36-401 Modern Regression HW #9 Solutions

DUE: 12/1/2017 at 3PM

Problem 1 [44 points]

(a) (7 pts.)

Let

$$SSE = \sum_{i=1}^{n}(Y_i - \beta X_i)^2.$$

$$\frac{\partial}{\partial\beta}SSE = -2\sum_{i=1}^{n}(Y_i - \beta X_i)X_i$$

Set

$$\frac{\partial}{\partial\beta}SSE = 0.$$

Then,

$$\begin{aligned}
-2\sum_{i=1}^{n}(Y_i - \beta X_i)X_i &= 0 \\
\sum_{i=1}^{n}(Y_iX_i - \beta X_i^2) &= 0 \\
\beta &= \frac{\sum_{i=1}^{n} Y_iX_i}{\sum_{i=1}^{n} X_i^2}.
\end{aligned}$$

And

$$\begin{aligned}
\frac{\partial^2}{\partial\beta^2}SSE &= 2\sum_{i=1}^{n} X_i^2 \\
&> 0
\end{aligned}$$

So

$$\frac{\sum_{i=1}^{n} Y_iX_i}{\sum_{i=1}^{n} X_i^2}$$

is indeed the unique least squares estimator, denoted $\widehat{\beta}$.

(b) (7 pts.)

Let

$$\begin{aligned}
WSSE &= \sum_{i=1}^{n}\frac{(Y_i - \beta X_i)^2}{\sigma_i^2} \\
&= \sum_{i=1}^{n}\left(\frac{Y_i - \beta X_i}{\sigma_i}\right)^2
\end{aligned}$$

$$\frac{\partial}{\partial\beta}WSSE = -2\sum_{i=1}^{n}\left(\frac{Y_iX_i - \beta X_i^2}{\sigma_i^2}\right)$$

Set

$$\frac{\partial}{\partial \beta} WSSE = 0.$$

Then,

$$\beta = \frac{\sum_{i=1}^{n} \frac{Y_i X_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{X_i^2}{\sigma_i^2}}$$

And

$$\begin{aligned} \frac{\partial^2}{\partial \beta^2} WSSE &= 2 \sum_{i=1}^{n} \frac{X_i^2}{\sigma_i^2} \\ &> 0 \end{aligned}$$

So

$$\frac{\sum_{i=1}^{n} \frac{Y_i X_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{X_i^2}{\sigma_i^2}}$$

is indeed the unique weighted least squares estimator, denoted $\widetilde{\beta}$.

(c) (7 pts.)

$$\begin{aligned} \mathbb{E}[\widehat{\beta}] &= \mathbb{E}\left[\frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}\right] \\ &= \frac{\sum_{i=1}^{n} X_i \mathbb{E}[Y_i]}{\sum_{i=1}^{n} X_i^2} \\ &= \frac{\beta \sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} X_i^2} \\ &= \beta \end{aligned}$$

$$\begin{aligned} \text{Var}(\widehat{\beta}) &= \text{Var}\left(\frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}\right) \\ &= \frac{\sum_{i=1}^{n} X_i^2 \sigma_i^2}{\left(\sum_{i=1}^{n} X_i^2\right)^2} \end{aligned}$$

$$\begin{aligned} \mathbb{E}[\widetilde{\beta}] &= \mathbb{E}\left[\frac{\sum_{i=1}^{n} \frac{Y_i X_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{X_i^2}{\sigma_i^2}}\right] \\ &= \frac{1}{\sum_{i=1}^{n} \frac{X_i^2}{\sigma_i^2}} \sum_{i=1}^{n} \frac{X_i \mathbb{E}[Y_i]}{\sigma_i^2} \\ &= \frac{\beta}{\sum_{i=1}^{n} \frac{X_i^2}{\sigma_i^2}} \sum_{i=1}^{n} \frac{X_i^2}{\sigma_i^2} \\ &= \beta \end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(\widetilde{\beta}) &= \mathrm{Var}\left(\frac{\sum_{i=1}^n \frac{Y_i X_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{X_i^2}{\sigma_i^2}}\right) \\
&= \frac{1}{\left(\sum_{i=1}^n \frac{X_i^2}{\sigma_i^2}\right)^2} \mathrm{Var}\left(\sum_{i=1}^n \frac{Y_i X_i}{\sigma_i^2}\right) \\
&= \frac{1}{\left(\sum_{i=1}^n \frac{X_i^2}{\sigma_i^2}\right)^2} \sum_{i=1}^n \frac{X_i^2 \mathrm{Var}(Y_i)}{\sigma_i^4} \\
&= \frac{\sum_{i=1}^n \frac{X_i^2}{\sigma_i^2}}{\left(\sum_{i=1}^n \frac{X_i^2}{\sigma_i^2}\right)^2} \\
&= \frac{1}{\sum_{i=1}^n \frac{X_i^2}{\sigma_i^2}}
\end{aligned}$$

(d) (8 pts.)

