Linear Regression

Part 2

Assumptions of the Linear Regression Model

- Errors follow normal distribution
- Variance of error
  - Homoscedasticity.
  - Heteroscedasticity.
- The error and independent variable are uncorrelated.
- Relationship between outcome and variable feature defined correctly.
-

MODEL DIAGNOSTICS

- Coefficient of determination (R-squared).
- Hypothesis test for the regression coefficient.
- Analysis of variance for overall model validity
- Residual analysis to validate the regression model assumptions.
- Outlier analysis

Co-efficient of Determination (R-Squared or R2)

$$\underbrace{Y_i}_{\text{Variation in } Y} = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Variation in } Y \text{ explained by the model}} + \underbrace{\varepsilon_i}_{\text{Variation in } Y \text{ not explained by the model}}$$

$$\underbrace{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2}_{SSE}$$

$$R\text{-Squared} = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

The value of R-squared lies between 0 and 1
R-squared (R_2) is square of correlation coefficient ($R_2 = r_2$),
Higher R-squared indicates better fit

Hypothesis Test for the Regression Coefficient

- Existence of a linear relationship between the outcome variable and the feature
- It is proved that sampling distribution of b1 is a t-distribution

$$H_0: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

$$t_{\alpha/2,n-2} = \frac{\widehat{\beta}_1}{S_e(\widehat{\beta}_1)}$$

$$S_e(\widehat{\beta}_1) = \frac{S_e}{\sqrt{(X_i - \overline{X})^2}}$$

standard error of estimate of the regression coefficient

$$S_e = \sqrt{\frac{(Y_i - \widehat{Y}_i)}{n-2}}$$

S_e is the standard error of the estimated value of Y_i

Hypothesis Test for the Regression Coefficient

Hypothesis Testing for the Regression Coefficient – Step-by-Step Explanation

1. Understanding the Regression Coefficient b_1

When we perform a linear regression, we are trying to find the relationship between a dependent variable Y (the outcome) and an independent variable X (the feature). The equation of a simple linear regression is:

$$Y = b_0 + b_1 X + \varepsilon$$

where:

- b_0 is the **intercept** (the value of Y when $X = 0$),
- b_1 is the **regression coefficient** (it tells us how much Y changes when X increases by 1 unit),
- ε is the **random error** (the part of Y that is not explained by X).

Hypothesis Test for the Regression Coefficient

2. What Does Hypothesis Testing for b_1 Mean?

We want to check if b_1 is **statistically different from zero.** If $b_1 = 0$, it means there is **no real linear relationship** between X and Y.

To test this, we set up two hypotheses:

- **Null Hypothesis** (H_0): $b_1 = 0$
  → There is **no significant linear relationship** between X and Y.
- **Alternative Hypothesis** (H_A): $b_1 \neq 0$
  → There **is** a significant linear relationship between X and Y.

Hypothesis Test for the Regression Coefficient

3. The Test Statistic for b_1

To test if b_1 is significantly different from 0, we calculate a test statistic t:

$$t = \frac{b_1}{\text{Standard Error of } b_1}$$

This formula tells us **how many standard errors away** our estimated b_1 is from 0. If the value of t is large (either positive or negative), it suggests that b_1 is significantly different from 0.

4. The Standard Error of b_1

The **standard error** of b_1, denoted as S_{b_1}, measures how much the estimated b_1 would vary if we repeated the study multiple times. It is given by:

$$S_{b_1} = \frac{S_e}{\sqrt{\sum (X_i - \bar{X})^2}}$$

where:

- S_e is the **standard error of the residuals**, which tells us how much the actual Y values vary around the predicted Y.
- $\sum (X_i - \bar{X})^2$ is the sum of squared deviations of X from its mean, which accounts for how spread out the X values are.

Hypothesis Test for the Regression Coefficient

5. The Distribution of the Test Statistic

The test statistic t follows a **t-distribution** with $n - 2$ degrees of freedom (since we estimate two parameters, b_0 and b_1, we lose 2 degrees of freedom).

- If $|t|$ is **large**, we reject H_0, meaning X has a significant effect on Y.
- If $|t|$ is **small**, we fail to reject H_0, meaning there is **not enough evidence** to say that X and Y are related.

6. Conclusion

The hypothesis test for b_1 helps us determine if our independent variable X is meaningfully related to Y. If we reject the null hypothesis, we conclude that there is a **statistically significant** relationship between X and Y. If not, we assume that any observed relationship might just be due to random chance.

