Lecture 1: Review of Normal Linear Regressions

(Reference: Chapter 2-3 of the book by Peter K. Dunn and Gordon K. Smyth)

Zhenxing Guo

Spring 2024

Table of Contents

Linear Regressions

Basics of regression:

- We observe a response or dependent variable Y
- With each Y, we also observe a set of regressors or predictors $\{x_1, ..., x_p\}$
- Goal: determine the mathematical relationship between response variables and regressors

$$Y = g(x) = g(x_1, ..., x_p)$$

In linear regressions, only linear functions of $x_1, ..., x_p$ are considered.

$$Y = g(x_1, ..., x_p) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p.$$

Ordinary Linear Regressions

In general, the former relationship may not hold exactly for the largely unobserved population of values of the independent and dependent variables; we call the unobserved deviations from the above equation the **errors**.

Definition (Ordinary Linear Regression)

With observed data $(\mathbf{x_1}, Y_1), \ldots, (\mathbf{x_n}, Y_n)$, we model the linear relationship as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

where, ϵ_1,...,ϵ_n are independent random errors, with $E[\epsilon_i] = 0$ and $var(\epsilon_i) = \sigma^2$.

For robustness of statistical inference, normal random error is usually assumed, i.e., $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$.

Weighted Linear regressions

In some cases, the homogeneity maybe violated by unequal variance among observations. A simple way to address this issue can be

$$var(\epsilon_i) = \sigma^2/\omega_i.$$

Definition (Weighted Linear regression models)

Consider linear regression models for modelling data $(\mathbf{x_1}, Y_1), \ldots, (\mathbf{x_n}, Y_n)$ with unequal variances among responses. A weighted linear regression model takes the following form,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

with $\epsilon_i \overset{independent}{\sim} N(0, \sigma^2/\omega_i)$. $\omega_i = 1$ (for all i) indicates ordinary linear regression model.

Necessity of weights

Example

A study that investigated the relationship between the mean birth weights and gestational ages of babies born to Caucasian mothers at St George's hospital, London, between August 1982 and March 1984.

	Age (x_i)	# of Births (N_i)	Mean weight (Y_i)
1	22	1	0.52
2	23	1	0.70
3	25	1	1.00
4	27	1	1.17
5	28	6	1.20
6	29	1	1.48
7	30	3	1.62
...			

In this example, if the birth weight of individual babies at gestational age x_i has a constant σ^2, then $var(Y_i) = \sigma^2/N_i$.

Table of Contents

OLS & WLS

The goal is to find estimates for β that "best" fit data points in the sense of least-squares: a line that minimizes the sum of squared residuals, or vertical the distance between predicted and observed responses.

Definition (OLS)

For ordinary linear regression, the OLS estimator is

$$\hat{\beta}_{OLS} = arg\min_{\beta} S(\beta) = arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 = \boldsymbol{(X^T X)^{-1} X^T Y}$$

Definition (WLS)

For weighted linear regression, the WLS estimator is

$$\hat{\beta}_{WLS} = arg\min_{\beta} \sum_{i=1}^{n} \omega_i (Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 = \boldsymbol{(X^T W X)^{-1} X^T W Y}$$

OLS & WLS

For OLR, OLS is the BLUE estimate based on Gauss–Markov theorem.

Theorem (Gauss–Markov theorem)

The OLS estimator has the lowest sampling variance within the class of linear unbiased estimators, if the errors in the linear regression model are uncorrelated, have equal variances and expectation value of zero.

For WLR, is WLS is its BLUE estimate?

Variance estimate

Take WLR as an example, and let $\hat{\beta} = \hat{\beta}_{WLS}$.

Definition (Residual sum-of-squares (RSS))

$$RSS = \sum_i \omega_i (Y_i - \hat{\mu}_i)^2 = \sum_i \omega_i (Y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2.$$

By definition, $\sigma^2/\omega_i = var[Y_i] = E[(Y_i - \mu_i)^2]$, then a sensible estimate for σ^2 is the average of the squared deviations $\frac{1}{n}\sum_i \omega_i (Y_i - \hat{\mu}_i)^2 = \frac{RSS}{n}$. In practice, a more frequently adopted one is the adjusted unbiased estimate:

$$\hat{\sigma}^2 = s^2 = \frac{RSS}{n - (p+1)},$$

with n and $p+1$ being the number of observations and regression coefficients.

