CHAPTER 2

Inference for Simple Linear Regression

2.1 Inference for Regression Slope

2.2 Partitioning Variability

Recall that in the car price example we consider a simple linear regression model to predict the price of the used Porsches based on mileage.

- How can we evaluate the effectiveness of this model?
- Are prices significantly related to mileage?
- How much of variability in Porsche prices can explain by knowing their mileage?
- If we are interested in a use Porsche with about 50,000 miles, how accurately can we predict its price?

General Steps of Hypothesis Testing:

1. Determine the null and alternative hypothesis.
2. Verify necessary data conditions, and if met, summarize the data into appropriate test statistic.
3. Assume the null hypothesis is true, find the *p-value*.
4. Decide whether or not the result is statistically significant based on the *p-value*.
5. Report the conclusion in the context of the situation.

Step 1: Determine the hypotheses:

The hypothesis is the statement created by researchers.

There are two types of hypothesis, which are

The **null hypothesis** **(H_0)** is a statement about a population parameter, such as the population mean, that is assume to be true.

- The null hypothesis is a starting point. We will test whether the value stated in the null hypothesis is likely to be true.

The alternative hypothesis (H_1) is a statement that directly contradicts a null hypothesis by stating that the actual value of a population parameter is less than (<), greater than (>), or not equal (≠) to that value stated in the null hypothesis.

- The alternative hypothesis states what we think is wrong about the null hypothesis.
- It is usually the reason data being collected.

Step 2: Collect Data and summarize with a test statistic:

Decision in hypothesis test based on single summary of data (test statistic).

Step 3: Determine how unlikely test statistic would be if null hypothesis true:

If null hypothesis true, how likely to observe sample results of this magnitude or larger (in direction of the alternative) just by chance?

P-value

Step 4: Make a statistical decision:

1. *P-value* not small enough to convincingly rule out chance. (Usually use 0.05 or 5%, so if *p-value* > 0.05).

- ➢ Can’t (fail to) reject the null hypothesis (H_0).
- ➢ No statistically significant difference or relationship.

2. *P-value* small enough to convincingly rule out chance. (Usually use 0.05, so if *p-value* ≤ 0.05).

- ➢ Reject the null hypothesis.
- ➢ There is a statistically significant difference.

How small is small enough?

Standard of 5% = 0.05 is called level of significance (α).

- Sometimes use a different value, like 0.01 or 0.10.

The level of significance (α): is the probability of rejection the null hypothesis (H_0) when it is true. For example, a significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference [Probability (wrong decision)].

P-value: is the probability of finding the observed, or more extreme, results when the null hypothesis (H_0) is true.

Step 5: Make a conclusion in the context of the situation:

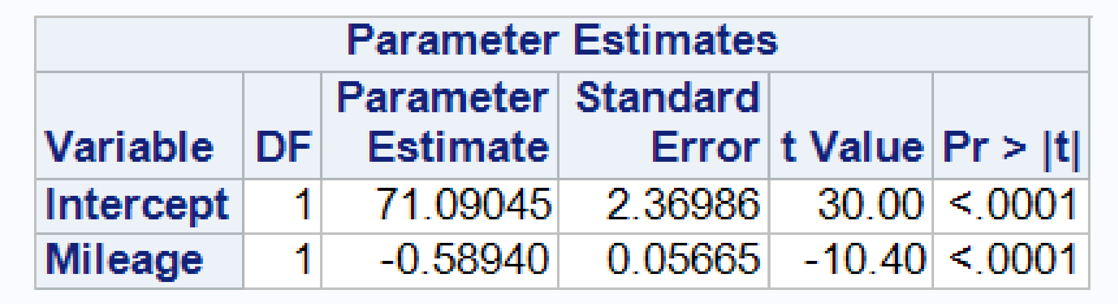
- It is important to answer the research question of interest.
- Do not just stop with whether or not to reject the null hypothesis – explain what that means in context.

