1. The following table contains data from a study concerning the distribution of party aliation in a neighborhood. The interest was whether there was an association between registered political parties and the race.

	Race				
Party	White	Black	Hispanic	Asian	Total
Democrat	431	190	180	95	896
Republican	390	248	159	106	902
Independent	406	177	164	82	828
Total	1227	615	503	282	2627

To determine whether there was an association between registered political party and the race, we would like to perform the Chi-square test.

a. We first calculate the table of expected counts. Table below lists the expected counts calculated in SAS. The table is incomplete as results for four cells in the lower-right corner of the table are removed and replaced by a; b; c; d respectively. Complete the table by calculating the values of a; b; c; d. Remember to show your calculation process.

	Race				
Party	White	Black	Hispanic	Asian	Total
Democrat	a	b	171.71	96.19	896
Republican	c	d	172.96	96.89	902
Independent	386.75	193.81	158.73	88.92	828
Total	1227	615	503	282	2627

a + b = (896 - 96.19 - 171.71) = 628.1
c + d = (902 - 96.89 - 172.96) = 632.15
a + c = (1227 - 386.75) = 840.25
b + d = (615 - 193.81) = 421.19
a = 628.1 - b
b = 421.19 - d
c = 840.25 - a
840.25 - a + d = 632.15
840.25 - 628.1 + b + d = 632.15
212.15 + b + d = 632.15
212.15 + 421.19 - d + d = 632.15
633.34 - d = 632.15
d = 633.34 - 632.15 = 1.19
b = 421.19 - 1.19 = 420

c + 1.19 = 632.15; c = 632.15 - 1.19 = 630.96
a + 630.96 = 840.25; a = 840.25 - 630.96 = 209.29
In summary:
a = 209.29
b = 420
c = 630.96
d = 1.19

b. What are the null hypothesis and alternative hypothesis of this test?
H_0 : There is no statistically significant relationship between registered political parties and race.
H_A : There is a statistically significant relationship between registered political parties and race.

c. Performing data analysis in SAS, we got the chi-square test statistic $\chi^2 = 16.50$ (with p-value 0.011). What are the degrees of freedom for this χ^2 test statistic?
df = (r - 1)(c - 1)
df = (3 - 1)(4 - 1)
df = (2)(3) = 6
There are 6 degrees of freedom for this chi-square test.

2. A survey is conducted to see whether people have a certain symptom X. The following table breaks down these results by gender:

Symptom X by Gender

having symptom X	Male	Female
Yes	9	36
No	39	63

a. Under the hypothesis of independence between row and column variables, calculate the expected cell count for the two cells in the column for women.
ECC = (row i total x column j total) / total count in table
ECC = (36) * (36 + 63) / (9 + 36 + 39 + 63)
ECC = (36 * 99) / 147
ECC = 3564 / 147 = 24.2449

b. What is the chi-square χ^2 statistic for this contingency table? Is it significant at the $\alpha = 0.05$ level? Justify your answer. (Use Chi-square distribution table in textbook)

```sas
/* Re-create contingency table */
data symptomx;
    input symptomX gender$ count;
datalines;
1 Male 9
0 Male 39
1 Female 36
0 Female 63
;
run;

proc freq data = symptomx;
    tables symptomX*gender / chisq expected nocol norow nopercent;
    weight count;
run;
```

The FREQ Procedure

Frequency Expected	Table of symptomX by gender		
	gender		
symptomX	Female	Male	Total
0	63 68.694	39 33.306	102
1	36 30.306	9 14.694	45
Total	99	48	147

Statistics for Table of symptomX by gender

Statistic	DF	Value	Prob
Chi-Square	1	4.7215	0.0298
Likelihood Ratio Chi-Square	1	4.9803	0.0256
Continuity Adj. Chi-Square	1	3.9287	0.0475
Mantel-Haenszel Chi-Square	1	4.6894	0.0303
Phi Coefficient		-0.1792	
Contingency Coefficient		0.1764	
Cramer's V		-0.1792	

