Stat 206: Linear Models

Lecture 6

October 14, 2019

Adjusted Coefficient of Determination R_a^2

- A modified measure for degree of linear association between X and Y:

$$R_a^2 = 1 - \frac{MSE}{MSTO} = 1 - \frac{n-1}{n-2}\frac{SSE}{SSTO}.$$

- $R_a^2 \leq R^2 = 1 - \frac{SSE}{SSTO}$.
- Heights.

$$R_a^2 = 1 - \frac{927}{926} \times \frac{4659}{5893} = 0.2085.$$

Model Diagnostics

- Assumptions of the simple linear model with Normal errors:
  -
  -
  -
  -
- Diagnostic plots can be used to examine the appropriateness of these assumptions.
  - **Residual plots**.
- Remedial measures: transformations.

Residual Plots

- Examine regression relation and error variance.
  - Residual vs. predictor variable or residual vs. fitted value.
  - Residual vs. omitted predictor variable(s). (Later)
- Examine error distributions.
  - Normality: normal probability plot (Q-Q plot) of residuals.

Detection of Nonlinearity

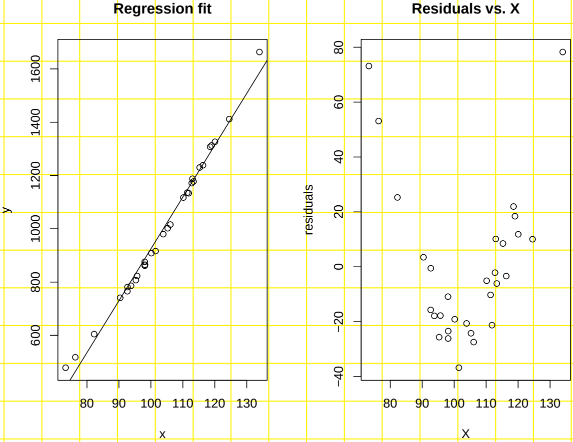
- **If the residual vs. predictor variable plot (or residual vs. fitted value plot) shows a**
  , then it is an indication of possible nonlinearity in regression relation.
- True model : $Y = 5 - X + 0.1X^2 + \varepsilon$.
  - 30 cases with $X \sim N(100, 16^2)$ and $\varepsilon \sim N(0, 10^2)$.
  - Summary statistics:

$$\overline{X} = 104.13, \overline{Y} = 1004.79, \sum_i X_i^2 = 330962.9, \sum_i Y_i^2 = 32466188, \sum_i X_i Y_i = 3249512.$$

  - Simple linear regression model was fitted to this data.

Coefficients	Estimate	Std. Error	t-statistic	P-value
Intercept	-1021.3803	40.0648	-25.49	$< 2 \times 10^{-16}$
Slope	19.4587	0.3814	51.01	$< 2 \times 10^{-16}$

$\sqrt{MSE} = 28.78$, $R^2 = 0.9894$, $R_a^2 = 0.989$.



Here, the scatter plot (left) is not very effective in showing the nonlinearity: the observations Y_i are close to the fitted values $\widehat{Y}_i$ due to the steep slope of the fitted regression line.

Detection of Nonconstancy in Variance

- **If the residual vs. predictor variable plot (or residual vs. fitted value plot) shows**

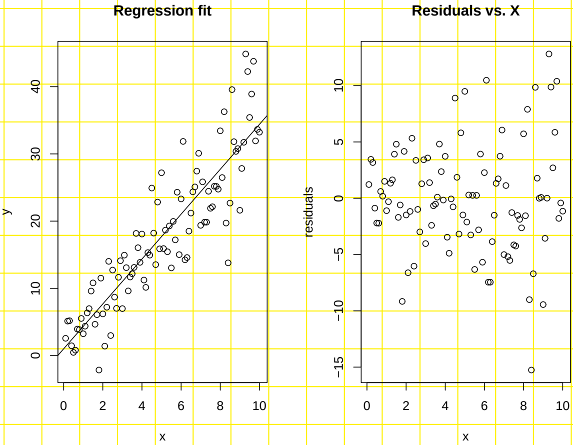
  , then this is an indication of unequal variance.

