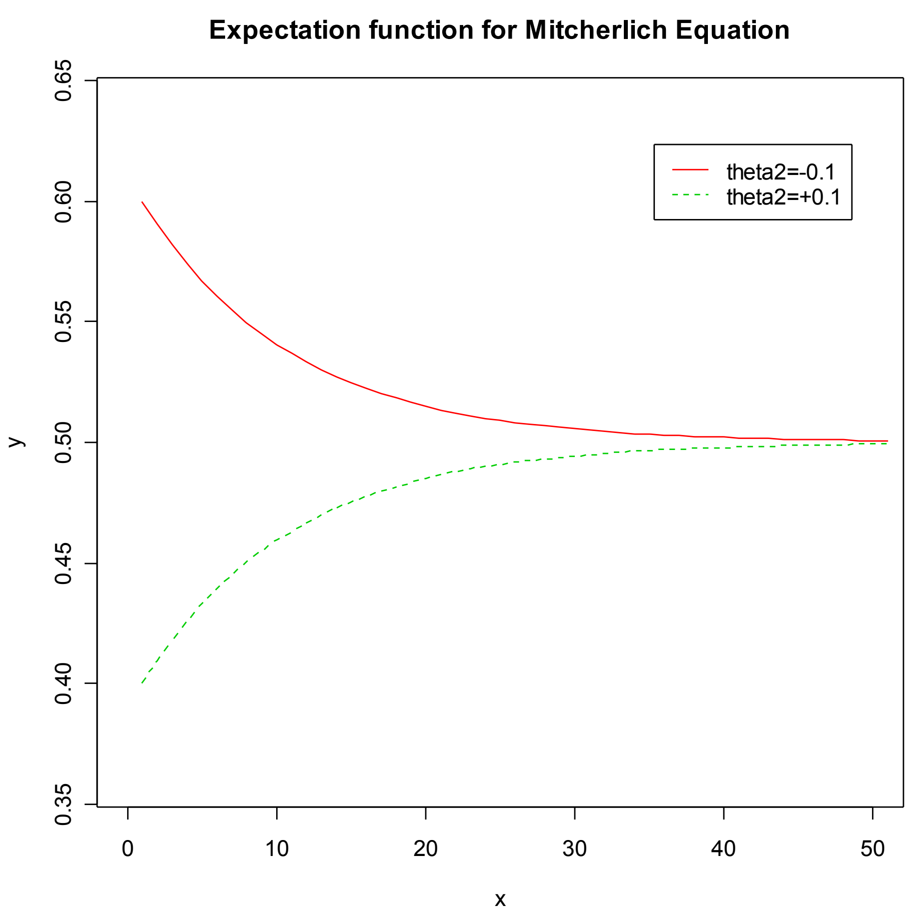
Statistics 203
Introduction to Regression and Analysis of Variance
Assignment #3 Solutions

Question 1:

a) This is a non-linear regression model since the parameter θ_3 does not enter the response function in a linear way and the model cannot be made linear.

b) and c)

```
> x <- 0:50
> y1 <- 0.5+0.1*exp(-0.1*x)
> y2 <- 0.5-0.1*exp(-0.1*x)
> plot(c(0,50), c(0.36, 0.64), type="n", xlab="x", ylab="y", main="Expectation
function for Mitcherlich Equation")
> lines(y1,col=2,lty=1)
> lines(y2,col=3,lty=2)
> legend(locator(1), lty=c(1,2), col=c(2,3), legend=c("theta2=-0.1",
"theta2=+0.1"))
```



As time tends to infinity, the yield for both substances tends to the limiting value θ_3=0.5. One curve can be transformed into the other one by flipping the image at this value.

