BIOSTAT/STAT 570: Coursework 6 Key

To be submitted to the course canvas site by 11:59pm Monday 29th November, 2021.

1. Table 5 reproduce data from Altham (1991) of counts of T_4 cells/mm^3 in blood samples from 20 patients in remission from Hodgkin's disease and 20 other patients in remission from disseminated malignancies. The question of interest here is: Is there a difference in the distribution of cell counts between the two diseases? A quantitative assessment of any difference is also desirable.

   (a) Carry out an exploratory data analysis and provide a summary of the two distributions (marks will be deducted for unnecessary plots/summaries).

   Solution: The table below provides descriptive statistics for each of the populations.

Group	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev.
Hodgkin's	171	397	682	823	1131	2415	566
Other	116	360	433	522	709	1252	293

Table 1: Summary statistics of cell counts by disease status.

We can see from the table that, on average, patients with Hodgkin's have T_4 cell counts that are higher than patients with Non-Hogkin's. The table also indicates that there is greater variability in T_4 cell counts among those with Hodgkin's compared to those with other disseminated malignancies. We can also see that the data is right-skewed (median $<$ mean) for both groups of patients. A boxplot (Figure 1) of the data also gives a sense of the distribution of the disease counts.

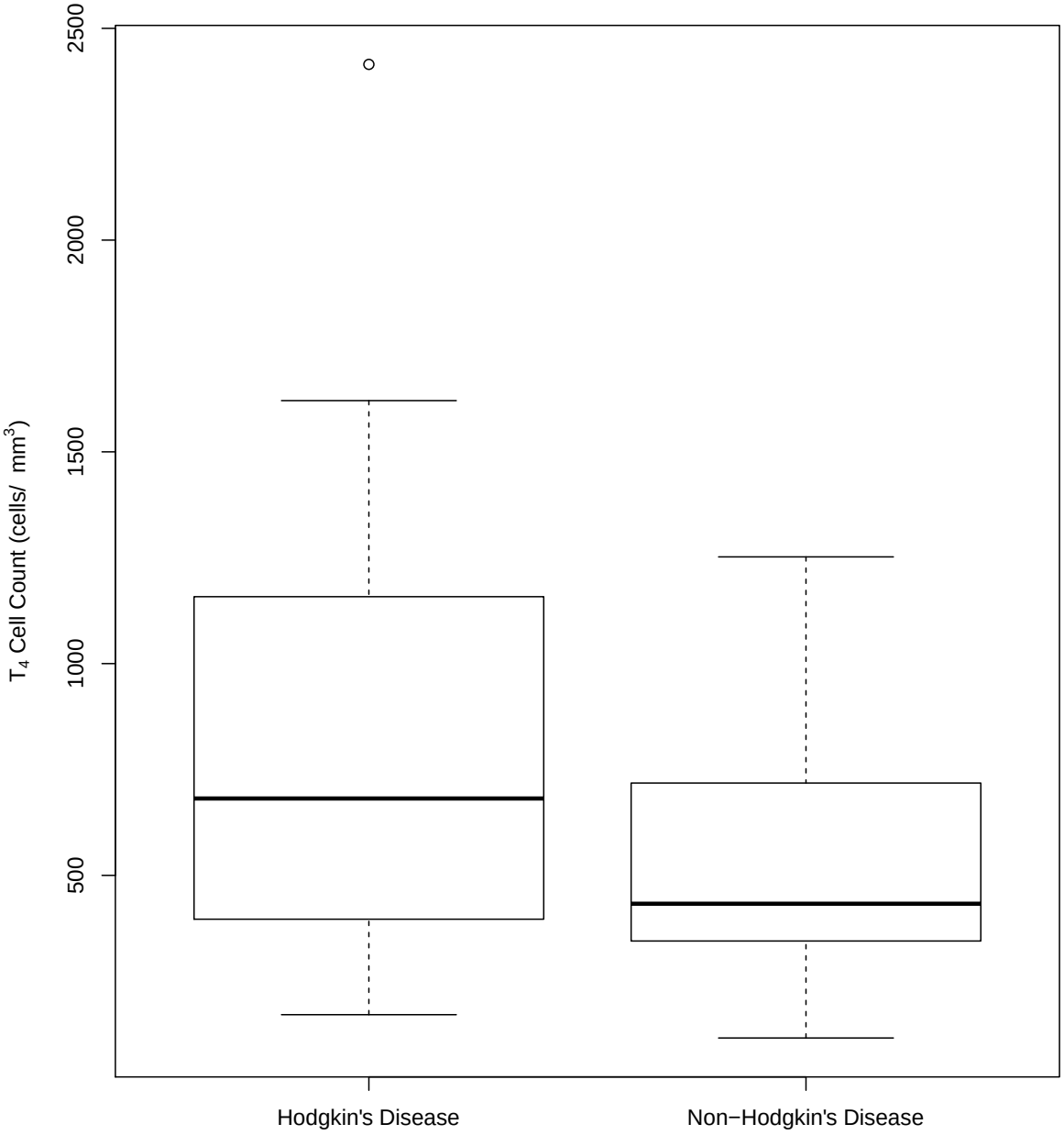

Figure 1: Boxplot of the distribution of T_4 cell counts by disease status.

(b) We now examine various approaches to obtaining a 90% confidence interval for differences (on some scale) between distributions. Obtain such a confidence intervals for these data when the data are: i) on their original scale; ii) $\log_e$ transformed; iii) square root transformed. What are the considerations when choosing a scale here? **Solution:** We will provide 90% CI for difference in the mean of (possibly) transformed outcomes comparing a group with Non-Hodgkin's Disease to those with Hodgkin's Disease.

- Original scale: $E[\text{cell count}] = \beta_0 + \beta_1 I_{\text{Other}}$
