Biostat 653 Hw1

David (Daiwei) Zhang

September 27, 2017

Problem 6

(1)

Our model is

$$Y \sim \beta_0 + \beta_1 I(\text{girl}) + \beta_2 I(\text{cur. expo.}) + \beta_3 I(\text{past. expo.}) + \beta_4 I(\text{cur. expo.}) I(\text{girl}) + \beta_5 I(\text{past. expo.}) I(\text{girl})$$

Then the three questions are hypothesis tests with

1.

$$H_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$
$$H_1 : \text{At least one of these is nonzero.}$$

2.

$$H_0 : \beta_4 = \beta_5 = 0$$
$$H_1 : \text{At least one of these is nonzero.}$$

3.

$$H_0 : \beta_2 < \beta_3 \quad \text{and} \quad \beta_2 + \beta_4 < \beta_3 + \beta_5$$
$$H_1 : \text{At least one of these is false.}$$

We fit the linear model:

```
leadiq <- read.table("~/leadiq.txt")
colnames(leadiq) <- c("id", "expo.cat", "gender", "age", "iq")
leadiq$expo.cat <- factor(leadiq$expo.cat)
leadiq.model <- lm(iq ~ gender + expo.cat + expo.cat * gender, data=leadiq)
summary(leadiq.model)
```

```
##
## Call:
## lm(formula = iq ~ gender + expo.cat + expo.cat * gender, data = leadiq)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.696  -9.385  -0.412   9.621  45.304
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        103.696      2.327  44.567   <2e-16 ***
## gender              -2.414      3.633  -0.665   0.5076
## expo.cat2           -6.284      4.479  -1.403   0.1633
## expo.cat3          -10.234      4.957  -2.065   0.0411 *
## gender:expo.cat2    -3.569      7.964  -0.448   0.6549
```

```
## gender:expo.cat3    4.064      7.747   0.525   0.6009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.78 on 118 degrees of freedom
## Multiple R-squared:  0.06598,    Adjusted R-squared:  0.02641
## F-statistic: 1.667 on 5 and 118 DF,  p-value: 0.1478
```

(2)

We do not need to include age as a covariate, because the investigator does not ask for adjustment for age. Moreover, if we include age in our model,

```
leadiq.model.withage <- lm(iq ~ age+ gender + expo.cat + expo.cat * gender, data=leadiq)
summary(leadiq.model.withage)
```

```
##
## Call:
## lm(formula = iq ~ age + gender + expo.cat + expo.cat * gender,
##     data = leadiq)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.207  -9.999   0.070  10.107  42.603
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      100.2531     4.3839  22.868   <2e-16 ***
## age                0.3860     0.4165   0.927   0.3560
## gender            -2.7997     3.6585  -0.765   0.4457
## expo.cat2         -5.6418     4.5351  -1.244   0.2160
## expo.cat3         -9.9365     4.9702  -1.999   0.0479 *
## gender:expo.cat2  -4.3856     8.0171  -0.547   0.5854
## gender:expo.cat3   4.2238     7.7540   0.545   0.5870
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.79 on 117 degrees of freedom
## Multiple R-squared:  0.07279,    Adjusted R-squared:  0.02524
## F-statistic: 1.531 on 6 and 117 DF,  p-value: 0.1741
```

We see that age is not a significant predictor at $\alpha = 0.05$.

(3)

β_0: The mean IQ of a boy with no lead exposure is 103.696 ($p < 2 \times 10^{-16}$).

β_1: The mean IQ of a girl is 2.414 ($p = 0.508$) lower than that of a boy, given that they have no lead exposure.

β_2: The mean IQ of a boy with current lead exposure is 6.284 ($p = 0.163$) lower than that with no lead exposure.

β_3: The mean IQ of a boy with past lead exposure is 10.234 ($p = 0.041$) lower than that with no lead exposure.

β_4: The mean difference in IQ from a girl with current lead exposure to that with no lead exposure is 3.57 ($p = 0.655$) lower than that of two boys in such situations.

β_5: The mean difference in IQ from a girl with past lead exposure to that with no lead exposure is 4.06 ($p = 0.601$) higher than that of two boys in such situations.

