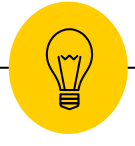
Linear Regression



Group Members:
Yangzhou Tang
Isabella Zhai
Monty Xu
Summer Zhang

Agenda

Package Introduction & Initial Regression - Yangzhou

Linear regression and python packages introduction

T-test Interpretation - Isabella

Interpretation on significance test and regression coefficients

Categorical Variables - Monty

Categorical variable in linear regression

ANOVA Tests - Summer

ANOVA test definition & three types of ANOVA analysis

1

Package Introduction & Initial Regression

Idea of Linear Regression

- A type of predictive analysis
- Simple Linear Regression: $y=c+b*x$
- Multiple Linear Regression: $y=c+b_1*x_1+b_2*x_2+b_3*x_3$

LR modeling in Python

- OLS
- package 1 : pandas
- package 2: statsmodels
- Store data in a dataframe

```
import pandas
import statsmodels.formula.api as smf
df_KBBD=pd.read_csv('KelleyBlueBookData.csv')
```

LR modeling in Python

- Choose predictors
- Choose dependent variable
- The coefficients we want

```
reg = smf.ols('Price ~ Mileage + Type + C(Cylinder) + Liter + Cruise + Sound + Leather', data=df_KBBD).fit()
reg.summary()
```

OLS Regression Results

Dep. Variable:	Price	**R-squared:**	0.738
Model:	OLS	**Adj. R-squared:**	0.734
Method:	Least Squares	**F-statistic:**	202.6
Date:	Tue, 05 Oct 2021	**Prob (F-statistic):**	1.51e-221
Time:	16:12:01	**Log-Likelihood:**	-7998.0
No. Observations:	804	**AIC:**	1.602e+04
Df Residuals:	792	**BIC:**	1.608e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.943e+04	608.357	18.300	0.000	2.63e+04	3.26e+04
Type[T.Coupe]	-1.857e+04	874.134	-21.244	0.000	-2.03e+04	-1.69e+04
Type[T.Hatchback]	-1.831e+04	085.146	-16.873	0.000	-2.04e+04	-1.62e+04
Type[T.Sedan]	-1.547e+04	799.994	-19.336	0.000	-1.7e+04	-1.39e+04
Type[T.Wagon]	-9452.1225	000.020	-9.452	0.000	-1.14e+04	-7489.120
C(Cylinder)[T.6]	1360.1311	075.453	1.265	0.206	-750.943	3471.205
C(Cylinder)[T.8]	1.416e+04	959.004	7.231	0.000	1.03e+04	1.8e+04
Mileage	-0.1871	0.022	-8.505	0.000	-0.230	-0.144
Liter	1115.8414	621.236	1.796	0.073	-103.622	2335.305
Cruise	4650.7921	473.626	9.820	0.000	3721.081	5580.504
Sound	14.7921	404.338	0.037	0.971	-778.909	808.493
Leather	1677.9449	433.225	3.873	0.000	827.541	2528.349

2

T-test Interpretation

Significance Test (t test)

- Linear relationship
- If slope ≠ 0

$$y=c+b_1*x_1+b_2*x_2+b_3*x_3$$

  - we will conclude that there is a significant relationship

Significance Test (t test)

- 1) State the Hypothesis
  - Ho: B1 = 0
  - Ha: B1 ≠ 0
- 2) Analysis Plan
  - alpha = 0.05
- 3) Analyze Data
  - P-value vs. Significance level
  - Leather: p<0.000
- 4) Interpret results

```
reg = smf.ols('Price ~ Mileage + Type + C(Cylinder) + Liter + Cruise + Sound + Leather', data=df_KBBD).fit()
reg.summary()
```

