The REG Procedure

When two variables are observed for each observation in a data set, one may be interested whether or not one variable can be used to predict the other. A simple linear regression line can be fit to the data. The values of the dependent and independent variables are given by (x_i, y_i). The statistical model for simple linear regression is given by:
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where i = 1, . . . n and y_i is the value of the dependent variable, β_0 is the y-intercept, β_1 is the slope of the line, x_i is the value of the independent variable, and ε_i is the unobservable random error about the regression line. The values of the y-intercept and slope are population parameters that can be estimated using ordinary least square regression techniques. The estimates of β_0 and β_1 are given by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. The estimated regression line is given by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ where $\hat{y}_i$ is the predicted value for y and a given x; that is, $\hat{y}_i$ is a point on the regression line.

Of interest is whether or not the regression line adequately models the data. That is, is x a good predictor of y? If so, then we'd expect that the value of the slope would be different from zero. (A line with a zero slope is a horizontal line, and is equivalent to using $\bar{y}$ to predicting the outcome for y.) Therefore, it is of interest to test the slope or compute a confidence interval for the slope. Similarly, the y-intercept can be tested or estimated.

Hypotheses	**Test Statistic**	**Reject H_0 if**	**$(1-\alpha)100\%$ CI for β_i**
$H_0: \beta_i = 0$ $H_1: \beta_i \neq 0$	$t = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$ where $se(\hat{\beta}_i)$ is the standard error of the $\hat{\beta}_i$ parameter estimate	$\lvert t \rvert \geq t_{\alpha/2,df}$ where df is the error degrees of freedom. In simple linear regression this is n – 2.	$\hat{\beta}_i \pm t_{\alpha/2,df}\ se(\hat{\beta}_i)$

Confidence intervals for the mean response at a given value of x and prediction intervals for an individual response at a given value of x can also be computed. These intervals are centered at the point prediction, $\hat{y}_i$.

When there is only one independent variable, the regression model is a simple linear regression model. More independent variables and hence, more regression parameters could be included in a regression analysis. The resulting model is a multiple regression model. If there are k different independent variables used to predict y, then there are be k slope parameters and a y-intercept (k + 1 parameters total). The above formulas for the hypotheses tests and the confidence intervals apply for each of the possible k + 1 parameters in the model.

When either simple linear or multiple regression model equations are to be estimated for a set of data, the REG procedure can provide estimates of the parameters, tests of their significance, residual analysis, plots of data, and much more.

The syntax of the REG procedure:

PROC REG DATA=*tablename* <options> ;
MODEL *dependents = regressors </options* > ;
PLOT < *yvariable*xvariable*> *<=symbol> </options >;*
RUN;

(BY and WHERE statements can be included, of course.) This is not a complete list of statements for the REG procedure. See SAS/STAT for more detailed information.

PROC REG Statement options include:

PLOTS= NONE | DIAGNOSTICS | FIT
: ODS Graphics must be enabled. There are many more ODS Graphics options available. See SAS Help and Documentation.

SIMPLE
: simple statistics for each variable in the MODEL statement are printed. (Mean, summation, uncorrected sum of squares, . . .)

MODEL Statement – In the MODEL statement the dependent and independent variables are identified. The dependent variables always occur to the left of the equality symbol, and the independents are always listed to the right of the equality symbol. If only one independent variable is identified, the result is a simple linear regression analysis as is taught in introductory statistics courses.

More than one MODEL statement can be used in a single regression procedure. Each MODEL statement can have different options also. MODEL statement options include:

ALPHA = p
: specifies the type I error rate for the confidence and prediction intervals that may be requested in the MODEL statement options. $0 < p < 1$. If the ALPHA option is not specified, all intervals are computed with the default setting $p = 0.05$.

CLI
: requests the (1-p)100% upper- and lower-confidence limits for an individual predicted value. (Note: CL**I** for **individual**.)

CLM
: requests the (1-p)100% upper- and lower-confidence limits for the expected value or mean response. (Note: CL**M** for **mean** response.)

CLB
: requests the (1-p)100% confidence limits for the regression parameters. Typically, regression parameters are identify as β's in statistical literature, hence the CL**B**.

P
: calculates the predicted values from the input data and the estimated model.

The following options are for those who have had more than an introduction to regression analysis.

COLLIN
: requests a detailed analysis of collinearity among the regressors.