- Log transformed: $E[\log(\text{cell count})] = \beta_0 + \beta_1 I_{\text{Other}}$
- Square root transformed: $E[\sqrt{\text{cell count}}] = \beta_0 + \beta_1 I_{\text{Other}}$

The table below gives the parameter estimates and 90% CI after the various transformations allowing for unequal variances.

Transformation of Outcome	$\hat{\beta}_1$ (90% CI)
Original Scale	-301.15 (-543.57, -58.73)
Log Transformed	-0.3983 (-0.7562, -0.0404)
Square root Transformed	-5.207 (-9.546, -0.867)

Table 2: Parameter estimates and 90% confidence intervals for the difference in mean (possibly) transformed T_4 cell counts comparing groups with different disease status.

One consideration would be the interpretation of the parameters. On the original scale, β_1 would be the difference in the expected cell count comparing patients with Non-Hodgkin's Disease to those with Hodgkin's Disease. If we $\log$ transform cell counts, the interpretation of e^{β_1} would be the ratio of geometric mean cell counts comparing patients with Non-Hodgkin's Disease to those with Hodgkin's Disease. The interpretation with square root transformed outcome is much more difficult. Another consideration would be whether or not the data are skewed. Since cell counts are likely to be skewed, we might prefer a $\log$ transformation.

(c) We will fit Poisson, gamma and inverse Gaussian models to the cell count data assuming the canonical link with $g(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta}$, where $\boldsymbol{x}_i = [1\ 0]$ for $i = 1, \ldots, n = 20$, and $\boldsymbol{x}_i = [1\ 1]$ for $i = n+1, \ldots, 2n = 40$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)$. The question of interest here is whether the means of the two groups are equal. Express this question in terms of β_0 and β_1. For what function of $\boldsymbol{\beta}$ is this question answered on the scale of the original data?

Solution: The underlying model is: $g(\mu) = \beta_0 + \beta_1 I_{\text{Other}}$. In order to test for a difference in cell counts between the two groups, we need only test $H_0 : \beta_1 =$

0 where β_1 is the coefficient associated with the indicator of being in the "Non-Hodgkin's Disease."

