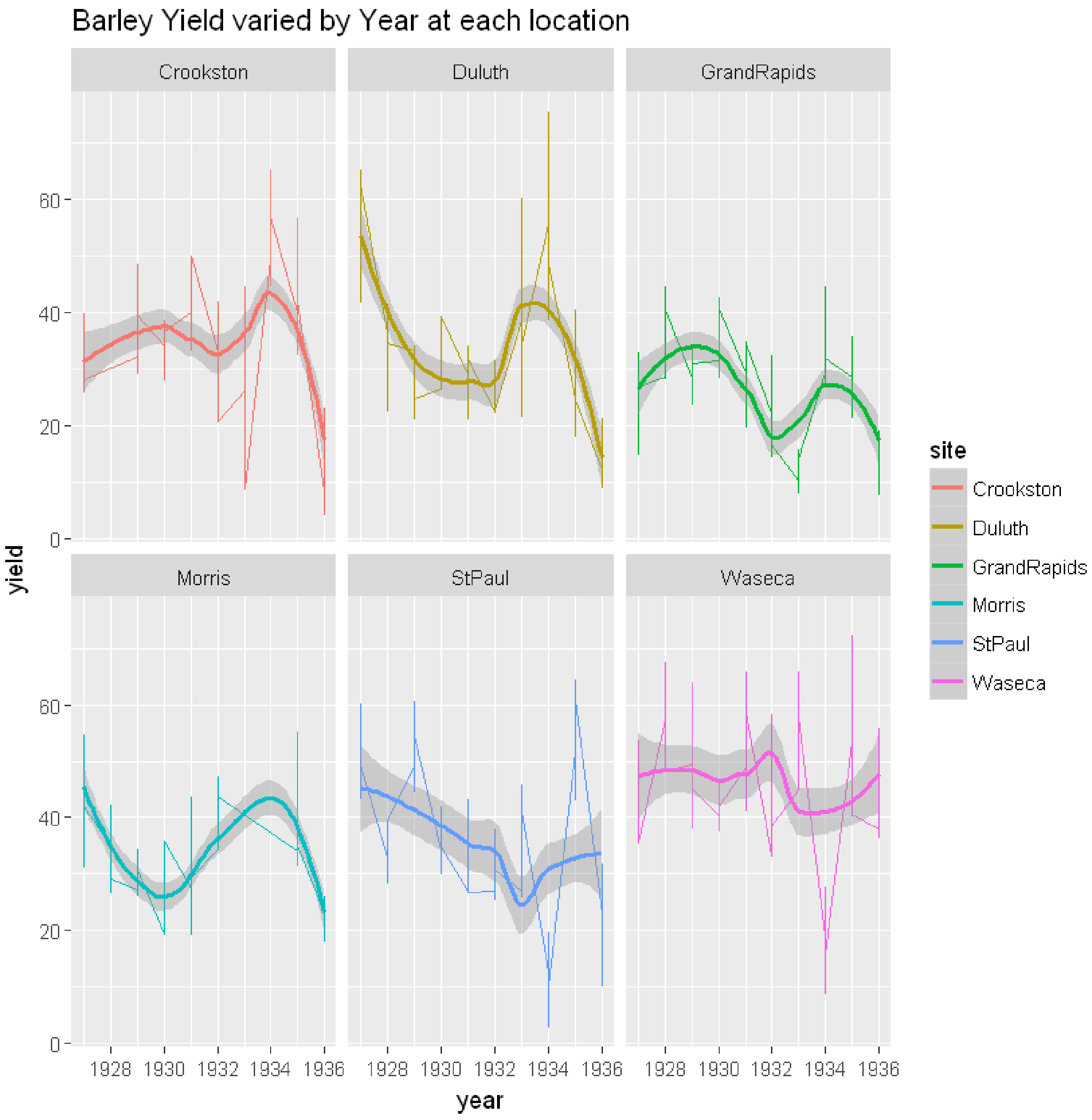
```
In [6]: df = read.table("minnesota.barley.yield.txt",header = T)
        library(ggplot2)
```

Q1

```
In [20]: ggplot(df,aes(x=year,y=yield,color = site)) + geom_line() + geom_smooth()+facet_wrap(~site,ncol=3) +
         ggtitle("Barley Yield varied by Year at each location")
```

```
`geom_smooth()` using method = 'loess'
```



The trend in barley yield is not identical at all the sites but we could identify three distinct trends where each trend is similar at couple of sites.

