ST552 HW 8

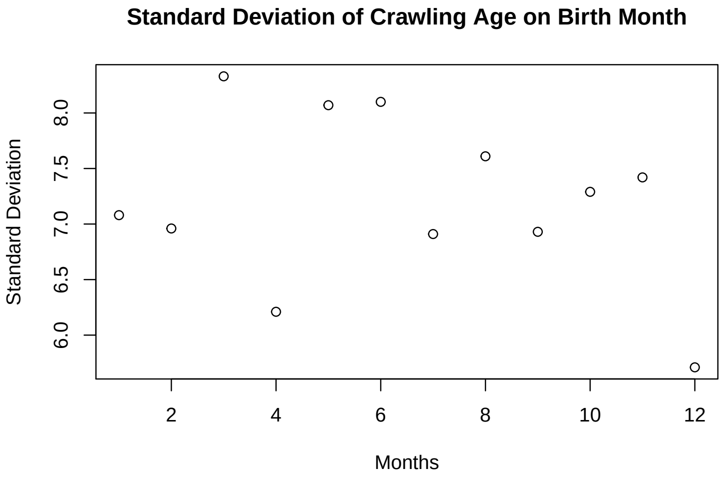
Nick Sun

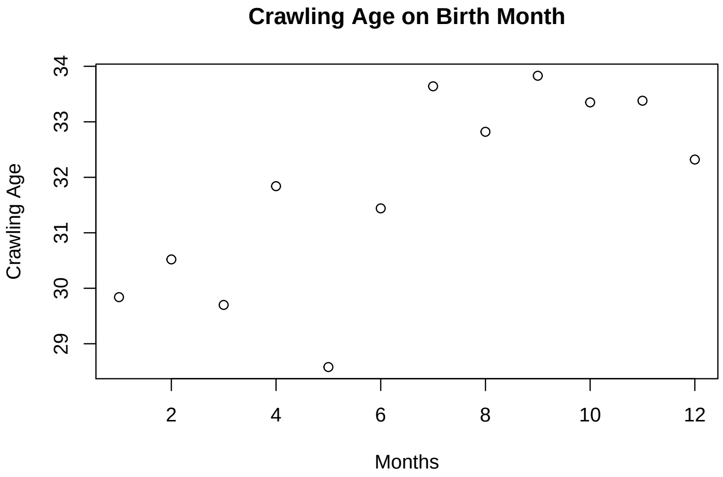
March 4, 2019

Faraway 8.7

Crawling in my skin, these woooooounds they will not heaaaal.

Let's examine some data of when babies started crawling and the average temperature 6 months after their birth. We have 12 data points here, one for each month. Each data point is an average of several babies for that month.





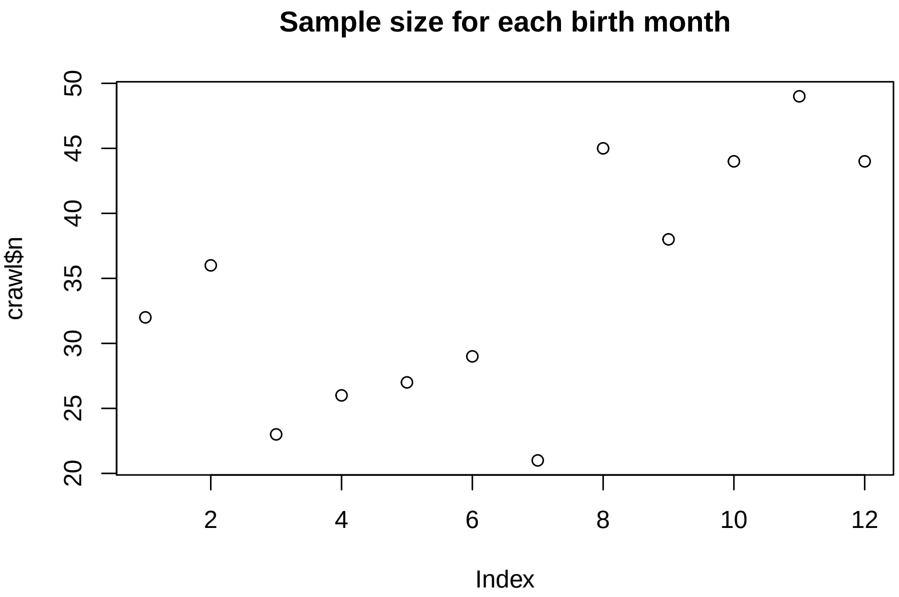
Looks like tere is some pretty significant changes in the variation of crawling age that is dependent upon the birth month. This problem is a great candidate for weighted least squares.

