Homework 7

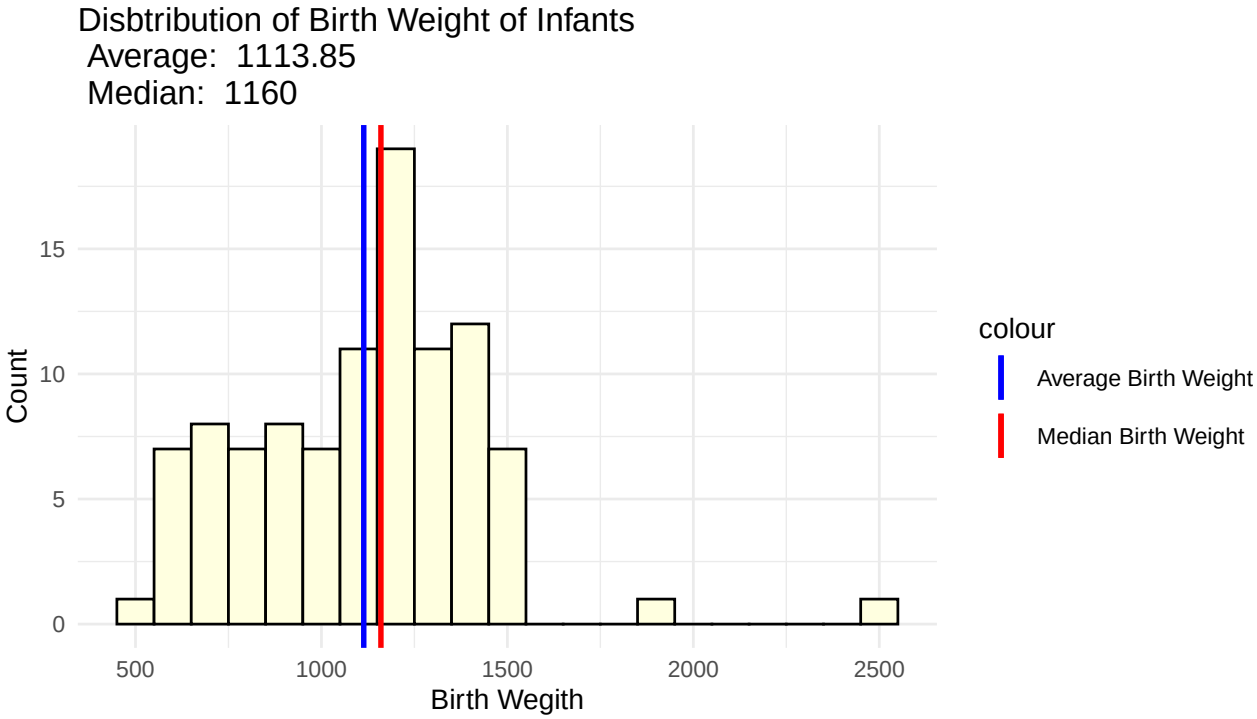
Denis Ostroushko

2022-11-01

14.2

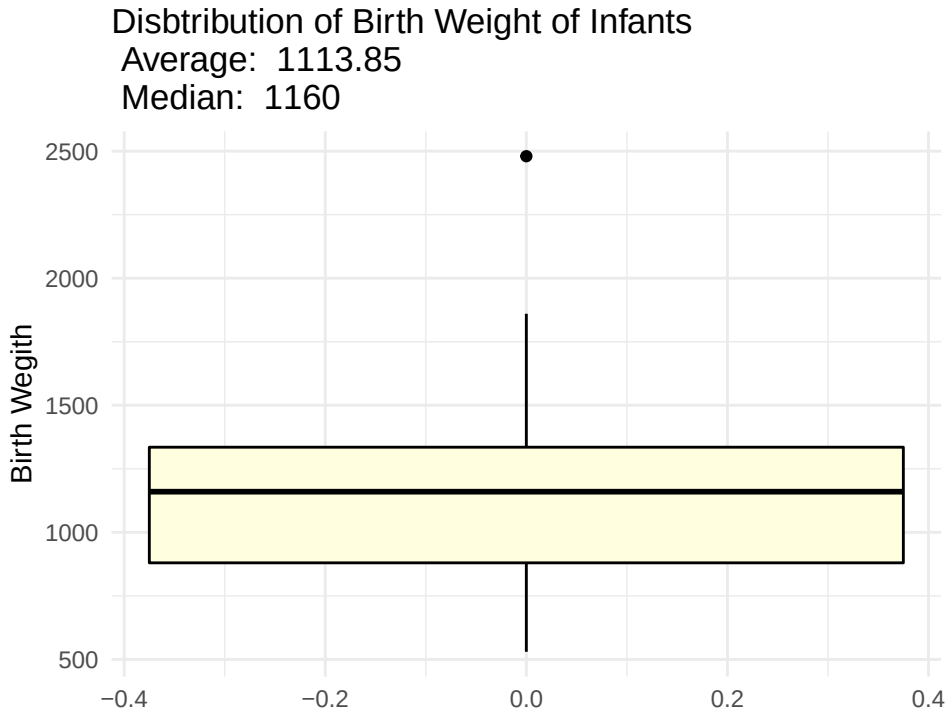
14.2 - A

First, we evaluate the distribution of birth weight measurements of infants. We will need to evaluate residuals and overall model fit, which are impacted by the shape of the response variable sample and outliers.



