MATH 321: Mathematical Statistics

Lecture 8: Regression

Applied Linear Statistical Models: Chapters 1-3

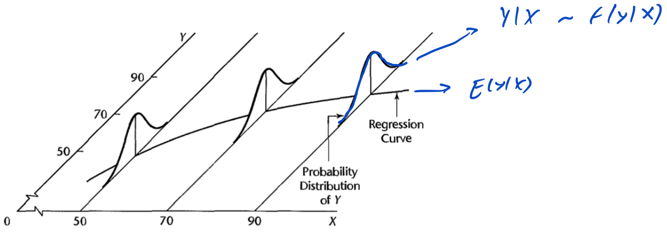
Introduction

Regression overview

- Goal is to determine if and how one variable is related to a set of other variables.
- Variables
  - Response variable, denoted Y, represents an outcome whose variation is being studied.
  - Explanatory variable, denoted X, represents the causes (i.e. potential reasons for variation).
- Two types of relationships
  - Functional (deterministic): There is an exact relation between two variables (have the form $y = ax + b$).
  - Statistical (probabilistic): There is not an exact relation because there are other variables that affect the relationship (have the form $y = ax + b + \underset{\text{Random}}{\epsilon}$).

Regression models and their uses

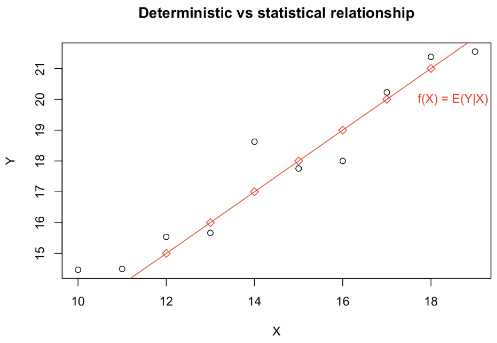
- Statistical models quantify the relationship between a response variable (i.e. a random variable) and explanatory variables, which are usually assumed to be deterministic (i.e. known exactly).
- Elements of a statistical regression model
  - In general, observations do not fall directly on the curve of a relationship.
    - $Y \mid X$ has a probability distribution.
    - $E(Y \mid X)$ varies deterministically with X.



- So the statistical model is:

$$\begin{aligned} Y &= E(Y \mid X) + \epsilon \\ &= f(X) + \epsilon, \qquad \text{where } \epsilon \text{ has some distribution} \end{aligned}$$

- Two components of a statistical model:
  1. $f(X) = E(Y \mid X)$: Defines relationship between Y and X; explains the average behavior of the response.
  2. ϵ: An element of randomness (i.e. error). This contains the variation that $f(X)$ cannot explain and/or that is of no interest.
- This means $f(X) = E(Y \mid X)$ will be the same for all samples with the same X values. The only thing that changes is the random error ϵ and as a result Y. Example $Y = 3 + 1X + \epsilon$:

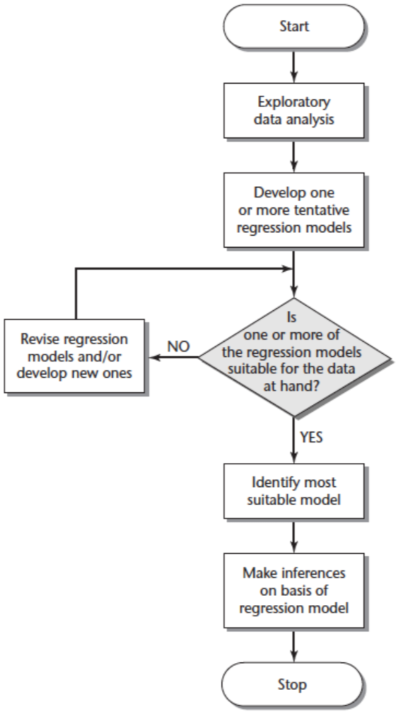


Construction of statistical regression models

1. Selection of predictor variables (how to decide which ones?).
   - Use of outside information, historical knowledge, and/or experience.
   - Exploratory data analysis.
   - Variable selection techniques: Find a subset of important variables (i.e. practical and easy to find).
2. Functional form of the regression relation (what is form of $f(X)$?).
   - $<$ based on same info as (1) $>$
   - If there is an abundance of data, maybe start with more complex models and then simplify.
3. Scope of model (when is the model useful?).
   - When the model best predicts or describes the relationship between response and predictor variables.

