AssessedCoursework1

Preliminary and reading data

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
library(car)

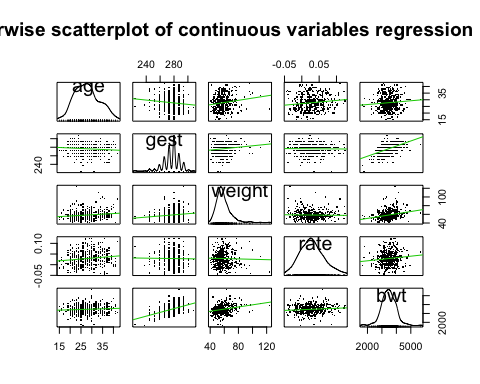
##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

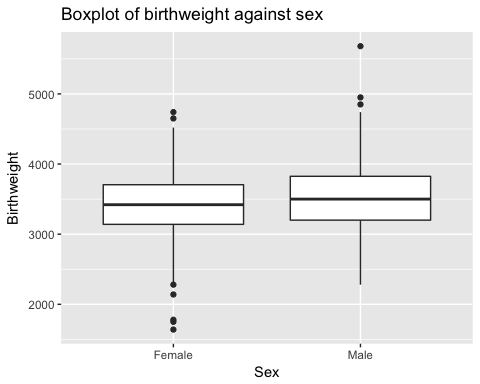
BirthData = read.table("BirthData.txt",header = TRUE)  
Births = filter( BirthData, !is.na(age), !is.na(gest), !is.na(sex), !is.na(smokes), !is.na(weight), !is.na(rate), !is.na(bwt))

Q1a)

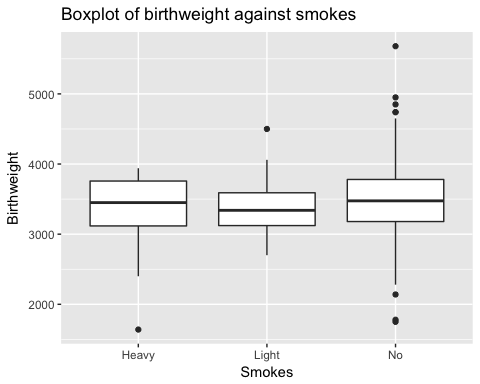
#Pairwise scatterplot of continuous variables regression line to observe correlation.  
Births1 <- Births[c(1, 2, 5, 6, 7)]   
scatterplotMatrix(Births1[,1:5], pch = ".",smoother=FALSE, main = "Pairwise scatterplot of continuous variables regression line" )



#Two boxplots of birthweight against sex and another at smokes.  
qplot( x = sex, y = bwt, data = Births, geom = 'boxplot', main = "Boxplot of birthweight against sex", xlab = "Sex" ,ylab = "Birthweight")



qplot(x = smokes, y = bwt, data = Births, geom = 'boxplot', main = "Boxplot of birthweight against smokes", xlab = "Smokes" ,ylab = "Birthweight")

 From pairwise scatterplot with the regression line, it is clear there is some positive correlation between birthweight and age, birthweight and gestation, and birthweight and rate of growth, with the strongest and most significant correlation being between birthweight and gestation, the second strongest between birthweight and weight, and the relationships of birthweight against the other two are mild. From the plot it is not evidently clear that there is a relationsip between birthweight and rate of growth, however, since the scales of the birthweight compared to the other variables is consideraby larger, a relationship may be present.

From observing the median from the boxplots, birthweight appears to be higher in males than females. In addition, birthweight apears to be highest when the mother did not smoke during the pregnancy, followed by 'heavy' smoking and then 'light' smoking. For mother's with 'light' smoing levels, the interquartile range is much smaller, as is the spread of the data compared to no smoking.

