TT23 Collection - Ushika Kidd

1a) Download the data file Dutch.csv and load the data into a Pandas dataframe called Dutch. [2]

In [3]:
```
# Set up Python Libraries
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
import pandas
import seaborn as sns
import statsmodels.api as sm
import statsmodels.formula.api as smf
sns.set_theme()

# Import and view the data - 'Dutch.csv'
from google.colab import drive
drive.mount('/content/drive')
Dutch = pandas.read_csv('/content/drive/MyDrive/QM/data/Dutch.csv')
Dutch
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, ca
ll drive.mount("/content/drive", force_remount=True).
```

Out[3]:

	l1	aaa	lor	enroll	speaking	sex	family
0	Russian	26	3	101	574	Female	Indo-European
1	Portugese	25	5	106	533	Female	Indo-European
2	Romanian	36	2	86	534	Female	Indo-European
3	Polish	31	2	100	494	Female	Indo-European
4	Spanish	35	4	118	480	Female	Indo-European
...	...	...	...	...	...	...	...
49676	Afrikaans	25	4	31	480	Male	Indo-European
49677	Czech	23	4	90	603	Male	Indo-European
49678	Albanian	23	3	85	504	Male	Indo-European
49679	Afrikaans	35	2	93	537	Male	Indo-European
49680	Yoruba	26	12	35	531	Male	Niger-Congo

49681 rows × 7 columns

b) What is the correlation coefficient for the following pairs of variables:

1. Age at arrival in Netherlands and speaking score.
2. Length of residence in the Netherlands and speaking score.
3. Enrolment in Dutch secondary schools and speaking score.

Comment on the direction and strength of the correlations. [6]

In [4]:
```
# 1b) Finding the correlation coefficient
Dutch.corr(method='spearman')
```

```
<ipython-input-4-34e492a2679f>:2: FutureWarning: The default value of numer
ic_only in DataFrame.corr is deprecated. In a future version, it will defau
lt to False. Select only valid columns or specify the value of numeric_only
to silence this warning.
  Dutch.corr(method='spearman')
```

Out[4]:

	aaa	lor	enroll	speaking
aaa	1.000000	0.006968	-0.023858	-0.149123
lor	0.006968	1.000000	-0.187745	-0.083722
enroll	-0.023858	-0.187745	1.000000	0.432986
speaking	-0.149123	-0.083722	0.432986	1.000000

bi) Age at arrival in Netherlands and speaking score

r = -0.15

bii) Length of residence in the Netherlands and speaing score

r = -0.084

biii) Enrolment in Dutch secondary schools and speaking score.

r = 0.43

The negative r value for bi) suggests that there is a negative correlation between the age at arrival in the Netherlands and the speaking score. This is a fairly weak correlation, being close to 0.

The negative value for bii) suggests there is a negative correlation between length of residence in the Netherlands and speaking score, but this is a very weak correlation, essentially no correlation.

There is a relatively strong positive correlation between enrollment in Dutch secondary schools and speaking score in biii).

c) Do men and women differ in their speaking scores?

ci) Find the mean speaking score for men and women

In [5]:
```
# Mean speaking score
Dutch.mean()
```

```
<ipython-input-5-33a20713598b>:2: FutureWarning: The default value of numer
ic_only in DataFrame.mean is deprecated. In a future version, it will defau
lt to False. In addition, specifying 'numeric_only=None' is deprecated. Sel
ect only valid columns or specify the value of numeric_only to silence this
warning.
  Dutch.mean()
```

Out[5]:
```
aaa          26.482418
lor           3.922828
enroll       81.151124
speaking    517.777037
dtype: float64
```

In [6]:
```
# Mean speaking score for men and women
mean_values = Dutch.groupby("sex")["speaking"].mean()
print(mean_values)
```

```
sex
Female    524.076221
Male      505.451586
Name: speaking, dtype: float64
```

cii) Test whether the difference between men and women is statistically significant, at the 95% level, with a t-test.

In [7]:
```
# Independent t-test = a parametric test used to test for a statistically si
# Alpha value is .05 by default
stats.ttest_ind(Dutch["speaking"][Dutch["sex"]=="Female"], Dutch["speaking"]
```

Out[7]:
```
Ttest_indResult(statistic=53.27989320857528, pvalue=0.0)
```

t = 53.28, p=0.0 If p < 0.05 then there is a statistically significant difference in the means between the two groups. Therefore we can reject the null hypothesis.

ciii) State the null and alternative hypotheses for the t-test.

