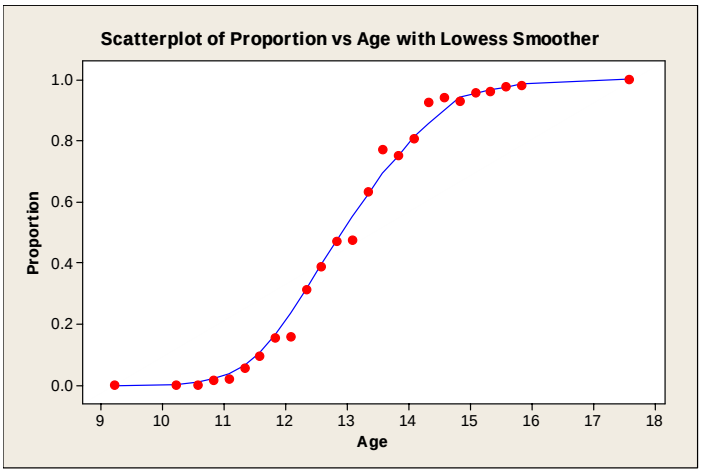
Menarche Data:

The data were entered as N, Freq, and Age as in the notes where N is sample size, Freq is #RM, and Age is the midpoint of the Age interval. Proportion was calculated from these as Freq/N.

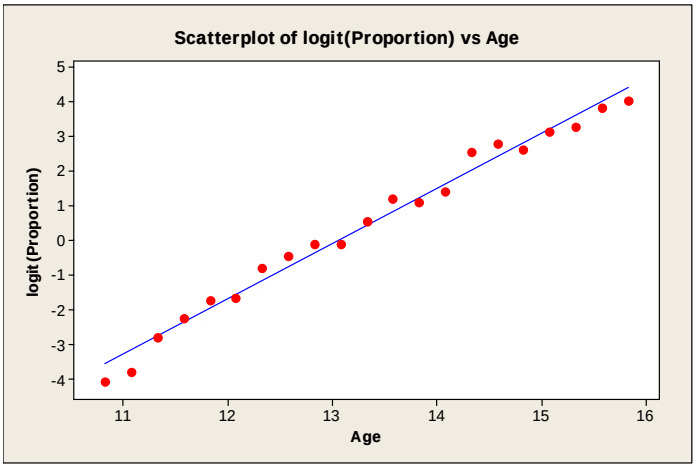
Data Display

Row	N	Freq	Age	Proportion
1	376	0	9.21	0.00000
2	200	0	10.21	0.00000
3	93	0	10.58	0.00000
4	120	2	10.83	0.01667
5	90	2	11.08	0.02222
6	88	5	11.33	0.05682
7	105	10	11.58	0.09524
8	111	17	11.83	0.15315
9	100	16	12.08	0.16000
10	93	29	12.33	0.31183
11	100	39	12.58	0.39000
12	108	51	12.83	0.47222
13	99	47	13.08	0.47475
14	106	67	13.33	0.63208
15	105	81	13.58	0.77143
16	117	88	13.83	0.75214
17	98	79	14.08	0.80612
18	97	90	14.33	0.92784
19	120	113	14.58	0.94167
20	102	95	14.83	0.93137
21	122	117	15.08	0.95902
22	111	107	15.33	0.96396
23	94	92	15.58	0.97872
24	114	112	15.83	0.98246
25	1049	1049	17.58	1.00000

We want a plot of proportion vs. age to see the relationship more clearly. Minitab offers an option in `Graph > Scatterplot > Simple > Data View > Smoother > Lowess` to construct a curve based only on the data with no functional form (such as logs or squares or ….). I set degree of smoothing to .4 instead of the default .5 in this case for a little better behavior. There are a lot of other smoothers out there, but this is all Minitab offers with scatterplots . The lazy-S or sigmoidal shape is clear from this plot. Of course the proportion cannot be smaller than 0 nor larger than 1 no matter what the Age value, so this is a necessary shape.

