STAT302 Methods of Data Analysis 1
Module 3: Assumptions in Regression

Austin Brown

September 18, 2024

Lecture 1: Introduction to sampling distributions

Learning goals

- Introduce sampling distributions
- Introduce R Markdown and basic data visualization

Lecture 1: Sampling distributions

Extending our current definitions

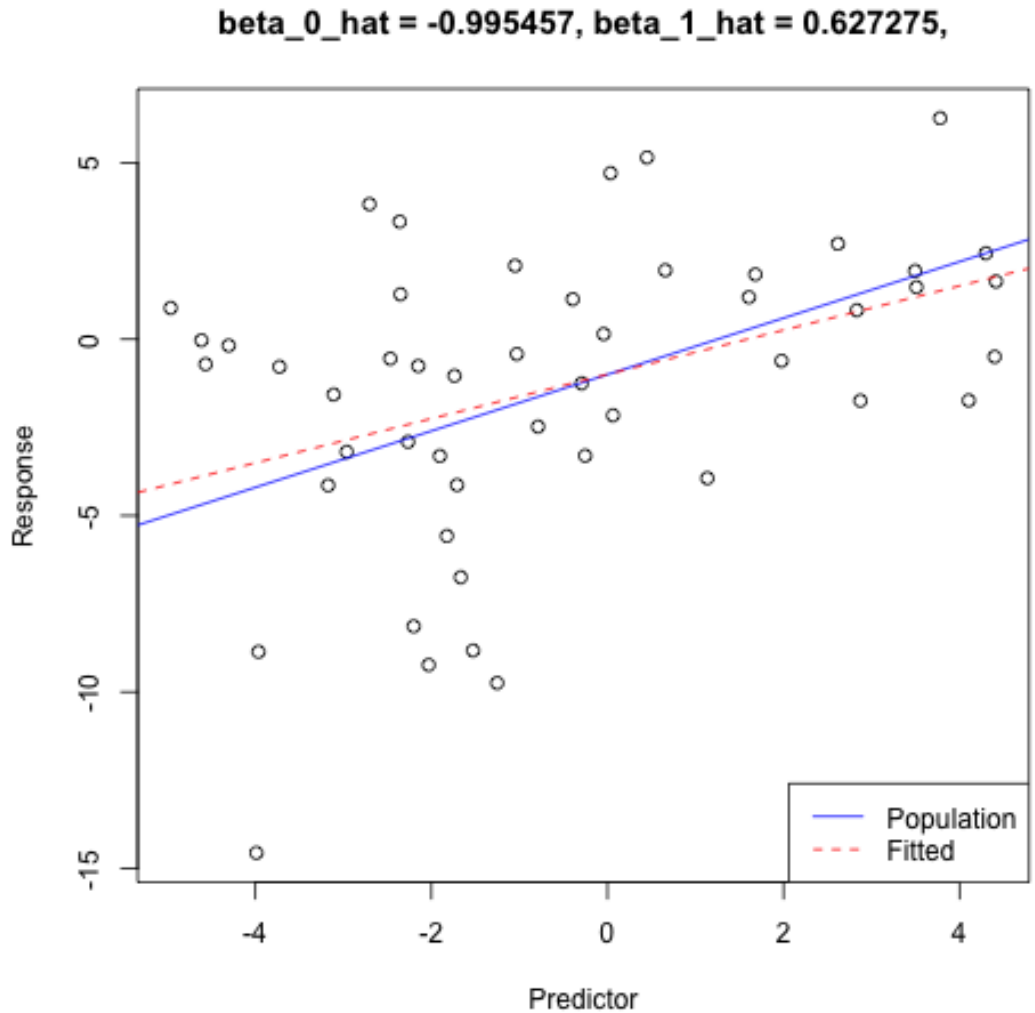
- We always assume our multivariate regression model.
- We have random **estimators** $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2$ depending on the regression model.
- We have computed **estimates** $\hat{\boldsymbol{\beta}} = \{value\}$, $\hat{\beta}_0 = \{value\}$ and $\hat{\beta}_1 = \{value\}$, $\hat{\sigma}^2 = \{value\}$ computed with the data.
- $\boldsymbol{b}^*$ is a computed value or realized value of $\hat{\boldsymbol{\beta}}$.

Terminology

- $\hat{\boldsymbol{\beta}}, \hat{\sigma}^2$ random depending on the regression model assumptions
- $\hat{\boldsymbol{\beta}} = \{\text{some value}\}$: The theoretical $\hat{\boldsymbol{\beta}}$ is computed at a value from a data.
- Given a dataset, compute $\hat{\boldsymbol{\beta}}, \hat{\sigma^2}$: means compute using the given data.

Sampling distribution visualization

Population: $Y_i = -1 + .8x + e_i, n = 50$ generated multiple times
Compute from the data: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$



Simple linear regression: sampling distributions

$Y_1, \ldots, Y_n$ are random from the regression model assumptions and $x_1, \ldots, x_n$ are fixed. The theoretical simple least squares solutions are **random**:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$$

The distribution of these estimators is their **sampling distribution**.

The computed $\hat{\beta}_0, \hat{\beta}_1$ from the data we know.

Multiple linear regression: sampling distributions

The regression model assumes $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Then $\boldsymbol{Y}$ is random, $\boldsymbol{X}$ is fixed and

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

is **random** with a **sampling distribution**.

The computed $\hat{\boldsymbol{\beta}}$ from the data is a vector that we know.

Estimator for σ^2

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_p x_{i,p}$ is **random** with a **sampling distribution**.

- We use $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ for estimating the standard deviation
- The computed $\hat{\sigma}^2$ from the data is a number we know.

Summary

- **Regression model:** $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$
- **Population parameters/vectors:** $\boldsymbol{\beta}$, σ^2
- **Theoretical estimators (Random):** $\hat{\boldsymbol{\beta}}, \hat{\sigma^2}$
- **Sampling distribution:** Distribution of $\hat{\boldsymbol{\beta}}, \hat{\sigma^2}$
- **Data:** $(x_1, y_1), \ldots, (x_n, y_n)$
- **Estimates from the data (Not random):** Compute values for $\hat{\boldsymbol{\beta}}$, $\hat{\sigma^2}$ using the data.

Remark: Compare this to basic statistics in Module 1.

Lecture 1: Running example

Menu pricing for a new restaurant in NYC

From [Sheather, 2009]:
Imagine that you have been asked to join the team supporting a new chef who plans to create a new Italian restaurant in Manhattan. The aims of the restaurant are to provide the highest quality Italian food utilizing state-of-the-art décor while setting a new standard for high-quality service in Manhattan. You have been told that the restaurant is going to be located no further south than the Flatiron District and it will be either east or west of Fifth Avenue.

You have been asked to determine the pricing of the restaurants dinner menu such that it is competitively positioned with other high-end Italian restaurants in the target area.

