CHAPTER 2

Inference for Simple Linear Regression

2.3 Regression and Correlation

2.4 Intervals for Predictions

2.5 Chapter Summary

2.3 Regression and Correlation :

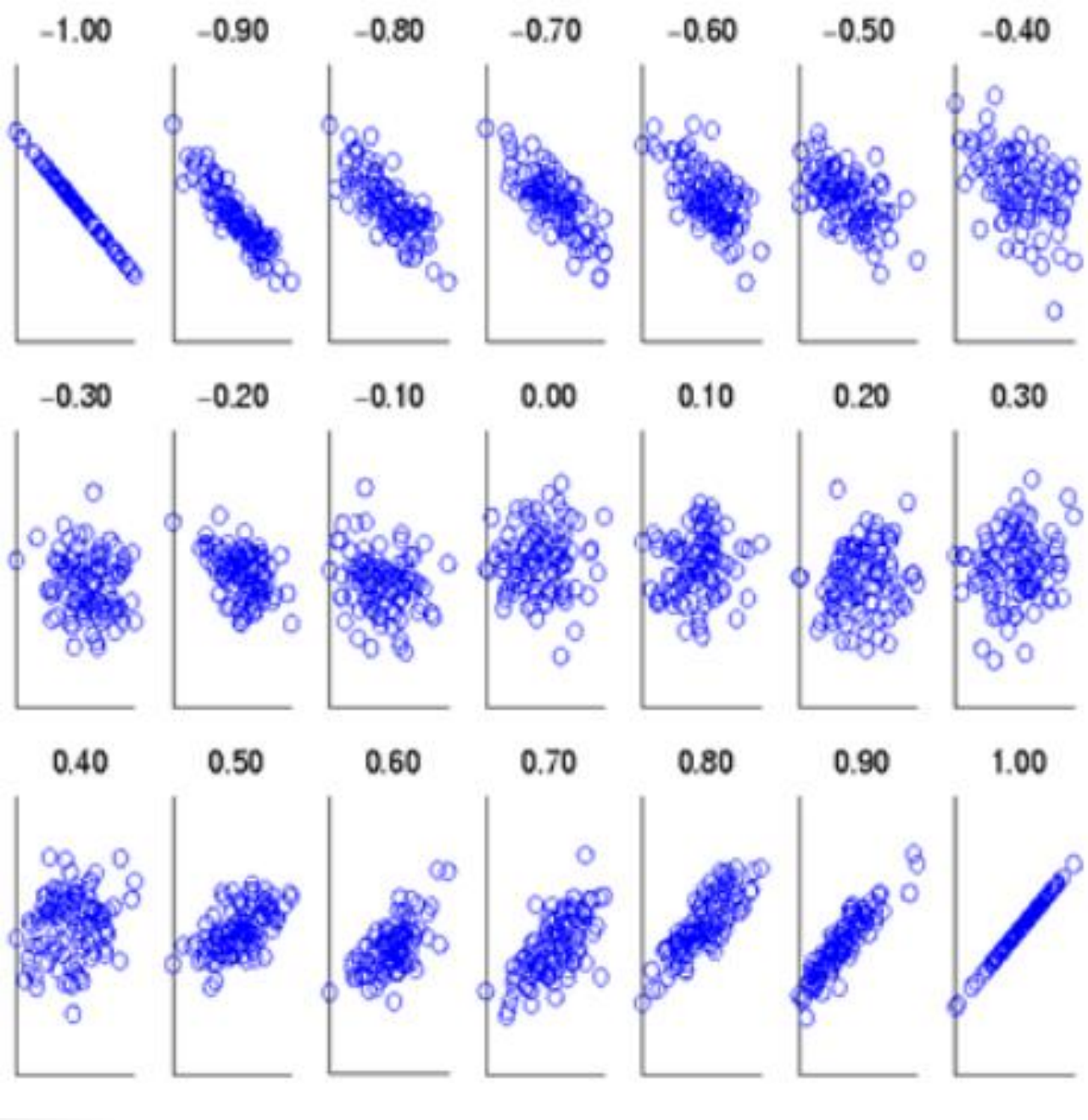
Recall that the sample correlation coefficient (r) is a number between (-1) and (+1) that measure the strength of the linear association between two quantitative variables. $(-1 \leq r \leq 1)$

- The correlation coefficient is useful for assessing the significant of a simple linear model.
- r measures the strength of the linear association.
- The correlation coefficient formula is:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2]\,[n(\sum y^2) - (\sum y)^2]}}$$

