SAS Simple Linear Regression Example

This handout gives examples of how to use SAS to generate a simple linear regression plot, check the correlation between two variables, fit a simple linear regression model, check the residuals from the model, and also shows some of the ODS (Output Delivery System) output in SAS.

Read in Raw Data

We first read in the raw data from the werner2.dat raw dataset, and set up the missing value codes using a data step, and then check descriptive statistics for the numeric variables, using Proc Means.

```
OPTIONS FORMCHAR="|----|+|---+=|-/\<>*";

libname b510 "C:\Users\kwelch\Desktop\B510";
DATA b510.werner;
   INFILE "C:\Users\kwelch\Desktop\B510\werner2.dat";
   INPUT ID 1-4 AGE 5-8 HT 9-12 WT 13-16
           PILL 17-20  CHOL 21-24 ALB 25-28 1
           CALC 29-32 1 URIC 33-36 1;

   IF HT = 999 THEN HT = .;
   IF WT = 999 THEN WT = .;
   IF CHOL = 600 THEN CHOL = .;
   IF ALB = 99 THEN ALB = .;
   IF CALC = 99 THEN CALC = .;
   IF URIC = 99 THEN URIC = .;
run;

/*Check the Data*/
title "DESCRIPTIVE STATISTICS";
proc means data=b510.werner;
run;
```

DESCRIPTIVE STATISTICS
The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
ID	188	1598.96	1057.09	3.0000000	3519.00
AGE	188	33.8191489	10.1126942	19.0000000	55.0000000
HT	186	64.5107527	2.4850673	57.0000000	71.0000000
WT	186	131.6720430	20.6605767	94.0000000	215.0000000
PILL	188	1.5000000	0.5013351	1.0000000	2.0000000
CHOL	187	235.1550802	44.5706219	50.0000000	390.0000000
ALB	186	4.1112903	0.3579694	3.2000000	5.0000000
CALC	185	9.9621622	0.4795556	8.6000000	11.1000000
URIC	187	4.7705882	1.1572312	2.2000000	9.9000000

Correlation

We now check the correlation between the response (or dependent) variable, CHOL, and the predictor (or independent) variable, AGE. It is positive, and significant (r = .369, p<.0001). Note that there are 188

observations for AGE, but only 187 for CHOL, and that the correlation is based on the 187 observations that have values for both variables.

```
title "Pearson Correlation";
proc corr data=b510.werner;
 var age chol;
run;
```

Pearson Correlation
The CORR Procedure

2 Variables: AGE CHOL

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
AGE	188	33.81915	10.11269	6358	19.00000	55.00000
CHOL	187	235.15508	44.57062	43974	50.00000	390.00000

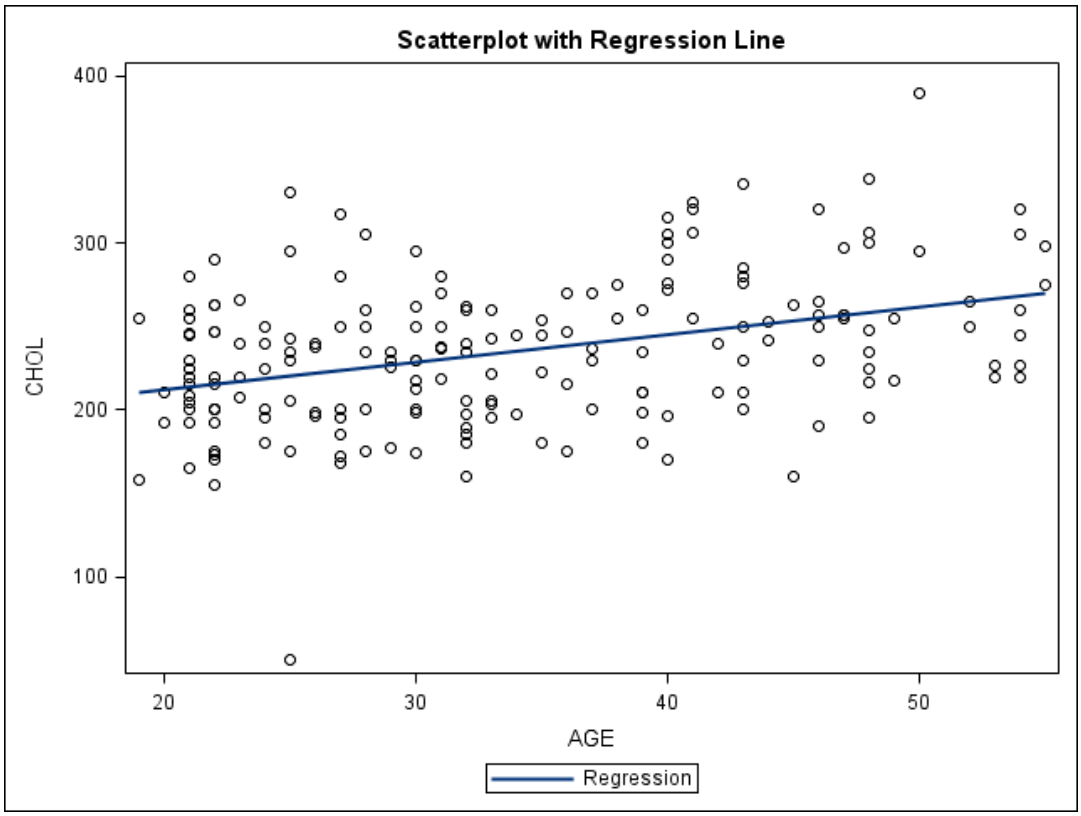
Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	AGE	CHOL
AGE	1.00000	0.36923
		<.0001
	188	187
CHOL	0.36923	1.00000
	<.0001	
	187	187

