ACT-2003 Modèles linéaires en actuariat

Exercices supplémentaires

Transformations pour stabiliser la variance et méthode de Box-Cox

Marie-Pier Côté

30 septembre 2013

1. Dans un graphiques des résidus en fonction des valeurs prédites, on observe de l'hétéroscédasticité. Après une analyse plus poussée, on note que la variance de $\hat{\varepsilon}_i$ est approximativement proportionnelle à $E[Y_i]^4$. Proposer une transformation g de la variable réponse qui permettra de stabiliser la variance.

2. Les données suivantes présentent le nombre moyen de bactéries vivantes dans une boîte de conserve de nourriture et le temps (en minutes) d'exposition à une chaleur de 300^oF.[1]

Nombre de bactéries	Temps d'exposition (min)
175	1
108	2
95	3
82	4
71	5
50	6
49	7
31	8
28	9
17	10
16	11
11	12

a) Tracer un nuage de points des données. Est-ce qu'un modèle de régression linéaire semble adéquat ?

b) Ajuster au données un modèle de régression linéaire. Calculer les statistiques sommaires et produire les graphiques de résidus. Quelles sont vos conclusions par rapport à la validité du modèle de régression ?

c) Identifier une transformation pour ces données afin d'utiliser adéquatement les méthodes de régression. Ajuster ce nouveau modèle et tester la validité de la régression.

[1] Source : D. Montgomery, E.A. Peck et G.G. Vining (2012). Introduction to Linear Regression Analysis. Fifth Edition. Wiley.

Solutions

1. Utiliser l'approximation de Taylor de premier ordre pour montrer que la variance de $g(Y) = 1/Y$ est approximativement constante.

2. (a) Figure 1 shows a scatter plot of the number of bacteria versus the minutes of exposure. The plot shows a straight line would be a reasonable model, but an even better model would be capturing the curvature. In fact, the plot shows that when the canned food is exposed to 300^o F for a long time, there is ultimately no bacteria left. This suggests a model that would capture the asymptotic behavior of the number of bacteria when the number of minutes of exposure increases. A linear model would continue to drive down the number of bacteria, eventually leading to negative values, which is nonsensical in this context.

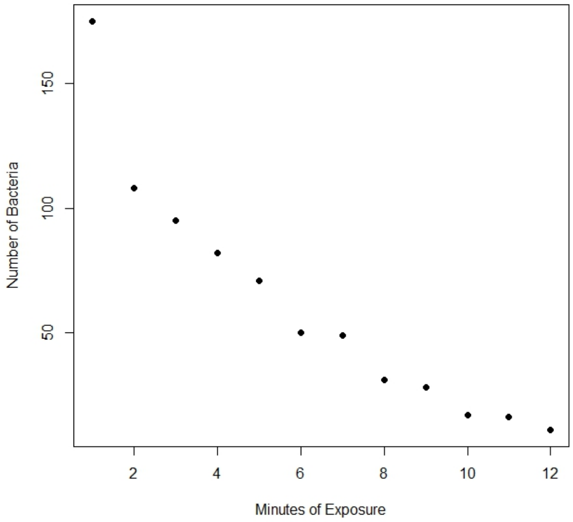

FIGURE 1: Scatter Plot of the Number of Bacteria versus the Minutes of Exposure to 300^o F

(b) A simple linear model is fitted to the data using `R`. Here is a summary of the model :

```
> fit1 <- lm(bact~min)
> summary(fit1)

Call:
lm(formula = bact ~ min)

Residuals:
    Min      1Q  Median      3Q     Max
-17.323  -9.890  -7.323   2.463  45.282
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   142.20      11.26  12.627 1.81e-07 ***
min           -12.48       1.53  -8.155 9.94e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.3 on 10 degrees of freedom
Multiple R-squared: 0.8693,Adjusted R-squared: 0.8562
F-statistic: 66.51 on 1 and 10 DF,  p-value: 9.944e-06
```

