Math 3080 § 1. Treibergs | **Speedway Example: Confidence Intervals in Simple Regression** | Name: Example Feb. 17, 2014

This note explores variability of statistics produced by simple regression. We investigate confidence intervals (and thus hypothesis tests) for the regression coefficients $\widehat{\beta_0}$, $\widehat{\beta_1}$, the mean response Y^* and prediction for next observation $\hat{Y}$.

This data is taken from Snedcore & Cochran, *Statistical Methods 6th ed.*, Iowa State University Press, Ames (1967) as quoted by Evans & Rosenthal, *Probability and Statistics: The Science of Uncertainty 2nd. ed.*, Freeman, New York (2010). This is an observational study which give the speed (in mph) of the winners of the Indianapolis Memorial Day car races 1911 - 1941 except 1917 - 1918. Year is number of years after 1911.

It is assumed that the response variable Y, "Speed"in this case, is normally distributed with a constant variance σ^2 and with a mean that depends linearly on the explanatory variable x, "Year" in this case

$$y = \beta_0 + \beta_1 x + \epsilon \qquad \text{where } \epsilon \sim N(0, \sigma^2).$$

The *mean response* when $x = x^*$ is the point on the fitted line

$$y^* = \widehat{\beta_0} + \widehat{\beta_1} x^*.$$

The *predicted value,* $\hat{y}$, is an estimator for the next observation when $x = x^*$. Its variability comes from both the error in the next observation and the variability of the response.

The estimated standard errors on the quantities.

$$\begin{aligned}
s &= \sqrt{MSE},\\
s_{\widehat{\beta_0}} &= s\sqrt{\frac{1}{n^2} + \frac{\bar{x}^2}{S_{xx}}},\\
s_{\widehat{\beta_1}} &= \frac{s}{\sqrt{S_{xx}}},\\
s_{Y^*} &= s\sqrt{\frac{1}{n^2} + \frac{(x^* - \bar{x})^2}{S_{xx}}},\\
s_{\hat{Y}} &= s\sqrt{1 + \frac{1}{n^2} + \frac{(x^* - \bar{x})^2}{S_{xx}}}
\end{aligned}$$

See my notes, `Truck Example: Simple Regression` from Feb. 15, 2014 for the derivations of formulas, which are longish but easily understandable by Math 3080 students.

In all cases, the computed statistics $\widehat{\beta_0}$, $\widehat{\beta_1}$, Y^* and $\hat{Y}$ are linear combinations of Y_i's thus are normally distributed. Their variance is a sum of squares of normal variables. Denote any of these by $\hat{\theta}$. Thus the standardized statistics

$$t = \frac{\hat{\theta} - \mu_{\hat{\theta}}}{s_{\hat{\theta}}}$$

have t-distribution with $n - 2$ degrees of freedom. We may use this statistic to make hypothesis tests. Equivalently we may use it to find confidence intervals. For example, if we wanted a two-sided CI for *theta* at the α significance level, the interval would be

$$\hat{\theta} \pm t_{\alpha/2, n-2} s_{\hat{\theta}}.$$

Data Set Used in this Analysis :

```
# M3080 - 1         Speedway Data           Feb. 17, 2014
# Treibergs
#
# Taken from Snedcore & Cochran, Statistical methods 6th ed., Iowa State
# University Press, Ames (1967) as quoted by Evans & Rosenthal, Probability
# and Statistics: The Science of Uncertainty 2nd. ed., Freeman, New York
# (2010).
#
# This is an observational study which give the speed (in mph) of the
# winners of the Indianapolis Memorial Day car races 1911 - 1941 except
# 1917 - 1918. Year is number of years after 1911.
"Year" "Speed"
 0   74.6
 1   78.7
 2   75.9
 3   82.5
 4   89.8
 5   83.3
 8   88.1
 9   88.6
10   89.6
11   94.5
12   91.0
13   98.2
14  101.1
15   95.9
16   97.5
17   99.5
18   97.6
19  100.4
20   96.6
21  104.1
22  104.2
23  104.9
24  106.2
25  109.1
26  113.6
27  117.2
28  115.0
29  114.3
30  115.1
```