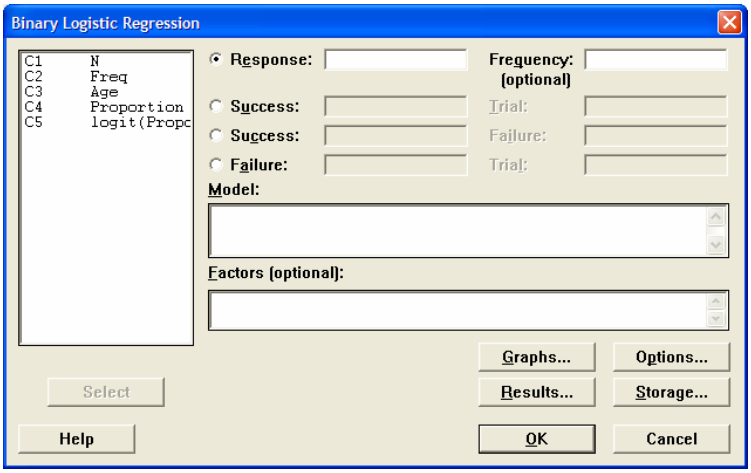


To demonstrate how much sense the logistic regression model makes in this case, I computed $\text{logit(Proportion)}=\log\left(\frac{\text{Proportion}}{\text{1-Proportion}}\right)$ and plotted with a simple regression fit (that is not the line that is fit via maximum likelihood in logistic regression). Minitab will return missing values when Proportion is 0 or 1 (why?).



These data do in fact look fairly linear in the logit scale.

Data formats: In order to fit a formal logistic regression by the methods in the notes, we need to make sure the data are in the correct form. Minitab is a lot more flexible than most packages in how you can structure the data for this analysis. Follow the path `Stat > Regression > Binary Logistic Regression` to bring up the following box.



The four buttons at the top define the four formats allowed (you get the same results for all four). If you click the Help button you get a description of the four options. The first is how "raw" data (one line per observation) is entered; the others are summarized forms (the first can be also). The four possibilities for the menarche data appear in the table below

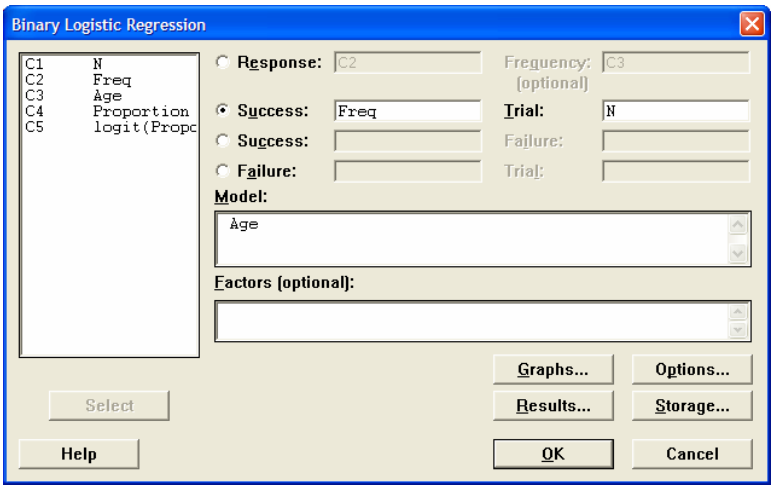
Resp - Freq Form			Succ - Trial Form			Succ-Failure Form			Failure - Trial Form		
Age	Resp.	Freq	Age	# RM	N	Age	# RM	# ~RM	Age	# ~RM	N
9.21	RM	0	9.21	0	376	9.21	0	376	9.21	376	376
9.21	~RM	376	10.21	0	200	10.21	0	200	10.21	200	200
10.21	RM	0	10.58	0	93	10.58	0	93	10.58	93	93
10.21	~RM	200	10.83	2	120	10.83	2	118	10.83	118	120
10.58	RM	0	11.08	2	90	11.08	2	88	11.08	88	90
10.58	~RM	93	11.33	5	88	11.33	5	83	11.33	83	88
10.83	RM	2	11.58	10	105	11.58	10	95	11.58	95	105
10.83	~RM	118	11.83	17	111	11.83	17	94	11.83	94	111
11.08	RM	2	12.08	16	100	12.08	16	84	12.08	84	100
11.08	~RM	88	12.33	29	93	12.33	29	64	12.33	64	93
11.33	RM	5	12.58	39	100	12.58	39	61	12.58	61	100
11.33	~RM	83	12.83	51	108	12.83	51	57	12.83	57	108
11.58	RM	10	13.08	47	99	13.08	47	52	13.08	52	99
11.58	~RM	95	13.33	67	106	13.33	67	39	13.33	39	106
11.83	RM	17	13.58	81	105	13.58	81	24	13.58	24	105
11.83	~RM	94	13.83	88	117	13.83	88	29	13.83	29	117
12.08	RM	16	14.08	79	98	14.08	79	19	14.08	19	98
12.08	~RM	84	14.33	90	97	14.33	90	7	14.33	7	97
12.33	RM	29	14.58	113	120	14.58	113	7	14.58	7	120
12.33	~RM	64	14.83	95	102	14.83	95	7	14.83	7	102
12.58	RM	39	15.08	117	122	15.08	117	5	15.08	5	122
12.58	~RM	61	15.33	107	111	15.33	107	4	15.33	4	111
12.83	RM	51	15.58	92	94	15.58	92	2	15.58	2	94
12.83	~RM	57	15.83	112	114	15.83	112	2	15.83	2	114
13.08	RM	47	17.58	1049	1049	17.58	1049	0	17.58	0	1049
13.08	~RM	52									
13.33	RM	67									
13.33	~RM	39									
13.58	RM	81									
13.58	~RM	24									
13.83	RM	88									
13.83	~RM	29									
14.08	RM	79									
14.08	~RM	19									
14.33	RM	90									
14.33	~RM	7									
14.58	RM	113									
14.58	~RM	7									
14.83	RM	95									
14.83	~RM	7									
15.08	RM	117									
15.08	~RM	5									
15.33	RM	107									
15.33	~RM	4									
15.58	RM	92									
15.58	~RM	2									
15.83	RM	112									
15.83	~RM	2									
17.58	RM	1049									
17.58	~RM	0									

