Multiple Linear Regression Models

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$
- Parameters included
  - β_0: intercept
  - β_k: regression coefficient (slope) for the kth explanatory variable
  - σ^2: variance of errors
- Observed (known): $y_i, x_{i1}, \ldots, x_{ip}$

  Unknown: $\beta_0, \beta_1, \ldots, \beta_p, \sigma^2, \epsilon_i$
- Random: ϵ_i, y_i

  Constants (not random): $\beta_k, \sigma^2, x_{ik}$

Linear Regression Models

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \cdots + \beta_p X_p + \epsilon$$

More generally, linear regression can be categorized as a model is linear in its parameters $\beta_0, \beta_1, \ldots, \beta_p$. For example, the following are linear regression models:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

$$Y = \beta_0 + \beta_1 \log(X) + \epsilon$$

even though the relationship between Y and X is not linear.

- Linear in parameters, not linear in predictors
- less restrictive than you might think

Which of the Following Models are Linear?

(a) $Y = \beta_0 + \beta_1 X + \epsilon$

(b) $Y = \beta_0 \beta_1^X \epsilon$

(c) $Y = \beta_0 + \beta_1 e^X + \epsilon$

(d) $Y = \beta_0 + \beta_1 X^2 + \beta_2 \log(X) + \epsilon$

Answer: (c) and (d)

Which of the following models can be turned linear after transformation?

Revisit: Assumptions about the Errors

The errors $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are

1. Independent

   Residual plot

2. with mean 0 and

   Since $E(Y|X) = X\beta$

3. constant variance σ^2, and

   Residual plot

4. (optional) normally distributed

   Check a normal probability plot

Assumptions about the Predictors

1. The predictors $X_1, X_2, \ldots, X_p$ are nonrandom fixed values.
2. The predictors $X_1, X_2, \ldots, X_p$ are measured withour error.
   - Hardly satisfied in real life.
   - Prediction are less accurate.
3. The predictors are linearly independent, i.e., no predictor can be expressed as a linear combination of others
   - No unique LS estimates for coefficients if there exist exact collinearity between predictors
   - Fine if there is no strong collinearity
   - Violation of this assumption is called multicollinearity, will be discussed later

Revisit: Assumptions about the observations

All observations are equally reliable and have an approximately equal role in determining the regression results and in influencing conclusions

Simple Linar Regression

When there is only one predictor X. The model is called simple linear regression model.

$$y = \beta_0 + \beta_1 x + \epsilon$$

- β_0 and β_1 are unknown population parameters, therefore are estimated from the data.
- ϵ is a random variable to represent noise/ uncertainty.

We can also write the simple linear regression equation as:

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

- $E(Y|X = x)$ is the expected value of Y for a given x value.

Multiple Linear Regression Models

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$
- Parameters included
  - β_0: intercept
  - β_k: regression coefficient (slope) for the kth explanatory variable
  - σ^2: variance of errors
- Observed (known): $y_i, x_{i1}, \ldots, x_{ip}$

  Unknown: $\beta_0, \beta_1, \ldots, \beta_p, \sigma^2, \epsilon_i$
- Random: ϵ_i, y_i

  Constants (not random): $\beta_k, \sigma^2, x_{ik}$

Recap: Errors and Residuals

- Error (ϵ_i) can not be directly computed,
  $\epsilon_i = y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip}$
- The errors ϵ_i can be estimatedby residuals e_i

  residual e_i = observed y_i - predicted y_i

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip})$$

- To estimate parameters, try to get $\hat{y}_i$ to be as close to y_i, so we want each $y_i - \hat{y}_i$ to be close to 0
- To make all of these residuals on average close to zero, consider minimizing the **sum of squares**

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Gauss–Markov Theorem (BLUE)

- Why should we focus on it and are there any better estimators?

  Under the Gauss–Markov model, the answer is definite: we focus on the OLS estimator because it is optimal in the sense of best linear unbiased estimator (BLUE).

  - It has the smallest covariance matrix among all linear unbiased estimators.

Theorem

Under the Gauss–Markov model, the OLS estimator $\hat{\beta}$ for β is the best linear unbiased estimator (BLUE) in the sense that $cov(\hat{\beta}) \preceq cov(\tilde{\beta})$ for any estimator $\tilde{\beta}$ satisfying

1. $\tilde{\beta} = AY$ *for some* $A \in \mathbb{R}^{(p+1)\times n}$ *not depending on* Y;
2. $\tilde{\beta}$ *is unbiased for* β.

Least Squares Solution to SLR

- To find $(\hat{\beta}_0, \hat{\beta}_1)$ that minimize:

$$L(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- One can set the derivatives of L with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ to 0

$$\frac{\partial L}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial L}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Least Squares Solution to SLR

- This results in the 2 equations with 2 unknowns:

$$n\hat{\beta}_0 + \hat{\beta}_1 \underbrace{\sum_{i=1}^{n} x_i}_{n\bar{x}} = \underbrace{\sum_{i=1}^{n} y_i}_{n\bar{y}} \xrightarrow{\text{divide by n}} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_0 \underbrace{\sum_{i=1}^{n} x_i}_{n\bar{x}} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \xrightarrow{\text{replace } \hat{\beta}_0} (\bar{y} - \hat{\beta}_1 \bar{x}) n\bar{x} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\iff \hat{\beta}_1 (\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) = (\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y})$$

$$\iff \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Least Squares Solution to MLR

- To find $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \beta_p)$ that minimize:

$$L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \beta_p) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2$$

- One can set the derivatives of L with respect to $\hat{\beta}_j$ to 0

$$\frac{\partial L}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}) = 0$$

$$\frac{\partial L}{\partial \hat{\beta}_k} = -2 \sum_{i=1}^{n} x_{ik} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}) = 0, \ k = 1, 2, \ldots, p$$

This results in a linear system of $(p+1)$ equations in $(p+1)$ unknowns.