R Session:

```
R version 2.13.1 (2011-07-08)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-apple-darwin9.8.0/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.41 (5874) i386-apple-darwin9.8.0]

[History restored from /Users/andrejstreibergs/.Rapp.history]

> tt=read.table("M3082DataSpeedway.txt",header=T)
> attach(tt)
> tt
   Year Speed
1     0  74.6
2     1  78.7
3     2  75.9
4     3  82.5
5     4  89.8
6     5  83.3
7     8  88.1
8     9  88.6
9    10  89.6
10   11  94.5
11   12  91.0
12   13  98.2
13   14 101.1
14   15  95.9
15   16  97.5
16   17  99.5
17   18  97.6
18   19 100.4
19   20  96.6
20   21 104.1
21   22 104.2
22   23 104.9
```

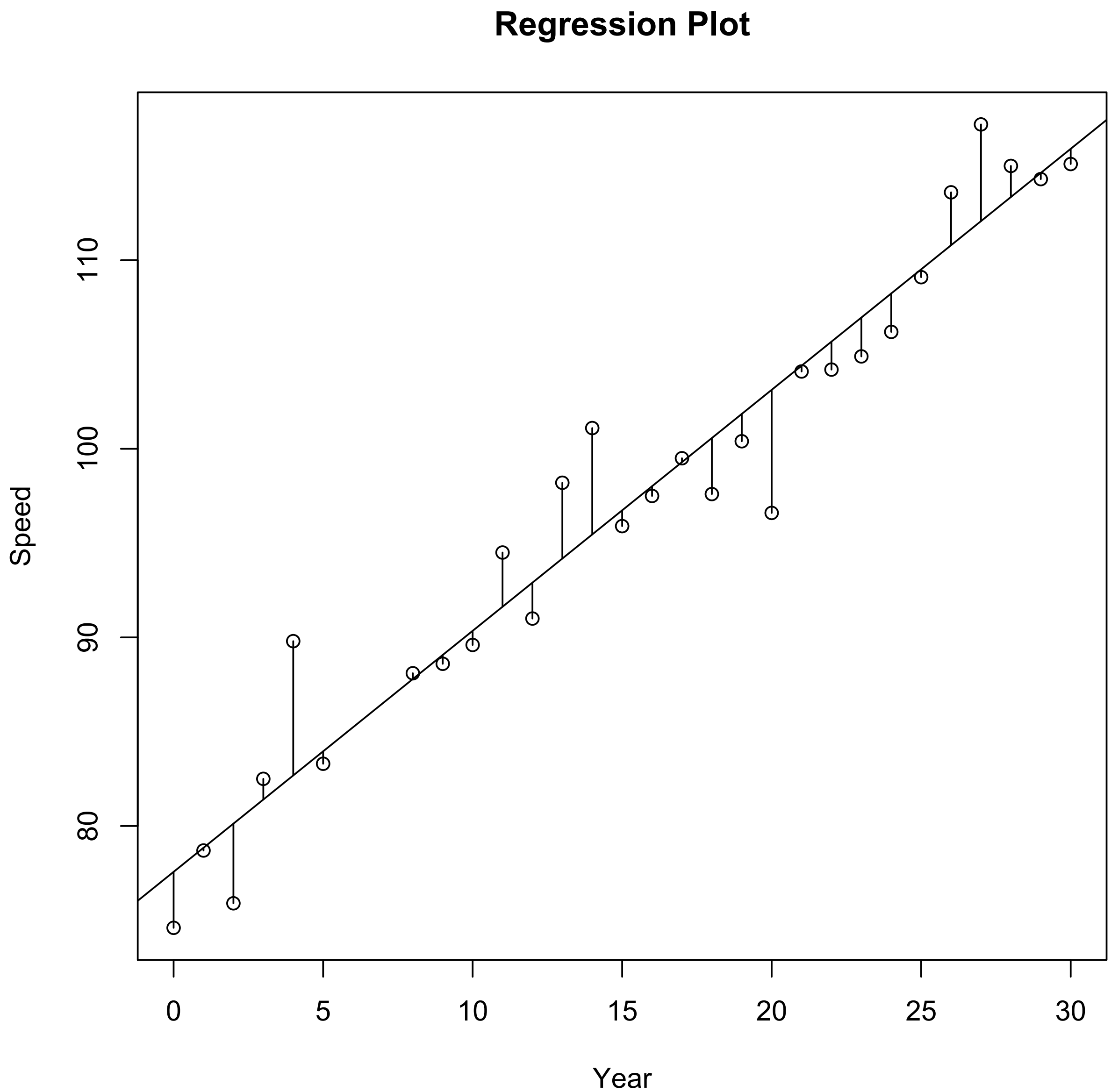
```
23   24 106.2
24   25 109.1
25   26 113.6
26   27 117.2
27   28 115.0
28   29 114.3
29   30 115.1

> ############### RUN ANOVA ####################################

> f1=lm(Speed ~ Year); summary(f1)

Call:
lm(formula = Speed ~ Year)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.5267 -1.4826 -0.4695  1.0981  7.1202 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 77.56810    1.11784   69.39   <2e-16 ***
Year         1.27793    0.06219   20.55   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 2.999 on 27 degrees of freedom
Multiple R-squared: 0.9399,Adjusted R-squared: 0.9377 
F-statistic: 422.3 on 1 and 27 DF,  p-value: < 2.2e-16 

> ##########  SCATTERPLOT WITH VARIATION SEGMENTS  #############
> plot(Speed~Year, main="Regression Plot"); abline(f1)
> segments(Year,fitted(f1),Year,Speed)
> 
> ##############  LIST FITTED AND RESIDUALS  ##################
> fitted(f1)
        1         2         3         4         5         6         7 
 77.56810  78.84604  80.12397  81.40190  82.67983  83.95776  87.79156 
        8         9        10        11        12        13        14 
 89.06949  90.34742  91.62535  92.90328  94.18121  95.45914  96.73707 
       15        16        17        18        19        20        21 
 98.01501  99.29294 100.57087 101.84880 103.12673 104.40466 105.68259 
       22        23        24        25        26        27        28 
106.96053 108.23846 109.51639 110.79432 112.07225 113.35018 114.62811 
       29 
115.90605 
```