Uses of statistical regression models

1. Determining whether an X “affects” Y or not.
2. Estimation of impact of a given X on the Y.
3. Estimation of the mean of Y for a given X value.
4. Prediction of a single value of Y for a given X value.


Start
Exploratory data analysis
Develop one or more tentative regression models
Is one or more of the regression models suitable for the data at hand?
NO
Revise regression models and/or develop new ones
YES
Identify most suitable model
Make inferences on basis of regression model
Stop

Simple linear regression (SLR)

Goal of SLR

- Investigate the relationship between Y and a single numeric independent variable X, assuming that, in the population, the mean of Y is linearly related to the value of X.
- Population relationship: $Y = \beta_0 + \beta_1 X + \epsilon$.
- Sample relationship: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

Data structure

- Both X and Y on a random sample of n individuals are collected from the population of interest. The resulting data has the form $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Model statement: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

- Y_i: Dependent (or response) variable value. These are independent, but not identically distributed (Conditional on X_i).
- X_i: Independent (or predictor) variable value. These are **not random variables**, rather Known constants .
- ϵ_i: Random error term, **assumed** to have mean zero and variance σ^2. $\mathrm{Cov}(\epsilon_i, \epsilon_j) = \mathrm{Corr}(\epsilon_i, \epsilon_j) = 0$ for all $i, j : i \neq j$. Often, the ϵ_i are assumed to be *iid*.
- β_0 and β_1: fixed, but unknown regression parameters that need to be estimated.
- σ^2: Another parameter that needs estimated, but it is technically not a "regression" parameter since it does not determine the relationship between Y and X (i.e. it only deals with randomness).
- Note that Y_i and ϵ_i are random variables and therefore have distributions. Thus, discussing their mean and variances are appropriate.

Some implications of above

- Mean of Y_i for given X_i

$$E(Y_i) = E(\beta_0 + \beta X_i + \varepsilon_i)$$
$$= E(\underbrace{\beta_0 + \beta_1 X_i}_{\text{constants}}) + \underbrace{E(\varepsilon_i)}_{= 0}$$
$$= \beta_0 + \beta_1 X_i$$

Variance of Y_i for given X_i

$$V(Y_i) = V(\underbrace{\beta_0 + \beta_1 X_i}_{\text{constants}} + \varepsilon_i)$$
$$= V(\varepsilon_i)$$
$$= \sigma^2$$

Interpretation of regression parameters (β_0, β_1)

- β_0: Y-intercept of the regression line and gives Y's mean when $X = 0$

$$E(Y \mid X = 0) = \beta_0 + \beta_1 0 \overset{\checkmark}{=} \beta_0$$

- β_1: Slope of the regression line and indicates the change in Y's **mean** when X increases by one unit

$$E(Y \mid X = x^* + 1) - E(Y \mid X = x^*) = [\beta_0 + \beta_1 (x^* + 1)] - [\beta_0 + \beta_1 x^*] \overset{\checkmark}{=} \beta_1$$

  - Determines whether a relationship exists between Y and X.
  - ☆ Note that regression **does not** substantiate or prove a **cause-effect** relationship. Rather it gives evidence that Y and X are related (but not that X "causes" the value of Y).

Estimation of the regression function

- Setup
  - For each point we have an observed value Y_i, a fitted value $\hat{Y}_i$ and a residual $\hat{\epsilon}_i$.
  - Fitted regression function: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, where $\hat{\beta}_0$ and $\hat{\beta}_0$ are estimators of β_0 and β_1, respectively.
- Goal
  - Goal is to estimate the two "regression" parameters β_0 and β_1.
  - There are several methods to do this.

