Tutorial Notes 7 of MATH3424

1 Summary of course material

1. We do transformation on variables to **ensure linearity**, to achieve **normality**, or to **stabilize the variance**

2. Transformations may be necessary for several reasons:

   (a) Theoretical consideration may specify the relationship are between two variables are non-linear.

   (b) The response variable Y may have a probability distribution whose variance is related to the mean.

   (c) The evidence comes from examining the residuals from the fit of a linear regression model using original variables.

3. For model $n_t = n_0 e^{\beta_1 t}\varepsilon_t'$, after log transformation we get

   $$\ln n_t = \beta_0 + \beta_1 t + \varepsilon_t$$

   with $\beta_0 = \ln n_0$, and we can fit a linear regression model to get $\hat{\beta}_0$. From discussion in class we know $\mathbb{E}(\hat{\beta}_0) = \beta_0$ but $\mathbb{E}(e^{\hat{\beta}_0}) > e^{\beta_0} = n_0$. A correction can be made to reduce the bias in the estimate of n_0, which is $e^{\hat{\beta}_0 - \mathrm{Var}(\hat{\beta}_0)/2}$ (why?)

4. Transformation stabilize the variance: Poisson, Binomial, Negative Binomial...

5. Detection, Removal (by transformation / weighted least square) of Heteroscedastic Errors

6. Logarithm transformation and Power Transformation

Comments:

1. At times it may be desirable to introduce a constant into a transformation of Y, such as when Y may be negative. For instance, $Y' = \log(Y + k)$, where k is an appropriately chosen constant.

2. When unequal error variances are present but the regression relation is linear, a transformation on Y may not be sufficient as such a transformation may stabilize the error variance, it will also change the linear relationship to a nonlinear one. A transformation on X may therefore also be required. This case can also be handled by using weighted least squares, see details in textbook.

$\beta_0 = \mathbb{E}(\hat{\beta}_0)$

$$\mathbb{E}(e^{\hat{\beta}_0}) = \mathbb{E}(e^{\beta_0} \cdot e^{\hat{\beta}_0 - \beta_0}) = e^{\beta_0}\,\mathbb{E}(e^{\hat{\beta}_0 - \beta_0})$$

$$= e^{\beta_0}\,\mathbb{E}\left(1 + (\hat{\beta}_0 - \beta_0) + \frac{(\hat{\beta}_0 - \beta_0)^2}{2} + o\left((\hat{\beta}_0 - \beta_0)^2\right)\right)$$

$$= e^{\beta_0}\left(1 + \frac{\mathbb{E}(\hat{\beta}_0 - \beta_0)^2}{2} + o\left(\mathbb{E}(\hat{\beta}_0 - \beta_0)^2\right)\right)$$

$$= e^{\beta_0}\left(1 + \frac{\mathrm{Var}(\hat{\beta}_0)}{2} + o\left(\mathbb{E}(\hat{\beta}_0 - \beta_0)^2\right)\right)$$

$$\frac{\mathbb{E}(e^{\hat{\beta}_0 - \mathrm{Var}(\hat{\beta}_0)/2})}{\mathbb{E}e^{\hat{\beta}_0}} = \frac{e^{\beta_0}\left(1 + \cancel{\frac{\mathrm{Var}(\hat{\beta}_0)}{2}} + o(1)\right)}{e^{\mathrm{Var}(\hat{\beta}_0)/2}\left(1 + \cancel{\frac{\mathrm{Var}(\hat{\beta}_0)}{2}} + o(1)\right)} \simeq e^{\beta_0}$$

2 Questions

2.1

One of the remarkable technological developments in computer industry has been the ability to store information densely on hard disk. The cost of storage has steadily declined. Table below shows the average price per megabyte in dollars from 1988 to 1998.

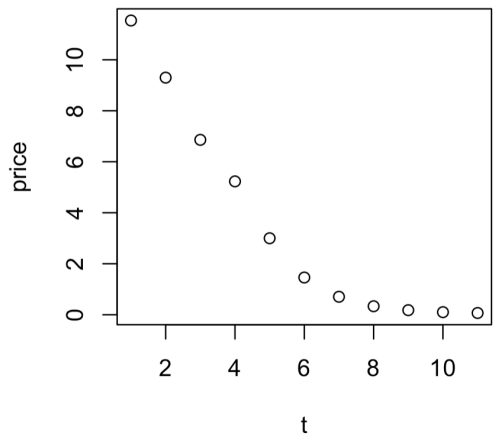
Table 6.20 Average Price Per Megabyte in Dollars from 1988 to 1998

Year	Price	Year	Price
1988	11.54	1994	0.705
1989	9.30	1995	0.333
1990	6.86	1996	0.179
1991	5.23	1997	0.101
1992	3.00	1998	0.068
1993	1.46		

Source: Kindly provided by Jim Porter, Disk/Trends in Wired April 1998.

(a) Does a linear time trend describe the data? Define a new variable t by coding 1988 as 1, 1989 as 2, and so forth.

