Lab week 8

Using ANOVA for testing the goodness-of-fit of regression

Exercise 1. The following results were obtained when each of a series of standard silver solutions was analysed by flame-absorption spectrometry.

Concentration(X)	0	5	10	15	20	25	30
Absorbance(Y)	0.003	0.127	0.251	0.390	0.498	0.625	0.763

I. Determine the slope and the intercept of the calibration plot and their confidence limits.

II. Use ANOVA to test the goodness-of-fit of the linear relationship.

We determine the slope and intercept of the calibration plot using the linear regression model function in R.

$> model < -lm(y \sim x)$

$> summary(model)$

Coefficients:

	Estimate	Std. Error	t value	Pr(> $\|t\|$)
(Intercept)	0.0021071	0.0047874	0.44	0.678
x	0.0251643	0.0002656	94.76	2.48e-09

We read the slope b = 0.0251643 and intercept a = 0.0021071 and we can express the relationship between absorbance and silver concentration as: y= 0.0021071 + 0.0251643 $\cdot$ x

absorbance = 0.0021071 + 0.0251643 concentration

We can extract the confidence intervals for the two estimates using the **confint** function.

$> \ confint(model)$

	2.5 %	97.5 %
(Intercept)	-0.01019924	0.01441353
x	0.02448165	0.02584692

The 95% CI for the intercept is $[-0.01019924, 0.01441353]$ and includes zero. This makes the intercept statistically not significant.

The 95% CI for the slope is $[0.02448165, 0.02584692]$ and it does not include zero. This makes the slope statistically significant.

II. We can also test for significance the slope of the linear model by means of ANOVA. The null hypothesis for the analysis of variance is $H_0 : \beta = 0$ and the alternative hypothesis is $H_a : \beta \neq 0$.

The ANOVA table for regression line is obtained by calculating the variation due to regression(SSR) and the variation around the regression line(SSE).

Source	Sum of Squares (SS)	DF	Mean Squares (MS)	F
Total	TSS= $\sum(y_i - \overline{y})^2$	n-1		
Regression	SSR= $\sum(\hat{y}_i - \overline{y})^2$	1	MSR= $\frac{SSR}{1}$	F= $\frac{MSR}{MSE}$
Error	SSE= $\sum(y_i - \hat{y}_i)^2$	n-2	MSE= $\frac{SSE}{n-2}$	

The $\hat{y}_i$ values are the values of y predicted using the y= 0.0021071 + 0.0251643x equation. These values can be extracted from the linear model using the **predict** function.

> pred < − predict(model)

The residuals, $y_i - \hat{y}_i$, can also be extracted from the model with the **residuals** function.

> resid < − residuals(model)

The total sum of squares TSS= $\sum(y_i - \overline{y})^2$.

> TSS < − sum((y-mean(y))ˆ2)

> TSS

[1] 0.4435157

The sum of squares due to regression SSR= $\sum(\hat{y}_i - \overline{y})^2$.

> SSR < − sum((pred-mean(y))ˆ2)

> SSR

[1] 0.4432689

The sum of squares around regression SSE= $\sum(y_i - \hat{y}_i)^2$.

> SSE < − sum((resid)ˆ2)

> SSE

[1] 0.0002468214

The mean sum of squares are :

MSR=$\frac{SSR}{1} = \frac{0.4432689}{1} = 0.4432689 = 0.44327$

MSE=$\frac{SSE}{n-2} = \frac{0.0002468214}{7-2} = 4.936428e-05 = 0.00005$

The test statistic $F = \frac{MSR}{MSE} = 8979.5$. This value is very significant since it is much greater than the critical value $F_{1,5;0.05}$= 0.001084778 and which can be obtained in R with qf(0.975,1,5,lower.tail=F).

We reject the null hypothesis that states $\beta = 0$.

The ANOVA table can be obtained for the regression model with the command: > anova(model)

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	0.44327	0.44327	8979.5	2.481e-09
Residuals	5	0.00025	0.00005		

The same conclusion can be drawn from the t-test for the slope:

Coefficients:

	Estimate	Std. Error	t value	Pr(> $\|t\|$)
(Intercept)	0.0021071	0.0047874	0.44	0.678
x	0.0251643	0.0002656	94.76	2.48e-09

The p-value = 2.48e-09 is less than 0.05 hence we reject the null hypothesis $\beta = 0$.