- Assuming a Poisson distribution, the canonical link is $g(\mu) = \log(\mu)$. If we want to answer the question on the scale of the original data we would use $g^{-1}(\cdot) = \exp(\cdot)$. So e^{β_1} is the difference on the original scale. The table below displays useful quantities that could be estimated from this model.
- Assuming a Gamma distribution, the canonical link is $g(\mu) = 1/\mu$. If we want to answer the question on the original scale we would use $g^{-1}(x) = 1/x$. Useful quantities that could be estimated from this model are displayed below.
- Assuming an Inverse Gaussian distribution, the canonical link is $g(\mu) = (1/\mu)^2$; hence, to answer the question on the original scale we would use $g^{-1}(x) = \sqrt{1/x}$. Useful quantities that could be estimated from this model are displayed below.

Quantity	Poisson	Gamma	Inverse Gaussian
Hodgkins mean cell count	e^{β_0}	$\frac{1}{\beta_0}$	$\frac{1}{\sqrt{\beta_0}}$
Other mean cell count	$e^{\beta_0+\beta_1}$	$\frac{1}{\beta_0+\beta_1}$	$\frac{1}{\sqrt{\beta_0+\beta_1}}$

Table 3: Useful quantities that can be estimated for the various regression models.

(d) Using the asymptotic distribution of the MLE, that is

$$\boldsymbol{I}(\hat{\boldsymbol{\beta}})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_d \mathsf{N}_2\left(0, I_2\right),$$

give 90% confidence intervals for each parameter. Under each of the distributional assumptions, would you conclude that the means of the two groups are equal?

Solution: For each distribution we fit the model described in question (3) and compute a 90% CI using the asymptotic distribution of the MLE. The table below displays the parameter estimates and confidence intervals.

Model	$\hat{\beta}_0$ (90% CI)	$\hat{\beta}_1$ (90% CI)
Poisson	6.7132 (6.7004, 6.7260)	-0.4554 (-0.4760, -0.4349)
Gamma	1.215e-03 (9.343e-04, 1.495e-03)	7.008e-04 (1.770e-04, 1.225e-03)
Inverse Gaussian	1.476e-06 (7.197e-07, 2.232e-06)	2.194e-06 (5.165e-07, 3.871e-06)

Table 4: Parameter estimates and 90% confidence intervals for the various regression models.

Since we are testing $H_0 : \beta_1 = 0$, we would reject the null hypothesis for all of the models considered since the confidence interval does not contain 0 in favor of a hypothesis that the mean cell counts for the two groups are different.

Hodgkin's Disease	Non-Hodgkin's Disease
396	375
568	375
1212	752
171	208
554	151
1104	116
257	736
435	192
295	315
397	1252
288	675
1004	700
431	440
795	771
1621	688
1378	426
902	410
958	979
1283	377
2415	503

Table 5: Counts of T_4 cells/mm^3 in blood samples from 20 patients in remission from Hodgkin's disease and 20 other patients in remission from disseminated malignancies.

2. The data in Table 6, taken from Wakefield et al. (1994), were collected following the administration of a single 30mg dose of the drug Cadralazine to a cardiac failure patient. The response y_i represents the drug concentration at time x_i, $i = 1, \ldots, 8$. The most straightforward model for these data is to assume

$$\log y_i = \mu(\boldsymbol{\beta}) + \epsilon_i = \log\left[\frac{D}{V}\exp(-k_e x_i)\right] + \epsilon_i,$$

where $\epsilon_i \sim_{iid} \mathsf{N}(0, \sigma^2)$, $\boldsymbol{\beta} = [V, k_e]$ and the dose is $D = 30$. The parameters are the volume of distribution $V > 0$ and the elimination rate k_e.

i	x_i	y_i
1	2	1.63
2	4	1.01
3	6	0.73
4	8	0.55
5	10	0.41
6	24	0.01
7	28	0.06
8	32	0.02

Table 6: Concentrations y_i of the drug Cadralazine (in mg/liter) as a function of time (in hours) x_i, for $i = 1, \ldots, 8$.

