IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

REGRESSION

Linear Regression-II

Dr. Ramesh Anbanandam

DEPARTMENT of Management Studies

Least Squares Method

- Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Sum of squares and sum of cross-products

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})\ (y_i - \bar{y})$$

Sum of squares and sum of cross-products

$$Slope(m) = \frac{S_{xy}}{S_{xx}}$$

$$\text{SSE} = \text{error sum of squares} = S_{yy} - \frac{S_{xy}}{S_{xx}}$$

Least Squares Method

y-Intercept for the Estimated Regression Equation

$$b_0 = \overline{y} - b_1\overline{x}$$

where:

x_i = value of independent variable for *i*th observation

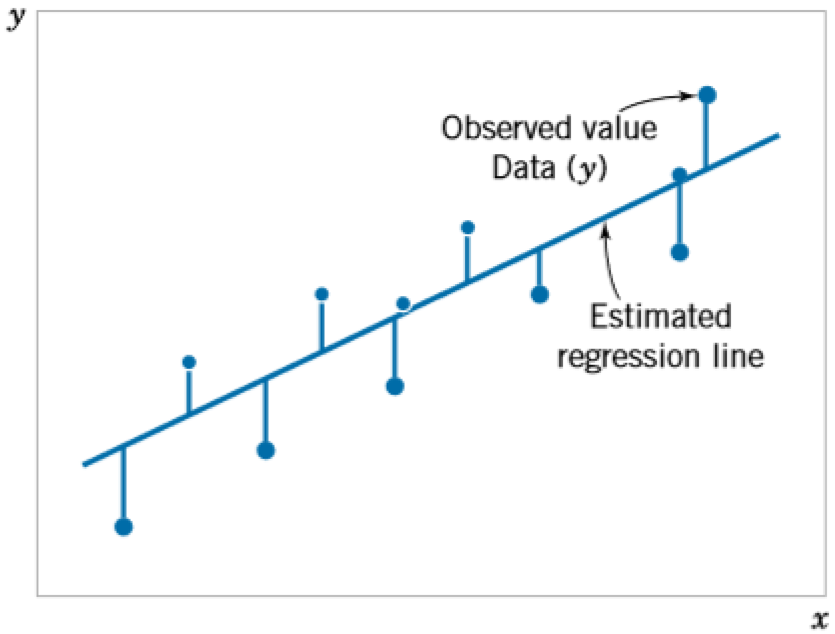
y_i = value of dependent variable for *i*th observation

$\overline{x}$ = mean value for independent variable

$\overline{y}$ = mean value for dependent variable

n = total number of observations

Simple Linear Regression



Deviation from the estimated regression model

Simple Linear Regression

Example: Auto Sales

An Auto company periodically has a special week-long sale.

As part of the advertising campaign runs one or more television commercials during the weekend preceding the sale.

Data from a sample of 5 previous sales are shown on the next slide.

Simple Linear Regression

Example: Auto Sales

Number of TV Ads	Number of Cars Sold
1	14
3	24
2	18
1	17
3	27

Estimated Regression Equation

Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{20}{4} = 5$$

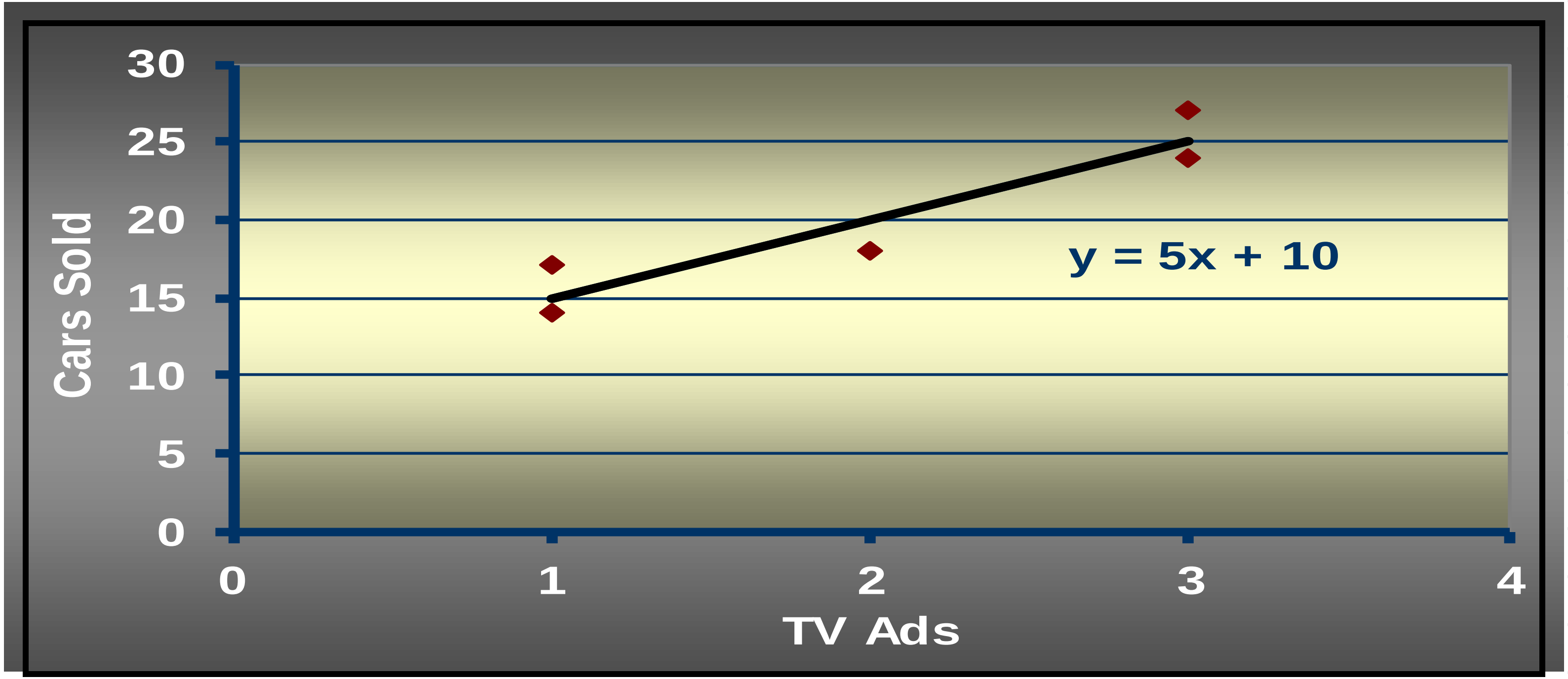
y-Intercept for the Estimated Regression Equation

