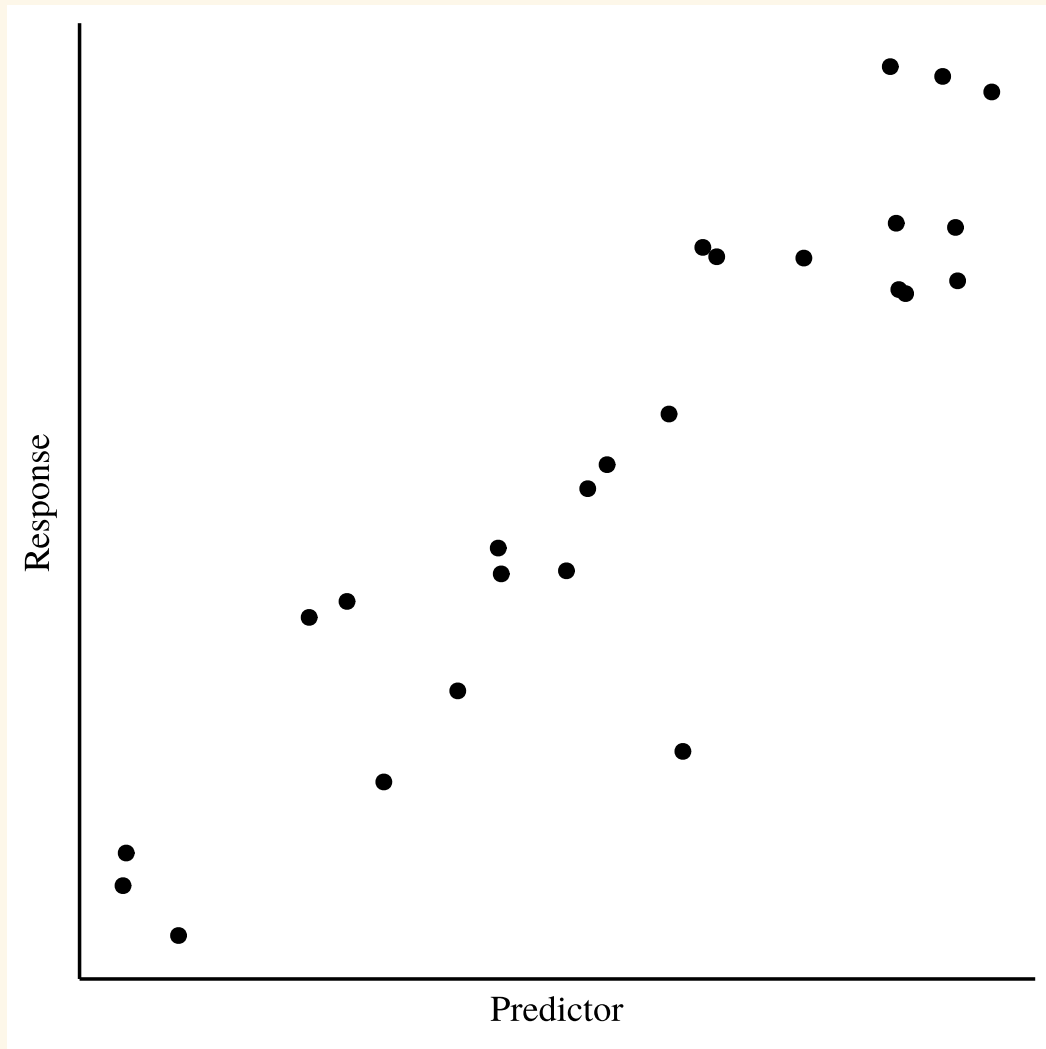# Simple Linear Regression (SLR) Model Inference

Stat 230

April 06 2022

Overview

Today:

Inference for the SLR model

- Residual plots
- Normal QQ plots
- Checks for Independence

The Simple Linear Regression (SLR) model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma)$$

- **L**inear mean function: varies linearly with x

- **I**ndependence of errors

- **N**ormality of errors

- **E**quality of variance of the errors

> "All models are wrong, but some are useful."
> Professor George Box (1919-2013)

- Statistical models are used to model or describe potentially complex physical phenomenon

Need to strike a balance between

- model simplicity: easy to derive estimates/inference methods and interpret
- model complexity: more challenging to derive estimates/inference methods and interpret

For any model: we need check whether our data "fits" the model assumptions

Model error terms

Reframe the theoretical model:

$$\epsilon_i = Y_i - (\beta_0 + \beta_1 x_i) \quad \epsilon_i \sim N(0, \sigma)$$

ϵ_i are the key!

- mean and SD don't depend on x,

- normally distributed

- and independent!

Problem: we only know ϵ_i if we know β_0 and β_1 !

Residuals

Residuals are what we get when we estimate β_0 and β_1 :

$$r_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) = y_i - \hat{y}_i$$

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ are called **fitted** values

Residuals can be thought of as the error or the "lack-of-fit" between the observed value y_i and the fitted value $\hat{y}_i$ on the regression line

R package `moderndive`

Can get nice tables with built-in functions in `moderndive`. Previously used `broom` and `tidy` to get a nice coefficients table.

```
# call the library
# install.packages("moderndive")
library(moderndive)
```

```
# Get regression table:
regression_table <- get_regression_table(cars_lm)
knitr::kable(regression_table, digits = 4, format = "html") # does not work in .pdf
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-17.579	6.758	-2.601	0.012	-31.168	-3.990
speed	3.932	0.416	9.464	0.000	3.097	4.768

Residuals

```
ggplot(regression_points, aes(x = speed, y = dist))
  geom_point() +
  theme(legend.position = "none") +
  geom_smooth(method = "lm", se = FALSE, color = "l
  geom_point(aes(y = dist_hat), shape = 1) +
  geom_segment(aes(xend = speed, yend = dist_hat),
  scale_color_continuous(low = "black", high = "red
  geom_point(aes(color = abs(residual))) +
  labs(x='Speed (in Miles Per Hour)',
       y='Stopping Distance (in Feet)',
       title='Regression of stopping distance on sp
  theme(plot.title = element_text(hjust=0.5, size=9
```


Regression of stopping distance on speed
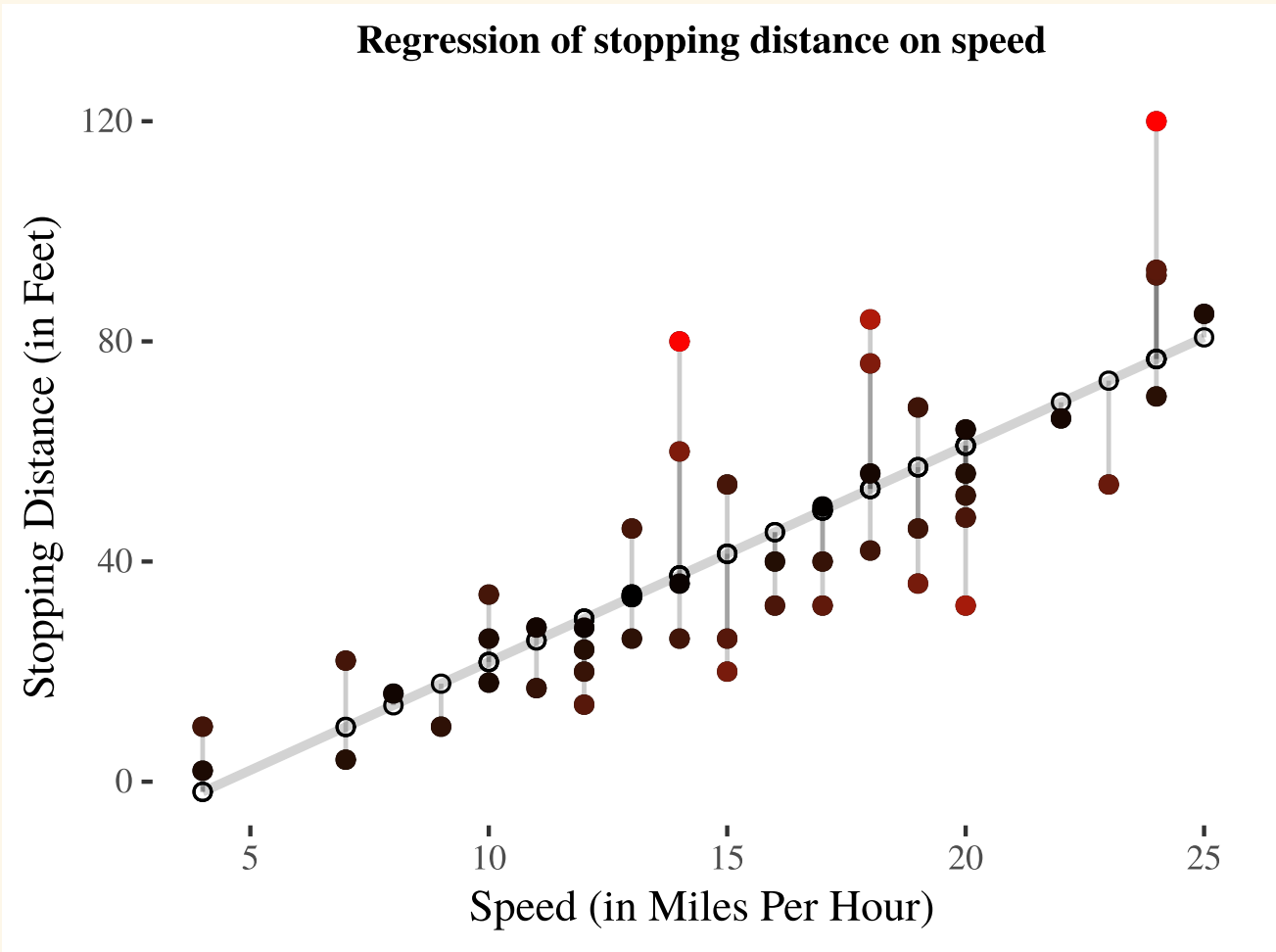
Residual plot (explanatory variable)