The fitted model is

$$\hat{y} = 142.20 - 12.48x,$$

where the parameters of the model are estimated by the best linear unbiased estimators. The ANOVA table is obtained using R :

```
> anova(fit1)
Analysis of Variance Table

Response: bact
          Df  Sum Sq Mean Sq F value    Pr(>F)
min        1 22268.8 22268.8  66.512 9.944e-06 ***
Residuals 10  3348.1   334.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In order to test for the significance of regression, we use the F-statistic. The F-statistic is 66.512, and it has 1 and 10 degrees of freedom, so the p-value is

$$P[F_{(1,10)} > 66.512] = 9.944 \times 10^{-6}.$$

Since the p-value is much smaller than 1%, there is enough evidence to reject the null hypothesis that $\beta_1 = 0$ at the 1% level. The simple linear model is significant.

The value of R^2 is 86.93%. This is a high coefficient of correlation, it means that about 87% of the variation in the number of bacteria in the canned food is explained by the minutes of exposure to 300^oF. The model seems to perform well.

The Q-Q Plot of the studentized residuals is shown in Figure 2. The line represents when the empirical quantiles are exactly equal to the standard normal quantiles. The normality assumption is seriously violated as the dots are clearly not on a straight line. This means there are serious flaws in the model, including the fact that the hypothesis tests are not reliable.

Figure 3 shows a plot of the studentized residuals versus the fitted values. The plot suggests a clear curve, which is usually an indicator of non-linearity. This is in line with the previous comments.

Finally, this model is inadequate and transformations on the response variables are required.

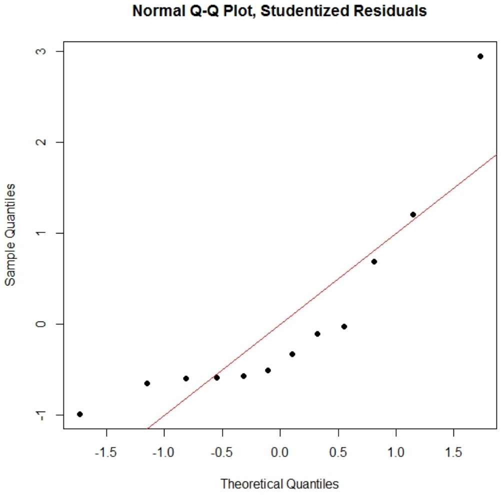

Figure 2: Q-Q Plot for Simple Linear Model in Problem 5-3

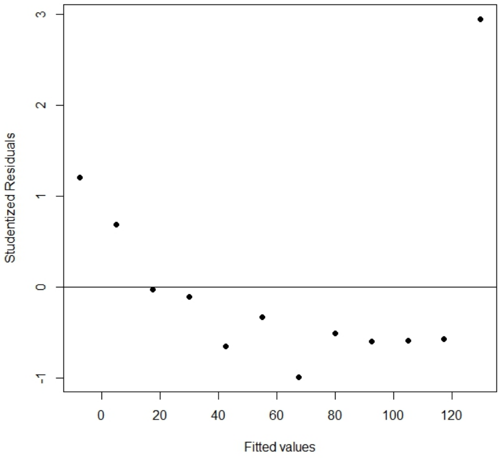

Figure 3: Residuals versus the Fitted Values for Simple Linear Model in Problem 5-3

(c) The Box-Cox method is used to determine which transformation is optimal. Figure 4 shows the plot of the log-likelihood function in terms of λ, for two different ranges of λ. It was obtained with the `R` commands :

```
boxcox(bact~minutes,lambda=seq(-2,2,len=20),plotit=TRUE)
boxcox(bact~minutes,lambda=seq(-0.2,0.5,len=20),plotit=TRUE)
```

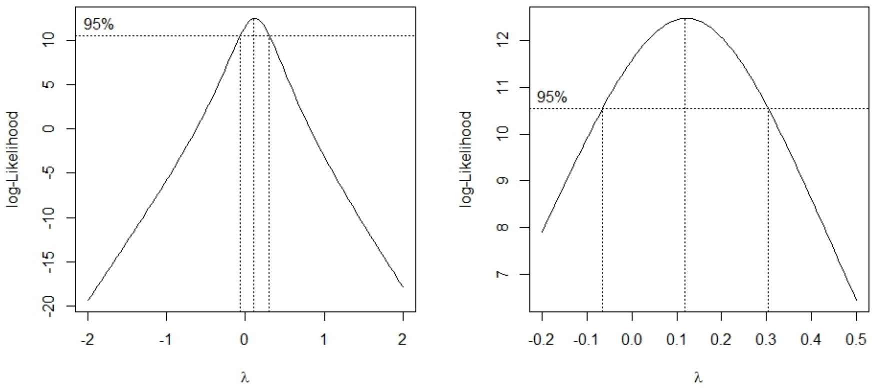

FIGURE 4: Log-likelihood versus λ in the Box-Cox method for Problem 5-3

Note that the maximum is around 0.1 and 0 is included in the 95% confidence interval for λ. Therefore, it is preferable to use 0 as this is a common transformation, it represents the logarithm transformation. Let $y^* = \ln(y)$. A simple linear model is fitted to the transformed data. The output is the following :

```
> fit2 <- lm(logbact~minutes)
> summary(fit2)

Call:
lm(formula = logbact ~ minutes)

Residuals:
      Min        1Q    Median        3Q       Max
-0.184303 -0.083994  0.001453  0.072825  0.206246

