Stats Inference - Assignment 4 Analysis / Plots

All Disasters

1) lmboot(All.Disasters.Count ~ delta.temp*Year - Year, NOAAGISSWD)

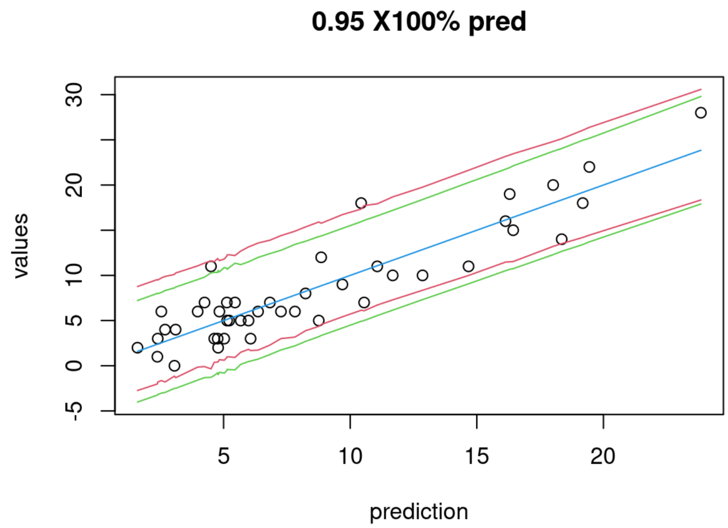
```
$coef
  (Intercept)      delta.temp delta.temp:Year
    7.0945835   -1477.3653920       0.7373628

$coef.point
    (Intercept) delta.temp delta.temp:Year
2.5%     4.022585 -2009.6656       0.4842644
97.5%   10.854360  -965.3773       0.9994218

$simultaneous
    (Intercept) delta.temp delta.temp:Year
[1,]    1.818941 -2408.2693       0.3487642
[2,]   15.147936  -691.1777       1.1954488

$conf
[1] 0.95

$PRESS
[1] 349.8295
```

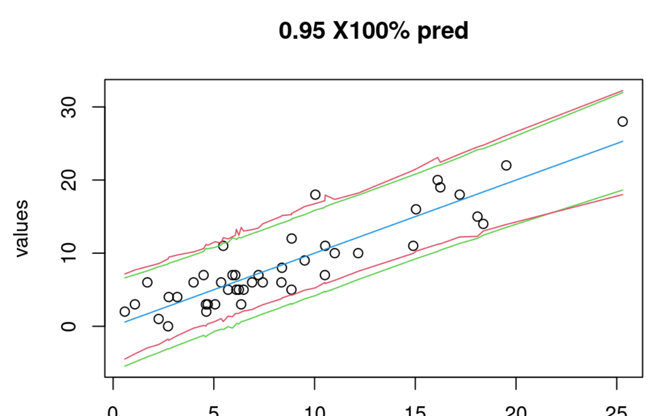


2) lmboot(All.Disasters.Count ~ I(delta.temp^2) + Year + delta.temp,NOAAGISSWD)

```
$coef
  (Intercept) I(delta.temp^2)          Year      delta.temp
 -701.1246857      28.7285652      0.3578963     -31.5661007

$coef.point
    (Intercept) I(delta.temp^2)      Year delta.temp
2.5%   -1137.1914        15.24162 0.1711554  -48.05608
97.5%   -333.4787        38.90234 0.5780951  -15.64465

$simultaneous
    (Intercept) I(delta.temp^2)       Year delta.temp
```



[1,] -1467.09027   1.596802 0.04848457 -75.597701
[2,]   -88.15436   52.213691 0.74595039 -3.641618

$conf
[1] 0.95

$PRESS
[1] 383.4326

Analysis: PRESS stands for Prediction Error of Sum of Squares, and typically, in linear regression models, the lower PRESS value indicates better predictive performance for a regression model on unseen data. For the All Disasters Count variable, we have two different plausible lmboot models that we are analyzing (code provided above), and we can see that the first model has a PRESS score of 349.8295 and the second model has a PRESS score of 383.4326. Because the first model has the lower press score, it is the better-suited regression model for All Disasters count. Analyzing the graph itself, we can see that the prediction model appears to be positively correlated, following an upward trend. Lower prediction values are advantageous because they prevent overfitting, which often tends to capture "noise" in training data rather than generalizing new data.