Menu pricing for a new restaurant in NYC

168 Italian restaurants in the target area from Zagat Survey 2001: New York City.

- Price = the price (in USD) of dinner (including one drink and a tip)
- Food = customer rating of the food (out of 30)
- Décor = customer rating of the decor (out of 30)
- Service = customer rating of the service (out of 30)
- East = 1 (0) if the restaurant is east (west) of Fifth Avenue

```
nyc_restaurant_data = read.csv("https://gattonweb.uky.
    edu/sheather/book//docs/datasets/nyc.csv", header =
     T)
```

R Markdown

- Introduce R Markdown

Lecture 2: Regression model diagnostics

Learning goals for Lecture 2

- Introduce residuals, their properties, and plotting residuals
- Checking the linearity assumption in regression models using the residuals

Reminder: the multivariate regression model

Recall the **multiple linear regression model** is

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + e_i$$

for $i = 1, \ldots, n$ where

- $\boldsymbol{\beta}$: parameter vector of dimension $p + 1$
- σ^2: error variance parameter
- $e_i \sim N(0, \sigma^2)$ i.i.d. errors

Regression model assumptions

- **Linearity assumption**:

$$E(Y|\boldsymbol{X} = \boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p$$

- **Constant error variance (homoscedasticity)**:

$$\mathsf{Var}(Y|\boldsymbol{X} = \boldsymbol{x}) = \sigma^2$$

- **Independent and normal errors**: e_i independent $e_i \sim N(0, \sigma^2)$.

Regression model assumptions: linearity

$$E(Y|\boldsymbol{X} = \boldsymbol{x}) =$$

Regression model assumptions: homoscedasticity

$$\mathsf{Var}(Y|\boldsymbol{X} = \boldsymbol{x}) =$$

No perfect diagnostic

- There is no perfect diagnostic for checking the regression modeling assumptions.
- Diagnostic plots may not capture violations
- Diagnostic plots may be misleading, especially when the sample size is small
- Diagnostic plots are generally more trustworthy when the number of observations is large
- Violation of one assumption may affect the diagnostic plot for a separate assumption.

Lecture 2: Residuals

Residuals

The theoretical **residual** under the multivariate linear regression model is random:

$$\hat{e}_i = Y_i - \hat{Y}_i.$$

The computed residual from the data is a number. i.e. $\hat{e}_i = \{\text{value}\}$.

Residuals

Given the predictors, the theoretical residuals satisfy

$$E(\hat{e}_i|\boldsymbol{X}) = 0$$
$$\text{Var}(\hat{e}_i|\boldsymbol{X}) = \sigma^2(1 - h_{i,i}).$$

- $h_{i,i}$ are the diagonal entries of the **hat matrix** $H = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.
- Generally $h_{i,i}$ are small
- The residuals $\hat{e}_i$ are not independent and are **correlated**.
- If the errors e_i are normally distributed, then the residuals are normally distributed (but not standard normal).

Compute the residuals with R:

```
resid(fit)
```

Standardized residuals

Standardized residuals are random

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{i,i}}}$$

- The standardized residuals are correlated.
- The standardized residuals should be **approximately** $N(0, 1)$ but not exactly.

Compute the standardized residuals with R:

```
rstandard(fit)
```

Residuals for regression diagnostics

- When the model is correct, the residuals are uncorrelated with the fitted values and predictors.
- When the model is correct, the scatterplot of residuals against the fitted values, or any predictors should have constant mean 0. The residual plot should look like a null plot (no systematic pattern).
- When the model is correct, the variance of the residuals is not quite constant but should be generally close to 1.
- When the model is correct, the residuals are correlated, but this is often not visible in residual plots.
- When the model is correct, the standardized residuals should constant mean 0, are correlated, and have roughly a normal distribution $N(0, 1)$.

Residual plots

The plot that are commonly used are:

- Plot each **residual** against its corresponding fitted value
- Plot each **standardized residual** against its corresponding fitted value

We could look at many other plots as well:

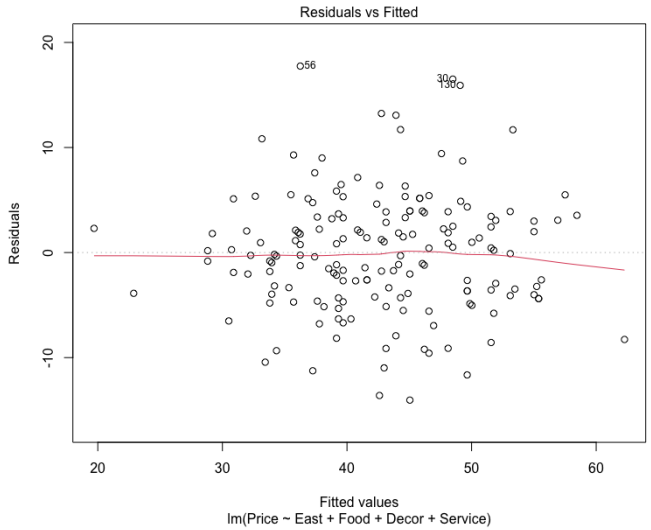
- Plot histograms of standardized residuals
- Plot each residual against a predictor variable
- Plot each residual
- Plot each residual against a predictor variable

Scatterplots

How to make a scatter plot

```
plot(x_values, y_values,
     main = "TITLE",
     xlab = "X LABEL",
     ylab = "Y LABEL")
```

NYC example: creating residual plots

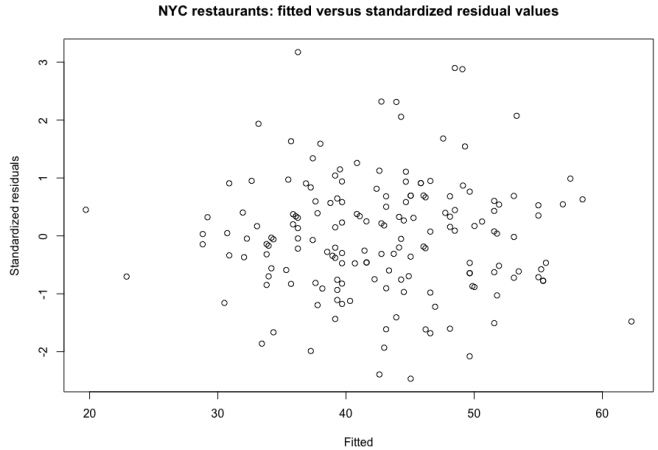



```
plot(fit, which=1)

fitted_values = fitted(fit)
residual_values = resid(fit)

plot(fitted_values, residual_values, main = "NYC
    restaurants: fitted versus residual values", xlab =
     "Fitted", ylab = "Residuals")
```

