**5.3 – Logistic Regression Case Study**

See 1:15-2:10 of [www.youtube.com/watch?v=j4JOjcDFtBE](http://www.youtube.com/watch?v=j4JOjcDFtBE)

and 3:31-4:22 of [www.youtube.com/watch?v=gEjXjfxoNXM](http://www.youtube.com/watch?v=gEjXjfxoNXM) (full text here:

<http://millercenter.org/scripps/archive/speeches/detail/3413>)

The January 18, 1986 explosion of the space shuttle Challenger was investigated by the Presidential Commission on the Space Shuttle Challenger Accident. The Commission's

1986 report attributed the explosion to a burn through of an O-ring seal at a field joint in one of the solid-fuel rocket boosters. This 1986 launch was the 25th space shuttle launch. After each of the previous 24 launches, the solid rocket boosters were inspected.

The following data are from the Commission's 1986 report, with the following variables:

|  |  |
| --- | --- |
| Flight | an identifier code for the launch |
| Temp | temperature (degrees F) at launch |
| Damage | indicator of damage to the field joint  (a missing value is recorded for one launch  where the solid rocket boosters were not  recovered) |

Note that seven of the 24 launches experienced field joint damage but did not explode like the Challenger. The Challenger launch was Flight STS51L (not in these data)

and the temperature was 31.

**/\* Define options \*/**

**ods html image\_dpi=300 style=journal;**

**/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/**

**/\* Read in the data and check it was read in properly \*/**

**data shuttle; input Flight $ Temp Damage $ @@; cards;**

**STS1 66 NO STS9 70 NO STS51B 75 NO STS2 70 YES**

**STS41B 57 YES STS51G 70 NO STS3 69 NO STS41C 63 YES**

**STS51F 81 NO STS4 80 . STS41D 70 YES STS51I 76 NO**

**STS5 68 NO STS41G 78 NO STS51J 79 NO STS6 67 NO**

**STS51A 67 NO STS61A 75 YES STS7 72 NO STS51C 53 YES**

**STS61B 76 NO STS8 73 NO STS51D 67 NO STS61C 58 YES**

**;**

**data shuttle; set shuttle;**

**if Damage = 'YES' | Damage = 'NO';**

**proc print data=shuttle;**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Obs** | **Flight** | **Temp** | **Damage** | | --- | --- | --- | --- | | **1** | STS1 | 66 | NO | | **2** | STS9 | 70 | NO | | **3** | STS51B | 75 | NO | | **4** | STS2 | 70 | YES | | **5** | STS41B | 57 | YES | | **6** | STS51G | 70 | NO | | **7** | STS3 | 69 | NO | | **8** | STS41C | 63 | YES | | **9** | STS51F | 81 | NO | | **10** | STS41D | 70 | YES | | **11** | STS51I | 76 | NO | | | **Obs** | **Flight** | **Temp** | **Damage** | | --- | --- | --- | --- | | **12** | STS5 | 68 | NO | | **13** | STS41G | 78 | NO | | **14** | STS51J | 79 | NO | | **15** | STS6 | 67 | NO | | **16** | STS51A | 67 | NO | | **17** | STS61A | 75 | YES | | **18** | STS7 | 72 | NO | | **19** | STS51C | 53 | YES | | **20** | STS61B | 76 | NO | | **21** | STS8 | 73 | NO | | **22** | STS51D | 67 | NO | | **23** | STS61C | 58 | YES | |

