

Regression Error

Grinnell College

Fall 2025

- ▶ 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.)

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

Assumptions:

- ▶ Linear relationship between X and y
- ▶ Error term is normally distributed, $\epsilon \sim N(0, \sigma)$
 - ▶ We needed Normal distributions for means when using t-tests
- ▶ Error *variance* should be the same for all values of X , i.e., roughly same error for all observations
 - ▶ otherwise something could be going horribly wrong

Graphing the residuals gives us a way to test the assumptions of our model

Residuals

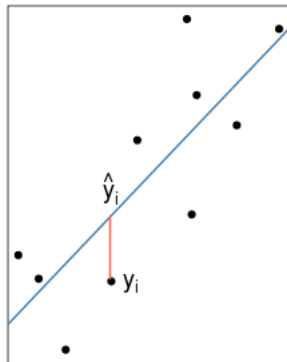
Visually, let's review what residuals look like

- residuals represent how far off our prediction is

Collection of (x, y) points



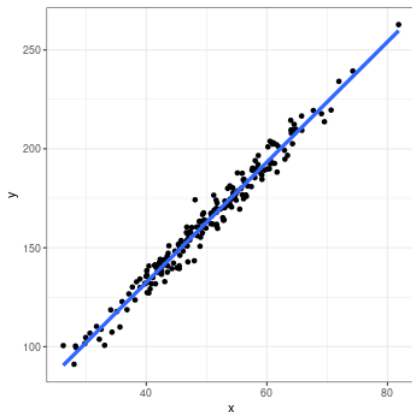
Fitted line with residual



Residuals and assumptions

Three common ways to investigate residuals visually:

1. Plot histogram of residuals (normality)
2. Plot residuals against a predictor (linear trend, changing variance)
3. Normal Quantile Plot – compares quantiles of residuals to quantiles of Normal distribution to see if they match

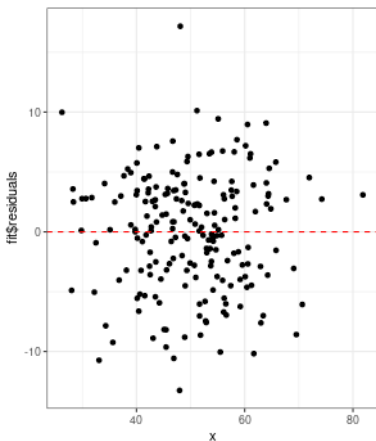
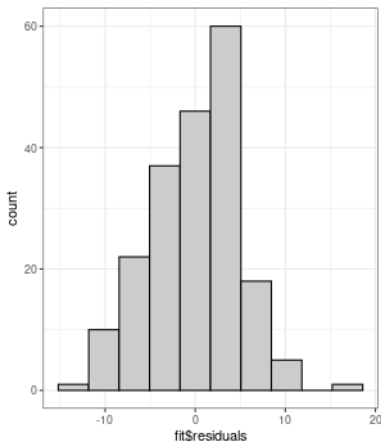


Checking Normality

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

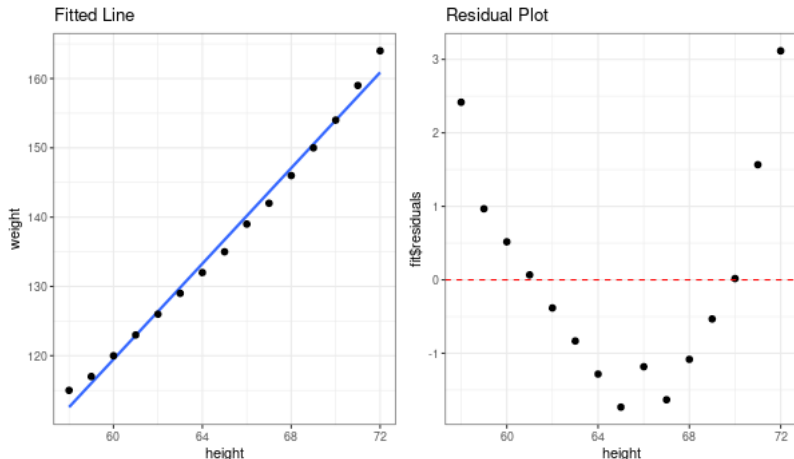
Residuals should not have a pattern other than 'blob of points' in a Resid. vs. Expl. Var. scatterplot