Q1b)

fit1 <- lm( bwt ~ sex, data = Births)  
fit1

##   
## Call:  
## lm(formula = bwt ~ sex, data = Births)  
##   
## Coefficients:  
## (Intercept) sexMale   
## 3411.6 103.2

#fit0 is the controlled linear model  
fit0 <- lm( bwt ~ 1, data = Births)  
model01 <- anova(fit0, fit1)  
model01

## Analysis of Variance Table  
##   
## Model 1: bwt ~ 1  
## Model 2: bwt ~ sex  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 426 101184262   
## 2 425 100054850 1 1129411 4.7974 0.02905 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

model <- data.frame(model.matrix(fit1))  
model[c(1:10), 1:2]

## X.Intercept. sexMale  
## 1 1 0  
## 2 1 0  
## 3 1 0  
## 4 1 1  
## 5 1 0  
## 6 1 0  
## 7 1 0  
## 8 1 1  
## 9 1 0  
## 10 1 0

The coding of the factor is denoted by 0 and 1 for females and males respectively. The estimated parameters for the model E[bwt] = a + b.x are a = 3411.6 units and b = 103.2 units. This means males have an increasing affect of 103.2 units onto the birthweight. The average birthweight is 3411.6 units which is the average birthweight for a female.The model of birthweight against sex is significant since the p-value is 0.02905 < 0.05, the p-value boundary for acceptance of a linear model. The design matrix is supposed to have dimentiosn (n x p) but here, we have a (3 x 4) matrix but the rank is 3 and so, since 4 =/= 3 the model is not apprpriate.

Q1c) The given form of the model is not appropriate since it has too many parameters. This causes a range of problems with regards to calculations. The design matrix Z (3 x 4) which corresponds to the given model is not of full rank due to this over parameterisation when we only have three linearly independent columns. This means (Z^T)Z is not invertible which is a condition that needs to be met and so the usual method does not work to get the estimated beta\_^.

Q1d)

fit3 <- lm( bwt ~ smokes, data = Births)  
summary(fit3)

##   
## Call:  
## lm(formula = bwt ~ smokes, data = Births)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1728.06 -298.06 0.26 317.53 2201.94   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3286.9 121.7 27.001 <2e-16 \*\*\*  
## smokesLight 112.9 165.2 0.683 0.495   
## smokesNo 191.2 124.2 1.539 0.124   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 486.9 on 424 degrees of freedom  
## Multiple R-squared: 0.006454, Adjusted R-squared: 0.001767   
## F-statistic: 1.377 on 2 and 424 DF, p-value: 0.2535

fit2 <- lm( bwt ~ 1, data = Births)  
model23 <- anova(fit2, fit3)  
model23

## Analysis of Variance Table  
##   
## Model 1: bwt ~ 1  
## Model 2: bwt ~ smokes  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 426 101184262   
## 2 424 100531269 2 652993 1.377 0.2535

The p-value is relatively high at 0.2535 when we use 0.05 as the typical boundary to accept whether a model is significant enough. This means that smokes does not have a strong association on birthweight. Moreover, for a mother who smokes at a 'heavy' level during her pregancy, the average birthweight of the child is 3286.9 units, a mother who smokes at a 'light' level has this average plus 112.9 units, and a mother who does not smoke during her pregancy has a child of the average weight stated plus 191.2 units.

Q1e)

fit5 <- lm( bwt ~ sex + gest, data = Births)  
summary(fit5)

##   
## Call:  
## lm(formula = bwt ~ sex + gest, data = Births)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -951.39 -278.26 -16.15 253.26 1936.89   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2474.153 487.479 -5.075 5.8e-07 \*\*\*  
## sexMale 108.804 40.678 2.675 0.00777 \*\*   
## gest 21.284 1.759 12.097 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 418.8 on 424 degrees of freedom  
## Multiple R-squared: 0.2649, Adjusted R-squared: 0.2614   
## F-statistic: 76.38 on 2 and 424 DF, p-value: < 2.2e-16

fit4 <- lm( bwt ~ gest, data = Births)  
anova45 <- anova(fit4, fit5)  
anova45

## Analysis of Variance Table  
##   
## Model 1: bwt ~ gest  
## Model 2: bwt ~ sex + gest  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 425 75639024   
## 2 424 74383946 1 1255078 7.1541 0.007768 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

There is significant evidence for strong association between sex and birthweight after controlling for gestation giving a p-value of 0.007768 << 0.05. The adjusted R^2 is also relatively low for fit5 at 0.2614 which is increases its significance as the variation of the fitted and observed values is low. After fixing gestation in the linear model males have an increasing affect of 108.804 grams which is clearly significant in the perspective of the model.

