

OLS Model Summary Reading

Ordinary Least Squares (OLS) is a widely used method for deriving a **linear regression equation**. The term "Least Squares" refers to the approach of minimizing the **sum of squared errors (SSE)**, which are the differences between actual and predicted values. The goal of OLS is to find the **best-fitting line** that minimizes these squared differences, making the model's predictions as accurate as possible. Graphically, the OLS regression line is the **closest possible line to all data points simultaneously**, reducing the overall error. The derivation of this **optimal regression line** is a **minimization problem** that relies on **calculus and linear algebra** to determine the **slope (β_1)** and **intercept (β_0)**. These parameters are chosen in a way that results in the **smallest sum of squared residuals**, ensuring the best possible fit to the data.

R-squared – R^2 measures how well the independent variables explain the variation in the dependent variable. It is calculated as SSR/SST . **SSR (Regression Sum of Squares)** - Difference between the predicted values and the mean of the dependent variable. **SST (Total Sum of Squares)** - Difference between the actual values and the mean of the dependent variable. **Interpretation:** 0 – no variability; 1 – all variability

Adjusted R-squared – Adjusted R^2 improves upon R^2 by adjusting for the number of independent variables in the model. Unlike R^2 , which always increases when new variables are added, Adjusted R^2 decreases if the new variable **does not contribute significantly** to the model. **Note:** Adjusted R^2 should only be used when comparing models trained on the **same dataset and target variable**.

F-statistic – The F-statistic helps determine if the independent variables together significantly impact the dependent variable. **Null Hypothesis (H_0)**: All regression coefficients (β values) = 0 (meaning no independent variable has an effect). **Alternative Hypothesis (H_1)**: At least one coefficient (β) $\neq 0$ (at least one independent variable significantly impacts the dependent variable). **Interpretation:** A higher F-statistic and a lower p-value indicate that at least one independent variable is important for prediction.

The **Durbin-Watson test** used to check autocorrelation; a value close to 2 means no autocorrelation. If the value is near 0 or 4, autocorrelation exists.

Model Summary

OLS Regression Results

Dep. Variable:

price

R-squared:

0.745

Model:

OLS

Adj. R-squared:

0.742

Method:

Least Squares

F-statistic:

285.9

Date:

Sun, 03 Nov 2024

Prob (F-statistic):

8.13e-31

Time:

11:17:07

Log-Likelihood:

-1198.3

No. Observations:

100

AIC:

2481.

Df Residuals:

98

BIC:

2486.

Df Model:

1

Covariance Type:

nonrobust

Coeff. table

coef

std err

t

P>|t|

[0.025

0.975]

→ const

1.019e+05

1.19e+04

8.550

0.000

7.83e+04

1.26e+05

→ size

223.1787

13.199

16.909

0.000

196.986

249.371

Omnibus:

6.262

Durbin-Watson:

2.267

Prob(Omnibus):

0.044

Jarque-Bera (JB):

2.938

Skew:

0.117

Prob(JB):

0.230

Kurtosis:

2.194

Cond. No.

2.75e+03

Some additional test

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly

[2] The condition number is large, 2.75e+03. This might indicate that there are

strong multicollinearity or other numerical problems.

Intercept b_0 of equation $y = b_0 + b_1x_1$
Variable of equation, i.e. b_1

std.err – It represents the average deviation between the predicted values and the actual values in the dataset. A **lower standard error** indicates that the model's predictions are closer to the actual data points, making it more reliable

t – t-statistic. This is used to test whether a regression coefficient (β) is **significantly different from zero**. **Null Hypothesis (H_0)**: **The coefficient (β) = 0** (meaning the variable has no effect). If: **β_0 (Intercept) = 0** The regression line passes through the origin; **β_1 (Slope) = 0** The line is flat, meaning the independent variable has no impact on the target variable. The **t-test helps determine whether the coefficients are statistically significant** and contribute meaningfully to the model.

P-value – The p-value tells us **how likely it is that the coefficient (β) is actually zero** based on the data. Interpretation: **$p < 0.05$** The coefficient is statistically significant, meaning it likely has an impact. **$p > 0.05$** The coefficient may not be important in predicting the dependent variable.

OLS Model Assumptions:

- Linearity:** The relationship between the dependent and independent variables should be **linear**. That means we assume linearity in data.
- No Endogeneity** (originated from Greek word Endogenous meaning Endo (within) generous (produce)) – Endogeneity occurs **when an independent variable is correlated with the error term**. Example – Larger properties are usually more expensive, but in some locations (e.g., downtown areas), even smaller properties can be costly. If location is not included in the model, it becomes part of the residuals. Hence, we assume all independent factors affecting dependent variable is considered. How to check? The **covariance** between independent variables and the error term should be zero.
- Normality and Homoscedasticity (can have effect on confidence intervals):** (i) Normality – The dataset should be such that the errors (unexplained variations) follow a normal pattern. This ensures that **statistical tests (like t-tests and confidence intervals) remain valid**. How to check? The **residuals** (differences between observed and predicted values) should be **normally distributed** (can be checked using histograms or Q-Q plots). (ii) Homoscedasticity (originated from Greek words homo (same) and skedastikos (scattered) meaning equal variance) The dataset should have a **consistent spread of variability** across all values of independent variables. This means the **predictive power of the model remains stable across different data points**. How to check? The **variance of residuals should remain constant** across all values of independent variables (checked using residual vs. fitted plots). Note if the variance is heterogenous then we use logarithm methods ($\log Y = b_0 + b_1(\log x_1)$) to generalize it.
- No Autocorrelation (can have effect on prediction):** Linear regression assumes that **data does not have a "Day of the Week" effect**, meaning past values should not influence future values. Residuals should be random and **not follow a time-dependent pattern**. How to check? Use the **Durbin-Watson test**—a value close to 2 means no autocorrelation. If the value is near 0 or 4, autocorrelation exists.
- No Multicollinearity:** Linear regression assumes that **independent variables are not highly correlated** with each other. If two or more variables convey similar information, the model cannot determine their individual effects correctly. Example: In a dataset predicting sales, if both **advertising spend** and **marketing budget** are included, but they are highly correlated, the model struggles to assign accurate weights to each variable. How to check? Check the **correlation matrix**—if independent variables have a correlation above **0.8**, multicollinearity is likely.