The null hypothesis is that there is no difference in mean speaking scores between men and women.

The alternative hypothesis is that there is a statistically significant difference in the mean speaking scores between men and women.

civ) Choosing an appropriate plot type, plot the relationship between sex and speaking score. Comment on your results. [10]

In [8]:
```
# Plotting the relationship between sex and speaking score
# Make separate data frames
dutch_Female = dutch[dutch["sex"]=="Female"]
dutch_Male = dutch[dutch["sex"]=="Male"]

sns.kdeplot(dutch_Female["sex"]=="Female"],  color='b', shade="True", bw_adj
sns.rugplot(dutch_Male["brother"],  color='b', height=0.1) # plot individual
```

```
  File "<ipython-input-8-69335189202b>", line 6
    sns.kdeplot(dutch_Female["sex"]=="Female"],  color='b', shade="True", b
w_adjust=1.0) # plot the KDE
                                    ^
SyntaxError: closing parenthesis ']' does not match opening parenthesis '('
```

Comment on your results

Conduct a linear regression analysis, predicting speaking score with the following x-variables:

- Age at arrival in Netherlands
- Length of residence in Netherlands
- Enrolment in Dutch secondary schools
- Sex

Report the results including direction of association, size of the coefficients, and the significance level. [10]

d) Interpreting the output

The regression equation is y = -0.721a + 537.0. The t-value is -31.265. There is a negative correlation between age at arrival and speaking score but this is not statistically significant.

In [13]:
```
#### Speaking score and age at arrival (aaa)

# run the regression model
# Tell statsmodels where to find the data and the explanatory variables
reg_formula = sm.regression.linear_model.OLS.from_formula(data = Dutch, form

# Fit the regression (work out the values of intercept and slope)
# the output is a structure which we will call reg_results
reg_results = reg_formula.fit()

# Summary of the regression results
reg_results.summary()
```

Out[13]:

OLS Regression Results

Dep. Variable:	speaking	**R-squared:**	0.019
Model:	OLS	**Adj. R-squared:**	0.019
Method:	Least Squares	**F-statistic:**	977.5
Date:	Wed, 26 Apr 2023	**Prob (F-statistic):**	1.63e-212
Time:	10:49:04	**Log-Likelihood:**	-2.5060e+05
No. Observations:	49681	**AIC:**	5.012e+05
Df Residuals:	49679	**BIC:**	5.012e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	537.0317	0.638	841.134	0.000	535.780	538.283
aaa	-0.7271	0.023	-31.265	0.000	-0.773	-0.681

Omnibus:	1415.351	**Durbin-Watson:**	1.897
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3520.070
Skew:	0.084	**Prob(JB):**	0.00
Kurtosis:	4.293	**Cond. No.**	104.

Notes:

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

d) Interpreting the output

The regression equation is y = 0.0553a + 517. The t-value is 1.392. This is > 1.96 which means the correlation is not statistically significant. While the coefficient is small, this shows the direction of the correlation is positive (but very weak).

In [9]:

```
#### Speaking score and length of residence (lor)

# run the regression model
# Tell statsmodels where to find the data and the explanatory variables
reg_formula = sm.regression.linear_model.OLS.from_formula(data = Dutch, form

# Fit the regression (work out the values of intercept and slope)
# the output is a structure which we will call reg_results
reg_results = reg_formula.fit()

# Summary of the regression results
reg_results.summary()
```

Out [9]:

OLS Regression Results

Dep. Variable:	speaking	**R-squared:**	0.000
Model:	OLS	**Adj. R-squared:**	0.000
Method:	Least Squares	**F-statistic:**	1.939
Date:	Wed, 26 Apr 2023	**Prob (F-statistic):**	0.164
Time:	10:41:34	**Log-Likelihood:**	-2.5108e+05
No. Observations:	49681	**AIC:**	5.022e+05
Df Residuals:	49679	**BIC:**	5.022e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	517.5600	0.231	2243.499	0.000	517.108	518.012
lor	0.0553	0.040	1.392	0.164	-0.023	0.133

Omnibus:	1364.315	**Durbin-Watson:**	1.898
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3321.716
Skew:	0.084	**Prob(JB):**	0.00
Kurtosis:	4.256	**Cond. No.**	7.98

Notes:

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

d) Interpreting the output

The regression equation is y = 0.555a + 473. The t-value is 89.4 which means this is a stastically significant correlation (speaking score and enrollment in Dutch secondary schools). Again, the correlation is very weakly positive.

