

Stat262 Hw1. Solution.

(by Pei Wang, 4/8/04)

Dear students:

Next time, please first summarize your answer, and then attach the code in the end of the hw file. Thanks.

Problem1.

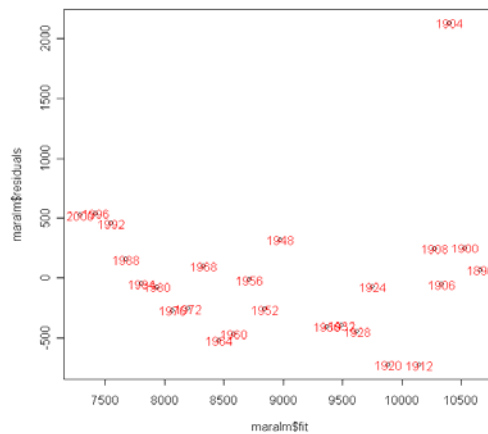
* Typo: the time for year 2000 appears as "2:10.11"... it probably means "2:10:11". I used this version for later analysis.

(a). Fit linear model, get

$$\text{Time} = 72193.80 - 32.45 * \text{Year}.$$

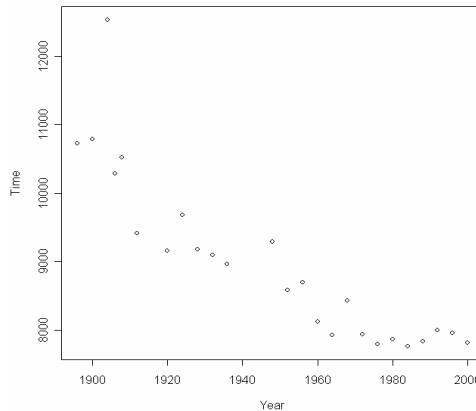
Here "Time" is in second units. From the model we can see that the winning time decreases around half a minute each year.

(b). Plot residuals v.s. fitted value:



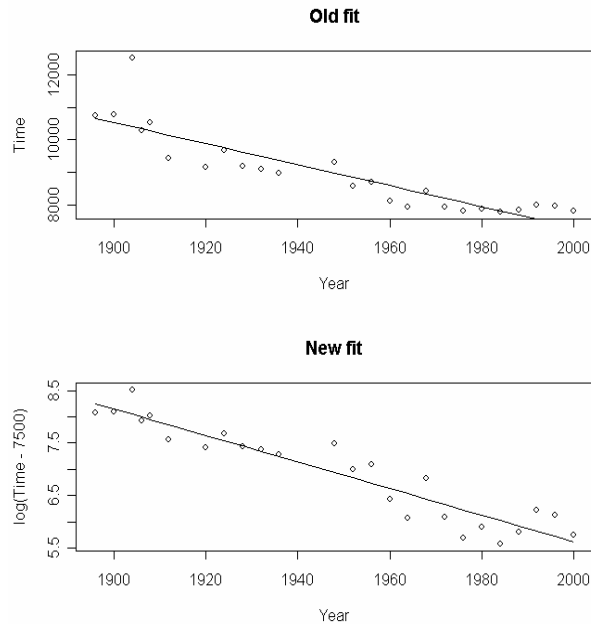
point of 1904 may be a potential outlier.

(c) Plot Year v.s. Time:



The plot doesn't look very linear, especially when Year goes beyond 1980.

So try other models. e.g. $\log(\text{Time}-7500) \sim \text{Year}$ (the shape of the previous plot suggests some exponential function) :



(d) For the old model, predicted time at year 2050 is 5666.245sec, that's roughly 1.57 hour, a little bit too fast for human beings.

For the new model, predicted time at year 2050 is 7577.54sec, that's roughly 2.10 hour, which makes more sense.

(e) Adding dummy variables of continent to the model in (a). The fit result is:

Call:

```
lm(formula = Time ~ Year + Year:continent)
```

Residuals:

Min	1Q	Median	3Q	Max
-731.45	-325.14	-34.14	184.91	1863.73

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.791e+04	7.480e+03	9.079	1.02e-08 ***
Year	-3.032e+01	3.804e+00	-7.970	8.74e-08 ***
Year:continent1	1.006e-02	1.489e-01	0.068	0.947
Year:continent2	2.561e-01	1.865e-01	1.373	0.184

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 575 on 21 degrees of freedom
Multiple R-Squared: 0.8081, Adjusted R-squared: 0.7807

F-statistic: 29.47 on 3 and 21 DF, p-value: 1.021e-07

We can see that two interact-term of dummy variables (Year:continent1 , Year:continent2) are not significant in this model. We can do a F-test to see weather the continent effect plays a role here:

Analysis of Variance Table

```
Model 1: Time ~ Year
Model 2: Time ~ Year + Year:continent
  Res.Df  RSS Df Sum of Sq  F      Pr(>F)
1     23 7874361
2     21 6941968 2   932393  1.4103 0.2663
```

So we conclude that there is no significant continent effect.

(Conclusion would be the same when we introduce dummy variables to the model in (c)).

Problem2.

(a) Weight= -105.011+ 1.018 * Height.