- ▶ don't want correlation between residuals and explanatory variables



Tests of linearity

Residual vs. Explanatory plot makes seeing non-linearity easier

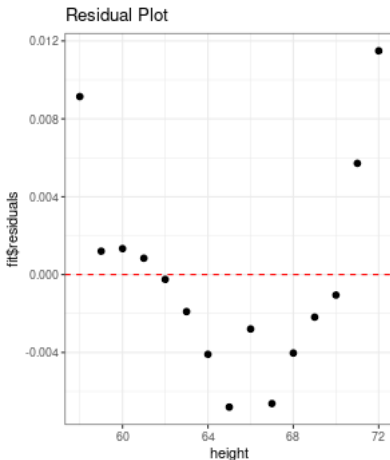
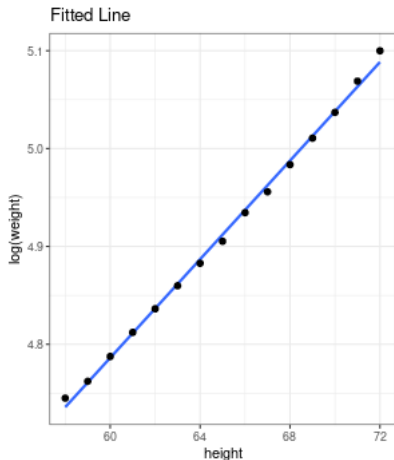


- ▶ linear regression could still be useful!
- ▶ but we could also look at doing something more complicated if we really cared

Tests of linearity

Sometimes a transformation of a variable can help correct trends $\rightarrow \log(\text{weight})$

- ▶ better, still have a funky Residual vs. Height plot

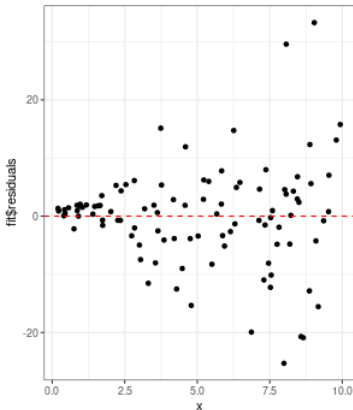
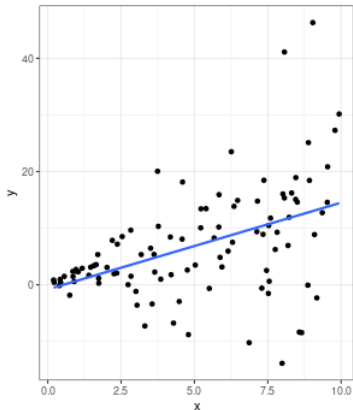


Heteroscedasticity / Homoscedasticity

Hetero- = different, Homo- = same, scedastic = random

We do not want variance of residuals to increase for really small or really large values of a predictor

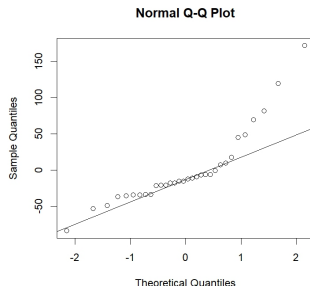
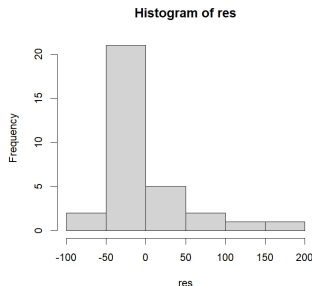
- ▶ This means our residuals start out small but then keep getting bigger → bad!
- ▶ predictions for small values of x are good, but predictions for large x are bad



Normal QQ Plot

A Normal Q-Q plot (Quantile - Quantile) is useful for seeing if our residuals follow a Normal distribution.