(a) For this model obtain expressions for:

i. The log-likelihood function $L(\boldsymbol{\beta}, \sigma^2)$.

ii. The score function $\boldsymbol{S}(\boldsymbol{\beta}, \sigma^2)$.

iii. The expected information matrix $\boldsymbol{I}(\boldsymbol{\beta}, \sigma^2)$.

Solution:

i. From the assumed model, we have $\log y_i \sim N(\log\left(\frac{D}{V}\exp\{-k_e x_i\}, \sigma^2\right)$. The likelihood of i-th sample is

$$l_i(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\Big\{-\frac{(\log y_i - \log D + \log V + k_e x_i)^2}{2\sigma^2}\Big\}$$

The log-likelihood of i-th sample is

$$L_i(\boldsymbol{\beta}, \sigma^2) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\big(\log y_i - \log D + \log V + k_e x_i\big)^2$$

The log-likelihood function is

$$L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(\log y_i - \log D + \log V + k_e x_i\big)^2$$

ii. The score function is

$$\begin{aligned}
\frac{\partial L(\boldsymbol{\beta}, \sigma^2)}{\partial V} &= -\frac{1}{V\sigma^2}\sum_{i=1}^{n}\Big(\log y_i - \log D + \log V + k_e x_i\Big) \\
\frac{\partial L(\boldsymbol{\beta}, \sigma^2)}{\partial k_e} &= -\frac{x_i}{\sigma^2}\sum_{i=1}^{n}\Big(\log y_i - \log D + \log V + k_e x_i\Big) \\
\frac{\partial L(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\Big(\log y_i - \log D + \log V + k_e x_i\Big)^2
\end{aligned}$$

iii. The expected information is a 3×3 matrix. We calculate each element as following:

$$\begin{aligned}
I_{VV} &= -\mathsf{E}\Big[\frac{\partial^2 L(\boldsymbol{\beta}, \sigma^2)}{\partial V^2}\Big] = -\mathsf{E}\Big[\frac{\sum_{i=1}^{n}\big(\log y_i - \log D + \log V + k_e x_i\big)}{V^2\sigma^2} - \frac{n}{V^2\sigma^2}\Big] \\
&= \frac{n}{V^2\sigma^2} \\
I_{Vk_e} &= -\mathsf{E}\Big[\frac{\partial^2 L(\boldsymbol{\beta}, \sigma^2)}{\partial V\,\partial k_e}\Big] = -\mathsf{E}\Big[-\frac{\sum_{i=1}^{n} x_i}{V\sigma^2}\Big] = \frac{\sum_{i=1}^{n} x_i}{V\sigma^2} \\
I_{V\sigma^2} &= -\mathsf{E}\Big[\frac{\partial^2 L(\boldsymbol{\beta}, \sigma^2)}{\partial V\,\partial \sigma^2}\Big] = -\mathsf{E}\Big[\frac{\sum_{i=1}^{n}\big(\log y_i - \log D + \log V + k_e x_i\big)}{V\sigma^2}\Big] = 0 \\
I_{k_e k_e} &= -\mathsf{E}\Big[\frac{\partial^2 L(\boldsymbol{\beta}, \sigma^2)}{\partial k_e^2}\Big] = \frac{\sum_{i=1}^{n} x_i^2}{\sigma^2} \\
I_{k_e\sigma^2} &= -\mathsf{E}\Big[\frac{\partial^2 L(\boldsymbol{\beta}, \sigma^2)}{\partial k_e \partial \sigma^2}\Big] = -\mathsf{E}\Big[\frac{\sum_{i=1}^{n} x_i\big(\log y_i - \log D + \log V + k_e x_i\big)}{\sigma^4}\Big] = 0 \\
I_{\sigma^2\sigma^2} &= -\mathsf{E}\Big[\frac{\partial^2 L(\boldsymbol{\beta}, \sigma^2)}{\partial k_e \partial \sigma^2}\Big] = -\mathsf{E}\Big[\frac{n}{2\sigma^4} - \frac{\sum_{i=1}^{n}\big(\log y_i - \log D + \log V + k_e x_i\big)^2}{\sigma^6}\Big] \\
&= \frac{n\sigma^2}{\sigma^6} - \frac{n}{2\sigma^4} = \frac{n}{2\sigma^4}
\end{aligned}$$

