Residuals

Grinnell College

May 7, 2025

Review

Below is a model of chicken weight in days since birth according to one of four separate diets

```
> lm(weight ~ Time + Diet, ChickWeight) %>% summary()

             Estimate Std. Error t value             Pr(>|t|)
(Intercept)   10.924      3.361    3.25               0.0012 **
Time           8.750      0.222   39.45 < 0.0000000000000002 ***
Diet2         16.166      4.086    3.96    0.00008556049098 ***
Diet3         36.499      4.086    8.93 < 0.0000000000000002 ***
Diet4         30.233      4.107    7.36    0.00000000000064 ***

Multiple R-squared:  0.745, Adjusted R-squared:  0.744
F-statistic:  419 on 4 and 573 DF,  p-value: <0.0000000000000002
```

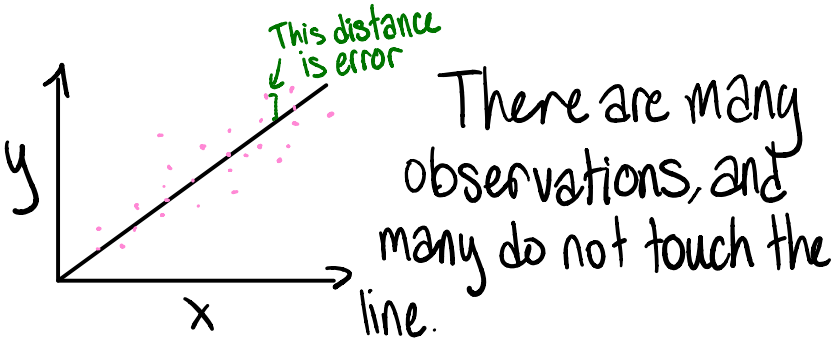
1. Write out the equation for this linear model
2. What proportion of the total variability in chick weight is described by this model?
3. Which diet resulted in the heaviest weight? The lightest?
4. After 10 days, what is the predicted difference in weight between a chicken on Diet 2 and Diet 3?

Today

- Regression is a *model* posits linear relationship between dependent variable y and independent variable X of the form

$$y = \beta_0 + \beta_1 X + \epsilon$$

- Expand this to include combinations of independent variables, both qualitiative and quantitative
- Today our focus is on the error term ϵ (epsilon)



Error Terms

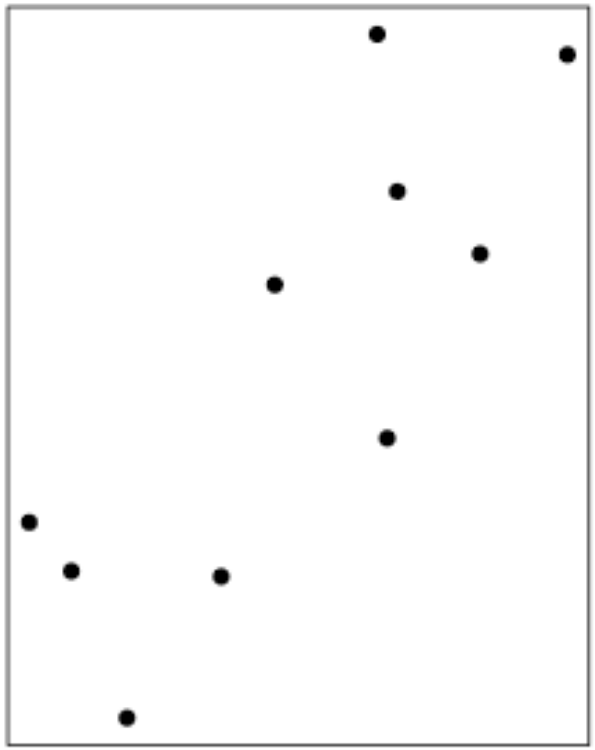
$$y = \beta_0 + \beta_1 X + \epsilon$$

Assumptions:

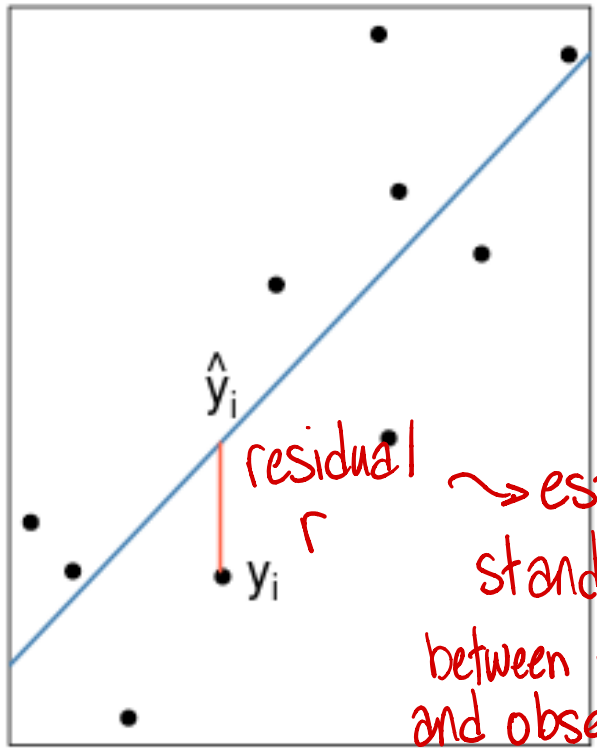
- Linear relationship between X and y
- Error term is normally distributed, $\epsilon \sim N(0, \sigma)$ Assuming residuals are centered around 0
- Variability in error should be the same for all values of X, i.e., error same for all observations

Analyzing the error terms gives us a way to test the assumptions of our model

Collection of (x, y) points



Fitted line with residual



residual
r

→ estimate of standard deviation between predicted and observed data

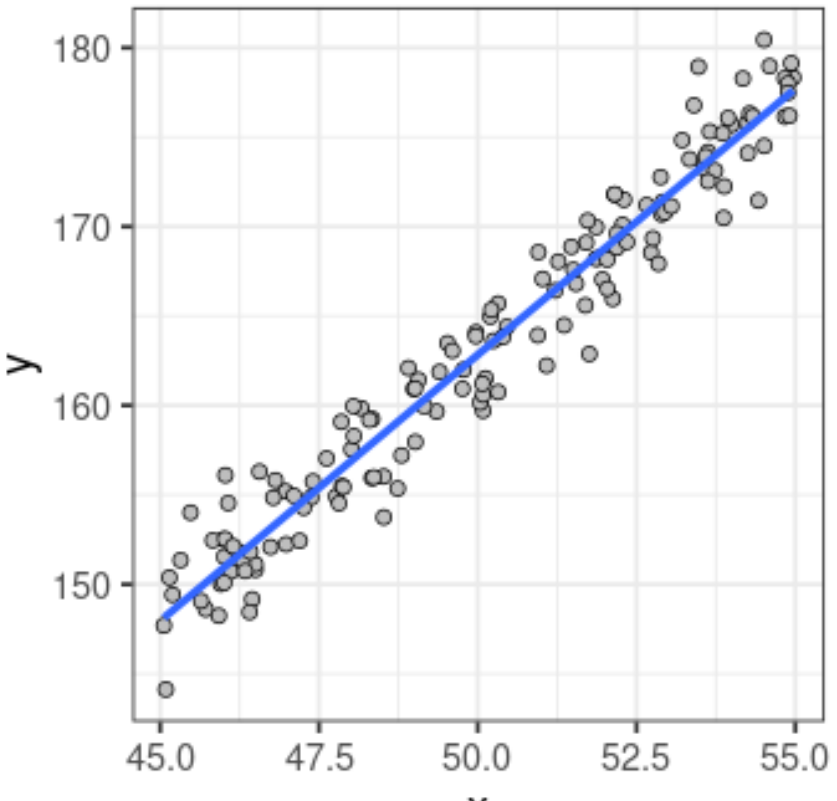
Part 1: Checking Assumptions

Residuals and assumptions

Check how far away each point is from the line.

Three common ways to investigate residuals visually:

1. Plot histogram of residuals (normality)
2. Plot residuals against covariate (linearity, constant variance)
3. Plot residuals against new covariates (pattern identification)

