Regression Analysis

Linear model

Recall a linear model :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbb{E}[\boldsymbol{\epsilon}] = 0$, $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_n$, and $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_p]$ is a design matrix which we assume to be of full rank.

Model Diagnostics

Having fitted a classical linear model, we assess the validity of the model using diagnostic tools.

- Examination of the model assumptions
  - 1. linearity of the predictors, 2. constant variance, 3. uncorrelated, and 4. normally distributed errors
- Outliers
  - "unusual" points in feature or response spaces
- Collinearity
  - highly correlated predictors

Residuals

Recall, for each $i = 1, \ldots, n$, the ith residual is defined as

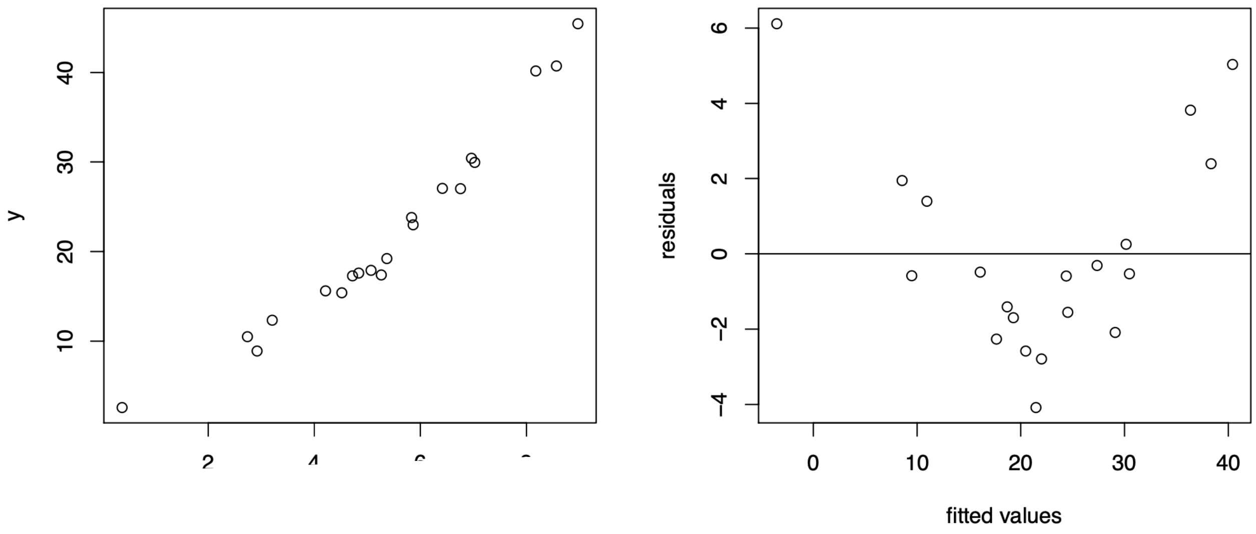
$$\widehat{\boldsymbol{\epsilon}}_i = \mathbf{Y}_i - \widehat{\mathbf{Y}}_i,$$

where $\widehat{\mathbf{Y}}_i = \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}$

Regression Diagnostics

1. linearity of the predictors

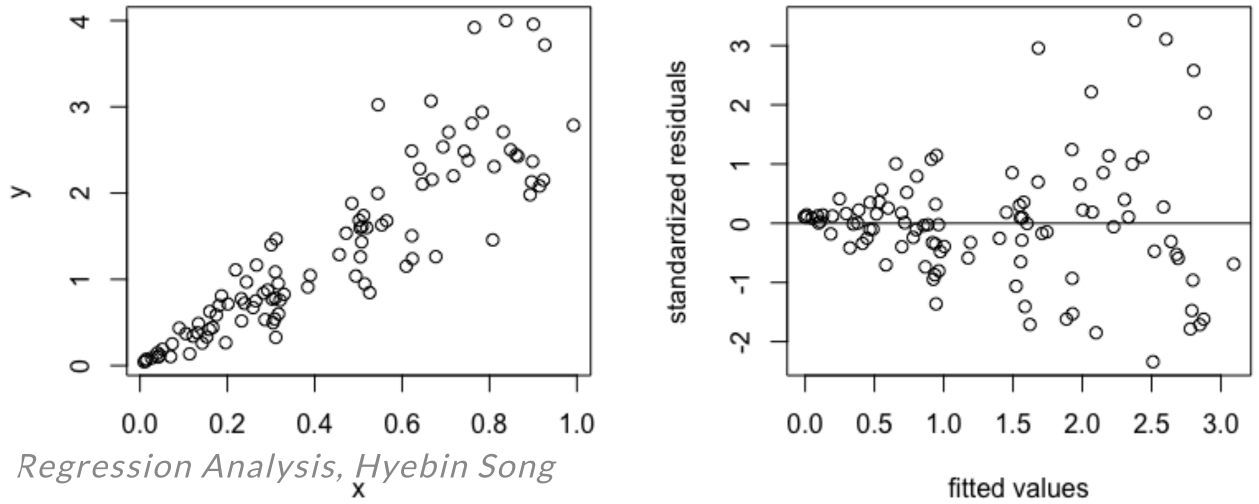
- plot $\hat{\boldsymbol{\epsilon}}_i$ against the fitted values $\widehat{\mathbf{Y}}_i$.
- we expect a plot with **no discernible trend or pattern**