Scatterplot

We now check a bivariate scatterplot to assess whether the relationship between CHOL and AGE appears to be linear, and to check for outliers. Although there is not a very tight relationship between these two variables, it does appear that the relationship is linear and increasing.

```
title "Scatterplot with Regression Line";
proc sgplot data=b510.werner;
   reg y=chol x=age;
run;
```



Simple Linear Regression

We now fit a linear regression model, with CHOL as the Y (dependent or outcome) variable and AGE as the X (independent or predictor) variable, using Proc Reg. We first illustrate the most basic Proc Reg syntax, and then show some useful options. The Quit statement is used to tell SAS that there are no more statements coming for this run of Proc Reg.

The output shows that there is a positive relationship between these two variables. When age increases by one year, average cholesterol is predicted to increase by 1.62 units, and this is a significant relationship ($t(185) = 5.40$, $p<.0001$). Note that the degrees of freedom for the t-test are 185, the same as the error degrees of freedom. The model R-square (.1368) is the square of the correlation between the two variables. There were 187 observations used in the regression model.

```
title "Simple Linear Regression Model with no options";
proc reg data=b510.werner;
     model chol = age;
run;quit;
```

```
                Simple Linear Regression Model with no options
                            The REG Procedure
                              Model: MODEL1
                        Dependent Variable: CHOL

Number of Observations Read                         188
Number of Observations Used                         187
Number of Observations with Missing Values            1
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	50373	50373	29.20	<.0001
Error	185	319123	1724.99020		
Corrected Total	186	369497			

Root MSE	41.53300	R-Square	0.1363
Dependent Mean	235.15508	Adj R-Sq	0.1317
Coeff Var	17.66196		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	179.96174	10.65564	16.89	<.0001
AGE	1	1.62897	0.30144	5.40	<.0001

Simple Linear Regression with Diagnostic Plots

We now include some diagnostic plots using Proc Reg. We also generate a new dataset called OUTREG1 that contains all of the original variables, plus the predicted value for each observation (PREDICT), the residual (RESID) and the studentized-deleted residual (RSTUD), and Cook's Distance (COOKD)..

