Regression Error

Grinnell College

December 9, 2024

Review

- Regression models a linear relationship between response variable y and explanatory variable X of the form

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

- Can expand this to include combinations of explanatory variables (quant. and cat.)

Error Terms

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

Assumptions:

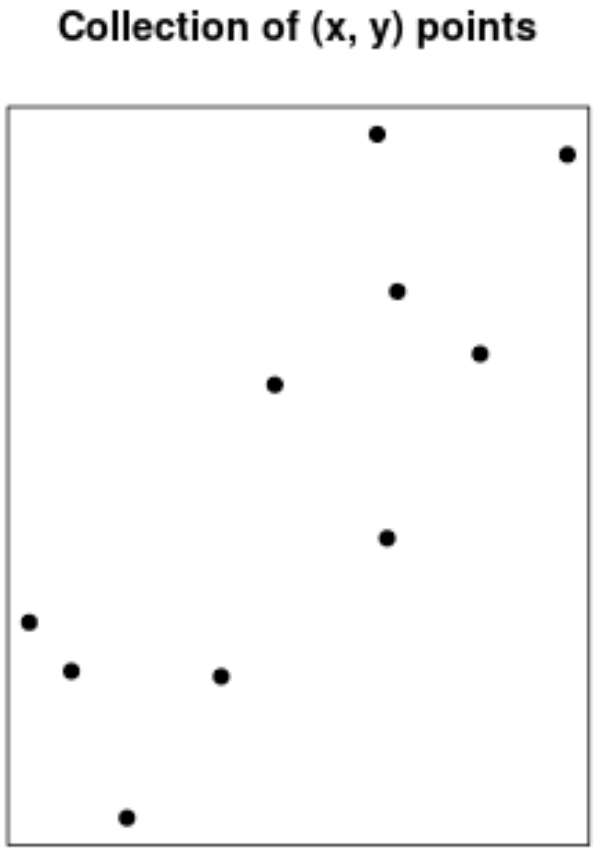
- Linear relationship between X and y
- Error term is normally distributed, $\epsilon \sim N(0, \sigma)$
- Error should be the same for all values of X, i.e., error same for all observations

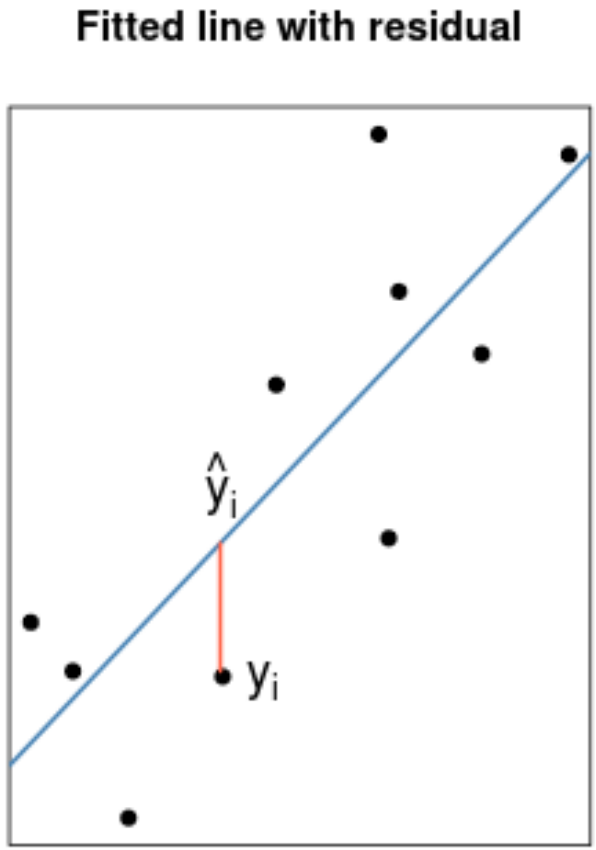
Analyzing the error terms gives us a way to test the assumptions of our model

Residuals

Visually, let's review what residuals look like

- residuals represent how far off our prediction is



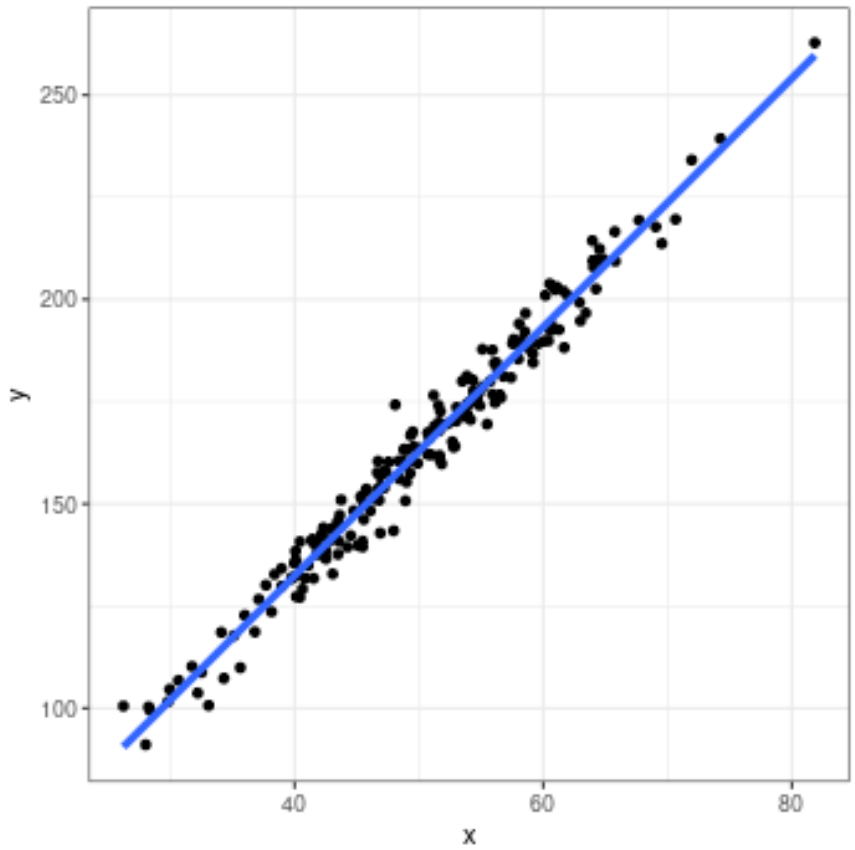


Part 1: Checking Assumptions

Residuals and assumptions

Three common ways to investigate residuals visually:

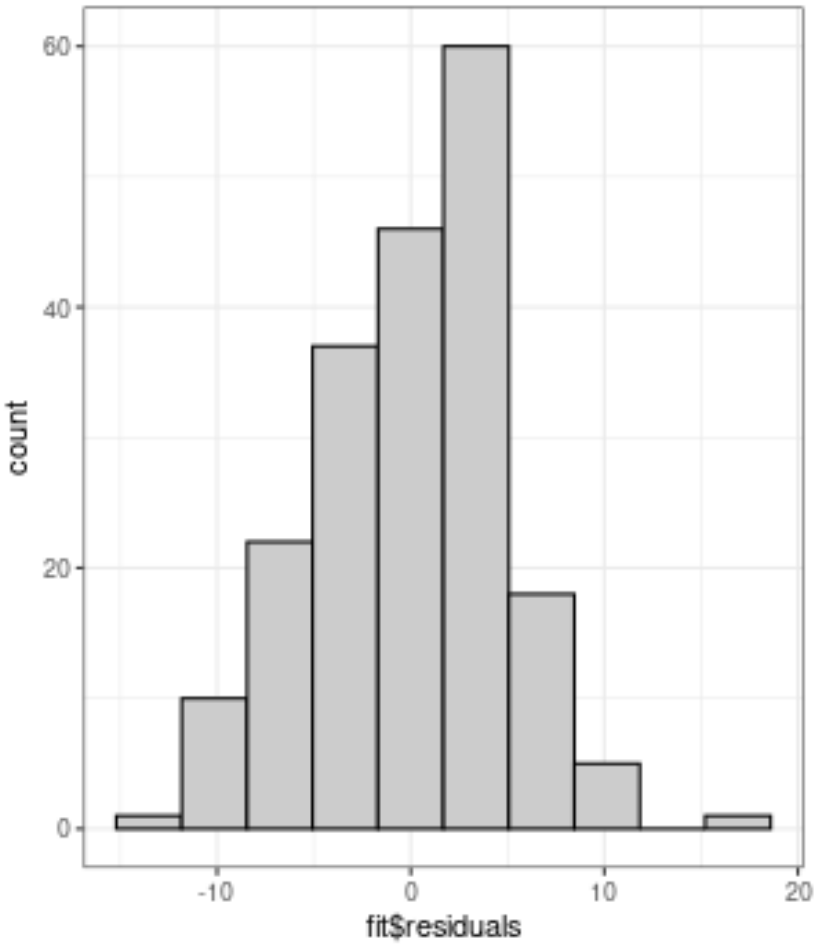
1. Plot histogram of residuals (normality)
2. Plot residuals against covariate (linear trend, changing variance)
3. Plot residuals against new covariates (pattern identification)

