Logistic Regression

Based on slides from *Linear Regression Analysis* 5E
Montgomery, Peck & Vining

Also see
https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/

Binary Response Variables

- Possible model:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \begin{cases} i = 1, 2, \ldots, n \\ y_i = 0 \text{ or } 1 \end{cases}$$

- The response y_i is a **Bernoulli** random variable

$$P(y_i = 1) = \pi_i \text{ with } 0 \le \pi_i \le 1$$

$$P(y_i = 0) = 1 - \pi_i$$

$$E(y_i) = \mu_i = \mathbf{x}_i'\boldsymbol{\beta} = \pi_i$$

$$Var(y_i) = \sigma^2_{y_i} = \pi_i(1 - \pi_i)$$

Problems With This Model

- The error terms take on only two values, so they can't possibly be normally distributed
- The variance of the observations is a function of the mean (see previous slide)
- A linear response function could result in predicted values that fall outside the 0, 1 range, and this is impossible because

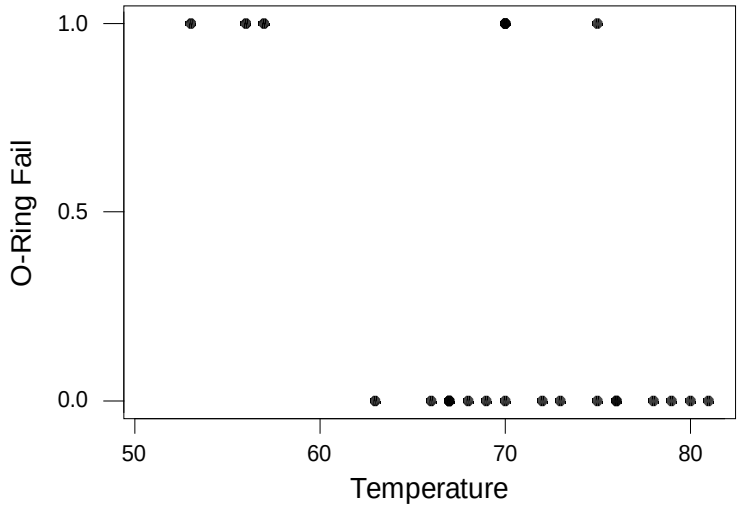
$$0 \leq E(y_i) = \mu_i = \mathbf{x}_i'\boldsymbol{\beta} = \pi_i \leq 1$$

E(Yi) = 0*P(Yi=0) + 1*P(Yi=1)

Binary Response Variables – The Challenger Data

n=24

Temperature at Launch	At Least One O-ring Failure	Temperature at Launch	At Least One O-ring Failure
53	1	70	1
56	1	70	1
57	1	72	0
63	0	73	0
66	0	75	0
67	0	75	1
67	0	76	0
67	0	76	0
68	0	78	0
69	0	79	0
70	0	80	0
70	1	81	0



Data for space shuttle launches and static tests prior to the launch of Challenger

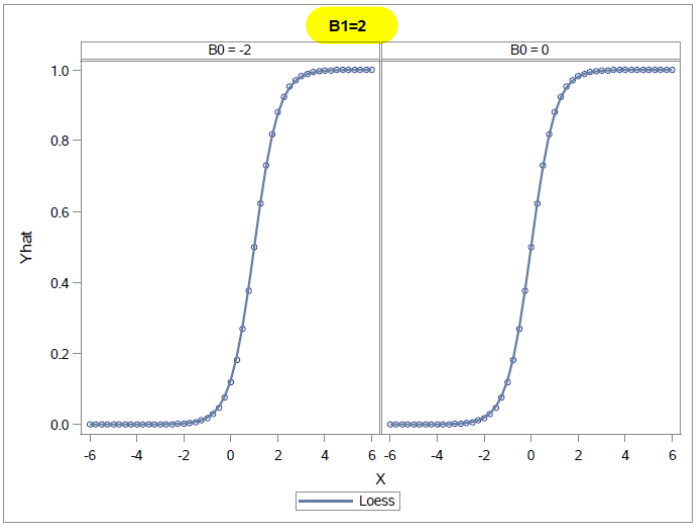
Tuesday, January 28, 1986: when the Space Shuttle Challenger broke apart 73 seconds into its flight, killing all seven crew members aboard. O-ring seals in the solid-rocket boosters were later determined to be the cause.

