Estimation and Assumptions

Sean Hellingman ©

Regression for Applied Data Science (ADSC2020)

shellingman@tru.ca

Winter 2025

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- Now that we know how to interpret model formulas, it is important to properly estimate the models themselves.
- There are different methods to estimate statistical models.
- In order to make valid inferences from estimated statistical models, various assumptions must hold.
- When the assumptions do not initially hold, there are steps we can take to fix our models.

Estimation

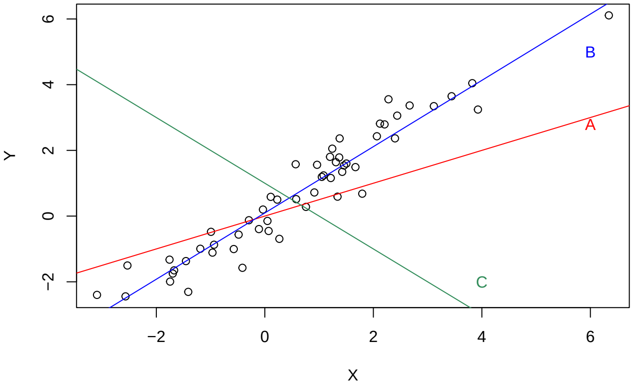
- **Simple linear regression model** (Expected value of Y):

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

- We do not know the true values of β_0 and β_1 because we do not have the entire population.
- We need to *estimate* these parameters the best we can using the data we do have.
- If we draw a straight line, we will never be able to include all of the data points unless we have a deterministic relationship.

Illustrative Example 1

- Which line should we choose to represent the linear relationship between X and Y?



Estimated Regression Line

- **The estimated simple linear regression equation is**:

$$\hat{Y} = b_0 + b_1 X.$$

  - b_0 and b_1 are estimates of β_0 and β_1.

- If (X_i, Y_i) is the i^{th} observation then $\hat{Y}_i = b_0 + b_1 X_i$ is the estimated value of Y for X_i.

Residuals

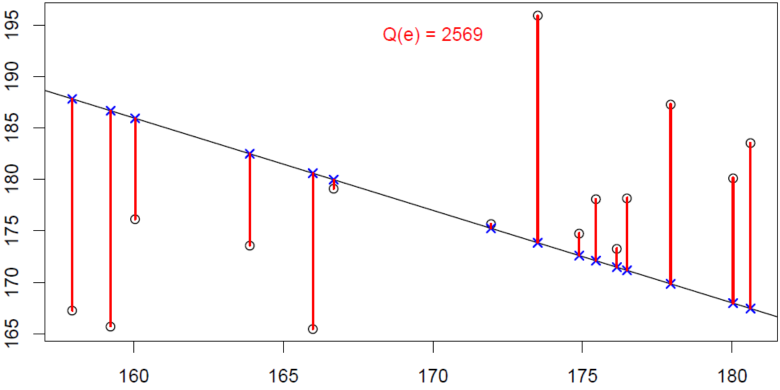
- One way to quantify the relationship between each point and the estimated regression equation is to measure the vertical distance between them.

- The **residuals** (observed errors) are defined as follows:

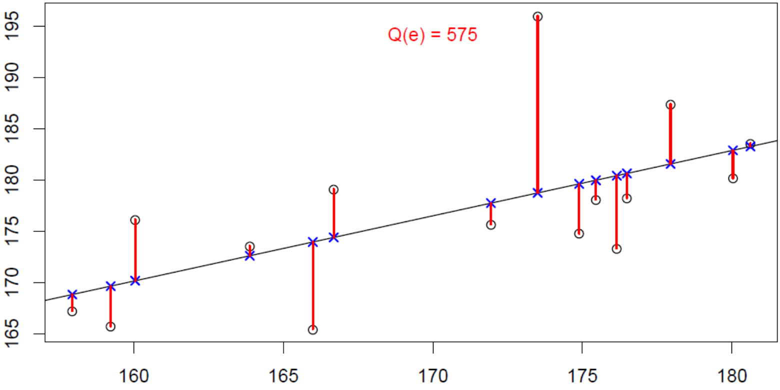
$$e_i = Y_i - \hat{Y}_i. \quad (1)$$

- The best-fitting line should minimize some measure of these errors.