This gives an expected information matrix of

$$\boldsymbol{I}(\boldsymbol{\beta}, \sigma^2) = \begin{bmatrix} \frac{n}{V^2\sigma^2} & \frac{\sum_{i=1}^{n} x_i}{V\sigma^2} & 0 \\ \frac{\sum_{i=1}^{n} x_i}{V\sigma^2} & \frac{\sum_{i=1}^{n} x_i^2}{\sigma^2} & 0 \\ 0 & 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

(b) Obtain the MLE, and give an asymptotic 95% confidence interval for each element of $\boldsymbol{\beta}$.
Solution: We can derive MLE by solving equation of $\boldsymbol{S}(\boldsymbol{\beta}, \sigma^2) = \mathbf{0}$. The MLEs are

$$\widehat{k_e} = \frac{\sum_{i=1}^{n} \log y_i (x_i - \bar{x})}{n\bar{x}^2 - \sum_{i=1}^{n} x_i^2 e}$$
$$\widehat{V} = \exp\left(\log D - \frac{1}{n} \log y_i - \widehat{k_e}\bar{x} \right)$$
$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\log y_i - \log D + \log \widehat{V} + \widehat{k_e} x_i \right)^2$$

Confidence intervals are from the expected information matrix in part 1(iii). The results are in Table 7:

parameter	MLE	95%CI lower	95%CI upper
V	16.66	4.58	28.75
k_e	0.15	0.11	0.19
σ^2	0.41	0.01	0.82

Table 7: MLE and 95% CI of $\boldsymbol{\beta}$

Alternatively, we can also take use of optim function in R to numerically derive the MLEs. Note that in this method, we should optimize all three parameters at the same time.

(c) Plot the data, along with the fitted curve. **Solution:** See Figure 2

(d) Using residuals, examine the appropriateness of the assumptions of the above model. Does the model seem reasonable for these data?

Solution: Figure 3 shows a plot of the residuals over time. We see that all but one of the residuals are positive, which indicates some lack of model fit. Additionally the first data point has a very large positive residual. Overall, the model provides an OK fit to the data.

(e) The clearance $Cl = V \times k_e$ and elimination half-life $x_{1/2} = \log 2 / k_e$ are parameters of interest in this experiment. Find the MLEs of these parameters along with asymptotic 95% confidence intervals.

Solution: By the invariance property of MLEs,we can simply plug-in V and k_e to obtain MLEs of clearance and half-life.