The first form is a little tricky since Minitab has to decide what is an event (Success) and what is not an event (Failure). Minitab's rule (reported from the help system) is:

> ***Reference event for the response variable***
>
> *Minitab needs to designate one of the response values as the reference event. Minitab defines the reference event based on the data type:*
>
> - *For numeric factors, the reference event is the greatest numeric value.*
> - *For date/time factors, the reference event is the most recent date/time.*
> - *For text factors, the reference event is the last in alphabetical order.*
>
> *You can change the default reference event in the Options subdialog box. To change the event, specify the new event value in the* ***Event*** *box.*

Other packages have different rules. Minitab is more flexible about all this than most other packages. If I had wanted ~RM to be the event instead of RM, I would enter "~RM" in the box described above.

The data as presented on page 1 are in the second form, so fill in the box as follows:



With these results:

Binary Logistic Regression: Freq, N versus Age

```
Link Function: Logit


Response Information
```

Variable	Value	Count
Freq	Success	2308
	Failure	1610
N	Total	3918

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-21.2264	0.770656	-27.54	0.000			
Age	1.63197	0.0589509	27.68	0.000	5.11	4.56	5.74

Log-Likelihood = -819.652
Test that all slopes are zero: G = 3667.180, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	21.8699	23	0.528
Deviance	26.7035	23	0.269
Hosmer-Lemeshow	6.7833	5	0.237

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

Value	Success		Failure		
Group	Observed	Expected	Observed	Expected	Total
1	0	2.8	576	573.2	576
2	9	14.2	382	376.8	391
3	72	64.7	337	344.3	409
4	204	198.7	209	214.3	413
5	338	338.7	79	78.3	417
6	432	435.0	23	20.0	455
7	1253	1253.9	4	3.1	1257

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	3599107	96.9	Somers' D	0.94
Discordant	90351	2.4	Goodman-Kruskal Gamma	0.95
Ties	26422	0.7	Kendall's Tau-a	0.46
Total	3715880	100.0		