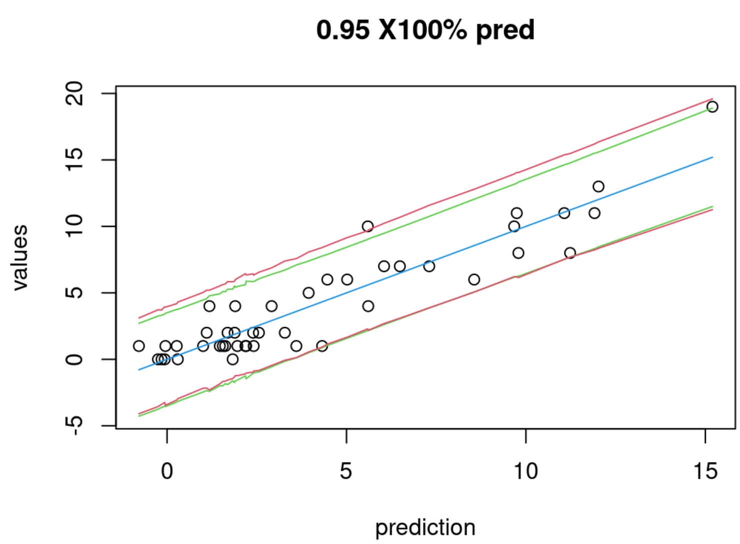
Severe Storm

1. **lmboot(Severe.Storm.Count ~ delta.temp*Year - Year,NOAAGISSWD)**

$coef
   (Intercept)   delta.temp delta.temp:Year
     3.5001463   -1097.0929693
0.5472525

$coef.point
     (Intercept) delta.temp delta.temp:Year
2.5%   1.809389 -1425.7908
0.4075687
97.5%   5.544866 -814.4526
0.7098020

$simultaneous
    (Intercept) delta.temp delta.temp:Year
[1,]   0.297717 -1663.0190   0.3277611



[2,] 8.473719 -651.6685 0.8266144

$conf
[1] 0.95

$PRESS
[1] 137.1605

2. **lmboot(Severe.Storm.Count ~ Year * I(delta.temp^2) - Year, NOAAGISS)**

$coef

(Intercept) I(delta.temp^2)

Year:I(delta.temp^2)

2.1536882 -1080.9807546

0.5402085

$coef.point

(Intercept) I(delta.temp^2)

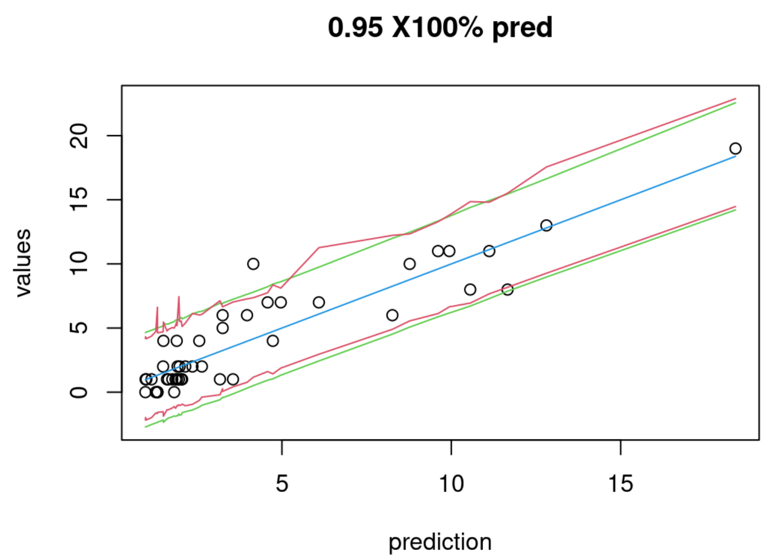
Year:I(delta.temp^2)

2.5% 1.170907 -1472.1001

0.3451501

97.5% 3.215336 -686.1084

0.7336645



$simultaneous

(Intercept) I(delta.temp^2) Year:I(delta.temp^2)

[1,] 0.6898658 -1807.8023 0.1965312

[2,] 4.3727934 -388.1616 0.8996514

$conf

[1] 0.95

$PRESS

[1] 148.1495

Analysis: Comparing the PRESS values for the Severe Storm variable lmboot plots, it is clear to see that the first plot has a lower PRESS value, meaning that it will be a better predictor of the given data in the NOAAGISSWD dataset, as well as unseen data. It will help avoid overfitting unnecessary data and create a plot from which we can make accurate conclusions. A model with lower PRESS is advantageous because it reduces the risk of overfitting, which occurs when it becomes too complex and starts to capture noise or random fluctuations in the training data rather than the underlying patterns. Overfitted models have a pattern of performing well on training data but fail to generalize new data, making them less reliable for more practical applications.

