SJSU SAN JOSÉ STATE UNIVERSITY

DATA 220: Mathematical Methods for Data Analytics

Dr. Mohammad Masum

Multiple Linear Regression

- Simple linear regression

$$\hat{y} = b_0 + b_1 * X_1$$

Multiple Linear Regression

- Simple linear regression

$$\underset{\text{Dependent Variable}}{\hat{y}} = \overset{\text{Constant}}{b_0} + \overset{\text{Coefficient}}{b_1} * \underset{\text{Independent Variable}}{X_1}$$

Multiple Linear Regression

- Years of Experience and Salary Data
  - Available on Kaggle - https://www.kaggle.com/rohankayan/years-of-experience-and-salary-dataset
  - Number of Observation: **30**

	YearsExperience	Salary
0	1.1	39343.0
1	1.3	46205.0
2	1.5	37731.0
3	2.0	43525.0
4	2.2	39891.0

Multiple Linear Regression

- Simple linear regression

$$\widehat{salary} = b_0 + b_1 * YearsExperience$$

Multiple Linear Regression

- Simple linear regression

$$residual = y - \hat{y}$$

$$\min \sum (y - \hat{y})^2$$

We can solve this minimization problem to achieve the optimal parameters using **gradient descent algorithm**

Multiple Linear Regression

- Simple linear regression

Steps

1. Find loss function
2. Set initial value of parameters and hyperparameters
3. Calculate the partial derivatives of the loss function with respect to β parameters
4. Define the update equation (GD)
5. Repeat the process

Multiple Linear Regression

- Normal Equation for Linear Regression
  - Closed form solution for linear regression – achieve optimal set of parameters

$$b_{opt} = \left(X^T X\right)^{-1} X^T y$$

Derivation in class

Simple Linear Regression

- Assumptions of SLR
  - Linearity – The relationship between independent (X: years of experience) and dependent (y: salary) variables
    - Check: Scatter Plot – linear relationship
  - Homoscedasticity (Equal Variance) – The variance of the residuals is constant for all values of X
    - Check: Residual vs. fits plot – randomly scattered
  - Independence of Error – No relationship between the residuals and fitted values (salary)
    - Check: Residual vs. fits plot – randomly scattered
  - Normality of errors – the residuals must be approximately normally distributed
    - Check: Q-Q plot – most data points fall close to the line

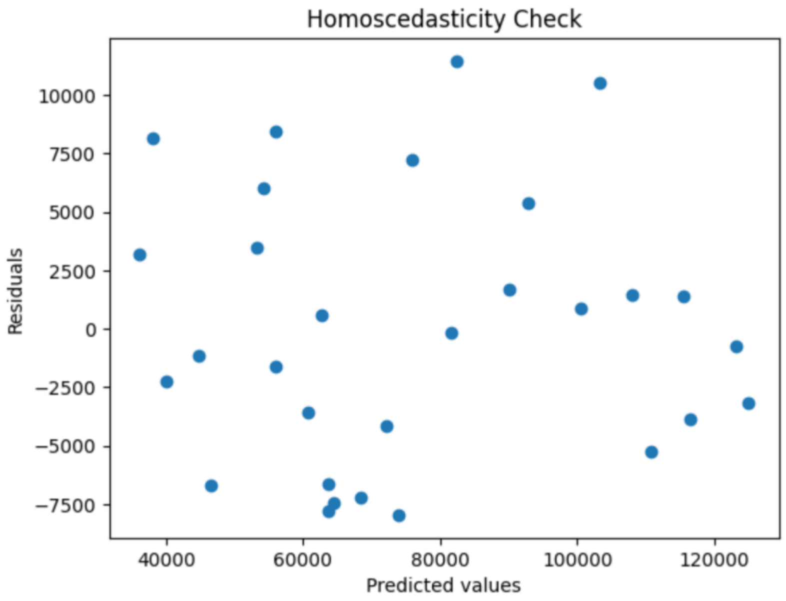
Simple Linear Regression

- Linearity – The relationship between independent (X: years of experience) and dependent (y: salary) variables
  - Check: Scatter Plot – linear relationship