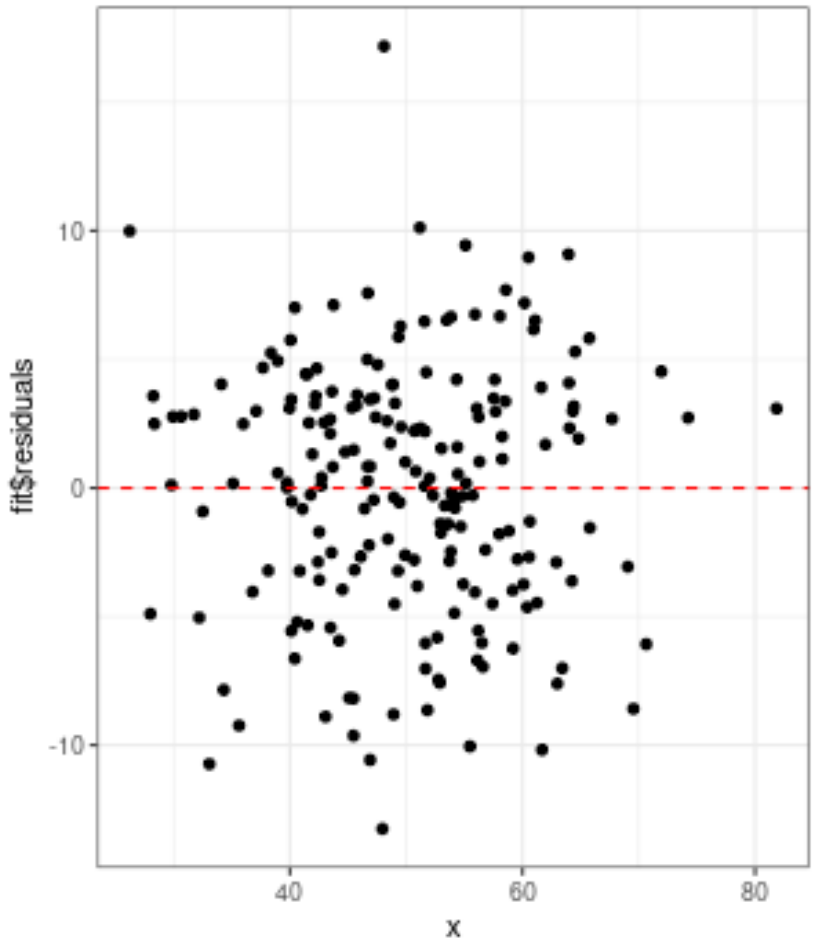


Checking Normality

Histogram of Residuals should be ≈Normal if our model is doing well

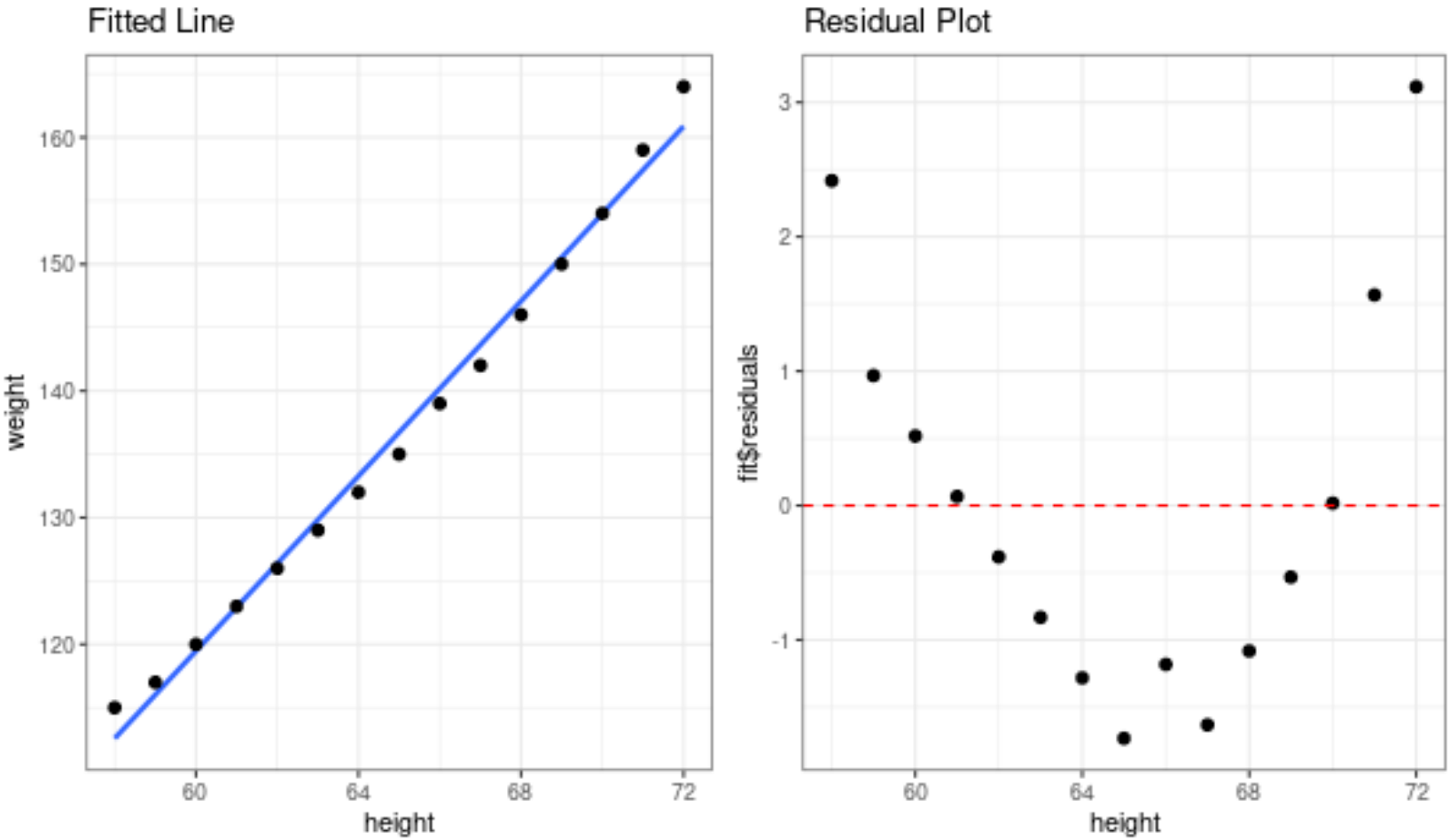
Residuals should not have a pattern other than 'blob of points' in a Resid. vs. Expl. Var. plot





Tests of linearity

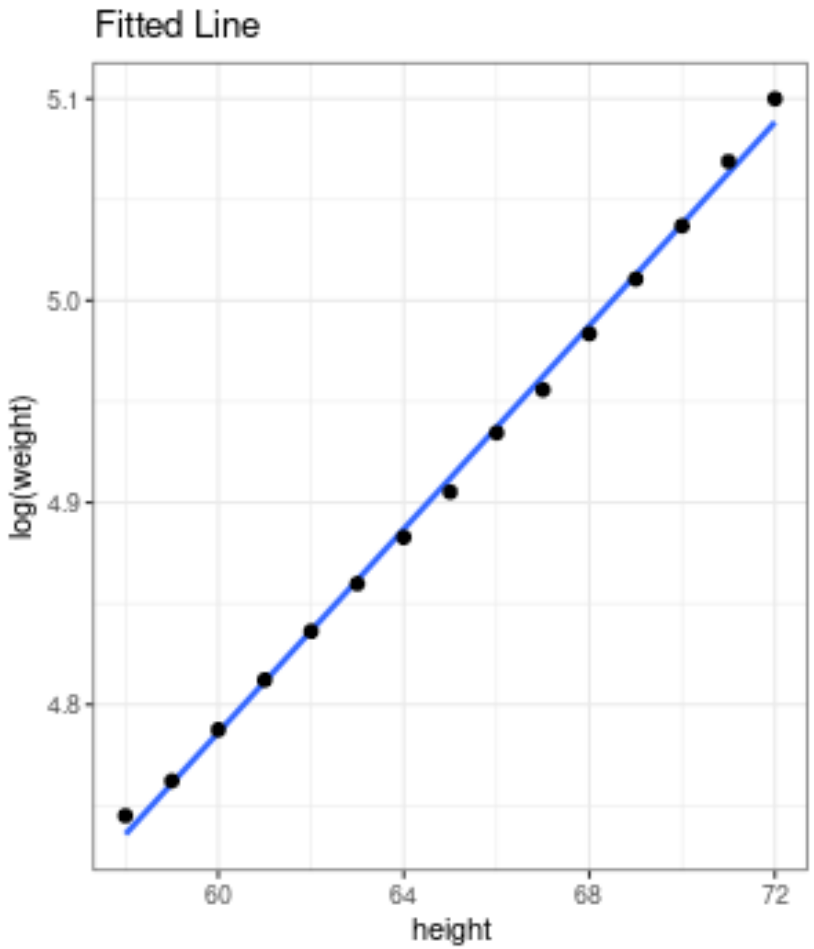
Residual vs. Explanatory plot makes seeing non-linearity easier

