REGRESSION ANALYSIS

FINANCE AND ECONOMICS IN SPORT

FALL 2022

UNDERSTANDING REGRESSION ANALYSIS

- Regression analysis is one of the most important methods in quantitative market research.

- Marketing and strategy departments of major companies use regression analysis regularly for decision-making

- Regression analysis can:
  - Indicate if independent variables have a significant relationship with a dependent variable.
  - Indicate the relative strength of different independent variables' effects on a dependent variable.
  - Make predictions.

UNDERSTANDING REGRESSION ANALYSIS

Notation:

For simple regression:

$$y = \alpha + \beta_1 x_1 + e$$

Y: Dependent variable

α: Intercept (constant)

β_1: Regression parameter (coefficient)

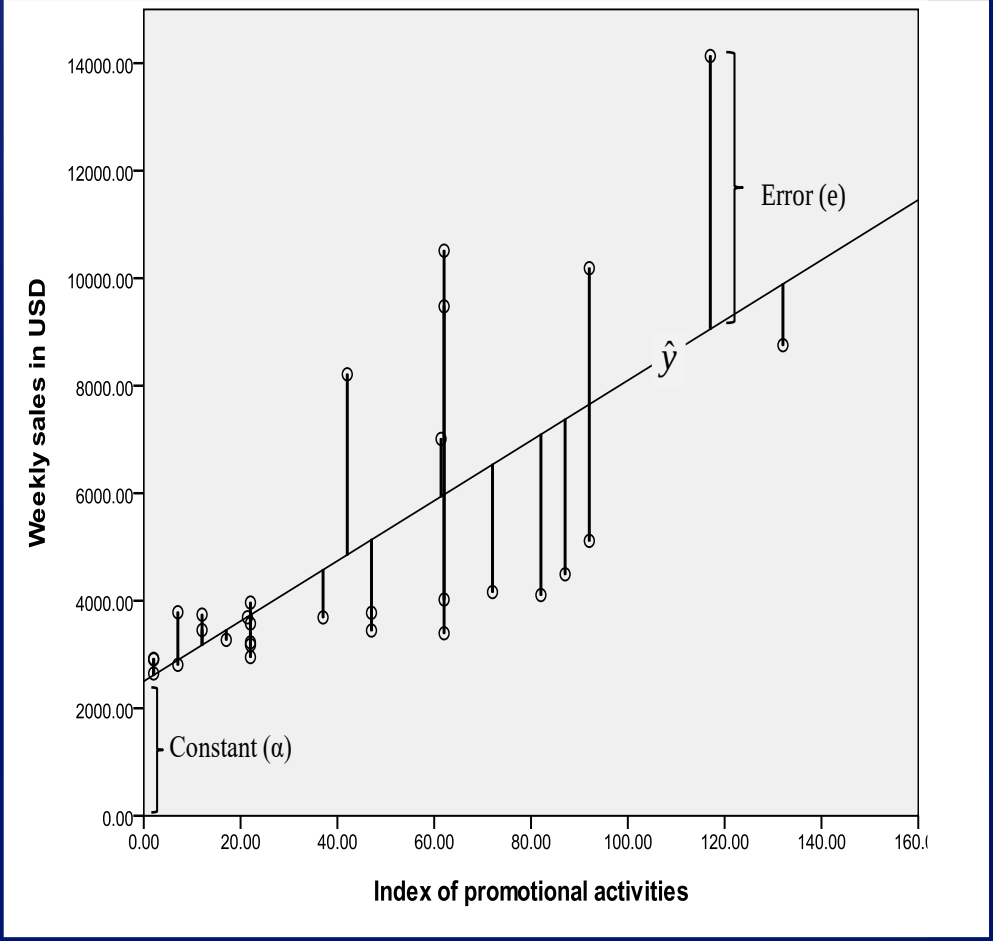
x_1: Independent variable

e: Error or residual

For multiple regression:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

Predicted (or estimated) value



UNDERSTANDING REGRESSION ANALYSIS

- $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e$
- $\textit{Total revenue} = \alpha + \beta_1\, \textit{Ticket} + \beta_2\, \textit{Media} + \beta_3\, \textit{Donation} + e$

- The regression estimation procedure attempts to fit a best possible line by using the observations. This line is optimal if the squared distances to all observation points are minimized.