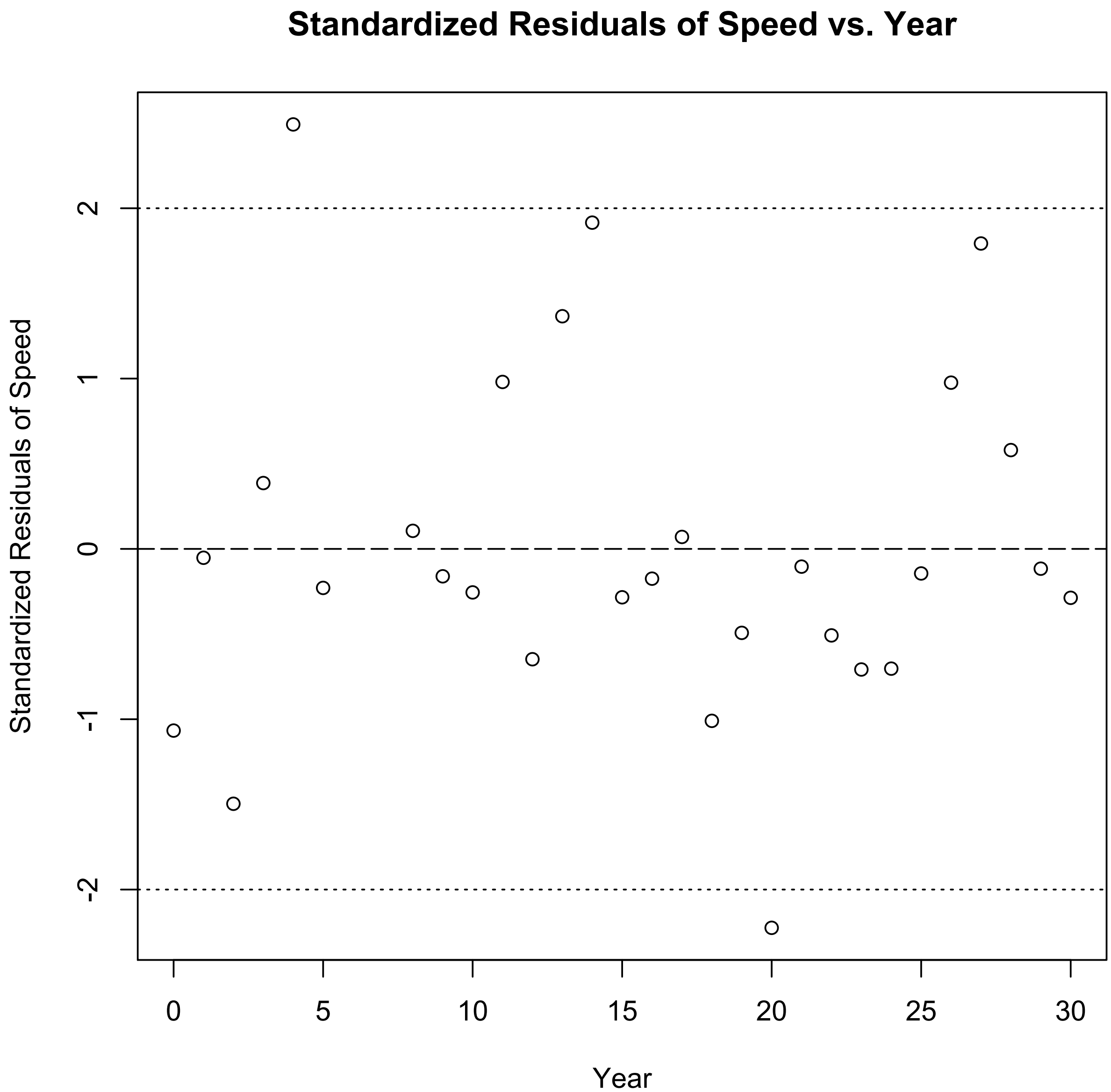
```
> resid(f1)
         1          2          3          4          5          6          7 
-2.9681043 -0.1460357 -4.2239670  1.0981016  7.1201703 -0.6577611  0.3084448 
         8          9         10         11         12         13         14 
-0.4694865 -0.7474179  2.8746507 -1.9032806  4.0187880  5.6408566 -0.8370747 
        15         16         17         18         19         20         21 
-0.5150061  0.2070626 -2.9708688 -1.4488002 -6.5267315 -0.3046629 -1.4825943 
        22         23         24         25         26         27         28 
-2.0605256 -2.0384570 -0.4163883  2.8056803  5.1277489  1.6498176 -0.3281138 
        29 
-0.8060451 
>
> ############## PLOT STD. RESIDUALS VS YEAR, QQ PLOT  ############

> plot(rstandard(f1)~Year,main="Standardized Residuals of Speed
 vs.Year",ylab="Standardized Residuals of Speed")
> abline(h=c(0,-2,2),lty=c(5,3,3))
> qqnorm(rstandard(f1),ylab="Standardized Residuals of Speed")
> abline(h=c(0,-2,2),lty=c(5,3,3)); abline(0,1,col=2)
>