Correlation Coefficient

Shows Strength & Direction of Correlation


-1.00
-0.90
-0.80
-0.70
-0.60
-0.50
-0.40
-0.30
-0.20
-0.10
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00

T-test for Correlation:

Let ρ (rho) denote the population correlation, the hypotheses are:

$$H_0: \rho = 0 \quad vs \quad H_1: \rho \neq 0$$

and the test statistic is

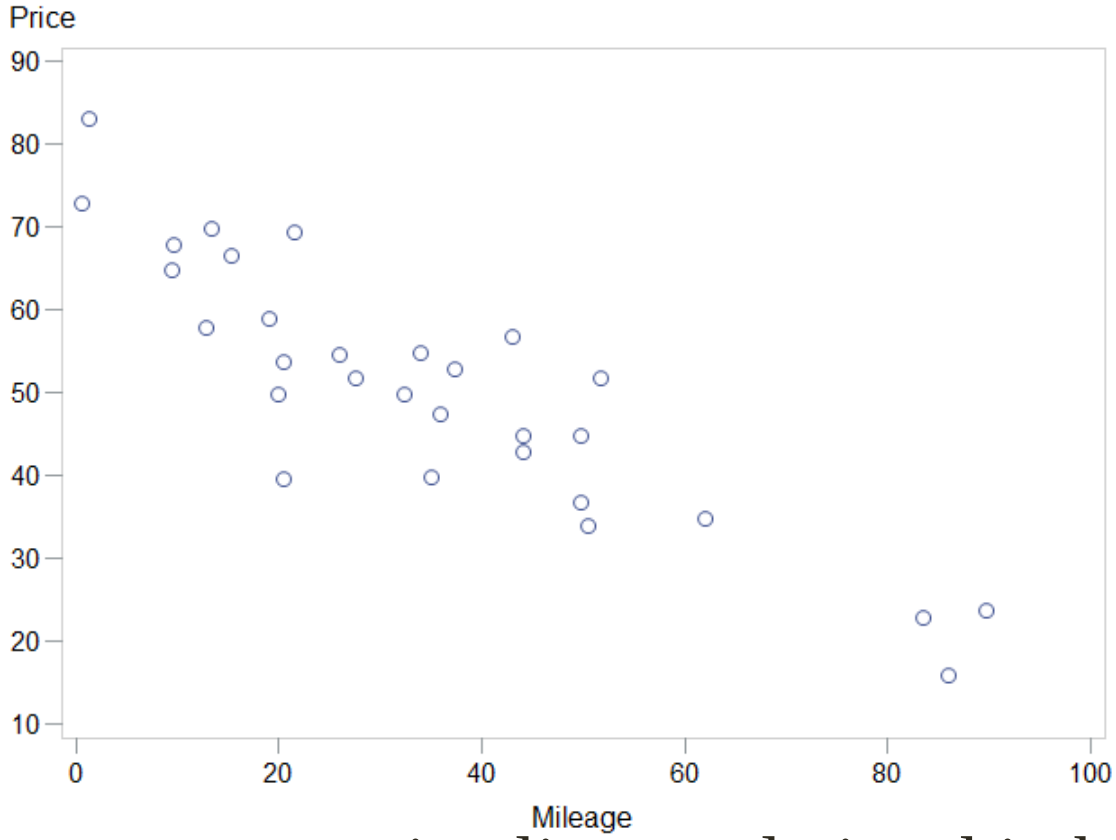
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Similar to any other hypothesis test, we find the *p-value* to test the hypotheses.

Example 1: *(Porsche prices)*

For the same dataset *Porsche prices.csv*.

1. Using SAS, graph the scatterplot of the mileage vs. price.



There is a strong negative linear relationship between the price and mileage.

In SAS, we can use the correlation procedure to calculate the correlation coefficient and graph the scatter plot.

```
proc corr data=PorschePrice plots=matrix(histogram);
var price Mileage;
run;
```

2. Using SAS, Calculate and interpret the correlation coefficient.

Pearson Correlation Coefficients, N = 30 Prob > |r| under H0: Rho=0	
	Price
Mileage	-0.89135
	<.0001

Since $r = -0.89$, so there is a strong negative linear relationship between the mileage and the price, which means that cars with high mileage will has lower price.