SJSU SAN JOSÉ STATE UNIVERSITY

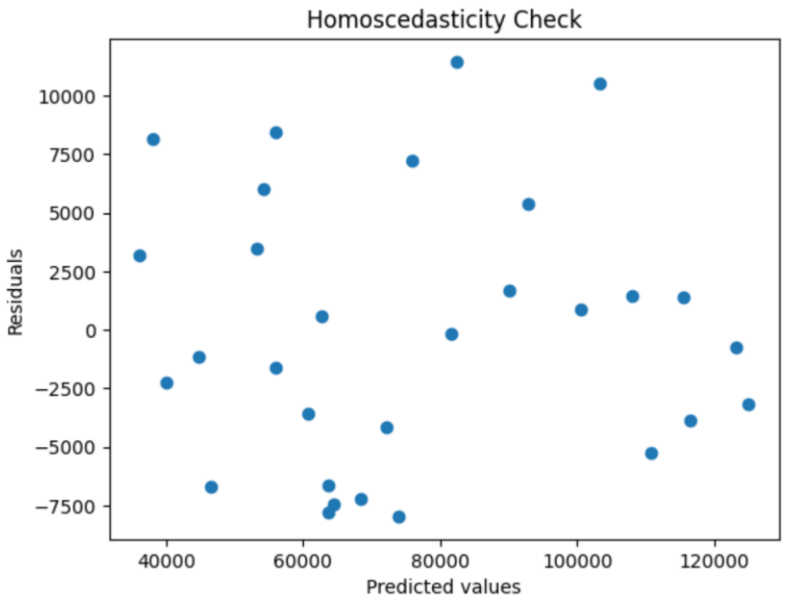
Simple Linear Regression

- Homoscedasticity (Equal Variance) – The variance of the residuals is constant for all values of X
  - Plot the residuals against the predicted values - if the plot shows a random scattering of points with no clear pattern, the assumption of homoscedasticity is met
  - However, if the plot shows a funnel shape or a systematic pattern, it indicates heteroscedasticity



Simple Linear Regression

- Independence of Error – No relationship between the residuals and fitted values (salary)
  - Check: Residual vs. fits plot – randomly scattered

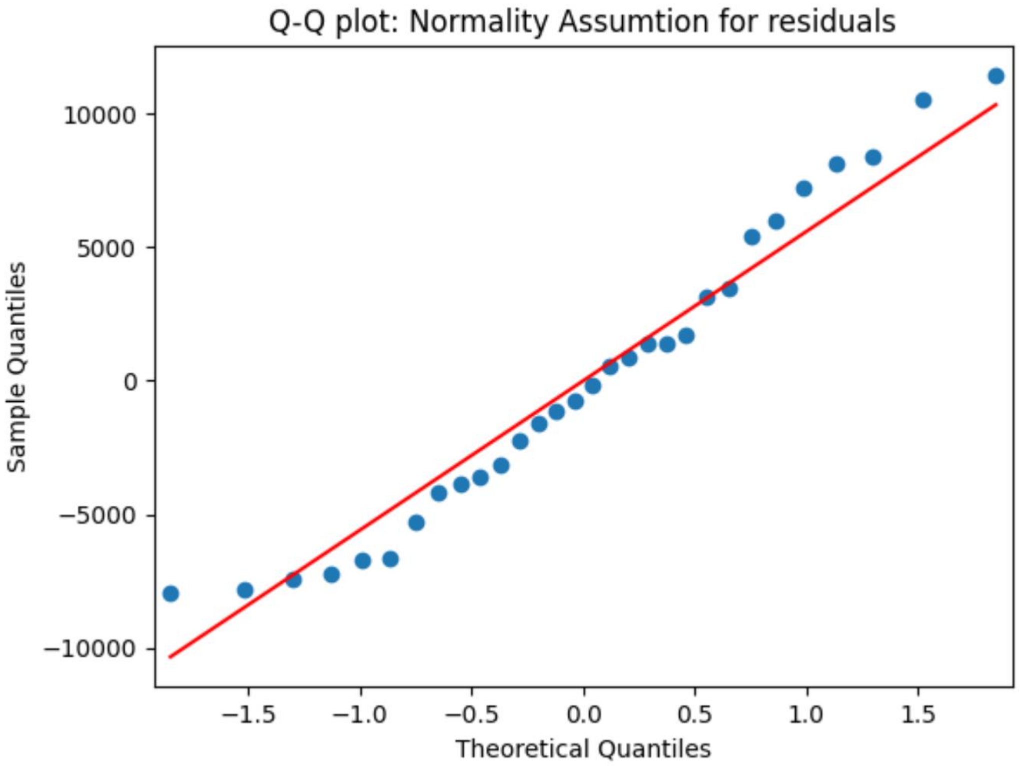


Simple Linear Regression

- Normality of errors – the residuals must be (approximately) normally distributed
  - Q-Q plot – quantile-quantile plot – visually assess the normality assumption of the residuals
    - It compares the distribution of the residuals to a normal distribution
      - Determines whether two samples are come from the same population
    - Process of constructing a Q-Q plot – plotting the ordered residuals against the expected values of a normal distribution with the same mean and variance as the residuals
  - If the points on the Q-Q plot form a straight line with a slope of 1 – indicating the residuals follow the normal distribution
  - If the residuals deviate significantly from the straight line, it suggests that the normality assumption is not met
  - In such a scenario, the linear regression model may not be appropriate for the given data and may require further investigation and modification