> #######  CANNED ESTIMATORS FOR MEAN RESPONSE WITH CI'S  ########
> predict(f1,interval="confidence")
         fit       lwr       upr
1   77.56810  75.27448  79.86173
2   78.84604  76.66213  81.02995
3   80.12397  78.04773  82.20020
4   81.40190  79.43095  83.37284
5   82.67983  80.81139  84.54827
6   83.95776  82.18857  85.72696
7   87.79156  86.29409  89.28902
8   89.06949  87.65116  90.48781
9   90.34742  89.00077  91.69407
10  91.62535  90.34167  92.90903
11  92.90328  91.67252  94.13404
12  94.18121  92.99198  95.37045
13  95.45914  94.29882  96.61946
14  96.73707  95.59210  97.88205
15  98.01501  96.87126  99.15876
16  99.29294  98.13625 100.44962
17 100.57087  99.38755 101.75419
18 101.84880 100.62605 103.07155
19 103.12673 101.85293 104.40053
20 104.40466 103.06953 105.73980
21 105.68259 104.27719 107.08800
22 106.96053 105.47719 108.44387
23 108.23846 106.67066 109.80626
24 109.51639 107.85860 111.17418
25 110.79432 109.04187 112.54677
26 112.07225 110.22117 113.92333
27 113.35018 111.39712 115.30324
28 114.62811 112.57021 116.68602
29 115.90605 113.74085 118.07124
```

