Contents

KATA PENGANTAR

Puji dan syukur penulis panjatkan ke hadirat Tuhan Yang Maha Esa karena atas rahmat dan karunia-Nya, penulis dapat menyelesaikan penulisan eBook ini yang berjudul "Analisis Data Kategori".

eBook ini disusun sebagai hasil pembelajaran dan pendalaman materi dari perkuliahan Analisis Data Kategori yang diampu oleh Dr. I Gede Nyoman Mindra Jaya, dosen Statistik di Universitas Padjadjaran. Selama proses perkuliahan, penulis banyak mendapatkan wawasan mendalam tentang prinsip-prinsip dasar hingga lanjutan dalam analisis data kategori, termasuk teori probabilitas diskrit, model log-linear, regresi logistik biner, multinomial, dan ordinal, serta berbagai uji asosiasi untuk data tabel kontingensi.

Tujuan penulisan eBook ini adalah untuk merangkum dan menyusun kembali seluruh materi yang telah dipelajari dalam format yang lebih sistematis, aplikatif, dan mudah dipahami oleh mahasiswa atau peneliti yang tertarik dengan analisis data kategorik. Beberapa pendekatan analisis dijelaskan lengkap dengan teori dasar, rumus matematis, contoh kasus nyata, serta interpretasi hasil menggunakan software statistik.

Penulis menyadari bahwa eBook ini masih memiliki banyak kekurangan, baik dari segi isi maupun penyajian. Oleh karena itu, kritik dan saran yang membangun sangat penulis harapkan untuk penyempurnaan edisi selanjutnya.

Akhir kata, penulis mengucapkan terima kasih yang sebesar-besarnya kepada Dr. I Gede Nyoman Mindra Jaya atas ilmu, bimbingan, dan inspirasi akademik yang sangat berarti. Semoga eBook ini dapat memberikan manfaat bagi siapa pun yang ingin memperdalam pemahaman tentang analisis data kategori dalam konteks akademik maupun penelitian terapan.

I. Pendahuluan

Dalam dunia nyata, banyak fenomena yang kita amati tidak berbentuk angka kontinu, melainkan berupa kategori—misalnya status merokok (ya/tidak), jenis kelamin (pria/wanita), tingkat kepuasan (puas/tidak puas), atau hasil diagnosis (positif/negatif). Data seperti ini disebut sebagai data kategori. Seiring berkem bangnya kebutuhan analisis pada bidang-bidang terapan seperti epidemiologi, ilmu sosial, ekonomi, dan kesehatan masyarakat, kebutuhan akan metode analisis yang sesuai untuk data non-kuantitatif pun se makin meningkat. Inilah yang melatarbelakangi berkembangnya analisis data kategori sebagai suatu cabang penting dalam statistik terapan.

1. Tujuan Analisis Data Kategori

Tujuan utama dari analisis data kategori adalah untuk memahami dan memodelkan hubungan antara variabel-variabel kategori, baik dalam bentuk perbandingan proporsi, asosiasi antar kategori, maupun model regresi yang disesuaikan untuk data kategori (seperti regresi logistik). Dengan menggunakan metode-metode ini, kita dapat menguji hipotesis, mengukur kekuatan asosiasi, serta memprediksi probabilitas kejadian berdasarkan karakteristik tertentu.

2. Definisi dan Ruang Lingkup Analisis Data Kategori

Analisis data kategori merupakan cabang statistik yang fokus pada pengolahan dan interpretasi data yang dinyatakan dalam bentuk kategori, bukan nilai numerik kontinu. Data ini umumnya diklasifikasikan sebagai nominal (tanpa urutan, misalnya warna atau jenis kelamin) atau ordinal (memiliki urutan, misalnya tingkat kepuasan: rendah, sedang, tinggi). Ruang lingkup analisis ini meliputi:

- Analisis tabel kontingensi (seperti tabel 2x2 dan RxC)
- Uji chi-kuadrat dan Fisher's exact test
- Pengukuran asosiasi (seperti risk difference, relative risk, dan odds ratio)
- Model regresi untuk data kategori, termasuk regresi logistik dan regresi poisson

3. Perbedaan dengan Data Kuantitatif

Berbeda dengan data kuantitatif yang dianalisis menggunakan teknik seperti regresi linier atau ANOVA, data kategori memerlukan pendekatan yang disesuaikan karena tidak memenuhi asumsi-asumsi seperti normalitas atau varians homogen. Data kategori tidak memiliki skala interval yang tetap, sehingga perhitungan seperti rata-rata seringkali tidak bermakna. Sebagai gantinya, analisis lebih berfokus pada frekuensi, proporsi, dan probabilitas. Model statistik yang digunakan pun lebih banyak menggunakan distribusi diskrit, seperti distribusi binomial dan multinomial, daripada distribusi kontinu seperti normal.

4. Manfaat Analisis Data Kategori dalam Berbagai Bidang

Analisis data kategori memiliki aplikasi luas di berbagai bidang:

- Kesehatan masyarakat & epidemiologi: untuk menganalisis hubungan antara paparan dan penyakit (misalnya pengaruh merokok terhadap kanker paru-paru).
- Ilmu sosial: mengevaluasi preferensi pilihan, perilaku voting, dan hubungan antar kelompok sosial.
- Pemasaran: memahami kategori perilaku konsumen dan preferensi produk.

- Psikologi: mengkaji data kuesioner dan skala sikap.
- Ilmu politik dan hukum: mengamati perbedaan putusan hukum berdasarkan kategori tertentu seperti jenis kelamin atau ras.

Analisis data kategori membantu pengambilan keputusan berbasis bukti (evidence-based decision making), memungkinkan perumusan kebijakan, dan pengembangan teori dalam berbagai disiplin ilmu.

II. Metode dalam Analisis Data Kategori

1. Analisis Tabel Kontingensi dan Chi-Square

Analisis tabel kontingensi digunakan untuk menggambarkan dan menguji asosiasi antara dua variabel kategori, baik nominal maupun ordinal. Metode ini sangat tepat saat data disusun dalam bentuk tabel frekuensi silang, misalnya hubungan antara jenis kelamin dan preferensi produk. Keunggulannya terletak pada kesederhanaan penggunaannya dan kemampuannya untuk mengidentifikasi pola asosiasi secara eksplisit dalam data.

Rumus Manual (Chi-kuadrat Pearson):

$\chi^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}}, \quad E_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$

Uji Hipotesis:

- H_0: Tidak ada asosiasi antara variabel
- H_1: Terdapat asosiasi antara variabel
- Kriteria: Tolak H_0 jika $\chi^2 \geq \chi^2_{\alpha,df}$

2. Fisher's Exact Test

Digunakan saat ukuran sampel kecil atau terdapat frekuensi sel < 5 dalam tabel 2x2. Metode ini sangat bermanfaat dalam situasi dengan data langka karena menghasilkan nilai p yang akurat tanpa bergantung pada distribusi asimtotik.

Rumus Manual:

$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$

Kriteria Uji:
Bandingkan p-value langsung dengan nilai .

3. Odds Ratio (OR), Relative Risk (RR), dan Risk Difference (RD)

Metode ini digunakan untuk mengukur kekuatan asosiasi antara dua variabel biner, khususnya dalam studi epidemiologi. OR sering digunakan dalam studi kasus-kontrol, sedangkan RR dan RD lebih umum pada studi kohort.

Tabel 2x2:

	Kasus (Y=1)	Bukan Kasus (Y=0)
Terpapar	a	b
Tidak Terpapar	c	d

- OR = (a * d) / (b * c)
- RR = [a / (a + b)] / [c / (c + d)]
- RD = [a / (a + b)] – [c / (c + d)]

4. Regresi Logistik Biner

Model ini digunakan untuk memodelkan variabel dependen biner (misalnya: ya/tidak) berdasarkan satu atau lebih variabel independen.

$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$

Uji Hipotesis:

- H_0: $\beta = 0$

$Z = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}\right)$

- Tolak H_0 jika $|Z| > Z_{(1-\alpha/2)}$

5. Regresi Logistik Multinomial

Model ini digunakan ketika variabel dependen memiliki lebih dari dua kategori nominal tanpa urutan. Memodelkan probabilitas dari setiap kategori dibandingkan dengan kategori referensi.

$\log\left(\frac{\pi_j}{\pi_{ref}}\right) = \beta_{0j} + \beta_{1j} X_1 + \cdots + \beta_{kj} X_k$

Uji Hipotesis:

H_0: $\beta_{ij} = 0$ untuk semua j dan i

$G^2 = -2[logL_0 - logL_1]$

Tolak H_0 jika $G^2 \quad \chi^2_{(df,\alpha)}$

6. Regresi Logistik Ordinal (Proportional Odds Model)

Model ini digunakan saat variabel dependen bersifat ordinal. Keunggulannya adalah efisiensi dalam menangkap informasi urutan.

$\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \alpha_j - \beta X$

Uji Hipotesis:

- H_0: $\beta = 0$

$Z = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}\right)$

Tolak H_0 jika $|Z| > Z_{(1-\alpha/2)}$

7. Cochran-Mantel-Haenszel (CMH)

Digunakan untuk menguji asosiasi antara dua variabel kategori setelah mengontrol satu variabel stratifikasi (confounder).

Rumus:

$\chi^2_{CMH} = \frac{\left[\sum_k (n_k - E_k)\right]^2}{\sum_k \text{Var}(n_k)}$

Uji Hipotesis:

H_0: Tidak ada asosiasi setelah mengontrol tabel

Tolak H_0 jika $\chi^2_{CMH} \geq \chi^2_{(1,\alpha)}$

8. Decision Tree (CART)

Decision tree digunakan untuk mengklasifikasikan kategori output berdasarkan hierarki pembelahan fitur. Cocok untuk eksplorasi data atau prediksi dengan interpretasi visual.

Rumus Manual:

- **Gini Index:**

$Gini(t) = 1 - \sum_{i=1}^{C} p_i^2$

Entropy:

$Entropy(t) = -\sum_{i=1}^{C} p_i \log_2(p_i)$

III. Distribusi Probabilitas Dalam Data Kategori

1. Distribusi Bernoulli

Distribusi Bernoulli digunakan untuk memodelkan data kategori biner, yaitu kejadian dengan dua hasil, seperti "ya/tidak" atau "sukses/gagal". Distribusi ini diterapkan saat hanya ada satu percobaan atau observasi, dan digunakan untuk membangun fondasi dari model yang lebih kompleks seperti regresi logistik. Keunggulannya adalah kesederhanaannya dan kemampuannya menjadi dasar dari berbagai pendekatan model statistik pada data biner.

Fungsi Massa Peluang (PMF):

$P(Y = y) = \pi^y (1 - \pi)^{1-y}, \quad y \in \{0, 1\}$

Keterangan:

- Y: variabel acak biner (0 atau 1)
- π: probabilitas keberhasilan $P(Y = 1)$

2. Distribusi Binomial

Distribusi Binomial digunakan untuk menghitung jumlah keberhasilan dalam n percobaan independen, dengan probabilitas sukses tetap. Distribusi ini sangat cocok untuk model proporsi dan data agregat biner.

Fungsi Massa Peluang (PMF):

$P(Y = y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}, \quad y = 0, 1, \ldots, n$

Keterangan:

- Y: jumlah sukses
- n: jumlah percobaan
- π: probabilitas sukses per percobaan

3. Distribusi Multinomial

Distribusi Multinomial digunakan ketika suatu kejadian dapat memiliki lebih dari dua kemungkinan hasil, seperti kategori warna, preferensi, atau pilihan. Distribusi ini digunakan saat kita melakukan n percobaan dan setiap percobaan menghasilkan satu dari k kategori.

Fungsi Massa Peluang (PMF):

$P(Y_1 = y_1, \ldots, Y_k = y_k) = \frac{n!}{y_1! \ldots y_k!}\pi_1^{y_1} \ldots \pi_k^{y_k}$

Keterangan:

- Y_i: frekuensi kategori ke-i
- π_i: probabilitas kategori ke-i

4. Distribusi Hipergeometrik

Distribusi Hipergeometrik digunakan dalam konteks pengambilan sampel tanpa pengembalian dari populasi terbatas, misalnya dalam uji Fisher's Exact Test. Cocok digunakan saat ukuran populasi diketahui dan pengambilan sampel tidak bersifat independen.

Fungsi Massa Peluang (PMF):

$P(Y = y) = \frac{\binom{M}{y}\binom{N-M}{n-y}}{\binom{N}{n}}, \quad y = \max(0, n-N+M), \ldots, \min(n, M)$

Keterangan:

- N: ukuran populasi
- M: jumlah sukses dalam populasi
- n: ukuran sampel
- y: jumlah sukses dalam sampel

5. Distribusi Poisson

Distribusi Poisson digunakan untuk memodelkan jumlah kejadian dalam interval waktu atau ruang tertentu dengan rata-rata kejadian λ. Sangat ideal untuk data count (jumlah kejadian) yang jarang terjadi.

Fungsi Massa Peluang (PMF):

$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y = 0, 1, 2, ...$

Keterangan:

- λ: rata-rata kejadian
- y: jumlah kejadian dalam interval

6. Distribusi Logistik (untuk model regresi)

Distribusi logistik digunakan dalam konteks **regresi logistik biner** dan memberikan model yang efisien untuk memprediksi probabilitas dari hasil biner.

Fungsi Probabilitas:

$P(Y = 1|X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$

Keterangan:

- X: prediktor
- β_0, β_1: koefisien regresi logistik

IV. Desain Sampling dalam Analisis Data Kategori

1. Prospective Design (Desain Prospektif)

Desain prospektif bertujuan untuk mengamati hubungan antara variabel prediktor dan outcome dengan memulai dari eksposur atau karakteristik awal, kemudian mengikuti subjek ke masa depan untuk melihat hasilnya.
Desain ini digunakan ketika peneliti memiliki kontrol penuh atas pemilihan subjek dan pengamatan variabel sejak awal studi.