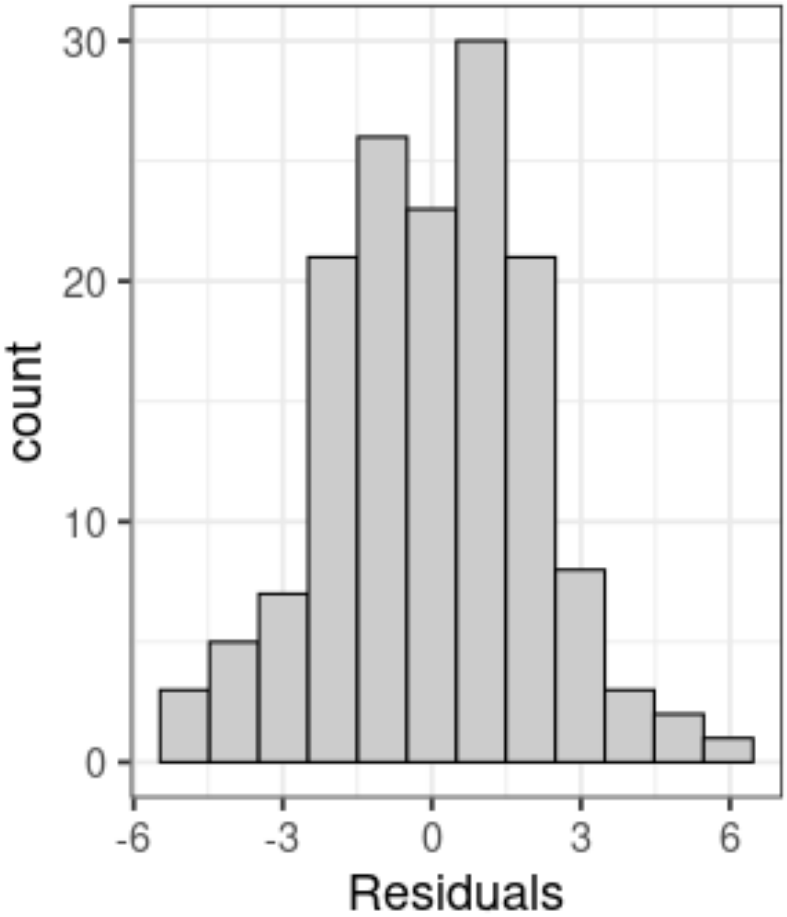


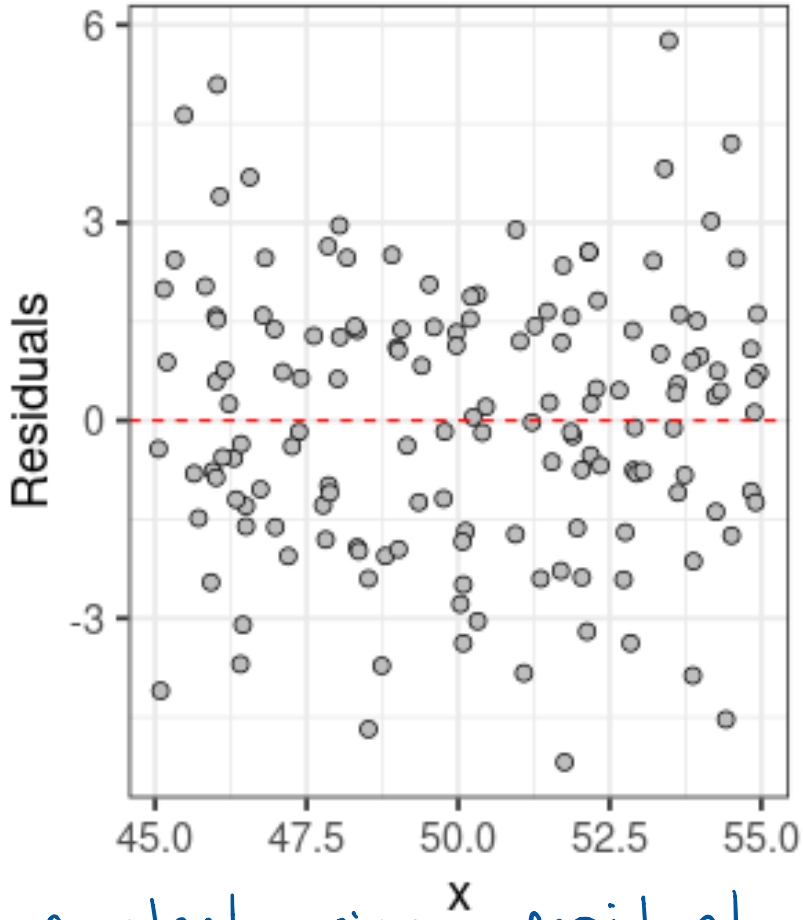
This distance is called residuals

Take the residuals and put them in a bag.

Are my errors normal?

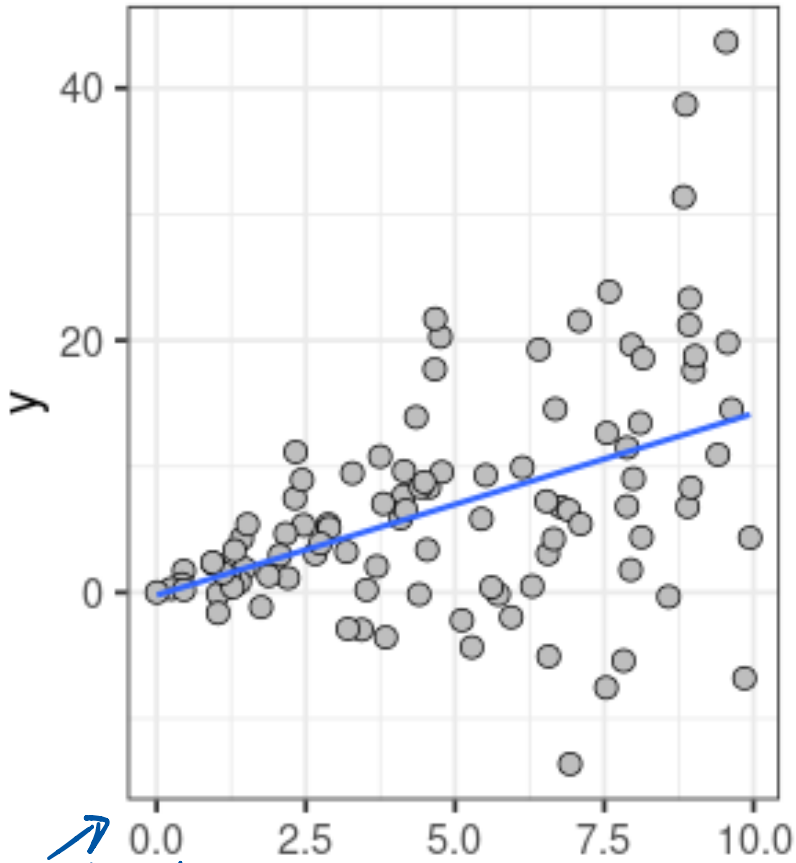
Is the variance the same at all levels of X?



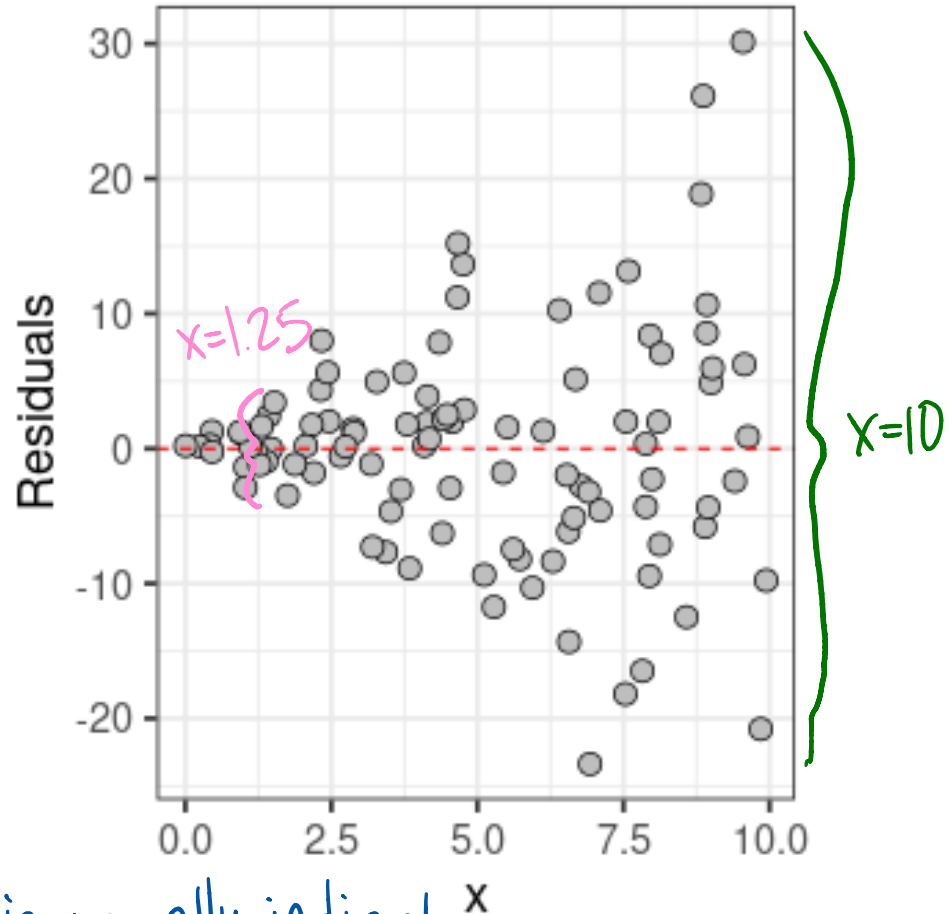


Constant variance: residuals have the same variance regardless of x. No association between x and error terms.

