22s:152 Applied Linear Regression

Ch. 14 (sec. 1) and Ch. 15 (sec. 1 & 4): Logistic Regression

Logistic Regression

- When the response variable is a binary variable, such as
  - 0 or 1
  - live or die
  - fail or succeed

  then we approach our modeling a little differently

- What is wrong with our previous modeling?
  - Let's look at an example...

- **Example**: Study on lead levels in children

  Binary response called *highbld*:

  > Either high lead blood level(1) or
  > low lead blood level(0)

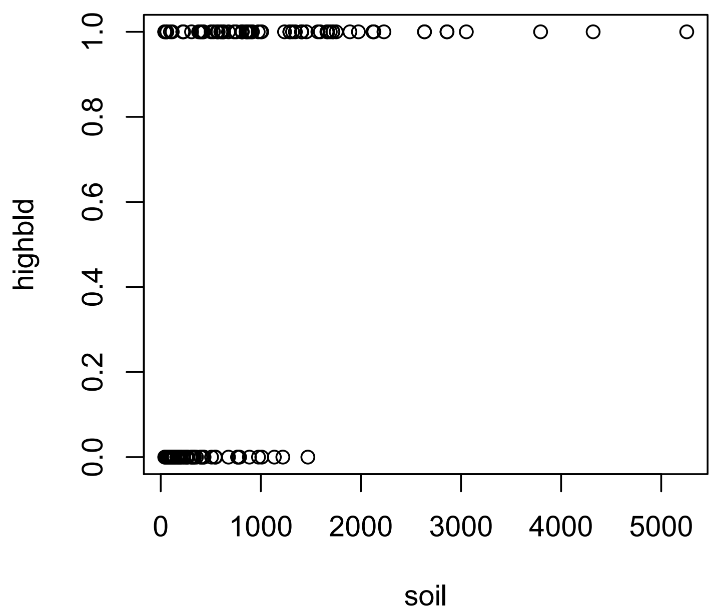
  One continuous predictor variable called *soil*:

  > Measure of the level of lead in the soil in the subject's backyard

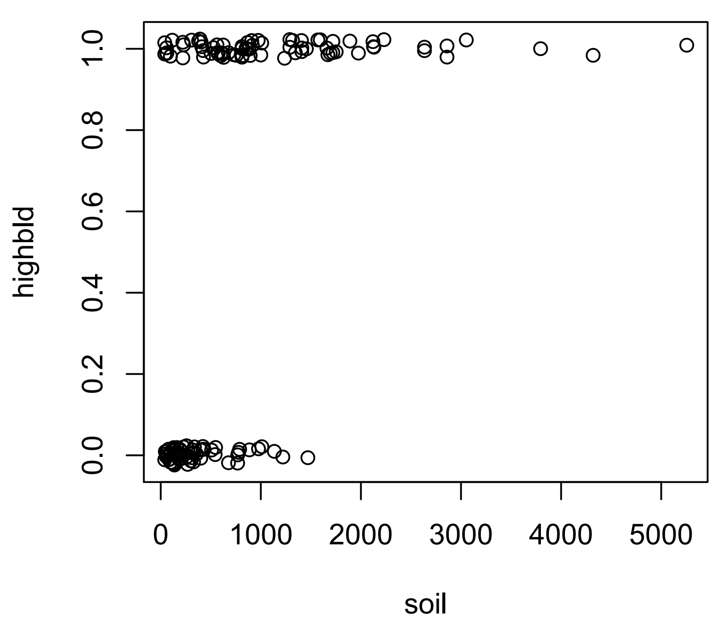
```
> sl=read.csv("soillead.csv")
> attach(sl)
> head(sl)

  highbld soil
1       1 1290
2       0   90
3       1  894
4       0  193
5       1 1410
6       1  410
```

The dependent variable plotted against the independent variable:



