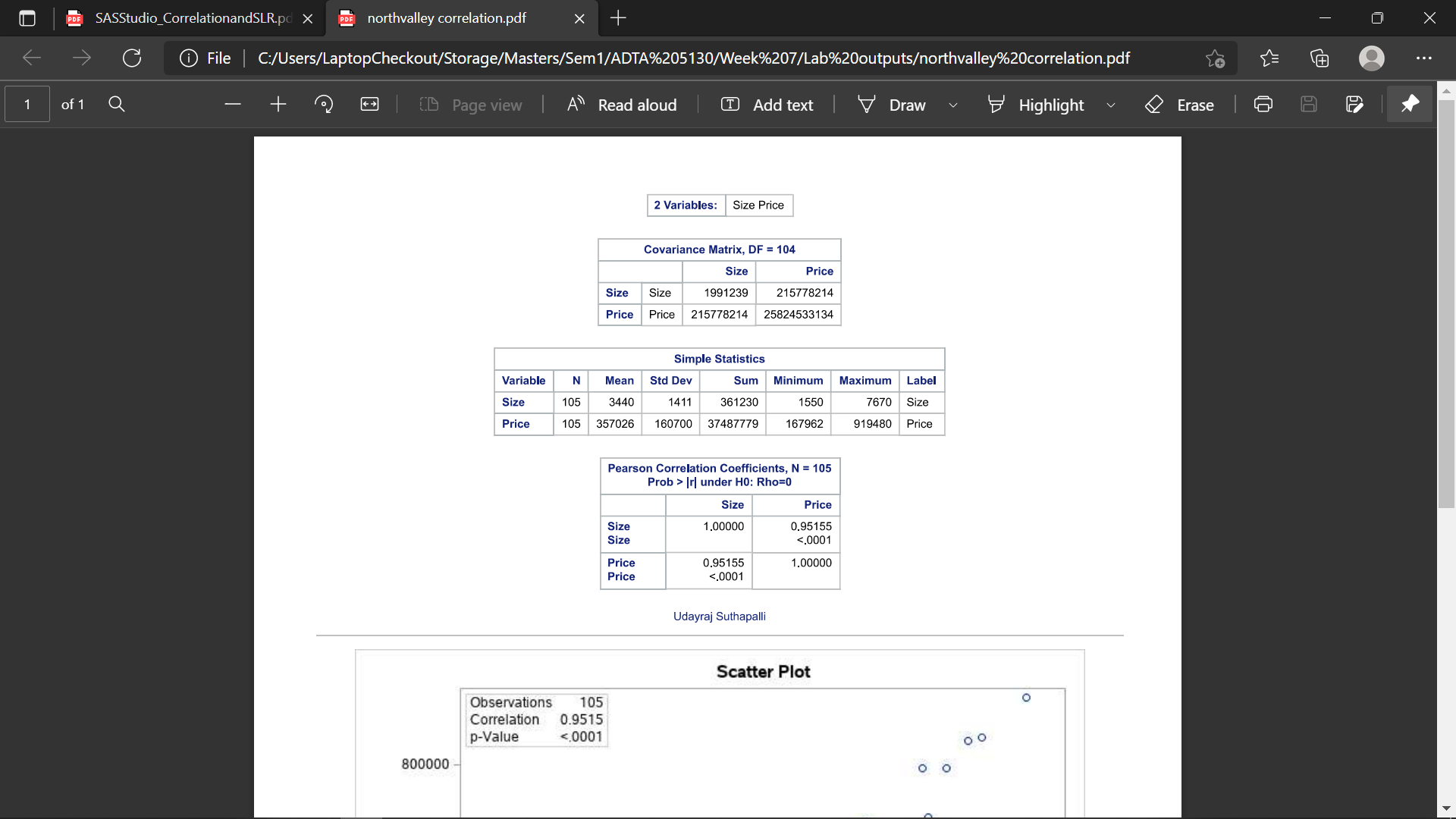
H0: - There is a 0 correlation between the variables price and size.

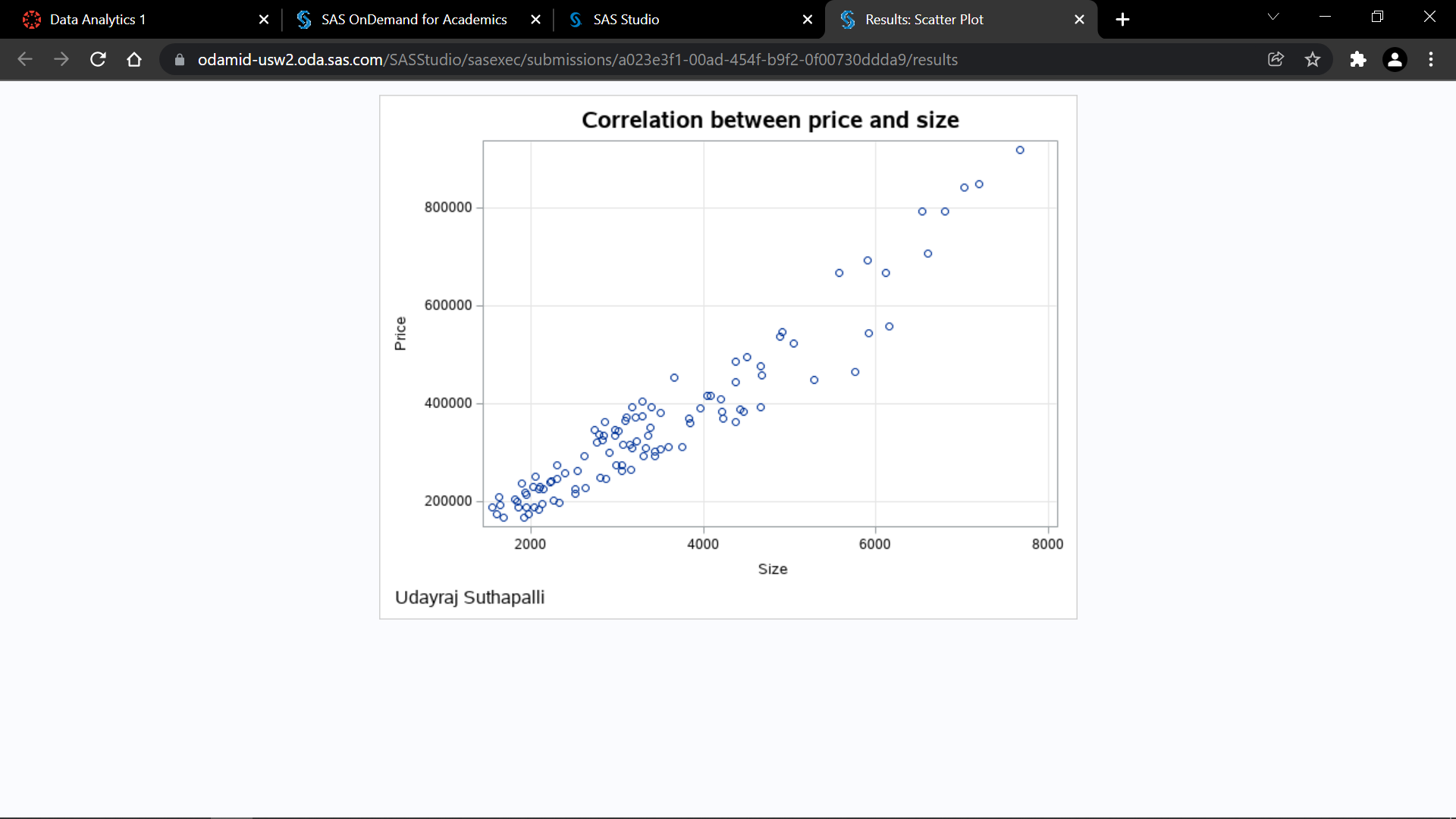
HA: - The correlation between the variables price and size of the dataset is not 0.