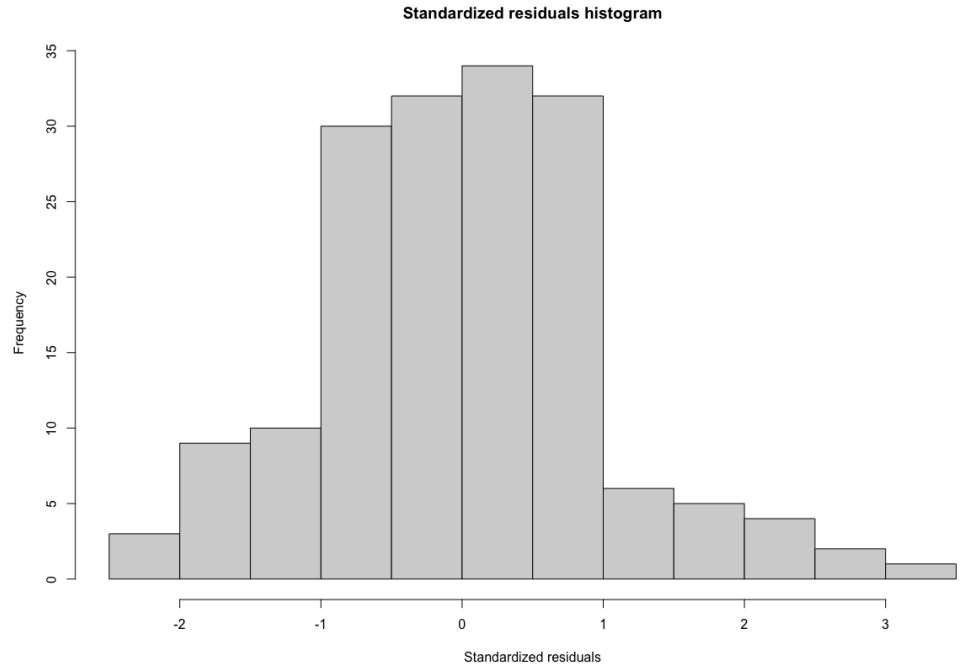
NYC example: creating standardized residual plots




```
fitted_values = fitted(fit)
sresidual_values = rstandard(fit)

plot(fitted_values, sresidual_values, main = "NYC
    restaurants: fitted versus standardized residual
    values", xlab = "Fitted", ylab = "Standardized
    residuals")
```

NYC example: creating histograms for standardized residual




```
sresidual_values = rstandard(fit)
hist(sresidual_values, xlab = "Standardized residuals",
     main = "Standardized residuals histogram")
```

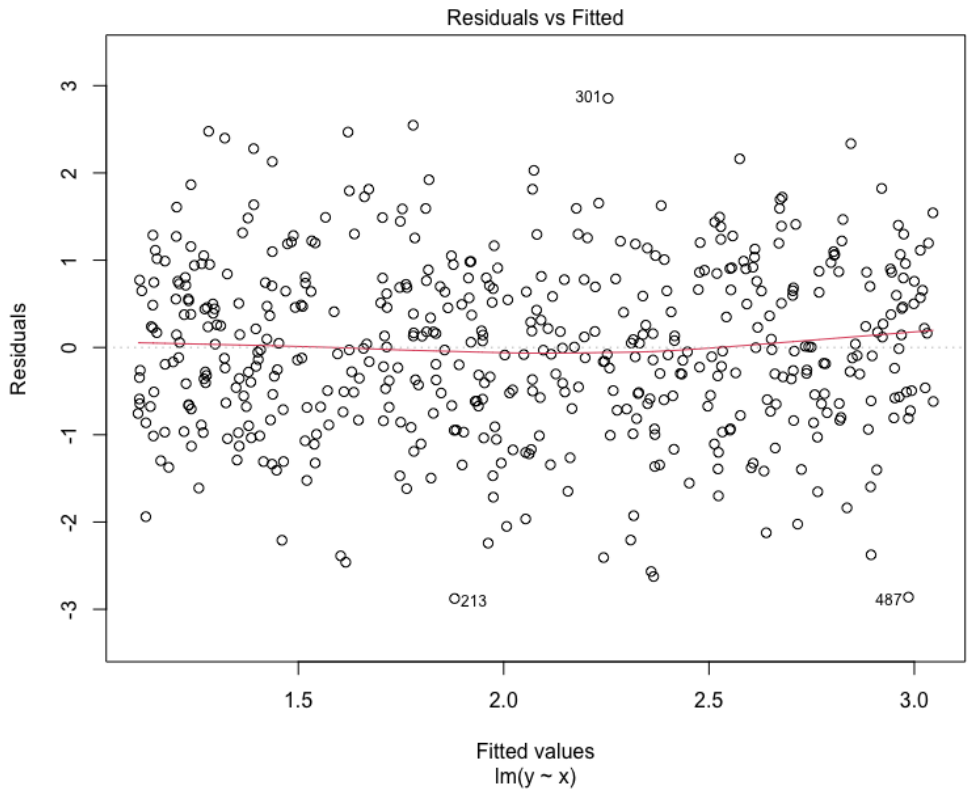
Lecture 2: Checking the linearity assumption

Types of plots

- Residuals against fitted values and against predictors
- Response against fitted values
- Response against predictors

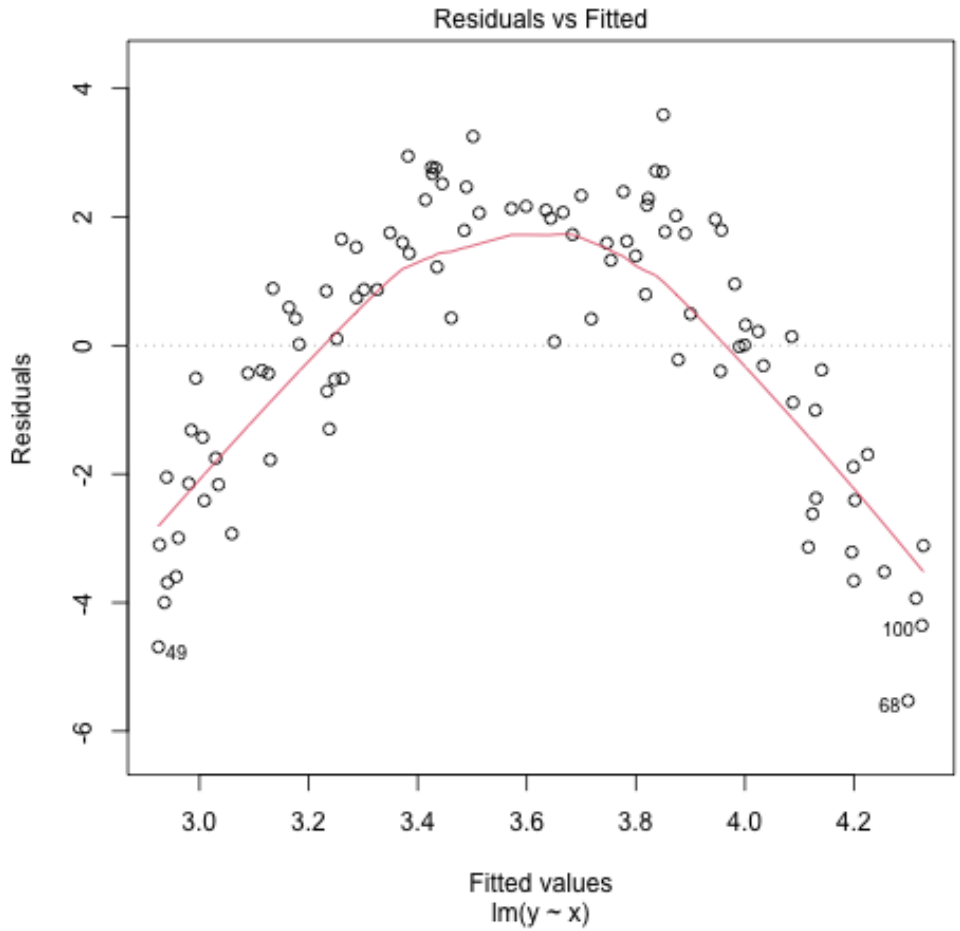
Plot: residuals against fitted values

Under the linearity assumption model, the residuals and the fitted values are uncorrelated. We should see approximately null plots (no systematic pattern) for residual versus fitted and standardized residual versus fitted.