Histogram suggests that there are potential outliers, with a few infants having very high birth weights.

We now refer to the box plot to see if those few observations are in fact outliers, or just appear as visually extreme observations.



The box plot suggests that the most extreme observation is in fact an outlier, hopefully it will not affect the model fit and estimates.

Model

We now fit a regression model, regression model is below:

$$E[Birth\ \ Weight_i] = \hat{\beta}_0 + \hat{\beta}_1 * Gestational\ \ Weeks + \hat{\beta}_2 * Mother'sAge + \hat{\beta}_3 * Toxemia\ \ Flag$$

Overall ANOVA

Before investigating individual coefficients and t-test for predictors, we want to look at the overall ANOVA table, and overall F-test. We want to see if the set of all predictors is helpful at explaining the variance of infants' birth weights and therefore we will know if some of all coefficients are statistically different from 0.

ANOVA table for the F-test is given below:

Source	SSR	DF	MS	F Statistic	P(F* > F)
Regression	4784507	3	1594835.50	30.61	0
Error	5001186	96	52095.69	NA	NA
Total	9785693	99	NA	NA	NA

- Null Hypothesis: $H_0 : \beta_1 = \beta_2 = ... = \beta_{p-1}$
- Alternative Hypothesis: H_a : Not all coefficients β_i are zero
- $F-$statistic: 30.61
- Cutoff F^*-statistic: 2.6994
- So, $F > F^*$, therefore we have enough evidence to reject the null hypothesis and conclude that some or all coefficients β_i are consistently different from zero.

- Moreover, $P(F^* < F) = 0$, so the results are very convincing here, we should see some very strong predictors in the model, especially considering that this set of predictors is able to explain 48.89% of variance in birth weights

Regression Coefficients

Coefficients for predictors, estimate standard errors and t-tests are given in the table below.

Predictor	Estiamte	Standard Error	T Value	P value
(Intercept)	-1551.14484	285.484477	-5.433377	0.000000
gest_weeks	96.21208	10.237777	9.397750	0.000000
mom_age	-2.07795	3.977342	-0.522447	0.602563
toxemia	-271.19265	61.499368	-4.409682	0.000027

It appears that the number of gestational weeks and toxemia flag are extremely strong predictors that we need to retain.

- An additional week of gestation adds an average of 96.21 pounds to infant's birth weight, after adjusting for other variables.
- Presence of toxemia on average reduces the birth weight by 271.19 pounds, after adjusting for other variables.
- Mom's age is not a string predictor with a coefficient close to 0, relative to the scale of outcome measurement and other coefficients' values. Therefore, we can remove this predictor in an applied research setting, unless we have a strong desire to keep it in the model.

14.2 - B

We use an added variable plot in order to evaluate the nature of the relationship between birth weights and the number of gestational weeks after adjusting for the other 2 predictors. We will need to obtain two sets of residuals from the two models:

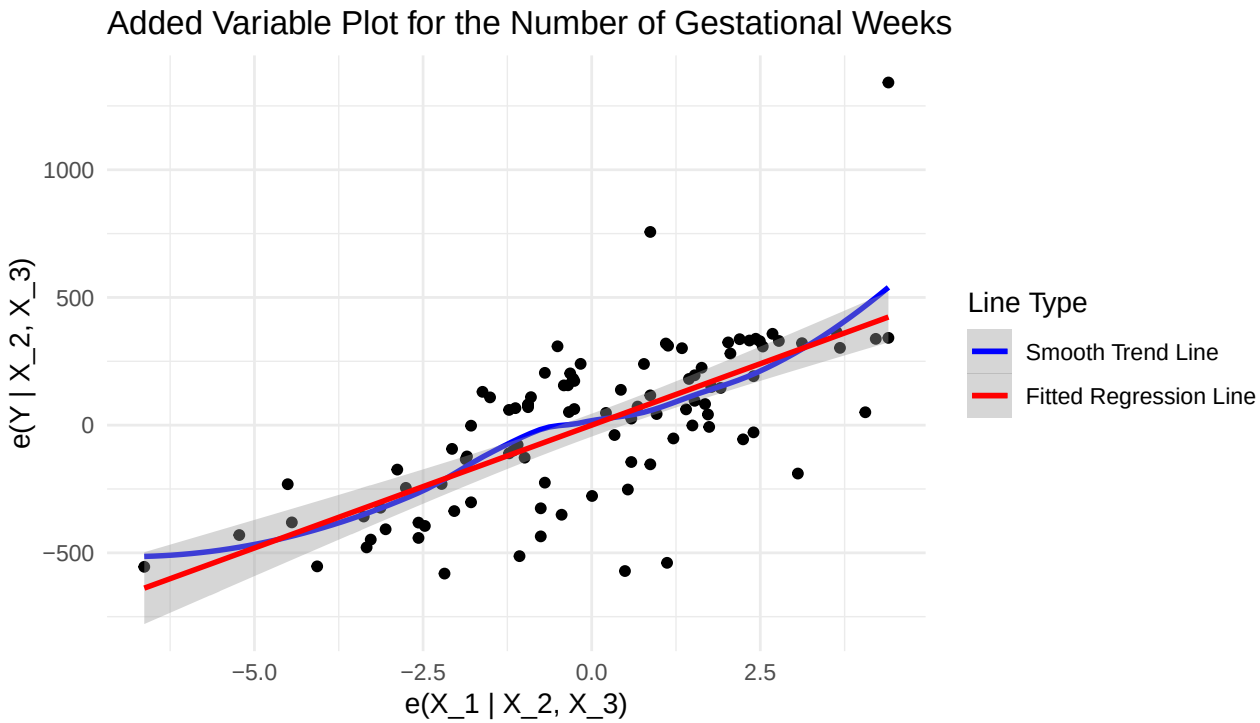
- Model 1: obtains residuals for Y = birth weight. We denote these residuals as $\epsilon_Y = e(Y|X_2, X_3)$:
  - $Y = \hat{\beta}_0 + \hat{\beta}_2 * Mom's\ \ Age + \hat{\beta}_3 * Toxemia\ \ Flag + \epsilon_Y$
- Model 2: obtains residuals for \$X_1 = \$ N of Gestational Weeks. We denote these residuals as $\epsilon_X = e(X_1|X_2, X_3)$:
  - $X_2 = \hat{\beta}_0 + \hat{\beta}_2 * Mom's\ \ Age + \hat{\beta}_3 * Toxemia\ \ Flag + \epsilon_x$

Plot below shows the relationship between the two sets of residuals:

X1 Added Variable Plot

```
y_reg <- lm(birthwght ~  mom_age + toxemia, data = infant_m)
x_reg <- lm(gest_weeks ~ mom_age + toxemia, data = infant_m)

d <-
  data.frame(
    y_res = y_reg$residuals,
    x_res = x_reg$residuals
  )
```



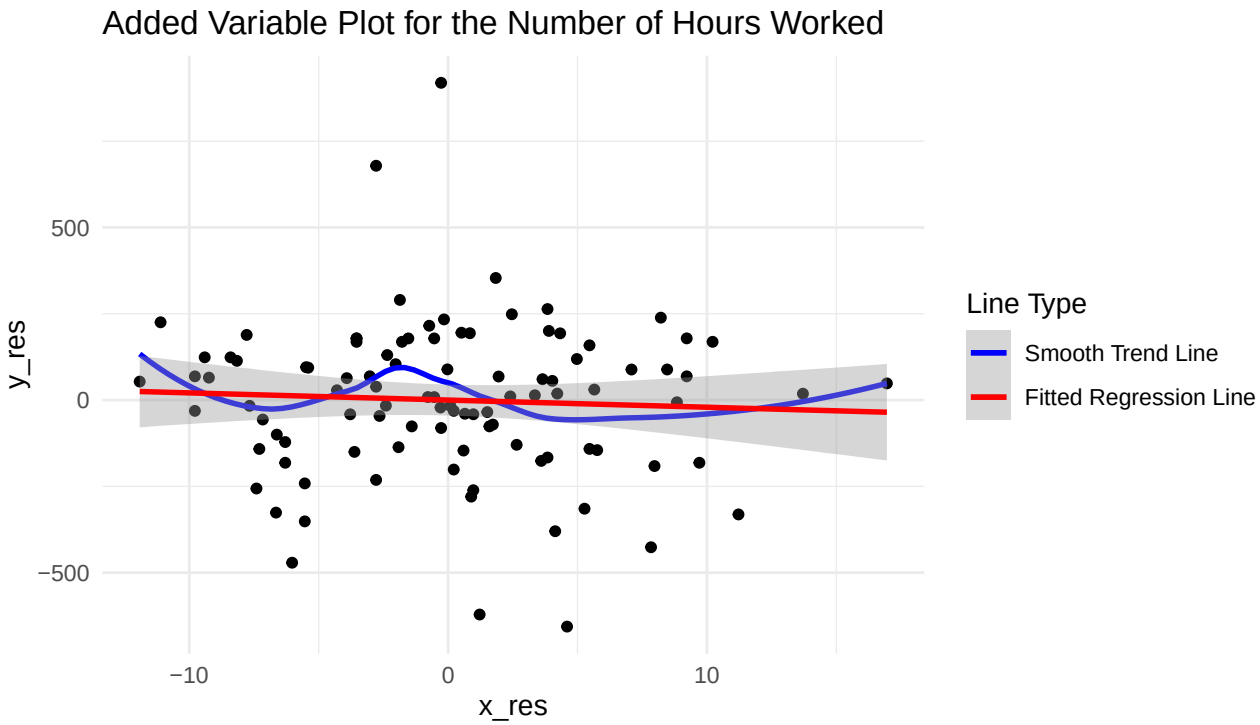
1. The relationship between X_1 and Y, after accounting for other predictors, is linear in its nature. We can see that the smooth trend line fluctuates randomly around the fitted regression line, suggesting that there is no consistent curved, or other non-linear relationship between the number of gestational weeks and birth weight.
2. The fitted regression line that confirm linear relationship has a positive, upward facing, slope, suggesting that the number of gestational weeks can be used a potentially useful predictor that help increase the percentage of variation in birth weight that this model explains.
3. One potentially troublesome conclusion is the issue with variance assumption. We can see that as values increase along the $x-$axis, so does the spread of data points around the average fitted line. We will investigate more in the later section.