```
ods graphics on;
title "Simple Linear Regression with Diagnostic Plots";
proc reg DATA=B510.werner;
     MODEL CHOL=AGE / stb clb;
     OUTPUT OUT=OUTREG1 P=PREDICT R=RESID RSTUDENT=RSTUDENT COOKD=COOKD;
run;quit;
ods graphics off;
```

The partial output below shows the standardized estimate (obtained with the STB option), which shows the estimated change in Y (in standard deviation units) when X is increased by one standard deviation. This estimate is 0.369. We also see the 95% Confidence limits for the parameter estimate, which are form 1.03 to 2.22.

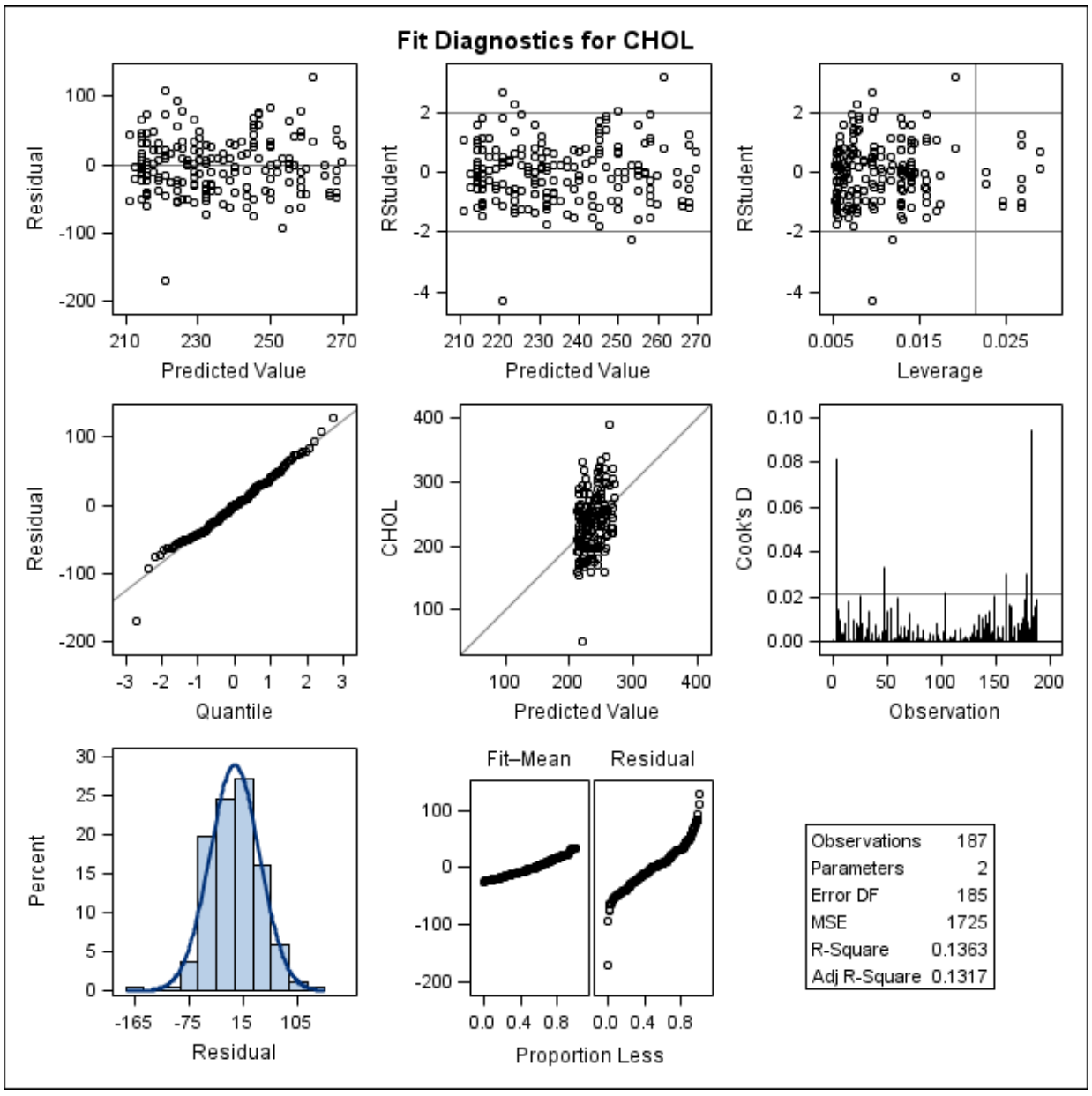
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	179.96174	10.65564	16.89	<.0001	0
AGE	1	1.62897	0.30144	5.40	<.0001	0.36923

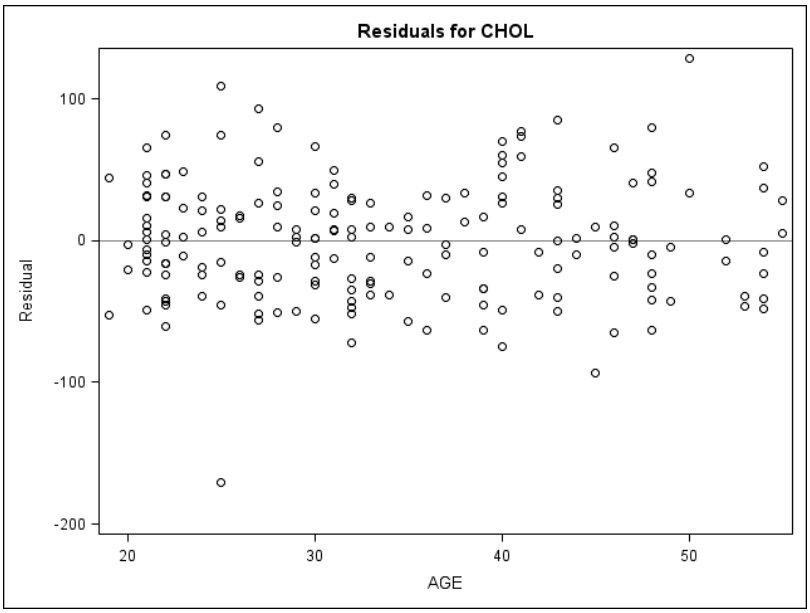
Parameter Estimates

Variable	DF	95% Confidence Limits	