**/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**Steps in this case study:**

**1. Visualize the data**

**2. Evaluate the probability of damage based on**

**temperature**

**3. Check for influential observations and outliers**

**4. Calculate the probability of damage at**

**temperature 31 (temperature at Challenger launch)**

**5. How is logistic regression different from ANOVA?**

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/**

**/\* 1. Visualize the data \*/**

**proc sort data=shuttle; by damage;**

**proc boxplot data=shuttle;**

**plot temp \* damage;**

**title1 'Full Data';**

**run;**

**proc sgplot data=shuttle;**

**scatter y=damage x=temp /**

**markerattrs=(symbol=CIRCLEFILLED size=2pt);**

**xaxis label='Temperature at launch';**

**yaxis label='Damage to field joint';**

**title1 'Full Data';**

**run;**

|  |  |
| --- | --- |
|  |  |

**/\* 2. Evaluate the probability of damage based**

**on temperature \*/**

**proc logistic data=shuttle plots(only)=(effect);**

**model damage (event='YES') = temp / lackfit;**

**title1 'Logistic Regression with Full Data';**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Logistic Regression with Full Data*** |   Probability modeled is Damage='YES'.     | **Model Convergence Status** | | --- | | Convergence criterion (GCONV=1E-8) satisfied. |      | **Analysis of Maximum Likelihood Estimates** | | | | | | | --- | --- | --- | --- | --- | --- | | **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** | | **Intercept** | 1 | 15.0429 | 7.3786 | 4.1563 | 0.0415 | | **Temp** | 1 | -0.2322 | 0.1082 | 4.6008 | 0.0320 |      | **Odds Ratio Estimates** | | | | | --- | --- | --- | --- | | **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | | | **Temp** | 0.793 | 0.641 | 0.980 |      | **Hosmer and Lemeshow Goodness-of-Fit Test** | | | | --- | --- | --- | | **Chi-Square** | **DF** | **Pr > ChiSq** | | 9.7032 | 7 | 0.2060 | |

**/\* 3. Check for influential observations and outliers \*/**

**proc logistic data=shuttle**

**plots(only label)=(phat influence dpc);**

**model damage (event='YES') = temp;**

**title1 'Graphical checks for influential observations';**

**run;**

|  |  |
| --- | --- |
| |  | | --- | | ***Graphical checks for influential observations*** | |