```
ggplot(regression_points, aes(x = speed, y = residu
  geom_point() +
  geom_segment(aes(xend = speed, yend = 0), alpha =
  theme(legend.position = "none") +
  scale_color_continuous(low = "black", high = "red
  geom_point(aes(color = abs(residual))) +
  geom_hline(yintercept = 0, col = "blue", size = 0
  labs(x='Speed (in Miles Per Hour)',
       y='Residuals',
       title='Residual plot for regression of stopp
  theme(plot.title = element_text(hjust=0.5, size=9
```


Residual plot for regression of stopping distance on speed
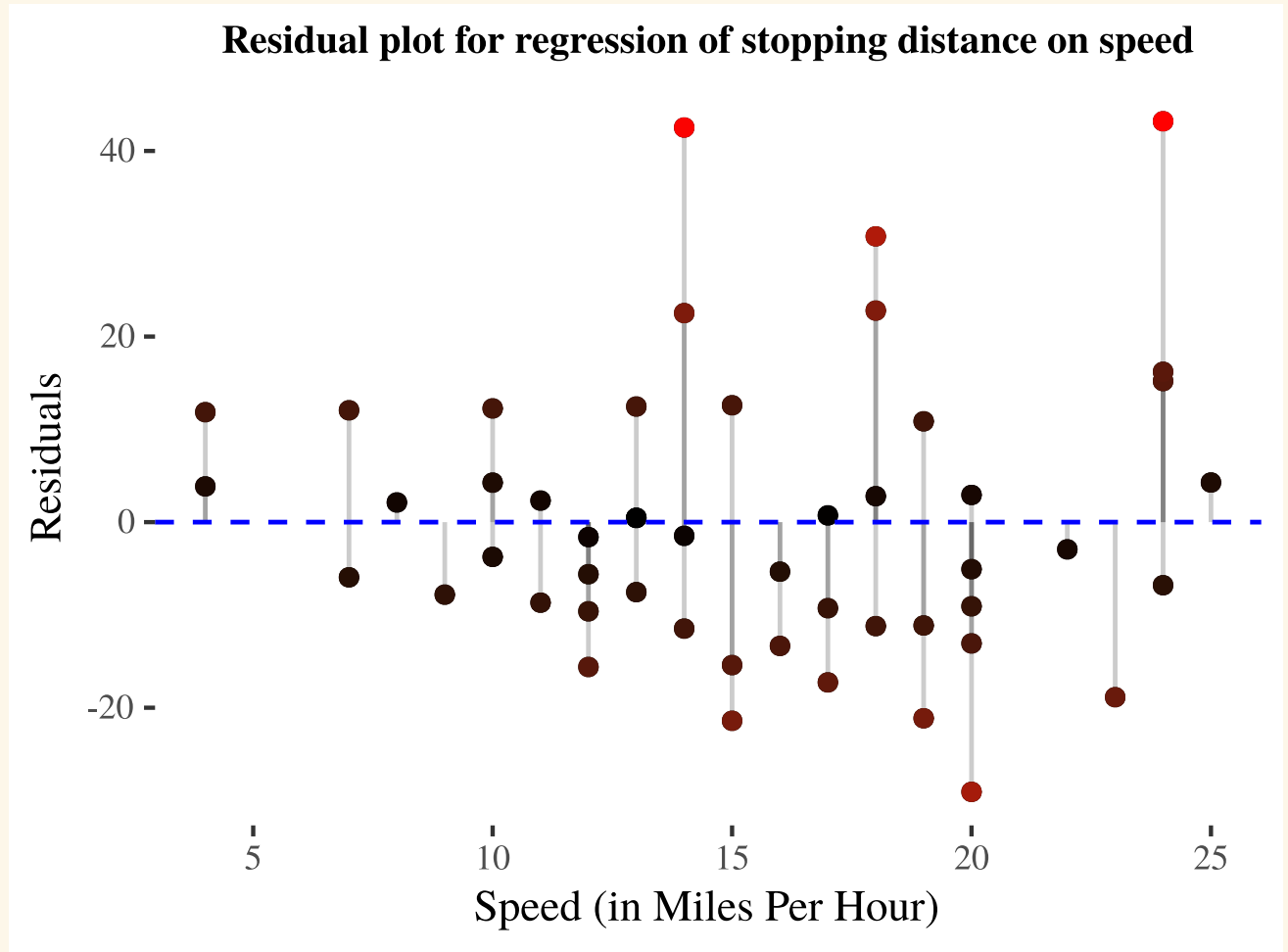
Scatterplot of r_i against $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
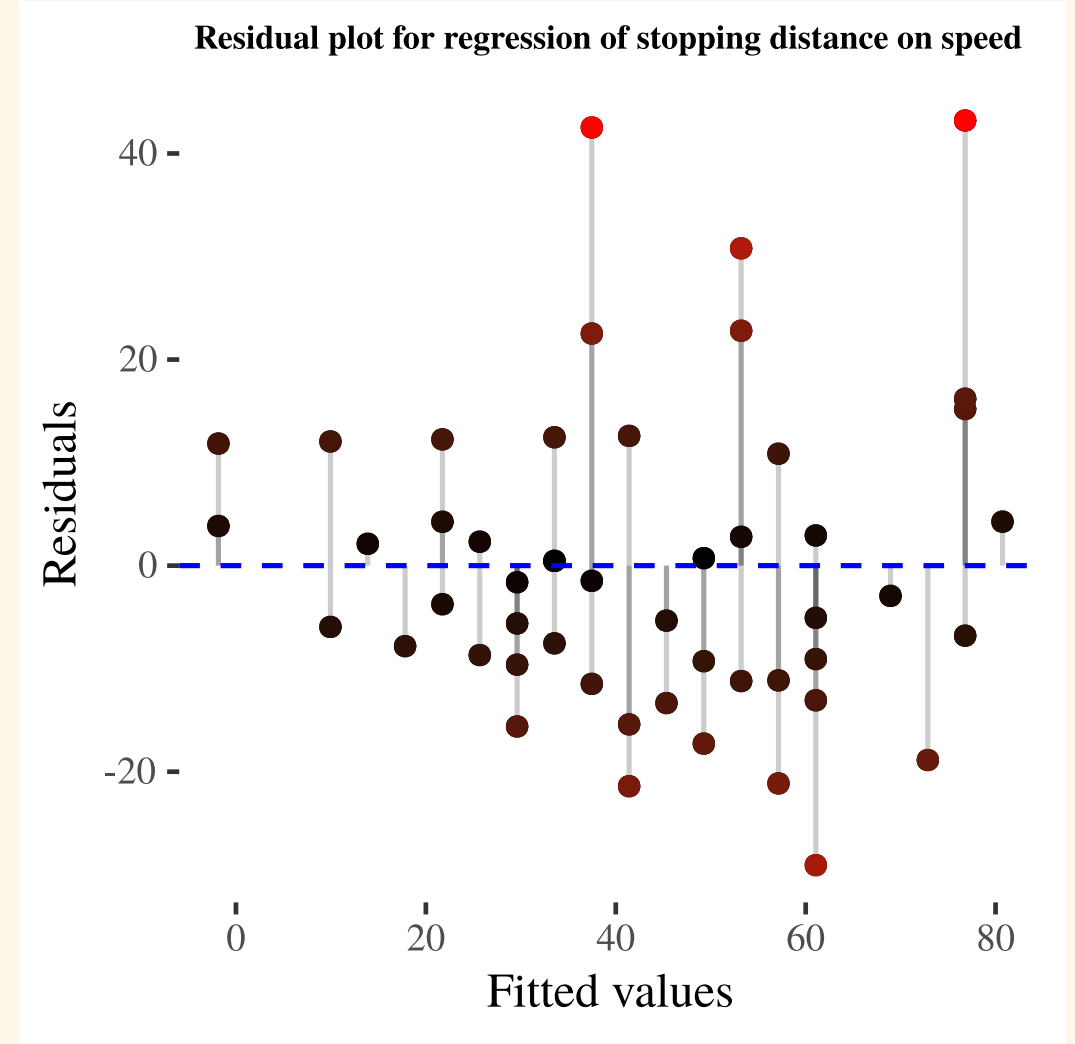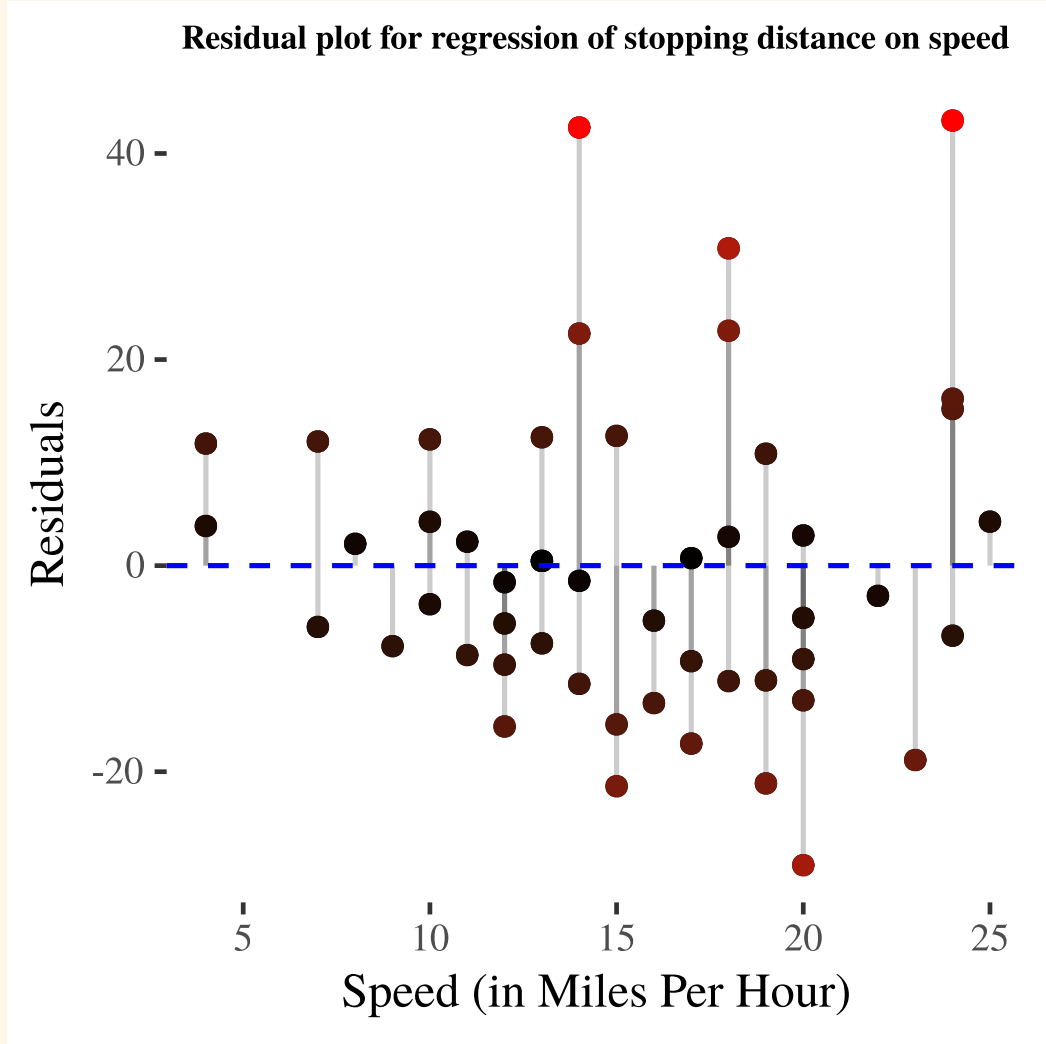
Residual plot: interpretation

SLR assumptions met if we see a "null" plot:

1. **Linearity met:** similar scatter above and below the 0-line, for all values of x

2. **Equal/Constant variance met:** magnitude of residual values around the 0-line doesn't depend on x