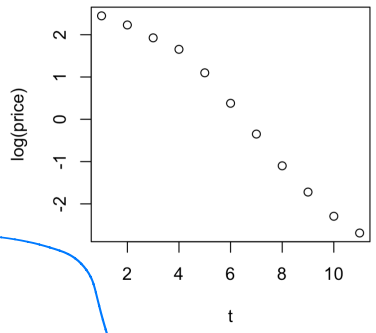
$P_t = \beta_0 + \beta t + \varepsilon_t$



No enough to describe the data.

(b) Fit the model $P_t = P_0 e^{\beta t}$, where P_t is the price in period t. Does this model describe the data?

$$\ln P_t = \underbrace{\ln P_0}_{\beta_0} + \beta t + \varepsilon.$$

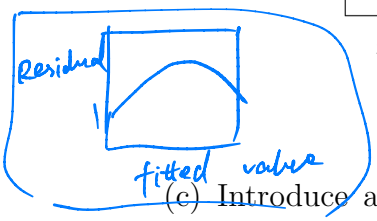



```
Call:
lm(formula = log(price) ~ t)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5006 -0.1693 -0.0296  0.1674  0.3942

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.50689    0.18527   18.93 1.47e-08 ***
t           -0.56050    0.02732  -20.52 7.24e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2865 on 9 degrees of freedom
Multiple R-squared:  0.9791,    Adjusted R-squared:  0.9767
F-statistic:   421 on 1 and 9 DF,  p-value: 7.242e-09
```



(c) Introduce an indicator variable which takes the value 0 for the years 1988-1991, and 1 for the remaining years. Fit a model to connecting $\log(P_t)$ with time t, the indicator variable, and the variable created by taking the product of time and the indicator variable. Interpret the coefficients of the fitted model.

```
Call:
lm(formula = log(price) ~ t + ind + t * ind)

Residuals:
      Min        1Q    Median        3Q       Max
-0.146274 -0.045015 -0.007797  0.036454  0.201542

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.73362    0.14445  18.924 2.86e-07 ***
t           -0.26785    0.05275  -5.078 0.001434 **
ind          1.47691    0.23377   6.318 0.000397 ***
t:ind       -0.37763    0.05726  -6.595 0.000306 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1179 on 7 degrees of freedom
Multiple R-squared:  0.9972,    Adjusted R-squared:  0.9961
F-statistic: 843.4 on 3 and 7 DF,  p-value: 2.565e-09
```

$$\log y = \beta_0 + \beta_1 x$$

$$y = e^{\beta_0} e^{\beta_1 x}$$

$$y' = e^{\beta_0} e^{\beta_1 (x+1)}$$

$$= e^{\beta_0} e^{\beta_1 x} e^{\beta_1}$$

$$= y e^{\beta_1}$$

$$\ln P_t = \beta_0 + \beta t + \gamma I + \alpha (I \cdot t)$$

$$= 2.73 - 0.27 t + 1.48 I - 0.378 (I \cdot t)$$

β: From 1988 to 1991, $\ln P_t$ decrease $|\beta| = 0.27$ every year.

α: From 1992 and onwards, $\ln P_t$ decrease $|\beta + \alpha| = 0.648$ every year.

γ: From 1991 to 1992, $\ln P_t$ decrease $|\beta + \gamma + 5\alpha| = \cdots$.

$t = 4$, $I = 0$

$t = 5$, $I = 1$

β: P_t would drop $(1-e^{\beta}) \times 100\% \approx \beta \times 100\%$

α: . . . $(1-e^{\beta+\alpha}) \times 100\%$

r: . . . . . $(1-e^{\beta+r+5\alpha}) \times 100\%$

$e^{\beta} \approx 1+\beta$

2.2

Wind Chill Factor: Table 6.18 gives the effective temperatures (W), which are due to the wind chill effect, for various values of the actual temperatures (T) in still air and wind speed (V). The zero-wind condition is taken as the rate of chilling when one is walking through still air [an apparent wind of four miles per hour (mph)]. The National Weather Service originally published the data; we have compiled it from a publication of the Museum of Science of Boston. The temperatures are measured in degrees Fahrenheit and the wind speed in mph.

Table 6.18 Wind Chill Factor (°F) for Various Values of Winds peed, V, in Miles/Hour, and Temperature (°F)

	Actual Air Temperature (T)											
V	50	40	30	20	10	0	−10	−20	−30	−40	−50	−60
5	48	36	27	17	5	−5	−15	−25	−35	−46	−56	−66
10	40	29	18	5	−8	−20	−30	−43	−55	−68	−80	−93
15	35	23	10	−5	−18	−29	−42	−55	−70	−83	−97	−112
20	32	18	4	−10	−23	−34	−50	−64	−79	−94	−108	−121
25	30	15	−1	−15	−28	−38	−55	−72	−88	−105	−118	−130
30	28	13	−5	−18	−33	−44	−60	−76	−92	−109	−124	−134
35	27	11	−6	−20	−35	−48	−65	−80	−96	−113	−130	−137
40	26	10	−7	−21	−37	−52	−68	−83	−100	−117	−135	−140
45	25	9	−8	−22	−39	−54	−70	−86	−103	−120	−139	−143
50	25	8	−9	−23	−40	−55	−72	−88	−105	−123	−142	−145