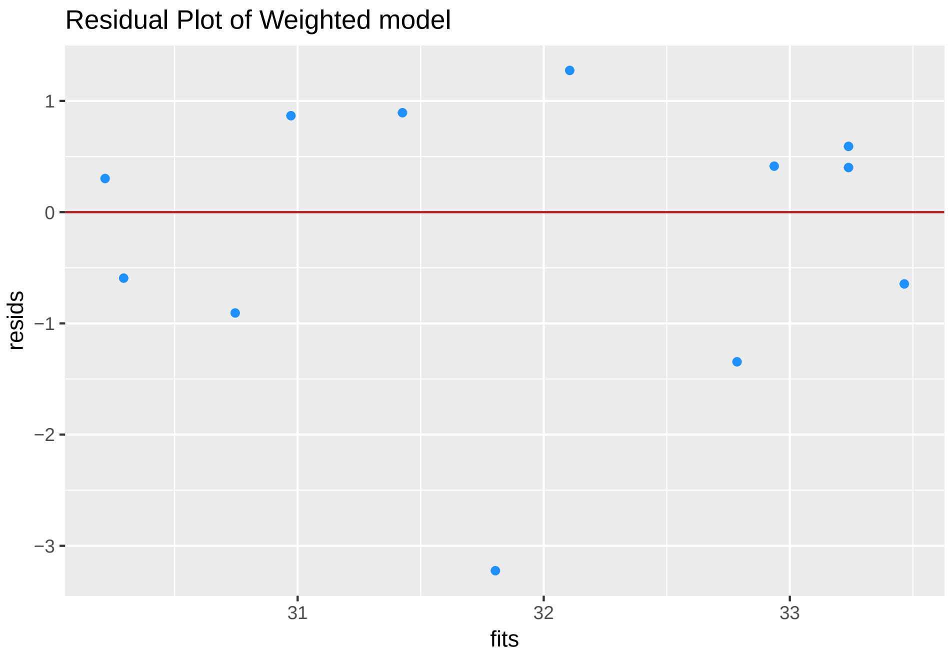
Our goal is to figure out Σ which here is a diagonal matrix where the w_{ii} are weights for each observation. Normally, since our observation Y_i are averages of n_i observations, then

$$var(y_i) = var(\epsilon_i) = \frac{\sigma^2}{n_i} \rightarrow w_i = n_i$$

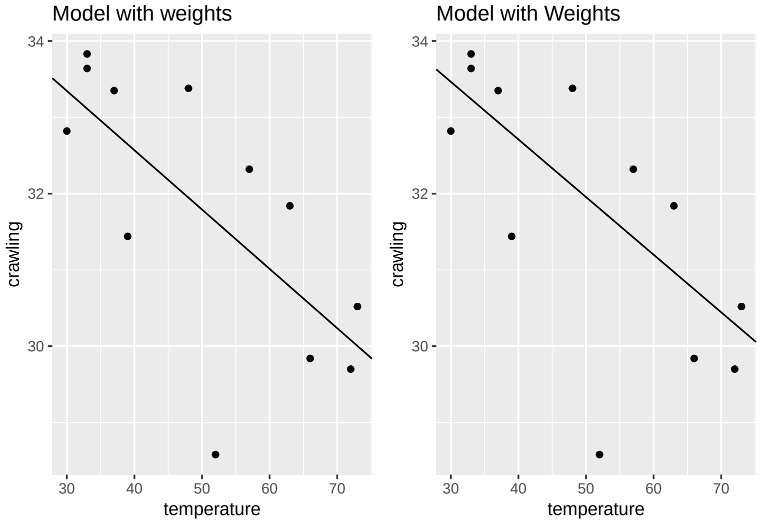
where n_i is the number of babies measured in that month.

