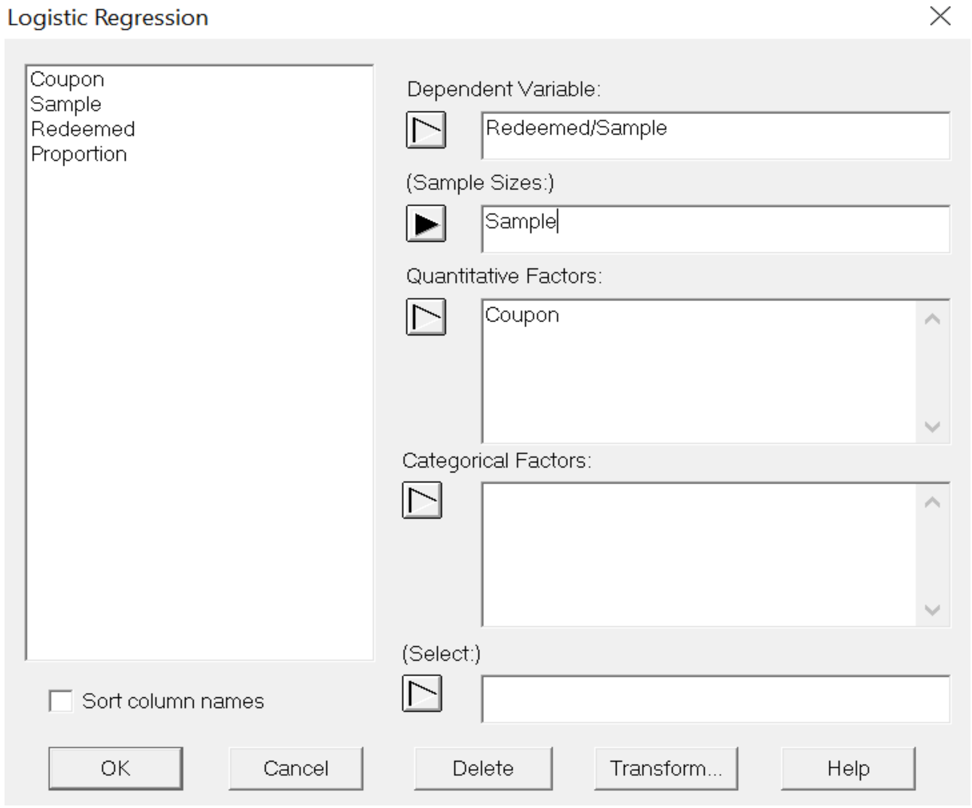
Logistic Regression with Repeated Observations

In an experimental setting, and in some applications, multiple observations may be conducted at only a handful of values of *X*. For example, Covid-19 patients in an ICU may be given an experimental antiviral drug at one of several doses. In an example such as this, a logistics model is run on the proportions of successes observed at each *X*, $\tilde{p}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$, where n_i observations are made at X_i, and $Y_{ij} = \begin{cases} 1 \text{ if } j^{\text{th}} \text{observation at } X_i \text{ is a success} \\ 0 \text{ if } j^{\text{th}} \text{observation at } X_i \text{ is a failure} \end{cases}$.

Example: In a study of the effectiveness of discount coupons, coupons offering a discount on a common household item were mailed to 1,000 households. 200 households were selected to receive coupons for $5, while similar numbers received discounts for $10, $15, $20, and $30. The response variable ***Y*** recorded whether a household used the coupon to purchase the product. The results appear in the file *Coupon Redemption*.

Coupon	Sample	Redeemed	Proportion
Dollars			
5	200	30	0.15
10	200	55	0.275
15	200	70	0.35
20	200	100	0.5
30	200	137	0.685

In Statgraphics, after selecting the menus *Relate* → *Attribute Data* → *Logistics Regression*, the data is entered as shown below. (Since the proportions have already been computed in the data spreadsheet, they can be entered directly in the *Dependent Variable* field, but you must still let Statgraphics know the sample sizes by completing the *Sample Sizes* field.)