Method of least squares

- Overview
  - For each observation (X_i, Y_i), this method considers the model error term, which is the deviation of Y_i from its expected value:

$$\epsilon_i = Y_i - E(Y_i) = Y_i - (\beta_0 + \beta_1 X_i)$$

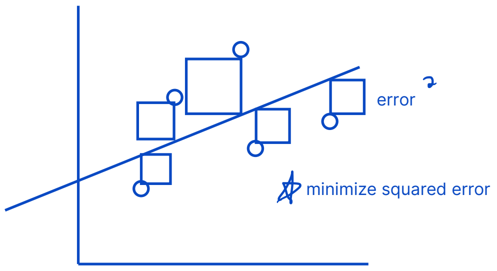
  - Then we minimize the sum of some function of these errors:

$$\begin{aligned} Q &= \sum_{i=1}^{n} \text{function of } \epsilon_i \\ &= \sum_{i=1}^{n} \text{function of } \big(Y_i - E(Y_i)\big) \\ &= \sum_{i=1}^{n} \text{function of } \big(Y_i - (\beta_0 + \beta_1 X_i)\big) \qquad < \text{for SLR} > \end{aligned}$$

  - For least squares method specifically, we consider the sum of the n squared errors (deviations). Thus we have:

$$\boxed{\begin{aligned} Q &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \end{aligned}}$$

  - And the point estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are the values that achieve the minimum Q. These can be found analytically.



- Results

$$\text{Intercept} \quad \hat{\beta}_0 = \frac{1}{n}\sum Y_i + \hat{\beta}_1 \frac{1}{n}\sum X_i = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\text{Slope} \quad \hat{\beta}_1 = \frac{\sum X_i Y_i - \frac{1}{n}\sum X_i Y_i}{\sum X_i^2 - \frac{1}{n}(\sum X_i)^2} \xrightarrow{\text{algebra}} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}$$

- Derivation

→ ① Goal: Minimize $Q = \sum_{i=1}^{n} (\overset{y_i}{y_i} - \underbrace{\beta_0 - \beta_1 x}_{E(y_i)})^2$

→ ② Take partial derivatives

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum (y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum x_i (y_i - \beta_0 - \beta_1 x_i)$$

→ ③ Set to zero & solve

$$\hat{\beta}_0 \to 0 = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$\downarrow = \sum y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum x_i$$

$$n\hat{\beta}_0 = \sum y_i - \hat{\beta}_1 \sum x_i$$

$$\hat{\beta}_0 = \frac{\sum y_i}{n} - \hat{\beta}_1 \frac{\sum x_i}{n}$$

$$\boxed{\downarrow = \bar{y} - \hat{\beta}_1 \bar{x}}$$

$$\hat{\beta}_1 \to 0 = -2 \sum x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$= \sum x_i y_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2$$

$$= \sum x_i y_i - \left(\frac{\sum y_i}{n} - \hat{\beta}_1 \frac{\sum x_i}{n}\right) \sum x_i - \hat{\beta}_1 \sum x_i^2$$

$$= \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} + \hat{\beta}_1 \left[\left(\frac{\sum x_i}{n}\right)^2 - \sum x_i^2 \right]$$

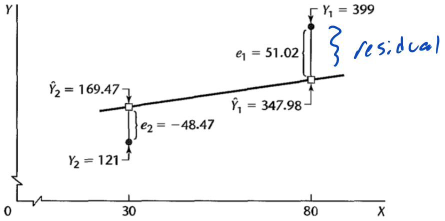
$$\boxed{\hat{\beta}_1 \overset{\checkmark}{=} \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}}$$

  - Note: We did not have to assume any distribution of the error term. These are the LSE estimators for any SLR model.

- Least squares estimators contain some optimal properties (Best Linear Unbiased Estimator), similar to how MLEs did.

Residuals and estimation of the error terms variance (useful for inference on model)

- Residuals Actual - Predicted (A-P)
  - $\hat{\epsilon}_i = e_i = Y_i - \hat{Y}_i$: This is a known, observable estimate of the unobservable model error. Measures the deviation of the observed value from the fitted regression function.



  - Residuals are very useful for studying whether the given regression model is appropriate for the data.

- Error terms variance
  - Need to estimate the variance σ^2 of the error terms ϵ_i in a regression model to get an indication of the variability of the probability distributions of Y.
  - Motivation: Very similar to variance of a single population $S^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}$, except we use the residuals as the deviations because each Y_i comes from a different probability distribution with different X (depends on the X_i level).

$$\boxed{\text{Error (residual) sum of squares} \quad SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2}$$

  - Then divide by the $df = n-2$ to the mean square (two dfs are lost when because β_0 and β_1 need to be estimated when getting the estimated means $\hat{Y}_i$).

$$\text{Error (residual) mean square} \quad S^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

  - It can be shown that MSE is an $\underline{\text{unbiased estimator for } \sigma^2\text{: } E(MSE) = \sigma^2}$.