Constant Variance

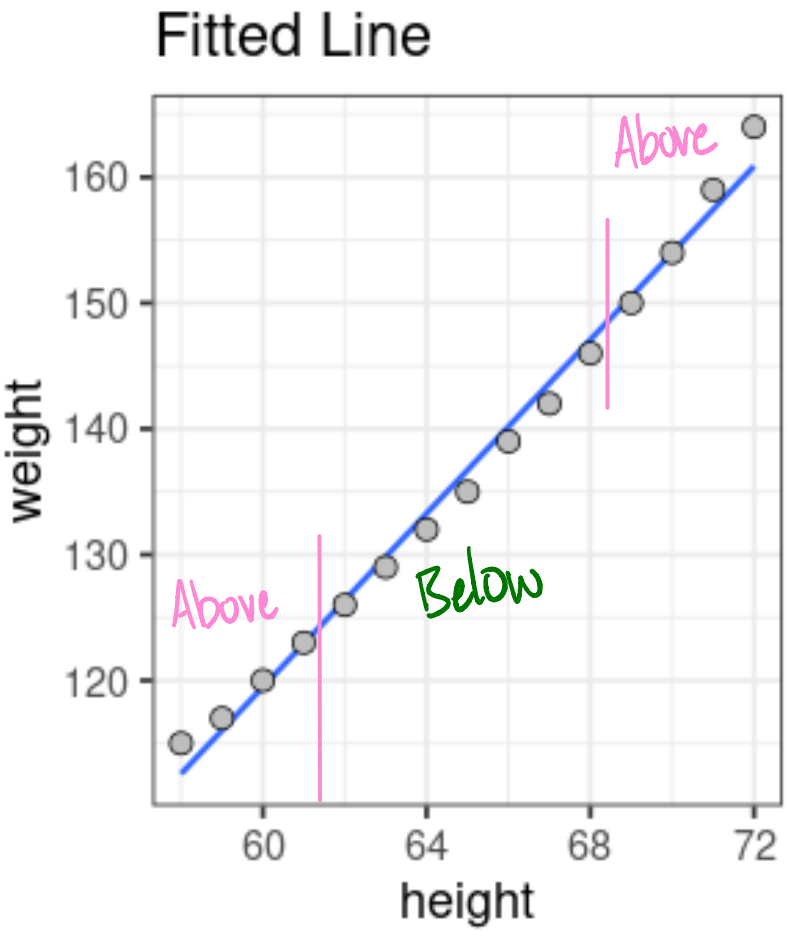


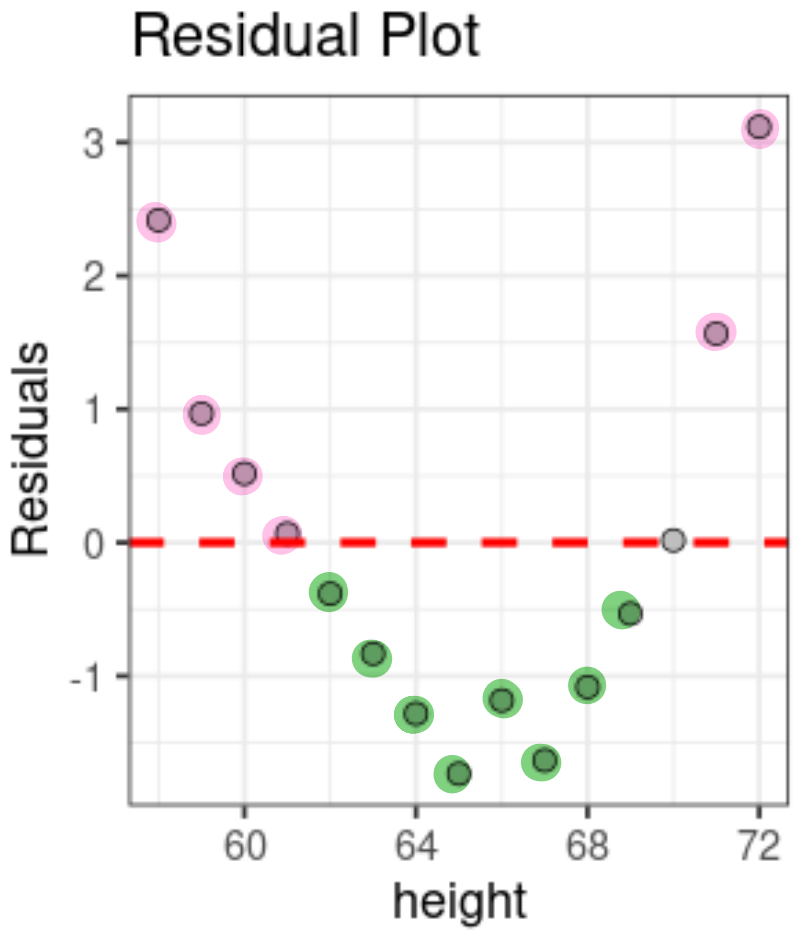
Here the relationship is linear but the variance is increasing with x.



This usually indicates we are missing some covariate associated with time.

Tests of linearity

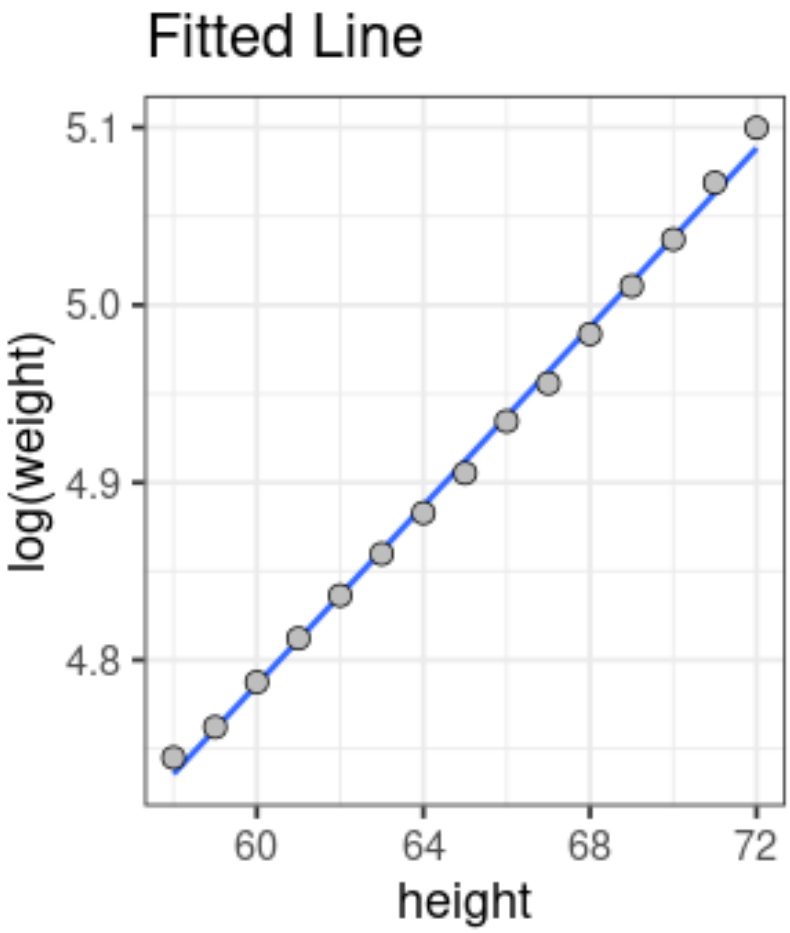


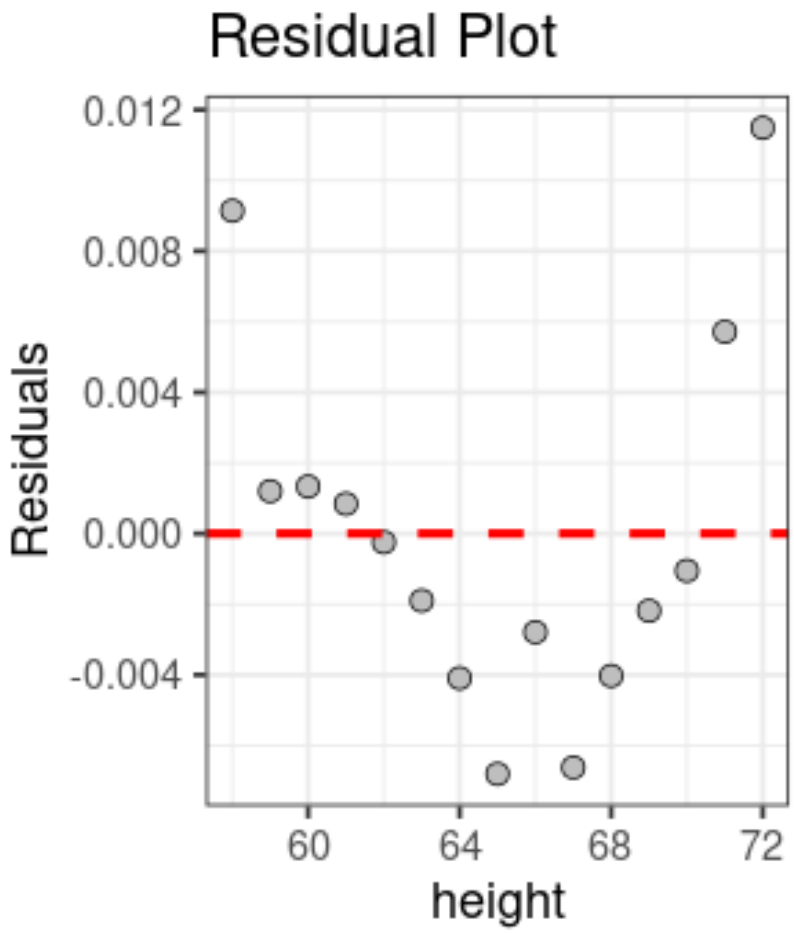


This is not a linear relationship. They may cancel out the the mean but each section of data is off.

