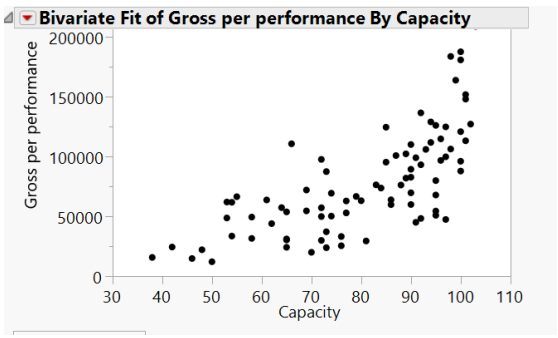
DSA 8010 - simple linear regression 2

Residual analysis

Example (Broadway)

Data were collected from a random sample of 95 Broadway shows. Each row in the data represents one week of data from a particular show. Two of the variables in the data set are `Capacity`, the percentage of the theater that was filled that week, and `Gross per performance`, the gross revenue from that show that week divided by the number of performances. Here is a scatterplot of the two variables:



I fit a regression line to predict `Gross per performance` from `Capacity`. The estimated regression line is

Gross per performance = -87268.18 + 2082.4557*Capacity

Statistical model for SLR

The statistical model for SLR is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where ϵ_i, $i = 1, \ldots, n$, are i.i.d. and approximately $N(0, \sigma^2)$. Three assumptions of SLR:

1. equality of variances
2. approximate normality
3. linear relationship between x and y

Residuals for SLR

The residual for simple linear regression is defined as

$$\begin{aligned}\widehat{\epsilon}_i &= y_i - \widehat{y}_i \\ &= y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right)\end{aligned}$$

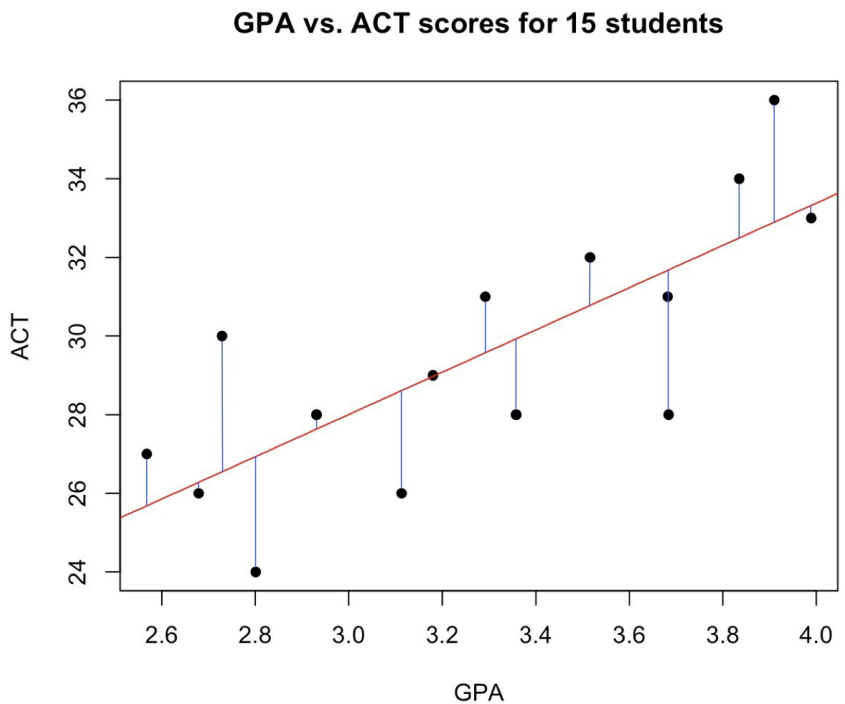


Image source: Christian Lee, `https://towardsdatascience.com/`

Residuals for SLR

The residual for simple linear regression is defined as

$$\begin{aligned}\widehat{\epsilon}_i &= y_i - \widehat{y}_i \\ &= y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right) \\ &= \text{observed - predicted}\end{aligned}$$