1. What is the value of correlation and its p-value? Support your answer with an SAS output. **Make sure to add your first and last name in the footnote**.

The correlation value between size and price is 0.95155 with p-value <0.001. From those we can conclude that there is a good positive relationship between size and price with a p-value of <0.001.

We can reject our null hypothesis from this outputs.





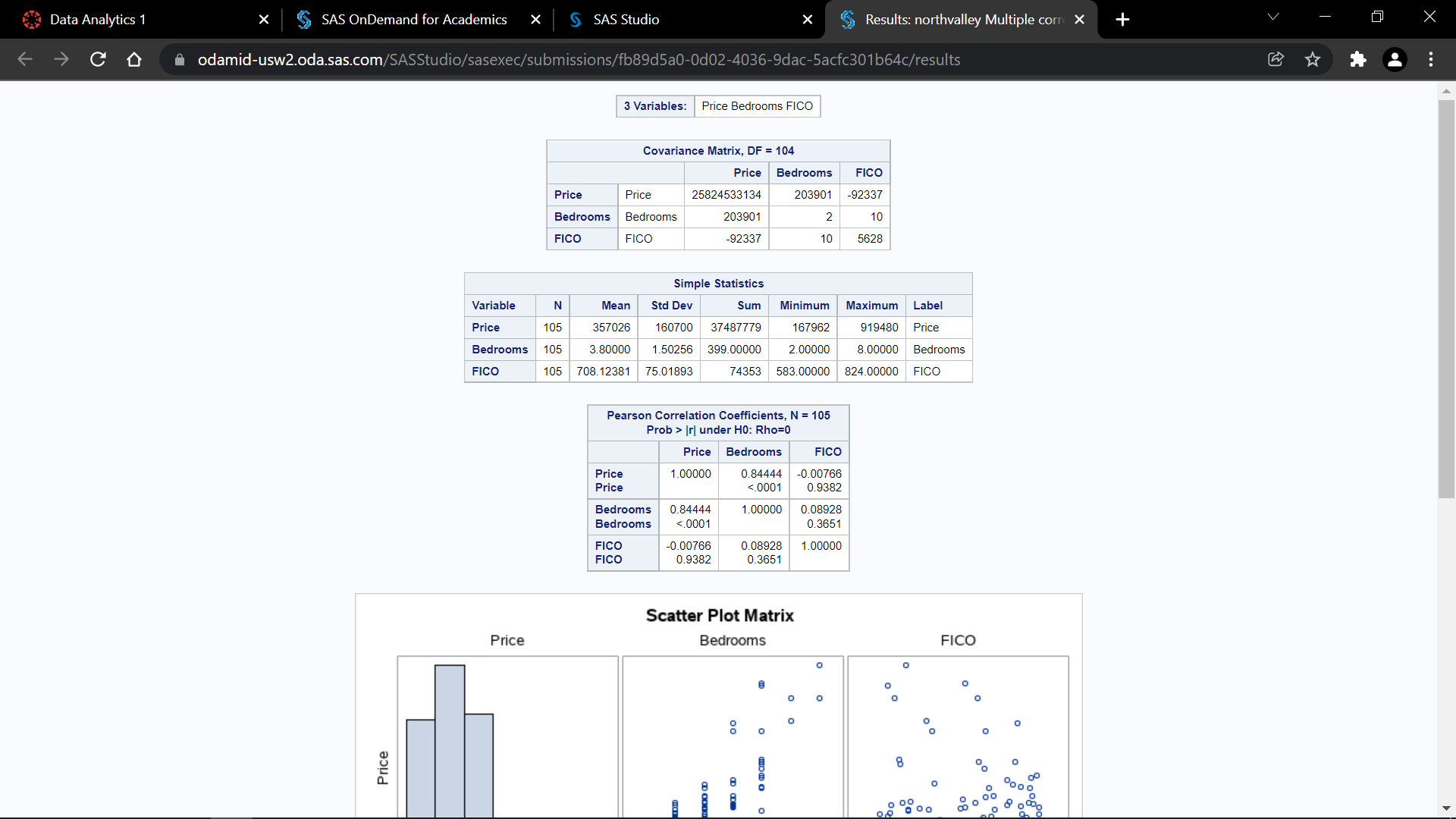
1. Create a pairwise correlation matrix for variables Price, Bedrooms, and FICO. Which pair has significant linear relationship and so we can reject null hypothesis of zero relationship between them?