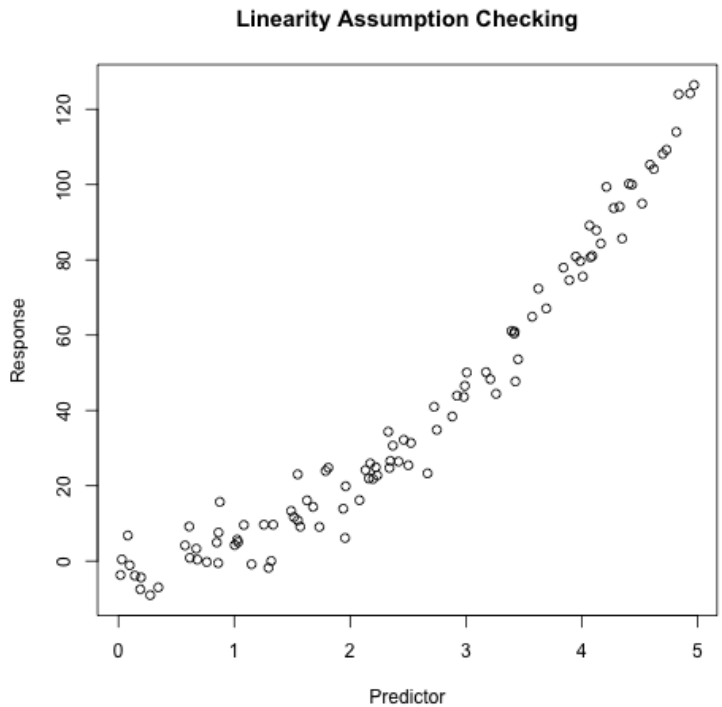
Plot: residuals against fitted values

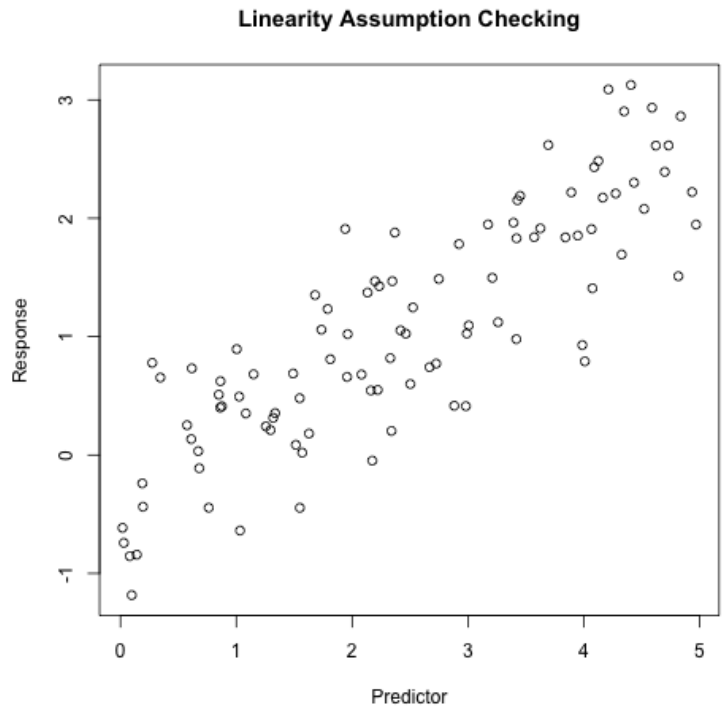
Does this provide evidence for or against the linearity modeling assumption?



Plot: response against predictor

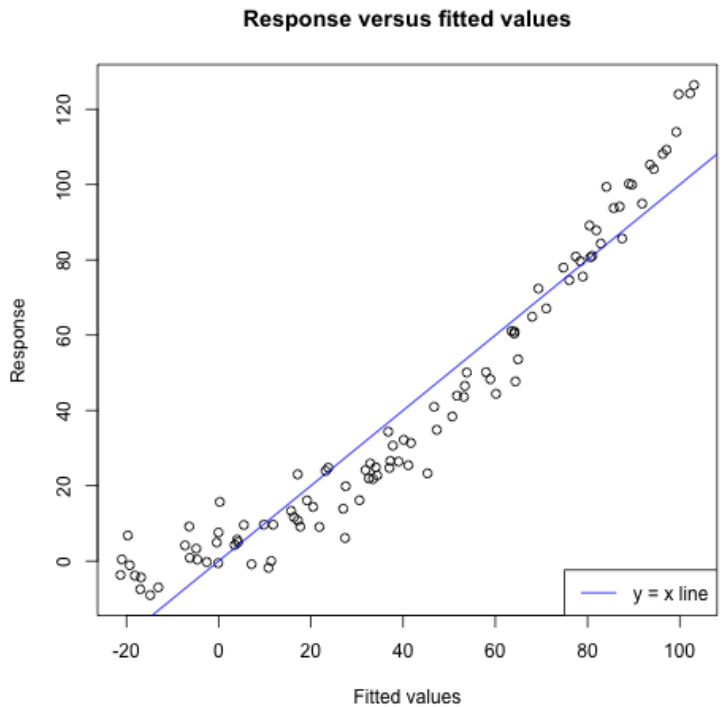
Which one of these provides evidence **against** the linearity assumption?

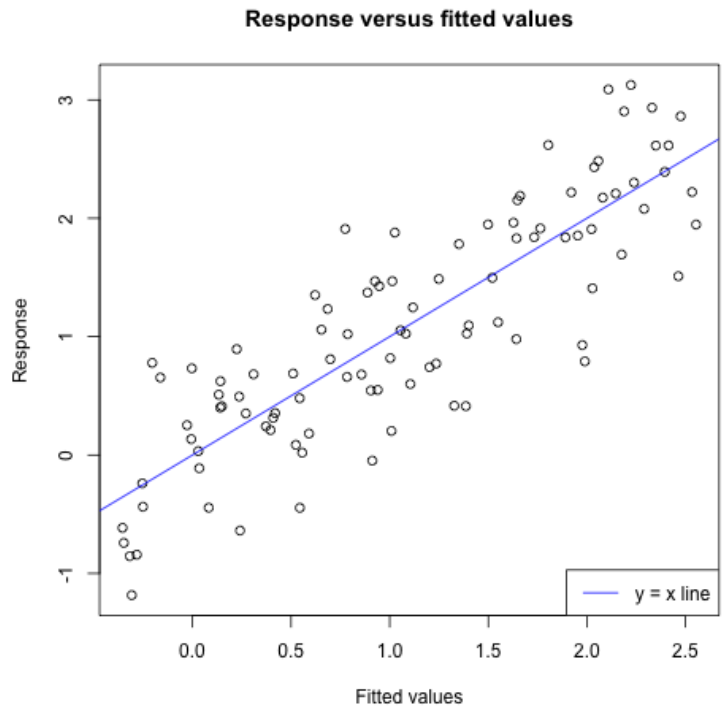




Plot response against fitted values

Which one of these provides evidence **against** the linearity assumption?