Original



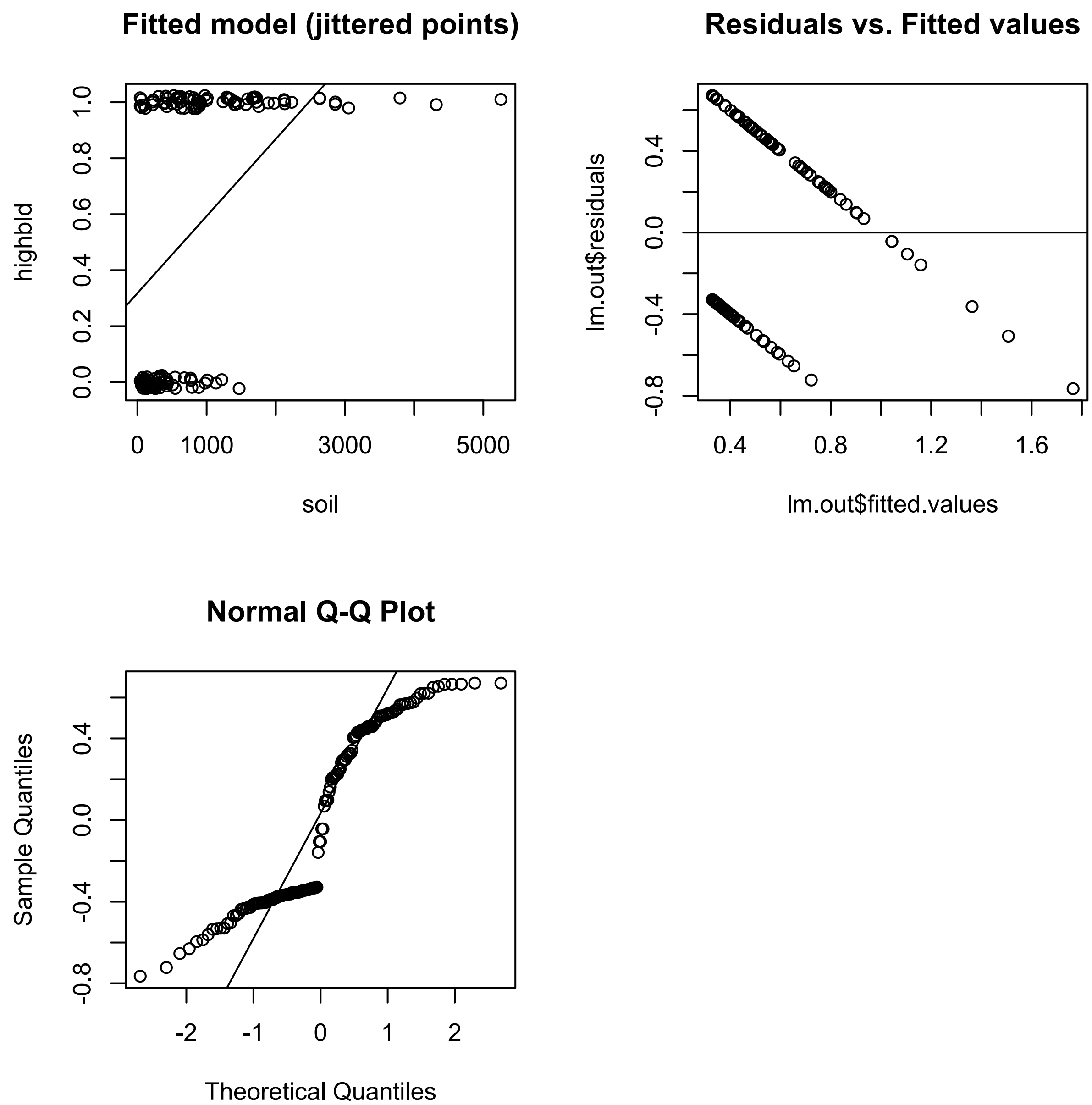
Jittered

Consider the usual regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

The fitted model and residuals:

$$\hat{Y} = 0.3178 + 0.0003x$$



All sorts of violations here...

Violations:

- Our usual $\epsilon_i \sim N(0, \sigma^2)$ isn't reasonable

  The errors are not normal, and they don't have the same variability across all x-values.