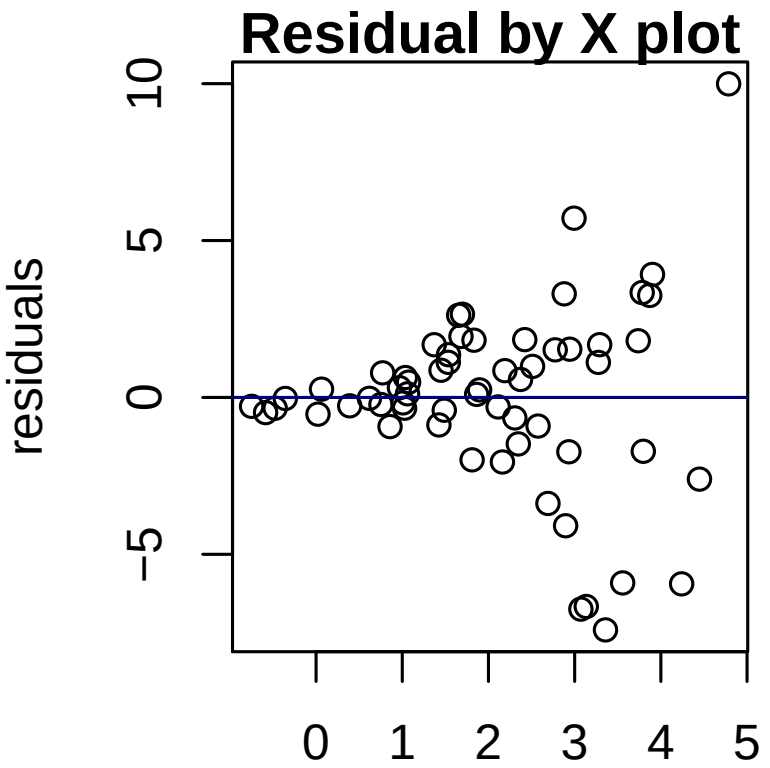
If the assumptions of the simple linear regression model are reasonable, the residuals should look approximately Normal and be randomly scattered around zero.

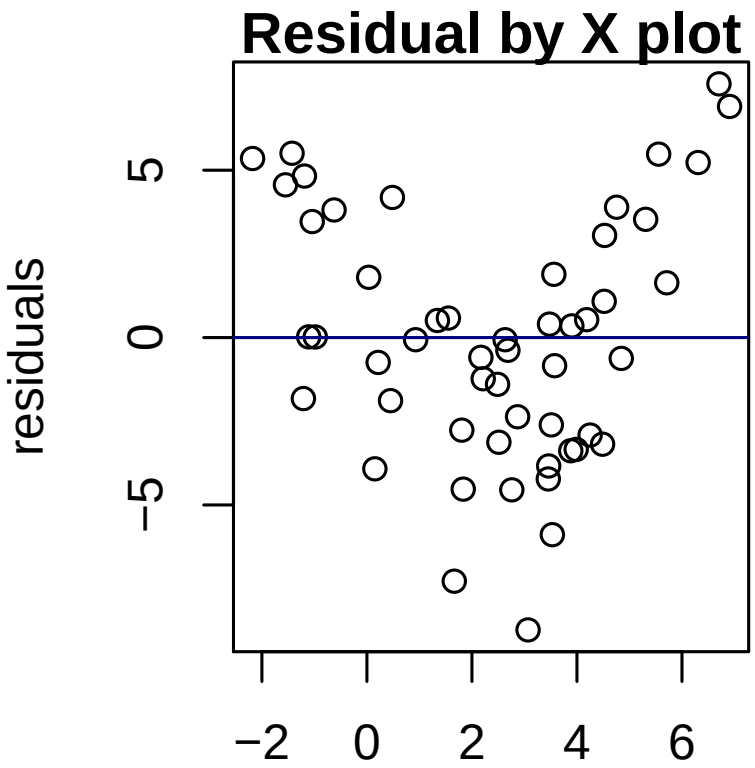
Residual analysis for SLR

- Create a Normal quantile plot of residuals to assess Normality
- Assess equality of variances and linearity with residual plots:
  1. Residual by X plot: x_i (x-axis) vs. $\widehat{\epsilon_i}$ (y-axis).
  2. Residual by predicted plot: $\widehat{y_i}$ (x-axis) vs. $\widehat{\epsilon_i}$ (y-axis).

     (The "residual by predicted" might also be called "residual by fitted" at times.)