Keunggulan:

- Struktur temporal yang jelas (eksposur mendahului outcome)
- Mengurangi bias recall
- Memungkinkan pengendalian variabel pengganggu secara langsung

Bidang Penggunaan:

- Epidemiologi
- Kedokteran klinis
- Kesehatan masyarakat
- Ilmu perilaku eksperimental

a. Eksperimen Acak (Randomized Experiment)

Eksperimen acak adalah desain di mana subjek secara acak dialokasikan ke dalam grup perlakuan atau kontrol. Tujuannya adalah menguji efek kausal dari suatu perlakuan atau intervensi terhadap outcome.

Teknik Sampling:

- Simple Random Sampling
- Random Allocation
- Stratified Randomization

Contoh:
Uji klinis obat baru: pasien dibagi secara acak ke kelompok yang menerima obat vs. plasebo, lalu diamati perkembangan kesehatannya.

b. Studi Kohort Prospektif (Prospective Cohort Study)

Studi observasional di mana sekelompok individu diklasifikasikan berdasarkan eksposur dan diikuti dari waktu ke waktu untuk mengamati apakah mereka mengalami outcome tertentu.

Teknik Sampling:

- Stratified Sampling (berdasarkan eksposur)
- Systematic Sampling (misal, daftar pekerja)
- Cluster Sampling (populasi geografis)

Contoh:
Mengamati kelompok perokok dan bukan perokok selama 10 tahun untuk melihat kejadian kanker paru-paru.

2. Retrospective Design (Desain Retrospektif)

Desain retrospektif bertujuan untuk menganalisis hubungan antara outcome dan eksposur dengan memulai dari hasil (outcome) yang sudah terjadi, lalu menelusuri kembali faktor-faktor risikonya.

Keunggulan:

- Efisiensi waktu dan biaya
- Efektif untuk penyakit langka atau waktu laten panjang

Bidang Penggunaan:

- Epidemiologi
- Rekam medis
- Penelitian kasus langka
- Analisis data sekunder

a. Studi Kasus-Kontrol (Case-Control Study)

Dimulai dari outcome yang telah terjadi (kasus), lalu membandingkan frekuensi eksposur dengan kelompok kontrol.

Teknik Sampling:

- Purposive Sampling (kasus berdasarkan diagnosis)
- Matched Sampling (umur, jenis kelamin)
- Convenience / Registry-based Sampling

Contoh:
Memilih pasien yang mengidap kanker (kasus) dan yang tidak (kontrol), lalu menelusuri riwayat merokok.

b. Studi Kohort Retrospektif (Retrospective Cohort Study)

Menggunakan data historis dari kohort yang pernah terpapar suatu kondisi, lalu meninjau kembali apakah mereka mengalami outcome tertentu.

Teknik Sampling:

- Archival Sampling (catatan rumah sakit, perusahaan)
- Complete Enumeration
- Time-based Stratification

Contoh:
Data karyawan pabrik tahun 1990-an ditinjau untuk melihat apakah paparan asbes berkaitan dengan kematian akibat kanker tahun 2000-an.

3. Tabel Perbandingan Jenis Sampling

Jenis Desain	Eksperimental	Observasional Prospektif	Observasional Retrospektif
Contoh	Uji klinis acak	Studi kohort prospektif	Studi kasus-kontrol, kohort retrospektif
Arah waktu	Maju (intervensi → outcome)	Maju (eksposur → outcome)	Mundur (outcome → eksposur)
Kontrol peneliti	Tinggi	Sedang	Rendah
Efisiensi biaya/waktu	Rendah	Sedang	Tinggi
Bias	Kecil	Sedang	Besar (recall, seleksi)
Cocok untuk	Intervensi kausal	Faktor risiko umum	Outcome langka, data historis

```
knitr::opts_chunk$set(echo = TRUE)
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.4.2
```

```r
library(kableExtra)
```

```
## Warning: package 'kableExtra' was built under R version 4.4.3
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:kableExtra':
##
##     group_rows
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

V. Tabel Kontringensi 2 Arah

Tabel kontingensi dua arah menyajikan frekuensi pengamatan untuk dua variabel kategorik. Tujuannya adalah untuk mengevaluasi adanya hubungan atau asosiasi antara dua variabel.

Kasus

Mengambil 1000 record dari data hasil survei *European Social Survey 11*, akan dianalisis hubungan antara kelelahan kerja dengan tekanan darah tinggi. Tabel berikut menunjukkan frekuensi:

Hipertensi	Ya	Tidak	Total
Kelelahan Ya	95	405	500
Kelelahan Tidak	81	419	500
Total	176	824	1000

1. Peluang Bersama

Peluang bersama (joint probability) adalah probabilitas dua kejadian terjadi sekaligus.

$P(X = x, Y = y) = \frac{n_{xy}}{n}$

Contoh:

$P(Kelelahan = Ya, Hipertensi = Ya) = \frac{95}{1000} = 0.095$

$P(Kelelahan = Ya, Hipertensi = Tidak) = \frac{405}{1000} = 0,405$ $P(Kelelahan = Tidak, Hipertensi = Ya) = \frac{81}{1000} = 0{,}081$ $P(Kelelahan = Tidak, Hipertensi = Tidak) = \frac{419}{1000} = 0{,}419$

Kasus

Data diambil dari European Social Survey (ESS) pada survey ke-11. Diambil secara acak 1000 sampel hasil survey. Kasus ditujukan untuk mengetahui kelelahan bekerja terhadap hipertensi seseorang.

```
# Total sampel
n <- 1000
# Data: (Kelelahan, Hipertensi)
n11 <- 95 # Ya, Ya
n10 <- 405 # Ya, Tidak
n01 <- 81 # Tidak, Ya
n00 <- 419 # Tidak, Tidak

# Matriks peluang bersama
P_joint <- matrix(c(n11, n10, n01, n00) / n, nrow = 2, byrow = TRUE,
dimnames = list(c("Kelelahan=Ya", "Kelelahan=Tidak"),
c("Hipertensi=Ya", "Hipertensi=Tidak")))
P_joint
```

```
##                 Hipertensi=Ya Hipertensi=Tidak
## Kelelahan=Ya            0.095            0.405
## Kelelahan=Tidak         0.081            0.419
```

2. Peluang Marjinal

Peluang marjinal adalah peluang terjadinya satu kejadian tanpa mempertimbangkan kejadian lainnya.

$P(X = x) = \sum_y \ P(X = x, Y = y)$

Contoh:

$P(Kelelahan = Ya) = 0{,}095 + 0{,}405 = 0{,}5$

$P(Hipertensi = Ya) = 0{,}095 + 0{,}081 = 0{,}176$

```
P_X <- rowSums(P_joint)
P_Y <- colSums(P_joint)
P_X
```

```
##    Kelelahan=Ya Kelelahan=Tidak
##             0.5             0.5
```

```
P_Y
```

```
##    Hipertensi=Ya Hipertensi=Tidak
##            0.176            0.824
```

3. Peluang Bersyarat

Peluang bersyarat menyatakan probabilitas suatu kejadian terjadi dengan syarat kejadian lain diketahui.

$P(Y = y \mid X = x) = \frac{P(X=x, Y=y)}{P(X=x)}$

Contoh:

$P(Hipertensi = Ya \mid Kelelahan = Ya) = \frac{95}{500} = 0.19$

```r
P_Y_given_X <- P_joint / P_X
dimnames(P_Y_given_X) <- list(c("X=1", "X=0"), c("Y=1", "Y=0"))
P_Y_given_X
```

```
##       Y=1   Y=0
## X=1 0.190 0.810
## X=0 0.162 0.838
```

4. Ukuran Asosiasi

Untuk dua variabel dikotomi, asosiasi dapat diukur dengan:

- **Risk Difference (RD)**:
  $RD = P(Y = 1 \mid X = 1) - P(Y = 1 \mid X = 0) = 0{,}19 - 0{,}162 = 0{,}028$

```r
RD <- P_Y_given_X["X=1", "Y=1"]- P_Y_given_X["X=0", "Y=1"]
RD
```

```
## [1] 0.028
```

- **Relative Risk (RR)**:
  $RR = \frac{P(Y=1|X=1)}{P(Y=1|X=0)} = \frac{0{,}19}{0{,}162} = 1{,}1728$

```r
RR <- P_Y_given_X["X=1", "Y=1"] / P_Y_given_X["X=0", "Y=1"]
RR
```

```
## [1] 1.17284
```

- **Odds Ratio (OR)**:
  $OR = \frac{(0{,}19/0{,}81)}{(0{,}162/0{,}838)} = 1{,}2133$

```r
OR <- (P_Y_given_X["X=1", "Y=1"] / P_Y_given_X["X=1", "Y=0"]) /
(P_Y_given_X["X=0", "Y=1"] / P_Y_given_X["X=0", "Y=0"])
OR
```

```
## [1] 1.213382
```

Kesimpulan:

- RD menunjukkan bahwa terdapat selisih probabilitas hipertensi sebesar 0,028 antara pekerja yang mengalami kelelahan dan yang tidak.
- RR sebesar 1,172 menunjukkan bahwa risiko hipertensi pada pekerja yang mengalami kelelahan adalah 1,172 kali lebih besar.
- OR sebesar 1,2133 menunjukkan peluang relatif hipertensi lebih besar bagi kelompok kelelahan kerja.

1) UJI PROPORSI

Uji proporsi membandingkan dua proporsi untuk menentukan apakah perbedaan yang diamati signifikan secara statistik.

Hipotesis:

- H_0: $p_1 = p_2$
- $H_1 : p_1 \neq p_2$

Statistik Uji:

$Z = \frac{p_1 - p_2}{SE}$

Dengan:

- $p_1 = \frac{95}{500} = 0{,}19$
- $p_2 = \frac{81}{500} = 0{,}162$
- $\hat{p} = \frac{95+81}{1000} = 0{,}176$
- $SE = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \approx 0{,}02426$

$Z = \frac{0{,}19 - 0{,}162}{0{,}02426} \approx 1{,}155$

Kesimpulan:

Statistik uji Z = 1,155 < Z tabel (1,96). Maka, **tidak terdapat perbedaan signifikan** antara dua proporsi kejadian hipertensi pada pekerja kelelahan dan tidak kelelahan.

```
data<- matrix(c(95,405,81,419), nrow = 2, byrow = TRUE)
dimnames(data) <- list("Kelelahan Kerja" = c("Ya", "Tidak"), "Tekanan Darah Tinggi" = c("Ya", "Tidak"))
print(data)
```

```
##                Tekanan Darah Tinggi
## Kelelahan Kerja Ya Tidak
##           Ya    95   405
##           Tidak 81   419
```

```
# Uji Proporsi dengan variabel yang benar
prop_test <- prop.test(x = c(data[1,1], data[2,1]),
n = c(sum(data[1,]), sum(data[2,])))
print(prop_test)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(data[1, 1], data[2, 1]) out of c(sum(data[1, ]), sum(data[2, ]))
## X-squared = 1.1653, df = 1, p-value = 0.2804
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.02117418  0.07717418
## sample estimates:
## prop 1 prop 2
##  0.190  0.162
```

Hasil Uji Proporsi (Chi-square)

- Nilai Chi-square: 1.165
- df: 1
- p-value: 0.2804
- Confidence Interval (95%) untuk selisih proporsi (0.190 - 0.162):
  [-0.0212, 0.0772]

Interpretasi

1. **Perbedaan Proporsi**:
   Perbedaan antara kelompok kelelahan dan tidak kelelahan adalah 0.028 (atau 2.8%).
2. **Confidence Interval (CI)**:
   Interval antara -2.1% hingga 7.7% mengandung nol, artinya: Tidak dapat disimpulkan secara statistik bahwa proporsi hipertensi berbeda antara dua kelompok.
3. **p-value = 0.2804 > 0.05**

   Ini menunjukkan bahwa perbedaan proporsi tidak signifikan secara statistik pada taraf 5%.
   gagal menolak H : tidak ada perbedaan proporsi.

Kesimpulan

Meskipun terdapat proporsi hipertensi sedikit lebih tinggi pada kelompok kelelahan kerja, namun tidak signifikan secara statistik (baik dari p-value maupun confidence interval).
Dengan demikian, tidak ada bukti kuat bahwa kelelahan kerja berasosiasi dengan hipertensi berdasarkan uji ini.

2) UJI ASOSIASI

Digunakan untuk mengetahui apakah dua variabel kategori saling berasosiasi atau independen. Umumnya menggunakan uji Chi-Square Pearson.

Hipotesis:

- H_0: Tidak ada asosiasi antara variabel X dan Y (independen)
- H_1: Ada asosiasi (terdapat hubungan)

```
n11 <- 95; n12 <- 405; n21 <- 81; n22 <- 419
n1. <- n11 + n12; n2. <- n21 + n22
prob1 <- n11/n1.
prob2 <- n21/n2.
# Risk Difference
rd <- (n11/n1.)- (n21/n2.)
se_rd <- sqrt(((prob1 * (1- prob1)) / n1.) + (((1- prob2) * prob2) / n2.))
z_rd <- rd / se_rd
# Relative Risk
rr <- (n11/n1.) / (n21/n2.)
se_ln_rr <- sqrt((1/n11)- (1/n1.) + (1/n21)- (1/n2.))
z_rr <- log(rr) / se_ln_rr
```