Residuals Image I



Residuals Image II



Zero sum of Residuals

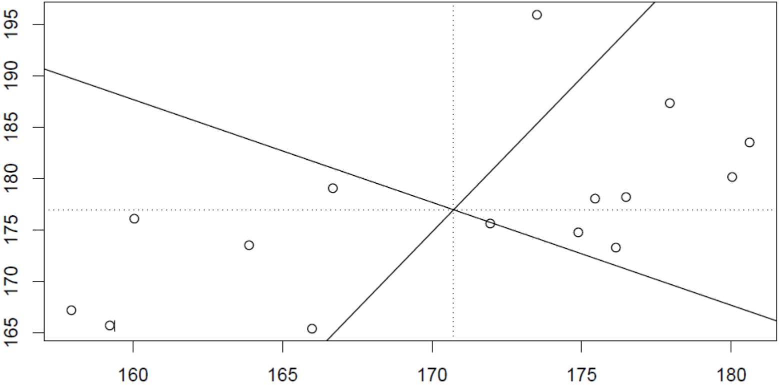
- Could we consider setting the sum of residuals to be 0?

$$\sum_{i=1}^{n} e_i = 0$$

- This criterion is satisfied if the line goes through the sample means of the variables Y and X.

- Some of the lines that satisfy the *zero sum of residuals* **do not** capture the relationship between Y and X.

Zero sum of Residuals Example



- The solid lines that satisfy the *zero sum of residuals*.

Least Absolute Deviations

- Could we consider minimizing the sum of the absolute deviations?

$$\text{Minimize} \sum_{i=1}^{n} |e_i| = \text{Minimize} \sum_{i=1}^{n} |Y_i - \hat{Y}_i|$$

- Least Absolute Deviations (LAD) Regression is actually fairly robust against the presence of outliers.
- However, **there may be more than one solution.**

Squared Residuals

- Because some of the residuals are positive and others are negative we square them (mathematical simplicity).
- We want to minimize the sum of the squared **residuals** (observed errors):

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \tag{2}$$

- The best-fitting line finds the intercept and slope that minimizes this sum (*Ordinary Least Squares (OLS) Regression*).
- **When $n > k$ (parameters) guaranteed a unique solution.**

Parameter Estimates

- Using calculus we can derive the following *least squares* estimates for a simple linear regression model:

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}. \tag{3}$$

$$b_0 = \bar{Y} - b_1\bar{X}. \tag{4}$$

- R has the functionality we need to estimate these parameters.

Illustrative Example 1

- Import the *Football22.csv* dataset in R and follow along with the example.

  1. Use the LAD method to estimate the following simple linear model: `Points ~ 1 + Goal_Differential`
  2. Use the OLS method to estimate the same linear model.
  3. Are the results different?

Multiple Linear Regression

- **Multiple linear regression models** contain more than one independent variable.
- Multiple linear regression models are used in many different real-world applications.
- Example: There may be multiple variables that explain someone's income level.
- We still use **Least Squares Regression** to estimate multiple linear regression models.

Estimated Regression Model

- **The estimated linear regression equation is**:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k. \tag{5}$$

  - b_0, b_1, b_2, ..., b_k, are estimates of β_0, β_1, β_2, ..., β_k.

- Where $\hat{Y}_i$ is the fitted (expected) value of Y_i.

Maximum Likelihood Estimation

- **Maximum Likelihood Estimation (MLE)** is a method of estimating parameters from an assumed probability distribution.
- *What parameter(s) values are most likely to have generated the observations.*
- Estimates are obtained by maximising a likelihood function based on an assumed distribution.
- **If the error term of a OLS regression model follows a normal distribution with a constant variance, the OLS estimates are the same as the MLE estimates.**

Properties

- If the assumptions about the error term are satisfied the Least Squares estimates (MLE) are the Best Linear Unbiased Estimators (BLUE).

    1. Unbiased: the expected value of the estimate is the population parameter.

    2. Minimum Variance: The sampling distribution of the estimate is the tightest around the population parameter.

- *The Best in BLUE refers to the sampling distribution with the minimum variance. That's the tightest possible distribution of all unbiased linear estimation methods*

MLE in R

- We can use R to estimate the parameters of assumed distributions of observed data using the *fitdistrplus* package:
    - `fitdist(data, distr = "name", method = "mle")`
- Some common distribution arguments:
    - Normal: `distr = "norm"`
    - Log-normal: `distr = "lnorm"`
    - Exponential: `distr = "exp"`
    - Poisson: `distr = "pois"`
    - Gamma: `distr = "gamma"`
    - Chi-squared: `distr = "chisq"`

Example 1

- Use the `fitdist()` to determine the Maximum Likelihood Estimates of the parameters that the samples in the example code are drawn from.

Model Assumptions

Estimation of the Variability of the Error Term

- We use $\hat{\sigma}^2$ for testing hypotheses about the coefficients and constructing confidence and prediction intervals.