INFLUENCE	requests a detailed analysis of the influence of each observation of the estimates and the predicted values.
NOINT	suppresses the intercept term that is otherwise included in the model.
PARTIAL	requests partial regression leverage plots for each regressor.
R	requests an analysis of the residuals.
SELECTION = *name*	specifies the method used to select the model, where name can be FORWARD (or F), BACKWARD (or B), STEPWISE, MAXR, MINR, RSQUARE, ADJRSQ, CP. Only one method can be specified in a MODEL statement.
SS1	prints the sequential sums of squares (Type I SS) along with the parameter estimates for each term in the model.
SS2	prints the partial sums of squares (Type II SS) along with the parameter estimates for each term in the model.

PLOT Statement

The PLOT statement prints scatter plots with y-variables on the vertical axes and x-variables on the horizontal axes. It uses symbols to mark points in the plots. The y-variables and x-variables can be any variables that appear in the MODEL statement. Y-variables and x-variables can also be keywords that are specified and followed by a period. Keywords include: residual. predicted. L95. U95. L95M. U95M.

The images produced by this plot statement are independent of ODS Graphics.

Changing symbols, colors and other attributes of the graph in the graphics window does require one to know graphics statements such as SYMBOLn which are covered later. The following objective demonstrates the PLOT statement within the REG procedure.

Objective 7: A study was conducted to examine the quality of fish after seven days in ice storage. Ten raw fish of the same kind and approximately the same size were caught and prepared for ice storage. Two of the fish were placed in storage immediately after being caught, two were placed in storage 3 hours after being caught, and two each were placed in storage at 6, 9, and 12 hours after being caught. A measure of fish quality after the seven days of storage was made. The data appear below.

Quality	8.5	8.4	7.9	8.1	7.8	7.6	7.3	7.0	6.8	6.7
Hours	0	0	3	3	6	6	9	9	12	12

- Compute simple summary statistics for each of the variables in the regression analysis. (SIMPLE option)
- Regress the dependent variable quality on the independent variable hours. (MODEL . . .)
- Predict the quality of the fish if it will be 10 hours before the fish is placed in storage. (P option, Note that 10 was not one and the original independent values in the experiment and is input as an independent variable and the corresponding dependent value is missing.)
- Compute the 95% prediction and confidence intervals for quality (CLI and CLM options)
- Compute the 95% confidence intervals for the regression parameters (CLB option)
- Plot the regression line and the data on the same set of axes. (PLOT statement)
- Observe the default ODS Graphics obtained.

```
DM 'LOG; CLEAR; ODSRESULTS; CLEAR; ';

DATA one;
INPUT quality hours;
DATALINES;
8.5  0
8.4  0
7.9  3
8.1  3
7.8  6
7.6  6
7.3  9
7.0  9
6.8  12
6.7  12
.    10
;
PROC PRINT DATA=one;
TITLE 'Simple Linear Regression Example';
TITLE2 'Objective 7';
PROC REG DATA=one SIMPLE;
MODEL quality = hours/P  CLM   CLI  CLB;
ID hours ;
PLOT quality*hours;

RUN;
QUIT;
```

Simple Linear Regression Example
Objective 7

Obs	quality	hours
1	8.5	0
2	8.4	0
3	7.9	3
4	8.1	3
5	7.8	6
6	7.6	6
7	7.3	9
8	7.0	9
9	6.8	12
10	6.7	12
11	.	10

Simple Linear Regression Example
Objective 7

The REG Procedure

Number of Observations Read	11
Number of Observations Used	10
Number of Observations with Missing Values	1

Descriptive Statistics

Variable	Sum	Mean	Uncorrected SS	Variance	Standard Deviation
Intercept	10.00000	1.00000	10.00000	0	0
hours	60.00000	6.00000	540.00000	20.00000	4.47214
quality	76.10000	7.61000	582.85000	0.41433	0.64369

Simple Linear Regression Example
Your Name

The REG Procedure
Model: MODEL1
Dependent Variable: quality

Number of Observations Read	11
Number of Observations Used	10
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3.61250	3.61250	248.07	<.0001
Error	8	0.11650	0.01456		
Corrected Total	9	3.72900			

Root MSE	0.12068	**R-Square**	0.9688
Dependent Mean	7.61000	**Adj R-Sq**	0.9649
Coeff Var	1.58574		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	**1**	8.46000	0.06610	127.99	<.0001	8.30758	8.61242
hours	**1**	-0.14167	0.00899	-15.75	<.0001	-0.16241	-0.12093