However, from the premise of the question we are assuming here that $weight_i \propto 1/Var(y_i)$. Using these weights:





When we use these weights, there really isn't that big of a difference. This is more apparent when you plot the weighted and unweighted coefficients side by side.


Model with weights

Model with Weights

That being said, if we know that there is a correlation between observations, we should still opt to use the weighted model for our final interpretations. Here is the summary of the weighted model.

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

(Intercept) 35.730840 1.271539 28.101 7.567e-11 temperature -0.075522 0.024021 -3.144 0.01044

n = 12, p = 2, Residual SE = 0.17466, R-Squared = 0.5

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	35.73	1.272	28.1	7.567e-11
temperature	-0.07552	0.02402	-3.144	0.01044

Table 2: Fitting linear model: crawling ~ temperature

Observations	Residual Std. Error	R^2	Adjusted R^2
12	0.1747	0.4971	0.4468

Using this model for inference, as temperature increases we see a decrease in the mean time before a baby starts crawling. The exact estimated effect is every degree increase in average temperature (F) is associated with a decrease of .07 weeks in mean age that crawling begins.

Faraway 8.4

Checking out the `cars` data set and using a lack of fit test! We being by fitting a simple regression model with `dist ~ speed`.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-17.58	6.758	-2.601	0.01232
speed	3.932	0.4155	9.464	1.49e-12

Table 4: Fitting linear model: dist ~ speed

Observations	Residual Std. Error	R^2	Adjusted R^2
50	15.38	0.6511	0.6438

We also create a saturated model where each level in `speed` is treated as its own factor. This model essentially just uses the mean of each `speed` level as its $\hat{y}$.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6	10.45	0.5744	0.5698
factor(speed)7	7	14.77	0.4739	0.6389
factor(speed)8	10	18.09	0.5527	0.5844
factor(speed)9	4	18.09	0.2211	0.8265
factor(speed)10	20	13.49	1.483	0.1481
factor(speed)11	16.5	14.77	1.117	0.2726
factor(speed)12	15.5	12.79	1.212	0.2348
factor(speed)13	29	12.79	2.267	0.03052
factor(speed)14	44.5	12.79	3.478	0.001518
factor(speed)15	27.33	13.49	2.027	0.05134
factor(speed)16	30	14.77	2.031	0.05092
factor(speed)17	34.67	13.49	2.571	0.01517
factor(speed)18	58.5	12.79	4.573	7.28e-05
factor(speed)19	44	13.49	3.263	0.002686
factor(speed)20	44.4	12.36	3.592	0.001117
factor(speed)22	60	18.09	3.316	0.002334
factor(speed)23	48	18.09	2.653	0.01247
factor(speed)24	87.75	12.79	6.859	1.094e-07
factor(speed)25	79	18.09	4.367	0.0001307

Table 6: Fitting linear model: dist ~ factor(speed)

Observations	Residual Std. Error	R^2	Adjusted R^2
50	14.77	0.7921	0.6714

We then use an F-test to compare these models. Doing so will help us determine if the estimate of model-free variance (represented by the saturated model) is significantly less than the regression standard error of our chosen model. A low p-value will indicate that there is a lack of fit in our chosen model.

Table 7: Analysis of Variance Table

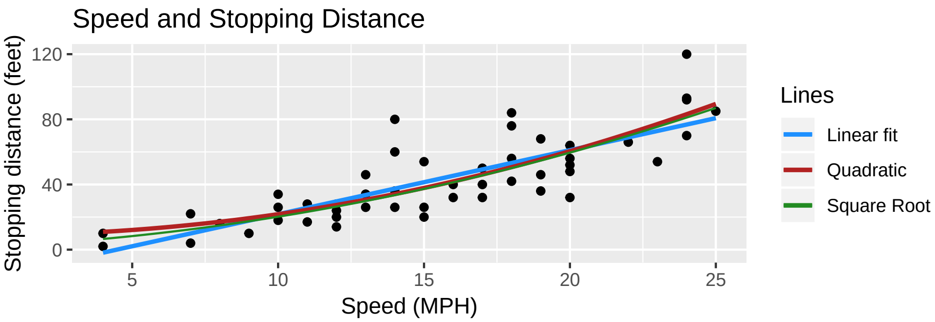
Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
48	11354	NA	NA	NA	NA
31	6765	17	4589	1.237	0.2948