Matrix Notation

$$\begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ip} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}x_{i1} & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{ik}x_{i1} & \sum_{i=1}^n x_{ik}x_{i2} & \cdots & \sum_{i=1}^n x_{ik}x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ip} & \sum_{i=1}^n x_{ip}x_{i1} & \sum_{i=1}^n x_{ip}x_{i2} & \cdots & \sum_{i=1}^n x_{ip}x_{ip} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_i \\ \vdots \\ \sum_{i=1}^n x_{ip}y_i \end{bmatrix}$$

- In matrix notation, the normal equation is $(X^T X)\hat{\beta} = X^T Y$
- And the least squares estimate is $\hat{\beta} = (X^T X)^{-1} X^T Y$

Properties of OLS estimator ($\hat{\beta}$)

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Theorem

Under the Gauss-Markov Model,

$$E(\hat{\beta}) = \beta$$

$$cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Proof. Because $E(Y) = X\beta$, we have

$$E(\hat{\beta}) = E\{(X^T X)^{-1} X^T Y\} = (X^T X)^{-1} X^T E(Y) = (X^T X)^{-1} X^T X \beta = \beta$$

Because $\text{cov}(Y) = \sigma^2 I_n$, we have

$$\begin{aligned} \text{cov}(\hat{\beta}) &= \text{cov}\{(X^T X)^{-1} X^T Y\} = (X^T X)^{-1} X^T \text{cov}(Y) X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1} \end{aligned}$$

Variance of estimation and prediction

- $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$
- $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 + \epsilon) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) + \sigma^2$
- As n gets larger,
  - $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0)$ would go down to zero
  - $\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0 + \epsilon)$ just goes down to σ^2

Statistical Inference for β_j

- $\hat{\beta} \sim \mathrm{N}(\beta, \sigma^2(X^T X)^{-1})$
- We can also denote it as $\hat{\beta}|X \sim \mathrm{N}(\beta, \sigma^2(X^T X)^{-1})$
- Therefore, for $j = 0, 1, \ldots, p$:

$$\hat{\beta}_j|X \sim \mathrm{N}(\beta_j, \mathrm{Var}(\hat{\beta}_j|X))$$

$\mathrm{Var}(\hat{\beta}_j|X)$: j+1 th diagonal entry of $\sigma^2(X^T X)^{-1}$

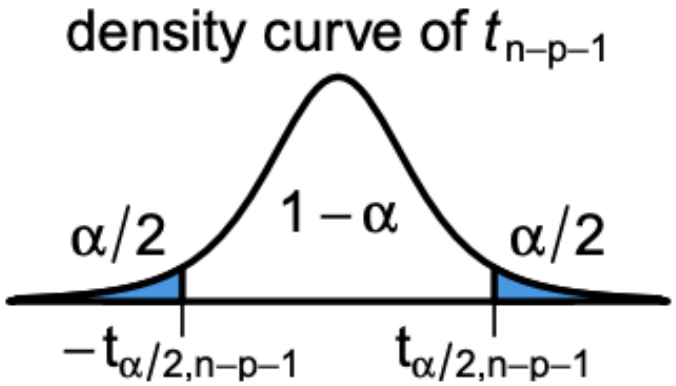
- $\dfrac{\hat{\beta}_j - \beta_j}{\sqrt{\mathrm{Var}(\hat{\beta}_j|X)}} \sim \mathrm{N}(0, 1)$
- Issue: σ^2 is unknown
- $\dfrac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\mathrm{Var}}(\hat{\beta}_j|X)}} = \dfrac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim t_{n-p-1}$

Confidence Interval

- $100(1-\alpha)\%$ confidence interval for β_j is:

$$\hat{\beta}_j \pm t_{n-p-1,\alpha/2} s.e.(\hat{\beta}_j)$$

where $t(n-p-1, \alpha/2)$ is the critical value for the t_{n-p-1} distribution at confidence level $1-\alpha$



which can be found using either of the following R commands:

```
qt(alpha/2, df=n-p-1, lower.tail=FALSE)
qt(1-alpha/2, df=n-p-1)
```

Mean and Covariance matrix of $\hat{Y}$ and $\hat{\epsilon}$

- Since $Y = \hat{Y} + \hat{\epsilon}$, where $\hat{Y} = X\hat{\beta} = HY$ and the residual vector is $\hat{\epsilon} = Y - \hat{Y} = (I_n - H)Y$, $H = X(X^TX)^{-1}X^T$
- Both H and $I_n - H$ are projection matrices. You can prove that $H(I_n - H) = (I_n - H)H = 0$

Theorem

Under the Gauss-Markov model,

$$E\begin{pmatrix}\hat{Y}\\ \hat{\epsilon}\end{pmatrix} = \begin{pmatrix}X\beta\\ 0\end{pmatrix}, \qquad cov\begin{pmatrix}\hat{Y}\\ \hat{\epsilon}\end{pmatrix} = \sigma^2\begin{pmatrix}H & 0_{n\times n}\\ 0_{n\times n} & I_n - H\end{pmatrix}$$

- So $\hat{Y}$ and $\hat{\epsilon}$ are uncorrelated.