**/\* outlier check using simulated envelope macro \*/**

**%macro simEnv(dataset, response, predictors, N); proc ...**

**data shuttle; set shuttle;**

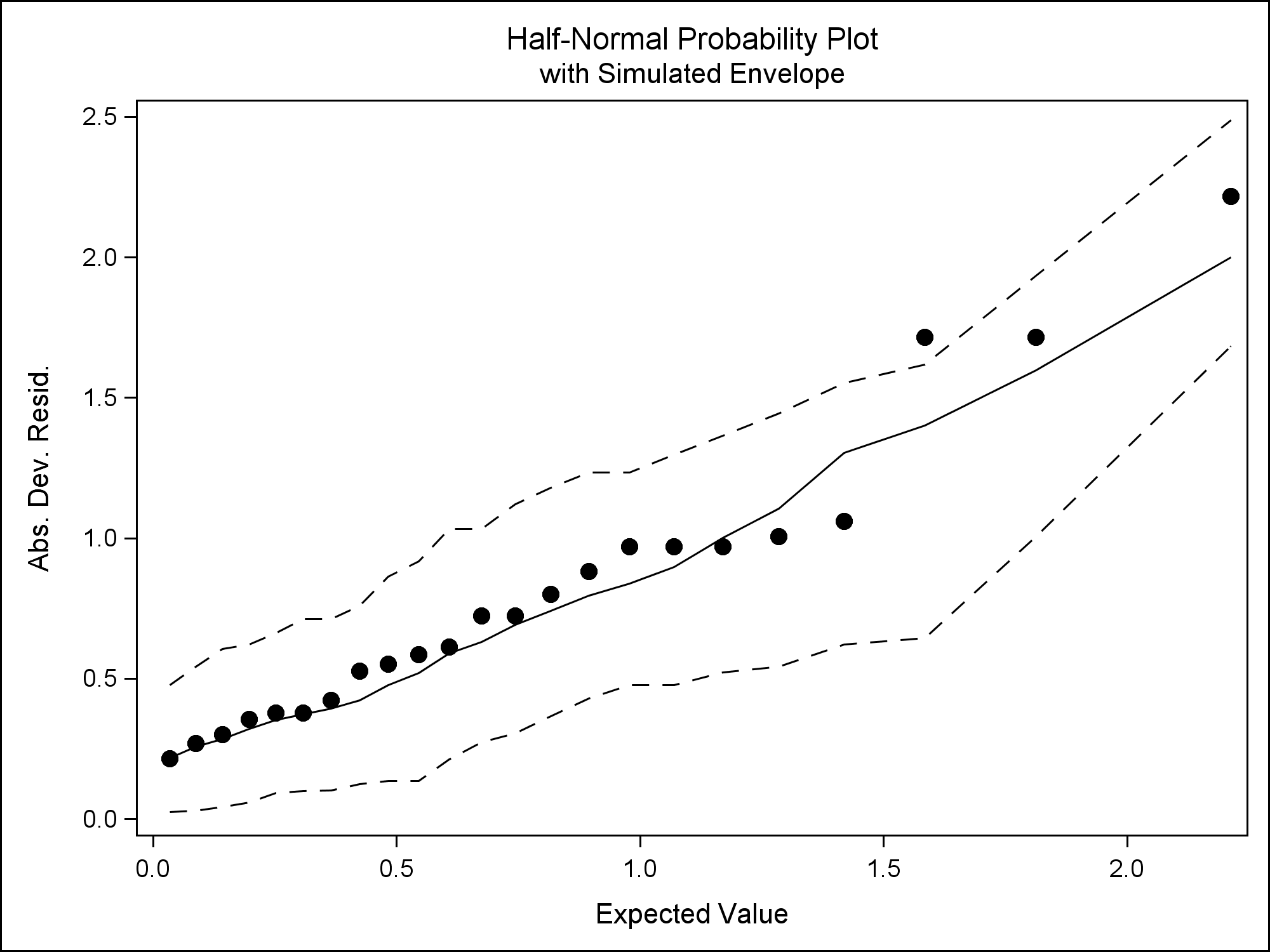
**damY=(damage='YES');**

**run;**

**%*simEnv*(dataset = shuttle, response = damY,**

**predictors = temp, N=23);**

**run;**



**data shuttle; set shuttle;**

**obs = \_n\_;**

**infl = (obs = 17 | obs = 20 | obs = 21);**

**run;**

**proc print data=shuttle;**

**where infl=1;**

**var Flight Temp Damage;**

**title1 'Suspect Observations';**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Suspect Observations*** |  | **Obs** | **Flight** | **Temp** | **Damage** | | --- | --- | --- | --- | | **17** | STS2 | 70 | YES | | **20** | STS41D | 70 | YES | | **21** | STS61A | 75 | YES | |