THE PROCESS OF CONDUCTING A REGRESSION ANALYSIS

1	Check the regression analysis data requirements
2	Specify and estimate the regression model
3	Test the regression analysis assumptions
4	Interpret the regression results
5	Use the regression model

CHECK THE REGRESSION ANALYSIS DATA REQUIREMENTS

Sample size

- Formal power analysis (e.g. using G*power)
- Green's (1991) rule of thumb: 104+k where k is the number of independent variables

Scale type of the dependent variable

- Interval scaled
- Ratio scaled

Collinearity

- Collinearity is a data issue that arises if two independent variables are highly correlated.
- The **VIF** (Variance Inflation Factor) can indicate collinearity. **Values over 10 indicate collinearity.**
- If collinearity is present either use factor analysis, re-specify the regression model, or simply accept the collinearity (noting its presence).

SPECIFY AND ESTIMATE THE REGRESSION MODEL

- The goal is to build a simple yet complete model. Some suggestions
- If independent variables have overlap then focus on the most important or distinct variable.
- If you need to use a model for different circumstances, ensure the independent variables are the same to allow comparison.
- Consider the type of advice you want to give.
- Consider the sample sizes and rules of thumb. Smaller datasets require fewer independent variables
- If variables are ordinal, try to use dummy variables.

TEST THE REGRESSION ANALYSIS ASSUMPTIONS

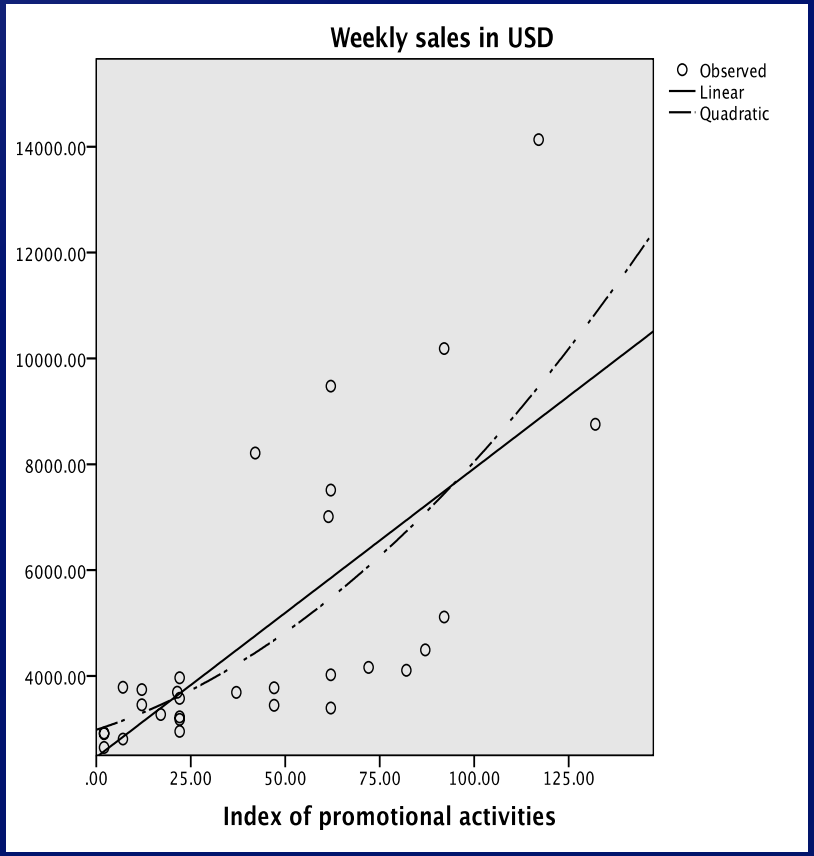
- The regression model can be expressed linearly.
- The regression model's expected mean error is zero.
- The errors' variance is constant (homoscedasticity).
- The errors need to be approximately normally distributed (normality).
- The errors are independent (no autocorrelation).

LINEARITY

- Linearity is an assumption that asks if we can write the regression model as $y = \alpha + \beta_1 x_1 + .. + \beta_z x_{z+} e$

- A separate issue is if the relationship between y and x is best expressed as a linear relationship. If not, we can use transformations:
  - x^2
  - Log (x)

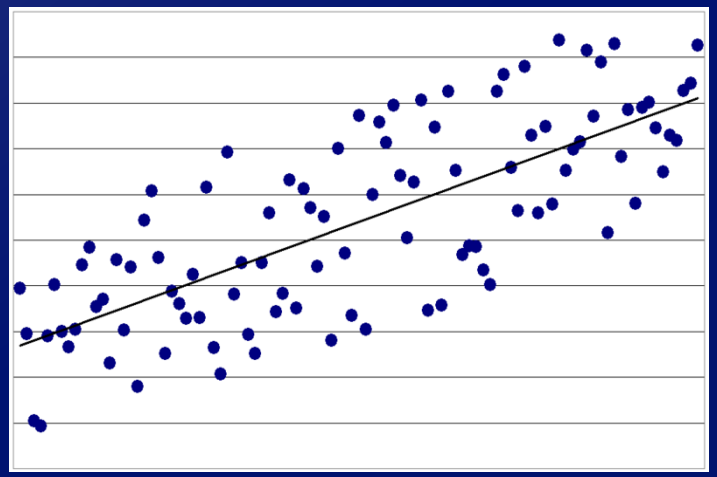
- After such transformations the model is still linear!



