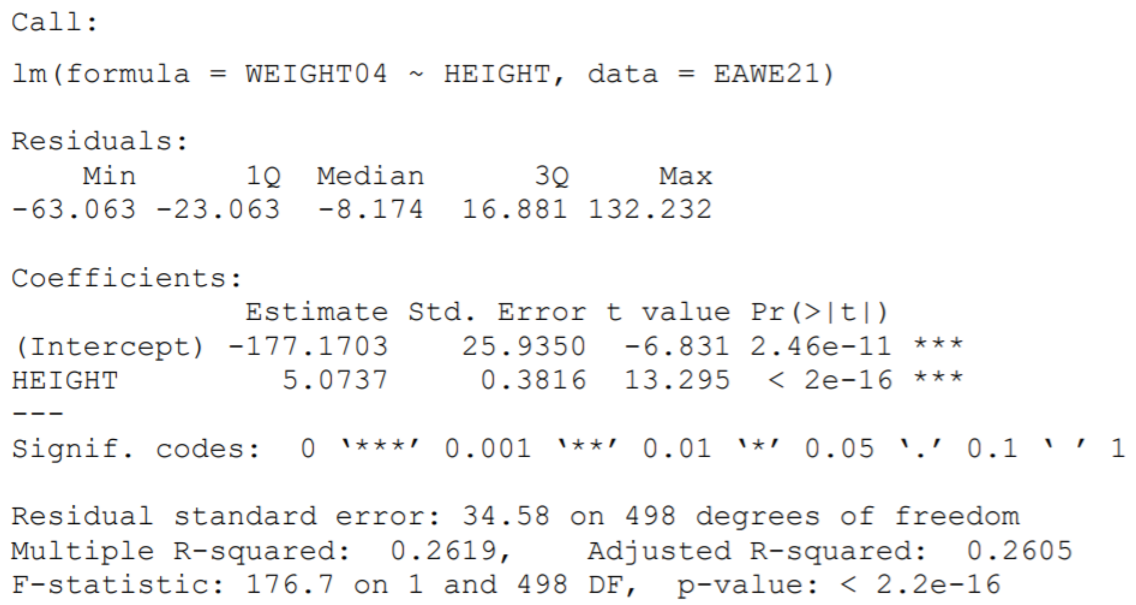
BS2280 – Econometrics I
Homework 3: Interpretation of coefficients and properties of OLS - Solution

1

The output shows the result of regressing the weight of the respondent in 2004, measured in pounds, on his or her height, measured in inches. Provide an interpretation of the coefficients. Does this model provide a good fit?



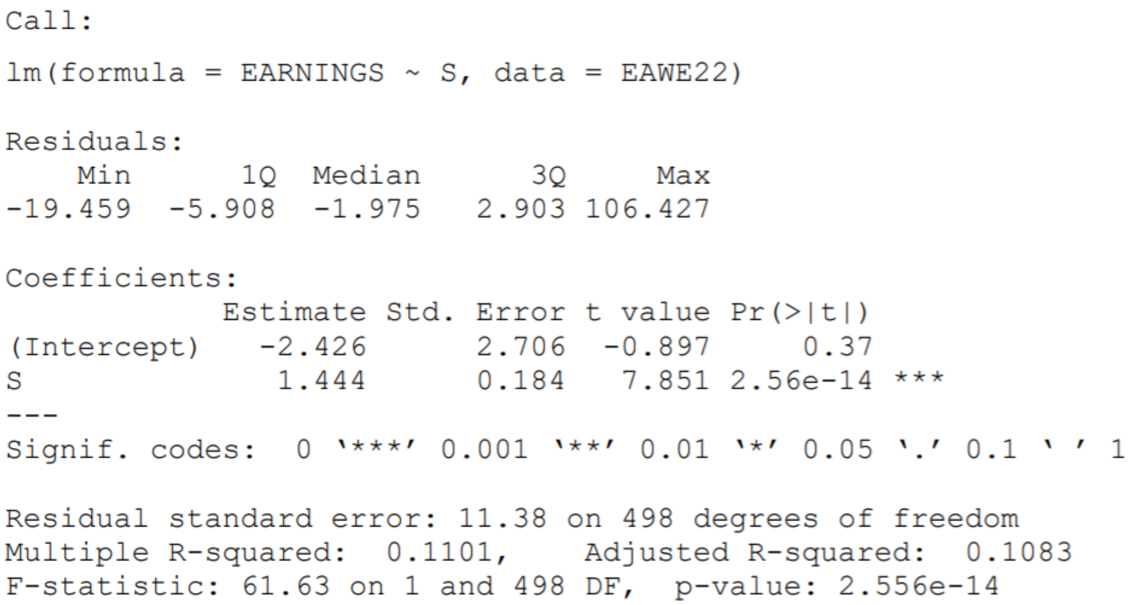
```
Call:
lm(formula = WEIGHT04 ~ HEIGHT, data = EAWE21)

Residuals:
    Min      1Q  Median      3Q     Max 
-63.063 -23.063  -8.174  16.881 132.232 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -177.1703    25.9350  -6.831 2.46e-11 ***
HEIGHT         5.0737     0.3816  13.295  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.58 on 498 degrees of freedom
Multiple R-squared:  0.2619,	Adjusted R-squared:  0.2605 
F-statistic: 176.7 on 1 and 498 DF,  p-value: < 2.2e-16
```

