SAS Tutorial: OLS Regression

Data Set: cigarette_consumption

Tutorial Instructions:

For this tutorial we demonstrate how to fit and interpret and OLS Regression Model and produce the default graphics using PROC REG.

(0) To begin we Set the Cigarette Consumption data set with a shorter name for reference

```
* Set the Cigarette Consumption data set;
data cig;
set mydata.cigarette_consumption;
run;
```

A description is available in our textbook and our data dictionary. For the convenience of this tutorial we will post that description and a sample of the observations here.

```
Regression Analysis By Example, 5th Edition ISBN 9780470905845
      - p. 86-89

A national insurance organization wanted to study the consumption
pattern of cigarettes in all 50 states plus the District of Columbia.
The data are from 1970 and the response variable of interest is per
capita cigarette sales.

State:      State abbreviation
Age:        median age of person living in a state
HS:         percentage of people over 25 who completed high school
Income:     per capita income in dollars
Black:      percentage of population that is black
Female:     percentage of population that is female
Price:      weighted average price (in cents) of a pack of cigarettes
Sales:      number of packs of cigarettes sold in a state per capita
```

Obs	state	age	hs	income	black	female	price	sales
1	AL	27.0	41.3	2948.0	26.2	51.7	42.7	89.8
2	AK	22.9	66.7	4644.0	3.0	45.7	41.8	121.3
3	AZ	26.3	58.1	3665.0	3.0	50.8	38.5	115.2
4	AR	29.1	39.9	2878.0	18.3	51.5	38.8	100.3
5	CA	28.1	62.6	4493.0	7.0	50.8	39.7	123.0
6	CO	26.2	63.9	3855.0	3.0	50.7	31.1	124.8

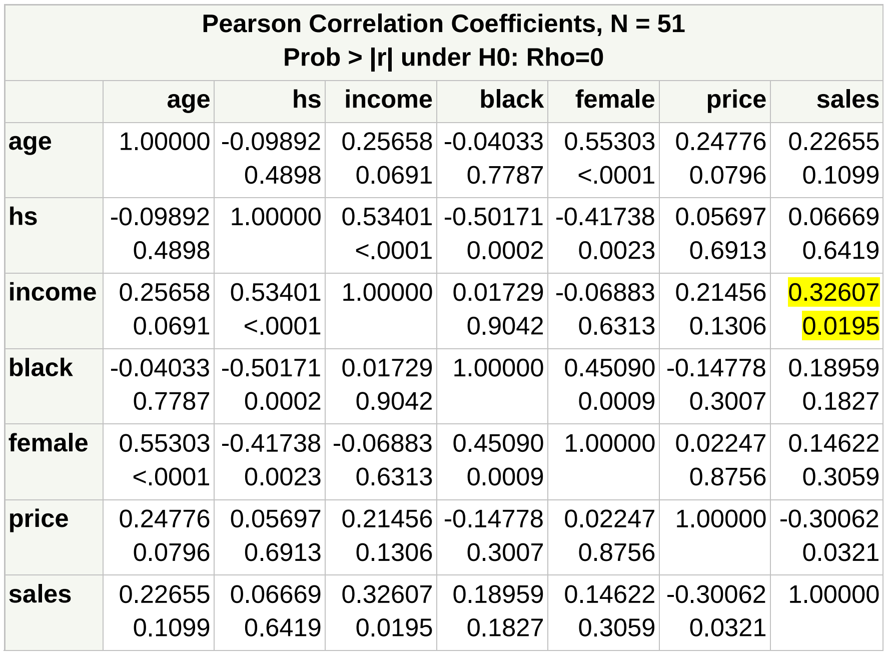
(1) We will use PROC CORR to produce the Pearson correlation coefficients for the dependent variable Sales and all other independent variables. The code also produces the scatter plot matrix for showing the relationships between all variables (see Figure 1).

```
Title "OLS Regression SAS Tutorial";
ods graphics on;
* Produce the scatterplot matrix;
Title2 "Scatterplot Matrix";
proc corr data=cig plot=matrix(histogram nvar=all);
run;
ods graphics off;
```

The output for the PROC CORR procedure is given in Table 1, Table 2, and Figure 1. The default numerical summaries produced by PROC CORR are the simple summary statistics provided in Table 1 and the estimates of the Pearson correlation coefficients between all variables in the data set. The default graphical summary produced by PROC CORR is the scatter plot matrix displayed in Figure 1.