Variance Estimation (σ^2)

- σ^2 is the variance of each ϵ_i, but the ϵ_i's are not observable
- We estimate ϵ by the residuals $\hat{\epsilon}_i = y_i - x_i^T \beta$
- $RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2$
  - We have shown that $\hat{\epsilon}_i$ has mean zero and variance $\sigma^2(1 - H_{ii})$.
  - So the mean of RSS (Residual Sum of Squares):

$$\sum_{i=1}^{n} \sigma^2(1 - H_{ii}) = \sigma^2[n - \text{trace}(H)] = \sigma^2[n - (p+1)]$$

Theorem

Define

$$\hat{\sigma}^2 = \frac{RSS}{n - (p+1)} = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n - (p+1)}$$

Then $E(\hat{\sigma}^2) = \sigma^2$ is an unbiased estimator under the Gauss Markov model.

Joint Distribution of $(\hat{\beta}, \hat{\sigma}^2)$

Theorem

Under the Gaussian linear model,

$$\begin{pmatrix} \hat{\beta} \\ \hat{\epsilon} \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} \beta \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} (X^T X)^{-1} & 0 \\ 0 & I_n - H \end{pmatrix} \right\}$$

so $\hat{\beta} \perp\!\!\!\perp \hat{\epsilon}$; $\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-(p+1)}$, *and* $\hat{\beta} \perp\!\!\!\perp \hat{\sigma}^2$

Proof.

$$\begin{pmatrix} \hat{\beta} \\ \hat{\epsilon} \end{pmatrix} = \begin{pmatrix} (X^T X)^{-1} X^T Y \\ (I_n - H) Y \end{pmatrix} = \begin{pmatrix} (X^T X)^{-1} X^T \\ I_n - H \end{pmatrix} Y$$

This is a linear transformation of Y , so they are jointly normal.

Joint distribution of $(\hat{Y}, \hat{\epsilon})$

Theorem

Under the Gaussian linear model,

$$\begin{pmatrix} \hat{Y} \\ \hat{\epsilon} \end{pmatrix} \sim \mathrm{N}\left\{ \begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} H & 0 \\ 0 & I_n - H \end{pmatrix} \right\}$$

so $\hat{Y} \perp\!\!\!\perp \hat{\epsilon}$

- Orthogonal: a linear algebra fact without assumptions.
- Independent: a statistical property under the Gaussian linear model.

Sum of Squares

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SSE}$$

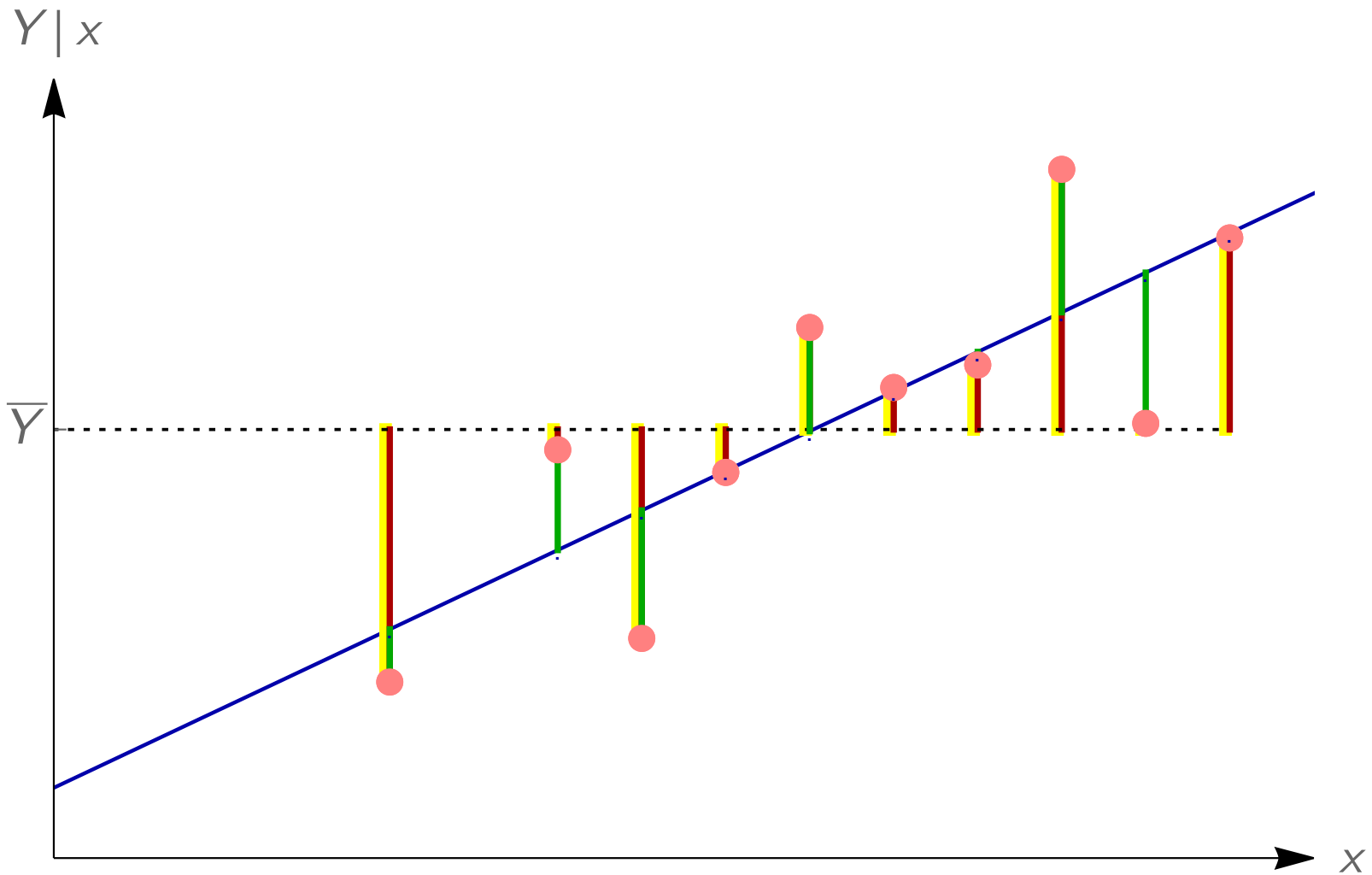
(Proof on board)