OLS Regression Results

Dep. Variable:	Price	**R-squared:**	0.738
Model:	OLS	**Adj. R-squared:**	0.734
Method:	Least Squares	**F-statistic:**	202.6
Date:	Tue, 05 Oct 2021	**Prob (F-statistic):**	1.51e-221
Time:	16:12:01	**Log-Likelihood:**	-7998.0
No. Observations:	804	**AIC:**	1.602e+04
Df Residuals:	792	**BIC:**	1.608e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.943e+04	1608.357	18.300	0.000	2.63e+04	3.26e+04
Type[T.Coupe]	-1.857e+04	874.134	-21.244	0.000	-2.03e+04	-1.69e+04
Type[T.Hatchback]	-1.831e+04	1085.146	-16.873	0.000	-2.04e+04	-1.62e+04
Type[T.Sedan]	-1.547e+04	799.994	-19.336	0.000	-1.7e+04	-1.39e+04
Type[T.Wagon]	-9452.1225	1000.020	-9.452	0.000	-1.14e+04	-7489.120
C(Cylinder)[T.6]	1360.1311	1075.453	1.265	0.206	-750.943	3471.205
C(Cylinder)[T.8]	1.416e+04	1959.004	7.231	0.000	1.03e+04	1.8e+04
Mileage	-0.1871	0.022	-8.505	0.000	-0.230	-0.144
Liter	1115.8414	621.236	1.796	0.073	-103.622	2335.305
Cruise	4650.7921	473.626	9.820	0.000	3721.081	5580.504
Sound	14.7921	404.338	0.037	0.971	-778.909	808.493
Leather	1677.9449	433.225	3.873	0.000	827.541	2528.349

3

Categorical Variable

Categorical Variable

- Definition of Categorical Variable
  - A categorical variable is one that has two or more categories.
    1) Gender - Female/ Male
    2) Interest Rate - Low/ Median / High

Categorical Variable

- Check Categorical Variable
  - Does “integer” mean “Cylinder” is numeric data ?
    We don’t know.

```
kelleydata.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 804 entries, 0 to 803
Data columns (total 12 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Price     804 non-null    float64
 1   Mileage   804 non-null    int64
 2   Make      804 non-null    object
 3   Model     804 non-null    object
 4   Trim      804 non-null    object
 5   Type      804 non-null    object
 6   Cylinder  804 non-null    int64
 7   Liter     804 non-null    float64
 8   Doors     804 non-null    int64
 9   Cruise    804 non-null    int64
 10  Sound     804 non-null    int64
 11  Leather   804 non-null    int64
dtypes: float64(2), int64(6), object(4)
memory usage: 75.5+ KB
```

Categorical Variable

- Check Categorical Variable
  - [column].value_counts()
  - “Cylinder” has 3 categories.

```
kelleydata['Cylinder'].value_counts()
```

```
4    394
6    310
8    100
Name: Cylinder, dtype: int64
```

Categorical Variable

- Recode Categorical Variable
  ---->> Dummy Variables
  - Using (k-1) dummy variables to model the categorical variable with K levels.

Cylinder_type	Dummy_variable1	Dummy_variable2
Cylinder4	0	0
Cylinder6	1	0
Cylinder8	0	1

C(Cylinder)[T.6]	1360.1311	1075.453	1.265	0.206	-750.943	3471.205
C(Cylinder)[T.8]	1.416e+04	1959.004	7.231	0.000	1.03e+04	1.8e+04
Mileage	-0.1871	0.022	-8.505	0.000	-0.230	-0.144
Liter	1115.8414	621.236	1.796	0.073	-103.622	2335.305
Cruise	4650.7921	473.626	9.820	0.000	3721.081	5580.504
Sound	14.7921	404.338	0.037	0.971	-778.909	808.493
Leather	1677.9449	433.225	3.873	0.000	827.541	2528.349

Categorical Variable

- Recode Categorical Variable

  --->> Dummy Variables

```
reg2 = smf.ols('Price ~ Mileage + Type + C(Cylinder) + Liter + Cruise + Sound + Leather', data = kelleydata).fit()
reg2.summary()
```

  - Adding a “C” in front of the categorical variable to “tell” the model is a categorical variable.

