In a seed germination test, seeds of two cultivars were planted in pots of two soil conditions. The following statements create the data set *seeds*, which contains the observed proportion of seeds that germinated for various combinations of cultivar and soil condition. The variable *n* represents the number of seeds planted in a pot, and the variable *r* represents the number germinated. The indicator variables *cult* and *soil* represent the cultivar and soil condition, respectively.

```
data seeds;
   input pot n r cult soil;
   datalines;
 1 16     8      0       0
 2 51    26      0       0
 3 45    23      0       0
 4 39    10      0       0
 5 36     9      0       0
 6 81    23      1       0
 7 30    10      1       0
 8 39    17      1       0
 9 28     8      1       0
10 62    23      1       0
11 51    32      0       1
12 72    55      0       1
13 41    22      0       1
14 12     3      0       1
15 13    10      0       1
16 79    46      1       1
17 30    15      1       1
18 51    32      1       1
19 74    53      1       1
20 56    12      1       1
;
```

First, we decide that logistic regression is a sensible model for this data as it is of the form "success" and "failure".

Looking at the data, there are many observations at each unique value of the X. Therefore, the individual success and failure trials get aggregated into number of successes out of a total number of trials. This is binomial distribution:

If for a fixed X, $(Y_i|X_i = X) \sim Bernoulli(\mu(X))$, then we can sum the total number of successes. If the trials are independent with the same probability, then:

$\sum(Y_i|X_i = X) \sim Binomial(n_i, \mu(X))$ where n_i is the total number of trials at that value of X

1. First, we can check whether the model fits with a goodness-of-fit test, which can be run as we have multiple observations at each value of X. This can be produced via "scale = none":

```
proc logistic data=seeds;
   model r/n=cult soil/scale=none;
run;
```

What value for the Pearson Chi squared do you get? What about the deviance (G^2)? What do you conclude?

2. There are two major reasons for model misfit with GLMs:

<u>The systematic component is incorrect:</u> This means that the transformation or inclusion of explanatory variables is required

<u>The random component is incorrect:</u> This can mean that the distribution isn't appropriate for the data, the underlying independence assumption is violated, or the distribution is generally appropriate but it isn't flexible enough

It can be difficult to separate these potential causes of model misfit. A general approach is as follows:

a. Consider whether the distribution for the random component is appropriate. For this problem, does using a binomial make sense? What kind of data sets would the binomial model be an appropriate model? Give an example of a data set in which the binomial model would not be appropriate.

b. For the "flexible enough" issue, we need to discuss overdispersion: Comparing multiple linear regression (MLR) to logistic regression, there is an important difference. Writing $\mathbb{E}[Y|X] = \mu(X)$
   - $Y|X \sim N(\mu(X), \sigma^2)$ and $\mu(X) = X^T\beta$
   - $Y|X \sim Binomial(n_i, \mu(X))$ and $\log\left(\frac{\mu(X)}{1-\mu(X)}\right) = X^T\beta$. The variance is $n_i\mu(X)\big(1-\mu(X)\big)$

Hence, there is a separate variance parameter in MLR but not in Logistic regression. This can mean that we might need to generalize the logistic regression model to allow for a separate variance, known as an overdispersion parameter.
To check for overdispersion, fit a model with a complex systematic component and look at the goodness-of-fit test. The idea here is that if we can rule out model misfit due to the systematic component being incorrect (and that the random component is reasonable as in a.), then we can reasonably conclude the misfit is coming from overdispersion. In this case, we could consider the interactive model, but in general you could include interactions and low order polynomial terms. What is the deviance goodness of fit test statistic and p-value for this model?

c.

d. For the "flexible enough" issue, we need to discuss overdispersion: Comparing multiple linear regression (MLR) to logistic regression, there is an important difference. Writing $\mathbb{E}[Y|X] = \mu(X)$
   - $Y|X \sim N(\mu(X), \sigma^2)$ and $\mu(X) = X^T\beta$
   - $Y|X \sim Binomial(n_i, \mu(X))$ and $\log\left(\frac{\mu(X)}{1-\mu(X)}\right) = X^T\beta$. The variance is $n_i\mu(X)\big(1-\mu(X)\big)$

   Hence, there is a separate variance parameter in MLR but not in Logistic regression. This can mean that we might need to generalize the logistic regression model to allow for a separate variance, known as an overdispersion parameter.
   To check for overdispersion, fit a model with a complex systematic component and look at the goodness-of-fit test. The idea here is that if we can rule out model misfit due to the systematic component being incorrect (and that the random component is reasonable as in a.), then we can reasonably conclude that the misfit is coming from overdispersion. In this case, we could consider the interactive model, but in general you could include interactions and low order polynomial terms. What is the deviance goodness of fit test statistic and p-value for this model?

3. The typical adjustment for overdispersion is to create a new parameter ψ and write the variance as $\psi^2 n_i\mu(X)\big(1-\mu(X)\big)$. ψ can be conveniently estimated as the square root of the deviance. Write down the value of the square root of the deviance from the previous problem (round it to 6 or so decimals). Now, fit the additive model, but with scale=**the square root of the deviance**. Note, no longer look at the goodness of fit test for this model nor the residuals (as these are computed from the goodness of fit test). Write a statistical conclusion for interpreting the estimated systematic component in this model. How might your approach be different if you are interested in making predictions with this model versus inference?

Vasoconstriction: Let's look at a separate data set. In a controlled experiment to study the effect of the rate and volume of air intake on a transient reflex vasoconstriction in the skin of the digits, 39 tests under various combinations of rate and volume of air intake were obtained. The researcher also records whether vasoconstriction occurred.

We can make similar statements about model fit and influential observations for GLMs as we did for linear models. Fit the model for the systematic component of being additive in log volume and log rate. Do there appear to be any observations that have outsized influence and/or do not seem to be fit by the model? If so, which one(s)?

We will need to delete these observations as we cannot investigate whether they were incorrectly recorded. Rerun the analysis without the identified observation(s) and write up an appropriate statistical conclusion.