- We can estimate σ^2 using the residual variance:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k}$$

- If the assumptions hold: $E[\hat{\sigma}^2] = \sigma^2$ (unbiased estimator)

Model Assumptions

The validity of any inferences from our linear regression models depend on the following assumptions:

1. **Linearity**
   - The relationship between the dependent and independent variable(s) needs to be linear.
2. **Normality** (multivariate normal for multiple independent variables)
   - In linear regression, all variables must be normally distributed (can be fixed).
3. **Homoscedasticity** (constant variance)
   - The variation about the regression line is constant for all values of the independent variable(s) (can be fixed).
4. **Independence**
   - There is little or no multicollinearity in the data (independent variables are too highly correlated with each other).

Model Assumption Violations

If Violated:

1. **Linearity**
   - May lead to serious inaccuracies when making predictions.

2. **Normality** (multivariate normal for multiple independent variables)
   - Causes problems in determining if model coefficients are significantly different from zero.
   - Also causes problems in any confidence interval estimation.

3. **Homoscedasticity**
   - As we are minimizing the residual sum of squares, extra *weight* may be given to observations with a higher variability during estimation.
   - Also causes problems with confidence intervals of predictions.

4. **Independence**
   - May lead to bias (over/under estimate) the nature of the linear relationship.

Diagnostics

- Due to the assumptions imposed on the error term (closely related to the residuals) we can use the residuals to help us check the model assumptions.
- We can also *standardize* the residuals to help with outlier identification and assumption checks.
- **Standardized residuals**: $\text{Residual}_i/\text{Standard Deviation of Residual}_i$
- In R:
  - `rstandard(linear.model)` from the *car* package

Linearity Check

- The **linearity** assumption may be checked in a few different ways:

  1. Examine the scatterplot(s) of the dependent and independent variable(s) (linear relationship?).

  2. Plot the residuals, they should be randomly scattered around zero with no apparent pattern.
     - In R: `plot(lm1$residuals)`
     - Add line at 0: `abline(h=0,col="blue")`

  3. Plot residuals vs fitted values, again should be randomly scattered around zero with no apparent pattern.
     - In R: `plot(model, 1)`

Normality Check

- The **normality** assumption may be checked in a few different ways (generally we focus on the residuals):

  1. Examine the histogram of the *standardized residuals* for approximate normality.
  2. Q-Q plot of the *standardized residuals*.
     - In R we can also use: `plot(model, 2)`
  3. We can use a Shapiro–Wilk test on the *standardized residuals*.
     - Shapiro–Wilk: `shapiro.test(standardized.residuals)`

- *Slight departures from normality may be acceptable and we can also take steps to fix departures from normality.*

Homoscedasticity Check

- The **Homoscedasticity** assumption may be checked in a few different ways (generally we focus on the residuals):

  1. Examine the scatterplot(s) of the *standardized residuals* for a constant variance.

  2. Examine the scatterplot(s) of the fitted values and the square root of the *standardized residuals* (*scale-location plot*). (It's good if you see a horizontal line with equally spread points)
     - In R: `plot(model, 3)`

  3. We can use the `ncvTest()` similar to a (Breusch–Pagan test) in R with the null hypothesis being a constant variance (Homoscedasticity).
     - In R (*car* package): `ncvTest(lm1)`
     - If we do not reject the null hypothesis (large p-value) we can assume Homoscedasticity.

- *We can also take steps to fix departures from Homoscedasticity.*

Independence Check

- *This assumption may be violated if we have time-dependent independent (explanatory) variables.*

- The **Independence** assumption may be checked in a few different ways (generally we focus on the residuals):

  1. Examine the scatterplot(s) of the *standardized residuals* for patterns or obvious clusters.

  2. We can use the Durbin Watson test in R with the null hypothesis being that the residuals are independent.
     - In R (*car* package): `durbinWatsonTest(lm1)`
     - If we do not reject the null hypothesis (large p-value) we can assume independence at one lag.

- *There are other tests like the Ljung-Box test that may be used for multiple lags.*

Example 2

- Import the *Duncan* dataset in R and complete the following tasks:

1. Use the defined functions in the `pairs()` function to examine the data.
2. Assuming *prestige* as the dependent variable, estimate a linear regression model.
3. Write out the model formula.
4. Conduct the appropriate model diagnostics, does your model pass?

