Esercizio 1

Data la tabella a doppia entrata seguente:

Tempo libero / Titolo di studio	Cinema	Teatro	Musica	Sport	Totale
Scuola media	25	12	18	45	100
Scuola superiore	76	58	49	67	250
Laurea	39	35	35	11	120
Totale	140	105	102	123	470

Verificare ad un livello di significatività del 99% se esiste una relazione tra titolo di studio e tipo di svago nel tempo libero.

Svolgimento:

Si tratta di eseguire un test χ^2 sull'indipendenza che consiste nel verificare la bontà dell'adattamento delle frequenze osservate con quelle teoriche.

Ho: Le variabili sono indipendenti;
H1: Le variabili non sono indipendenti.

Le frequenze assolute teoriche equivalgono a:

$$\hat{n}_{ij} = \frac{n_{i.}n_{.j}}{n}; \forall (i, j)$$

Si costruisce, pertanto, la seguente tabella:

	Cinema	Teatro	Musica	Sport	totale
Scuola media	29,79	22,34	21,70	26,17	100
Scuola superiore	74,47	55,85	54,26	65,43	250
Laurea	35,74	26,81	26,04	31,40	120
totale	140	105	102	123	470

ove, ad esempio, all'incrocio tra la modalità *scuola media* e la modalità *cinema* il valore 29,79 è dato da (140*100)/470.

Se è vera Ho, la statistica $\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{h}\frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$ si distribuisce come una V.C. χ^2 con (k-1)(h-1) gradi di libertà.

Si può costruire un'ulteriore tabella le cui celle sono formate dalla differenza al quadrato tra frequenze osservate e teoriche diviso la corrispondente frequenza teorica:

	Cinema	Teatro	Musica	Sport
Scuola media	0,77	4,79	0,63	13,55
Scuola superiore	0,03	0,08	0,51	0,04
Laurea	0,30	2,50	3,08	13,26

ove, ad esempio, all'incrocio tra la modalità *scuola media* e la modalità *cinema* il valore 0,77 è dato da $(25-29,79)^2/29,79$.

Si procede alla somma di tutti gli incroci relativi a tale tabella e si ottiene:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{h}\frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = 39,53$$

Il valore assunto dalla statistica va confrontato con quello riportato dalle tavole del χ^2 facendo attenzione ai gradi di libertà, che devono essere (4-1)(3-1)=6. Da queste si evince che

$$\chi^2_{(6;0,01)} = 16,8119 < X^2 = 39,53$$

Pertanto si rigetta l'ipotesi nulla.

Esercizio 2

Un'azienda produttrice di dentifricio ha deciso di comprare spazi pubblicitari in punti strategici della città. Nella seguente tabella sono riportate le vendite (in centinaia) e i manifesti pubblicitari affissi in città.

Vendite	**Numero di manifesti**
18	30
24	35
25	38
11	20
32	45
15	25
40	60
12	21
35	49
21	33

a) Verificare l'esistenza di dipendenza tra il numero di manifesti affissi e il numero di dentifrici venduti.
b) Verificare la significatività dei coefficienti della regressione

Svolgimento:

a).

D'ora in poi chiameremo Y le vendite e X il numero di manifesti.
Per procedere alle stime dei parametri della retta di regressione è conveniente costruire la seguente tabella:

	Y	X	Y^2	X^2	XY	$\hat{Y}$
	18	30	324	900	540	18,998
	24	35	576	1225	840	22,823
	25	38	625	1444	950	25,118
	11	20	121	400	220	11,348
	32	45	1024	2025	1440	30,473
	15	25	225	625	375	15,173
	40	60	1600	3600	2400	41,948
	12	21	144	441	252	12,113
	35	49	1225	2401	1715	33,533
	21	33	441	1089	693	21,293
Totali	233	356	6305	14150	9425	

Innanzitutto si calcolano le medie di Y e di X:

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 23,3 \qquad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = 35,6$$

Le stime dei parametri della retta si ottengono con il metodo dei mini quadrati.

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \qquad \hat{\beta}_1 = \frac{Cod_{Y,X}}{Dev_X} = \frac{\sum_{i=1}^{n} y_i x_i - n\overline{Y}\overline{X}}{\sum_{i=1}^{n} x_i^2 - n(\overline{X})^2}$$

$$Cod_{Y,X} = 9425 - 10 * 23,3 * 35,6 = 1130,2$$

$$Dev_X = 14150 - 10 * 35,6^2 = 1476,4$$

$$\hat{\beta}_1 = \frac{1130,2}{1476,4} = 0,765511$$

$$\hat{\beta}_0 = 23,3 - 0,765511 * 35,6 = -3,95218$$

La retta stimata è, quindi, la seguente:

$$\hat{y}_i = -3,95218 + 0,765511 x_i + \hat{e}_i$$

b). Inferenza sul coefficiente angolare

Per fare inferenza sul coefficiente angolare della retta di regressione occorre conoscere la varianza degli errori. Poiché questa è ignota bisogna stimarla dai dati campionari.
La stima della varianza degli errori è:

$$S_u^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{e}_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$S_u^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{e}_i^2 = \frac{1}{8} * 10,9199 = 1,3650$$

Avendo dovuto stimare la varianza degli errori, la statistica da usare per il test è:

$$T_1 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{Var(\hat{\beta}_1)}}$$ che si distribuisce come una t di Student con n-2 g.d.l.

