

## Lecture 5. Logistic Regression

Suzer-Gurtekin

February 2025

# Overview

- 1 [Lecture 4 Review](#)
- 2 [Predictive Accuracy](#)
- 3 [Graphical Presentation of Data and Predictions](#)
- 4 [Applications](#)

## Linear Model and Binary Outcome

In some cases, the linear model might not produce a bad fit, but there are some issues.

- 1 Linear regression assumes residuals are normally distributed.
- 2 Linear regression also assumes constant variance.
- 3 Linear model can produce predicted values outside the restricted range of a probability  $(0,1)$ .

Violations of assumptions mean linear model won't always/usually be appropriate.

# Logistic Regression

We need some function  $\eta_i = g(p_i)$  such that  $\eta_i$  can be predicted in a linear model.

It must also satisfy  $0 \leq g^{-1}(\eta) \leq 1$ .

The most common function is called the logit function:  $\eta = \log \left( \frac{p}{1-p} \right)$ .

# Logistic Regression

Logistic regression fits the following model:

$$\eta = \beta_0 + \beta_1 x_{i1} + \dots + \beta_q x_{iq}$$

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_q x_{iq}$$

$$Pr(Y_i = 1) = \text{logit}^{-1}(\beta_0 + \beta_1 x_{i1} + \dots + \beta_q x_{iq})$$

$$Pr(Y_i = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \dots + \beta_q x_{iq})}}$$

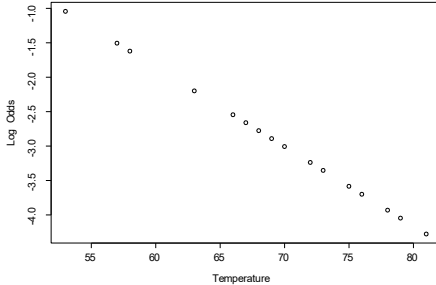
## Logistic Regression

If we fit the model  $\eta = \beta_0 + \beta_1 x_1$ , where  $\eta = \log \left( \frac{p}{1-p} \right)$  and  $x$  is the temperature, we get the following:

Parameter	DF	Estimate	Standard Error
Intercept	1	5.0850	3.0525
temp	1	-0.1156	0.0470

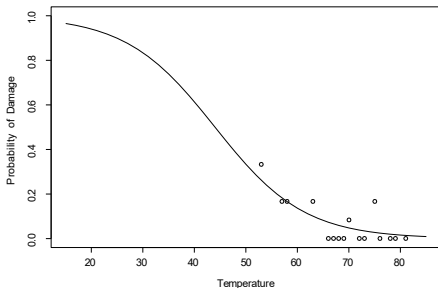
In this case  $\eta = 5.0850 - 0.1156 * TEMP$ . Let's plot the predicted values from this model.

# Logistic Regression



# Logistic Regression

We could also convert this to the probability scale:





# Logistic Regression

It is possible to use many of the techniques that you learned last semester to predict probabilities in this setup.

- Transformations
- Dummy variables
- Interactions

# Interpretation of Coefficients

Coefficients from this model don't translate easily to the probability scale. How do we interpret them?

- 1 Graphic presentation of predicted values.
- 2  $e^{\beta_1}$  can be interpreted as the increase in the odds-ratio that would result from a 1-unit increase in  $x_{i1}$  with all other  $x$ 's being held constant.
- 3 Present the probability at the mean of the variable of interest.
- 4 Present the probability at several points.
- 5 The 'divide by 4 rule'.

## Interpretation of Coefficients

We say that the exponentiated coefficient can be interpreted as an odds ratio.

Sometimes, people interpret the odds ratio as "Males are 2.25 times more likely to have disease than females."

Recall that this statement is more correct for relative risk. The odds ratio and relative risk are close when  $p$  is small

# OR and RR

	Yes	No	Row Total
Male	$\pi_{11}$	$\pi_{12}$	$\pi_{1+}$
Female	$\pi_{21}$	$\pi_{22}$	$\pi_{2+}$
Column Total	$\pi_{+1}$	$\pi_{+2}$	

	Disease	
	Yes	No
Male	0.15	0.35
Female	0.10	0.40

We want the **conditional probability** that you have disease given that you are male:  $PR(D|M) = \pi_{1|1} = \frac{\pi_{11}}{\pi_{11} + \pi_{12}}$ .

$$PR(D|M) = \pi_{1|1} = \frac{0.15}{0.15 + 0.35} = 0.3$$

$$PR(D|F) = \pi_{1|2} = \frac{0.10}{0.10 + 0.40} = 0.2$$

# Relative Risk

Now, define the relative risk.

$$\text{Relative Risk (Response Category 1)} = \frac{\pi_{1|1}}{\pi_{1|2}}$$

For example:

$$\frac{\pi_{1|1}}{\pi_{1|2}} = \frac{\Pr(D|M)}{\Pr(D|F)} = \frac{.3}{.2} = 1.5$$