Estimates in matrix form

Denote the $n \times 1$ vector of responses as Y, and the $n \times p'$ $(p' = p + 1)$ matrix of explanatory variables, called the model matrix, as $\boldsymbol{X} = [\boldsymbol{x_0}, \boldsymbol{x_1}, ..., \boldsymbol{x_p}]$, where $\mathbf{x}_j$ contains $n \times 1$ observed values for variable x_j. We write $\mathbf{x}_0$ for the vector of ones (the constant term) for convenience.

$$\begin{cases} \boldsymbol{Var[Y|X]} = \boldsymbol{W^{-1}\sigma^2} \\ \boldsymbol{E[Y|X]} = \mu = \boldsymbol{X}\beta. \end{cases}$$

where $\mathbf{W}^{-1}$ is a diagonal matrix with the diagonal elements (i, i) being $1/\omega_i$ and the off-diagonal elements are zero.

Estimates in matrix form

Matrix form for related terms,

- Coefficient estimate:

$$S(\beta) = (\boldsymbol{Y} - \boldsymbol{X}\beta)^T \boldsymbol{W}(\boldsymbol{Y} - \boldsymbol{X}\beta),$$
$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{Y} \tag{1}$$

- Variance estimate

$$s^2 = \frac{(\boldsymbol{Y} - \hat{\mu})^T \boldsymbol{W}(\boldsymbol{Y} - \hat{\mu})}{n - p'} = \frac{RSS}{n - p'}$$

- Standard error of coefficient

$$Var(\hat{\beta}) = Var((\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{Y}) = \sigma^2(\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X})^{-1}$$
$$\hat{Var}(\hat{\beta}) = s^2(\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X})^{-1}$$

- Variance of Fitted Values. Suppose x_g is the row vector of design matrix X, then

$$Var[\hat{\mu}_g] = [\boldsymbol{x_g}\hat{\beta}] = \boldsymbol{x_g}(\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{x_g}^T\sigma^2.$$

$$\hat{SE}[\hat{\mu}_g] = s\sqrt{\boldsymbol{x_g}(\boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{x_g}^T}.$$

Table of Contents

Distribution of $\hat{\beta}$

Note that $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{Y}$, a linear combination of $\mathbf{Y}$. If $\mathbf{Y}$ is normally distributed, then

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1})$$

All notes bellow for hypothesis testing and confidence interval estimate are based on normal assumption of $\mathbf{Y}$. When $\mathbf{Y}$ is not normally distributed, the asymptotic normal distribution of $\hat{\boldsymbol{\beta}}$ still holds for large sample size n (variance structure may be in a different form).

Simple hypothesis test

Consider

$$H_0 : \beta_j = \beta_j^0 \quad vs. \quad H_A : \beta_j \neq \beta_j^0.$$

Above β_j^0 is certain hypothesized value of β_j (usually zero).

Note that $\hat{\beta}_j \sim N(\beta_j, var(\hat{\beta}_j))$ with $var(\hat{\beta}_j)$ being a function of σ^2. To test the H_0, we usually use the statistics

$$T = \frac{\hat{\beta}_j - \beta_j^0}{se(\hat{\beta}_j)}.$$

- When σ^2 is known
  $T = \frac{\hat{\beta}_j - \beta_j^0}{se(\hat{\beta}_j)} \sim N(0,1)$, or, $[\hat{\beta}_j - u^*_{\frac{\alpha}{2}} se(\hat{\beta}_j), \quad \hat{\beta}_j + u^*_{\frac{\alpha}{2}} se(\hat{\beta}_j)]$ being the (1-α)CI under H_0.
- when σ^2 is unknown
  $T = \frac{\hat{\beta}_j - \beta_j^0}{\hat{se}(\hat{\beta}_j)} \sim t_{n-p'}$, or, $(1-\alpha)CI$: $[\hat{\beta}_j - t^*_{\frac{\alpha}{2},np'} \hat{se}(\hat{\beta}_j)$, $\hat{\beta}_j + t^*_{\frac{\alpha}{2},np'} \hat{se}(\hat{\beta}_j)]$ being the (1-α)CI under H_0.