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**
age	51	27.46667	1.87698	1401	22.90000	32.30000
hs	51	53.14902	8.00118	2711	37.80000	67.30000
income	51	3764	594.71564	191949	2626	5079
black	51	9.99216	12.64832	509.60000	0.20000	71.10000
female	51	50.95098	1.11146	2599	45.70000	53.50000
price	51	38.07451	4.12858	1942	29.00000	45.50000
sales	51	121.54118	32.07037	6199	65.50000	265.70000

Table 1: Proc Corr Output – Simple Summary Statistics



Pearson Correlation Coefficients, N = 51
Prob > |r| under H0: Rho=0

	age	hs	income	black	female	price	sales
age	1.00000	-0.09892 0.4898	0.25658 0.0691	-0.04033 0.7787	0.55303 <.0001	0.24776 0.0796	0.22655 0.1099
hs	-0.09892 0.4898	1.00000	0.53401 <.0001	-0.50171 0.0002	-0.41738 0.0023	0.05697 0.6913	0.06669 0.6419
income	0.25658 0.0691	0.53401 <.0001	1.00000	0.01729 0.9042	-0.06883 0.6313	0.21456 0.1306	0.32607 0.0195
black	-0.04033 0.7787	-0.50171 0.0002	0.01729 0.9042	1.00000	0.45090 0.0009	-0.14778 0.3007	0.18959 0.1827
female	0.55303 <.0001	-0.41738 0.0023	-0.06883 0.6313	0.45090 0.0009	1.00000	0.02247 0.8756	0.14622 0.3059
price	0.24776 0.0796	0.05697 0.6913	0.21456 0.1306	-0.14778 0.3007	0.02247 0.8756	1.00000	-0.30062 0.0321
sales	0.22655 0.1099	0.06669 0.6419	0.32607 0.0195	0.18959 0.1827	0.14622 0.3059	-0.30062 0.0321	1.00000

Table 2: Proc Corr Output – Pearson Correlation Coefficients

We can see that the highest correlation coefficient estimate with our response variable Sales is with the predictor variable Income. We will use this variable to begin creating a simple regression model in the next step of the tutorial.

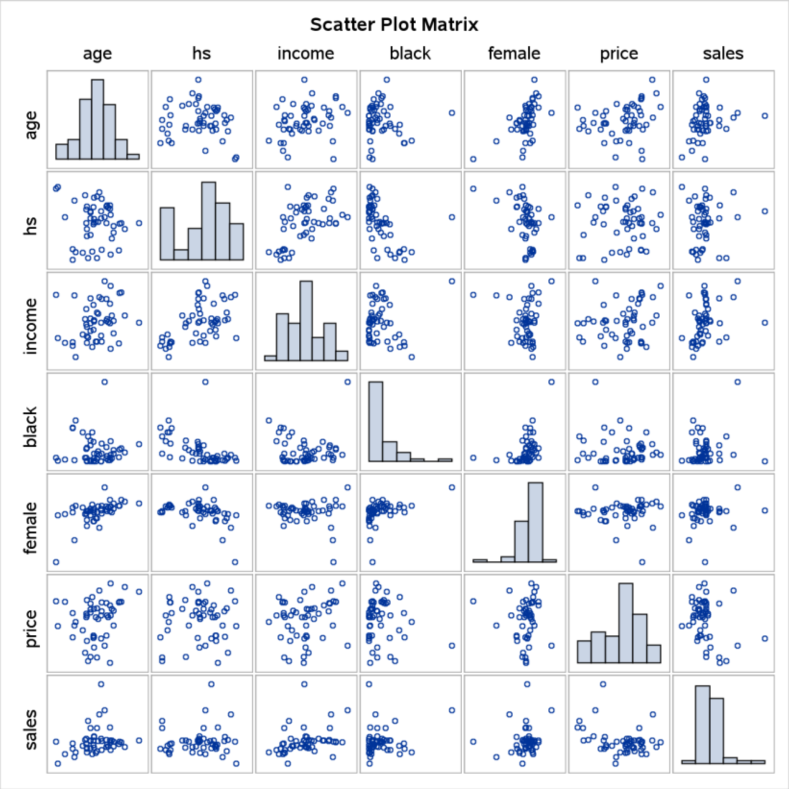
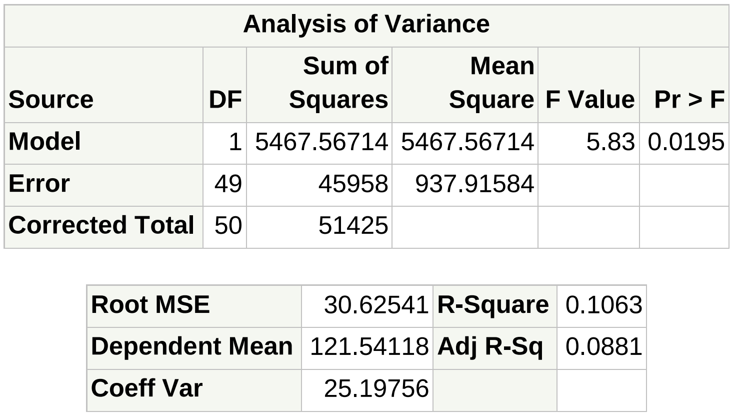