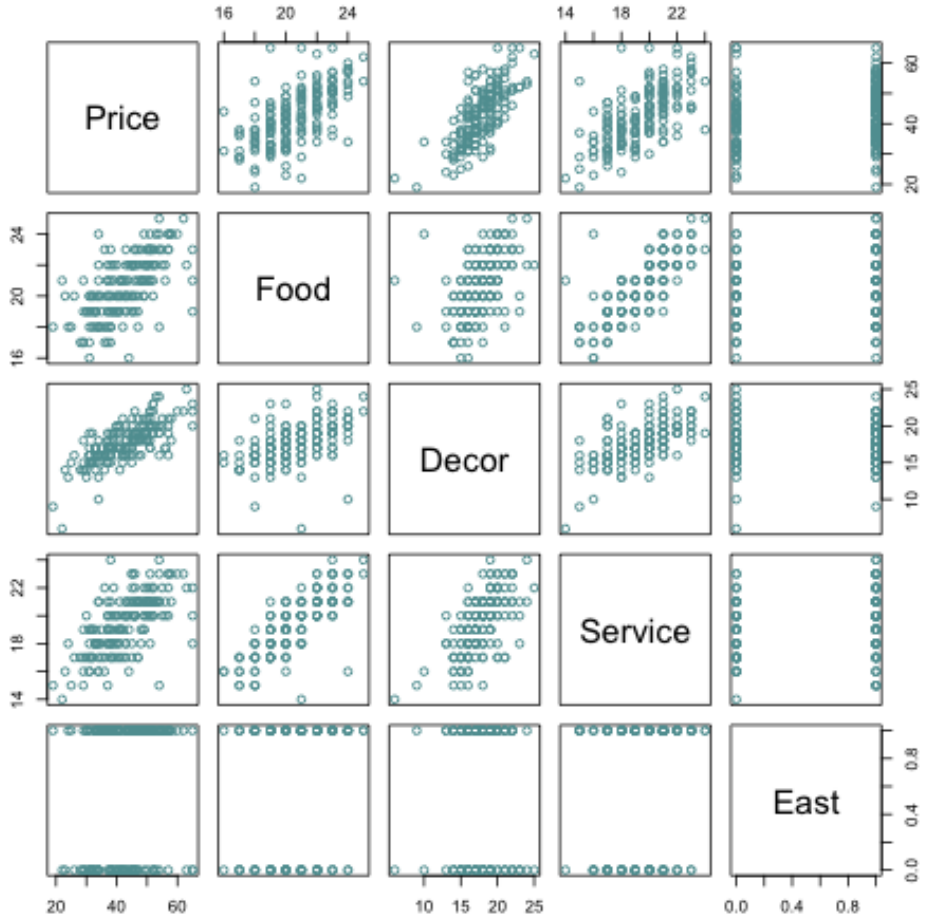


Example NYC data

Let's apply this to our example.

Scatterplot matrix

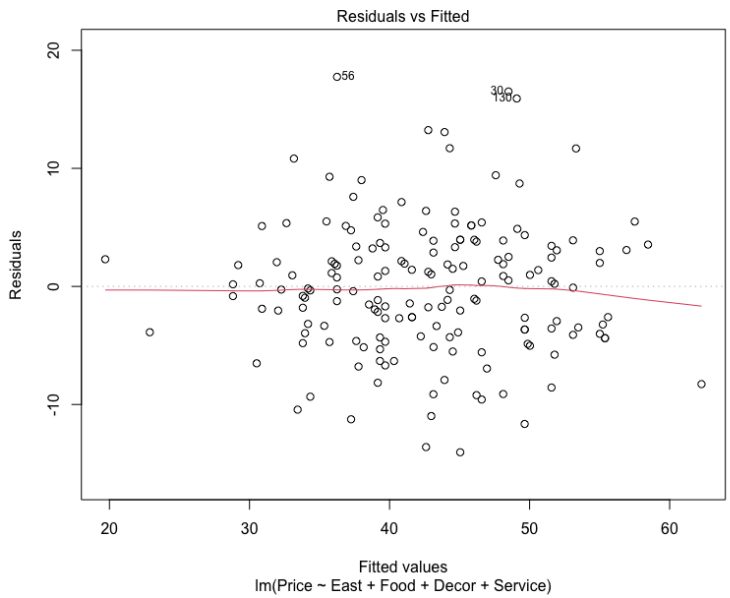
Do we see non-linear trends?

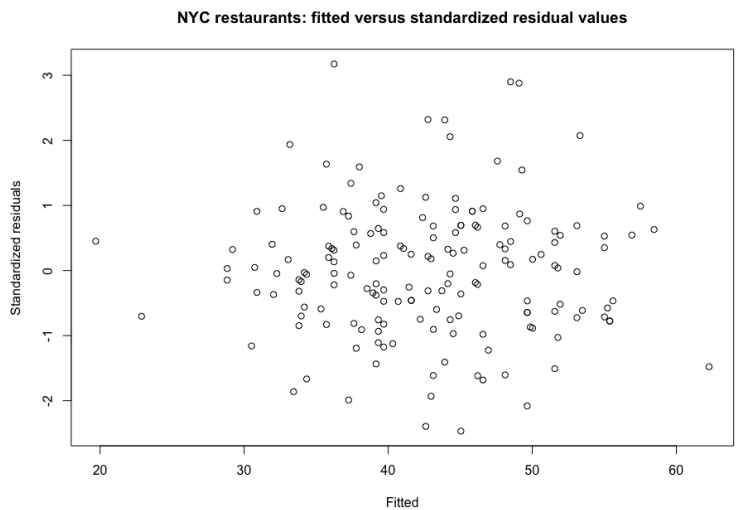



```
plot(nyc_restaurant_data[, c(3, 4, 5, 6, 7)], col="
    cadetblue")
```