Composite hypothesis test

Consider

$$H_0 : C^T\beta = C^T\beta^0 \quad vs. \quad H_A : C^T\beta \neq C^T\beta^0.$$

where $\mathbf{C}$ is a $(p+1)$ dimensional numerical vector.

- When σ^2 is known
  $T = \frac{C^T(\hat{\beta}-\beta^0)}{se(C^T\hat{\beta})} \sim N(0,1)$, or, $[C^T\hat{\beta} - u^*_{\frac{\alpha}{2}} se(C^T\hat{\beta}),$
  $C^T\hat{\beta} + u^*_{\frac{\alpha}{2}} se(C^T\hat{\beta})]$ being the (1-α)CI under H_0.
- when σ^2 is unknown
  $T = \frac{C^T(\hat{\beta}-\beta^0)}{se(C^T\hat{\beta})} \sim t_{n-p'}$, or, $(1-\alpha)CI$: $[C^T\hat{\beta} - t^*_{\frac{\alpha}{2},np'}\hat{se}(C^T\hat{\beta}),$
  $C^T\hat{\beta} + t^*_{\frac{\alpha}{2},np'}\hat{se}(C^T\hat{\beta})]$ being the (1-α)CI under H_0.

Table of Contents

ANOVA

For $\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon$, $\epsilon \sim N(0, \sigma^2 I_n)$, $\boldsymbol{X} = [\boldsymbol{x_0}, \boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_p}] \in R^{n \times (p+1)}$. After model fitting, each observation can be separated into a component predicted by the model, and the remainder or residual that is left over, as

$$Y_i = \hat{\mu}_i + (Y_i - \hat{\mu}_i).$$

$$Y_i - \bar{Y} = (\hat{\mu}_i - \bar{Y}) + (Y_i - \hat{\mu}_i).$$

Further, it easy to show that

$$\sum_i (Y_i - \bar{Y})^2 = \sum_i (\hat{\mu}_i - \bar{Y})^2 + \sum_i (Y_i - \hat{\mu}_i)^2.$$

$$SST = SSReg + RSS$$

Where SST is *total sum of square*, SSReg is *regression sum of squares* and RSS is *residual sum of squares*.

ANOVA

Above decomposition can be used to test whether the regression sum of squares SSReg is larger than would be expected due to random variation, or, whether the explanatory variables are useful predictors of the responses. Formally,

$$H_0 : \beta_1 = ... = \beta_j = ... = \beta_p = 0.$$

vs.

$$H_A : \exists j \text{ such that } \beta_j \neq 0.$$

To test H_0, we will need the following information

$$F = \frac{SSReg/p}{RSS/(n-(p+1))} = \frac{MSReg}{MSE} \sim F(p, n-(p+1)).$$

ANOVA

Table: The general form of an analysis of variance table for a linear regression model

Source of variation	Sums of squares	df	Mean square	F
Systematic comp	SSReg	p	MSReg: $\frac{SSReg}{p}$	$\frac{MSReg}{MSE}$
Random comp	RSS	n-(p+1)	MSE: $\frac{RSS}{n-(p+1)} = s^2$	
Total variation	SST	$n-1$		

Table of Contents

Compare nested model

Definition (Nested models)

Model A is nested in Model B if Model A can be obtained from Model B by setting some parameter(s) in Model B to zero or, more generally, if Model A is a special case of Model B.

Example

$$\begin{cases} ModelA : \mu_A = \beta_0 + \beta_1 x_1 + ... + \beta_{p_A} x_{p_A}; \\ ModelB : \mu_B = \beta_0 + \beta_1 x1 + ... + \beta_{p_A} x_{p_A} + ..., + \beta_{p_B} x_{p_B}. \end{cases}$$

Assuming $H_0 : \beta_{p_A+1} = ... = \beta_{p_B} = 0$ is true, the models are identical, which can be tested using statistic,

$$F = \frac{(RSS_A - RSS_B)/(p_B - p_A)}{s^2} = \frac{SS_B/(p_B - p_A)}{RSS_B/(n - p_B - 1)}.$$

SS_B measures the reduction in the RSS gained by using the more complex Model B. A P-value is deduced by referring to an $F(p_B - p_A, n - (p_B + 1))$.

Compare non-nested models

When the compared the models do not have nesting structure, criteria such as AIC and BIC be applied for model selection to balance accuracy and parsimony.