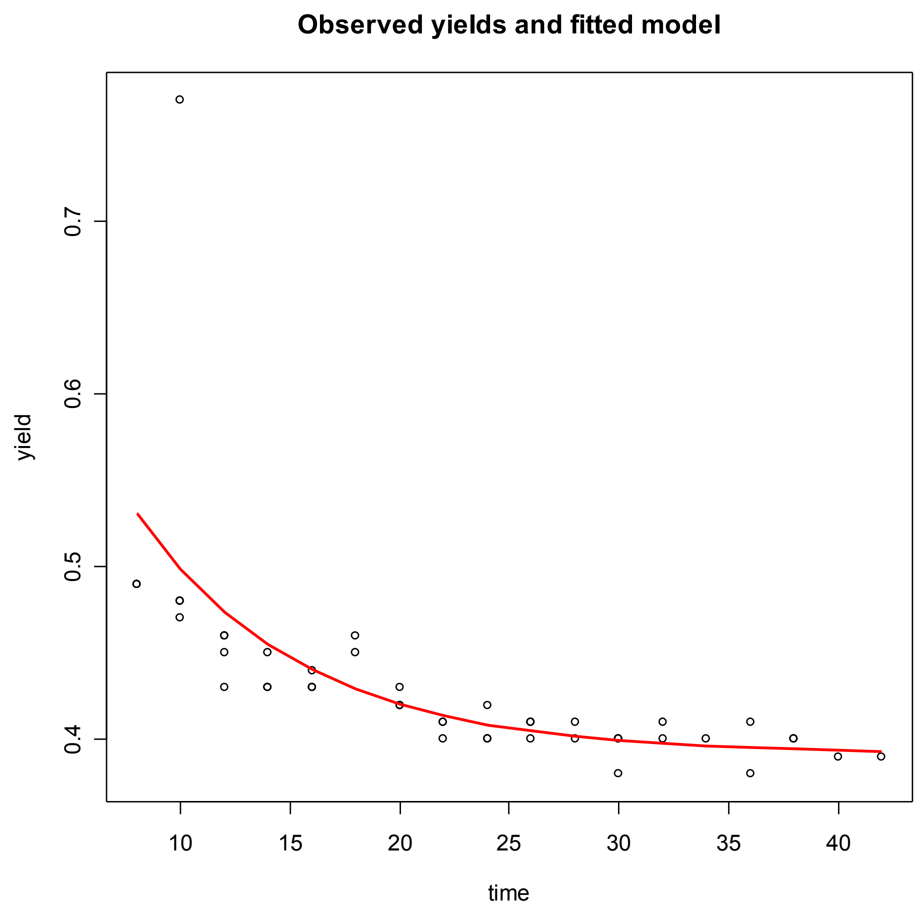
d)

```
> data<-read.table("chlorine_table.txt", header=T, sep=",")
> attach(data)
> nls1<-nls(chlorine~t1-t2*exp(-t3*time), start=c(t1=0.1, t2=-0.1, t3=0.1),
trace=T, control=nls.control(maxiter=200))
> summary(nls1)

Formula: chlorine ~ t1 - t2 * exp(-t3 * time)

Parameters:
   Estimate Std. Error t value Pr(>|t|)
t1  0.39185    0.01696  23.101   <2e-16 ***
t2 -0.39866    0.21451  -1.858   0.0703 .
t3  0.13173    0.06128   2.149   0.0375 *

> plot(time, chlorine, xlab="time", ylab="yield", main="Observed yields and fitted
model")
> lines(time, predict(nls1), col=2, lwd=2)
```



e)

```
> library(MASS)
> confint(nls1)
          2.5%      97.5%
t1  0.29961127  0.4175069
t2 -1.46600307 -0.2052151
t3  0.03146043  0.2829473
```

These are the approximate 95% confidence intervals for the parameters.

Question 2:

(a) If we use the design matrix

$$X = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

we get

$$\left(X'X\right)^{-1} X'Y = \frac{1}{n}X'Y = \frac{1}{n}\sum_{i=1}^{n} Y_i = \widehat{\mu}.$$

(b) To solve for μ_λ, we differentiated the equation of the minimization problem and set it equal to zero:

$$0 = \frac{\partial}{\partial \mu_\lambda}\left(\sum_{i=1}^{n}(Y_i - \mu_\lambda)^2 + \lambda\mu_\lambda^2\right) = -2\sum_{i=1}^{n}(Y_i - \mu_\lambda) + 2\lambda\mu_\lambda.$$

This gives

$$\widehat{\mu} = \sum_{i=1}^{n} Y_i = \lambda\mu_\lambda + n\mu_\lambda$$

or equivalently

$$\widehat{\mu}_\lambda = \frac{\widehat{\mu}}{1+\frac{\lambda}{n}}.$$

(c) We can interpret the term $\lambda\mu_\lambda^2$ is the minimization problem as $\left(y_{n+1} - x_{n+1}\right)^2$ of an additional observation. This suggests

$$X(\lambda) = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ \sqrt{\lambda} \end{pmatrix} \qquad \text{and} \qquad Y(\lambda) = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \\ 0 \end{pmatrix}.$$