$$\begin{aligned}
\mathrm{Var}(\widetilde{\beta}) &= \frac{1}{\sum_{i=1}^n \frac{X_i^2}{\sigma_i^2}} \\
&= \frac{\sum_{i=1}^n X_i^2 \sigma_i^2}{\sum_{i=1}^n \frac{X_i^2}{\sigma_i^2} \sum_{i=1}^n X_i^2 \sigma_i^2} \\
&= \frac{\sum_{i=1}^n X_i^2 \sigma_i^2}{\sum_{i=1}^n \left(\frac{X_i}{\sigma_i}\right)^2 \sum_{i=1}^n \left(X_i \sigma_i\right)^2} \\
&\le \frac{\sum_{i=1}^n X_i^2 \sigma_i^2}{\left(\sum_{i=1}^n X_i^2\right)^2} \\
&= \mathrm{Var}(\widehat{\beta}),
\end{aligned}$$

where the inequality comes from Cauchy-Schwartz.

(e) (7 pts.)

Above we showed

$$\widetilde{\beta} = \frac{1}{\sum_{i=1}^n \frac{X_i^2}{\sigma_i^2}} \sum_{i=1}^n \frac{X_i}{\sigma_i^2} Y_i.$$

$\widetilde{\beta}$ is a linear combination of Normal random variables Y_i, and thus also Normally distributed. We have already found the mean and variance in part (c). Therefore,

$$\widetilde{\beta} \sim N\left(\beta,\ \frac{1}{\sum_{i=1}^n \frac{X_i^2}{\sigma_i^2}}\right)$$

and a $1-\alpha$ confidence interval for β is

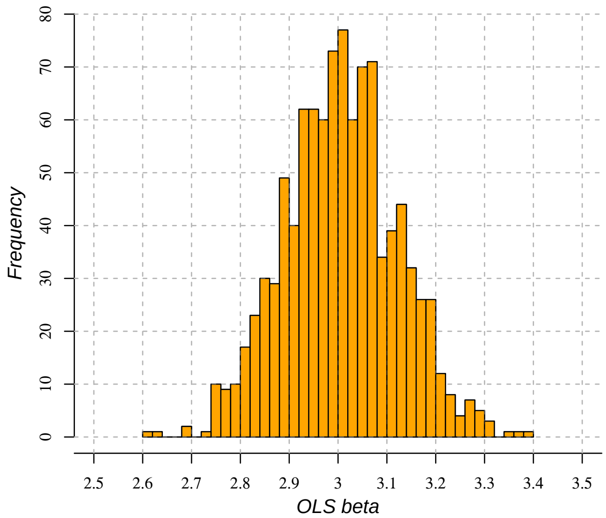
$$\widetilde{\beta} \pm z_{\alpha/2} \sqrt{\frac{1}{\sum_{i=1}^n \frac{X_i^2}{\sigma_i^2}}}.$$

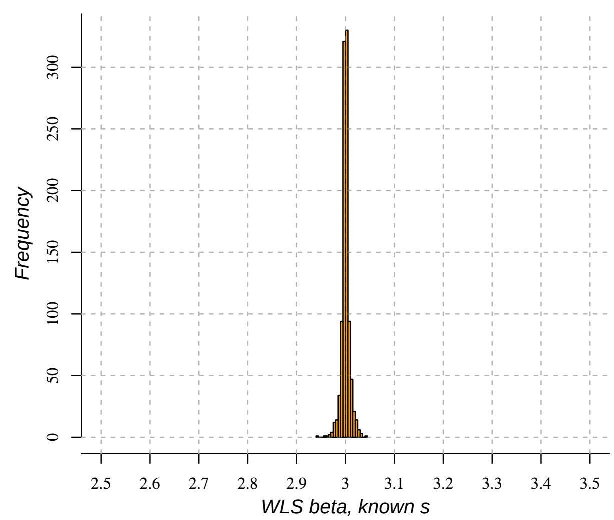
(f) (8 pts.)

```r
set.seed(100)
n = 100
b.OLS <- rep(NA,1000)
b.WLS.known.var <- rep(NA,1000)
b.WLS.unknown.var <- rep(NA,1000)
for (itr in 1:1000){
  x = runif(n)
  s = x^2
  y = 3*x + rnorm(n, mean = 0, sd = s)

  out <- lm(y ~ x - 1)
  b.OLS[itr] <- out$coefficients[1]
  out2 <- lm(y ~ x - 1, weights = 1/s^2)
  b.WLS.known.var[itr] <- out2$coefficients[1]
  u = log((resid(out))^2)
  tmp = loess(u ~ x)
  s2 = exp(tmp$fitted)

  w = 1/s2
  out3 = lm(y ~ x - 1,weights=w)
  b.WLS.unknown.var[itr] <- out3$coefficients[1]
}
```