- SST = total sum of squares
  - total variability of Y
  - depends on the response Y only, not on the form of the model
- SSR = regression sum of squares
  - variability of Y explained by $X_1, \ldots, X_p$
- SSE = error (residual) sum of squares
  - variability of Y not explained by $X_1, \ldots, X_p$

Fundamental Sum-of-Squares Error Decomposition

$$\sum(\text{yellow lengths})^2 = \sum(\text{green lengths})^2 + \sum(\text{red lengths})^2$$

SST SSE SSR



Confidence Intervals VS Prediction Intervals

- $100(1-\alpha)\%$ confidence interval for $\beta_0 + \beta_1 x_0$ is:

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- $100(1-\alpha)\%$ prediction interval for $Y = \beta_0 + \beta_1 x_0 + \epsilon$ is:

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{n-2,\alpha/2}\hat{\sigma}\sqrt{1+\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Summary of Model Comparison Methods

- Nested models
  - F-test (Come along with P-values)
- Ant two models with the same response (no P-values)
  - MSE (Mean Squared Error = SSE/(n-p-1))
  - AIC (Akaike Information Criterion)
  - BIC (Bayesian Information Criterion)
  - Mallow's C_p (has fallen out of favor)
- Any two models with the same response (but possibly differently transformed)
  - Adjusted R^2

Confidence Intervals and Prediction Intervals

- $100(1-\alpha)\%$ confidence interval for $E(Y|X_0) = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p}$ is:

$$\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p} \pm t_{n-p-1,\alpha/2} s.e.(E(\hat{Y}|X_0))$$

- $100(1-\alpha)\%$ prediction interval for $Y = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} + \epsilon$ is:

$$\hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p} \pm t_{n-p-1,\alpha/2} s.e.(\hat{Y}|X_0)$$

Multiple R^2 and Adjusted R^2

- Multiple R^2 also called the coefficient of determination, is defined as

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- Interpretation: proportion of variability in Y explained by $X_1, \ldots, X_p$
- $0 \leq R^2 \leq 1$
- For SLR, $R^2 = r_{xy}^2$ is the square of the correlation between X and Y. So multiple R^2 is a generalization of the correlation
- For MLR, R^2 is the square of the correlation between Y and $\hat{Y}$
- When more terms are added into a model, R^2 may increase or stay the same but never decrease.
    - Is large R^2 always preferable?

Adjusted R^2

Since R^2 always increases as we add terms to the model, some people prefer to use an adjusted R^2 defined as

$$R^2_{adj} = 1 - \frac{SSE/df_{SSE}}{SST/dt_{SST}} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$
$$= 1 - \frac{n-1}{n-p-1}(1-R^2)$$

- $-\frac{p}{n-p-1} \leq R^2_{adj} \leq R^2 \leq 1$
- R^2_{adj} can be negative
- R^2_{adj} does not always increase as more variables are added. In fact, if unnecessary terms are added, R^2_{adj} may decrease.

General Framework for Testing Nested Models

H_0: Reduced model is true v.s. H_1: Full model is true

- Simplicity or Accuracy?
  - The full model fits the data better (with a smaller SSE) but it is more complex.
  - The reduced model doesn't fit as well but it is simpler.
  - If $\text{SSE}_{reduced} \approx \text{SSE}_{full}$: one can sacrifice a bit of accuracy in exchange for simplicity.
  - If $\text{SSE}_{reduced} \gg \text{SSE}_{full}$:it would sacrifice too much in accuracy in exchange for simplicity. The full model is preferred.
- How to quantify?
  - F statistic!

The F-Statistic

$$F = \frac{(\text{SSE}_{reduced} - \text{SSE}_{full})/(df_{reduced} - df_{full})}{\text{MSE}_{full}}$$

- Under H_0, F statistic has an F-distribution with $(df_{reduced} - df_{full}, df_{full})$ degrees of freedom
- $F \geq 0$ since $\text{SSE}_{reduced} \geq \text{SSE}_{full}$
- The smaller the F-statistic, the more the reduced model is favored

Example 1: Testing All Coefficients Equal to Zero

$H_0 : \beta_1 = \beta_2 \cdots = \beta_p = 0$ v.s. H_a: not all $\beta_1, \ldots, \beta_p = 0$