Some Solutions

Linearity

1. Apply a nonlinear transformation to the dependent and/or the independent variables.
    - Logarithmic transformations only to the dependent/response variable: *Assumes that the response grows/decays exponentially as a function of the independent variables.*
    - Logarithmic transformation of both the dependent and the independent variables implies that the effects are multiplicative rather than additive. *small percentage change in one of the independent variables induces a proportional percentage change in the expected value of the dependent variable*

2. Consider adding *nonlinear* functions of the explanatory variables X_j^z
    - Can use visuals (scatterplots with LOWESS curves to identify degree of polynomial)

3. Identify missing variables (including interactions) and add them to your regression model.

Independence

1. Examine the Variance Inflation Factors (VIF) and consider removing one or more of the variables identified as having a high VIF.
    - Consider removing when the result of `vif()`: $GVIF \wedge (1/(2 * Df)) > 5$
2. Re-consider any transformations that you have already made to your data.
    - Transformations *may* actually increase the chance of violating the independence assumption.
3. If your data contain time-dependent variables, consider time-series/econometric models.
    - May be able to include lagged values of dependent variable in your model (not generally encouraged).

Homoscedasticity

1. Box-Cox Power Transformation
   - If the variance increases with the mean, choose $\lambda < 1$.
   - If the variance decreases as the mean increases, choose $\lambda > 1$
2. Weighted Least Squares (If the variance is not constant and is **not** related to the factor-level means we can use weighted least squares.)
   - A weight is assigned to each observation based on the variability.

     $W_{ii} = \frac{1}{\sigma_i^2}$
3. Try modelling with more complex regression models
   - *We will cover some of these models later in the course*

Box-Cox Transformation in R

- We can use the *MASS* package to obtain the *best* λ.
- In R:
    - `bc <- boxcox(model)`
    - `lambda <- bc$x[which.max(bc$y)]`

      To transform your data:
    - `transform.df <- transform(df, variable = variable^(lambda))`
- Then you can estimate your model again.
- *The Box-Cox transformation can usually help with the normality assumption.*

Weighted Least Squares in R

- `model <- lm(response ~ explanatory1 + ..., data = data)` `# Estimate Model`
- **Perform diagnostics to determine this is needed.**
- ```
  wt <- 1/lm(abs(model$residuals) ~
      model$fitted.values)$fitted.values^2 # Obtain weights
  ```
- ```
  wls_model <- lm(response ~ explanatory1 + ...,
      data = data, weights = wt) # Estimate WLS model
  ```
- ```
  summary(wls_model) # Check results
  ```
- *Be sure to perform diagnostics on your new model*
  ```

Normality

1. Apply a nonlinear transformation to the dependent and/or the independent variables.
    - *The Box-Cox transformation can usually help with the normality assumption.*

2. Examine your data for outliers.
    - Sometimes, outlying observations may contribute heavily to a violation of this assumption.

3. Try modelling with more complex regression models
    - *We will cover some of these models later in the course*

Outliers in Linear Regression

- **Regression outliers** are those observations whose values (of the response and explanatory variables) deviate from the regression relationship which holds for the majority of observations.

- **Cook's distance** is a measure of influence which measures the difference between the fitted values in the model with all the observations and the model with observation i removed.
  - We can plot in R: `plot(model,3)` AND `plot(model,4)`
  - To get the numeric values in R: `cooks.distance(model)`

- The `plot(model)` visualizations in R usually do a good job of identifying possible outliers.

Removing Outliers in R

- Can just remove specific rows by number: `data[-c(1,4,6),]`
- By name: `rows.to.remove <- c("row1","row2")`
  data[!(row.names(data) %in% rows.to.remove),]
- OR, you can save the Cook's distance as a variable and `filter` rows based on a threshold.

Illustrative Example 2

- Examine the scatterplots in R and identify which *outlier* has the most influence on the estimated regression line.

Example 3

- Using the data in *Illustrative Example 2*:

  1. Estimate the linear regression model for each pair of Y and X (three models).
  2. Use the `plot()` and `cooks.distance()` functions to obtain the observations with the largest Cook's Distance.
  3. How would you deal with each situation?

Example 4

- Based on the diagnostics we conducted on the linear regression model for the *Duncan* data which solutions might we apply to improve the model?

- Apply these solutions.

- Does the model improve with regards to the assumptions?

Example 5

- Does using the weighted least squares estimation method to estimate the linear regression model from Example 2 improve things?

- Compare the model formulas, how do they differ?

Exercise 1

- Using the *Wages.csv* dataset:
  - Estimate a linear model for the dependent variable *Salary*.
    - Express your results using a model formula and describe the meaning of all of the coefficients.
  - Follow all the steps covered in these slides to obtain a model that passes the diagnostic tests.
    - Do the model formulas differ?
    - Are you more comfortable making inferences from your final model?

References & Resources

1. Kaplan, Daniel T. (2017). *Statistical Modelling: A Fresh Approach. (Second Edition).* Retrieved from https://dtkaplan.github.io/SM2-bookdown/

2. Fox, J. (2015). *Applied regression analysis and generalized linear models (Third Edition).* Sage Publications.

- LAD Models
- `fitdist()`
- `vif()`
- `boxcox()`
- `cooks.distance()`