X2 Added Variable Plot

We use the same procedure as described previously. We will skip the explanation and obtain the plot.

```
y_reg <- lm(birthwght ~  gest_weeks + toxemia, data = infant_m)
x_reg <- lm(mom_age ~  gest_weeks + toxemia, data = infant_m)

d <-
  data.frame(
    y_res = y_reg$residuals,
    x_res = x_reg$residuals
  )
```



1. The relationship between X_2 and Y, after accounting for other predictors, is linear in its nature. We can see that the smooth trend line fluctuates randomly around the fitted regression line, suggesting that there is no consistent curved, or other non-linear relationship between mom's age and birth weight.
2. This plot supports the statement that mom's age is not a useful predictors of birth weight in the context of this model. We can see that there are definitely some outliers, where babies are way heavier than what is predicted by the model. These outliers are perhaps due to other factors such as mom's height or other size measurements.
3. This plot highlights less issues with the constant variance assumption, which we should not pay much attention to, since this variable should be removed from the model

14.2 - C

We obtain Variable Inflation Factors(VIFs) by creating three regression models. We follow these steps for each regression model:

1. Use predictor i as a response variable
2. Use all other predictors, except for i, as predictors of variable i
3. Obtain R^2 from the model, high value implies some predictors from the independent set are related to predictor i
4. Obtain VIF $= (1 - R^2)^{-1}$

First, let's investigate the correlation matrix between all three predictors. Note that *toxemia* is a binary indicator, so this correlation coefficient is not meaningful for interpretation, but we can still use it to see the degree of 'correlation' between toxemia and the other two predictors

	Gestational Weeks	Mom's Age	Toxemia
Gestational Weeks	1.0000000	0.2658396	0.4119675
Mom's Age	0.2658396	1.0000000	0.1141187
Toxemia	0.4119675	0.1141187	1.0000000

It appears that Toxemia and Gestational Weeks count are moderately correlated, so we should expect their VIF values to be above 1. Mom's age is weakly correlated with the other two predictors, which would be favorable if Mom's age was an actually useful predictor.

```
r_sq_gest_weeks <- summary(lm(gest_weeks ~ mom_age  + toxemia, data = infant))$r.squared
r_sq_mom_age <- summary(lm(mom_age ~ gest_weeks  + toxemia, data = infant))$r.squared
r_sq_toxemia <- summary(lm(toxemia ~ mom_age  +  gest_weeks, data = infant))$r.squared

vif_d <-
  data.frame(
    var = c("Gestational Weeks", "Mom's Age", "Toxemia"),
    r_sq = c(r_sq_gest_weeks, r_sq_mom_age, r_sq_toxemia),
    std_e = summary(full)$coefficients[2:4,2]
  )

rownames(vif_d) <- NULL

vif_d$vif <- 1/(1-vif_d$r_sq)
vif_d <- vif_d %>% select(-r_sq)
vif_d %>%
  kbl(booktabs = T, align = c('l','c', 'c', 'c', 'c'),
      col.names = c("Variable", "Standard Error", "VIF")) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))
```