Wildfire

1. **logitboot(Wildfire.Count~delta.temp+I(Year^2)-1, NOAAGISSWD)**

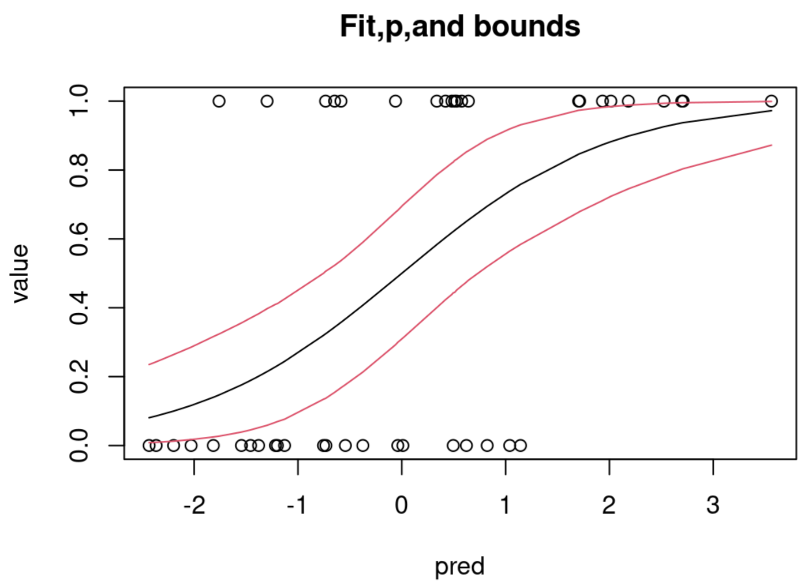
```
$alpha
[1] 0.05

$aic
[1] 49.50986

$coef
   delta.temp     I(Year^2)
 5.824457e+00 -7.950709e-07

$pointwiseCI
      delta.temp     I(Year^2)
2.5%    3.290233 -1.540879e-06
97.5%  11.063659 -4.098987e-07

$simultaneousCI
     delta.temp     I(Year^2)
[1,]   2.436985 -3.905683e-06
[2,]  25.229098 -2.410760e-07
```

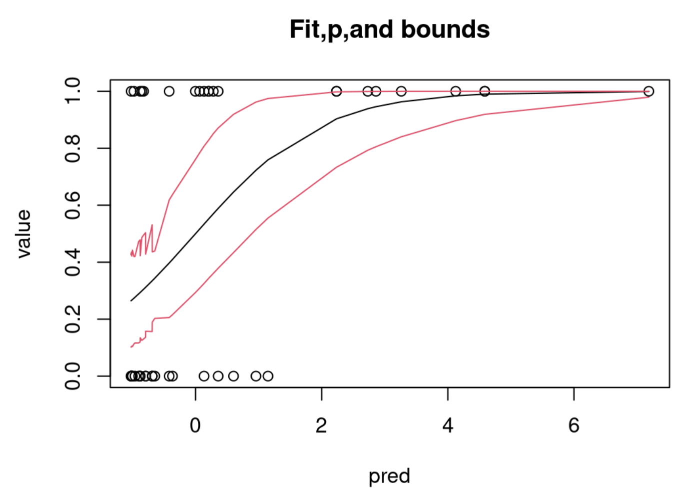


2. **logitboot(Wildfire.Count ~ delta.temp + I(delta.temp^2) - 1, NOAAGISSWD)**

```
$alpha
[1] 0.05

$aic
[1] 50.92199

$coef
     delta.temp I(delta.temp^2)
      -6.701762       10.978013
```



$pointwiseCI
    delta.temp I(delta.temp^2)
2.5% -14.225147   5.525354
97.5% -2.729194   24.454052

$simultaneousCI
    delta.temp I(delta.temp^2)
[1,] -24.870535   3.687148
[2,] -1.335035   43.611328