**proc sgplot data=shuttle;**

**where infl ne 1;**

**scatter y=damage x=temp /**

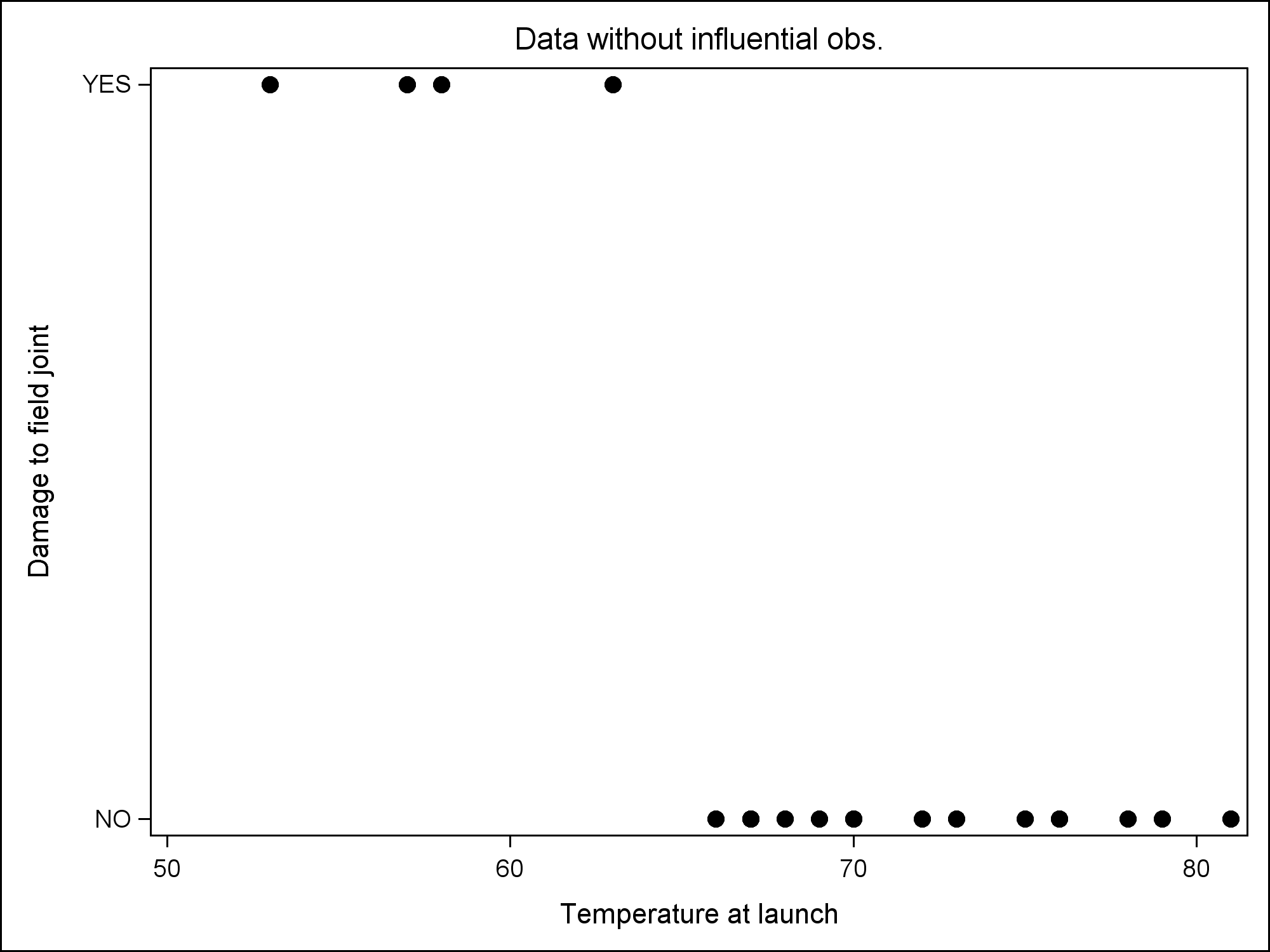
**markerattrs=(symbol=CIRCLEFILLED size=2pt);**

**xaxis label='Temperature at launch';**

**yaxis label='Damage to field joint';**

**title1 'Data without influential obs.';**

**run;**



**/\* Try refitting without these three points**

**(just for example here) \*/**

**data shuttle1; set shuttle;**

**if flight ne 'STS2' & flight ne 'STS41D'**

**& flight ne 'STS61A';**

**proc logistic data=shuttle1 plots(only)=(effect);**

**model damage(event='YES') = temp;**

**title1 'Logistic Regression with Separation of Points';**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Logistic Regression with Separation of Points*** |   Probability modeled is Damage='YES'.     | **Model Convergence Status** | | --- | | Complete separation of data points detected. |   Warning: The maximum likelihood estimate does not exist.   |  |  | | --- | --- | | Warning: | The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable. |        | **Analysis of Maximum Likelihood Estimates** | | | | | | | --- | --- | --- | --- | --- | --- | | **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** | | **Intercept** | 1 | 214.5 | 350.2 | 0.3752 | 0.5402 | | **Temp** | 1 | -3.3232 | 5.3974 | 0.3791 | 0.5381 |      | **Odds Ratio Estimates** | | | | | --- | --- | --- | --- | | **Effect** | **Point Estimate** | **95% Wald Confidence Limits** | | | **Temp** | 0.036 | <0.001 | >999.999 | |