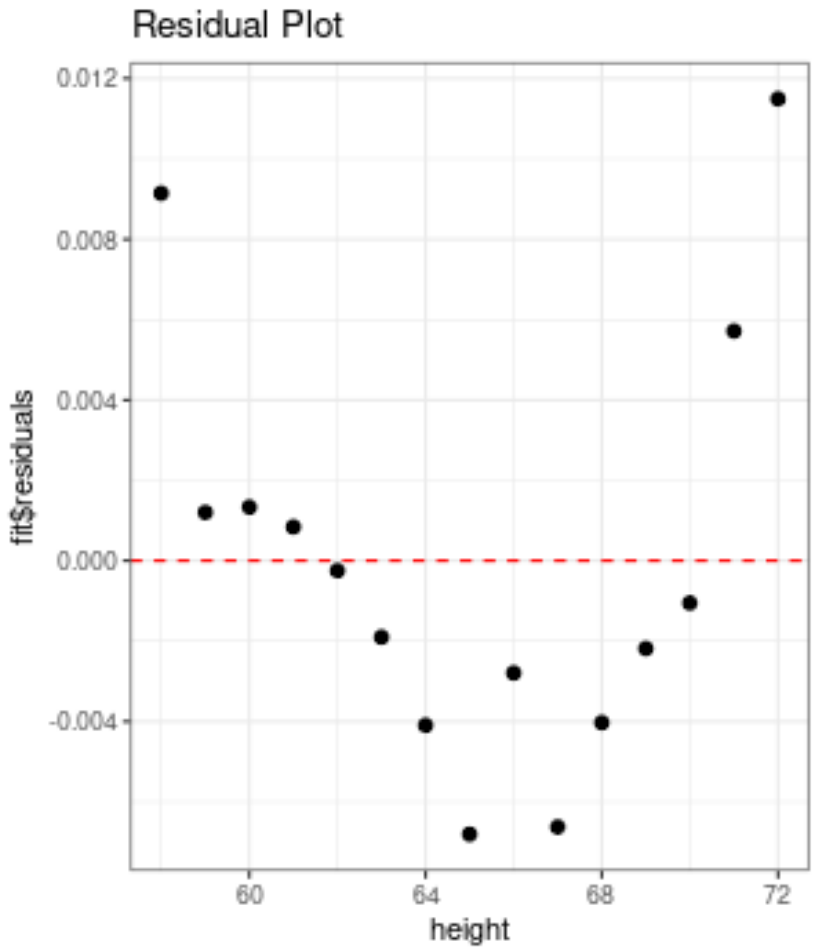


- linear regression could still be useful!
- but we could also look at doing something more complicated if we really cared

Tests of linearity

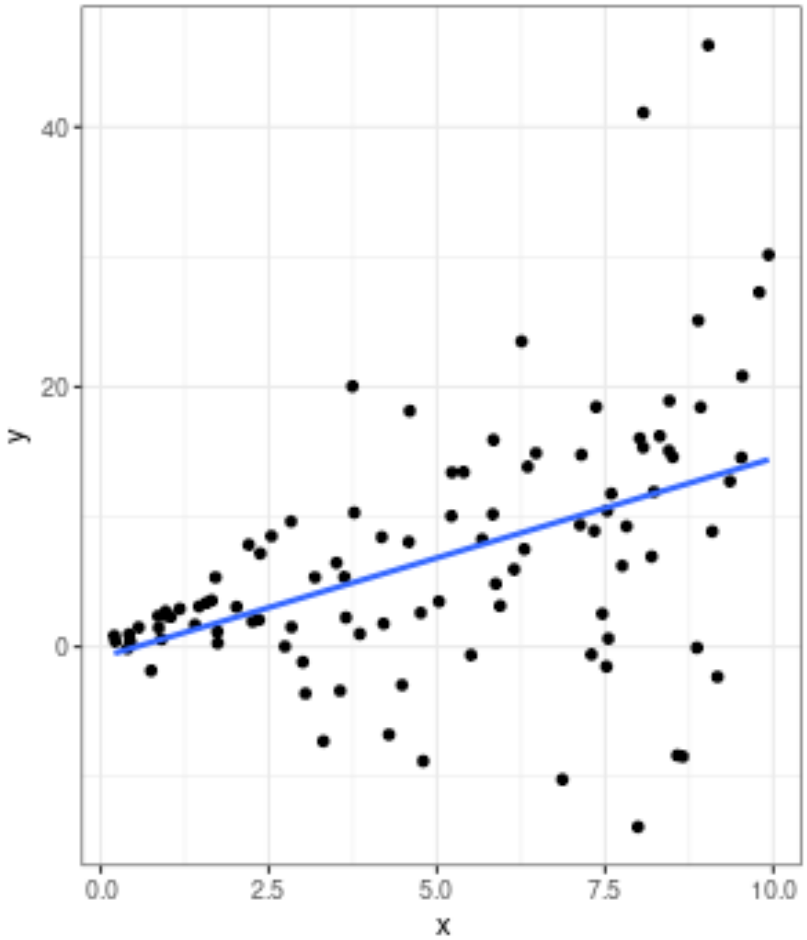
Sometimes a transformation of a variable can help correct trends (log(weight))