(b) Introduce gender as dummy variable. The result of the fit:

Call:

```
lm(formula = Weight ~ Height * Gender)
```

Residuals:

```
  Min     1Q  Median     3Q    Max
-20.187 -5.957 -1.439  4.955 43.355
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -43.81929   13.77877  -3.180  0.00156 **
Height       0.63334    0.08351   7.584 1.63e-13 ***
Gender1     -17.13407   19.56250  -0.876  0.38152
Height:Gender1 0.14923   0.11431   1.305  0.19233
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.795 on 503 degrees of freedom

Multiple R-Squared: 0.5682, Adjusted R-squared: 0.5657

F-statistic: 220.7 on 3 and 503 DF, p-value: < 2.2e-16

The dummy variable of Gender as well as the interaction term of Gender*Height do not have significant p-values in the above summary. But we can do a F-test to investigate whether the bigger model improves the fit or not:

Analysis of Variance Table

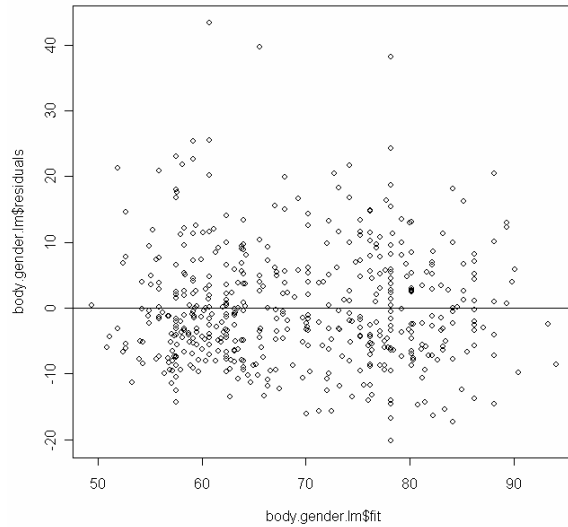
```
Model 1: Weight ~ Height
Model 2: Weight ~ Height * Gender
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	505 43753				
2	503 38912	2	4841 31.292	1.553e-13	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

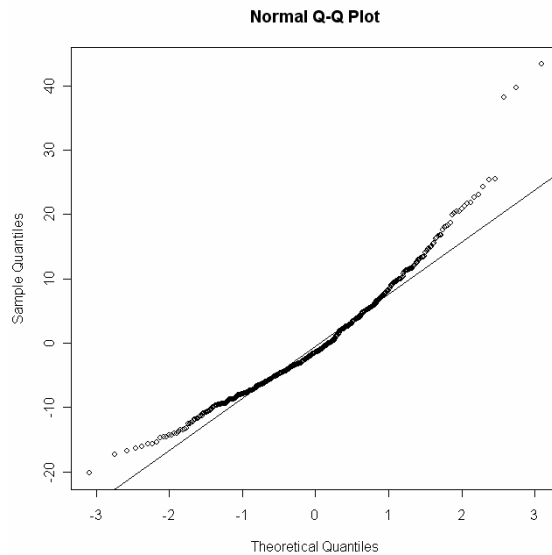
The answer is YES, the big model does improve the fit significantly.

(c). Plot the residuals v.s. fitted weight:



There is no obvious pattern for higher order effect. The variance is quite constant.

(d) qqplot:



It's a little skewing to the right. Especially the three points at the up-right corner is not consistent with the normality assumption.

```
##### Code in R #####
```

```
## Problem1.
```

```
##(a)
```

```
marathon<-read.table("marathon.txt", header=T, sep="\t")  
maralm<-lm(Time~Year, data=marathon)
```

```
##(b)
```

```
plot(maralm$fit, maralm$residuals)  
text(maralm$fit, maralm$residuals, marathon$Year, col=2)
```

```
##(c)
```

```
attach(marathon)  
plot(Year, Time)
```

```
maralm2<-lm(log(Time-7500)~Year)  
par(mfrow=c(2,1))  
plot(Year, Time)  
points(Year, maralm$fit, type="l")  
title(main="Old fit")  
plot(Year, log(Time-7500))  
points(Year, maralm2$fit, type="l")  
title(main="New fit")
```

```
##(d)
```

```
predict.lm(maralm, data.frame(Year=2050))  
exp(predict.lm(maralm2, data.frame(Year=2050)))+7500
```

```
##(e)
```

```
## make factor vector of continents. Since only 25 samples, I didn't struggle to find a  
## smart way(maybe possible with perl), but did this manually. As suggested in the
```

```
##problem: 0--- Africa; 1---Europe, Asia; 2---America  
continent<-c(1,1,2,2,2,0,1,1,1,2,1,2,1,1,0,0,0,2,1,1,1,1,0,0)  
continent<-as.factor(continent)
```

```
## For model in (a)
```

```
maralm3<-lm(Time~Year+Year:continent)  
summary(maralm3)
```

```
anova(maralm, maralm3)
```

```
# Problem2.
```

```
##(a)
```

```
Bodydata<-read.table("body_table.txt", sep="," , head=T)
```

```
attach(Bodydata)
bodylm<-lm(Weight~Height)
summary(bodylm)
```

```
 #(b)
Gender<-as.factor(Gender)
body.gender.lm<-lm(Weight~Height*Gender)
summary(body.gender.lm)
```

```
anova(bodylm, body.gender.lm)
```

```
 #(c)
plot(body.gender.lm$fit, body.gender.lm$residuals)
abline(h=0)
```

```
 #(d)
qqnorm(body.gender.lm$residuals)
qqline(body.gender.lm$residuals)
```