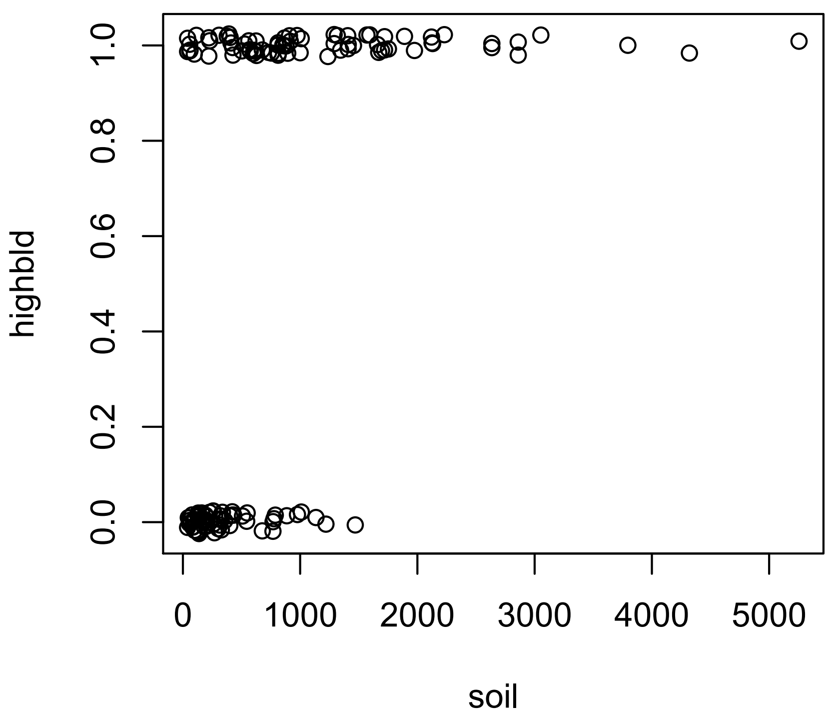
- predictions for observations can be outside [0,1].

  For x=3000, $\hat{Y} = 1.14$, and this is not possibly an average of the 0s and 1s at x=3000 (like a conditional mean given x).

  $\hat{Y}_{x=3000} = 1.14$ which is not in [0,1].

- What is a better way to model a 0-1 response using regression?

- At each x value, there's a certain chance of a 0 and a certain chance of a 1.



- For example, in this data set, it looks more likely to get a 1 at higher values of x.

- Or, $P(Y = 1)$ changes with the x-value.

- Conditioning on a given x, we have $P(Y = 1)$, and $P(Y = 1) \in (0, 1)$.

- We will consider this probability of getting a 1 given the x-value(s) in our modeling...

$$\pi_i = P(Y_i = 1 | X_i)$$

  The previous plot shows that $P(Y = 1)$ depends on the value of x.

- It is actually a transformation of this probability (π_i) that we will use as our response in the regression model.

Odds of an Event

- First, let's discuss the <u>odds of an event</u>.

  When two fair coins are flipped,
  P(two heads)=1/4
  P(not two heads)=3/4

  The odds in favor of getting two heads is:

  $$\text{odds} = \frac{P(2\ heads)}{P(not\ 2\ heads)} = \frac{1/4}{3/4} = 1/3$$

  or sometimes referred to as 1 to 3 odds.

  You're 3 times as likely to *not get 2 heads* as you are to *get 2 heads*.

- For a binary variable Y (2 possible outcomes),

  odds in favor of Y=1 is $\frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1-P(Y=1)}$

- For example, if P(heart attack)=0.0018, then the odds of a heart attack is

$$\frac{0.0018}{0.9982} = \frac{0.0018}{1-0.0018} = 0.001803$$

- The ratio of the odds for two different groups is also a quantity of interest.

  For example, consider heart attacks for "male nonsmoker vs. male smoker"

  Suppose P(heart attack)=0.0036 for a male smoker, and P(heart attack)=0.0018 for a male nonsmoker.