```r
# Odds Ratio
or <- (n11 * n22) / (n12 * n21)
se_ln_or <- sqrt((1/n11) + (1/n12) + (1/n21) + (1/n22))
z_or <- log(or) / se_ln_or
# Hasil
list(RD = rd, SE_RD = se_rd, Z_RD = z_rd, RR = rr, SE_Ln_RR = se_ln_rr, Z_RR = z_rr, OR = or, SE_Ln_OR =
```

```
## $RD
## [1] 0.028
##
## $SE_RD
## [1] 0.0240689
##
## $Z_RD
## [1] 1.163327
##
## $RR
## [1] 1.17284
##
## $SE_Ln_RR
## [1] 0.1373754
##
## $Z_RR
## [1] 1.160526
##
## $OR
## [1] 1.213382
##
## $SE_Ln_OR
## [1] 0.1665166
```

Hasil Perhitungan Statistik:

1. Risk Difference (RD)

Nilai: RD = 0.028
SE: 0.0241
Z: 1.163

Interpretasi: Risiko hipertensi pada pekerja yang mengalami kelelahan lebih tinggi sebesar 2,8% dibandingkan yang tidak mengalami kelelahan.
Namun, nilai Z = 1.16 < 1.96, artinya tidak signifikan secara statistik ($p > 0.05$).

2. Risk Ratio (RR)

Nilai: RR = 1.1728
SE ln(RR): 0.1374
Z: 1.161

Interpretasi: Pekerja yang mengalami kelelahan memiliki risiko hipertensi 1,17 kali lebih besar dibandingkan yang tidak kelelahan.
Karena Z < 1.96, ini tidak signifikan.

3. Odds Ratio (OR)

Nilai: OR = 1.213
SE ln(OR): 0.1665

Interpretasi: Odds (peluang relatif) hipertensi pada yang kelelahan adalah 1.21 kali lebih besar dibandingkan yang tidak.
Namun, karena tidak diberikan nilai Z/CI lengkap, dan SE-nya tidak kecil, kemungkinan tidak signifikan secara statistik.

Kesimpulan

- Semua ukuran asosiasi (RD, RR, OR) menunjukkan asosiasi positif lemah antara kelelahan kerja dan tekanan darah tinggi.
- Namun, tidak ada yang signifikan secara statistik pada tingkat signifikansi 5% ($Z < 1.96$).
- Ini bisa jadi karena ukuran efek kecil atau data belum cukup kuat (power kecil).

3) UJI INDEPENDENSI

Uji independensi pada tabel kontingensi dilakukan untuk menguji apakah dua variabel kategori saling bebas.

Hipotesis:

- H_0: X dan Y independen
- H_1 : X dan Y tidak independen

Statistik Uji:

$\chi^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}}, \quad E_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$

Dimana:

- O_{ij} = frekuensi pengamatan
- E_{ij} = frekuensi harapan jika tidak ada asosiasi

Kriteria Keputusan:

Tolak H_0 jika nilai $\chi^2 \geq \chi^2_{\alpha,df}$

Contoh Kasus

Sebuah rumah sakit ingin mengetahui apakah penggunaan masker medis saat kunjungan pasien di ruang tunggu berhubungan dengan kejadian infeksi saluran pernapasan akut (ISPA) seminggu setelah kunjungan.

Dari total 160 pengunjung, mereka dikelompokkan menjadi dua berdasarkan apakah menggunakan masker saat di ruang tunggu atau tidak, dan dicatat apakah mengalami ISPA atau tidak seminggu kemudian.

Data yang Terkumpul

	ISPA: Ya	ISPA: Tidak	Total
Pakai	35	55	90
Tidak	35	45	80

Total	70	100	170

Langkah 1: Hitung Nilai Harapan

$E_{ij} = \frac{(\text{baris total}) \times (\text{kolom total})}{\text{grand total}}$

- $E_{11} = \frac{90 \times 70}{170} = 37.06$
- $E_{12} = \frac{90 \times 100}{170} = 52.94$
- $E_{21} = \frac{80 \times 70}{170} = 32.94$
- $E_{22} = \frac{80 \times 100}{170} = 47.06$

Langkah 2: Hitung Statistik Chi-Square

$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Substitusi:

$= \frac{(35-37.06)^2}{37.06} + \frac{(55-52.94)^2}{52.94} + \frac{(35-32.94)^2}{32.94} + \frac{(45-47.06)^2}{47.06}$

Hitung:

$= \frac{4.2436}{37.06} + \frac{4.2436}{52.94} + \frac{4.2436}{32.94} + \frac{4.2436}{47.06} = 0.1145 + 0.0801 + 0.1289 + 0.0902 = 0.4137$

Langkah 3: Tentukan Derajat Bebas dan P-Value

- df = (2-1)(2-1) = 1
- Nilai kritis χ^2 tabel df=1, =0.05 = 3.841

Karena 0.4137 < 3.841, maka tidak tolak H_0

```r
data_independensi <- matrix(c(35, 55, 35, 45), nrow = 2, byrow = TRUE)
dimnames(data_independensi) <- list("Masker" = c("Pakai", "Tidak"),
                        "ISPA" = c("Ya", "Tidak"))
# Uji Chi-Square
chisq_test <- chisq.test(data_independensi)
print(chisq_test)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  data_independensi
## X-squared = 0.23687, df = 1, p-value = 0.6265
```

Karena p-value < 0.05 dan $\chi^2_{hitung} < \chi^2_{tabel}$, maka H_0 diterima sehingga tidak terdapat hubungan signifikan antara penggunaan masker medis di ruang tunggu dengan kejadian ISPA seminggu setelah kunjungan

4) PARTISI CHI-SQUARE

Partisi dilakukan untuk memecah tabel menjadi bagian yang lebih kecil untuk mengidentifikasi sumber asosiasi.

Tabel observasi:

Gender	Karyawan	Karyawan di bisnis keluarga
Pria	11783	238
Wanita	11858	225
Total	**23641**	**463**

Partisi Chi-Square digunakan untuk:

- Menguraikan total statistik χ^2 dari tabel besar menjadi beberapa bagian yang lebih kecil
- Mengidentifikasi interaksi atau perbedaan utama antar kategori yang menyumbang pada asosiasi keseluruhan

Langkah-langkah pengujian chi-square pada tabel partisi:

- Tentukan sub-tabel (2x2) dari tabel kontingensi asli (disebut partisi)
- Hitung frekuensi harapan (expected frequencies) untuk setiap sel
- Hitung nilai chi-square untuk tiap partisi:
  $\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$
- Bandingkan nilai $\chi^2\chi$ dengan nilai kritis dari tabel chi-square (df = 1)

```r
# Data Observasi
data2 <- matrix(c(11783, 238, 2486, 11858, 225, 1212), nrow = 2, byrow = TRUE)
colnames(data2) <- c("Karyawan", "Karyawan di bisnis keluarga", "Wirausahawan")
rownames(data2) <- c("Male", "Female")
# Uji Chi-Square
chi_test <- chisq.test(data2)
# Hasil
list(Chi_Square = chi_test$statistic, P_Value = chi_test$p.value, Decision = ifelse(chi_test$p.value < 0
```

```
## $Chi_Square
## X-squared
##  387.4097
##
## $P_Value
## [1] 7.499786e-85
##
## $Decision
## [1] "Tolak H0"
```

INTERPRETASI:

Berdasarkan perhitungan uji hipotesis Chi-Kuadrat, didapatkan p-value sebesar 7.499786e-85.

P-value tersebut lebih kecil daripada alpha (0.05), sehingga dapat disimpulkan bahwa terdapat hubungan gender dengan tipe perusahaan tempat bekerja.

Partisi 1

Gender	Karyawan	Karyawan di bisnis keluarga
Pria	11783	238
Wanita	11858	225
Total	**23641**	**463**

$\chi^2 \approx \frac{7.417 \times 10^{14}}{4.329 \times 10^{10}} \approx 17136.1$

Partisi 2

Gender	Karyawan + Bisnis Keluarga	Wirausahawan
Pria	12021	2486
Wanita	12083	1212
Total	**24104**	**3698**

$\chi^2 \approx \frac{3.799 \times 10^{14}}{9.838 \times 10^{11}} \approx 386.3$

```r
# Data Observasi
data2_part1 <- matrix(c(11783, 238, 2486, 11858), nrow = 2, byrow = TRUE)
colnames(data2_part1) <- c("Karyawan", "Karyawan di bisnis keluarga")
rownames(data2_part1) <- c("Male", "Female")
# Uji Chi-Square Partisi 1
chi_test1 <- chisq.test(data2_part1)
# Uji Chi-Square Partisi 2
# Data Partisi 2
data2_part2 <- matrix(c(12021, 2486, 12083, 1212), nrow = 2, byrow = TRUE)
colnames(data2_part2) <- c("Karyawan+Karyawan di bisnis keluarga", "Wirausahawan")
rownames(data2_part2) <- c("Male", "Female")
# Hasil
chi_test2 <- chisq.test(data2_part2)
list(Chi_Square_Partisi1 = chi_test1, Chi_Square_Partisi2 = chi_test2)
```

```
## $Chi_Square_Partisi1
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  data2_part1
## X-squared = 17145, df = 1, p-value < 2.2e-16
##
##
## $Chi_Square_Partisi2
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  data2_part2
## X-squared = 386.27, df = 1, p-value < 2.2e-16
```

INTERPRETASI:

- Pada partisi pertama: p-value $< 0.05 \rightarrow$ beda signifikan antara karyawan dan bisnis keluarga
- Pada partisi kedua: p-value $< 0.05 \rightarrow$ beda signifikan antara kelompok karyawan (gabungan) dan wirausahawan

5) UJI LIKELIHOOD RATIO

Uji G^2 digunakan untuk:

- Menguji independensi antara dua variabel kategorik dalam tabel kontingensi (misal 2x2)
- Alternatif dari uji chi-square Pearson, berbasis likelihood

Hipotesis:

- H_0: Variabel kelelahan kerja dan hipertensi saling independen
- H_1: Variabel kelelahan kerja dan hipertensi tidak independen

Statistik uji:

- Hitung nilai ekspektasi:
  $E_{ij} = \frac{(\text{baris total}) \cdot (\text{kolom total})}{\text{total keseluruhan}}$
- Hitung G^2:
  $G^2 = 2 \sum O_{ij} \cdot \ln\left(\frac{O_{ij}}{E_{ij}}\right)$
- Bandingkan dengan:
  $\chi^2_{\text{kritis}} = \chi^2(1, 0.05) = 3.8414$

```
# Hitung Frekuensi Ekspektasi
data_expected <- chisq.test(data)$expected
# Hitung Statistik G²
G2 <- 2 * sum(data * log(data / data_expected))
critical_value <- qchisq(0.95, df = 1)
# Hasil
# Nilai kritis chi-square untuk df = 1 dan alpha = 0.05
list(G2 = G2, Critical_Value = critical_value, Decision = ifelse(G2 > critical_value, "Reject H0", "Fai
```

```
## $G2
## [1] 1.352689
##
## $Critical_Value
## [1] 3.841459
##
## $Decision
## [1] "Fail to Reject H0"
```

KESIMPULAN:
Nilai $G^2 = 1.3527 < 3.8414$ Gagal tolak H_0 tidak ada hubungan signifikan antara kelelahan kerja dan tekanan darah tinggi

6) UJI EXACT FISHER

Uji digunakan untuk tabel 2x2 kecil (n < 20 atau $E_{ij} < 5$)

Keunggulan:

- Akurat untuk sampel kecil
- Tidak bergantung pada distribusi asimtotik
- Valid saat ada nilai ekspektasi kecil
- Menghasilkan p-value eksak
- Tidak sensitif terhadap ketidakseimbangan data

Kelemahan:

- Komputasi berat jika tabel besar
- Kurang efisien untuk sampel besar
- Terbatas untuk tabel 2x2
- Sulit dilakukan secara manual (butuh enumerasi semua konfigurasi)

Hipotesis:

- H_0: Tidak ada hubungan (independen)
- H_1: Ada hubungan (tidak independen)

Rumus probabilitas Hipergeometrik:

$$P(A = a) = \frac{\binom{c_1}{a} \cdot \binom{c_2}{r_1 - a}}{\binom{n}{r_1}}$$

KASUS

Berdasarkan survei *European Social Survey 11*, diambil **30 sampel**. Variabel yang digunakan:

- **Kelelahan Kerja**
  - Ya
  - Tidak
- **Tekanan Darah Tinggi**
  - Ya
  - Tidak

Tabel kontingensi:

Tekanan Darah Tinggi	Ya	Tidak	Total
Kelelahan: Ya	1	14	15

Tekanan Darah Tinggi	Ya	Tidak	Total
Kelelahan: Tidak	2	13	15
Total	3	27	30

Langkah penghitungan:

- Probabilitas konfigurasi tabel berdasarkan distribusi hipergeometrik:

$\mathrm{P}(A = 1) = \frac{\binom{3}{1} \cdot \binom{27}{14}}{\binom{30}{15}} = 0.387931$

Semua konfigurasi lain dihitung:

Ya	Tidak	P(A = a)
0	15	0.112069
1	14	0.387931
2	13	0.387931
3	12	0.112069

```
# Definisi parameter
N <-30#Totalpopulasi
K <-3 # Jumlahkategorisukses(TekananDarahTinggi)
n <-15#Jumlahsampeldiambil
x <-1 # Jumlahsuksesdalamsampel
# HitungprobabilitasP(X=1)
dhyper(x,m=K, n=N-K,k=n)
```

