36-401 Modern Regression HW #6 Solutions

Problem 1 [32 points]

(a) (4 pts.)

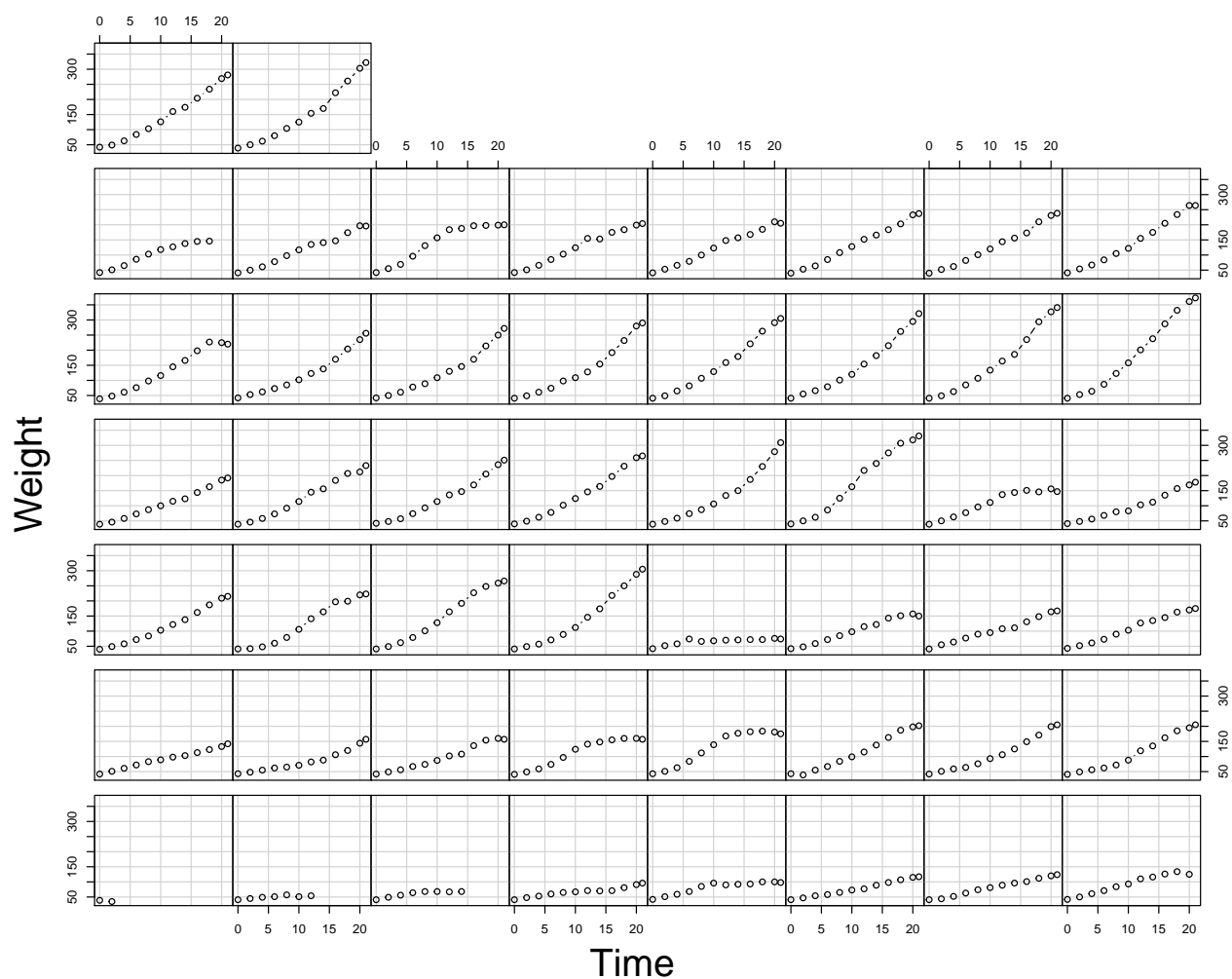Figure 1: Weight Progression of 50 Chicks in the ChickWeight Data Set

(b) (9 pts.)

The results of a naive simple linear fit of Chick `Weight` on `Time` are shown in Figure 2. The right panel shows very significant residual correlation. Therefore, the model does not fit well.
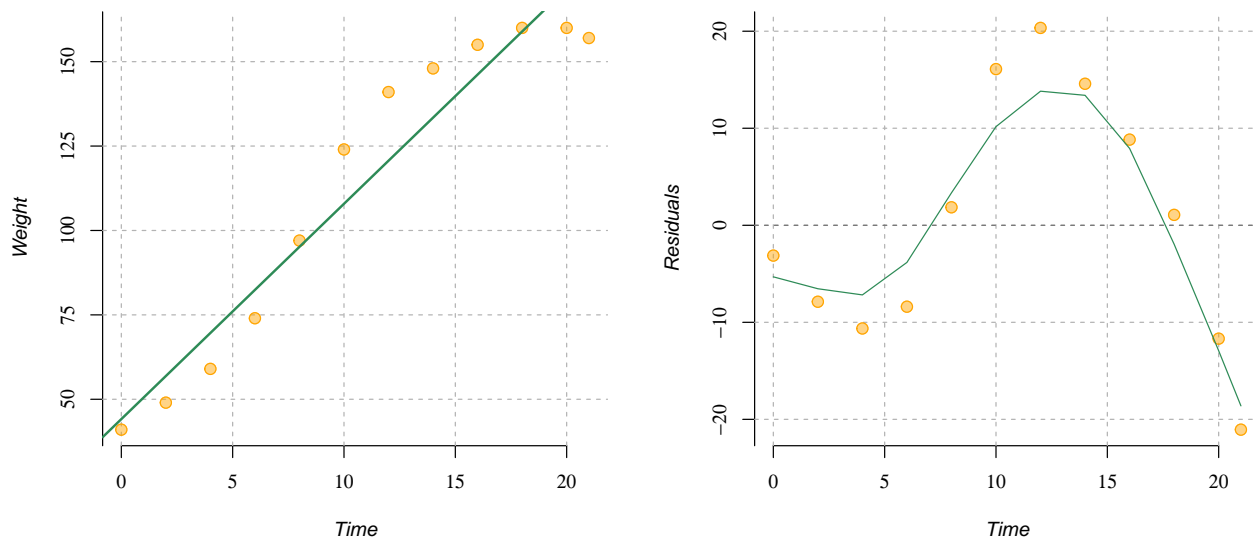

Figure 2: Results of Naive Linear Regression of Chick 6 Weight on Time

Figure 3 displays the results of a degree four[1] polynomial regression of `Weight` on `Time`. The lack of visible correlation in the residuals suggests this is a much more suitabe fit.
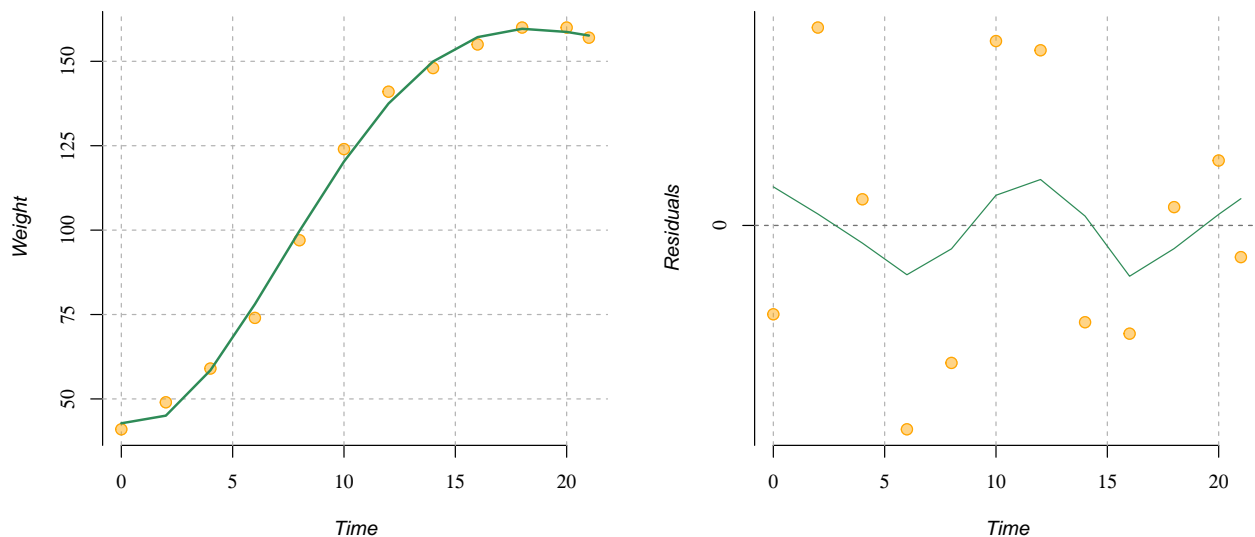

Figure 3: Results of Polynomial Regression of Chick 6 Weight on Time

---

[1]We chose this heuristically.

Table 1: Summary of Polynomial Regression for Chick 6

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.7758402	3.2118729	13.318036	0.0000032
Time	-2.6030894	2.3499713	-1.107711	0.3045914
I(Time^2)	2.1242507	0.4898299	4.336711	0.0034098
I(Time^3)	-0.1324975	0.0359324	-3.687415	0.0077831
I(Time^4)	0.0023642	0.0008494	2.783231	0.0271716

Not surprisingly, given the polynomial shape of the weight progression of Chick 6, our polynomial regression does a much better job fitting the data than the naive linear regression. The most notable aspect of our model parameters is that `Time` is no longer statistically significant given its higher order terms.
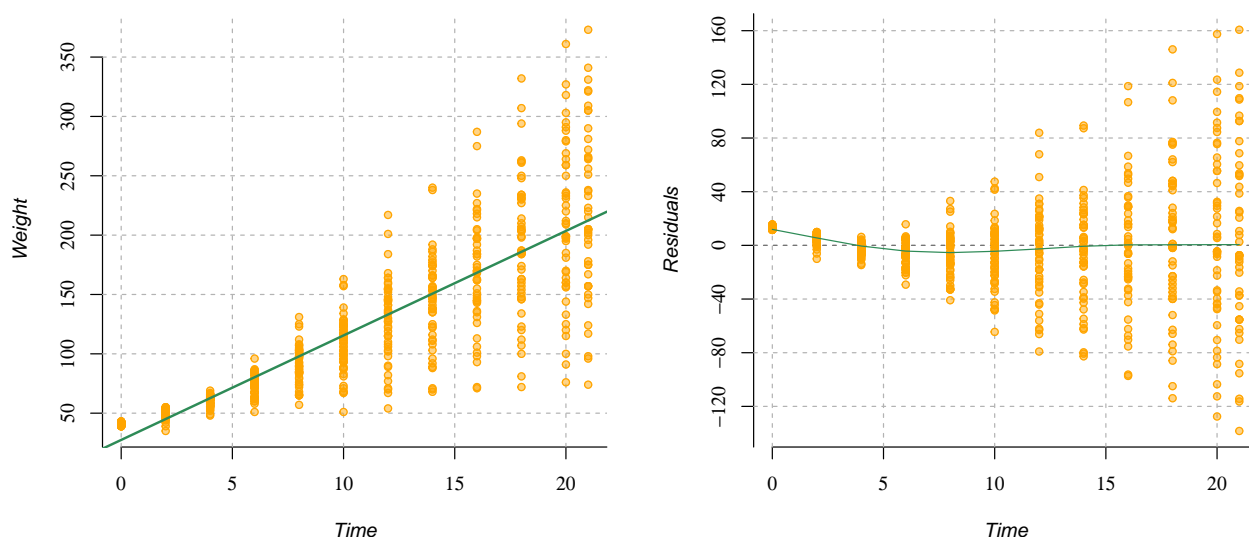
(c) (9 pts.)