Then, the odds ratio ($O.R.$) for a heart attack in nonsmoker vs. smoker is

$$O.R. = \frac{\text{odds of a heart attack for non-smoker}}{\text{odds of a heart attack for smoker}}$$

$$= \frac{\left(\frac{Pr(\text{heart attack}|\text{non-smoker})}{1-Pr(\text{heart attack}|\text{non-smoker})}\right)}{\left(\frac{Pr(\text{heart attack}|\text{smoker})}{1-Pr(\text{heart attack}|\text{smoker})}\right)}$$

$$= \frac{\left(\frac{0.0018}{0.9982}\right)}{\left(\frac{0.0036}{0.9964}\right)} = 0.4991$$

- Interpretation of the odds ratio for binary 0-1
  - $O.R. \geq 0$
  - If $O.R. = 1.0$, then $P(Y = 1)$ is the same in both samples
  - If $O.R. < 1.0$, then $P(Y = 1)$ is less in the numerator group than in the denominator group
  - $O.R. = 0$ if and only if $P(Y = 1) = 0$ in numerator sample

Back to Logistic Regression...

- The response variable we will model is a transformation of $P(Y_i = 1)$ for a given X_i.

- The transformation is the *logit* transformation
$$logit(a) = ln\left(\frac{a}{1-a}\right)$$

- The response variable we will use:
$$logit[P(Y_i = 1)] = ln\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right)$$
This is the $\log_e$ of the odds that $Y_i = 1$.

Notice: $P(Y_i = 1) \in (0, 1)$

$$\left(\frac{P(Y_i=1)}{1-P(Y_i=1)}\right) \in (0, \infty)$$

and $-\infty < ln\left(\frac{P(Y_i=1)}{1-P(Y_i=1)}\right) < \infty$

- The logistic regression model:

$$ln\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$$

  - This response on the left isn't 'bounded' by [0,1] eventhough the Y-values themselves are (having the response bounded by [0,1] was a problem before).

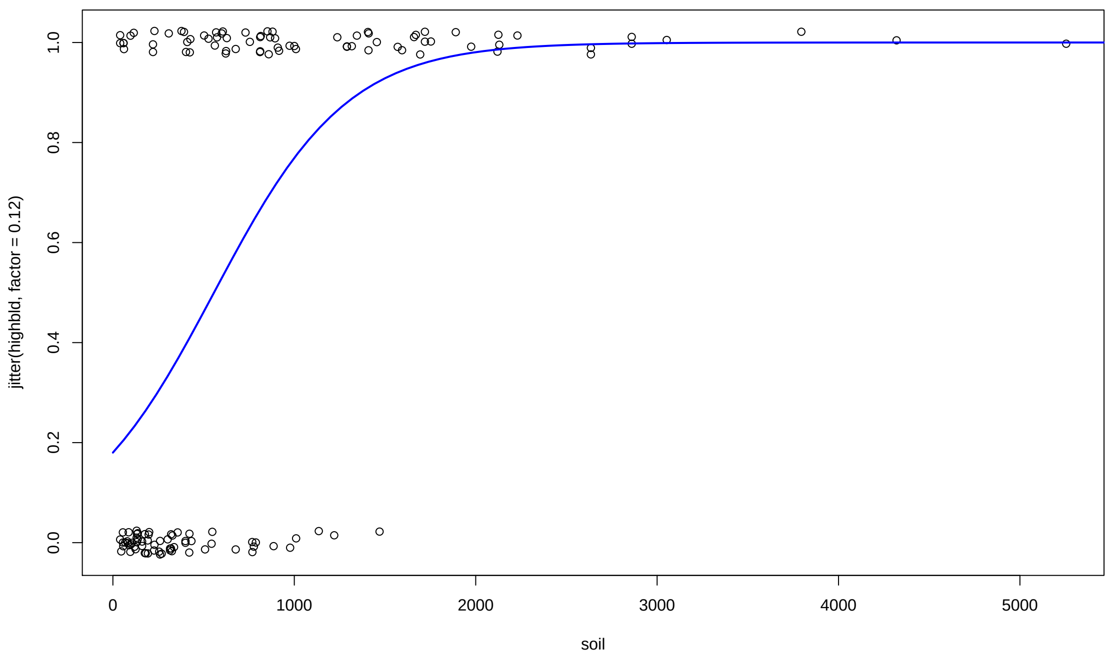
  - The response on the left can feasibly be any positive or negative quantity.