The rise and fall in the barley yield appears similar at (Crookston & GrandRapids), (Duluth & Morris) and (StPaul & Waseca).

At the first couple of sites, the barley yield reached a local minimum of 30 bushels at Crookston and 18 bushels at GrandRapis in the year 1932. The yield gradually surges to a peak in the year 1934.

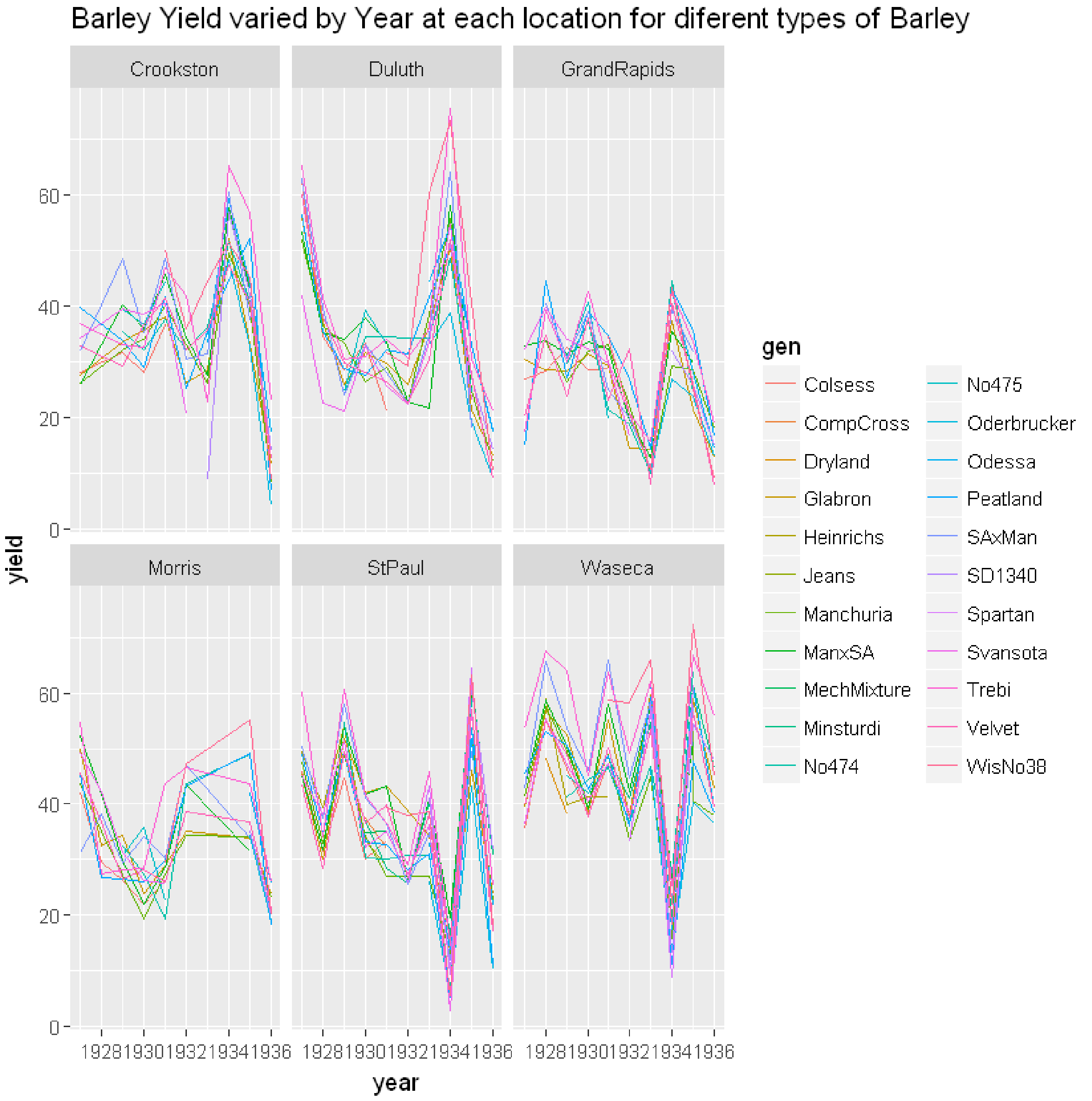
At the second couple of sites, the trend in yield appears like a similar curve except that the yeild starts dipping from 1926 which picks up again in the year 1932 but in the case of Morris the yield reaches a local minimum in the year 1930 which gradually increases to a maximum in the year 1934.