Figure 4: Results of Naive Linear Regression of Chick Weight on Time

As we saw with Chick 6, and as we could infer from Figure 1, the weight progression of individual chicks is often more suitable to model with a polynomial than a simple linear regression. However, when the data are aggregated over all chicks, a simple linear model is actually not a poor choice. The model fit and residuals are shown in Figure 4. For the most part the line fits the data fairly well. However, some residual curvature is apparent in the earlier stages of the chick lifetime. Therefore, adding a quadratic term will likely benefit the model. Figure 5 shows that the extra term does indeed level out the residuals nicely.
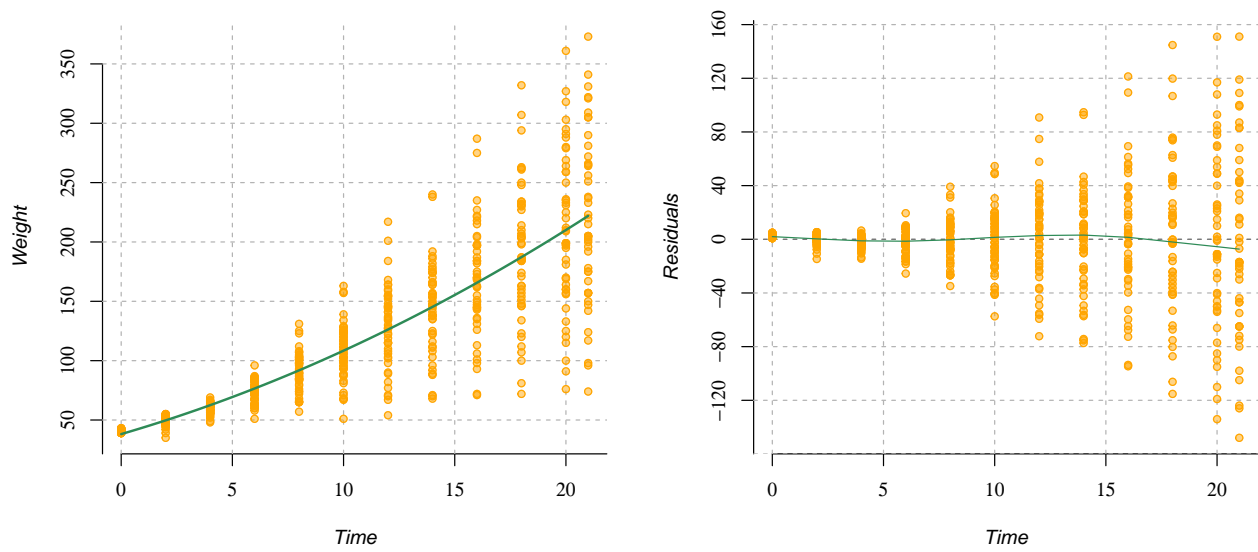
Figure 5: Results of Polynomial Regression of Chick Weight on Time
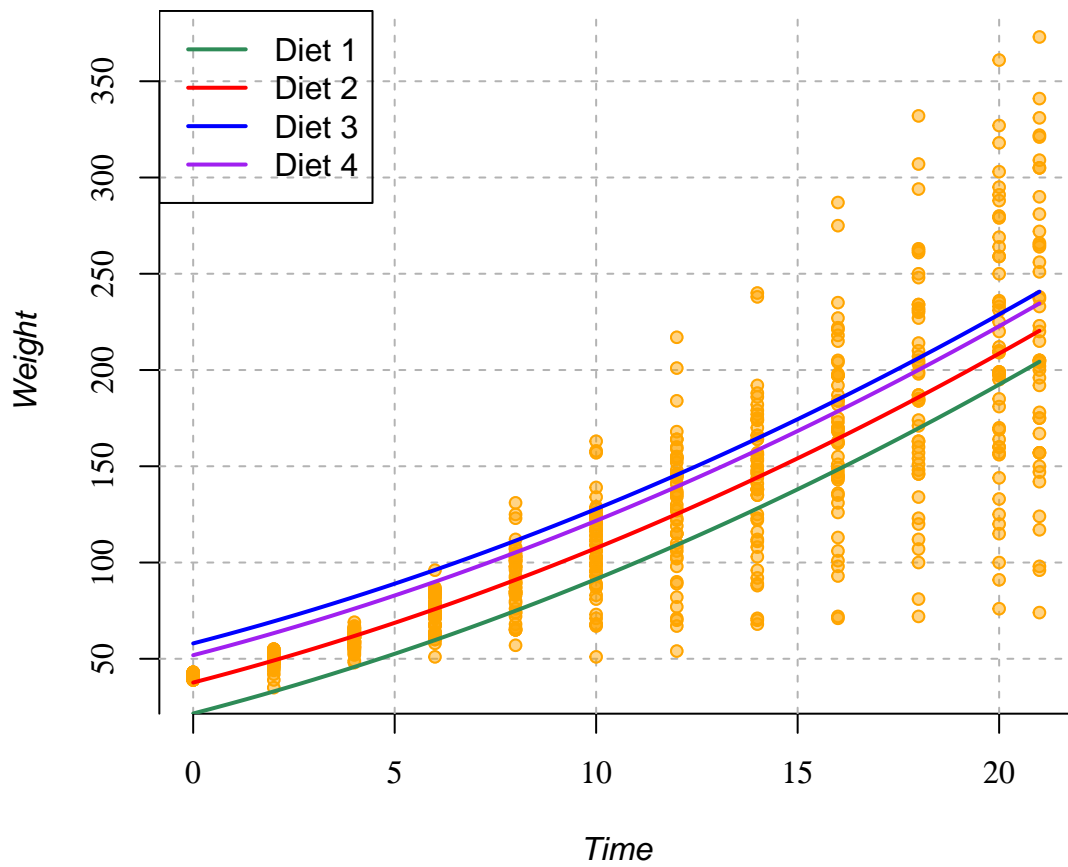
(d) (9 pts.)

Figure 6: Results of Polynomial Regression of Chick Weight on Time and Diet

Simply adding `Diet` as a predictor without interacting it with `Time` amounts to the growth curve for each diet group having the same shape but possibly different intercepts. As we can see in Figure 6, such a model specification is not suitable in this setting. Given the HW instructions, it is fine if you do this as long as you comment on the ill-suitedness of the model.

Since a chick's birthweight is independent of the diet it is thereafter fed, it is more sensible to force each diet group to have the same intercept, but possibly differing curvatures. The fits of this model are shown in Figure 7 (superposed) and Figure 8 (separated into respective Diet groups).
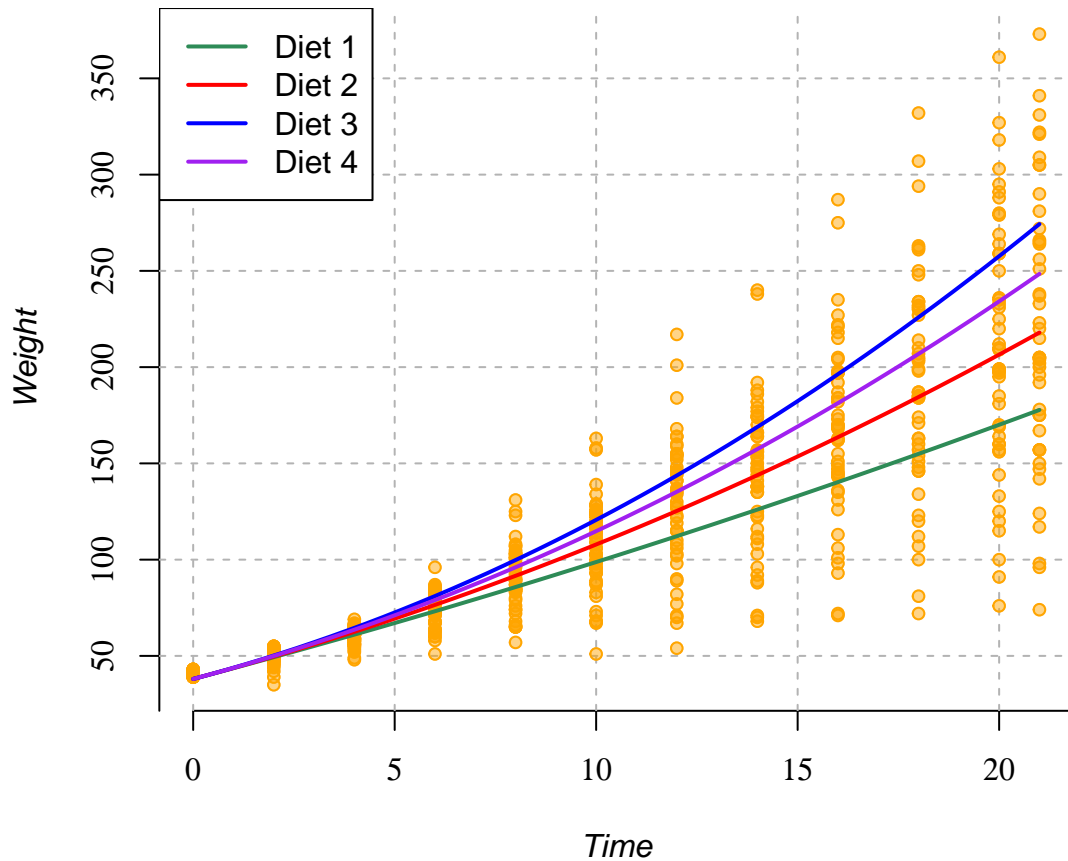