Q1f)

#Variance of errors  
summaryfit5 <- summary(fit5)  
summaryfit5

##   
## Call:  
## lm(formula = bwt ~ sex + gest, data = Births)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -951.39 -278.26 -16.15 253.26 1936.89   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2474.153 487.479 -5.075 5.8e-07 \*\*\*  
## sexMale 108.804 40.678 2.675 0.00777 \*\*   
## gest 21.284 1.759 12.097 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 418.8 on 424 degrees of freedom  
## Multiple R-squared: 0.2649, Adjusted R-squared: 0.2614   
## F-statistic: 76.38 on 2 and 424 DF, p-value: < 2.2e-16

summaryfit5$sigma^2

## [1] 175433.8

#Variance-covariance matrix  
vcov(fit5)

## (Intercept) sexMale gest  
## (Intercept) 237636.1223 -1120.2126511 -856.0957571  
## sexMale -1120.2127 1654.7385263 0.8141519  
## gest -856.0958 0.8141519 3.0957873

#Deviance  
deviance(fit5)

## [1] 74383946

The deviance of the model fit is 74383946.

Q1g)

newdata <- data.frame(sex = "Female" , age = 30 , gest = 275)  
#A 95% confidence interval  
predict(fit5, newdata, interval = "confidence", level = 0.95)

## fit lwr upr  
## 1 3378.896 3319.851 3437.941

#A 95% prediction interval  
predict(fit5, newdata, interval = "prediction", level = 0.95)

## fit lwr upr  
## 1 3378.896 2553.504 4204.288

This means at the 5% significance level, data in the interval (3319.851 , 3437.941), are significant. Also, we predict that 95% of the data lies in the interval (3378.896 , 4204.288).

Q1h)

lm1 <- lm( bwt ~ age + gest + sex + smokes + weight + rate, data = Births)  
summary1 <- summary(lm1)  
summary1

##   
## Call:  
## lm(formula = bwt ~ age + gest + sex + smokes + weight + rate,   
## data = Births)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1053.68 -276.97 -2.73 238.20 1834.00   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3159.053 499.707 -6.322 6.63e-10 \*\*\*  
## age 7.953 3.652 2.178 0.02998 \*   
## gest 20.609 1.736 11.870 < 2e-16 \*\*\*  
## sexMale 105.842 39.332 2.691 0.00741 \*\*   
## smokesLight 75.476 137.529 0.549 0.58344   
## smokesNo 134.967 103.480 1.304 0.19285   
## weight 8.172 1.979 4.130 4.38e-05 \*\*\*  
## rate 1647.892 603.635 2.730 0.00660 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 404.1 on 419 degrees of freedom  
## Multiple R-squared: 0.3236, Adjusted R-squared: 0.3123   
## F-statistic: 28.64 on 7 and 419 DF, p-value: < 2.2e-16

deviance(lm1)

## [1] 68436886

#Testing significance of smokes by controlling all other factors  
lm1.1 <- lm( bwt ~ age + gest + sex + weight + rate, data = Births)  
summary1.1 <- summary(lm1.1)  
summary1.1

##   
## Call:  
## lm(formula = bwt ~ age + gest + sex + weight + rate, data = Births)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1044.89 -278.37 -2.68 238.04 1840.68   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3068.400 494.367 -6.207 1.30e-09 \*\*\*  
## age 8.102 3.651 2.219 0.02699 \*   
## gest 20.735 1.733 11.963 < 2e-16 \*\*\*  
## sexMale 108.035 39.301 2.749 0.00624 \*\*   
## weight 8.138 1.975 4.121 4.55e-05 \*\*\*  
## rate 1599.635 602.616 2.654 0.00824 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 404.2 on 421 degrees of freedom  
## Multiple R-squared: 0.3204, Adjusted R-squared: 0.3123   
## F-statistic: 39.69 on 5 and 421 DF, p-value: < 2.2e-16

anova(lm1.1,lm1)

## Analysis of Variance Table  
##   
## Model 1: bwt ~ age + gest + sex + weight + rate  
## Model 2: bwt ~ age + gest + sex + smokes + weight + rate  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 421 68767746   
## 2 419 68436886 2 330860 1.0128 0.3641

deviance(lm1.1)

## [1] 68767746

We remove the smokes factor from the model as it has a high p-value at 0.58344 and 0.19285. Additionally, when we compare the p-value of 0.3641 from the anova output when we control all the factors except smokes, we can see that this is high and so significant. The adjusted R^2 has not changed so the affect of smokes is not significant. Also, from earlier sections we established that gestation has a significant affect on birthweight so we shall keep this factor in the model. From part 1e) when we tested for the affect of sex on birthweight we found that sex has a significant affect on birthweight for gestation being controlled. Although, the deviance has increased by a margin of 330860. This is relatively small and so does not have a big affect on the model so we will still remove smokes.