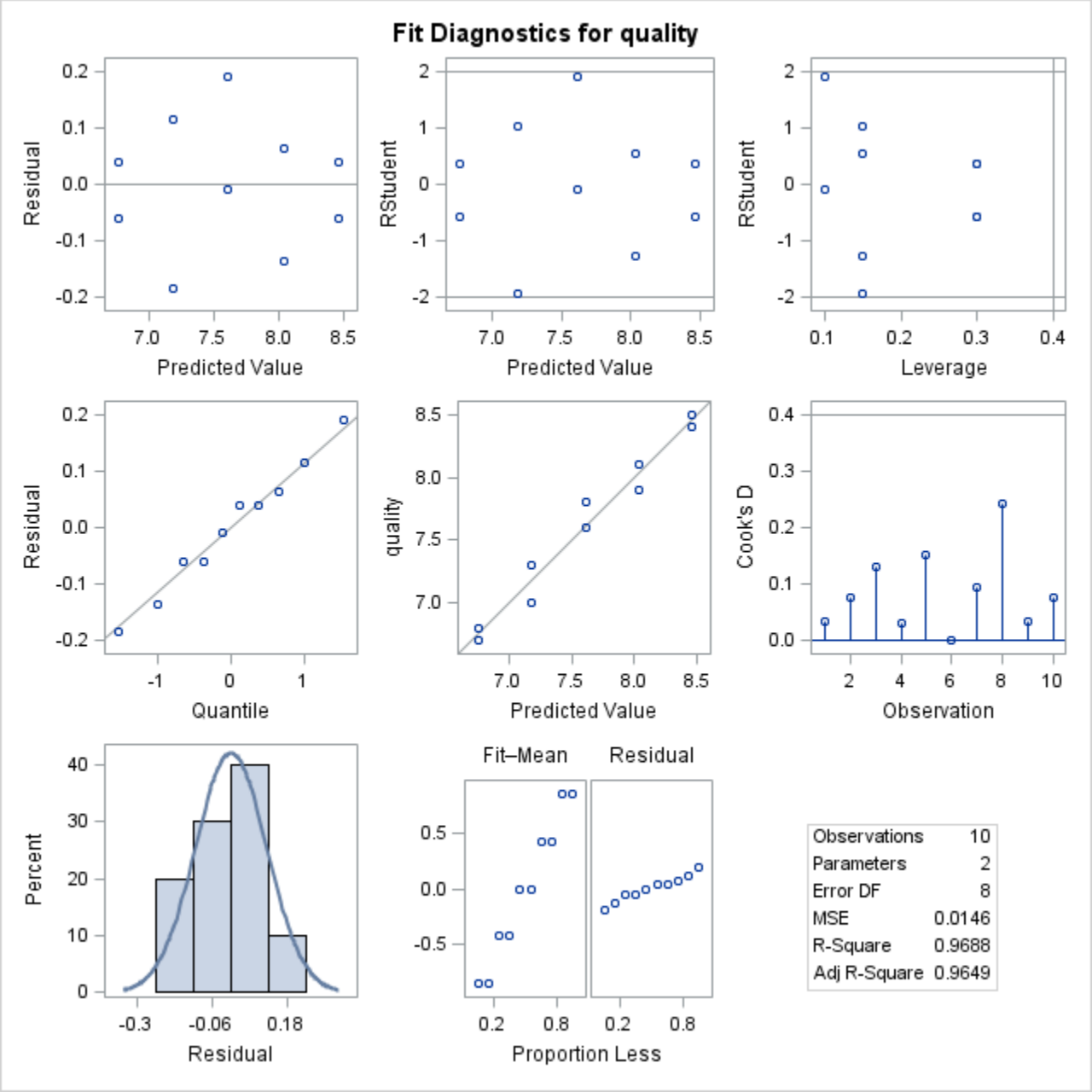
Simple Linear Regression Example
Objective 7