3

ANOVA Test

Partial ANOVA Test

- Definition of ANOVA test
- Package
- Types of ANOVA tests

Partial ANOVA Test

- Definition of ANOVA test
  - ANOVA (analysis of variance), is a statistical method that separates observed data into two parts:
    1) systematic and - statistical influence
    2) random factors - no statistical influence
  - Purpose: determine the influence that independent variables have on the dependent variables

Partial ANOVA Test

- Definition of ANOVA test
- Package
  - import statsmodels.api as sm
  - calling function: stats.anova_lm
  - sm.stats.anova_lm (data, typ=1, or 2 or 3)

Partial ANOVA Test

- Definition of ANOVA test
- Package
- Types of ANOVA Tests
  - Type 1: sm.stats.anova_lm (data, typ=1)

	df	sum_sq	mean_sq	F	PR(>F)
Type	4.0	2.409164e+10	6.022911e+09	231.858421	1.033148e-131
C(Cylinder)	2.0	2.901814e+10	1.450907e+10	558.542116	4.880181e-152
Mileage	1.0	1.730471e+09	1.730471e+09	66.616337	1.294144e-15
Liter	1.0	2.334144e+08	2.334144e+08	8.985537	2.806297e-03
Cruise	1.0	2.408427e+09	2.408427e+09	92.714963	7.920459e-21
Sound	1.0	1.607876e+07	1.607876e+07	0.618969	4.316660e-01
Leather	1.0	3.896844e+08	3.896844e+08	15.001317	1.163062e-04
Residual	792.0	2.057353e+10	2.597668e+07	NaN	NaN

H0: Y ~ Type
H 1: Y ~ Type + Cylinder

Partial ANOVA Test

- Definition of ANOVA test
- Package
- Types of ANOVA Tests
  - Type 1: sm.stats.anova_lm (data, typ=2)

	sum_sq	df	F	PR(>F)
Type	1.381333e+10	4.0	132.939757	7.297205e-87
C(Cylinder)	5.348559e+09	2.0	102.949252	1.807044e-40
Mileage	1.879060e+09	1.0	72.336414	8.996777e-17
Liter	8.380606e+07	1.0	3.226204	7.284949e-02
Cruise	2.504757e+09	1.0	96.423312	1.485766e-21
Sound	3.476617e+04	1.0	0.001338	9.708262e-01
Leather	3.896844e+08	1.0	15.001317	1.163062e-04
Residual	2.057353e+10	792.0	NaN	NaN

H0: Y ~ Type + Mileage + Liter + Cruise + Sound + Leather + Residual
H 1: Y ~ Type + Cylinder + Mileage + Liter + Cruise + Sound + Leather + Residual

Partial ANOVA Test

- Definition of ANOVA test
- Package
- Types of ANOVA Tests
  - Type 1: sm.stats.anova_lm (data, typ=3)

	sum_sq	df	F	PR(>F)
Intercept	8.699699e+09	1.0	334.904231	1.154930e-62
Type	1.381333e+10	4.0	132.939757	7.297205e-87
C(Cylinder)	5.348559e+09	2.0	102.949252	1.807044e-40
Mileage	1.879060e+09	1.0	72.336414	8.996777e-17
Liter	8.380606e+07	1.0	3.226204	7.284949e-02
Cruise	2.504757e+09	1.0	96.423312	1.485766e-21
Sound	3.476617e+04	1.0	0.001338	9.708262e-01
Leather	3.896844e+08	1.0	15.001317	1.163062e-04
Residual	2.057353e+10	792.0	NaN	NaN

H0: Y ~ Intercept + Type + Mileage + Liter + Cruise + Sound + Leather + Residual
H 1: Y ~ Intercept + Type + Cylinder + Mileage + Liter + Cruise + Sound + Leather + Residual

Thanks!

Q & A