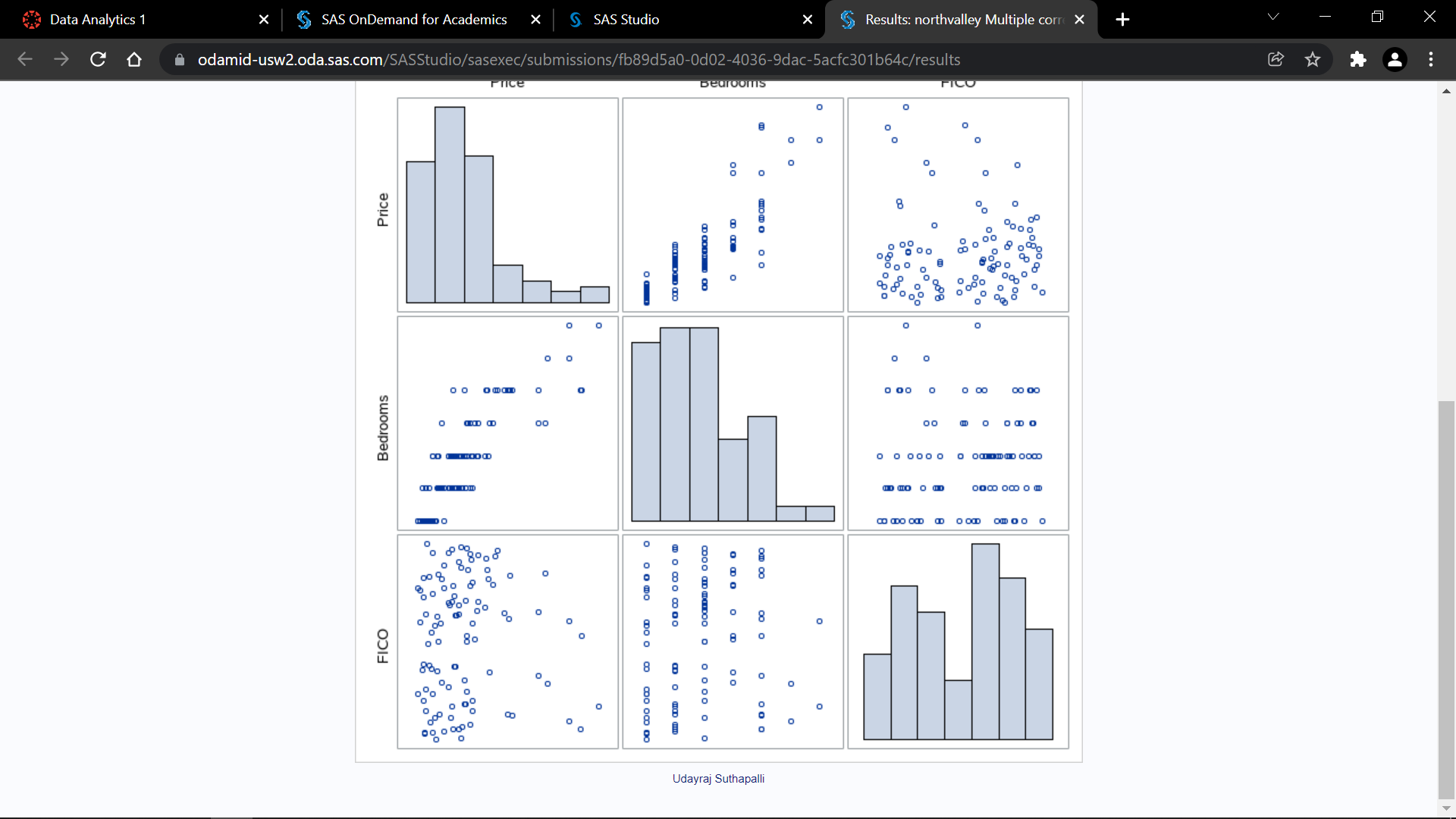
The relationship between price and bedroom is very good having correlation value of 0.8444 with a p-value <0.0001.

The relationship between price and FICO is negatively low with a correlation value of -0.00766 and p-value of 0.9382.

The relationship between bedroom and FICO is also low but positively low with a correlation value of 0.08928 with a p-value of 0.3651.

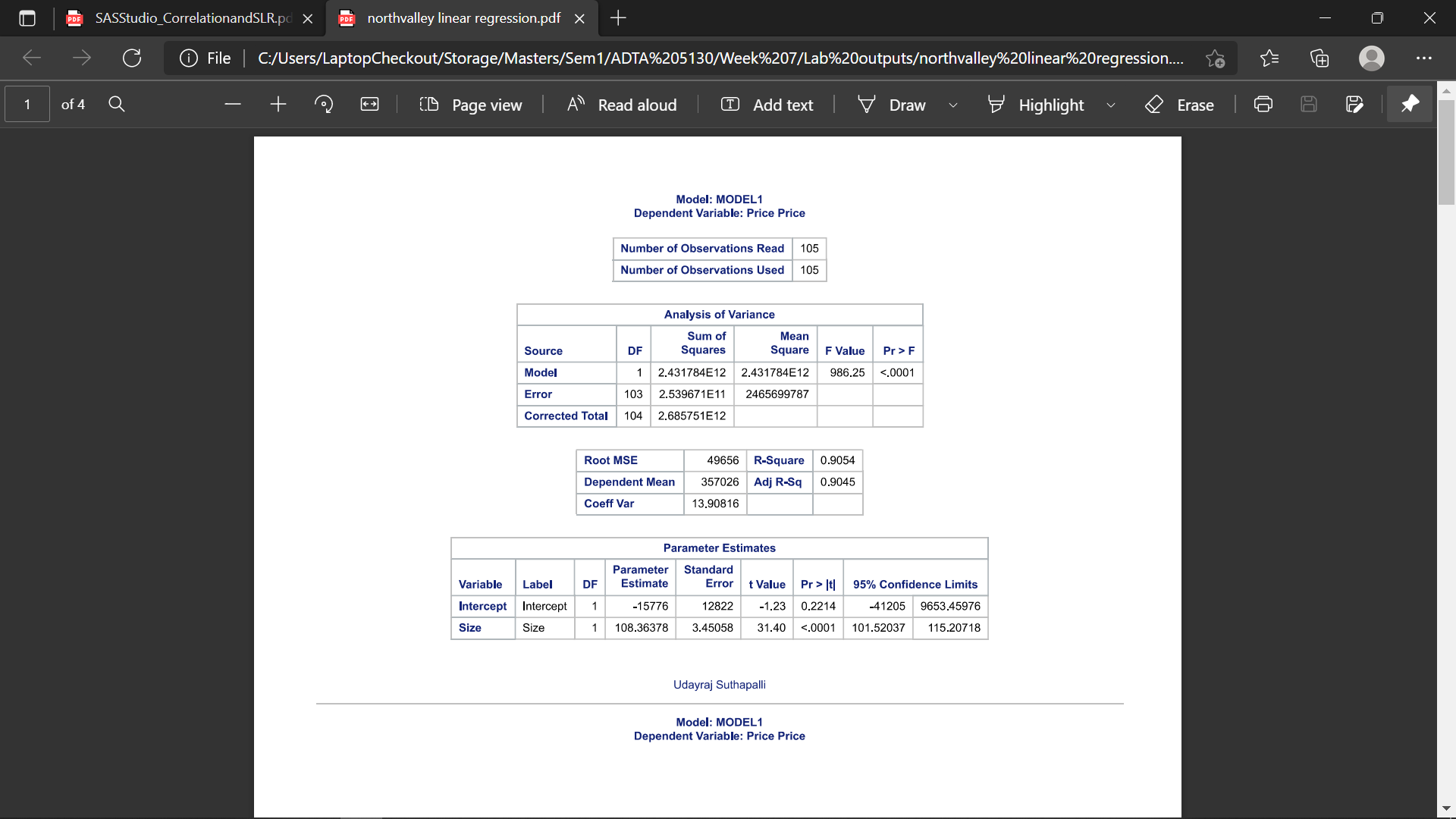
So, we cannot reject null hypothesis for variables bedrooms and FIC0, price and FICO. We can reject null hypothesis for variables bedrooms and price.