Intercept	1	158.93955	200.98392
AGE	1	1.03426	2.22368

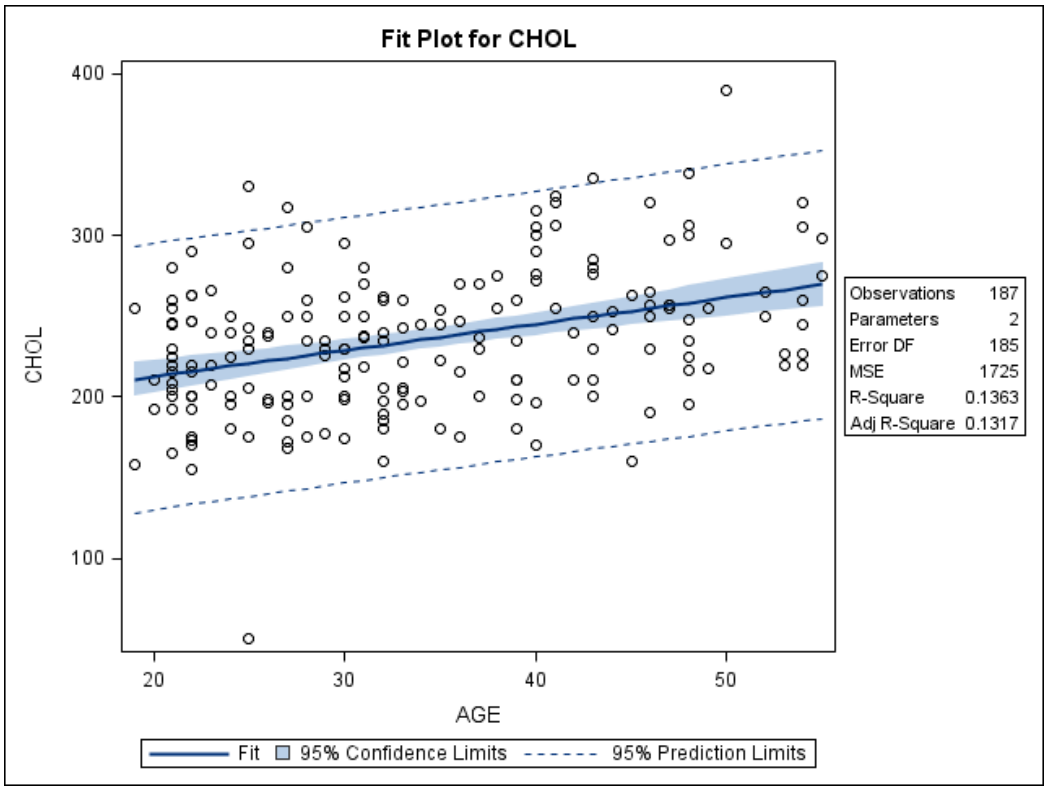
The diagnostic panel shows a series of diagnostic plots for this regression model.



The residual plot below shows a scatterplot with the residuals on the Y-axis and AGE on the X-axis. We want to look for a lack of pattern in these residuals. We can see that there is one low outlier, at about age 25.



The fit plot shown below shows the regression model fit, and summarizes some of the statistics for the model.



Check the output dataset

We now check the output dataset, using Proc Print. We also request that Proc Print display the labels for the each variable, by using the Label option. We print selected variables for those observations with the absolute value of the studentized deleted residuals being greater than or equal to 3, using a Where statement.