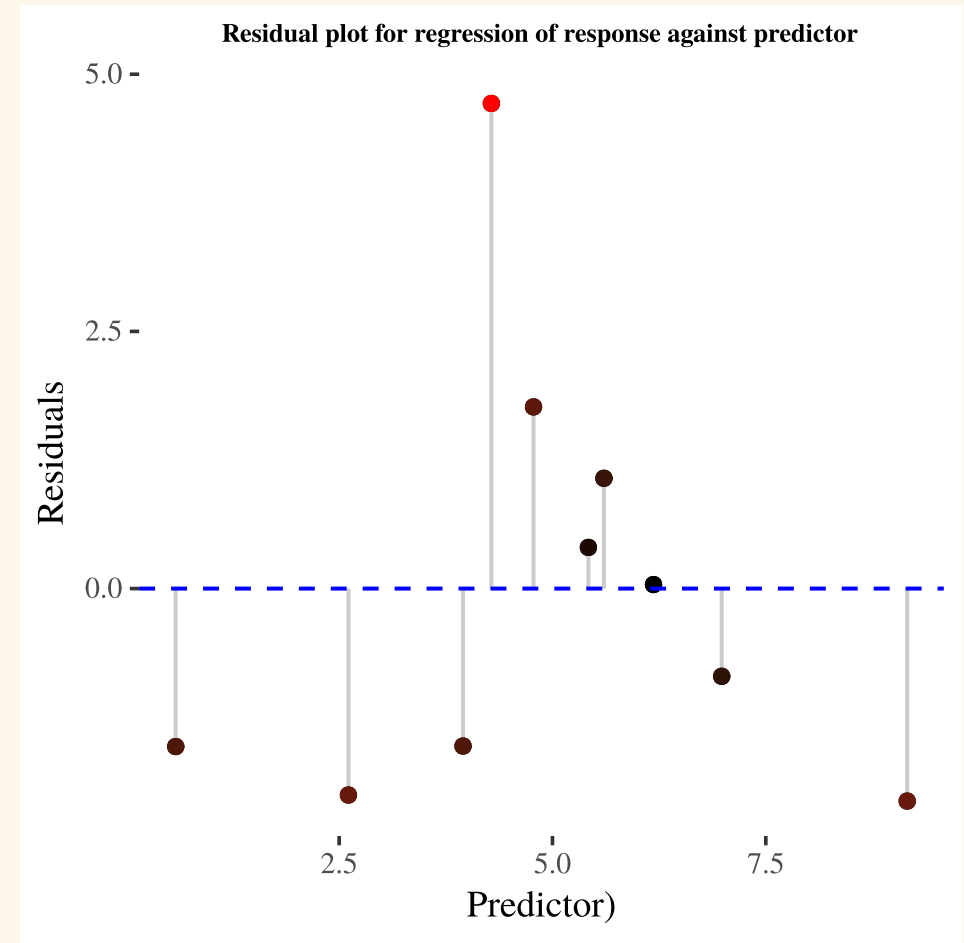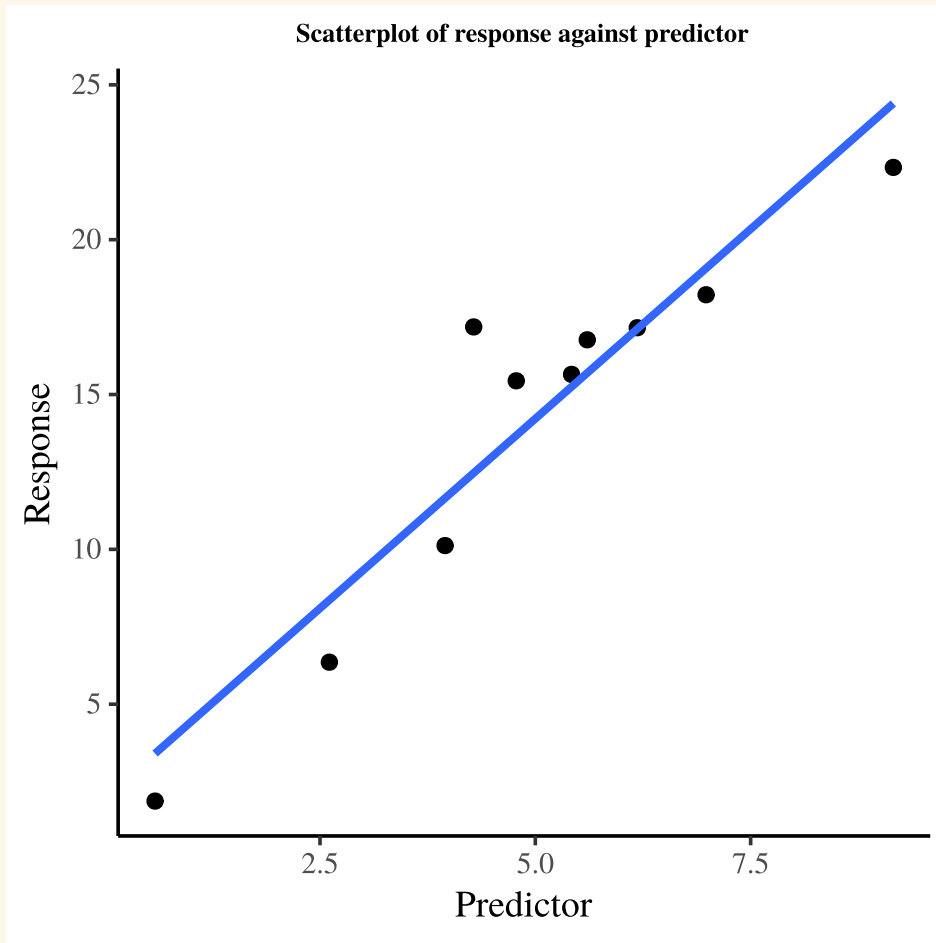
Watch out for:

- **Curvature** - indicates mean function may not be linear

- **Fan shapes** - indicates nonconstant variance

- **Outliers** (may have large effect on estimates)

Residual plot: null plot

Scatterplot of response against predictor

Residual plot for regression of response against predictor
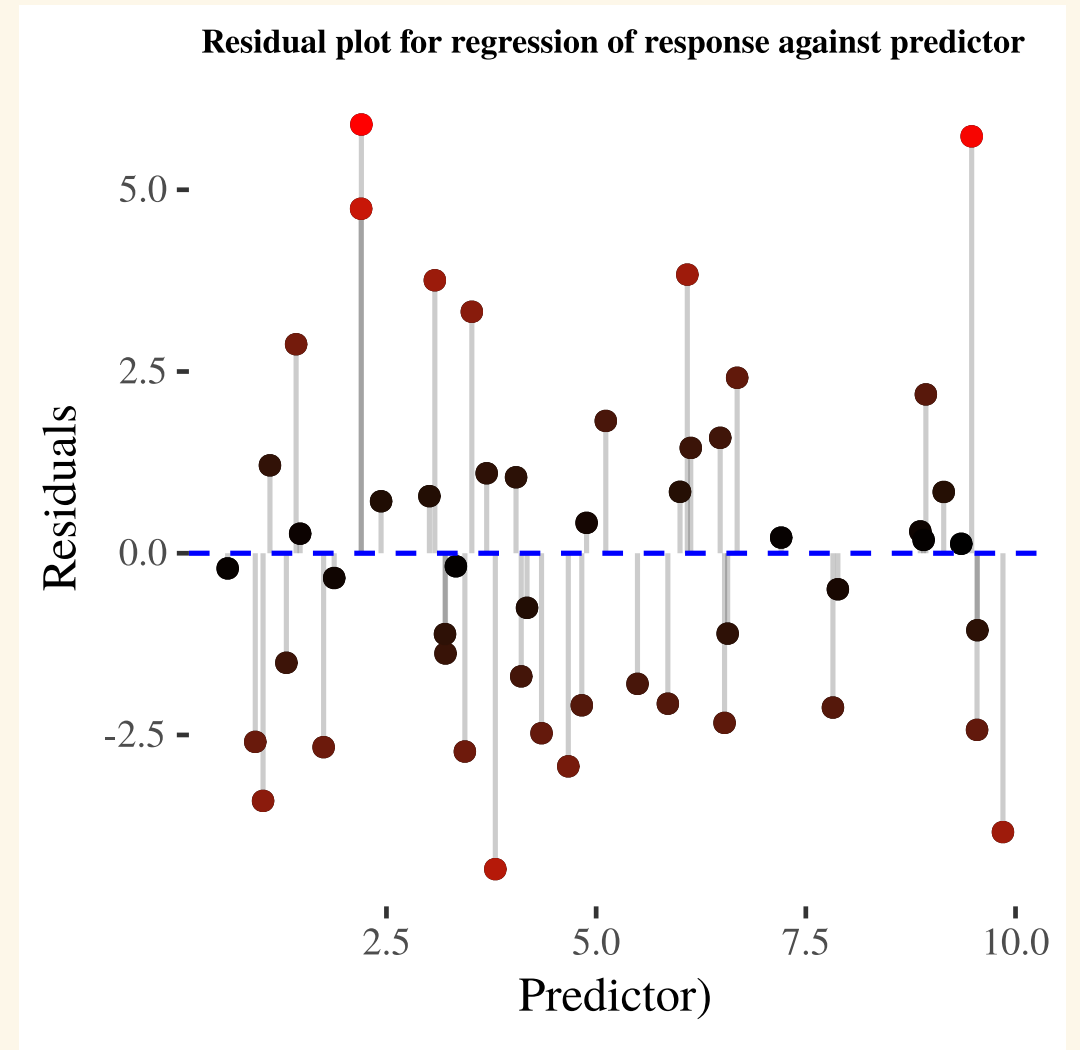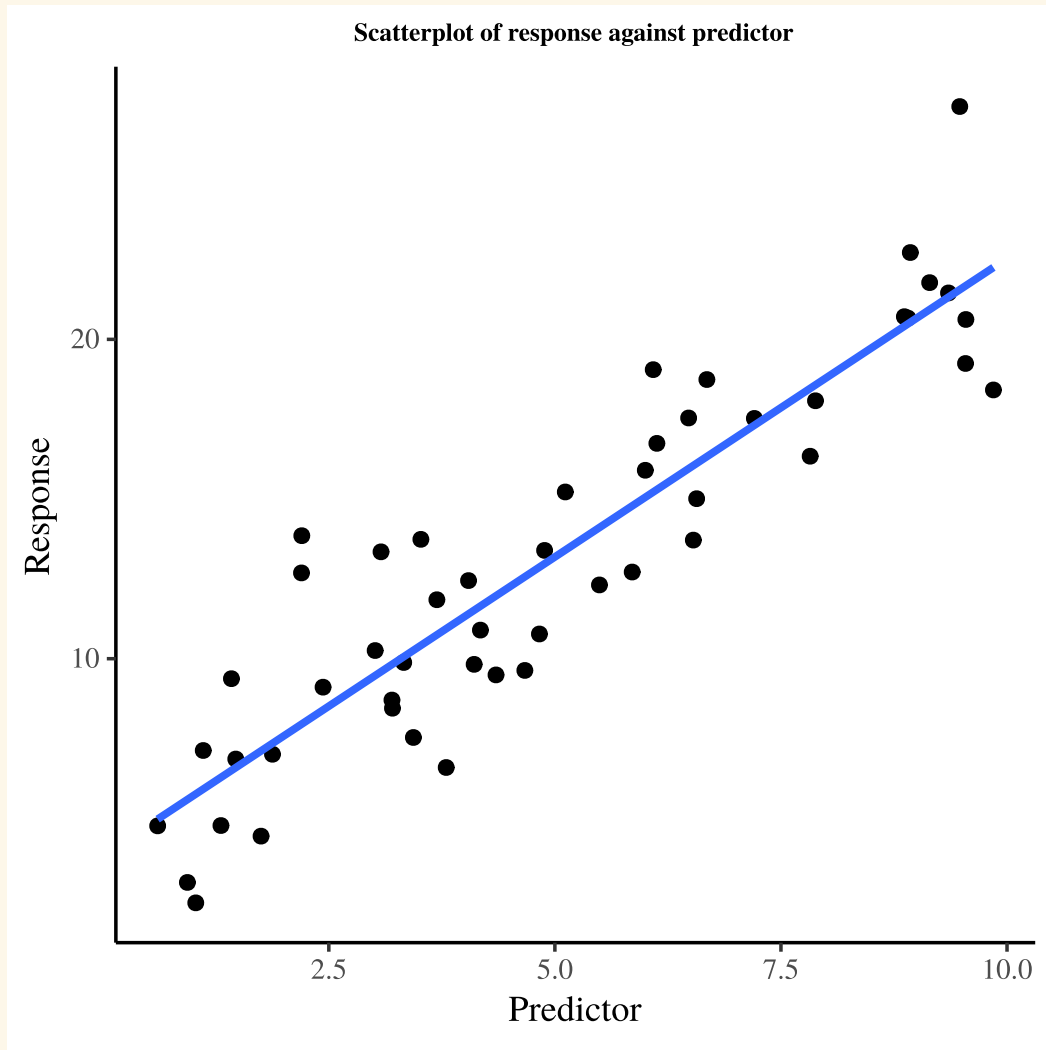