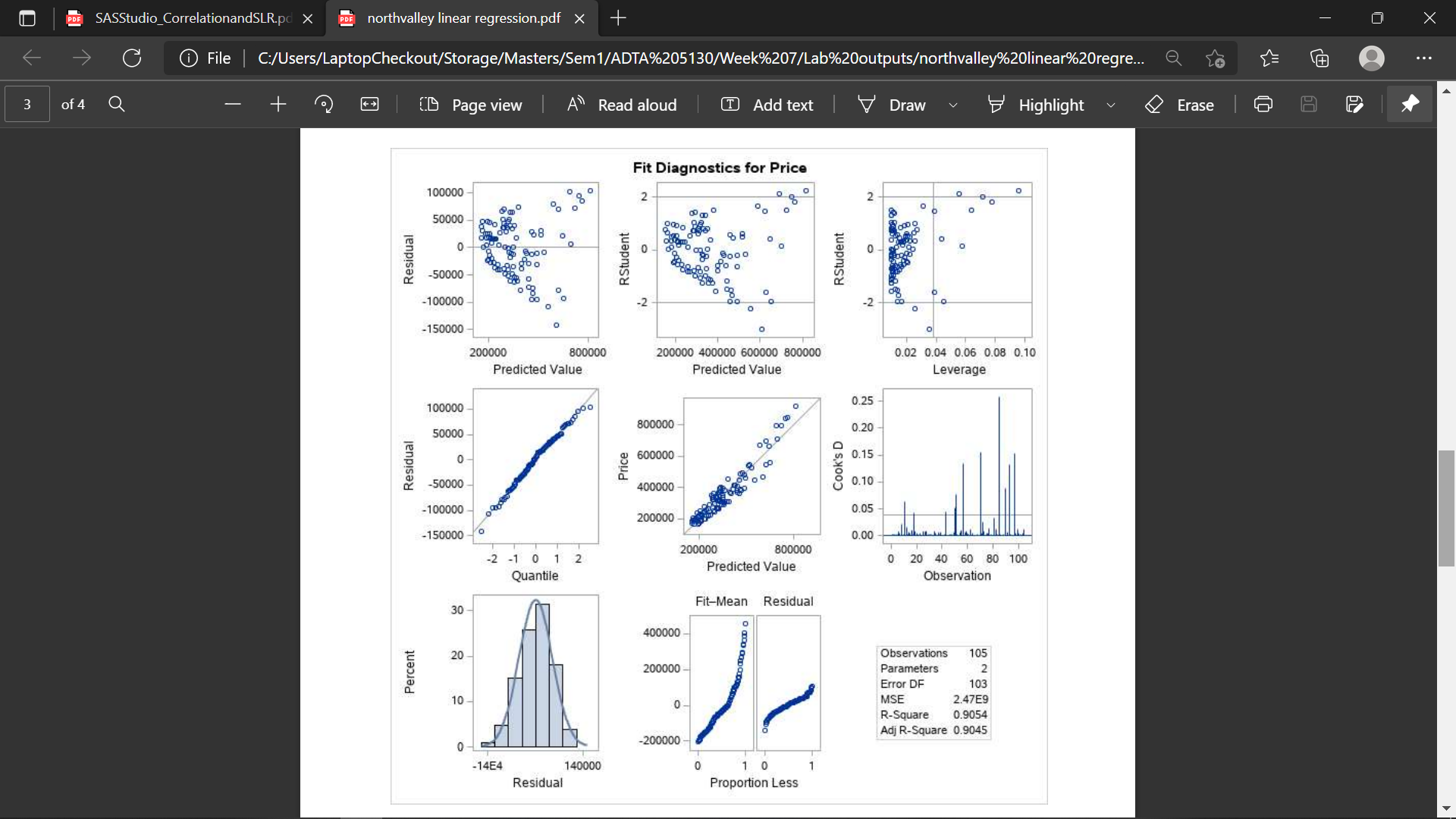
1. Build a simple linear relationship between Price and Size. What is the R-square, parameter estimate and confidence interval? Take a screenshot of the output showing R-square, parameter estimate and confidence interval. **Make sure to add your first and last name as the footnote**.

The R-square value is 0.9054, The parameter estimate is 108.36378 with a CIs of 101.5207 to 115.20718



1. Take a screenshot of the diagnostic plot (the panel of plots) of the model between Price and Size. Are the assumptions satisfied?

We can confirm the variance of residuals are constant for both predicted values and they are distributing normally with a mean of 0. The data points for the cook’s D had lot of points very much higher than the cutoff. So, the assumptions are not satisfied completely, we can improve more.



1. Because some assumptions are not satisfied, we decide to transform the dependent variable Price. Under Tasks, expand Data, select Transform Data. In the opened Transform Data window, under Data tab, select data, and choose ‘Square root’ transform on the Price variable as pictures shown below. Save transformed results to the default library and data. Click run and you will see a column sqrt\_Price attached to the original data which is the square root transformed Price.