```
## [1] 0.387931
```

```
choose(3, 3) * choose(27, 12) /choose(30, 15)
```

```
## [1] 0.112069
```

```
choose(3, 2) * choose(27, 13) /choose(30, 15)
```

```
## [1] 0.387931
```

```
choose(3, 1) * choose(27, 14) / choose(30, 15)
```

```
## [1] 0.387931
```

```
choose(3, 0) * choose(27, 15) / choose(30, 15)
```

```
## [1] 0.112069
```

```r
p.value<-0.112069+0.387931+0.387931+0.112069
p.value
```

```
## [1] 1
```

```r
data3 <- matrix(c(1, 14, 2, 13), nrow = 2, byrow = TRUE);data3
```

```
##      [,1] [,2]
## [1,]    1   14
## [2,]    2   13
```

```r
fisher.test(data3)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  data3
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   0.007359892 10.178680403
## sample estimates:
## odds ratio
##  0.4759965
```

Total p-value:

p-value $= 0.112069 + 0.387931 + 0.387931 + 0.112069 = 1$

Kesimpulan:

- p-value $= 1 > 0.05$, tidak cukup bukti untuk menolak H_0
- **Tidak terdapat hubungan signifikan** antara kelelahan kerja dengan tekanan darah tinggi.

7) ANALISIS RESIDUAL DAN OUTLIER PADA TABEL KONTINGENSI

Residual adalah selisih antara nilai yang diamati (observed) dan nilai harapan (expected) berdasarkan model independensi.

Jenis Residual:

- **Raw Residual**:
  $r_{ij} = O_{ij} - E_{ij}$
- **Pearson Residual**:
  $r_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$
- **Standardized Residual** (adjusted):
  $r_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}(1-\pi_{i.})(1-\pi_{.j})}}$

Tujuan Analisis Residual:

- Mengetahui kontribusi sel terhadap ketidaksesuaian model
- Mengidentifikasi sel yang menyimpang dari ekspektasi
- Mendeteksi outlier

Kriteria Outlier:

- Jika $|r| > 2$, **indikasi outlier**
- Jika $|r| > 3$, **indikasi kuat (strong) outlier**

KASUS

Berdasarkan survei *European Social Survey 11*, diambil 30 sampel dengan variabel:

- **Gender**:
  - Pria
  - Wanita
- **Vote** (apakah melakukan pemilu terakhir):
  - Ya
  - Tidak

Tabel kontingensi:

Gender	Ya	Tidak	Total
Pria	9	6	15
Wanita	8	7	15
Total	17	13	30

```
data3<-matrix(c(9,6,8,7),nrow=2,byrow=T)
colnames(data3)<-c("Ya","Tidak")
rownames(data3)<-c("Male","Female")
# Hitung nilai ekspektasi
expected <- chisq.test(data3)$expected
# Pearson Residual
pearson_residual <- (data3- expected) / sqrt(expected)
# Standardized Residual
row_sum <- rowSums(data3)
col_sum <- colSums(data3)
total_sum <- sum(data3)
standardized_residual <- (data3- expected) / sqrt(expected * (1- row_sum / total_sum) * (1- col_sum / t
# Menampilkan hasil
list(
Observed = data3,
Expected = expected,
Pearson_Residual = pearson_residual,
Standardized_Residual = standardized_residual
)
```

```
## $Observed
##        Ya Tidak
## Male    9     6
## Female  8     7
##
## $Expected
##         Ya Tidak
## Male   8.5   6.5
## Female 8.5   6.5
##
## $Pearson_Residual
##                Ya      Tidak
## Male    0.1714986 -0.1961161
## Female -0.1714986  0.1961161
##
## $Standardized_Residual
##                Ya      Tidak
## Male    0.3684381 -0.4213250
## Female -0.3221897  0.3684381
```

Ekspektasi (Expected Counts):

Gender	Ya	Tidak
Pria	8.5	6.5
Wanita	8.5	6.5

Pearson Residual:

Gender	Ya	Tidak
Pria	0.1715	-0.1961
Wanita	-0.1715	0.1961

Standardized Residual:

Gender	Ya	Tidak
Pria	0.3684	-0.4213
Wanita	-0.3222	0.3684

Interpretasi:

- Semua nilai residual < 2, tidak ada outlier
- Semua nilai mendekati ekspektasi model, tidak ada penyimpangan signifikan
  Tidak ada sel dalam tabel yang memberikan kontribusi signifikan terhadap ketidaksesuaian model independensi.

VI. Tabel Kontingensi 3 Arah

Tabel kontingensi 3 arah adalah tabel frekuensi yang menyajikan hubungan antara tiga variabel kategorik, misalnya **X**, **Y**, dan **Z**. Biasanya disajikan dalam bentuk beberapa tabel dua arah (antara X dan Y) yang masing-masing dikondisikan pada kategori dari variabel ketiga Z (*confounder/pengganggu*).

Tujuan

- Untuk mengevaluasi interaksi dan ketergantungan antara dua variabel (misal X dan Y) dengan mengontrol pengaruh variabel ketiga Z.
- Mendeteksi efek *confounding* atau interaksi (*efek moderasi*) dari Z terhadap hubungan antara X dan Y.

Kasus

Berdasarkan survei *European Social Survey 11*, diambil 1000 sampel secara acak untuk dilakukan analisis tabel kontingensi 2 arah dengan:

- **Z = Gender**
- **X = Kelelahan kerja**
- **Y = Tekanan darah tinggi (Ya/Tidak)**

Tabel Kontingensi 3 Arah (X, Y, Z):

Gender	Kelelahan Kerja	Hipertensi = Ya	Hipertensi = Tidak	Total
Pria	Ya	57	227	284
	Tidak	37	195	232
Wanita	Ya	38	178	216
	Tidak	44	224	268
Total		**176**	**824**	**1000**

1. Tabel Parsial

Tabel dua arah antara X dan Y yang dikondisikan pada setiap nilai dari Z. Ini disebut stratifikasi berdasarkan Z dan digunakan untuk memeriksa apakah hubungan antara X dan Y tetap konsisten di seluruh strata Z.

Tujuan:

- Menilai hubungan kondisional antara X dan Y di setiap lapisan Z.
- Menentukan apakah Z merupakan *confounding* (jika menyebabkan perubahan hubungan antara X dan Y) atau *efek interaksi*.

```
data4<-array(c(57,37,227,195,38,44,178,224),
 dim = c(2,2,2),
 dimnames=list(
 Kelelahan_Kerja =c("Ya","Tidak"),
 Tekanan_Darah_Tinggi = c("Ya","Tidak"),
 Gender = c("Pria","Wanita")
 ));data4
```

```
## , , Gender = Pria
##
##                 Tekanan_Darah_Tinggi
## Kelelahan_Kerja Ya Tidak
##           Ya    57   227
##           Tidak 37   195
##
## , , Gender = Wanita
##
##                 Tekanan_Darah_Tinggi
## Kelelahan_Kerja Ya Tidak
##           Ya    38   178
##           Tidak 44   224
```

```r
#Ekstrak tabel parsial berdasarkan usia
freq_parsial_pria <- data4[,,"Pria"]
freq_parsial_wanita <- data4[,,"Wanita"]
#Tampilkan Hasil
freq_parsial_pria
```

```
##                 Tekanan_Darah_Tinggi
## Kelelahan_Kerja Ya Tidak
##           Ya    57   227
##           Tidak 37   195
```

2. Tabel Marginal

Tujuan:

- Memberikan gambaran hubungan agregat antara X dan Y, tanpa mempertimbangkan potensi pengganggu Z.
- Digunakan untuk membandingkan hasil sebelum dan sesudah kontrol terhadap Z untuk melihat efek *confounding*.

Rumus:

$O_{ij}^{\text{marginal}} = \sum_k O_{ijk}$

Tabel Marginal Kelelahan Kerja (X):

Kelelahan Kerja	Hipertensi = Ya	Hipertensi = Tidak	Total
Ya	95	405	500
Tidak	81	419	500

Tabel Marginal Gender (Z):

Gender	Hipertensi = Ya	Hipertensi = Tidak	Total
Pria	94	422	516
Wanita	82	402	484

```r
#Hitung frekuensi marginal
freq_marginal_X <- apply(data4,1,sum)
freq_marginal_Z <- apply(data4,3,sum)
#Tampilkan hasil
freq_marginal_X
```

```
##    Ya Tidak
##   500   500
```

```r
freq_marginal_Z
```

```
##   Pria Wanita
##    516    484
```

3. Peluang Bersama

Peluang bersama menunjukkan probabilitas bahwa tiga kejadian terjadi secara simultan.
Dalam tabel kontingensi 3 arah (variabel X, Y, dan Z), peluang bersama menggambarkan frekuensi relatif dari kombinasi $X = i, Y = j, Z = k$.

Tujuan:

- Menyatakan hubungan probabilistik keseluruhan antar tiga variabel.
- Dasar untuk menghitung peluang bersyarat dan marginal.

Rumus:

$P(X = i, Y = j, Z = k) = \frac{O_{ijk}}{N}$

dimana:

- O_{ijk}: frekuensi observasi untuk kombinasi i, j, k
- N: total seluruh observasi

Contoh Tabel Probabilitas Bersama:

Gender	Kelelahan Kerja	Tekanan Darah Tinggi = Ya	Tidak
Pria	Ya	0.057	0.227
	Tidak	0.037	0.195
Wanita	Ya	0.038	0.178
	Tidak	0.044	0.224

```r
#Hitung probabilitas bersama
total <- sum(data4)
joint_prob <- data4/total
ftable(joint_prob)
```

```
##                                    Gender  Pria Wanita
## Kelelahan_Kerja Tekanan_Darah_Tinggi
## Ya              Ya                        0.057  0.038
##                 Tidak                     0.227  0.178
## Tidak           Ya                        0.037  0.044
##                 Tidak                     0.195  0.224
```

4. Peluang Bersyarat

Peluang bersyarat menyatakan probabilitas suatu kejadian dengan asumsi bahwa kejadian lain telah terjadi. Dalam konteks tabel 3 arah, kita sering menghitung peluang $P(X = i \mid Y = j, Z = k)$, dan sebaliknya.

Tujuan:

- Mengukur pengaruh satu variabel terhadap yang lain dalam kondisi variabel ketiga tetap (kontrol Z).
- Menunjukkan hubungan yang lebih terkontrol dibanding peluang bersama.

Rumus:

$P(X = i \mid Y = j, Z = k) = \frac{P(X=i,Y=j,Z=k)}{P(Y=j,Z=k)} = \frac{O_{ijk}}{O_{.jk}}$

$P(Y = j \mid X = i, Z = k) = \frac{O_{ijk}}{O_{i.k}}$

```
#Hitung total
total<-sum(data4)
#Hitung probabilitasgabungan(joint)
joint_prob<-data4 /total
#Hitung marjinalP(X,Y)denganmenjumlahkandidimensike-3(gender)
margin_XY <-apply(joint_prob, c(1, 2),sum)
#Hitung peluangbersyaratGender|X,Ytanpaloop
#Arrayhasilakanpunyadimensiyangsamadengandata3
P_gender_given_XY <-sweep(joint_prob, c(1, 2),margin_XY, FUN= "/")
#Tampilkanhasil
round(P_gender_given_XY, 3) #dibulatkanbiarrapi
```

```
## , , Gender = Pria
##
##                 Tekanan_Darah_Tinggi
## Kelelahan_Kerja    Ya Tidak
##           Ya    0.600 0.560
##           Tidak 0.457 0.465
##
## , , Gender = Wanita
##
##                 Tekanan_Darah_Tinggi
## Kelelahan_Kerja    Ya Tidak
##           Ya    0.400 0.440
##           Tidak 0.543 0.535
```

5. Ukuran Asosiasi

Ukuran asosiasi pada tabel kontingensi bertujuan mengukur seberapa kuat hubungan antara dua variabel kategorik. Untuk tabel 3 arah (X, Y, Z), kita bisa mengukur asosiasi antara X dan Y dalam tiap level Z.

Tujuan

Mengetahui apakah terdapat hubungan antara kelelahan kerja (X) dan tekanan darah tinggi (Y), pada masing-masing gender (Z).

Odds Ratio:

Rumus (Odds Ratio)

Untuk tabel 2x2:

$OR = \frac{a \cdot d}{b \cdot c}$

Dimana:
- a = frekuensi (X = Ya, Y = Ya)
- b = frekuensi (X = Ya, Y = Tidak)
- c = frekuensi (X = Tidak, Y = Ya)
- d = frekuensi (X = Tidak, Y = Tidak)

```
# Data berdasarkan tabel
data5 <- array(c(
57, 227, # Pria, X=Ya
37, 195, # Pria, X=Tidak
38, 178, # Wanita, X=Ya
44, 224 # Wanita, X=Tidak
), dim = c(2, 2, 2),
dimnames = list(
Y = c("Ya", "Tidak"),
X = c("Ya", "Tidak"),
Z = c("Pria", "Wanita")
))
# Hitung Risk Difference
p1 <- data5[1, 1,] / sum(data5[1,, ])
p2 <- data5[2, 1,] / sum(data5[2,, ])
RD <- p1- p2
RD
```

```
##          Pria        Wanita
##  0.0483781995 -0.0001103266
```

```
# Hitung Relative Risk
RR <- p1 / p2
RR
```

```
##      Pria    Wanita
## 1.1756107 0.9994893
```

```
# Hitung Odds Ratio
odds1 <- data5[1, 1,] / data5[1, 2,]
odds2 <- data5[2, 1,] / data5[2, 2,]
OR <- odds1 / odds2
OR
```