Logistic Regression

Dependent Variable: Redeemed/Sample

(Sample Sizes:) Sample

Quantitative Factors: Coupon

Categorical Factors:

(Select:)

□ Sort column names

The output below is then interpreted similarly to the earlier example, with a few adaptations.

Analysis Summary window: According to the output below, the redemption amount of a coupon is significant, i.e., is related to the probability that a coupon will be redeemed for the product. The residual *P*-value indicates that a model with coupon value as the only predictor cannot be significantly improved over the current model. The percentage of deviance explained by the model is quite large at 98.55%, which is obviously good. We need to be careful, however, about comparing the percentage of deviance explained in this model to the percentage explained in the logistic model from the previous lecture. Repeating observations on a few levels of a predictor has an averaging effect on the resulting proportions. Assuming a logistics model is appropriate, as suggested by a scatterplot of the proportions and the model and residual *P*-values, the logistic curve fitted to the proportions is likely to fit them much better than a logistics curve attempting to fit data of 0 – 1 outcomes. The percentage of deviance explained will therefore tend to be greater for the aggregated data.

Estimated Regression Model (Maximum Likelihood)

		Standard	*Estimated*
Parameter	*Estimate*	*Error*	*Odds Ratio*
CONSTANT	-2.04435	0.160976	
Coupon	0.0968336	0.00854912	1.10168

Analysis of Deviance

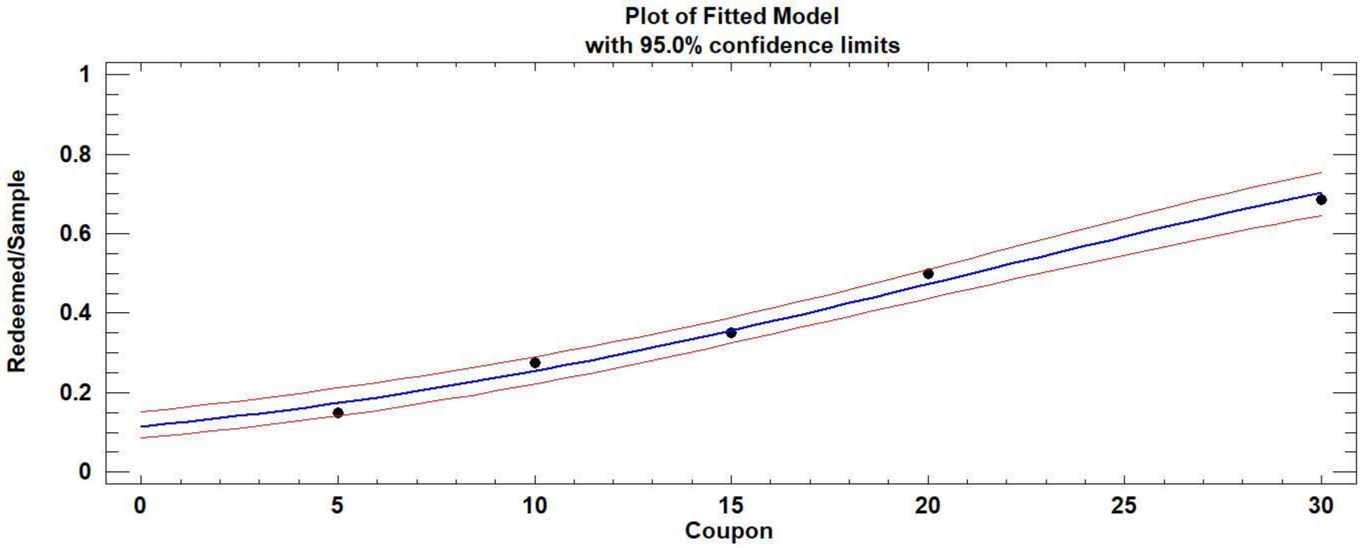
Source	*Deviance*	*Df*	*P-Value*
Model	147.296	1	0.0000
Residual	2.16682	3	0.5385
Total (corr.)	149.463	4	