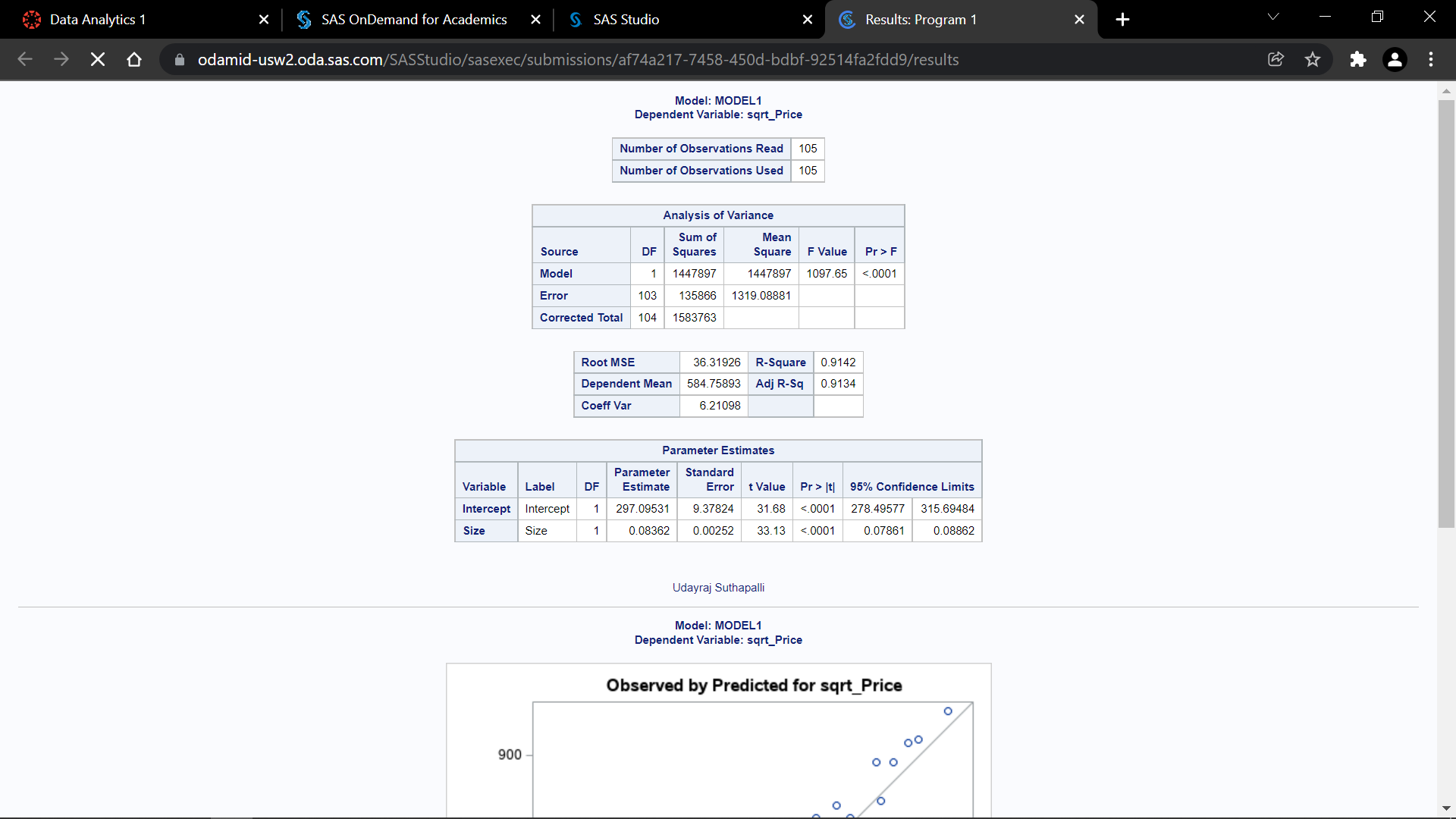
Graphical user interface, application

Description automatically generated Graphical user interface, text, application

Description automatically generated

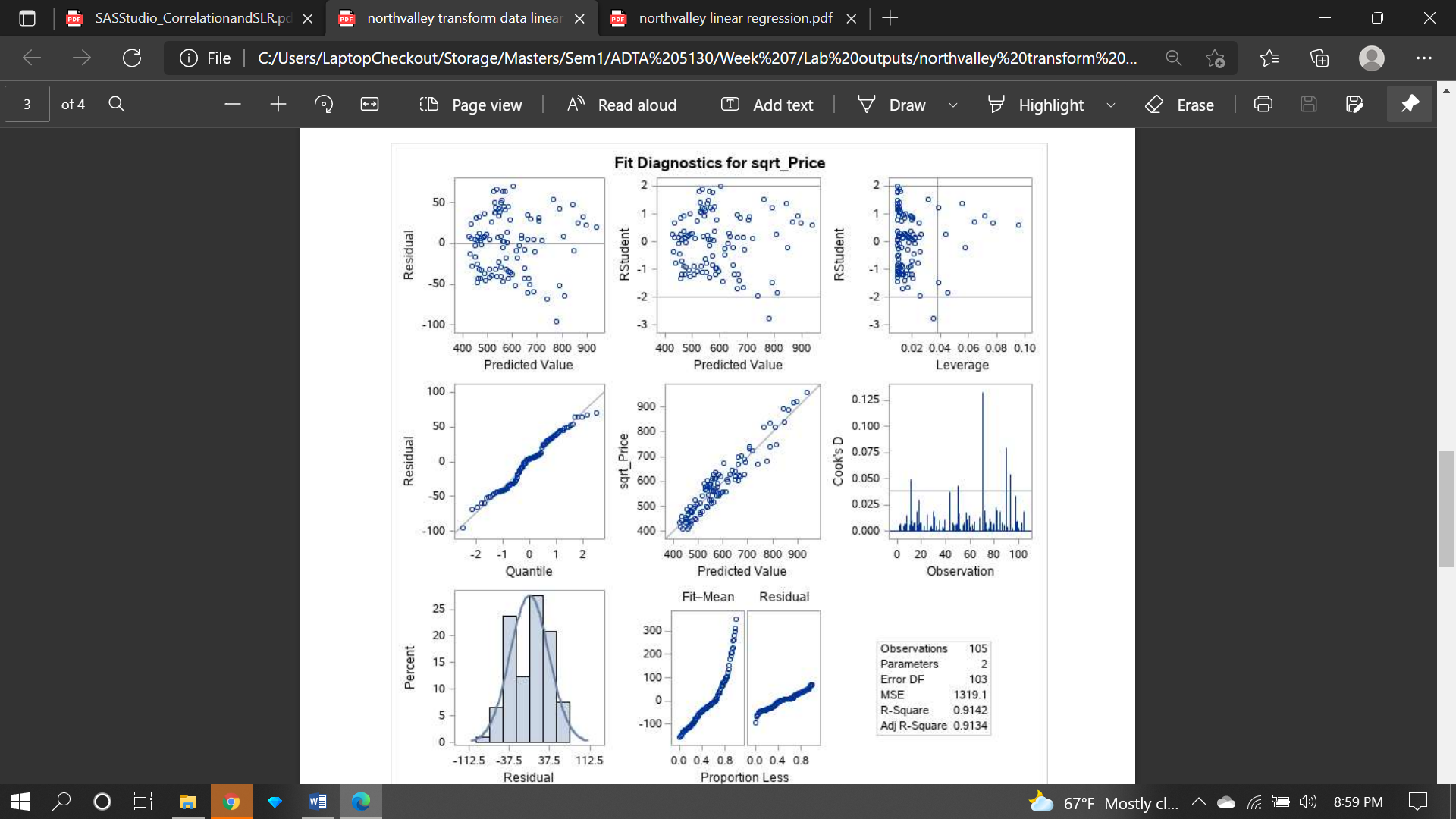
Now you can build a model between sqrt\_Price and size. What is the R-square, parameter estimate and confidence interval? Take a screenshot of the output showing R-square, parameter estimate and confidence interval. **Make sure to add your first and last name as the footnote.**

The R-square value is 0.9142, The parameter estimate is 0.08362 with a CIs of 0.07861 to 0.08862



1. Take a screenshot of the improved diagnostic plot (the panel of plots) of the model between sqrt\_Price and Size. Are the assumptions better satisfied?

The assumptions are satisfied better than using the normal price. we can observe the change of regression graph between sqrt\_price and predicted value. The sqrt\_price is much closer to the regression line with the predicted values. And also, we have low number of cook-d values that crosses the cutoff mark.