```
##     Pria   Wanita
## 1.323372 1.086823
```

Perhitungan Berdasarkan Tabel:

Gender	X = Ya	X = Tidak
Y = Ya	57	37
Y = Tidak	227	195

Gender	X = Ya	X = Tidak
Y = Ya	38	44
Y = Tidak	178	224

Risk Difference:

$RD = P(Y = 1|X = 1) - P(Y = 1|X = 0)$

Relative Risk:

$RR = \frac{P(Y=1|X=1)}{P(Y=1|X=0)}$

Interpretasi Risk Difference

- Untuk pria, kelelahan kerja meningkatkan risiko tekanan darah tinggi sebesar 4.8% dibanding yang tidak lelah.
- Untuk wanita, perbedaannya sangat kecil dan negatif → kelelahan kerja tidak berdampak signifikan terhadap peningkatan risiko tekanan darah tinggi (bahkan sedikit menurun, tapi sangat kecil dan tidak berarti secara praktis).

Interpretasi Relative Risk

- Pada pria, risiko tekanan darah tinggi 1.18 kali lebih besar bagi yang mengalami kelelahan kerja dibanding yang tidak. Ini menunjukkan hubungan positif ringan.
- Pada wanita, RR mendekati 1, tidak ada peningkatan risiko antara yang lelah dan tidak lelah.

Interpretasi Odds Ratio

- Untuk pria, odds tekanan darah tinggi 1.32 kali lebih tinggi pada mereka yang mengalami kelelahan kerja dibanding yang tidak → asosiasi lemah hingga sedang.
- Untuk wanita, OR = 1.09, asosiasi sangat lemah dan nyaris tidak ada.

6. Cochran-Mantel-Haenszel (CMH)

CMH adalah uji statistik untuk mengevaluasi hubungan antara dua variabel kategorik (X dan Y) setelah mengontrol pengaruh variabel ketiga (Z).
CMH memungkinkan pengujian asosiatif konsisten di seluruh strata dari Z.

Tujuan

- Menentukan apakah hubungan antara X dan Y tetap signifikan setelah menyesuaikan untuk pengaruh Z.

- Menghindari kesalahan akibat *confounding*.

Rumus

$\chi^2_{CMH} = \frac{\left(\sum_k (n_{11k} - \mu_{11k})\right)^2}{\sum_k Var(n_{11k})}$

Kasus

Berdasarkan survei *European Social Survey 11*, diambil 225 sampel secara acak untuk dilakukan analisis Cochran-Mantel-Haenszel dengan:

- **Z** = Tipe organisasi/perusahaan tempat bekerja
- **X** = Gender
- **Y** = Kelelahan kerja (Ya/Tidak)

Tabel Kontingensi CMH:

Tipe Organisasi	Gender	Ya	Tidak
Pemerintahan Pusat dan Daerah	Pria	5	11
	Wanita	6	23
Sektor publik (akademik dll)	Pria	8	8
	Wanita	10	19
Badan Usaha Milik Negara (BUMN)	Pria	5	19
	Wanita	2	19
Perusahaan Swasta	Pria	13	15
	Wanita	4	13
Wirausaha	Pria	20	15
	Wanita	3	7

```
#ArrayCMH: dimensi(Gender,Kelelahan,Organisasi)
data_cmh<-array(c(
#PemerintahanPusatdanDaerah
5, 11,6, 23,
#SektorPublik
8, 8, 10, 19,
#BUMN
5, 19,2, 19,
#Swasta
13,15, 4, 13,
#Wirausaha
20,15, 3, 7
),dim =c(2,2, 5),dimnames=list(
Gender= c("Pria", "Wanita"),
Kelelahan= c("Ya","Tidak"),
Organisasi= c(
"Pemerintah", "SektorPublik", "BUMN",
"Swasta","Wirausaha"
)
))
numerator <-0
denominator <-0
```

```r
for (k in 1:dim(data_cmh)[3]){
  tab<-data_cmh[,,k]
  n11k <-tab[1,1] #Gender=Pria, Kelelahan=Ya
  n1k. <-sum(tab[1,]) #Totalpria
  n.1k <-sum(tab[,1]) #Totalkelelahan
  n..k <-sum(tab)
  #Ekspektasisel(1,1)
  mu11k<-(n1k. *n.1k) /n..k
  #Variansisel(1,1)
  var11k<-(n1k.* (n..k-n1k.)* n.1k * (n..k-n.1k))/
    (n..k^2 *(n..k-1))
  numerator <-numerator +(n11k-mu11k)
  denominator<-denominator+ var11k
}
#StatistikCMH
CMH_stat<-(numerator^2) / denominator
p_value <-pchisq(CMH_stat,df= 1,lower.tail= FALSE)
CMH_stat
```

```
## [1] 6.853964
```

```r
p_value
```

```
## [1] 0.008844485
```

```r
#Otomatisasimenggunakanpackage
mantelhaen.test(data_cmh,correct=FALSE)
```

```
##
##  Mantel-Haenszel chi-squared test without continuity correction
##
## data:  data_cmh
## Mantel-Haenszel X-squared = 6.854, df = 1, p-value = 0.008844
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.236316 4.360335
## sample estimates:
## common odds ratio
##          2.321799
```

Hasil Statistik:

$\chi^2_{CMH} = 6.854, \quad df = 1, \quad p\text{-value} = 0.008844$

Interpretasi:

Terdapat asosiasi yang signifikan secara statistik antara kelelahan kerja dan tekanan darah tinggi setelah mengontrol variabel gender.
Dengan odds ratio sekitar 2.32, dapat disimpulkan bahwa kelelahan kerja merupakan faktor risiko penting terhadap hipertensi, terlepas dari jenis kelamin responden.

```r
org_names <-dimnames(data_cmh)[[3]]
#Hitung odds ratio(correct cell mapping)
or_values <- sapply(seq_along(org_names), function(k) {
 tab <- data_cmh[,,k]
 # Gender: baris | Kelelahan: kolom
 a <- tab["Pria", "Ya"]
 b <- tab["Pria", "Tidak"]
 c <- tab["Wanita", "Ya"]
 d <- tab["Wanita", "Tidak"]
 or <- (a * d) / (b * c)
 return(or)
})
or_table <- data.frame(
 Organisasi = org_names,
 Odds_Ratio = round(or_values, 9)
)
print(or_table, row.names = FALSE)
```

```
##    Organisasi Odds_Ratio
##    Pemerintah   1.742424
##  SektorPublik   1.900000
##          BUMN   2.500000
##        Swasta   2.816667
##     Wirausaha   3.111111
```

7. Conditional Independence

Conditional independence adalah kondisi di mana dua variabel (X dan Y) tidak saling tergantung setelah dikontrol oleh variabel ketiga (Z).

Tujuan:

Mengetahui apakah hubungan X dan Y masih signifikan setelah mengendalikan pengaruh Z (gender).

Rumus:

$P(X, Y \mid Z) \approx P(X \mid Z) \cdot P(Y \mid Z)$

Contoh (Pria):

- $P(\text{Hipertensi} \mid \text{Kelelahan, Pria}) = \frac{57}{57+227} = 0.2007$
- $P(\text{Hipertensi} \mid \text{Tidak Kelelahan, Pria}) = \frac{37}{37+195} = 0.1598$
- $P(\text{Hipertensi} \mid \text{Pria}) = \frac{57+37}{284} = \frac{94}{284} = 0.33099$

Jika $P(\text{Hipertensi} \mid \text{Kelelahan, Pria})$ tidak berbeda jauh dari $P(\text{Hipertensi} \mid \text{Pria})$, maka conditional independence berlaku.

Namun dalam kasus ini, perbedaannya masih tergolong cukup jauh, sehingga tidak terdapat conditional independence antara kelelahan kerja dan hipertensi setelah dikontrol berdasarkan gender.

8. Odds Ratio Bersama (Odds Ratio Mantel-Haenszel)

Odds Ratio (OR) mengukur kekuatan asosiasi antara dua variabel kategorik (misalnya X dan Y), dengan mengendalikan variabel ketiga Z.
Odds Ratio Bersama disebut juga Mantel-Haenszel OR, yaitu gabungan OR dari tiap strata Z.

Tujuan

- Mengukur hubungan global antara X dan Y, dengan penyesuaian terhadap Z
- Mendeteksi adanya confounding jika OR marjinal dan OR bersama berbeda jauh

Rumus Mantel-Haenszel (untuk 2x2xK):

$$OR_{MH} = \frac{\sum_k \frac{a_k d_k}{n_k}}{\sum_k \frac{b_k c_k}{n_k}}, \quad n_k = a_k + b_k + c_k + d_k$$

```
K <-dim(data_cmh)[3]
#Hitung nilai odds ratio per strata
or_vec <-numeric(K)
var_log_or_vec <-numeric(K)
for (k in 1:K){
 n11<-data_cmh[1,1,k]
 n12<-data_cmh[1,2,k]
 n21<-data_cmh[2,1,k]
 n22<-data_cmh[2,2,k]

 or<-(n11 * n22) /(n12 * n21)
 or_vec[k] <-or

 #Varians log(OR) untuk tiap strata(asumsi independen)
 var_log_or_vec[k] <-1/n11+ 1/n12+ 1/n21+ 1/n22
}

#Hitung log(ORMH)sesuai metode Robins-Breslow-Greenland
#Langkah1: Dapatkan bobot
weight_vec<-1 / var_log_or_vec

#Langkah2: Log ORMH berbobot
log_or_mh_weighted <-sum(weight_vec* log(or_vec)) / sum(weight_vec)

#Langkah3: Variansi dari log(ORMH)
var_log_or_mh<-1/ sum(weight_vec)
se_log_or_mh <-sqrt(var_log_or_mh)

#Interval Kepercayaan95%
z_alpha <-qnorm(0.975)
lower_log <-log_or_mh_weighted-z_alpha * se_log_or_mh
upper_log <-log_or_mh_weighted + z_alpha * se_log_or_mh

#Kembali ke skala OR
or_mh <- exp(log_or_mh_weighted)
ci_lower <- exp(lower_log)
ci_upper <- exp(upper_log)
# Output
cat("Manual (Robins-Breslow-Greenland):\n")
```

```
## Manual (Robins-Breslow-Greenland):
```

```
cat("Odds Ratio MH =", round(or_mh, 6), "\n")
```

```
## Odds Ratio MH = 2.303656
```

```
cat("Log Odds Ratio MH =", round(log_or_mh_weighted, 6), "\n")
```

```
## Log Odds Ratio MH = 0.834497
```

```
cat("Standard Error log(OR) =", round(se_log_or_mh, 6), "\n")
```

```
## Standard Error log(OR) = 0.323875
```

```
cat("95% CI untuk OR MH: [", round(ci_lower, 6), ",", round(ci_upper, 6), "]\n")
```

```
## 95% CI untuk OR MH: [ 1.221055 , 4.346103 ]
```

Interpretasi

Berdasarkan hasil perhitungan manual dengan pendekatan Robins-Breslow-Greenland, diperoleh nilai Odds Ratio Mantel-Haenszel sebesar 2.30.

- Variabel yang dianalisis:
  - X = Gender
  - Y = Kelelahan Kerja
  - Z = Tipe organisasi/perusahaan tempat bekerja (pengontrol)

Artinya: Setelah dikontrol berdasarkan tipe tempat bekerja, pekerja pria memiliki kemungkinan sekitar 2.3 kali lebih besar mengalami kelelahan kerja dibandingkan pekerja wanita.

- Nilai log odds ratio = 0.834 dan standard error = 0.324 menunjukkan estimasi yang cukup presisi.
- Interval kepercayaan 95% [1.22, 4.35] tidak mencakup nilai 1, maka asosiasi ini signifikan secara statistik.

9. Uji Breslow-Day untuk Homogenitas Odds Ratio

Uji Breslow-Day digunakan untuk menguji apakah odds ratio antar semua strata Z adalah homogen (sama). Uji ini dilakukan sebelum menggunakan OR Mantel-Haenszel, karena OR_MH mengasumsikan homogenitas OR.

Tujuan

- Untuk mengetahui apakah OR dalam tiap strata dapat digabungkan secara valid.
- Mendeteksi interaksi antara X dan Z atau Y dan Z.

Statistik uji Breslow-Day:

$Q = \sum_{k=1}^{K} \frac{(a_k - Ea_k)^2}{\text{Var}_{a_k|OR_{MH}}}$

Dimana:

- Ea_k: ekspektasi dari aka_kak berdasarkan OR_{MH}
- Var_{a_k}: variansi dari aka_kak di strata ke-k

Hipotesis

- H_0: OR sama untuk semua strata Z (homogen)
- H_1: OR tidak sama untuk semua strata Z (tidak homogen)

Statistik Uji dan Kriteria

- Statistik Q mengikuti distribusi chi-square (χ^2) dengan derajat bebas $df = K - 1$
- Tolak H_0 jika $Q > \chi^2_{\alpha,K-1}$ atau jika p-value $< \alpha$

```
#Fungsi menghitungORgabunganMH
OR_MH_manual <-function(data_cmh){
K <-dim(data_cmh)[3]
num<-0
den<-0
for(j in 1:K){
 tab<-data_cmh[,,j]
 a<-tab[1,1];b <-tab[1,2];c <-tab[2,1];d <-tab[2,2]
 n<-a + b + c+ d
 num<-num + (a*d) /n
 den<-den + (b*c) /n
}
 return(num / den)
}
breslow_day <-function(x){
 K <-dim(x)[3]
 or_hat_mh <-as.numeric(mantelhaen.test(x)$estimate)
 X2_HBD<-0
 a <-tildea<-var_a<-numeric(K)

for(j in 1:K){
 tab<-x[,,j]
 mj<-apply(tab,1,sum)
 nj<-apply(tab,2,sum)
 #Koefisienkuadrat:polyrootpakaiurutanC,B,A
 coef<-c(-mj[1]*nj[1]* or_hat_mh,nj[2]-mj[1] + or_hat_mh* (nj[1] + mj[1]), 1-or_hat_mh)
 roots<-Re(polyroot(coef))
 valid_root<-roots[roots >0 & roots<=min(nj[1],mj[1])]
 if(length(valid_root) == 0) stop(paste("Novalidrootatstrata",j))
 tildeaj <-valid_root[1]
 aj<-tab[1,1]
 tildebj <-mj[1]-tildeaj
```