    An estimator of the $\underline{\text{standard deviation is simply } S = \sqrt{MSE}}$.

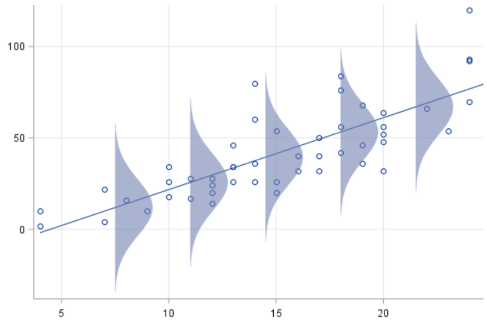
    ↳ R gives this

Normal error regression model

- These assumptions on ϵ_i are needed to set up interval estimates and make tests.
- The standard assumption is that the error terms are normally distributed. This greatly simplifies the theory of regression analysis and is justifiable in many real-world situations where regression analysis is applied.

$$\boxed{\text{New regression model:} \qquad Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad \text{where} \quad \epsilon_i \overset{iid}{\sim} \text{Normal}\,(0, \sigma^2)}$$

- This model means $\underline{Y_i \overset{\perp\!\!\!\perp}{\sim} \text{Normal with } E(Y_i) = \beta_0 + \beta_1 X_i \text{ and } V(Y_i) = \sigma^2}$.



- Justification of the normality assumption
  - Error terms frequently represent the effects of factors omitted from the model that affect the response to some extent and that vary at random without reference to the variable X.
  - These random effects have a degree of mutual independence, the composite error term representing all these factors tends to normal as the number of factors becomes large (by the CLT).
  - Also, the estimation and testing procedures shown later are based on the t distribution and are usually only sensitive to large departures from normality. So, unless the departures from normality are serious, particularly with respect to skewness, the actual confidence coefficients and risks of errors will be close to the levels for exact normality.

Estimation of parameters by method of maximum likelihood

- We can also estimate the parameters β_0, β_1, and σ^2 using maximum likelihood estimation.

Parameter	MLE	
β_0	$\hat{\beta}_0$	Same as LSE
β_1	$\hat{\beta}_1$	Same as LSE
σ^2	$\hat{\sigma}^2 = \dfrac{\sum(Y_i - \hat{Y}_i)^2}{n}$	

- Because of these results, $\hat{\beta}_0$ and $\hat{\beta}_1$ also possess the optimal properties of MLEs like consistency and minimum variance in the class of unbiased estimators;

Inference

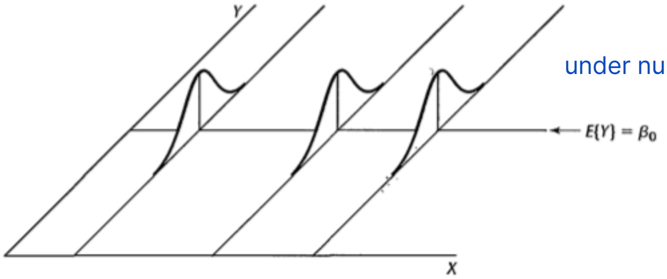
- For the rest of this section, assume the normal error regression model from above is applicable.

Inferences concerning β_1

- Overview
  - We often want to make inferences about β_1. A common test on β_1 has the form below.
  - If $\beta_1 = 0 \implies$ Regression line in horizontal, which means there is no linear association between Y and X, and even more no relation of any type because all probability distributions of Y are identical at all levels of X: normal with $E(Y) = \beta_0 + (0)X = \beta_0$ and variance σ^2.