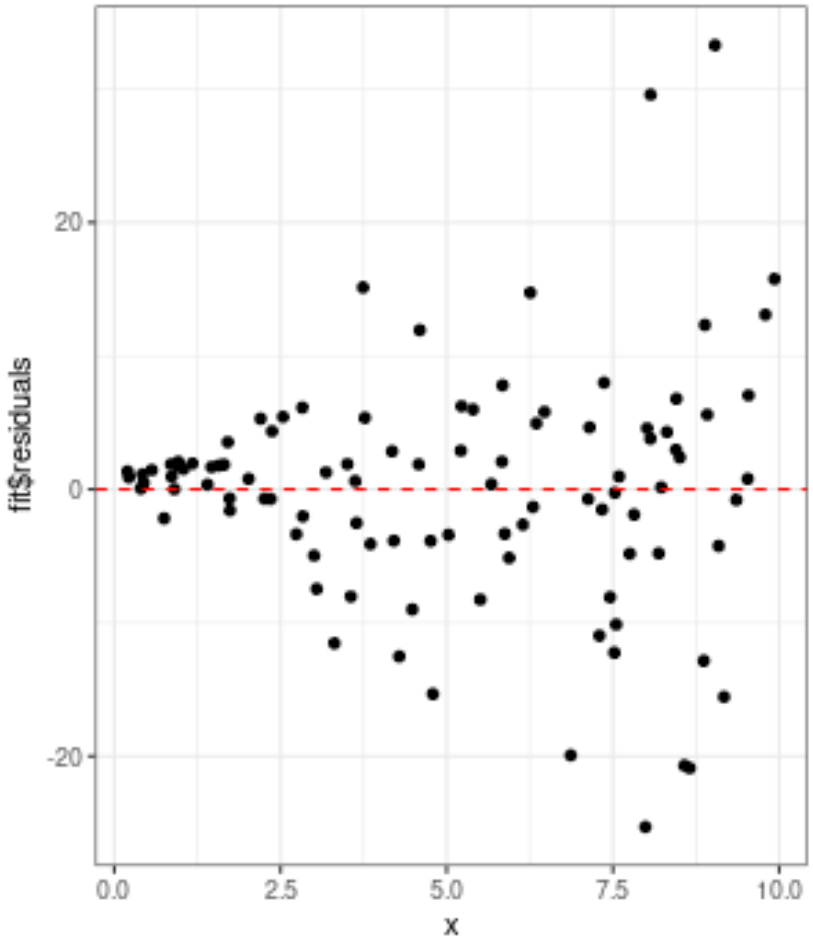




Heteroscedasticity

Hetero = different, scedastic = random



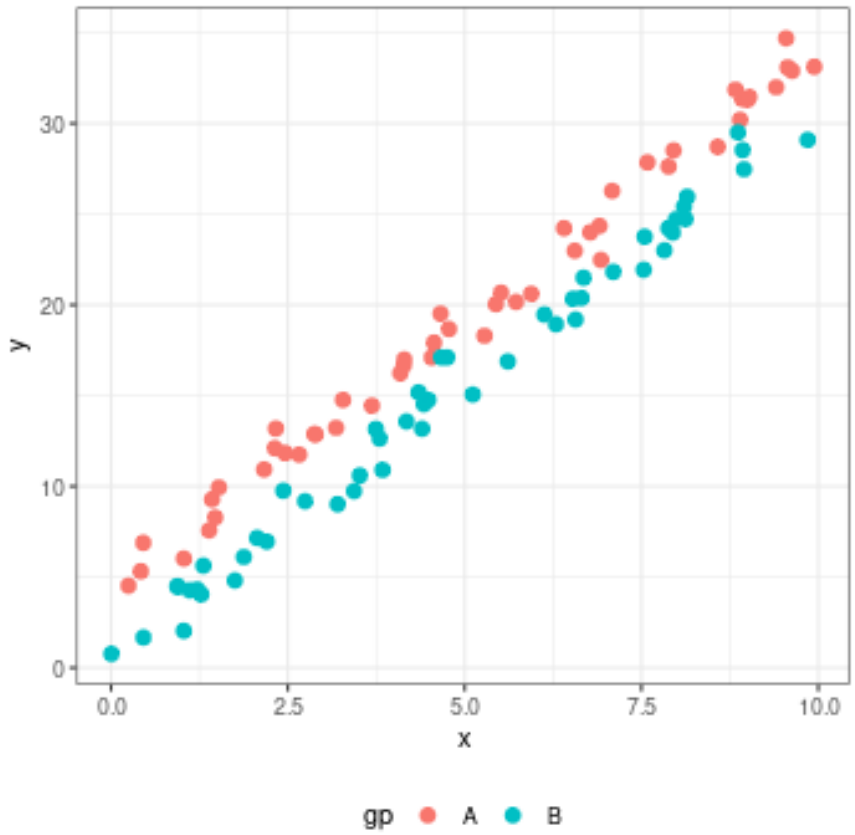


Part 2: Investigating Patterns

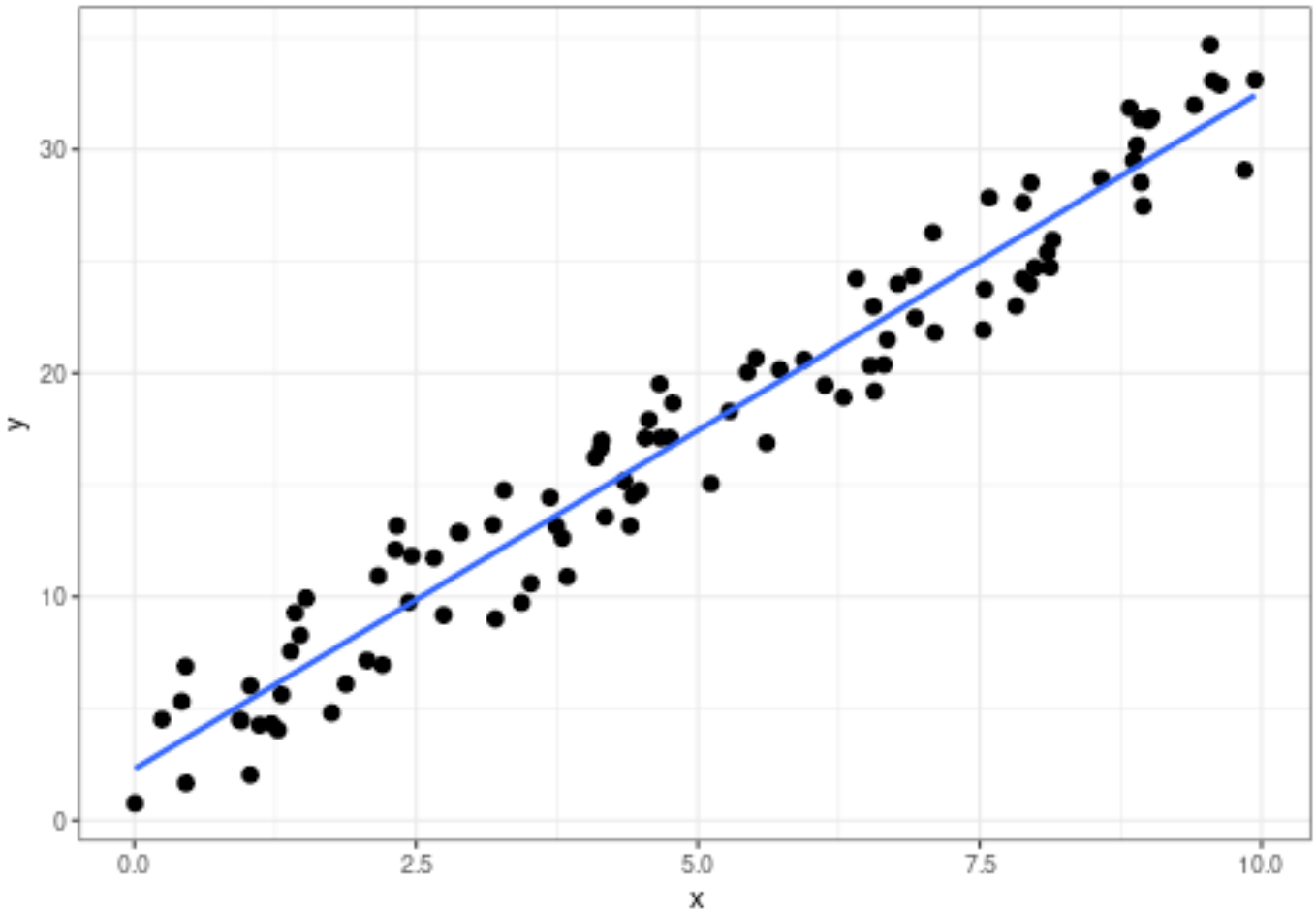
Considering new covariates

Suppose I have:

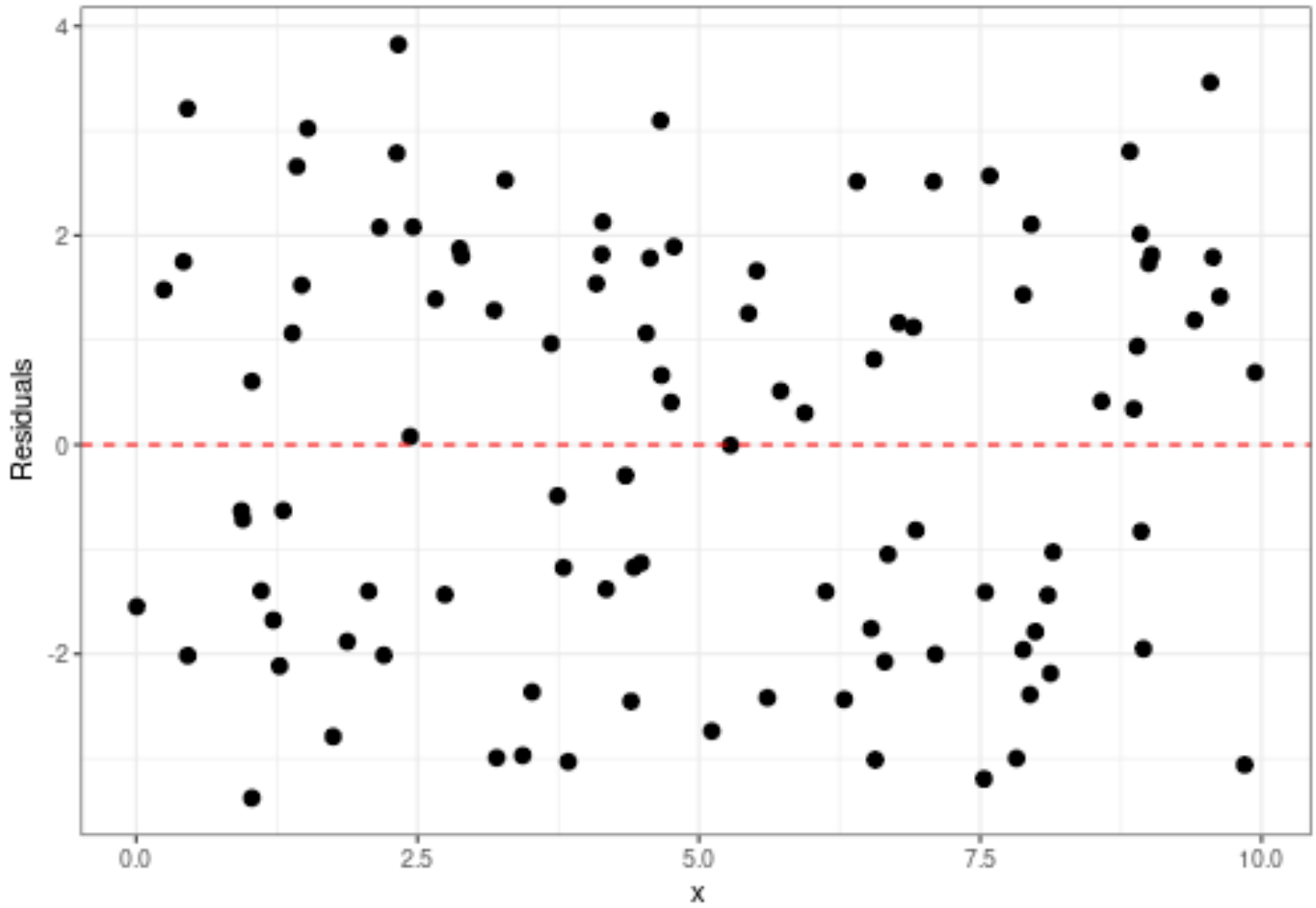
- Quantitative outcome y
- Quantitative predictor X
- Categorical predictor gp