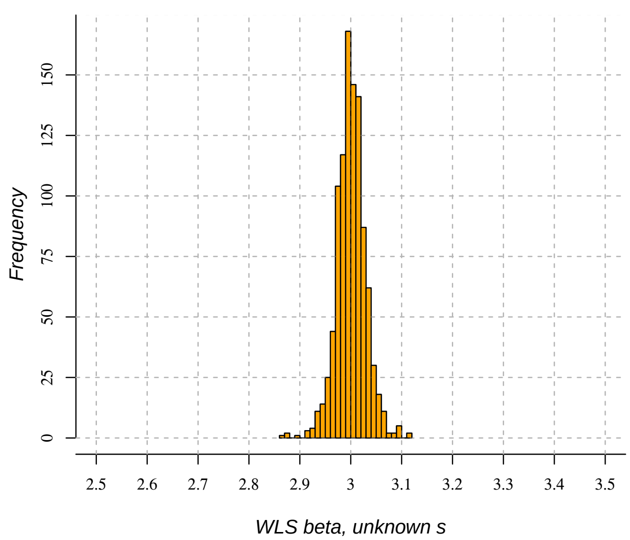


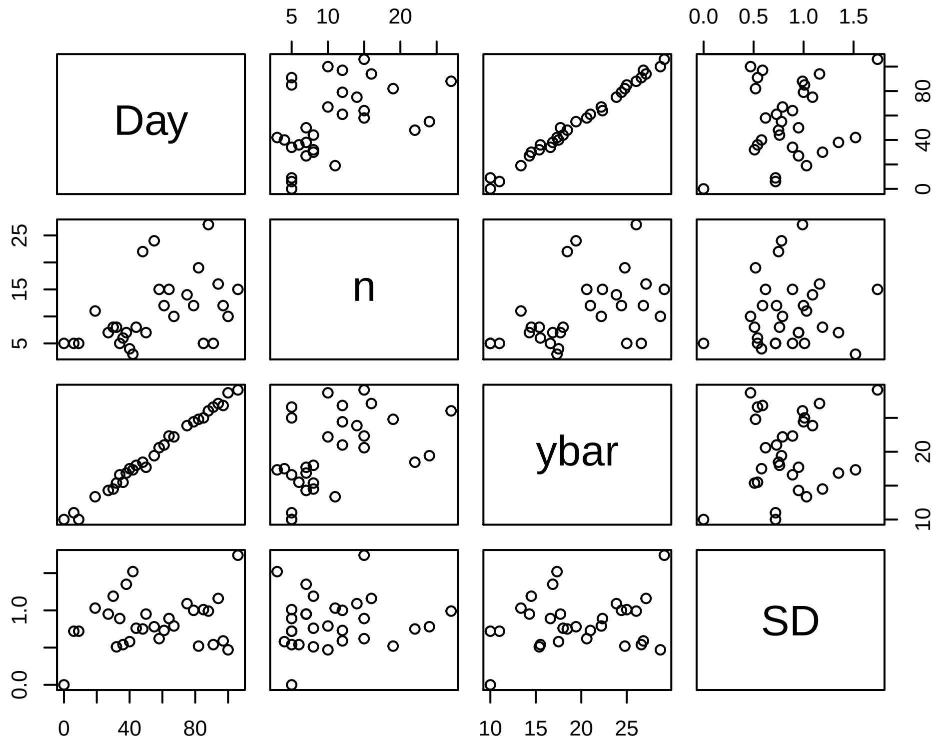
Table 1: Variances of Estimators

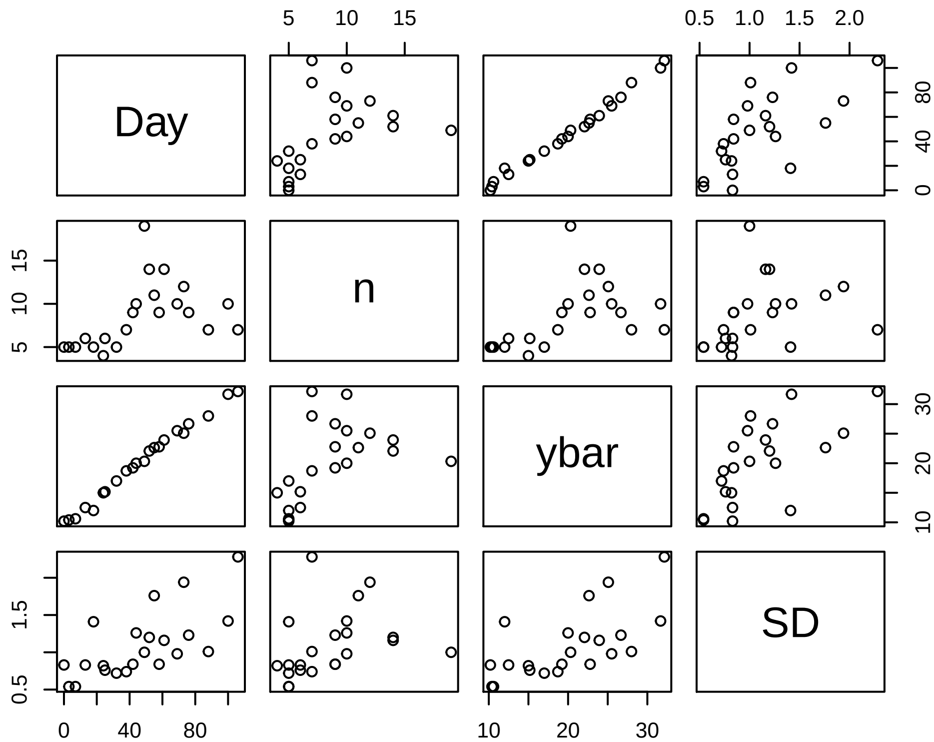
OLS	WLS.known.s	WLS.unknown.s
0.0131942	6.93e-05	0.0008159

