Assignment 7

108048110

2022-12-31

Assignment 7

Problem 1

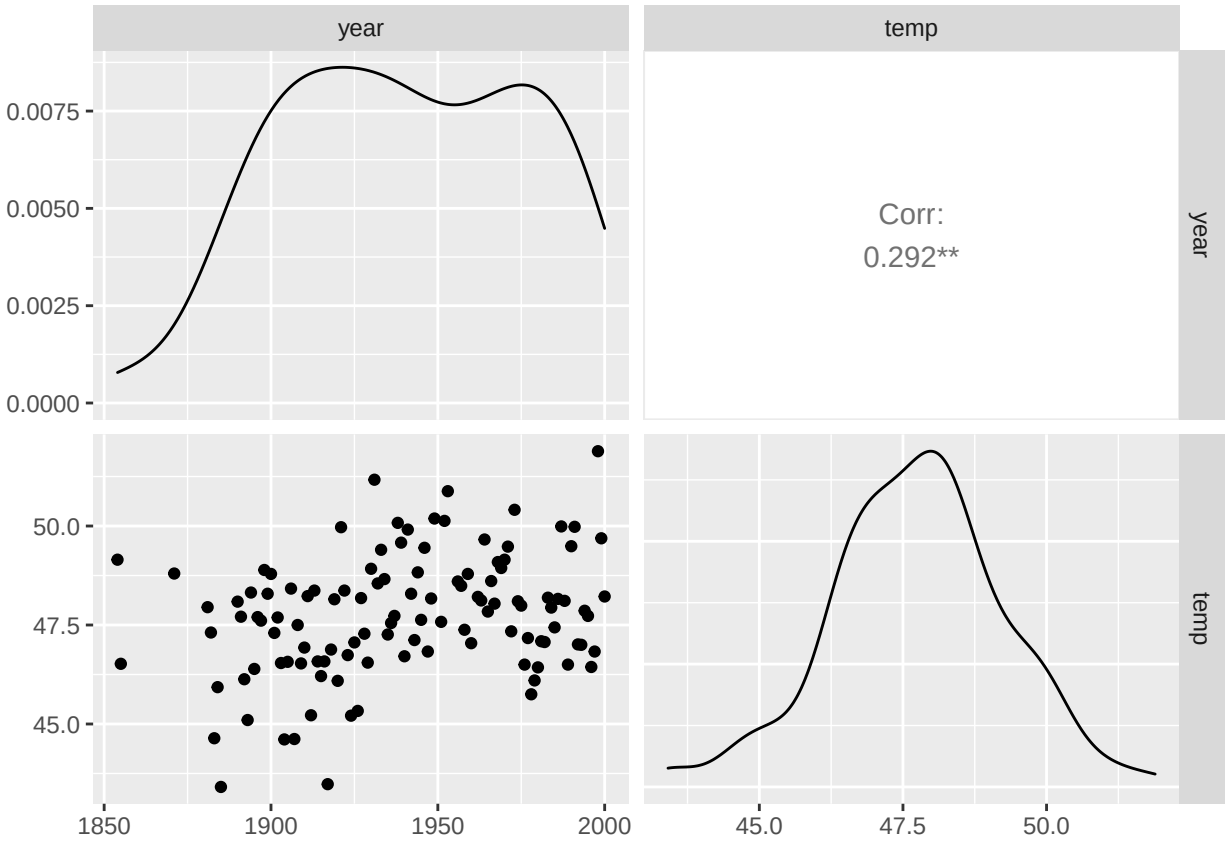
Data Source: US Historical Climatology Network

Data Description: Annual mean temperatures (F) in Ann Arbor.

```
data1 = read.table('./datasets/climatology network.txt', header=T)
summary(data1)
```

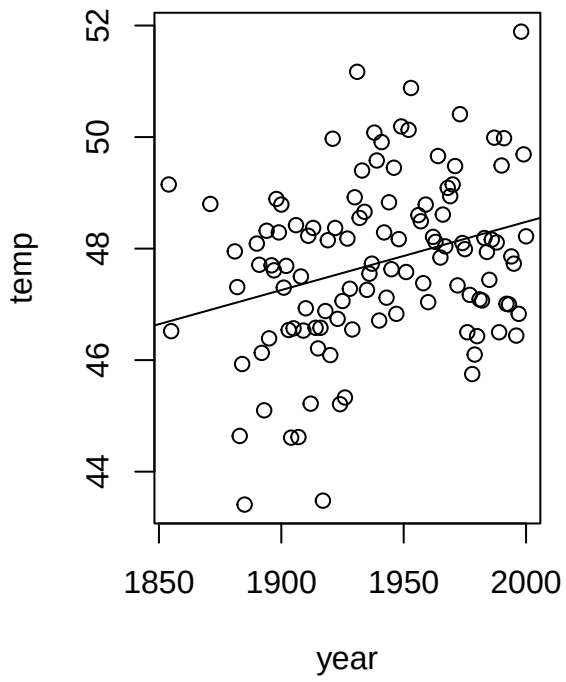
```
##       year           temp
##  Min.   :1854   Min.   :43.41
##  1st Qu.:1910   1st Qu.:46.78
##  Median :1939   Median :47.73
##  Mean   :1940   Mean   :47.74
##  3rd Qu.:1972   3rd Qu.:48.63
##  Max.   :2000   Max.   :51.89
```

```
ggpairs(data1)
```




```
attach(data1)
```

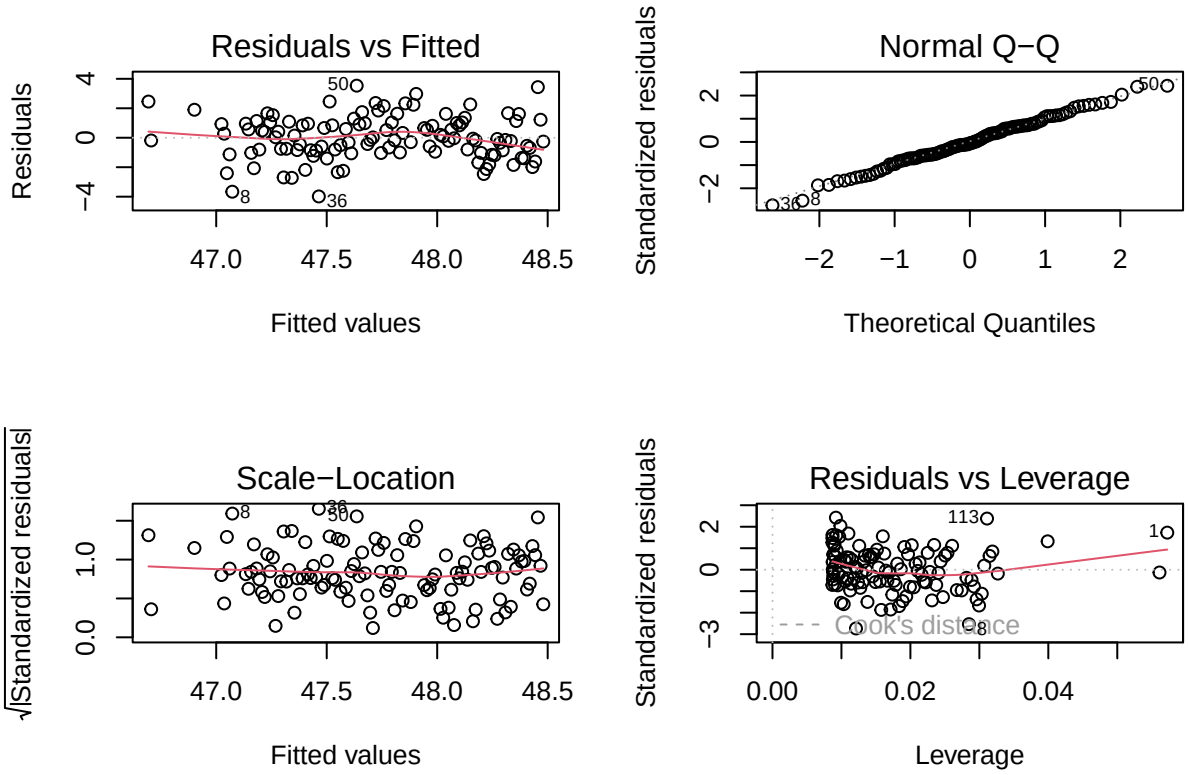
***i.* Is there a linear trend?**



- The plot suggests that local mean is increasing.

Perform diagnostic over the regression model to detect potential problems and to check whether the assumptions made by the linear model are met.

```
par(mfrow=c(2,2))
plot(model1)
```



- The diagnostic plots show residuals in 4 different ways.
  - Residual vs Fitted

    Used to check the linear relationship assumptions. The plot indicates a horizontal line, which is an indication for a linear relationship.
  - Normal QQ plot

    Used to examine whether the residuals are normally distributed. The plot shows that the residual points follow the straight dashed line, which is an indication of normally distributed residuals.
  - Scale-location

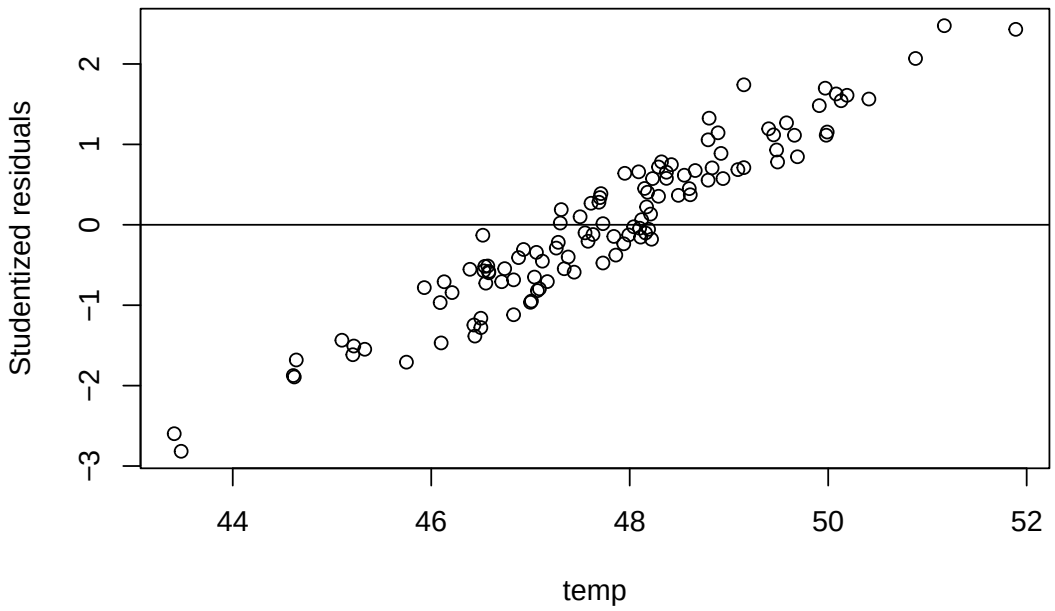
    Used to check the homogeneity of variance of the residuals. The plot has a horizontal line with equally spread points, which is an indication of homogeneity.
  - Residual vs. leverage

    1. Outliers

    In practice, any observation in a dataset that has a studentized residual greater than an absolute value of 3 is an outlier.

```
new_data[which(new_data$`studres(model1)`>3)]
```

```
## data frame with 0 columns and 115 rows
```



None of the observations have a studentized residual with an absolute value greater than 3, indicating there are no clear outliers in the dataset.