- ▶ Normal QQ Plot compares the quantiles of our residuals to what we would expect of a Normal distribution that has the same variance as our residuals ($\sigma^2 = \text{MSE}$)



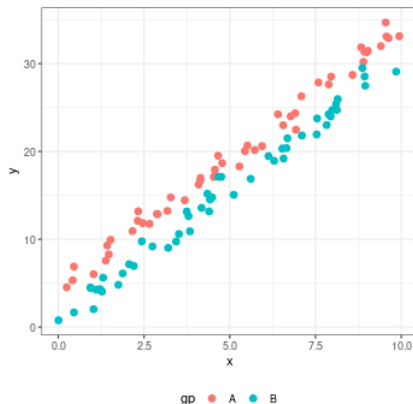
- ▶ Skewed residuals → most of the time residuals are positive/negative (bad), sometimes **really** far off in the other direction (very bad)
- ▶ straight line → Normal distribution seems OK

Part 2: Investigating Patterns

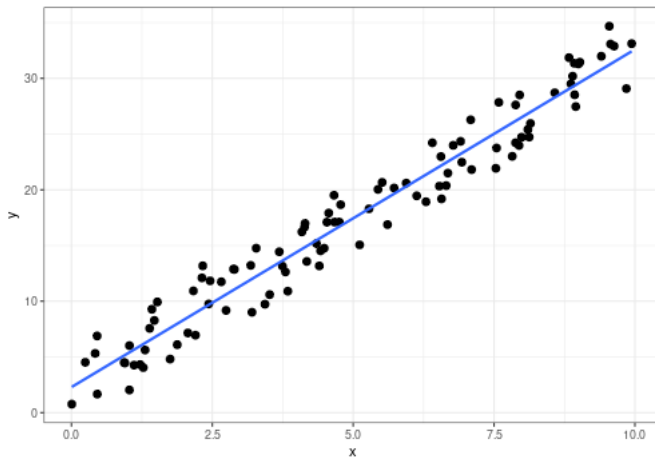
Considering new covariates

Suppose I have:

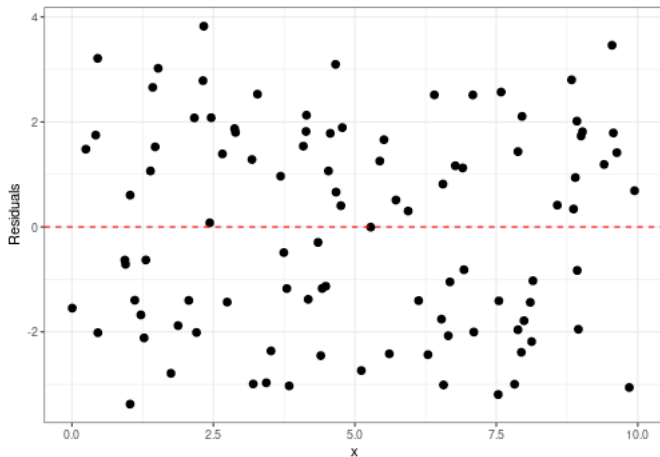
- ▶ Quantitative outcome y
- ▶ Quantitative predictor X
- ▶ Categorical predictor gp