Regression Diagnostics

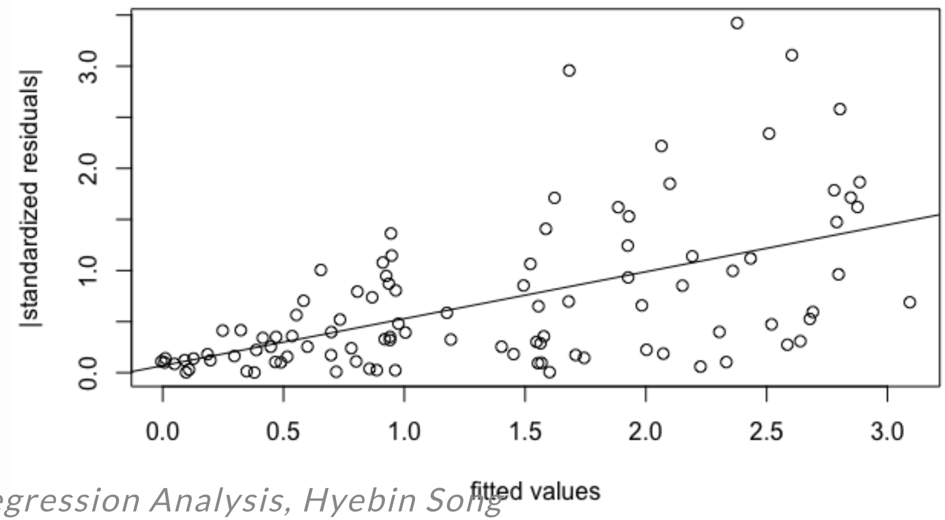
2. heteroscedasticity

- plot the standardized residuals r_i against the fitted values (or predictors)
- we expect **a "horizontal band" centered at zero.**



2. heteroscedasticity

- we may also plot $|r_i|$ against the fitted values $\widehat{\mathbf{Y}}_i$ or r_i^2 against the fitted values $\widehat{\mathbf{Y}}_i$.
  - If we add a least-squares line to this plot, we see whether there is any tendency for $|r_i|$ to increase or decrease with $\widehat{\mathbf{Y}}_i$.



Regression Diagnostics

3. Nonindependence of the error terms

- Possible forms of nonindependence:
  - Observations collected over time and/or across space.
  - Study done on sets of related subjects
- Correlated errors may indicate misspecification (especially through omitted explanatory variables and unnoticed nonlinearity). We should first examine whether model specification can be improved.