Residual Analysis

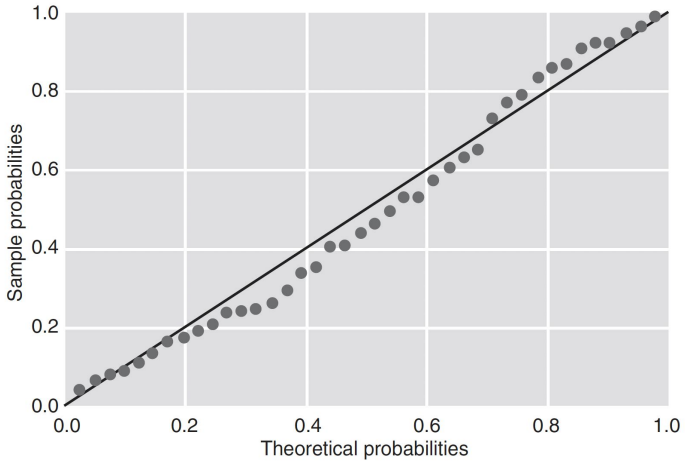
- Residuals are normally distributed
- Variance of a residual is constant
- The functional form of regression is correctly specified
- no outliers.

Residuals are normally distributed

Check for Normal Distribution of Residual

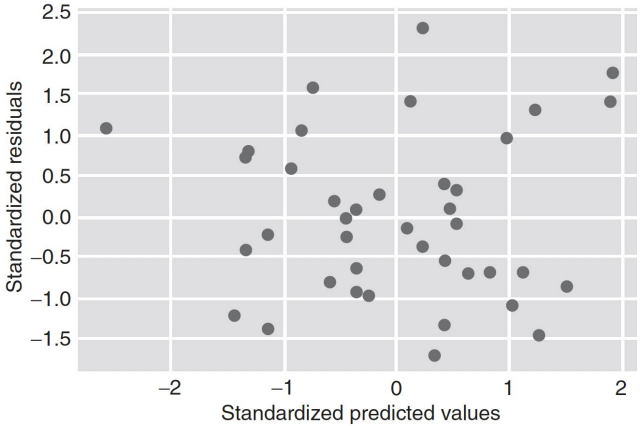
- probability−probability plot
- cumulative distribution function of two probability distributions against each other.

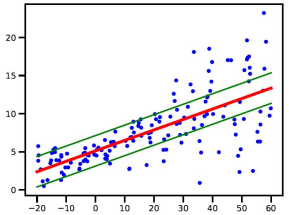
  - Normal Distribution
  - Distribution of residuals

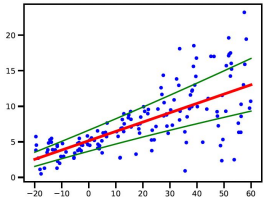


Test of Homoscedasticity

- Residuals have constant variance (homoscedasticity) across different values of the predicted value (Y).
- drawing a residual plot
  - Plot between standardized residual value vs standardized predicted value







Dataset

EXAMPLE **Predicting MBA Salary from Grade in 10th Marks**

Table 4.1 contains the salary of 50 graduating MBA students of a Business School in 2016 and their corresponding percentage marks in grade 10 (File name: *MBA Salary.csv*). Develop an SLR model to understand and predict salary based on the percentage of marks in Grade 10.

TABLE 4.1 Salary of MBA students versus their grade 10 marks

S. No.	Percentage in Grade 10	Salary	S. No.	Percentage in Grade 10	Salary
1	62.00	270000	26	64.60	250000
2	76.33	200000	27	50.00	180000
3	72.00	240000	28	74.00	218000
4	60.00	250000	29	58.00	360000
5	61.00	180000	30	67.00	150000
6	55.00	300000	31	75.00	250000
7	70.00	260000	32	60.00	200000
8	68.00	235000	33	55.00	300000

Outlier Analysis

- Outlier = large deviation from mean value
- Problem → influence the value of regression coefficient
- Important to identify them
- Use distance measures to identify them
  - Z-Score
  - Mahalanobis distance
  - Cook’s Distance
  - Leverage Values

Z-Score

- Z Score = standardized distance of an observation from its mean value.

$$Z = \frac{Y_i - \bar{Y}}{\sigma_Y}$$

Any observation more than 3 z-score is an outlier.

```
from scipy.stats import zscore
```

```
mba_salary_df['z_score_salary'] = zscore( mba_salary_df.Salary )
```

```
mba_salary_df[ (mba_salary_df.z_score_salary > 3.0) | (mba_salary_
df.z_score_salary < -3.0) ]
```

S. No.	Percentage in Grade 10	Salary	z_score_salary

Cook’s Distance

- The change in predicted value of dependent variable when a particular sample is excluded from the sample used for estimating parameters.
- Cook’s distance value of more than 1 indicates highly influential observation.

```
import numpy as np

mba_influence = mba_salary_lm.get_influence()
(c, p) = mba_influence.cooks_distance

plt.stem(np.arange( len( train_X) ),
         np.round( c, 3 ),
         markerfmt=”,” );
plt.title( “Figure 4.3 - Cooks distance for all observations in MBA
           Salaray data set” );
plt.xlabel(“Row index”)
plt.ylabel(“Cooks Distance”);
```

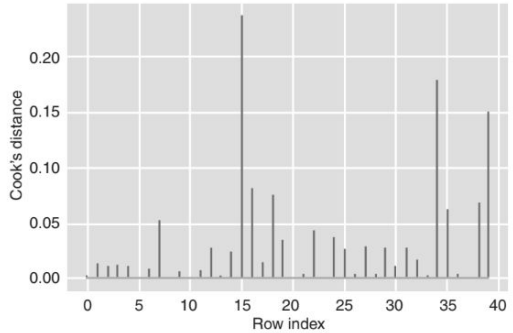


FIGURE 4.3 Cook’s distance for all observations in MBA salary dataset.