2.1 Inference for Regression Slope:

Example 1: *(Porsche prices)*

For the same dataset *Porsche prices.csv*

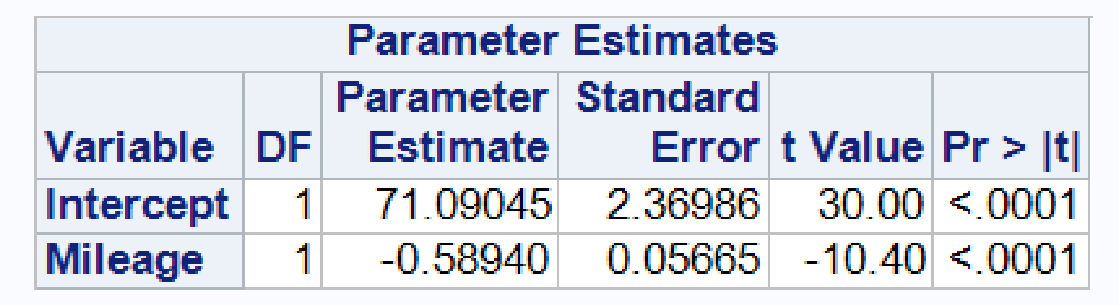
1. Using SAS, find the parameter estimates table.



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	71.09045	2.36986	30.00	<.0001
Mileage	1	-0.58940	0.05665	-10.40	<.0001

Parameter estimates:

$$\hat{\beta}_0 = 71.09045 \quad and \quad \hat{\beta}_1 = -0.58940$$



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	71.09045	2.36986	30.00	<.0001
Mileage	1	-0.58940	0.05665	-10.40	<.0001

The standard error (SE) for the parameters estimates:

- The standard error of the estimate is a measure of the accuracy of predictions.

$$SE_{\widehat{\beta}_0} = 2.36986 \quad and \quad SE_{\widehat{\beta}_1} = 0.05665$$

The (test statistic) T-test values for the parameters:

$$t = \frac{\hat{\beta}_0}{SE_{\widehat{\beta}_0}} = \frac{71.09045}{2.36986} = 30.00 \quad and$$

$$t = \frac{\hat{\beta}_1}{SE_{\widehat{\beta}_1}} = \frac{-0.58940}{0.05665} = -10.40$$

Example 1: *(Porsche prices)*

2. Describe the results of the significance test for the slope, using α = 0.05. Give the hypothesis being tested, the test statistics, the p-value, and your conclusion in sentence form.

i. The hypotheses: $H_0: \beta_1 = 0 \quad vs \quad H_1: \beta_1 \neq 0$

ii. The test statistic (t-test): $t = -10.40$

iii. P-value: $p - value < 0.0001$

iv. The conclusion:

since p-value < 0.0001 < 0.05 = α, so we reject H_0 which means that there is a significant linear relationship between the price and mileage of the car.

Confidence Interval for the Slope of a Simple Linear Model:

Estimation refers to the process by which one makes inferences about a population, based on information obtained from a sample.

Statisticians use sample statistics to estimate population parameters.

Two types of estimation:

Point estimate: A point estimate of a population parameter is a single value of a statistic.

Interval estimate: An interval estimate is defined by two numbers, between which a population parameter is said to lie.

Confidence Interval for the Slope of a Simple Linear Model:

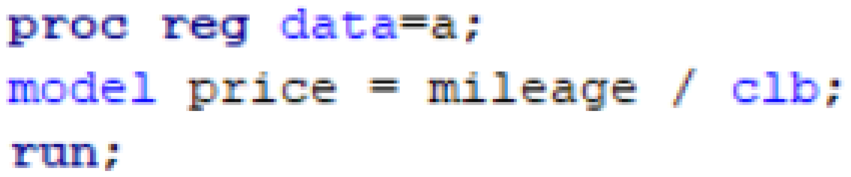
The confidence interval is a range of values we are fairly sure our true value lies in.

The confidence interval for β_1 has the form.

$$\hat{\beta}_1 \pm t_{(n-2)} . SE_{\hat{\beta}_1}$$

where t is the critical value from t table.

- The slope of the population (β_1) is usually the most important parameter in a simple regression problem.

