Why can’t we use regular linear regression when the outcome variable is binary?

Example: \( Y_i = \begin{cases} 
1 & \text{favor school closing} \\
0 & \text{oppose school closing} 
\end{cases} \\
X_i = \text{no. of years lived in town} \\
\text{OLS model} \\
\hat{Y}_i = 5.94 - .008X_i \\
gives negative predicted values if \( X_i \) is large enough \\
but \( \hat{Y}_i \) has to be between 0 and 1 because it is the mean of a bunch of 0’s and 1’s

- impossible predicted values
- violations of assumptions

### Odds ratios

odds ratio - the ratio of odds in 2 different groups

Example:
Suppose \( \text{Pr(heart attack in next 12 mos.)} \) is .01 for male smokers

Then odds ratio for heart attacks in next 12 mos. in nonsmoker vs. smoker males is

\[
O.R. = \frac{\frac{\text{Pr(heart attack|non-smoke)}}{1-\text{Pr(heart attack|non-smoke)}}}{\frac{\text{Pr(heart attack|smoke)}}{1-\text{Pr(heart attack|smoke)}}}
\]

\[
= \frac{.005}{.995} = 0.497
\]

### Interpretation of odds ratio

- \( O.R. \geq 0 \)
- if \( O.R. = 1.0 \), then \( \text{Pr}(Y = 1) \) is the same in both samples
- if \( O.R. < 1.0 \), then \( \text{Pr}(Y = 1) \) is less in numerator group than in denominator group
- \( O.R. = 0 \) if and only if \( \text{Pr}(Y = 1) = 0 \) in numerator sample
Logistic regression

Response variable is log odds of \( Y = 1 \)

\[
L_i = \log_e \left( \frac{\Pr(Y_i = 1)}{1 - \Pr(Y_i = 1)} \right)
\]

\[
L_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{K-1} X_{i,K-1} ;
\]

linear!

difference from linear regression with transformations:

We give SAS the untransformed binary \( Y_i \) values and it does the transforming for us.

Converting logistic regression coefficients to odds ratios

O.R. for increase of 1 unit in \( X_j = e^{\beta_j} \)

Example:

O.R. for 1 year increase in age

\[
\frac{Pr(Y=1|age=X+1)}{1-Pr(Y=1|age=X)} = \frac{Pr(Y=1|age=X+1)}{1-Pr(Y=1|age=X)} = e^{0.0971} = 1.102
\]
Converting logistic linear predictors to probabilities

\[ L_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki} \]

\[ P(Y_i = 1) = \frac{1}{1 + e^{-L_i}} \]

For example, for a 33-year old woman

\[ L_i = -3.4377 + 0.0971(33) = -0.2334 \]

\[ P(Y_i = 1) = \frac{1}{1 + e^{0.2334}} = 0.442 \]
Assumptions of logistic regression

- linearity in the logit
  - doesn’t need to be checked for dummy-variable predictors
- independence
- no perfect multicollinearity
- errors have logistic distribution
  - isn’t checked directly

Hypothesis testing

- overall \( \chi^2 \) test
  - corresponds to overall \( F \) test in linear regression
  - tests whether all the variables in the model taken together are useful in predicting \( Y \)

- \( \chi^2 \) test for individual predictors
  - corresponds to \( t \) test in linear regression
  - tests for significance of individual predictor variable after controlling for other predictors in the model

- \( \chi^2 \) test for a set of predictors

```
The LOGISTIC Procedure

Data Set: WING.INST
Response Variable: RD
Response Levels: 2
Number of Observations: 500
Link Function: Logit

Response Profile
Ordered Value   1   2   Count
           1       2     50
           2       3     50

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion          Intercept  Only Intercept and Covariates
                     140.623  31.728   34.517
SC                  143.235  44.751
-2 LOG L            138.623  21.725  116.904 with 4 DF (p=0.0001)
Score               .        .       74.208 with 4 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