# Odds Ratio

Within row 1 the *odds* (**not odds ratio**) that the response is in column 1 instead of column 2 is defined as:

$$Odds_1 = \frac{\pi_{1|1}}{\pi_{2|1}} = \frac{\pi_{11}}{\pi_{12}} \quad \frac{\text{probability of you are in column 1 given that you are in row 1}}{\text{probability of you are in column 2 given that you are in row 1}}$$

From our example, this could be written:

$$\frac{\Pr(\underline{D} | M)}{\Pr(\underline{D} | M)} = \frac{\Pr(D | M)}{1 - \Pr(D | M)}$$


	Yes	No	Row Total
Male	$\pi_{11}$	$\pi_{12}$	$\pi_{1+}$
Female	$\pi_{21}$	$\pi_{22}$	$\pi_{2+}$
Column Total	$\pi_{+1}$	$\pi_{+2}$	

Continuing the example, the odds that a man will have the disease are  $\frac{.3}{1-.3} = .43$ . For women, this odds are  $\frac{.2}{1-.2} = .25$

## Interpretation of Coefficients

Let's consider a special case. We have two predictors. Each predictor is categorical.

For example, we might have a 2x3 table.

		J		
		1	2	3
I	1			
	2			

## Interpretation of Coefficients

Each “cell” has a proportion with the condition.

The logistic regression model for this setup takes the following form:

$$\text{logit}(p_{ij}) = \log \left( \frac{p_{ij}}{1-p_{ij}} \right) = \beta_0 + a_i + \beta_j$$

The log-odds ratio for the cell (1,2) versus the cell (1,3) is:

$$\begin{aligned} \log \left( \frac{p_{12}/(1-p_{12})}{p_{13}/(1-p_{13})} \right) &= (\beta_0 + a_1 + \beta_2) - (\beta_0 + a_1 + \beta_3) \\ &= \beta_2 - \beta_3 \end{aligned}$$



## Interpretation of Coefficients

The corresponding odds ratio is  $e^{(\beta_2 - \beta_3)}$ .

This is easier to see in the two-level factor cases. Then, one of the categories is a reference category.

For example  $\beta_1$  and  $\beta_2$  are the two levels. Normally (effect vs. reference coding), we set  $\beta_1 = 0$ . Then,  $e^{\beta_2}$  is the odds-ratio for 2 relative to 1.

Note: We had to hold  $i$  constant. Every other predictor has to be held constant.

Won't work when interactions are present between the factor under consideration and any other predictor in the model.

# In-Class Exercise I

## Recall ORINGS data

```
> orings <- read.table("D:/Teaching/Survmeth686/Winter 2024/Lectures/04 Logistic/s
pacshu.dat", quote="\")
> orings[1:10,]
  v1 v2
1  53  1
2  53  1
3  53  0
4  53  0
5  53  0
6  53  0
7  57  1
8  57  0
9  57  0
10 57  0
> dim(orings)
[1] 138  2
```

# In-Class Exercise I

## Recall ORINGS data

```
> table(orings)
```

```
      damage  
temp  0  1  
  53  4  2  
  57  5  1  
  58  5  1  
  63  5  1  
  66  6  0  
  67 18  0  
  68  6  0  
  69  6  0  
  70 22  2  
  72  6  0  
  73  6  0  
  75 10  2  
  76 12  0  
  78  6  0  
  79  6  0  
  81  6  0
```

16x2 Table

# In-Class Exercise I

## Recall ORINGS data

```
#create an empirical logit plot.
#creating bins
orings2<-orings
orings2$temp.grp[orings2$temp< 60] <- 55
orings2$temp.grp[orings2$temp >= 60 & orings2$temp <65 ] <- 60
orings2$temp.grp[orings2$temp >= 65 & orings2$temp <70 ] <- 65
orings2$temp.grp[orings2$temp >= 70 & orings2$temp <75 ] <- 70
orings2$temp.grp[orings2$temp>=75] <- 75
```

	1	0	Row Total
Lower than 60	4	14	18
Equal and higher than 60 and lower than 65	1	5	6
Equal and higher than 65 and lower than 70	0	36	36
Equal and higher than 70 and lower than 75	2	34	36
Equal and higher than 75	2	40	42
<b>Column Total</b>	<b>9</b>	<b>129</b>	<b>138</b>

5x2 Table

## In-Class Exercise I

Compute:

1. Cell probabilities (for example:  $\pi_{11} = \frac{4}{138} = 0.029$ )
2. Relative Risk of “Lower than 60 degrees” compared to “Equal or higher than 75.”
3. Odds Ratio of oring failure at “Lower than 60 degrees”
4. Odds Ratio of oring failure at “Equal or higher than 75”
5. What is the difference between Odds Ratio of oring failure at “Equal or higher than 75” and “Lower than 60 degrees”
6. Fit a logistic regression model by creating the categorical variable for Temp:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(X = 55) + \beta_2(X = 60) + \beta_3(X = 65) + \beta_4(X = 70) + \beta_5(X = 75)$$

# Group Assignments

## Group Student

- |   |                            |
|---|----------------------------|
| 1 | Einolf, Zach Scott         |
| 1 | Fan, Zhaoyun               |
| 1 | Mishra, Rohin Prem         |
| 1 | DesJardins, Grace          |
| 2 | Adeniyi, Kehinde           |
| 2 | Lugu, Nicholas Reign       |
| 2 | LU, Aria                   |
| 2 | Gunderson, Jeremy          |
| 3 | Wenner, Theodore D         |
| 3 | Zhou, Zhenjing             |
| 3 | Kim, Jay                   |
| 3 | Bei, Rongqi                |
| 4 | Beshaw, Yael Dejene        |
| 4 | Hoglund, Quentin Michael   |
| 4 | Jiang, Yujing              |
| 4 | Jiang, Weishan             |
| 5 | Popky, Dana                |
| 5 | Sani, Jamila               |
| 5 | O'Connell, Greg Al         |
| 5 | Saucedo, Valeria Castaneda |

## Group Student

- |    |                      |
|----|----------------------|
| 6  | Hussein, Aya Moham   |
| 6  | Zou, Jianing         |
| 6  | Wang, Zixin          |
| 6  | Chakravarty, Sagnik  |
| 7  | Valmidiano, Megan    |
| 7  | Glidden, Sarah Acton |
| 7  | Sun, Yao             |
| 7  | Blakney, Aaron       |
| 8  | Xu, Kailin           |
| 8  | Linares, Kevin       |
| 8  | Odei, Doris          |
| 8  | Nana Mba, Line       |
| 9  | Zhou, Huan           |
| 9  | Meng, Lingchen       |
| 9  | Lin, Xinyu           |
| 9  | Ge, Feiran           |
| 10 | Liu, Xiaoqing        |
| 10 | Lu, Angelina         |
| 10 | Baez-Santiago, Felix |
| 10 | Ma, Ruisi            |

## Group Student

- |    |                           |
|----|---------------------------|
| 11 | Ding, Yuchen              |
| 11 | Shrivastava, Namit        |
| 11 | Kakiziba, Johnia Johansen |
| 11 | Cranmer, Evan Koba        |

## In-Class Exercise

The low birth weight problem produced this partial output...

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.3611	1.1046	1.5184	0.2179
AGE	1	-0.0295	0.0370	0.6367	0.4249
<...>					
HT	1	1.8633	0.6975	7.1356	0.0076
<...>					

- 1 What is the odds ratio for someone with HT (history of hypertension) relative to someone without HT?
- 2 What is the odds ratio for someone age 21 relative to someone age 20?
- 3 What is the odds ratio for someone age 30 relative to someone age 20?

## In-Class Exercise

The low birth weight problem produced this partial output...

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.3611	1.1046	1.5184	0.2179
AGE	1	-0.0295	0.0370	0.6367	0.4249
<...>					
HT	1	1.8633	0.6975	7.1356	0.0076
<...>					

- 1 What is the odds ratio for someone with HT relative to someone without HT?  $e^{1.8633} = 6.4450$
- 2 What is the odds ratio for someone age 21 relative to someone age 20?  $e^{-0.0295 \times 1} = 0.9709$
- 3 What is the odds ratio for someone age 30 relative to someone age 20?  $e^{-0.0295 \times 10} = 0.7445$



## More on Residuals

We saw that residuals have an unusual pattern since the outcome is either 0 or 1.

$$e_i = y_i - \hat{p}_i$$

Other types of residuals are possible with logistic regression. Pearson residual (raw residual divided by binomial standard deviation):

$$r_i = \frac{e_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$$

## More on Residuals

Another is the deviance residual. This is a case-level version of the deviance residuals.

$$d_i = \text{sign}(e_i)[2(y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i))]^{1/2}$$

Squaring these residuals and summing them will give the deviance.