$$\widehat{Cl}_{MLE} = \widehat{V} \times \widehat{k_e}$$

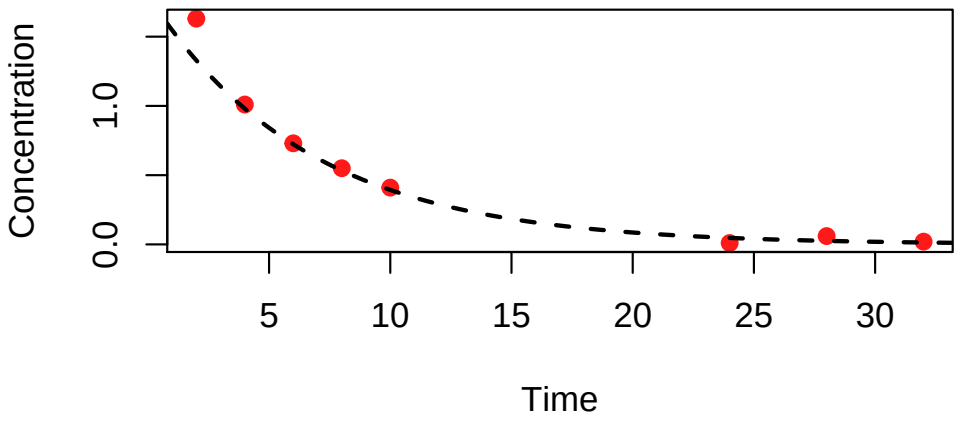

Figure 2: Data and fitted curve

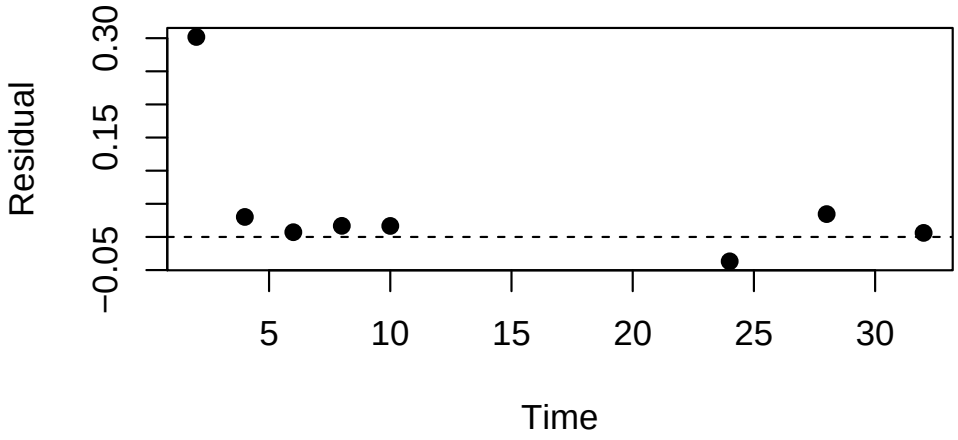

Figure 3: Data and fitted curve

$$\widehat{x_{1/2}}_{MLE} = \log 2/\hat{k}_e$$

We can use the delta method to obtain asymptotic confidence intervals. Here $g_1 = V \times k_e$ and $g_2 = \log 2/k_e$ with $\nabla g_1 = (k_e, V)$and$\nabla g_2 = (0, -log2/k_e^2)$. We know that

$$\sqrt{n}\left(\begin{pmatrix}\hat{V}\\ \hat{k}_e\end{pmatrix} - \begin{pmatrix}V\\ k_e\end{pmatrix}\right) \rightarrow N(\mathbf{0}, \mathbf{\Sigma})$$

and hence

$$\sqrt{n}(\widehat{Cl} - Cl) \rightarrow_d N(0, \nabla g_1 \mathbf{\Sigma} \nabla g_1^T)$$
$$\sqrt{n}(\widehat{x_{1/2}} - x_{1/2}) \rightarrow_d N(0, \nabla g_2 \mathbf{\Sigma} \nabla g_2^T)$$