- Akaike's An Information Criterion (AIC) balances accuracy and parsimony, by measuring the accuracy using the RSS but penalizing the complexity of the model as measured by the number of estimated parameters. For a normal linear regression model, with σ^2 unknown,

$$AIC = n \log(RSS/n) + 2(p+1)$$

- Bayesian information criterion (BIC). For a normal linear regression model, with σ^2 unknown,

$$BIC = n \log(RSS/n) + (p+1) \log n,$$

Compare non-nested models

AIC vs. BIC

- Both AIC and BIC combine a term reflecting how well the model fits the data with a term that penalizes the number of parameters.
- The BIC requires stronger evidence for including more explanatory variables, so it prefers simpler models.
- Neither aic nor bic are formal testing methods, so no test statistics or P-values can be produced.

Table of Contents

Leverage

For

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 \mathbf{W}^{-1}), \mathbf{W} = diag\{\omega_1, ..., \omega_n\}.$$

- If $\omega_i = 1$ or $\mathbf{W} = I_n$, denote $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Then

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

or

$$\hat{Y}_i = h_{i1}Y_1 + h_{i2}Y_2 + ... + h_{in}Y_n = \sum_j h_{ij}Y_j$$

where h_{ij} is the coefficient applied to y_j to obtain $\hat{\mu}_i$. We call h_{ii} *leverages*, written $h_i = h_{ii}$, which quantifies the contribution of Y_i to its own prediction.

Leverage

- If $\omega_i \neq 1$.
  Let $\mathbf{Y}_\omega = \mathbf{W}^{1/2}\mathbf{Y}$, and $\mathbf{X}_\omega = \mathbf{W}^{1/2}\mathbf{X}$. Then $E[\mathbf{Y}_\omega] = \mu_\omega = \mathbf{X}_\omega\beta$, and $Var[\mathbf{Y}_\omega] = \sigma^2 I_n$. Then we have the hat matrix as,

$$\mathbf{H} = \mathbf{X}_\omega(\mathbf{X}_\omega^T\mathbf{X}_\omega)^{-1}\mathbf{X}_\omega^T = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}^\mathbf{T}\mathbf{W}\mathbf{X})^{1}\mathbf{X}^\mathbf{T}\mathbf{W}^{1/2}.$$

For diagnostics of unusual points:

- Leverages h_i depends on $\mathbf{X}$ and $\mathbf{W}$, not on $\mathbf{Y}$, and it quantifies the contribution of Y_i to its predicted value $\hat{Y}_i$.
- Small h_i indicates that many observations are contributing to the estimation of the fitted value $\hat{Y}_i$, while large h_i suggests $\hat{Y}_i$ will largely rely on its observation.
- In practice, large h_i means unusual combinations of the explanatory variables on this observation i, or i is an unusual point.
  - Observation i is declared high leverage if $h_i > 3p'/n$.

Raw residuals

For $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2\mathbf{W}^{-1})$, $\mathbf{W} = diag\{\omega_1, ..., \omega_n\}$, and $\boldsymbol{X} = [\boldsymbol{J_n}, \boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_p}] \in R^{n\times(p+1)}$.
By definition, residual

$$r_i = Y_i - \hat{\mu}_i.$$

Let $\mathbf{Y}_\omega = \mathbf{W}^{1/2}\mathbf{Y}$, and $\mathbf{X}_\omega = \mathbf{W}^{1/2}\mathbf{X}$, $\boldsymbol{\mu}_\omega = \boldsymbol{W}^{1/2}\boldsymbol{\mu}$.

$$\begin{aligned} Var[\mathbf{r}] &= Var(\boldsymbol{Y} - \hat{\boldsymbol{\mu}}) \\ &= Var[\boldsymbol{W}^{-1/2}(\boldsymbol{Y}_\omega - \boldsymbol{X}_\omega\beta_\omega)] \\ &= \boldsymbol{W}^{-1/2} var(\boldsymbol{Y}_\omega - \boldsymbol{H}\boldsymbol{Y}_\omega)\boldsymbol{W}^{-1/2} \\ &= \sigma^2\boldsymbol{W}^{-1/2}(\boldsymbol{I_n} - \boldsymbol{H})\boldsymbol{W}^{-1/2} \end{aligned}$$