Note. Hence, although the residual vs leverage plot highlights the 3 most extreme points with standardized residuals whose absolute values are above 2, there is no outliers that exceed 3 standard deviations.

2. Leverage
   We observe a data with high leverage (usually > 2) if it has a high leverage value.

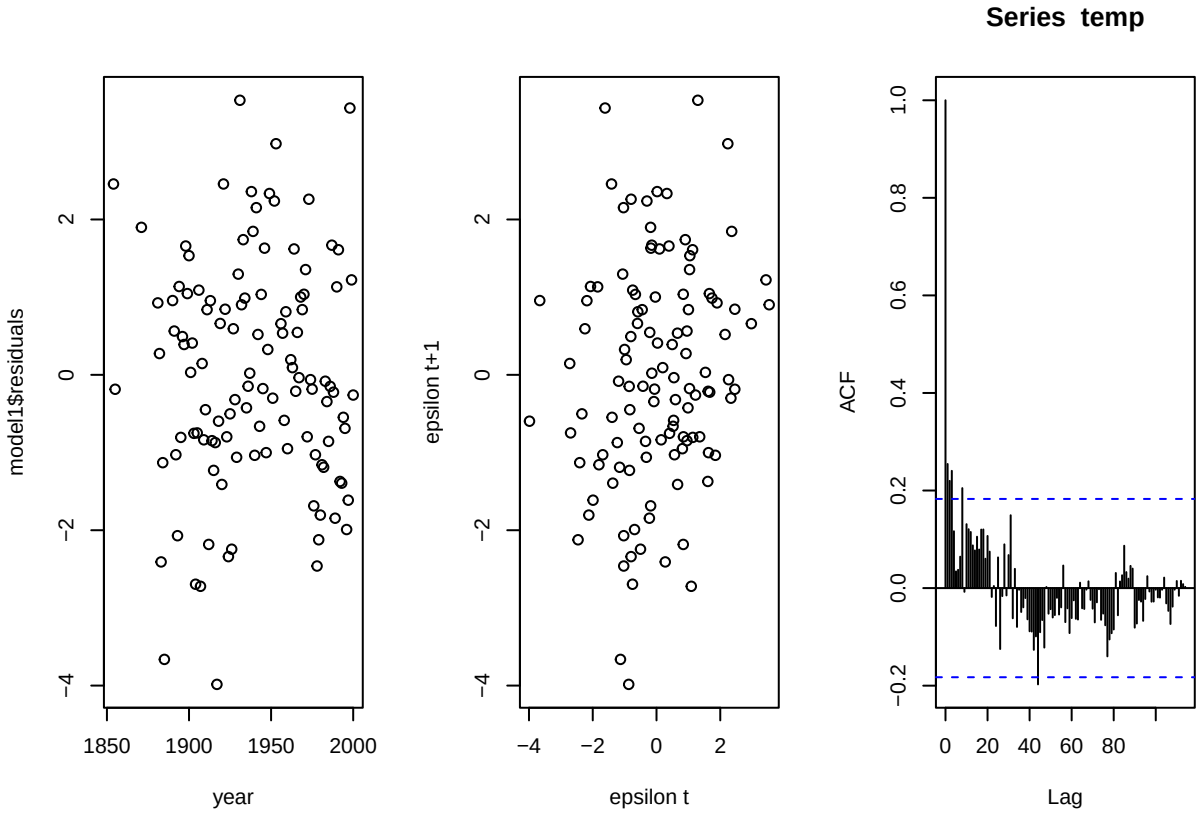
```
##          1
## 0.05724977
```

   The largest leverage value is 0.05, this is way smaller than 2, indicating that none of the observations in our dataset have high leverage.

Ans. Data shows a linear trend, but curvature seems to appear in model1.

***ii.* Observations in successive years may be correlated. Fit a model and estimates this correlation. Linear trend?**

- Check for correlated errors



- The plot seems to displaying correlated errors.
- And the autocorrelation plot indicates that data has the highest correlation when lag=1.
- DW test.

```
##
##  Durbin-Watson test
##
## data:  model1
## DW = 1.6177, p-value = 0.01524
## alternative hypothesis: true autocorrelation is greater than 0
```

- The p-value=0.015, when the significance level of α *is* 0.05, we have enough evidence to reject the null hypothesis, that is, correlated errors exist in the data.
- Assuming the correlated errors follow the autocorrelation structure of order 1.

```
model2 = gls(temp~year, correlation=corAR1(form=~year))

summary(model2)
```

```
## Generalized least squares fit by REML
##   Model: temp ~ year
##   Data: NULL
##        AIC     BIC    logLik
##   426.5694 437.479 -209.2847
```

```
##
## Correlation Structure: ARMA(1,0)
##  Formula: ~year
##  Parameter estimate(s):
##      Phi1
## 0.2303887
##
## Coefficients:
##                Value Std.Error  t-value p-value
## (Intercept) 25.18407  8.971864 2.807006  0.0059
## year         0.01164  0.004626 2.516015  0.0133
##
##  Correlation:
##      (Intr)
## year -1
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.7230803 -0.6321970 -0.0520135  0.6645795  2.3775123
##
## Residual standard error: 1.475718
## Degrees of freedom: 115 total; 113 residual
```

```
intervals(model2, which='var-cov')
```

```
## Approximate 95% confidence intervals
##
##  Correlation structure:
##           lower      est.     upper
## Phi1 0.02920118 0.2303887 0.4136364
##
##  Residual standard error:
##    lower     est.    upper
## 1.284091 1.475718 1.695942
```

- $\epsilon_{t+1} = \rho * \epsilon_t + \delta_t$,
- $\delta_t \sim N(0, \sigma^2)$
- So, under the AR(1) assumption, the estimate of correlation ρ is 0.19.
- After calculating C.I. of parameters for ρ, we can see that ρ's confidence interval do contain 0, that is to say, ρ is significantly different from 0.

iii. **Fit a polynomial model with degree 10 and use backward elimination to reduce the degree of the model. Plot your fitted model on top of the data. Use this model to predict the temperature in 2020.**

- Fit a $10^{th} - order$ polynomial model.

```
mean(year)
```

```
## [1] 1939.739
```

```r
model10 = lm(temp~poly(year, degree=10))
model9 = lm(temp~poly(year, degree=9))
model8 = lm(temp~poly(year, degree=8))
model7 = lm(temp~poly(year, degree=7))
model6 = lm(temp~poly(year, degree=6))
model5 = lm(temp~poly(year, degree=5), data=data1)
model4 = lm(temp~poly(year, degree=4))
model3 = lm(temp~poly(year, degree=3))
model2 = lm(temp~poly(year, degree=2))
model1 = lm(temp~year)
```

```r
summary(model10)
```

```
##
## Call:
## lm(formula = temp ~ poly(year, degree = 10))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4987 -0.8641 -0.1745  1.1450  3.4255
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 47.7426     0.1319 361.927  < 2e-16 ***
## poly(year, degree = 10)1     4.7616     1.4146   3.366  0.00107 **
## poly(year, degree = 10)2    -0.9071     1.4146  -0.641  0.52277
## poly(year, degree = 10)3    -3.3132     1.4146  -2.342  0.02108 *
## poly(year, degree = 10)4     2.4383     1.4146   1.724  0.08774 .
## poly(year, degree = 10)5     3.3824     1.4146   2.391  0.01860 *
## poly(year, degree = 10)6     1.2124     1.4146   0.857  0.39337
## poly(year, degree = 10)7    -0.9373     1.4146  -0.663  0.50908
## poly(year, degree = 10)8    -1.1011     1.4146  -0.778  0.43812
## poly(year, degree = 10)9     1.3994     1.4146   0.989  0.32483
## poly(year, degree = 10)10    0.3474     1.4146   0.246  0.80652
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.415 on 104 degrees of freedom
## Multiple R-squared:  0.2165, Adjusted R-squared:  0.1411
## F-statistic: 2.873 on 10 and 104 DF,  p-value: 0.003335
```

```r
summary(model9)
```

```
##
## Call:
## lm(formula = temp ~ poly(year, degree = 9))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4836 -0.8831 -0.2156  1.1354  3.3936
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               47.7426     0.1313 363.557  < 2e-16 ***
## poly(year, degree = 9)1    4.7616     1.4083   3.381  0.00101 **
## poly(year, degree = 9)2   -0.9071     1.4083  -0.644  0.52089
## poly(year, degree = 9)3   -3.3132     1.4083  -2.353  0.02050 *
## poly(year, degree = 9)4    2.4383     1.4083   1.731  0.08631 .
## poly(year, degree = 9)5    3.3824     1.4083   2.402  0.01807 *
## poly(year, degree = 9)6    1.2124     1.4083   0.861  0.39123
## poly(year, degree = 9)7   -0.9373     1.4083  -0.666  0.50716
## poly(year, degree = 9)8   -1.1011     1.4083  -0.782  0.43605
## poly(year, degree = 9)9    1.3994     1.4083   0.994  0.32265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.408 on 105 degrees of freedom
## Multiple R-squared:  0.216,  Adjusted R-squared:  0.1488
## F-statistic: 3.215 on 9 and 105 DF,  p-value: 0.001769
```

```
summary(model8)
```

```
##
## Call:
## lm(formula = temp ~ poly(year, degree = 8))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6086 -0.8600 -0.2385  1.0608  3.3975
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               47.7426     0.1313 363.579  < 2e-16 ***
## poly(year, degree = 8)1    4.7616     1.4082   3.381  0.00101 **
## poly(year, degree = 8)2   -0.9071     1.4082  -0.644  0.52085
## poly(year, degree = 8)3   -3.3132     1.4082  -2.353  0.02047 *
## poly(year, degree = 8)4    2.4383     1.4082   1.732  0.08626 .
## poly(year, degree = 8)5    3.3824     1.4082   2.402  0.01805 *
## poly(year, degree = 8)6    1.2124     1.4082   0.861  0.39118
## poly(year, degree = 8)7   -0.9373     1.4082  -0.666  0.50713
## poly(year, degree = 8)8   -1.1011     1.4082  -0.782  0.43600
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.408 on 106 degrees of freedom
## Multiple R-squared:  0.2086, Adjusted R-squared:  0.1489
## F-statistic: 3.494 on 8 and 106 DF,  p-value: 0.001284
```

```
summary(model7)
```

```
##
## Call:
## lm(formula = temp ~ poly(year, degree = 7))
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.5922 -0.9032 -0.2322  0.9880  3.2941
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               47.7426     0.1311 364.241  < 2e-16 ***
## poly(year, degree = 7)1    4.7616     1.4056   3.388 0.000988 ***
## poly(year, degree = 7)2   -0.9071     1.4056  -0.645 0.520083
## poly(year, degree = 7)3   -3.3132     1.4056  -2.357 0.020234 *
## poly(year, degree = 7)4    2.4383     1.4056   1.735 0.085672 .
## poly(year, degree = 7)5    3.3824     1.4056   2.406 0.017828 *
## poly(year, degree = 7)6    1.2124     1.4056   0.863 0.390303
## poly(year, degree = 7)7   -0.9373     1.4056  -0.667 0.506341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.406 on 107 degrees of freedom
## Multiple R-squared:  0.2041, Adjusted R-squared:  0.152
## F-statistic: 3.919 on 7 and 107 DF,  p-value: 0.0007651
```

```
summary(model6)
```

```
##
## Call:
## lm(formula = temp ~ poly(year, degree = 6))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6846 -0.8825 -0.1428  0.9388  3.2950
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               47.7426     0.1307 365.181  < 2e-16 ***
## poly(year, degree = 6)1    4.7616     1.4020   3.396 0.000957 ***
## poly(year, degree = 6)2   -0.9071     1.4020  -0.647 0.518996
## poly(year, degree = 6)3   -3.3132     1.4020  -2.363 0.019905 *
## poly(year, degree = 6)4    2.4383     1.4020   1.739 0.084851 .
## poly(year, degree = 6)5    3.3824     1.4020   2.413 0.017527 *
## poly(year, degree = 6)6    1.2124     1.4020   0.865 0.389067
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.402 on 108 degrees of freedom
## Multiple R-squared:  0.2008, Adjusted R-squared:  0.1564
## F-statistic: 4.522 on 6 and 108 DF,  p-value: 0.0003978
```

```
summary(model5)
```

```
##
## Call:
## lm(formula = temp ~ poly(year, degree = 5), data = data1)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.7142 -0.9198 -0.1420  0.9903  3.2364
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              47.7426     0.1306 365.604  < 2e-16 ***
## poly(year, degree = 5)1   4.7616     1.4004   3.400 0.000942 ***
## poly(year, degree = 5)2  -0.9071     1.4004  -0.648 0.518500
## poly(year, degree = 5)3  -3.3132     1.4004  -2.366 0.019749 *
## poly(year, degree = 5)4   2.4383     1.4004   1.741 0.084470 .
## poly(year, degree = 5)5   3.3824     1.4004   2.415 0.017384 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.4 on 109 degrees of freedom
## Multiple R-squared:  0.1952, Adjusted R-squared:  0.1583
## F-statistic: 5.289 on 5 and 109 DF,  p-value: 0.0002176
```

```
summary(model4)
```

```
##
## Call:
## lm(formula = temp ~ poly(year, degree = 4))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0085 -0.9618 -0.0913  0.9926  3.7370
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              47.7426     0.1334 357.827  < 2e-16 ***
## poly(year, degree = 4)1   4.7616     1.4308   3.328  0.00119 **
## poly(year, degree = 4)2  -0.9071     1.4308  -0.634  0.52741
## poly(year, degree = 4)3  -3.3132     1.4308  -2.316  0.02243 *
## poly(year, degree = 4)4   2.4383     1.4308   1.704  0.09117 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.431 on 110 degrees of freedom
## Multiple R-squared:  0.1522, Adjusted R-squared:  0.1213
## F-statistic: 4.936 on 4 and 110 DF,  p-value: 0.001068
```

```
summary(model3)
```

```
##
## Call:
## lm(formula = temp ~ poly(year, degree = 3))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8557 -0.9646 -0.1552  1.0485  4.1538
##
## Coefficients:
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               47.7426     0.1346 354.796   <2e-16 ***
## poly(year, degree = 3)1   4.7616     1.4430   3.300   0.0013 **
## poly(year, degree = 3)2  -0.9071     1.4430  -0.629   0.5309
## poly(year, degree = 3)3  -3.3132     1.4430  -2.296   0.0236 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.443 on 111 degrees of freedom
## Multiple R-squared:  0.1298, Adjusted R-squared:  0.1063
## F-statistic: 5.518 on 3 and 111 DF,  p-value: 0.001436
```