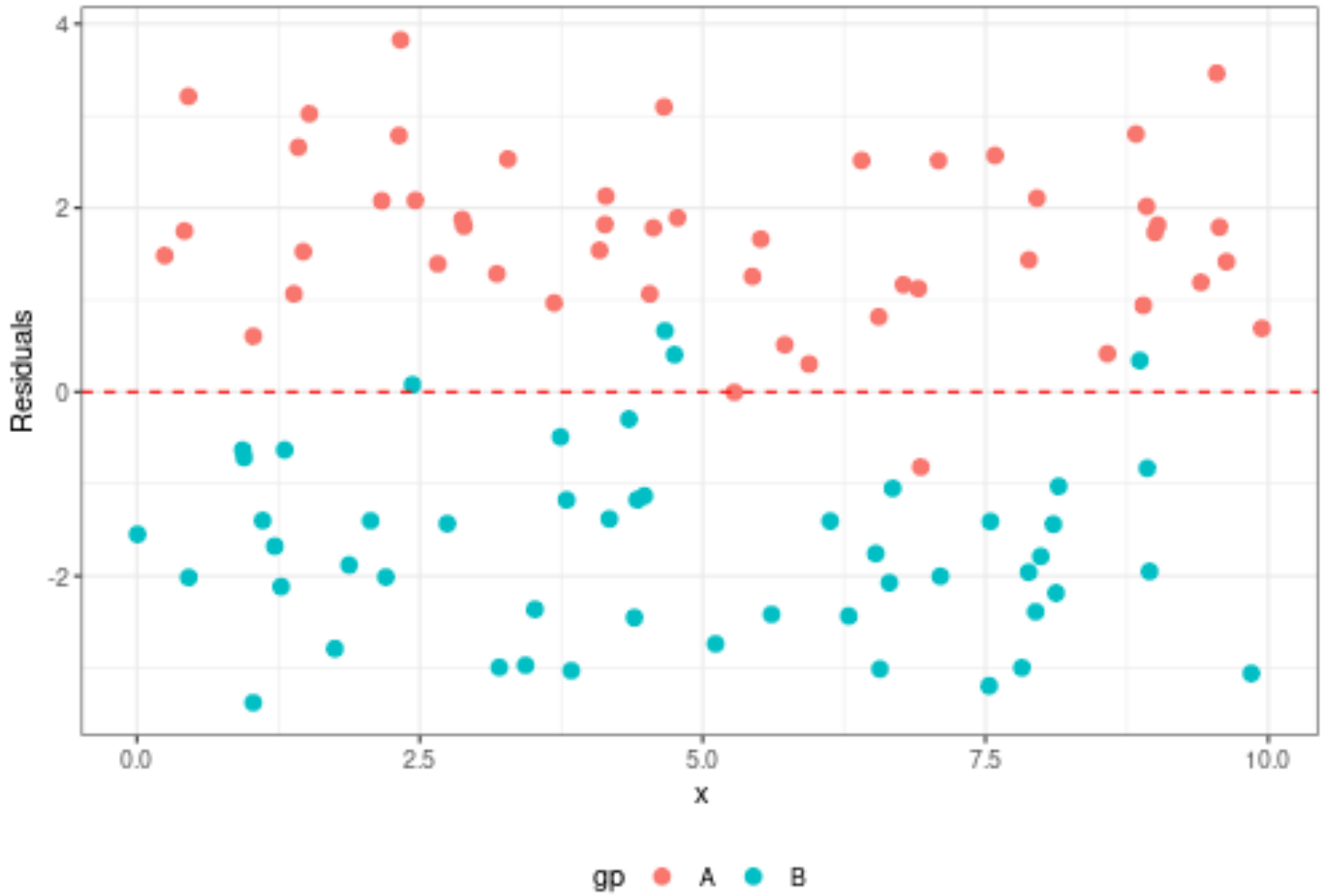
Considering new covariates



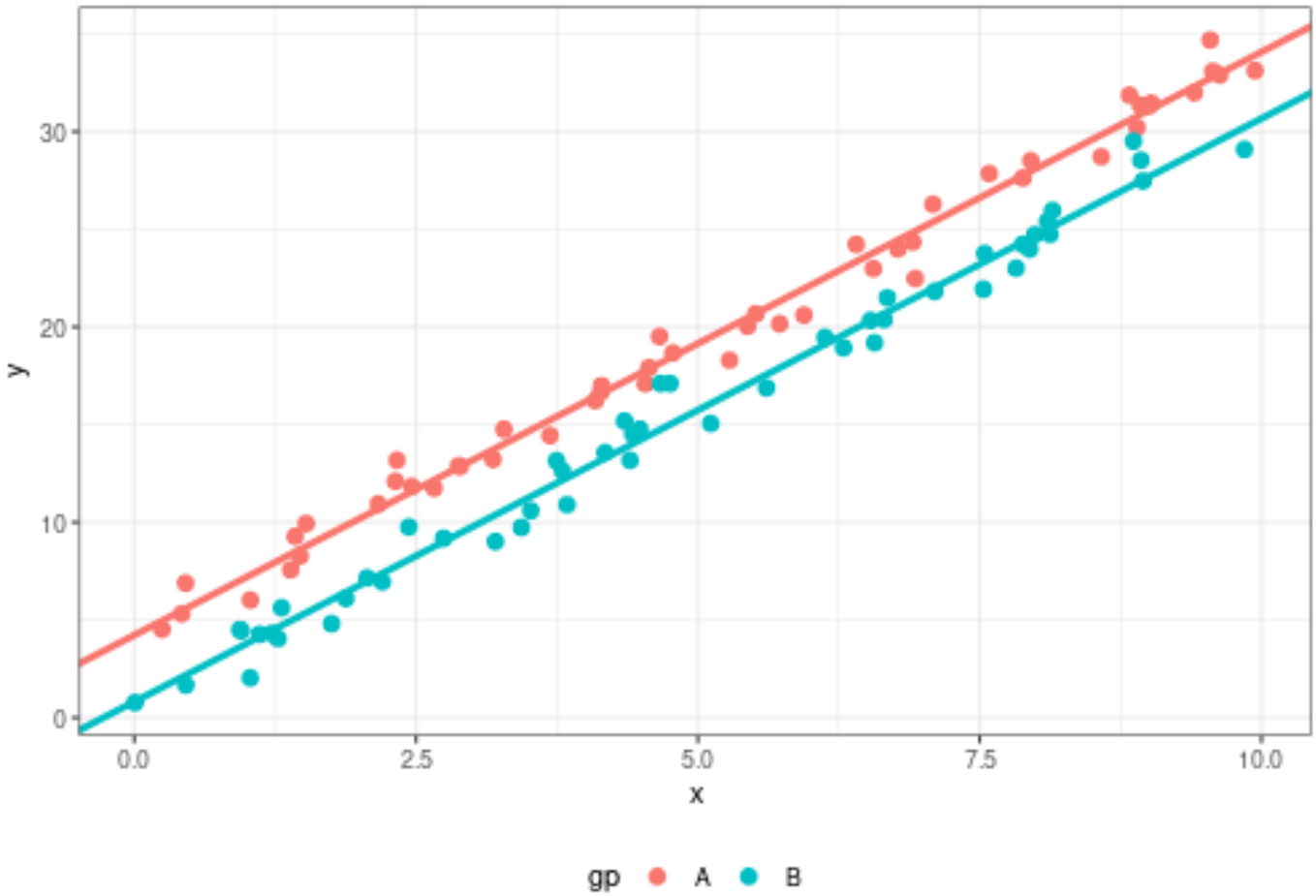
Considering new covariates



Considering new covariates

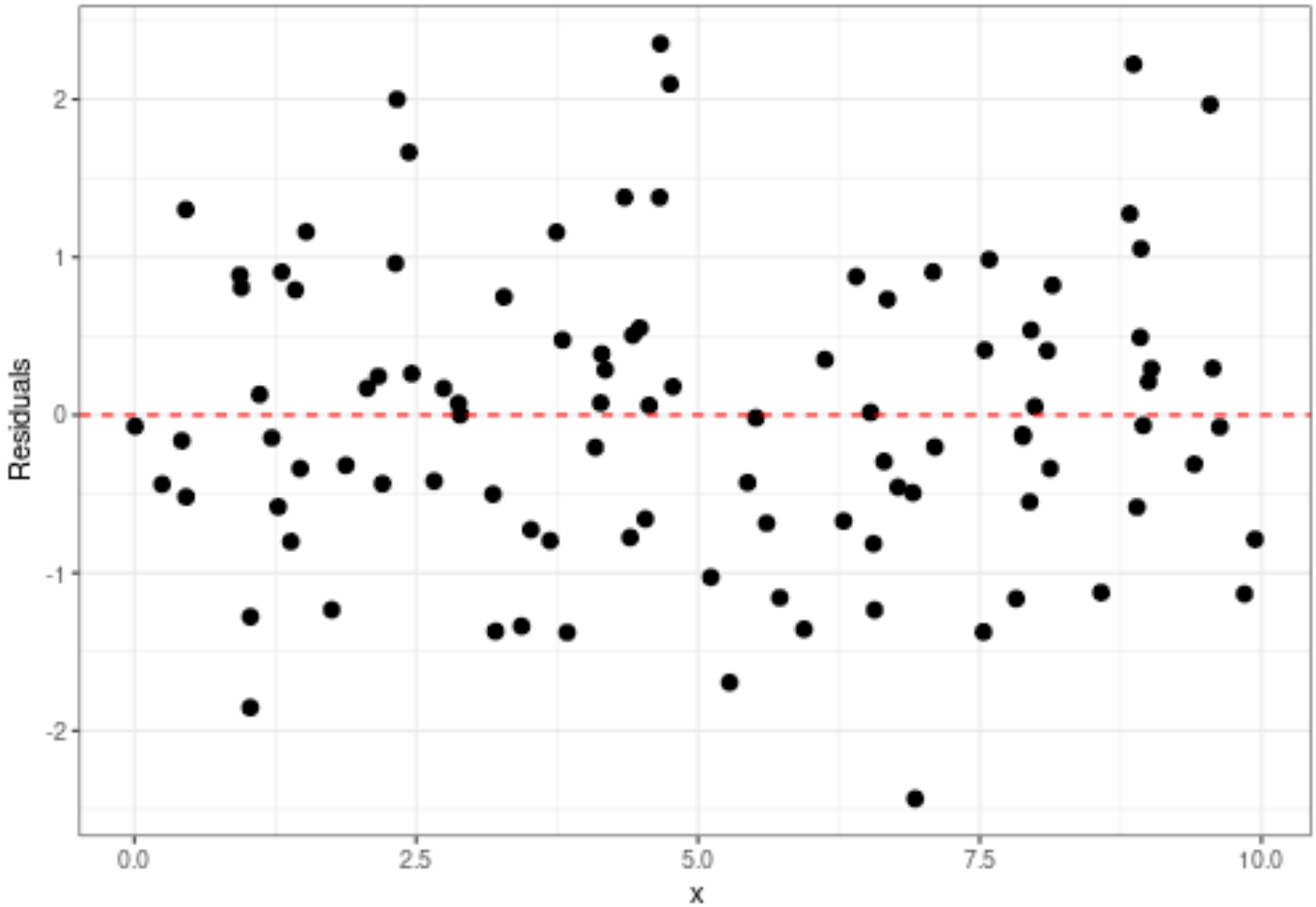


Considering new covariates



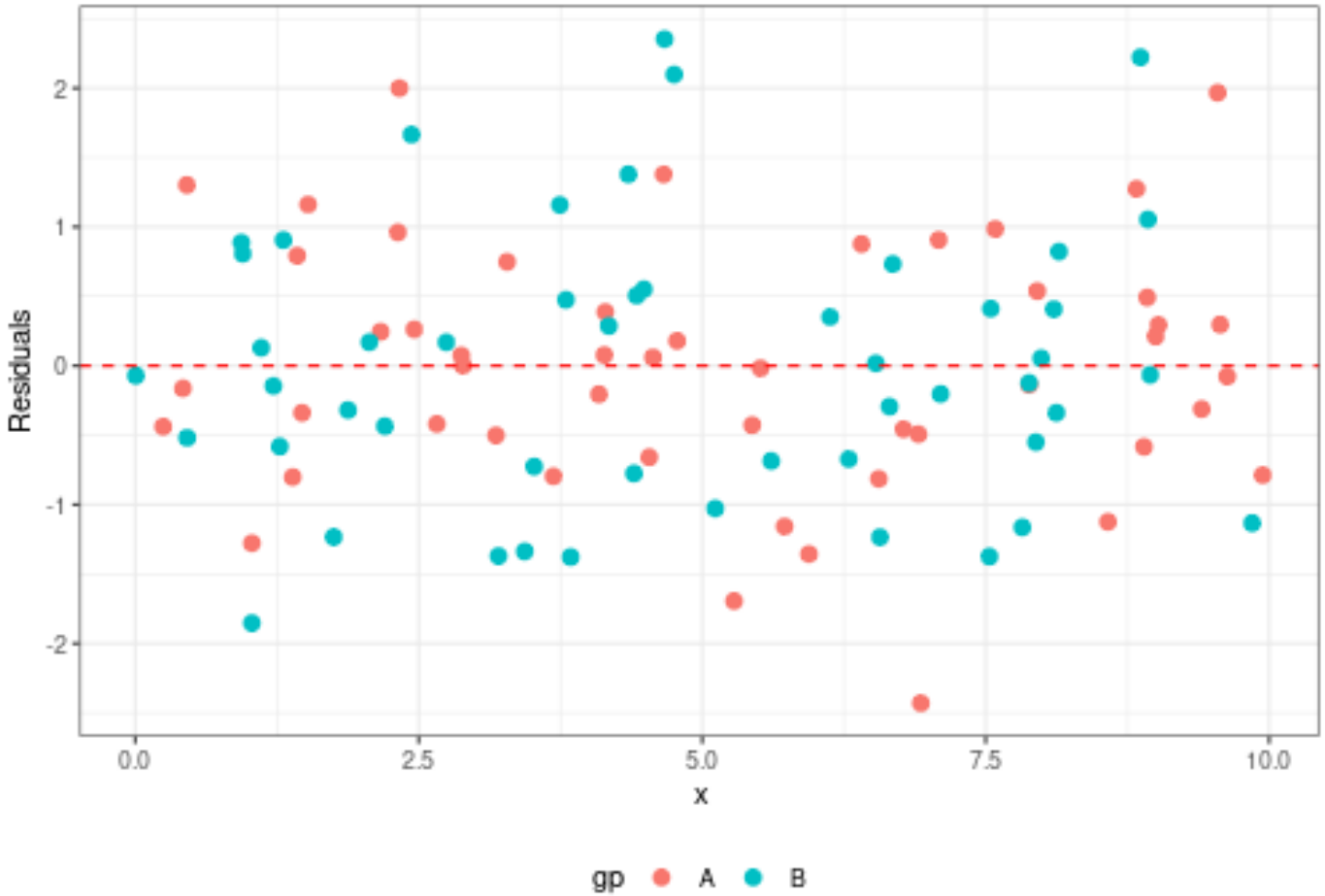
Considering new covariates

these residuals are from the model that *also* includes the gp variable



Considering new covariates

if we color by 'gp' we see that the pattern is now random about 0