```
 tildecj <-nj[1]-tildeaj
 tildedj <-mj[2]-tildecj
 var_aj<-1 / (1/tildeaj + 1/tildebj + 1/tildecj + 1/tildedj)
 X2_HBD<-X2_HBD + (aj-tildeaj)^2/ var_aj
 a[j]<-aj
 tildea[j]<-tildeaj
 var_a[j]<-var_aj
}

#Taronecorrection
X2_HBDT <-X2_HBD-(sum(a-tildea))^2 / sum(var_a)
p_value <-pchisq(X2_HBDT, df= K-1,lower.tail= FALSE)
result<-list(
X2_HBD= X2_HBD,
X2_HBDT= X2_HBDT,
p= p_value
)

class(result)<-"bdtest"
return(result)
}
print.bdtest <-function(x){
cat("BreslowandDaytest(withTaronecorrection):\n")
cat("Breslow-DayX-squared=",x$X2_HBD, "\n")
cat("Breslow-Day-TaroneX-squared=",x$X2_HBDT, "\n\n")
cat("TestfortestofacommonOR:p-value=",x$p,"\n\n")
}
#Hitung nilaistatistikujiBreslow-Daymanual
breslow_day(data_cmh)
```

```
## BreslowandDaytest(withTaronecorrection):
## Breslow-DayX-squared= 0.4961228
## Breslow-Day-TaroneX-squared= 0.4959237
##
## TestfortestofacommonOR:p-value= 0.9738966
```

```
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 4.4.3
```

```
#UjiBreslow-Day
breslow_test <-BreslowDayTest(data_cmh)
print(breslow_test)
```

```
##
##  Breslow-Day test on Homogeneity of Odds Ratios
##
## data:  data_cmh
## X-squared = 0.49612, df = 4, p-value = 0.9739
```

Interpretasi

Berdasarkan hasil uji Breslow-Day untuk homogenitas odds ratio:

- Nilai $X^2 = 0.49612$
- Derajat bebas = 4
- p-value = 0.9739

Kesimpulan:

Karena p-value sangat besar ($0.9739 > 0.05$), maka tidak ada cukup bukti untuk menolak H_0.

Artinya:

- Odds ratio antara gender dan kelelahan kerja dianggap homogen di seluruh tipe organisasi/perusahaan.
- Hubungan antara gender dan kelelahan kerja cenderung konsisten, baik di instansi pemerintah, sektor publik akademik, BUMN, perusahaan swasta, maupun wirausaha.
- Hasil ini juga mendukung validitas penggunaan Odds Ratio Mantel-Haenszel sebagai estimasi gabungan, karena tidak ditemukan perbedaan signifikan antar strata tipe perusahaan.

Studi Kasus 1

Penelitian ini menggunakan data dari European Social Survey (ESS) ke-11. Dari 1.000 responden yang diambil secara acak, diteliti apakah ada hubungan antara kelelahan akibat pekerjaan dan kejadian hipertensi.

Responden dibagi dua:

- Mengalami kelelahan kerja
- Tidak mengalami kelelahan kerja

Kemudian dicatat apakah mereka pernah didiagnosis hipertensi.

Tujuan penelitian:

Mengetahui apakah kelelahan kerja berhubungan dengan risiko terkena hipertensi (Uji Independensi chi-square).

	Hipertensi: Ya	**Hipertensi: Tidak**	**Total**
Kelelahan: Ya	95	405	500
Kelelahan: Tidak	81	419	500
Total	176	824	1000

Nilai Harapan (Expected Value)

$E_{ij} = \frac{(baris \times kolom)}{grand\ total}$

Kelelahan: Ya, Hipertensi: Ya

$E_{11} = \frac{500 \times 176}{1000} = 88$

Kelelahan: Ya, Hipertensi: Tidak

$E_{12} = \frac{500 \times 824}{1000} = 412$

Kelelahan: Tidak, Hipertensi: Ya

$E_{21} = \frac{500 \times 176}{1000} = 88$

Kelelahan: Tidak, Hipertensi: Tidak

$E_{22} = \frac{500 \times 824}{1000} = 412$

Hitung Chi-Square

$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(95-88)^2}{88} + \frac{(405-412)^2}{412} + \frac{(81-88)^2}{88} + \frac{(419-412)^2}{412}$

Hitung selisih kuadrat:

- $(95 - 88)^2 = 49$
- $(405 - 412)^2 = 49$
- $(81 - 88)^2 = 49$
- $(419 - 412)^2 = 49$

Lalu bagi dengan E:

- $\frac{49}{88} = 0.5568$
- $\frac{49}{412} = 0.1189$
- $\frac{49}{88} = 0.5568$
- $\frac{49}{412} = 0.1189$

Jumlahkan:

$\chi^2 = 0.5568 + 0.1189 + 0.5568 + 0.1189 = 1.3514$

Derajat Kebebasan

$df = (2-1)(2-1) = 1$

Chi-Square Tabel

$\chi^2_{tabel} = 3.841$

```
studi_kasus1 <- matrix(c(95, 405, 81, 419), nrow = 2, byrow = TRUE)
dimnames(studi_kasus1) <- list("Kelelahan" = c("Ya", "Tidak"),
                        "Hipertensi" = c("Ya", "Tidak"))

# Uji Chi-Square
chisq_test2 <- chisq.test(studi_kasus1)
print(chisq_test)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  data_independensi
## X-squared = 0.23687, df = 1, p-value = 0.6265
```

Kesimpulan

Berdasarkan uji independensi, didapatkan bahwa p-value > 0.05 dan $\chi^2_{hitung} < \chi^2_{tabel}$ sehingga H_0 diterima. Hal ini mengakibatkan kesimpulan bahwa secara statistik kelelahan bekerja tidak berhubungan signifikan dengan hipertensi seseorang.

VII. Generalized Linear Model (GLM)

Generalized Linear Model (GLM) adalah perluasan dari model regresi linear klasik yang memungkinkan penggunaan distribusi error yang tidak harus normal, serta menggunakan fungsi link untuk menghubungkan rata-rata dari distribusi respons dengan kombinasi linear dari prediktor. Tujuan utama GLM adalah untuk memodelkan hubungan antara satu atau lebih variabel prediktor dengan variabel respons yang dapat berbentuk biner, count, atau kontinu tapi tidak normal.

Komponen utama dari GLM

1. **Komponen sistematik (linear predictor)**:
   $\eta = X\beta$
2. **Komponen probabilistik (distribusi data)**: Berasal dari keluarga eksponensial (misalnya binomial, Poisson, dll)
3. **Fungsi link**: Fungsi yang menghubungkan ekspektasi $E(Y)$ dengan η, misalnya:
   - Fungsi logit untuk regresi logistik
   - Fungsi log untuk regresi Poisson

Keunggulan GLM

- Mampu menangani berbagai jenis data (biner, count, proporsi)
- Lebih fleksibel dibanding regresi linear biasa
- Bisa memasukkan lebih dari satu prediktor (multivariat)
- Interpretasi model tetap berbasis pada probabilitas atau ekspektasi

Kelemahan GLM

- Asumsi distribusi dari keluarga eksponensial harus sesuai
- Interpretasi koefisien tidak selalu intuitif (karena dalam skala link)
- Sensitif terhadap pencilan (outlier) dan multikolinearitas

1. Exponential Family

Distribusi yang digunakan dalam GLM berasal dari keluarga eksponensial, yaitu kelas distribusi yang bentuk umum densitas atau mass function-nya dapat dituliskan sebagai:

$f_Y(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$

Distribusi seperti normal, binomial, Poisson, gamma termasuk dalam keluarga ini.
Pemilihan distribusi mempengaruhi fungsi link yang digunakan.

2. Model Regresi Logistik

Model regresi logistik digunakan saat variabel dependen bersifat biner (misal: memiliki anak atau tidak). Link function yang digunakan adalah logit, dan distribusi error-nya adalah binomial.

Rumus

$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$

di mana:

- p adalah probabilitas sukses (misalnya, memiliki anak)

Kasus

Berdasarkan survei European Social Survey 11, diambil 100 sampel acak untuk dilakukan analisis model regresi logistik.
Variabel:

- X_1 = Gender (faktor)
- X_2 = Umur
- Y = Memiliki Anak (Ya/Tidak)

```
data7 <- data.frame(
  Gender = c(1,2,1,2,2,1,2,2,2,1,2,2,2,1,1,1,1,1,2,2,2,2,1,1,1,2,2,1,2,2,
             2,2,1,2,2,1,1,1,1,2,2,1,2,1,1,1,2,2,1,1,2,2,1,1,2,1,2,2,2,2,
             2,2,1,2,2,1,1,1,1,1,1,1,2,2,2,2,1,2,2,2,2,1,2,1,2,1,1,1,1,
             2,2,1,2,1,1,1,1,1,1,1),

  Umur = c(53,78,25,55,80,78,57,29,52,33,59,27,29,33,35,45,28,59,81,65,
           65,21,25,30,32,58,19,58,63,44,69,72,67,70,66,24,51,70,36,61,
           59,45,33,44,52,68,69,52,71,57,37,19,57,31,69,67,69,56,18,24,
           54,53,68,23,26,73,28,64,26,88,69,50,66,40,88,34,78,72,27,27,
           32,30,58,72,70,58,59,57,63,62,73,64,22,82,24,28,72,74,28,69),

  Memiliki_Anak = c(1,1,0,1,1,1,1,0,1,0,1,0,0,0,0,1,0,0,1,1,0,0,0,0,0,1,
                    0,0,1,1,1,0,0,1,1,0,0,1,0,1,1,1,0,1,1,1,1,1,1,1,0,0,1,
                    0,1,1,1,1,0,0,0,1,0,0,0,0,0,1,0,0,1,0,1,0,0,0,1,1,0,0,
                    0,0,1,1,1,1,1,0,1,0,1,1,0,1,1,0,1,1,0,1)
)



# Buat model regresi logistik
model <- glm(Memiliki_Anak ~ Umur, family = binomial(link = "logit"), data = data7)
summary(model)
```

```
##
## Call:
## glm(formula = Memiliki_Anak ~ Umur, family = binomial(link = "logit"),
##     data = data7)
##
## Coefficients:
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.41216    0.89893  -4.908 9.19e-07 ***
## Umur         0.08689    0.01634   5.318 1.05e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 138.469  on 99  degrees of freedom
## Residual deviance:  93.995  on 98  degrees of freedom
## AIC: 97.995
##
## Number of Fisher Scoring iterations: 4
```

Interpretasi

Model regresi logistik dibangun dengan rumus:

$\log\left(\frac{p}{1-p}\right) = -4.412 + 0.087 \cdot \text{Umur}$

- Intercept = -4.412:
  Ketika Umur = 0, log odds memiliki anak adalah -4.412 → sangat kecil (tidak realistis secara praktis, tetapi penting secara statistik).
- Koefisien Umur = 0.087:
  Untuk setiap kenaikan 1 tahun umur, log odds memiliki anak meningkat sebesar 0.087. Dalam bentuk odds:

  $e^{0.08689} \approx 1.090777$

  Artinya, setiap kenaikan 1 tahun usia meningkatkan kemungkinan memiliki anak sebesar ~9.07%, jika variabel lain dianggap konstan.

Uji Signifikansi

- p-value untuk umur = 1.05e-07
  Sangat signifikan (***)nunjukkan bahwa umur adalah prediktor yang signifikan secara statistik terhadap kemungkinan memiliki anak.

Evaluasi Model:

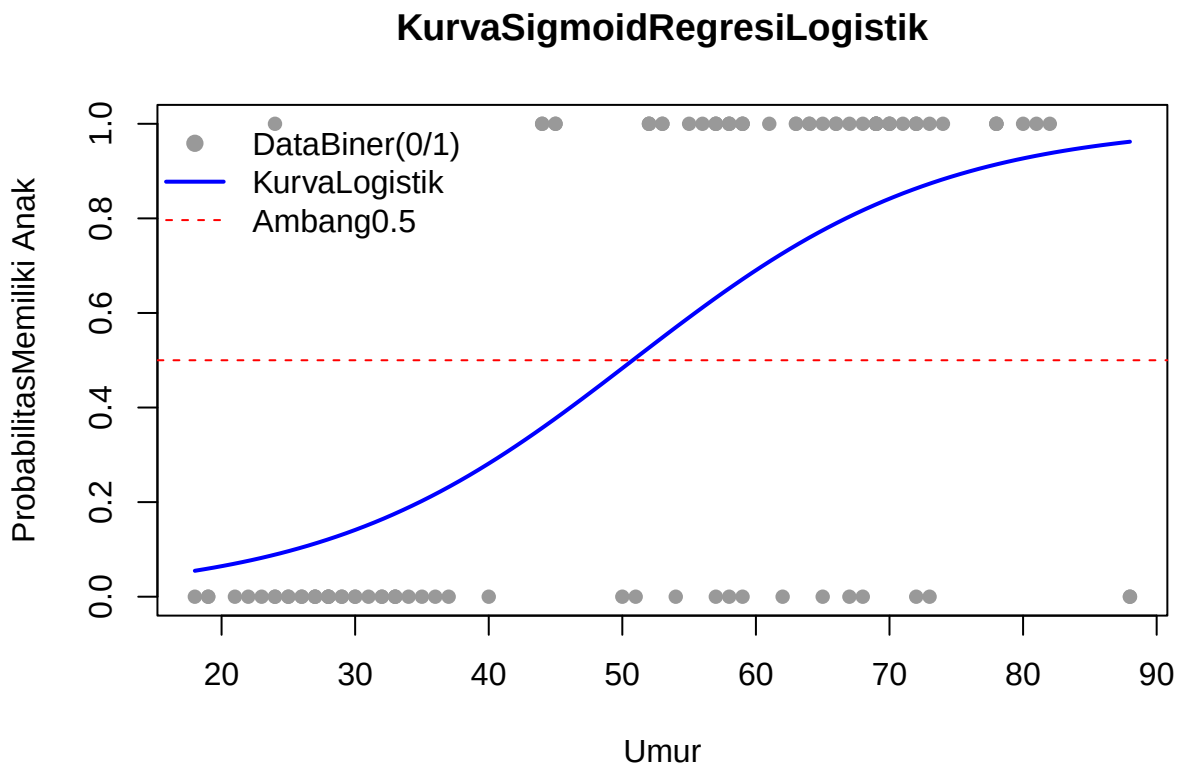
- Null deviance = 138.47
- Residual deviance = 93.995
- AIC = 97.995

Uji Likelihood Ratio:

Deviance = 44.475

Model dengan prediktor Umur secara signifikan lebih baik dibandingkan model kosong