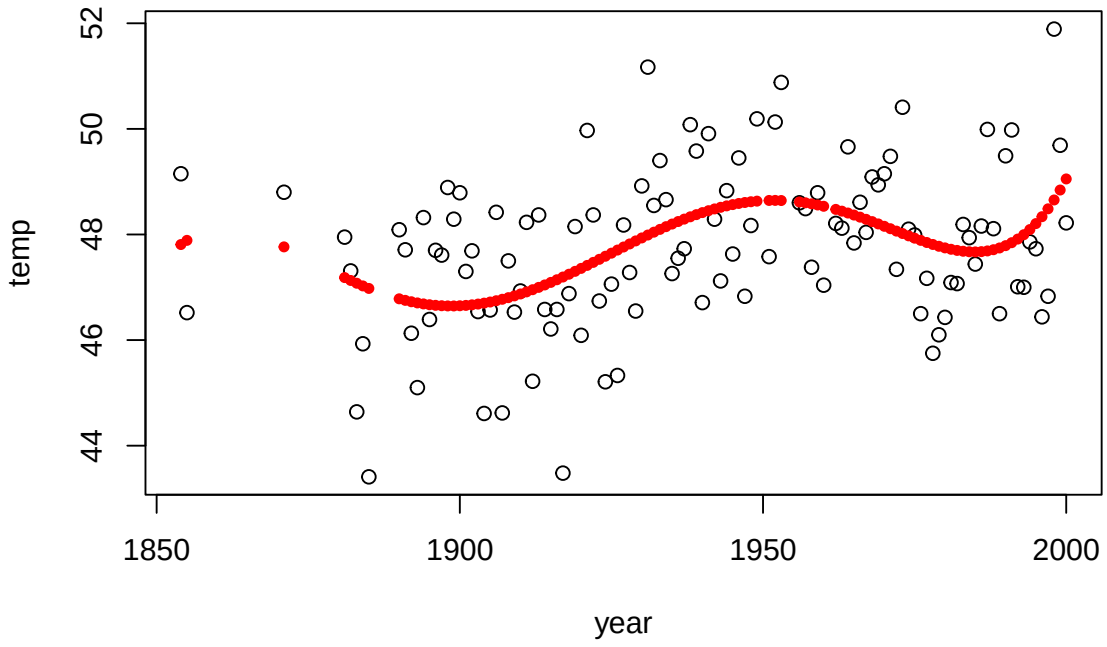
```
summary(model2)
```

```
##
## Call:
## lm(formula = temp ~ poly(year, degree = 2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0412 -0.9538 -0.0624  0.9959  3.5820
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               47.7426     0.1371 348.218  < 2e-16 ***
## poly(year, degree = 2)1   4.7616     1.4703   3.239  0.00158 **
## poly(year, degree = 2)2  -0.9071     1.4703  -0.617  0.53851
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.47 on 112 degrees of freedom
## Multiple R-squared:  0.08846,    Adjusted R-squared:  0.07218
## F-statistic: 5.434 on 2 and 112 DF,  p-value: 0.005591
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = temp ~ year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9843 -0.9113 -0.0820  0.9946  3.5343
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.005510   7.310781   3.284  0.00136 **
## year         0.012237   0.003768   3.247  0.00153 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.466 on 113 degrees of freedom
## Multiple R-squared:  0.08536,    Adjusted R-squared:  0.07727
## F-statistic: 10.55 on 1 and 113 DF,  p-value: 0.001533
```

- Seems like model5 is the highest polynomial model.
- Plotting fitted model5.



Predicting data=2020

```
##        1
## 47.80958
```

- 47.81

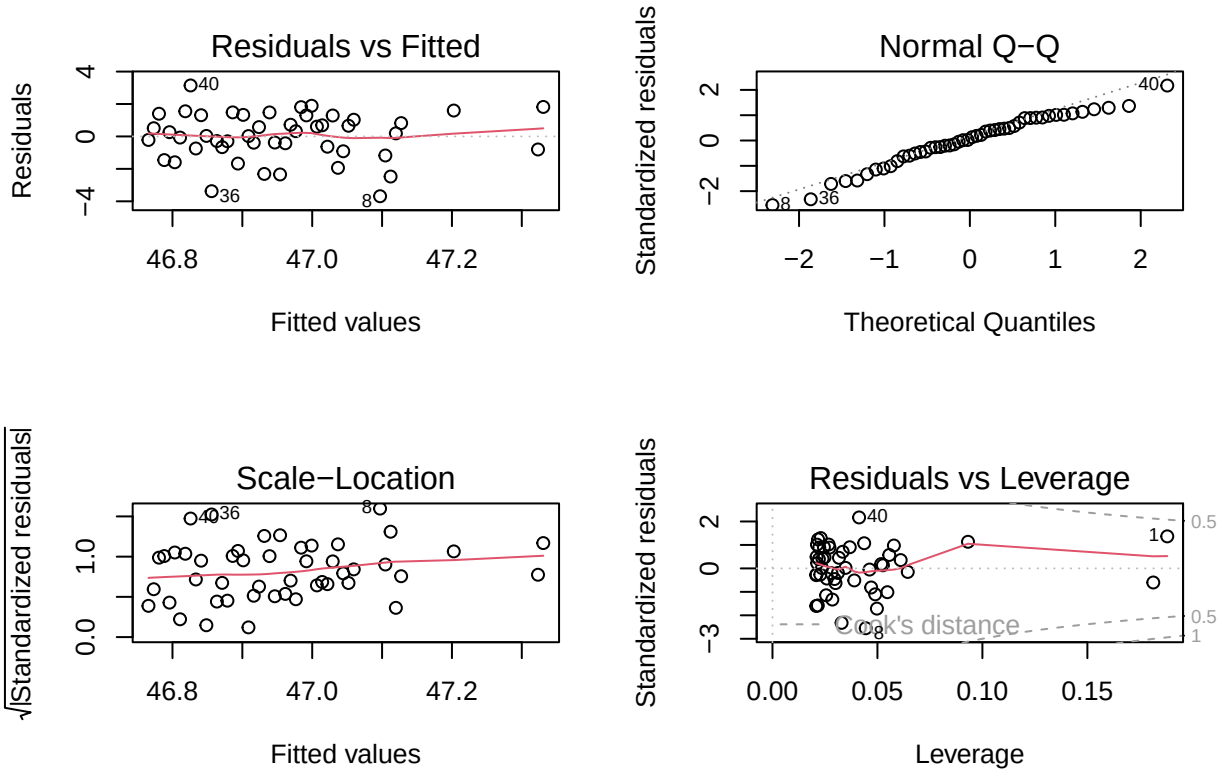
iv. **Suppose *temp* was constant until 1930 and then began a linear trend. Fit a model correspond to the claim. What does the fitted model tell?**