This gives

$$\left(X(\lambda)' X(\lambda)\right)^{-1} X(\lambda)' Y(\lambda) = (n+\lambda)^{-1}\sum_{i=1}^{n} Y_i = \frac{1}{n+\lambda}n\widehat{\mu} = \frac{\widehat{\mu}}{1+\frac{\lambda}{n}}.$$

(d) We can generalize this to constrained regression with multiple predictors

$$\widehat{\beta}_\lambda = \arg\min_\beta \left(\sum_{i=1}^{n}(Y_i - \sum_{j=0}^{p-1} X_{ij}\beta_j)^2 + \sum_{j=0}^{p-1}\lambda_j\beta_j^2\right) = \arg\min_\beta RSS(\lambda)$$

Differentiating with respect to gives

$$\frac{\partial}{\partial\beta}RSS(\lambda) = -2(Y - X\beta)'X + 2\lambda'\beta$$

Setting this equal to zero gives

$$Y'X = \lambda'\beta + \beta'X'X = \left(X'X + diag\left(\lambda\right)\right)\beta$$

which is amounts to

$$\widehat{\beta}_\lambda = \left(X'X + diag\left(\lambda\right)\right)^{-1} X'Y.$$

(e) The following function takes a 'lm' object and a vector of λ-values as an input argument an returns $\widehat{\beta}_{\lambda}$:

```
lm.shrinkage <- function(lm1, lambda) {
    X <- model.matrix(lm1)
    Y <- lm1$fitted.values+lm1$residuals
    beta_lambda <- solve(t(X) %*% X + diag(lambda)) %*% t(X) %*% Y
    beta_lambda
}
```

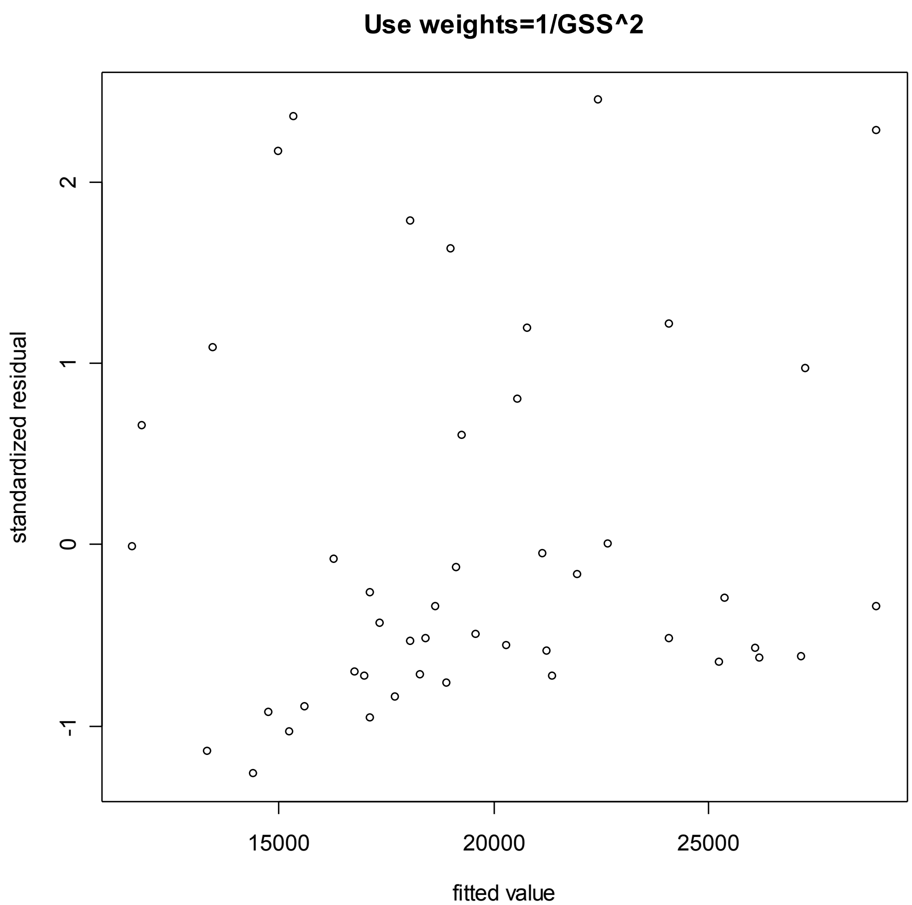
Question 3

a) First we remove the missing value and the outlier, i.e. observations 23 and 28.

```
> data <- read.table("vl_table.txt", header=T)
> data <- data[-23,]
> attach(data)
```