Likelihood Ratio Tests

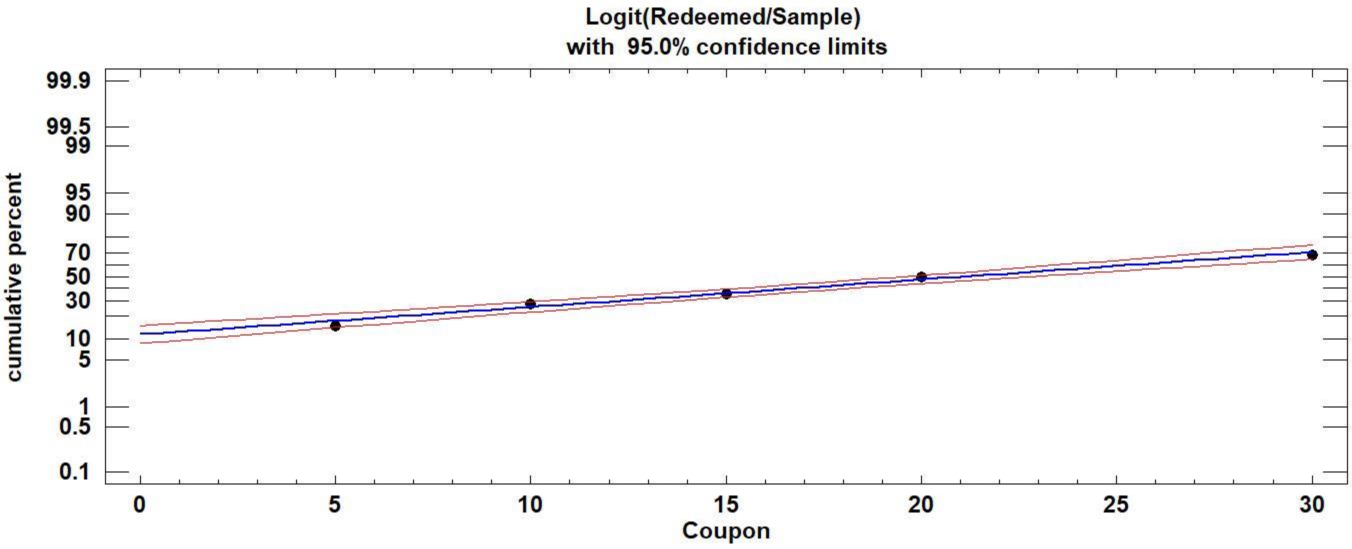
Factor	*Chi-Square*	*Df*	*P-Value*
Coupon	147.296	1	0.0000

Percentage of deviance explained by model = 98.5503
Adjusted percentage = 95.874

Plot of Fitted Logistics Model, showing 95% confidence bands:



Plot of the estimated Logit Function $\hat{\pi} = ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 X$, showing 95% confidence bands:



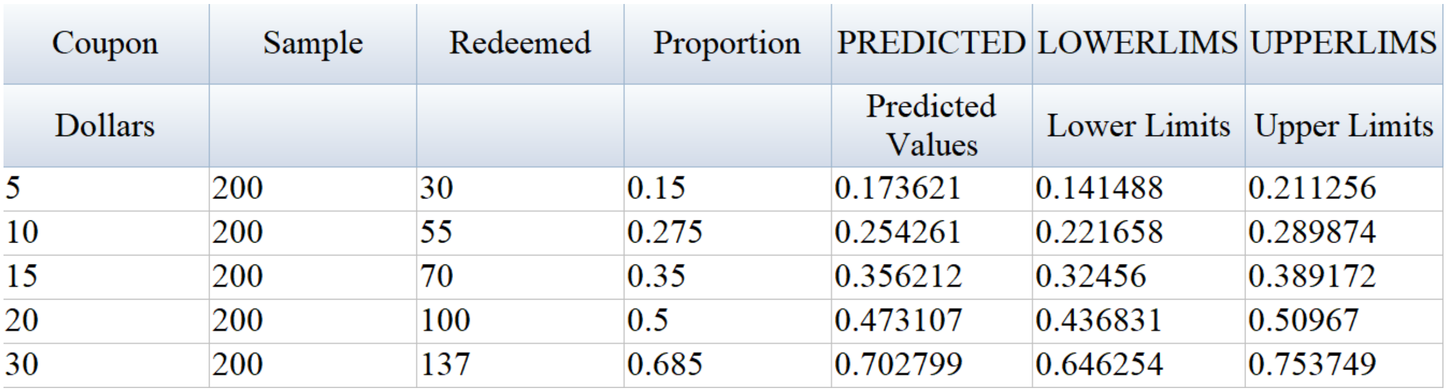
Example: Write out the equation of estimated logistic regression curve, and use it to estimate the probability that a $10 coupon will be redeemed for the product.