Figure 7: Polynomial Regression of Chick Weight with Time^2 and Diet Interacted
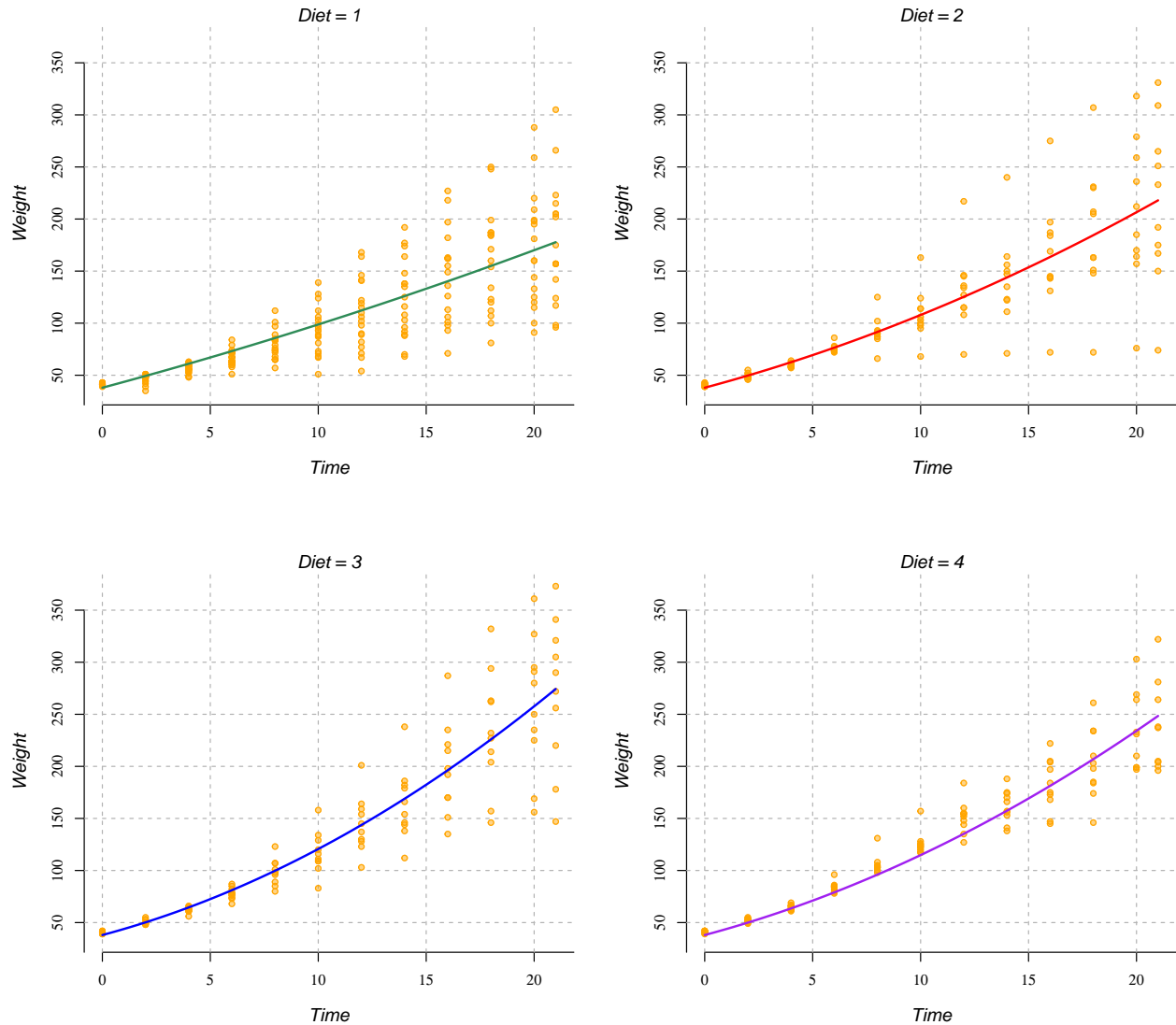
Figure 8: Polynomial Regression of Chick Weight with Time^2 and Diet Interacted

The residuals of the model are plotted vs. Time in Figure 9 (superposed) and Figure 10 (separated into respective Diet groups). The residuals look pretty healthy overall, but there is some clear correlation in Diet groups 1 and 4. A box plot can also be used to do diet-wise diagnostics of the residuals (see Figure 11). However, the significant heteroskedasticity in all diet groups renders this plot next to useless. Instead we are much better off examining Figure 10. Notice we are only a few predictor interactions shy of performing four separate regressions any way.
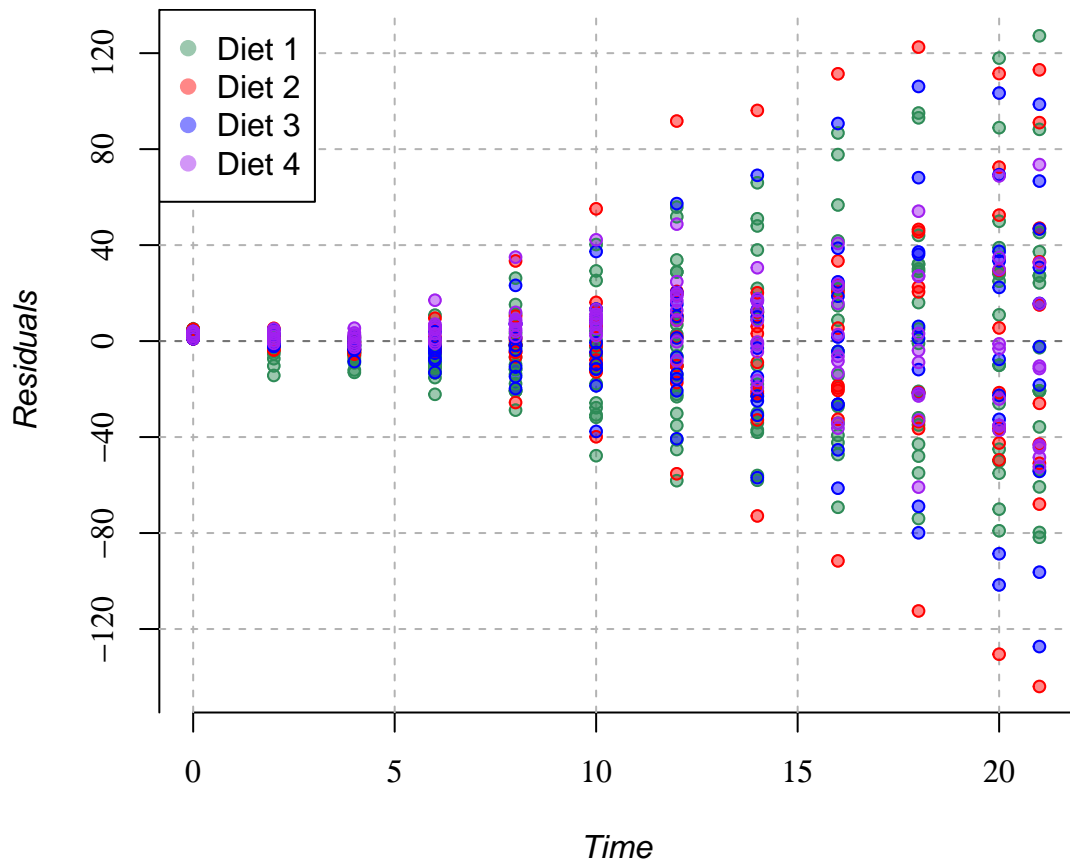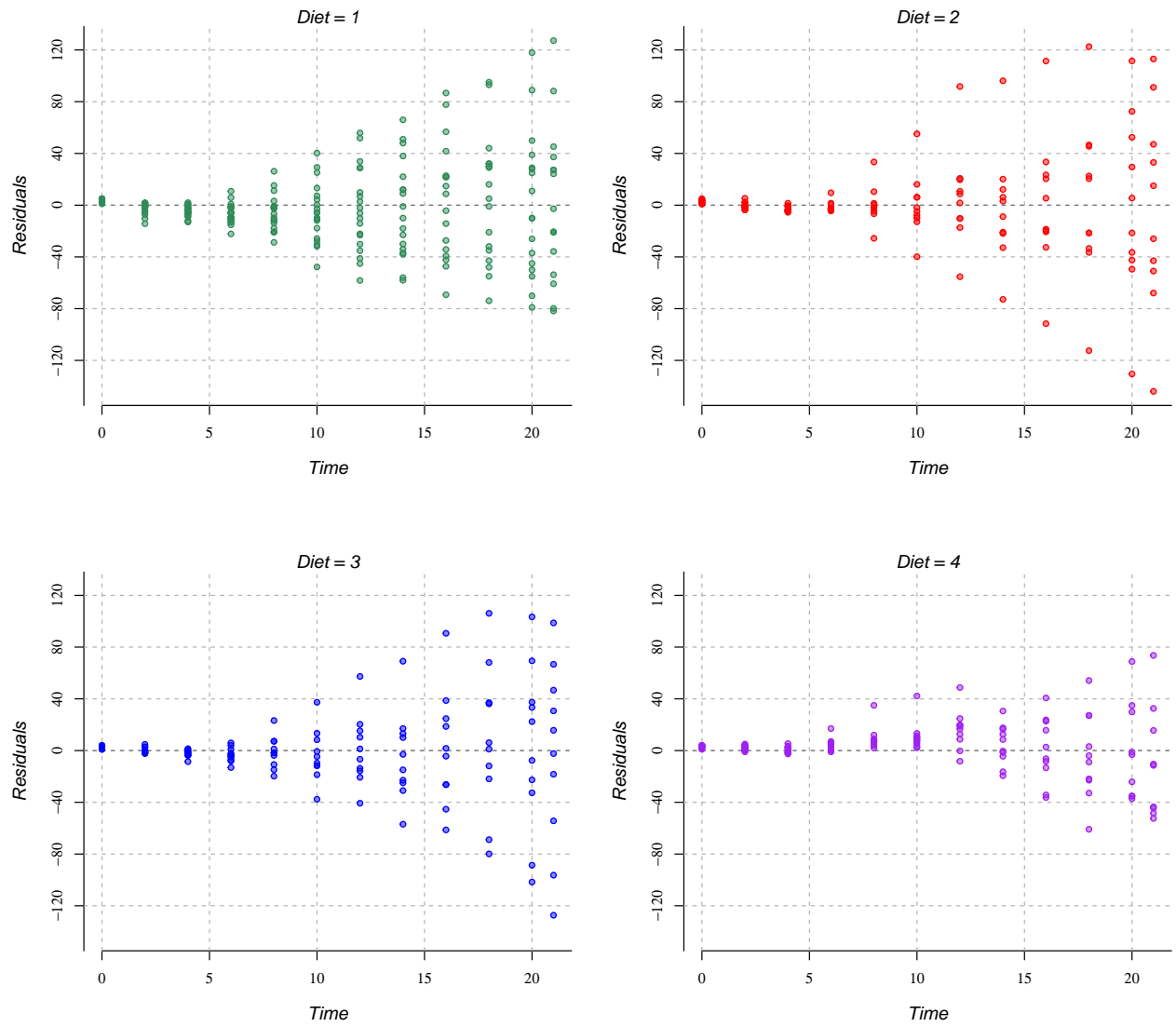
Figure 9: Residuals of chosen model
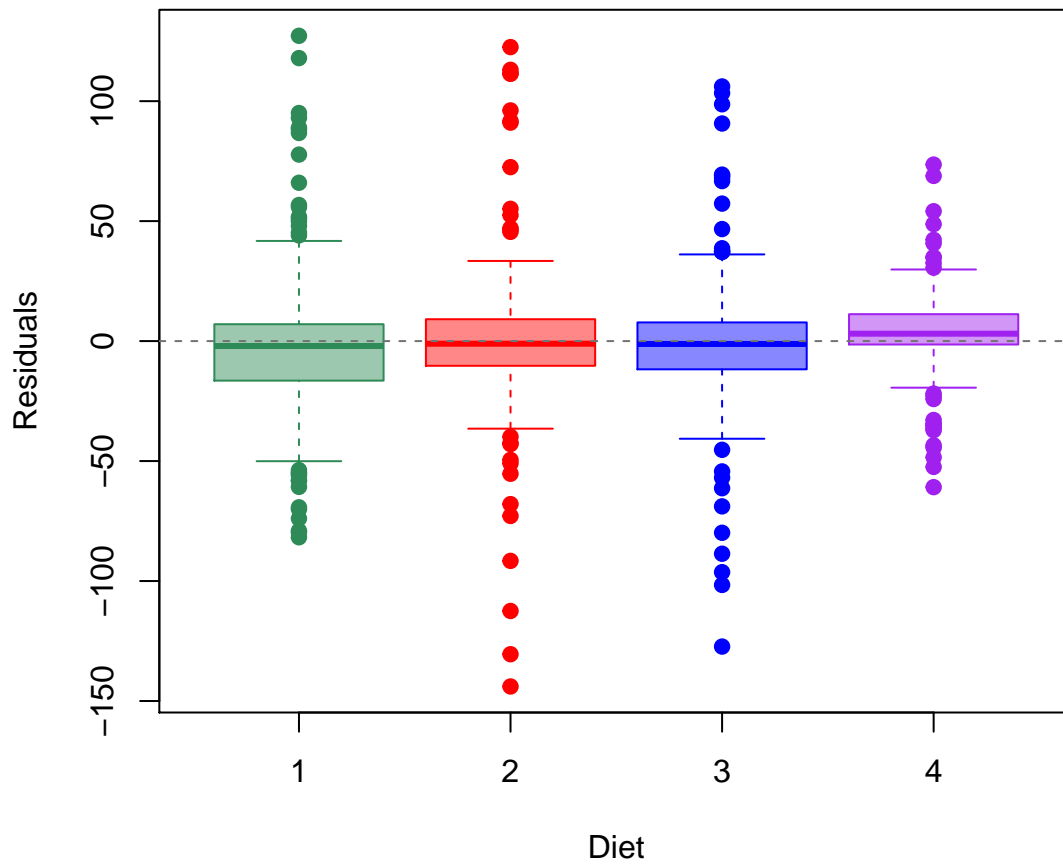
Figure 10: Residuals of chosen model

Figure 11: Diet-wise Distribution of Residuals

Problem 2 [36 points]

(a) (9 pts.)

