Problem 2.12 Table 2.10 refers to applicants to graduate school at the University of California at Berkeley, for fall 1973. It presents admissions decisions by gender of applicant for the six largest graduate departments.

Denote the three variables by:

A=whether admitted,
G=gender, and
D=department.

Table 2.10

	Whether Admitted				Totals		
	Male		Female				
Department	Yes	No	Yes	No	Male	Female	All
A	512	313	89	19	825	108	933
B	353	207	17	8	560	25	585
C	120	205	202	391	325	593	918
D	138	279	131	244	417	375	792
E	53	138	94	299	191	393	584
F	22	351	24	317	373	341	714
	1198	1493	557	1278	2691	1835	4526

Find the sample AG conditional odds ratios and the marginal odds ratio. Interpret, and explain why they give such different indications of the AG association.

Dept	Gender	Yes	No	OR		
A	Male	512	313	0.35	P(A=Yes| G=M, D=A)=512/825 = 0.621	P(A=Yes| G=F, D=A) = 89/108 = 0.824
A	Female	89	19			
B	Male	353	207	0.80	P(A=Yes| G=M, D=B)=512/560 = 0.630	P(A=Yes| G=F, D=B) = 17/25 = 0.680
B	Female	17	8			
C	Male	120	205	1.13	P(A=Yes| G=M, D=C)=512/325 = 0.369	P(A=Yes| G=F, D=C) =202/593 = 0.341
C	Female	202	391			
D	Male	138	279	0.92	P(A=Yes| G=M, D=D)=512/417 = 0.331	P(A=Yes| G=F, D=D) =131/375 = 0.349
D	Female	131	244			
E	Male	53	138	1.22	P(A=Yes| G=M, D=E)=512/191 = 0.277	P(A=Yes| G=F, D=E) = 94/393 = 0.239
E	Female	94	299			
F	Male	22	351	0.83	P(A=Yes| G=M, D=F)=512/373 = 0.059	P(A=Yes| G=F, D=F) = 24/341 = 0.070
F	Female	24	317			
All	Male	1198	1493	1.84	P(A=Yes| G=M)=1198/2691 = 0.445	P(A=Yes| G=F) = 557/1835 = 0.304
All	Female	557	1278			

Example calculations: (The table above was generated in Excel, using methods below)

Marginal Odds Ratio = P(M=Y)/P(F=Y) * (1-P(F=Y))/(1-P(M=Y))
= 0.445/0.304 * (1-0.304)/(1-0.445) = 1.84

Partial Table Odds Ratio for Department A =
P(A=Yes| G=M, D=A)/ P(A=Yes| G=F, D=A) * (1-P(F=Y))/(1-P(M=Y))
= 0.621/0.824 * (1-0.824)/(1-0.621) = 0.35

" … *Partial tables* display the XY (*Gender-Admitted*) relationship while removing the effect of Z (*Department*) by holding its value constant. The associations in partial tables are called *conditional associations*, because they refer to the effect of *X* on *Y* conditional on fixing *Z* at some level.

The two-way contingency table obtained by combining the partial tables is called the XY (*Gender-Admitted) marginal table. In a marginal table,* each cell count is a sum of counts from the same

> location in the partial tables. The marginal table, rather than controlling Z (Department), ignores it.
>
> Conditional associations in partial tables can be quite different from associations in marginal tables. In fact, it can be misleading to analyze only marginal tables of a multiway contingency table …"

In general, there are more male applicants (59%) to these departments than female applicants (41%). The marginal odds ratio is 1.84, which shows that when you ignore individual departments, the odds that male applicants would be accepted was 1.84 times the odds that female applicants would be accepted.

However, when you look at individual departments, you will see that in some departments, like C, the number of female applicants (65%) is higher than the number of male applicants while in others, like A, the situation is reversed. The odds ratio by department generally shows that the odds of acceptance favors the underrepresented gender. This phenomenon is due to the strong associations between A and G and between A and D (You should calculate the odds ratios to see it). These associations may possibly be due to an admission policy that favors underrepresented individuals.