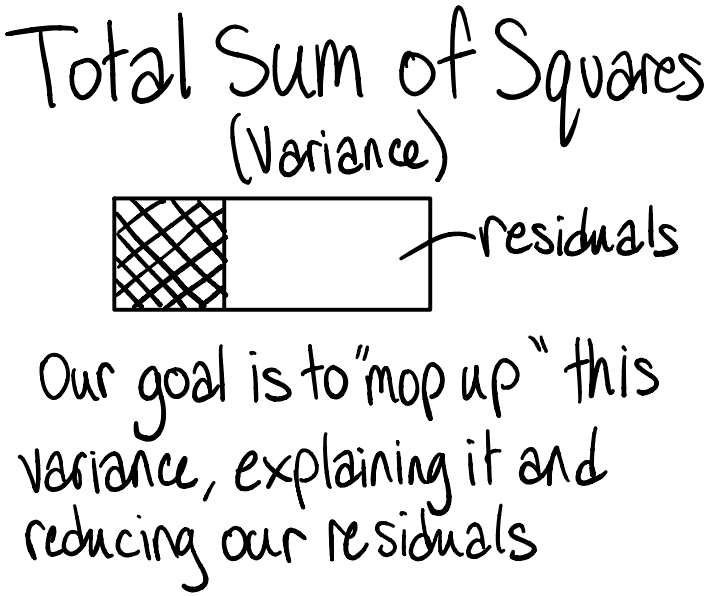
Tests of linearity

Sometimes a transformation of a variable (in this case, log(weight)) can help correct trends





Part 2: Investigating Patterns



Correlated Covariates

Consider a simple linear model in which a covariate X is used to predict some value y

$$\hat{y} = \hat{\beta}_0 + X\hat{\beta}_1$$

The residuals associated with this describe the amount of variability that *is yet to be explained*

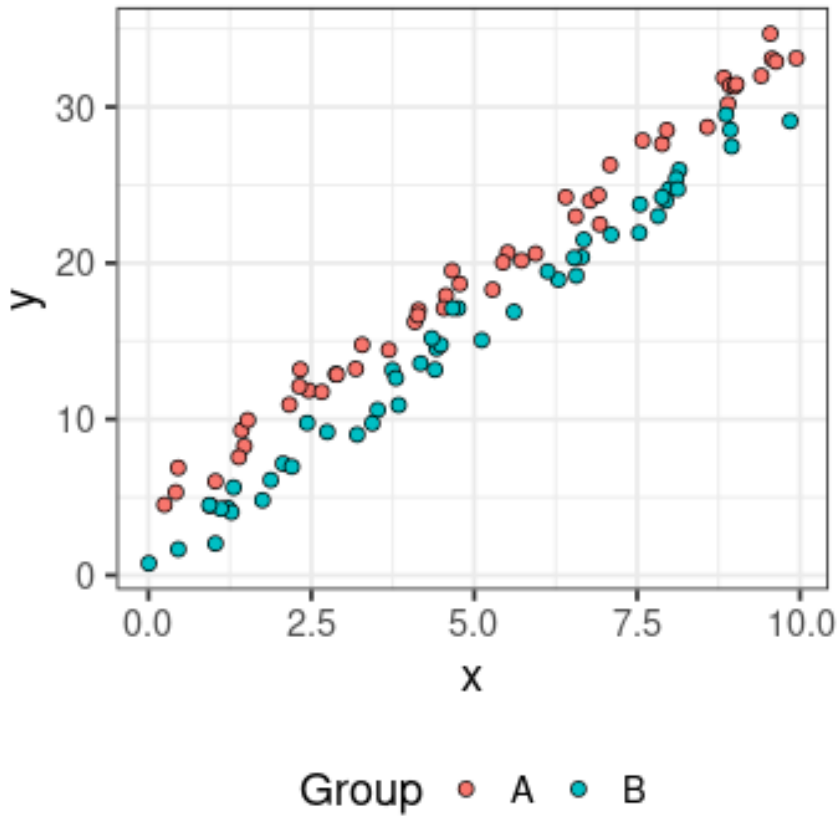
$$r = \hat{y} - y$$

The idea is to find new covariates *associated* with this residual, in effect "mopping up" the remaining uncertainty

Considering new covariates (categorical)

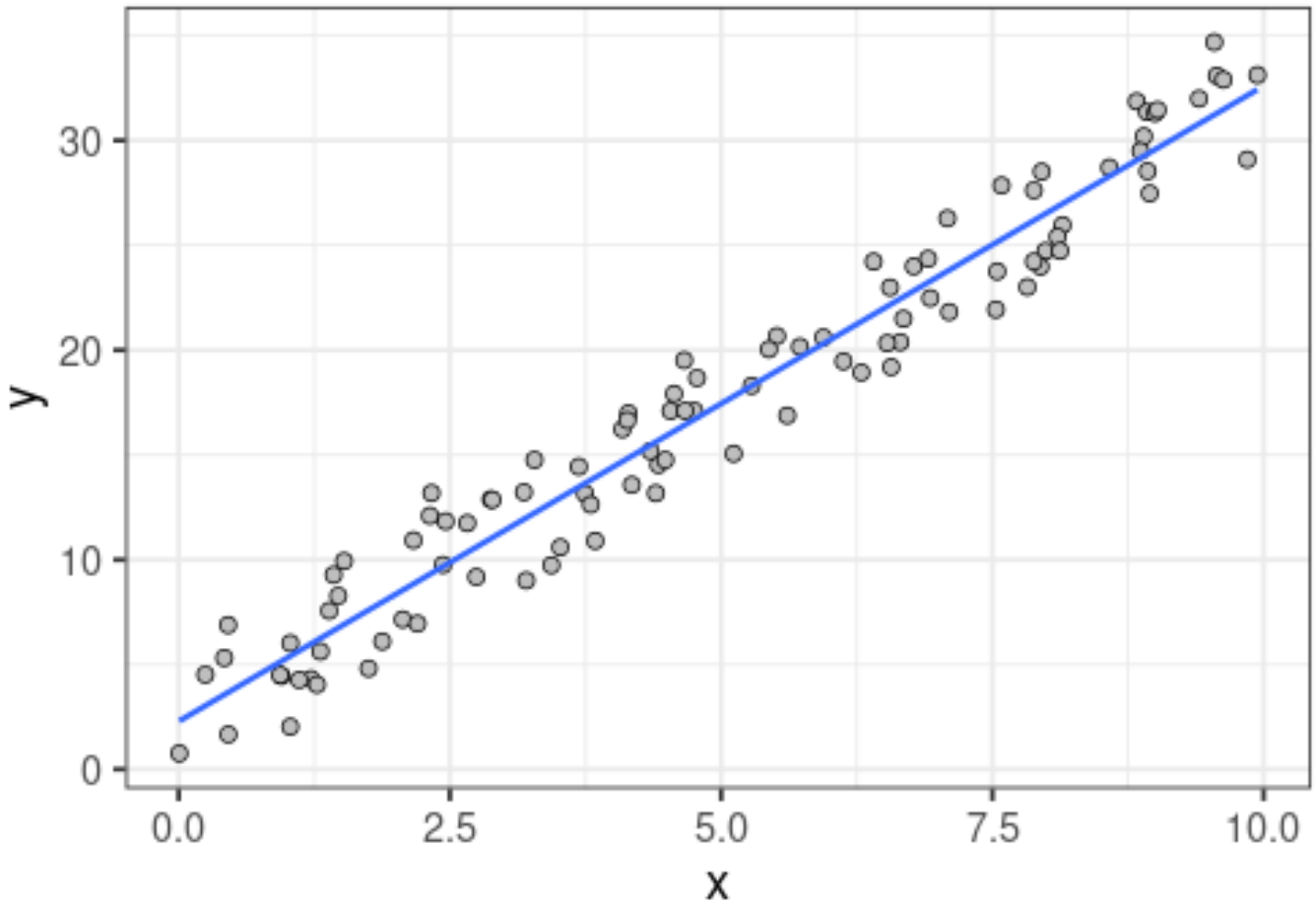
Suppose I have:

- Quantitative outcome y
- Quantitative predictor X
- Categorical predictor indicating group

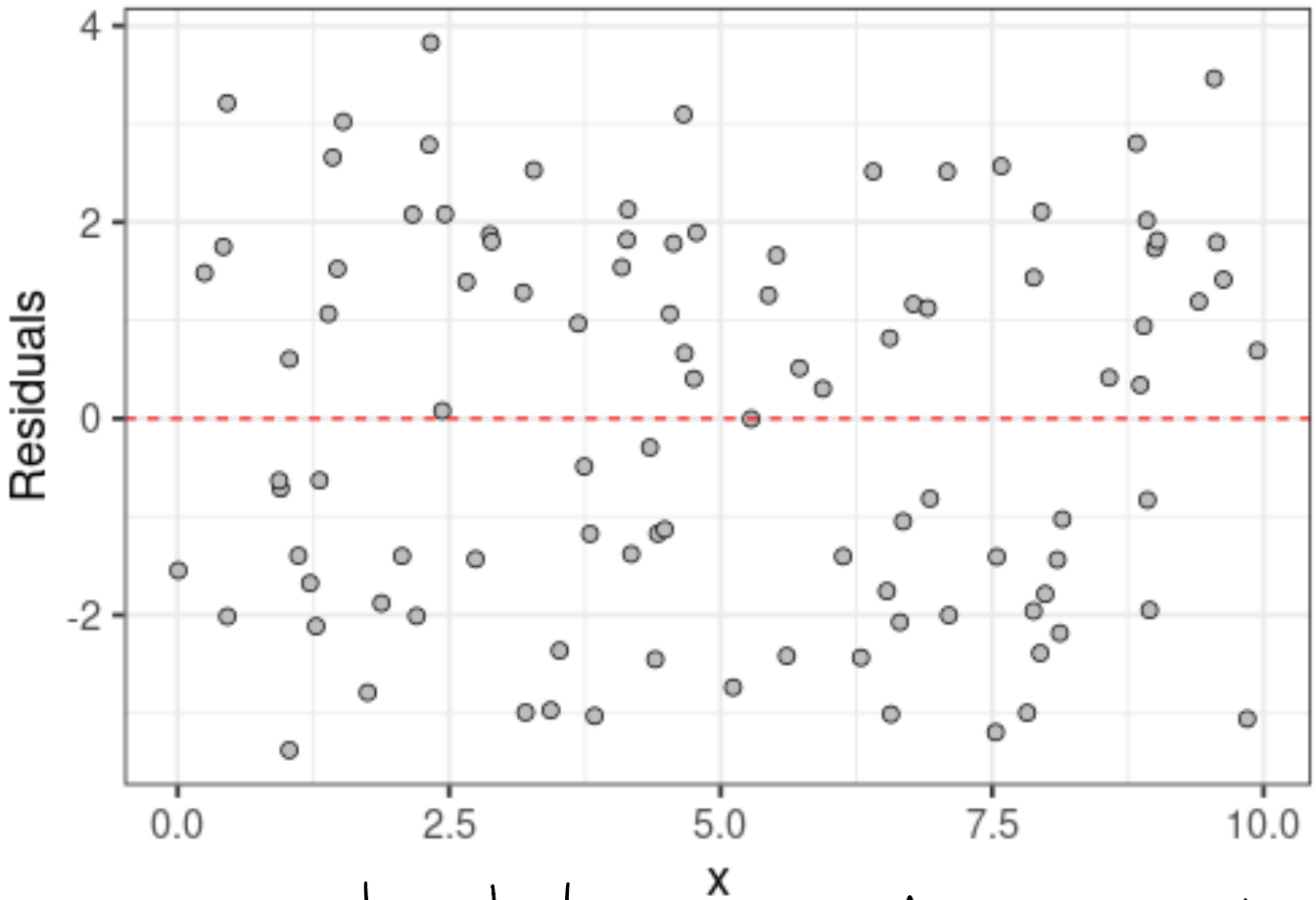


Considering new covariates (categorical)