#New and improved model2 (having removed smokes)  
lm2 <- lm( bwt ~ age + gest + sex + weight + rate, data = Births)  
summary2 <- summary(lm2)  
summary2

##   
## Call:  
## lm(formula = bwt ~ age + gest + sex + weight + rate, data = Births)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1044.89 -278.37 -2.68 238.04 1840.68   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3068.400 494.367 -6.207 1.30e-09 \*\*\*  
## age 8.102 3.651 2.219 0.02699 \*   
## gest 20.735 1.733 11.963 < 2e-16 \*\*\*  
## sexMale 108.035 39.301 2.749 0.00624 \*\*   
## weight 8.138 1.975 4.121 4.55e-05 \*\*\*  
## rate 1599.635 602.616 2.654 0.00824 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 404.2 on 421 degrees of freedom  
## Multiple R-squared: 0.3204, Adjusted R-squared: 0.3123   
## F-statistic: 39.69 on 5 and 421 DF, p-value: < 2.2e-16

deviance(lm2)

## [1] 68767746

Here all p-values in the summary appear to be significant so no additional chanes will be made here until further tests.

#Testing significance of age by controlling all other factors  
lm2.1 <- lm( bwt ~ gest + sex + weight + rate, data = Births)  
summary2.1 <- summary(lm2.1)  
summary2.1

##   
## Call:  
## lm(formula = bwt ~ gest + sex + weight + rate, data = Births)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1037.44 -268.52 -1.01 238.03 1852.83   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2750.273 475.326 -5.786 1.41e-08 \*\*\*  
## gest 20.206 1.725 11.715 < 2e-16 \*\*\*  
## sexMale 112.168 39.439 2.844 0.00467 \*\*   
## weight 8.917 1.953 4.567 6.51e-06 \*\*\*  
## rate 1794.935 598.923 2.997 0.00289 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 406 on 422 degrees of freedom  
## Multiple R-squared: 0.3124, Adjusted R-squared: 0.3059   
## F-statistic: 47.94 on 4 and 422 DF, p-value: < 2.2e-16

model2.1..2 <- anova(lm2,lm2.1)  
model2.1..2

## Analysis of Variance Table  
##   
## Model 1: bwt ~ age + gest + sex + weight + rate  
## Model 2: bwt ~ gest + sex + weight + rate  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 421 68767746   
## 2 422 69572342 -1 -804596 4.9258 0.02699 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

deviance(lm2.1)

## [1] 69572342

The p-value from the anova output shows that age appears to have a signficant association with birthweight with a p-value of 0.02699. Similarly, the adjusted R^2 has decreased by 0.0064 to 0.3059 are low showing that the observed and fitted values are very close, which is ideal for a model. Additionally, by removing this factor the deviance increases by a large degree of 1135456, which is ideal and so, provides more evidence to keep age.

#New and improved model3 (after having removed smokes and included age)  
lm3 <- lm( bwt ~ age + gest + sex + weight + rate, data = Births)  
summary3 <- summary(lm3)  
summary3

##   
## Call:  
## lm(formula = bwt ~ age + gest + sex + weight + rate, data = Births)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1044.89 -278.37 -2.68 238.04 1840.68   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3068.400 494.367 -6.207 1.30e-09 \*\*\*  
## age 8.102 3.651 2.219 0.02699 \*   
## gest 20.735 1.733 11.963 < 2e-16 \*\*\*  
## sexMale 108.035 39.301 2.749 0.00624 \*\*   
## weight 8.138 1.975 4.121 4.55e-05 \*\*\*  
## rate 1599.635 602.616 2.654 0.00824 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 404.2 on 421 degrees of freedom  
## Multiple R-squared: 0.3204, Adjusted R-squared: 0.3123   
## F-statistic: 39.69 on 5 and 421 DF, p-value: < 2.2e-16

deviance(lm3)

## [1] 68767746

Here all p-values appear to be significant so no additional changes will be made here until further tests.