Problem 3.9 Table 3.12 shows the diagnosis and if drugs were the recommended treatment for a sample of psychiatric patients.

Table 3.12

Diagnosis	Drugs	No Drugs	Totals	$\hat{\mu}_{ij}$ (Sect. 3.2.1) Drugs	No Drugs	Pearson χ^2 (Eq. 3.10)	df (I-1)(J-1)	p-value
Schizophrenia	105	8	113	74.51	38.49	84.19	4	2.25E-17
Affective disorder	12	2	14	9.23	4.77			
Neurosis	18	19	37	24.40	12.60			
Personality disorder	47	52	99	65.28	33.72			
Special symptoms	0	13	13	8.57	4.43			
Totals	182	94	276					

Note: Since some cell frequencies are below 5, χ^2 is preferred over G^2. The Pearson χ^2 test has a very small p-value (below 0.001) indicating that we can reject the H_o: of independence.

3.9a) Calculate the standardized Pearson residuals for independence.

Pearson standardized residuals (eq. 3.13): $e^s_{ij} = \dfrac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-p_{i+})(1-p_{+j})}}$

(Calculations done in Excel and shown in the table below)

Diagnosis	e^s_{ij} (Eq. 3.13) Drugs	No Drugs
Schizophrenia	7.87	-7.87
Affective disorder	1.60	-1.60
Neurosis	-2.39	2.39
Personality disorder	-4.84	4.84
Special symptoms	-5.14	5.14

The absolute value of the Standardized Pearson residuals shows a difference in magnitude between the different types of diagnoses. This gives some additional evidence that the recommended treatment may not be consistent between diagnoses. *The patients of diagnoses with positive residuals in the column of drugs (Schizophrenia and Affective disorder) are more likely to be treated by drugs than those of the other diagnoses.* The text suggests that a threshold of 2 to 3 be used to reject H_o: independence. The residuals for Schizophrenia, Personality Disorders, and Special Symptoms all exceed 3, which gives some additional evidence that the differences in treatment frequencies for some diagnoses are greater than H_o: independence predicts.

3.9b) Partition chi-squared into three components to describe differences and similarities among the diagnoses, by comparing:

i. The first two rows, (Calculations done in Excel, results shown in the table below)

Table i

Diagnosis	Drugs	No Drugs	Totals	$\hat{\mu}_{ij}$ (Sect. 3.2.1) Drugs	No Drugs	L-R χ^2 (Eq. 3.11)	df (I-1)(J-1)	p-value
Schizophrenia	105	8	113	104.1	8.90	0.7530	1	0.3855
Affective disorder	12	2	14	12.9	1.10			
Totals	117	10	127					

ii. The third and fourth rows, (Calculations done in Excel, results shown in the table below) and,

Table ii

Diagnosis	Drugs	No Drugs	Totals	$\hat{\mu}_{ij}$ (Sect. 3.2.1) Drugs	No Drugs	L-R χ^2 (Eq. 3.11)	df (I-1)(J-1)	p-value
Neurosis	18	19	37	17.68	19.32	0.0149	1	0.9029
Personality disorder	47	52	99	47.32	51.68			
Totals	65	71	136					

iii. The last row to the first and second rows combined and the third and fourth rows combined. (Calculations done in Excel, results shown in the table below.)