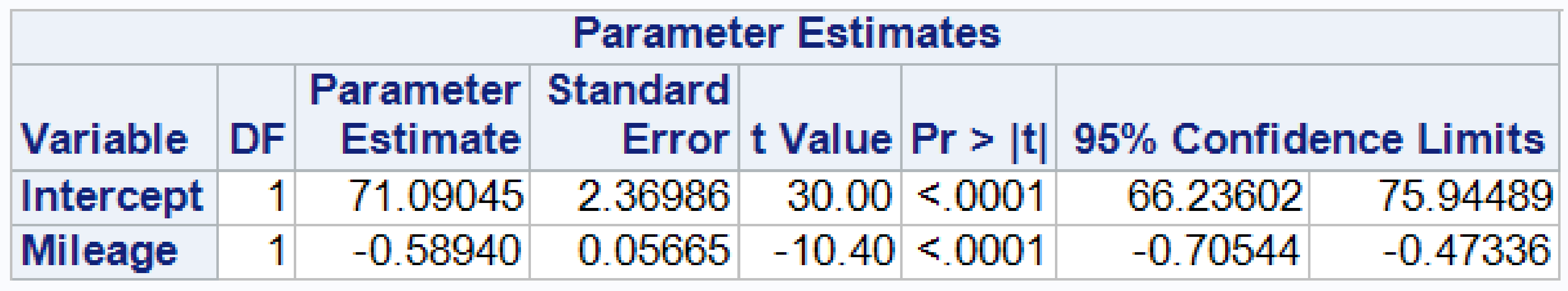


```
proc reg data=a;
model price = mileage / clb;
run;
```

Example 2: *(Porsche prices)*

For the same dataset *Porsche prices.csv*

1. Using SAS, find the 95% confidence interval for the population slope (mileage).



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	71.09045	2.36986	30.00	<.0001	66.23602	75.94489
Mileage	1	-0.58940	0.05665	-10.40	<.0001	-0.70544	-0.47336

the 95% confidence interval for the population slope is

$$\hat{\beta}_1 \pm t_{(n-2)} . SE_{\hat{\beta}_1}$$

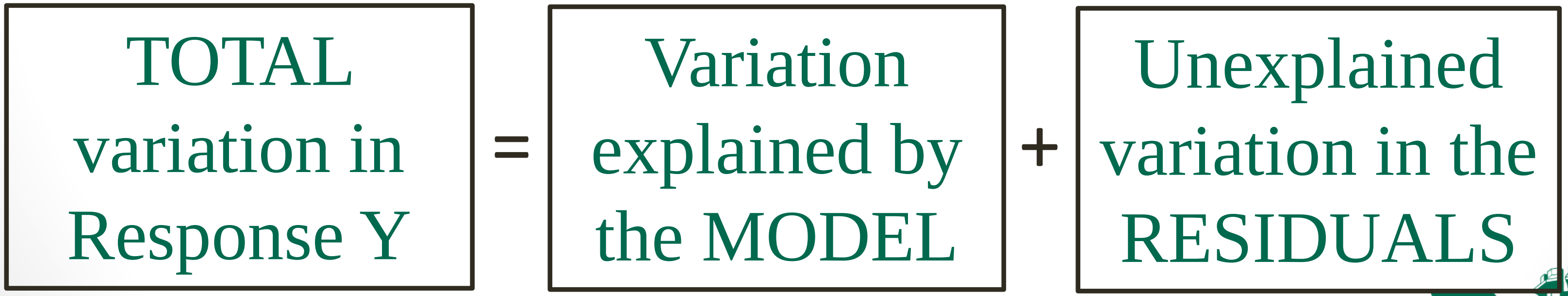
$$= -0.5894 \pm (2.05)(0.05665)$$

$$= (-0.70544, -0.47336)$$

2.2 Partition Variability - ANOVA:

Another way to assess the effectiveness of a model is to measure how much of the variability in the response variable is explained by the predictions based on the fitted model. This general technique is know in statistics as analysis of variance - ANOVA.

The basic idea is to partition the total variability in the responses into two pieces.



TOTAL variation in Response Y = Variation explained by the MODEL + Unexplained variation in the RESIDUALS

We summarize the partition with the following notation (SS means sum squares):

$$SSTotal = SSModel + SSError$$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