True model : $Y = 2 + 3X + \sigma(X)\varepsilon$, where $\log \sigma^2(X) = 1 + 0.1X$.

- 100 cases with $X_i = \frac{i}{10}$ and $\varepsilon_i \sim N(0, 1)$, $i = 1, \ldots, 100$.
- Simple linear regression model was fitted to this data.

Coefficients	Estimate	Std. Error	t-statistic	P-value
Intercept	1.0074	0.9729	1.035	0.303
Slope	3.3382	0.1673	19.958	$< 2 \times 10^{-16}$

$\sqrt{MSE} = 4.828$, $R^2 = 0.8026$.



Note the spread of the residuals with the value of X.

Detection of Nonnormality

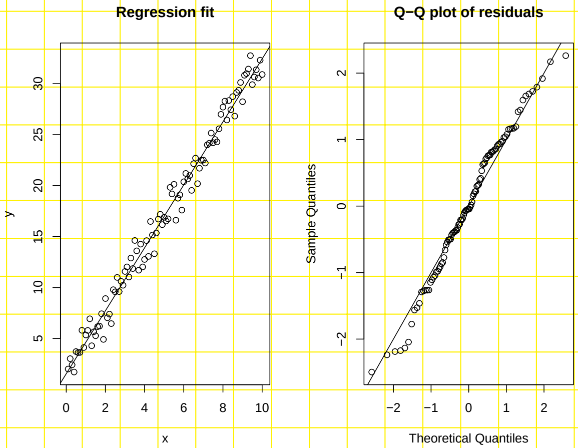
- **Normality of the errors can be examined by a normal probability plot, a.k.a. Q-Q plot.**
  - $z_{(k)}$'s: the *theoretical quantiles* under Normality
  - $e_{(k)}$'s: the *sample quantiles or empirical quantiles*.
  - Q-Q plot is simply a scatter plot of $e_{(k)}$'s vs. $z_{(k)}$'s.

Notes: Q-Q stands for quantile-quantile.

How to Read a Q-Q Plot?

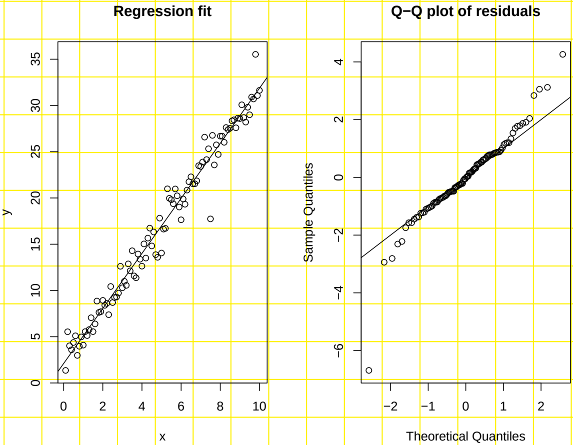
- If the errors are indeed normally distributed, then the points on the Q-Q plot should be .
- Departures from that could indicate **skewed** (non-symmetry) or **heavy-tailed** (probability mass in tails than a Normal distribution) distributions.
- Other types of departures (e.g., nonlinearity) may affect the distribution of the residuals and render them non-normal. **Thus it is better to examine other types of departures before checking normality.**

True model : $Y = 2 + 3X + \varepsilon$. $\varepsilon \sim N(0, 1)$.



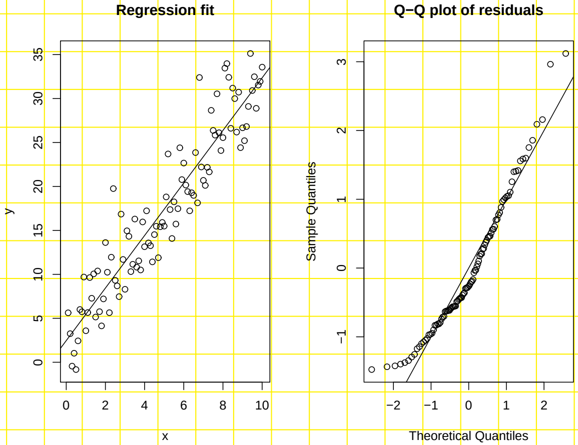
Q-Q plot shows a pattern.