Simple Linear Regression

- Normality of errors – the residuals must be (approximately) normally distributed



Multiple Linear Regression

- 50 Startups Data
  - Available on Kaggle- https://www.kaggle.com/amineoumous/50-startups-data
  - Predict which companies to invest for maximizing profit
  - Number of Observation **50**

Profit	R&D Spend	Administration	Marketing Spend	State
192261.83	165349.20	136897.80	471784.10	New York
191792.06	162597.70	151377.59	443898.53	California
191050.39	153441.51	101145.55	407934.54	Florida
182901.99	144372.41	118671.85	383199.62	New York
166187.94	142107.34	91391.77	366168.42	Florida

Multiple Linear Regression

Simple Linear Regression

$$\hat{y} = b_0 + b_1 * X_1$$

Dependent Variable

Independent Variable

Multiple Linear Regression

$$\hat{y} = b_0 + b_1 * X_1 + b_2 * X_2 + \dots + b_n * X_n$$

Constant

Coefficient

Multiple Linear Regression

Profit	R&D Spend	Administration	Marketing Spend	State
105733.54	75328.87	144135.98	134050.07	Florida
105008.31	72107.60	127864.55	353183.81	New York
103282.38	66051.52	182645.56	118148.20	Florida
101004.64	65605.48	153032.06	107138.38	New York
99937.59	61994.48	115641.28	91131.24	Florida

$$\hat{y} = b_0 + b_1 * X_1 + b_2 * X_2 + b_3 * X_3 + b_4 * X_4$$

Multiple Linear Regression

Dummy Variable

State
New York
California
Florida
New York
Florida

→

State_California	State_Florida	State_New York
0	0	1
1	0	0
0	1	0
0	0	1
0	1	0

$$\hat{y} = b_0 + b_1 * X_1 + b_2 * X_2 + b_3 * X_3 + \boldsymbol{b_4} * \boldsymbol{D_1} + \boldsymbol{b_5} * \boldsymbol{D_2}$$

Multiple Linear Regression

- Coefficient of determination (R^2) – a statistical measure that provide information about the goodness of fit of a model
  - Represents the proportion of the variance in the dependent variable that is explained by the independent variables in a regression model
  - It is a value between 0 and 1, with higher values indicating a better fit of the regression model

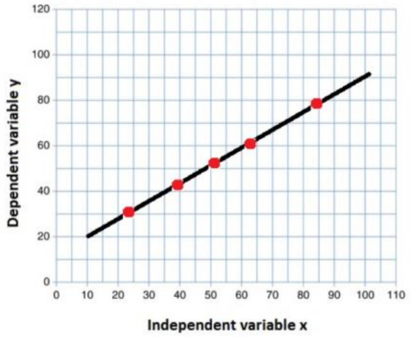
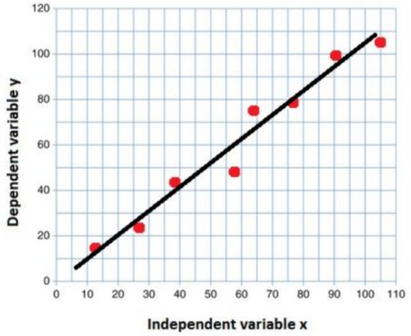
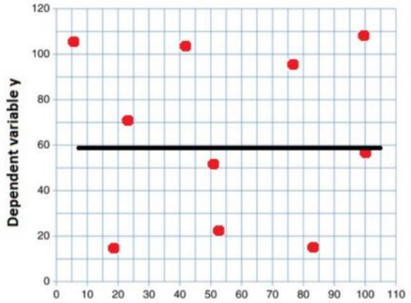
$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$
$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

- For example, a MLR model predicts house prices based on the area and number of bedrooms: an R^2 value of 0.8 means that 80% of the variation in house prices can be explained by the area and number of bedrooms in the model, while the remaining 20% is due to other factors that are not included in the model

https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html#:~:text=The%20coefficient%20of%20determination%2C%20or,line%20approximates%20the%20actual%20data.