$\hat{p} = \frac{e^{-2.04435+0.0968336x}}{1+e^{-2.04435+0.0968336x}}$. For a $10 coupon, $\hat{p} = \frac{e^{-2.04435+0.0968336\times 10}}{1+e^{-2.04435+0.0968336\times 10}} = 0.25426$. So, we expect about one quarter of $10 coupons to be redeemed for the product.

Example: Interpret $\hat{\beta}_1$. The estimated coefficient of the predictor *Coupon* is $\hat{\beta}_1 = 0.0968336$, so the estimated odds-ratio is $e^{\hat{\beta}_1} = e^{0.0968336} \simeq 1.10168$ (which I could simply have looked up in the *Estimated Regression Model* table in the *Analysis Summary* window). Thus, the model predicts that the odds a coupon is redeemed increase by about 10.2% for each additional $1 in coupon value.

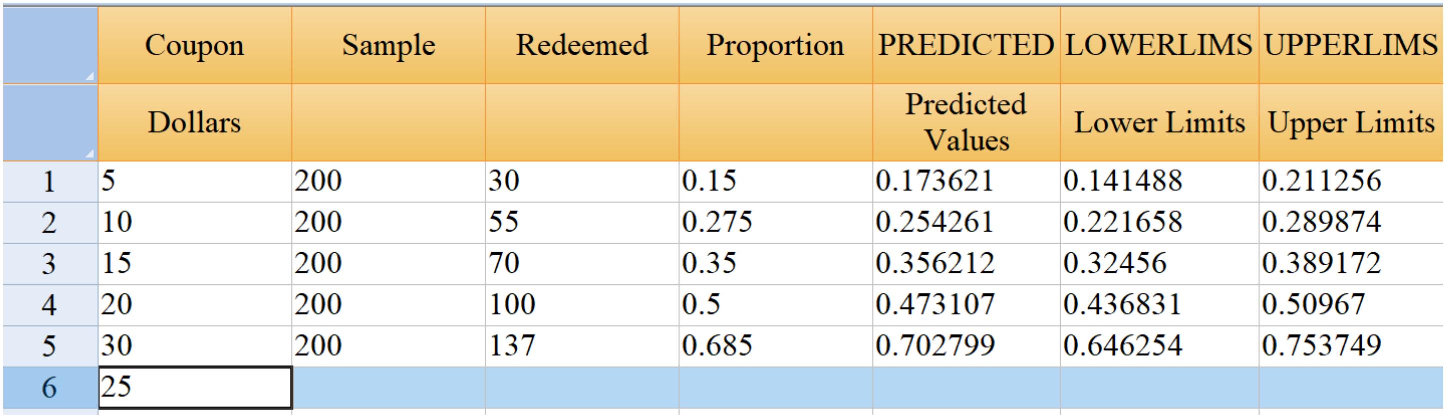
Example: For the coupon redemption data, save the confidence intervals for the proportion of coupons redeemed for different coupon values, and predict the redemption rate for coupons worth $25.

Confidence Intervals for Proportions (Expected ***Y***): *Save Results* was used to save the predicted proportions of coupons redeemed, as well as the lower and upper limits of 95% confidence intervals for the proportions, for the different coupon redemption amounts evaluated (other levels of confidence are available through *Pane Options*).



