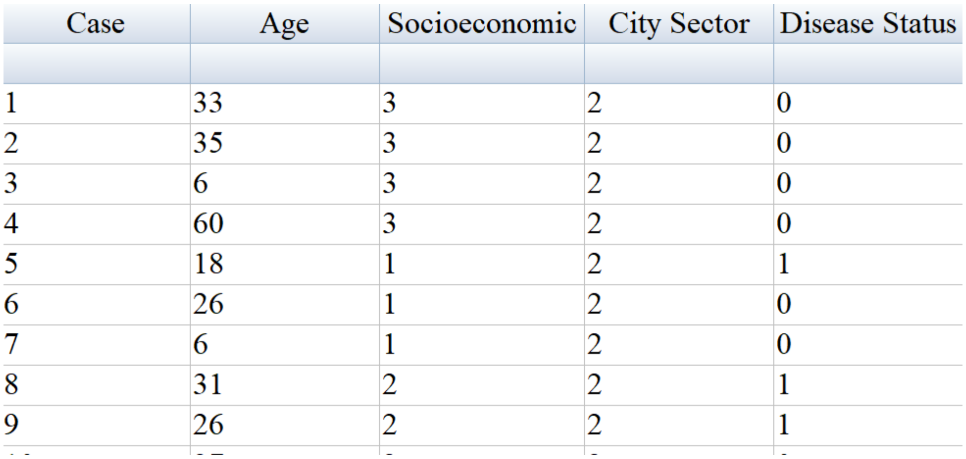
Multiple Logistic Regression

Our next example is borrowed from the text *Applied Linear Statistical Models*.

In the study of an epidemic spread by mosquitoes, people were randomly selected from two sectors of a city. In addition to their sector of residence, their age and socioeconomic status were determined, as well as whether they displayed symptoms consistent with having the disease. The data for 196 individuals is stored in the file *Epidemic*. I will follow the analysis in *Applied Linear Statistical Models* by splitting the data in half, as we would if we wished to test the predictions on half the data made by a model constructed using the other half.

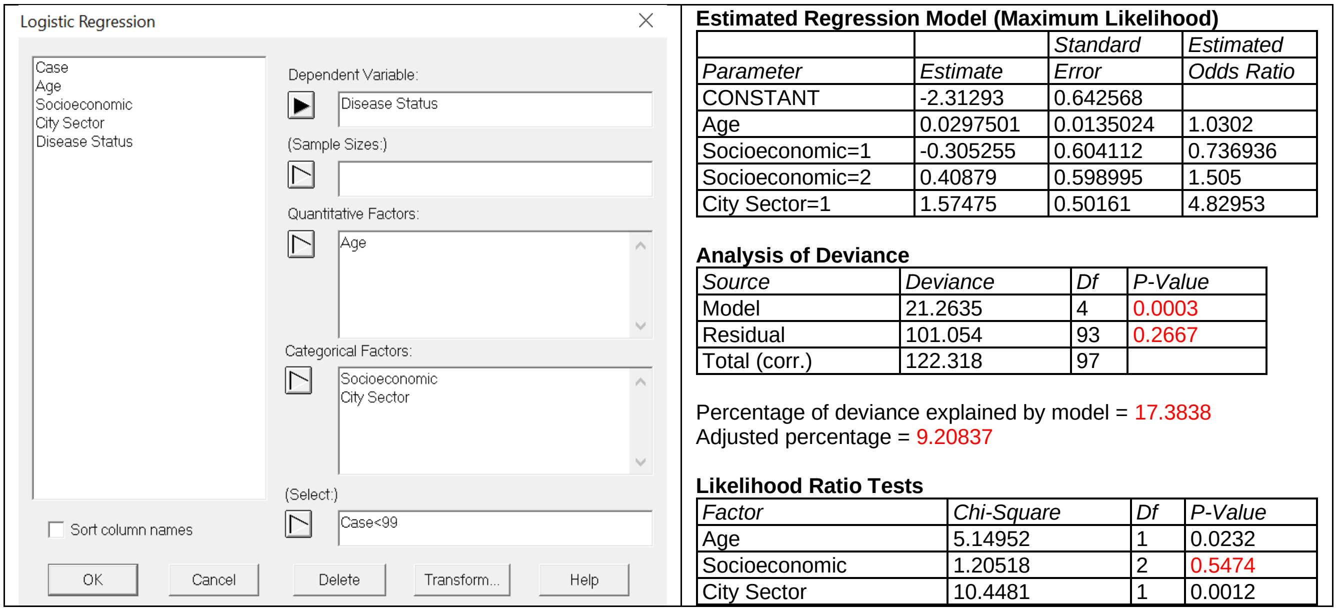
This data introduces two new elements to the analysis. Information on several independent variables was collected, and two of the variables, *Socioeconomic* and *City Sector*, are categorical. The coding used for the variables appears below the portion of the spreadsheet reproduced.



