Directions:

In the weather disasters data:

Model the Y variables drought and wild fire as binomial functions using the logit link and consider delta.temp, year, and delta.temp*year as the x variables (delta.temp*Year includes all 3).

Use Akaike score and deviance differences to look at which is best model for each of the two.

Choose best models for each of drought and wildfire.

---

1. GLM() models
    a. glm() stands for generalized linear model. It is a linear regression method that is used for response variables that do not follow a normal distribution. They are useful when the model between the two variables are not particularly linear.
2. 'Akaike' score
    a. The Akaike score is a measure of the quality of a statistical model for a set of data. The AIC is based on the fact that a simpler model that represents a set of data well enough is preferable to a more intricate model. The AIC score can be used to compare models even if they have a different set of data.

In the context of GLM and Akaike scores, the AIC score can be used to compare different models (even between different variables). Usually, a lower AIC value is chosen to be the better model for the data as it balances goodness of fit and the simplicity of the model. In our case, we would be able to use the AIC score to compare GLM models with different sets of variables, and those that have a lower AIC score should be preferred.

3. Other factors to Consider
    a. Best models for drought and wildfire cannot be chosen just on the basis of the lowest Akaike score. Instead, we have to consider other factors that play into the decision of choosing the best model in linear regression. Two such factors that we will also be considering would be Degrees of Freedom and pchisq().
    b. pchisq(q, df)
        i. This is the cumulative distribution function (CDF) of the chi-squared distribution.

ii. q: This represents the quantile value which is the value in which we want the CDF to be evaluated
iii. df: This represents the degrees of freedom. The degrees of freedom is what determines the shape and/or behavior of the distribution.

Drought.Count vs. Year, delta.temp, Year * delta.temp, and Year + delta.temp

```
# fit a binomial logit model with drought as the response and delta.temp, year, and
#delta.temp*year as predictors
drought_model_multiplicative <- glm(Drought.Count ~ delta.temp * Year, data = weather_data,
family = "binomial")

# fit a second model without the interaction term
drought_model_additive <- glm(Drought.Count ~ delta.temp + Year, data = weather_data, family
= "binomial")

AIC_drought_interaction <- AIC(drought_model_interaction)
AIC_drought_additive <- AIC(drought_model_additive)

#output: AIC for Drought.Count ~ delta.temp + Year: 52.99482
#output: AIC for Drought.Count ~ delta.temp * Year: 54.52863
```

In this case, we can see that Drought.Count ~ delta.temp + Year has a lower AIC score. However, this is not the only thing we should rely on. We should use the pchisq() function to double check this before checking which model is the better fit.

```
res_dev2<-46.995
res_dev1<-46.529
devdiff.Drought.Count<res_dev2 -  res_dev1
1-pchisq(devdiff.Drought.Count,1)'

#the output of this is approximately 0.4948
```

Since this value is greater than 0.01, we can confidently say that Year + delta.temp is the better model for linear regression and that Year * delta.temp is not as good of a representation of the dataset.

GLM Model for Drought.Count ~ Year + delta.temp

glmDeltaTempPlusYear <- glm(formula = Drought.Count~Year + delta.temp,family=binomial(link=logit),data = weather)

summary(glmDeltaTempPlusYear)

```
Call:
glm(formula = Drought.Count ~ Year + delta.temp, family = binomial(link = logit),
    data = weather)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -87.14359  148.59803  -0.586    0.558
Year          0.04364    0.07521   0.580    0.562
delta.temp    1.49853    3.78420   0.396    0.692

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 53.413  on 43  degrees of freedom
Residual deviance: 46.995  on 41  degrees of freedom
AIC: 52.995

Number of Fisher Scoring iterations: 4
```

GLM Model for Drought.Count ~ Year * delta.temp

glmDeltaTempTimesYear <- glm(formula = Drought.Count~Year*delta.temp, family = binomial(link=logit), data = weather)

summary(glmDeltaTempTimesYear)

```
Call:
glm(formula = Drought.Count ~ Year * delta.temp, family = binomial(link = logit),
    data = new_weather)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -164.34308  184.34723  -0.891    0.373
Year               0.08222    0.09286   0.885    0.376
delta.temp       181.03552  259.47423   0.698    0.485
Year:delta.temp   -0.08949    0.12922  -0.693    0.489

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 53.413  on 43  degrees of freedom
Residual deviance: 46.529  on 40  degrees of freedom
AIC: 54.529

Number of Fisher Scoring iterations: 4
```