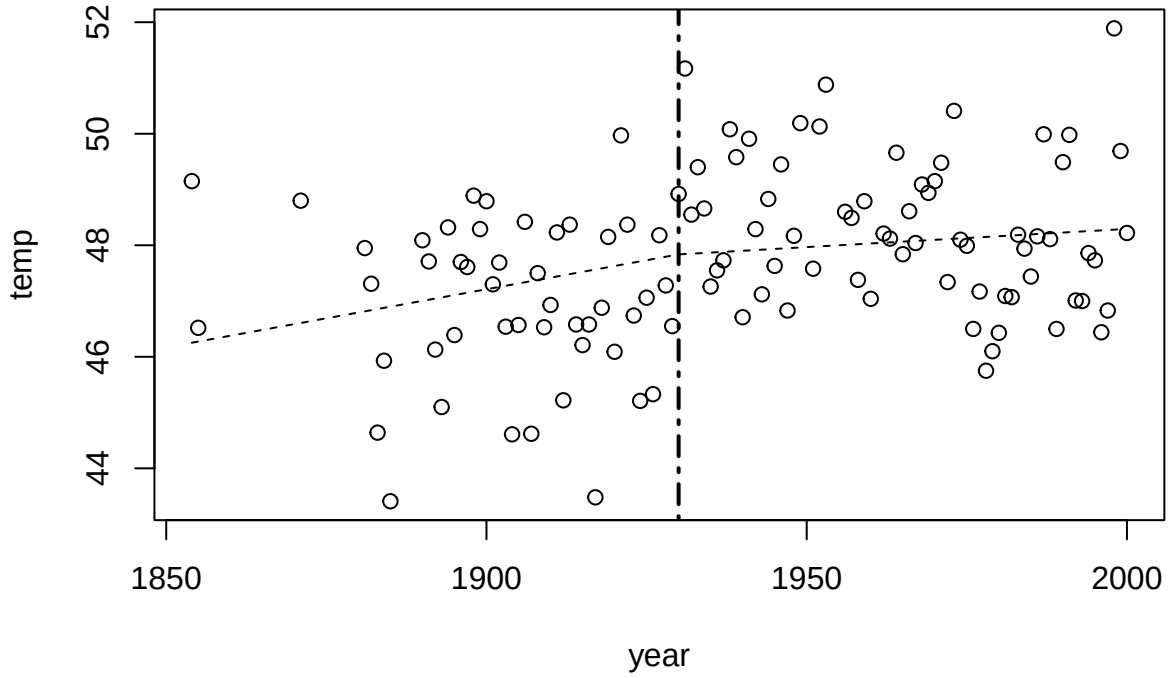
```
model1 = lm(temp~year, subset = (year<1930))
model2 = lm(temp~year, subset=year>=1930)

par(mfrow=c(2,2))
summary(model1)
```

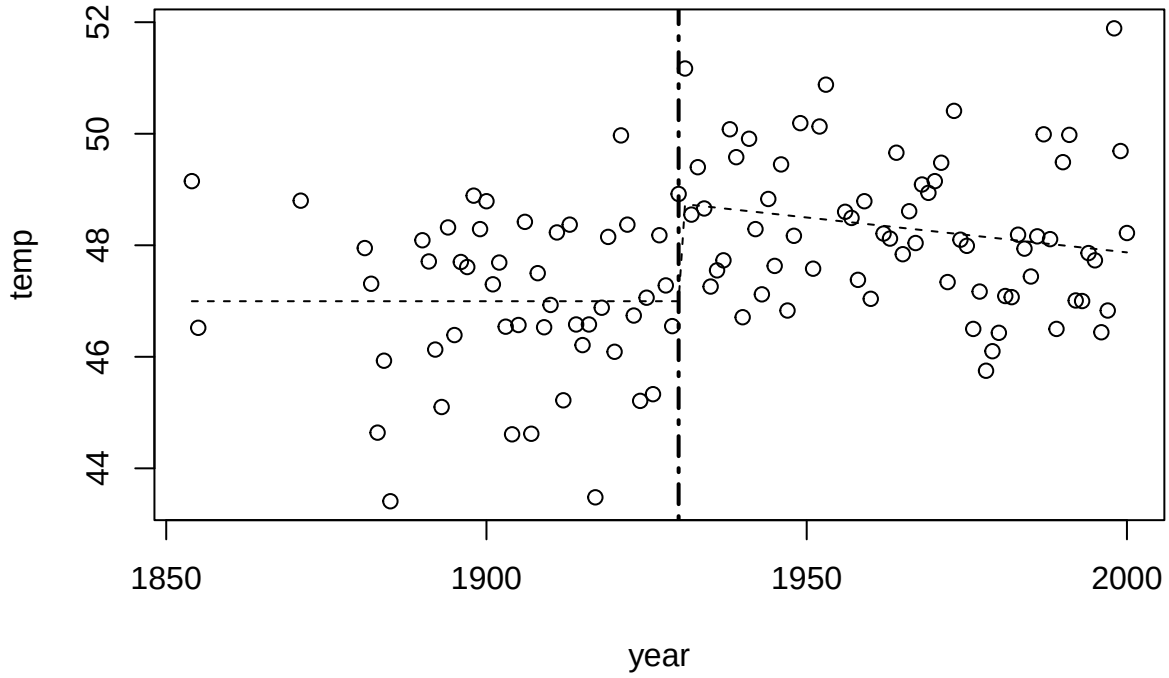
```
##
## Call:
## lm(formula = temp ~ year, subset = (year < 1930))
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.6872 -0.7584  0.1109  1.2926  3.1441
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 61.300439  23.189506   2.643   0.0112 *
## year        -0.007535   0.012181  -0.619   0.5392
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.479 on 46 degrees of freedom
## Multiple R-squared:  0.008249,   Adjusted R-squared:  -0.01331
## F-statistic: 0.3826 on 1 and 46 DF,  p-value: 0.5392
```

```
plot(model1)
```