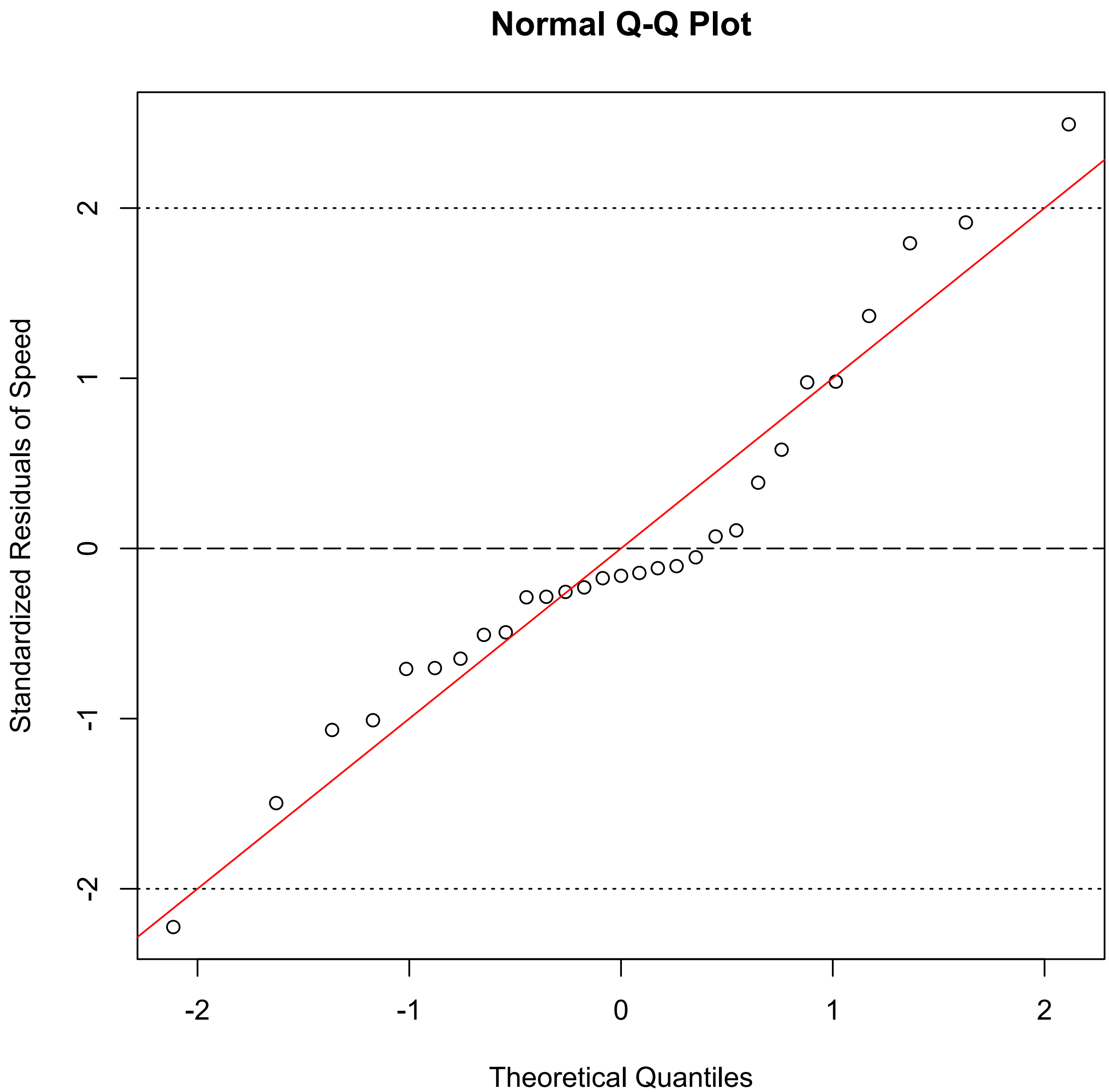
```
> #######  CANNED ESTIMATORS FOR PREDICTIONS WITH CI'S  ########
> predict(f1,interval="prediction")
         fit       lwr       upr
1   77.56810  71.00178  84.13443
2   78.84604  72.31723  85.37485
3   80.12397  73.63038  86.61755
4   81.40190  74.94121  87.86259
5   82.67983  76.24967  89.10999
6   83.95776  77.55574  90.35979
7   87.79156  81.45923  94.12388
8   89.06949  82.75541  95.38356
9   90.34742  84.04906  96.64578
10  91.62535  85.34015  97.91055
11  92.90328  86.62868  99.17789
12  94.18121  87.91462 100.44780
13  95.45914  89.19797 101.72031
14  96.73707  90.47873 102.99542
15  98.01501  91.75689 104.27313
16  99.29294  93.03244 105.55343
17 100.57087  94.30540 106.83634
18 101.84880  95.57576 108.12184
19 103.12673  96.84354 109.40992
20 104.40466  98.10875 110.70057
21 105.68259  99.37141 111.99378
22 106.96053 100.63153 113.28952
23 108.23846 101.88913 114.58778
24 109.51639 103.14425 115.88853
25 110.79432 104.39690 117.19174
26 112.07225 105.64711 118.49739
27 113.35018 106.89493 119.80544
28 114.62811 108.14037 121.11586
29 115.90605 109.38347 122.42862
Warning message:
In predict.lm(f1, interval = "prediction") :
  Predictions on current data refer to _future_ responses

> pp = predict(f1, int="p")
Warning message:
In predict.lm(f1, int = "p") :
  Predictions on current data refer to _future_ responses
> pc = predict(f1, int="c")

> ####### PLOT CI FOR MEAN RESPONSE AND PREDICTION ##########

> plot(Year, Speed, ylim=range(Speed,pp), main="CI on Mean
Response and Prediction of Speed")
> matlines(Year,pc,lty=c(1,2,2), col=2)
> matlines(Year,pp,lty=c(1,3,3), col=4)
> legend(5,120,legend=c("Mean Response","Prediction"),
  fill=c(2,4),title="Confidence Intervals")
>
```