NYC data: residual plots

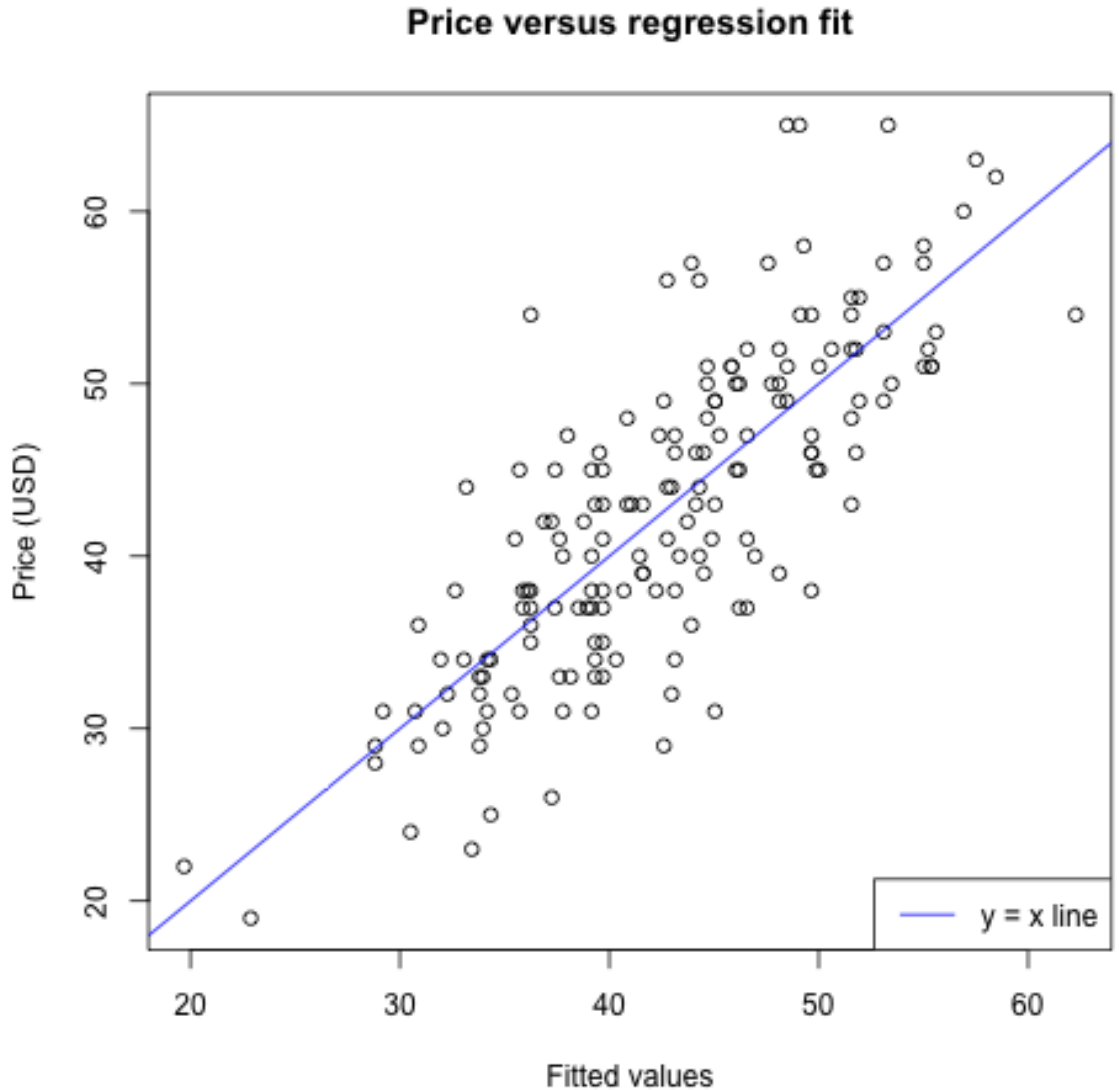
Evidence against the linearity assumption?





NYC data: response versus fitted values

Evidence against the linearity assumption?



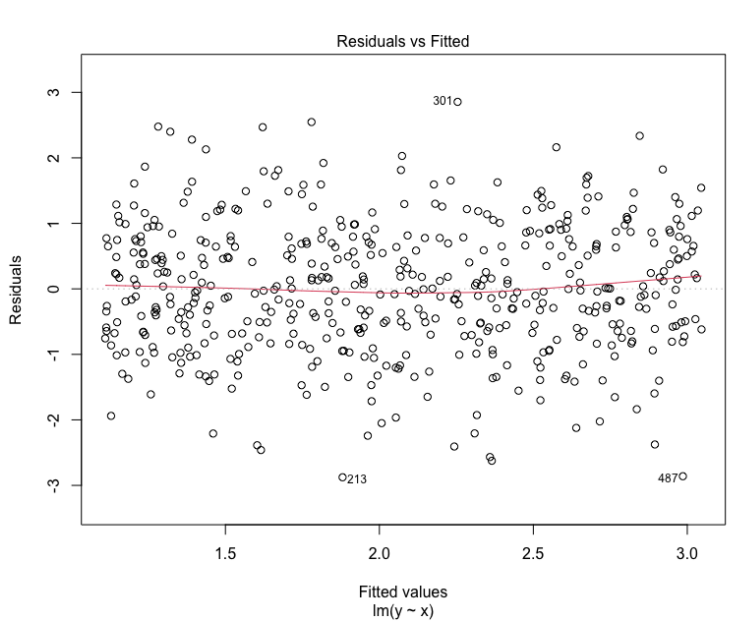
Lecture 3: More regression diagnostics

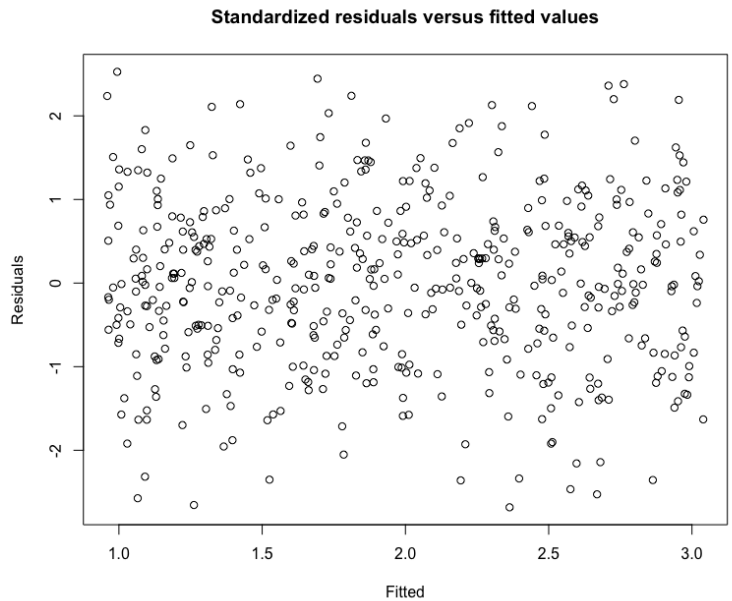
Learning goals for Lecture 3

- Checking the homoscedasticity assumption for regression models using the residuals
- Introduction to QQ plots

Checking the constant variance assumption

Under the regression model, the **residuals** $\text{Var}(\hat{e}_i) = \sigma^2(1 - h_{i,i})$ and the the **standardized residuals** r_i should have unit variance.

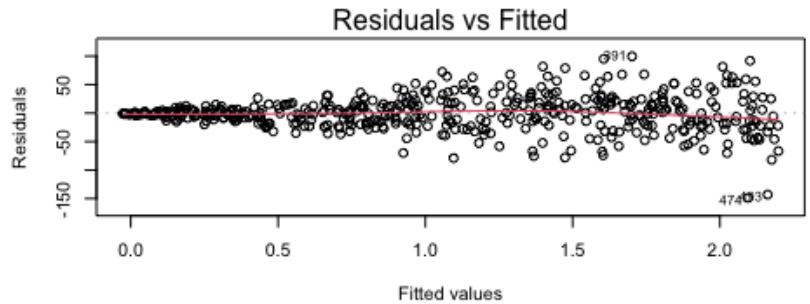


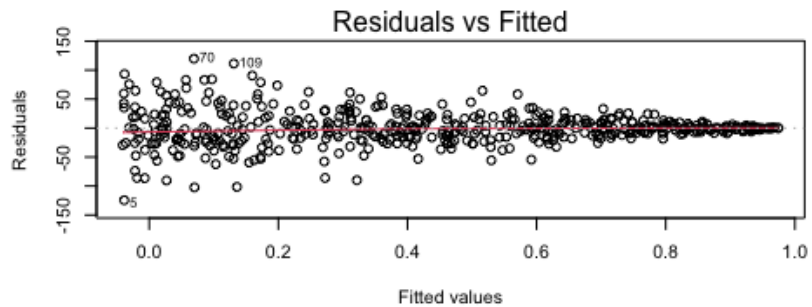


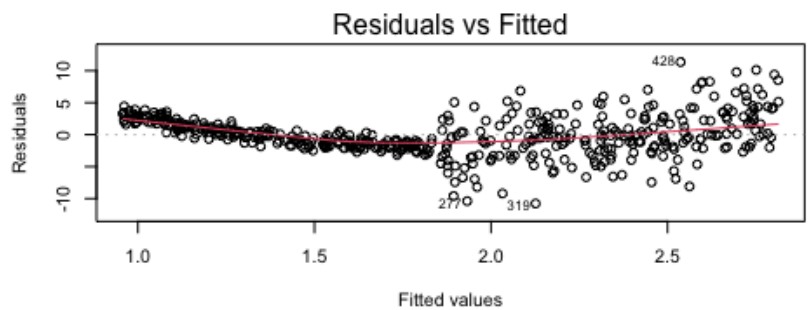
The plot of the residuals should vary without obvious patterns around 0.