Case	Age	Socioeconomic	City Sector	Disease Status
1	33	3	2	0
2	35	3	2	0
3	6	3	2	0
4	60	3	2	0
5	18	1	2	1
6	26	1	2	0
7	6	1	2	0
8	31	2	2	1
9	26	2	2	1

Socioeconomic: 1 = Lower, 2 = Middle, 3 = Upper: City Sector = 1 or 2

When entering the data into the logistic regression program in Statgraphics, there is the option of entering the categorical variables as is into the *Categorical Factor* field, or recoding them as dummy variables and entering them into the field for *Quantitative Factors*. If we wish to conduct tests of significance for individual levels of the socioeconomic variable, it is more convenient to recode them as dummies. Below, I'll compare the initial model both ways.



Dependent Variable: Disease Status

(Sample Sizes:)

Quantitative Factors: Age

Categorical Factors: Socioeconomic
City Sector

(Select:) Case<99

Estimated Regression Model (Maximum Likelihood)

		Standard	Estimated
Parameter	*Estimate*	*Error*	*Odds Ratio*
CONSTANT	-2.31293	0.642568	
Age	0.0297501	0.0135024	1.0302
Socioeconomic=1	-0.305255	0.604112	0.736936
Socioeconomic=2	0.40879	0.598995	1.505
City Sector=1	1.57475	0.50161	4.82953

Analysis of Deviance

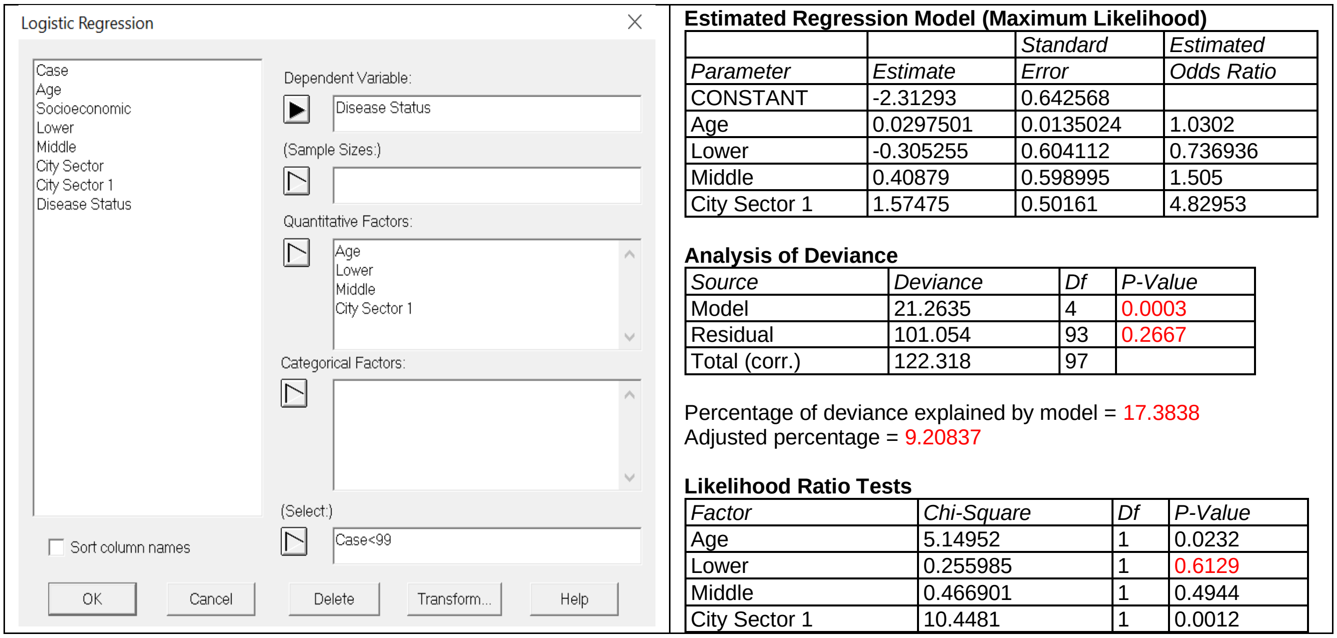
Source	*Deviance*	*Df*	*P-Value*
Model	21.2635	4	0.0003
Residual	101.054	93	0.2667
Total (corr.)	122.318	97	