There are about 100 ways to do this. Here is one.

Let

$$X_0 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix},$$

i.e. the intercept predictor, and

$$X = \begin{bmatrix} X_0 & X_1 & X_2 & X_3 \end{bmatrix}.$$

Now notice $X_0 = X_1 + X_2 + X_3$. That is, the columns of X are linearly dependent. Now,

$$\text{Columns of } X \text{ are LD} \implies \det(X) = 0.$$

And,

$$\det(X^T X) = \det(X^T) \cdot \det(X) = \det(X) \cdot \det(X) = 0^2 = 0.$$

$\det(X^T X) = 0$ so $X^T X$ is not invertible.

(b) (9 pts.)

```
Y <- c(33,36,35,35,31,29,31,29,37,39,36,36)
X0 <- rep(1,12)
X1 <- c(rep(1,4),rep(0,8))
X2 <- c(rep(0,4),rep(1,4),rep(0,4))
X3 <- c(rep(0,8),rep(1,4))
X <- cbind(X0,X1,X2) # leave out X3

model7 <- lm(Y ~ X - 1) # leave out default intercept since we have already included one
tmp <- summary(model7)$coefficients
rownames(tmp) <- c("(Intercept)","France","Italy")
kable(tmp, caption = "Income Regression Summary")
```

Table 2: Income Regression Summary

	Estimate	Std. Error	t value	Pr($>$|t|)
(Intercept)	37.00	0.6400955	57.803877	0.0000000
France	-2.25	0.9052317	-2.485552	0.0346742
Italy	-7.00	0.9052317	-7.732827	0.0000290

All coefficients in the regression are significant, signalling a statistically significant difference between the mean incomes of France, Italy, and the USA.

(c) (9 pts.)

As we defined the model in part (b), we have

$$\widehat{\mathbb{E}}[\texttt{Income} \mid \texttt{France}] = \widehat{\beta}_0 + \widehat{\beta}_1 \quad = 34.75 \text{ (thousand)}$$
$$\widehat{\mathbb{E}}[\texttt{Income} \mid \texttt{Italy}] = \widehat{\beta}_0 + \widehat{\beta}_2 \quad = 30.00 \text{ (thousand)}$$
$$\widehat{\mathbb{E}}[\texttt{Income} \mid \texttt{USA}] = \widehat{\beta}_0 \qquad = 37.00 \text{ (thousand)}$$

(d) (9 pts.)

Continuing with the same notation, the true mean income of France is given by $\beta_0 + \beta_1$. Assuming the noise is normally distributed implies

$$\widehat{\beta}_0 \sim N\big(\beta_0, \ \mathrm{Var}(\widehat{\beta}_0)\big) \ \text{ and } \ \widehat{\beta}_1 \sim N\big(\beta_1, \ \mathrm{Var}(\widehat{\beta}_1)\big)$$

and thus

$$\widehat{\beta}_0 + \widehat{\beta}_1 \sim N\big(\beta_0 + \beta_1, \ \mathrm{Var}(\widehat{\beta}_0 + \widehat{\beta}_1)\big)$$
$$N\big(\beta_0 + \beta_1, \ \mathrm{Var}(\widehat{\beta}_0) + \mathrm{Var}(\widehat{\beta}_1) + 2 \cdot \mathrm{Cov}(\widehat{\beta}_0 + \widehat{\beta}_1)\big).$$

Therefore, a 95% confidence interval for $\beta_0 + \beta_1$ (mean income of France) is given by

$$(\widehat{\beta}_0 + \widehat{\beta}_1) \pm z_{0.025} \cdot \sqrt{\mathrm{Var}(\widehat{\beta}_0 + \widehat{\beta}_1)}.$$

$\mathrm{Var}(\widehat{\beta}_0 + \widehat{\beta}_1)$ is not known so we replace it with its unbiased estimator and use the t-distribution. This yields the 95% confidence interval

$$(\widehat{\beta}_0 + \widehat{\beta}_1) \pm t_{n-1}(0.025) \cdot \sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_0 + \widehat{\beta}_1)}$$
$$\implies (\widehat{\beta}_0 + \widehat{\beta}_1) \pm t_3(0.025) \cdot \sqrt{\widehat{\mathrm{Var}}(\widehat{\beta}_0) + \widehat{\mathrm{Var}}(\widehat{\beta}_1) + 2 \cdot \widehat{\mathrm{Cov}}(\widehat{\beta}_0, \widehat{\beta}_1)}$$
$$\implies 34.75 \pm 2.037,$$

where we have gotten the estimated errors from the `vcov` function.
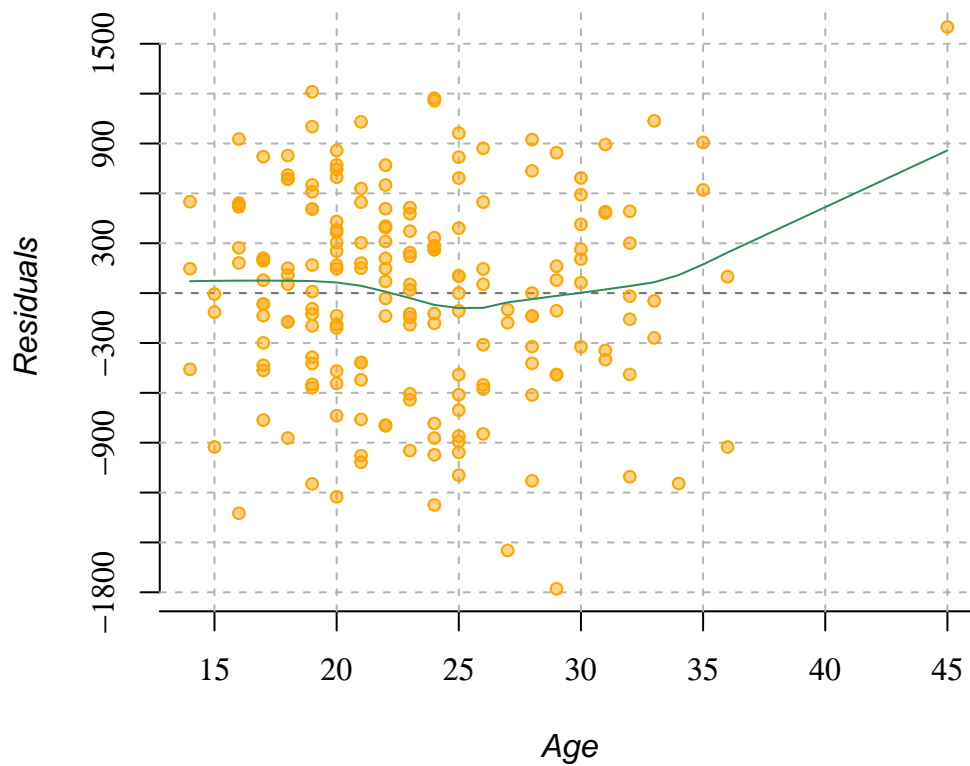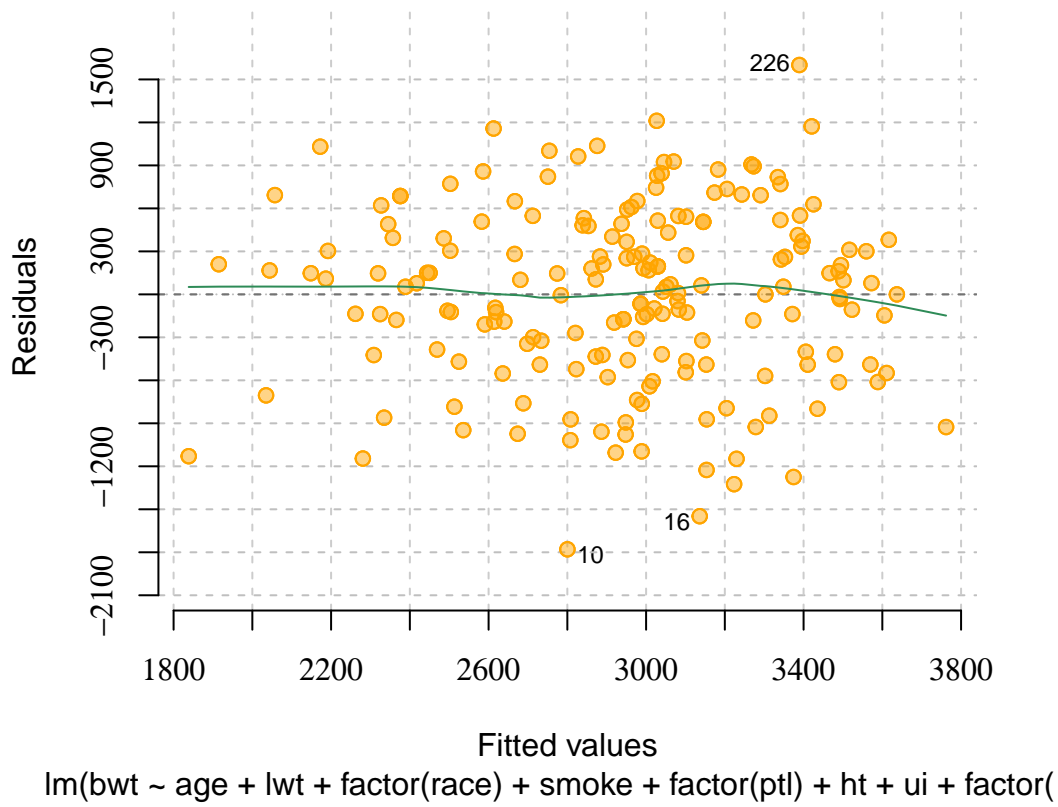
Problem 3 [32 points]