Coupon	Sample	Redeemed	Proportion	PREDICTED	LOWERLIMS	UPPERLIMS
Dollars				Predicted Values	Lower Limits	Upper Limits
5	200	30	0.15	0.173621	0.141488	0.211256
10	200	55	0.275	0.254261	0.221658	0.289874
15	200	70	0.35	0.356212	0.32456	0.389172
20	200	100	0.5	0.473107	0.436831	0.50967
30	200	137	0.685	0.702799	0.646254	0.753749

To predict the proportion of coupons redeemed for a discount value of $25, I entered **25** in the next available row in the datasheet while leaving other columns blank. This is similar to the procedure for fitting values and constructing confidence intervals in multiple regression.



	Coupon	Sample	Redeemed	Proportion	PREDICTED	LOWERLIMS	UPPERLIMS
	Dollars				Predicted Values	Lower Limits	Upper Limits
1	5	200	30	0.15	0.173621	0.141488	0.211256
2	10	200	55	0.275	0.254261	0.221658	0.289874
3	15	200	70	0.35	0.356212	0.32456	0.389172
4	20	200	100	0.5	0.473107	0.436831	0.50967
5	30	200	137	0.685	0.702799	0.646254	0.753749
6	25						

At the bottom of the *Predictions* window in Statgraphics we find the following fitted probability that a $25 coupon is redeemed for the product, as well as a confidence interval for the probability. *Pane Options* can be used to obtain confidence intervals at other levels of confidence.

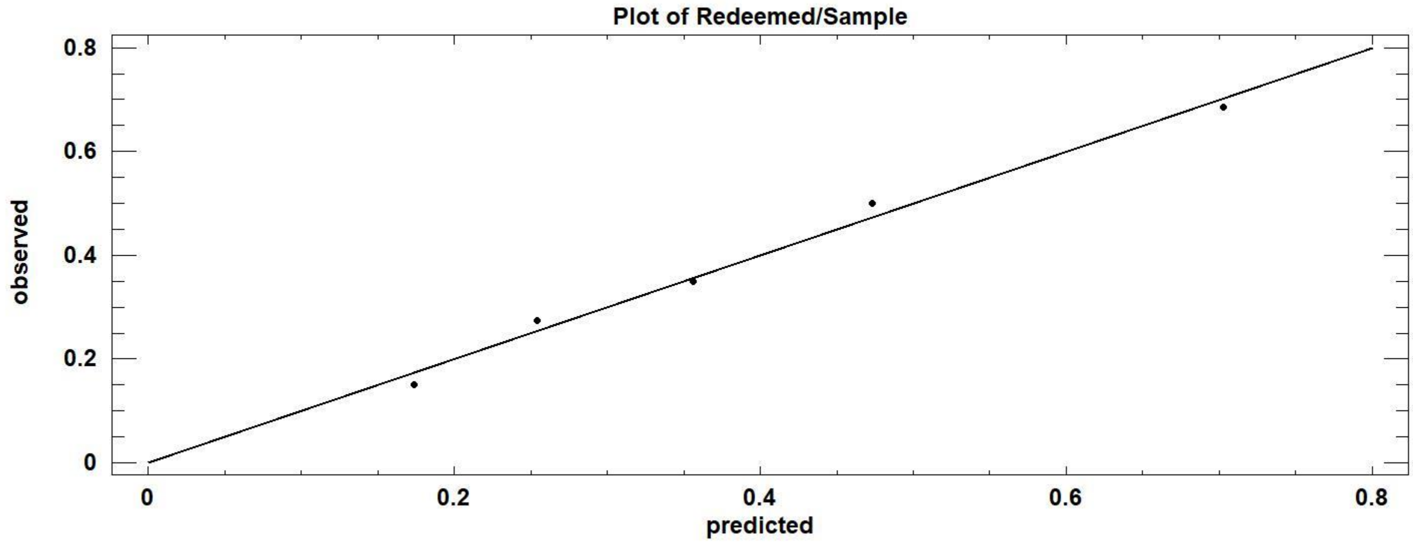
Predictions for Proportion

	Observed	*Fitted*	*Lower 95.0%*	*Upper 95.0%*
Row	*Value*	*Value*	*Conf. Limit*	*Conf. Limit*
6		0.593027	0.545914	0.638488