Percentage of deviance explained by model = 17.3838
Adjusted percentage = 9.20837

Likelihood Ratio Tests

Factor	*Chi-Square*	*Df*	*P-Value*
Age	5.14952	1	0.0232
Socioeconomic	1.20518	2	0.5474
City Sector	10.4481	1	0.0012

Variables *Socioeconomic* and *City Sector* as *Categorical Factors*



Dependent Variable: Disease Status

(Sample Sizes:)

Quantitative Factors: Age, Lower, Middle, City Sector 1

Categorical Factors:

(Select:) Case<99

Estimated Regression Model (Maximum Likelihood)

		Standard	*Estimated*
Parameter	*Estimate*	*Error*	*Odds Ratio*
CONSTANT	-2.31293	0.642568	
Age	0.0297501	0.0135024	1.0302
Lower	-0.305255	0.604112	0.736936
Middle	0.40879	0.598995	1.505
City Sector 1	1.57475	0.50161	4.82953

Analysis of Deviance

Source	*Deviance*	*Df*	*P-Value*
Model	21.2635	4	0.0003
Residual	101.054	93	0.2667
Total (corr.)	122.318	97	

Percentage of deviance explained by model = 17.3838
Adjusted percentage = 9.20837

Likelihood Ratio Tests

Factor	*Chi-Square*	*Df*	*P-Value*
Age	5.14952	1	0.0232
Lower	0.255985	1	0.6129
Middle	0.466901	1	0.4944
City Sector 1	10.4481	1	0.0012

Analysis using dummy variables *Lower*, *Middle*, and *City Sector 1*

Since I've gone to the trouble of defining dummy variables, I'll continue with the second approach.

The first thing to note is that neither of the dummy variables for socioeconomic class are significant, so I will rely on the simpler model based on age and the city sector of residence. Key output in the Analysis Summary window is shown below.

Estimated Regression Model (Maximum Likelihood)

		Standard	*Estimated*
Parameter	*Estimate*	*Error*	*Odds Ratio*
CONSTANT	-2.33515	0.511117	
Age	0.02929	0.0131702	1.02972
City Sector 1	1.67345	0.487332	5.33054

Analysis of Deviance

Source	*Deviance*	*Df*	*P-Value*
Model	20.0583	2	0.0000
Residual	102.259	95	0.2871
Total (corr.)	122.318	97	