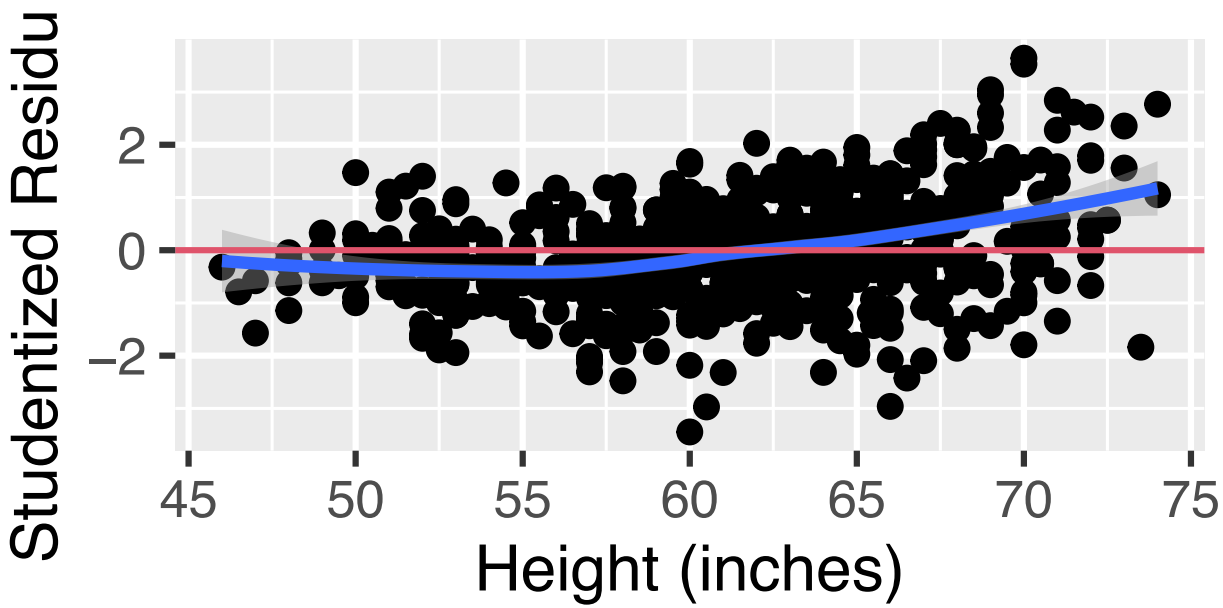
- The is a test to evaluate the **overall significance** of a model.
  - Full: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$
  - Reduced: $y_i = \beta_0 + \epsilon_i$ (All predictors are unnecessary)
- The OLS estimate for β_0 in the reduced model is $\hat{\beta}_0 = \bar{y}$, so $\text{SSE}_{reduced} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \text{SST}_{full}$
- $F = \dfrac{(\text{SSE}_{reduced} - \text{SSE}_{full})/(df_{reduced} - df_{full})}{\text{MSE}_{full}} = \dfrac{(\text{SST}_{full} - \text{SSE}_{full})/(n-1-(n-p-1))}{\text{MSE}_{full}} = \dfrac{\text{SSR}_{full}/p}{\text{MSE}_{full}} = \dfrac{\text{MSR}_{full}}{\text{MSE}_{full}}$
- Moreover, $F \sim F_{p,n-p-1}$ under $H_0 : \beta_1 = \beta_2 \cdots = \beta_p = 0$.

Residuals v.s. Potential Predictors

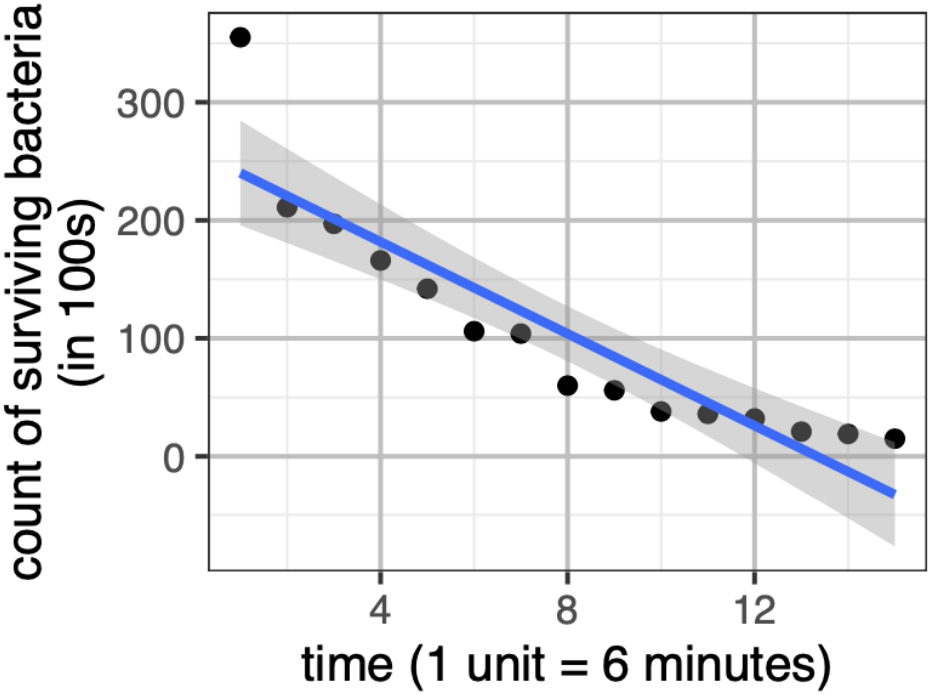
Recall the model lm1 doesn't not include ht (Height) as a predictor. Let's plot the residuals of lm1 against ht.