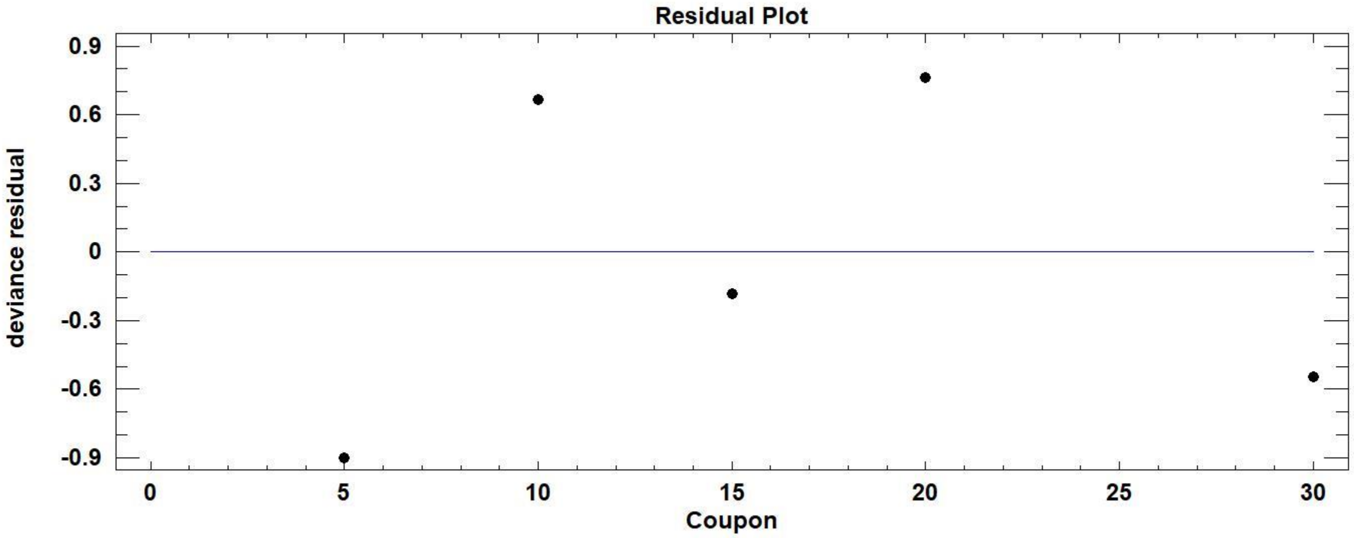
Diagnostics

We will not be conducting many diagnostics in logistic regression, but below I've run a few that are available in Statgraphics. I won't require that you run them as part of a logistic regression analysis in this course, but I discuss them briefly for the sake of completeness, and because they are discussed in the Statgraphics pdf for logistic regression that you may have downloaded.