**/\* How to deal with complete separation of points?**

**Rather than maximum likelihood, use penalized maximum**

**likelihood. Solution fairly recent: Heinze, G. &**

**Schemper, M. (2002). A solution to the problem of**

**separation in logistic regression. Statistics in**

**Medicine 21, 2409–2419. Convenient implementation even**

**more recent -- SAS 9.2 or later: FIRTH option in PROC**

**LOGISTIC**

**\*/**

**proc logistic data=shuttle1 plots(only)=(effect);**

**model damage(event='YES') = temp / firth**

**clparm=pl clodds=pl; /\* Note PL for profile-likelihood,**

**which is more accurate (likelihood ratio-based)**

**than WALD (asymptotic normal approx.) for**

**small sample sizes \*/**

**title1 'Logistic Regression with Separation of Points';**

**title2 '(using FIRTH option for pen. max. lik.)';**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Logistic Regression with Separation of Points*** | | ***(using FIRTH option for pen. max. lik.)*** |   Probability modeled is Damage='YES'.     | **Model Convergence Status** | | --- | | Convergence criterion (GCONV=1E-8) satisfied. |      | **Testing Global Null Hypothesis: BETA=0** | | | | | --- | --- | --- | --- | | **Test** | **Chi-Square** | **DF** | **Pr > ChiSq** | | **Likelihood Ratio** | 13.0618 | 1 | 0.0003 | | **Score** | 11.8077 | 1 | 0.0006 | | **Wald** | 3.6517 | 1 | 0.0560 |      | **Analysis of Maximum Likelihood Estimates** | | | | | | | --- | --- | --- | --- | --- | --- | | **Parameter** | **DF** | **Estimate** | **Standard Error** | **Wald Chi-Square** | **Pr > ChiSq** | | **Intercept** | 1 | 30.4123 | 16.5141 | 3.3915 | 0.0655 | | **Temp** | 1 | -0.4833 | 0.2529 | 3.6517 | 0.0560 |      | **Parameter Estimates and Profile-Likelihood Confidence Intervals** | | | | | --- | --- | --- | --- | | **Parameter** | **Estimate** | **95% Confidence Limits** | | | **Intercept** | 30.4123 | 8.4220 | 162.1 | | **Temp** | -0.4833 | -2.2770 | -0.1448 |      | **Odds Ratio Estimates and Profile-Likelihood Confidence Intervals** | | | | | | --- | --- | --- | --- | --- | | **Effect** | **Unit** | **Estimate** | **95% Confidence Limits** | | | **Temp** | 1.0000 | 0.617 | 0.103 | 0.865 | |

**/\* 4. Calculate the probability of damage at**

**temperature 31 \*/**

**data comp; phat = 1/(1+exp(-(30.4123-0.4833\*31)));**

**proc print data=comp;**

**var phat;**

**title1 'Prob. of Damage at Temp=31';**

**run;**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Prob. of Damage at Temp=31*** |  | **Obs** | **phat** | | --- | --- | | **1** | 1.00000 | |

**/\* 5. How is logistic regression different from ANOVA? \*/**

**proc reg data=shuttle;**

**model temp = damY;**

**title1 'ANOVA';**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***ANOVA*** |  | **Analysis of Variance** | | | | | | | --- | --- | --- | --- | --- | --- | | **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** | | **Model** | 1 | 344.47360 | 344.47360 | 9.63 | 0.0054 | | **Error** | 21 | 751.17857 | 35.77041 |  |  | | **Corrected Total** | 22 | 1095.65217 |  |  |  |      | **Parameter Estimates** | | | | | | | --- | --- | --- | --- | --- | --- | | **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | | **Intercept** | **1** | 72.12500 | 1.49521 | 48.24 | <.0001 | | **damY** | **1** | -8.41071 | 2.71030 | -3.10 | 0.0054 | | |
|  |  |