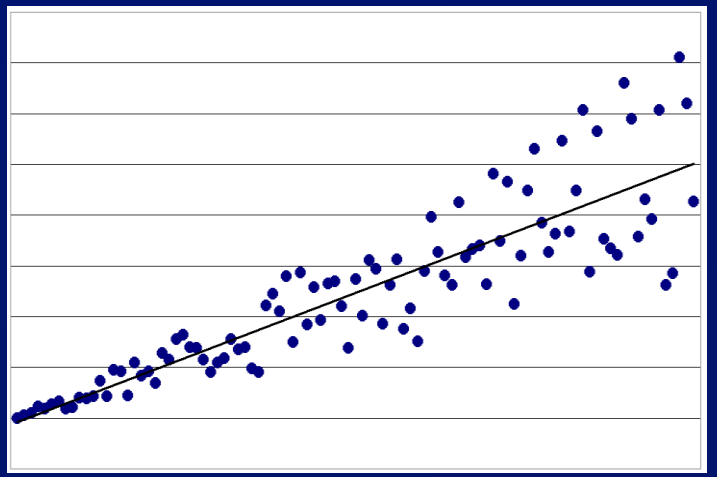
HOMOSKEDASTICITY
IS THE ERRORS' VARIANCE CONSTANT?

- Conduct the Breusch-Pagan test
- If p-value is less than 0.05, reject the null hypothesis of constant errors' variance).

Homoskedastic distribution



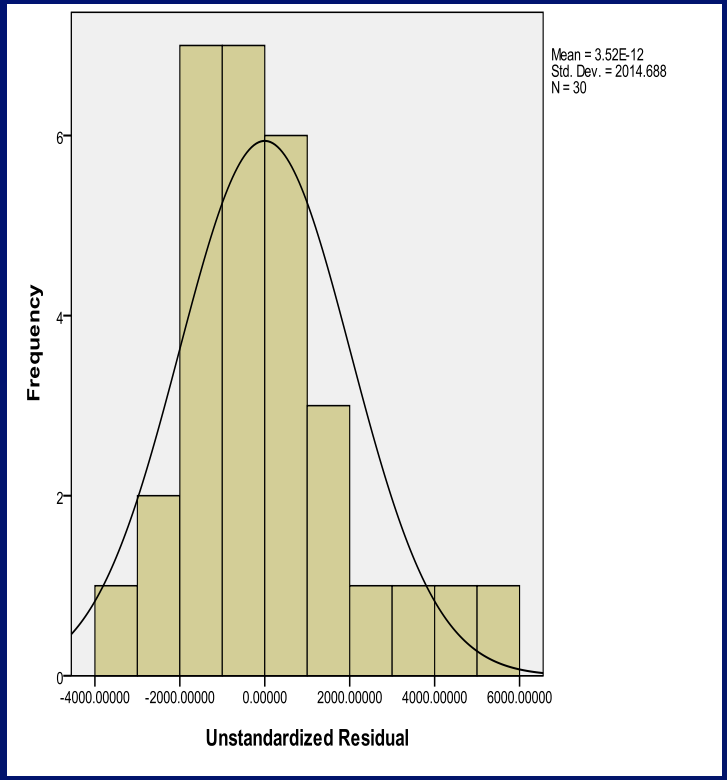
Heteroskedastic distribution



NORMALITY

ARE THE RESIDUALS APPROXIMATELY NORMALLY DISTRIBUTED?

- This assumption can be visualized using a histogram of the residuals.
- **Skewness: -2 to 2**
- **Kurtosis: -7 to 7**
- Conduct the **Shapiro-Wilk test** on the residuals
  - If *p-value* is less than 0.05, we can reject the null hypothesis of normality.



AUTOCORRELATION
ARE THE ERRORS INDEPENDENT?

- Does one observation influence another?
- Use the **Durbin-Watson test**
- The test can have four conclusions:
  - The errors may be positively related (positive autocorrelation).
  - If positive errors are commonly followed by negative errors and negative errors by positive errors, we have negative autocorrelation.
  - If no systematic pattern of errors occurs, we have no autocorrelation. This absence of autocorrelation is required to estimate standard (OLS) regression models.
  - The D-W values may fall between the lower and upper critical value. If this occur, the test is inconclusive.