(a) [5 pts.]
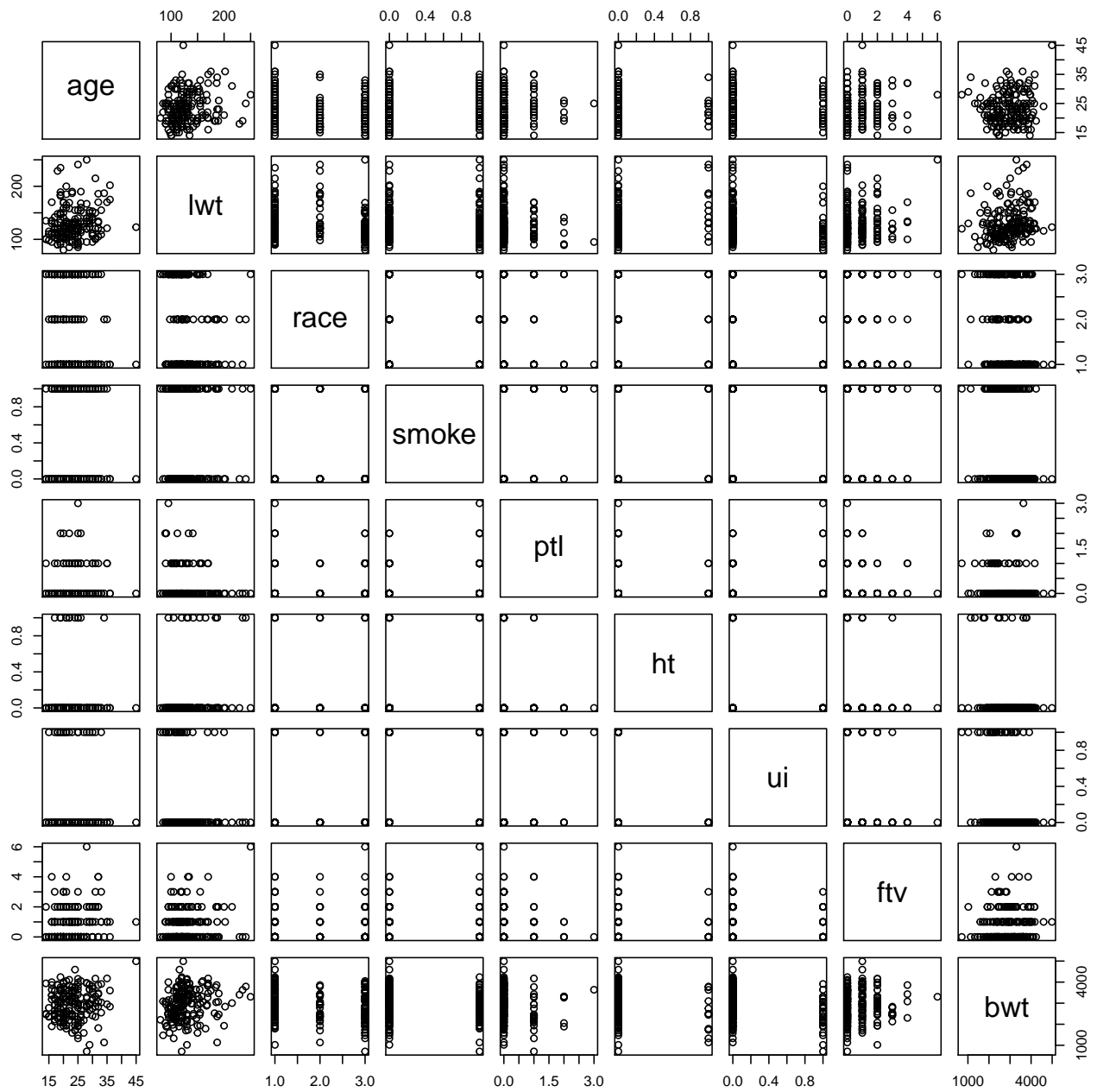
(b) [9 pts.]
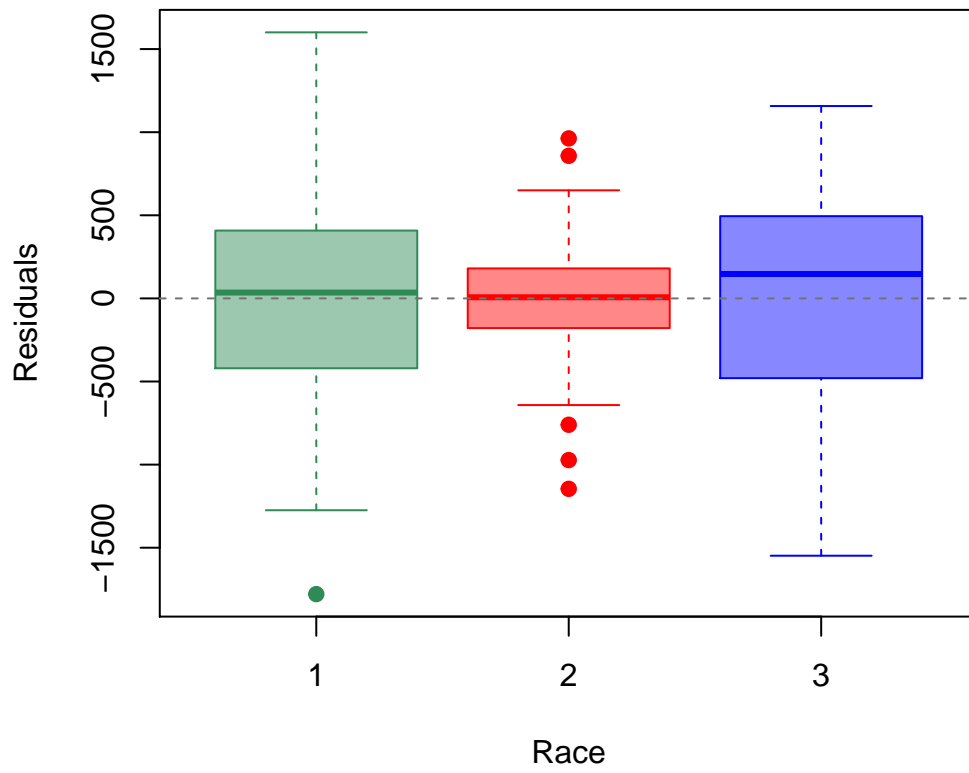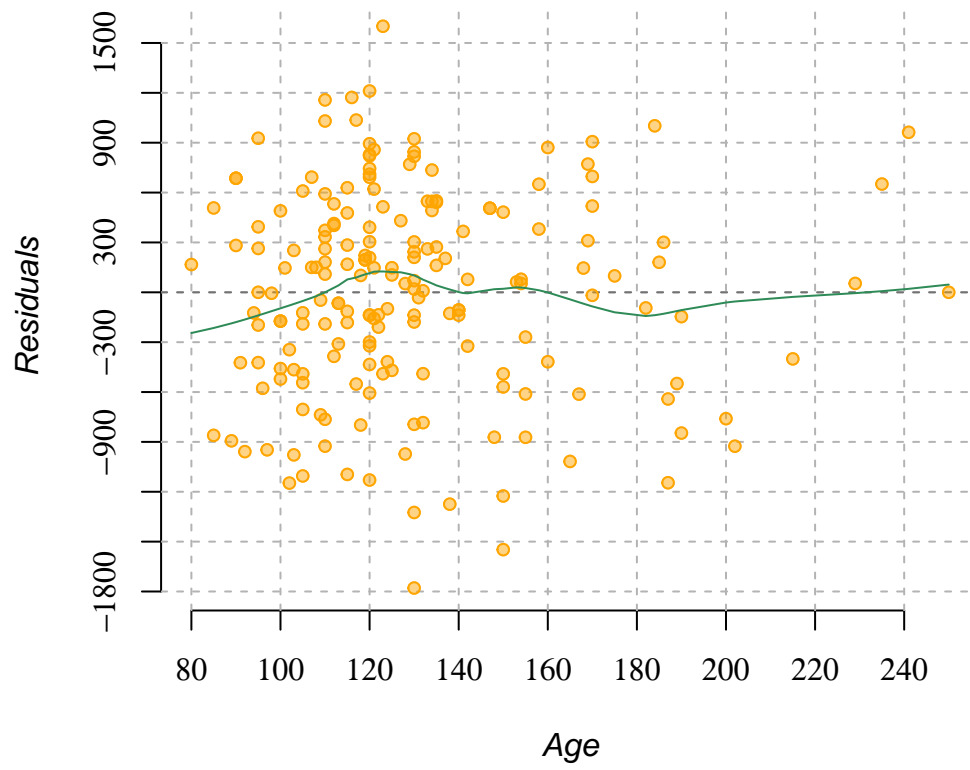
Table 3: Summary of Birth Weight Regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2841.040247	328.894986	8.6381379	0.0000000
age	-2.955912	9.604270	-0.3077706	0.7586271
lwt	4.380344	1.834985	2.3871276	0.0180591
factor(race)2	-442.329304	148.926061	-2.9701269	0.0033999
factor(race)3	-282.141331	120.748135	-2.3366103	0.0206050
smoke	-283.589210	111.748150	-2.5377531	0.0120397
factor(ptl)1	-357.092406	152.324861	-2.3442818	0.0201992
factor(ptl)2	-68.778992	297.501669	-0.2311886	0.8174415
factor(ptl)3	1260.027892	661.664516	1.9043305	0.0585268
ht	-553.036031	202.192322	-2.7351980	0.0068840
ui	-522.713818	138.164206	-3.7832796	0.0002130
factor(ftv)1	142.479511	122.824933	1.1600211	0.2476386
factor(ftv)2	-1.868773	138.659587	-0.0134774	0.9892624
factor(ftv)3	-319.779655	253.984296	-1.2590529	0.2097074
factor(ftv)4	244.941583	336.965702	0.7269036	0.4682675
factor(ftv)6	15.369792	700.858349	0.0219300	0.9825291

(c) [9 pts.]

Overall, the residuals look quite good plotted against the fitted values and the numerical predictors. Case 226 (a 45 year old) bears substantial leverage on the fit since all other cases are much younger. Furthermore, this data point also yields a very large outlier. In practice, we would therefore discard this observation when building a predictive model. In regard to the categorical predictors, there is some interclass heteroskedasticity; however, the varied sample sizes can also make is challenging to deduce this strictly from these box plots. That is, categorical levels containing a large number of observations can give the appearance of a wider distribution while levels with small sample sizes natural appear to have low variance. To normalize these comparisons with respect to sample size we could, for example, perform an F-test for equality of variances or (more sophisticated) a Kolmorogov-Smirnov test for equality of distributions.