```

```
\[ \chi^2 \] test for a set of predictors

- corresponds to partial \( F \) test in linear regression \( v \)

\[ \chi^2_{df} = -2 [\text{log likelihood}_\text{full model} - \text{log likelihood}_\text{reduced model}] \]

- degrees of freedom = \( H \)
  - how many more predictors there are in the full model than in the reduced model

To test whether the variables
- \text{sepwidth}
- \text{seplen}

are significant after
- \text{petwidth}
- \text{petwidth} are already in the model

\[ \chi^2 = 26.887 - 21.725 = 5.162 \]

Compare to \( \chi^2_{0.05} = 5.99 \)

5.162 < 5.99, so the set of variables is not significant at the .05 level.

Statistical problems in logistic regression
- multicollinearity
• poor model fit
• influential observations

High *discrimination* or *separation*

• what it is
  - all (or almost all) of the observations with a particular value of a predictor variable have the same value of the response variable

• what it does
  - coefficients cannot be estimated at all if discrimination is perfect
  - inflates standard errors of coefficients
Hosmer-Lemeshow $\chi^2$-test for model fit

- groups observations in dataset by quantiles of predicted probability of "yes" response
- within each group, compares the number of predicted positive responses to the number of observed positive responses
- should use at least 6 groups
- small p-value indicates poor fit of model

Evaluating the model’s predictive ability

- sensitivity
- specificity
- positive predictive value
- negative predictive value
Using the `c`table option