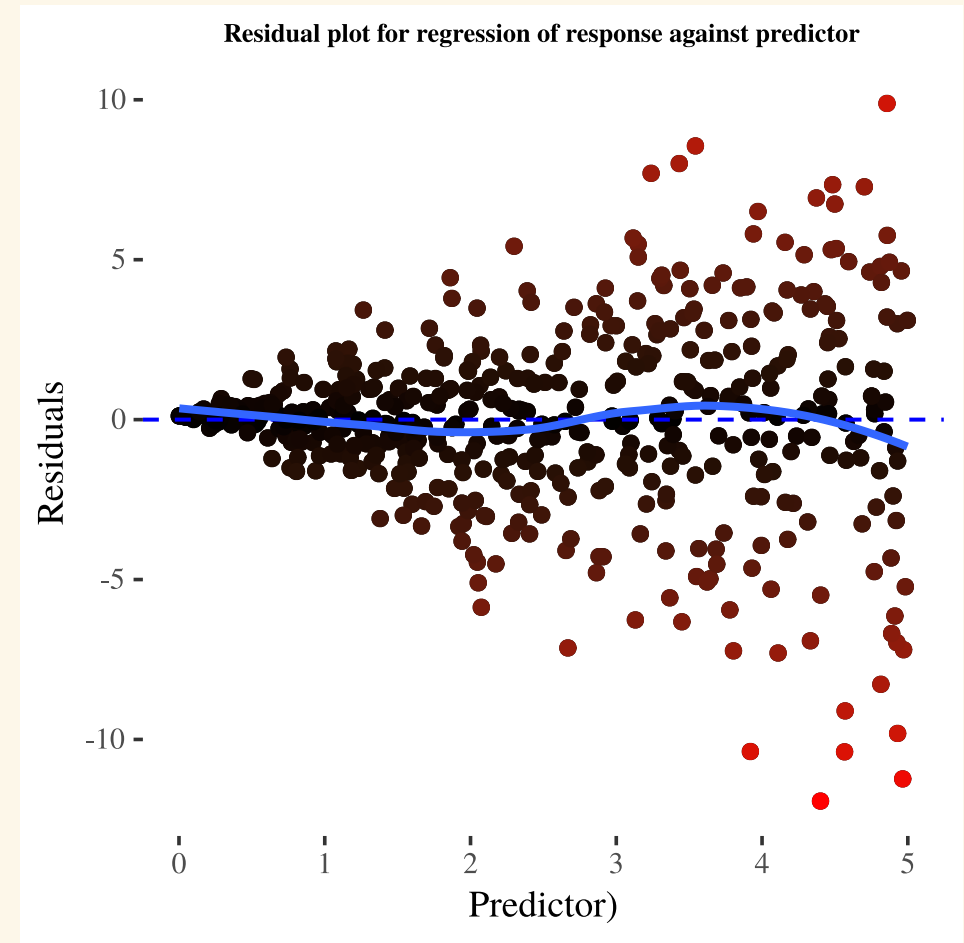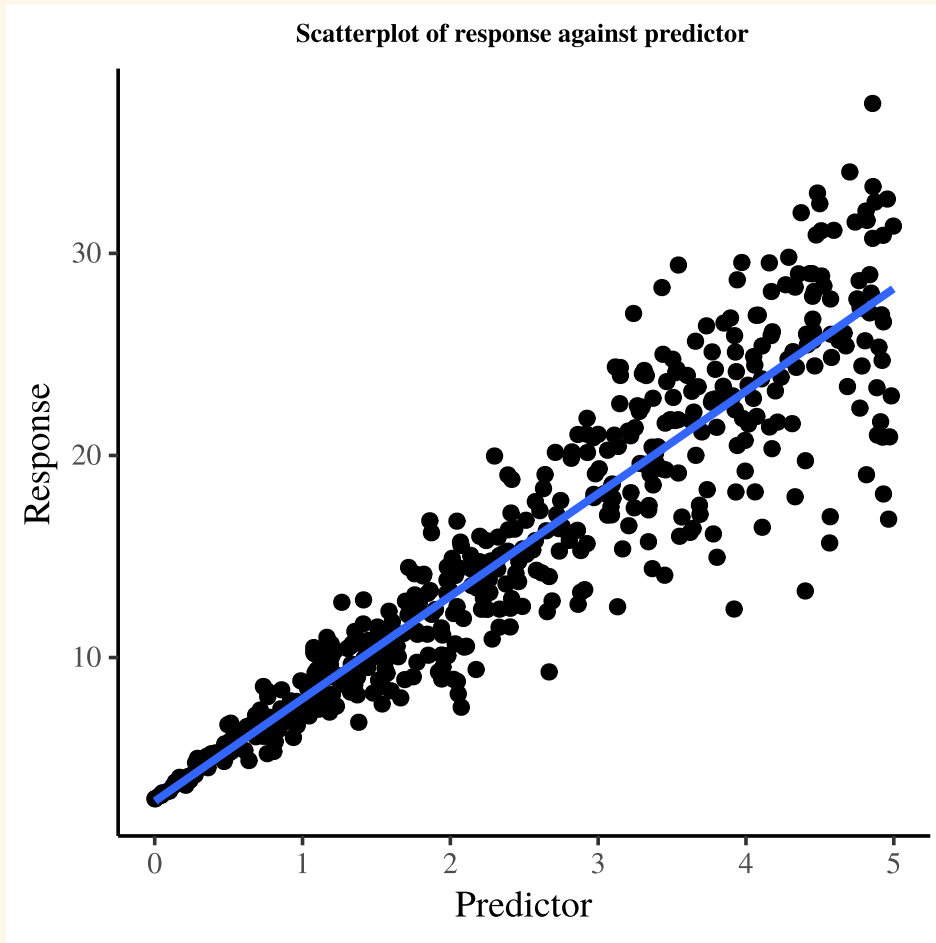# Residual plot: null plot (n=10)

Warning: residual plots can be hard to interpret if you don't have a lot of data!!

Residual plot: null plot (n=10)

Scatterplot of response against predictor

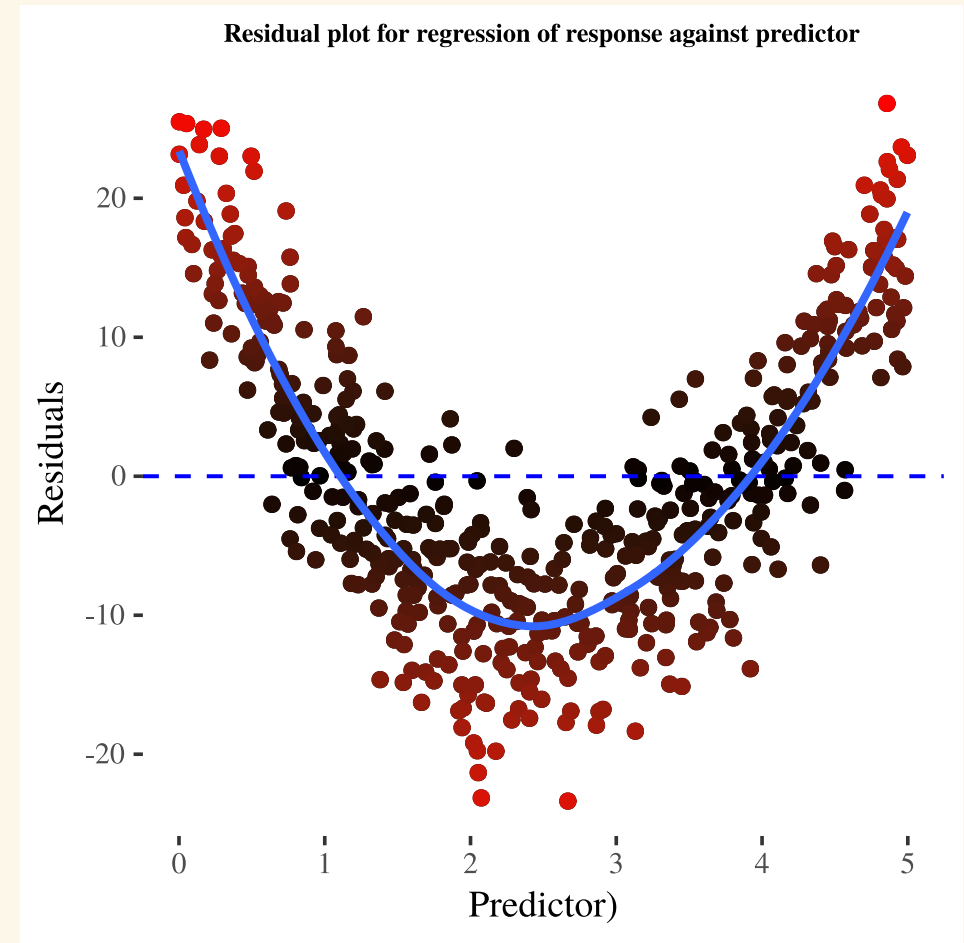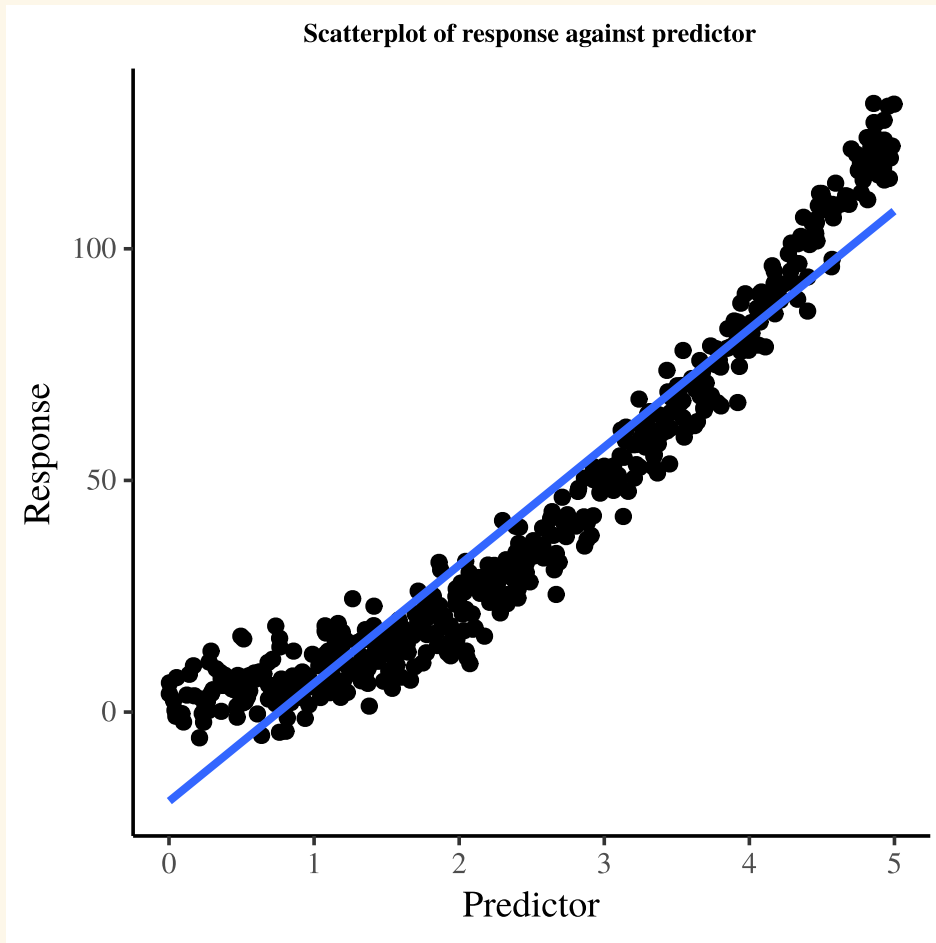Residual plot for regression of response against predictor
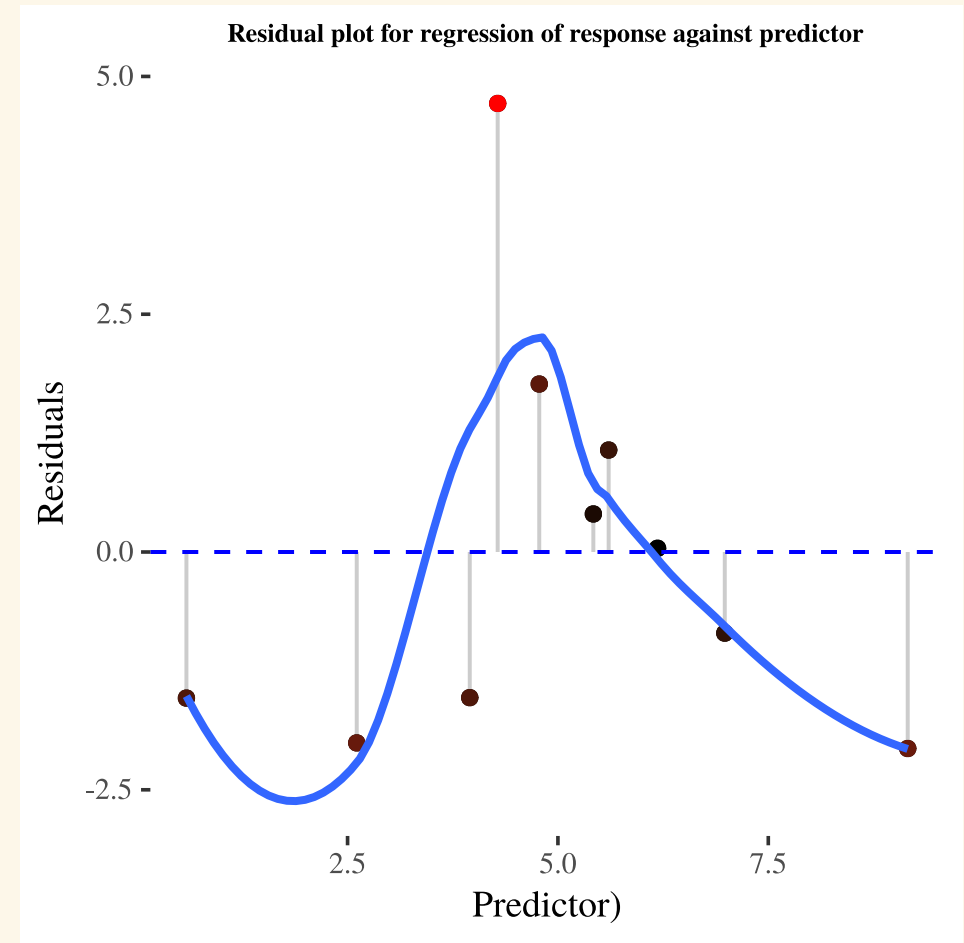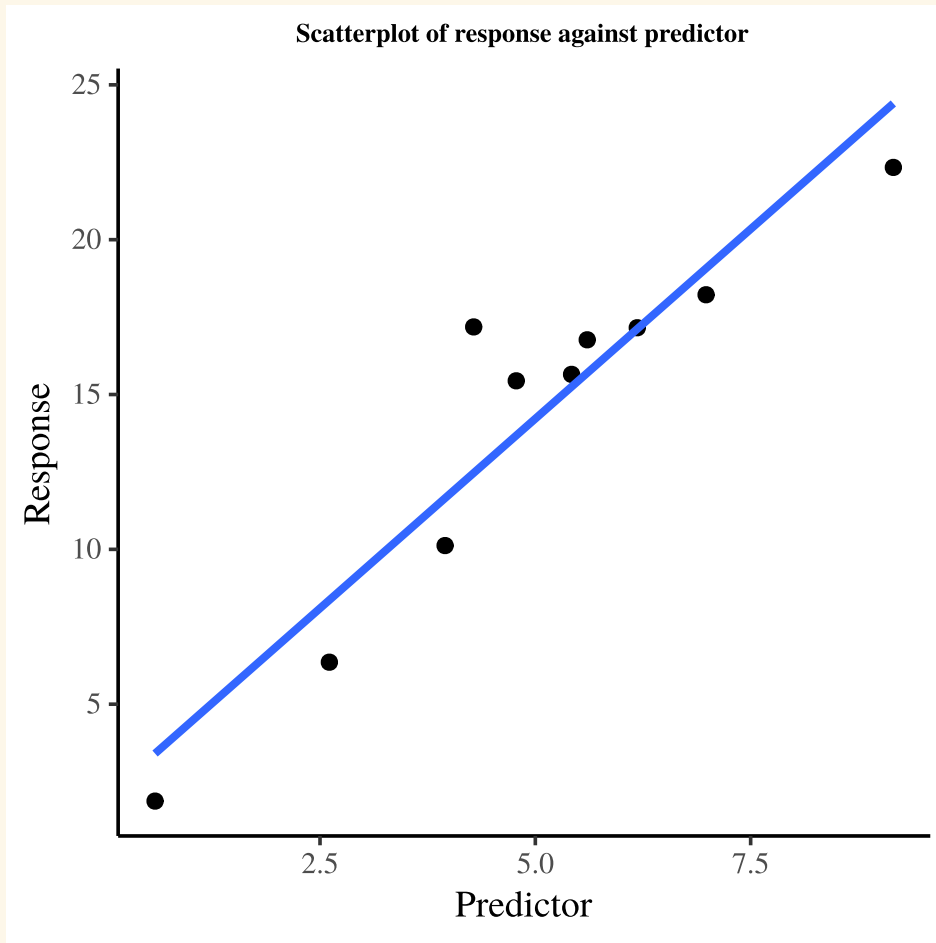