$H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$



under null hypothesis

- Sampling distribution of $\hat{\beta}_1$
  - Refers to distribution of $\hat{\beta}_1$ from repeated sampling when the levels of the predictor variable X are held constant from sample to sample.
  - Recall $\hat{\beta}_1 = \dfrac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$; this is the point estimator.
  - Distribution of $\hat{\beta}_1$ is Normal with mean and variance:

$$\hat{\beta}_1 \sim \text{Normal} \left(\begin{array}{l} E(\hat{\beta}_1) = \beta_1 \\ V(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum(X_i - \bar{X})^2} \end{array} \right)$$

  - Then we can estimate the variance by replacing the parameter σ^2 with MSE, the unbiased estimator of σ^2. This gives us $S^2_{\hat{\beta}_1}$, which is an unbiased estimator for the variance of the sampling distribution of $\hat{\beta}_1$. And we can take the positive square root to give us $S_{\hat{\beta}_1}$, which is the point estimator of $\sigma_{\hat{\beta}_1}$.

$$S^2_{\hat{\beta}_1} = \frac{MSE}{\sum(X_i - \bar{X})^2} = \frac{MSE}{S_{XX}} \qquad \longrightarrow \qquad s_{\hat{\beta}_1} = \sqrt{\frac{MSE}{S_{XX}}} = \frac{S}{\sqrt{S_{XX}}}$$

  - Thus, $S^2_{\hat{\beta}_1}$ is an unbiased estimator for the variance of the sampling distribution of $\hat{\beta}_1$.

- Sampling distribution of standardized $\hat{\beta}_1$: $(\hat{\beta}_1 - \beta_1)/S_{\hat{\beta}_1}$

$$\star \quad \boxed{\frac{\hat{\beta}_1 - E(\hat{\beta}_1)}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE/S_{XX}}} \sim \mathrm{t}_{n-2}}$$

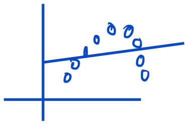
2 parameters estimated

Tests on β_1

- The test shown below is called a test of utility of the model for SLR.
- If reject: We conclude that X does contribute information for the prediction of Y when using the straight-line model.

  If fail to reject: Then we conclude there is no linear relationship between Y and X (horizontal model). But keep in mind:

  - Additional data might indicate that β_1 differs from zero.
  - A more complex relationship between Y and X may exist, which would require fitting a model other than the straight-line model.
  - All assumptions about the error terms ($\epsilon_i \overset{iid}{\sim} \text{Normal}\,(0, \sigma^2)$) should be satisfied.



- Two-tailed test (most common)
  - Hypotheses

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

  - Test statistic

$$TS = t^* = \frac{\hat{\beta}_1 - 0}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\sqrt{MSE/S_{XX}}}$$

  - Rejection region and p-value

$$RR = \{|t| > t_{\alpha/2,n-2}\}$$
$$p\text{-value} = 2 \cdot P(t_{n-2} \geq |t|)$$

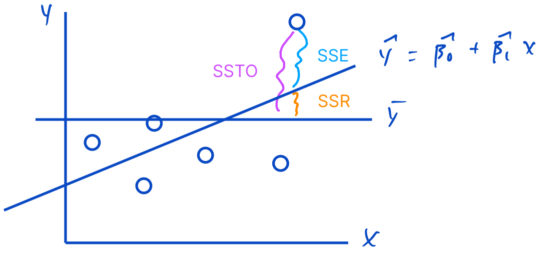
  - Decision
    - Reject H_0 and conclude H_A if $\quad TS \in RR \quad \iff \quad p\text{-value} \leq \alpha$
    - Fail to reject H_0 if $\quad TS \notin RR \quad \iff \quad p\text{-value} > \alpha$
    - Can also look at the $100(1-\alpha)\%$ CI for β_1 to see if contains 0.
  - Conclusion / Interpretation

    At the α significance level, we < have / do not have > sufficient evidence of a significant linear relationship between < Y context > and < X context >. < if yes... > This is a < positive / negative > linear relationship, indicating that as < X context > increases, < Y context > < increases / decreases >, on average.