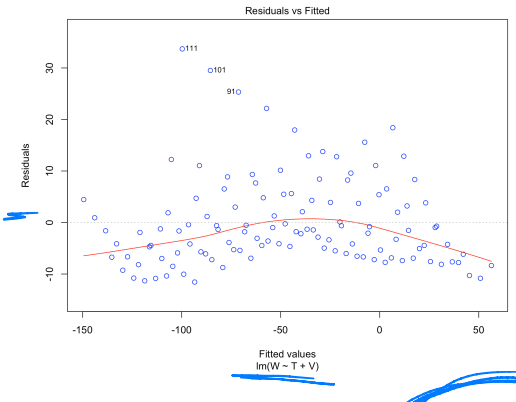
(a) Fit a linear relationship between W, T, and V. The pattern of residuals should indicate the inadequacy of the linear model.

```
Call:
lm(formula = W ~ T + V, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-11.560  -6.106  -1.791   4.336  33.704

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.05664    1.71960  -5.267  6.4e-07 ***
T            1.41867    0.02301  61.661  < 2e-16 ***
V           -1.10545    0.05530 -19.989  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.7 on 117 degrees of freedom
Multiple R-squared:  0.9729,    Adjusted R-squared:  0.9724
F-statistic:  2101 on 2 and 117 DF,  p-value: < 2.2e-16
```



model <- lm(. .)

plot(model)

plot(model, which=1)

(b) Fit the model

$$W = \beta_0 + \beta_1 T + \beta_2 V + \beta_3 \sqrt{V} + \varepsilon \tag{1}$$

Does the fit of this model appear adequate? The W numbers were produced by the National Weather Service according to the formula (except for rounding errors)

$$W = 0.0817(3.17\sqrt{V} + 5.81 - 0.25V)(T - 91.4) + 91.4 \tag{2}$$

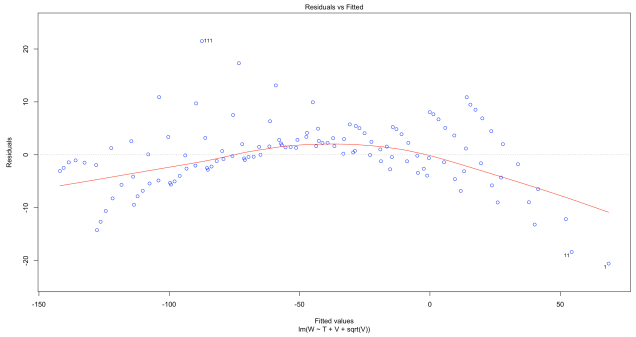
$W = \beta_0 + \beta_1 T + \beta_2 V + \beta_3 \sqrt{V} + \beta_4 V \cdot T + \beta_5 \sqrt{V} \cdot T + \varepsilon$

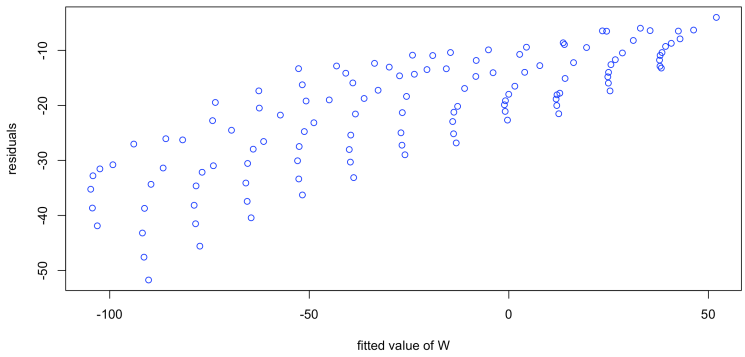
```
Call:
lm(formula = W ~ T + V + sqrt(V), data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-20.565  -2.873   0.044   3.154  21.489

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.90935    5.88953   8.304 2.11e-13 ***
T             1.41867    0.01687  84.076  < 2e-16 ***
V             1.65055    0.27651   5.969 2.66e-08 ***
sqrt(V)     -26.62313    2.64225 -10.076  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.381 on 116 degrees of freedom
Multiple R-squared:  0.9856,    Adjusted R-squared:  0.9852
F-statistic:  2638 on 3 and 116 DF,  p-value: < 2.2e-16
```





(c) Can you suggest a model better than those in (1) and (2)?

$W = \beta_0 + \beta_1 T + \beta_2 \sqrt{V} + \beta_3 \sqrt{V} \cdot T$

```
Call:
lm(formula = W ~ T + sqrt(V) + sqrt(V) * T, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1564 -2.9000 -0.3898  2.1509 14.8049

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.09130    1.36457   9.594   <2e-16 ***
T             0.85344    0.03912  21.815   <2e-16 ***
sqrt(V)     -10.45926    0.26021 -40.195   <2e-16 ***
T:sqrt(V)     0.11250    0.00746  15.081   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.24 on 116 degrees of freedom
Multiple R-squared:  0.9936,    Adjusted R-squared:  0.9935
F-statistic:  6023 on 3 and 116 DF,  p-value: < 2.2e-16
```