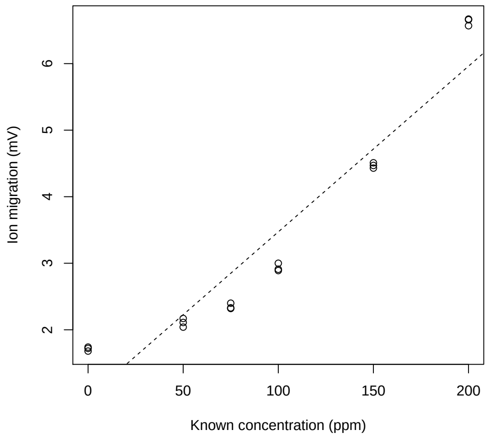
STAT 420 Spring 2014
Homework 6: Solutions

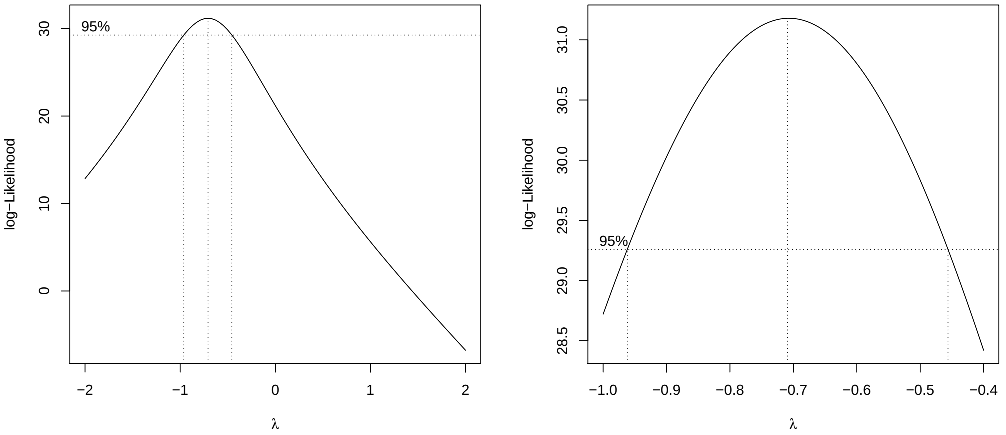
Exercise 1

(a) The scatterplot of the data (with best fitting linear regression line) is given below:



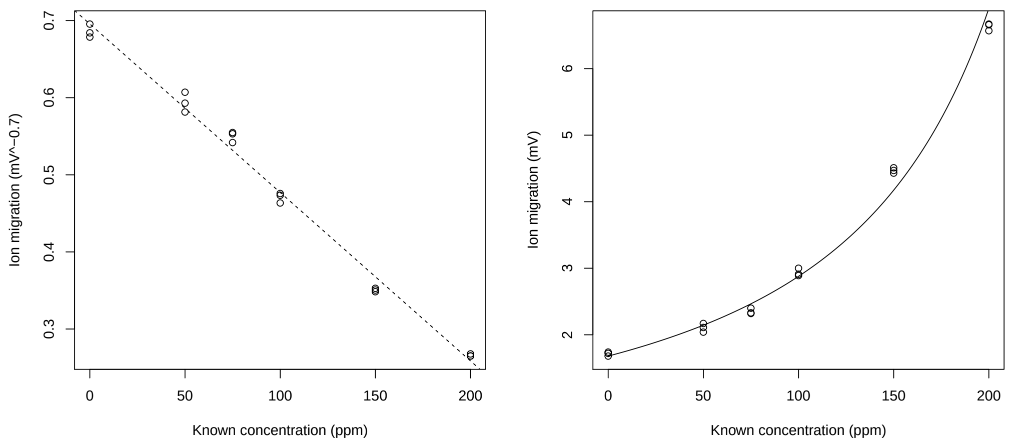
Note that the linear model does NOT seem appropriate here.

(b) The plots of the optimal Box-Cox power transformation are given below:



From the plots, you can see that the optimal transformation value is about $\lambda = -0.7$.