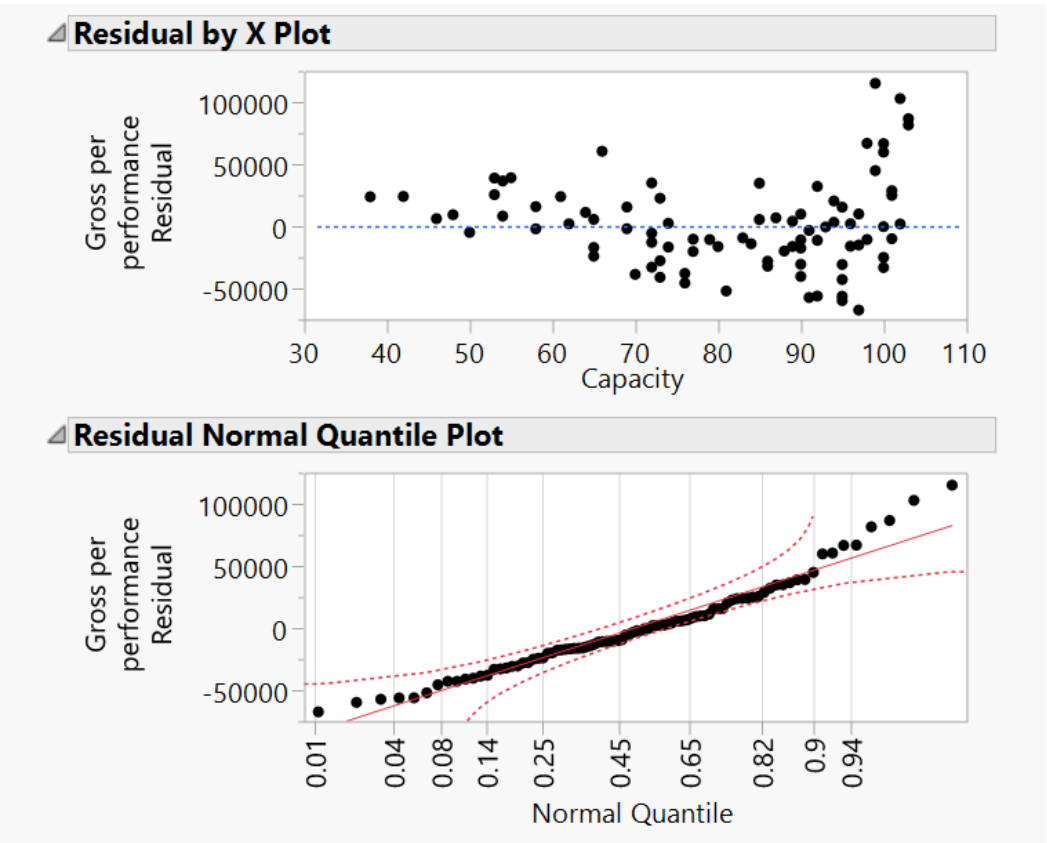
  Systematic patterns in the residuals as x_i or $\widehat{y_i}$ changes are indications that equality of variances and/or linearity does not hold.


Residual by X plot
residuals
10
5
0
−5
0
1
2
3
4
5


Residual by X plot
residuals
5
0
−5
−2
0
2
4
6

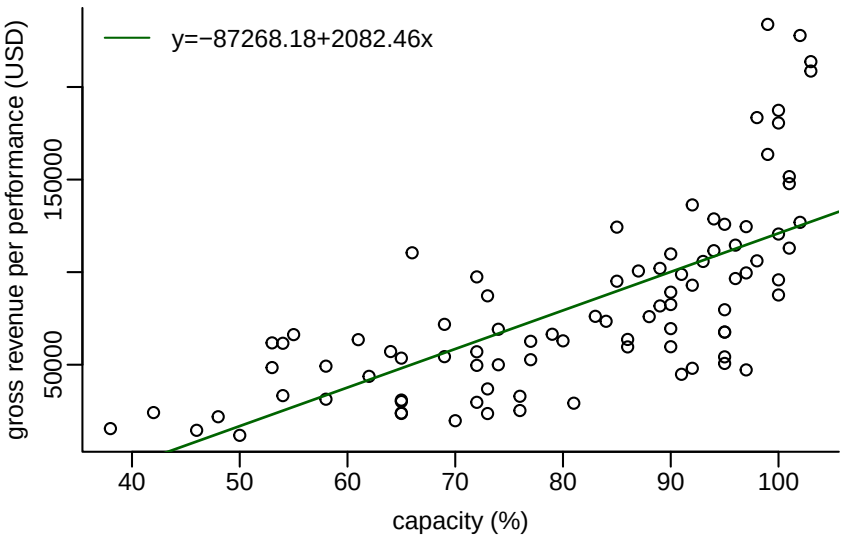
Example: Broadway

Here is a Normal quantile plot and residual vs. X plot for the Broadway data. What do they tell indicate about the modeling assumptions?