#Testing significance of weight by controlling all other factors  
lm3.1 <- lm( bwt ~ age + gest + sex + rate, data = Births)  
summary3.1 <- summary(lm3.1)  
summary3.1

##   
## Call:  
## lm(formula = bwt ~ age + gest + sex + rate, data = Births)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -915.4 -267.4 -14.9 241.1 1862.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2994.245 503.306 -5.949 5.67e-09 \*\*\*  
## age 10.775 3.660 2.944 0.00342 \*\*   
## gest 21.937 1.741 12.603 < 2e-16 \*\*\*  
## sexMale 104.378 40.028 2.608 0.00944 \*\*   
## rate 1476.216 613.161 2.408 0.01649 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 411.7 on 422 degrees of freedom  
## Multiple R-squared: 0.293, Adjusted R-squared: 0.2863   
## F-statistic: 43.71 on 4 and 422 DF, p-value: < 2.2e-16

model3.1..3 <- anova(lm3,lm3.1)  
model3.1..3

## Analysis of Variance Table  
##   
## Model 1: bwt ~ age + gest + sex + weight + rate  
## Model 2: bwt ~ age + gest + sex + rate  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 421 68767746   
## 2 422 71541235 -1 -2773490 16.98 4.552e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

deviance(lm2.1)

## [1] 69572342

Observing the p-value which is 4.552e-05, and so very small, meaning weight has a strong association with birthweight and we should keep it in the model. The adjusted R^2 has decreased by 0.026, which is very small and so can be ignored as the significance from the p-value has high. The deviance has increased but 804596 but again, the significance of the p-value is stronger so we will keep weight.

#New and improved model3 (after having removed smokes and included age)  
lm5 <- lm( bwt ~ age + gest + sex + weight + rate, data = Births)  
summary5 <- summary(lm5)  
summary5

##   
## Call:  
## lm(formula = bwt ~ age + gest + sex + weight + rate, data = Births)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1044.89 -278.37 -2.68 238.04 1840.68   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3068.400 494.367 -6.207 1.30e-09 \*\*\*  
## age 8.102 3.651 2.219 0.02699 \*   
## gest 20.735 1.733 11.963 < 2e-16 \*\*\*  
## sexMale 108.035 39.301 2.749 0.00624 \*\*   
## weight 8.138 1.975 4.121 4.55e-05 \*\*\*  
## rate 1599.635 602.616 2.654 0.00824 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 404.2 on 421 degrees of freedom  
## Multiple R-squared: 0.3204, Adjusted R-squared: 0.3123   
## F-statistic: 39.69 on 5 and 421 DF, p-value: < 2.2e-16

deviance(lm5)

## [1] 68767746

#Testing significance of rate by controlling all other factors  
lm5.1 <- lm( bwt ~ age +gest + sex + weight, data = Births)  
summary5.1 <- summary(lm5.1)  
summary5.1

##   
## Call:  
## lm(formula = bwt ~ age + gest + sex + weight, data = Births)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -985.37 -257.24 -9.05 257.49 1888.43   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3054.580 497.868 -6.135 1.96e-09 \*\*\*  
## age 9.517 3.637 2.617 0.00920 \*\*   
## gest 20.767 1.746 11.897 < 2e-16 \*\*\*  
## sexMale 106.579 39.578 2.693 0.00736 \*\*   
## weight 7.877 1.987 3.965 8.61e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 407 on 422 degrees of freedom  
## Multiple R-squared: 0.309, Adjusted R-squared: 0.3024   
## F-statistic: 47.18 on 4 and 422 DF, p-value: < 2.2e-16

model5.1..5 <- anova(lm5,lm5.1)  
model5.1..5

## Analysis of Variance Table  
##   
## Model 1: bwt ~ age + gest + sex + weight + rate  
## Model 2: bwt ~ age + gest + sex + weight  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 421 68767746   
## 2 422 69918712 -1 -1150967 7.0463 0.008244 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

deviance(lm5.1)

## [1] 69918712

Here, the p-value of 0.008244 is also very low, giving reason to keep rate in the model. The adjusted R^2 has also increased by 0.0099 which is quite and low which means it can be ignored. The deviance increased by 1,150,966 which although is high, is not as high in the wider picture, regardless we will keep rate in the model.