True model : $Y = 2 + 3X + \varepsilon$. $\varepsilon \sim t_{(5)}$ – symmetrical but heavy-tailed errors.



Q-Q plot shows
than a Normal distribution.

True model : $Y = 2 + 3X + \varepsilon$. $\varepsilon \sim \chi^2_{(5)}$ – right-skewed errors.



Q-Q plots shows

.

Transformations to Treat Unequal Variance and Nornormality

- Unequal variance and nornormality often appear together.
- Transformations on Y may fix the error distributions.
  - $Y' = \sqrt{Y}$
  - $Y' = \log Y$
  - $Y' = 1/Y$
  - Sometimes, add a constant to the transformation, e.g., $Y' = \log(c + Y)$, to avoid negative or nearly zero values.
- A member from the family of power transformations may be chosen automatically by the **Box-Cox** procedure.
- Sometimes, a simultaneous transformation on X may be needed to maintain a linear relationship.

Box-Cox Procedure

- For each $\lambda \in R$, standardize Y_i^λ such that the magnitude of SSE does not depend on λ:

$$Y_i^* = \begin{cases} K_1 \frac{Y_i^\lambda - 1}{\lambda}, & if, \quad \lambda \neq 0 \\ K_2 \log(Y_i), & if, \quad \lambda = 0 \end{cases}$$

with

$$K_2 = (\prod_{i=1}^{n} Y_i)^{1/n}, \quad K_1 = 1/K_2^{\lambda-1}.$$

- Notes: $\lambda = 0$ corresponds to the logarithm transformation.
- For each λ, fit a regression model on the transformed data Y_i^* and derive $SSE(\lambda)$ (or maximum loglikelihood).
- Find the λ that maximizes loglikelihood.

(*Notes:* Read the lab 2 handout on Box-Cox procedure.)

Simple Linear Regression in Matrix Form

The regression equations:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \cdots n$$

can be written in a compact matrix form:

- **Response vector Y** and **error vector** : $n \times 1$ column vectors

- **Design matrix**: an $n \times 2$ matrix:

- **Coefficient vector**: a 2×1 column vector:

- Model assumptions:

$$E(\epsilon_i) = 0, \quad Var(\epsilon_i) = \sigma^2, \text{ for all } i = 1, \cdots, n$$

$$Cov(\epsilon_i, \epsilon_j) = 0, \quad \text{for all } i \neq j.$$

- Matrix form:

- In terms of the response vector **Y**:

- $\mathbf{0}_n$ is the $n \times 1$ zero vector, $\mathbf{I}_n$ is the $n \times n$ identity matrix.
- Mean of the error vector:

$$\mathbf{E}\{\boldsymbol{\epsilon}\} := \begin{bmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_n) \end{bmatrix} =$$

Variance-covariance matrix of the error vector:

$$\begin{aligned}\boldsymbol{\sigma}^{\mathbf{2}}\{\boldsymbol{\epsilon}\}: \quad &= \begin{bmatrix} Var(\epsilon_1) & Cov(\epsilon_1, \epsilon_2) & \cdots & Cov(\epsilon_1, \epsilon_n) \\ Cov(\epsilon_2, \epsilon_1) & Var(\epsilon_2) & \cdots & Cov(\epsilon_2, \epsilon_n) \\ \vdots & \vdots & \vdots & \vdots \\ Cov(\epsilon_n, \epsilon_1) & Cov(\epsilon_n, \epsilon_2) & \cdots & Var(\epsilon_n) \end{bmatrix} \\ &= \end{aligned}$$

Mean response vector: an $n \times 1$ column vector

$$\mathbf{E}\{\mathbf{Y}\} = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_i) \\ \vdots \\ E(Y_n) \end{bmatrix} =$$

Summary: Simple Linear Regression in Matrix Form

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times 2}{\mathbf{X}} \underset{2\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}.$$

- $\boldsymbol{\epsilon}$ is a random vector with $\mathbf{E}\{\boldsymbol{\epsilon}\} = \mathbf{0}_n$, $\boldsymbol{\sigma}^2\{\boldsymbol{\epsilon}\} = \sigma^2\mathbf{I}_n$.
- Normal error model: $\boldsymbol{\epsilon} \sim \mathrm{Normal}_n(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$.