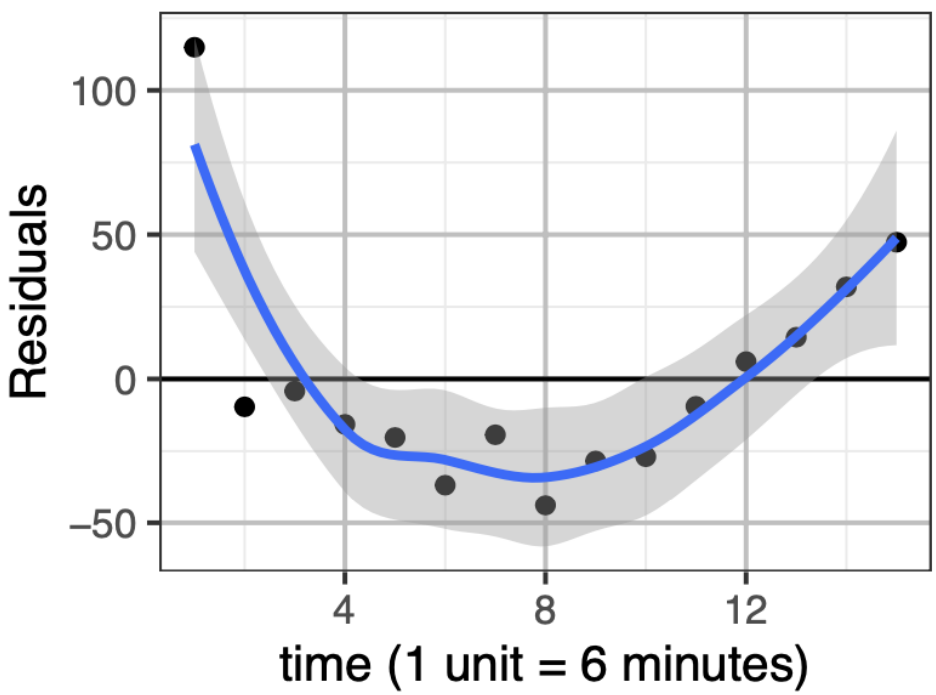
```
ggplot(fevdata, aes(x=ht, y=rstudent(lm1))) + geom_point()+geom_smooth(method='loess
  labs(x="Height (inches)", y="Studentized Residuals") +
  geom_hline(yintercept = 0, col=2)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The residuals clearly have a positive nonlinear relation with height, meaning ht should be included in the model.





What is your finding?

- When not knowing what transformation to make, we would begin by looking at the scatterplot (left). In this case, the scatterplot is obviously non-linear.
- In some cases, non-linearity may not be obvious in the scatterplot. Should always check the residual plot (right). We see that non-linearity is more obvious in the residual plot than in the scatterplot.

Polynomial Models

- Fitting the polynomial model:

$$\text{fev} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{age}^2 + \epsilon$$

doesn't mean we believe it is correct. It is just a decent approximation to the true underlying nonlinear model:

$$\text{fev} = f(\text{age}) + \epsilon$$

- One can try higher-order polynomials

$$\text{fev} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{age}^2 + \cdots + \beta_k (\text{age})^k + \epsilon$$

if lower-order ones don't capture the nonlinear pattern well.

Ordinal Categorical Predictors; Salary Data

- An ordinal variable is a categorical variable with ordered categories.
  - e.g., E = education (HS only, BA or BS, advance degree) in the salary survey data is an ordinal variable
- When we create indicators for E and include them in a model, we ignore the fact that the 3 education levels are ordered. Therefore, the estimated salary might not be ordered by education levels.
- We can incorporate the ordinal info of a ordinal predictor by assigning a score to each its category like

$$E = \begin{cases} 1 & \text{if HS only} \\ 2 & \text{if Bachelor's degree} \\ 3 & \text{if Advanced degree} \end{cases}$$

Check Interactions of Two Numerical Variables

The model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

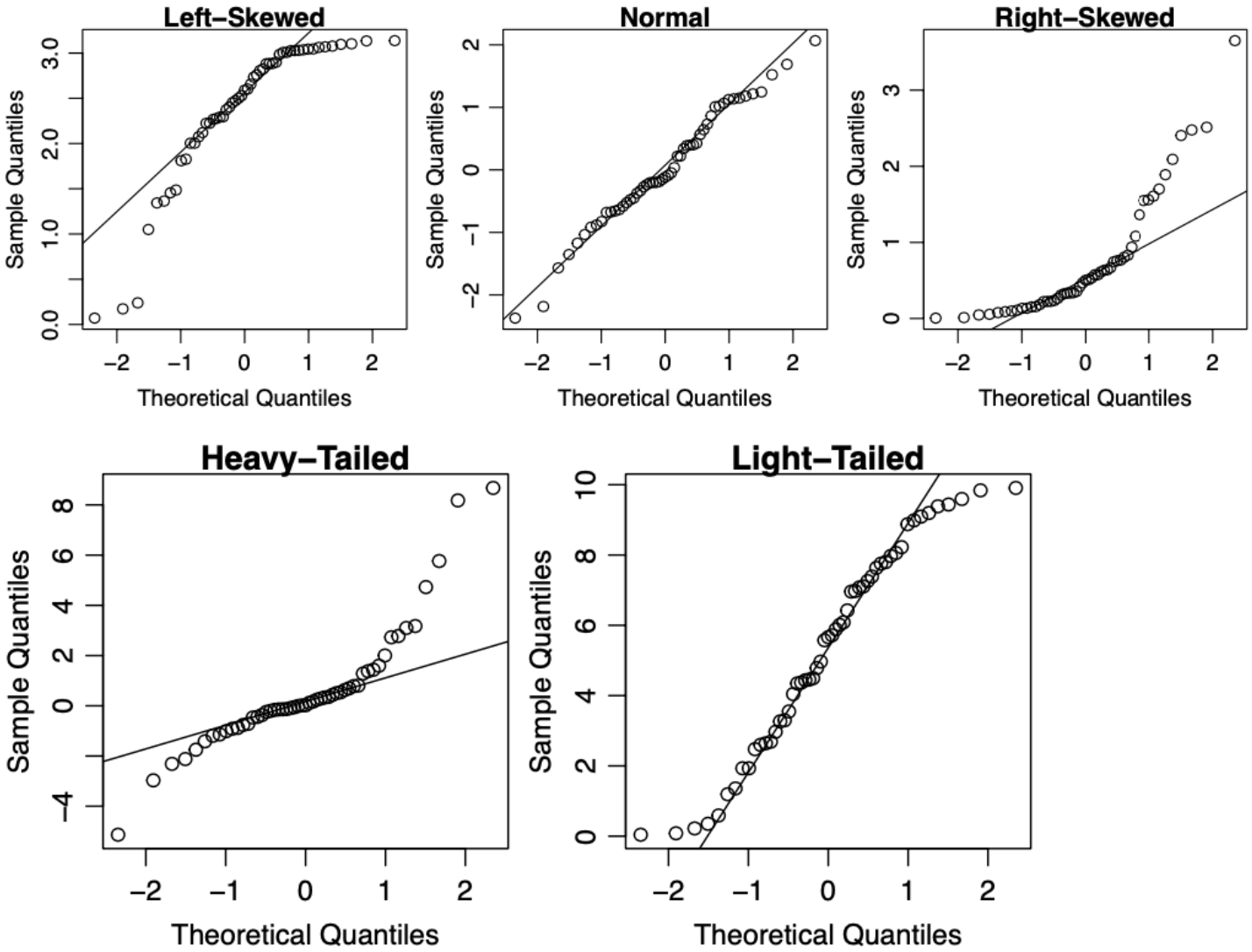
No interaction assumes that

- Y is linear in X_1 for each given X_2 and the slope of X_1 doesn't change with X_2
- Y is linear in X_2 for each given X_1 and the slope of X_2 doesn't change with X_1

Linearizable Models

Function	Transformation	Linear Form
$Y = \alpha X^{\beta}$	$Y' = \log Y, \quad X' = \log X$	$Y' = \log \alpha + \beta X'$
$Y = \alpha e^{\beta X}$	$Y' = \log Y$	$Y' = \log \alpha + \beta X$
$Y = \alpha + \beta \log X$	$X' = \log X$	$Y = \alpha + \beta X'$
$Y = \frac{X}{\alpha X - \beta}$	$Y' = \frac{1}{Y}, \quad X' = \frac{1}{X}$	$Y = \alpha - \beta X'$
$Y = \frac{e^{\alpha+\beta X}}{1 + e^{\alpha+\beta X}}$	$Y' = \log \frac{Y}{1 - Y}$	$Y' = \alpha + \beta X$

Different QQ Plots



Transformations to Reduce Skewness

Skewness can often be ameliorated by a power transformation.

$$f_\lambda(y) = \begin{cases} y^\lambda, & \text{if } \lambda \neq 0 \\ \log(y), & \text{if } \lambda = 0 \end{cases}$$