Analysis: AIC stands for Akaike Information Criterion, which is essential in making conclusions and determining the quality of a model for a given dataset. Generally, lower AIC values indicate a better model in terms of goodness of fit and complexity. There will be lower residuals as well. The AIC value for the first plot is 49.50986, and the AIC value for the second plot is 50.92199. Though the values are close, the AIC value for the first plot is lower than that of the second model. This indicates that the first model is better at being fit and complex, from which more accurate conclusions can be drawn about the wildfire count in the NOAAGISSWD data set.

logitboot(Wildfire.Count~delta.temp,NOAAGISS)

$alpha

[1] 0.05

$aic

[1] 49.40551

$coef

(Intercept) delta.temp

-3.117966   5.698753

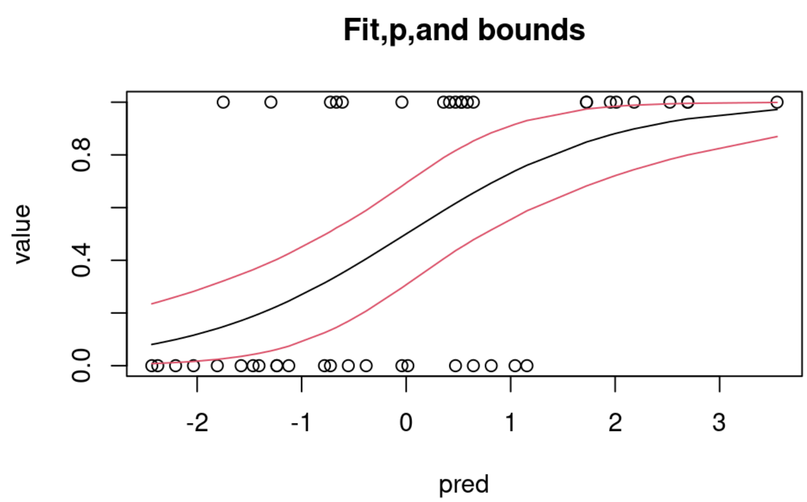
$pointwiseCI

(Intercept) delta.temp

2.5%   -6.088326   3.15400

97.5%   -1.599687   10.88487

$simultaneousCI



(Intercept) delta.temp

[1,] -16.558125 2.335633

[2,] -1.193882 30.365885

4. logitboot(Wildfire.Count~Year,NOAAGISS)

$alpha

[1] 0.05

$aic

[1] 47.78727

$coef

(Intercept) Year

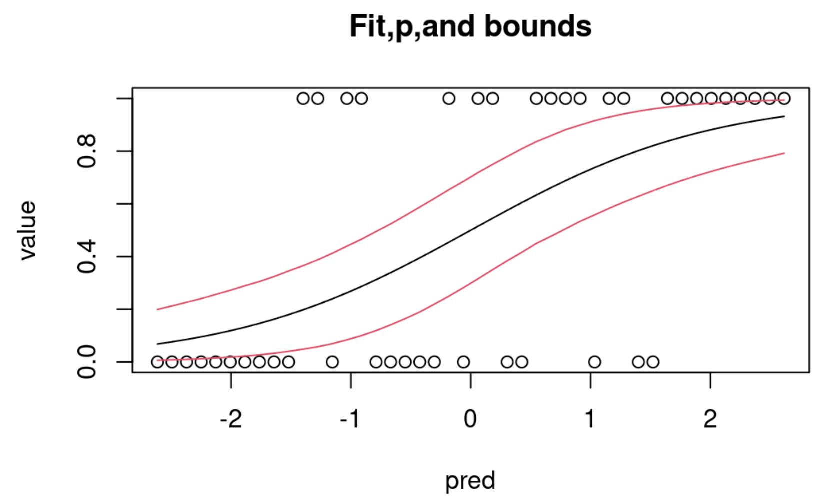
-243.4478775 0.1216327

$pointwiseCI

	(Intercept)	Year
2.5%	-449.7443	0.07024042
97.5%	-140.5188	0.22452427

$simultaneousCI

	(Intercept)	Year
[1,]	-1081.0364	0.05942513
[2,]	-119.1621	0.54015260



Analysis: The AIC value for the first plot representing the Wildfire.Count data, was 49.40551. The AIC value for the second plot was 47.78727. Since 47.78727 < 49.40551, we can conclude that the second plot is a better predictor of the data considering the variables goodness of fit as well as complexity. This way, the conclusions made about the patterns in the dataset are more reliable.