  - This a nice characteristic because the right side of the equation can 'potentially' give any possible predicted value $-\infty$ to ∞.

- The logistic regression model is a **GENERALIZED LINEAR MODEL**. Linear model on the right, something other than the usual continuous Y on the left.

- Let's look at the fitted logistic regression model for the lead level data with one X covariate.

$$ln\left(\frac{P(Y_i=1)}{1-P(Y_i=1)}\right)=\hat{\beta}_0+\hat{\beta}_1 X_i=-1.5160+0.0027\ X_i$$



The curve represents the $E(Y_i|x_i)$.

And...

$$E(Y_i|x_i)=0\cdot P(Y_i=0|x_i)+1\cdot P(Y_i=1|x_i)$$

$$=P(Y_i=1|x_i)$$

- We can manipulate the regression model to put it in terms of $P(Y_i = 1) = E(Y_i)$

- If we take this regression model

$$ln\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = -1.5160 + 0.0027\ X_i$$

  and solve for $P(Y_i = 1)$ we get...

$$P(Y_i = 1) = \frac{exp(-1.5160 + 0.0027\ X_i)}{1 + exp(-1.5160 + 0.0027\ X_i)}$$

  - The value on the right is bounded to [0,1].

  - Because our β_1 is positive,
    - As X gets larger, $P(Y_i = 1)$ goes to 1.
    - As X gets smaller, $P(Y_i = 1)$ goes to 0.

  - The fitted curve, i.e. the function of X_i on the right, is S-shaped (sigmoidal)

- In the general model with one covariate:

$$ln\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \beta_1 \, X_i$$

which means

$$P(Y_i = 1) = \frac{exp(\beta_0 + \beta_1 \, X_i)}{1 + exp(\beta_0 + \beta_1 \, X_i)}$$

$$= \frac{1}{1 + exp[-(\beta_0 + \beta_1 \, X_i)]}$$

- If β_1 is positive, we get the S-shape that goes to 1 as X_i goes to ∞ and goes to 0 as X_i goes to $-\infty$ (as in the lead example).

- If β_1 is negative, we get the opposite S-shape that goes to 0 as X_i goes to ∞ and goes to 1 as X_i goes to $-\infty$.

- Another way to think of logistic regression is that we're modeling a Bernoulli random variable occurring for each X_i, and the Bernoulli parameter π_i depends on the covariate value.

Then, $\quad Y_i|X_i \sim Bernoulli(\pi_i)$

where $\quad \pi_i$ represents $P(Y_i = 1|X_i)$

$E(Y_i|X_i) = \pi_i$ and

$V(Y_i|X_i) = \pi_i(1 - \pi_i)$

We're thinking in terms of the conditional distribution of $Y|X$ (we don't have constant variance across the X values, mean and variance are tied together).

Writing π_i as a function of X_i,

$$\pi_i = \frac{exp(\beta_0 + \beta_1\ X_i)}{1 + exp(\beta_0 + \beta_1\ X_i)}$$

Interpretation of parameters

For the lead levels example.

- Intercept β_0:

  When $X_i = 0$, there is no lead in the soil in the backyard. Then,

$$ln\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0$$

  So, β_0 is the log-odds that a randomly selected child with no lead in their backyard has a high lead blood level.

  Or, $\pi_i = \frac{e^{\beta_0}}{1+e^{\beta_0}}$ is the chance that a kid with no lead in their backyard has high lead blood level.

  For this data,

$$\hat{\beta}_0 = -1.5160 \quad \text{or} \quad \hat{\pi}_i = 0.1800$$

Interpretation of parameters

For the lead levels example.

- Coefficient β_1:

  Consider the $\log_e$ of the odds ratio $(O.R.)$ of having high lead blood levels for the following 2 groups...

  1) those with exposure level of x
  2) those with exposure level of $x+1$

$$
\begin{aligned}
log_e\left(\frac{\frac{\pi_2}{1-\pi_2}}{\frac{\pi_1}{1-\pi_1}}\right) &= log_e\left(\frac{e^{\beta_0}e^{\beta_1(x+1)}}{e^{\beta_0}e^{\beta_1 x}}\right) \\
&= log_e\left(e^{\beta_1}\right) \\
&= \beta_1
\end{aligned}
$$

  β_1 is the $\log_e$ of $O.R.$ for a 1 unit increase in x.