3. Nonindependence of the error terms

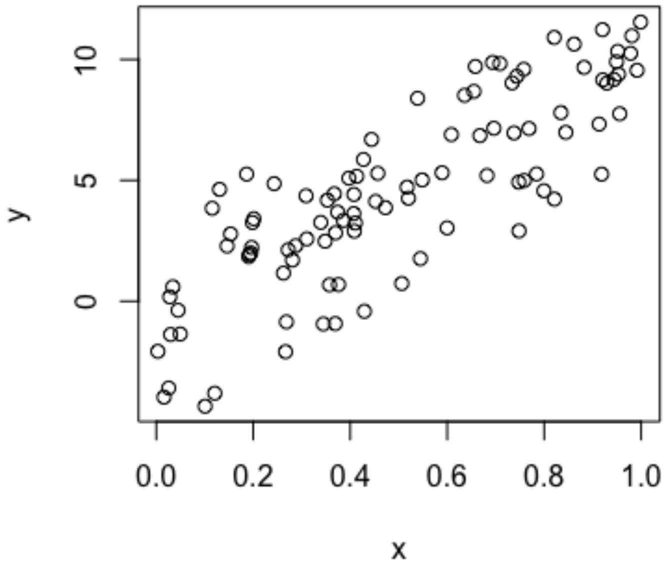
Example: Corn Yield

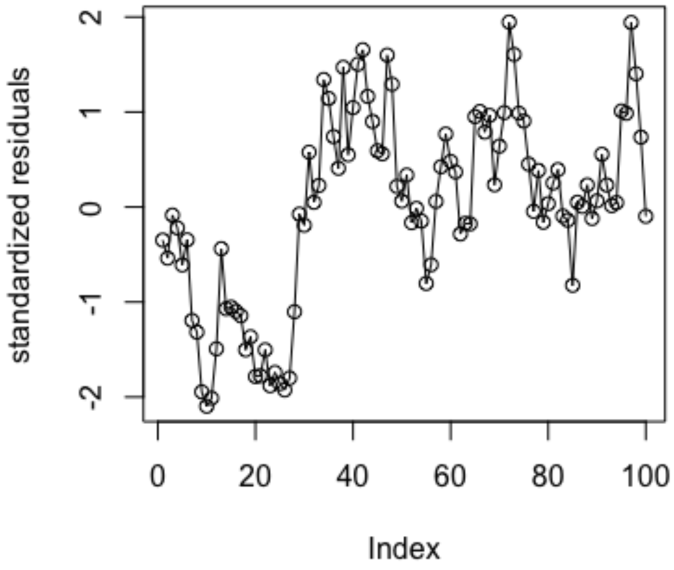
For $i = 1, ..., n$,

- i = the index of a $2m^2$ patch planted to corn and the patches are arranged in a long line at the edge of a field.
- x_i = the amount of fertilizer applied to the ith patch.
- Y_i = the corn yield in the ith patch.

3. Nonindependence of the error terms

Example: Corn Yield





Regression Diagnostics

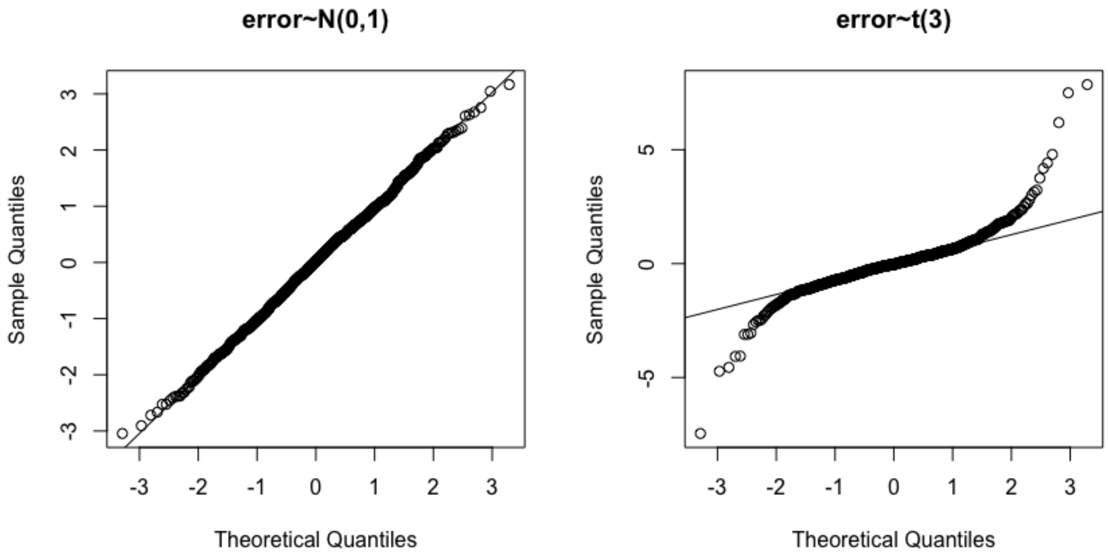
4. Nonnormality of error terms

- Normal QQ (quantile-quantile) plot of the standardized residuals.
    - **Idea**: if two distributions are similar, their quantiles should be similar.
        - We compare empirical quantiles of the standardized residuals with theoretical quantiles from N(0,1)
    - If the residuals are approximately normal, the normal QQ plot should be **approximately linear**

4. Nonnormality of error terms

Example: two Normal QQ plots for the standardized residuals

- Model1: $Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \epsilon_i \sim N(0, 3)$, iid
- Model2: $Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \epsilon_i \sim t(3)$, iid



Types of outliers

- An outlier is a data point which comes from a different distribution than the rest of the data
  - a big difference between the explanatory vector $\mathbf{x}_i$ for the ith case and the rest of the $\mathbf{x}$-data : "outlier in the x-direction" or "high-leverage point"
  - a large difference between the response Y_i and the mean $\mathbf{x}_i^\top \boldsymbol{\beta}$: "outlier in the y-direction", "error outlier", or "outlier"
- We say a point is a high-influence point if the regression result changes markedly after refitting without the observation of interest