More on the SLR model

Example: Broadway



What does the model predict the gross revenue per performance to be for a theater filled to 10% of capacity?

Extrapolation

Example: Broadway. The estimated regression line predicting `Gross_per_performance` using `Capacity` is

$$\text{Gross per performance} = -87268.18 + 2082.4557 * \text{Capacity}.$$

What does the model predict the gross revenue per performance to be for a theater filled to 10% of capacity?

Extrapolation refers to making predictions of Y for values of X outside the range of the data.

Extrapolation

Extrapolation refers to making predictions of Y for values of X outside the range of the data.

- Extrapolation is risky since we do not have information about the relationship between the explanatory and response variables in the region in which we are making predictions.
- Avoid using a regression line to predict Y for X-values outside the range of X in the data.

From *Open Intro Statistics, 4th edition, Diez et al* page 322:

> Applying a model estimate to values outside of the realm of the original data is called **extrapolation**. Generally, a linear model is only an approximation of the real relationship between two variables. If we extrapolate, we are making an unreliable bet that the approximate linear relationship will be valid in places where it has not been analyzed.

Coefficient of determination (R^2)

- $0 \leq R^2 \leq 1$
- Interpretation: R^2 is the proportion of variability in the response variable (Y) that can be explained by the regression line.
- The higher R^2 is, the more accurately X can be used to predict Y. Sometimes we say that the regression model "fits well" if R^2 is high.
- R^2 is closely related to the sample correlation, r.

$$R^2 = r^2, \ |r| = \sqrt{R^2}.$$

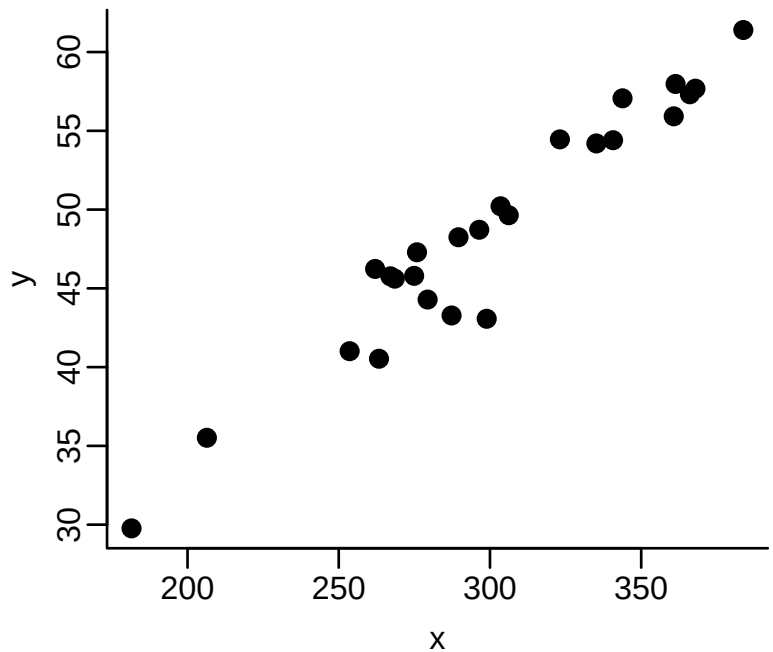
Example: Broadway

In the Broadway example, what is R^2 and its interpretation?

```
Call:
lm(formula = Gross.per.performance ~ Capacity, data = broadway)

Residuals:
   Min     1Q Median     3Q    Max
-67536 -23212  -2791  19124 114926

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -87268.2    18846.5   -4.63 1.25e-05 ***
Capacity      2082.5      227.3    9.16 1.92e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36160 on 88 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.4881,    Adjusted R-squared:  0.4823
F-statistic: 83.91 on 1 and 88 DF,  p-value: 1.919e-14
```