$$\hat{y} = \hat{\beta}_0 + X\hat{\beta}_1$$

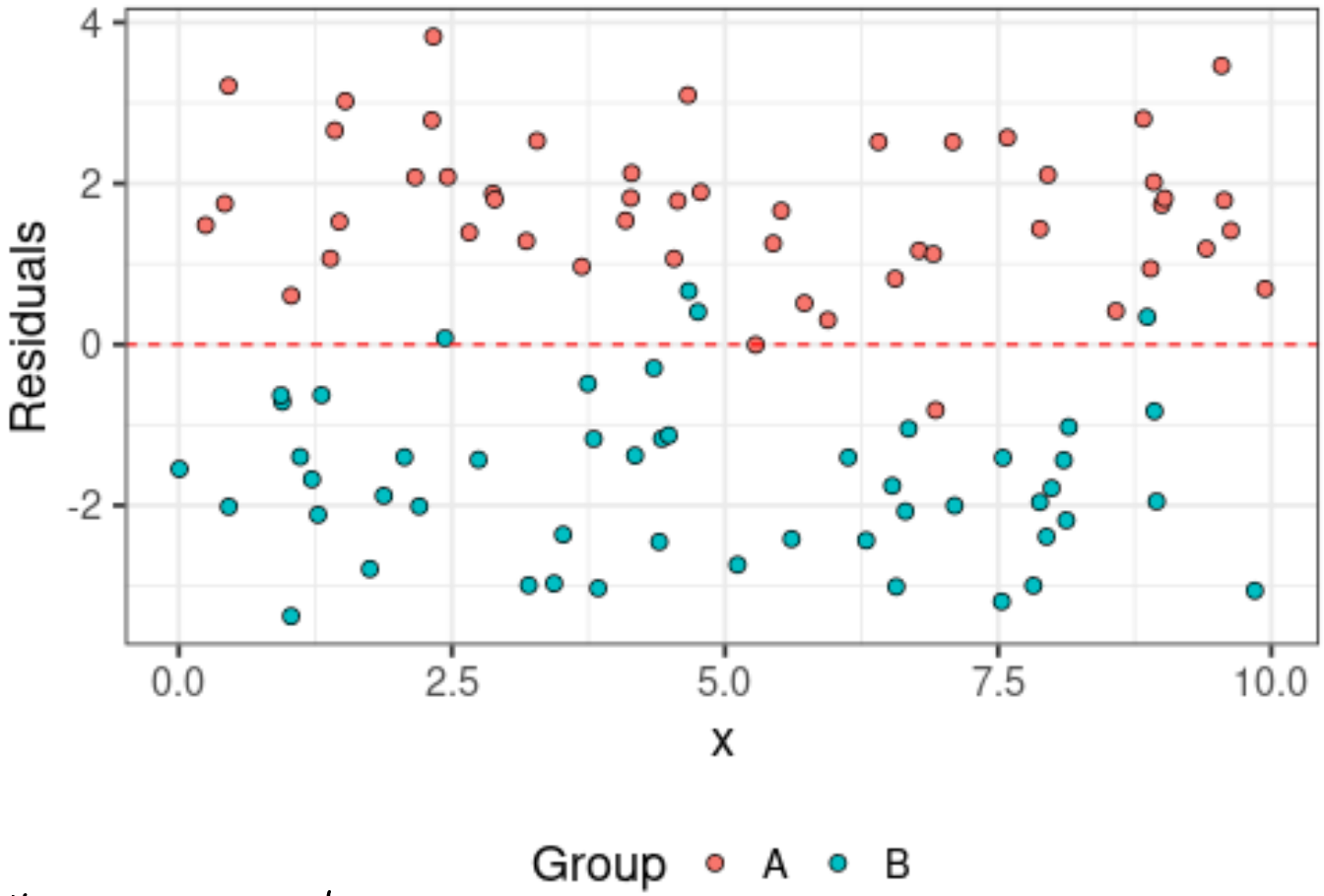


Considering new covariates (categorical)



Looks like no association between x and residuals, but we are still ignoring the categorical variable

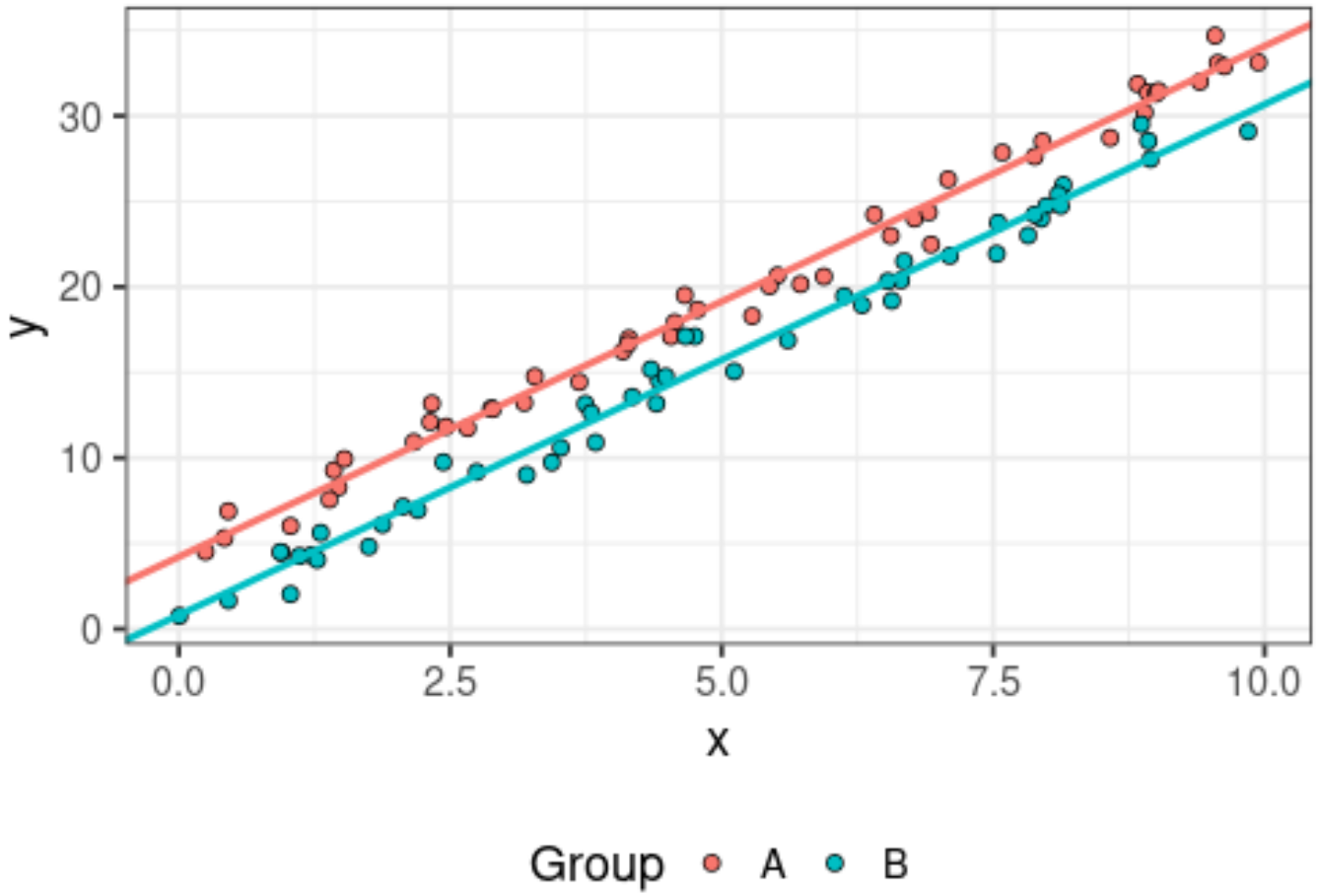
Considering new covariates (categorical)