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*

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

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

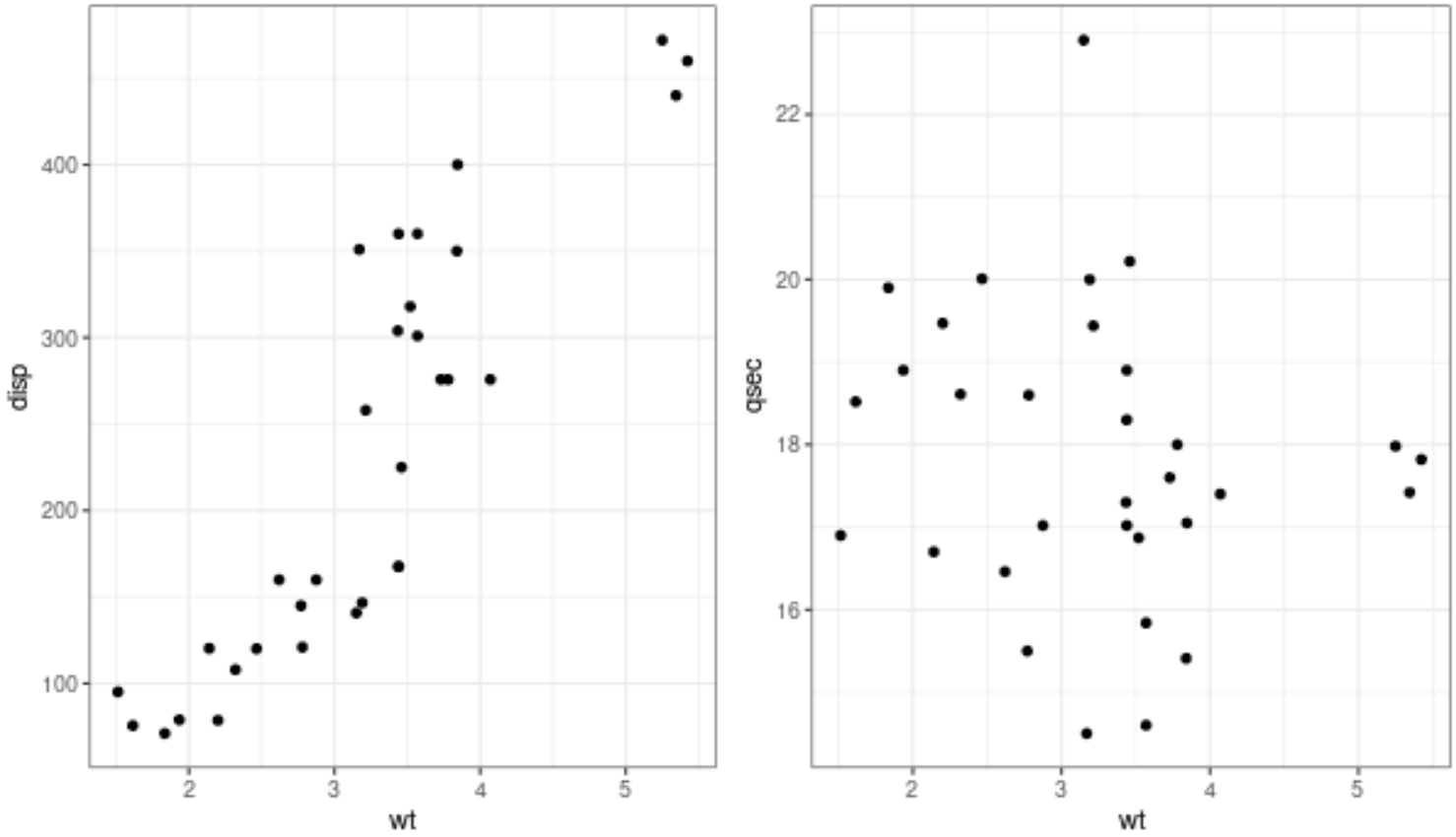
Considering new covariates

Last week (Friday) we considered an example predicting vehicle fuel economy (mpg) with three separate models:

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

Correlated Covariates