Based on the SAS output, the χ^2 statistic for this table is 4.7215 with a p-value of 0.0298. df = (2 - 1)(2 - 1) = (1)(1) = 1. The critical value for $\alpha = 0.05$ and 1 degree of freedom is 3.841. Because the χ^2 statistic is greater than the critical value, it is statistically significant, and gives us the evidence to reject the null hypothesis.

c. Calculate the relative odds (i.e., the OR) of having symptom X, for men to women.
   $\hat{OR}$ = ad / bc = (9 * 63) / (36 * 39) = 567 / 1404 = 0.4038

d. Calculate 95% confidence interval for the relative odds (i.e., the OR) in part (c).
   log($\hat{OR}$) = ln(0.4038) = -0.9068
   se(log($\hat{OR}$)) = √[(1/9) + (1/63) + (1/36) + (1/39)] = 0.4247
   Z(95%) = 1.96
   Lower = -0.9068 - 1.96 * 0.4247 = -1.7392
   Upper = -0.9068 + 1.96 * 0.4247 = -0.0744
   exp(-1.7392), exp(-0.0744) = (0.1757, 0.9283)

3. The dataset lowbwt.sas7bdat contains information for the sample of 100 low birth weight infants born in Boston, Massachusetts. The variable grmhem is a dichotomous random variable indicating whether an infant experienced a germinal matrix hemorrhage. The value 1 indicates that a hemorrhage occurred and 0 that it did not. The infants' five-minute apgar scores are saved under the name apgar5, and indicators of toxemia - where 1 represents a diagnosis of toxemia during pregnancy for the child's mother and 0 no such diagnosis - under the variable name tox.
    a. Write down the equation for a logistic regression model where germinal matrix hemorrhage is the response and five-minute apgar score is the predictor, using β_1 to represent the regression coefficient of apgar score.
    $\log\{p / (1 - p)\} = \beta_0 + \beta_1 apgar5$
    Where p represents the grmhem variable, β_0 represents the intercept, β_1 represents the regression coefficient of the apgar score, and apgar5 represents the predictor variable apgar score.
    b. Fit the logistic regression model in part (a). What is $\hat{\beta}_1$, the estimated regression coefficient of apgar score? What's the interpretation of $\hat{\beta}_1$?

```
17 libname ADS534 '/home/u63483466/ADS534/Data';
18     data work.lowbwt;
19     set ADS534.lowbwt;
20 run;
21
22 /* Fit the logistic regression model */
23 proc logistic data = work.lowbwt;
24     model grmhem = apgar5;
25 run;
```

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.3037	0.6191	0.2407	0.6237
apgar5	1	0.2496	0.1044	5.7206	0.0168

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
apgar5	1.284	1.046	1.575

    The value of $\hat{\beta}_1$ is 0.2496. This means that for every 1-unit increase in the five-minute apgar score, the likelihood of grmhem increases by 0.2496. While grmhem has a value of 1 or 0, this means that higher apgar scores increase the likelihood of a germinal matrix hemorrhage changing from 0 to 1, or occurring.
    c. If a particular child has a five-minute apgar score of 3, what is the predicted probability that this child will experience a brain hemorrhage?

```
27 /* Predict the probability for Apgar score of 3 */
28 /* Calculate predicted probabilities */
29 proc logistic data = work.lowbwt;
30    model grmhem = apgar5 / LINK = LOGIT;
31    output out = PredictedProb1 predicted = p_GrmHem;
32 run;
33
34 /* Calculate predicted probability for apgar5 = 3 */
35 data PredictedProb1;
36    set PredictedProb1;
37    if apgar5 = 3;
38 run;
39
40 /* Display predicted probability */
41 proc print data = PredictedProb1 noobs;
42    var apgar5 p_GrmHem;
43 run;
```

apgar5	p_GrmHem
3	0.74126

With an apgar5 score of 3, the probability of experiencing a germinal matrix hemorrhage is 0.74126, or 74.126%.