Checking the constant variance assumption

Evidence against constant variance?



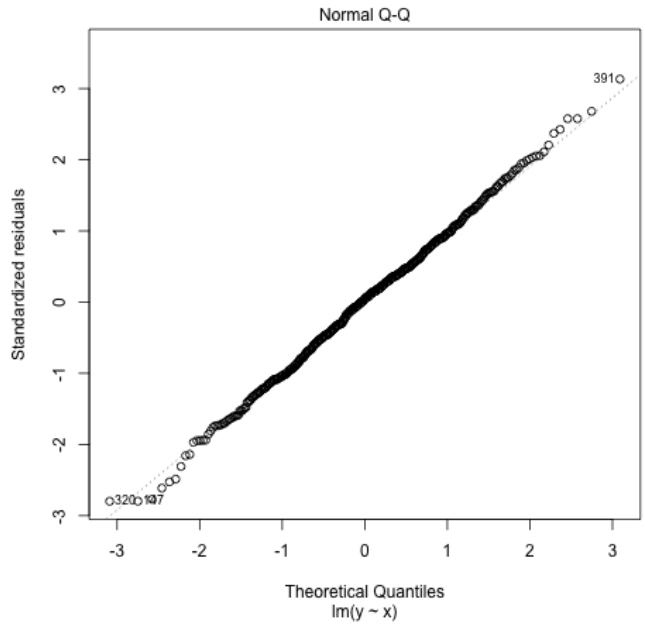




Lecture 3: Quantile-quantile plots

Checking for normal errors

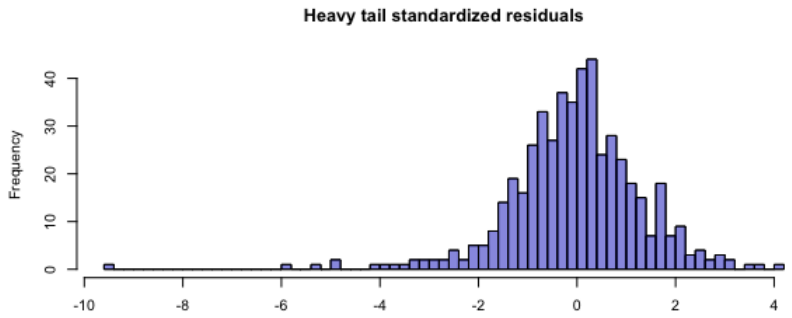
Under the regression model, the **standardized residuals** should be approximately normal. The **QQ plot** should be look approximately like a line (Intuitively quantile of sresidual $= 1 \cdot$ ideal quantile value).

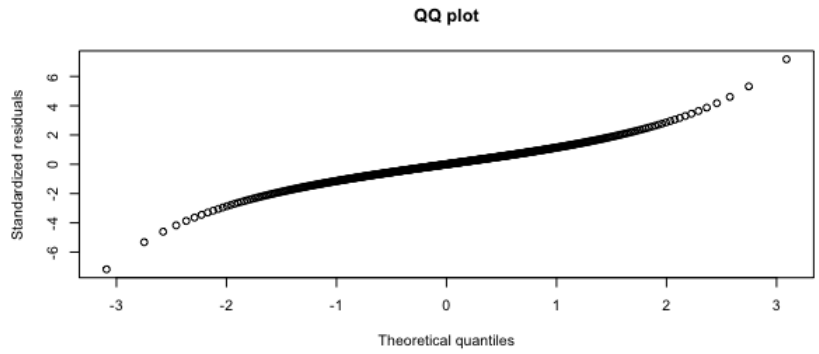


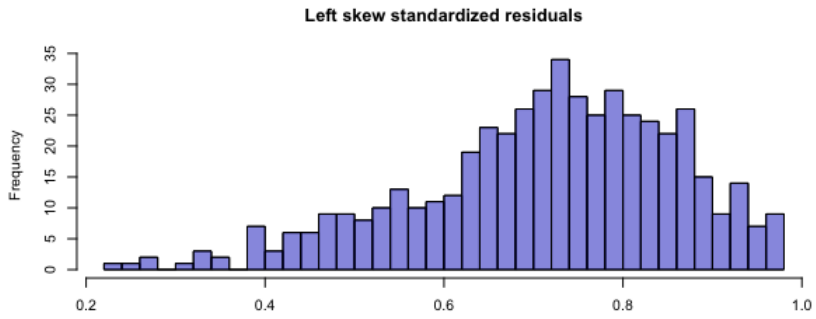
- Plots theoretical quantiles of standardized residuals on the x-axis and the actual standardizes residual quantiles on the y-axis.

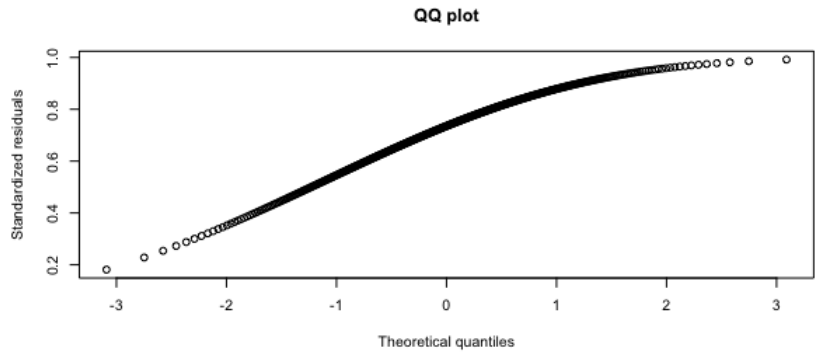
```
plot(fit, which=2) # Plot with R
```