**Multiple Linear Regression (20 pts)**

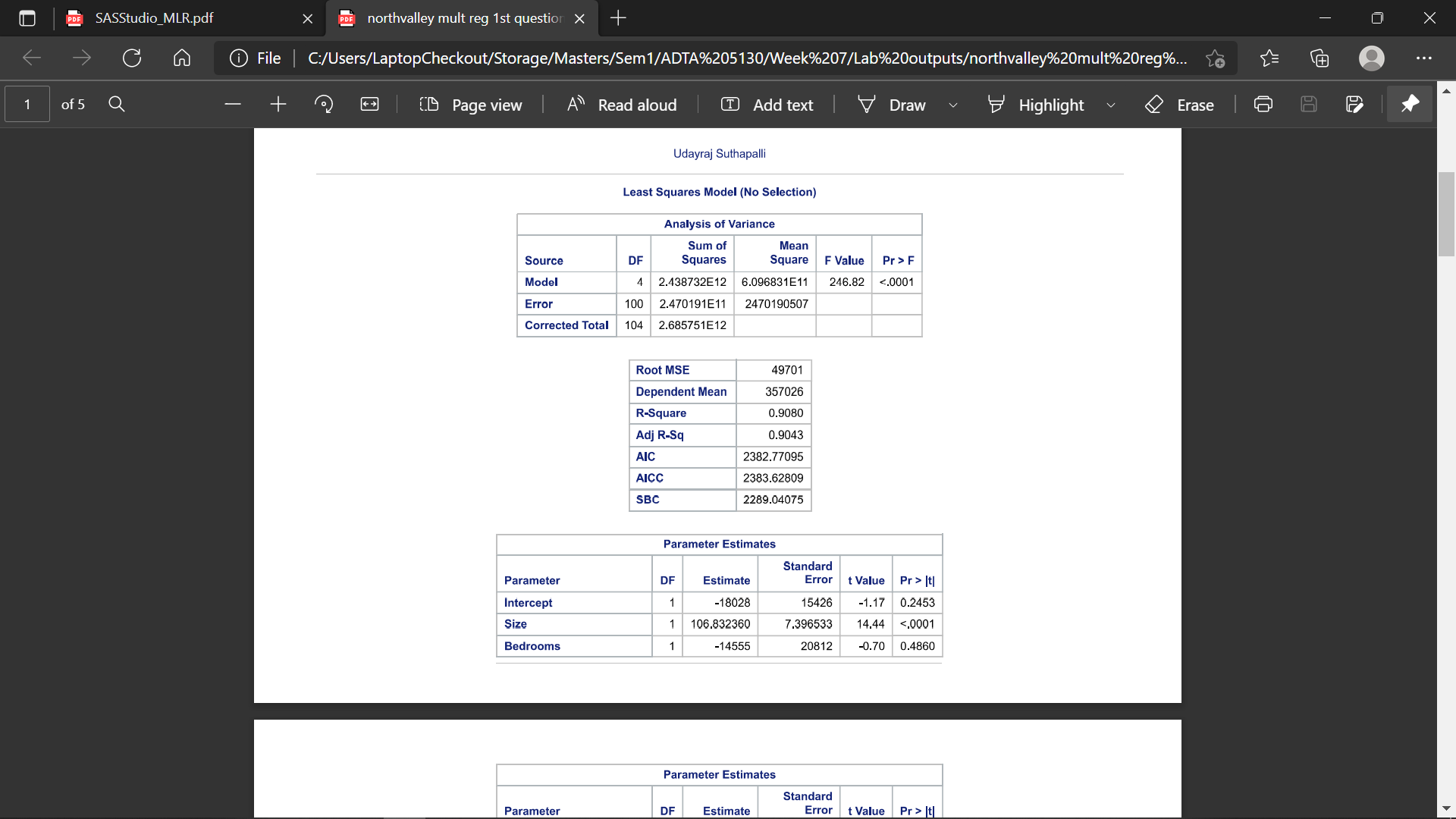
1. Build a model using Price as dependent variable, size, bedrooms, Baths, and mortgage type as predictors with NO selection method. Is the overall model significant? What is the degree of freedom of the model and so how many parameters are estimated in the model according to the DF? What is the sample size according to the DF of corrected total? What is the R-square and adjusted R-square? Support you answers with appropriate screenshots. What is the benefit of using adjusted R-square in multiple linear regression generally?

The degree of freedom for the model is 4 excluding the intercept from the independent variables and the independent variables are size, bedrooms, baths, mortgage type adjustable.

The sample size according to corrected total is 104 which is n-1 i.e., 105-1 = 104

The value of R-square for the model build is 0.9080

The value of Adjusted R-square for the model build is 0.9043

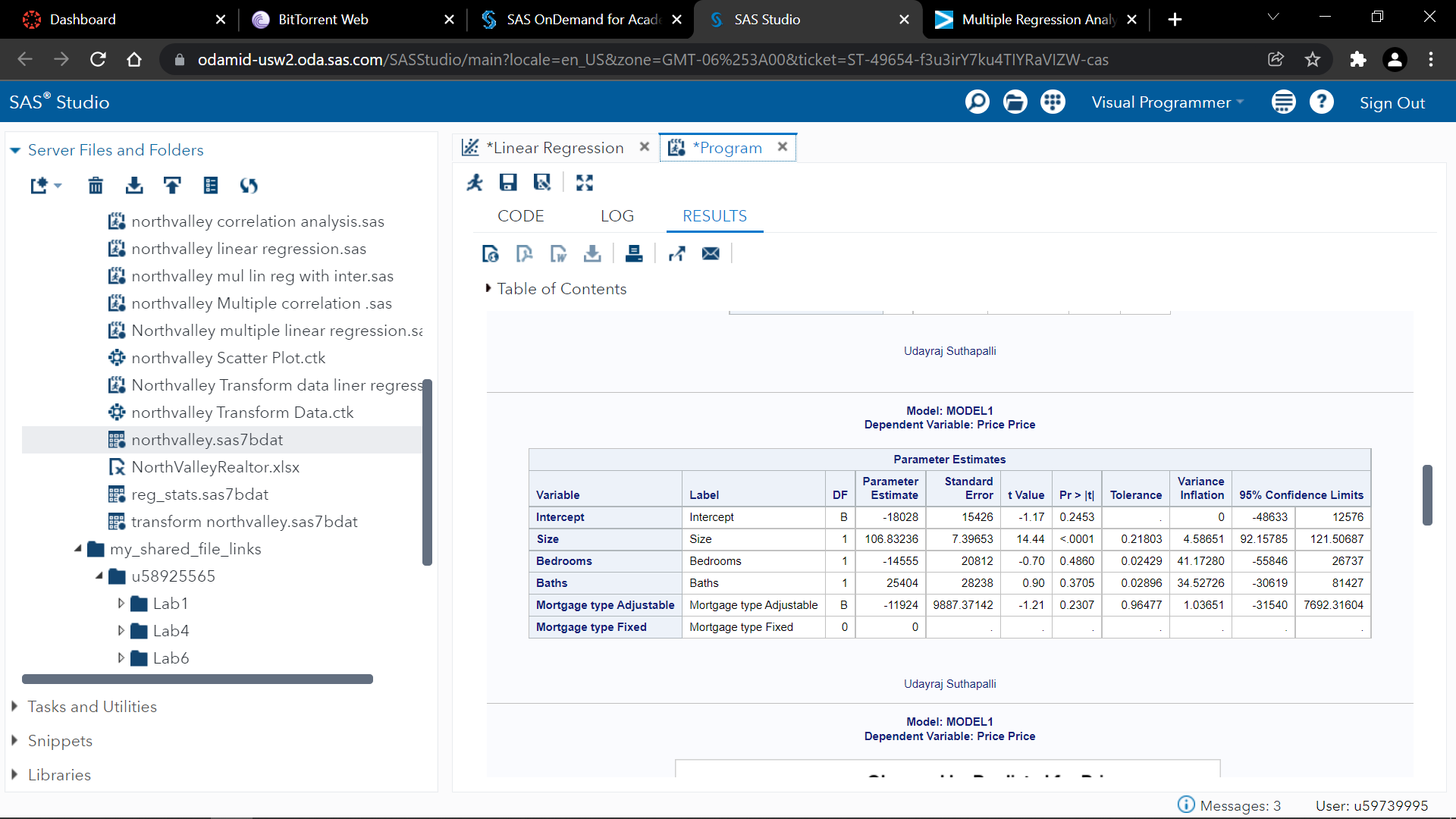


We will use the Adjusted R-square to reduce the overfitting of model to the test dataset. This adjusted R-square is much used in case of multiple linear regression as it have lot of variables, there is more chance of overfitting of model if we didn’t choose the attributes that should be used to build a model.