3. State your hypotheses. Using SAS, find p-value and interpret it. Use significance level $\alpha = 5\%$

Pearson Correlation Coefficients, N = 30 Prob > |r| under H0: Rho=0	
	Price
Mileage	-0.89135
	<.0001

$$H_0: \rho = 0 \quad vs \quad H_1: \rho \neq 0$$

Since $p - value < 0.0001 < 0.05$, so we reject H_0, the correlation coefficient of the population doesn't equal to 0. Which means there is a significant strong negative relationship between the mileage and the price (cars with high mileage will has lower price).

Coefficient of Determination (r^2):

Tells us how much variation in the response variable Y we explain by using the explanatory variable X in the regression model.

$$r^2 = \frac{Variability\ explained\ by\ the\ model}{Total\ variability\ in\ Y}$$

$$= \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2} = \frac{SSModel}{SSTotal}$$

Example 2: *(Porsche prices)*

For the same dataset *Porsche prices.csv*.

1. Using SAS, find the coefficient of determination and interpret it.

Root MSE	7.17029	R-Square	0.7945
Dependent Mean	50.53667	Adj R-Sq	0.7872
Coeff Var	14.18829		

$$r^2 = (-0.89135)^2 = 0.7945 \approx 79.45\%$$

Or

$$r^2 = \frac{SSModel}{SSTotal} = \frac{5565.68453}{7005.24967} = 0.7945$$

Interpretation: 79.5% of the variability in the price of the Porsches in this sample can be explained by the linear model based on their mileages.

Three Tests for a Linear Relationship?

We now have three distinct ways to test for a significant linear relationship between two quantitative variables.

- The t-test for slope.
- The ANOVA for regression.
- The t-test for correlation.

The three procedures are exactly equivalent in the case of simple linear regression.

Note: We need those tests for multiple regression.

2.4 Intervals for Predictions:

One of the most common reasons to fit a line to data is to predict the response for a particular value of the explanatory variable.

- To estimate the mean response μ_Y, we use a **confidence interval** for the mean.
- To estimate an individual response Y, we use a **prediction interval**.

Confidence and Prediction Intervals for a Simple Linear Regression Response:

- A confidence Interval for the mean response μ_Y when X takes the value x^* is:

$$\hat{y} \pm t^*.SE_{\hat{\mu}}$$

where

$$SE_{\hat{\mu}} = \hat{\sigma}_{\in}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})}{\sum(x - \bar{x})^2}}$$

- A Prediction Interval for a single observation of Y when X takes the value x^* is:

$$\hat{y} \pm t^*.SE_{\hat{y}}$$

where

$$SE_{\hat{y}} = \hat{\sigma}_{\in}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})}{\sum(x - \bar{x})^2}}$$

Example 3: *(Porsche prices)*

1. Predict the price of Porsche with 50,000 mileage.

Since there is no car with 50,000 mileage on it in the dataset, so we need to add another car to the dataset using the following code.

```
data PorschePrice_1;
set PorschePrice end=eof;
output;
if eof then do;
   Price = '.';
   Age = '.';
   Mileage = 50;
   output;
  end;
run;
```

Obs	Price	Age	Mileage
31	.	.	50

1. Predict the price of Porsche with 50,000 mileage.

$$\widehat{Price} = 71.09 - 0.5894\,(50) = 41.62$$

The average price is about \$41,620 (when $mileage = 50{,}000$).

```
proc reg data = PorschePrice_1;
model price = mileage ;
output out = results_Porschel p = predicted;
run;

proc print data = results_Porschel (firstobs = 31);
run;
```

Obs	Price	Age	Mileage	predicted
31	.	.	50	41.6204

2. Determine a 95% confidence interval for the average price of Porsche with 50,000 mileage.

```
proc reg data = PorschePrice_1;
model price = mileage ;
output out = results_Porsche2 LCLM = Lower_M UCLM = Upper_M;
run;

proc print data = results_Porsche2 (firstobs = 31);
run;
```

Obs	Price	Age	Mileage	Lower_M	Upper_M
31	.	.	50	38.4154	44.8255

The 95% confidence interval for the average price of all used Porsche with 50,000 mileage is somewhere between $38,429 and $44,830

($38,420, $44,830)

3. Determine a 95% prediction interval for the mean price of Porsche with 50,000 mileage.

```
proc reg data = PorschePrice_1;
model price = mileage ;
output out = results_Porsche3 LCL = LCL_Pred UCL = UCL_Pred;
run;

proc print data = results_Porsche3 (firstobs = 31);
run;
```