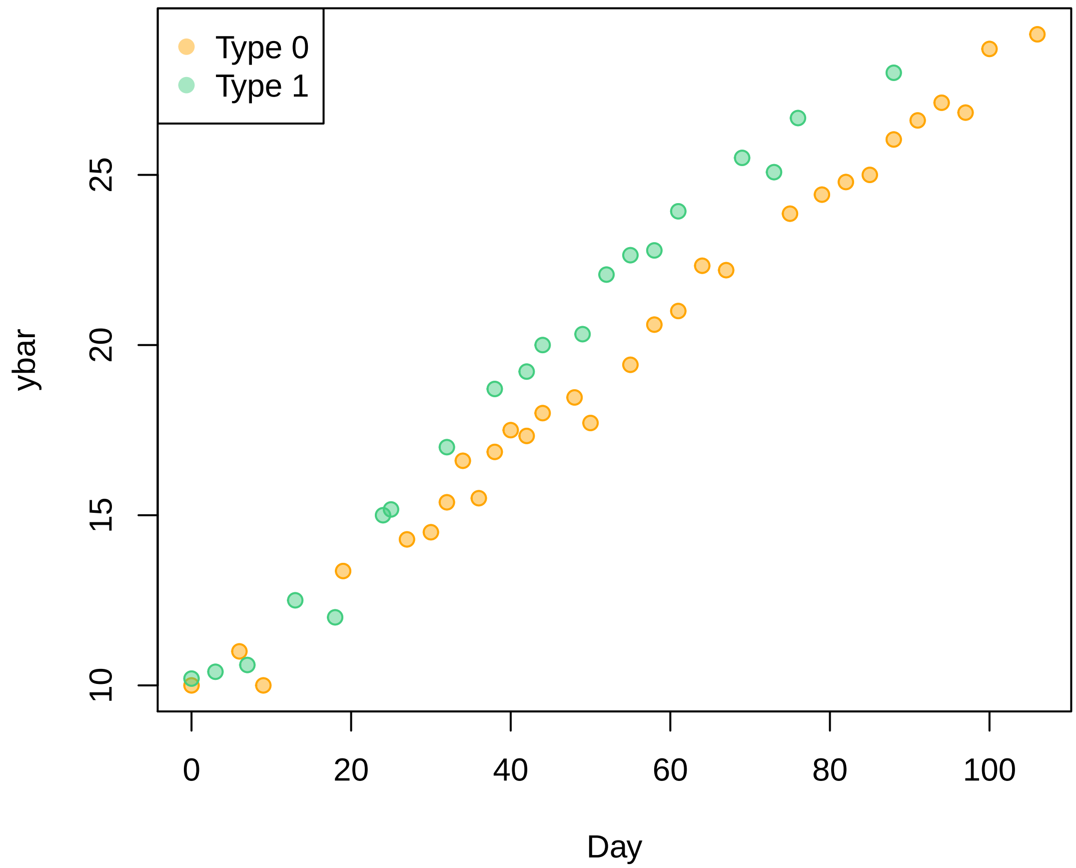
If we are dealing with a highly heteroskedastic data set such as this, and do not know the variance of the noise, using weighted least squares based on estimated variances is a better strategy.

Problem 2 [28 points]

(a) (7 pts.)