```r
#Buatgrid umur
umur_grid <-seq(min(data7$Umur), max(data7$Umur),length.out= 100)
#Prediksi probabilitasberdasarkangrid
data7$prediksi <-predict(model, newdata= data.frame(Umur= umur_grid),type= "response")
#   INILAH plotutamayangWAJIBadasebelumlines()
plot(data7$Umur,data7$Memiliki_Anak,
pch= 16, col= "gray60",
xlab= "Umur", ylab= "ProbabilitasMemiliki Anak",
main= "KurvaSigmoidRegresiLogistik")
#Tambahkankurvasigmoid
lines(umur_grid,data7$prediksi, col= "blue", lwd=2)
#Garisambangbatas0.5
abline(h= 0.5, col="red", lty= 2)
#Legenda
legend("topleft",
legend= c("DataBiner(0/1)", "KurvaLogistik", "Ambang0.5"),
col= c("gray60", "blue", "red"),
pch= c(16,NA, NA),
lty= c(NA,1,2),
lwd= c(NA,2,1),
pt.cex= 1.2,
bty= "n")
```



Interpretasi

- Pada usia 20–30 tahun, probabilitas memiliki anak sangat rendah (mendekati nol)

- Usia 35–45 tahun: Probabilitas meningkat tajam
- Usia 50 tahun: Probabilitas mencapai ambang 0.5
- Usia 60 tahun ke atas: Probabilitas mendekati 1 (mayoritas sudah memiliki anak)

```
# Klasifikasi dan confusion matrix
data7$pred_class <- ifelse(data7$prediksi > 0.5, 1, 0)
table(Predicted = data7$pred_class, Actual = data7$Memiliki_Anak)
```

```
##           Actual
## Predicted  0  1
##         0 22 25
##         1 26 27
```

Akurasi Model:

Akurasi $= \frac{TP+TN}{n} = \frac{27+22}{100} = 0.49$

Model memprediksi dengan benar hanya 49% dari seluruh observasi.

3. Model Regresi Poisson

Model regresi Poisson digunakan untuk memodelkan data hitung (count) atau jumlah kejadian dalam suatu interval waktu atau ruang. Distribusi Poisson cocok digunakan ketika variabel respons berupa jumlah kejadian (misalnya jumlah kecelakaan, kelahiran, kunjungan dokter, dll).

Rumus Model

$\log(\lambda_i) = \beta_0 + \beta_1 X_i$

Dengan:

- $\lambda_i = E(Y_i)$: ekspektasi jumlah kejadian pada unit ke-i
- β_0: intercept
- β_1: koefisien variabel prediktor X

Distribusi probabilitas Poisson:

$P(Y_i = y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$

Tujuan Model Regresi Poisson

- Memahami hubungan antara jumlah kejadian dan prediktor (usia, jenis kelamin, dll)
- Menaksir pengaruh prediktor terhadap jumlah kejadian
- Melakukan prediksi nilai respons

Kasus

Terdapat data jumlah kunjungan dokter berdasarkan umur pasien.

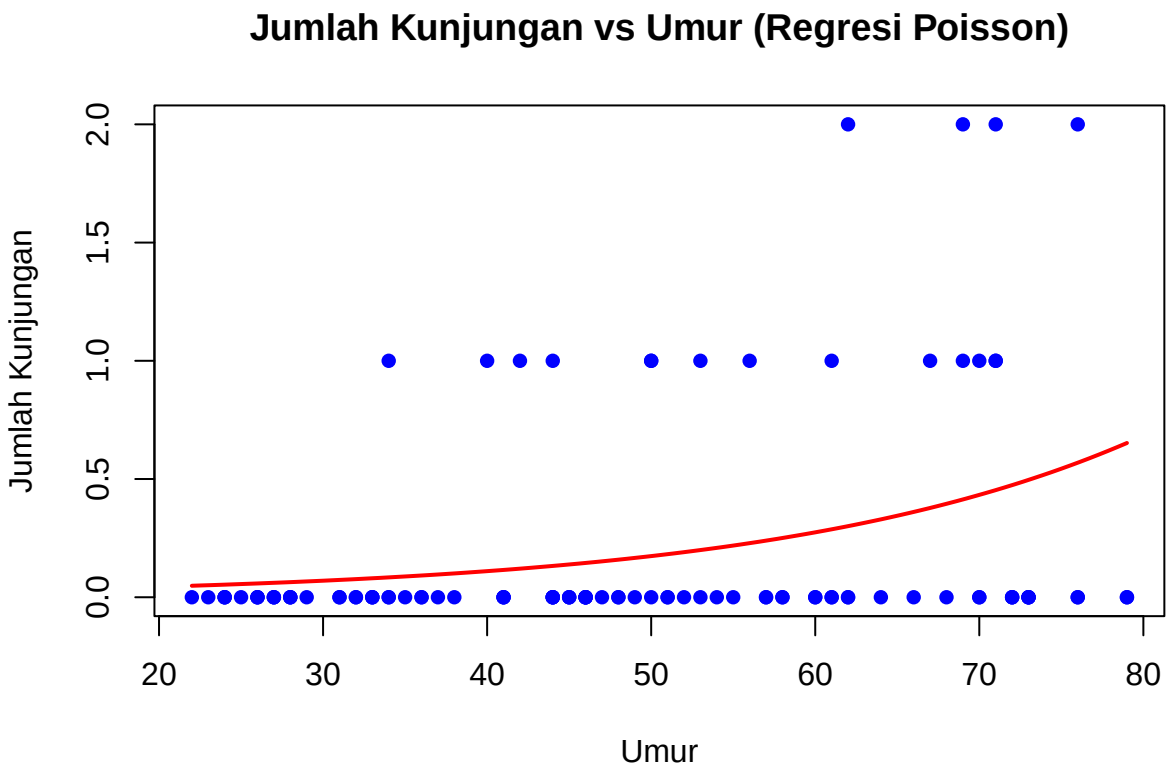
```r
# Membuat data dummy
set.seed(123)
n <- 100 # jumlah observasi
umur <- sample(20:80, n, replace = TRUE) # umur antara 20 dan 80 tahun
jumlah_kunjungan <- rpois(n, lambda = exp(0.03 * umur- 3)) # jumlah kunjungan sebagai fungsi log dari u
# Data
data_poisson <- data.frame(umur, jumlah_kunjungan)
# Model regresi Poisson
model_pois <- glm(jumlah_kunjungan ~ umur, family = poisson(link = "log"), data = data_poisson)
# Ringkasan model Poisson
summary(model_pois)
```

```
##
## Call:
## glm(formula = jumlah_kunjungan ~ umur, family = poisson(link = "log"),
##     data = data_poisson)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.02683    0.91582  -4.397  1.1e-05 ***
## umur         0.04556    0.01469   3.102  0.00192 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 77.712  on 99  degrees of freedom
## Residual deviance: 66.704  on 98  degrees of freedom
## AIC: 109.16
##
## Number of Fisher Scoring iterations: 6
```

Interpretasi

- Umur berpengaruh signifikan terhadap jumlah kunjungan dokter
- Setiap kenaikan usia 1 tahun $\rightarrow$ peningkatan 4.56% dalam ekspektasi kunjungan (karena $\exp(0.04556) \approx 1.0456$)
- Model fit cukup baik dengan penurunan deviance dan nilai AIC yang lebih kecil dibanding model kosong

```r
# Prediksi dengan umur baru untuk kurva
umur_grid <- seq(min(data_poisson$umur), max(data_poisson$umur), length.out = 100)
prediksi <- predict(model_pois, newdata = data.frame(umur = umur_grid), type = "response")
# Plot data asli dan kurva model Poisson
plot(data_poisson$umur, data_poisson$jumlah_kunjungan,
xlab = "Umur", ylab = "Jumlah Kunjungan",
main = "Jumlah Kunjungan vs Umur (Regresi Poisson)",
pch = 16, col = "blue")
lines(umur_grid, prediksi, col = "red", lwd = 2)
```



Interpretasi

Grafik menunjukkan jumlah kunjungan terhadap umur (titik-titik biru), dengan kurva Poisson (merah) yang menunjukkan prediksi model. Beberapa hal yang bisa diperhatikan:

- Pada umur muda, jumlah kunjungan relatif rendah dan tidak begitu bervariasi.
- Pada usia yang lebih tua (40 tahun ke atas), jumlah kunjungan mulai meningkat, dan variasi jumlah kunjungan juga meningkat. Hal ini menunjukkan bahwa jumlah kunjungan meningkat lebih cepat seiring bertambahnya umur.

```r
set.seed(87)
n <- 200
x <- rnorm(n)
lambda <- exp(0.3 + 0.6 * x)
y <- rpois(n, lambda)
data8 <- data.frame(y, x);data8
```

```
##     y           x
## 1   0 -2.142691355
## 2   0 -1.722811714
## 3   0 -1.871820865
## 4   0 -0.915033205
## 5   7  1.835613179
## 6   0 -0.104414368
## 7   1 -0.909716938
```

```
## 8   1 -0.675888905
## 9   1  1.009122608
## 10  1 -0.549772016
## 11  0 -0.665159169
## 12  0  0.001127092
## 13  3 -0.198474313
## 14  2  0.107947744
## 15  2  1.571038585
## 16  2  0.535314884
## 17  1  0.073722761
## 18  0 -0.693768667
## 19  4  1.385474614
## 20  0 -0.481392076
## 21  2  0.317030886
## 22  0 -0.813745438
## 23  2 -0.719092368
## 24  0 -0.850325647
## 25  0  0.317274695
## 26  1 -0.549250079
## 27  4  1.846108402
## 28  0 -2.132163385
## 29  0  0.547838042
## 30  1 -0.568938452
## 31  1  0.043982214
## 32  2  1.414463861
## 33  0 -0.803564692
## 34  0 -0.999601848
## 35  1 -0.443770450
## 36  1 -0.249763475
## 37  2  0.626142673
## 38  0 -0.633021693
## 39  1 -0.955683892
## 40  5  0.709884272
## 41  0  0.840026822
## 42  1 -1.132563957
## 43  0 -0.478722074
## 44  0 -0.008028463
## 45  4  1.468111512
## 46  3  0.182140544
## 47  2  1.382883836
## 48  0 -2.412223198
## 49  2 -1.246828273
## 50  2 -0.808755097
## 51  2 -0.553885910
## 52  4  1.491011274
## 53  0  1.414699476
## 54  5  0.633713130
## 55  2  0.399494136
## 56  5  0.370988027
## 57  0  0.985338005
## 58  1  0.385793070
## 59  0 -0.757083278
## 60  0 -0.749293066
## 61  1 -1.104788014
```

```
## 62  0 -1.819473818
## 63  2 -0.246109530
## 64  2  0.213879914
## 65  0  1.068431843
## 66  1  0.554276358
## 67  1 -0.671833225
## 68  0 -0.844349503
## 69  2  0.263086343
## 70  1 -1.837741048
## 71  4  1.288236361
## 72  0 -1.680539457
## 73  1 -0.416320660
## 74  3  0.300317631
## 75  1  0.340944361
## 76  1  0.274357283
## 77  6  0.866024844
## 78  1 -1.223824997
## 79  0 -0.187732944
## 80  1  0.330200091
## 81  1  0.205645062
## 82  3  1.179701270
## 83  2 -0.163456923
## 84  0 -1.760054132
## 85  3  0.626063449
## 86  1 -1.419444528
## 87  4  0.795984610
## 88  2  0.843358715
## 89  2  1.224892025
## 90  4  0.488997279
## 91  2  0.571813999
## 92  2  0.185807514
## 93  0 -1.579820459
## 94  5  0.094116311
## 95  2  0.016076585
## 96  1 -0.861586680
## 97  3  0.227468197
## 98  1  0.148452739
## 99  3  1.568513757
## 100 2  1.038339921
## 101 1 -0.646386920
## 102 0 -1.172680982
## 103 1 -1.910614130
## 104 2  0.739350953
## 105 1  0.321802420
## 106 0 -1.545607111
## 107 2 -0.282476328
## 108 2  0.729739871
## 109 1 -0.412584674
## 110 1 -0.032168128
## 111 2 -1.154244889
## 112 0 -0.856995716
## 113 6  0.766985027
## 114 0 -0.078728670
## 115 1  0.861131404
```