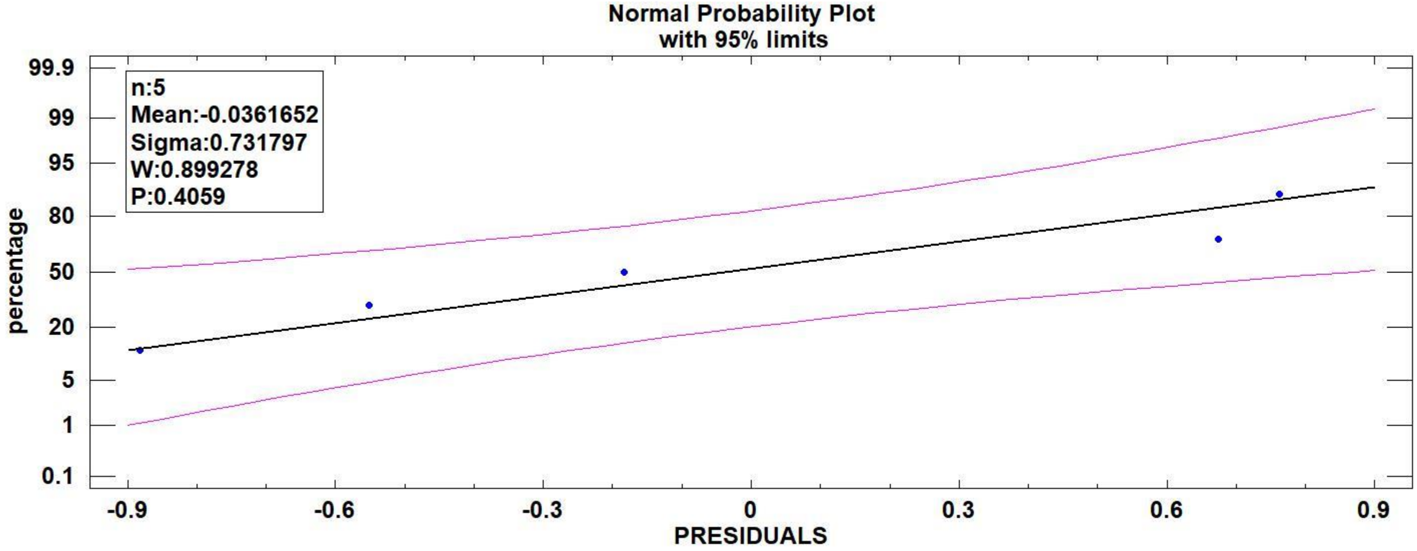
If the model fits well, we expect the points on the plot of observed versus predicted proportions to be randomly scattered about the diagonal line, as they appear to be in the plot below.



Plot of Residuals:



Normal Probability Plot: The normal probability plot of Pearson residuals, Shapiro-Wilk test, and standardized skewness and kurtosis scores indicate that an assumption of normally distributed residuals is reasonable. Although errors in logistic regression are not normally distributed, where repeated observations involving large samples are involved, the sample (binomial) proportions may be approximaely normal by an application of the central limit theorem. Then the i^{th} sample proportion may be at least approximately distributed as $\tilde{p}_i \sim N\left(p_i, \frac{p_i(1-p_i)}{n_i}\right)$. The Pearson residuals r are standarized, $r_i = \frac{e_i}{\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n_i}}}$, where $e_i = \tilde{p}_i - \hat{p}_i$ equals the observed minus the predicted proportion at X_i. For large sample sizes, n_i , Pearson residuals may be approximately standard normal. In the coupon redemption example, the sample sizes of 200 for each treatment (coupon redemption value) are relatively large, and a normal probability plot of the residuals may be used to evaluate wether the Pearson residuals are plausibly derived from conditional normal distributions.



The *P*-value for the Shapiro-Wilkes tests supports the assumption of normally distributed error

Stnd. skewness	0.117508
Stnd. kurtosis	-1.18215

The StatAdvisor: The standardized skewness and standardized kurtosis are within the range expected for data from a normal distribution.

All of which begs the question, do we care that the residuals of sample proportions are normally distributed, i.e., why does Statgraphics produce a normal probability plot of Pearson residuals. I suspect that we don't much care if we are using maximum likelihood estimation to fit the logistic regression function and create confidence intervals for $p = E\left(\boldsymbol{Y}\right)$, since the key assumptions of MLE aren't the normality or constant variance of the errors, but their independence. However, if we are using weighted least squares (WLS) to fit the model and create the intervals, then inference such as the computed *P*-values of hypothesis tests and the creation of confidence intervals is conducted on the assumption of normality. In the coupon redemption example there are only five proportions, so it is important the verify that the assumption that sample proportions are normally distributed is reasonable. (WLS reweights the sample proportions to compensate for their unequal variance, but inference still requires normally distributed residuals.)