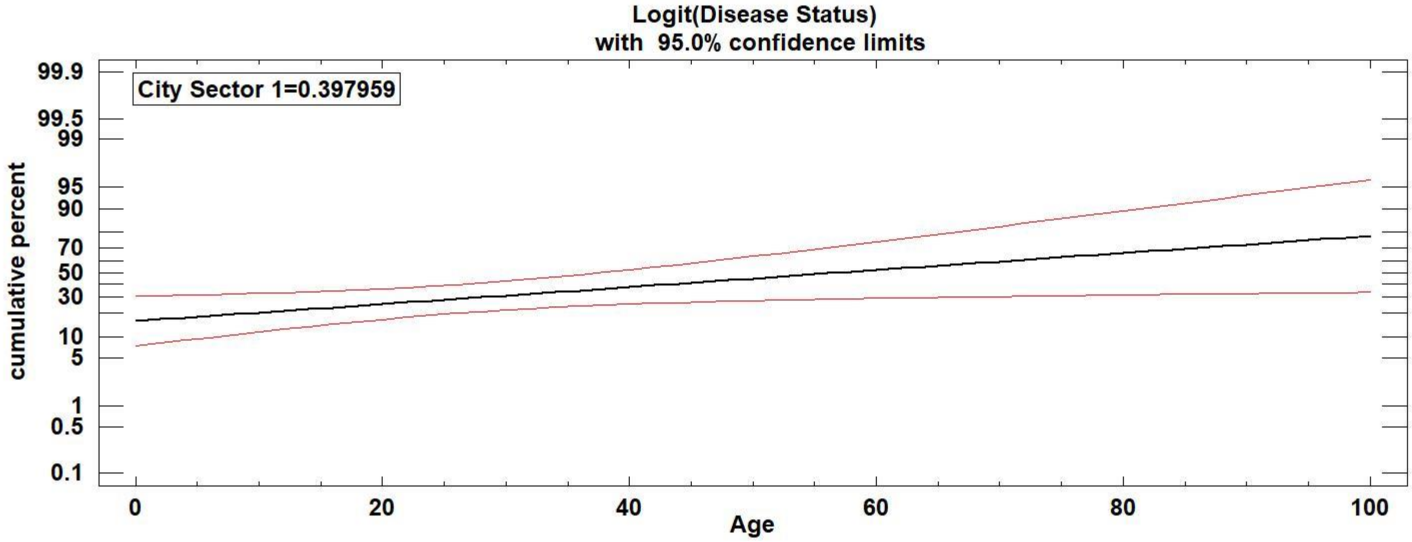
Percentage of deviance explained by model = 16.3985
Adjusted percentage = 11.4933

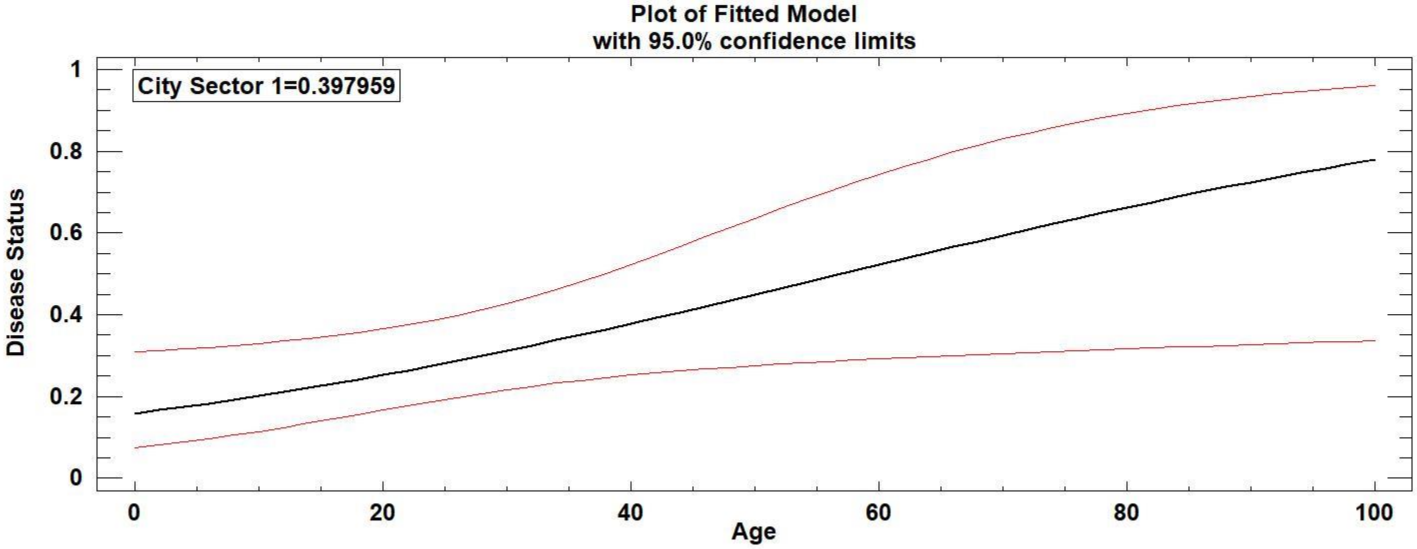
Likelihood Ratio Tests

Factor	*Chi-Square*	*Df*	*P-Value*
Age	5.27448	1	0.0216
City Sector 1	12.6533	1	0.0004

Example: In the final model, minus the socioeconomic variable, both *Age* and *City Sector 1* are significant. The *P*-value for the residual deviance is 0.2871, indicating that the model is a satisfactory fit. Only 16.4% of the deviance is explained by the model, and that drops to 11.4% after adjusting for the number of independent variables included in the model.