At the last couple of sites, the trend is more or less similar except for a slight contrast in the period of 1932 to 1934 at StPaul and 1931 to 1933 at Waseca. StPaul experiences a sharp fall in the barley yield to 25 bushels in the year 1933 which is otherwise characterized by a steady fall till 1932. the yield again starts to increase steadily from 1933. On a contrasting note, Waseca experiences a steady increase reaching its peak in the year 1932 and reaches its minimum in 1933.

To sum up, the time between 1932 to 1934 is the period when major changes appear to happen in the barley production at all the sites.

Q2

In [51]:
```
ggplot(df,aes(x=year,y=yield,color = gen)) + geom_line()+facet_wrap(~site,ncol
=3) +
ggtitle("Barley Yield varied by Year at each location for diferent types of Ba
rley")
```



As we can observe that there is some missing data for some type of Barley.We can interpret that as at that particular site they have gone for a different crop or the data is missing for that particular time period.

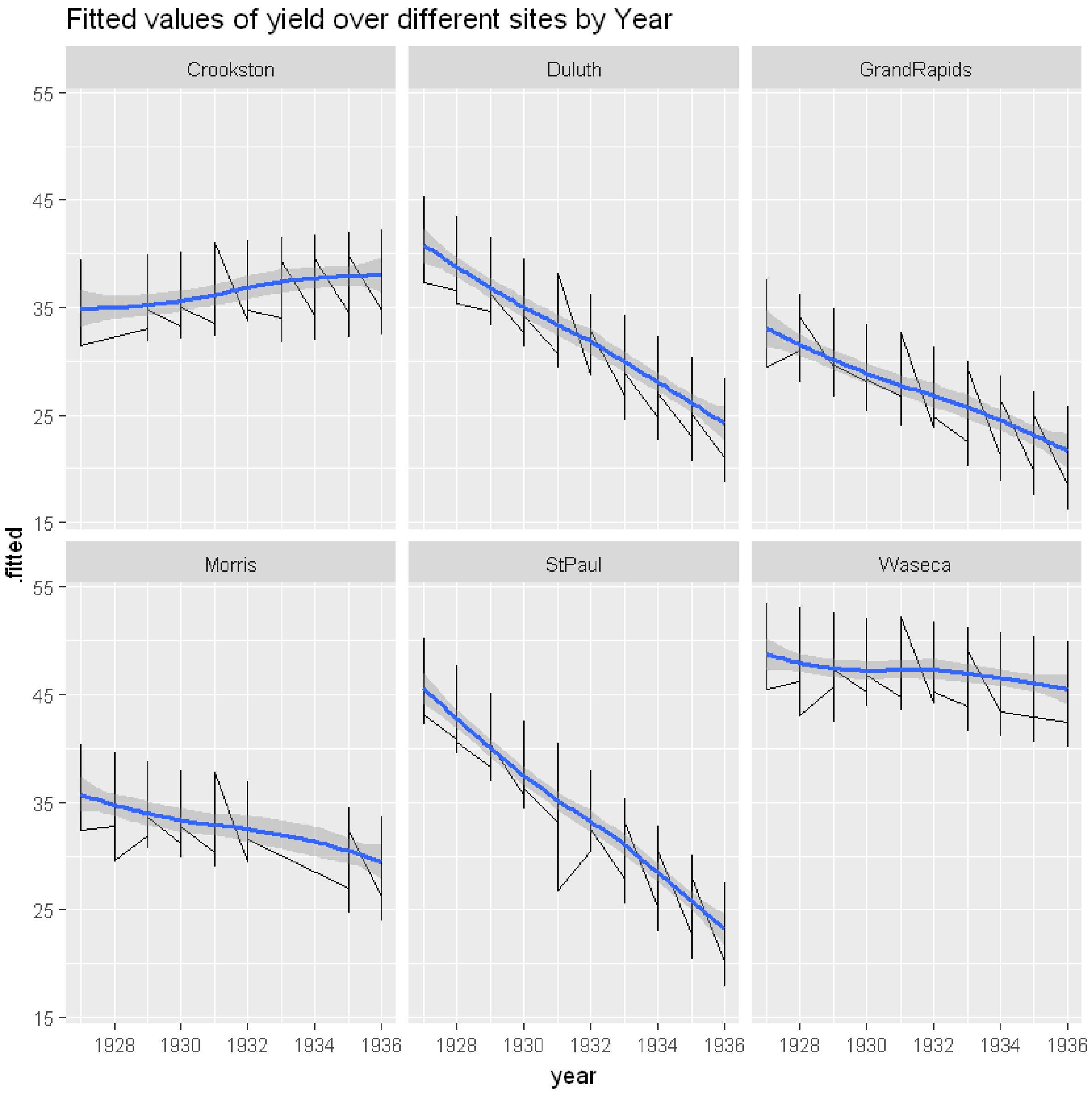
We can also say that the the direction of change of yield for different types of Barley is same with in a site over the time. So, for simplicity we can group and aggregate them together.