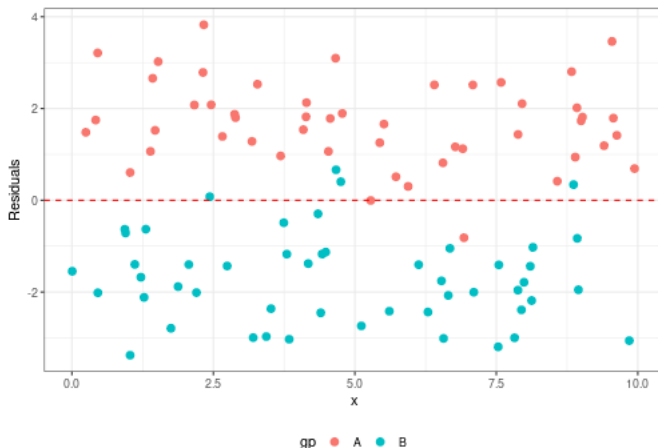
Considering new covariates



Considering new covariates

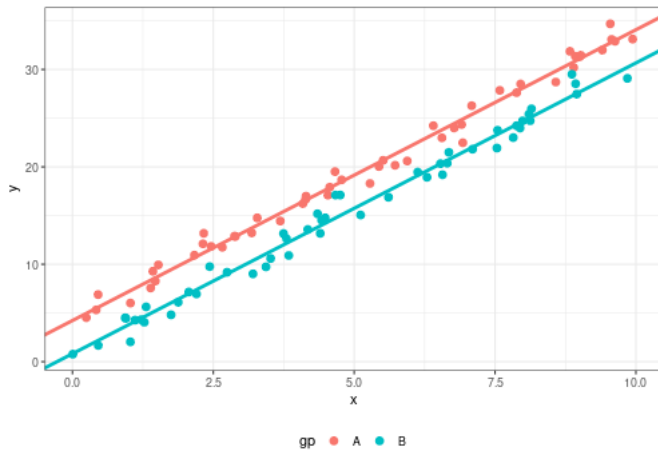


Considering new covariates



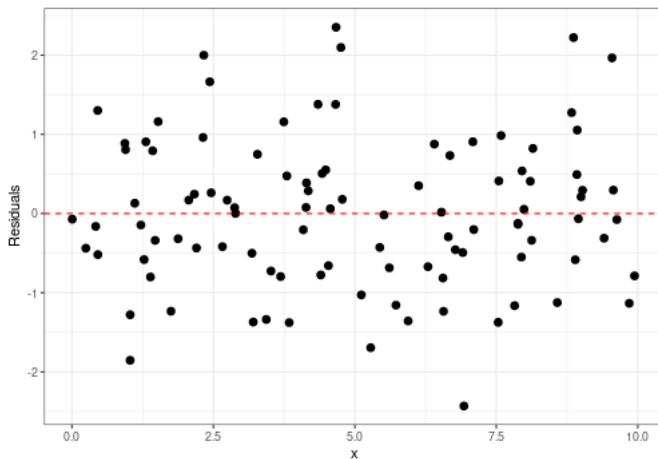
- ▶ Nearly all 'A' observations are overpredicted, all 'B' residuals underpredicted
- ▶ we could use the original scatterplot + color by gp to see pattern
 - ▶ residual plot is easier to quickly cycle through other variables to see patterns