GLM Model for Drought.Count ~ delta.temp

glmDroughtAndDeltaTemp <- glm(formula = Drought.Count~delta.temp, family = binomial(link = logit), data = weather)

summary(glmDroughtAndDeltaTemp)

```
Call:
glm(formula = Drought.Count ~ delta.temp, family = binomial(link = logit),
    data = new_weather)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9172     0.8204  -1.118   0.2636
delta.temp    3.5189     1.5922   2.210   0.0271 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 53.413  on 43  degrees of freedom
Residual deviance: 47.335  on 42  degrees of freedom
AIC: 51.335

Number of Fisher Scoring iterations: 4
```

GLM Model for Drought.Count ~ Year

glmDroughtAndYear <- glm(formula = Drought.Count~Year, family = binomial(link = logit), data = weather)

summary(glmDroughtAndYear)

```
Call:
glm(formula = Drought.Count ~ Year, family = binomial(link = logit),
    data = new_weather)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -141.19884   61.85816  -2.283   0.0225 *
Year           0.07106    0.03097   2.294   0.0218 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 53.413  on 43  degrees of freedom
Residual deviance: 47.153  on 42  degrees of freedom
AIC: 51.153

Number of Fisher Scoring iterations: 4
```

1. Model with Drought.Count ~ Year + delta.temp:
    a. AIC: 52.995
    b. Residual deviance: 46.995
2. Model with Drought.Count ~ Year * delta.temp
    a. AIC: 54.529
    b. Residual deviance: 46.529
3. Model with Drought.Count ~ delta.temp:
    a. AIC: 51.335
    b. Residual deviance: 47.335
4. Model with Drought.Count ~ Year:
    a. AIC: 51.153
    b. Residual deviance: 47.153

We can start off by analyzing the last two models first, the one with Drought.Count vs delta.temp and Drought.Count vs Year. Since these models are made by using independent variables, we do not need to use the pchisq() function on it and can just compare the models using the AIC values. In this case, since Drought.Count ~ Year has a lower AIC value (51.153 < 51.335), we can say that Drought.Count ~ Year is the better model for Drought.Count.

If we want to look at the deviance values, we are looking for the lowest deviance as it would be the model that best explains the variability in the data. Drought.Count ~ Year * delta.temp has the lowest residual deviance and thus, we can say that it accounts for the most variability in the data sets and can be considered the best fit for the data. However, just because the residual deviance is low does not mean it is the best model. The model has a high AIC score and thus, we cannot say that residual deviance is the only thing that creates a good model for the best representation of the data. If we want to consider both the AIC value and the residual deviances values, the model with Drought.Count ~ Year seems to be the best one, as it has the lowest AIC score as well as a somewhat decent residual difference.

Based solely on the AIC values, Model with Drought.Count ~ Year has the lowest AIC, which makes it seem like it may be the best option for the model while considering all the other ones. However, Drought.Count ~ delta.temp has an AIC value that is extremely close to the model of Drought.Count ~ Year. Thus, the two models can both be considered good, depending on the context.

In conclusion, after taking all factors into account (residual deviation, AIC values, etc.), Drought.Count ~ Year may be the preferred model that we should consider.

Wildfire.Count vs. Year, delta.temp, Year * delta.temp, and Year + delta.temp

GLM Model for Wildfire.Count and Year + delta.temp

```
glmWildfireDeltaTempPlusYear <- glm(formula = Wildfire.Count~Year +
delta.temp,family=binomial(link=logit),data = weather)

summary(glmWildfireDeltaTempPlusYear)
```

```
Call:
glm(formula = Wildfire.Count ~ Year + delta.temp, family = binomial(link = logit),
    data = weather)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -212.80370  169.82370  -1.253    0.210
Year           0.10610    0.08587   1.236    0.217
delta.temp     0.81269    4.12303   0.197    0.844

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 60.997  on 43  degrees of freedom
Residual deviance: 43.748  on 41  degrees of freedom
AIC: 49.748

Number of Fisher Scoring iterations: 4
```