ASSUMPTION CHECKS

Normality Tests

	Statistic	p
Shapiro-Wilk	0.875	< .001
Kolmogorov-Smirnov	0.146	< .001
Anderson-Darling	59.8	< .001

Note. Additional results provided by *moretests*

Heteroskedasticity Tests

	Statistic	p
Breusch-Pagan	471	< .001
Goldfeld-Quandt	0.878	0.974
Harrison-McCabe	0.532	0.978

Note. Additional results provided by *moretests*

P-value <0.05, suggesting that the errors are not normally distributed (i.e., reject the null hypothesis of normality).

P-value <0.05, suggesting that the error variance is not constant (i.e., reject the null hypothesis of constant error variance).

ASSUMPTION CHECKS

Durbin–Watson Test for Autocorrelation

Autocorrelation	DW Statistic	p
0.546	0.908	< .001

Collinearity Statistics

	VIF	Tolerance
media	2.07	0.484
donor	2.00	0.501
sponsor	1.81	0.553
stfee	1.15	0.869

DW statistic <1.352, suggesting that positive autocorrelation exists (i.e., reject the null hypothesis of no autocorrelation).

VIF < 10, suggesting no collinearity.

INTERPRET THE REGRESSION RESULTS

- Overall model fit
  - F-test: If *p-value* is less than 0.05, the model is significant and acceptable.
  - R^2 or Adjusted R^2 : range 0 - 1

$$R^2 = \frac{SS_R}{SS_E + SS_R} = \frac{SS_R}{SS_T}$$

INTERPRET THE REGRESSION RESULTS

Model Fit Measures

				Overall Model Test			
Model	R	R^2	Adjusted R^2	F	df1	df2	p
1	0.799	0.639	0.638	786	4	1779	< .001

For comparing different models

63.8% of variance are explained

Significant F-value ($p<0.05$) suggesting a significant model.

INTERPRET THE REGRESSION RESULTS

- To find the effects of individual variables:
- Consider the significance of the parameter of each variable (not the intercept) in turn. The hypothesis tested for each variable is
  - H_0: $\beta_1 = 0$ (line is horizontal)
  - H_1: $\beta_1 \neq 0$ (regression line is not horizontal)
- If the *p* column of the model coefficients table in *Jamovi* indicates a value below 0.05, we consider this to be significant.

Model Coefficients - tickets

Predictor	Estimate	SE	t	p	Stand. Estimate
Intercept	4.09e+6	334181.2499	12.25	< .001	
media	0.0909	0.0160	5.68	< .001	0.116
donor	0.3269	0.0179	18.23	< .001	0.367
sponsor	0.8096	0.0444	18.22	< .001	0.349
stfee	-0.3655	0.0340	-10.76	< .001	-0.164

Significant p-value ($p<0.05$) suggesting a significant independent variables.

INTERPRET THE REGRESSION RESULTS

For significant variables also consider:

- The sign
- The βs. (direction of the regression line)
  - Unstandardized: expresses the effect of a one-unit change in the independent variable on the dependent.
  - The standardized effect: the effect relative to the other variables. Standardized effects can be compared but not for dummy variables

USE THE REGRESSION RESULTS FOR PREDICTION

Model Coefficients - tickets

Predictor	Estimate	SE	t	p	Stand. Estimate
Intercept	4.09e+6	334181.2499	12.25	< .001	
media	0.0909	0.0160	5.68	< .001	0.116
donor	0.3269	0.0179	18.23	< .001	0.367
sponsor	0.8096	0.0444	18.22	< .001	0.349
stfee	-0.3655	0.0340	-10.76	< .001	-0.164

- Prediction of effects or interpreting effects.
- Prediction: consider the model and results obtained:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$$

What is our predicted tickets revenues if we set the media to $4 million, the donor to $3 million, the sponsor to $2 million, and the student fee (stfee) to $1 million?

4,090,000 + 0.0909 (4,000,000) + 0.3269 (3,000,000) + 0.8096 (2,000,000)

– 0.3655 (1,000,000)

= $6,688,000

SUMMARY

- Regression analysis is one of the most important methods in empirical economic research.
- The regression analysis puts an optimal line through the observed points
- Assumptions should be checked.
- Interpretation of results
  - Consider model fit and the effects of individual variables.

CLASS ACTIVITY

- Review Quiz
- Class assignment
  - Regression analysis assumptions
  - Regression model fit
  - Effects of independent variables
  - Interpretation of the results
  - Prediction