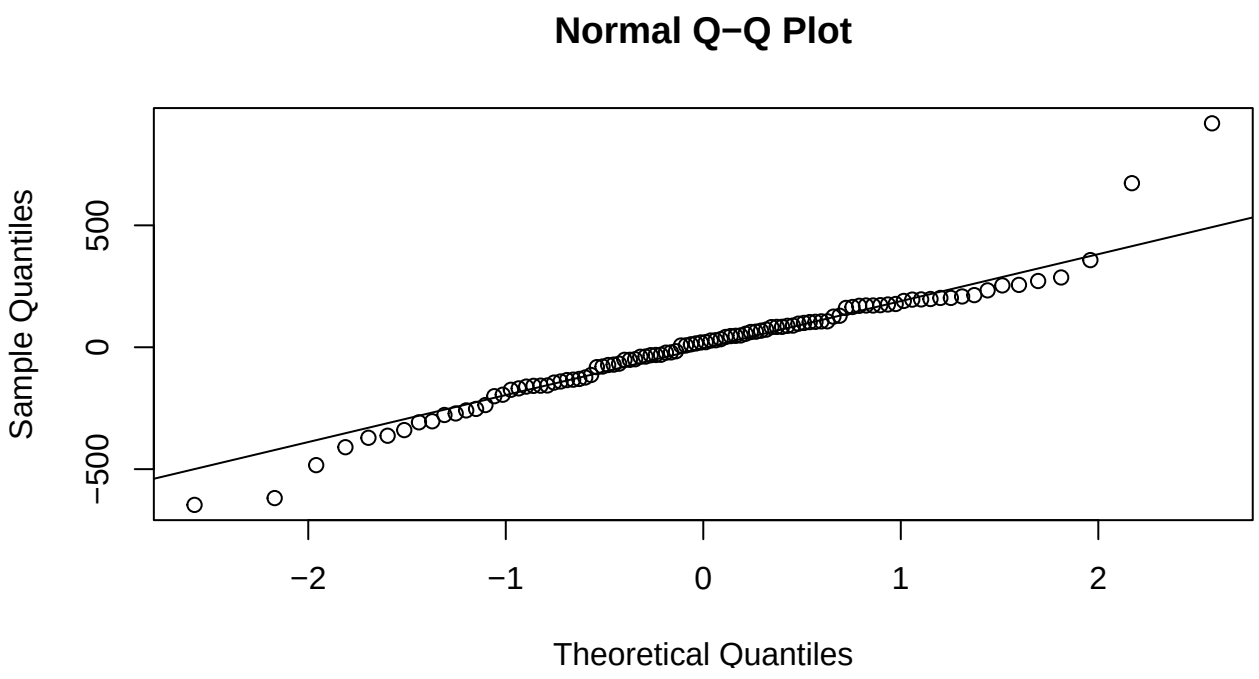
Variable	Standard Error	VIF
Gestational Weeks	10.237777	1.279155
Mom's Age	3.977342	1.076074
Toxemia	61.499368	1.204442

- Overall, these values of VIF are not a cause for concern
  - We would investigate a certain predictor is VIF value exceeded 5
  - We would say we have strong evidence for a multicollinearity issue if a single VIF value here would be greater than 10
- A VIF value of 1.2791552 for Gestational Weeks means that the variance for this predictor was inflated by 27.9155238%
  - For coefficients $\hat{\beta}_i$ variance is given by $Var(\hat{\beta}_i) = se(\hat{\beta}_i)^2$
  - using variance property $Var(aX) = a^2 var(X)$, we know that the for the predictor i standard error was inflated by a factor $\sqrt{VIF_i}$
  - for example, standard error for Gestational Weeks is 10.2377773, which was inflated by 13.0997453% due to correlation with other predictors
  - Overall, this should not be a cause for concern, from a practical point of view if we were to use this model for inference and recommendations
- Since the greatest VIF value is not a cause for concert, the other two values also follow the same conclusion.