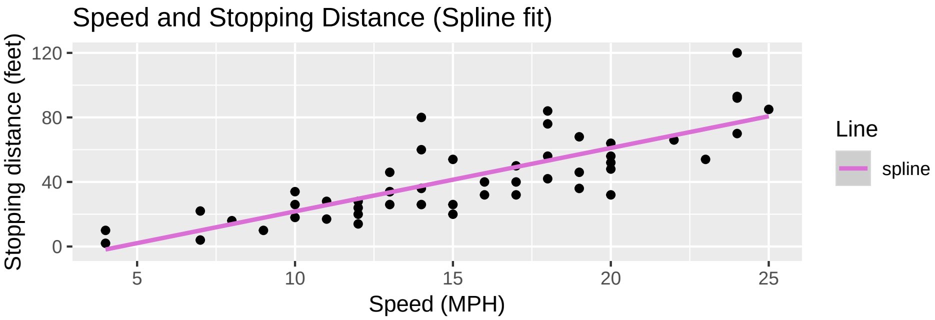
Our p-value is too large to reject the H_0 and we have no evidence to say that there is a lack of fit with the first model.

Faraway 9.8

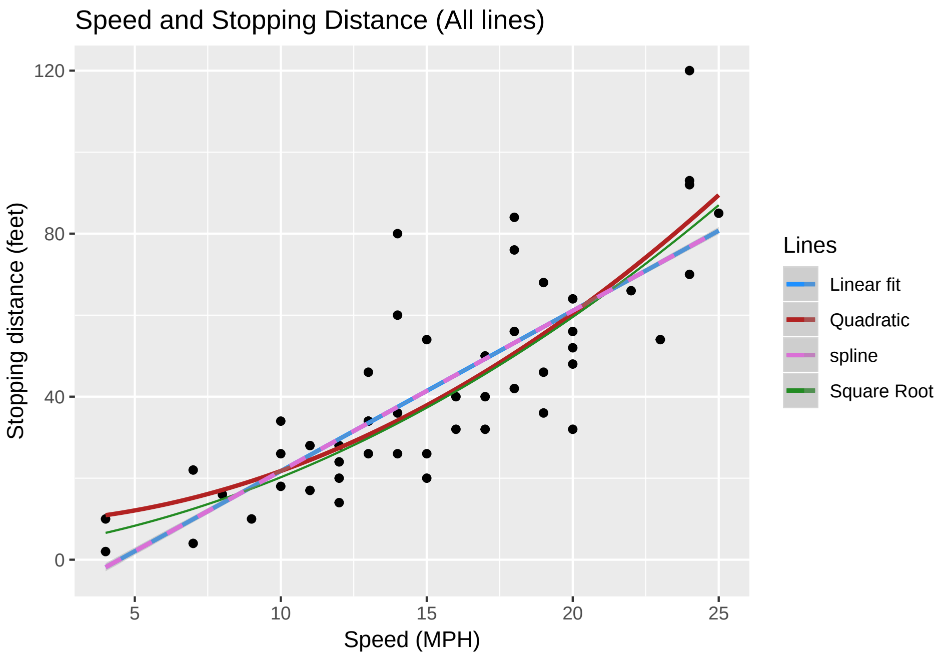
Using the cars data again! We are graphically comparing a linear fit to a quadratic fit to a backtransformed square-root!

Then we are comparing all of those fits to spline fit which we get using the `smooth.spline()` function. Let's see what we get.





The spline is pretty close to the other fitted lines. In fact, If we plot all the lines together, we can see that the spline fits almost right on top of the linear fit.


Speed and Stopping Distance (All lines)
Stopping distance (feet)
120
80
40
0
5
10
15
20
25
Speed (MPH)
Lines
Linear fit
Quadratic
spline
Square Root

AFQT Analysis

Introduction

The AFQT score is a measure of intelligence derived from the Armed Forces Vocational Aptitude Battery Test (ASVAB) that is given to all prospective US military members. We will seek to answer **if there is any evidence that the mean salary of males exceeds the mean salary for females after accounting for education and AFQT scores**. The study followed individuals from 1981 when they first took the ASVAB to 2006 when their salary was recorded.