Binary Response Variables

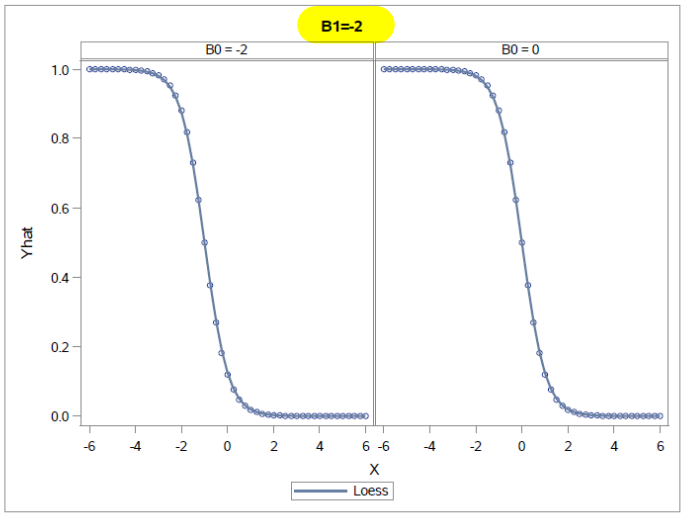
- There is a lot of empirical evidence that the response function should be nonlinear; an "S" shape is quite logical
- See the scatter plot of the Challenger data
- The **logistic response function** is a common choice note: 0 <= E(y) <= 1

$$E(y) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{x}'\boldsymbol{\beta})}$$

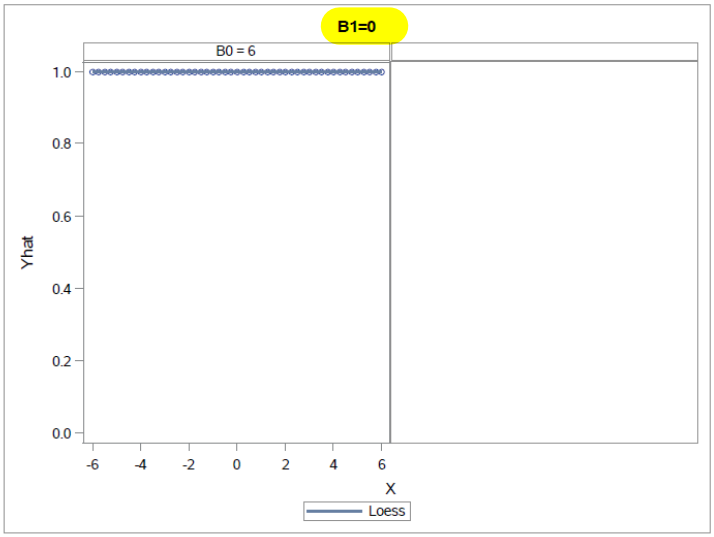
For model with 1 predictor



B1 = 2



B1 = - 2



The Logistic Response Function

- The logistic response function can be easily linearized. Let:

$$\eta = \mathbf{x}'\boldsymbol{\beta} \text{ and } E(y) = \pi$$

- Define

$$\eta = \ln \frac{\pi}{1-\pi}$$

=B0 +B1X in simple case

- This is called the **logit** transformation

Logistic Regression Model

- Model:

$$y_i = E(y_i) + \varepsilon_i$$

where

$$E(y_i) = \pi_i$$

$$= \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})}$$

- The model parameters are estimated by the method of **maximum likelihood (MLE)**

A Logistic Regression Model for the Challenger Data (Using Minitab)

Binary Logistic Regression: O-Ring Fail versus Temperature

Link Function: Logit

Response Information

Variable	Value	Count	
O-Ring F	1	7	(Event)
	0	17	
	Total	24	

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	10.875	5.703	1.91	0.057			
Temperat	-0.17132	0.08344	-2.05	0.040	0.84	0.72	0.99