where

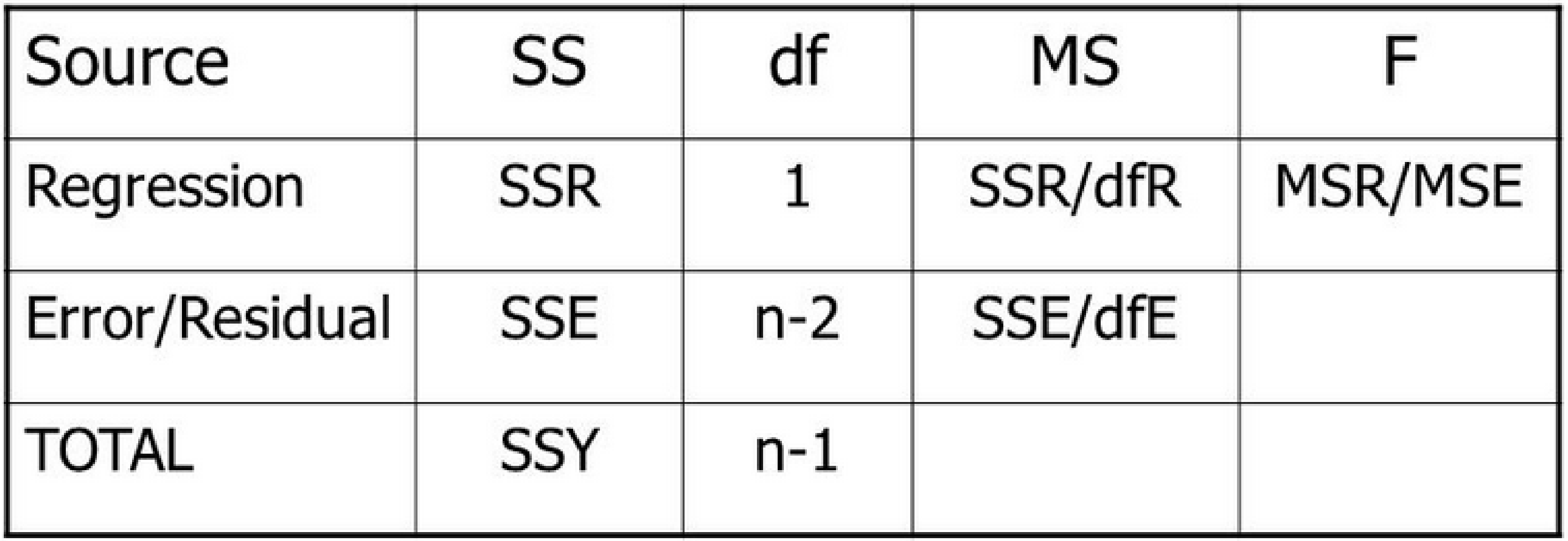
Y_i: the original values of the response variable.

$\hat{Y}_i$: the predicted values of the response variable using the regression model.

$\bar{Y}$: the mean (average) of the original values of the response variable.

ANOVA table for simple linear regression is used to test the following hypothesis:

$$H_0: \beta_1 = 0 \quad vs \quad H_1: \beta_1 \neq 0$$



Source	SS	df	MS	F
Regression	SSR	1	SSR/dfR	MSR/MSE
Error/Residual	SSE	n-2	SSE/dfE	
TOTAL	SSY	n-1		

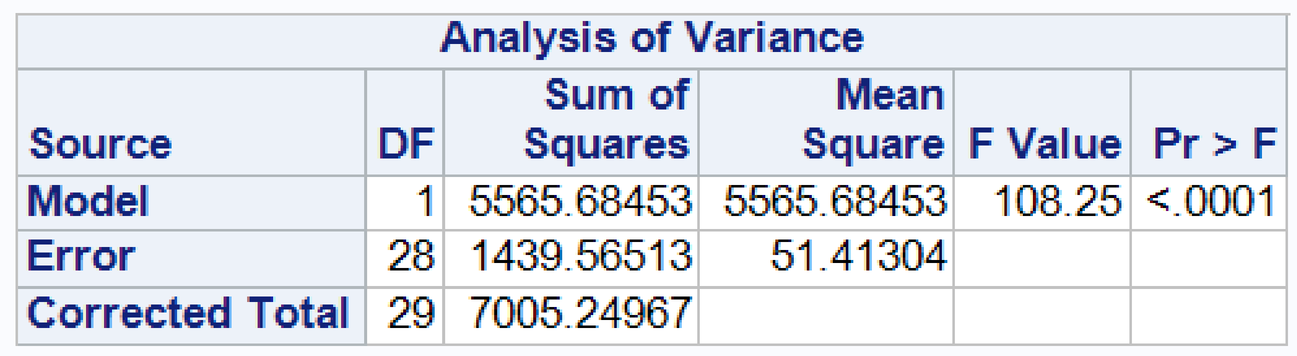
where

n: the sample size (# of observations in the sample)