14.2 - D

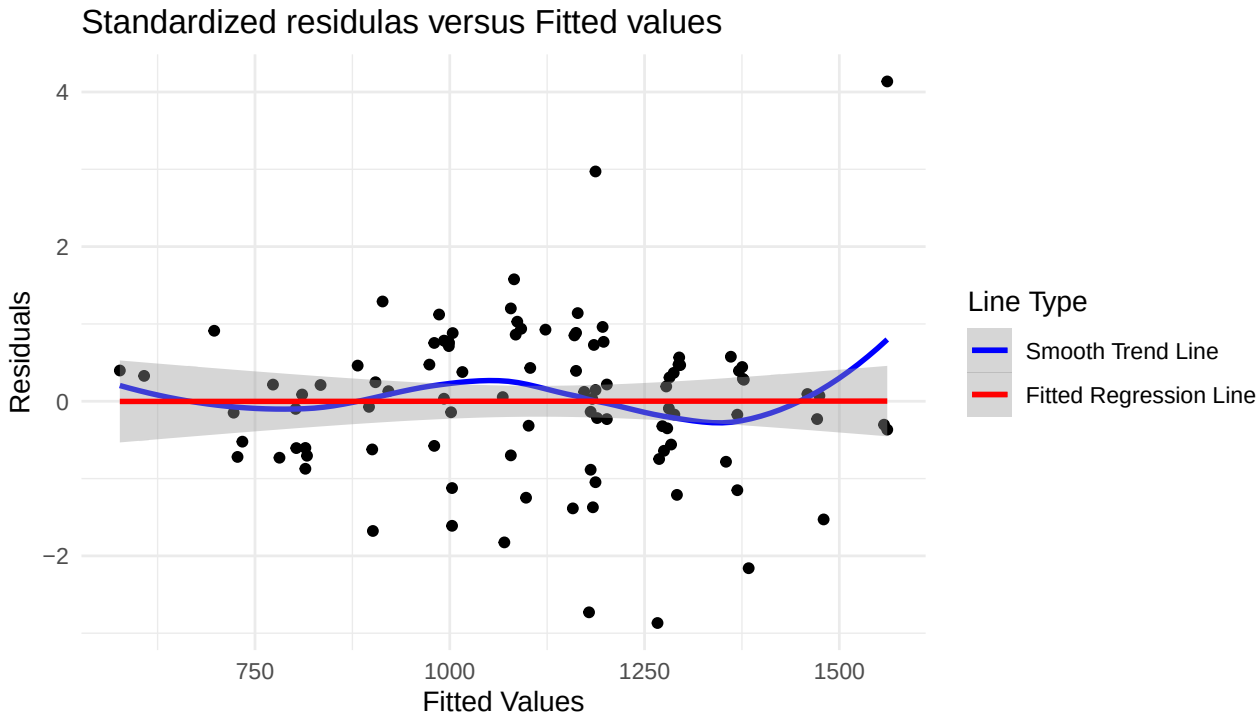
Residuals

First, we plot residuals against the expected quadrilles under the normal distribution to verify that the residuals are in fact normally distributed.



Overall, it seems that for the most part there are no issues, residuals should be approximately normally distributed. There are some heavier than expected tails present in the data. Overall, nothing that will be a severe issue.

We plot standardized residuals against fitted values and obtain a residual plot:



We can see multiple issues with this plot:

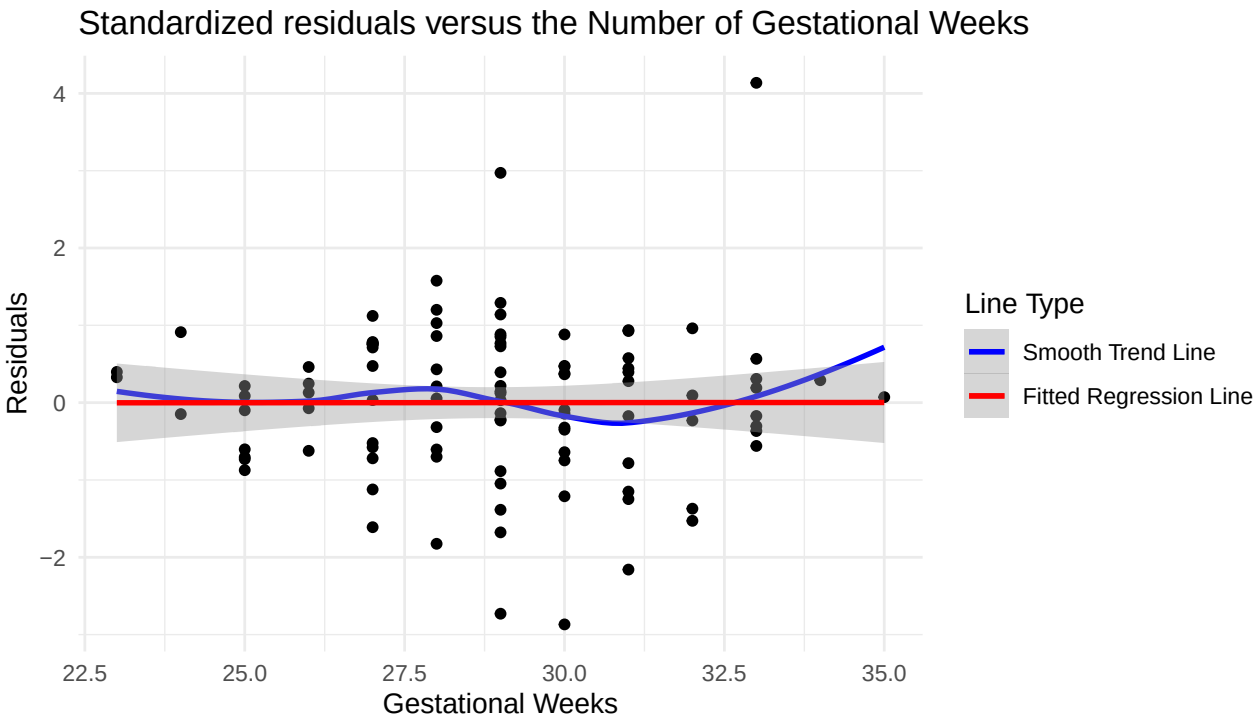
- The variance of residuals in not constant across different values of fitted values. We have a megaphone shape. So, as the predicted values of birth weight increase, so does the variance of residuals, so the

model does not perform well for babies with higher birth weights.