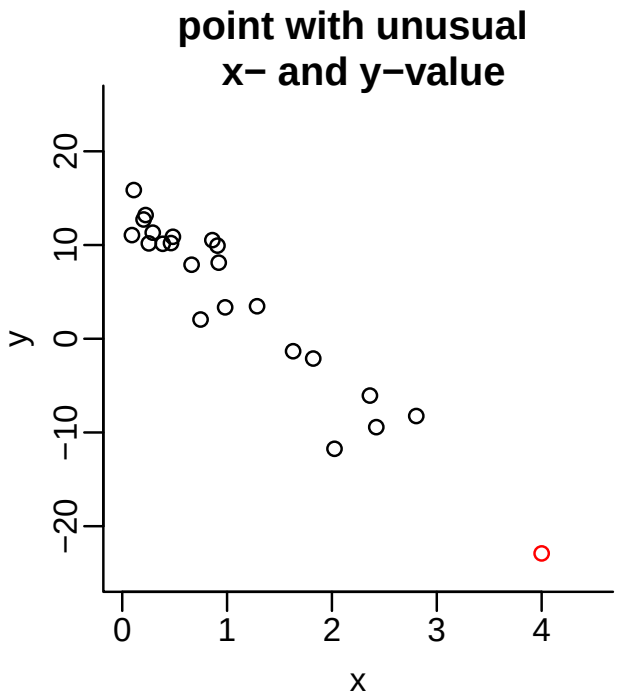
Outliers and influential points

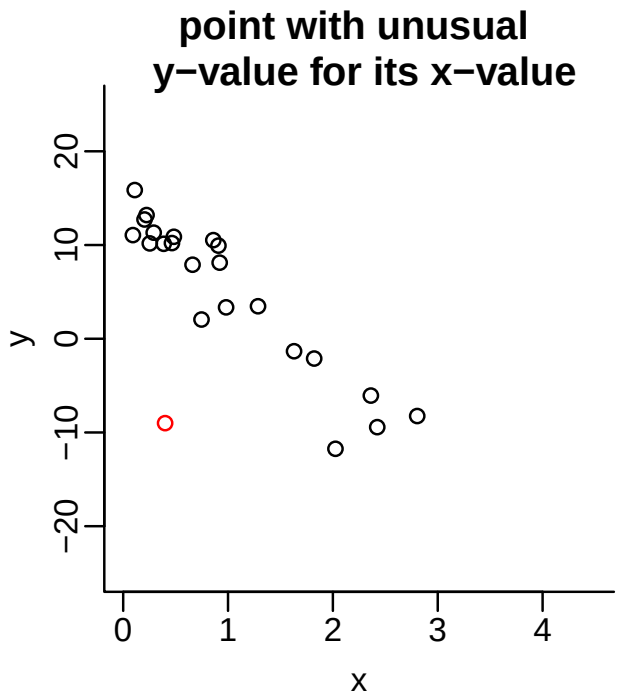
Recap: the least squares criterion is used to find the SLR line. It seeks to make the line as close as possible to all points simultaneously.



Outliers and influential points

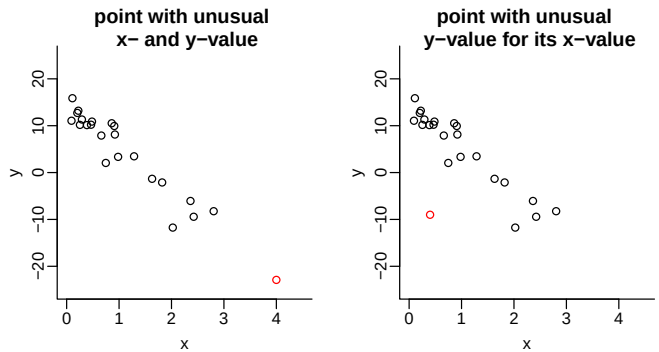
Outlier. A point whose x- or y-value is unusual, or whose y-value is unusual considering its x value.





Outliers and influential points

Outlier. A point whose x- or y-value is unusual, or whose y-value is unusual considering its x value.