Table iii

Diagnosis	Drugs	No Drugs	Totals	$\hat{\mu}_{ij}$ (Sect. 3.2.1) Drugs	No Drugs	L-R χ^2 (Eq. 3.11)	df (I-1)(J-1)	p-value
Schizo. + Aff. Dis.	117	10	127	83.75	43.25	95.7691	2	<.0001
Neuro. + Per. Dis.	65	71	136	89.68	46.32			
Special symptoms	0	13	13	8.57	4.43			
Totals	182	94	276					

These three tests demonstrate interesting relationships between the type of diagnosis and the recommended treatment. In part i, we are comparing Schizophrenia to Affective Disorder, and in both cases a drug treatment is recommended in a high proportion of the cases (93% and 86%, respectively). The L-R χ^2 test statistic is not significant (p-value = 0.39), and we fail to reject the H_o: independence for these two diagnoses. In part ii, we are comparing Neurosis and Personality Disorder, and again the L-R χ^2 test statistic is not significant (p-value = 0.90), and we fail to reject the H_o: independence for these two diagnoses. In the last case (iii), the L-R χ^2 test statistic is highly significant (p-value $\cong$ 0), showing can reject H_o: independence when comparing the three groupings of diagnoses. This is not surprising given the patterns observed in parts i and ii where drug treatments were predominant and evenly divided, respectively. When examining the Special Symptoms treatments we see that drugs were not recommended. Thus the three groups of diagnoses show quite different treatment patterns and the rejection of H_o is logical.

Here the three L-R χ^2 test statistics sum to the L-R χ^2 test statistic of the full 5x2 table since the three partitioning components are independent.

Problem 3.13 "Table 3.13 shows the results of a retrospective study comparing radiation therapy with surgery in treating cancer of the larynx. The response indicates whether the cancer was controlled for at least two years following treatment."

(The data and SAS code to generate Table 3.14 can be found in the Appendix.)

```
SAS Output for Table 3.14 in Problem 3.13                David Slaughter        14
                                                   15:45 Saturday, October 17, 2009

                 The FREQ Procedure

             Table of Treatment by Outcome

Treatment              Outcome(Cancer Controlled?)

Frequency
Expected              Yes       No        Total
--------------------------------------------
Surgery                 21         2         23
                       20.2       2.8
--------------------------------------------
Radiation Therapy       15         3         18
                       15.8       2.2
--------------------------------------------
Total                   36         5         41

       Statistics for Table of Treatment by Outcome

Statistic                        DF        Value       Prob
------------------------------------------------------------
Chi-Square                        1       0.5992     0.4389
Likelihood Ratio Chi-Square       1       0.5948     0.4406

WARNING: 50% of the cells have expected counts less
         than 5. Chi-Square may not be a valid test.

          Fisher's Exact Test
----------------------------------
Cell (1,1) Frequency (F)        21
Left-sided Pr <= F          0.8947
Right-sided Pr >= F         0.3808

Table Probability (P)       0.2755
Two-sided Pr <= P           0.6384

   Odds Ratio (Case-Control Study)
-----------------------------------
Odds Ratio                   2.1000

Asymptotic Conf Limits
95% Lower Conf Limit         0.3116
95% Upper Conf Limit        14.1523

Exact Conf Limits
95% Lower Conf Limit         0.2089
95% Upper Conf Limit        27.5522

Sample Size = 41
```

3.13a) Explain how the *p*-values for Fisher's exact test are calculated, and report and interpret the results of: H_o: independence, odds ratio $\theta = 1$, H_a: $\theta > 1$.

As described in section 3.5.1 the *p*-value is based upon the hypergeometric distribution. The formula for Fisher's Exact Test expresses the distribution of $\{n_{ij}\}$ in terms of only n_{11} because, "given the marginal totals, n_{11} determines the other three cell counts".