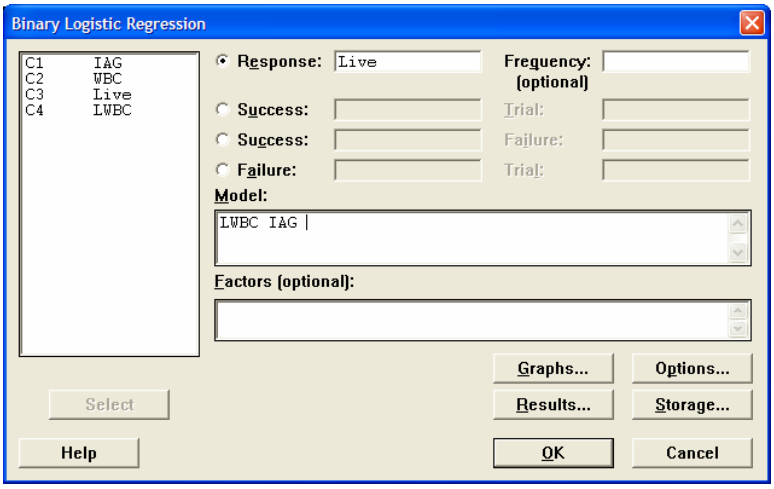
Leukemia Data

Data Display

Row	IAG	WBC	Live	LWBC
1	1	75	1	4.31749
2	1	230	1	5.43808
3	1	430	1	6.06379
4	1	260	1	5.56068
5	1	600	0	6.39693
6	1	1050	1	6.95655
7	1	1000	1	6.90776
8	1	1700	0	7.43838
9	1	540	0	6.29157
10	1	700	1	6.55108
11	1	940	1	6.84588
12	1	3200	0	8.07091
13	1	3500	0	8.16052
14	1	5200	0	8.55641
15	1	10000	1	9.21034
16	1	10000	0	9.21034
17	1	10000	0	9.21034
18	0	440	1	6.08677
19	0	300	1	5.70378
20	0	400	0	5.99146
21	0	150	0	5.01064
22	0	900	0	6.80239
23	0	530	0	6.27288
24	0	1000	0	6.90776
25	0	1900	0	7.54961
26	0	2700	0	7.90101
27	0	2800	0	7.93737
28	0	3100	0	8.03916
29	0	2600	0	7.86327
30	0	2100	0	7.64969
31	0	7900	0	8.97462
32	0	10000	0	9.21034
33	0	10000	0	9.21034

Note that the data are not in the summarized form of the previous data set. Here there is one row of data for each individual in the data set. This is the most common structure for data. The data are coded so that IAG=1 for AG+ and IAG=0 for AG-, while Live=0 if pt died in less than a year and Live=1 if pt lived at least one year. We are trying to model probability of living at least one year, and since Minitab uses the largest numerical value as the reference category, this is the coding we want. If we were using JMP-IN, we would reverse the coding on Live since that package uses the lowest numerical value as the event.

In order to fit the desired model in this form, fill in the dialog box as follows:



Binary Logistic Regression: Live versus LWBC, IAG

```
Link Function: Logit


Response Information

Variable  Value  Count
Live      1         11  (Event)
          0         22
          Total     33


Logistic Regression Table

                                                 Odds      95% CI
Predictor      Coef   SE Coef      Z      P  Ratio  Lower   Upper
Constant    5.54335   3.02242   1.83  0.067
LWBC       -1.10876  0.460948  -2.41  0.016   0.33   0.13    0.81
IAG         2.51956   1.09068   2.31  0.021  12.42   1.46  105.35


Log-Likelihood = -13.416
Test that all slopes are zero: G = 15.177, DF = 2, P-Value = 0.001


Goodness-of-Fit Tests

Method           Chi-Square  DF      P
Pearson             19.8094  27  0.839
Deviance            23.0136  27  0.684
Hosmer-Lemeshow      7.0303   8  0.533


Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)
```

					Group						
Value	1	2	3	4	5	6	7	8	9	10	Total
1											
Obs	0	0	0	1	0	1	1	3	1	4	11
Exp	0.0	0.1	0.1	0.4	0.5	0.8	1.6	1.8	2.2	3.5	
0											
Obs	3	3	3	3	3	2	3	0	2	0	22
Exp	3.0	2.9	2.9	3.6	2.5	2.2	2.4	1.2	0.8	0.5	
Total	3	3	3	4	3	3	4	3	3	4	33

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	210	86.8	Somers' D	0.74
Discordant	30	12.4	Goodman-Kruskal Gamma	0.75
Ties	2	0.8	Kendall's Tau-a	0.34
Total	242	100.0		