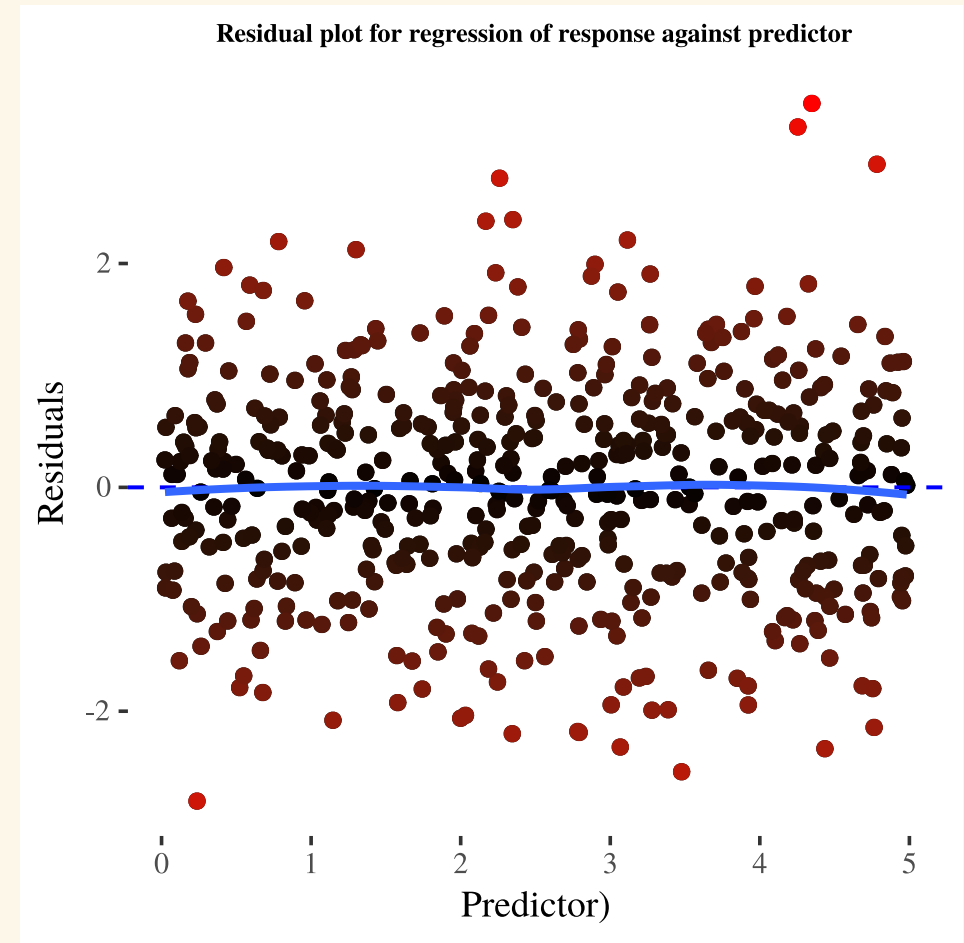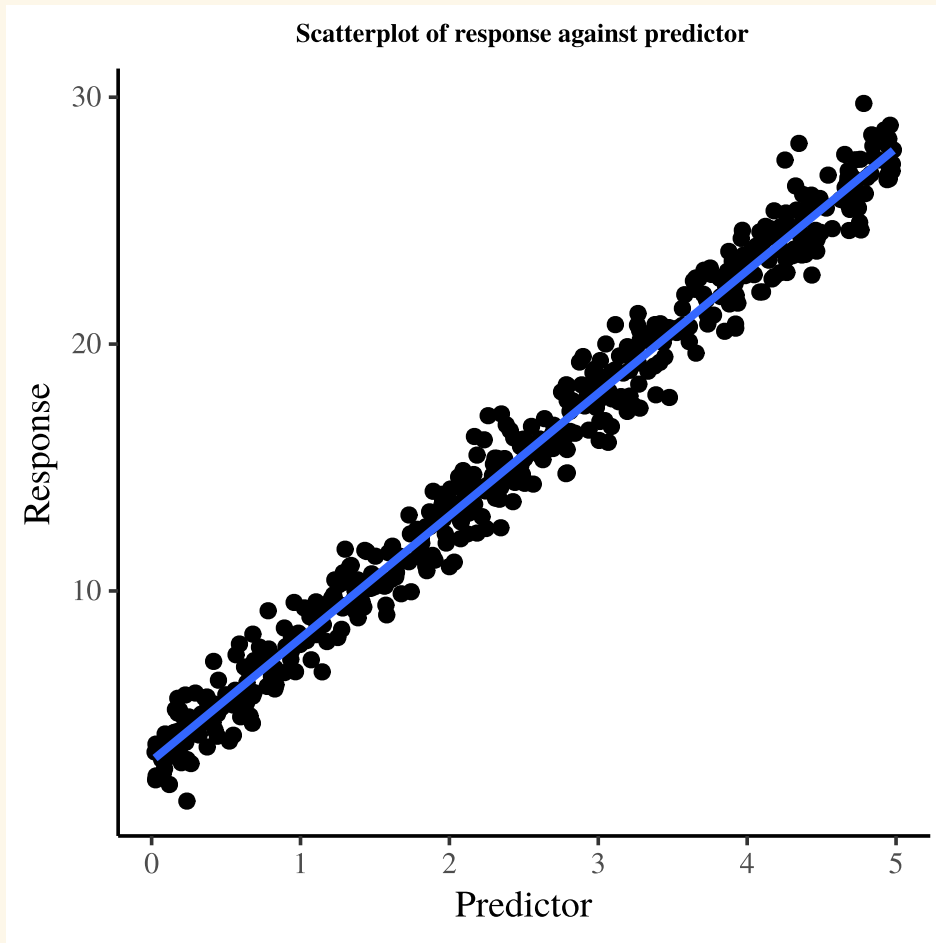Smoothers are very sensitive to individual points when data is sparse

Violation 1: Guess the assumption violation ...
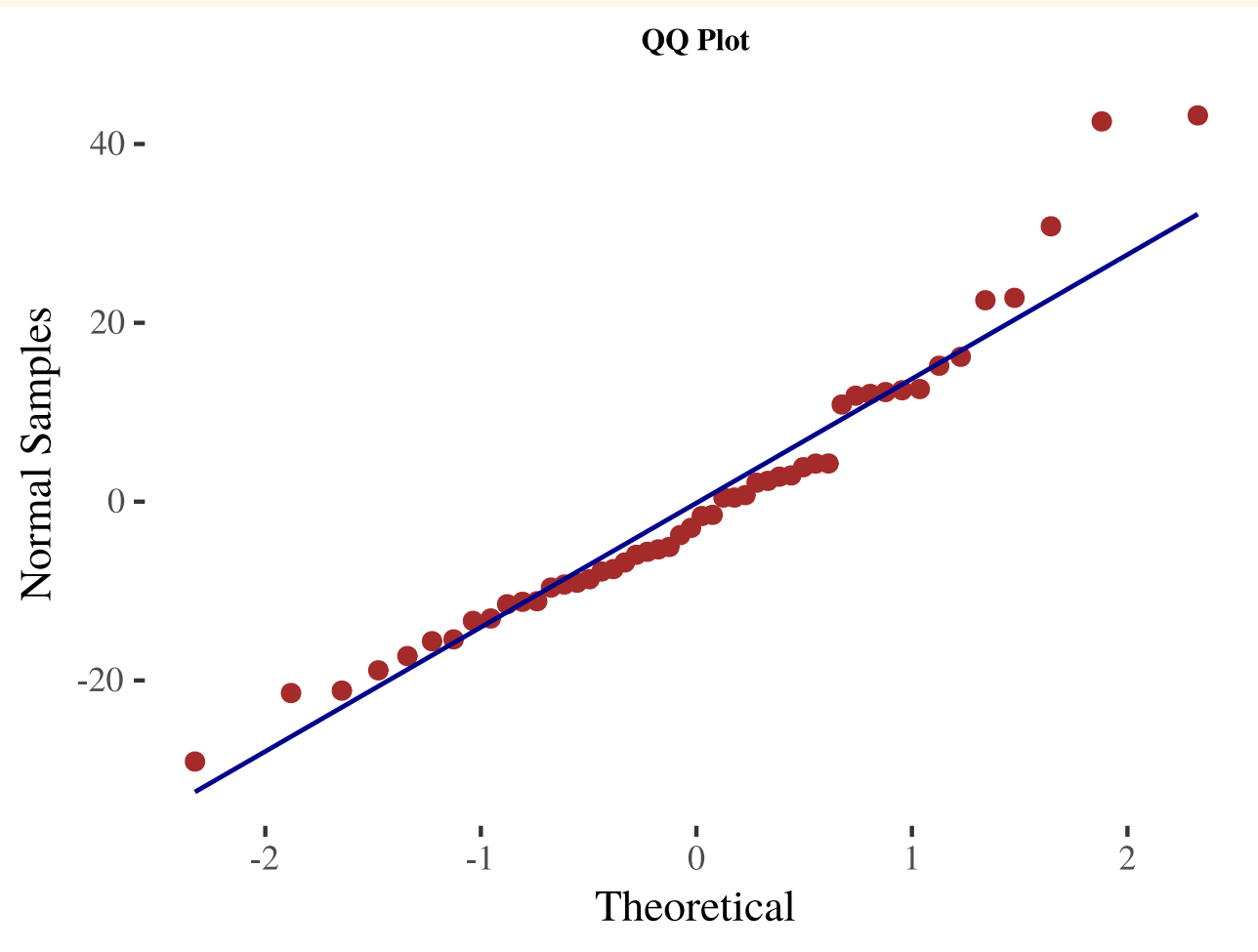
Violation 2: Guess the assumption violation ...
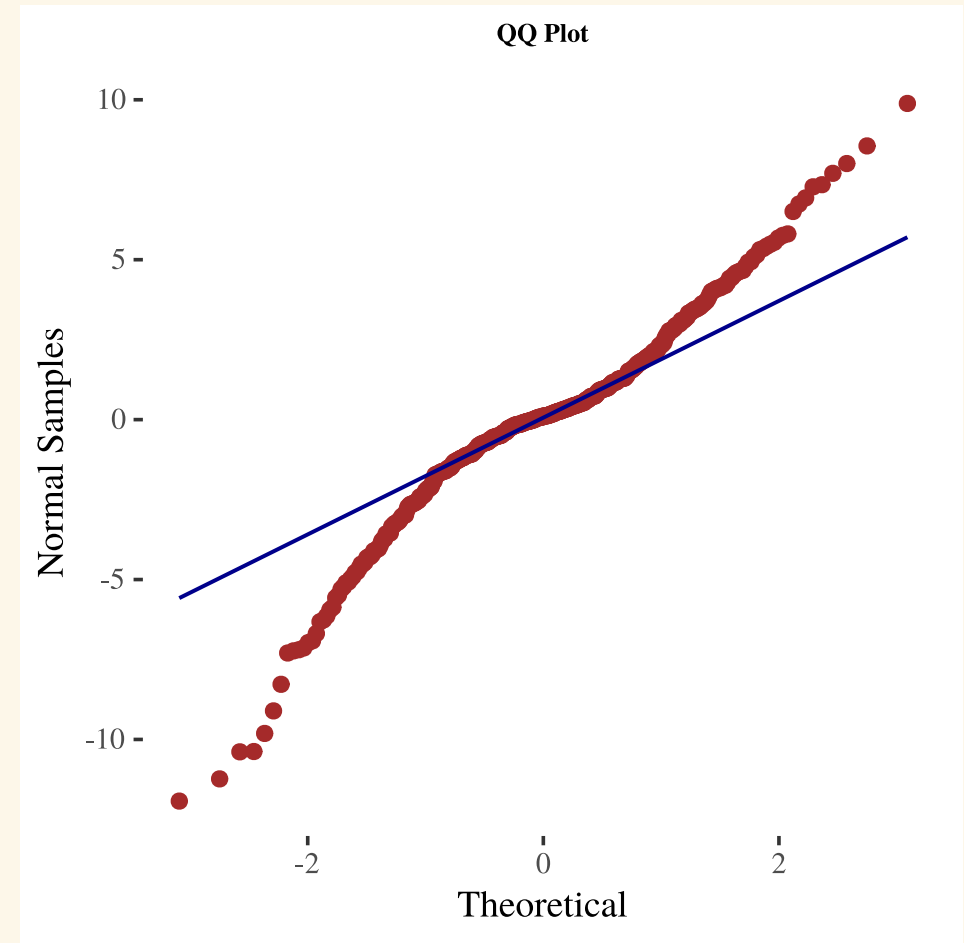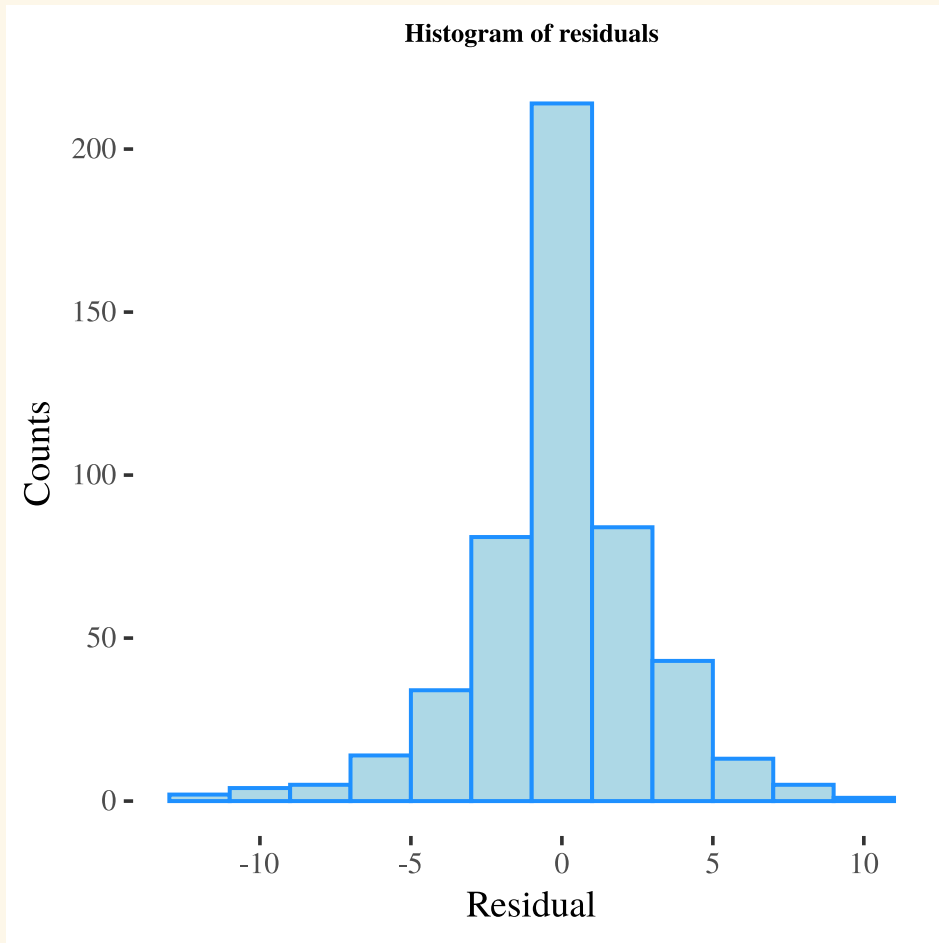
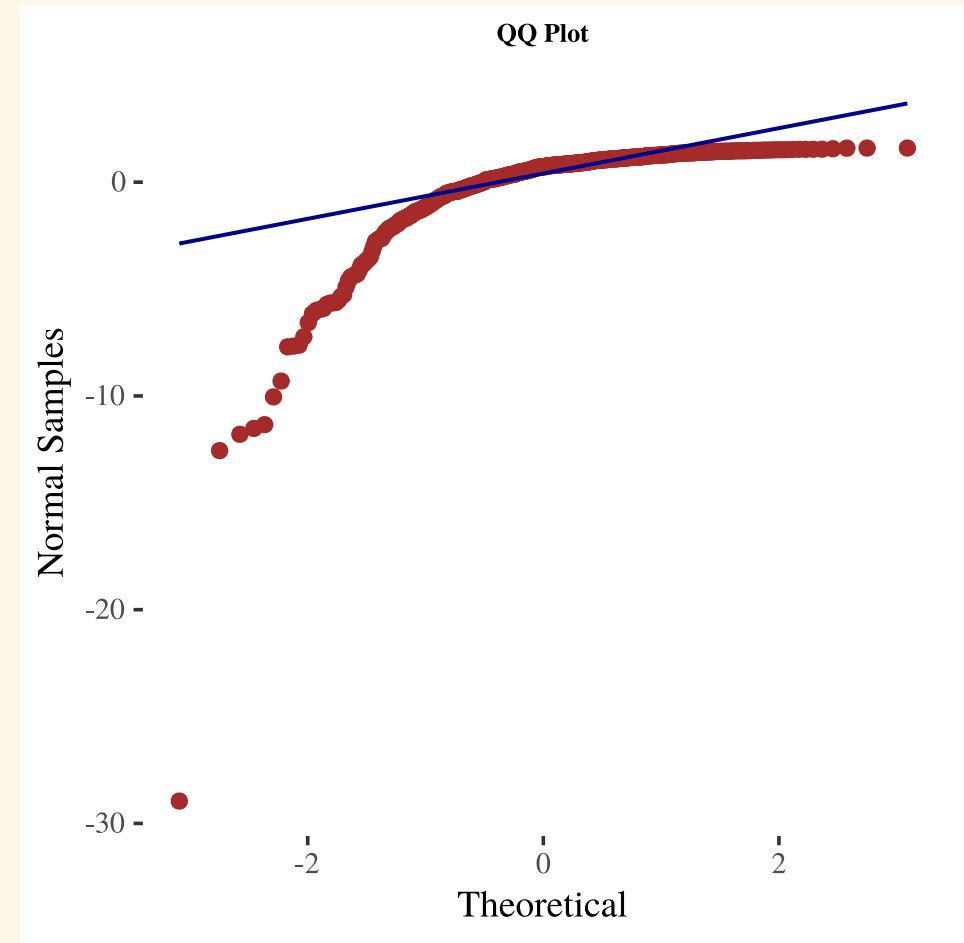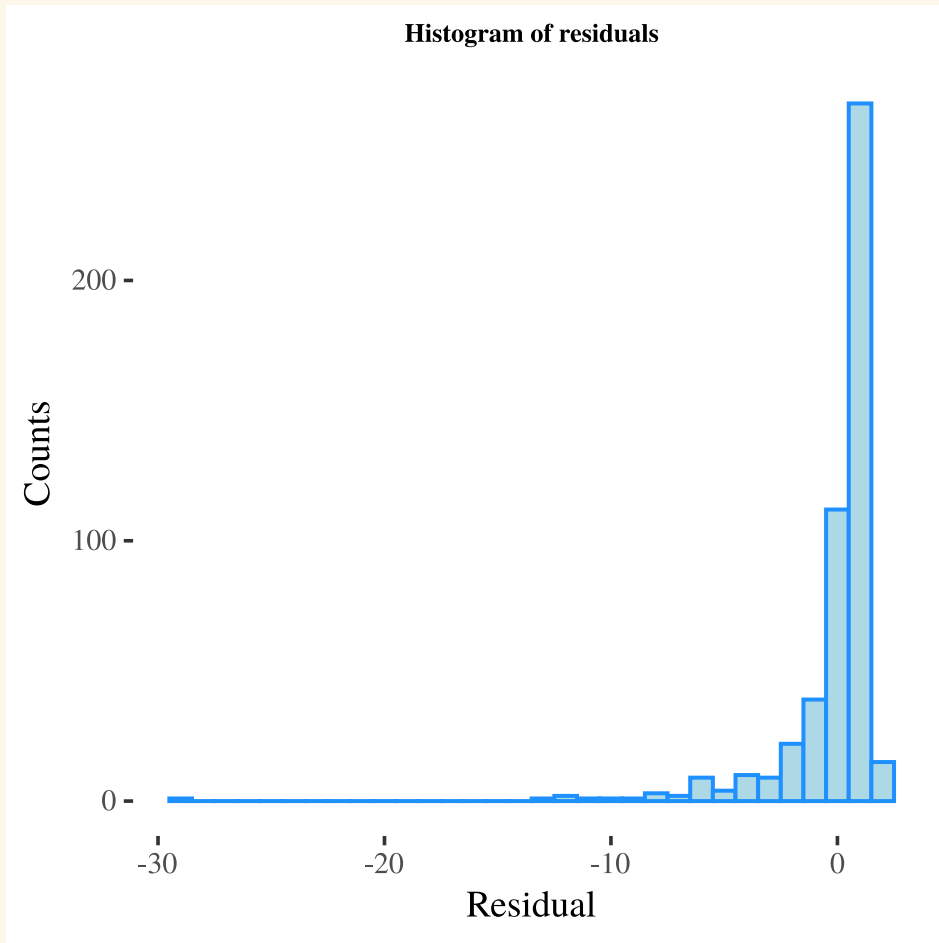# Good model

Q-Q Plot: Cars dataset

