PROJECT REPORT
IE7280 Statistical Methods in Engineering

Hypothesis Testing and Multi-way ANOVA

Akshay Anandbabu(001277319)
Rohit Appandaraju(001279555)

Project Description:

In the recent years, nutrition and diet have been proven to play an important role in weight loss for young adults. This has resulted in controlling the intake of various food constituents and has allowed various nutritionists to perform detailed analysis to enhance decision making. The objective of our project is to perform hypothesis testing on various factors to decide the significance between each factor based on 3 different diet plans. We derive at our conclusions using the statsmodel api python library.

Solution Design:

Our project is divided into the following sections:

- Data Collection
- Exploratory Analysis
- Multi-way ANOVA
- Model Building (Linear Regression w/ Ordinary Least Squares)
- Conclusions/ Statistical Inferences

Data Collection:

The publicly available data was collected from the University of Sheffield website. The data set ***'Diet.sav'*** contains information on 78 people who undertook one of three diets. There is background information such as age, gender and height as well as weight lost on the diet (a positive value means they lost weight). The period during which a person underwent a diet plan is 6 weeks after which their weight was noted. The weight loss was calculated by subtracting a person's weight before and after the diet. The dataset is non-partitioned and the single data set is used to perform analysis.

Metadata Description:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 78 entries, 0 to 77
Data columns (total 8 columns):
Person          78 non-null int64
gender          78 non-null int64
Age             78 non-null int64
Height          78 non-null int64
pre.weight      78 non-null int64
Diet            78 non-null int64
weight6weeks    78 non-null float64
Weight_Loss     78 non-null float64
dtypes: float64(2), int64(6)
memory usage: 5.0 KB
```

Person :	Index for each person
Gender :	Male=0 Female =1
Age :	Age of a person in years
Height :	height in cm
Diet :	Type of the Diet (1,2,3)
pre.weight :	Weight before diet in Kg
weight6weeks:	Weight after taking diet in Kg
Weight_loss:	pre.weight - weight6weeks in Kg (Dependant Variable)

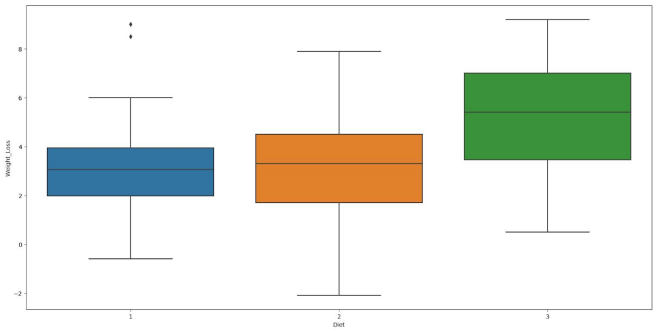
Data Summary

The below table gives an overall description of the data giving quantitative measures like mean,median and quartile ranges

Index	Person	gender	Age	Height	pre.weight	Diet	weight6weeks	Weight_Loss
count	78	78	78	78	78	78	78	78
mean	39.5	0.423077	39.1538	170.821	72.5256	2.03846	68.6808	3.84487
std	22.6605	0.497245	9.81528	11.2766	8.72334	0.81292	8.9245	2.55148
min	1	0	16	141	58	1	53	-2.1
25%	20.25	0	32.25	164.25	66	1	61.85	2
50%	39.5	0	39	169.5	72	2	68.95	3.6
75%	58.75	1	46.75	174.75	78	3	73.825	5.55
max	78	1	60	201	103	3	103	9.2

Box Plot for Distribution of Weight Loss for 3 Diet types:

The below plot shows the range of weight losses for different diet types



Inferences obtained from the boxplot:

1. We can see that the medians of weight loss for diet 1 and 2 are similar, but there is a significant increase in weight loss for diet 3 and has a higher range of values.
2. Diet type 1 has 2 outliers (8.5,9)

Correlation Plot

The Following heatmap shows the correlation between the different variables and weight loss

We can see that the diet type,weight before/after 6 weeks,gender,height have the high correlation to weight loss(dependant variable)

Hypothesis Testing and ANOVA:

3 way anova with Diet type,gender,height as the 3 factors along with interaction between the factors.

Parameter of Interest :

αi - Effect of each factor under consideration

Null Hypothesis:

Ho : $\alpha 1 = \alpha 2 = \alpha 3 \ldots = \alpha i$

$\beta 1 = \beta 2 = \beta 3 \ldots = \beta j$

$\gamma 1 = \gamma 2 = \gamma 3 \ldots . = \gamma k$

Where, αi- Effect of factor diet at level i(1,2,3)

βj- Effect of factor Age at level j(1,2)

γk- Effect of factor Height at level k

Alternate Hypothesis:

H1 : Atleast one $\alpha i, \beta j, \gamma k$ is unequal

We perform ANOVA using the OLS model we created with Height , Age and Diet as the predictor variables.

ANOVA Output:

	sum_sq	df	F	PR(>F)
Diet	33.305768	1.0	5.587845	0.020865
gender	8.953545	1.0	1.502173	0.224445
Diet:gender	6.457909	1.0	1.083470	0.301504
Height	1.993772	1.0	0.334503	0.564876
Diet:Height	12.740743	1.0	2.137567	0.148205
gender:Height	1.454627	1.0	0.244049	0.622844
Diet:gender:Height	5.001958	1.0	0.839199	0.362771
Residual	417.227735	70.0	NaN	NaN

From the ANOVA Table ,
P value (Diet) =0.020865 < 0.05 So we fail to accept null hypothesis resulting in our conclusion that weight loss is affected by diet

Following Anova we are performing Tukey's Procedure to identify significantly different diet type since its the only categorical variable under consideration

Multiple Comparison of Means - Tukey HSD,FWER=0.05

group1	group2	meandiff	lower	upper	reject
1	2	-0.2741	-1.8806	1.3325	False
1	3	1.8481	0.2416	3.4547	True
2	3	2.1222	0.5636	3.6808	True

We can see that diet types 1,3 and 1,2 have absolute mean difference greater than *W (critical value)* .Hence we can clearly say that diet type 3 is significantly different from 1 and 2.

Model Building:

From our exploratory analysis we found out that the most significant predictors for regression were Diet, Age and Height. We perform an OLS regression to fit our predicted values using these independent variables.

Intercepts of linear regression:

```
Out[10]: LinregressResult(slope=0.9485260770975061, intercept=1.9113378684807245,
rvalue=0.3022076049332828, pvalue=0.007164023698650191, stderr=0.34319450027189763)
```

Model Output:

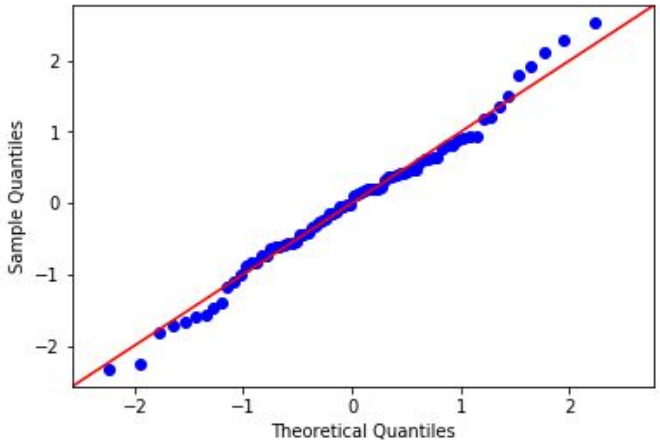
OLS Regression Results

Dep. Variable:	Weight_Loss	R-squared:	0.722
Model:	OLS	Adj. R-squared:	0.711
Method:	Least Squares	F-statistic:	64.87
Date:	Wed, 15 Aug 2018	Prob (F-statistic):	8.67e-21
Time:	20:06:02	Log-Likelihood:	-179.90
No. Observations:	78	AIC:	365.8
Df Residuals:	75	BIC:	372.9
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Age	0.0031	0.028	0.110	0.913	-0.054	0.060
Height	0.0094	0.008	1.171	0.245	-0.007	0.025
Diet	1.0307	0.337	3.059	0.003	0.359	1.702

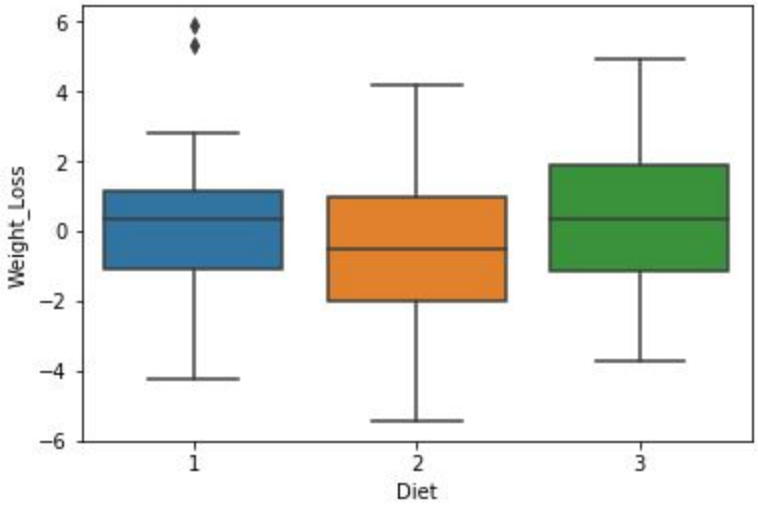
Omnibus:	0.140	Durbin-Watson:	1.835
Prob(Omnibus):	0.932	Jarque-Bera (JB):	0.028
Skew:	-0.046	Prob(JB):	0.986
Kurtosis:	2.987	Cond. No.	211.

Quantile-Quantile Plot



We can see that the residuals in the qq plot follow linearity thus satisfying the assumption of normality between predictor variables

Boxplot of Residuals:



The boxplot depicts the range of residuals of each data type. We can confidently say that the error residual median for diet 3 is positive whereas for diet 2, median is negative suggesting positive and negative difference between the fitted weight loss and actual weight loss respectively

Statistical Inferences Obtained:

From the Above Analysis we have found Following Outcomes

- 3 Way Anova concluded that weight loss (dependant variable) is related to diet type. (other variables do effect weight loss for this dataset)
- Tukey's Procedure shows that diet type 3 is significantly different from 1 and 2 diet types based on the absolute mean differences.