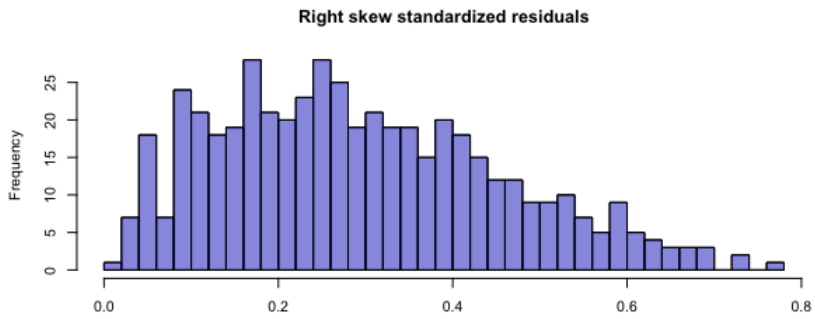
QQ plots

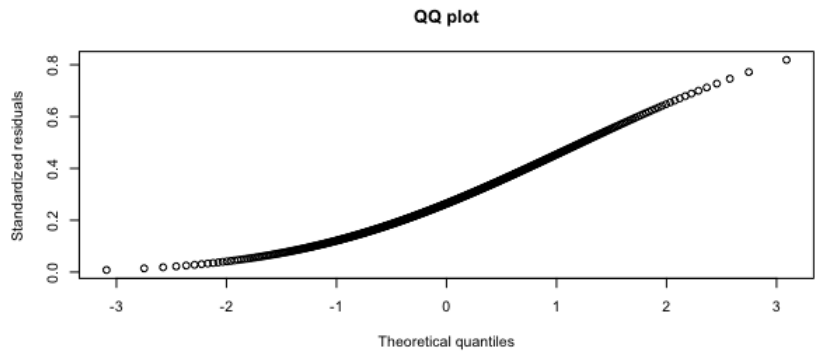

Heavy tail standardized residuals
Frequency
0
10
20
30
40
-10
-8
-6
-4
-2
0
2
4


QQ plot
Standardized residuals
-6
-4
-2
0
2
4
6
-3
-2
-1
0
1
2
3
Theoretical quantiles


Left skew standardized residuals
Frequency
0
5
10
15
20
25
30
35
0.2
0.4
0.6
0.8
1.0

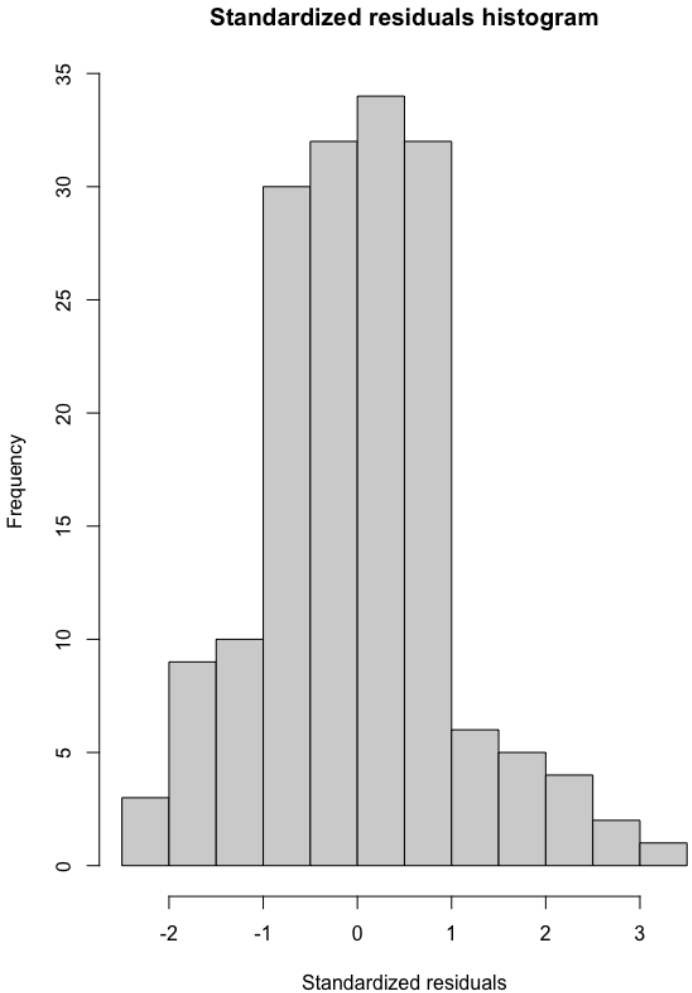

QQ plot
Standardized residuals
0.2
0.4
0.6
0.8
1.0
-3
-2
-1
0
1
2
3
Theoretical quantiles

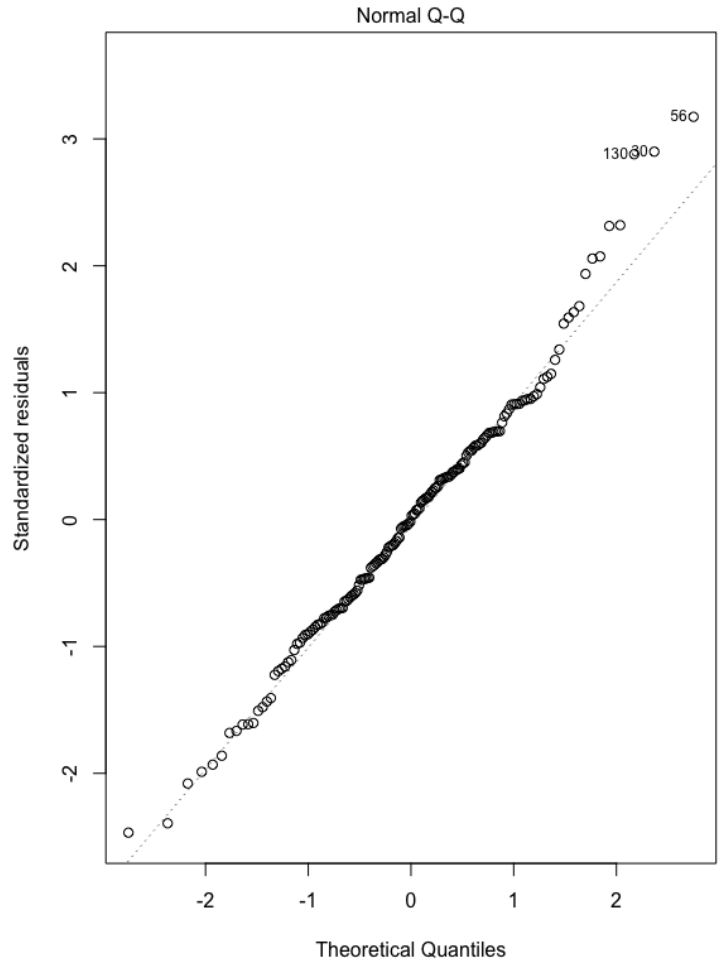

Right skew standardized residuals
Frequency
0
5
10
15
20
25
0.0
0.2
0.4
0.6
0.8


QQ plot
Standardized residuals
0.0
0.2
0.4
0.6
0.8
-3
-2
-1
0
1
2
3
Theoretical quantiles

NYC example

What can we interpret from these plots?





Summary: regression model diagnostics

1. Choose a which predictors to use for the regression model
2. Fit the model
3. Always plot (extensively) to check the assumptions are valid

Module takeaways

1. What are the assumptions of linear regression and what do they mean?
2. What tools are used to assess whether each assumption holds?
3. What do we look for to know whether each assumption holds?
4. Why is it important to check the conditions in multiple linear regression models?

Lecture 3: Activity

Activity

- Get in groups
- Organize your final project groups in Quercus
- Use RMarkdown and create some diagnostic plots for the NYC data

References I

Simon Sheather. *A modern approach to regression with R*. Springer Science & Business Media, 2009.