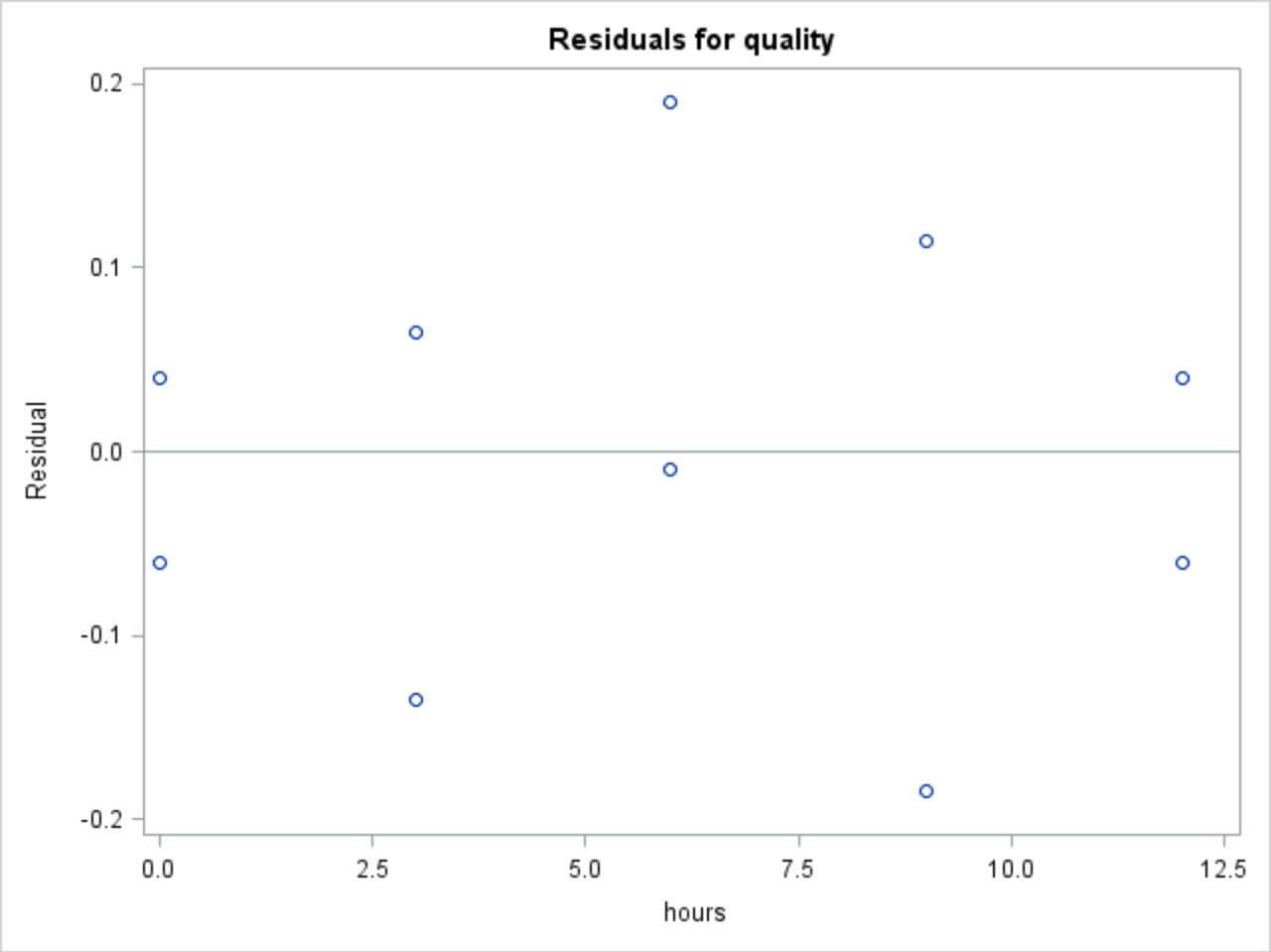
The REG Procedure
Model: MODEL1
Dependent Variable: quality