Drought

1. **logitboot(Drought.Count ~ Year, NOAAGISS, 500)**

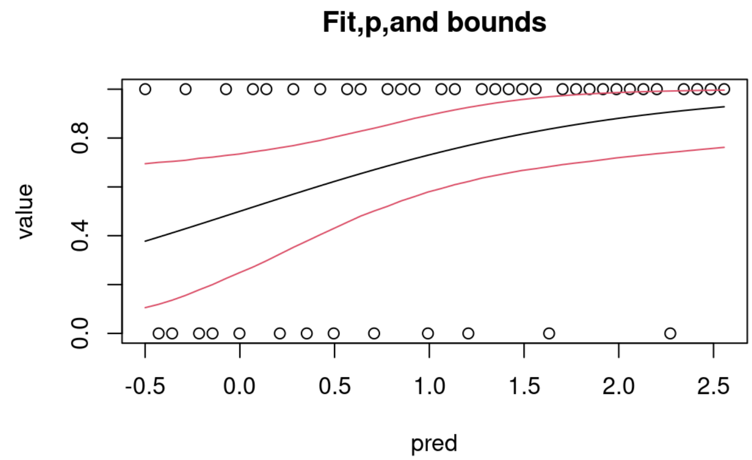
$alpha
[1] 0.05

$aic
[1] 51.15299

$coef
(Intercept) Year
-141.19883593 0.07106025

$pointwiseCI
(Intercept) Year
2.5% -323.88167 0.01565903
97.5% -30.66282 0.16266751

$simultaneousCI
(Intercept) Year
[1,] -694.40352 0.00776318
[2,] -14.33518 0.34983868

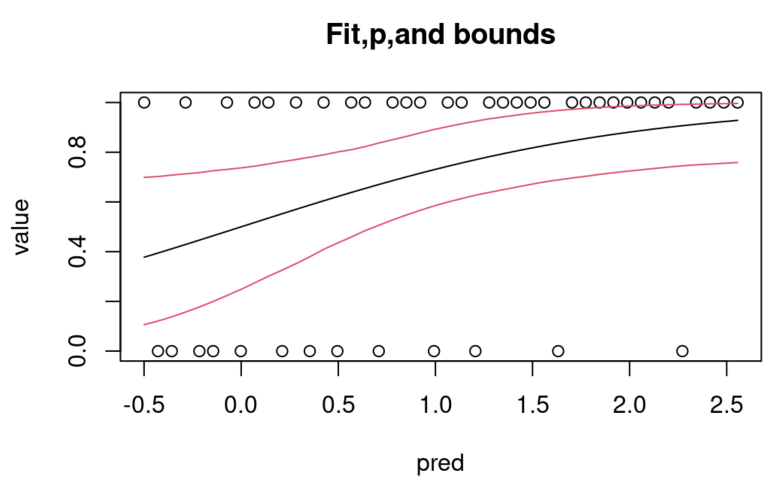


2. **logitboot(Drought.Count ~ Year, NOAAGISS)**

$alpha
[1] 0.05

$aic
[1] 51.15299

$coef
(Intercept) Year
-141.19883593 0.07106025



```
$pointwiseCI
      (Intercept)       Year
2.5%   -336.95313 0.01483237
97.5%   -28.79483 0.16915918

$simultaneousCI
       (Intercept)        Year
[1,] -1108.8423712 0.001020183
[2,]    -0.8200136 0.556546019
```

Analysis: After attempting 50+ models with bootstrapping at different values, using as.factor, and various different polynomial equations with the delta.temp and Year variables, we had found out that bootstrapping from no values to 500 is what gave us two plausible models with no zeros. Since both models use the same formula (Drought.Count ~ Year), they have identical AIC values. AIC comparison helps to compare models against each other, and in this case, it doesn't really matter since both models have the same formula. AIC can be helpful in differentiating between two different models. It's challenging to compare them using PRESS, as it's primarily a measure used for least-squares regression. For Model 1, we have the $aic value as 51.15299. The "lmboot" function from the regboot.pck package provides "aic" under the logitboot() function. For Model 2, we had calculated the AIC using the "AIC()" function for the glm object.