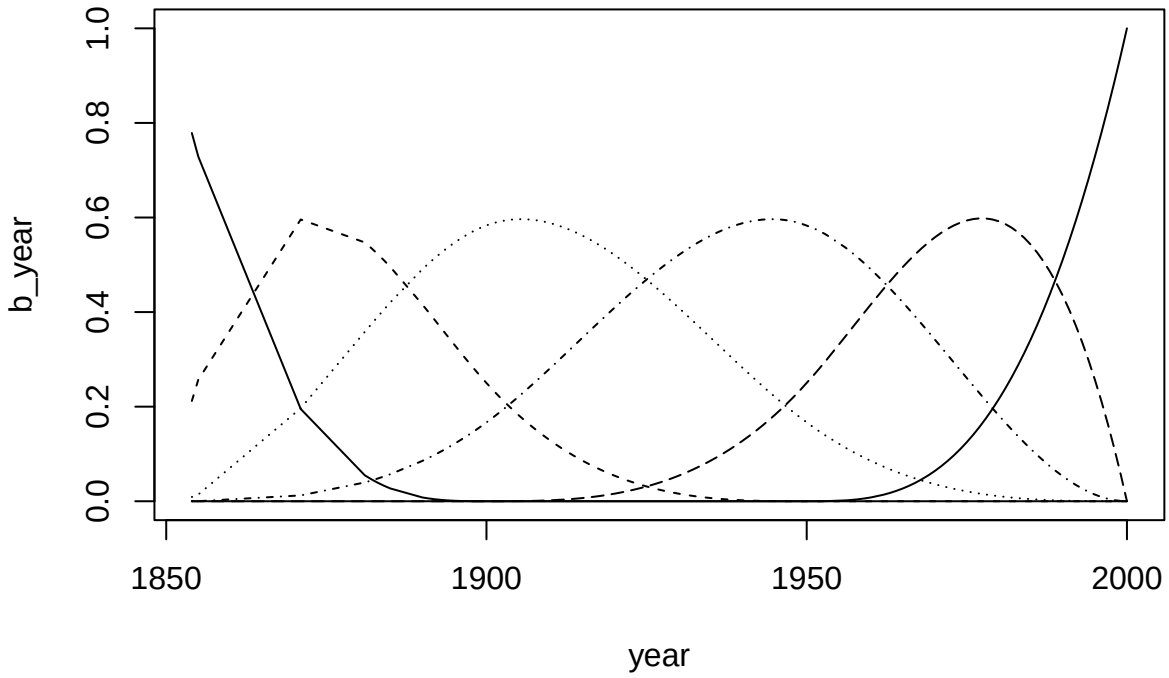


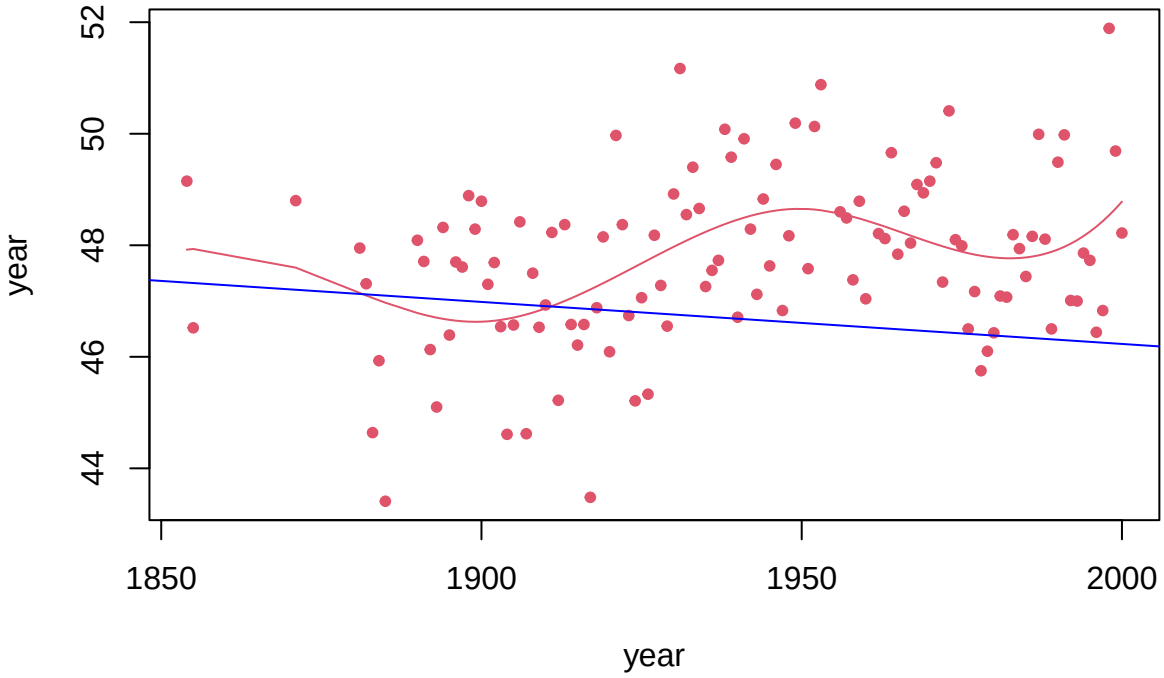

temp
52
50
48
46
44
1850
1900
1950
2000
year

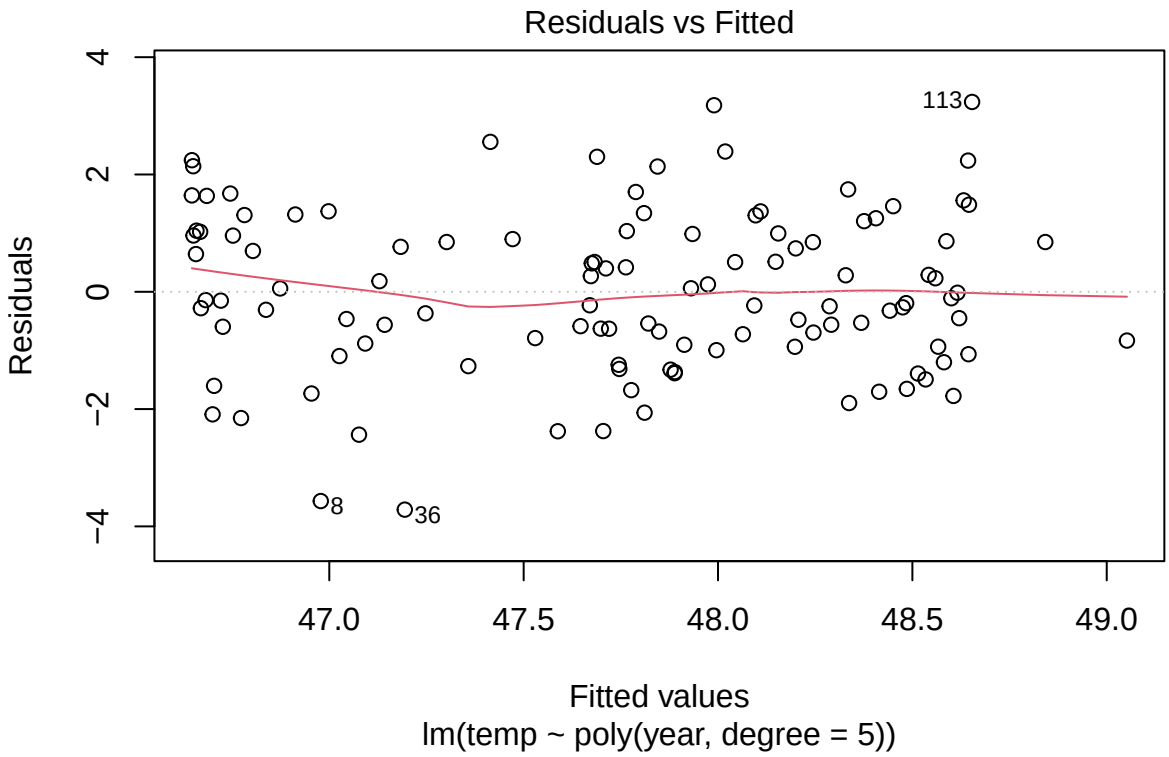


v. Make a cubic spline fit with 6 basis functions evenly spaced on the range. Visualize this basis functions. Plot the fit in comparison to the previous fits. Does this model fit better than the straight line model?

```
## Loading required package: splines
```


b_year
0.0
0.2
0.4
0.6
0.8
1.0
1850
1900
1950
2000
year


year
52
50
48
46
44
1850
1900
1950
2000
year



- As we can observe from the plot, by complicating the model, we ameliorate the effect of curvature.
- Fits better

Problem 2

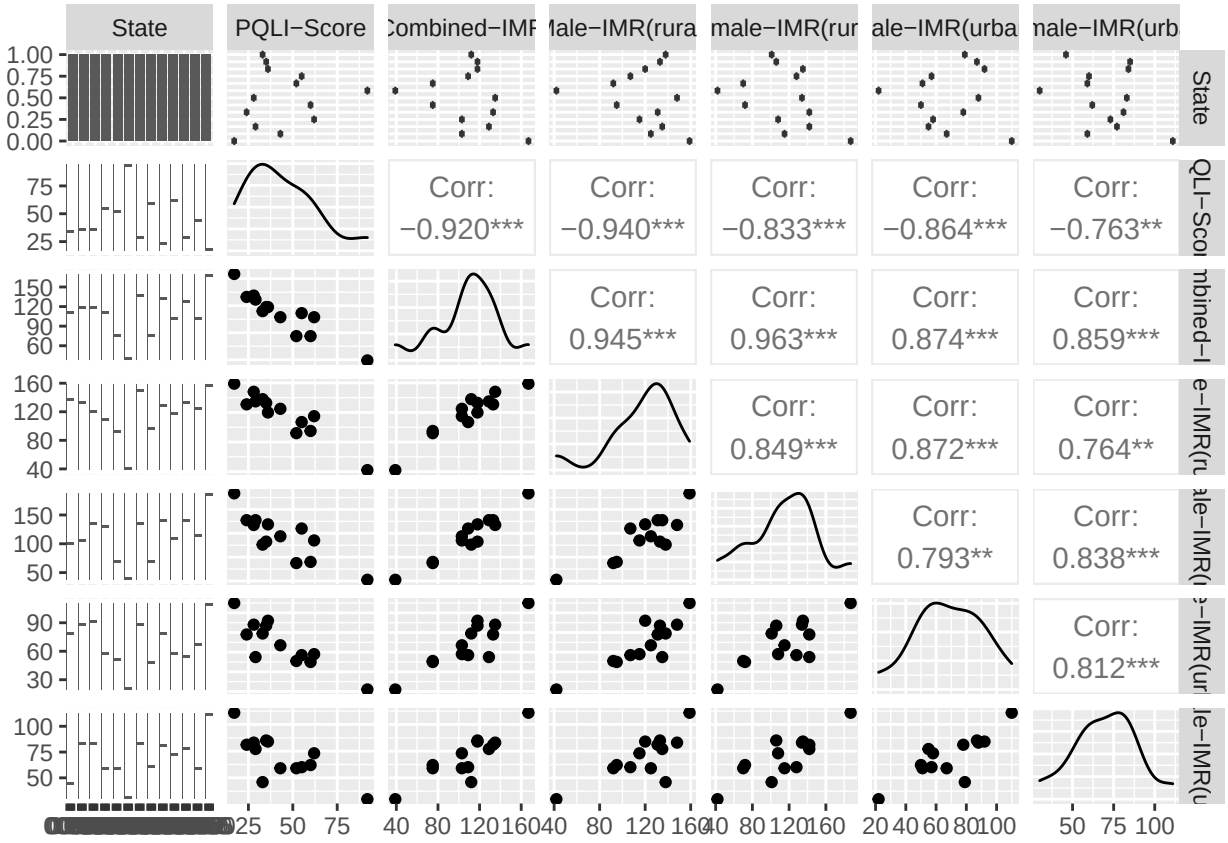
Data Overview

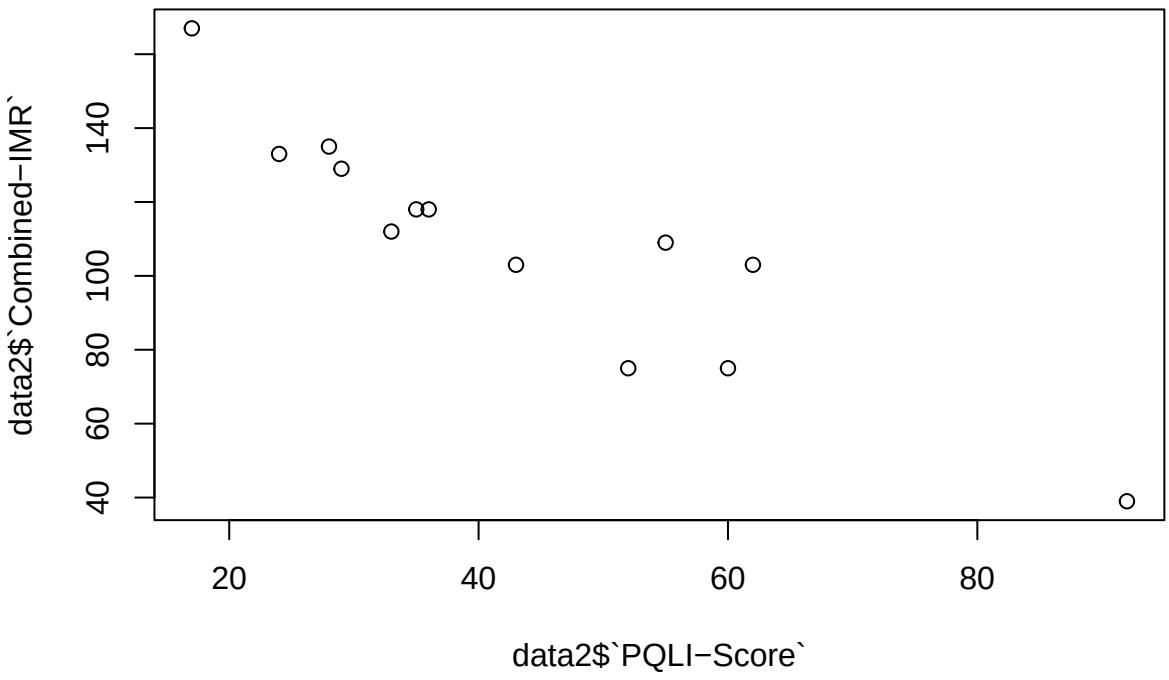
```
summary(data2)
```

```
##     State              PQLI-Score     Combined-IMR   Male-IMR(rural)
##  Length:13          Min.   :17.00   Min.   : 39.0   Min.   : 42.0
##  Class :character   1st Qu.:29.00   1st Qu.:103.0   1st Qu.:107.0
##  Mode  :character   Median :36.00   Median :112.0   Median :125.0
##                     Mean   :43.54   Mean   :108.9   Mean   :118.5
##                     3rd Qu.:55.00   3rd Qu.:129.0   3rd Qu.:135.0
##                     Max.   :92.00   Max.   :167.0   Max.   :159.0
##  Female-IMR(rural) Male-IMR(urban)  Female-IMR(urban)
##  Min.   : 42       Min.   : 22.00   Min.   : 30
##  1st Qu.:101       1st Qu.: 55.00   1st Qu.: 59
##  Median :115       Median : 67.00   Median : 73
##  Mean   :114       Mean   : 68.77   Mean   : 70
##  3rd Qu.:135       3rd Qu.: 87.00   3rd Qu.: 83
##  Max.   :187       Max.   :110.00   Max.   :111
```

```
ggpairs(data2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





- they seems to have a strong negative correlation.

Construct a single model, using dummy variables to distinguish rural-urban and male-female difference. Investigate whether there is a male-female and rural-urban difference in IMR after adjusting for other covariates.

- Construct new data, setting dummy variables for gender and location.
- Gender: 0=male; 1=female
- Rural: 0=rural; 1=urban

newstate	newScore	newIMR	gender	rural	data
UTTAR PRAD.	17	167	-1	-1	159
MADHYA PRAD.	28	135	-1	-1	148
ORISSA	24	133	-1	-1	131
RAJASTHAN	29	129	-1	-1	135
GUJARAT	36	118	-1	-1	120
ANDHRA PRAD.	33	112	-1	-1	138