d. What is the estimated odds ratio of suffering a germinal matrix hemorrhage associated with 1 unit increase in five-minute apgar score?

```
33 /* Determine the OR of germ with 1-unit increase */
34 proc logistic data = work.lowbwt;
35     model grmhem = apgar5 / LINK = LOGIT;
36 run;
```

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
apgar5	1.284	1.046	1.575

Based on the output from SAS, a 1-unit increase in apgar5 score will increase the odds of suffering a germinal matrix hemorrhage by 1.284, and the 95% confidence limits suggest that the value will most likely fall between 1.046 and 1.575.

e. What is the estimated odds ratio of suffering a germinal matrix hemorrhage associated with 3 units increase in five-minute apgar score?

```
/* Predict the probability for Apgar score of 3 */
/* Calculate predicted probabilities */
proc logistic data = work.lowbwt;
   model grmhem = apgar5 / LINK = LOGIT;
   output out = PredictedProb1 predicted = p_GrmHem;
run;

/* Calculate predicted probability for apgar5 = 3 */
data PredictedProb1;
   set PredictedProb1;
   if apgar5 = 3;
run;

/* Display predicted probability */
proc print data = PredictedProb1 noobs;
   var apgar5 p_GrmHem;
run;
```

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
apgar5	1.284	1.046	1.575

Based on the SAS output, the estimated odds ratio associated with a 3-unit increase in apgar5 is 1.284, with 95% confidence limits of (1.046, 1.575).

f. Write down the equation for a logistic regression model where germinal matrix hemorrhage is the response and toxemia status is the predictor, using β_2 to represent the regression coefficient of toxemia status.

$\log\{p / (1 - p)\} = \beta_0 + \beta_2 tox$

In this equation, p represents the grmhem variable, β_0 represents the intercept, and $\beta_2 tox$ represents the change in grmhem given a 1-unit increase in toxemia status.

g. Fit the logistic regression model in part (f) in SAS. What is $\hat{\beta}_2$, the estimated regression coefficient of toxemia status? What's the interpretation of $\hat{\beta}_2$?

```
/* Fit the logistic regression model */
proc logistic data = work.lowbwt;
    model grmhem = tox / LINK = LOGIT;
run;
```

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.5353	0.2946	27.1530	<.0001
tox	1	1.4604	1.0662	1.8761	0.1708

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
tox	4.308	0.533	34.818

$\hat{\beta}_2$, the estimated regression coefficient of toxemia status, is 1.4604. This indicates that an increase from 0 to 1 for toxemia will result in a 1.4604 increase

in grmhem. Given that grmhem is also on a scale from 0 to 1, this means that higher tox scores will likely result in higher odds of a germinal matrix hemorrhage occurring.

h. For a child whose mother was diagnosed with toxemia during pregnancy, what is the predicted probability of experiencing a germinal matrix hemorrhage?

```
43 /* Calculate predicted probabilities */
44 proc logistic data = work.lowbwt;
45    model grmhem = tox / LINK = LOGIT;
46    output out = PredictedProb predicted = p_GrmHem;
47 run;
48
49 /* Calculate predicted probability for toxemia status = 1 */
50 data PredictedProb;
51    set PredictedProb;
52    if tox = 1;
53 run;
54
55 /* Display predicted probability */
56 proc print data = PredictedProb noobs;
57    var tox p_GrmHem;
58 run;
```

tox	p_GrmHem
1	0.95238

Based on a toxemia status of '1', the predicted probability is 0.95238, or 95.238%. This indicates a high likelihood of a germinal matrix hemorrhage occurring with a diagnosis of toxemia.

i. What are the estimated odds of suffering a germinal matrix hemorrhage for children whose mothers were diagnosed with toxemia relative to children whose mothers were not?

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
tox	4.308	0.533	34.818

The odds ratio suggests that a change in toxemia status is met with odds being 4.308 times more likely to suffer a germinal matrix hemorrhage compared to no toxemia being present.