- If **right-skewed**, take $\lambda < 1$

$$y \longrightarrow \frac{1}{y}, \log(y), \sqrt{y}, \ldots (y^\lambda, \lambda < 1)$$

- If **left-skewed**, take $\lambda > 1$

$$y \longrightarrow y^2, y^3, \ldots (y^\lambda, \lambda > 1)$$

Residual Plus Component Plot

A Residual Plus Component Plot is a scatterplot of

$$(e + \hat{\beta}_j X_j) \quad \text{versus} \quad X_j$$

where e and $\hat{\beta}_j$ are from the regression of Y on all predictors, including X_j.

- The slope of the graph is $\hat{\beta}_j$
- This plot is useful to detect non-linearity in the partial relationship between Y and X_j.
- It adds a line indicating where the line of best fit lies

Box-Cox Method

Box-Cox method is an automatic procedure to select the "best" power λ that make the residuals of the model

$$Y^{\lambda} = \beta_0 + \beta_1 X1 + \cdots + \beta_p X_p + \epsilon$$

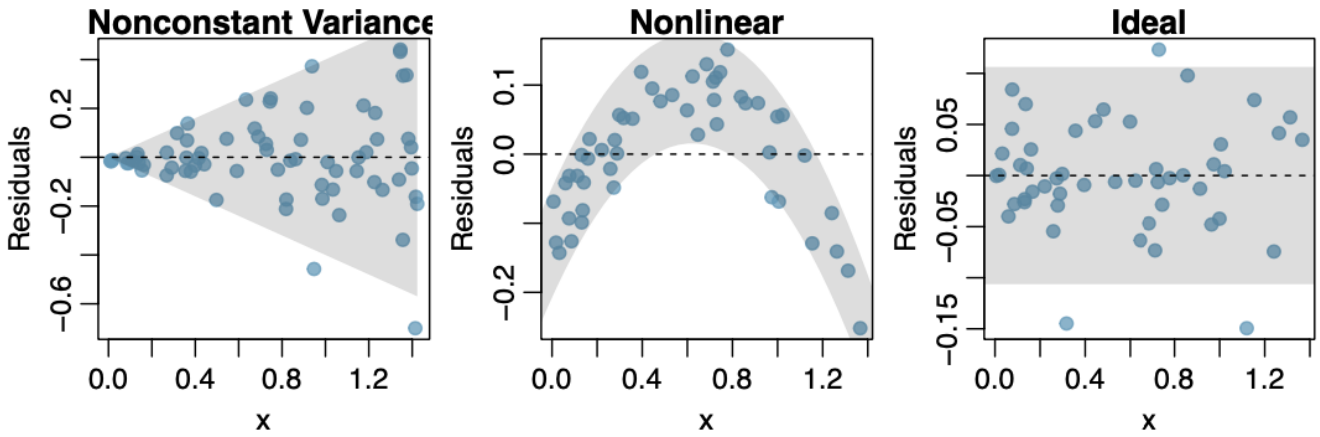
closest to **normal** and **constant** variability.

- We usually round the optimalλ to a convenient power like $-1, -1/2, -1/3, 0, 1/3, 1/2, 1, 2, \ldots$ since the practical difference of $y^0.5827$ and $y^0.5$ is usually small, but the square-root transformation is much easier to interpret.
- A confidence interval for the optimal λ can also be obtained (formula and details omitted). We usually select a convenient power λ^* in this C.I.

Various Kinds of Residual Plots

- Residuals v.s. fitted values
- Residuals v.s. each predictor
- Residuals v.s. potential predictors not yet included in the model
- Residuals v.s. several predictors using `ggplot()`
- Residuals v.s. time if the data are collected over time
- Residuals v.s. ...

In all the plots above, points should scatter evenly above and below the zero line in a band of constant width.



Weighted Least Squares Estimates

- The WLS estimate of $(\beta_0, \beta_1, \ldots, \beta_p)$ that minimizes the weighted sum of squares: $\sum_{i=1}^{n} w_i(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$
- Matrix Notation:

$$RSS = (Y - X\beta)^T W (Y - X\beta)$$

where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n1} & \ldots & x_{np} \end{bmatrix}, W = \begin{bmatrix} w_1 & & & & \\ & w_2 & & & \\ & & w_3 & & \\ & & & \ddots & \\ & & & & w_n \end{bmatrix}$$

and W is an $n \times n$ matrix with $(w_1, w_2, \ldots, w_n)$ on the diagonal and 0 elsewhere.

Leverage

$$\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \dots & h_{nn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$\hat{Y} = HY$ means every predicted value $\hat{y}_i$ is a linear combination of $y_1, \dots, y_n$

$$\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{in}y_n$$

and h_{ij} is the (i, j) th element of the matrix H, and is completely determined by the predictors X as $H = X(X^TX)^{-1}X^T$

- h_{ij} is the weight given to y_j in predicting $\hat{y}_i$.
- h_{ii} is the weight given to y_i in predicting $\hat{y}_i$, is called the **leverage** of i-th observation.

Leverage