- There are some obvious positive outliers.

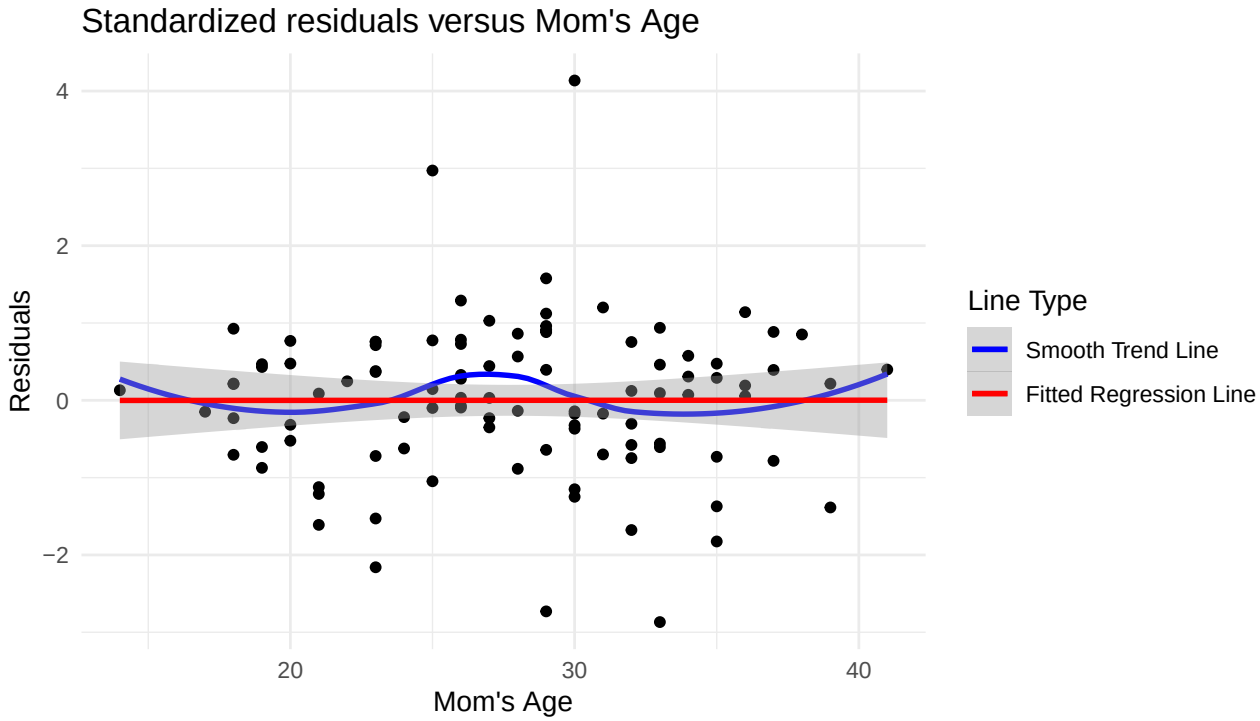
We can see that some assumptions also held:

- The average values of residuals is 0
- There is no linear trend in residuals against fitted values, so the two are not correlated.



The plot of residuals versus the number of gestational weeks follows the same conclusions as the plot of residuals versus fitted values. In fact, the plots look extremely similar. The same two outliers are visible on the upper end of residuals. Variance of residuals is even greater on for residuals whose values are below zero.

And lastly we present the residual plot of standardized residuals versus the mom's age



This residual plot looks different from the previous two.

- First of all, the issue with non constant variance is not apparent from this plot, unlike the plot versus gestational weeks. Given that mom's age is not a significant predictor, while gestational weeks is, makes me believe that gestational weeks is somehow the root cause of non-constant variance of residuals
- The fitted line shows slight upward trend, however, the regression line confidence bound also includes a hypothetical line with slope 0, so this trend is not a cause for concern

Overall

The residual plot shows that there is a non-constant variance problem with the model. We can address it using WLS regression model.

There are a few outliers that we need to address. However, that should be done after we re-fit the WLS model.

Other assumptions violations are not apparent.