The Logit Function: the **Logit Function** is $ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1X_1+\beta_2X_2$, where X_1 is age, in years, and X_2 equals 1 for a resident in sector 1. The estimated logit is $\hat{\pi}=ln\left(\frac{\hat{p}}{1-\hat{p}}\right)=-2.335+0.029X_1+1.673X_2$, and the estimated probability that $\boldsymbol{Y}=1$ is $\hat{p}=\frac{e^{-2.335+0.029X_1+1.673X_2}}{1+e^{-2.335+0.029X_1+1.673X_2}}=\left(1+e^{2.335-0.029X_1-1.673X_2}\right)^{-1}$. Graphs of $\hat{\pi}$ and $\hat{p}$ are displayed by independent variable. Since a plot versus a dummy variable isn't informative, only the plots against age are reproduced. Statgraphics includes the mean of other variables. In this case, $\bar{x}_2=0.397959$ because 39 of the 98 individuals used to build the model live in Sector 1.





Example: Interpret $\hat{\beta}_1$ and $\hat{\beta}_2$. As in multiple linear regression, the estimated coefficients must be interpreted marginally. Using the column in the *Analysis Summary* window containing the estimated odds-ratios.

$\hat{\beta}_1$: Within a city sector, the odds that a person has contracted the disease increase by 3.0% per additional year of age.

$\hat{\beta}_2$: For a given age, the odds that an individual in Sector 1 has contracted the disease is more than 5 times (a whopping 533.1%) that of a person in Sector 2.

Example: The 99th person in the sample is 16 and lives in sector 2. What is the estimated infection rate for such people, i.e., the probability they have the disease? What are the 95% confidence limits for the rate?

Recall that 196 people were originally sampled, but we constructed the model from the first 98. Predictions for the remainder of the sample can be obtained in the second table in the *Predictions* window, part of which is shown below. The 99th observation corresponds to a 16-year-old from sector 2 (who didn't have the disease). The final model estimates the probability that such an individual has (symptoms of) the disease is $\hat{p} = \left(1+e^{2.335-0.029(16)-1.673(0)}\right)^{-1} = 0.134$, within the 95% confidence interval (0.067, 0.248). **Note:** The 95% confidence interval for $E(\boldsymbol{Y}) = p$ is not symmetric about the point estimate $\hat{p} = 0.134$.

Predictions for Disease Status

	Observed	*Fitted*	*Lower 95.0%*	*Upper 95.0%*
Row	*Value*	*Value*	*Conf. Limit*	*Conf. Limit*
99	**0.0**	**0.133946**	**0.067462**	**0.248491**
100	0.0	0.0906387	0.0359443	0.210396
101	0.0	0.103455	0.0449622	0.220476
102	0.0	0.175908	0.0963242	0.299455
103	0.0	0.167576	0.0910288	0.288092
104	0.0	0.140888	0.0725446	0.255856