`sign` returns a vector with the signs of the corresponding elements of `x` (the sign of a real number is 1, 0, or -1 if the number is positive, zero, or negative, respectively)

## More on Residuals

In R:

```
> orings$p_hat<-predict(logitmod,orings,type="response")
> orings$e<-orings$damage-orings$p_hat
> orings$d <- sign(orings$e)*sqrt(-2*(orings$damage*log(orings$p_hat) +
  (1 - orings$damage)*log(1 - orings$p_hat)))
> orings$d.alt <- residuals(logitmod)
> head(orings)
```

	temp	damage	p_hat	e	d	d.alt
1	53	1	0.2607865	0.7392135	1.6395445	1.6395445
2	53	1	0.2607865	0.7392135	1.6395445	1.6395445
3	53	0	0.2607865	-0.2607865	-0.7773912	-0.7773912
4	53	0	0.2607865	-0.2607865	-0.7773912	-0.7773912
5	53	0	0.2607865	-0.2607865	-0.7773912	-0.7773912
6	53	0	0.2607865	-0.2607865	-0.7773912	-0.7773912

```
> sum(orings$d^2)
[1] 60.39634
> deviance(logitmod)
[1] 60.39634
```

## More on Residuals

Recall the LRT for comparing observed and expected we learned in Weeks 1 and 2:

$$G^2 = 2 \sum_{i=1}^k O_i \ln \left( \frac{O_i}{E_i} \right)$$

In fact, the deviance residual is a reformulation of this same idea.

$$G^2 = 2 \sum_{i=1}^n \left[ y_i \log \left( \frac{y_i}{\hat{p}_i} \right) + (n_i - y_i) \log \left( \frac{n_i - y_i}{n_i - \hat{p}_i} \right) \right]$$

## Predictive Accuracy

For logistic regression, the prediction is a probability.

What if we want a prediction that can be observed – that is, (0, 1)?

We could choose a cutoff value ( $\pi_0$ ).

Every prediction  $\hat{\pi}_i > \pi_0$  is set to 1,  $\hat{y}_i = 1$ .

Every prediction  $\hat{\pi}_i < \pi_0$  is set to 0,  $\hat{y}_i = 0$ .

I will show a confusion matrix. I wouldn't normally do this unless I have a specific decision in mind where I know the costs of misclassification. I'm using this as a way to understand AUC.

## Predictive Accuracy

We need to define *sensitivity* and *specificity*.

Sensitivity  $\Rightarrow P(\hat{y} = 1 | y = 1)$ ;  $\frac{TP}{TP+FN}$ , aka “true positive rate.”

Specificity  $\Rightarrow P(\hat{y} = 0 | y = 0)$ ;  $\frac{TN}{FP+TN}$ , aka “true negative rate.”

	$y = 1$	$y = 0$
$\hat{y} = 1$	TP	FP
$\hat{y} = 0$	FN	TN

## Predictive Accuracy

Related concepts are **false positive rate** (FPR) and **false negative rate** (FNR).

$$FPR = \frac{FP}{FP+TN} \text{ or Specificity} = 1 - FPR$$

$$FNR = \frac{FN}{TP+FN} \text{ or Sensitivity} = 1 - FNR$$

	$y = 1$	$y = 0$
$\hat{y} = 1$	TP	FP
$\hat{y} = 0$	FN	TN

## Predictive Accuracy

To understand this, let's look at the extreme cutoff values first.

		Specificity	Sensitivity
$\pi_0 = 0$	All $\hat{y}_i = 1$	0	1
$\pi_0 = 1$	All $\hat{y}_i = 0$	1	0



## Predictive Accuracy

```
#Confusion matrix
# Use the model to make predictions, these are probabilities
pdata <- predict(lowbwt.mod, type = "response")

# use caret and compute a confusion matrix
#First, categorize probabilities into 0/1 using a cutoff value
pclass.64 = as.factor(as.numeric(pdata>0.64))

confusionMatrix(pclass.64,
  reference=as.factor(lowbwt$LOW),
  positive="1")
```

## Predictive Accuracy

From Example 5. If we set  $\pi_0 = 0.64$ , we get the following classification table (notice that 1,1 is in lower right corner):

```
> confusionMatrix(pclass.64,  
+                 reference=as.factor(lowbwt$LOW),  
+                 positive="1")  
Confusion Matrix and Statistics
```

	Reference	
Prediction	0	1
0	128	49
1	2	10

```
Accuracy : 0.7302  
95% CI : (0.6609, 0.792)  
No Information Rate : 0.6878  
P-Value [Acc > NIR] : 0.1187
```

```
Kappa : 0.197
```

## Predictive Accuracy

And the following output:

Sensitivity : 0.16949

Specificity : 0.98462

	$y = 1$	$y = 0$	
$\hat{y} = 1$	10	2	12
$\hat{y} = 0$	49	128	177

$$\text{Sensitivity} = \frac{10}{10+49} = 0.169$$

$$\text{Specificity} = \frac{128}{128+2} = 0.985$$

$$1 - \text{Sensitivity} = \text{FNR} = \frac{49}{10+49} = 0.831$$

$$1 - \text{Specificity} = \text{FPR} = \frac{2}{2+128} = 0.015$$

## In-Class Exercise

Calculate the sensitivity and specificity of the classification of the data in each of the following tables.