In homework 2 we have seen that the variance of VL increases linearly with the fitted values (or GSS). We therefore use weights of the form 1/GSS$^{\alpha}$. By trial and error we see that a value of α=2 stabilizes the variance of the residuals.

```
> w <- 1/GSS^2
> lm1 <- lm(VL~GSS, weights=w)
> plot(lm1$fitted.values, rstudent(lm1), xlab="fitted value", ylab="standardized
residual", main="Use weights=1/GSS^2")
```



b)

```
> lm2 <- lm(VL~GSS)
> summary(lm1)
> summary(lm2)
```

From summary(lm1) we see that the parameters of the new model are given by:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10288.8     2864.3   3.592 0.000822 ***
```

```
GSS           1179.6      687.1   1.717 0.093035 .
```

Whereas the old model has the form

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     2802       9296   0.301   0.7646
GSS             2271       1060   2.142   0.0378 *
```

The 95% confidence intervals are given by

```
> confint(lm1)
                2.5 %    97.5 %
(Intercept) 4516.2417 16061.434
GSS         -205.0845  2564.298
> confint(lm2)
                 2.5 %    97.5 %
(Intercept) -15933.580 21536.613
GSS            134.198  4407.684
```

The two models seem to be fundamentally different and we cannot directly compare the size of the coefficients. Note however that the model using weighted least squares gives narrower confidence intervals for parameters, since they can be estimated more efficiently.

Question 4

a)

```
> data <- read.table("car_table.txt", header=T)
> attach(data)
> glm1 <- glm(purchase~income+age, family="binomial")
> summary(glm1)

Call:
glm(formula = purchase ~ income + age, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6189  -0.8949  -0.5880   0.9653   2.0846

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.73931    2.10195  -2.255   0.0242 *
income       0.06773    0.02806   2.414   0.0158 *
age          0.59863    0.39007   1.535   0.1249
```

b)

The form of the response function is

$$\mathrm{logit}(p) = \log(p/(1-p)) = \beta_0 + \beta_{income}\cdot income + \beta_{age}\cdot age$$

or equivalently

$$p = \exp(\beta_0 + \beta_{income}\cdot income + \beta_{age}\cdot age) \,/\, (1 + \exp(\beta_0 + \beta_{income}\cdot income + \beta_{age}\cdot age))$$

where p denotes the probability of buying a new car during the next 12 months.

c)

We have that exp(β_hat_income) = 1.07 and exp(β_hat_age) = 1.82. The estimated odds ratio for purchasing a new car is multiplied by this factor when income (respectively age) increases by one unit. This means that the odds of buying a new car increase by 7% for every additional unit ($1,000) of income and 82% for every additional unit of age (supposedly years).