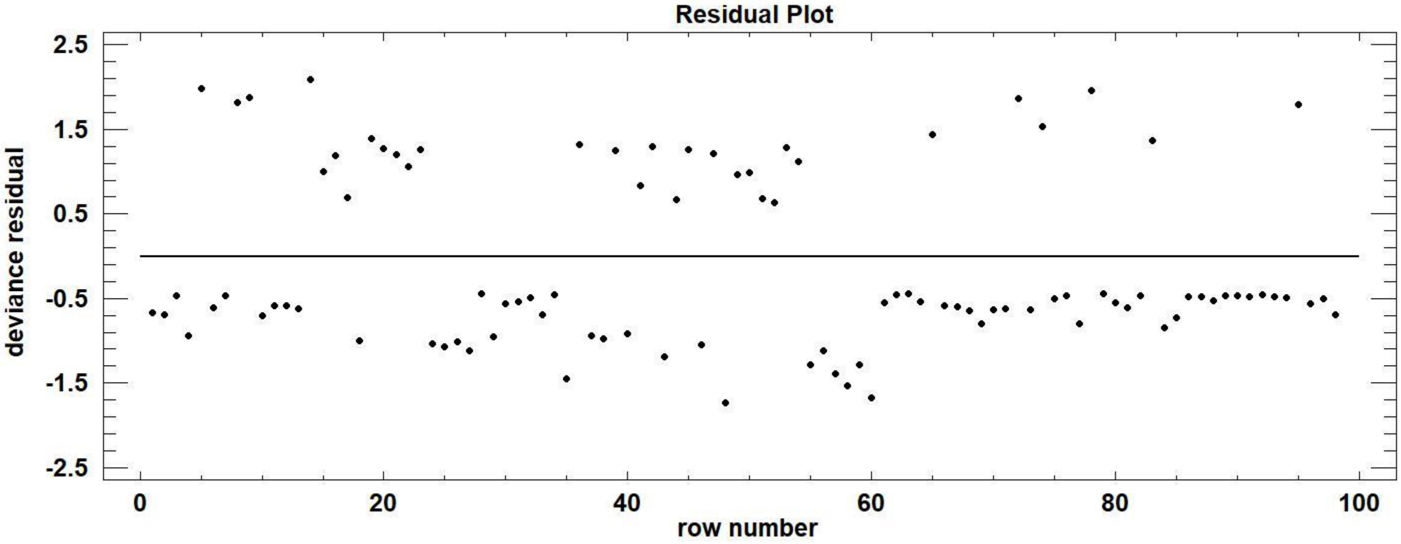
Example: Derive a rule for predicting $\boldsymbol{Y}$. The table below allows us to select a cutoff value for making predictions. Statgraphics has flagged a cutoff of $\hat{p} = 0.4$ as the best overall choice because it maximizes the percent of correct predictions over the 98 people included in the model. Of course, I'm not a doctor or health professional (I just play one in class), but I notice that a cutoff of $\hat{p} = 0.35$ does a much better job of identifying people with the disease with only a small increase in the overall error rate, so that may be a better choice. Then the decision rule would be: If $\hat{p} > 0.35$ predict the person has the disease, otherwise predict they don't. (Using this rule for the 16-year-old from sector 2 with $\hat{p} = 0.134$ correctly predicts he/she doesn't have the disease.) OK, I'll admit this seems silly. You wouldn't predict a person has the disease based on this analysis (you would consider the presence or absence of symptoms), but you might use this analysis to decide how to target testing (sound familiar). Of course, other choices of a cutoff may be reasonable.

Prediction Performance - Percent Correct

Cutoff	*TRUE*	*FALSE*	*Total*
0.0	100.00	0.00	31.63
0.05	100.00	0.00	31.63
0.1	100.00	7.46	36.73
0.15	90.32	40.30	56.12
0.2	77.42	56.72	63.27
0.25	77.42	65.67	69.39
0.3	77.42	68.66	71.43
0.35	74.19	71.64	72.45
0.4	64.52	79.10	74.49
0.45	51.61	85.07	74.49
0.5	32.26	88.06	70.41
0.55	29.03	89.55	70.41
0.6	25.81	92.54	71.43
0.65	16.13	94.03	69.39
0.7	16.13	97.01	71.43
0.75	12.90	97.01	70.41
0.8	6.45	100.00	70.41
0.85	0.00	100.00	68.37
0.9	0.00	100.00	68.37
0.95	0.00	100.00	68.37
1.0	0.00	100.00	68.37

I’m not sure that inverse predictions make sense here, especially with regards to the city sector, so I won’t investigate them.

The only diagnostic I’ll run is the plot of deviance residuals versus row number to look for conspicuous outliers. While the graph looks peculiar due to the binary nature of residuals, no residuals stand out as having unusually large (in absolute value) deviance residuals.



Final Discussion: The analysis conducted in this lecture didn’t consider interactions between the variables or polynomial terms for *Age*. While the *P*-value of 0.2871 for the residual deviance indicated that the model is an adequate fit, that doesn’t preclude the possibility that the model could be improved with further effort. I won’t expect you to try an improve on logistic models, however, as simply understanding the role they play in statistical modeling and their use and interpretation is sufficient for an introductory course such as this.