Descriptive measures of linear association between X and Y

- Overview
  - There is no one single measure to completely describe the usefulness of a regression model for a particular application.
  - If the goal is estimation of parameters and means and predicting new observations, usefulness of estimates or predictions depends upon the width of the interval and the user's needs for precision. This can vary from one application to another.
  - Rather than making inferences, goals could be to describe the degree of linear association between Y and X.
- Coefficient of determination R^2
  - A very common measure because of its simplicity is the coefficient of determination R^2, which is a measure of the effect of X in reducing the uncertainty in predicting Y. This reduction in sum of squares ($SSTO - SSE = SSR$) gets expressed as a proportion:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}, \qquad \text{range:} \quad 0 \leq R^2 \leq 1$$



  - Interpretation: $< R^2 * 100 >$% of the variation in $< Y$ context $>$ can be explained by the linear relationship between $< Y$ context $>$ and $< X$ context $>$.
    - So, the larger R^2 is, the more the total variation of Y is reduced by introducing the predictor variable $X \Longleftrightarrow$ greater degree of linear association between Y and X.
    - Practically, this indicates the quality of the fit by measuring the proportion of variability explained by the fitted model.
  - Facts about R^2:
    - $0 \leq R^2 \leq 1$, which ranges from horizontal regression line to perfect fit.
    - Usefulness in prediction: A high coefficient of determination does not necessarily indicate that useful (precise) predictions can be made.
    - Overfitting: R^2 can be artificially inflated by including additional model terms (adding extra predictors). This is because SSR always increases with more predictors, even if they are completely unrelated to the response variable.

Diagnostics

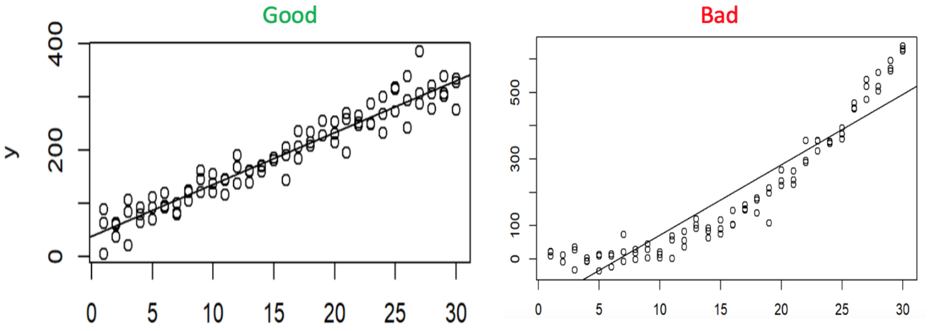
Overview

- Diagnostics are methods to check whether our model is reasonable for our data and representative of the system that we are studying (i.e. assumption checking).
- Why do we need to check the model?
  - The goal of building a model is to **learn something** about the real world or **predict outcomes** in the real world.
- To use a model successfully, we need to know its limitations:
  - Does it adequately describe the functional relationship of interest?
  - Is there reason to worry that inferences about the parameters might be flawed?
  - Is the error distribution appropriate?
- All of these are checked via **residual analysis**, whose goal is to assess the aptness of a statistical model.
- Why residuals?
  - Direct diagnostic plots for the response variable Y are ordinarily not too useful in regression analysis because the response variable observations are a function of the level of predictor variable.
  - So, instead we look at diagnostics for Y indirectly by examining the residuals.
- For our regression model, we assume $\epsilon_i \overset{iid}{\sim}$ Normal $(0, \sigma^2)$.

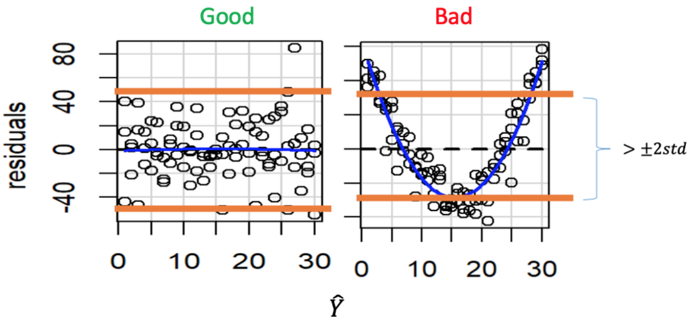
  So if the model is appropriate for the data at hand, the residuals should reflect these properties.