Leverage Values

- Measures an observation's influence on the regression fit.
- Related to **Mahalanobis Distance** (distance from the mean in a multivariate space).
- High leverage points can significantly affect the regression model.
- An observation is highly influential if its leverage value exceeds
    - $3(k+1)/n$
        - k= Number of features,
        - n = Sample size.
- Such points can disproportionately affect the regression model.

Leverage Values

```
from statsmodels.graphics.regressionplots import influence_plot

fig, ax = plt.subplots( figsize=(8,6) )
influence_plot( mba_salary_lm, ax = ax )
plt.title("Figure 4.4 - Leverage Value Vs Residuals")
plt.show();
```

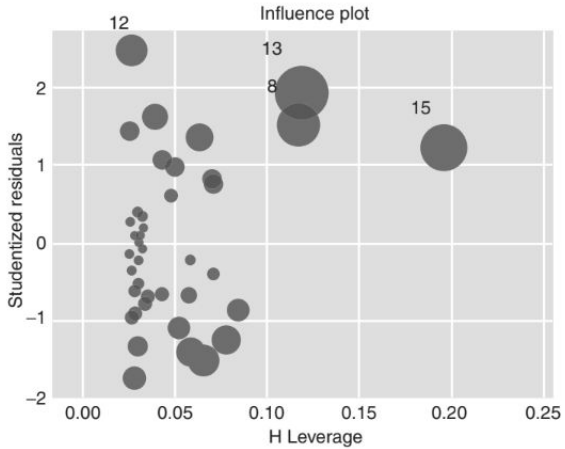


FIGURE 4.4 Leverage value versus residuals.

Making Prediction and Measuring Accuracy

Predicting using the Validation Set

```
pred_y = mba_salary_lm.predict( test_X )
```

Making Prediction and Measuring Accuracy

- Mean Square Error (MSE),
- Root Mean Square Error (RMSE)
- Mean Absolute Percentage Error (MAPE)

```
from sklearn.metrics import r2_score, mean_squared_error
```

```
np.abs(r2_score(test_y, pred_y))
```

```
0.15664584974230378
```

So, the model only explains 15.6% of the variance in the validation set.

```
import numpy
```

```
np.sqrt(mean_squared_error(test_y, pred_y))
```

```
73458.043483468937
```

Making Prediction and Measuring Accuracy

Calculating Prediction Intervals

- **Regression equation** provides a **point estimate** for the outcome variable.
- **Interval estimate** gives a range for YiY_iYi for a given explanatory variable.
- `wls_prediction_std()` returns the **prediction interval** during prediction.
- Requires **significance value (α\alphaα)** to compute the interval.
- α=0.1\alpha = 0.1α=0.1 gives a **90% confidence interval** for predictions.

Making Prediction and Measuring Accuracy

Calculating Prediction Intervals

```
from statsmodels.sandbox.regression.predstd import wls_prediction_std

# Predict the y values
pred_y = mba_salary_lm.predict( test_X )
```

```
# Predict the low and high interval values for y
_, pred_y_low, pred_y_high = wls_prediction_std( mba_salary_lm,
                    test_X,
                    alpha = 0.1)

# Store all the values in a dataframe
pred_y_df = pd.DataFrame( { 'grade_10_perc': test_X['Percentage in
                                                    Grade 10'],
                            'pred_y': pred_y,
                            'pred_y_left': pred_y_low,
                            'pred_y_right': pred_y_high } )
```

```
pred_y_df[0:10]
```

Making Prediction and Measuring Accuracy

Calculating Prediction Intervals

	grade_10_perc	pred_y	pred_y_left	pred_y_right
6	70.0	279828.402452	158379.832044	401276.972860
36	68.0	272707.227686	151576.715020	393837.740352
37	52.0	215737.829560	92950.942395	338524.716726
28	58.0	237101.353858	115806.869618	358395.838097
43	74.5	295851.045675	173266.083342	418436.008008
49	60.8	247070.998530	126117.560983	368024.436076
5	55.0	226419.591709	104507.444388	348331.739030
33	78.0	308313.101515	184450.060488	432176.142542
20	63.0	254904.290772	134057.999258	375750.582286
42	74.4	295494.986937	172941.528691	418048.445182

MULTIPLE LINEAR REGRESSION