lm(bwt ~ age + lwt + factor(race) + smoke + factor(ptl) + ht + ui + factor(

Number of physician visits during the first trimester.
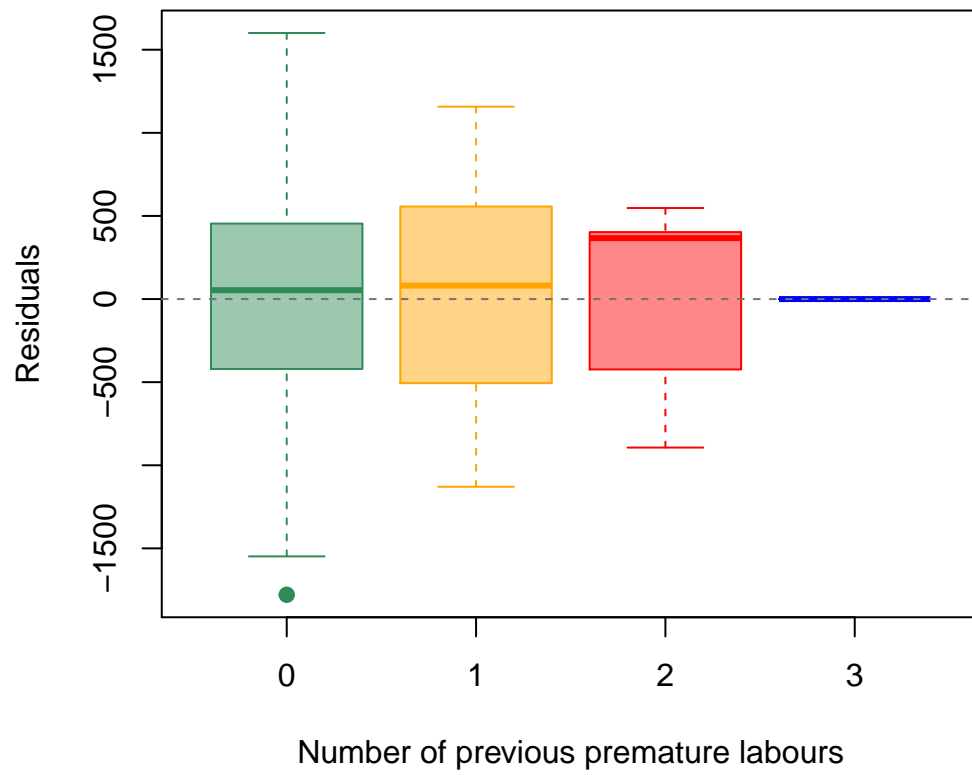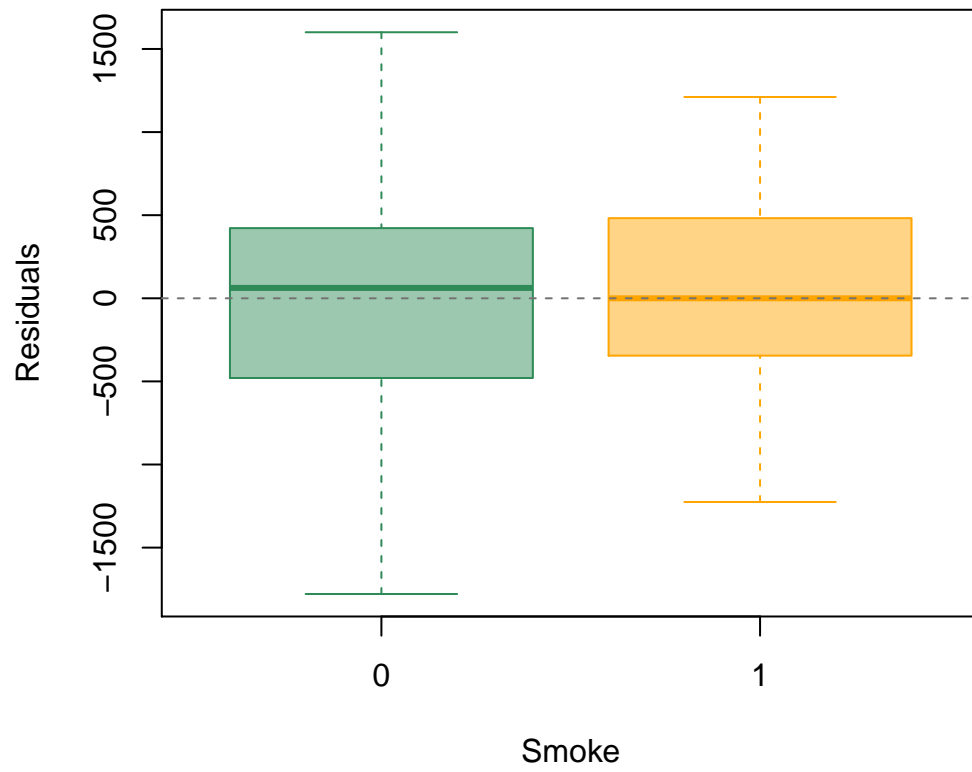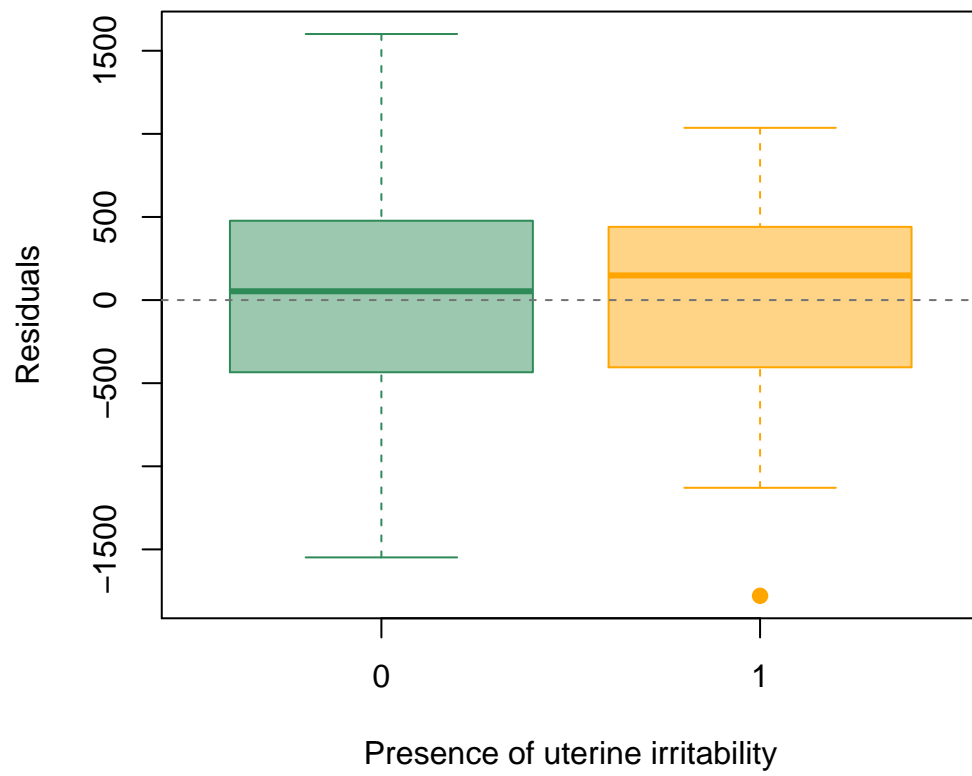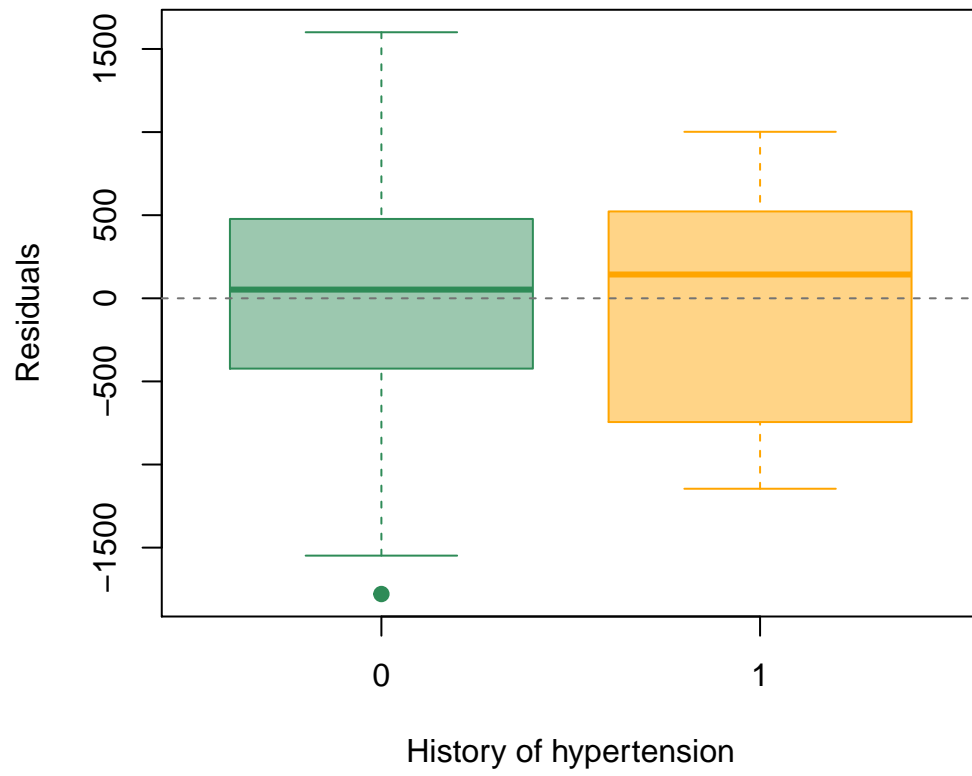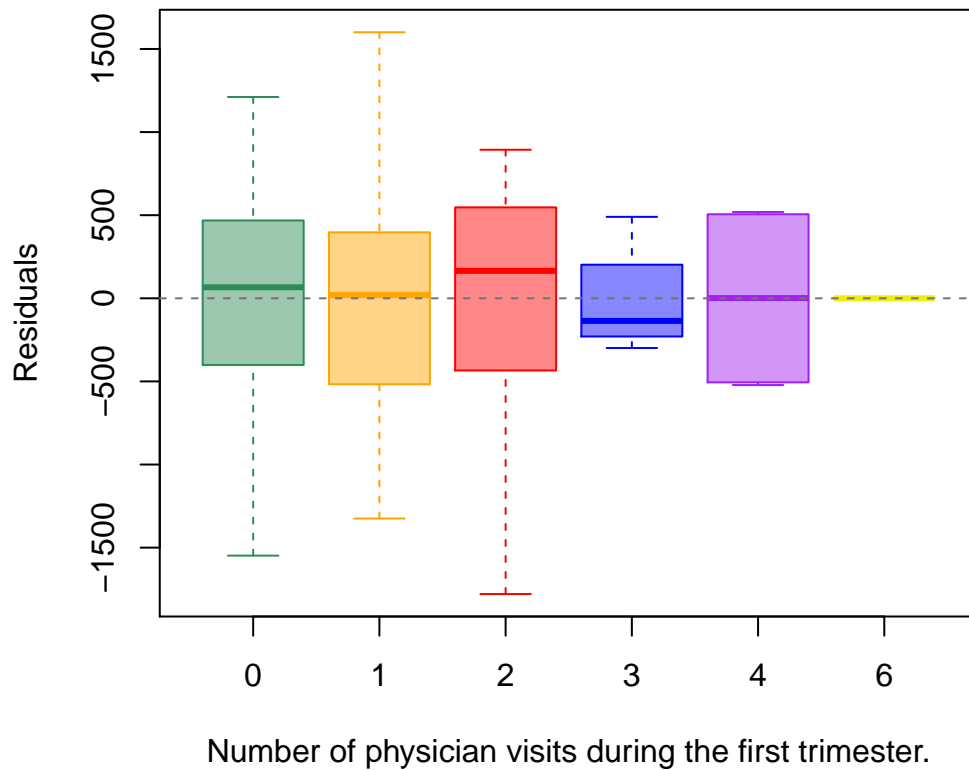
(d) [9 pts.]

Table 3 (part b) suggests we may have included more predictors than are useful for predicting `Birth Weight`. A more predictive model would likely reduce the dimension of our covariates. We could, for example, proceed by performing a sequence of partial F-tests (equivalently stepwise regression; 36-402) or perhaps implementing a LASSO regression (also 36-402)[2] Given the full set of predictors, the multiple linear regression suggests that a baby's birth weight is significantly associated with the mother's weight, race, smoking status, history of hypertension, and presence of uterine iritability (and possibly the number of previous premature labours).