Residual analysis (LINE)

- Linearity
  - Can look at the scatterplot of Y vs X from the initial EDA to see if a linear model is appropriate:



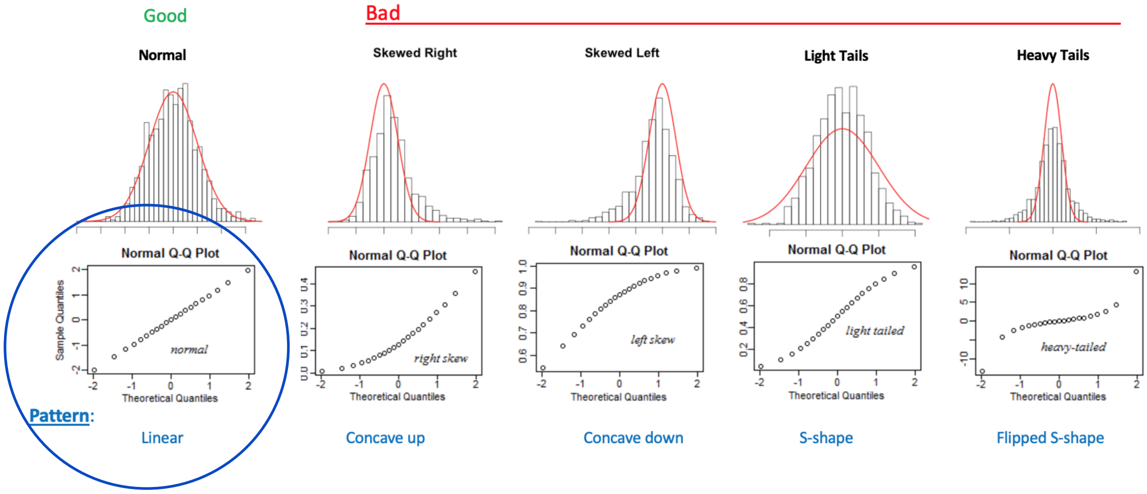
  - The preferred plot is the **residual plot against the fitted values** plot.



    - When a linear regression model is appropriate, the residuals then fall within a horizontal band centered around 0, displaying no systematic tendencies to be positive and negative (randomly scattered around 0).
    - When the linearity assumption is violated, there are systematic deviations.

- Independence
  - Ideally, any potential source of dependence is handled at the experimental design stage (or the sampling scheme), so that it is either eliminated by randomization or explicitly included in the data and we have one observation per subject.
  - There is a plot to look at this, but we will not cover it.

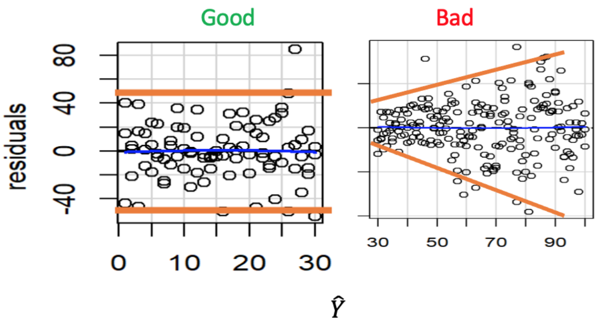
- Normality
  - Small departures from normality do not create any serious problems, but major departures should be of concern.
  - We can check the normality of error terms in a variety of ways:
    - Normal probability (QQ) plot of residuals.



    - Distribution plots of residuals: Boxplots should be symmetric and histograms should be roughly normal.
  - Difficulties in assessing normality: The analysis for model departures regarding normality is often more difficult than departures of other types because...
    - Random variation can be particularly mischievous when studying the nature of a probability distribution unless the sample size is quite large.
    - Even worse, other types of departures can and do affect the distribution of the residuals.

    e.g. Residuals may appear to be not normally distributed because an inappropriate regression function is used or because the error variance is not constant.
  - So, it is usually a good strategy to investigate these other types of departures first, before assessing the normality of the error terms.

- Equal variance
  - We can again look at the residual plot against the fitted values.



    * When there is a constant error variance, points again should fall within a horizontal band. So there is a constant spread of the residuals as move across the scope of fitted (or X) values.
    * "Tipped over tornado" effect of the points indicates a non-constant variance (e.g. as the fitted values increase, the residuals vary more, or vice versa).

      A nonconstant variance in called *heteroscedasticity* (the assumption is a *homoscedastic* error variance).