Literally the regression implies that, for every extra inch of height, an individual tends to weigh an extra 5.1 pounds. The intercept, which literally suggests that an individual with no height would weigh -177 pounds, has no meaning.

The overall fit of the model (R^2) is rather low, as only 26% of the variation in weight can be explained by the variation in height.

2

Do earnings depend on education? Use the output table below to give an interpretation of the coefficients. Comment also on R^2.



```
Call:
lm(formula = EARNINGS ~ S, data = EAWE22)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.459  -5.908  -1.975   2.903 106.427 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -2.426      2.706  -0.897     0.37    
S              1.444      0.184   7.851 2.56e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.38 on 498 degrees of freedom
Multiple R-squared:  0.1101,	Adjusted R-squared:  0.1083 
F-statistic: 61.63 on 1 and 498 DF,  p-value: 2.556e-14
```

Of course earnings depend (partly) on education. The regression indicates that each additional year of education increases hourly earnings by $1.44. The interpretation of the constant literally implies that an individual with no years of schooling would earn –$2.43 per hour. There are two ways out of this nonsense.

One is to say that the fitted relationship can claim to be valid only for the range of values of S in the data set (7–20 in this one) and nothing can be inferred for values outside this range.

The other is to say that the relationship may be nonlinear and the negative intercept is an artefact caused by forcing a linear relationship on the observations. The overall fit of the model (R2) is rather low, as only 11% of the variation in earnings can be explained with the variation in years of schooling.

3

What are the 6 main assumptions of OLS? For each assumption explain the implications if it does not hold.

Assumption 1: Model is linear in parameters and correctly specified.
Implications: Poor fit of model
Assumption 2: There is some variation in the X variable.
Implications: betas cannot be estimated
Assumption 3: Disturbance term has zero expectation
Implications: Interpretation of intercept will change
Assumption 4: Disturbance term is homoscedastic
Implications: inefficient estimates/ unreliable results
Assumption 5: Values of disturbance term have independent distributions

Implications: inefficient estimates/ unreliable results
Assumption 6: The disturbance term has a normal distribution
Implications: cannot undertake standard hypothesis tests

A more detailed discussion can be found on Lecture slides for lecture 3.

4

The OLS estimator is BLUE. Explain what BLUE stands for and why OLS is referred to BLUE. (Hint: you can link your answer to your answer of question 3).

BLUE: Best linear unbiased estimator! The Gauss-Markov Theorem states that provided the assumptions hold, OLS estimators will have the lowest variance amongst all possible estimators An estimator with low variance means it comes from a sampling distribution where most of the values are concentrated around the true but unknown population value of the characteristic it is estimating.

See explanation on lecture slides for more information on why OLS estimates are unbiased and efficient.

5

Referring to the equation below, explain what factors determine the variance of $\hat{\beta}_2$. Furthermore, use this formula to explain why OLS will be the most efficient estimator.

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma^2_{u_i}}{nMSD(X)}$$

Three factors will determine variance of $\hat{\beta}_2$

1. Number of observations (n): The larger n, the smaller $\sigma^2_{\hat{\beta}_2}$
2. Variance of residuals: Larger residuals reduce precision of estimates
3. Larger MSD(X) leads to smaller $\sigma^2_{\hat{\beta}_2}$. Low variations in MSD(X) means low variations in X_i → more variations from u_i → higher variance (lower precision) in $\sigma^2_{\hat{\beta}_2}$

OLS is the estimator that minimises the size of the residuals. Therefore it will have a smaller $\sigma^2_{\hat{\beta}_2}$ than any other estimator.