Table:  $\pi_0 = 0.3$

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	46	13
$y = 0$	49	81

Table:  $\pi_0 = 0.7$

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	2	57
$y = 0$	1	129

## In-Class Exercise

Calculate the sensitivity and specificity of the classification of the data in each of the following tables.

Table:  $\pi_0 = 0.3$

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	46	13
$y = 0$	49	81

$$\text{Sensitivity} = \frac{46}{46+13} = 0.780$$

$$\text{Specificity} = \frac{81}{81+49} = 0.623$$

Table:  $\pi_0 = 0.7$

	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	2	57
$y = 0$	1	129

$$\text{Sensitivity} = \frac{2}{2+57} = 0.034$$

$$\text{Specificity} = \frac{129}{129+1} = 0.992$$

## Predictive Accuracy

We can create many tables with different values for  $\pi_0$ .

Or, we can use the ROC curve to display them all.

ROC curve plots the sensitivity and (1-specificity) for all possible  $\pi_0$  values.

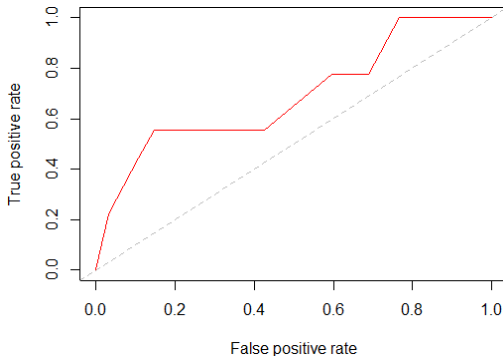
## Predictive Accuracy

The R code for the sapce shuttle example to produce an ROC curve:

```
#ROC curve.  
library(ROCR)  
spacshu$m1.yhat <- predict(logitmod, spacshu, type = "response")  
spacshu$m1.yhat  
m1.scores <- prediction(spacshu$m1.yhat, spacshu$damage)  
  
plot(performance(m1.scores, "tpr", "fpr"), col = "red")  
abline(0,1, lty = 8, col = "grey")
```

# Predictive Accuracy

Using R, the Space Shuttle ROC curve:



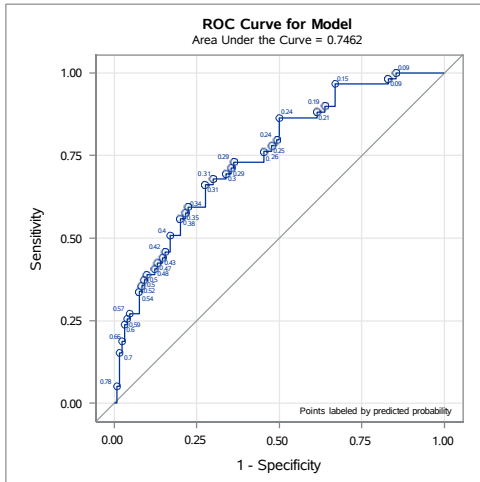


## Predictive Accuracy

The SAS code for the low birthweight example to produce an ROC curve:

```
ods graphics on;  
ods pdf file="C:\temp\example5roc.pdf";  
ods select ROCCurve;  
proc logistic data=lecture8.lowbwt descending plots=roc(id=prob);  
class race(ref='3') /param = ref ;  
model low=age lwt race smoke ptl ht ui ftv / rsq lackfit ;  
run;  
ods pdf close;  
ods graphics off;
```

# Predictive Accuracy



## Predictive Accuracy

A summary of the ROC curve is known as the **Area Under the Curve** (AUC).

Literally the area under the ROC curve.

AUC ranges from 0.5 (random guessing) to 1.0 (each case correctly classified 0,1).

# Predictive Accuracy

In R, the following code:

```
library(ROCR)
library(AUC)
spacshu$ml.yhat <- predict(logitmod, spacshu, type = "response")
spacshu$ml.yhat

ml.scores <- prediction(spacshu$ml.yhat, spacshu$damage)
auc(roc(ml.scores@predictions[[1]],ml.scores@labels[[1]]))
```

.... produces the following output:

```
> auc(roc(ml.scores@predictions[[1]],ml.scores@labels[[1]]))
[1] 0.6912145
```

## Predictive Accuracy

In SAS, the AUC is called “c.” Automatically produced by PROC LOGISTIC:

Association of Predicted Probabilities and Observed

Percent Concordant	65.4	Somers' D	0.382
Percent Discordant	27.1	Gamma	0.413
Percent Tied	7.5	Tau-a	0.047
Pairs	1161	c	0.691

# Graphical Display

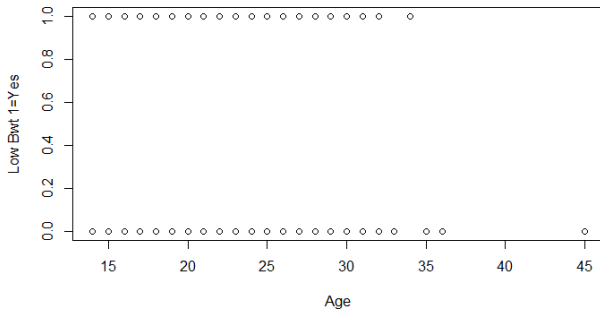
The binary nature of the outcome raises issues with graphical displays.