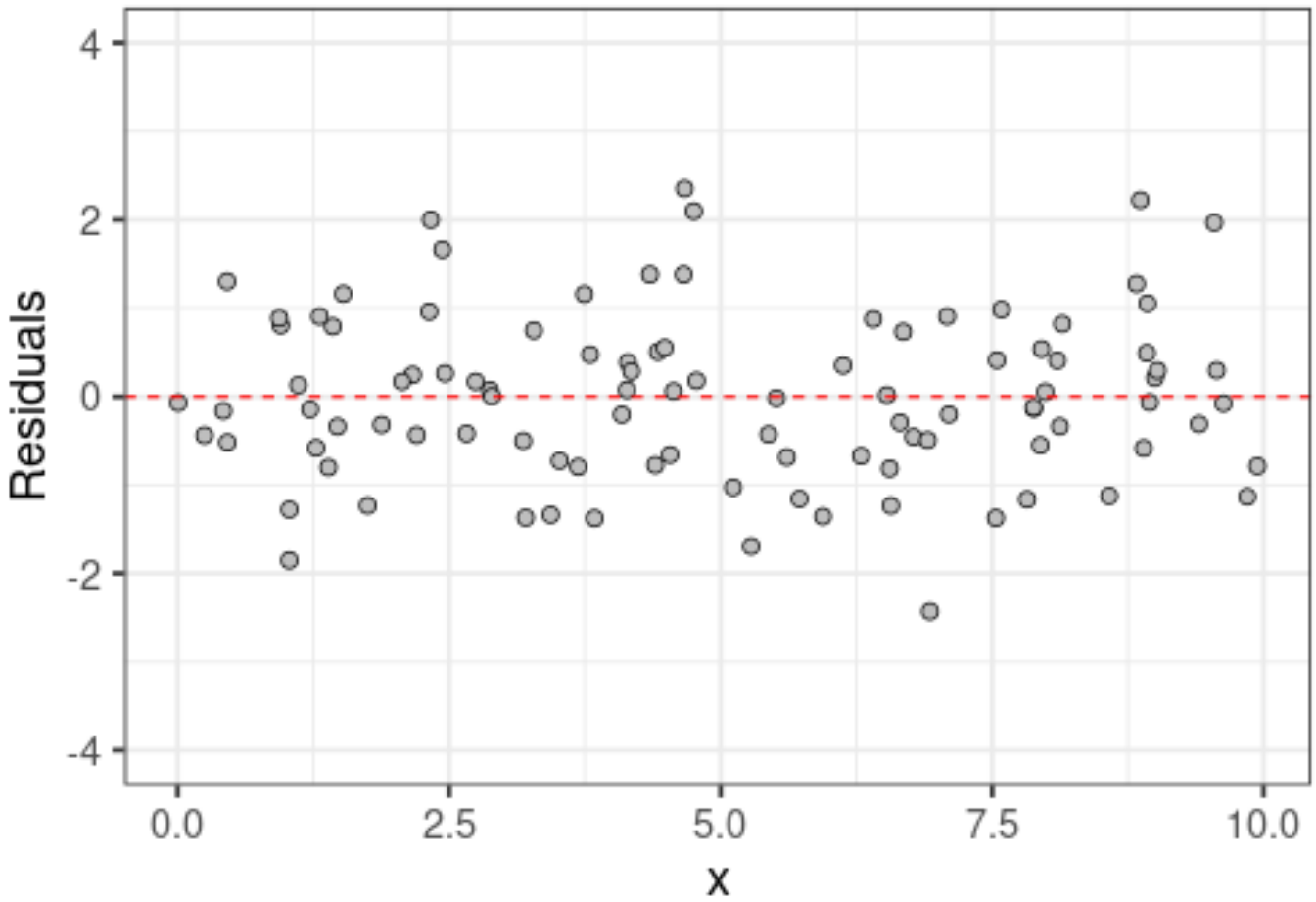
We see there is a clear association between a missing variable and the residuals. So we add that variable!

Considering new covariates (categorical)

$$\hat{y} = \hat{\beta}_0 + X\hat{\beta}_1 + \mathbb{1}_A\hat{\beta}_2$$

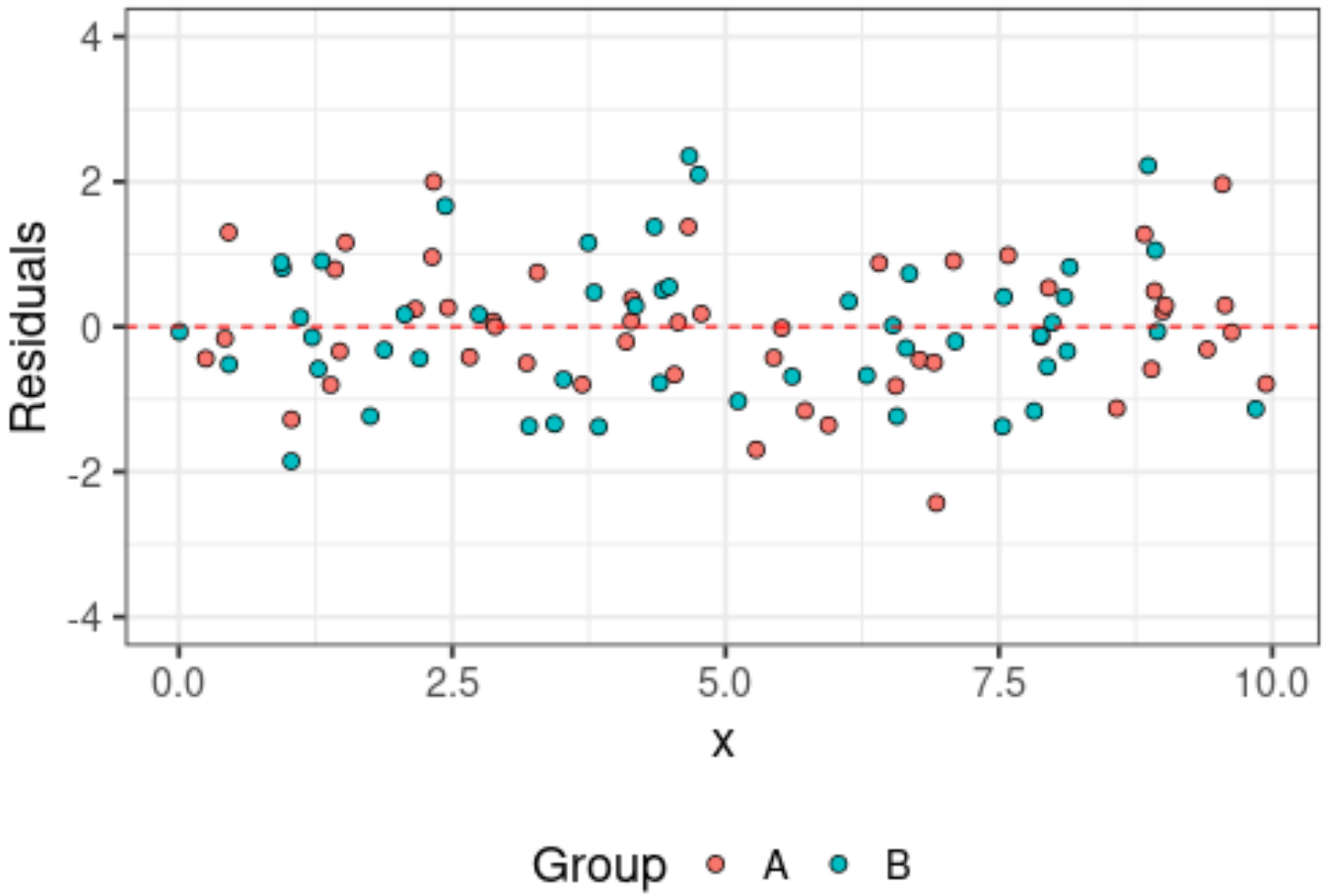


Considering new covariates (categorical)



The range of our residuals really shrunk! This is good!

Considering new covariates



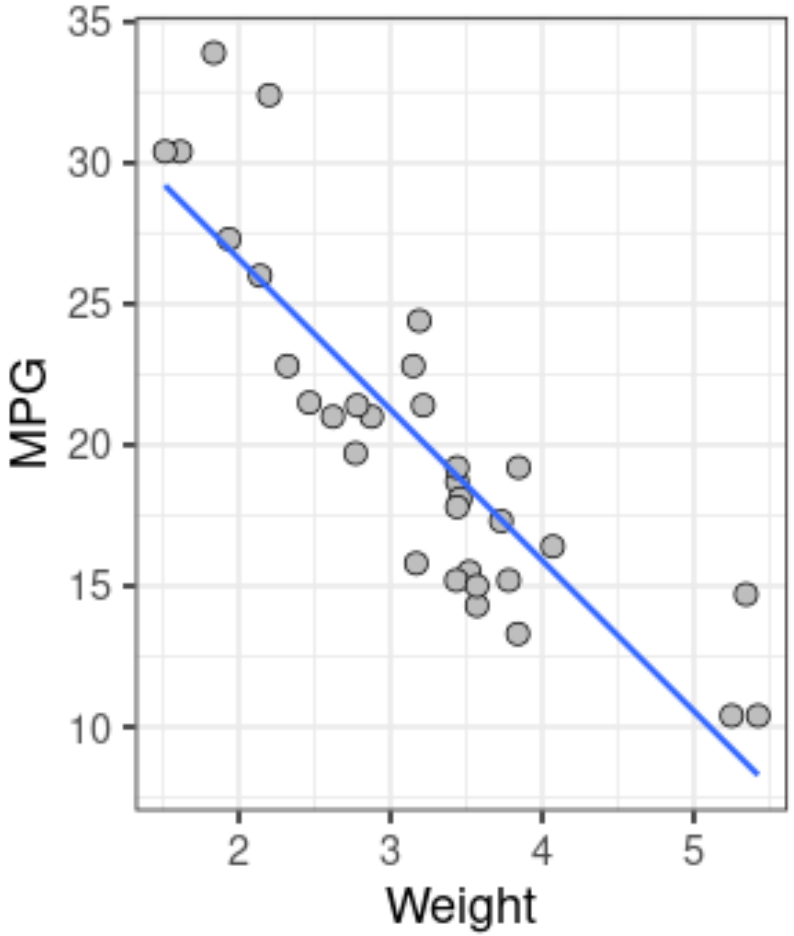
Considering new covariates (quantiative)