The scatterplot of the transformed data (with $\text{mV}^{-0.7}$) is given below (left):



Note that the linear model does seem appropriate here (left plot). The right plot gives the nonlinear relationship (on the original data scale) that is implied by the linear relationship on the transformed data scale.

Exercise 2

```
> library(faraway)
> data(uswages)
> w1mod=lm(wage~educ+exper,data=uswages)
> w1mod$coef
(Intercept)        educ       exper
-242.799412   51.175268    9.774767
> w2mod=lm(log(wage)~educ+exper,data=uswages)
> w2mod$coef
(Intercept)        educ       exper
 4.65031905  0.09050628  0.01807855
```

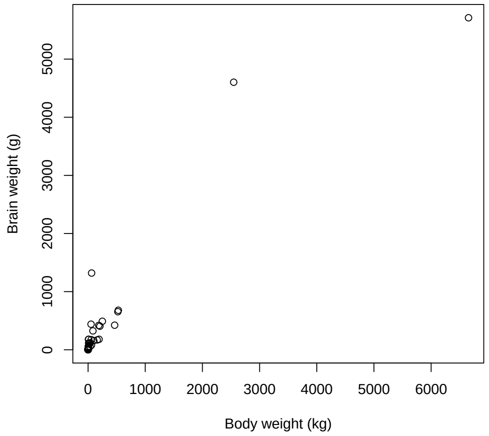
For original (non-transformed) model, we would expect to earn an additional $51.18 dollars per week for every 1-year increase in education (with our experience fixed).

For log-transformed model, we would expect weekly wages to increase $e^{0.09051} = 1.094728$ times for every 1-year increase in education (with our experience fixed), i.e., expect to earn about 9.5% more for every 1 additional year of education (with experience fixed).

Second model makes more sense because estimated wages will be constrained to be positive.

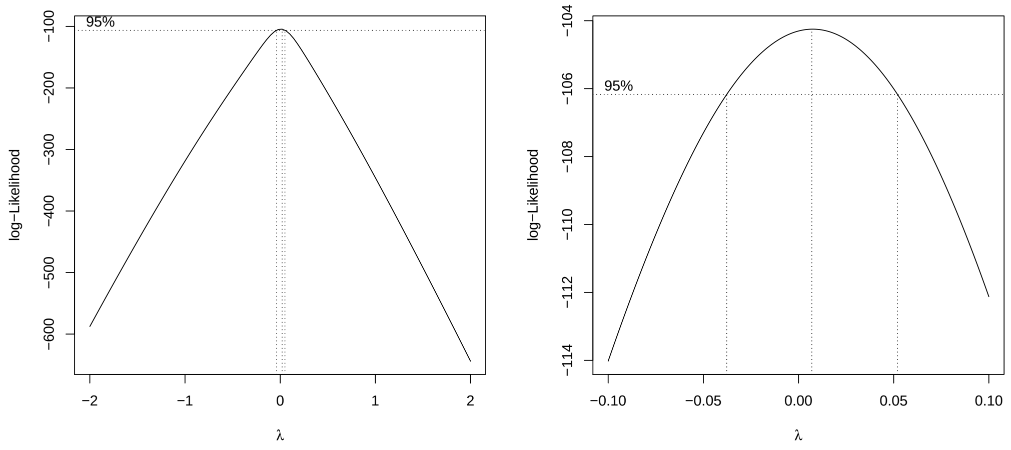
Exercise 3

(a) The scatterplot of the `mammals` data is given below:



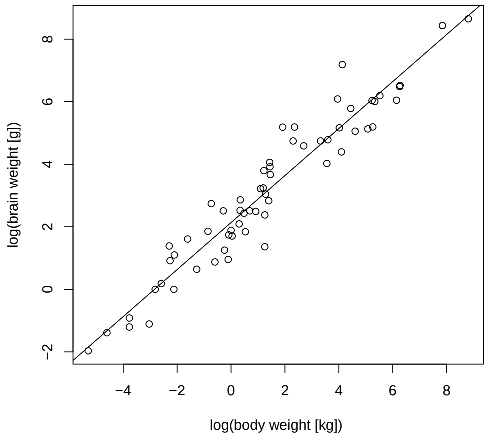
Note that the linear model does NOT seem appropriate here.