1. Is there any multicollinearity issue and how do you know? Support your answer with appropriate SAS output.

If there is multicollinearity, remove the predictor with highest VIF from the model. Rerun the new model. Is there still multicollinearity?

From the below image we can observe that VI s high for variables Bedrooms and Baths. This gives us enough proof that there is multicollinearity issue in the model we build. First, we need to remove the variable with the highest VI value rebuild and then check the value again.

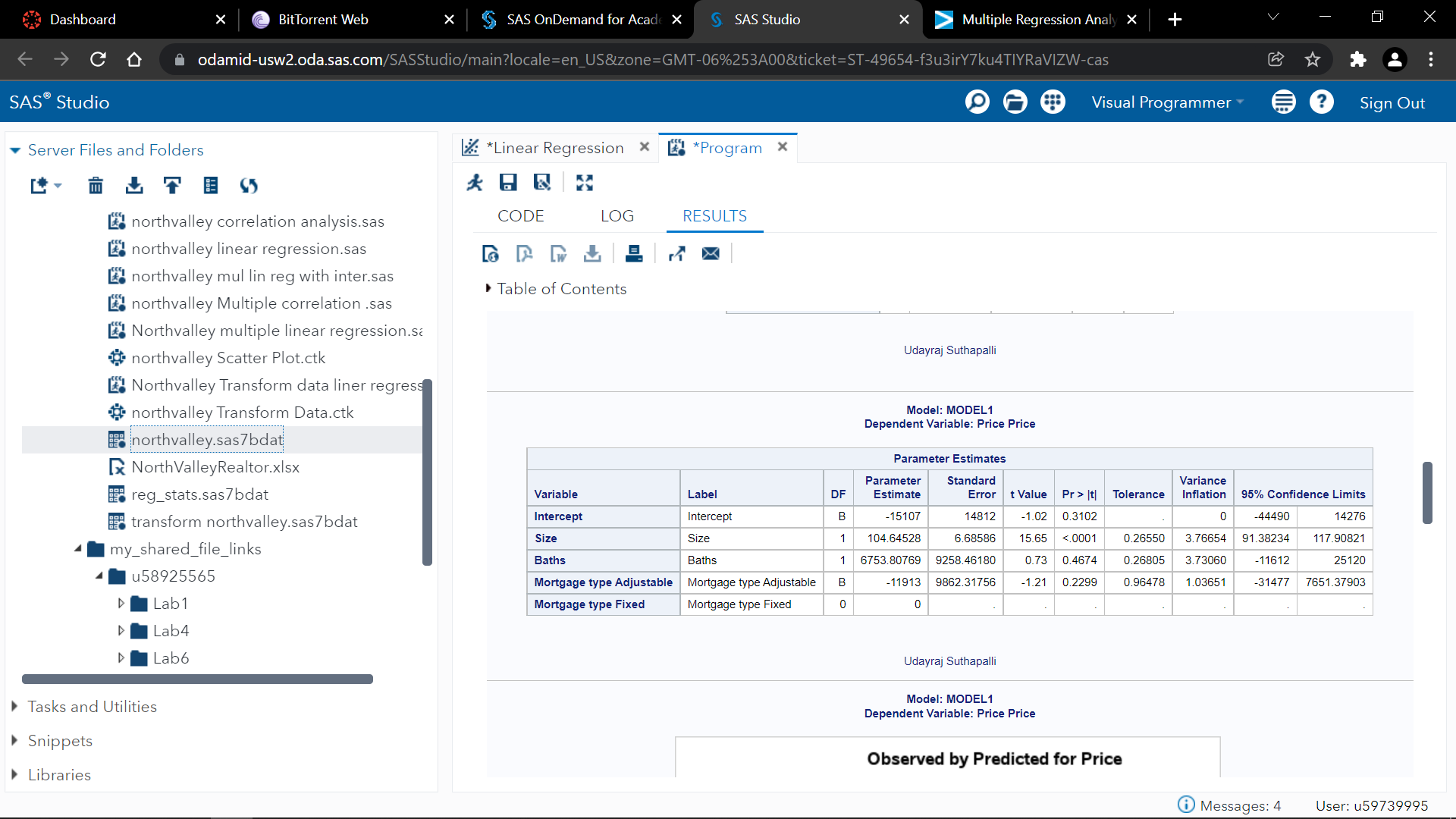


As the variable Bedrooms VI is higher that the VI of Baths we will remove the Bedrooms variable from the model and build again.

After removing the bedrooms and rebuild the model we got the values for VI and tolerance as shown below.

From the below value we can say that by removing the variable bedrooms we removed the multicollinearity from the entire model without removing the baths variable which is also highest when the bedrooms variables is there in the model.

Now the model doesn’t have any multicollinearity.



1. Build a model with Price as response, size, Baths, Mortgage type and interaction of size\*Baths as predictors with NO selection method. Write the linear model in an equation. Which variable(s) is/are significant predictor(s) at significance level of 0.05 and what is/are the confidence interval(s)? Support your answers with a screenshot. (You don’t need to output collinearity measurements VIF or Tolerance here)

The linear regression can be written as ----- **122367 + 58.37Size + (-35024Baths) + (12.963Size\*Baths)**