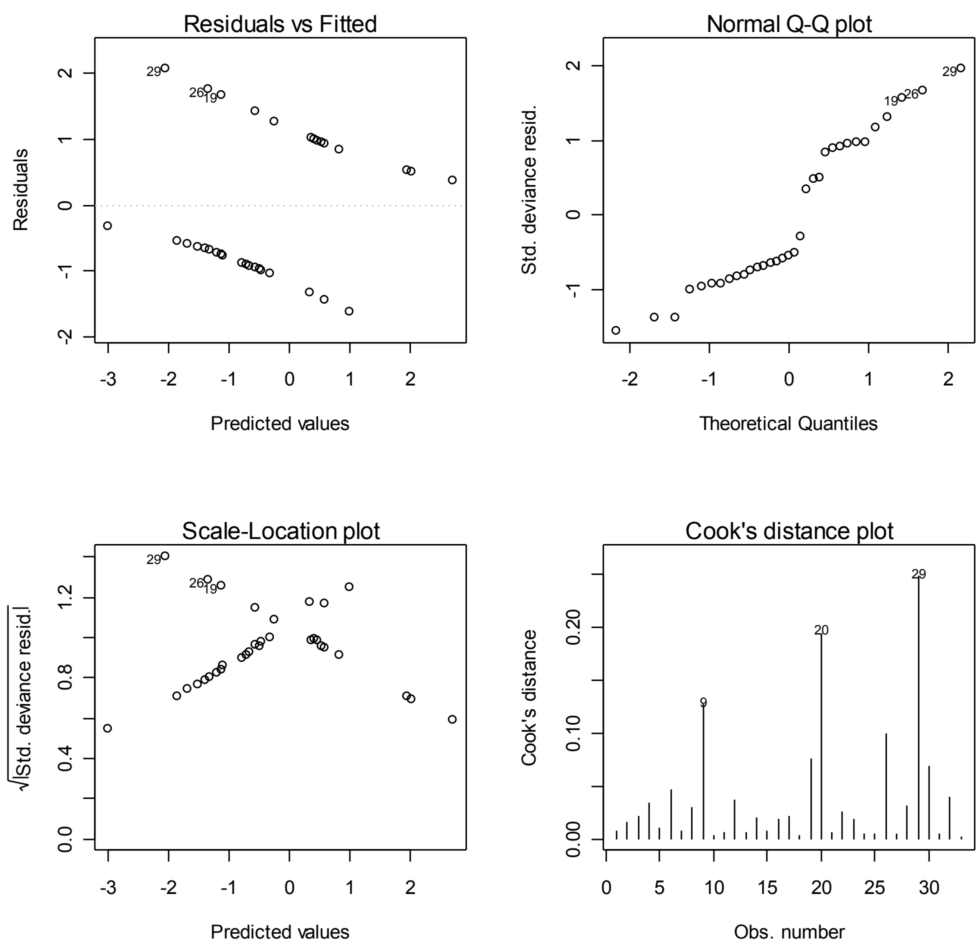
d)

Plugging into the formula given in part b) or using R we get an estimated probability of 60.9%.

```
> predict(glm1, data.frame(age=3, income=50), type="response")
[1] 0.6090245
```

e)

```
> par(mfrow=c(2,2))
> plot(glm1)
```



The diagnostic plots show that the residuals do not follow a normal distribution as expected since the response variable is binary. There are no obvious outliers and overall fit of the model seems to be quite good.

f)

```
> glm2<-update(glm1, .~.-age)
> anova(glm2, glm1,test="Chisq")
Analysis of Deviance Table

Model 1: purchase ~ income
Model 2: purchase ~ income + age
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        31     39.305
2        30     36.690  1    2.615     0.106
```

To see whether the variable 'age' can be dropped from the model, we perform a partial deviance test. The test statistic is 2.615. Comparing this with a chi-squared distribution with one degree of freedom gives an approximate (because the assumption of normally distributed errors is not satisfied) p-value of 0.106. We therefore reject the null hypothesis that age has no influence at confidence level α=0.15.

g)

```
> glm3<-update(glm1, .~.+age:income)
> summary(glm3)

Call:
glm(formula = purchase ~ income + age + income:age, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6096  -0.8222  -0.5334   0.8731   1.9924

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.372993   2.862477  -0.829    0.407
income       0.001326   0.064770   0.020    0.984
age         -0.303860   0.890512  -0.341    0.733
income:age   0.028860   0.026493   1.089    0.276

> anova(glm3, glm1)
Analysis of Deviance Table

Model 1: purchase ~ income + age + income:age
Model 2: purchase ~ income + age
  Resid. Df Resid. Dev Df Deviance
1        29     35.404
2        30     36.690 -1   -1.286

> 1-pchisq(1.286, 1)
[1] 0.2567864
```

We again use the partial deviance test to determine whether the interaction term of 'age' and 'income' can be dropped. The test statistics is -1.286 with one degree of freedom. This corresponds to an approximated p-value of 0.257. We therefore do not reject the null hypothesis at level α=0.05 that the interaction age:income is not relevant.

h)

```
> ptest <- function (F.glm, R.glm, phi=1)
{
        pear.F <- sum(resid(F.glm, type="pearson")^2)
        pear.R <- sum(resid(R.glm, type="pearson")^2)
        df <- R.glm$df.residual - F.glm$df.residual
```

```
        pp <- (pear.R-pear.F)/phi
        pval <- 1-pchisq(pp, df)
        return(data.frame(pp, df, pval))
}
> ptest(glm3, glm1)
         pp df      pval
1 0.1901181  1 0.6628184
```

Using a test based on the Pearson's X^2 we get an approximate p-value of 0.663 (or 1-0.663=0.337). We therefore do not reject the null hypothesis that the interaction effect between 'age' and 'income' is not important.