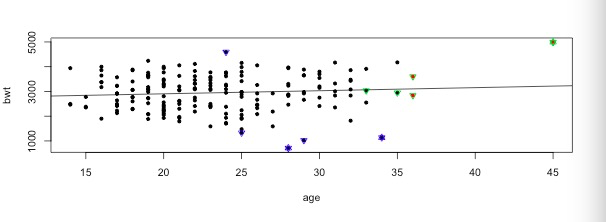
Q3.

(a) (b)



Call:

lm(formula = birthwt$bwt ~ birthwt$age)

Residuals:

Min 1Q Median 3Q Max

-2294.78 -517.63 10.51 530.80 1774.92

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2655.74 238.86 11.12 <2e-16 \*\*\*

birthwt$age 12.43 10.02 1.24 0.216

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 728.2 on 187 degrees of freedom

Multiple R-squared: 0.008157, Adjusted R-squared: 0.002853

F-statistic: 1.538 on 1 and 187 DF, p-value: 0.2165

(d) What distinguishes high (>0) form low (<0) DEFITSt values?

The fitted regression line distinguishes the DEFITSt values, measuring the influence of case i on response value y. The points above the fitted regression line is considered to be high (>0), and the points below the line is considered to be low (<0)

(e) Is there any flag criterion which dominates, that is, one which is flagged when any of the others is (if the answer is yes, this might, or might not, mean that we only need that criterion)?

Base on plot, there is always covariance ratio is always flagged, whenever the other influenced points are flagged.

(f) What distinguishes high (>1) from low (<1) covariance ratio values? Refer to the plot, and also to the original definition in Section 6.5

According to the plot, if the points’ covariance value is high, they are far away from the sample mean. If the estimated standard error is big, the covariance ratio values is high (>1), otherwise, it is low (<1).

(g) > X1

X1

X1 0.3181818 -0.4545455

-0.4545455 0.9090909

> X2

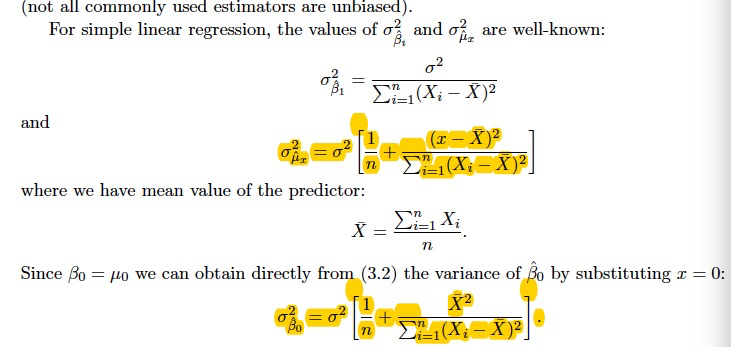
X2

X2 0.2 -0.2

-0.2 0.4

(h) Examine the form of the variances in Equations (3.1), (3.3). What effect does the sample variance of the predictor variable have on these quantities?

According to the formula in (3.1) and (3.3) below, when sample variance is bigger, σβ1^2 and σβ2^2 will be bigger.

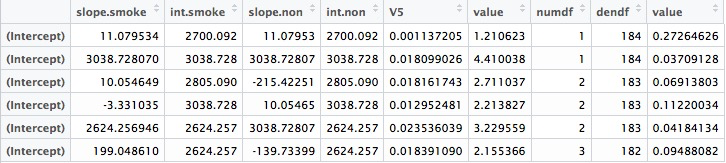


(i) Would it be a good idea to rely exclusively on the covariance ratio to flag anomalies?

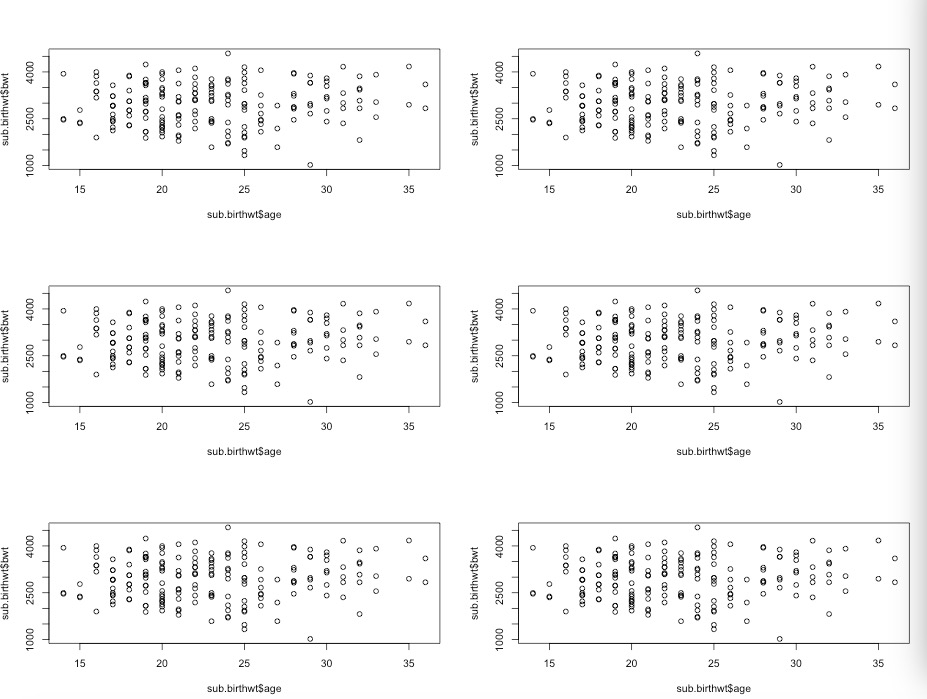
It would not be a good idea to rely exclusively on the covariance ratio to flag anomalies, because relying on the covariance ratio can only help us to detect non-linear situations, we also need to consider other anomalies.

Q4.

(b)



(c)



(d) Which model has the highest R^2(adj)? Write explicitly the estimated linear relationship between bwt and age separately for smokers and nonsmokers.

(e)

> anova(lm(model.list[[5]]), lm(model.list[[1]]))

Analysis of Variance Table

Model 1: sub.birthwt$bwt ~ sub.birthwt$smoke + sub.birthwt$age:I(1 - sub.birthwt$smoke)

Model 2: sub.birthwt$bwt ~ sub.birthwt$age

Res.Df RSS Df Sum of Sq F Pr(>F)

1 183 84509173

2 184 86920097 -1 -2410924 5.2207 0.02347 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>