Note: The summary table has lot of parameters, but this document contains explanation to only few important parameters.
For more details on additional parameters refer: [how-to-interpret-result-from-linear-regression](#)

OLS Model Summary Reading

OLS (Ordinary Least Squares) is common method to derived linear regression equation. The least squares stands for the minimum square error meaning estimating linear equation with minimum sum of squared errors. Graphically, it is one closest to all point simultaneously. The derivation of optimal line or linear equation is a minimization problem that uses calculus and linear algebra to determined the slope and intercept of line.

R-squared – gives variability explained by model vs total variability, i.e. $SSR(\text{err. btw.}(\text{mean to predicted}))/SST(\text{err. btw.}(\text{mean to actual}))$. R-squared value gives btw 0 to 1; 0 – no variability, 1- explained total variability.

Adjusted R-squared – The adjusted R-squared is always smaller than R-Squared as it penalizes excessive use of variables, meaning formula accounts no. of variables information. Note: It only make sense to compare 2 models considering same dataset and dependent variable

F-statistic – As mentioned statistic in its name used for test for F-distribution. Null hypothesis: if all betas are equal to zero. Alternative hypothesis is at least one of the beta is not equal to zero

Model Summary

OLS Regression Results

Dep. Variable: price
Model: OLS
Method: Least Squares
Date: Sun, 03 Nov 2024
Time: 11:17:07
No. Observations: 100
DF Residuals: 98
DF Model: 1
Covariance Type: nonrobust

R-squared: 0.745
Adj. R-squared: 0.742
F-statistic: 285.9
Prob (F-statistic): 8.13e-31
Log-Likelihood: -1198.3
AIC: 2481.
BIC: 2486.

Coeff. table

	coef	std err	t	P> t	[0.025	0.975]
const	1.019e+05	1.19e+04	8.550	0.000	7.83e+04	1.26e+05
size	223.1787	13.199	16.909	0.000	196.986	249.371

Omnibus: 6.262

Prob(Omnibus): 0.044

Skew: 0.117

Kurtosis: 2.194

Durbin-Watson: 2.267

Jarque-Bera (JB): 2.938

Prob(JB): 0.230

Cond. No.: 2.75e+03

Some additional test

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 2.75e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Intercept b0 of equation $y = b_0 + b_1x$

Variable of equation, i.e. b1

std.err – error between predicted data points and original data points (lower the better)

t – t-statistic. There hypothesis involved here ,i.e. null hypothesis: is coefficient equal to 0 ? Note: if B0 is 0 then line starts from origin and B1 is 0 then predicted line is flat and of no use This test is asking question like, do we have enough statistical evidence that intercept defer from zero

P-value – the position of 0 in the t-distribution, hence it outside 95% (considering 1-tail) of the area, i.e. below 0.05, we are good as we don't want 0 as coefficient especially for B1. Please B0 if its more 0.05 also it doesn't matter

OLS Model Assumptions:

- Linearity:** The relationship between the dependent and independent variables should be **linear**. That means we assume linearity in data.
- No Endogeneity** (originated from Greek word Endogenous meaning Endo (within) generous (produce)) – Endogeneity occurs **when an independent variable is correlated with the error term**. Example – Larger properties are usually more expensive, but in some locations (e.g., downtown areas), even smaller properties can be costly. If location is not included in the model, it becomes part of the residuals. Hence, we assume all independent factors affecting dependent variable is considered. How to check? The **covariance** between independent variables and the error term should be zero.
- Normality and Homoscedasticity (can have effect on confidence intervals):** (i) Normality – The dataset should be such that the errors (unexplained variations) follow a normal pattern. This ensures that **statistical tests (like t-tests and confidence intervals) remain valid**. How to check? The **residuals** (differences between observed and predicted values) should be **normally distributed** (can be checked using histograms or Q-Q plots).(ii) Homoscedasticity (originated from Greek words homo (same) and skedastikos (scattered) meaning equal variance) The dataset should have a **consistent spread of variability** across all values of independent variables. This means the **predictive power of the model remains stable across different data points**. How to check? The **variance of residuals should remain constant** across all values of independent variables (checked using residual vs. fitted plots). Note if the variance is heterogenous then we use logarithm methods ($\log Y = b_0 + b_1(\log x_1)$) to generalize it.
- No Autocorrelation (can have effect on prediction):** Linear regression assumes that **data does not have a "Day of the Week" effect**, meaning past values should not influence future values. Residuals should be random and **not follow a time-dependent pattern**. How to check? Use the **Durbin-Watson test**—a value close to 2 means no autocorrelation. If the value is near 0 or 4, autocorrelation exists.
- No Multicollinearity:** Linear regression assumes that **independent variables are not highly correlated** with each other. If two or more variables convey similar information, the model cannot determine their individual effects correctly. Example: In a dataset predicting sales, if both **advertising spend** and **marketing budget** are included, but they are highly correlated, the model struggles to assign accurate weights to each variable. How to check? Check the **correlation matrix**—if independent variables have a correlation above **0.8**, multicollinearity is likely.

Note: The summary table has lot of parameters, but this document contains explanation to only few important parameters. For more details on additional parameters refer: [how-to-interpret-result-from-linear-regression](#)