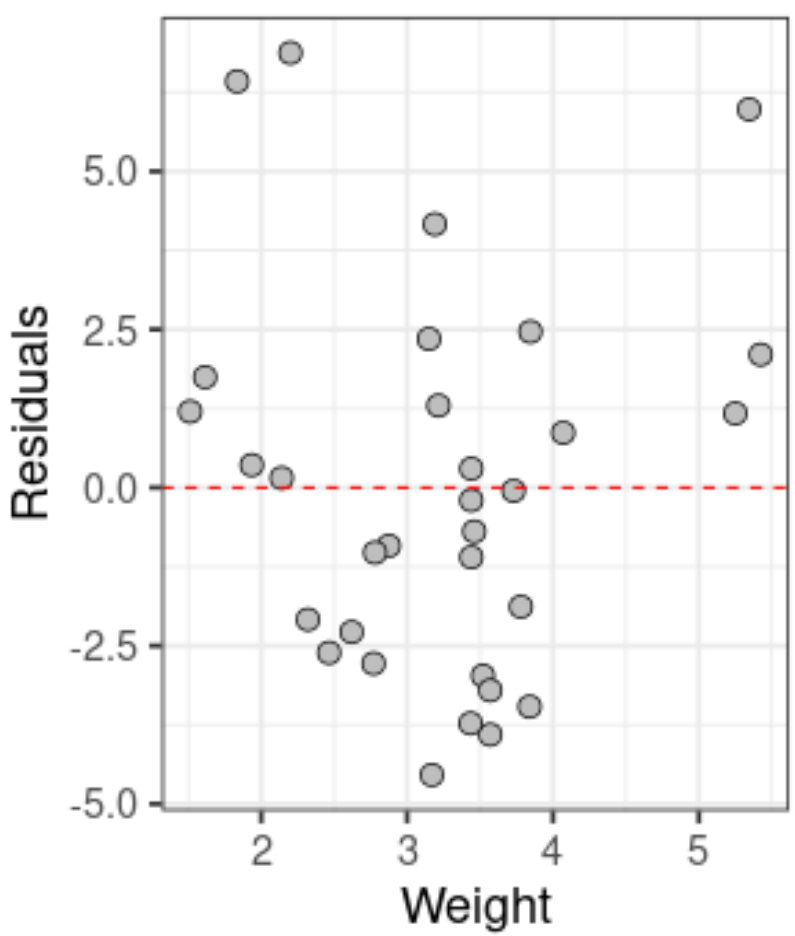
Now consider a situation in which we wish to predict fuel economy with three separate models:

1. Using weight
2. Using weight and engine displacement
3. Using weight and quarter mile time

Starting with the first model, we can consider the relationship of the residuals with both engine displacement and quarter mile time

Considering new covariates (quantiative)





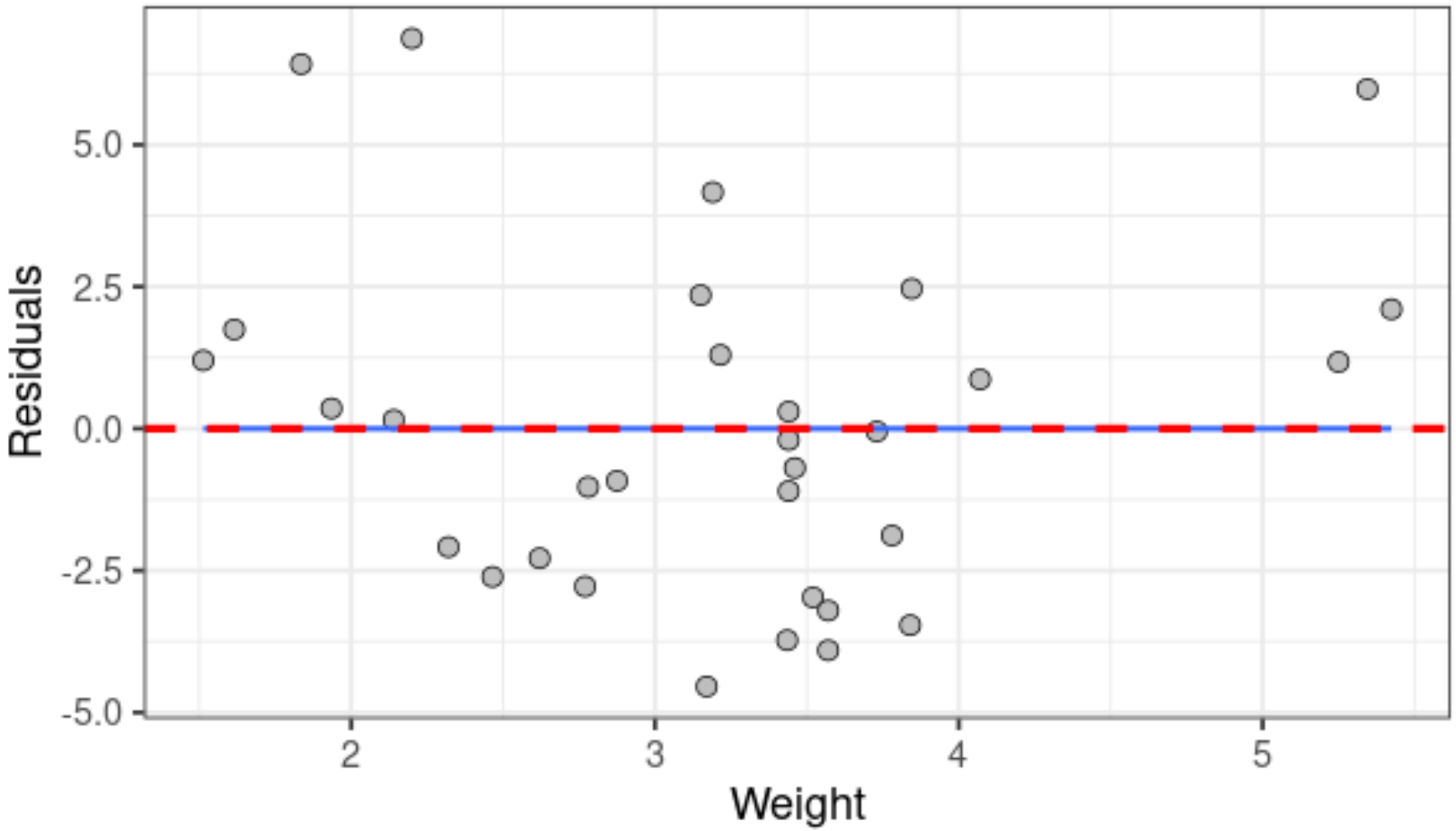
This model is decent, but we want to explain some of the leftover residual. What variable do we add?

Considering new covariates (quantiative)

An interesting thing occurs when we try to create a regression model of the residuals with the original variable:

$$\text{residuals} = \hat{\beta}_0 + \hat{\beta}_1 \text{Weight} = 0$$

Skip



Considering new covariates (quantiative)

skip

When considering adding new variables to our regression model, we want to add those that will "mop up" the residuals that are left after considering weight

This brings us to the idea of **correlated variables**, or variables that have evidence of a *linear relationship* with one another

How correlated our variables may be will impact how much of the residual they are able to account for

Correlated Covariates

skip

We can consider two extremes: if two quantitative variables are perfectly correlated, knowing the value of one variable means we also know the value of the other

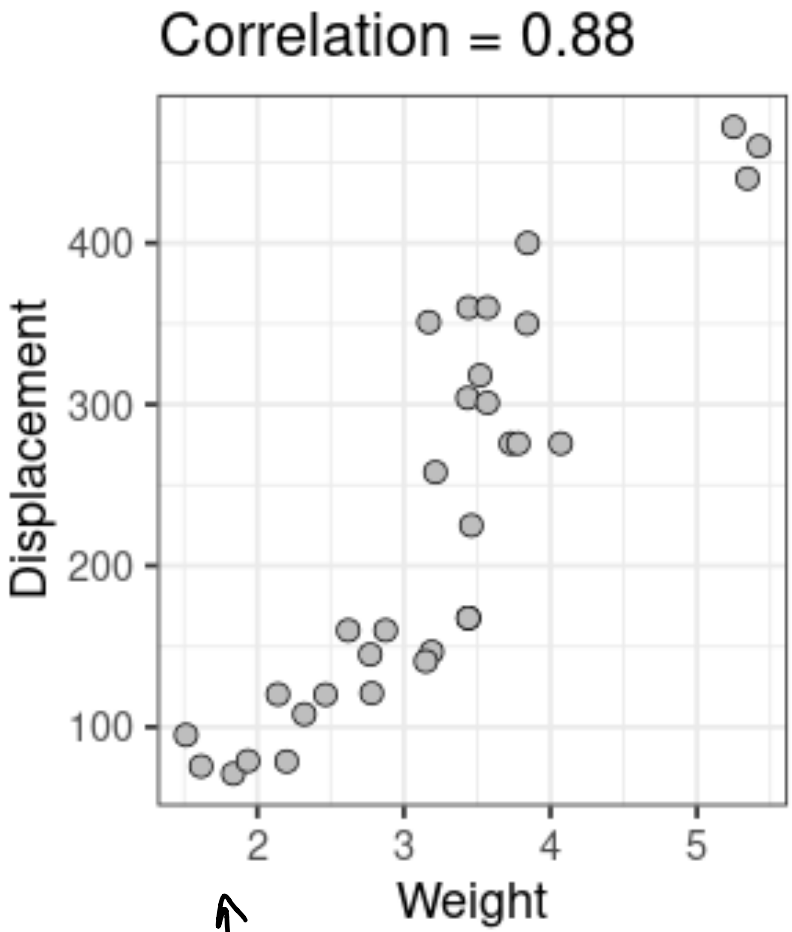
This means that, in terms of predicting an outcome, adding a highly correlated variable to our model will contribute little new information and will not be very useful

By constrast, if two variables have perfectly uncorrelated, then knowing the value of one tells us nothing about the value of another