Our data consists of 2583 individuals: 1278 females and 1306 males. There are 5 total variables: **Subject ID**, **gender**, **AFQT score** (percentile), **years of education**, and **salary** (USD in 2005).

Methods

To answer our question, we will fit the following linear model:

$$salary_i = \beta_0 + \beta_1 gender_i + \beta_2 afqt_i + \beta_3 education_i + \epsilon_i$$

The parameter estimate for β_1 should correspond to the mean salary difference between males and females after accounting for education and AFQT score.

Like with any linear model, inference on this model comes with the following assumptions: a linear relationship between the predictors and the response, independently distributed errors, and normality of those errors. We will check linearity graphically, and the normality of errors using residual plots and the Shapiro-Wilk test. If errors are not normal or identically distributed, we may have to use a response transformation.

Results

Let's first begin with a numerical summary.

Table 8: Income (USD) by Gender

Gender	min	Q1	median	Q3	max	mean	sd
female	147	16000	29811	45000	253043	35211	28776
male	63	32000	50000	78000	703637	63319	55861

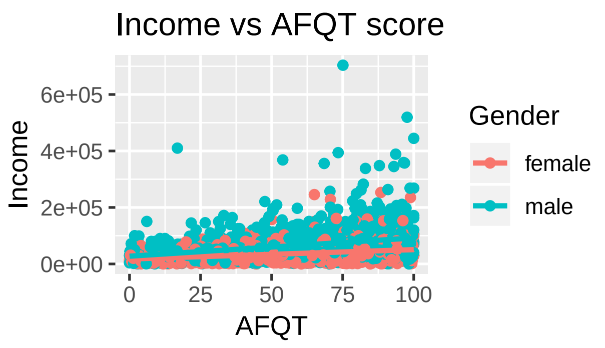
Table 9: Education (Years) by Gender

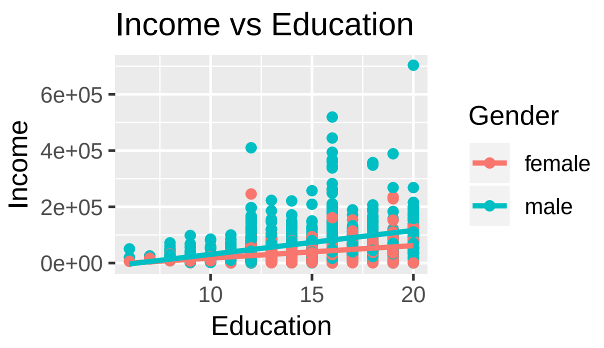
Gender	min	Q1	median	Q3	max	mean	sd
female	6	12	13	16	20	13.97	2.412
male	6	12	13	16	20	13.81	2.588

Table 10: AFQT Score by Gender

Gender	min	Q1	median	Q3	max	mean	sd
female	0	31.59	54.92	76.62	100	53.41	26.89
male	0	31.38	58.99	79.75	100	55.45	28.57

From this numerical summary, we notice that there is income disparity between males and females, even though education and AFQT score does not differ dramatically between males and females. Let's investigate the linearity assumption for a linear model using scatterplots.





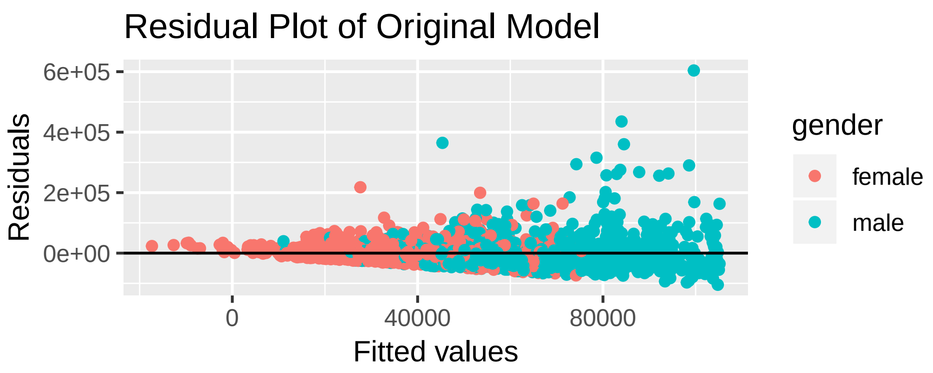
There does seem to be linearity here, but we also notice that the variability of the data seems to change along the predictors. When we fit the linear model, we see how this affects our model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-48760	4868	-10.02	3.417e-23
Gendermale	28463	1621	17.56	2.776e-65
Education	5158	402.7	12.81	1.811e-36
AFQT	223	36.32	6.139	9.558e-10