- If the leverage of i-th observation, h_{ii}, is large (close to 1), then this i-th observation is called a leverage point. It means the prediction of $\hat{y}_i$ depends a lot on the observation y_i itself and relatively less on other observations.
- Recall in SLR, we can solve H analytically:

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n} (x_k - \bar{x})^2}$$

And the leverage in SLR is given by

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n} (x_k - \bar{x})^2}$$

Standardized Residuals

- Raw residuals have different variances $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$, therefore we cannot identify outliers by comparing the magnitude of raw residuals.
- We standardize the i-th residual e_i as

$$z_i = \frac{e_i}{\sigma\sqrt{1 - h_{ii}}}$$

- When we estimate σ by $\hat{\sigma} = \sqrt{MSE}$, we get the standardized residual or internally studentized residuals:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

- r_i has mean zero and standard deviation 1
- Observations w/ large $|r_i|$ (over 2 or 3 or 4) are potential outliers

Studentized Residuals = Externally Studentized Residuals

Externally studentized residuals or studentized residuals are defined as:

$$r_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

- e_i is still computed using all the data but
- $\hat{\sigma}_{(i)}$ is computed from the MSE of the model that uses all the data EXCEPT the i-th observation
- The subscript "(i)" means "all but the ith observation".
- Externally studentized residuals r_i^* can be calculated from internally studentized residuals r_i via $r_i^* = r_i\sqrt{\frac{n-p-2}{n-p-1-r_i^2}}$

  If an observation is not an outlier, $r_i^* \approx r_i$.

Comparisons of 3 Types of Residuals

- e_i's add up to 0, r_i's and r_i^*'s do not add up to 0
- e_i's have unequal variance, r_i's and r_i^*'s have variance 1
- r_i^*'s has a t-distribution with df=n-p-2, but r_i does not have a t-distribution.
- With a large enough sample, r_i's and r_i^*'s are approximately $N(0, 1)$
- None of the 3 types of residuals are strictly independent, but the dependence can be ignored with large enough samples

Influential Points

- An influential point has an unduly large effect on the model. The fitted model changes drastically when it is included.
- Observations whose removal will cause major changes in the regression analysis are called influential points.
- Changes such as the predicted responses, the estimated slope coefficients, or the hypothesis test results can be considered
- Influential points are not necessarily outliers
- A point can be influential, an outlier, or both

Cook's Distance

Cook's distance measures the difference between the fitted model values between the full data set and the -(i) data set.

$$C_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{\hat{\sigma}^2(p+1)}$$

for $i = 1, 2, \ldots, n$

- A more computationally efficient way to C_i:

$$C_i = \frac{1}{p+1} \cdot r_i^2 \cdot \frac{h_{ii}}{1 - h_{ii}}$$

- The second term $h_{ii}/(1 - h_{ii})$ is called the potential.
- Influential points have high a C_i compared to the other points

Outliers vs Influential Points