  It compares the groups with x exposure and $(x+1)$ exposure.

Or, un-doing the log,

$$e^{\beta_1} \text{ is the } O.R. \text{ comparing the two groups.}$$

For this data,

$$\hat{\beta}_1 = 0.0027 \quad \text{or} \quad e^{0.0027} = 1.0027$$

A 1 unit increase in X increases the odds of having a high lead blood level by a factor of 1.0027.

Though this value is small, the range of the X values is large(40 to 5255), so it can have a substantial impact when you consider the full spectrum of x values.

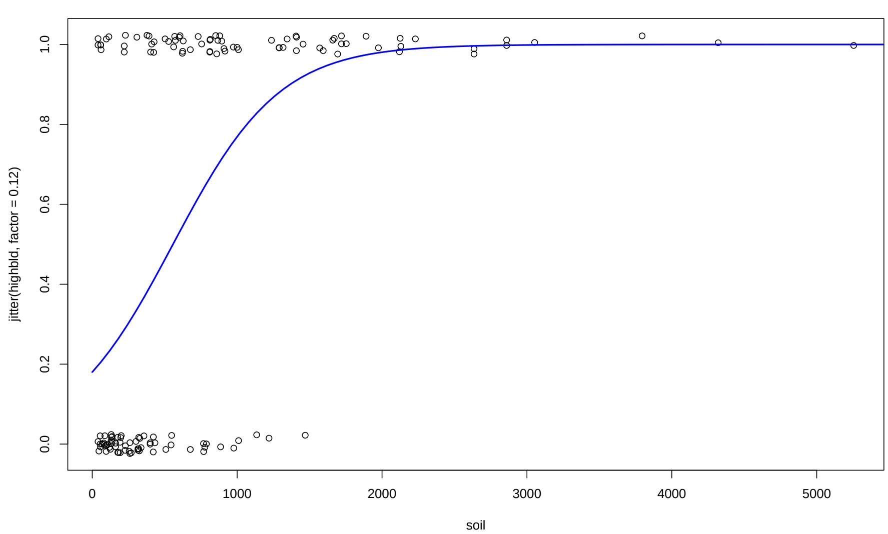
Prediction

- What is the predicted probability of high lead blood level for a child with $X = 500$.

$$P(Y_i = 1) = \frac{exp(-1.5160 + 0.0027\ X_i)}{1 + exp(-1.5160 + 0.0027\ X_i)}$$
$$= \frac{1}{1 + exp[-(-1.5160 + 0.0027\ X_i)]}$$
$$= \frac{1}{1 + exp[1.5160 - 0.0027 \times 500]} \;=\; 0.4586$$

- What is the predicted probability of high lead blood level for a child with $X = 4000$.

$$P(Y_i = 1) = \frac{1}{1 + exp[1.5160 - 0.0027 \times 4000]} \;=\; 0.9999$$



- What is the predicted probability of high lead blood level for a child with $X = 0$.

$$P(Y_i = 1) = \frac{1}{1 + exp[1.5160]} \;=\; 0.1800$$

So, there's still a chance of having high lead blood level even when the backyard doesn't have any lead. This is why we don't see the low-end of our fitted S-curve go to zero.

Testing Significance of Covariate

```
> glm.out=glm(highbld~soil,family=binomial(logit))
> summary(glm.out)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.5160568  0.3380483  -4.485 7.30e-06 ***
soil         0.0027202  0.0005385   5.051 4.39e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Soil is a significant predictor in the logistic regression model.

A couple things...

- Method of **Maximum Likelihood** is used to fit the model.

  Estimates $\hat{\beta}_j$ are asymptotically normal, so **R** uses Z-tests (or Wald tests) for covariate significance in the output. [NOTE: squaring a z-statistic gets you a chi-squared statistic with 1 df.]

- General concepts easily extended to logistic regression with many covariates.