Considering new covariates



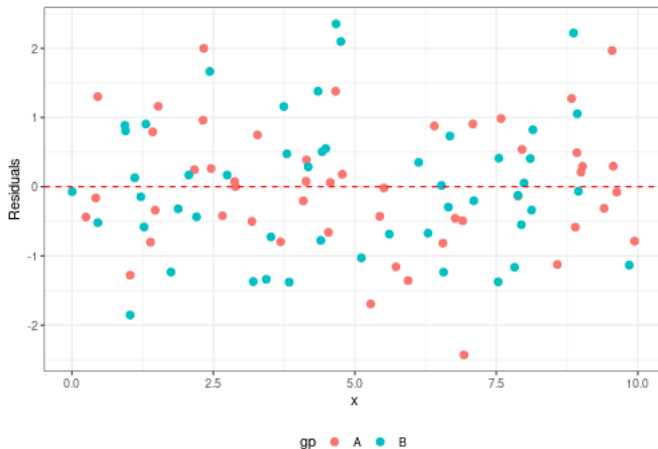
Considering new covariates

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



► indicates this model does better than previous one

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

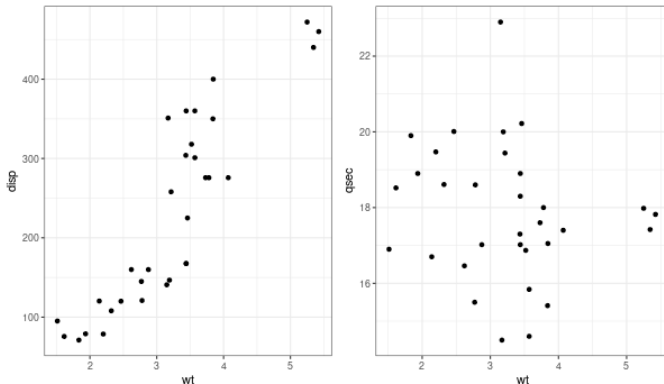
Considering new covariates

Wednesday 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 (in seconds)

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 → probably not much improvement if we add it
- ▶ rephrased: knowing about wt already gives us a good idea of disp values → disp is not useful if we are already using wt

Correlated Covariates

Predicting mpg with wt

```
1 > lm(mpg ~ wt, mtcars) %>% summary()
2           Estimate Std. Error t value      Pr(>|t|)
3 (Intercept)   37.285      1.878   19.86 < 0.000002 ***
4 wt           -5.344      0.559   -9.56  0.000013 ***
5 R-squared = 0.75
```

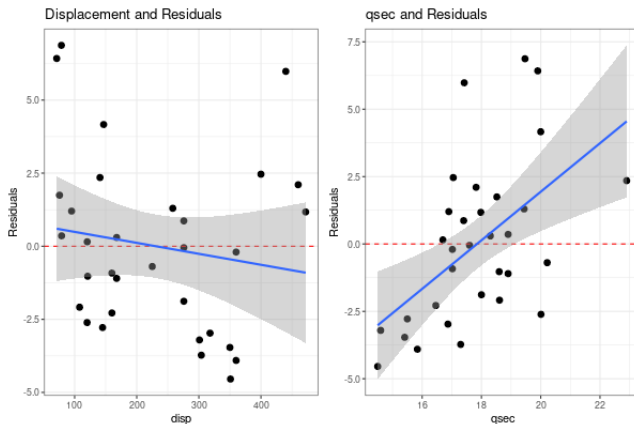
Add displacement to original

```
1 > lm(mpg ~ wt + disp, mtcars) %>% summary()
2           Estimate Std. Error t value      Pr(>|t|)
3 (Intercept)  34.96055      2.16454   16.15 0.0000000049 ***
4 wt          -3.35083      1.16413    -2.8   0.0074 **
5 disp        -0.01772      0.00919    -1.93  0.0636 .
6 R-squared = 0.78
```

Add qsec to original

```
1 > lm(mpg ~ wt + qsec, mtcars) %>% summary()
2           Estimate Std. Error t value      Pr(>|t|)
3 (Intercept)   19.746      5.252    3.76   0.00077 ***
4 wt           -5.048      0.484   -10.43 0.000000000025 ***
5 qsec          0.929      0.265    3.51   0.00150 **
6 R-squared = 0.82
```

Residual Plots



- ▶ both of these residuals are made with model that does not use either disp or qsec
- ▶ We just saw 'qsec' would be better to add to the model → corresponds to a linear pattern in the residuals

Key Takeaways

1. Number of assumptions for linear model
 - ▶ Linearity
 - ▶ Normal errors
 - ▶ Homoscedasticity
2. Residual plots can help determine which new variables to add to model
3. Examining errors is an effective way to test assumptions