Types of outliers

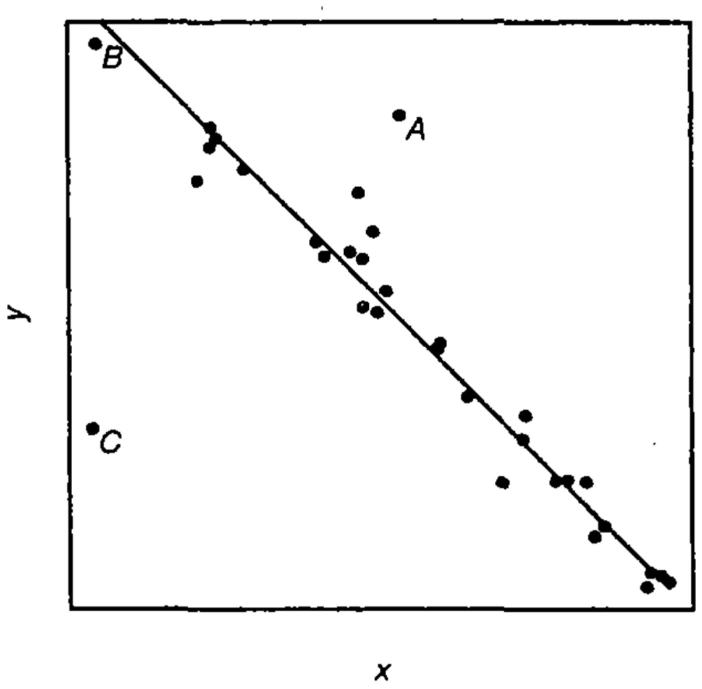


Figure credit: SL 9.1

Outliers (in the y-direction)

Say Y_n is an outlier. That is, we assume, $Y_i \sim N(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2), \quad i = 1, 2, \cdots, n-1$ but $Y_n \sim N(\Delta + \mathbf{x}_n^\top \boldsymbol{\beta}, \sigma^2)$ for some $\Delta \neq 0$.

1. diagnose by looking at absolute values of standardized/studentized residuals

- Recall both standardized and studentized residuals approximately follow $N(0,1)$ distribution.
- Any $|r_i| \geq 3$ or $|r_i^*| \geq 3$ are potential outliers

Outliers (in the y-direction)

Remarks

- If there are outliers detected, do not just throw them away, but search for an explanation (an effect not modeled, error in data entry, etc.)
- Not all outliers have a very strong influence on the fitted regression mean. However, an observation which is both outlier and high-leverage point may have a strong effect on the response mean.
- In the presence of an outlier, estimate model parameters **with** and **without** the outliers. Report both results.

Cook's Distance

Recall a $100(1-\alpha)\%$ confidence ellipsoid for $\boldsymbol{\beta}$ is

$$\left\{\mathbf{b} : (\mathbf{b} - \widehat{\boldsymbol{\beta}})^{\top} \mathbf{X}^{\top} \mathbf{X}(\mathbf{b} - \widehat{\boldsymbol{\beta}}) \leq pS^2 F_{p,n-p,\alpha}\right\}$$

Cook [1977] suggested measuring the distance of $\widehat{\boldsymbol{\beta}}_{(i)}$ from $\widehat{\boldsymbol{\beta}}$ by using the measure

$$D_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^{\top} \mathbf{X}^{\top} \mathbf{X}(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{pS^2}$$

Cook's Distance

Note,

$$D_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})^\top \mathbf{X}^\top \mathbf{X}(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{pS^2} = \frac{\|\widehat{\mathbf{Y}}_{(i)} - \widehat{\mathbf{Y}}\|^2}{pS^2}$$