```
## The following objects are masked _by_ .GlobalEnv:
##
##     data, gender, newIMR, newScore, newstate, rural
```

Construct single model

```
model1 = lm(data~newScore+gender+rural+newScore:gender+newScore:rural+gender:rural)

summary(model1)
```

```
##
## Call:
## lm(formula = data ~ newScore + gender + rural + newScore:gender +
##     newScore:rural + gender:rural)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.095  -5.626   0.054   6.561  34.363
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     143.91349    5.15626  27.910  < 2e-16 ***
## newScore         -1.17381    0.10808 -10.860 3.68e-14 ***
## gender           -1.43219    5.15626  -0.278  0.78247
## rural           -36.32821    5.15626  -7.045 8.73e-09 ***
## newScore:gender   0.01434    0.10808   0.133  0.89502
## newScore:rural    0.29641    0.10808   2.742  0.00872 **
## gender:rural      1.42308    2.10764   0.675  0.50300
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.2 on 45 degrees of freedom
## Multiple R-squared:  0.8472, Adjusted R-squared:  0.8269
## F-statistic:  41.6 on 6 and 45 DF,  p-value: < 2.2e-16
```

- *p-value of newScore, rural, and interaction term between newScore and rural are significant.*
  - $C = c1 : \mu = E(data|d_1 = -1, d_2 = 1) = (\beta_0 - \beta_2 + \beta_3 - \beta_6) + (\beta_1 - \beta_4 + \beta_5)x$
  - $C = c2 : \mu = E(data|d_1 = -1, d_2 = -1) = (\beta_0 - \beta_2 - \beta_3 + \beta_6) + (\beta_1 - \beta_4 - \beta_5)x$
  - $C = c3 : \mu = E(data|d_1 = 1, d_2 = 1) = (\beta_0 + \beta_2 + \beta_3 + \beta_6) + (\beta_1 + \beta_4 + \beta_5)x$
  - $C = c4 : \mu = E(data|d_1 = 1, d_2 = -1) = (\beta_0 + \beta_2 - \beta_3 - \beta_6) + (\beta_1 + \beta_4 - \beta_5)x$
  - Constant terms, d1, d2 are orthogonal when there are equal number of observations in each categories.
- Forward Filtering

```
model1 = lm(data~newScore)
summary(model1)
```

```
##
## Call:
## lm(formula = data ~ newScore)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -59.18 -22.53  -5.05  25.41  63.04
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 143.9135     9.6918  14.849  < 2e-16 ***
## newScore     -1.1738     0.2032  -5.778 4.84e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.57 on 50 degrees of freedom
## Multiple R-squared:  0.4004, Adjusted R-squared:  0.3884
## F-statistic: 33.38 on 1 and 50 DF,  p-value: 4.837e-07
```

```
model2 = lm(data~newScore+I(rural))
summary(model2)
```

```
##
## Call:
## lm(formula = data ~ newScore + I(rural))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.299  -8.702   1.386   9.605  39.618
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 143.9135     5.3697   26.80  < 2e-16 ***
## newScore     -1.1738     0.1126  -10.43 4.91e-14 ***
## I(rural)    -23.4231     2.1949  -10.67 2.23e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.83 on 49 degrees of freedom
## Multiple R-squared:  0.8196, Adjusted R-squared:  0.8122
## F-statistic: 111.3 on 2 and 49 DF,  p-value: < 2.2e-16
```

```
model3 = lm(data~newScore+I(rural)+newScore:rural)
summary(model3)
```

```
##
## Call:
## lm(formula = data ~ newScore + I(rural) + newScore:rural)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.790  -5.237   0.175   7.759  31.752
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    143.9135     5.0268  28.629  < 2e-16 ***
## newScore        -1.1738     0.1054 -11.140 6.58e-15 ***
## I(rural)       -36.3282     5.0268  -7.227 3.30e-09 ***
## newScore:rural   0.2964     0.1054   2.813  0.00709 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.82 on 48 degrees of freedom
## Multiple R-squared:  0.8451, Adjusted R-squared:  0.8355
## F-statistic: 87.32 on 3 and 48 DF,  p-value: < 2.2e-16
```

- Add rural variable.

```
model4 = lm(data~newScore+I(rural)+newScore:rural+I(gender))
summary(model4)
```

```
##
## Call:
## lm(formula = data ~ newScore + I(rural) + newScore:rural + I(gender))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.983  -5.691   0.175   7.929  32.560
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    143.9135     5.0718  28.375  < 2e-16 ***
## newScore        -1.1738     0.1063 -11.041 1.19e-14 ***
## I(rural)       -36.3282     5.0718  -7.163 4.63e-09 ***
## I(gender)       -0.8077     2.0731  -0.390  0.69859
## newScore:rural   0.2964     0.1063   2.788  0.00763 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.95 on 47 degrees of freedom
## Multiple R-squared:  0.8456, Adjusted R-squared:  0.8325
## F-statistic: 64.37 on 4 and 47 DF,  p-value: < 2.2e-16
```

```
model5 = lm(data~newScore+I(rural)+newScore:rural+newScore:gender)
summary(model5)
```

```
##
## Call:
## lm(formula = data ~ newScore + I(rural) + newScore:rural + newScore:gender)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.112  -5.810   0.143   7.921  31.974
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     143.91349    5.07515  28.357  < 2e-16 ***
## newScore         -1.17381    0.10638 -11.034 1.22e-14 ***
## I(rural)        -36.32821    5.07515  -7.158 4.71e-09 ***
## newScore:rural    0.29641    0.10638   2.786  0.00767 **
## newScore:gender  -0.01306    0.04349  -0.300  0.76533
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.96 on 47 degrees of freedom
## Multiple R-squared:  0.8454, Adjusted R-squared:  0.8323
## F-statistic: 64.27 on 4 and 47 DF,  p-value: < 2.2e-16
```

As we can see, neither gender nor the interaction term, newScore:gender are significant to the model.

- Backward Filtering
- Full model: $IMR = \beta_0 + \beta_1 * newScore + \beta_2 * gender + \beta_3 * rural + \beta_4 * newScore * gender + \beta_5 * newScore * rural + \beta_6 * gender * rural + \epsilon$

```
model2 = lm(data~newScore+gender+rural+newScore:gender+newScore:rural+gender:rural)
summary(model2)
```

```
##
## Call:
## lm(formula = data ~ newScore + gender + rural + newScore:gender +
##     newScore:rural + gender:rural)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.095  -5.626   0.054   6.561  34.363
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     143.91349    5.15626  27.910  < 2e-16 ***
## newScore         -1.17381    0.10808 -10.860 3.68e-14 ***
## gender           -1.43219    5.15626  -0.278  0.78247
## rural           -36.32821    5.15626  -7.045 8.73e-09 ***
## newScore:gender   0.01434    0.10808   0.133  0.89502
## newScore:rural    0.29641    0.10808   2.742  0.00872 **
## gender:rural      1.42308    2.10764   0.675  0.50300
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.2 on 45 degrees of freedom
## Multiple R-squared:  0.8472, Adjusted R-squared:  0.8269
## F-statistic:  41.6 on 6 and 45 DF,  p-value: < 2.2e-16
```

- Removing elements with least significant p-value from the model.
- Tried removing newScore:gender

```
model3 = lm(data~newScore+I(gender)+I(rural)+newScore:rural+gender:rural)

summary(model3)
```

```
##
## Call:
```

```
## lm(formula = data ~ newScore + I(gender) + I(rural) + newScore:rural +
##     gender:rural)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.246  -5.853   0.132   6.642  33.983
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    143.9135     5.1009  28.213  < 2e-16 ***
## newScore        -1.1738     0.1069 -10.978 1.93e-14 ***
## I(gender)       -0.8077     2.0850  -0.387  0.70026
## I(rural)       -36.3282     5.1009  -7.122 5.98e-09 ***
## newScore:rural   0.2964     0.1069   2.772  0.00801 **
## rural:gender     1.4231     2.0850   0.683  0.49833
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.04 on 46 degrees of freedom
## Multiple R-squared:  0.8472, Adjusted R-squared:  0.8306
## F-statistic:    51 on 5 and 46 DF,  p-value: < 2.2e-16
```

```
anova(model3, model2)
```

```
## Analysis of Variance Table
##
## Model 1: data ~ newScore + I(gender) + I(rural) + newScore:rural + gender:rural
## Model 2: data ~ newScore + gender + rural + newScore:gender + newScore:rural +
##     gender:rural
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     46 10399
## 2     45 10395  1    4.0681 0.0176  0.895
```

- First remove *gender:rural* to see if it affect gender's parameter.

```
model4= lm(data~newScore+I(gender)+I(rural)+newScore:rural)
summary(model4)
```

```
##
## Call:
## lm(formula = data ~ newScore + I(gender) + I(rural) + newScore:rural)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.983  -5.691   0.175   7.929  32.560
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    143.9135     5.0718  28.375  < 2e-16 ***
## newScore        -1.1738     0.1063 -11.041 1.19e-14 ***
## I(gender)       -0.8077     2.0731  -0.390  0.69859
## I(rural)       -36.3282     5.0718  -7.163 4.63e-09 ***
```

```
## newScore:rural   0.2964     0.1063   2.788  0.00763 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.95 on 47 degrees of freedom
## Multiple R-squared:  0.8456, Adjusted R-squared:  0.8325
## F-statistic: 64.37 on 4 and 47 DF,  p-value: < 2.2e-16
```

- Gender is still not significant.
- Removing Gender