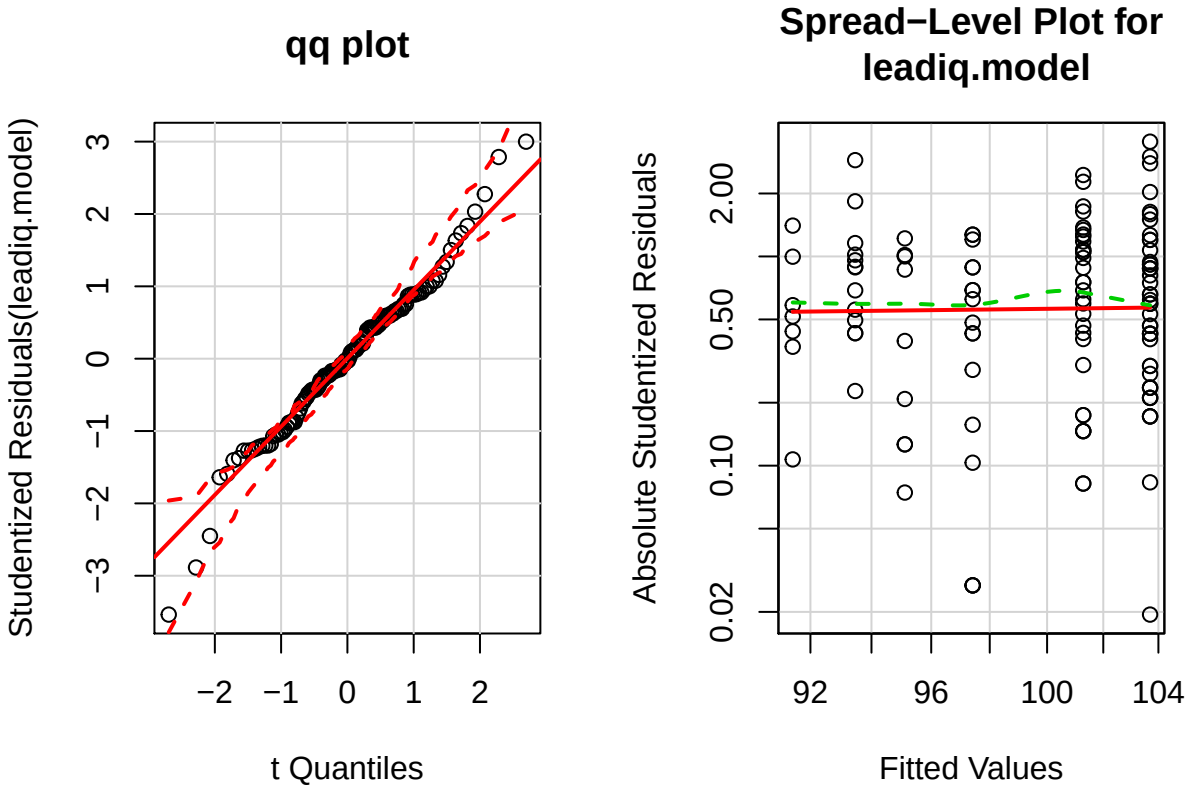
(4)

IQ of boy without lead exposure: $\beta_0 = 103.696$. IQ of boy with current lead exposure: $\beta_0 + \beta_2 = 97.412$. IQ of boy with past lead exposure: $\beta_0 + \beta_3 = 93.462$.

(5)

As seen from the qq plot below, the observed data is close to the straight line, so which supports the linearity assumption. However, in the absolute standardized residual plot, we see that the absolute residuals increase with the fitted values, so the constant variance assumption is slightly violated.

```
par(mfrow=c(1,2))
qqPlot(leadiq.model, main = "qq plot")
spreadLevelPlot(leadiq.model)
```




```
##
## Suggested power transformation:  0.6396838
```

Problem 7

(1)

```
lead <- read.table("~/lead.txt")
colnames(lead) <- c("id", "week.0", "week.1", "week.4", "week.6")
lead.time <- c(0,1,4,6)
lead.mat <- as.matrix(lead[,-1])
lead.mean <- colMeans(lead.mat)
lead.sd <- colSds(lead.mat)
lead.var <- lead.sd^2
print(lead.mean)
```

```
## week.0 week.1 week.4 week.6
## 26.540 13.522 15.514 20.762
```

```
print(lead.sd)
```

```
## [1] 5.020936 7.672487 7.852207 9.246332
```