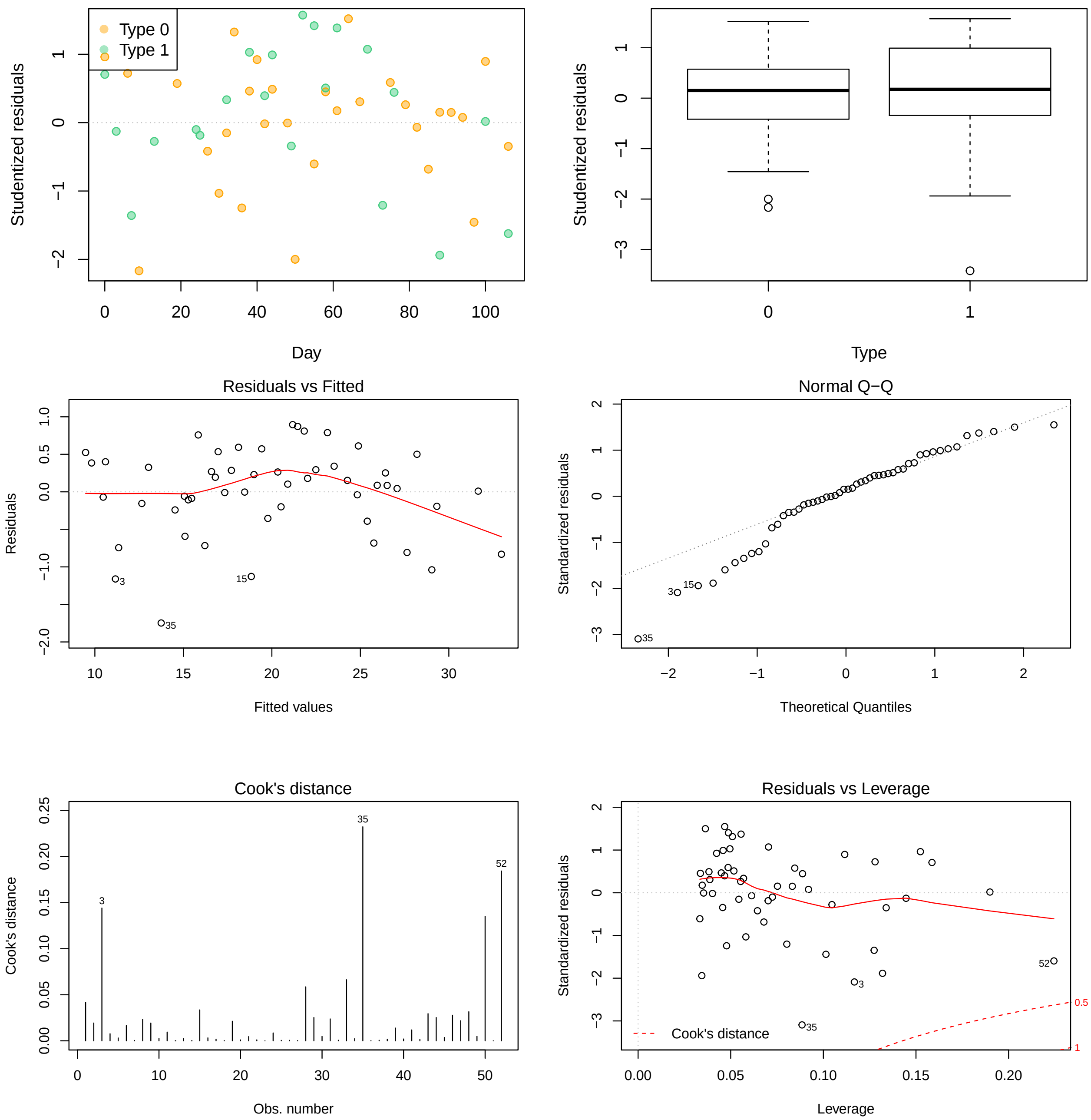
A common intercept looks feasible, however, `ybar` appears to increase at a faster rate in Type 1.

(b) (7 pts.)

```
##
## Call:
## lm(formula = ybar ~ Day * Type, data = allshoots)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74747 -0.21000  0.08631  0.35212  0.89507
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.475879   0.230981  41.025  < 2e-16 ***
## Day         0.187238   0.003696  50.655  < 2e-16 ***
## Type        0.339406   0.329997   1.029    0.309
## Day:Type    0.031217   0.005625   5.550 1.21e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5917 on 48 degrees of freedom
## Multiple R-squared:  0.9909, Adjusted R-squared:  0.9903
## F-statistic:  1741 on 3 and 48 DF,  p-value: < 2.2e-16
```

Table 2: 90% confidence intervals for regression coefficeints

	5 %	95 %
(Intercept)	9.0884732	9.8632855
Day	0.1810386	0.1934377
Type	-0.2140739	0.8928853
Day:Type	0.0217825	0.0406507