```r
ggplot(regression_points, aes(sample = residual)) +
 stat_qq(col = "brown") +
  stat_qq_line(col = "darkblue")+
  labs(
    title = "QQ Plot",
    x = "Theoretical",
    y = "Normal Samples") +
  theme(plot.title = element_text(hjust=0.5, size=7
```

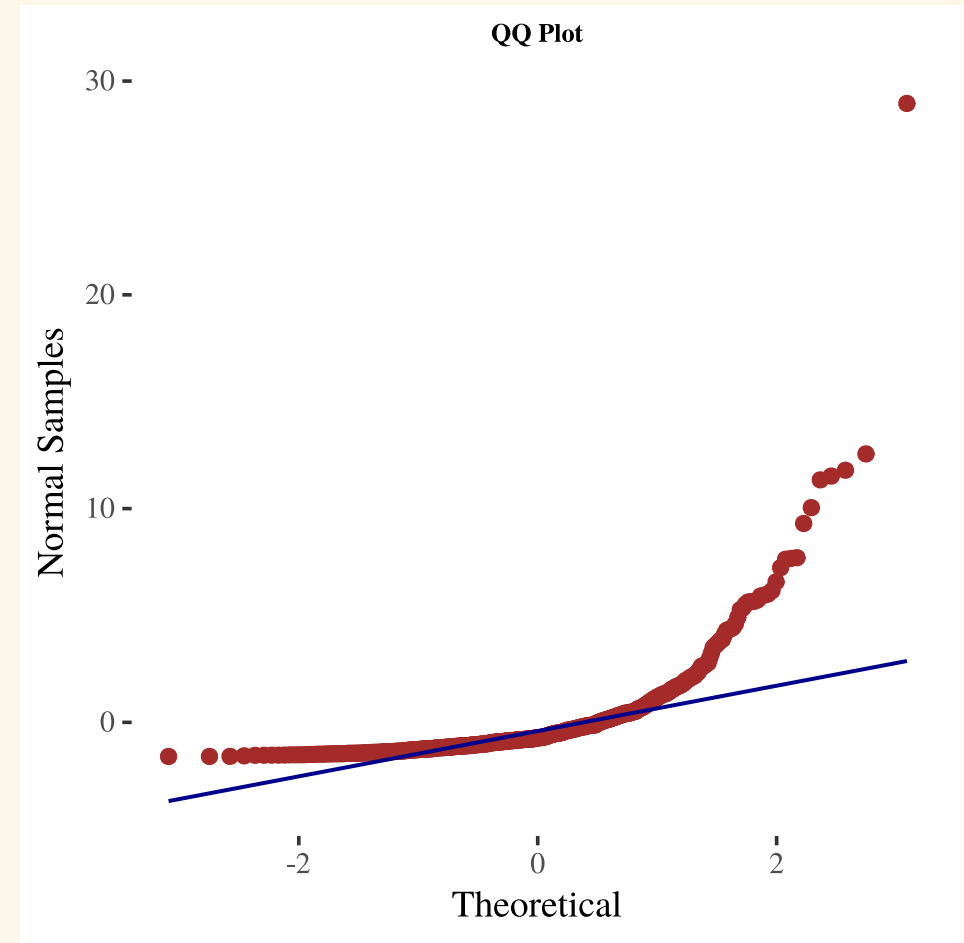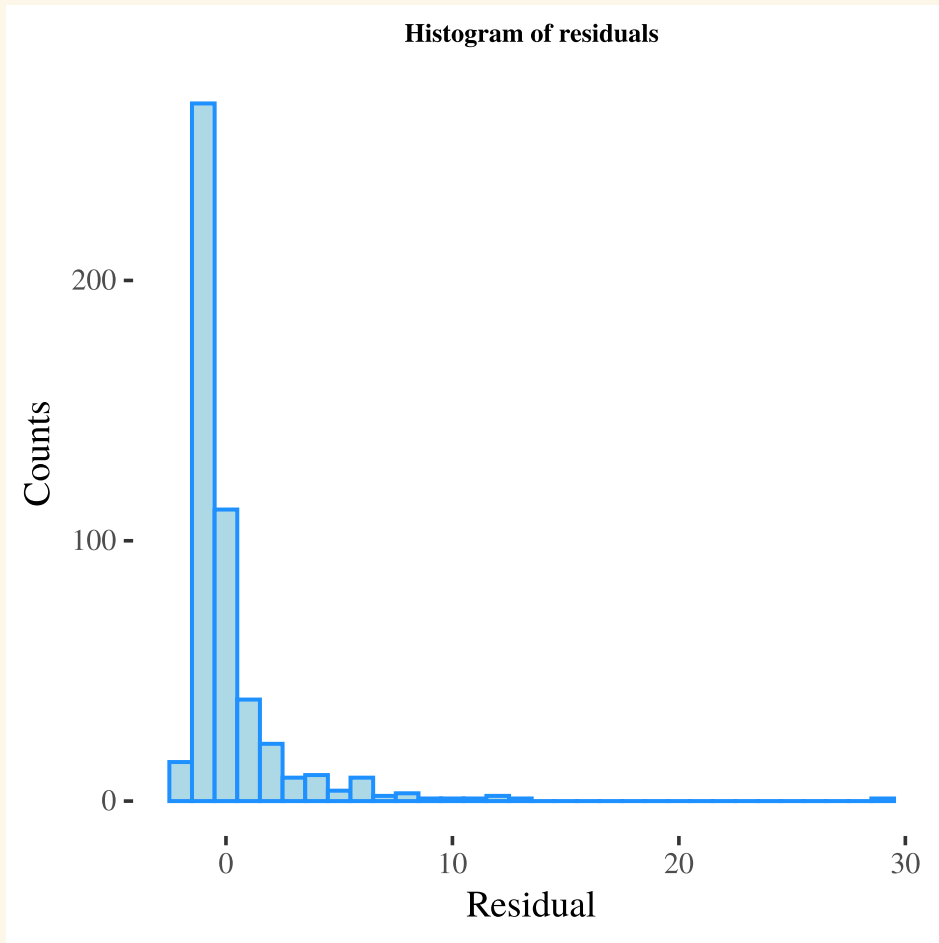# Normal Q-Q plot: Potential violations

Normal Q-Q plot: Potential violations
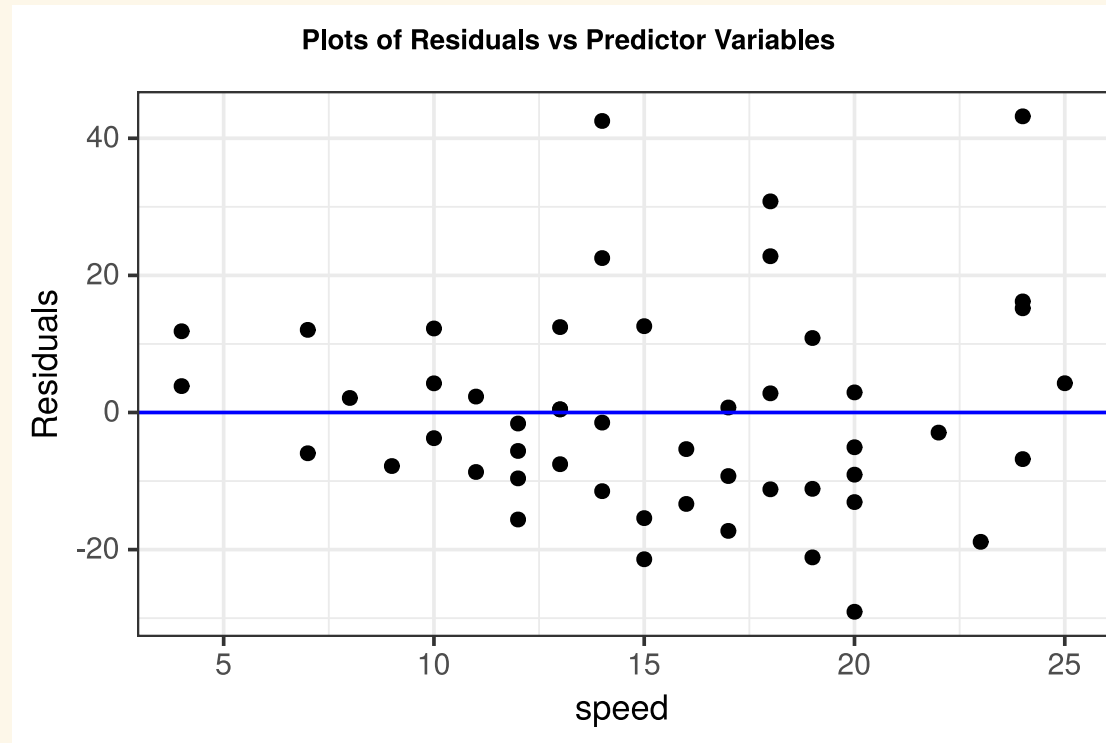

Concave down means left skewed!

Normal Q-Q plot: Potential violations

Concave up means right skewed!
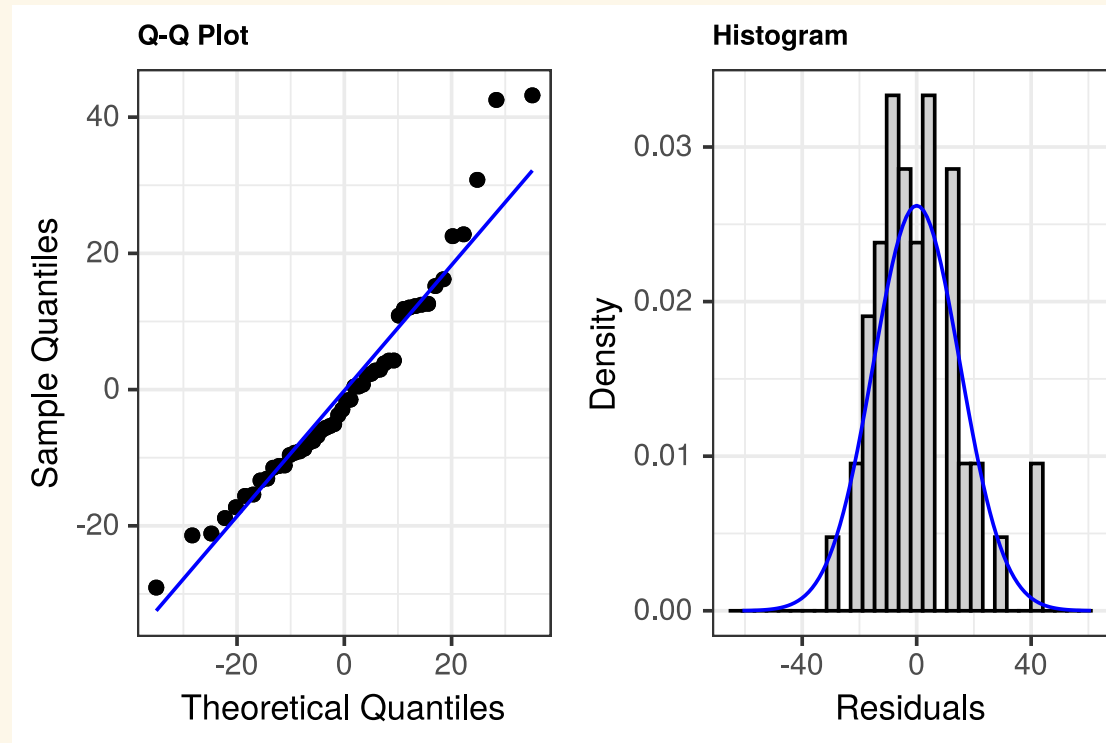
Diagnostics using `ggResidpanel` package