```
title "Partial Listing of Output Dataset";
proc print data=outreg1;
   where abs(rstud) >=3;
   VAR ID AGE CHOL PREDICT RESID RSTUD COOKD LCL UCL LCLM UCLM;
run;
```

Partial Listing of Output Dataset

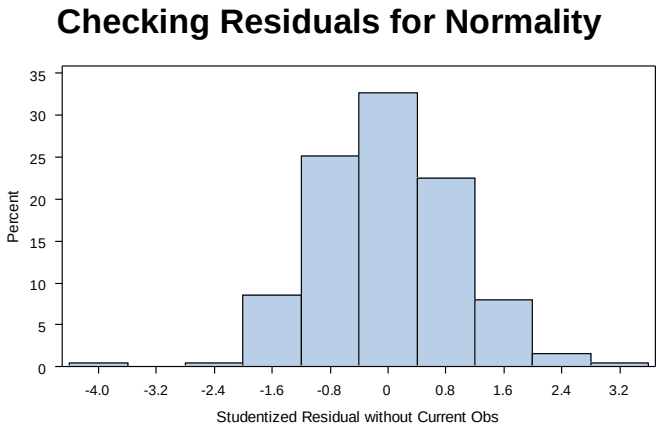
Obs	ID	AGE	CHOL	PREDICT	RESID	RSTUD	COOKD	LCL	UCL	LCLM	UCLM
4	1797	25	50	220.686	-170.686	-4.32214	0.081802	138.358	303.014	212.698	228.674
182	3134	50	390	261.410	128.590	3.20326	0.094792	178.695	344.126	250.106	272.714

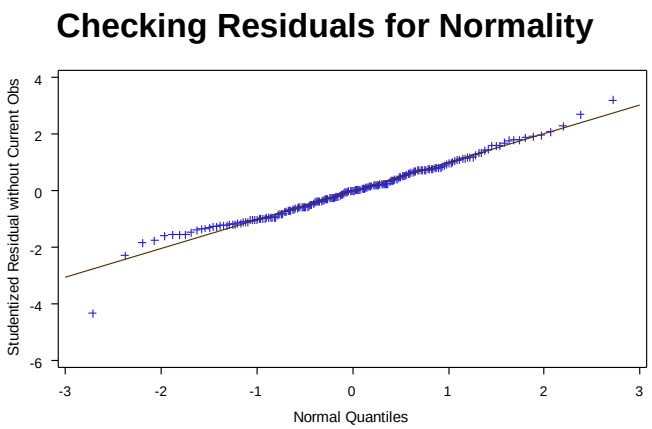
Check the residuals for normality

We now check the studentized residuals for normality, using Proc Univariate. This is similar to the output from the ODS graphics that was shown in the earlier panel.

```
title "Checking Residuals for Normality";
proc univariate data=outreg1 PLOT NORMAL;
     var rstud;
       histogram / normal;
       qqplot / normal(mu=est sigma=est);
run;
```

The residuals appear to be fairly normally distributed, but there is at least one very low outlier, which we identified earlier, when we checked the values in the output dataset.





Refit the regression model without the cases in question

We now refit the model, but without the two outliers being included, by using a Where statement..

```
ods graphics on;
title "Rerun the model without two obs";
proc reg data=b510.WERNER;
```

```
    where id not in (1797, 3134);
    model chol=age;
run;quit;
ods graphics off;
```

We can see the changes in the parameter estimates from the output below.

Dependent Variable: CHOL

Number of Observations Read	186
Number of Observations Used	185
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	38478	38478	25.82	<.0001
Error	183	272754	1490.46158		
Corrected Total	184	311232			

Root MSE	38.60650	R-Square	0.1236
Dependent Mean	235.31892	Adj R-Sq	0.1188
Coeff Var	16.40603		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	186.70039	9.98091	18.71	<.0001
AGE	1	1.43658	0.28274	5.08	<.0001