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.33878    0.07409   72.05 6.47e-15 ***
minutes     -0.23617    0.01007  -23.46 4.49e-10 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.1204 on 10 degrees of freedom
```

```
Multiple R-squared: 0.9822,Adjusted R-squared: 0.9804
F-statistic: 550.3 on 1 and 10 DF,  p-value: 4.489e-10
```

The fitted model is

$$\hat{y}^* = 5.33878 - 0.23617x,$$

where the parameters of the model are estimated by the best linear unbiased estimators. Figure 5 is a scatter plot of the transformed response variable versus the covariate, along with the fitted line. The scatter plot looks much more linear now than in (a).

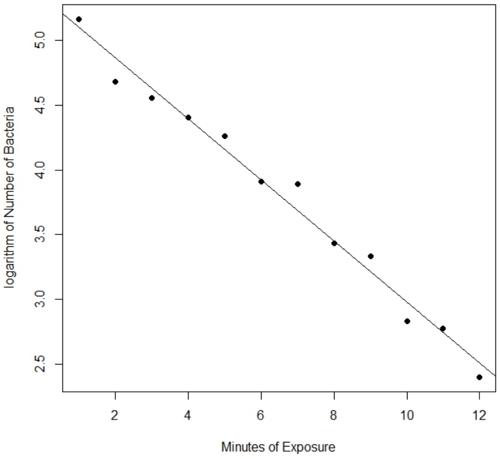

FIGURE 5: Scatter Plot of the Logarithm of the Number of Bacteria versus the Minutes of Exposure to 300^o F

The ANOVA table is obtained using `R` :

```
> anova(fit2)
Analysis of Variance Table

Response: logbact
          Df Sum Sq Mean Sq F value    Pr(>F)
minutes    1 7.9761  7.9761  550.33 4.489e-10 ***
Residuals 10 0.1449  0.0145
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The F-statistic for the test of significance of regression is 550.33, and it has 1 and 10 degrees of freedom, so the p-value is

$$P[F_{(1,10)} > 550.33] = 4.489 \times 10^{-10}.$$

Since the p-value is much smaller than 1%, there is enough evidence to reject the null hypothesis that $\beta_1 = 0$ at the 1% level. This model is significant.

The value of R^2 is very high at 98.22%. This means that about 98% of the variation in the log of the number of bacteria in the canned food is explained by the minutes of exposure to 300^oF. The model seems to perform very well, better than the model proposed in (b).

The Q-Q Plot of the studentized residuals is shown in Figure 6. The dots are beautifully aligned with the standard normal quantiles. The normality assumption is appropriate. Figure 7 shows a plot of the studentized residuals versus the fitted values. The dots can be contained in horizontal bands and looks randomly scattered.

Finally, this model is adequate and the transformation used on the response variables fixed the problems in the model.

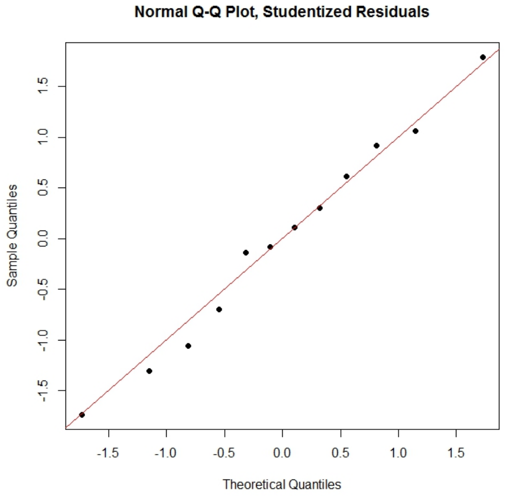

FIGURE 6: Q-Q Plot of Model for the Logarithm of the Number of Bacteria in Problem 5-3

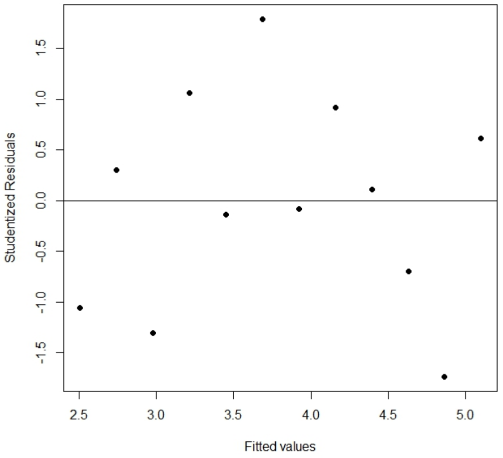

Figure 7: Residuals versus the Fitted Values for Model for the Logarithm of the Number of Bacteria in Problem 5-3