Log-Likelihood = -11.515

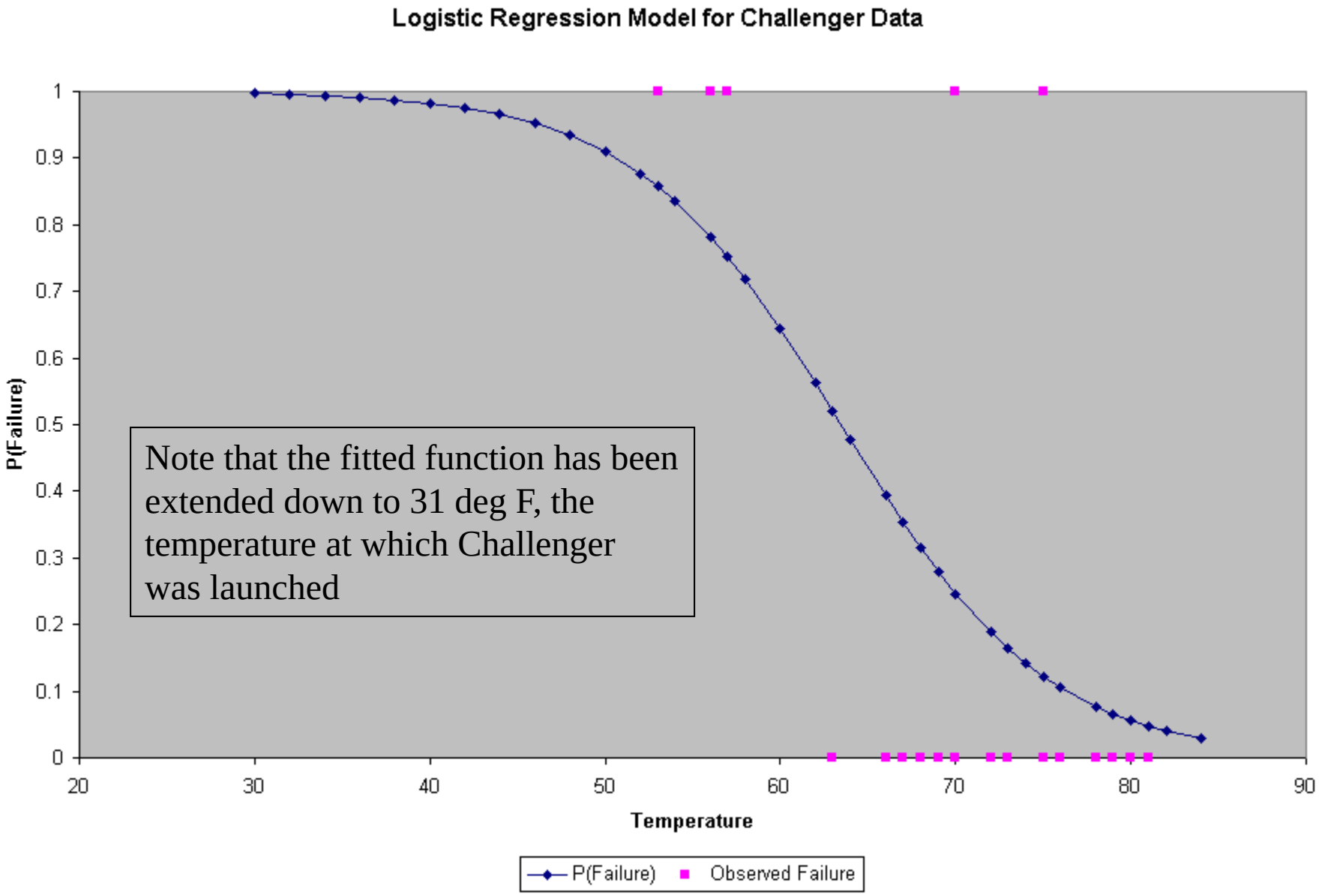
A Logistic Regression Model for the Challenger Data

Test that all slopes are zero: G = 5.944, DF = 1, P-Value = 0.015

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	14.049	15	0.522
Deviance	15.759	15	0.398
Hosmer-Lemeshow	11.834	8	0.159

$$\hat{y} = \frac{\exp(10.875 - 0.17132x)}{1 + \exp(10.875 - 0.17132x)}$$


Logistic Regression Model for Challenger Data
P(Failure)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
20
30
40
50
60
70
80
90
Temperature
Note that the fitted function has been extended down to 31 deg F, the temperature at which Challenger was launched
P(Failure)
Observed Failure

Maximum Likelihood Estimation in Logistic Regression

- The distribution of each observation y_i is

$$f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, i = 1, 2, ..., n$$

- The **likelihood function** is

$$L(\mathbf{y}, \boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

- We usually work with the log-likelihood:

$$\ln L(\mathbf{y}, \boldsymbol{\beta}) = \ln \prod_{i=1}^{n} f_i(y_i) = \sum_{i=1}^{n} \left[y_i \ln \left(\frac{\pi_i}{1-\pi_i} \right) \right] + \sum_{i=1}^{n} \ln(1-\pi_i)$$

Maximum Likelihood Estimation in Logistic Regression

- The maximum likelihood estimators (MLEs) of the model parameters are those values that maximize the likelihood (or log-likelihood) function
- ML has been around since the first part of the previous century
- Often gives estimators that are intuitively pleasing
- MLEs have nice **properties**; unbiased (for large samples), minimum variance (or nearly so), and they have an approximate normal distribution when *n* is large