Example 3: *(Porsche prices)*

For the same dataset *Porsche prices.csv*

1. Using SAS, find ANOVA table.



Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	5565.68453	5565.68453	108.25	<.0001
Error	28	1439.56513	51.41304		
Corrected Total	29	7005.24967			

$n = 30$

df of Model = number of predictors = 1.

df of Error = $n - 2(sample\ size\ - \#of\ predictors\ - 1) = 28.$

df of Total = $n - 1(sample\ size\ - 1) = 29.$

Mean square (Model) = $\dfrac{SS(Model)}{df(Model)} = \dfrac{5565.68453}{1} = 5565.68453$

Mean square (Error) = $\dfrac{SS(Error)}{df(Error)} = 51.41304$

F-Value = $\dfrac{MS(Model)}{MS(Error)} = \dfrac{5565.68453}{51.41304} = 108.25$

Degrees of Freedom:

Degrees of freedom aren’t easy to explain. They come up in many different contexts in statistics- some advanced and complicated.

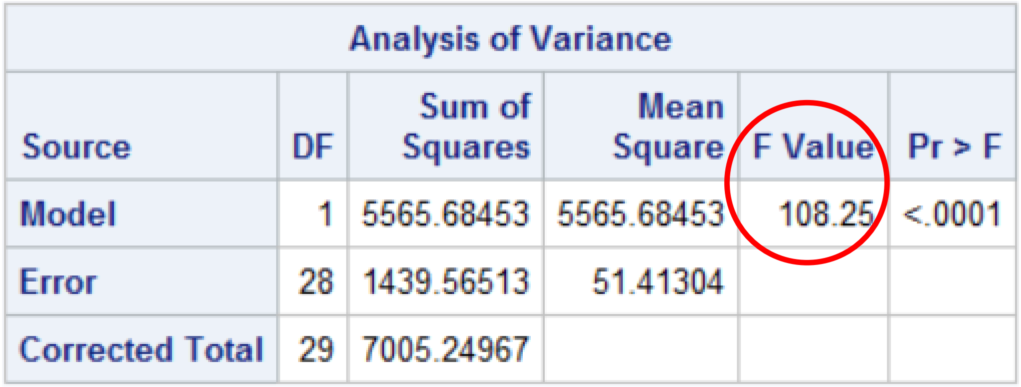
Degrees of freedom generally equals the number of observations (or pieces of information) minus the number of parameters estimated.

Example 3: *(Porsche prices)*

2. Using ANOVA table, interpret *p-value*. Write your conclusion.

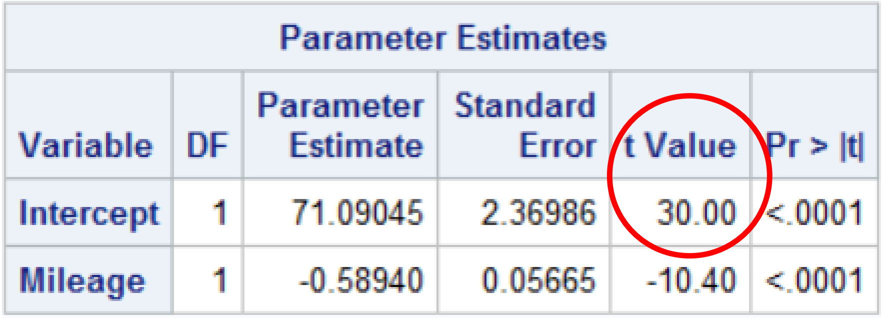
since p-value < 0.0001 < 0.05 = α, so we reject H_0 which means that there is a significant linear relationship between the price and mileage of the car.

Comparing the results of Analysis of Variance table and Parameter Estimates table: *(Porsche prices)*



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5565.68453	5565.68453	108.25	<.0001
Error	28	1439.56513	51.41304		
Corrected Total	29	7005.24967			



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	71.09045	2.36986	30.00	<.0001
Mileage	1	-0.58940	0.05665	-10.40	<.0001

$$F\ value = (t\ value)^2 = (-10.404)^2 = 108.25$$

So we can use F value to make a decision.

Note: That can happen only in the simple linear regression.

Reading Assignment

Read section 2.1 - 2.2