```plaintext
proc logistic descending;
model hidel = age newburb termite / lackfit ctable;
output out = Ifits predicted = pred lower=LCL upper=UCL;
run;
```

The LOGISTIC Procedure

Classification Table

<table>
<thead>
<tr>
<th>Prob Level</th>
<th>Correct Event</th>
<th>Incorrect Event</th>
<th>Sensi- tivity</th>
<th>Speci- city</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>False POS</td>
</tr>
<tr>
<td>0.000</td>
<td>16</td>
<td>27</td>
<td>0</td>
<td>37.2</td>
</tr>
<tr>
<td>0.020</td>
<td>16</td>
<td>23</td>
<td>0</td>
<td>41.9</td>
</tr>
<tr>
<td>0.040</td>
<td>15</td>
<td>21</td>
<td>1</td>
<td>48.8</td>
</tr>
<tr>
<td>0.060</td>
<td>15</td>
<td>19</td>
<td>1</td>
<td>53.5</td>
</tr>
<tr>
<td>0.080</td>
<td>15</td>
<td>17</td>
<td>1</td>
<td>58.1</td>
</tr>
<tr>
<td>0.100</td>
<td>15</td>
<td>16</td>
<td>1</td>
<td>60.5</td>
</tr>
<tr>
<td>0.120</td>
<td>15</td>
<td>16</td>
<td>1</td>
<td>62.8</td>
</tr>
<tr>
<td>0.140</td>
<td>14</td>
<td>14</td>
<td>2</td>
<td>62.8</td>
</tr>
<tr>
<td>0.160</td>
<td>14</td>
<td>14</td>
<td>2</td>
<td>62.8</td>
</tr>
<tr>
<td>0.180</td>
<td>14</td>
<td>14</td>
<td>2</td>
<td>66.1</td>
</tr>
<tr>
<td>0.200</td>
<td>14</td>
<td>14</td>
<td>2</td>
<td>65.1</td>
</tr>
<tr>
<td>0.220</td>
<td>14</td>
<td>14</td>
<td>2</td>
<td>65.1</td>
</tr>
<tr>
<td>0.240</td>
<td>14</td>
<td>14</td>
<td>2</td>
<td>65.1</td>
</tr>
<tr>
<td>0.260</td>
<td>14</td>
<td>14</td>
<td>2</td>
<td>65.1</td>
</tr>
<tr>
<td>0.280</td>
<td>14</td>
<td>15</td>
<td>2</td>
<td>67.4</td>
</tr>
<tr>
<td>0.300</td>
<td>14</td>
<td>15</td>
<td>2</td>
<td>67.4</td>
</tr>
<tr>
<td>0.320</td>
<td>13</td>
<td>16</td>
<td>3</td>
<td>67.4</td>
</tr>
<tr>
<td>0.340</td>
<td>11</td>
<td>16</td>
<td>5</td>
<td>62.8</td>
</tr>
<tr>
<td>0.360</td>
<td>11</td>
<td>18</td>
<td>5</td>
<td>67.4</td>
</tr>
<tr>
<td>0.380</td>
<td>11</td>
<td>18</td>
<td>5</td>
<td>67.4</td>
</tr>
<tr>
<td>0.400</td>
<td>11</td>
<td>20</td>
<td>5</td>
<td>72.1</td>
</tr>
<tr>
<td>0.420</td>
<td>10</td>
<td>20</td>
<td>7</td>
<td>69.8</td>
</tr>
<tr>
<td>0.440</td>
<td>10</td>
<td>20</td>
<td>7</td>
<td>69.8</td>
</tr>
<tr>
<td>0.460</td>
<td>10</td>
<td>21</td>
<td>6</td>
<td>72.1</td>
</tr>
</tbody>
</table>

| 0.480      | 10            | 22              | 5            | 74.4      | 62.5      | 81.5     | 33.3    | 21.4     |
| 0.500      | 10            | 22              | 5            | 74.4      | 62.5      | 81.5     | 33.3    | 21.4     |
| 0.520      | 10            | 22              | 5            | 74.4      | 62.5      | 81.5     | 33.3    | 21.4     |
| 0.540      | 8             | 22              | 5            | 69.8      | 50.0      | 81.5     | 38.5    | 26.7     |
| 0.560      | 8             | 23              | 4            | 72.1      | 50.0      | 85.2     | 33.3    | 26.8     |
| 0.580      | 8             | 23              | 4            | 72.1      | 50.0      | 85.2     | 33.3    | 26.8     |
| 0.600      | 7             | 23              | 4            | 69.8      | 43.8      | 85.2     | 36.4    | 28.1     |
| 0.620      | 7             | 24              | 3            | 72.1      | 43.8      | 88.9     | 30.0    | 27.3     |
| 0.640      | 5             | 24              | 3            | 72.1      | 31.3      | 88.9     | 37.5    | 31.4     |
| 0.660      | 5             | 24              | 3            | 72.1      | 31.3      | 88.9     | 37.5    | 31.4     |
| 0.680      | 5             | 24              | 3            | 72.1      | 31.3      | 88.9     | 37.5    | 31.4     |
| 0.700      | 5             | 24              | 3            | 72.1      | 31.3      | 88.9     | 37.5    | 31.4     |
| 0.720      | 4             | 24              | 3            | 72.1      | 25.0      | 88.9     | 42.9    | 33.3     |
| 0.740      | 4             | 24              | 3            | 72.1      | 25.0      | 88.9     | 42.9    | 33.3     |
| 0.760      | 4             | 24              | 3            | 72.1      | 25.0      | 88.9     | 42.9    | 33.3     |
| 0.780      | 1             | 24              | 3            | 72.1      | 6.3       | 88.9     | 75.0    | 38.5     |
| 0.800      | 1             | 24              | 3            | 72.1      | 6.3       | 88.9     | 75.0    | 38.5     |
| 0.820      | 1             | 24              | 3            | 72.1      | 6.3       | 88.9     | 75.0    | 38.5     |
| 0.840      | 1             | 24              | 3            | 72.1      | 6.3       | 88.9     | 75.0    | 38.5     |
| 0.860      | 0             | 26              | 1            | 60.5      | 0.0       | 96.3     | 100.0   | 38.1     |
| 0.880      | 0             | 26              | 1            | 60.5      | 0.0       | 96.3     | 100.0   | 38.1     |
| 0.900      | 0             | 26              | 1            | 60.5      | 0.0       | 96.3     | 100.0   | 38.1     |
| 0.920      | 0             | 27              | 0            | 62.8      | 0.0       | 100.0    | 37.2    | .        |