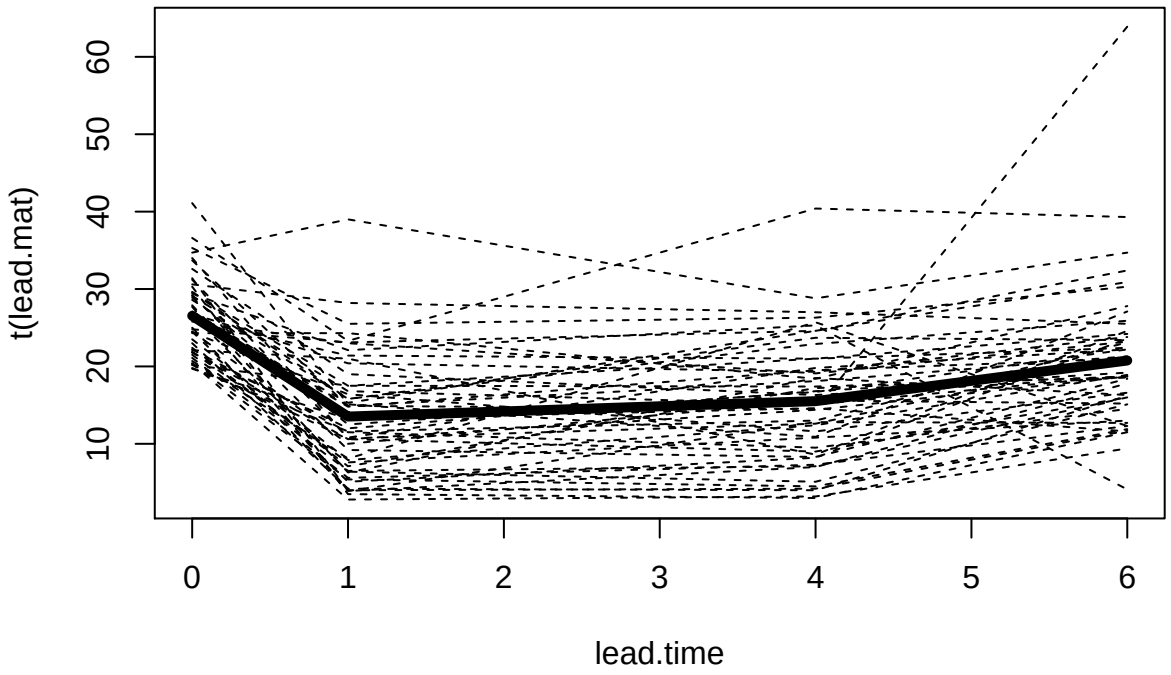
```
print(lead.var)
```

```
## [1] 25.20980 58.86706 61.65715 85.49465
```

(2)

Overall, the lead level decrease sharply from the baseline to the first week, and then increases slowly till the sixth week.

```
matplot(lead.time, t(lead.mat), type="l", lty=2, lwd=1, col=1)
points(lead.time, lead.mean, type="l", lty=1, lwd=5, col=1)
```



(3)

The two calculations of the variance are identical.

```
lead.cov <- cov(lead.mat)
lead.cor <- cor(lead.mat)
print(lead.cov)
```

```
##          week.0   week.1   week.4   week.6
## week.0 25.20980 15.46543 15.13800 22.98543
## week.1 15.46543 58.86706 44.02907 35.96596
## week.4 15.13800 44.02907 61.65715 33.02197
## week.6 22.98543 35.96596 33.02197 85.49465
```

```
print(lead.cor)
```

```
##           week.0    week.1    week.4    week.6
## week.0 1.0000000 0.4014589 0.3839654 0.4951063
## week.1 0.4014589 1.0000000 0.7308221 0.5069743
## week.4 0.3839654 0.7308221 1.0000000 0.4548224
## week.6 0.4951063 0.5069743 0.4548224 1.0000000
```