ACTL3142

Week 2: Linear Regression 1

Tetian Madfouni

t.madfouni@unsw.edu.au

Announcements

Your first storywall is up on Moodle

Your assignment is also up on Moodle, have a look now and try to get the gist of it.

This week will be very much focussed on theoretical stuff. This primarily won’t be examinable (proofs and stuff rarely are because you’re doing things on Inspera). BUT it’s important to understand.

Next week will be a lot more focussed on coding exercises and applications.

Overview: Simple Linear Regression

In general, we want to predict a quantitative response Y based on a single predictor variable X. We now let the f from $y = f(x) + \epsilon$ be some linear function

$$Y = \beta_0 + \beta_1 X + \epsilon$$

We use training data to estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ and then can predict Y_i (given $X_i = x$)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

This assumes a few things

- Weak assumptions: $E(\epsilon_i | \boldsymbol{X} = \boldsymbol{x}) = 0,\ Var(\epsilon_i | \boldsymbol{X} = \boldsymbol{x}) = \sigma^2$ and $Cov\left(\epsilon_i, \epsilon_j \middle| \boldsymbol{X} = \boldsymbol{x}\right) = 0$
- Strong assumption

$$\epsilon_i | \boldsymbol{X} = \boldsymbol{x} \sim N(0, \sigma^2)$$

Overview: Fitting a Regression

Two primary methods, what are they?

- Least Squares Estimates (LSE)
- Maximum Likelihood Estimates (MLE)

How does LSE work?

- Minimising the value of some error function, typically $RSS = \sum(y_i - \hat{y}_i)^2$

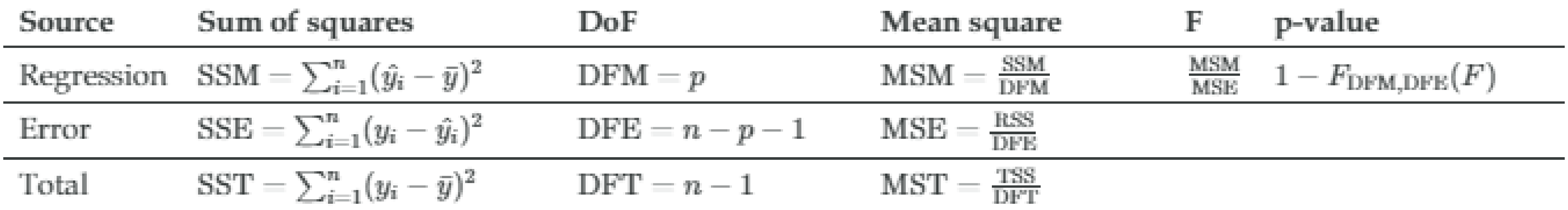
How does MLE work?

- Maximising the likelihood of the data occurring conditional on the parameters, i.e. maximising $L(y; \beta_0, \beta_1, \sigma)$. So then what is log-likelihood and how does it fit?

Are they equivalent?

Overview: ANOVA Table

The table below gives some standard measurements to be used in an ANOVA table



Source	Sum of squares	DoF	Mean square	F	p-value
Regression	$\mathrm{SSM} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$\mathrm{DFM} = p$	$\mathrm{MSM} = \frac{\mathrm{SSM}}{\mathrm{DFM}}$	$\frac{\mathrm{MSM}}{\mathrm{MSE}}$	$1 - F_{\mathrm{DFM},\mathrm{DFE}}(F)$
Error	$\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$\mathrm{DFE} = n - p - 1$	$\mathrm{MSE} = \frac{\mathrm{RSS}}{\mathrm{DFE}}$		
Total	$\mathrm{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$\mathrm{DFT} = n - 1$	$\mathrm{MST} = \frac{\mathrm{TSS}}{\mathrm{DFT}}$		

What is an ANOVA table/what do each of the measurements measure?

- It gives an analysis of variance ☺. It gives various breakdowns of the variance. SST tells us how much the data itself varies. SSM tells us how much of that variance is explained by our model. SSE tells us how much of the variance is remaining after our model does its work.