---

[2]Gee, I bet you guys can't wait until 36-402.

Appendix

```r
addTrans <- function(color,trans)
{
  # This function adds transparancy to a color.
  # Define transparancy with an integer between 0 and 255
  # 0 being fully transparant and 255 being fully visable
  # Works with either color and trans a vector of equal length,
  # or one of the two of length 1.

  if (length(color)!=length(trans)&!any(c(length(color),length(trans))==1)){
    stop("Vector lengths not correct")
  }
  if (length(color)==1 & length(trans)>1) color <- rep(color,length(trans))
  if (length(trans)==1 & length(color)>1) trans <- rep(trans,length(color))

  num2hex <- function(x)
  {
    hex <- unlist(strsplit("0123456789ABCDEF",split=""))
    return(paste(hex[(x-x%%16)/16+1],hex[x%%16+1],sep=""))
  }
  rgb <- rbind(col2rgb(color),trans)
  res <- paste("#",apply(apply(rgb,2,num2hex),2,paste,collapse=""),sep="")
  return(res)
}
```

Problem 1 [32 points]

```r
data(ChickWeight)
names(ChickWeight)
attach(ChickWeight)
```

(a) (4 pts.)

```r
coplot(weight ~ Time | Chick, data = ChickWeight,type = "b", show.given = FALSE,
       xlab = "", ylab = "", main = "")
mtext(side = 1, text = "Time", cex = 2, line = 3.75)
mtext(side = 2, text = "Weight", cex = 2, line = 2.25)
```

(b) (9 pts.)

```r
Chick6 <- ChickWeight[ChickWeight$Chick == 6,]

par(mfrow=c(1,2))
with(Chick6, plot(Time, weight, col = NA, pch = 19, cex = 1.25,
                  axes = FALSE, xlab="",ylab=""))
axis(side = 1, at = c(0,5,10,15,20), as.character(c(0,5,10,15,20)),
     font = 5)
```

```
axis(side = 2, at = seq(0,300,25), as.character(seq(0,300,25)),
     font = 5)
abline(v = c(0,5,10,15,20), h = seq(50,300,25), col = "gray70",
       lty = 2)
mtext(side = 2, text = "Weight", font = 3, line = 3)
mtext(side = 1, text = "Time", font = 3, line = 3)
with(Chick6, points(Time, weight, col = addTrans("orange",120),
                    cex = 1.25, pch = 19))
with(Chick6, points(Time, weight, col = "orange", pch = 1,
                    cex = 1.25))
model1 <- lm(weight ~ Time, data = Chick6)
abline(model1, col = "seagreen", lwd = 2)

plot(Chick6$Time, residuals(model1), col = NA, axes = FALSE,
     xlab= "Time", ylab = "Residuals", font.lab = 3)
axis(side = 1, at = c(0,5,10,15,20), as.character(c(0,5,10,15,20)),
     font = 5)
axis(side = 2, at = seq(-70,30,10), labels = as.character(seq(-70,30,10)),
     font = 5)
abline(h = seq(-70,30,10), v = c(0,5,10,15,20), col = "gray70",
       lty = 2)
abline(0,0, lty = 2, col = "gray45")
points(Chick6$Time, residuals(model1), col = addTrans("orange",120),
       pch = 19, cex = 1.25)
points(Chick6$Time, residuals(model1), col = "orange", cex = 1.25)
panel.smooth(Chick6$Time, residuals(model1), col = NA,cex = 0.5,
             col.smooth = "seagreen", span = 0.5, iter = 3)
```

```
par(mfrow=c(1,2))
with(Chick6, plot(Time, weight, col = NA, pch = 19, cex = 1.25,
                  axes = FALSE, xlab="",ylab=""))
axis(side = 1, at = c(0,5,10,15,20), as.character(c(0,5,10,15,20)),
     font = 5)
axis(side = 2, at = seq(0,300,25), as.character(seq(0,300,25)),
     font = 5)
abline(v = c(0,5,10,15,20), h = seq(50,300,25), col = "gray70",
       lty = 2)
mtext(side = 2, text = "Weight", font = 3, line = 3)
mtext(side = 1, text = "Time", font = 3, line = 3)
with(Chick6, points(Time, weight, col = addTrans("orange",120),
                    cex = 1.25, pch = 19))
with(Chick6, points(Time, weight, col = "orange", pch = 1, cex = 1.25))
model2 <- lm(weight ~ Time + I(Time ^ 2) + I(Time ^ 3) + I(Time ^ 4),
             data = Chick6)
lines(Chick6$Time, fitted(model2), col = "seagreen", lwd = 2)

plot(Chick6$Time, residuals(model2), col = NA, axes = FALSE, xlab= "Time",
     ylab = "Residuals", font.lab = 3)
axis(side = 1, at = c(0,5,10,15,20), as.character(c(0,5,10,15,20)),
     font = 5)
axis(side = 2, at = seq(-70,30,10), labels = as.character(seq(-70,30,10)),
     font = 5)
abline(h = seq(-70,30,10), v = c(0,5,10,15,20), col = "gray70", lty = 2)
abline(0,0, lty = 2, col = "gray45")
```

```
points(Chick6$Time, residuals(model2), col = addTrans("orange",120),
       pch = 19, cex = 1.25)
points(Chick6$Time, residuals(model2), col = "orange", cex = 1.25)
panel.smooth(Chick6$Time, residuals(model2), col = NA,cex = 0.5,
             col.smooth = "seagreen", span = 0.5, iter = 3)
```

```
library(knitr)
kable(summary(model2)$coefficients, caption = "Summary of Polynomial Regression for Chick 6")
```

(c) (9 pts.)

```
par(mfrow=c(1,2))
with(ChickWeight, plot(Time, weight, col = NA, pch = 19, cex = 0.8,
                       axes = FALSE, xlab="",ylab=""))
axis(side = 1, at = c(0,5,10,15,20,25), as.character(c(0,5,10,15,20,25)),
     font = 5)
axis(side = 2, at = seq(0,350,50), as.character(seq(0,350,50)),
     font = 5)
abline(v = c(0,5,10,15,20,25), h = seq(50,350,50), col = "gray70",
       lty = 2)
mtext(side = 2, text = "Weight", font = 3, line = 3)
mtext(side = 1, text = "Time", font = 3, line = 3)
with(ChickWeight, points(Time, weight, col = addTrans("orange",120),
                         cex = 0.8, pch = 19))
with(ChickWeight, points(Time, weight, col = "orange", pch = 1,
                         cex = 0.8))
model3 <- lm(weight ~ Time, data = ChickWeight)
abline(model3, col = "seagreen", lwd = 2)

plot(ChickWeight$Time, residuals(model3), col = NA, axes = FALSE,
     xlab= "Time", ylab = "Residuals", font.lab = 3)
axis(side = 1, at = c(0,5,10,15,20,25), as.character(c(0,5,10,15,20,25)),
     font = 5)
axis(side = 2, at = seq(-200,200,40), labels = as.character(seq(-200,200,40)),
     font = 5)
abline(h = seq(-200,200,40), v = c(0,5,10,15,20,25), col = "gray70",
       lty = 2)
abline(0,0, lty = 2, col = "gray45")
points(ChickWeight$Time, residuals(model3), col = addTrans("orange",120),
       pch = 19, cex = 0.8)
points(ChickWeight$Time, residuals(model3), col = "orange", cex = 0.8)
panel.smooth(ChickWeight$Time, residuals(model3), col = NA,cex = 0.5,
             col.smooth = "seagreen", span = 0.5, iter = 3)
```

```
par(mfrow=c(1,2))
with(ChickWeight, plot(Time, weight, col = NA, pch = 19, cex = 0.8,
                       axes = FALSE, xlab="",ylab=""))
axis(side = 1, at = c(0,5,10,15,20,25), as.character(c(0,5,10,15,20,25)),
     font = 5)
axis(side = 2, at = seq(0,350,50), as.character(seq(0,350,50)), font = 5)
abline(v = c(0,5,10,15,20,25), h = seq(50,350,50), col = "gray70", lty = 2)
mtext(side = 2, text = "Weight", font = 3, line = 3)
```

```
mtext(side = 1, text = "Time", font = 3, line = 3)
with(ChickWeight, points(Time, weight, col = addTrans("orange",120), cex = 0.8, pch = 19))
with(ChickWeight, points(Time, weight, col = "orange", pch = 1, cex = 0.8))
model4 <- lm(weight ~ Time + I(Time ^ 2), data = ChickWeight)
lines(seq(0,max(ChickWeight$Time),length = 50),
      predict(model4, newdata=data.frame(Time = seq(0,max(ChickWeight$Time),length = 50))),
      col = "seagreen", lwd = 2)

plot(ChickWeight$Time, residuals(model4), col = NA, axes = FALSE, xlab= "Time",
     ylab = "Residuals", font.lab = 3)
axis(side = 1, at = c(0,5,10,15,20,25), as.character(c(0,5,10,15,20,25)), font = 5)
axis(side = 2, at = seq(-200,200,40), labels = as.character(seq(-200,200,40)),
     font = 5)
abline(h = seq(-200,200,40), v = c(0,5,10,15,20,25), col = "gray70", lty = 2)
abline(0,0, lty = 2, col = "gray45")
points(ChickWeight$Time, residuals(model4), col = addTrans("orange",120),
       pch = 19, cex = 0.8)
points(ChickWeight$Time, residuals(model4), col = "orange", cex = 0.8)
panel.smooth(ChickWeight$Time, residuals(model4), col = NA,cex = 0.5,
             col.smooth = "seagreen", span = 0.5, iter = 3)
```