Then

$$Var[r_i] = \sigma^2(1 - h_i)/\omega_i.$$

Standardized residuals

As r_i essentially is a linear combination of y, then for normal linear regressions, r_i is still normally distributed with mean 0 and variance $\sigma^2(1-h_i)/\omega_i$, or, $r_i \sim N(0, \sigma^2(1-h_i)/\omega_i)$. Then

$$r_i^* = \frac{\sqrt{\omega_i}(Y_i - \hat{\mu}_i)}{\sqrt{(1-h_i)}} \sim N(0, \sigma^2).$$

Then the standardized residuals

$$r_i' = \frac{r_i^*}{s} = \frac{\sqrt{\omega_i}(Y_i - \hat{\mu}_i)}{s\sqrt{(1-h^{i)}} \sim t_{n-(p+1)} \text{ with } \quad s^2 = \frac{RSS}{n-p-1}}$$

Approximately, r_i' follows standard normal distribution for large n.

Standardized residuals

Use residuals for overall diagnostics:

- Check model fits: Residual plot against x_j
  If the model fits well, the residuals should show no pattern, just constant variability around zero for all values of x_j. Any systematic trend in the residuals (e.g., a quadratic curve) would suggest that the residuals are correlated with x_j (i.e, not independently random), or the current function form of x_j is insufficient to predict Y.
  - Try to transform x_j or to include extra terms in the linear model.
- Check constant variance of response: Residual plot against μ_j
  An increasing or decreasing trend in the variability of the residuals about the zero line suggests that there are some dependence between the mean and variance of response.
  - Try to transform or change the scale of the response variable to achieve constant variance.
- Normality assumption: Q-Q plot for standardized residuals.

Outliers and Influential Points

- **Outliers** are observations inconsistent with the rest of the data set, which are located by identifying the unusual large (positive or negative) Studentized Residual:

$$r_i'' = \frac{\sqrt{\omega_i}(y_i - \hat{\mu}_{i(i)})}{s_{(i)}\sqrt{(1-h_i)}}$$

  where $\hat{\mu}_{i(i)}$ is the fitted value for Observation i computed from the model fitted without Observation i, $s^2_{(i)}$ is estimate of variance without observation i.
  - Here $\hat{\boldsymbol{\mu}}_{(i)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\omega}_{(i)}$, with $\hat{\boldsymbol{\beta}}^{\omega}_{(i)}$ is estimated without observation i.
- **Influential observations** are observations that substantially change the fitted model when omitted from the data set.

Outliers and Influential Points

Identification of influential observations relies on both large residuals and high leverage (i.e., outliers with high leverage), which can be done using the following criteria:

- Cook distance:

$$D_i = \frac{(r_i')^2}{p+1} \frac{h_i}{1-h_i}$$

  - Observation i is declared influential when D_i exceeds the 50th percentile of $F(p+1, n-p-1)$.

- DFFITS: measures how much $\hat{\mu}_i$ differs from $\hat{\mu}_{i(i)}$

$$DEFFITS_i = \frac{\hat{\mu}_i - \hat{\mu}_{i(i)}}{s_{(i)}} = r_i'' \frac{h_i}{1-h_i}$$

  - Observation i is declared influential when

$$|DFFITS_i| > 3/\sqrt{(p+1)/(n-p-1)}.$$

Outliers and Influential Points

- DFBETAS: is a coefficient-specific version of DFFITS that measures how much the estimates of each individual regression coefficient changes with/without observation i.

$$DFBETAS_i = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{se(\hat{\beta}_{j(i)})}$$

  - Observation i is declared influential when $|DFBETAS_i| > 1$.
- Covariance ratio (CR): measures the increase in uncertainty about the regression coefficients when Observation i is omitted.

$$CR = \frac{1}{1-h_i}\{\frac{n-p}{n-p-1+(r_i'')^2}\}^p$$

where r_i'' is the Studentized residual

  - Observation i is declared influential when $CR_i > 3(p+1)/(n-p-1)$.

Table of Contents

Transformation of response

Transformation: convert the response variable to a different measurement scale.

$$Y^* = h(Y)$$

$h()$ is some inversible function. After transforming the response, the basic linear regression model structure remains the same

$$\begin{cases} Y_i^* \sim N(\mu_i, \sigma^2/\omega_i) \\ \mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}. \end{cases}$$

Reason of transformation

Why do we want to transfer the response?