```r
library(ggResidpanel) # need to install if you aren't using maize
resid_xpanel(cars_lm, axis.text.size = 10, title.text.size = 8, scale = 1) # residual against
```

Diagnostics using ggResidpanel package

```r
# QQ plot and histogram of residuals
resid_panel(cars_lm, plots = c("qq", "hist"), axis.text.size = 10, title.text.size = 8, scale
```

Assessing independence

Can only verified through an understanding of how the data was collected

- do measurements vary over time?

- do measurements vary over a space (geographic region)?

- are measurements naturally clustered together?

- If the answer is "yes", then try plotting residuals against any variables that measure these attributes (time/space/clustering)
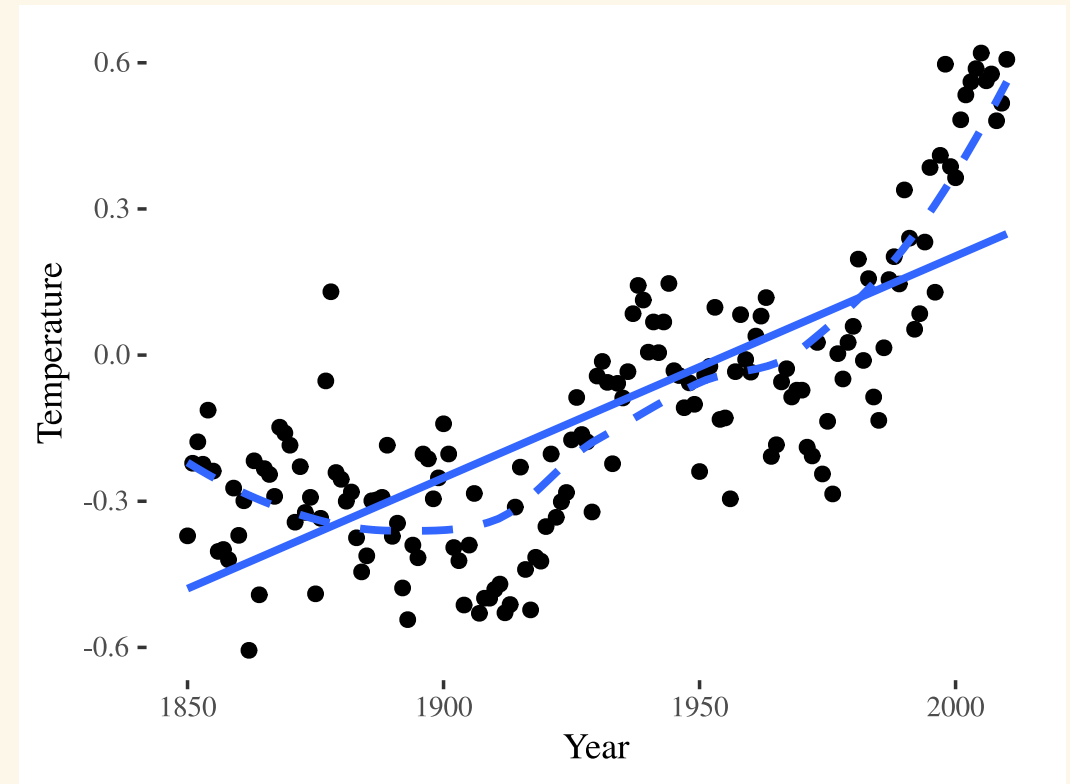
05:00
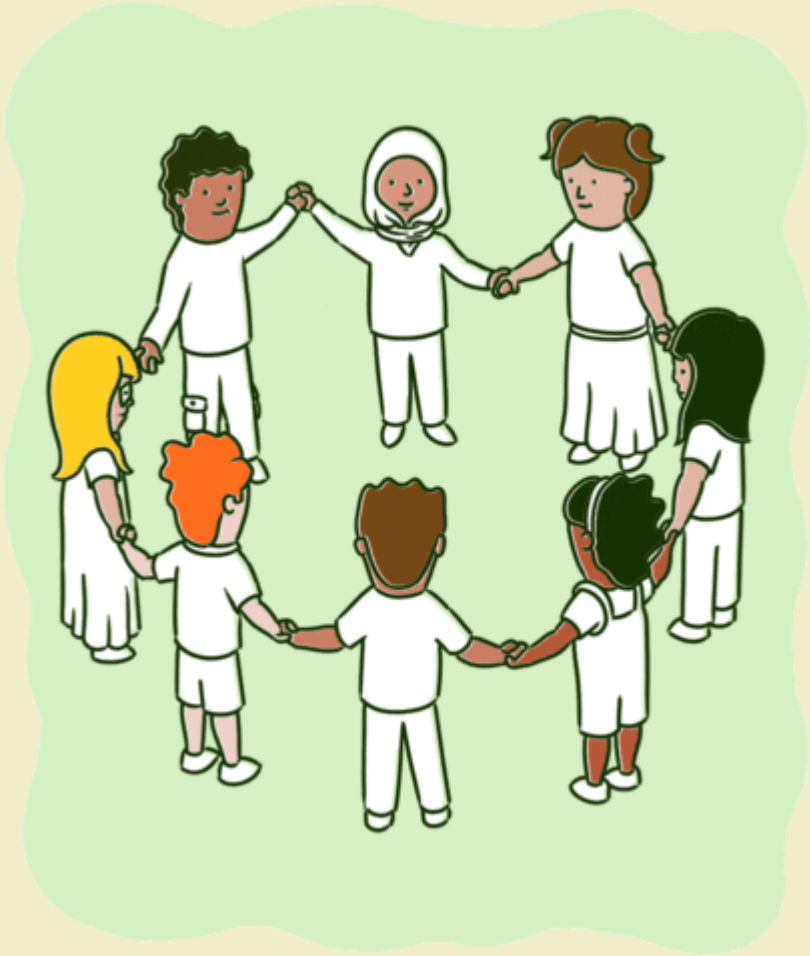
- Get the in class activity file from moodle

- We will do a case study of global warming

- Please skim through the activity .Rmd file in your group

Case Study 15.2 Global Warming

- $y =$ annual mean global temp

- $x =$ year

- $n = 161$ years from 1850 to 2010

- Model: SLR of global temp on year

Case Study 15.2 Global Warming

Regression of Temperature on Year
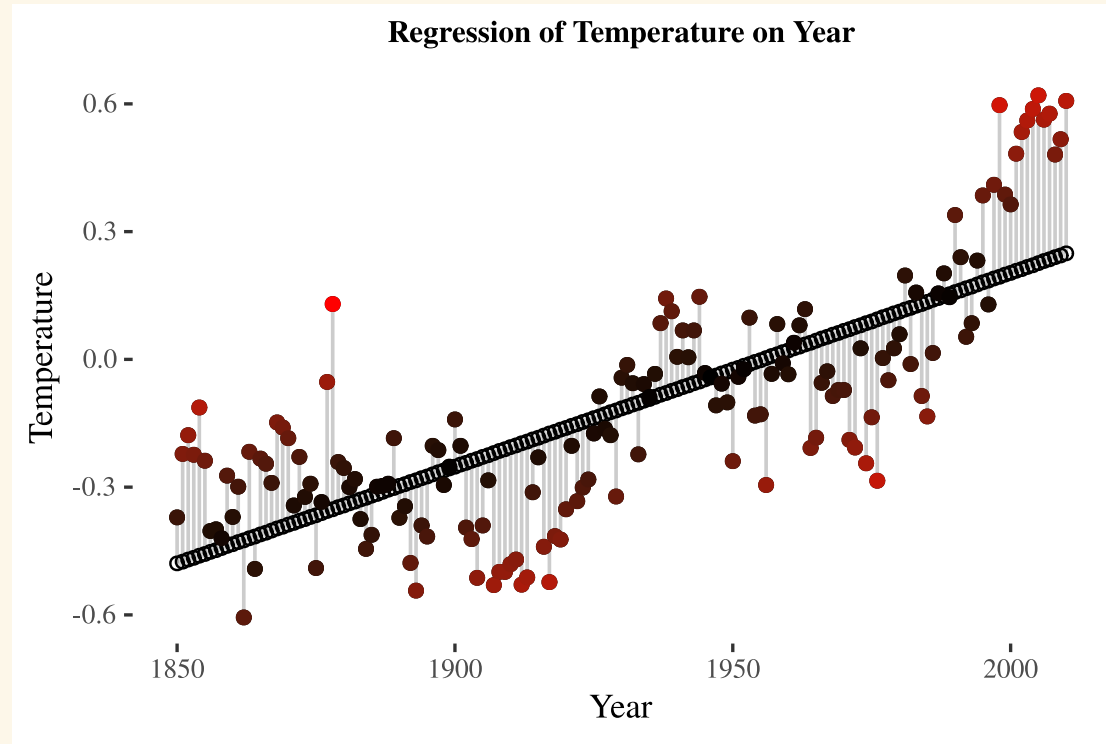
Residual plot for SLR of global temp on year

Regression of Temperature on Year
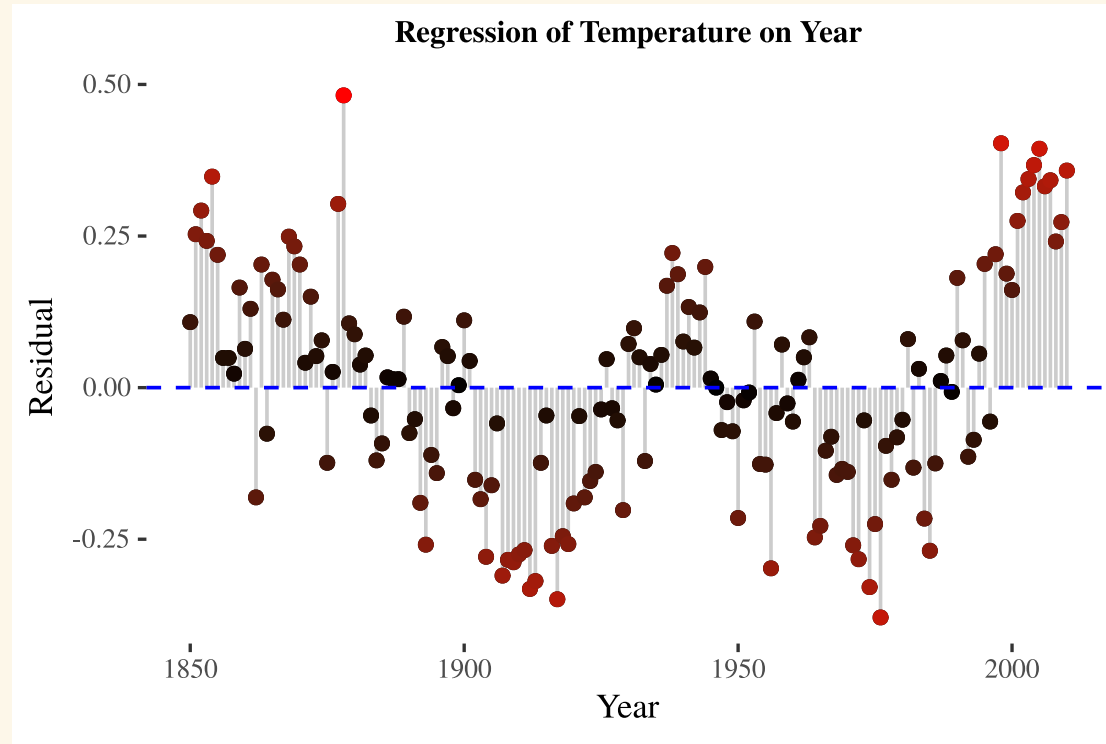
Residual plot for regression of global temp on $year$ and $year^2$
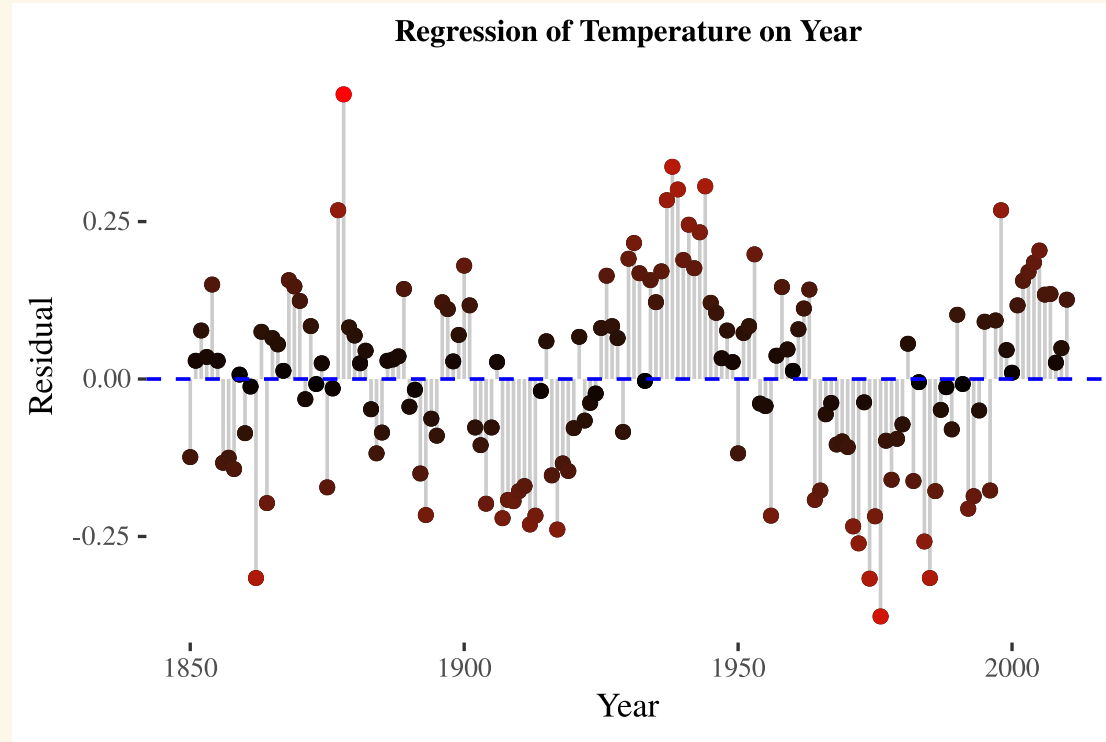
- Residuals that are 1 year apart have a(n) (auto)correlation coefficient of 0.59!
- Even after accounting for year, temps are still correlated. A more complex model is needed!

How "robust" is regression against violation of the assumptions?

Robust = can violate an assumption and still get valid inference results

SLR models are robust against violations of

- **Normality:** the t-tests and CI for model parameters and the mean response are saved by the Central Limit Theorem when n is large, even if your subpopulation of responses are not normally distributed.

How "robust" is regression against violation of the assumptions?

SLR models are not robust against violations of

- **Linearity:** if the mean function is wrong then your estimated effects, mean response, or predicted response will be biased!

- **Equal/Constant variance and independence:** if you are not correctly modeling your response variability, then your SEs will not be an accurate reflection of your actual uncertainty (meaning CIs/tests might be misleading)

- **Normality when computing prediction intervals:** these intervals need the normal subpopulation assumption to hold

What if model assumptions are violated?

Start by finding the correct mean function form!

- **Linearity, Variance, Normality:** Transform one or both variables
- **Linearity:** change mean function, use non-linear regression
- **Variance:** weighted regression, "robust SEs"
- **Independence:** use a mixed-effects model (Stat 330), times series (Stat 320) or spatial regression models