(d) (9 pts.)

```
with(ChickWeight, plot(Time, weight, col = NA, pch = 19, cex = 0.8,
                       axes = FALSE, xlab="",ylab=""))
axis(side = 1, at = c(0,5,10,15,20,25), as.character(c(0,5,10,15,20,25)), font = 5)
axis(side = 2, at = seq(0,350,50), as.character(seq(0,350,50)), font = 5)
abline(v = c(0,5,10,15,20,25), h = seq(50,350,50), col = "gray70", lty = 2)
mtext(side = 2, text = "Weight", font = 3, line = 3)
mtext(side = 1, text = "Time", font = 3, line = 3)
with(ChickWeight, points(Time, weight, col = addTrans("orange",120),
                         cex = 0.8, pch = 19))
with(ChickWeight, points(Time, weight, col = "orange", pch = 1, cex = 0.8))
model5 <- lm(weight ~ Time + I(Time ^ 2) + factor(Diet), data = ChickWeight)
lines(seq(0,max(ChickWeight$Time),length = 50),
      predict(model5, newdata=data.frame(Time = seq(0,max(ChickWeight$Time),length = 50),
                                         Diet = 1)),
      col = "seagreen", lwd = 2)
lines(seq(0,max(ChickWeight$Time),length = 50),
      predict(model5, newdata=data.frame(Time = seq(0,max(ChickWeight$Time),length = 50),
                                         Diet = 2)),
      col = "red", lwd = 2)
lines(seq(0,max(ChickWeight$Time),length = 50),
      predict(model5, newdata=data.frame(Time = seq(0,max(ChickWeight$Time),length = 50),
                                         Diet = 3)),
      col = "blue", lwd = 2)
lines(seq(0,max(ChickWeight$Time),length = 50),
      predict(model5, newdata=data.frame(Time = seq(0,max(ChickWeight$Time),length = 50),
                                         Diet = 4)),
      col = "purple", lwd = 2)
legend(x = "topleft", legend = c("Diet 1", "Diet 2", "Diet 3", "Diet 4"),
       col = c("seagreen","red","blue","purple"), lwd = 2)
```

```r
with(ChickWeight, plot(Time, weight, col = NA, pch = 19, cex = 0.8, axes = FALSE,
                       xlab="",ylab=""))
axis(side = 1, at = c(0,5,10,15,20,25), as.character(c(0,5,10,15,20,25)), font = 5)
axis(side = 2, at = seq(0,350,50), as.character(seq(0,350,50)), font = 5)
abline(v = c(0,5,10,15,20,25), h = seq(50,350,50), col = "gray70", lty = 2)
mtext(side = 2, text = "Weight", font = 3, line = 3)
mtext(side = 1, text = "Time", font = 3, line = 3)
with(ChickWeight, points(Time, weight, col = addTrans("orange",120), cex = 0.8,
                         pch = 19))
with(ChickWeight, points(Time, weight, col = "orange", pch = 1, cex = 0.8))
model6 <- lm(weight ~ Time + I(Time ^ 2) : factor(Diet), data = ChickWeight)
lines(seq(0,max(ChickWeight$Time),length = 50),
      predict(model6, newdata=data.frame(Time = seq(0,max(ChickWeight$Time),length = 50),
                                          Diet = 1)),
      col = "seagreen", lwd = 2)
lines(seq(0,max(ChickWeight$Time),length = 50),
      predict(model6, newdata=data.frame(Time = seq(0,max(ChickWeight$Time),length = 50),
                                          Diet = 2)),
      col = "red", lwd = 2)
lines(seq(0,max(ChickWeight$Time),length = 50),
      predict(model6, newdata=data.frame(Time = seq(0,max(ChickWeight$Time),length = 50),
                                          Diet = 3)),
      col = "blue", lwd = 2)
lines(seq(0,max(ChickWeight$Time),length = 50),
      predict(model6, newdata=data.frame(Time = seq(0,max(ChickWeight$Time),length = 50),
                                          Diet = 4)),
      col = "purple", lwd = 2)
legend(x = "topleft", legend = c("Diet 1", "Diet 2", "Diet 3", "Diet 4"),
       col = c("seagreen","red","blue","purple"), lwd = 2, bg = "white")
```

```r
par(mfrow=c(2,2))
cols <- c("seagreen","red","blue","purple")
for (itr in 1:4){
  with(ChickWeight, plot(Time, weight, col = NA, pch = 19, cex = 0.8,
                         axes = FALSE, xlab="",ylab=""))
  axis(side = 1, at = c(0,5,10,15,20,25), as.character(c(0,5,10,15,20,25)),
       font = 5)
  axis(side = 2, at = seq(0,350,50), as.character(seq(0,350,50)),
       font = 5)
  abline(v = c(0,5,10,15,20,25), h = seq(50,350,50), col = "gray70",
         lty = 2)
  mtext(side = 2, text = "Weight", font = 3, line = 3)
  mtext(side = 1, text = "Time", font = 3, line = 3)
  mtext(side = 3, text = paste0("Diet = ",itr), font = 3)
  with(ChickWeight[which(ChickWeight$Diet==itr),], points(Time, weight,
                                                col = addTrans("orange",120),
                                                cex = 0.8, pch = 19))
  with(ChickWeight[which(ChickWeight$Diet==itr),], points(Time, weight,
                                                col = "orange",
                                                pch = 1, cex = 0.8))
  model6 <- lm(weight ~ Time + I(Time ^ 2) : factor(Diet),
               data = ChickWeight)
  lines(seq(0,max(ChickWeight$Time),length = 50),
        predict(model6, newdata=data.frame(Time = seq(0,max(ChickWeight$Time),
```

```
                                                    length = 50), Diet = itr)),
        col = cols[itr], lwd = 2)
}
```

```
plot(ChickWeight$Time, residuals(model6), col = NA, axes = FALSE,
     xlab= "Time", ylab = "Residuals", font.lab = 3)
axis(side = 1, at = c(0,5,10,15,20,25), as.character(c(0,5,10,15,20,25)),
     font = 5)
axis(side = 2, at = seq(-200,200,40), labels = as.character(seq(-200,200,40)),
     font = 5)
abline(h = seq(-200,200,40), v = c(0,5,10,15,20,25), col = "gray70", lty = 2)
abline(0,0, lty = 2, col = "gray45")
points(ChickWeight$Time, residuals(model6),
       col = addTrans(c("seagreen","red","blue","purple")[ChickWeight$Diet],120),
       pch = 19, cex = 0.8)
points(ChickWeight$Time, residuals(model6),
       col = c("seagreen","red","blue","purple")[ChickWeight$Diet], cex = 0.8)
legend(x = "topleft", legend = c("Diet 1", "Diet 2", "Diet 3", "Diet 4"),
       col = addTrans(c("seagreen","red","blue","purple"),120), pch = 19, bg = "white")
```

```
par(mfrow=c(2,2))
cols <- c("seagreen","red","blue","purple")
for (itr in 1:4){
  with(ChickWeight, plot(Time, residuals(model6), col = NA,
                         pch = 19, cex = 0.8, axes = FALSE, xlab="",ylab=""))
  axis(side = 1, at = c(0,5,10,15,20,25), as.character(c(0,5,10,15,20,25)),
       font = 5)
  axis(side = 2, at = seq(-200,200,40), labels = as.character(seq(-200,200,40)),
       font = 5)
  abline(h = seq(-200,200,40), v = c(0,5,10,15,20,25), col = "gray70", lty = 2)
  abline(0,0, lty = 2, col = "gray45")
  mtext(side = 2, text = "Residuals", font = 3, line = 3)
  mtext(side = 1, text = "Time", font = 3, line = 3)
  mtext(side = 3, text = paste0("Diet = ",itr), font = 3)
  with(ChickWeight, points(Time[which(ChickWeight$Diet==itr)],
                           residuals(model6)[which(ChickWeight$Diet==itr)],
                           col = addTrans(c("seagreen","red","blue","purple"),120)[itr],
                           cex = 0.8, pch = 19))
  with(ChickWeight, points(Time[which(ChickWeight$Diet==itr)],
                           residuals(model6)[which(ChickWeight$Diet==itr)],
                           col = c("seagreen","red","blue","purple")[itr], pch = 1,
                           cex = 0.8))
}
```

```
boxplot(residuals(model6) ~ ChickWeight$Diet,
        col = addTrans(c("seagreen","red","blue","purple"),120),
        border = c("seagreen","red","blue","purple"),
        xlab = "Diet",ylab = "Residuals", pch = 19)
abline(0,0, lty = 2, col = "gray45")
```

Problem 3 [32 points]

```
library(MASS)
```

(a) [5 pts.]

```
pairs(birthwt[,2:10])
```

(b) [9 pts.]

```
model8 <- lm(bwt ~ age + lwt + factor(race) + smoke + factor(ptl) + ht
             + ui + factor(ftv), data = birthwt)
kable(summary(model8)$coefficients,
      caption = "Summary of Birth Weight Regression")
```

(c) [9 pts.]

```
plot(model8, which = 1, col = NA, pch = 19, axes = FALSE,
     add.smooth = FALSE, caption = "")
abline(h = seq(-2100,1500,300), col = "gray75", lty = 2)
abline(v = seq(1800,3800,200), col = "gray80", lty = 2)
abline(0,0, lty = 2, col = "gray45")
axis(side = 1, at = seq(1800,3800,200),
     as.character(seq(1800,3800,200)), font = 5)
axis(side = 2, at = seq(-2100,1500,300),
     labels = as.character(seq(-2100,1500,300)), font = 5)
points(fitted(model8), residuals(model8),
       col = addTrans("orange",120), pch = 19)
points(fitted(model8), residuals(model8), col = "orange")
panel.smooth(fitted(model8), residuals(model8), col = "orange",
             cex = 1, col.smooth = "seagreen", span = 2/3, iter = 3)
```

```
plot(birthwt$age, residuals(model8), col = NA, axes = FALSE, xlab= "Age",
     ylab = "Residuals", font.lab = 3)
axis(side = 1, at = seq(10,50,5), as.character(seq(10,50,5)), font = 5)
axis(side = 2, at = seq(-1800,1500,300),
     labels = as.character(seq(-1800,1500,300)), font = 5)
abline(h = seq(-1800,1500,300), v = seq(10,50,5), col = "gray70", lty = 2)
abline(0,0, lty = 2, col = "gray45")
points(birthwt$age, residuals(model8),
       col = addTrans("orange",120), pch = 19, cex = 0.8)
points(birthwt$age, residuals(model8), col = "orange", cex = 0.8)
panel.smooth(birthwt$age, residuals(model8), col = NA,cex = 0.5,
             col.smooth = "seagreen", span = 2/3, iter = 3)
```

```
plot(birthwt$lwt, residuals(model8), col = NA, axes = FALSE,
     xlab= "Age", ylab = "Residuals", font.lab = 3)
axis(side = 1, at = seq(80,250,20), as.character(seq(80,250,20)), font = 5)
axis(side = 2, at = seq(-1800,1500,300),
```

```
      labels = as.character(seq(-1800,1500,300)), font = 5)
abline(h = seq(-1800,1500,300), v = seq(80,250,20), col = "gray70", lty = 2)
abline(0,0, lty = 2, col = "gray45")
points(birthwt$lwt, residuals(model8), col = addTrans("orange",120),
       pch = 19, cex = 0.8)
points(birthwt$lwt, residuals(model8), col = "orange", cex = 0.8)
panel.smooth(birthwt$lwt, residuals(model8), col = NA,cex = 0.5,
             col.smooth = "seagreen", span = 2/3, iter = 3)
```

```
boxplot(residuals(model8) ~ birthwt$race, col = addTrans(c("seagreen","red","blue"),120),
        border = c("seagreen","red","blue"), xlab = "Race",ylab = "Residuals", pch = 19)
abline(0,0, lty = 2, col = "gray45")
```

```
boxplot(residuals(model8) ~ birthwt$smoke, col = addTrans(c("seagreen","orange"),120),
        border = c("seagreen","orange"), xlab = "Smoke",ylab = "Residuals", pch = 19)
abline(0,0, lty = 2, col = "gray45")
```

```
boxplot(residuals(model8) ~ birthwt$ptl,
        col = addTrans(c("seagreen","orange","red","blue"),120),
        border = c("seagreen","orange","red","blue"),
        xlab = "Number of previous premature labours",
        ylab = "Residuals", pch = 19)
abline(0,0, lty = 2, col = "gray45")
```

```
boxplot(residuals(model8) ~ birthwt$ht, col = addTrans(c("seagreen","orange"),120),
        border = c("seagreen","orange"), xlab = "History of hypertension",
        ylab = "Residuals", pch = 19)
abline(0,0, lty = 2, col = "gray45")
```

```
boxplot(residuals(model8) ~ birthwt$ui, col = addTrans(c("seagreen","orange"),120),
        border = c("seagreen","orange"), xlab = "Presence of uterine irritability",
        ylab = "Residuals", pch = 19)
abline(0,0, lty = 2, col = "gray45")
```

```
boxplot(residuals(model8) ~ birthwt$ftv,
        col = addTrans(c("seagreen","orange","red","blue","purple","yellow2","pink"),120),
        border = c("seagreen","orange","red","blue","purple","yellow2","pink"),
        xlab = "Number of physician visits during the first trimester.",
        ylab = "Residuals", pch = 19)
abline(0,0, lty = 2, col = "gray45")
```