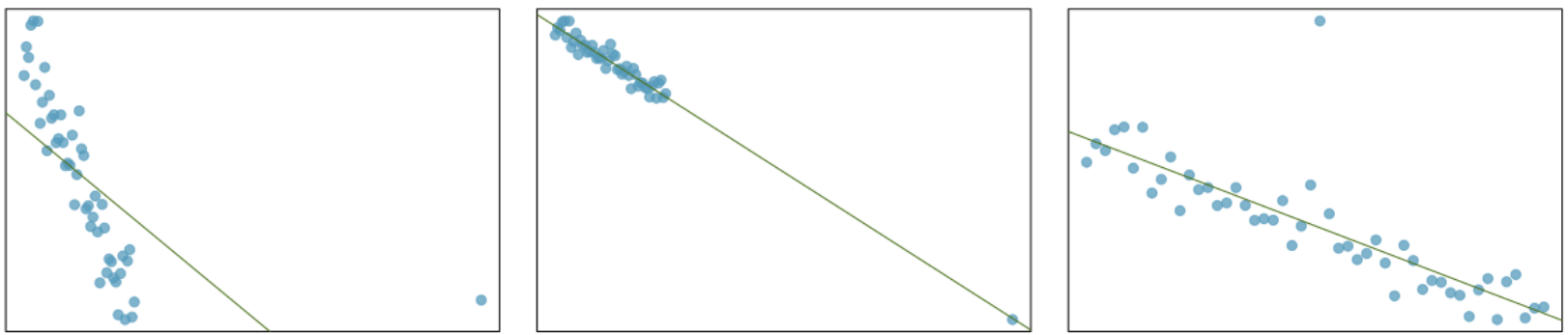


Leverage. A point whose x value is unusual is said to have high leverage. This means that it has the potential to pull the regression line towards itself.

Influential point. A point whose presence in the dataset has a strong effect on the regression line.

Outliers and influential points

From *Open Intro Statistics, 4th edition, Diez et al*.



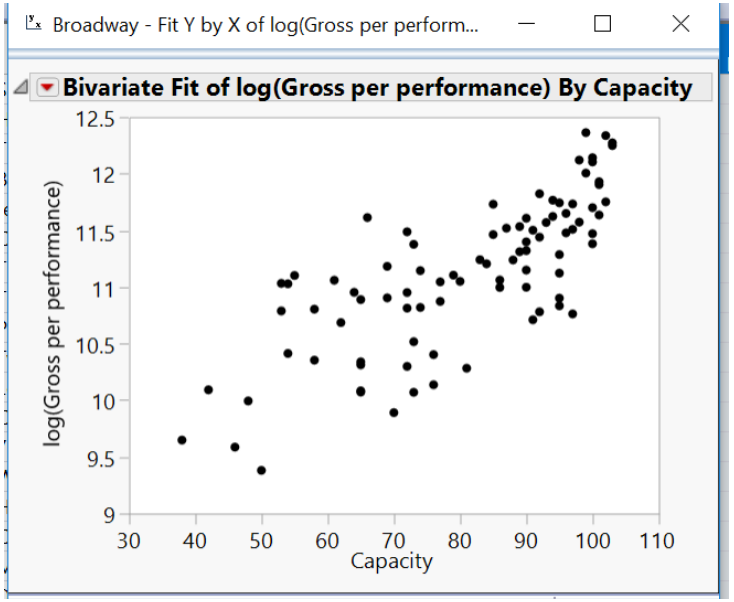
Outliers and influential points

Outliers typically provide important information about the population and should not be eliminated from the data or ignored without good reason. However, it is helpful to identify outliers that are influential and, perhaps, report results of the analyses with and without them.

What if X and Y do not have a linear relationship?