Figure 1: Proc Corr Scatter Plot Matrix

(2) We will use PROC REG to fit the single variable regression model using the highest correlated independent variable Income.

```
ods graphics on;
* Best Single Variable model from Correlation Matrix;
proc reg data=cig PLOTS(ONLY)=(DIAGNOSTICS FITPLOT RESIDUALS);
model sales = income;
title2 'Base Model';
run;
ods graphics off;
```

The output for the model produces an Analysis of Variance (ANOVA) table, parameter estimates, and the goodness-of-fit model diagnostics. Let's begin by taking a look at the information provided by the ANOVA table. The primary pieces of information of interest are the Overall F-test for a regression effect and the goodness-of-fit metrics R-Squared and Adjusted R-Squared. From Table 2 we can see that the the F value is statistically significant at the $p < 0.05$ level, indicating that Income has predictive power in explaining the variability in Sales. In addition we see that the R-Squared value of 0.1063 indicates that 10.63% of the variability in cigarette sales can be explained by the state per capita income. In general this is a very weak relationship. Here we have an example of a model with statistically significant predictor variable but not much predictive power.

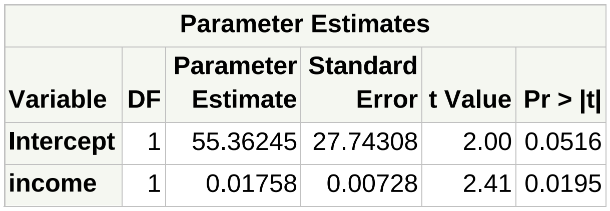


Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	5467.56714	5467.56714	5.83	0.0195
Error	49	45958	937.91584		
Corrected Total	50	51425			

Root MSE	30.62541	**R-Square**	0.1063
Dependent Mean	121.54118	**Adj R-Sq**	0.0881
Coeff Var	25.19756		

Table 2: Analysis of Variance

The parameter estimates can be found in Table 3. Plugging these estimates into the simple linear regression model yields the model: Sales = 55.36 + 0.01758*Income. We interpret this model as: each dollar increase in per capita income will increase the per capita sales of cigarette packs by 0.01758 packs. A better interpretation may state that each $100 increase in per capita income will increase per capital cigarette sales by 1.7 packs of cigarettes.