$$b_0 = \overline{y} - b_1 \overline{x} = 20 - 5(2) = 10$$

Estimated Regression Equation

$$\hat{y} = 10 + 5x$$

Scatter Diagram and Trend Line



Jupyter Code

```
In [2]: import numpy as np
        import matplotlib.pyplot as plt
```

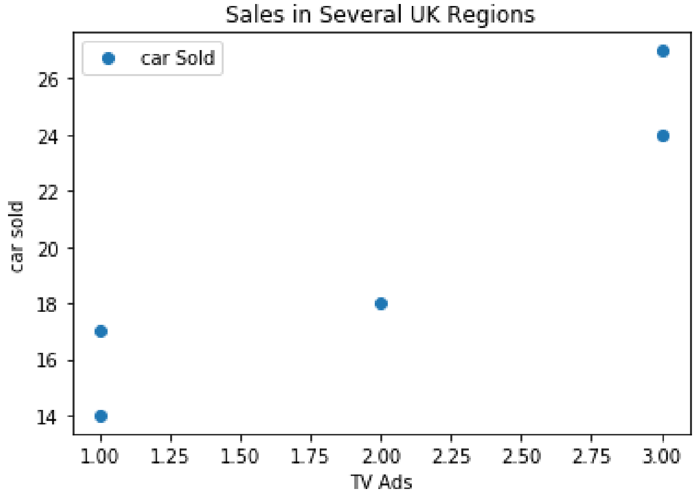
```
In [3]: import seaborn as sns
```

```
In [4]: import pandas as pd
        import matplotlib as mpl
        import statsmodels.formula.api as sm
        from sklearn.linear_model import LinearRegression
        from scipy import stats
```

```
In [5]: tbl = pd.read_excel('C:/Users/Somi/Documents/regr.xlsx')
```

Jupyter Code

In [6]:
```
tbl.plot('TV Ads', 'car Sold', style='o')
plt.ylabel('car sold')
plt.title('Sales in Several UK Regions')
plt.show()
```



Jupyter code

In [5]:
```python
t= tbl['TV Ads']
c= tbl['car Sold']
```

In [8]:
```python
import statsmodels.api as s
t = s.add_constant(t)
model1 = sm.OLS(c,t)
result1 = model1.fit()
print(result1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               car Sold   R-squared:                       0.877
Model:                            OLS   Adj. R-squared:                  0.836
Method:                 Least Squares   F-statistic:                     21.43
Date:                Fri, 30 Aug 2019   Prob (F-statistic):             0.0190
Time:                        08:31:20   Log-Likelihood:                -9.6687
No. Observations:                   5   AIC:                             23.34
Df Residuals:                       3   BIC:                             22.56
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         10.0000      2.366      4.226      0.024       2.469      17.531
TV Ads         5.0000      1.080      4.629      0.019       1.563       8.437
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   1.214
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.674
Skew:                           0.256   Prob(JB):                        0.714
Kurtosis:                       1.276   Cond. No.                         6.33
==============================================================================
```

Example Problem- II

- The data in the file hardness.xls provide measurements on the hardness and tensile strength for 35 specimens of die-cast aluminum.
- It is believed that hardness (measured in Rockwell E units) can be used to predict tensile strength (measured in thousands of pounds per square inch).

a. Construct a scatter plot.

b. Assuming a linear relationship, use the least-squares method to find the regression coefficients b_0 and b_1.

c. Interpret the meaning of the slope, b_1, in this problem.

d. Predict the mean tensile strength for die-cast aluminum that has a hardness of 30 Rockwell E units.

Tensile strength	Hardness
53	29.31
70.2	34.86
84.3	36.82
55.3	30.12
78.5	34.02
63.5	30.82
71.4	35.4
53.4	31.26
82.5	32.18
67.3	33.42
69.5	37.69
73	34.88
55.7	24.66
85.8	34.76
95.4	38.02
51.1	25.68
74.4	25.81
54.1	26.46
77.8	28.67
52.4	24.64
69.1	25.77
53.5	23.69
64.3	28.65
82.7	32.38
55.7	23.21
70.5	34
87.5	34.47
50.7	29.25
72.3	28.71
59.5	29.83
71.3	29.25
52.7	27.99
76.5	31.85
63.7	27.65
69.2	31.7

```
In [30]: mean_squared_error(y_test, y_predict)
Out[30]: 35.71053398209997

In [31]: reg.score(x_test, y_test)
Out[31]: 0.5362243730094254

In [32]: reg.score(x_train, y_train)
Out[32]: 0.4500146647765303
```

Thank You