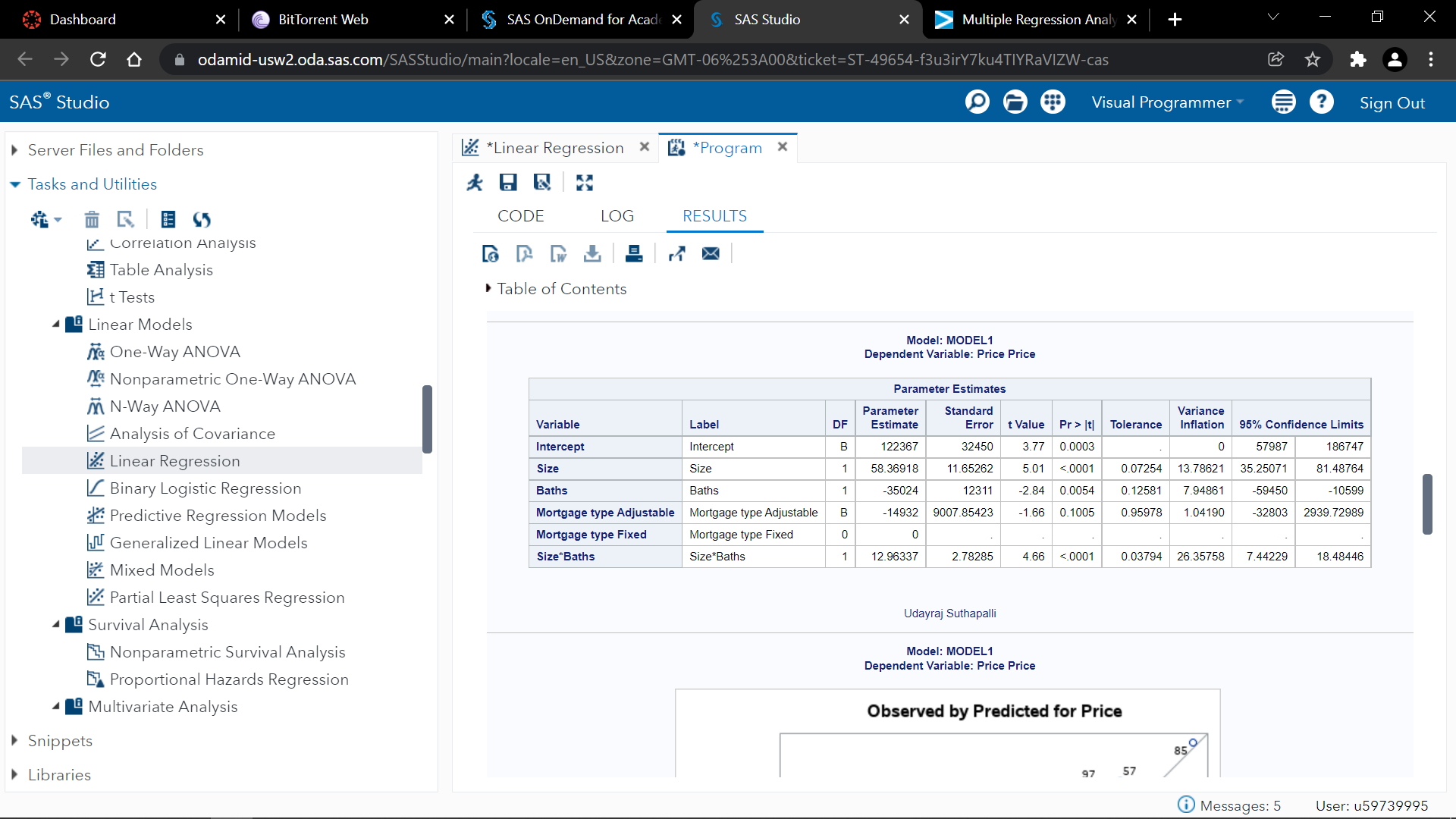
From the below image we can observe that “Mortgage type Adjustable” is not a significant variable for the model with a significance level of 0.05.

The CI for the variable Size is (35.25071, 81.48764).

The CI for the variable Baths is (-59450, -10599).

The CI for the variable Mortgage type is (-32803,2939.7289).

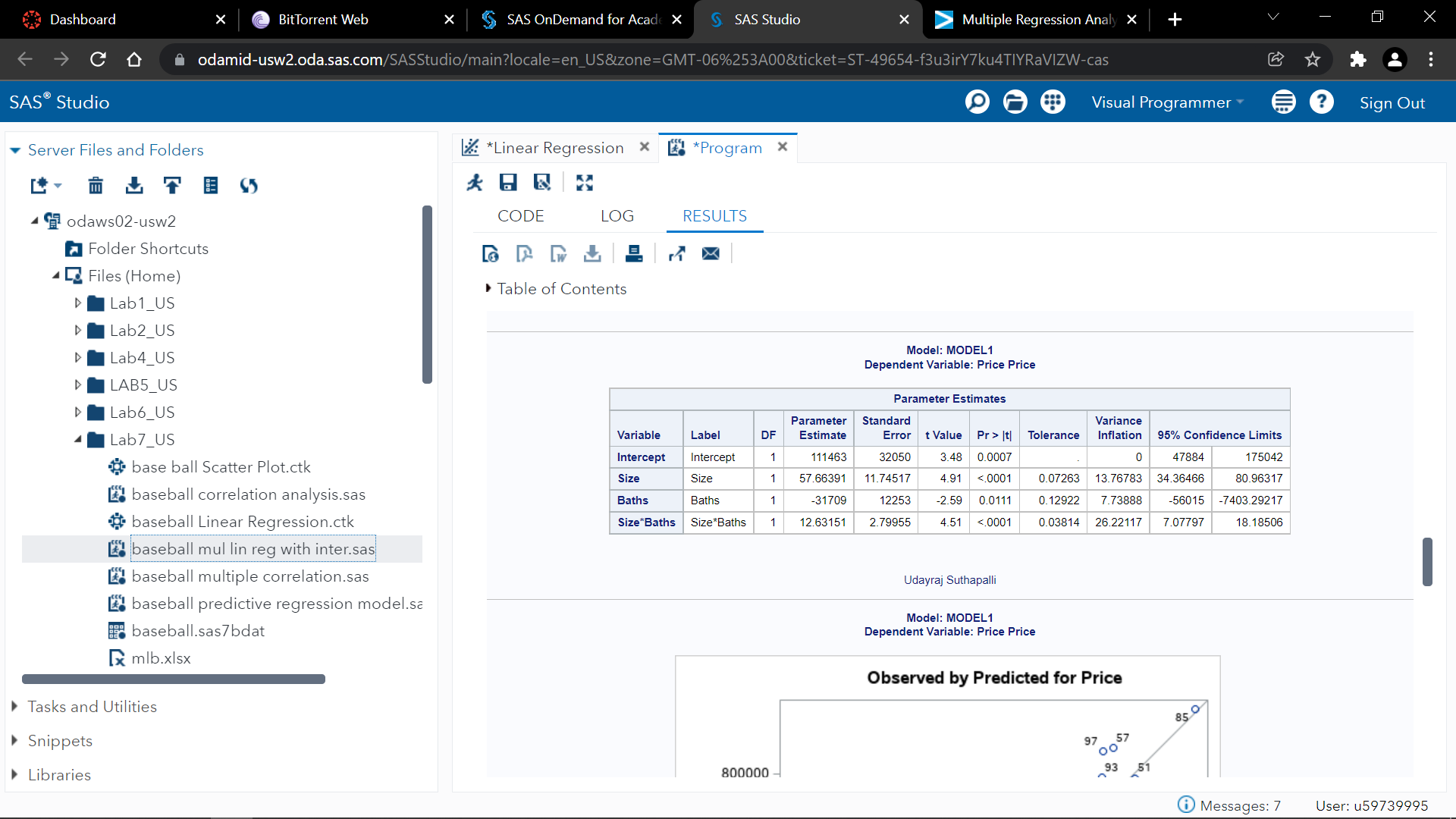
The CI for the variable Size\*Baths is (7.4429, 18.48446).



1. Use stepwise selection method, add/remove effects with significance levels at 0.05 to automatically select a final model. Write the equation with estimated values of parameters of the final model. Support your answer with the appropriate output

The equation after applying the selection to the linear regression model becomes like

**111463 + 57.664Size + (-31709\*Baths) + (12.632Size\*Baths).**



1. For the above final selected model, is any assumption violated according to the diagnostic panel of plots? Attach the diagnostic panel of plots.

The residuals are normally distributed with constant variance of residuals and mean as zero. There are some outliers according to Cook’s D plot. But there only 7 of them crossed the cutoff value 4/n which is 4/105=0.0381 out of 105 which is very low and only 3 of them are very far to the cutoff value.

All the assumptions are satisfied better than the previous one with no selection and the current model can be accepted as all the assumptions are satisfied from the diagnostic plots.