Obs	Price	Age	Mileage	LCL_Pred	UCL_Pred
31	.	.	50	26.5871	56.6537

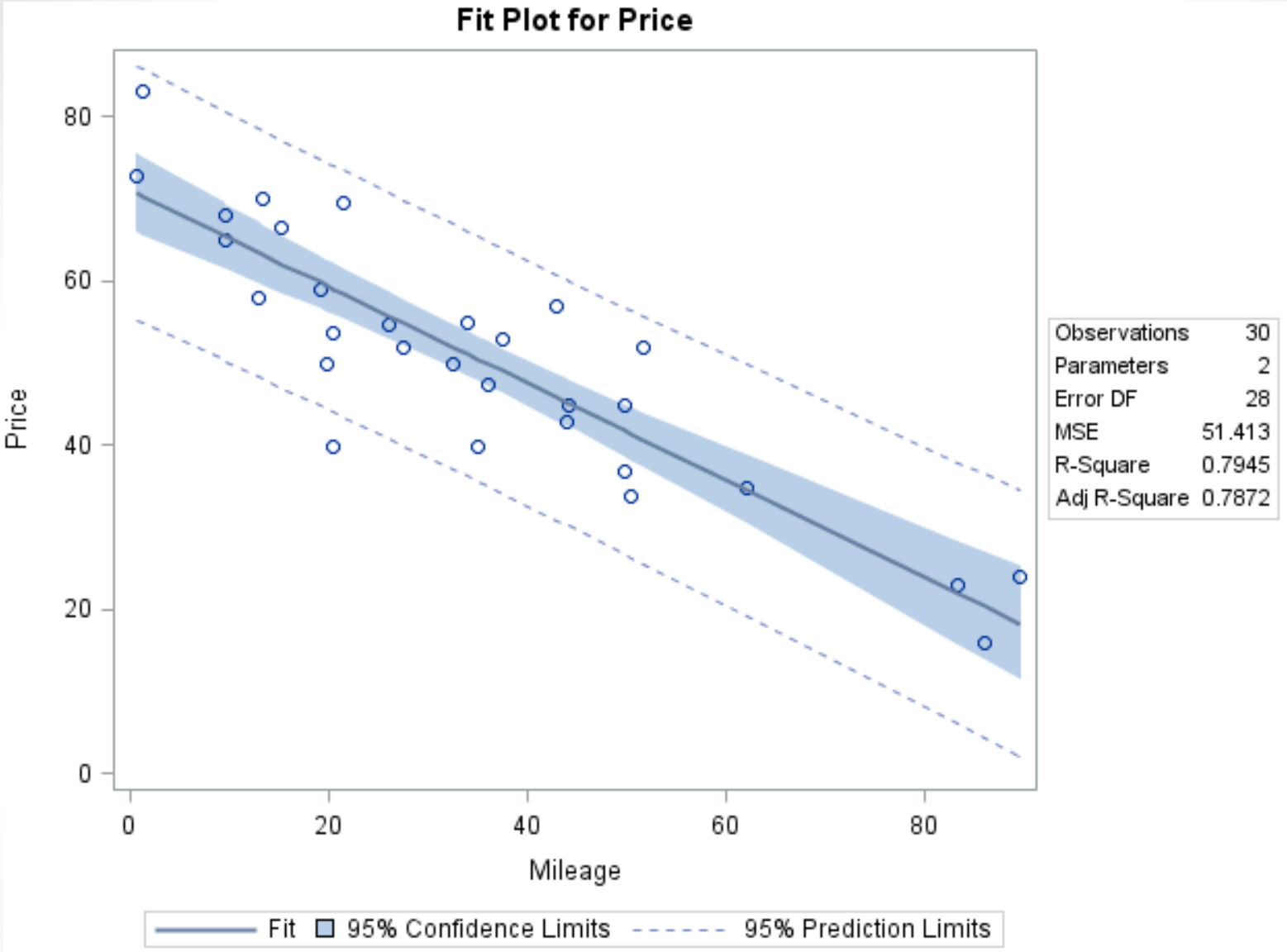
The 95% prediction interval tells us that we should to expect about %95 of those Porsches with 50,00 miles to be priced between $26,590 and $56,650

($26,590, $56,650)

Note: we can also use "clm" and "cli" function in the model statement to find the confide and prediction interval.

```
proc reg data = PorschePrice_1;
model price = mileage / clm cli;
run;
```

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
31	.	41.6204	1.5647	38.4154	44.8255	26.5871	56.6537	.



Note: The prediction interval (PI) is always wider than the confidence interval (CI).

Reading Assignment

Read section 2.3 - 2.5