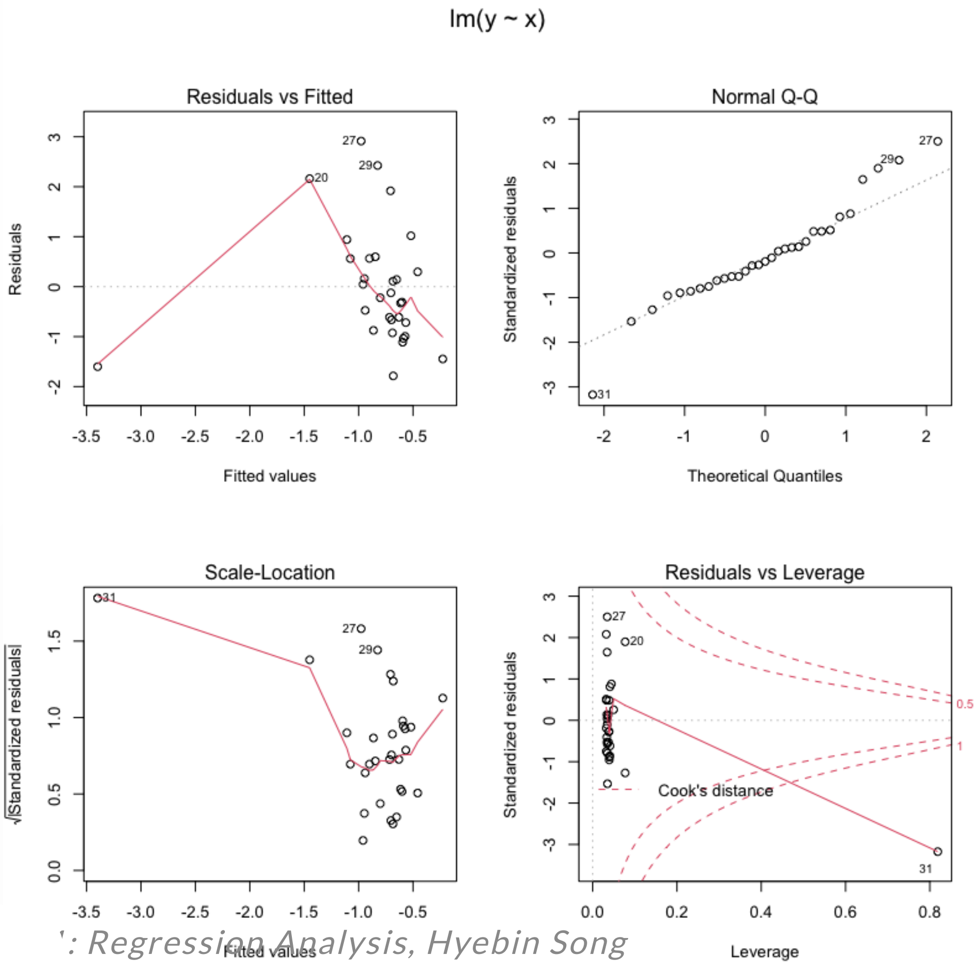
- Some recommend considering points for which $D_i > 1$ to be influential. Others suggest points for which $D_i > F_{p,n-p,.1}$

We can show that

$$D_i = r_i^2 \frac{\mathbf{H}_{ii}}{p\left(1 - \mathbf{H}_{ii}\right)}$$

- captures the distance of the ith observation from the other points in both x and y directions

Example `plot(lm(fit))`



Colinearity

One important assumption in the classical linear model is that $\mathbf{X}$ is full rank. In practice, however, the columns of $\mathbf{X}$ could be *almost* linearly dependent or colinear.

- In such case, $\mathbf{X}^\top \mathbf{X}$ is close to singular, i.e., the smallest eigenvalue of the matrix $\mathbf{X}^\top \mathbf{X} \approx 0$
- Since $\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1})$, the precision of $\hat{\boldsymbol{\beta}}$ is determined by $(\mathbf{X}^\top \mathbf{X})^{-1}$
    - near collinearity = large variances of the estimated coefficients

Remedies

Examination of the model assumptions $\rightarrow$ Revise a model assumption

1. linearity of the predictors
   - transform the response or predictors
   - include additional predictors into the model to account for the non-linear relationships
   - non-parametric regression (e.g., polynomial regression, splines)
2. constant variance
   - variable transformation (choose h so that $\mathrm{Var}(h(\mathbf{Y})) \approx \mathrm{const}$)
   - weighted least squares

Remedies

Examination of the model assumptions $\rightarrow$ Revise a model assumption

3. uncorrelated errors
    - model error variances ("general" linear model)
4. normally distributed errors
    - mild non-normality can be safely ignored (especially with light-tailed distributions)
    - for heavy-tailed distributions, base the inference on the assumption of another distribution, or use resampling methods for the inference

Remedies

Outliers, leverage points, and influential points $\rightarrow$ Examine each point

- do not throw them away, but search for an explanation
- estimate model parameters **with** and **without** the observations of interest. Report both results.

Remedies

- Collinearity $\rightarrow$ Revise $\mathbf{X}$ or use shrinkage methods
  - drop variables, ideally based on scientific knowledge.
  - obtain more data if possible.
  - create composite variables of collinear predictors, e.g., using a principle component analysis.
  - use regularization techniques.