we are considering only year and site interaction with gen as yield is almost uniform for all the varieties of Barley. We are going for robust linear model as this will take care of all the outliers.

Q3

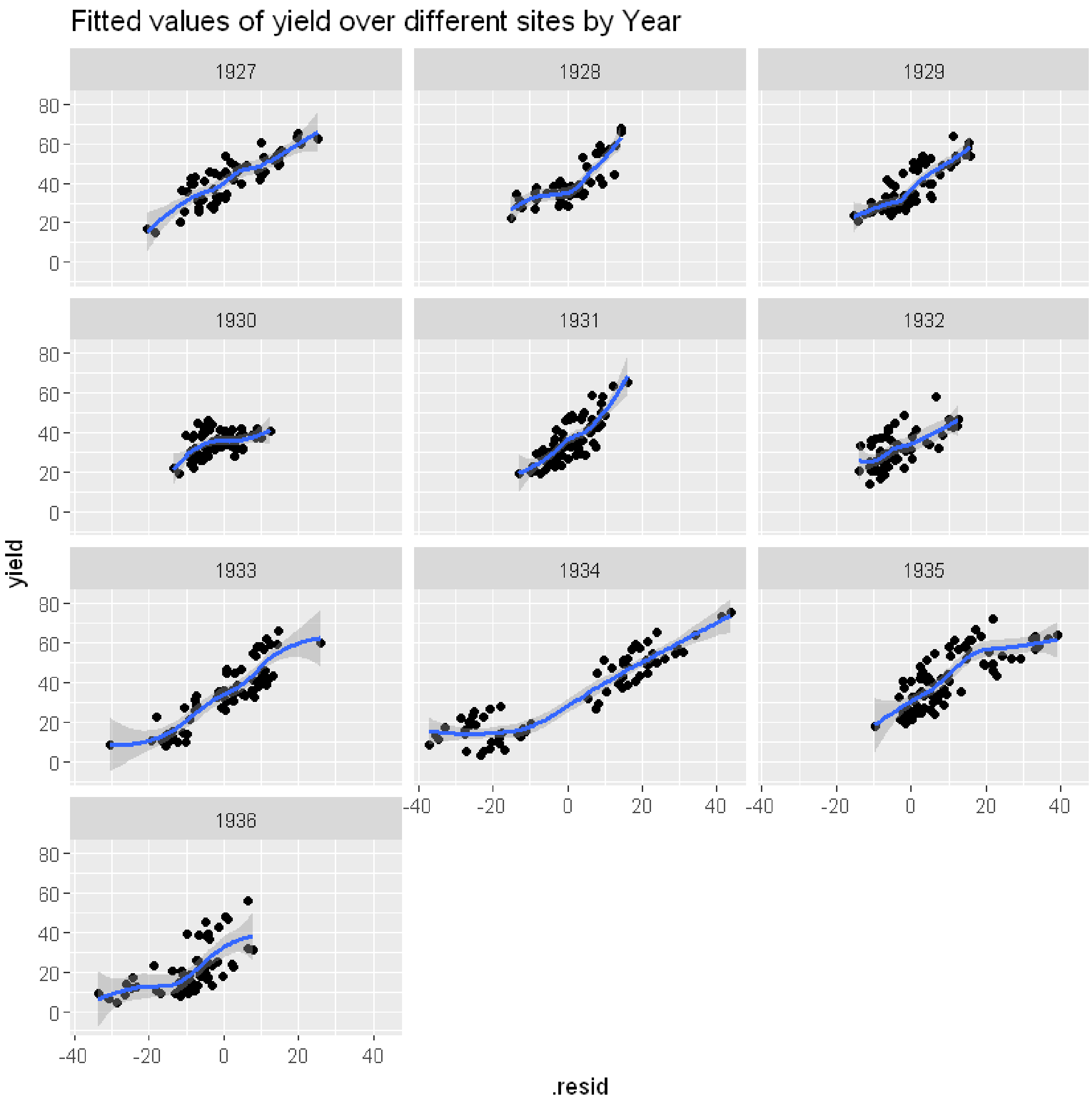
In [32]:
```
library(MASS)
df.rlm = rlm(yield ~ gen+(year*site), psi = psi.bisquare,data = df)

library(broom)
df.rlm.au = augment(df.rlm)

ggplot(df.rlm.au, aes(y= .fitted, x= year)) + geom_line()+ geom_smooth(span
=0.75)+
  facet_wrap(~site,ncol=3) + ggtitle("Fitted values of yield over different
 sites by Year")
```

```
`geom_smooth()` using method = 'loess'
```



In [43]:
```
ggplot(df.rlm.au, aes(y= yield, x= .resid)) + geom_point()+ geom_smooth()+
  facet_wrap(~year,ncol=3) + ggtitle("Fitted values of yield over different si
tes by Year")
```

```
`geom_smooth()` using method = 'loess'
```



When we look at the yield values for different residuals there is pattern difference from 1930-1931-1932 and for the years before and after that the increase in the trend is same. So, this will explain that there is an anamoly during 1931 and 1932. We might think that the yield trend at 1931-32 is an anamoly because of drought or something or they have made a mistake while recording the data. Let's check this by interchanging the values of 1931 and 1932.