In [10]:

```
#### Speaking score and enrollment in Dutch secondary schools (enroll)

# run the regression model
# Tell statsmodels where to find the data and the explanatory variables
reg_formula = sm.regression.linear_model.OLS.from_formula(data = Dutch, form

# Fit the regression (work out the values of intercept and slope)
# the output is a structure which we will call reg_results
reg_results = reg_formula.fit()

# Summary of the regression results
reg_results.summary()
```

Out[10]:

OLS Regression Results

Dep. Variable:	speaking	**R-squared:**	0.139
Model:	OLS	**Adj. R-squared:**	0.139
Method:	Least Squares	**F-statistic:**	8008.
Date:	Wed, 26 Apr 2023	**Prob (F-statistic):**	0.00
Time:	10:43:40	**Log-Likelihood:**	-2.4737e+05
No. Observations:	49681	**AIC:**	4.947e+05
Df Residuals:	49679	**BIC:**	4.948e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	472.7264	0.528	896.020	0.000	471.692	473.760
enroll	0.5551	0.006	89.487	0.000	0.543	0.567

Omnibus:	1664.078	**Durbin-Watson:**	1.979
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4695.343
Skew:	0.053	**Prob(JB):**	0.00
Kurtosis:	4.502	**Cond. No.**	284.

Notes:

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

d) Interpreting the output

The regression equation is y = -18.624 + 524. The t-value is -53.280. The correlation according to the equation suggests that sex has a negative effect on the speaking score, but this is not statistically significant.

In [11]:

```
#### Speaking score and sex

# run the regression model
# Tell statsmodels where to find the data and the explanatory variables
reg_formula = sm.regression.linear_model.OLS.from_formula(data = Dutch, form

# Fit the regression (work out the values of intercept and slope)
# the output is a structure which we will call reg_results
reg_results = reg_formula.fit()

# Summary of the regression results
reg_results.summary()
```

Out[11]:

OLS Regression Results

Dep. Variable:	speaking	**R-squared:**	0.054
Model:	OLS	**Adj. R-squared:**	0.054
Method:	Least Squares	**F-statistic:**	2839.
Date:	Wed, 26 Apr 2023	**Prob (F-statistic):**	0.00
Time:	10:45:29	**Log-Likelihood:**	-2.4970e+05
No. Observations:	49681	**AIC:**	4.994e+05
Df Residuals:	49679	**BIC:**	4.994e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	524.0762	0.203	2577.934	0.000	523.678	524.475
sex[T.Male]	-18.6246	0.350	-53.280	0.000	-19.310	-17.939

Omnibus:	1566.188	**Durbin-Watson:**	2.007
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3757.253
Skew:	0.148	**Prob(JB):**	0.00
Kurtosis:	4.314	**Cond. No.**	2.41

Notes:

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

e) Conduct a second linear regression analysis, including 'family' as an additional x-variable. Making note of the reference category, interpret the results. [4]

In [12]:

```
#### Second linear regression analysis with 'family' as additional x-variabl
#### Speaking score (reference category), age at arrival, and family

# run the regression model
# Tell statsmodels where to find the data and the explanatory variables
reg_formula = sm.regression.linear_model.OLS.from_formula(data = Dutch, form

# Fit the regression (work out the values of intercept and slope)
# the output is a structure which we will call reg_results
reg_results = reg_formula.fit()

# Summary of the regression results
reg_results.summary()
sp_aaa_fam = reg_results.summary()
sp_aaa_fam
```

Out[12]:

OLS Regression Results

Dep. Variable:	speaking	**R-squared:**	0.131
Model:	OLS	**Adj. R-squared:**	0.131
Method:	Least Squares	**F-statistic:**	680.9
Date:	Wed, 26 Apr 2023	**Prob (F-statistic):**	0.00
Time:	10:48:34	**Log-Likelihood:**	-2.4759e+05
No. Observations:	49681	**AIC:**	4.952e+05
Df Residuals:	49669	**BIC:**	4.953e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	521.3207	0.708	736.581	0.000	519.933	522.708
family[T.Altaic]	-0.5466	0.764	-0.715	0.474	-2.044	0.951
family[T.Austro-Asiatic]	-8.2126	2.651	-3.098	0.002	-13.409	-3.016
family[T.Austronesian]	3.1618	0.919	3.439	0.001	1.360	4.964
family[T.Dravidian]	22.9854	5.065	4.538	0.000	13.058	32.913
family[T.Indo-European]	27.1179	0.453	59.875	0.000	26.230	28.006
family[T.Japanese]	3.7220	2.160	1.723	0.085	-0.511	7.955
family[T.Korean]	0.1517	4.436	0.034	0.973	-8.543	8.846
family[T.Niger-Congo]	-3.8933	1.705	-2.284	0.022	-7.235	-0.552
family[T.Sino-Tibetan]	-6.1995	1.239	-5.002	0.000	-8.629	-3.770
family[T.Uralic]	39.0025	1.143	34.122	0.000	36.762	41.243
aaa	-0.8723	0.022	-39.465	0.000	-0.916	-0.829

Omnibus:	1563.172	**Durbin-Watson:**	1.956
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3859.460
Skew:	0.132	**Prob(JB):**	0.00
Kurtosis:	4.340	**Cond. No.**	878.

Notes:

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

f) Report the R-squared for the two models (the model specified in part d) and the model specified in part e)) and give an interpretation in words. [5]

di) R-squared is 0.019. This means the age at arrival in the Netherlands explains 0.019% of the variability in speaking scores. As this is close to 0, this means the least squares line is not very effective in predicting y (explained very little of the variation in the dependent variable). dii) R-squared is 0.000. This means that the length of residence does not explain any of the variability in speaking scores. diii) R-squared is 0.139. This means that the enrollment in Dutch secondary schools explains 0.139% of variability in speaking scores. div) R-squared is 0.054. This means that sex explains 0.054% of variability in speaking scores.

e) R-squared is 0.131. This means age at arrial in the Netherlands and family explain 0.131% of the variability in speaking scores.

g) Save the residuals from the second model (as specified in part e)) as a new variable and plot them in a histogram. Do you think the assumption of normally distributed residuals has been met? [5]

In [ ]:

```
# Save residuals as new variable - speaking score, age at arrival, family
sm.graphics.plot_regress_exog(reg_formula,
                              'speaking,')
```

h) Find out of there is a significant interaction between sex and length of residence on speaking score. Run a new regression model with the interaction term and interpret. [5]

In [14]:

```
#### Third linear regression analysis with 'family' as additional x-variable
#### Speaking score (reference category), sex, and length of residence

# run the regression model
# Tell statsmodels where to find the data and the explanatory variables
reg_formula = sm.regression.linear_model.OLS.from_formula(data = Dutch, form

# Fit the regression (work out the values of intercept and slope)
# the output is a structure which we will call reg_results
reg_results = reg_formula.fit()

# Summary of the regression results
reg_results.summary()
sp_aaa_fam = reg_results.summary()
sp_aaa_fam
```

Out[14]:

OLS Regression Results

Dep. Variable:	speaking	**R-squared:**	0.054
Model:	OLS	**Adj. R-squared:**	0.054
Method:	Least Squares	**F-statistic:**	1430.
Date:	Wed, 26 Apr 2023	**Prob (F-statistic):**	0.00
Time:	10:53:46	**Log-Likelihood:**	-2.4969e+05
No. Observations:	49681	**AIC:**	4.994e+05
Df Residuals:	49678	**BIC:**	4.994e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	523.4337	0.250	2095.404	0.000	522.944	523.923
sex[T.Male]	-18.7114	0.350	-53.454	0.000	-19.397	-18.025
lor	0.1713	0.039	4.424	0.000	0.095	0.247

Omnibus:	1547.989	**Durbin-Watson:**	2.006
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3706.656
Skew:	0.145	**Prob(JB):**	0.00
Kurtosis:	4.306	**Cond. No.**	13.3

Notes:

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

i) The Root MSE in model 1 (part e)) was 34.486. Without doing any calculations, explain how we can interpret this value. [3]

The least squares regression model provides an estimate of the variability in y- values at each value of x, called the root mean square error (RMSE). This is one measure of how well the regression model fits the data (spread of y-values around the regression line).