```
model5 = lm(data~newScore+I(rural)+newScore:rural)
summary(model5)
```

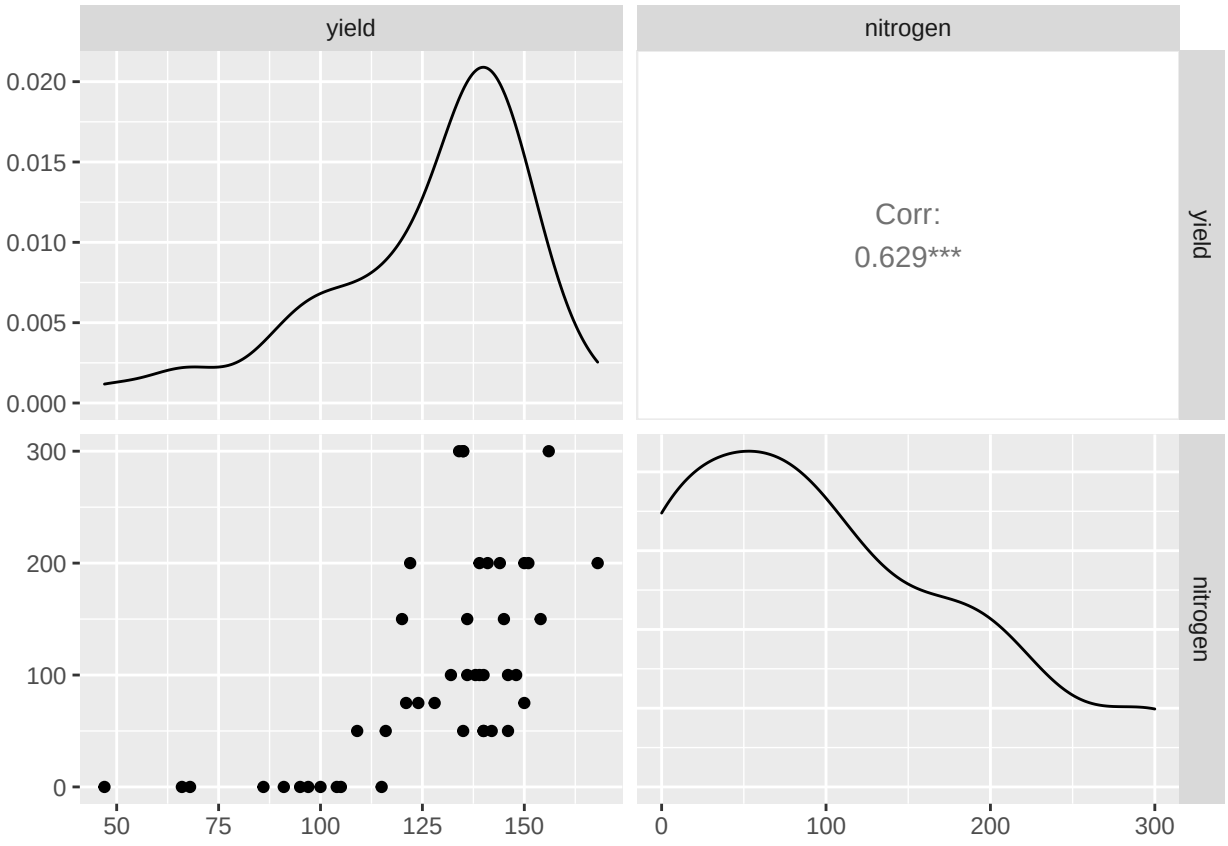
```
##
## Call:
## lm(formula = data ~ newScore + I(rural) + newScore:rural)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.790  -5.237   0.175   7.759  31.752
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    143.9135     5.0268  28.629  < 2e-16 ***
## newScore        -1.1738     0.1054 -11.140 6.58e-15 ***
## I(rural)       -36.3282     5.0268  -7.227 3.30e-09 ***
## newScore:rural   0.2964     0.1054   2.813  0.00709 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.82 on 48 degrees of freedom
## Multiple R-squared:  0.8451, Adjusted R-squared:  0.8355
## F-statistic: 87.32 on 3 and 48 DF,  p-value: < 2.2e-16
```

- To sum up, the conclusion drawn by the backward approach corresponded to the forward method, that is, location was significant while gender was not.

Problem 3

Data Overview

```
##      yield          nitrogen
##  Min.   : 47.0   Min.   :  0.0
##  1st Qu.:113.5   1st Qu.: 37.5
##  Median :135.0   Median : 87.5
##  Mean   :125.8   Mean   :103.4
##  3rd Qu.:142.5   3rd Qu.:162.5
##  Max.   :168.0   Max.   :300.0
```

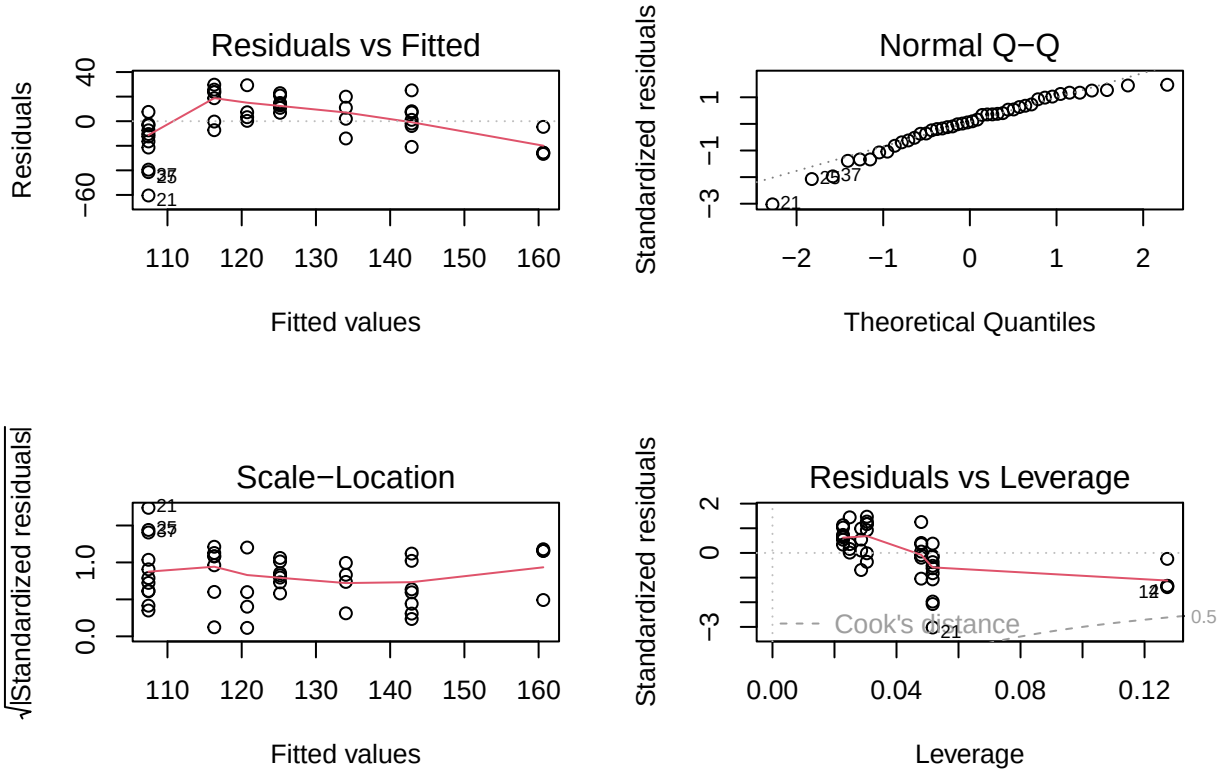


Use transformations to find a good model for predicting yield from nitrogen. Use goodness of fit to check your model.

```
model1 = lm(yield~nitrogen)
summary(model1)
```

```
##
## Call:
## lm(formula = yield ~ nitrogen)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.439 -10.939   1.534  14.082  29.697
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 107.43864    4.66622   23.02  < 2e-16 ***
## nitrogen      0.17730    0.03377    5.25 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.53 on 42 degrees of freedom
## Multiple R-squared:  0.3962, Adjusted R-squared:  0.3818
## F-statistic: 27.56 on 1 and 42 DF,  p-value: 4.713e-06
```

- No outier
- Check for the assumption of constant variance.

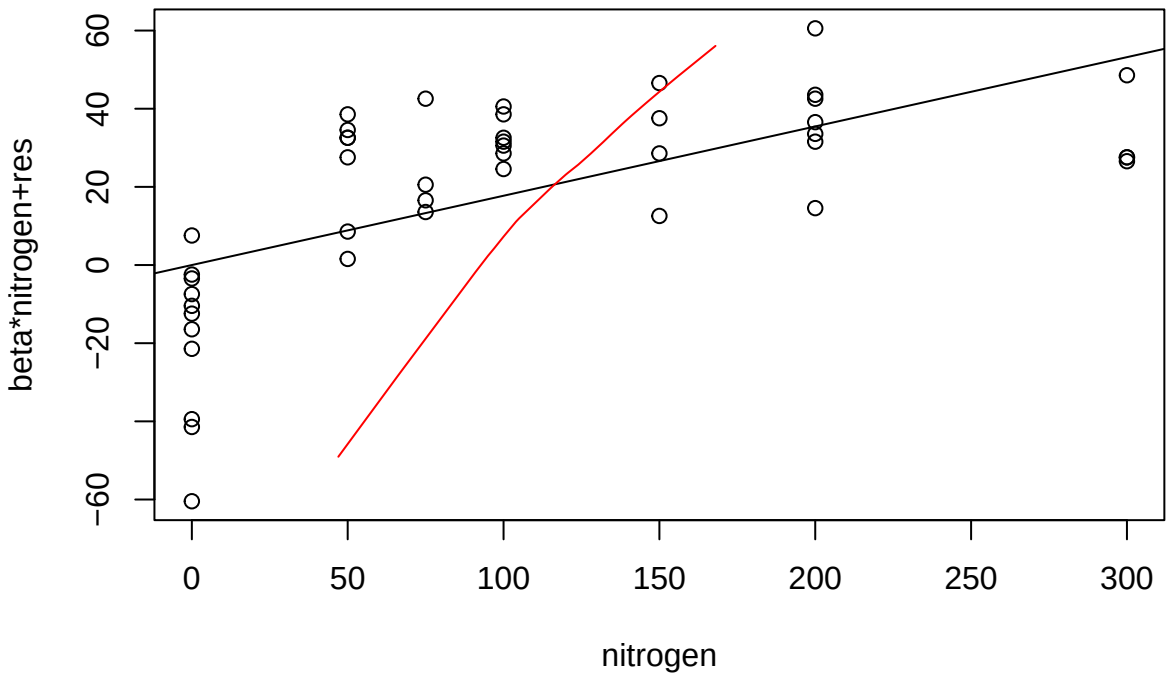


- Curvature and non-constant variance formal test

```
ncvTest(model1)
```

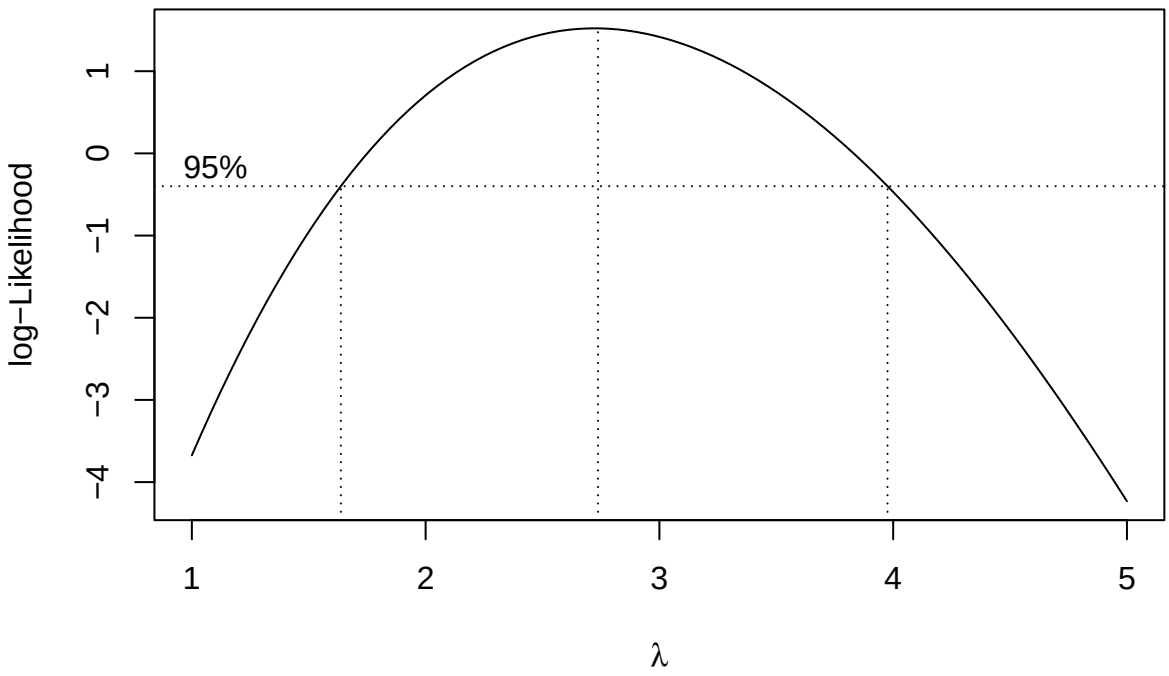
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.899564, Df = 1, p = 0.16813
```

- p-value=0.17, which is larger than the significance level of 0.05, we do not have enough evidence to reject the null hypothesis, that is, this model does not violate the assumption of constant variance.