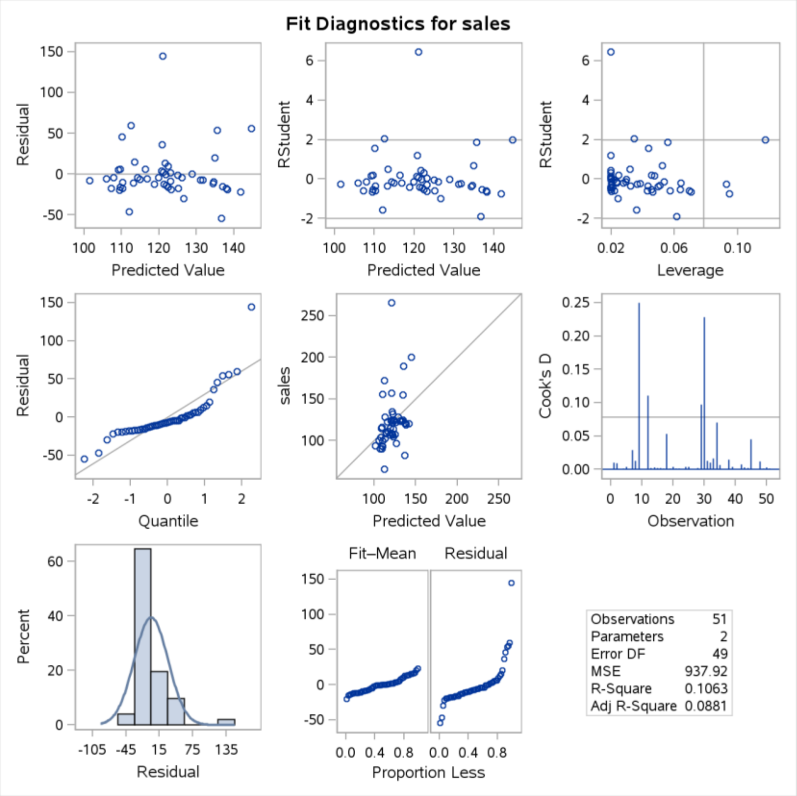


Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	55.36245	27.74308	2.00	0.0516
income	1	0.01758	0.00728	2.41	0.0195

Table 3: Parameter Estimates for Single Variable Model

The SAS procedure PROC REG will produce a panel of diagnostic plots to help you examine the quality of your regression model. Two plots of particular interest are the Quantile-Quantile Plot (or QQ-Plot) and the Cook’s D plot. Both of these plots are available in the matrix of plots provided in Figure 2.

We can use the QQ-Plot of the residuals to examine the normality assumption. In the QQ plot we want to check that the residuals hug the 45-degree line, from which we can conclude that the quantiles of the residuals match the quantiles of a normal distribution. We can see in the QQ-Plot that there are appear to be several outliers in the data set. We can explore these outliers further by looking at the Cook’s D measure. From the Cook’s D fit diagnostic plot we can see the observation number associated with the potential outliers by identifying the observations that are measured above the threshold (horizontal line in the plot).



Observations	51
Parameters	2
Error DF	49
MSE	937.92
R-Square	0.1063
Adj R-Square	0.0881

Figure 2: Proc Reg Fit Diagnostics for Single Variable Model

A final model diagnostic of interest is to plot the model residuals against the model predictor variables to check the assumption of homoscedasticity. When we examine this plot, the residuals should not contain any discernible pattern. Residuals that do not have any discernible pattern are considered to be "random".