Maximum Likelihood Estimation in Logistic Regression

- If we have n_i trials at each observation, we can write the log-likelihood as

$$\ln L(\mathbf{y},\boldsymbol{\beta}) = \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - \sum_{i=1}^{n} n_i \ln[1+\exp(\mathbf{x}_i'\boldsymbol{\beta})]$$

- The derivative of the log-likelihood is

$$\frac{\partial \ln L(\mathbf{y},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} - \sum_{i=1}^{n}\left[\frac{n_i}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right]\exp(\mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i$$

$$= \mathbf{X}'\mathbf{y} - \sum_{i=1}^{n}\left[n_i\pi_i\right]\mathbf{x}_i$$

$$= \mathbf{X}'\mathbf{y} - \mathbf{X}'\boldsymbol{\mu} \text{ (because } \mu_i = n_i\pi_i)$$

Maximum Likelihood Estimation in Logistic Regression

- Setting this last result to zero gives the **maximum likelihood score equations**

$$\mathbf{X}'(\mathbf{y}-\boldsymbol{\mu})=\mathbf{0}$$

- These equations look easy to solve…we've actually seen them before in **linear** regression:

$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\boldsymbol{\mu}+\boldsymbol{\varepsilon}$

$\mathbf{X}'(\mathbf{y}-\boldsymbol{\mu})=\mathbf{0}$ results from OLS or ML with normal errors

Since $\boldsymbol{\mu}=\mathbf{X}\boldsymbol{\beta}$, $\mathbf{X}'(\mathbf{y}-\boldsymbol{\mu})=\mathbf{X}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})=\mathbf{0}$,

$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}=\mathbf{X}'\mathbf{y}$, and $\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ (OLS or the normal-theory MLE)

Maximum Likelihood Estimation in Logistic Regression

- Solving the ML score equations in logistic regression isn't quite as easy, because

$$\mu_i = \frac{n_i}{1+\exp(-\mathbf{x}_i'\boldsymbol{\beta})}, i = 1, 2, \ldots, n$$

- Logistic regression is a nonlinear model
- It turns out that the solution is actually fairly easy, and is based on **iteratively reweighted least squares** or **IRLS**
- An iterative procedure is necessary because parameter estimates must be updated from an initial "guess" through several steps
- Weights are necessary because the variance of the observations is not constant
- The weights are functions of the unknown parameters

Interpretation of the Parameters in Logistic Regression

- The **log-odds** at x is

$$\hat{\eta}(x) = \ln \frac{\hat{\pi}(x)}{1-\hat{\pi}(x)} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The log-odds at $x + 1$ is

$$\hat{\eta}(x+1) = \ln \frac{\hat{\pi}(x+1)}{1-\hat{\pi}(x+1)} = \hat{\beta}_0 + \hat{\beta}_1 (x+1)$$

- The difference in the log-odds is

$$\hat{\eta}(x+1) - \hat{\eta}(x) = \hat{\beta}_1$$

Interpretation of the Parameters in Logistic Regression

- The **odds ratio** is found by taking antilogs:

$$\hat{O}_R = \frac{Odds_{x+1}}{Odds_x} = e^{\hat{\beta}_1}$$

- **The odds ratio is interpreted as the estimated increase in the probability of "success" associated with a one-unit increase in the value of the predictor variable**

Odds Ratio for the Challenger Data

$$\hat{O}_R = e^{-0.17132} = 0.84$$

Bhat = -0.17132

This implies that every decrease of one degree in temperature increases the odds of O-ring failure by about 1/0.84 = 1.19 or 19 percent

The temperature at Challenger launch was 22 degrees below the lowest observed launch temperature, so now

$$\hat{O}_R = e^{22(-0.17132)} = 0.0231$$

This results in an increase in the odds of failure of 1/0.0231 = 43.34, or about 4200 percent!!

There's a big extrapolation here, but if you knew this prior to launch, what decision would you have made?

Inference on the Model Parameters

Multiple logistic regression

Likelihood Ratio Tests

A likelihood ratio test can be used to compare a "full" model with a "reduced" model that is of interest. This is analogous to the "extra-sum-of-squares" technique that we have used previously to compare full and reduced models. The likelihood ratio test procedure compares twice the logarithm of the value of the likelihood function for the full model (FM) to twice the logarithm of the value of the likelihood function of the reduced model (RM) to obtain a test statistic, say

$$LR = 2\ln\frac{L(FM)}{L(RM)} = 2[\ln L(FM) - \ln L(RM)]$$

For large samples, when the reduced model is correct, the test statistic LR follows a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the full and reduced models. Therefore, if the test statistic LR exceeds the upper α percentage point of this chi-square distribution, we would reject the claim that the reduced model is appropriate.