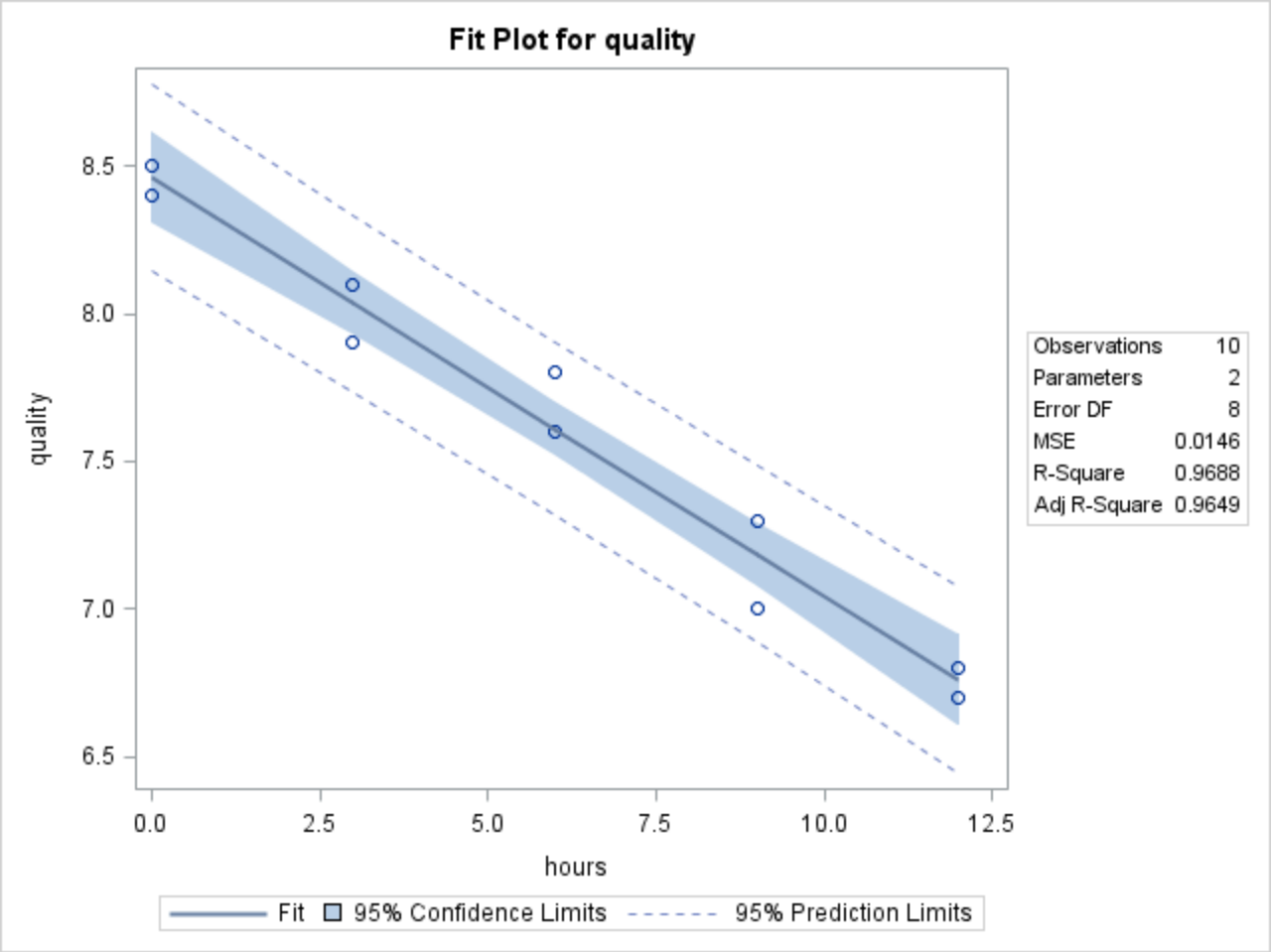
Output Statistics

Obs	hours	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	**0**	8.5000	8.4600	0.0661	8.3076	8.6124	8.1427	8.7773	0.0400
2	**0**	8.4000	8.4600	0.0661	8.3076	8.6124	8.1427	8.7773	-0.0600
3	**3**	7.9000	8.0350	0.0467	7.9272	8.1428	7.7366	8.3334	-0.1350
4	**3**	8.1000	8.0350	0.0467	7.9272	8.1428	7.7366	8.3334	0.0650
5	**6**	7.8000	7.6100	0.0382	7.5220	7.6980	7.3181	7.9019	0.1900
6	**6**	7.6000	7.6100	0.0382	7.5220	7.6980	7.3181	7.9019	-0.0100
7	**9**	7.3000	7.1850	0.0467	7.0772	7.2928	6.8866	7.4834	0.1150
8	**9**	7.0000	7.1850	0.0467	7.0772	7.2928	6.8866	7.4834	-0.1850
9	**12**	6.8000	6.7600	0.0661	6.6076	6.9124	6.4427	7.0773	0.0400
10	**12**	6.7000	6.7600	0.0661	6.6076	6.9124	6.4427	7.0773	-0.0600
11	**10**	.	7.0433	0.0524	6.9224	7.1643	6.7399	7.3468	.

Sum of Residuals	0
Sum of Squared Residuals	0.11650
Predicted Residual SS (PRESS)	0.16266


Fit Diagnostics for quality
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
quality
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 10
Parameters 2
Error DF 8
MSE 0.0146
R-Square 0.9688
Adj R-Square 0.9649


Residuals for quality
Residual
0.2
0.1
0.0
-0.1
-0.2
0.0
2.5
5.0
7.5
10.0
12.5
hours


Fit Plot for quality
quality
8.5
8.0
7.5
7.0
6.5
0.0
2.5
5.0
7.5
10.0
12.5
hours
Observations 10
Parameters 2
Error DF 8
MSE 0.0146
R-Square 0.9688
Adj R-Square 0.9649
Fit
95% Confidence Limits
95% Prediction Limits

The REG Procedure

Simple Linear Regression Example

Objective 7