Least Squares Estimation in Matrix Form

- Least squares criterion:

$$Q(b_0, b_1) = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2 .$$

- Matrix form :

$$Q(\mathbf{b}) = (\mathbf{Y} - \mathbf{X}\mathbf{b})' (\mathbf{Y} - \mathbf{X}\mathbf{b}) = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}.$$

- Differentiate Q with respect to $\mathbf{b}$:

$$\frac{\partial}{\partial \mathbf{b}} Q = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}.$$

- Set the gradient to zero $\Longrightarrow$ normal equation:

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}. \tag{1}$$

Least-square estimators are the solutions of equation (1).

- Multiply both sides of equation (1) by $(\mathbf{X}'\mathbf{X})^{-1}$:

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

- The left hand side becomes

- **LS estimators**:

$$\mathbf{X}'\mathbf{X} =$$

Note that $\mathbf{X}'\mathbf{X}$ *and* $(\mathbf{X}'\mathbf{X})^{-1}$ *are* *matrices.*

- LS estimators:

$$\begin{bmatrix}\hat{\beta}_0 \\ \hat{\beta}_1\end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \qquad\qquad\qquad\qquad ,$$

provided that X_is are .

- $n \times 1$ vector of fitted values:

$$\widehat{\mathbf{Y}} = \qquad\qquad\qquad\qquad ,$$

where $\mathbf{H} := \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the .

- $n \times 1$ vector of residuals:

$$\mathbf{e} = \qquad\qquad\qquad\qquad .$$

- Fitted values vector $\widehat{\mathbf{Y}}$ and residuals vector $\mathbf{e}$ are of the observations vector $\mathbf{Y}$.

Hat Matrix

$$\underset{n\times n}{\mathbf{H}} := \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

$\mathbf{H}$ and $\mathbf{I}_n - \mathbf{H}$ are **projection matrices**.

- Symmetric:
- **Idempotent**:
- Moreover, $rank(\mathbf{H}) =$, $rank(\mathbf{I}_n - \mathbf{H}) =$.

Column Space of the Design Matrix X

- Let $\mathbf{1}_n$ denote the $n \times 1$ vector of ones and $\mathbf{x} = (X_1, \cdots, X_n)^T$ denote the $n \times 1$ vector of design points.
- The design matrix
$$\mathbf{X} = (\mathbf{1}_n, \mathbf{x}).$$
- $\langle X \rangle$ is the
- $\langle X \rangle =$

Geometric Interpretation of Linear Regression

The hat matrix $\mathbf{H}$ projects a vector in $\mathbf{R}^n$ to the column space $\langle X \rangle$ of the design matrix $\mathbf{X}$: for any $\mathbf{w} \in \mathbf{R}^n$

- $\mathbf{Hw} \in \langle X \rangle$, i.e., there exists $c_0, c_1 \in \mathbf{R}$ such that $\mathbf{Hw} = c_0 \mathbf{1}_n + c_1 \mathbf{x}$.
- $\mathbf{w} - \mathbf{Hw} \perp \langle X \rangle$, i.e., for any $\mathbf{v} \in \langle X \rangle$, the inner product $< \mathbf{w} - \mathbf{Hw}, \mathbf{v} >= (\mathbf{w} - \mathbf{Hw})^T \mathbf{v} = 0$.

- $\widehat{\mathbf{Y}} = \mathbf{HY}$: the fitted values vector is the column space of $\mathbf{X}$:
- $\mathbf{e} = \mathbf{Y} - \mathbf{HY}$: the residuals vector is the column space of $\mathbf{X}$.
- So

Figure: Orthogonal projection of response vector **Y** onto the linear subspace of $\mathbb{R}^n$ generated by the columns of the design matrix **X**.