Let's say I have a regression using wt to predict mpg. We are looking for a new variable to add to the model. Which of these would be better to use?



- because wt and disp are correlated, much of the info in disp is already contained within wt $\rightarrow$ probably not much improvement if we add it

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
```

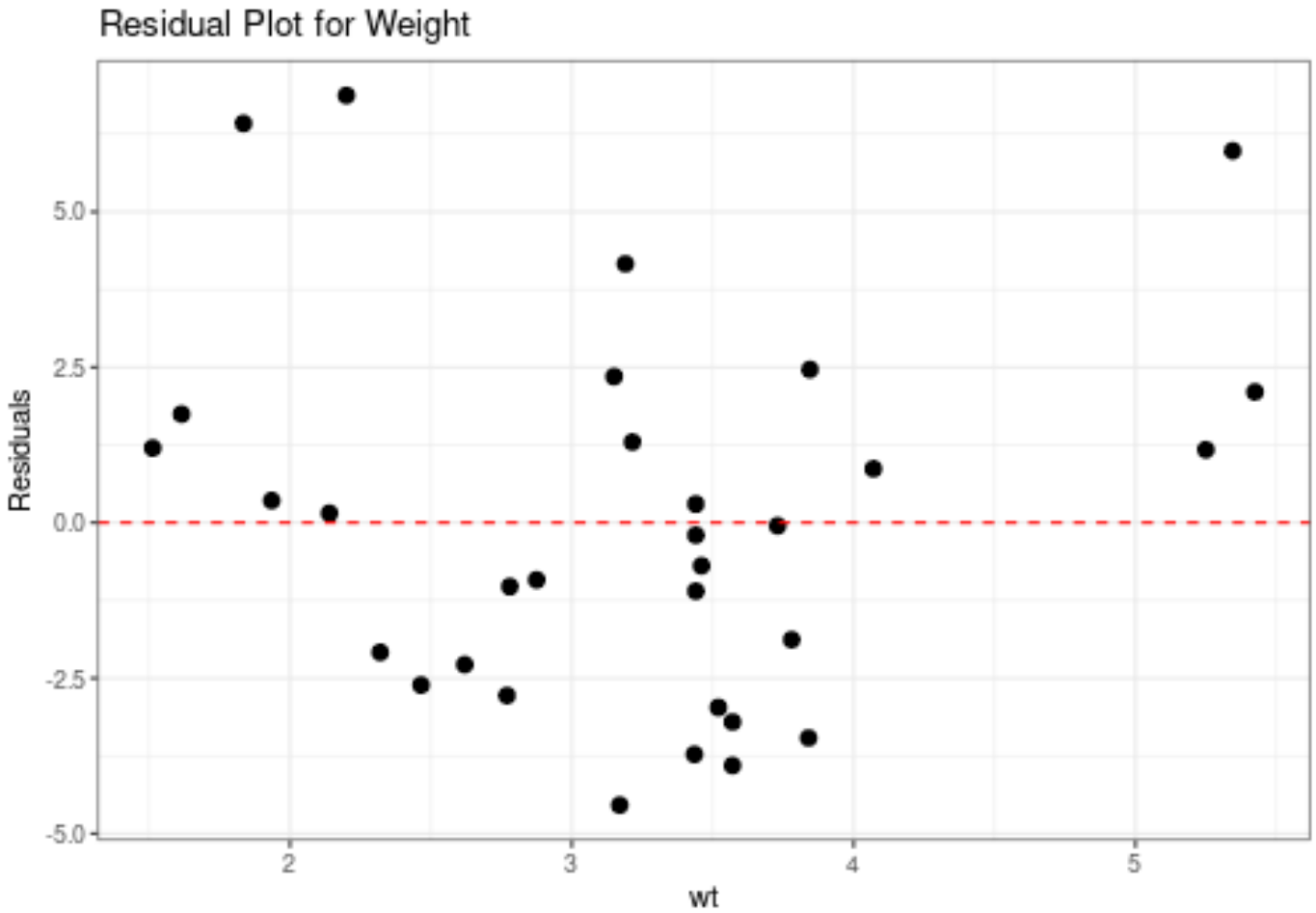
```
> lm(mpg ~ wt + disp, mtcars) %>% summary()