GLM Model for Wildire.Count and Year * delta.temp

glmWildfireDeltaTempTimesYear <- glm(formula = Wildfire.Count~Year*delta.temp, family = binomial(link=logit), data = weather)

summary(glmWildfireDeltaTempTimesYear)

```
Call:
glm(formula = Wildfire.Count ~ Year * delta.temp, family = binomial(link = logit),
    data = weather)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -188.74474  218.72821  -0.863    0.388
Year               0.09409    0.11004   0.855    0.393
delta.temp       -52.78261  306.29915  -0.172    0.863
Year:delta.temp    0.02670    0.15263   0.175    0.861

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 60.997  on 43  degrees of freedom
Residual deviance: 43.717  on 40  degrees of freedom
AIC: 51.717

Number of Fisher Scoring iterations: 4
```

GLM Model for Wildfire.Count vs delta.temp

glmWildfireDeltaTemp <- glm(formula = Wildfire.Count~delta.temp, family = binomial(link = logit), data = weather)

summary(glmWildfireDeltaTemp)

```
Call:
glm(formula = Wildfire.Count ~ delta.temp, family = binomial(link = logit),
    data = weather)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.118      1.033  -3.020  0.00253 **
delta.temp     5.699      1.799   3.168  0.00154 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 60.997  on 43  degrees of freedom
Residual deviance: 45.406  on 42  degrees of freedom
AIC: 49.406

Number of Fisher Scoring iterations: 4
```

GLM Model for Wildfire.Count vs Year

glmWildfireAndYear <- glm(formula = Wildfire.Count~Year, family = binomial(link = logit), data = weather)

summary(glmWildfireAndYear)

```
Call:
glm(formula = Wildfire.Count ~ Year, family = binomial(link = logit),
    data = weather)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -243.44788   72.16830  -3.373 0.000743 ***
Year           0.12163    0.03606   3.373 0.000743 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 60.997  on 43  degrees of freedom
Residual deviance: 43.787  on 42  degrees of freedom
AIC: 47.787

Number of Fisher Scoring iterations: 4
```

1. Model with Wildfire.Count ~ Year + delta.temp:
    a. AIC: 49.748
    b. Residual deviance: 43.748
2. Model with Wildfire.Count ~ Year * delta.temp:
    a. AIC: 51.717
    b. Residual deviance: 43.717
3. Model with Wildfire.Count ~ delta.temp:
    a. AIC: 49.406
    b. Residual deviance: 45.406
4. Model with Wildfire.Count ~ Year:
    a. AIC: 47.787
    b. Residual deviance: 43.787

The model Wildfire.Count ~ Year has the lowest AIC value among the models under consideration, indicating that it might be the superior model. In terms of residual deviation, Wildfire.Count ~ Year * delta.temp has the lowest residual deviation, but you also have to consider the AIC value for that model. Since the AIC value is not the lowest, it would not inevitably imply that this model is superior to others. Still, Wildfire.Count ~ Year appears to be

the best model (due to the lowest AIC score), suggesting it is the best among the other models when taking into account both the AIC and the deviation values.

When we look at models with the deviance, which describes the variability in the data with the fewest unexplained or residual components, we can see that the model with Wildfire.Count ~ Year * delta.temp has the lowest residual deviance. That shows how the model is a good fit for the data as it uses the explanatory variables to account for the variability in the number of wildfire counts. However, its high AIC score of 51.717 shows that a model with a lower residual deviance does not necessarily translate to being a suitable fit for linearization.

If we want to consider only the low AIC score, Wildfire.Count ~ Year is the strongest predictor for linearization. We do not need to use the pchisq() function because Year is the only variable that we are considering. When we compare the other AIC values to this one, we can see that none of the other models have an AIC value that is relatively close enough to the AIC value for the Wildfire.Count ~ Year model. This shows us that for Wildfire, the Year model would be the best fit.

In conclusion, for both Drought and Wildfire, the model with Year as the variable seemed to return models with the lowest AIC scores. Thus, we can conclude that the variable “Year” seems to be a valuable tool in determining the optimal linear regression model and it would be advantageous to include “Year” in our models to accurately capture the trends in the dataset.