It may be difficult to look at the observed values against potential covariates.

How do we check model fit? How do we present results?

## Relationship to Covariates

From example 5, we can display age by the outcome – low birth weight:



Difficult to read as points are “stacked.”

## Relationship to Covariates

One solution is to “jitter” the points. Add some small noise to each point to “unstack” them. In R:

```
plot(lowbwt$AGE, jitter(lowbwt$LOW))
```

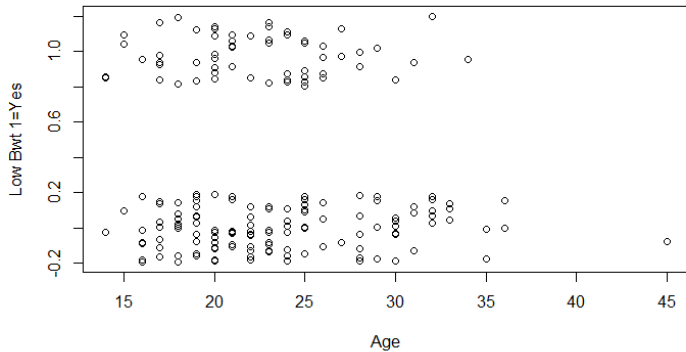
In SAS:

```
data low4;  
set low3;  
low_jit=low+rannor(0)/20;  
run;
```



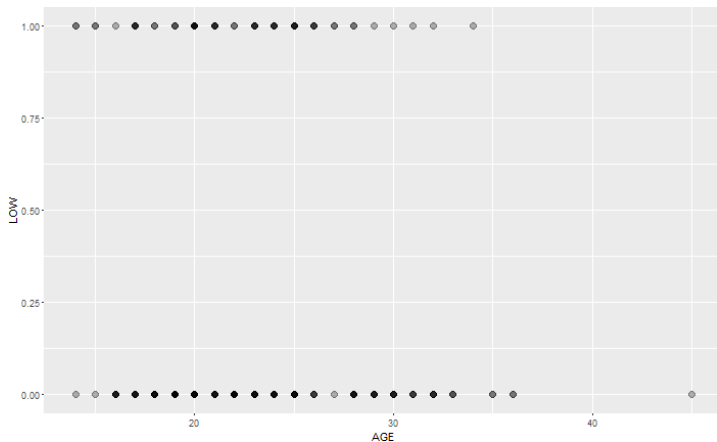
## Relationship to Covariates

Easier to read:



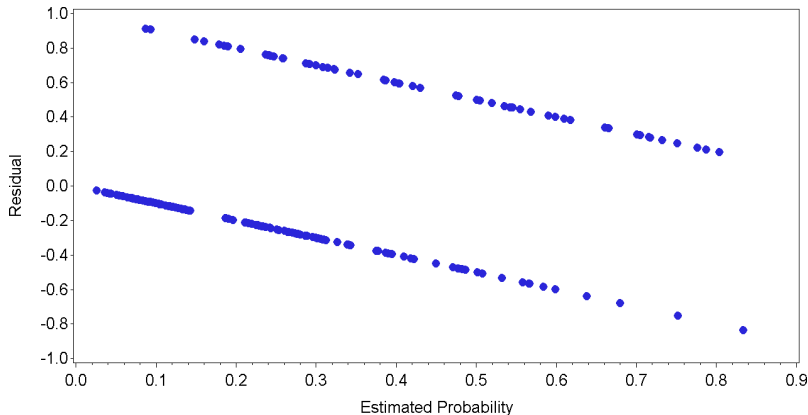
# Relationship to Covariates

Another option:



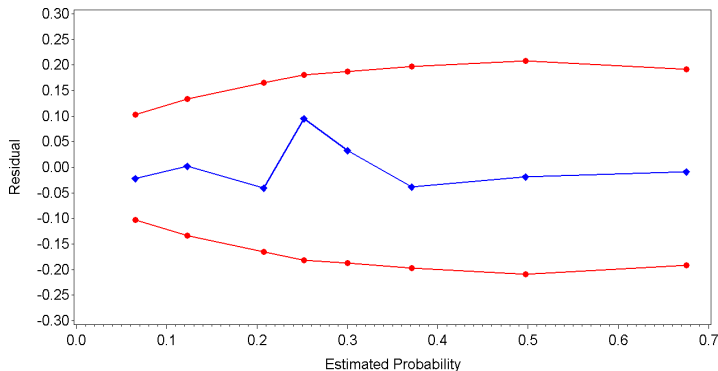
# Model Fit

We can also look at the residuals. Here are the residuals plotted against the fitted values:



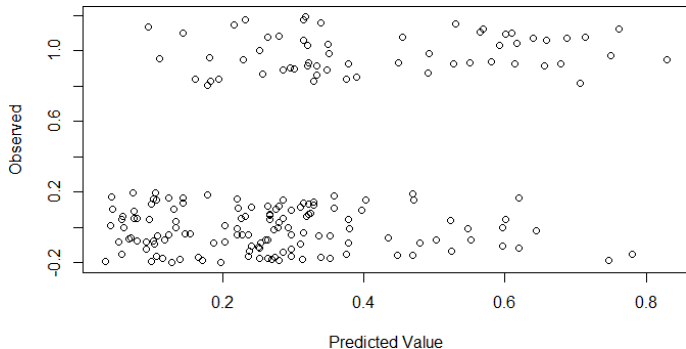
# Model Fit

- This is not very helpful as any given predicted value can only take on two values for the residual.
- One method around this is to group cases, average their residuals and fitted values, and plot those. "Binned residual plot."



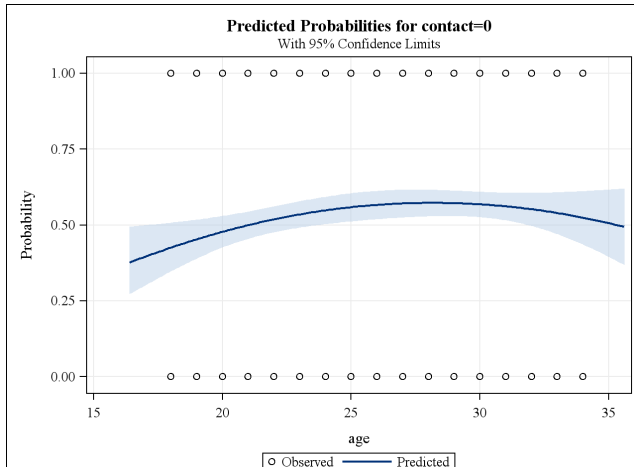
## Model Fit

We can look at the (jittered) observed versus the predicted:



# Model Fit

We can also look at predicted values against a predictor (from a model we will look at later):



## Model Fit

Collinearity can be issue. Often see very large standard errors for coefficients when this is a problem.

Use linear regression to check for collinearity.

```
#Look at collinearity using VIF  
#Doesn't matter that the outcome is binary.  
#Outcome not used in VIF  
lowbwt.lm<-lm(LOW ~ AGE + LWT + factor(RACE) + SMOKE +  
              PTL + HT + UI + FTV,data=lowbwt)  
vif(lowbwt.lm)
```

## Nonresponse Weighting Adjustments

One use for logistic regression is to predict the probability of response and use the inverse as a nonresponse adjustment.

The method models the response process. If this is unrelated to the survey data, then the weights will just add noise.

*(Little and Vartivarian, 2005)*

Since we *estimate* the response propensities, these estimates are often smoothed using response-propensity stratification.

*(Little, 1986)*



## Example: Study of Self-Employed Persons

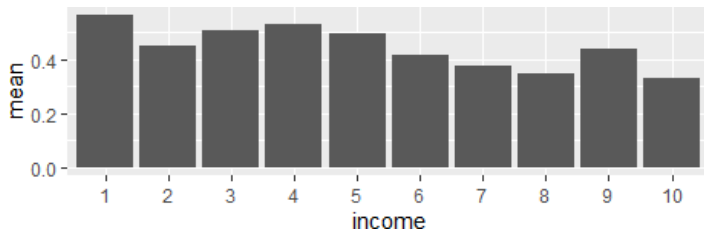
We have an example dataset. There were 953 completes out 2,302 sampled cases.  $RR=41.4\%$

On the sampling frame, we have the following variables:

Variable	Description
REGION	Census Region (1-4)
AGE	Age categories (1-13)
RACE	4 Groups (1=white,2=black,3=hispanic,4=other)
INCOME	10 Groups
MARITAL_STATUS	1=Married,2=Living w/Partner,3=Divorced, 4=Separated,5=Widowed,6=Never Married
QSEX	1=Male, 2=Female
EDUC	7 Categories

## Example: Study of Self-Employed Persons

First, look at bivariate relationships with COMPLETE.