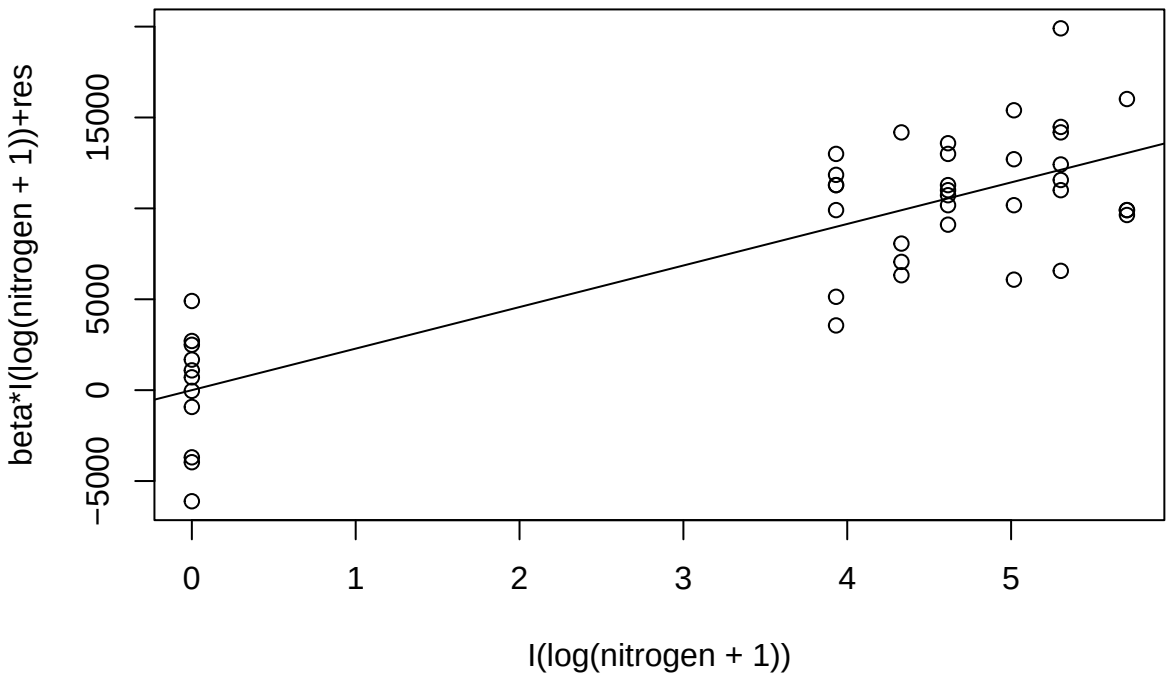


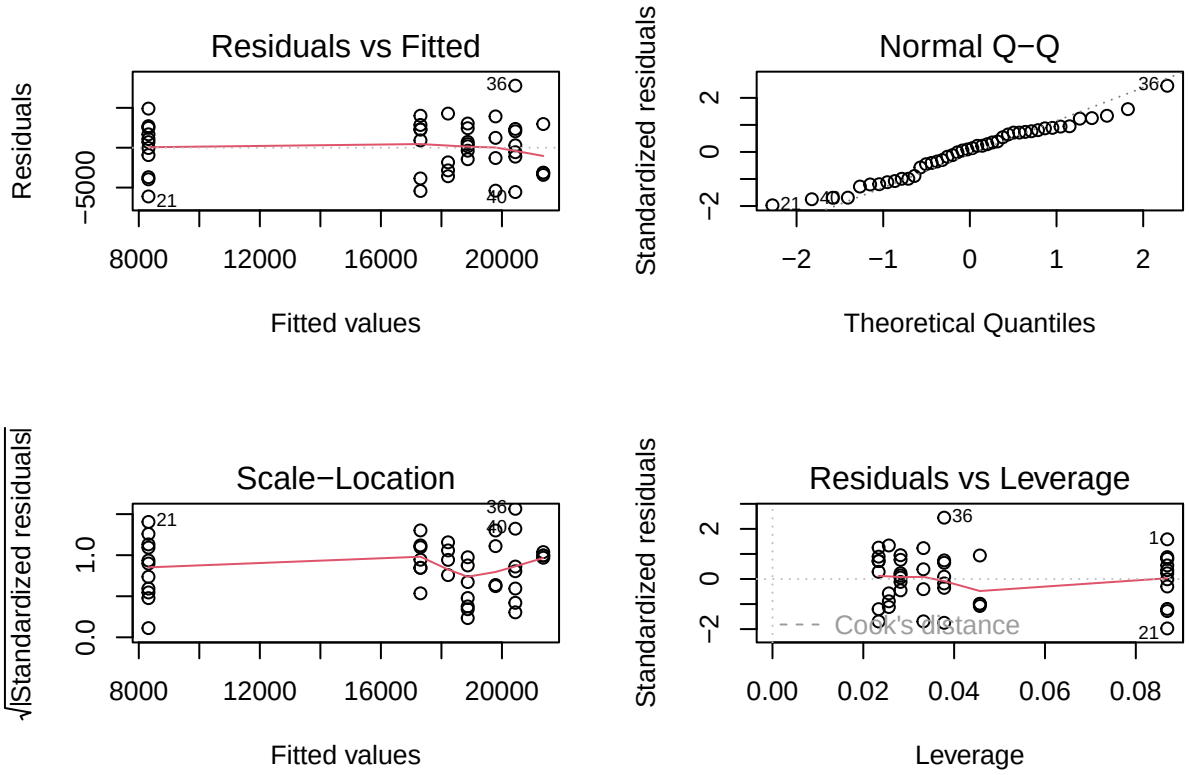
- Boxcox

```
boxcox(model1, seq(1,5))
```



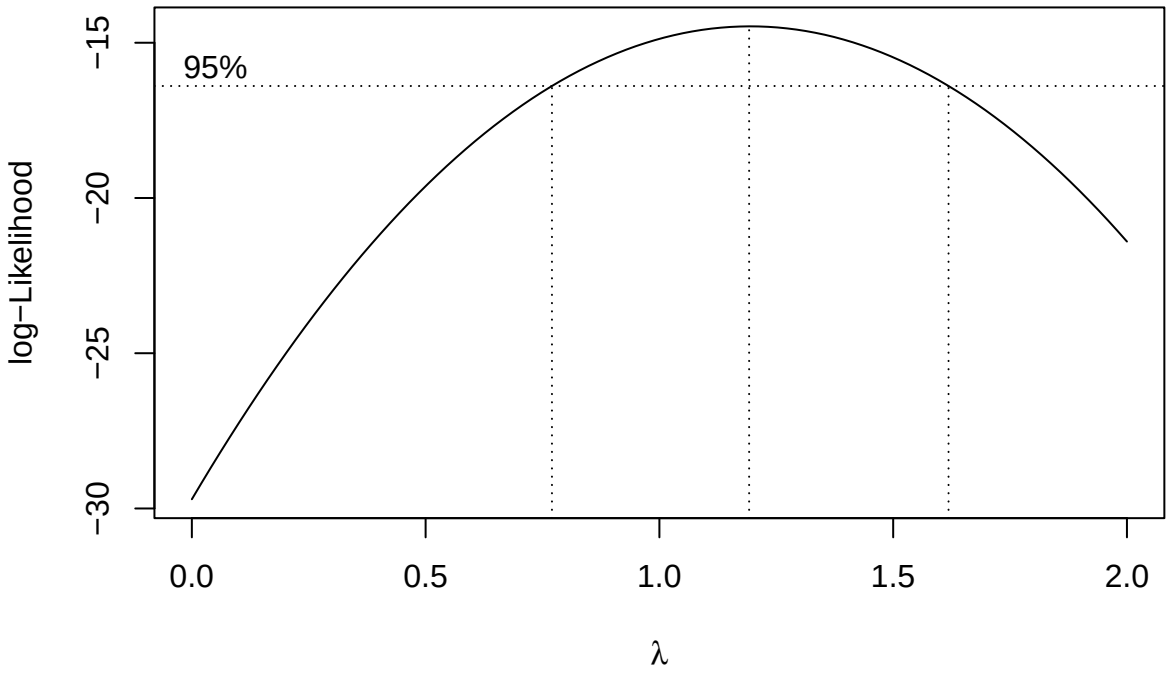
- Not only lambda value and confidence interval did not include value1, but prplot shows quadratic effect, so we do need to perform power transform the predictors or response, hence I decided to preform power transformation on response and log transformation to the variable in the model.
- Since there were several 0 in nitrogen data, I added 1 in order to perform log transformation.


beta*I(log(nitrogen + 1))+res
−5000
0
5000
15000
0
1
2
3
4
5
I(log(nitrogen + 1))



- mean curvature no longer exists in variable `nitrogen`.
- Residual plot became flatter as well.

```
boxcox(model2, seq(0,2))
```



- Value 1 was now lying in the C.I. range of lambda.
- No further transformation is required.

```
summary(model1); summary(model2)
```

```
##
## Call:
## lm(formula = yield ~ nitrogen)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.439 -10.939   1.534  14.082  29.697
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 107.43864    4.66622   23.02  < 2e-16 ***
## nitrogen      0.17730    0.03377    5.25 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.53 on 42 degrees of freedom
## Multiple R-squared:  0.3962, Adjusted R-squared:  0.3818
## F-statistic: 27.56 on 1 and 42 DF,  p-value: 4.713e-06
```

```
##
## Call:
## lm(formula = newyield ~ I(log(nitrogen + 1)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6110.8 -2917.4   372.8  2380.8  7781.7
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            8319.8      955.4   8.708 5.89e-11 ***
## I(log(nitrogen + 1))   2285.8      229.9   9.944 1.33e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3241 on 42 degrees of freedom
## Multiple R-squared:  0.7019, Adjusted R-squared:  0.6948
## F-statistic: 98.89 on 1 and 42 DF,  p-value: 1.325e-12
```

- By simply transforming the predictor and response, we enhance R^2 to about 70%.