SLR: Conceptual Q1

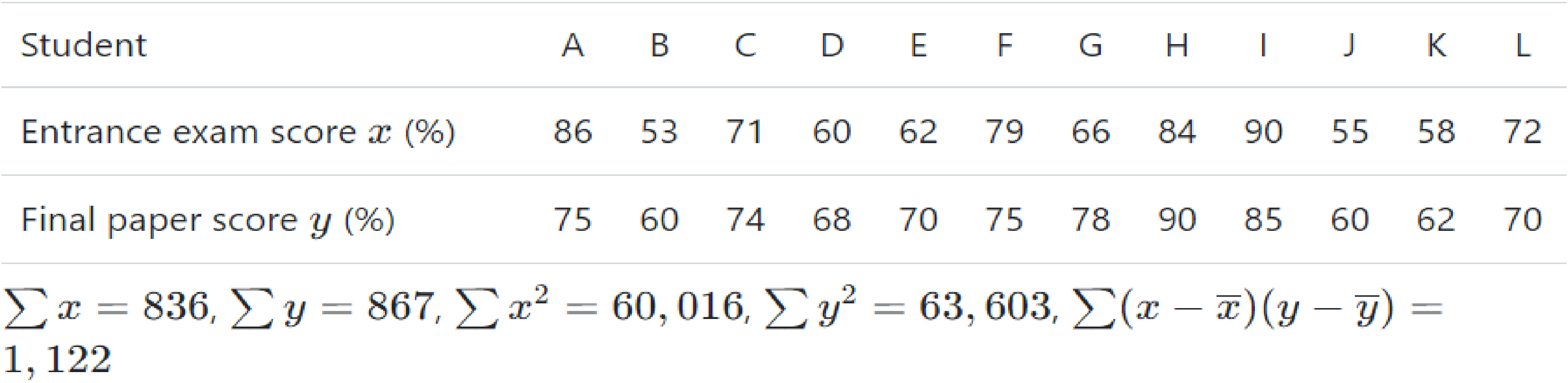
Prove that the Least Squared coefficient estimates (LSE) for β_0 and β_1 are:

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}}$$

SLR: Conceptual Q10

Below are students' scores on entrance exams and final papers. We want to analyse how their entrance exam score influences their final paper score



Student	A	B	C	D	E	F	G	H	I	J	K	L
Entrance exam score x (%)	86	53	71	60	62	79	66	84	90	55	58	72
Final paper score y (%)	75	60	74	68	70	75	78	90	85	60	62	70

$\sum x = 836$, $\sum y = 867$, $\sum x^2 = 60,016$, $\sum y^2 = 63,603$, $\sum (x - \overline{x})(y - \overline{y}) = 1,122$

a) Calculate the fitted linear regression equation of y?

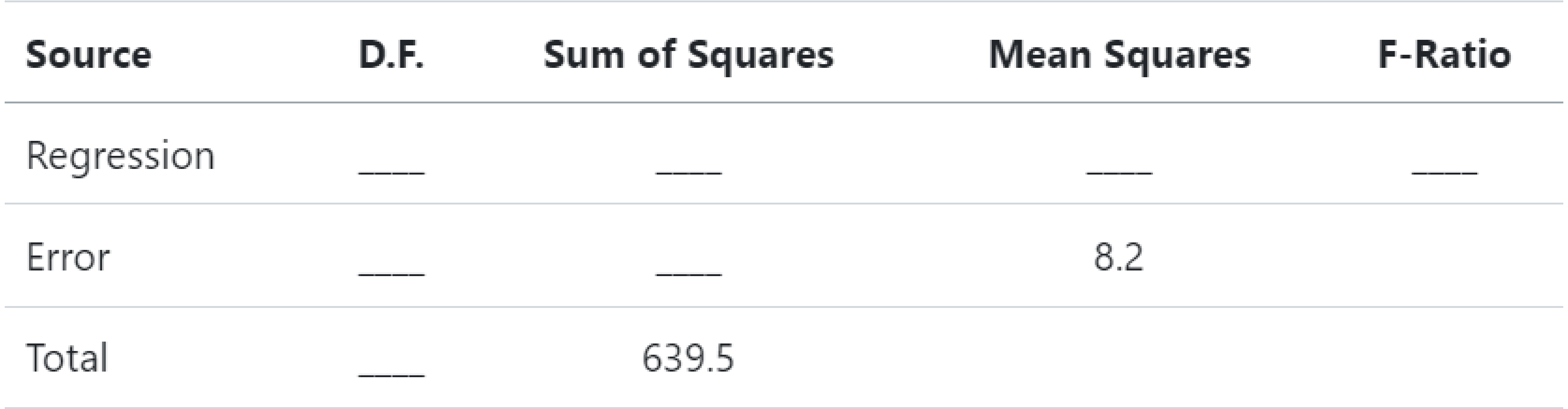
b) Assuming the full normal model, calculate an estimate of the error variance σ^2 and obtain a 90% confidence interval for σ^2.

c) Test whether this data comes from a population with a correlation coefficient equal to 0.75.

d) Calculate the proportion of variance explained by the model. Hence, comment on the fit of the model

SLR: Conceptual Q11

Complete the following ANOVA table for a simple linear regression with 60 observations.