Table 12: Fitting linear model: Income ~ Gender + Education + AFQT

Observations	Residual Std. Error	R^2	Adjusted R^2
2584	41081	0.228	0.2271

While all of our predictors are signifcant and the F statistic for this model has an incredibly small p-value ($< .00001$), we see that our R^2 is not great. From our graphical displays above, we should investigate the error assumptions of this model using a residual plot.



We see a very noticeably funnel shape, which violates our assumption of constant variance. Additionally, our

Shapiro-Wilk test also reports a *very* tiny p-value ($< .00001$) indicating that these errors are not normal. It would be a good idea to transform this model so that the errors fit our assumptions.

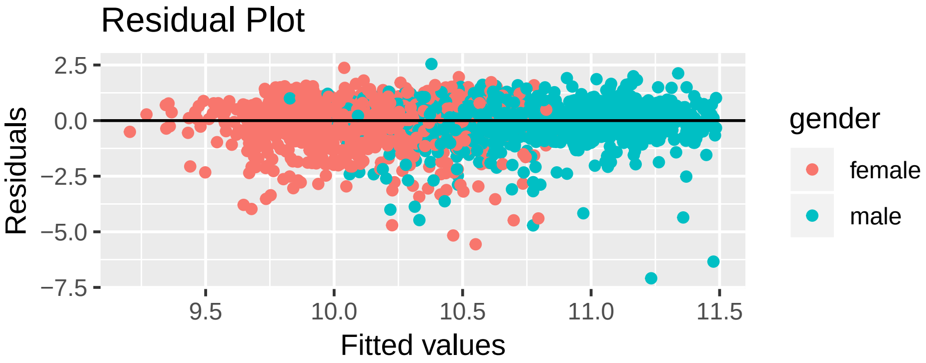
We will use a log transform on the response and recheck the residuals. Our model is now:

$$log(income_i) = \beta_0 + \beta_1 Gender_i + \beta_2 AFQT_i + \beta_3 Education_i$$

After the response transformation and model refit, we now have the following coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.731	0.1026	85.08	0
Gendermale	0.6245	0.03417	18.27	2.986e-70
Education	0.07695	0.008489	9.065	2.403e-19
AFQT	0.005914	0.0007657	7.724	1.602e-14

The R^2 values have not changed an appreciable amount from the original model ($R^2 = .21$), but what we are mainly concerned with are the residuals.



While t-tests on studentized residuals detected 15 outliers, removing them does not appreciably change the estimated coefficents of the model. The residual plot of our log transformed model better fits our assumption of constant variance. The Shapiro-Wilk test still report a small p-value ($< .0001$), indicating that our residuals are still not normal.

However being able to deal with at least one of the error assumptions is better than nothing and the normality of errors is less important than constant variance when the sample size is large. We will now use the coefficient estimates from this model to answer our question.

Conclusion

An advantage of using the log transform is that we can still obtain an interpretation on the original response scale by back transforming using exponentiation.

$$income_i = e^{\beta_0} e^{\beta_1 Gender_i} e^{\beta_2 AFQT_i} e^{\beta_3 Education_i}$$

Since for each individual i, *gender_i* is an indicator variable (0 for female, 1 for male), being male adds a *multiplicative effect* of e^{β_1} on income. From our model, we calculated a statistically significant $\hat{\beta_1} = .6245$ so the estimated effect of gender on income is that being male multiplies median income by 1.867 after accounting for education and AFQT score.