Interchanging the values

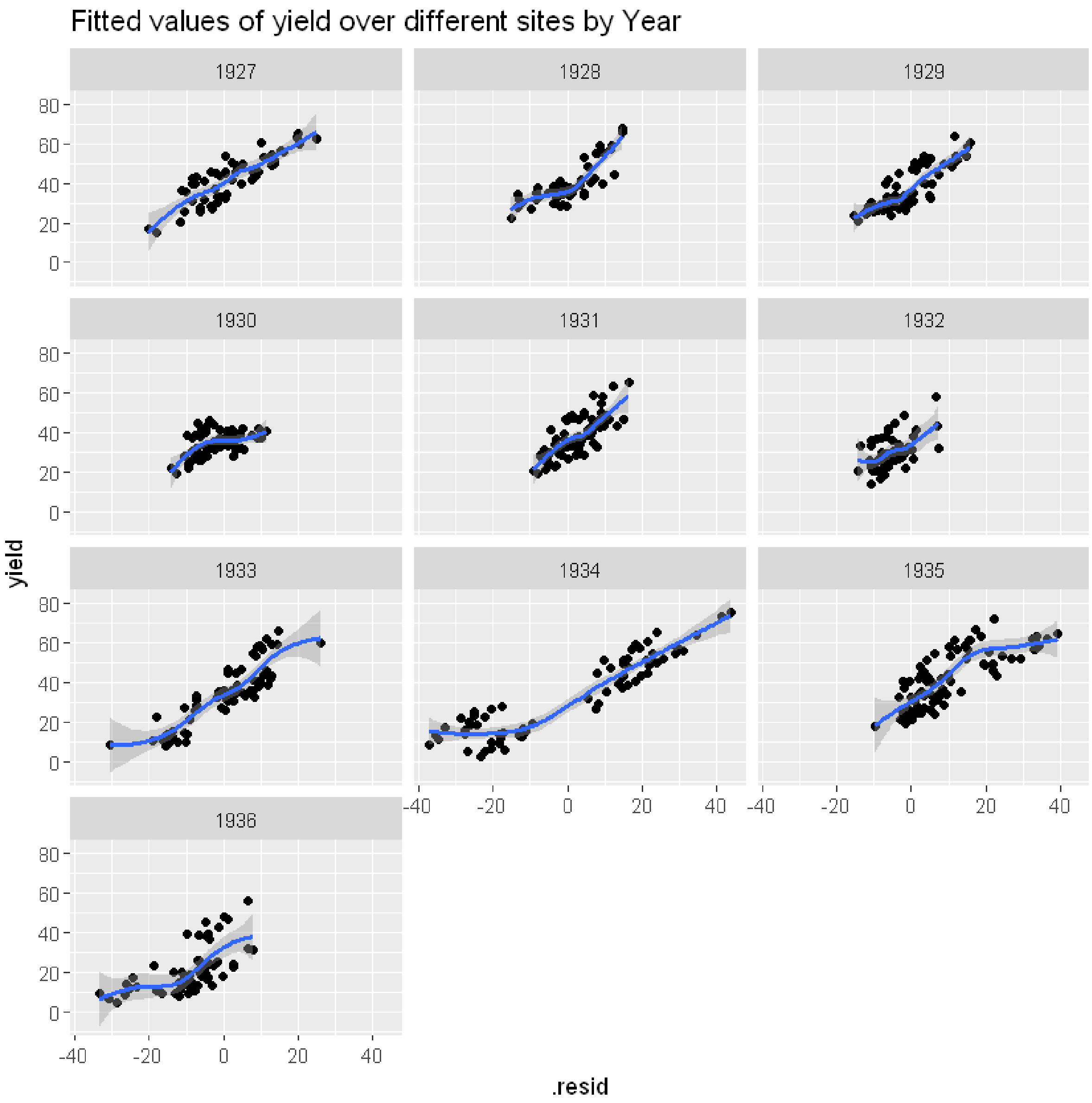
In [48]:

```
barley = df
morris1932fixed = barley$yield[barley$site == "Morris" & barley$year == 193
1]
morris1931fixed = barley$yield[barley$site == "Morris" & barley$year == 193
2]
barley.fixed = barley
barley.fixed$yield[barley$site == "Morris" & barley$year == 1932] = morris1
932fixed
barley.fixed$yield[barley$site == "Morris" & barley$year == 1931] = morris1
931fixed

barley.rlm = rlm(yield ~ gen + year * site, psi = psi.bisquare, data = barl
ey.fixed)

barley.rlm.au = augment(barley.rlm)
```

```
Warning message in barley.fixed$yield[barley$site == "Morris" & barley$year =
= 1932] = morris1932fixed:
"number of items to replace is not a multiple of replacement length"Warning m
essage in barley.fixed$yield[barley$site == "Morris" & barley$year == 1931] =
morris1931fixed:
"number of items to replace is not a multiple of replacement length"
```

```
In [49]: ggplot(barley.rlm.au, aes(y= yield, x= .resid)) + geom_point()+ geom_smooth()+
           facet_wrap(~year,ncol=3) + ggtitle("Fitted values of yield over different si
         tes by Year")
```

```
`geom_smooth()` using method = 'loess'
```



As explained above there is an anomaly in the original data at 1931- 1932.