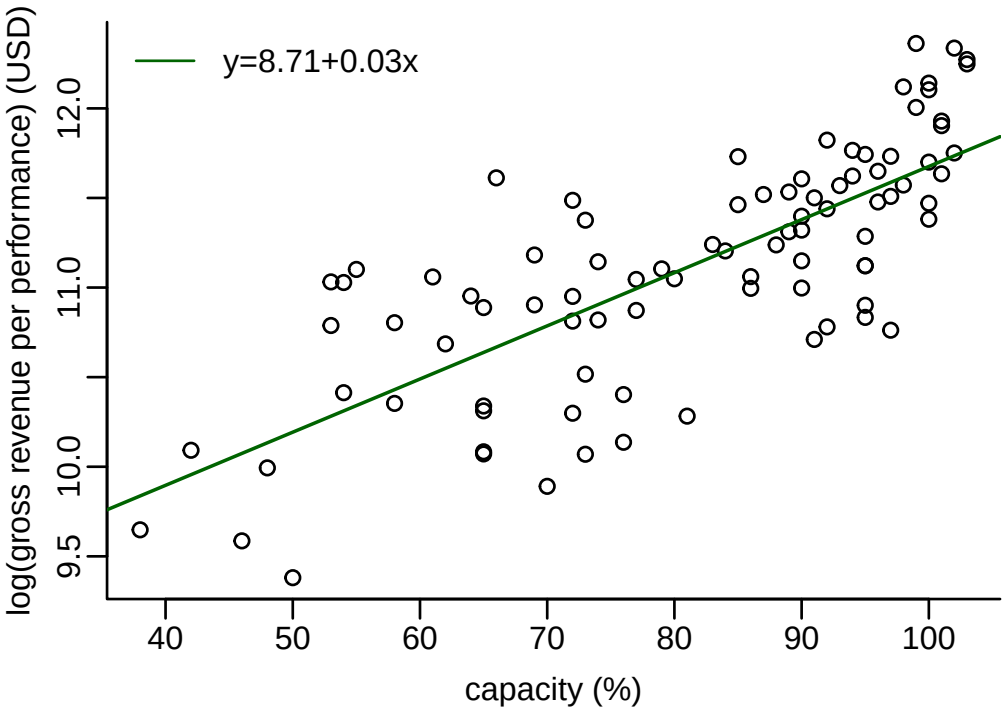
Transformations. Transforming one or more variable might produce a linear relationship.

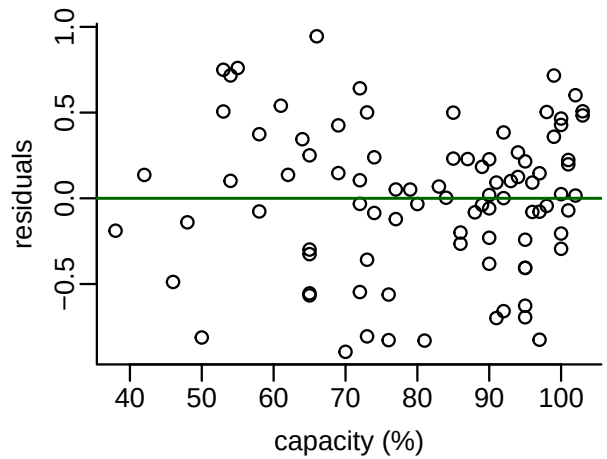
Example: This plot shows a transformation of the Broadway data. The response variable is log(*Grossperperformance*).

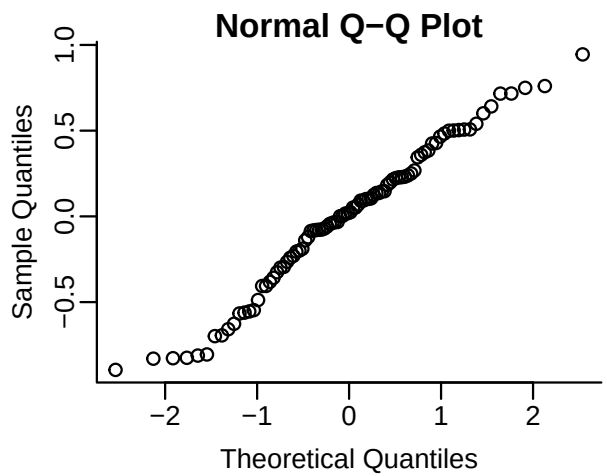


What if X and Y do not have a linear relationship?

Transformations.

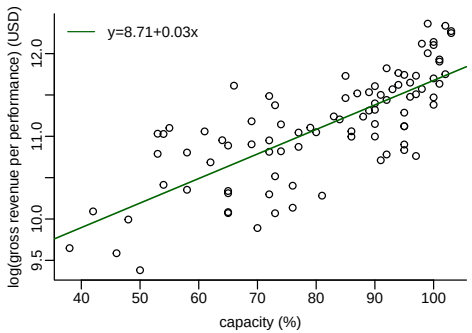






What if X and Y do not have a linear relationship?

Transformations. Transforming one or more variable might produce a linear relationship.



Non-linear regression. There are also methods for non-linear regression.