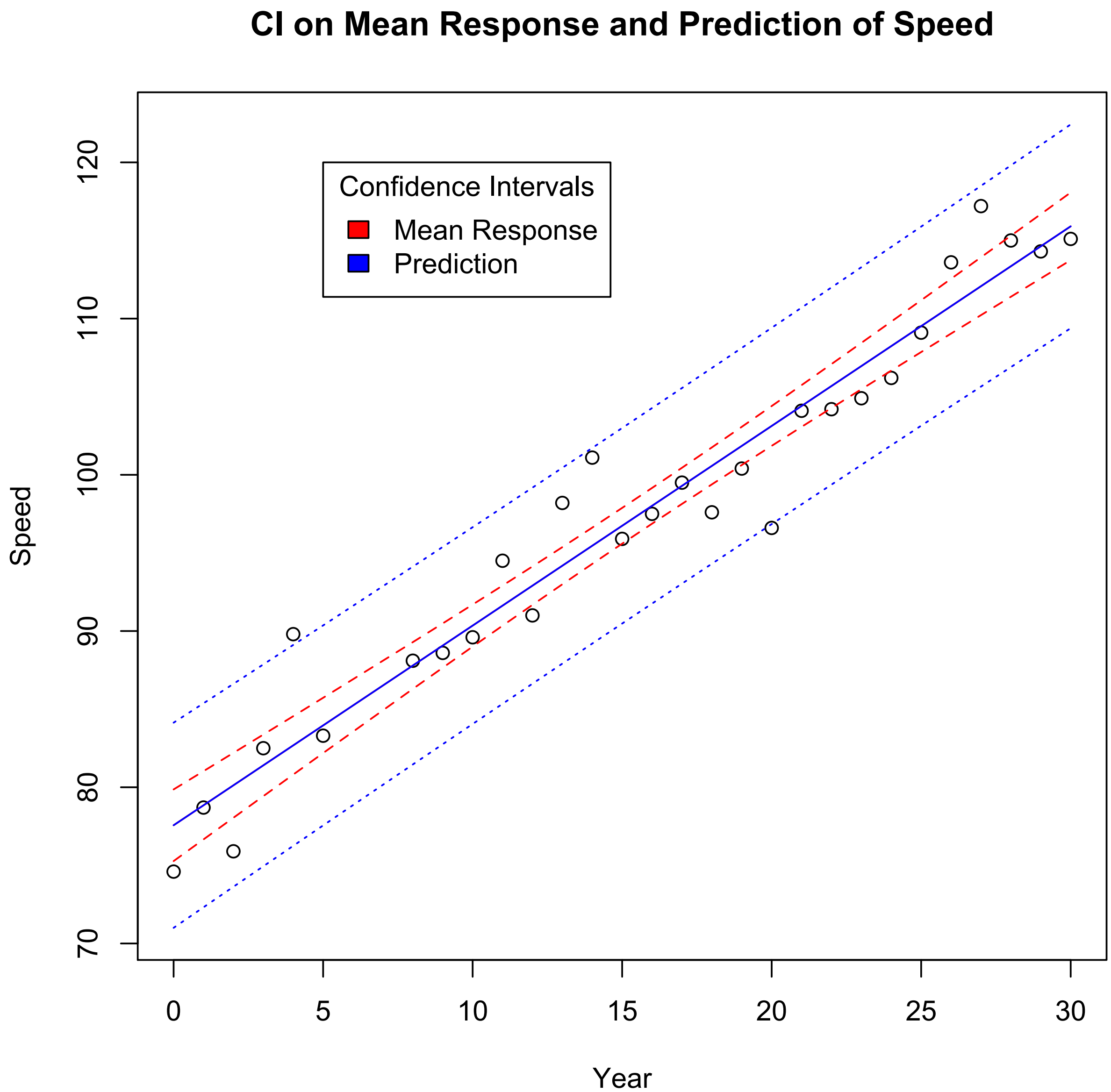
```
> anova(f1)
Analysis of Variance Table

Response: Speed
          Df Sum Sq Mean Sq F value    Pr(>F)
Year       1 3797.0    3797  422.27 < 2.2e-16 ***
Residuals 27  242.8       9
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> ################  DO REGRESSION "BY HAND"  #################

> ybar = mean(Speed); ybar
[1] 97.48621
> xbar = mean(Year); xbar
[1] 15.58621
> n = length(Speed); n
[1] 29
> Sxx = (n-1)*var(Year); Sxx
[1] 2325.034
> Sxy = sum((Speed-ybar)*(Year-xbar)); Sxy
[1] 2971.234
> beta1 = Sxy/Sxx; beta1
[1] 1.277931
> beta0 = ybar-beta1*xbar; beta0
[1] 77.5681
> SSE = sum(Speed^2)-beta0*sum(Speed)-beta1*sum(Speed*Year); SSE
[1] 242.7808
> SST = sum((Speed-ybar)^2); SST
[1] 4039.814
> SSR=SST-SSE; SSR
[1] 3797.034
> MSE=SSE/(n-2); MSE
[1] 8.99188
> R2=SSR/SST; R2; sqrt(R2)
[1] 0.939903
[1] 0.9694859
> R2adj = 1-(n-1)*SSE/((n-2)*SST); R2adj
[1] 0.9376772
> ((n-1)*R2-1)/(n-2)
[1] 0.9376772
>
```

```
> ###  COMPUTE STANDARD ERRORS. COMPARE TO CANNED SUMMARY  ####
> s=sqrt(MSE); s
[1] 2.998646
> s1=s/sqrt(Sxx); s1
[1] 0.06218857
> s0=s*sqrt(1/n + xbar^2/Sxx); s0
[1] 1.117844
> ### STD. ERROR FOR MEAN RESPONSE AND PREDICTION FOR X*=7 ####
> xstar=7
> sstar=s*sqrt(1/n + (xbar-xstar)^2/Sxx); sstar
[1] 0.7714806
> spred=s*sqrt(1+ 1/n + (xbar-xstar)^2/Sxx); spred
[1] 3.096298
>
> ############# 0.05 LEVEL TWO-SIDED CI'S FOR STATISTICS  ######
> tcrit=qt(.025,n-2,lower.tail=F); tcrit
[1] 2.051831
> ################## 0.05 LEVEL TWO-SIDED CI FOR  BETA1  #######