Source	D.F.	Sum of Squares	Mean Squares	F-Ratio
Regression	____	____	____	____
Error	____	____	8.2	
Total	____	639.5		

SLR: Conceptual Q12

Using the output and that $\bar{x} = 2.338, \bar{y} = 40.21, s_x = 2.004,\ s_y = 21.56$ and that $s_x^2 = \frac{S_{xx}}{n-1}$

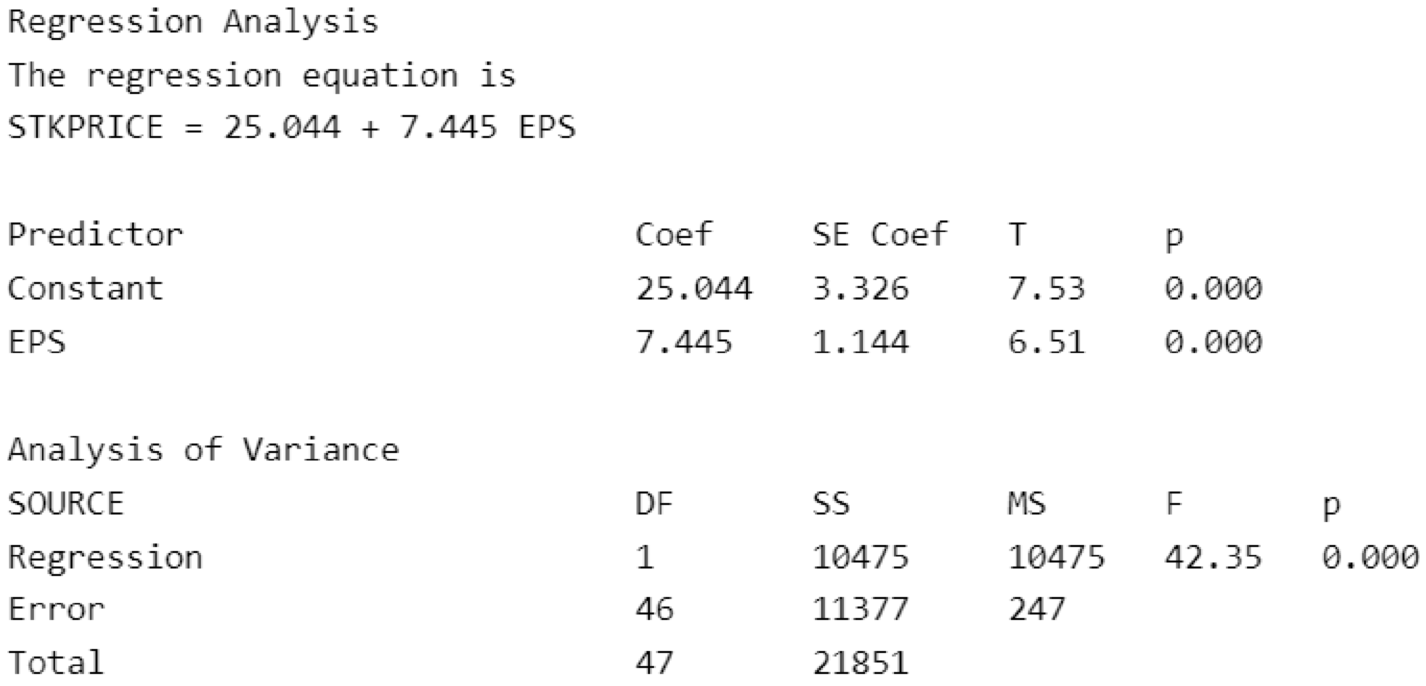
a. Calculate the correlation coefficient of EPS and STKPRICE

b. Estimate STKPRICE given EPS = \$2 with a 95% CI

c. Compute s and R^2 (note s is the estimate for σ^2)

d. How could we check if the errors have constant variance? (One of the weak assumptions)

e. Test the significance of EPS as a predictor at 5% significance level



```
Regression Analysis
The regression equation is
STKPRICE = 25.044 + 7.445 EPS

Predictor                       Coef      SE Coef    T        p
Constant                        25.044    3.326      7.53     0.000
EPS                             7.445     1.144      6.51     0.000

Analysis of Variance
SOURCE                          DF        SS         MS       F        p
Regression                      1         10475      10475    42.35    0.000
Error                           46        11377      247
Total                           47        21851
```

MLR: Conceptual Q3

Suppose we have $n = 100$ observations with a single predictor and a quantitative response. Then we fit a linear regression model to the data and a cubic regression i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

a. Suppose the true relationship is linear. Would the training RSS for the linear vs. cubic differ? If so which would be higher or do we not have enough information to tell?

b. Do the same for test RSS

c. Suppose the relationship is non-linear, but we don’t know exactly what it is. Would the training RSS for the linear vs. cubic differ? If so which would be higher or do we not have enough information to tell?

d. Do the same for test RSS