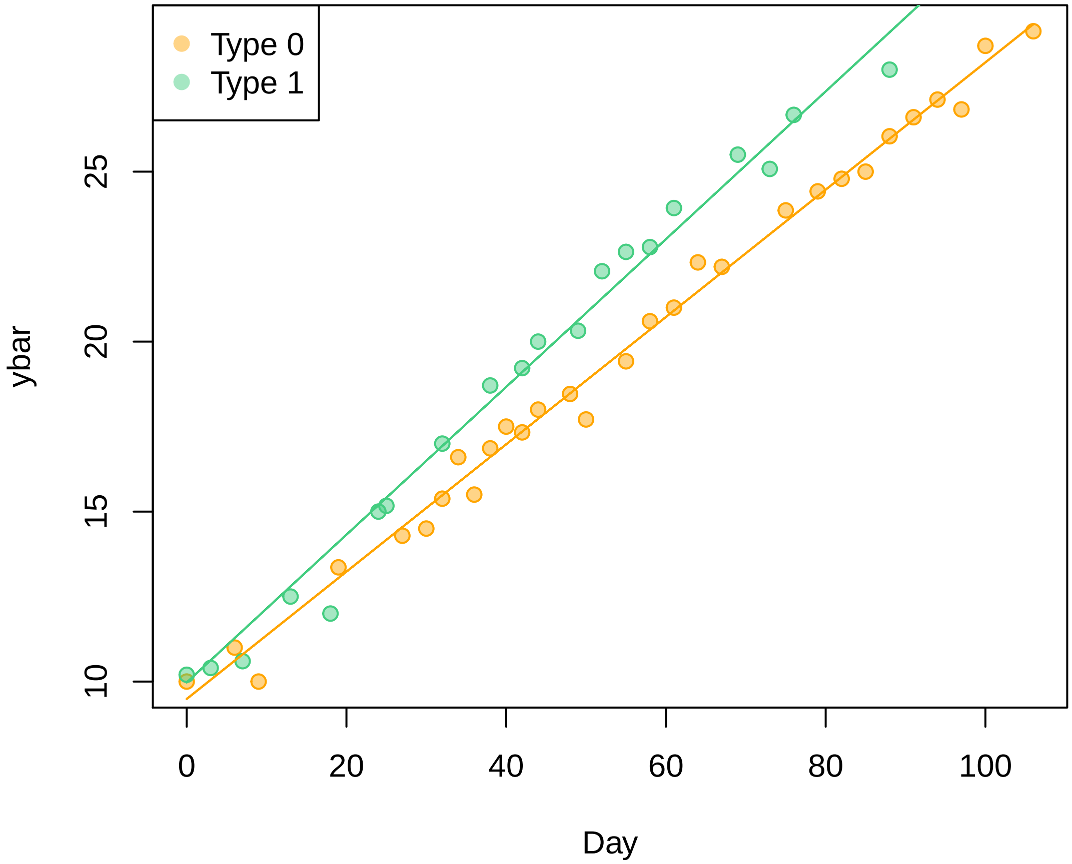
(c) (7 pts.)

```
##
## Call:
## lm(formula = ybar ~ Day * Type, data = allshoots, weights = n)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2166 -0.8300  0.1597  0.9882  3.3196
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.488374   0.238615  39.764  < 2e-16 ***
## Day         0.187258   0.003486  53.722  < 2e-16 ***
## Type        0.485380   0.362496   1.339    0.187
## Day:Type    0.030072   0.005800   5.185 4.28e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.675 on 48 degrees of freedom
## Multiple R-squared:  0.9906, Adjusted R-squared:  0.9901
## F-statistic:  1695 on 3 and 48 DF,  p-value: < 2.2e-16
```

Table 3: 90% confidence intervals for weighted regression coefficeints

	5 %	95 %
(Intercept)	9.0881641	9.8885842
Day	0.1814118	0.1931043
Type	-0.1226072	1.0933663
Day:Type	0.0203446	0.0397999

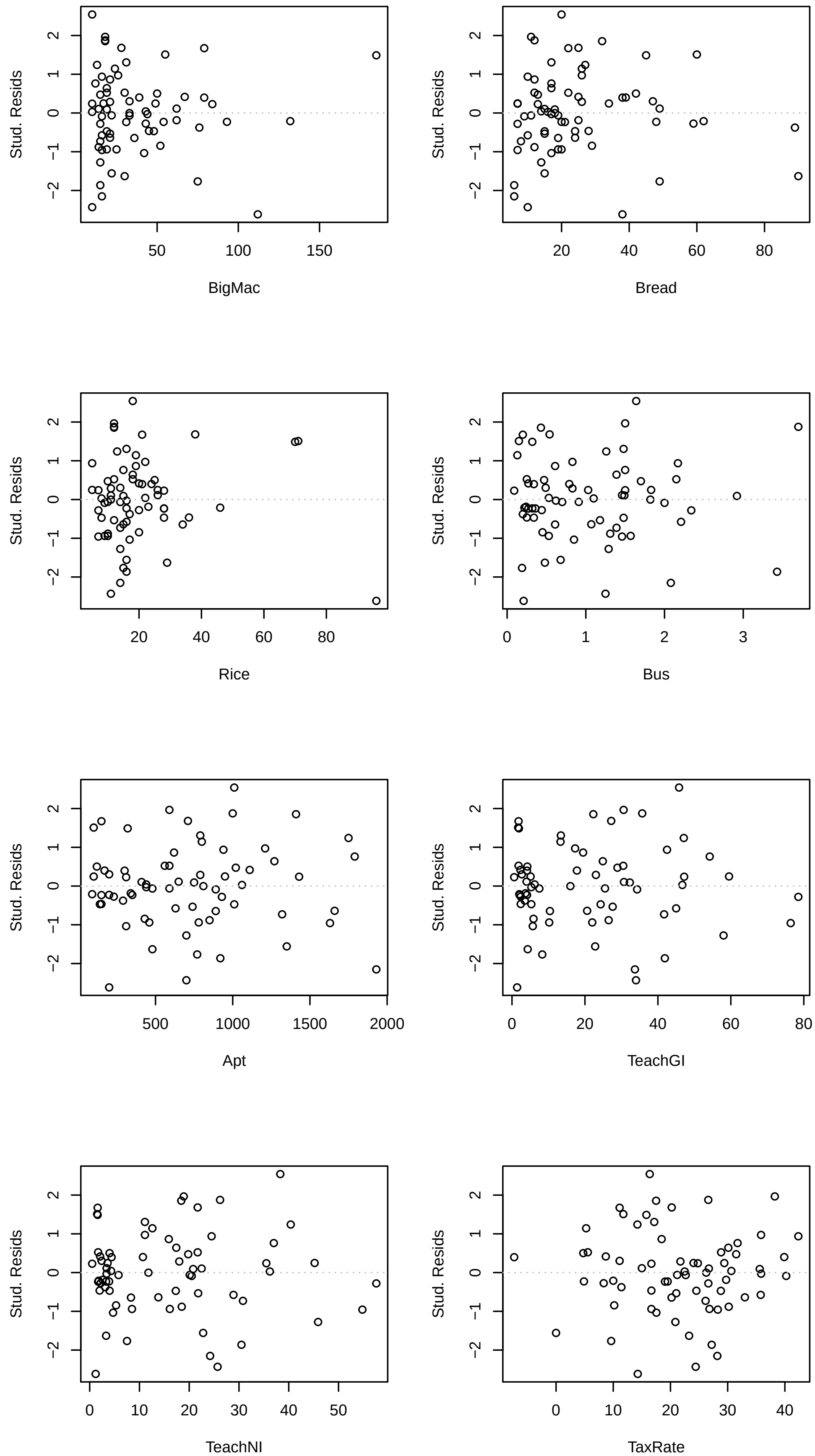
(d) (7 pts.)