> c(beta1-s1*tcrit,beta1+s1*tcrit)
[1] 1.150331 1.405532
> ################## 0.05 LEVEL TWO-SIDED CI FOR  BETA0  #######
> c(beta0-s0*tcrit,beta0+s0*tcrit)
[1] 75.27448 79.86173
> ##### 0.05 LEVEL TWO-SIDED CI FOR MEAN RESPONSE AT  X*=7  ####
> c(beta0+beta1*xstar-sstar*tcrit,beta0+beta1*xstar+sstar*tcrit)
[1] 84.93068 88.09657
> ##### 0.05 LEVEL TWO-SIDED CI FOR PREDICTION AT  X*=7  #######
> c(beta0+beta1*xstar-spred*tcrit,beta0+beta1*xstar+spred*tcrit)
[1] 80.16054 92.86670
> ### COMPARE TO CANNED CI MEAN RESPONSE AT YEAR[19]=20  #######
> xstar=Year[19]; xstar
[1] 20
> sstar=s*sqrt(1/n + (xbar-xstar)^2/Sxx); sstar
[1] 0.6208125
> c(beta0+beta1*xstar-sstar*tcrit,beta0+beta1*xstar+sstar*tcrit)
[1] 101.8529 104.4005
> pc[19,1:3]
     fit      lwr      upr
103.1267 101.8529 104.4005
> ###### COMPARE TO CANNED CI PREDICTION AT YEAR[19]=20  #######
> spred=s*sqrt(1+ 1/n + (xbar-xstar)^2/Sxx); spred
[1] 3.062236
> c(beta0+beta1*xstar-spred*tcrit,beta0+beta1*xstar+spred*tcrit)
[1]  96.84354 109.40992
> pp[19,1:3]
      fit       lwr       upr
103.12673  96.84354 109.40992
```


Regression Plot
Speed
80
90
100
110
0
5
10
15
20
25
30
Year


Standardized Residuals of Speed vs. Year
Standardized Residuals of Speed
Year
2
1
0
-1
-2
0
5
10
15
20
25
30


Normal Q-Q Plot
Standardized Residuals of Speed
Theoretical Quantiles
-2
-1
0
1
2
-2
-1
0
1
2


CI on Mean Response and Prediction of Speed
Confidence Intervals
Mean Response
Prediction
Speed
120
110
100
90
80
70
Year
0
5
10
15
20
25
30