(b) The plots of the optimal Box-Cox power transformation are given below:



From the plot, you can see that the log transformation ($\lambda = 0$) is within the 95% confidence interval of "recommended" transformation values.

(c) The scatterplot of the log-transformed `mammals` data is given below:



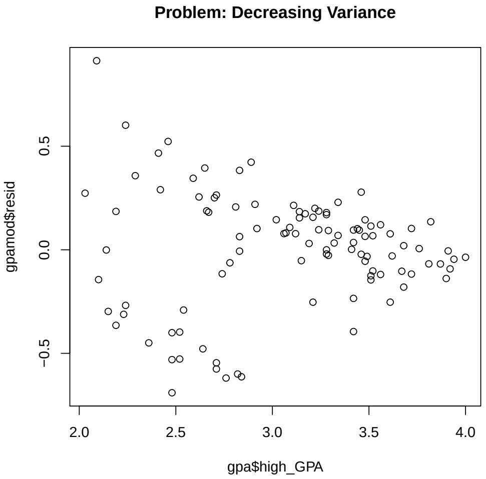
Note that the linear model does seem appropriate here. To predict the Siberian tiger's brain weight use the R code:

```
> bmod=lm(log(brain)~log(body),data=mammals)
> newdata=data.frame(body=227)
> tiger=predict(bmod,newdata,interval="prediction")
> tiger
       fit      lwr     upr
1 6.212647 4.793485 7.63181
> exp(tiger)
       fit      lwr     upr
1 499.0206 120.7213 2062.78
```

Prediction is $\hat{y} = 499.02$ g; the 95% prediction interval is [120.72 g; 2062.78 g].

Exercise 4

(a) The residual plot is given below:



Note that the variance of the residuals decreases as high school GPA increases.

(b) You can use the Shapiro-Wilk test. The null hypothesis is $H_0 : Y \sim \mathrm{N}(0, \sigma^2)$ and the alternative is $H_1 : Y \not\sim \mathrm{N}(0, \sigma^2)$.

```
> shapiro.test(gpamod$resid)

        Shapiro-Wilk normality test

data:  gpamod$resid
W = 0.9659, p-value = 0.008387
```

Observed test statistic is $W = 0.9659$, and the p-value is $p = 0.0084 < \alpha = 0.01$. So we **Reject** the null hypothesis that the residuals are normally distributed.

You could also use the Looney-Gulledge test. The null hypothesis is $H_0 : \rho_{LG} = 1$ and the alternative is $H_1 : \rho_{LG} < 1$, where ρ_{LG} is the correlation between the empirical and theoretical (normal) quantiles. This is equivalent to testing $H_0 : Y \sim \mathrm{N}(0, \sigma^2)$ versus the alternative $H_1 : Y \not\sim \mathrm{N}(0, \sigma^2)$.

```
> set.seed(123)
> LGnormtest(gpamod$resid)
$rho
[1] 0.9816301
```

```
$cval
       5%
0.9874272

$pval
[1] 0.008
```

The observed sample correlation is $\rho_{LG} = 0.9816$ and the corresponding p-value (based on 10,000 Monte Carlo samples) is $p = 0.008 < \alpha = 0.01$. So we **Reject** the null hypothesis that the residuals are normally distributed.

(c) You can use the Breusch-Pagan test. The null hypothesis is $H_0 : V(e_i) = \sigma^2$ and the alternative is $H_1 : V(e_i) \neq \sigma^2$, where $V(e_i)$ denotes the variance of the residuals.

```
> BPtest(gpamod)
$BP
[1] 29.55316

$df
[1] 1

$pval
[1] 5.440381e-08
```

Observed test statistic is $\chi^2_{BP} = 29.55$, and the p-value is $p \approx 0 < \alpha = 0.01$. So we **Reject** the null hypothesis that the residuals have constant variance.