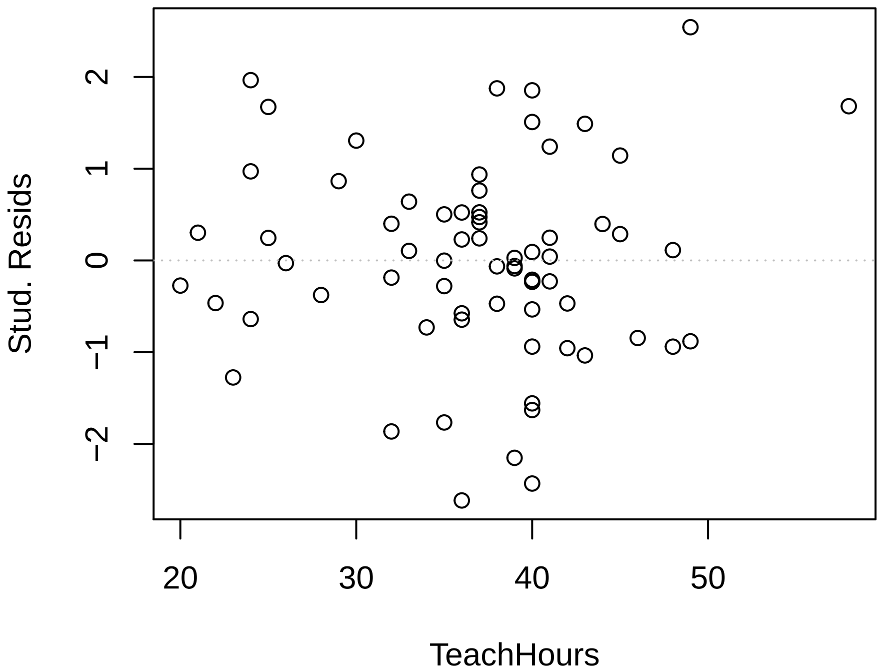


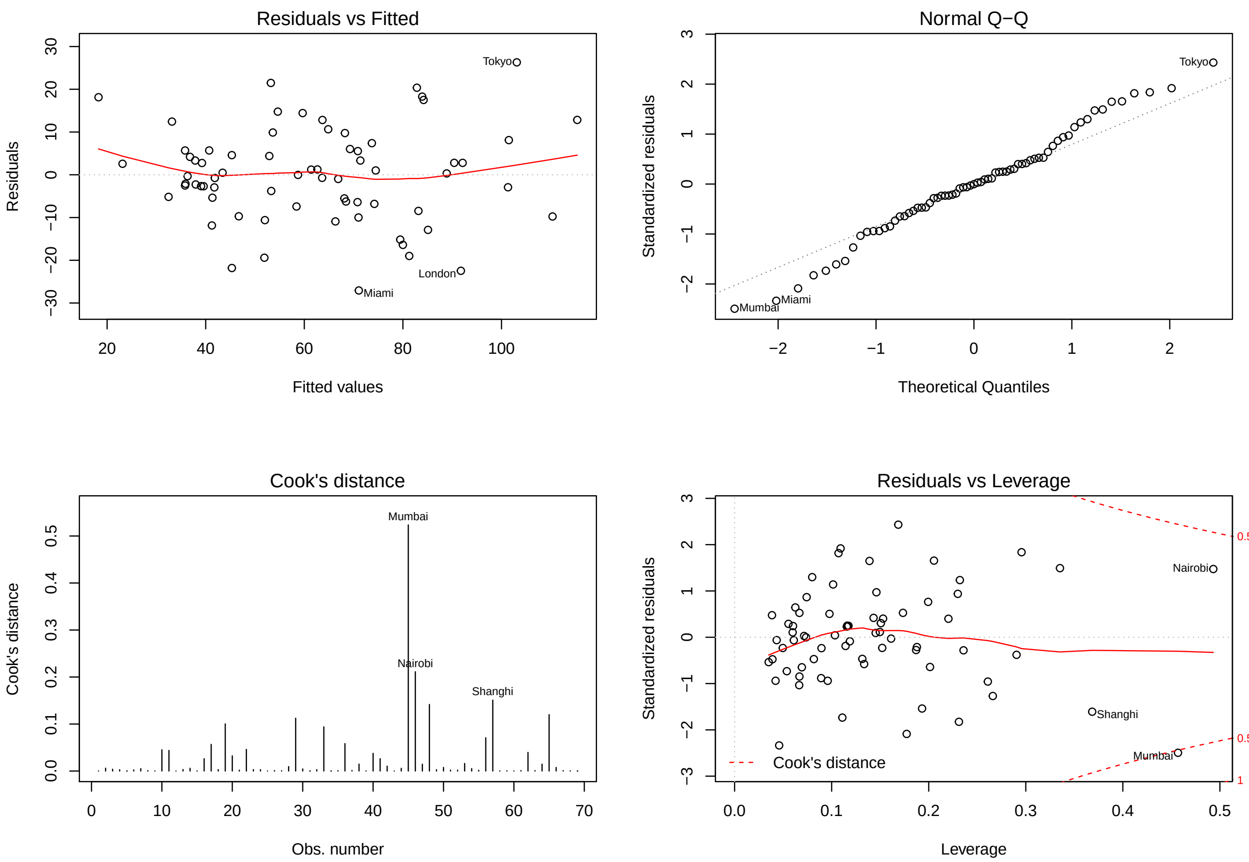
Problem 3 [28 points]

(a) (7 pts.)

```
##
## Call:
## lm(formula = FoodIndex ~ ., data = BigMac2003)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.0642  -6.3965  -0.0262   5.6928  26.3002
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.09968   11.19872  -0.098   0.9221
## BigMac      -0.20569    0.07798  -2.638   0.0107 *
## Bread        0.44383    0.10564   4.201 9.11e-05 ***
## Rice         0.26881    0.13597   1.977   0.0527 .
## Bus          3.59014    2.83317   1.267   0.2101
## Apt          0.01825    0.00434   4.204 9.02e-05 ***
## TeachGI     -0.97768    0.86750  -1.127   0.2643
## TeachNI      2.22275    1.13819   1.953   0.0556 .
## TaxRate      0.26530    0.25724   1.031   0.3066
## TeachHours   0.48015    0.20478   2.345   0.0224 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.86 on 59 degrees of freedom
## Multiple R-squared:  0.7981, Adjusted R-squared:  0.7673
## F-statistic: 25.91 on 9 and 59 DF,  p-value: < 2.2e-16
```


Stud. Resids
2
1
0
−1
−2
50
100
150
BigMac
20
40
60
80
Bread
20
40
60
80
Rice
0
1
2
3
Bus
500
1000
1500
2000
Apt
0
20
40
60
80
TeachGI
0
10
20
30
40
50
TeachNI
0
10
20
30
40
TaxRate


Stud. Resids
TeachHours


Residuals vs Fitted
Residuals
Fitted values