(i) For H_a: $\theta > 1$ the *p*-value equals $P(n_{11} \geq t_o)$ where t_o denotes the observed value of n_{11}. The *p*-value is calculated as shown in equation 3.16, which for H_a: $\theta > 1$ becomes the sum:

$$P(n_{11} \geq t_o) = \sum_{t=to}^{n_{1+}} \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{+1}-t}}{\binom{n}{n_{+1}}}$$

In this example,

$$P(n_{11} = 21) = \frac{\binom{23}{21}\binom{18}{15}}{\binom{41}{36}} = .276 \text{ (table probability in SAS output)}$$

$$P(n_{11} = 22) = \frac{\binom{23}{22}\binom{18}{14}}{\binom{41}{36}} = .0939$$

$$P(n_{11} = 23) = \frac{\binom{23}{23}\binom{18}{13}}{\binom{41}{36}} = .0114$$

So, $P(n_{11} \geq 21) = .276 + .0939 + .0114 = .3813$. Since the p-value is large, we fail to reject the null hypothesis of independence. That is, surgery and radiation therapy perform similarly in controlling cancer of the larynx.

(ii) For H_a: θ is not equal to 1, the 2-sided *p*-value equals $P(n_{11} \geq t_o) + P(n_{11} \leq t_1)$ where t_1 is the largest integer t such that $t < t_o$ and its probability $P(n_{11} = t) <$ the table probability $P(n_{11} = t_o)$. It is equal to .6384 as reported in SAS output. Since the p-value is large, we fail to reject the null hypothesis of independence. That is, surgery and radiation therapy perform similarly in controlling cancer of the larynx.

b) The 95% (exact) confidence interval is (.2089, 27.55), which means that we are 95% sure that the odds ratio is somewhere between .21 and 27.55. Since it includes 1, there is no sufficient evidence that surgery and radiation therapy are different in controlling cancer of the larynx. Note that the asymptotic confidence interval is not appropriate here due to the small sample size.

Problem 4.1

a) Because the intercept is close to 0, the ratio of π to x is approximately equal to the slope. So we can conclude as follows: The estimated proportion vote for Buchanan in 2000 was roughly 3% (.0304) of that for Perot in 1996.
b) The fitted value at x=.0774 is -.0003+.0304(.0774) = .00205, which is 3.5 times larger than the observed value .0079. So Palm Beach county appears to be an outlier.
c) With the logit link, the fitted value is now $\hat{\pi} = \dfrac{\exp[-7.164+12.219(.0774)]}{1+\exp[-7.164+12.219(.0774)]} = .001989$, which is very close to the fitted value in part b. However, in terms of logit function, the fitted logit is -7.164+12.219(.0774)= -6.2182, while the observed is log(.0079/(1-.0079)= -4.8330. The ratio of the fitted to the observed logit is now 6.22/4.83=1.29, not as large as in part b. Therefore, in the scatterplot of logit(π) vs. x, the county Palm Beach seems not an as strong outlier as in the scatterplot of π vs. x

Problem 1. The data in Table 4.2, is from an epidemiological survey of 2484 subjects to investigate snoring as a risk factor for heart disease.

The FREQ Procedure

Table of Snoring by Heart_Disease

Snoring Frequency Expected Row Pct	Heart_Disease Yes	No	Total
Never	24 61.1 1.74	1355 1317.9 98.26	1379
Occasionally	35 28.3 5.49	603 609.8 94.51	638
Almost Daily	21 9.4 9.86	192 203.6 90.14	213
Daily	30 11.2 11.81	224 242.8 88.19	254

a) Will you fit the data by the linear probability model (identity link) or the logistic regression (logit link)? Why?

The *link function* specifies the function used by the model to relate the explanatory variable(s) to the response variable. In general, when the random component (response variable) is binomial the Logit link is used. The "structural defect" of the Identity link for the binomial random component is the possibility of improper fitted probabilities exceeding the range (0,1). *However, if the identity link fits much better than logit link in the sampled range (for example, Table 4.2 on page 121), the identity link could be a good choice.*

b) So long as the relative spacing is not changed, different scores do not affect the fitted values for the categories, and thus the chi-squared test statistics and fitted values for all cells are the same. Note that stretching the scale does affect the model coefficient. Just like changing unit from cm to inches, the number changes but the real size does not.