La varianza dello stimatore $\hat{\beta}_1$, come è noto, è:

$$Var\left(\hat{\beta}_1\right) = \frac{S_u^2}{Dev(X)} = \frac{1,3650}{1476,4} = 0,0009245$$

A questo punto si può procedere al test:

Ho: $\beta_1 = 0$
H1: $\beta_1 \neq 0$

$$T_1 = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{Var(\hat{\beta}_1)}} = \frac{0,765511 - 0}{\sqrt{0,0009245}} = 25,1762$$

Confrontando tale valore con quello ottenuto dalle tavole della V.C. t di student con 8 gradi di libertà ad un livello di significatività, ad esempio, dell'1% (pari a 3,3553) si può rigettare l'ipotesi nulla.

b.2) Inferenza sull'intercetta.

Partendo dalla varianza dello stimatore $\hat{\beta}_0$:

$Var(\hat{\beta}_0) = S_u^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{Dev(X)} \right] =$ sviluppando il prodotto si ha:

$\frac{S_u^2}{n} + \frac{S_u^2 \overline{X}^2}{Dev(X)} =$ mettendo in evidenza $\frac{S_u^2}{n}$ si ha:

$\frac{S_u^2}{n} \left[1 + \frac{n\overline{X}^2}{Dev(X)} \right]$

QUINDI le formule della varianza dello stimatore $\hat{\beta}_0$ sono identiche.

$$Var\left(\hat{\beta}_0\right) = \frac{S_u^2}{n} \left[1 + \frac{n(\overline{x})^2}{dev(x)} \right]$$

Quindi $Var\left(\hat{\beta}_0\right) = \frac{S_u^2}{n} \left[1 + \frac{n(\overline{x})^2}{dev(x)} \right] = \frac{1,3650}{10} \left[1 + \frac{10(35,6)^2}{1476,4} \right] = 1,3082$

$H_0 : \beta_0 = 0$
$H_1 : \beta_0 \neq 0$

$$T_0 = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{Var(\hat{\beta}_0)}} \sim t_{(n-2)}$$

$$T_0 = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{Var(\hat{\beta}_0)}} = \frac{-3,95218}{\sqrt{1,3082}} = -3,4554$$

$$t_{0,005;8} = \pm 3,3553$$

Siccome -3,4454 < -3,3553 si rigetta l’ipotesi nulla

INDICE DI DETERMINAZIONE LINEARE

Effettuiamo il calcolo della devianza di Y:

$$Dev_Y = \sum_{i=1}^{n} y_i^2 - n(\overline{Y})^2 = 6305 - 10 * 23,3^2 = 876,1$$

$$R^2 = \hat{\beta}_1^2 \frac{Dev_X}{Dev_Y} = (0,765511)^2 \times \frac{1476,4}{876,1} = 0,9875$$

$$H_0 : R^2 = 0$$

$$H_1 : R^2 > 0$$

$$F = \frac{R^2}{1-R^2}(n-2) = 632$$

confrontato con $F_{\alpha=0,01; n_{Num}=1; n_{Den}=n-2} = 11,26$ si ha che

$F > F_c$ per cui si rifiuta l’ipotesi nulla.

PRECISAZIONE:

Nel modello di regressione lineare semplice si dimostra che $\rho_{y,x}^2 = R^2$, e siccome $\rho_{y,x} = \hat{\beta}_1 \frac{s_x}{s_y}$, fare inferenza sull’indice di determinazione lineare equivale a fare inferenza sul coefficiente angolare della regressione.

Infatti:

$$(T_1)^2 = \frac{\left(\hat{\beta}_1\right)^2}{Var\left(\hat{\beta}_1\right)} = \frac{\left(\hat{\beta}_1\right)^2 nVar_X}{S_u^2} = \frac{\left(Cov_{Y,X} / Var_X\right)^2 nVar_X}{\frac{n}{n-2} Var_Y \left(1 - \left[\rho_{Y,X}\right]^2\right)} = (n-2)\frac{\left(\rho_{Y,X}\right)^2}{1-\left(\rho_{Y,X}\right)^2} = (n-2)\frac{R^2}{1-R^2}$$

Da cui $$R^2 = \frac{(T_1)^2}{(T_1)^2 + n - 2} = \left[1 + \frac{n-2}{(T_1)^2}\right]^{-1}$$

Risulta, quindi, equivalente eseguire il test sull'indice di bontà di adattamento.