Tokyo
Miami
London
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Tokyo
Mumbai
Miami
Cook's distance
Cook's distance
Obs. number
Mumbai
Nairobi
Shanghai
Residuals vs Leverage
Standardized residuals
Leverage
Nairobi
Shanghai
Mumbai
Cook's distance

(b) (7 pts.)

	0.5 %	99.5 %
BigMac	-0.4132679	0.0018835

The confidence interval includes 0 so, given all the other variables, we cannot conclude the price of a BigMac has a significant association with Food Index at level $\alpha = 0.01$.

(c) (7 pts.)

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
67	27532.922	NA	NA	NA	NA
59	8299.912	8	19233.01	17.08975	0

We are testing the hypothesis that all other variables are conditionally uncorrelated with Food Index, given the price of a BigMac. The ANOVA table shows there is very strong evidence in favor of the alternative (we reject).

(d) (7 pts.)

```r
library(DAAG)
out1 <- cv.lm(df = BigMac2003, form.lm = formula(FoodIndex ~ .), m = 10, plotit = F)
out2 <- cv.lm(df = BigMac2003, form.lm = formula(FoodIndex ~ BigMac), m = 10, plotit = F)
```

The predictive MSEs of each model are estimated by 10-fold cross validation. We conclude the model only utilizing `BigMac` has better predictive accuracy.

$$\widehat{\text{Err}}_{full} = 1764, \quad \widehat{\text{Err}}_{BigMac} = 472$$