- Convert response to a scale where the linear predictor is unconstrained.

Example

$y* = \log(FEV)$, FEV is Forced expiratory volume, and > 0
$y* = \log(y + 0.5)$ or $y* = \log(y + 1)$ for counts with zeros.

- Make the distribution nearly normally distributed, more symmetric

Example

$y* = \log(y)$, $y* = y^{\lambda}(\lambda < 1)$ for right skewed data, $y* = y^{\lambda}(\lambda > 1)$ for left skewed data.

Reason of transformation

- (Most fundamental) Remove the mean–variance relationship to achieve close to constant variance across all observations. Suppose that Y has a mean–variance relationship defined by the function $V(\mu)$, with $Var[Y] = \phi V(\mu)$. Then, consider a transformation $Y^* = h(Y)$. A first-order Taylor series expansion of $h(Y)$ about μ gives $Y^* = h(Y) \approx h(\mu) + h'(\mu)(Y - \mu)$. Then

$$Var[Y^*] = Var[h(Y)] \approx h'(\mu)^2 Var[Y]$$

Then the variance will be stabilized if $h'(\mu)$ is proportional to $Var[Y]^{-1/2} = V(\mu)^{1/2}$.

Example

If $V(\mu) = \mu^2$ then $h'(\mu) = 1/\mu$. If $V(\mu) = \mu$ then $h'(\mu) = 1/\mu^{1/2}$

Box-Cox transformation

For non-negative response Y, the Box-Cox transformation is defined as

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda}, & \text{for } \lambda \neq 0 \\ \log(Y) & \text{for } \lambda = 0 \end{cases}$$

When can we try?

- The data is not normally distributed
- Residuals are not normally distributed or they don't have constant variance
- All data values must be greater than 0, and be continuous.

How to choose λ?

Suppose that after transformation $Y^{(\lambda)}$ is normally distributed, with mean $X\beta$ and variance σ^2. Then the data log-likelihood of $Y^{(\lambda)}$ is

$$logf_{\lambda}(Y_1^{(\lambda)}, ..., Y_n^{(\lambda)}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(Y^{(\lambda)} - X\beta)^T(Y^{(\lambda)} - X\beta)$$

Box-Cox transformation

Because $\frac{\partial Y_i^{(\lambda)}}{\partial y_i} = y_i^{\lambda-1}$ and $\frac{\partial y_i^{(\lambda)}}{\partial y_j, j \neq i} = 0$, then $|(\frac{\partial y_i^{(\lambda)}}{\partial y_j})_{i,j}| = \prod_i y_i^{\lambda-1}$, then

$$f(Y_1, ..., Y_n) = f_\lambda(Y_1^{(\lambda)}, ..., Y_n^{(\lambda)})|(\frac{\partial Y_i^{(\lambda)}}{\partial Y_j})_{i,j}| = f_\lambda(Y_1^{(\lambda)}, ..., Y_n^{(\lambda)}) \prod_i Y_i^{\lambda-1}$$

Then, the data likelihood of $Y = (Y_1, ..., Y_n)$ is

$$l(\lambda, \beta, \sigma^2) = logf(Y_1, ..., Y_n) = logf_\lambda(Y_1^{(\lambda)}, ..., Y_n^{(\lambda)}) + (\lambda - 1) \sum_i \log Y_i$$

$$= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(Y^{(\lambda)} - X\beta)^T(Y^{(\lambda)} - X\beta) + (\lambda - 1) \sum_i \log Y_i$$

Given λ, we have $\hat{\beta}(\lambda) = (X^T X)^{-1} X^T Y^{(\lambda)}$, $\hat{\sigma}^2(\lambda) = \frac{RSS(\lambda)}{n}$. Then

$$\lambda^* = arg \max_\lambda l(\lambda, \hat{\beta}(\lambda), \hat{\sigma}^2(\lambda))$$

$$= arg \max_\lambda \{-\frac{n}{2} \log[RSS(\lambda)/n] + (\lambda - 1) \sum_i \log(y_i)\}$$

Transformation not always work

- Hard to find a transformation that can remove the mean-variance dependence: $var(Y) = g(\mu)$, and g also contains other unknown parameters.
- Data is not continuous: binary, binomial, ordinal response.
- Survival outcome, incomplete observation.
- Further, data with longitudinal/repeated measurements where some responses are correlated.