Interpreting results in context

COGS 108 Fall 2024
Samyak Mehta
Discussion 5

sam012@ucsd.edu
OH: WED 3-4.50 pm

Materials adapted from Sam Lau WI20

D5 results

While height and weight may affect your soccer prowess, how you perform is what ultimately matters. Let's determine the relationship (French vs. Brazil) between nationality and `overall` score in `df_sub` using `OLS` from `statsmodels`.

The null hypothesis here would be that there is no relationship between nationality and `overall` score.

Store the results from your analysis in the variable `res_1`.

```
outcome_1, predictors_1 = patsy.dmatrices('overall ~ nationality', df_sub)
mod_1 = sm.OLS(outcome_1, predictors_1)
res_1 = mod_1.fit()
```

HOW AN OLS TABLE LOOKS LIKE

OLS Regression Results

Dep. Variable:	overall	R-squared:	0.078
Model:	OLS	Adj. R-squared:	0.077
Method:	Least Squares	F-statistic:	151.9
Date:	Mon, 03 May 2021	Prob (F-statistic):	1.43e-33
Time:	14:22:58	Log-Likelihood:	-5930.3
No. Observations:	1808	AIC:	1.186e+04
Df Residuals:	1806	BIC:	1.188e+04
Df Model:	1		
Covariance Type:	nonrobust		

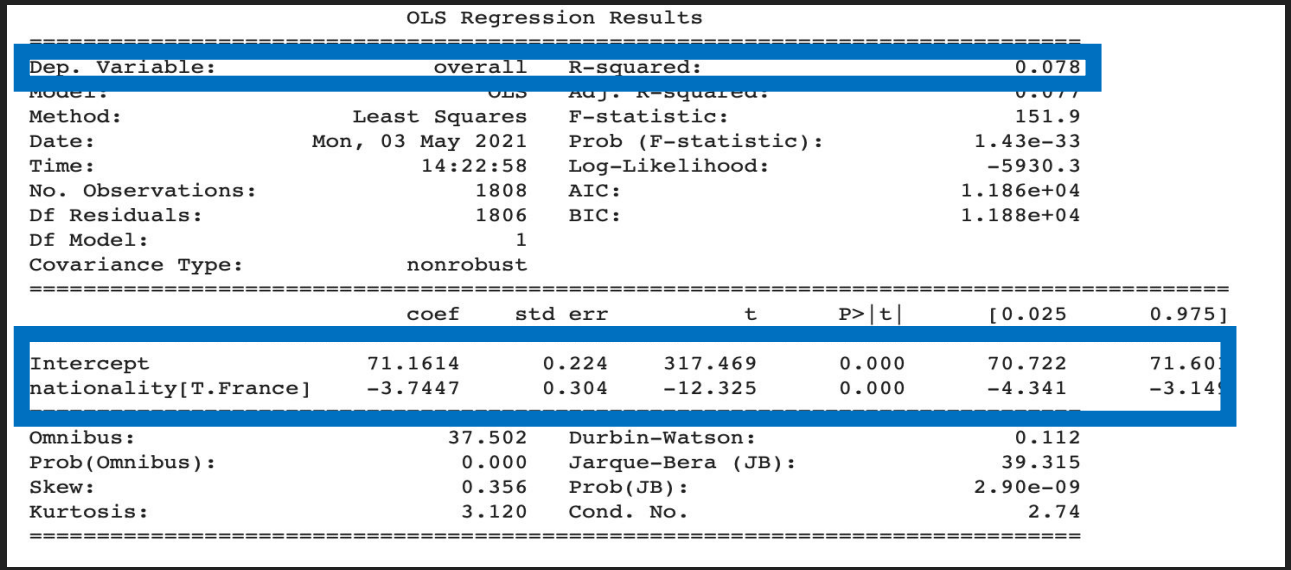
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	71.1614	0.224	317.469	0.000	70.722	71.601
nationality[T.France]	-3.7447	0.304	-12.325	0.000	-4.341	-3.149

Omnibus:	37.502	Durbin-Watson:	0.112
Prob(Omnibus):	0.000	Jarque-Bera (JB):	39.315
Skew:	0.356	Prob(JB):	2.90e-09
Kurtosis:	3.120	Cond. No.	2.74

R-squared measures the percentage of how much of the variance is described by the model

An R-squared of 0 means that the model explains none of the variability, while an R-squared of 1 means that the model explains all of it. In practice, a higher R-squared indicates a better fit, but it does not necessarily mean the model is correct.

Coef: This represents the estimated effect of each independent variable on the dependent variable. Positive coefficients indicate a positive relationship, while negative coefficients indicate a negative one.



OLS Regression Results

Dep. Variable:	overall	R-squared:	0.078
Model:	OLS	Adj. R-squared:	0.077
Method:	Least Squares	F-statistic:	151.9
Date:	Mon, 03 May 2021	Prob (F-statistic):	1.43e-33
Time:	14:22:58	Log-Likelihood:	-5930.3
No. Observations:	1808	AIC:	1.186e+04
Df Residuals:	1806	BIC:	1.188e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	71.1614	0.224	317.469	0.000	70.722	71.60
nationality[T.France]	-3.7447	0.304	-12.325	0.000	-4.341	-3.14

Omnibus:	37.502	Durbin-Watson:	0.112
Prob(Omnibus):	0.000	Jarque-Bera (JB):	39.315
Skew:	0.356	Prob(JB):	2.90e-09
Kurtosis:	3.120	Cond. No.	2.74

std error: This measures the precision of the estimated coefficients. A smaller standard error indicates more confidence in the estimate, while a larger one suggests that the coefficient estimate is less precise.

t: is the t-statistic value. It is a measure of how significant the coefficient is.

P>|t|: This represents the probability that the coefficient is different from zero purely by chance. A small p-value (typically ≤ 0.05) indicates that there is strong evidence against the null hypothesis (i.e., the coefficient is likely to be different from zero and is considered statistically significant). A larger p-value suggests that there is not enough evidence to conclude that the coefficient is different from zero.

Correlations

- Values in correlation table are correlations between pairs of variables.
- Most correlated = correlation furthest away from 0. Not always the most positive value!
- Most correlated with age? Steps
- Most correlated with income? Age

	id	age	steps	income	income10
id	1.00e+00	-6.85e-03	5.56e-03	-0.03	-7.75e-03
age	-6.85e-03	1.00e+00	-2.82e-01	0.27	1.03e-01
steps	5.56e-03	-2.82e-01	1.00e+00	0.05	2.78e-02
income	-2.57e-02	2.67e-01	5.11e-02	1.00	4.70e-01
income10	-7.75e-03	1.03e-01	2.78e-02	0.47	1.00e+00