             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
```

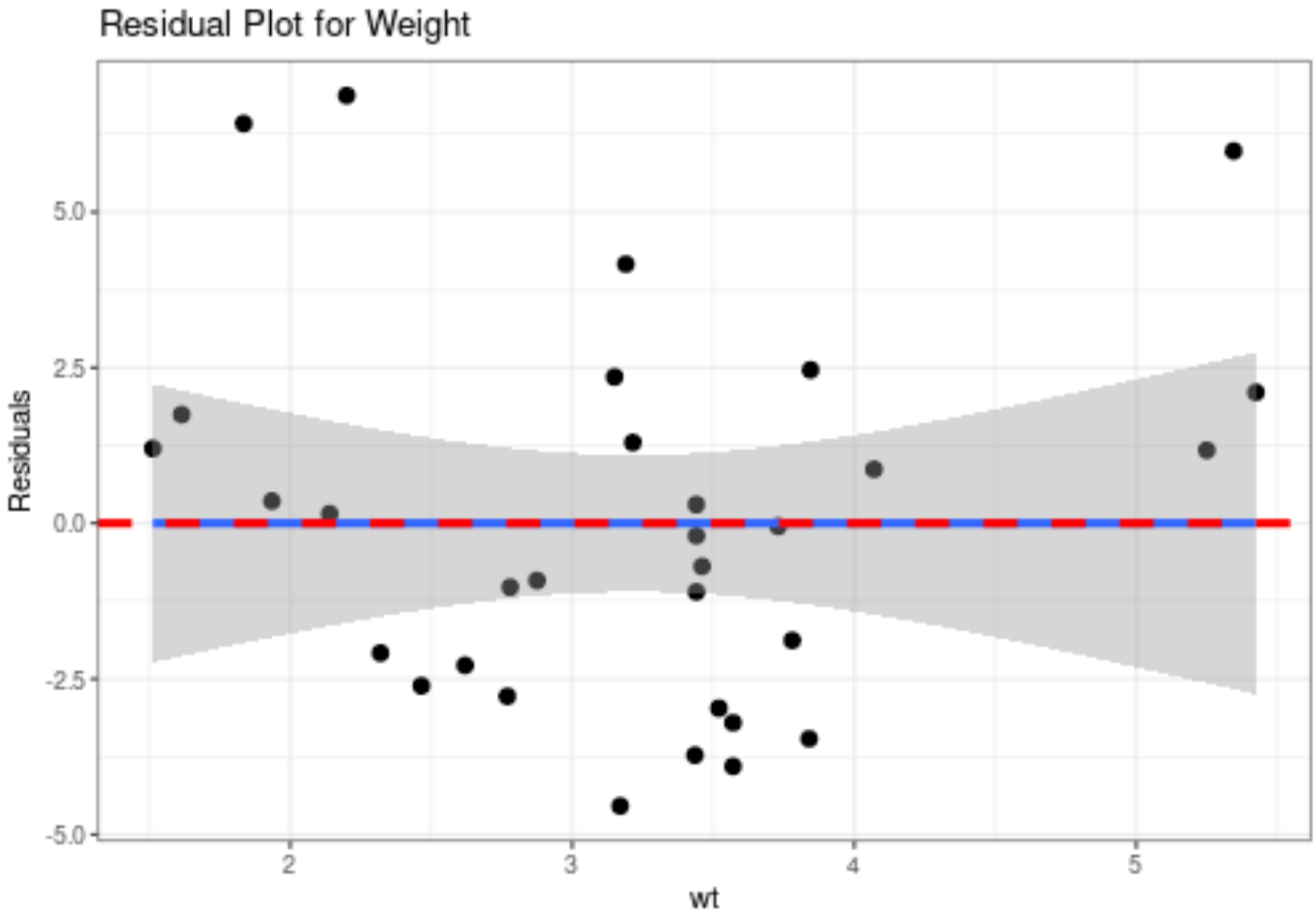
```
> lm(mpg ~ wt + qsec, mtcars) %>% summary()

             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
```

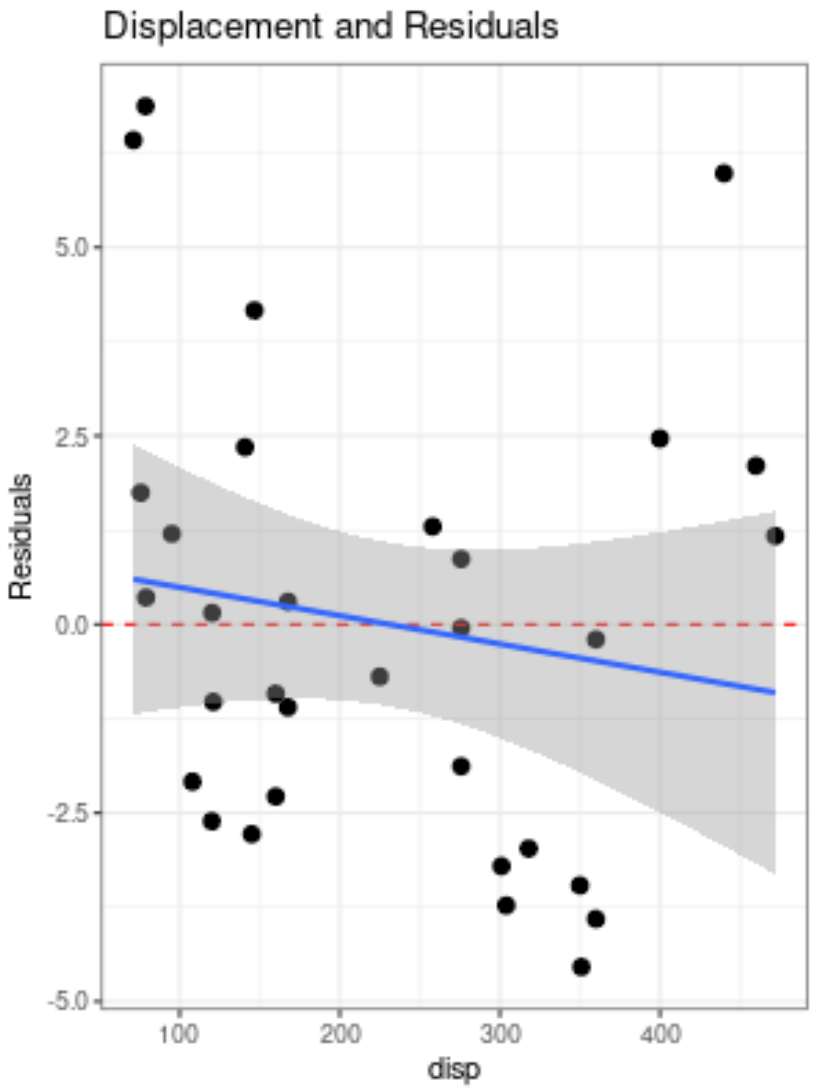
Residual Plots

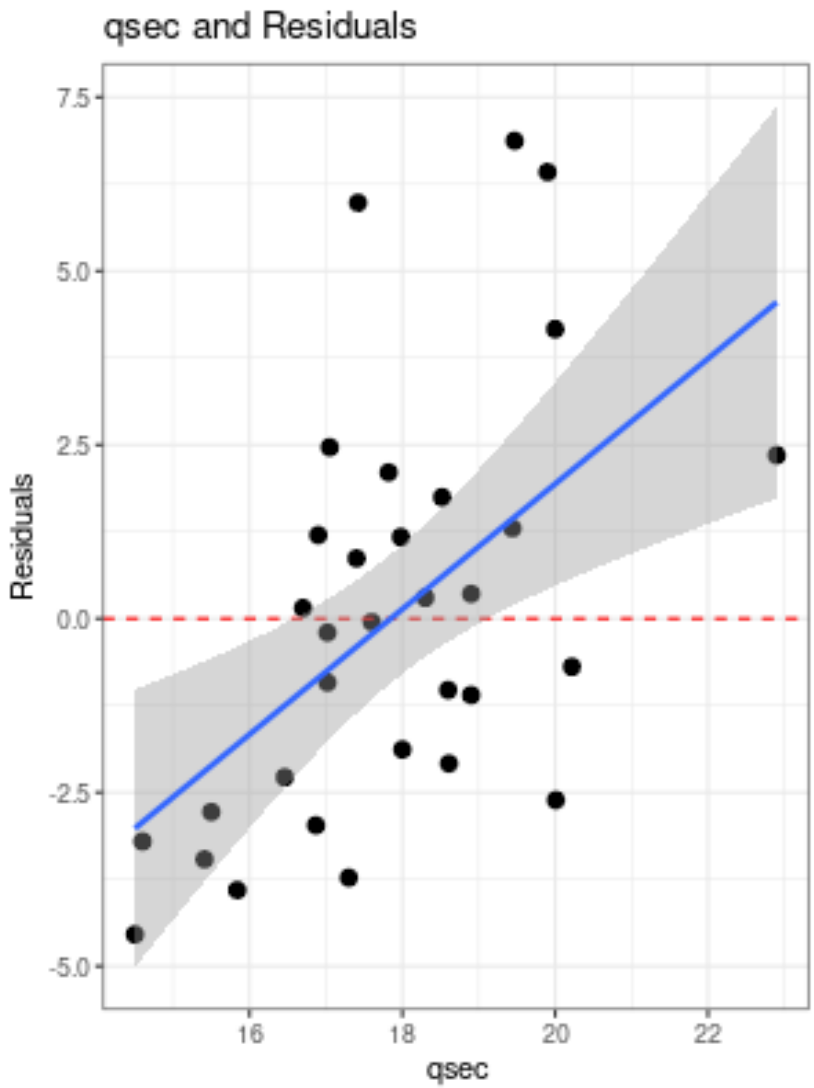


Residual Plots



Residual Plots





Key Takeaways

1. Number of assumptions for linear model
   - Linearity
   - Normal errors
   - Homoscedasticity
2. Need way to determine which new variables to add to model
3. Examining errors effective way to test assumptions and investigate new covariates