15.3

15.3 - A

We look at the number of unique levels for categorical predictors to calculate the total number of covariates. If a categorical variable has p levels, then it can be represented using $p - 1$ binary flag variables.

	class	n_unique
tumor_cat	factor	6
stageT	factor	5
stageN	factor	5
drinker	factor	4
smoker	factor	3
margins	factor	3
nodes_pos	factor	3
gender	factor	2
stageM	factor	2
diabetes	factor	2
heart_dx	factor	2
stroke	factor	2
lung_dx	factor	2
arthritis	factor	2
psych	factor	2
ctx	factor	2
income	numeric	144
age_diag	numeric	52

- For example, tumor category can be represented with 5 variables

The total number of predictors in the logistic regression model will be 32

15.3 - B

All coefficients are given below, $*$ mark predictors that are significant at the 0.05 level

Model Term	Estimate	Std. Error	T-value	P-value	Significance
age_diag	0.013	0.016	0.792	0.428	
income	0.000	0.000	2.158	0.031	*
gender1	0.314	0.430	0.731	0.465	
tumor_cat2	-1.442	0.483	-2.988	0.003	*
tumor_cat3	0.138	0.864	0.160	0.873	
tumor_cat4	-0.770	1.728	-0.445	0.656	
tumor_cat5	1.297	0.921	1.408	0.159	
tumor_cat6	0.850	1.586	0.536	0.592	
stageT2	-0.011	0.555	-0.019	0.984	
stageT3	0.842	0.625	1.349	0.177	
stageT4	-0.401	0.557	-0.720	0.471	
stageT9	-1.481	1.712	-0.865	0.387	
stageN1	-1.189	0.609	-1.953	0.051	
stageN2	-0.278	0.473	-0.587	0.557	
stageN3	0.028	1.343	0.021	0.984	
stageN9	-13.642	1455.399	-0.009	0.993	
diabetes1	-0.663	1.217	-0.545	0.586	
heart_dx1	0.235	0.625	0.375	0.707	
stroke1	0.714	1.633	0.437	0.662	
lung_dx1	-1.327	0.703	-1.888	0.059	
arthritis1	-1.029	0.702	-1.467	0.142	
smoker1	-0.638	0.569	-1.122	0.262	
smoker2	-0.777	0.626	-1.240	0.215	
drinker1	0.542	0.603	0.899	0.368	
drinker2	1.032	0.602	1.715	0.086	
drinker9	-17.277	1455.399	-0.012	0.991	
psych1	0.289	0.595	0.485	0.627	
margins1	-0.528	0.521	-1.013	0.311	
margins9	-14.264	1455.398	-0.010	0.992	
nodes_pos1	0.760	0.523	1.451	0.147	
nodes_pos9	14.115	1455.398	0.010	0.992	
ctx1	1.433	0.443	3.233	0.001	*

Income is statistically significant, but the coefficient is very small

Tumor category 2 and chemotherapy flag are also two statistically significant predictors.

We have a lot of simultaneous t-test, and therefore a high chance of detecting a false positive result. Use bonferroni adjustment to see what predictors are actually useful from a large set of predictors.

Bonferroni

Significance level with the bonferroni adjustment is 0.0015625

Model Term	P-value	Significant at Adj. Level
age_diag	0.428	
income	0.031	
gender1	0.465	
tumor_cat2	0.003	
tumor_cat3	0.873	
tumor_cat4	0.656	
tumor_cat5	0.159	
tumor_cat6	0.592	
stageT2	0.984	
stageT3	0.177	
stageT4	0.471	
stageT9	0.387	
stageN1	0.051	
stageN2	0.557	
stageN3	0.984	
stageN9	0.993	
diabetes1	0.586	
heart_dx1	0.707	
stroke1	0.662	
lung_dx1	0.059	
arthritis1	0.142	
smoker1	0.262	
smoker2	0.215	
drinker1	0.368	
drinker2	0.086	
drinker9	0.991	
psych1	0.627	
margins1	0.311	
margins9	0.992	
nodes_pos1	0.147	
nodes_pos9	0.992	
ctx1	0.001	*

Chemotherapy flag is the only useful predictor, after adjusting for other variance, at the bonferroni adjusted level

Hochberg

To verify the results, we also look at the Hochberg adjustments.

Model Term	P-value	Comparison P-value	Significant at Adj. Level
stageN9	0.993	0.050	
margins9	0.992	0.025	
nodes_pos9	0.992	0.017	
drinker9	0.991	0.013	
stageT2	0.984	0.010	
stageN3	0.984	0.008	
tumor_cat3	0.873	0.007	
heart_dx1	0.707	0.006	
stroke1	0.662	0.006	
tumor_cat4	0.656	0.005	
psych1	0.627	0.005	
tumor_cat6	0.592	0.004	
diabetes1	0.586	0.004	
stageN2	0.557	0.004	
stageT4	0.471	0.003	
gender1	0.465	0.003	
age_diag	0.428	0.003	
stageT9	0.387	0.003	
drinker1	0.368	0.003	
margins1	0.311	0.002	
smoker1	0.262	0.002	
smoker2	0.215	0.002	
stageT3	0.177	0.002	
tumor_cat5	0.159	0.002	
nodes_pos1	0.147	0.002	
arthritis1	0.142	0.002	
drinker2	0.086	0.002	
lung_dx1	0.059	0.002	
stageN1	0.051	0.002	
income	0.031	0.002	
tumor_cat2	0.003	0.002	
ctx1	0.001	0.002	*

It again appears that chemotherapy status is the only important predictor