Multiple Linear Regı

- Coefficient of determination (R^2)

R^2 Values	Interpretation	Graph
$R^2 = 1$	All the variation in the y values is accounted for by the x values	
$R^2 = 0.83$	83% of the variation in the y values is accounted for by the x values	
$R^2 = 0$	None of the variation in the y values is accounted for by the x values	

https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/coefficient-of-determination-r-squared.html#:~:text=The%20coefficient%20of%20determination%2C%20or,line%20approximates%20the%20actual%20data.

SJSU SAN JOSÉ STATE UNIVERSITY

Multiple Linear Regression

- Coefficient of determination (R^2)

```
import statsmodels.api as sm

X_constant = sm.add_constant(X)
lin_reg = sm.OLS(y,X_constant).fit()
lin_reg.summary()
```

OLS Regression Results

Dep. Variable:	Profit	**R-squared:**	0.951
Model:	OLS	**Adj. R-squared:**	0.945
Method:	Least Squares	**F-statistic:**	169.9
Date:	Thu, 20 Apr 2023	**Prob (F-statistic):**	1.34e-27
Time:	22:16:19	**Log-Likelihood:**	-525.38
No. Observations:	50	**AIC:**	1063.
Df Residuals:	44	**BIC:**	1074.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	5.013e+04	6884.820	7.281	0.000	3.62e+04	6.4e+04
R&D Spend	0.8060	0.046	17.369	0.000	0.712	0.900
Administration	-0.0270	0.052	-0.517	0.608	-0.132	0.078
Marketing Spend	0.0270	0.017	1.574	0.123	-0.008	0.062
State_Florida	198.7888	3371.007	0.059	0.953	-6595.030	6992.607
State_New York	-41.8870	3256.039	-0.013	0.990	-6604.003	6520.229

Multiple Linear Regression

- Assumptions of Multiple Linear Regression
  - Linearity – Relationship between independent and dependent variables
  - Homoscedasticity (Equal Variance) – Variance of the residuals is constant
  - Independence of Error – No relationship between the residuals and fitted values
  - Normality of errors – Residuals must be approximately normally distributed
  - Multicollinearity – High correlation among independent variables

Multiple Linear Regression

- Multicollinearity – High correlation among independent variables
  - Correlation matrix of the independent variables
  - Variation Inflation Factor (VIF) – measures how much the variance of the estimated regression coefficient is increased due to multicollinearity
- A VIF value of 5 or more is generally considered as an indication of multicollinearity
- Solutions – removing one of the correlated independent variables, transforming the variables, or using regularization techniques like Ridge or Lasso regression

VIF for j^{th} predictor (independent variable),

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the R^2-value obtained by regressing the j^{th} predictor on the remaining predictors

Multiple Linear Regression

	feature	VIF
0	R&D Spend	8.451019
1	Administration	4.950277
2	Marketing Spend	8.092278
3	State_Florida	2.004519
4	State_New York	1.928836

Correlation coefficients

	R&D Spend	Administration	Marketing Spend	State_Florida	State_New York
R&D Spend	1.000000	0.241955	0.724248	0.105711	0.039068
Administration	0.241955	1.000000	-0.032154	0.010493	0.005145
Marketing Spend	0.724248	-0.032154	1.000000	0.205685	-0.033670
State_Florida	0.105711	0.010493	0.205685	1.000000	-0.492366
State_New York	0.039068	0.005145	-0.033670	-0.492366	1.000000

Multiple Linear Regression

- Coefficient of determination (R^2)

	feature	VIF
0	R&D Spend	3.917998
1	Administration	4.863405
2	State_Florida	1.881915
3	State_New York	1.909248

OLS Regression Results

Dep. Variable:	Profit	**R-squared:**	0.948
Model:	OLS	**Adj. R-squared:**	0.943
Method:	Least Squares	**F-statistic:**	205.0
Date:	Thu, 20 Apr 2023	**Prob (F-statistic):**	2.90e-28
Time:	22:32:12	**Log-Likelihood:**	-526.75
No. Observations:	50	**AIC:**	1064.
Df Residuals:	45	**BIC:**	1073.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.46e+04	6371.060	8.571	0.000	4.18e+04	6.74e+04
R&D Spend	0.8609	0.031	27.665	0.000	0.798	0.924
Administration	-0.0527	0.050	-1.045	0.301	-0.154	0.049
State_Florida	1091.1075	3377.087	0.323	0.748	-5710.695	7892.910
State_New York	-39.3434	3309.047	-0.012	0.991	-6704.106	6625.420