For any link function g: $g[\pi(x)] = \alpha_1 + \beta_1 x$

It is clear that any linear transformation of $x = k_1 + k_2 * x_o$ in either of these equations will result in a linear model with $\alpha_2 = \alpha_1 + k_1 * \beta_1$.,and $\beta_2 = k_2 * \beta_1$ when x is replaced by x_o . Thus we expect that the goodness of fit and the fitted values will remain the same.

On the following pages are the selected SAS output of the Logit model fitted with the three different scoring systems for snoring (i) (0, 2, 6, 8) and (ii) (1,2,4,5). As you can see from the SAS output, the performance of the two Logit models using the two different scoring systems produced identical scaled deviance and predicted values (in bold). There was however, a change in the estimated intercept and β_1 parameters as expected. So long as the relative spacing is not changed, different scores do not affect the fitted values for the categories, and thus the chi-squared test statistics and fitted values for all cells are the same. Note that stretching the scale does affect the model fitting. Just like changing unit from cm to inches, the number changes but the real size does not.

Snoring score (i): (0, 2, 6, 8)

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-3.7126	0.1496	-4.0058	-3.4194	615.78	<.0001
score	1	0.2318	0.0296	0.1739	0.2897	61.49	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Observation Statistics

Observation	hd	count	**Pred**	Xbeta	Std	HessWgt
1	YES	24	0.0238322	-3.712599	0.149612	0.5583404
2	NO	1355	0.0238322	-3.712599	0.149612	31.52297
3	YES	35	0.0373632	-3.248989	0.1123253	1.2588531
4	NO	603	0.0373632	-3.248989	0.1123253	21.68824
5	YES	21	0.0893361	-2.321769	0.1185527	1.7084578
6	NO	192	0.0893361	-2.321769	0.1185527	15.620185
7	YES	30	0.1349178	-1.858159	0.1589318	3.5014501
8	NO	224	0.1349178	-1.858159	0.1589318	26.144161

Snoring score (ii): (1, 2, 4, 5)

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-4.1762	0.1978	-4.5640	-3.7884	445.59	<.0001
score	1	0.4636	0.0591	0.3477	0.5795	61.49	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Observation Statistics

Observation	hd	count	**Pred**	Xbeta	Std	HessWgt
1	YES	24	0.0238322	-3.712599	0.149612	0.5583404
2	NO	1355	0.0238322	-3.712599	0.149612	31.52297
3	YES	35	0.0373632	-3.248989	0.1123253	1.2588531
4	NO	603	0.0373632	-3.248989	0.1123253	21.68824
5	YES	21	0.0893361	-2.321769	0.1185527	1.7084578
6	NO	192	0.0893361	-2.321769	0.1185527	15.620185
7	YES	30	0.1349178	-1.858159	0.1589318	3.5014501
8	NO	224	0.1349178	-1.858159	0.1589318	26.144161

c) Using the score (iii) (0,2,4,5), the fitted logit model is as listed on textbook, page 123. From the SAS output of score (i) and score (ii), their fitted values (or the predicted probabilities) are slightly different. This is because the relative spacings of the two score sets are different.

Snoring score (iii): (0, 2, 4, 5)

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-3.8662	0.1662	-4.1920	-3.5405	541.06	<.0001
score	1	0.3973	0.0500	0.2993	0.4954	63.12	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Observation Statistics

Observation	hd	count	**Pred**	Xbeta	Std	HessWgt
1	YES	24	0.0205074	-3.866248	0.1662144	0.4820847
2	NO	1355	0.0205074	-3.866248	0.1662144	27.217698
3	YES	35	0.0442951	-3.071575	0.104568	1.4816569
4	NO	603	0.0442951	-3.071575	0.104568	25.526832
5	YES	21	0.0930541	-2.276902	0.1193745	1.7722959
6	NO	192	0.0930541	-2.276902	0.1193745	16.203848
7	YES	30	0.1324388	-1.879565	0.1530077	3.446964
8	NO	224	0.1324388	-1.879565	0.1530077	25.737331