## Example: Study of Self-Employed Persons

	Resid.	Dev	Pr(>Chi)	
NULL	3122.8			
as.factor(region)	3121.4	0.709223		
age	3112.8	0.003441	**	
as.factor(race)	3109.6	0.354505		
income	3069.9	2.94e-10	***	
as.factor(marital_status)	3060.3	0.088085	.	
as.factor(qsex)	3058.1	0.139015		
as.factor(educ)	3052.8	0.498987		

# Example: Study of Self-Employed Persons

Coefficients:

	Estimate	Std. Error	z value
(Intercept)	-0.15643	0.18330	-0.853
age	0.06221	0.01786	3.484
as.factor(race)2	-0.02473	0.14521	-0.170
as.factor(race)3	-0.32088	0.15363	-2.089
as.factor(race)4	-0.02421	0.23913	-0.101
income	-0.09382	0.01690	-5.551
as.factor(marital_status)2	-0.24979	0.22519	-1.109
as.factor(marital_status)3	0.46292	0.29954	1.545
as.factor(marital_status)4	0.13811	0.14239	0.970
as.factor(marital_status)5	-0.50757	0.27019	-1.879
as.factor(marital_status)6	0.11023	0.11978	0.920
as.factor(qsex)2	0.13313	0.09111	1.461

	Pr(> z )
(Intercept)	0.393416
age	0.000494 ***
as.factor(race)2	0.864761
as.factor(race)3	0.036735 *
as.factor(race)4	0.919347
income	2.84e-08 ***
as.factor(marital_status)2	0.267333
as.factor(marital_status)3	0.122236
as.factor(marital_status)4	0.332059
as.factor(marital_status)5	0.060306 .
as.factor(marital_status)6	0.357417
as.factor(qsex)2	0.143988

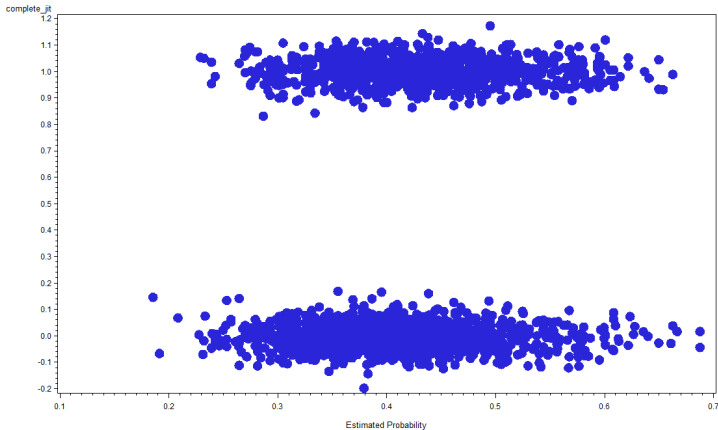
## Example: Study of Self-Employed Persons

Coefficients:

	Estimate	Std.Error	z	value	Pr(> z )
(Intercept)	-0.09433	0.14665	-0.643	0.520056	
age	0.05689	0.01669	3.409	0.000652	***
as.factor(hisp)1	-0.31525	0.15111	-2.086	0.036960	*
income	-0.09568	0.01623	-5.897	3.71e-09	***
as.factor(newmarital)2	0.17539	0.12821	1.368	0.171333	
as.factor(newmarital)3	-0.51105	0.26863	-1.902	0.057119	.
as.factor(qsex)2	0.12098	0.09028	1.340	0.180231	

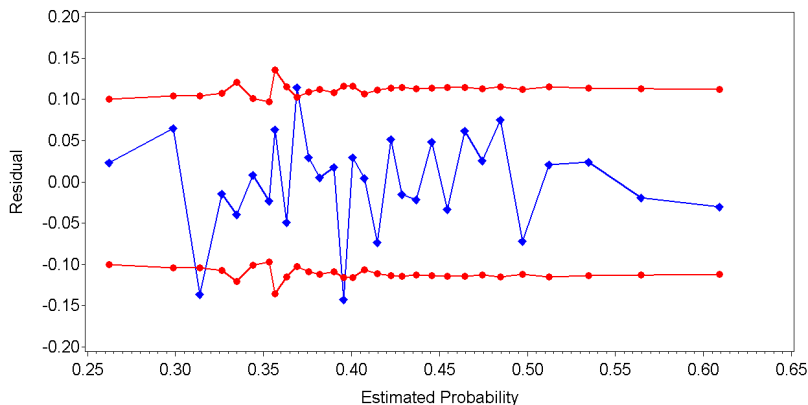
## Example: Study of Self-Employed Persons

Compare the (jittered) observed and predicted values.



## Example: Study of Self-Employed Persons

Check the binned residuals:



## Example: Study of Self-Employed Persons

nr_weight	Frequency	Percent	Cumulative Frequency
1.7829457364	129	13.54	129
1.9827586207	116	12.17	245
2.0720720721	111	11.65	356
2.2673267327	101	10.60	457
2.3365384615	104	10.91	561
2.4489795918	98	10.28	659
2.7317073171	82	8.60	741
2.8933333333	75	7.87	816
3.095890411	73	7.66	889
3.640625	64	6.72	953