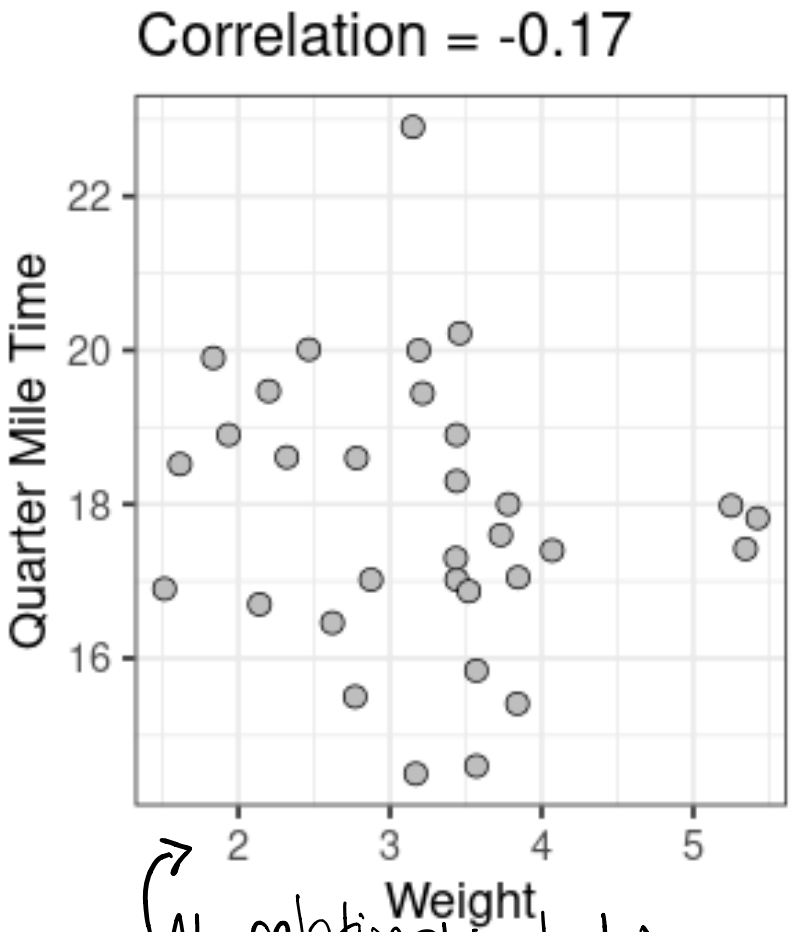
Adding an uncorrelated variable to our model thus offers more potential to "mop up" the variability that was not explained by the first variable

Correlated Covariates

multicolinearity: two or more independent variables are correlated and it causes problems because the share information



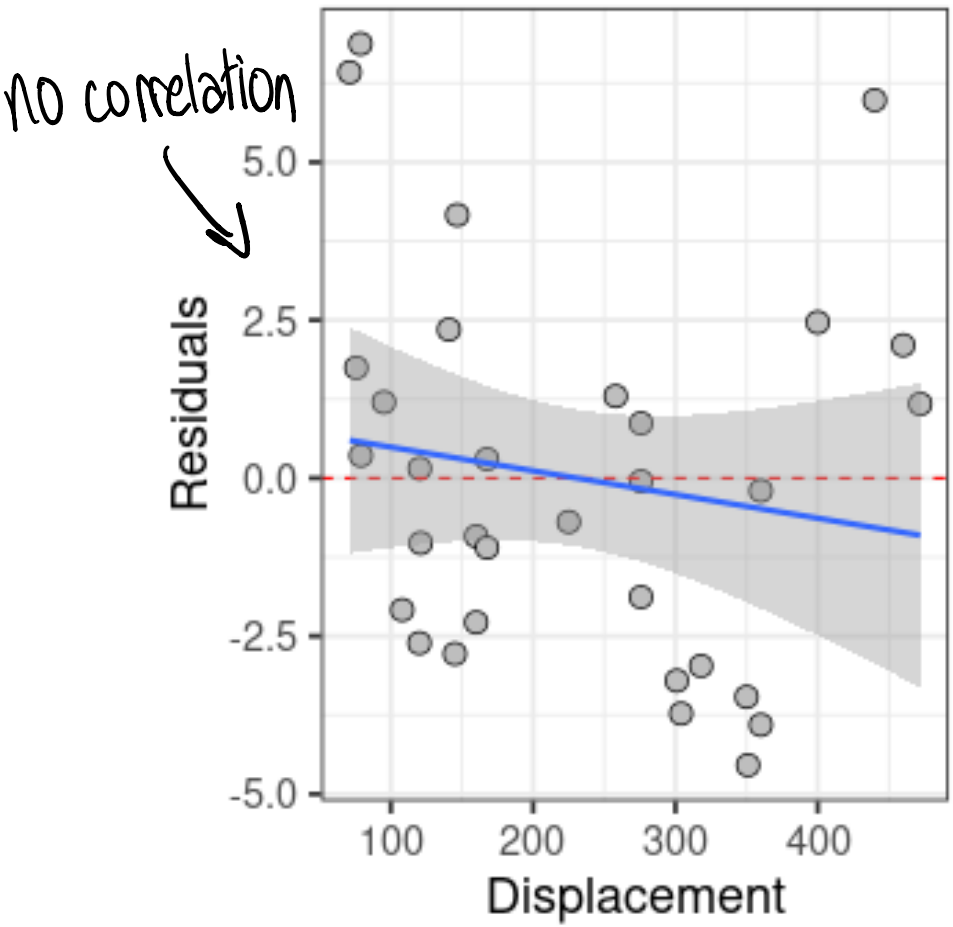
Clear relationship between weight and displacement, adding displacement does not add much new info.

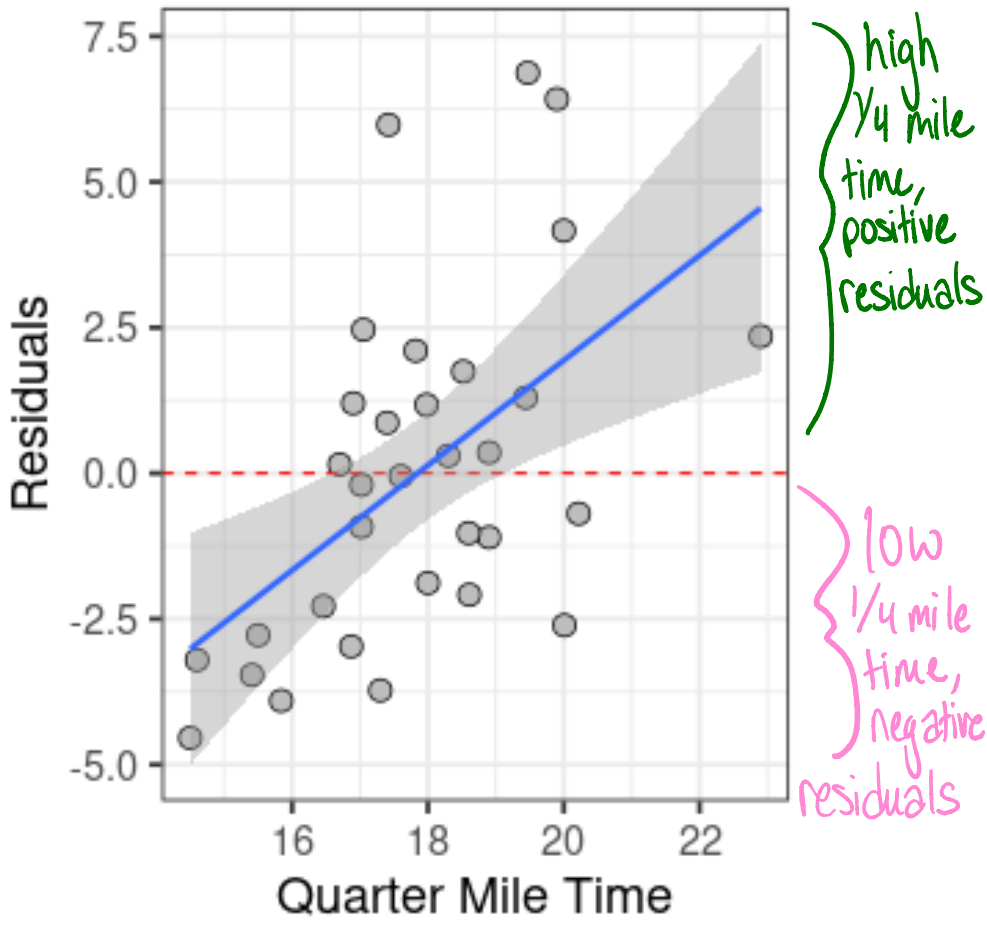


No relationship between quarter mile time and weight so it adds new info to our model

Residual Plots

Displacement has little association with the residuals, while quarter mile time has quite a bit of association. This suggests that adding QM time will "mop up" more unexplained variance





Correlated Covariates

```
> lm(mpg ~ wt, mtcars) %>% summary()

            Estimate Std. Error t value     Pr(>|t|)
(Intercept)   37.285      1.878   19.86    < 0.000002 ***
wt            -5.344      0.559   -9.56      0.000013 ***
R-squared = 0.75

> lm(mpg ~ wt + disp, mtcars) %>% summary()

            Estimate Std. Error t value      Pr(>|t|)
(Intercept) 34.96055    2.16454   16.15  0.000000049 ***
wt          -3.35083    1.16413   -2.8         0.0074 **
disp        -0.01772    0.00919   -1.93        0.0636 .
R-squared = 0.78

> lm(mpg ~ wt + qsec, mtcars) %>% summary()

            Estimate Std. Error t value       Pr(>|t|)
(Intercept)   19.746      5.252    3.76        0.00077 ***
wt            -5.048      0.484  -10.43 0.000000000025 ***
qsec           0.929      0.265    3.51        0.00150 **
R-squared = 0.82
```

adding displacement changed our weight estimate and increased sd

Adding ¼ mile time barely changed our weight estimate and decrease sd

Key Takeaways

1. Number of assumptions for linear model
   - Linearity
   - Normal errors
   - Constant Variance
2. Need way to determine which new variables to add to model
3. Examining errors effective way to test assumptions and investigate new covariates
4. Relationship between correlation of predictors and residual analysis