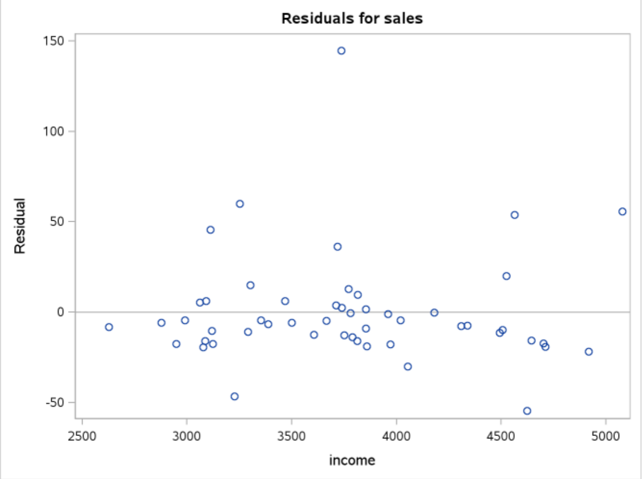
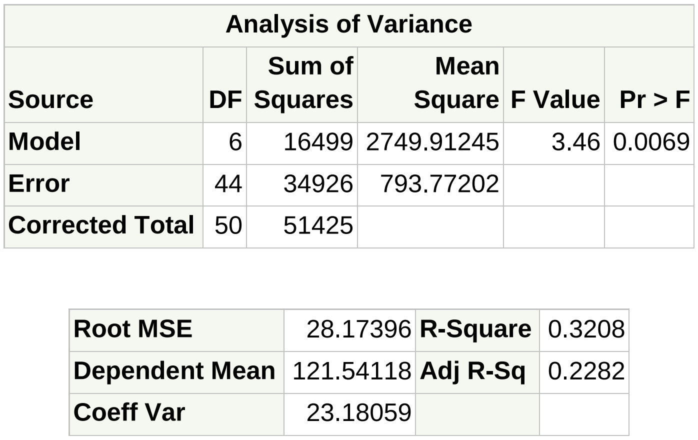


Figure 3: Plot of model residuals against the predictor variable Income.

(3) We will know use PROC REG to fit a multiple regression model, or a regression model with more than one predictor variable. In particular we will reproduce the regression model from p. 86 in *Regression Analysis by Example*.

```
ods graphics on;
* OLS Model using Part e on pp. 87 RABE Variables;
proc reg data=cig PLOTS(ONLY)=(diagnostics residuals fitplot);
model sales = age income price / vif;
title2 'Optimal Model';
output out=fitted_model pred=yhat residual=resid ucl=ucl lcl=lcl;
run;
ods graphics off;
```

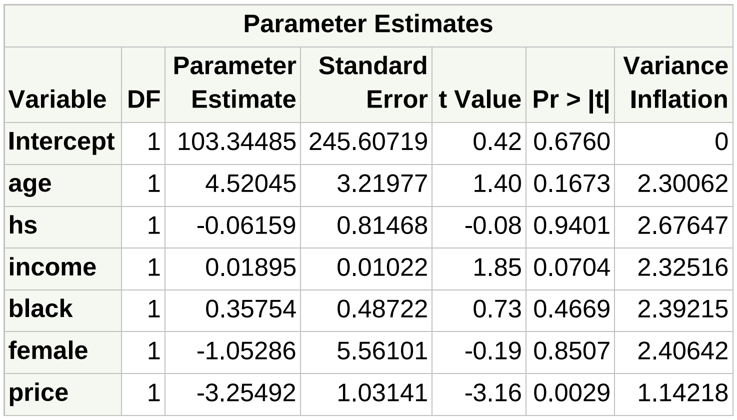
Notice that we have also added the Variance Inflation Factor (VIF), a multicollinearity diagnostic, to the model output. VIF values greater than 3 indicate multicollinearity. We have provided the table output in Tables 4 and 5, but we will leave it to you to reproduce and examine the full output. Be sure to compare your results with those in the text book, and make sure that you understand any interpretations that they are presenting.



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	16499	2749.91245	3.46	0.0069
Error	44	34926	793.77202		
Corrected Total	50	51425			

Root MSE	28.17396	R-Square	0.3208
Dependent Mean	121.54118	Adj R-Sq	0.2282
Coeff Var	23.18059		

Table 4: Analysis of Variance



Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	103.34485	245.60719	0.42	0.6760	0
age	1	4.52045	3.21977	1.40	0.1673	2.30062
hs	1	-0.06159	0.81468	-0.08	0.9401	2.67647
income	1	0.01895	0.01022	1.85	0.0704	2.32516
black	1	0.35754	0.48722	0.73	0.4669	2.39215
female	1	-1.05286	5.56101	-0.19	0.8507	2.40642
price	1	-3.25492	1.03141	-3.16	0.0029	1.14218

Table 5: Parameter Estimates