Numerical results are in Table 8

Parameter	MLE	95%CI lower	95%CI upper
Cl	2.53	1.16	3.91
$x_{1/2}$	4.56	3.35	5.76

Table 8: MLE and 95% CI of clearance and half-life

R Code

```
### PROBLEM 1 ###

# Input data
hodgkins <- c(396, 568, 1212, 171, 554, 1104, 257, 435, 295, 397, 288, 1004, 431,
              795, 1621, 1378, 902, 958, 1283, 2415)
non.hodgkins <- c(375, 375, 752, 208, 151, 116, 736, 192, 315, 1252, 675, 700, 440,
                  771, 688, 426, 410, 979, 377, 503)

# Print summaries
round(c(summary(hodgkins), sd(hodgkins)))
round(c(summary(non.hodgkins), sd(non.hodgkins)))

# Put data in regression form
y <- c(hodgkins, non.hodgkins)
x <- c(rep(0,length(hodgkins)), rep(1,length(non.hodgkins)))

# Boxplots
pdf("plot1.pdf", height=9.5, width=8.5)
par(mar=c(5,5,3,2)+0.1)
boxplot(y ~ x, names=c("Hodgkin's Disease", "Non-Hodgkin's Disease"),
        ylab=expression("T"[4]*" Cell Count (cells/"*"  mm"^3*")"))
dev.off()

# Linear models with transformations
lm.id <- t.test(y ~ x, conf.level=0.90)
lm.log <- t.test(log(y) ~ x, conf.level=0.90)
lm.sqrt <- t.test(sqrt(y) ~ x, conf.level=0.90)

# Print estimates and intervals
format(c(diff(lm.id$estimate), rev(-lm.id$conf.int)), digits=4)
format(c(diff(lm.log$estimate), rev(-lm.log$conf.int)), digits=4)
format(c(diff(lm.sqrt$estimate), rev(-lm.sqrt$conf.int)), digits=4)

# Fit GLMs
glm.pois <- glm(y ~ x, family=poisson)
glm.gamma <- glm(y ~ x, family=Gamma)
glm.invgauss <- glm(y ~ x, family=inverse.gaussian)
```

```
# Compute intervals
CI.pois <- coef(glm.pois) +
           outer(sqrt(diag(vcov(glm.pois))) * qnorm(1-0.10/2), c(-1,1))
CI.gamma <- coef(glm.gamma) +
            outer(sqrt(diag(vcov(glm.gamma))) * qnorm(1-0.10/2), c(-1,1))
CI.invgauss <- coef(glm.invgauss) +
               outer(sqrt(diag(vcov(glm.invgauss))) * qnorm(1-0.10/2), c(-1,1))

# Print estimates and intervals
format(cbind(coef(glm.pois), CI.pois), digits=4)
format(cbind(coef(glm.gamma), CI.gamma), digits=4, scientific=T)
format(cbind(coef(glm.invgauss), CI.invgauss), digits=4, scientific=T)


### PROBLEM 2 ###

x <- c(2,4,6,8,10,24,28,32)
y <- c(1.63,1.01,0.73,0.55,0.41,0.01,0.06,0.02)
log.y <- log(y)
n <- length(x)
D <- 30

ke.hat <- sum(log.y*(x-mean(x)))/(-sum(x^2)+n*mean(x)^2)
V.hat <- exp(log(D)-mean(log.y)-ke.hat*mean(x))
sig2.hat <- sum((log.y-log(D)+log(V.hat)+ke.hat*x)^2)/n
Info <- matrix(c(n/(V.hat^2*sig2.hat), sum(x)/(V.hat*sig2.hat), 0,
                    sum(x)/(V.hat*sig2.hat), sum(x^2)/sig2.hat, 0,
                    0, 0, n/(2*sig2.hat^2)),3,3)
se.hat <- sqrt(diag(solve(Info)))
tbl <- c(V.hat,ke.hat, sig2.hat) + as.vector(se.hat)%o%c(0,-1.96,1.96)
xtable(tbl, digits = 2)

#residuals and plotting
mu.hat <- function(V,k,x){ exp(log(D/V*exp(-k*x))) }
pdf("q3c.pdf",height=3,width=5)
plot(y~x,pch=19,col=rgb(1,0,0,0.9),xlab="Time",ylab="Concentration")
x.seq <- seq(0,35,length=80)
lines(x.seq,mu.hat(V.hat,ke.hat,x.seq),lty=2,lwd=2)
dev.off()
```

```
fit <- mu.hat(V.hat,ke.hat,x)
resid <- y-fit
pdf("q3d.pdf",height=3,width=5)
plot(resid~x,pch=19,xlab="Time",ylab="Residual")
abline(h=0,lty=2)
dev.off()

# interval for clearance
Sigma <- solve(Info[1:2,1:2])
grad1 <- c(ke.hat, V.hat)
grad2 <- c(0,-log(2)/ke.hat^2)


ci1 <- (V.hat * ke.hat) + sqrt(t(grad1)%*%Sigma%*%grad1)[1] %o% c(0,-1.96,1.96)

ci2 <- log(2)/ke.hat + sqrt(t(grad2)%*%Sigma%*%grad2)[1] %o% c(0,-1.96,1.96)
tbl <- rbind(ci1, ci2)
xtable(tbl, digits = 2)
```