```
## 116 0 -0.475344169
## 117 3  0.416467294
## 118 3  0.759838796
## 119 1 -0.133580705
## 120 0 -0.774795100
## 121 1  0.317737905
## 122 1  0.589830740
## 123 0 -0.312523814
## 124 2 -0.542111407
## 125 3  1.836848960
## 126 2 -0.228159108
## 127 5  1.381456686
## 128 1 -1.938380643
## 129 2 -0.054564917
## 130 4  1.510475175
## 131 3  0.477060395
## 132 3  1.537913482
## 133 1 -1.215329462
## 134 2  0.769625931
## 135 1 -0.376710706
## 136 2 -0.443919847
## 137 3 -0.943730898
## 138 2  0.579798996
## 139 1 -0.386790525
## 140 0 -0.417426547
## 141 1  1.080449578
## 142 3  0.237126979
## 143 0  1.250079006
## 144 1 -2.186009588
## 145 1 -0.576810195
## 146 4  0.138325140
## 147 0 -0.524387303
## 148 5  1.420716635
## 149 0  0.227827664
## 150 4  1.750439750
## 151 6  2.512304856
## 152 2  1.455004525
## 153 1 -1.343665159
## 154 0 -0.441168461
## 155 0 -0.236810141
## 156 2  0.904903783
## 157 1  0.231107098
## 158 1  1.589089586
## 159 0 -0.109647026
## 160 3  1.039541274
## 161 0 -2.577509009
## 162 3  0.382524686
## 163 0 -0.237945306
## 164 3  1.444222394
## 165 0 -0.923774212
## 166 0  0.490384553
## 167 0  0.034452147
## 168 1 -0.137302974
## 169 1 -1.667665527
```

```
## 170 0 -1.606760971
## 171 4 -0.028489847
## 172 0 -0.121322692
## 173 3  0.352214046
## 174 4  1.283316002
## 175 0 -0.639741463
## 176 1 -0.678109849
## 177 3 -0.315799131
## 178 0 -0.609786705
## 179 4  1.190250952
## 180 1 -1.739174402
## 181 4 -0.495863940
## 182 2  0.541971690
## 183 0 -0.659389042
## 184 2  0.854471981
## 185 3  0.369913683
## 186 2  0.053295087
## 187 2 -0.852540199
## 188 1  0.106266998
## 189 1  0.734034987
## 190 2 -0.435943168
## 191 0 -1.067064170
## 192 2  1.381813549
## 193 0 -0.206077078
## 194 2 -0.612407887
## 195 3  1.104730246
## 196 1  0.035830747
## 197 1  0.780090084
## 198 2  1.155116922
## 199 0 -2.616790189
## 200 7  2.329403334
```

```
poisson_model<-glm(y~ x,data= data8,family= poisson)
summary(poisson_model)
```

```
##
## Call:
## glm(formula = y ~ x, family = poisson, data = data8)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.30127    0.06502   4.634 3.59e-06 ***
## x            0.60616    0.05965  10.161  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 330.76  on 199  degrees of freedom
## Residual deviance: 221.43  on 198  degrees of freedom
## AIC: 589.23
##
## Number of Fisher Scoring iterations: 5
```

Interpretasi

Nilai estimasi untuk variabel x adalah 0.60616. Karena model menggunakan fungsi link log, maka setiap peningkatan satu satuan pada variabel x akan meningkatkan nilai ekspektasi (mean) dari y sebesar:

$exp(0.60616) \approx 1.833$

Artinya, setiap kenaikan 1 satuan pada x diharapkan meningkatkan jumlah kejadian y sebesar 83,3%

Uji Kesesuaian Model

- Null deviance: 330.76 (model tanpa prediktor)
- Residual deviance: 221.43 (model dengan x)
- Penurunan deviance menunjukkan bahwa penambahan variabel x secara signifikan meningkatkan kecocokan model.
- AIC (Akaike Information Criterion): 589.23 Semakin rendah nilai AIC, semakin baik model dalam hal keseimbangan antara kecocokan dan kompleksitas.

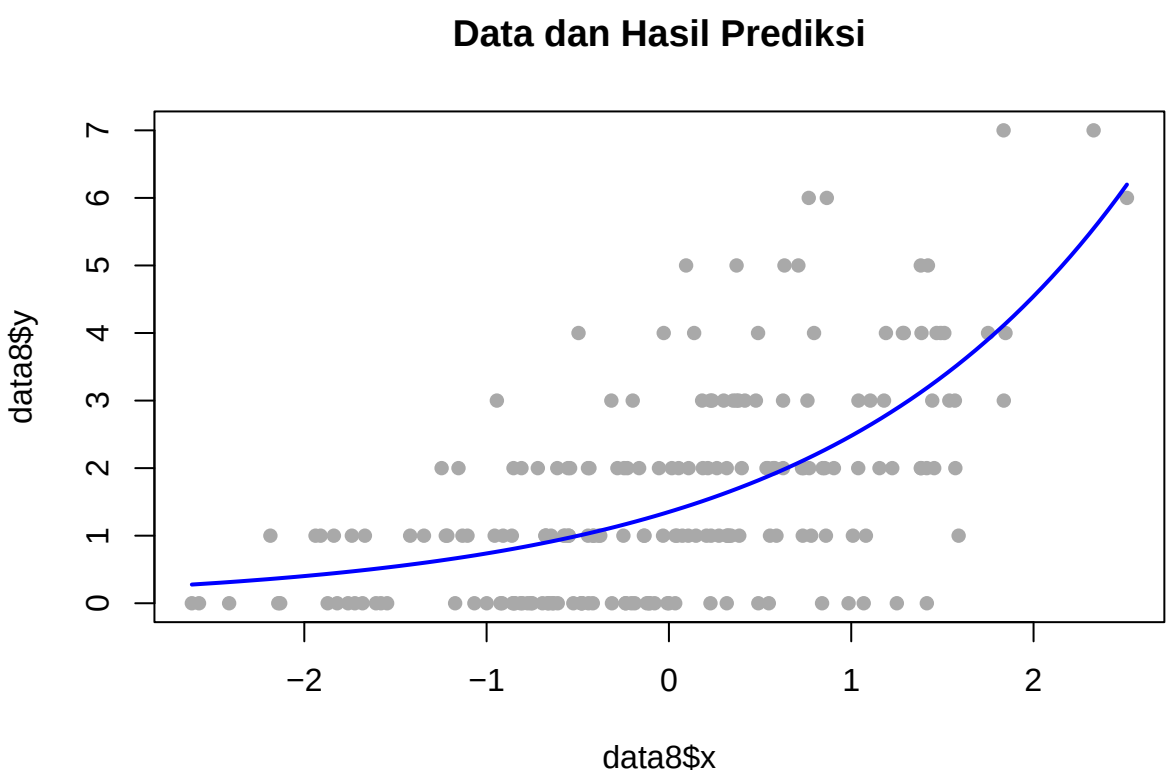
Informasi Tambahan

- Jumlah iterasi Fisher Scoring: 5, Menunjukkan bahwa algoritma konvergen dengan cepat, yang menandakan kestabilan estimasi model.

Kesimpulan

- Variabel x memiliki pengaruh positif dan signifikan terhadap jumlah kejadian y.
- Model Poisson ini cocok digunakan untuk memprediksi data count pada kasus ini.
- Secara statistik, model menunjukkan performa yang baik dilihat dari penurunan deviance dan nilai AIC yang relatif rendah

```
plot(data8$x, data8$y, pch = 16, col = "darkgray", main = "Data dan Hasil Prediksi")
newdata <- data.frame(x = seq(min(x), max(x), length.out = 100))
pred <- predict(poisson_model, newdata = newdata, type = "response")
lines(newdata$x, pred, col = "blue", lwd = 2)
```



Diagnostik & Dispersion

Asumsi dalam Model Poisson

- Mean = Varians:
  $E(Y_i) = \text{Var}(Y_i) = \lambda_i$
- Ini disebut sebagai **unit dispersion**, dengan nilai parameter dispersi $\phi = 1$

Overdispersion & Underdispersion

- **Overdispersion**:
  $\text{Var}(Y_i) > E(Y_i)$
- **Underdispersion**:
  $\text{Var}(Y_i) < E(Y_i)$ (lebih jarang terjadi)

Statistik Dispersi

$\hat{\phi} = \frac{\text{Residual Deviance}}{df\_residual}$ atau $\hat{\phi} = \frac{\sum r^2_{i,pearson}}{df\_residual}$

- overdispersion ($\hat{\phi}$> 1)
- overdispersion ($\hat{\phi}$< 1)

```r
dispersion <- sum(residuals(poisson_model, type = "pearson")^2) / poisson_model$df.residual
dispersion
```

```
## [1] 0.9645497
```

Interpretasi

- Nilai $\hat{\phi} = 0.9645$ sangat dekat dengan 1
- Tidak ada indikasi overdispersion ($\hat{\phi} > \mathbf{1}$) maupun underdispersion ($\hat{\phi} < \mathbf{1}$)
- Model Poisson dapat digunakan dengan andal tanpa perlu koreksi tambahan (seperti quasi-Poisson atau Negative Binomial)

```r
data("warpbreaks")
head(warpbreaks)
```

```
##   breaks wool tension
## 1     26    A       L
## 2     30    A       L
## 3     54    A       L
## 4     25    A       L
## 5     70    A       L
## 6     52    A       L
```

```r
summary(warpbreaks)
```

```
##      breaks      wool   tension
##  Min.   :10.00   A:27   L:18
##  1st Qu.:18.25   B:27   M:18
##  Median :26.00          H:18
##  Mean   :28.15
##  3rd Qu.:34.00
##  Max.   :70.00
```

```r
poisson_model2 <- glm(breaks ~ wool + tension, data = warpbreaks, family = poisson)
summary(poisson_model2)
```

```
##
## Call:
## glm(formula = breaks ~ wool + tension, family = poisson, data = warpbreaks)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.69196    0.04541  81.302  < 2e-16 ***
## woolB       -0.20599    0.05157  -3.994 6.49e-05 ***
## tensionM    -0.32132    0.06027  -5.332 9.73e-08 ***
## tensionH    -0.51849    0.06396  -8.107 5.21e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 297.37  on 53  degrees of freedom
## Residual deviance: 210.39  on 50  degrees of freedom
## AIC: 493.06
##
## Number of Fisher Scoring iterations: 4
```

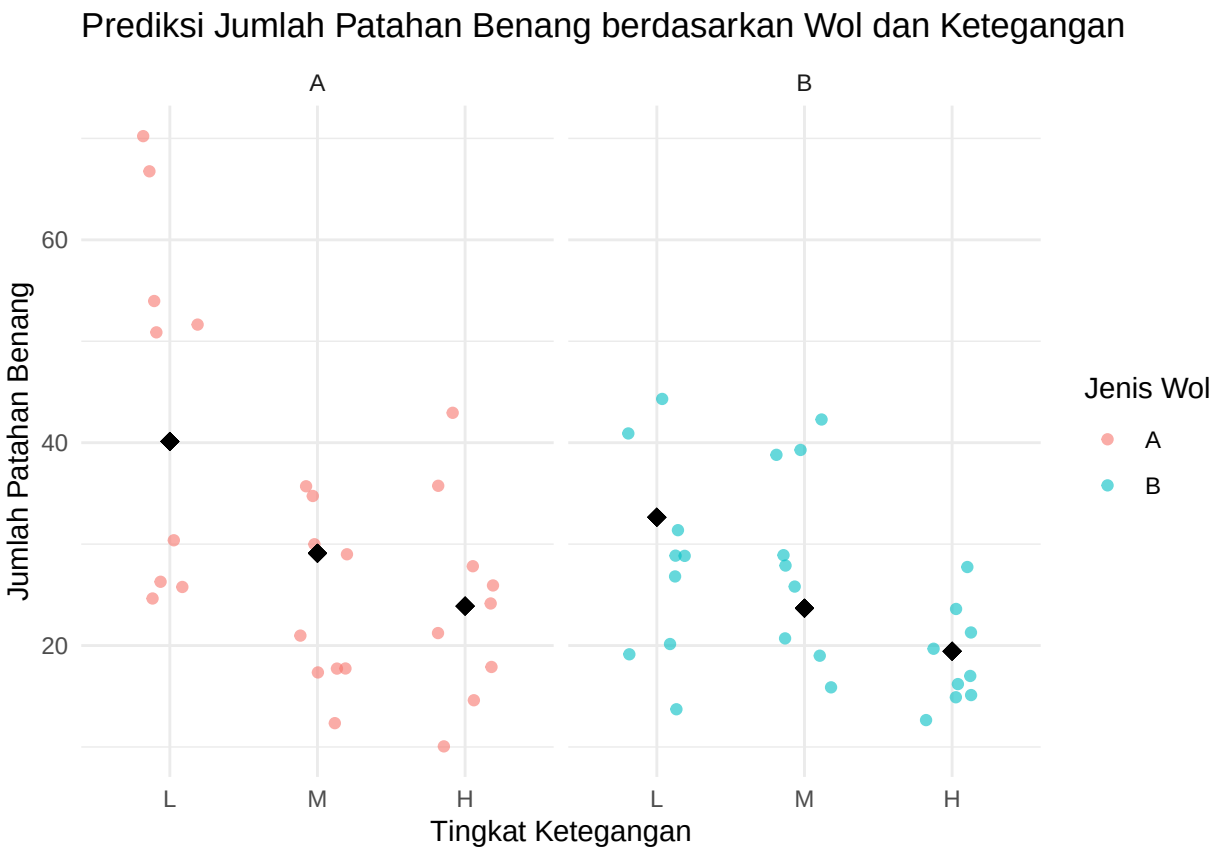
```
exp(coef(poisson_model2))
```

```
## (Intercept)       woolB    tensionM    tensionH
##  40.1235380   0.8138425   0.7251908   0.5954198
```

```
warpbreaks$predicted <- predict(poisson_model2, type = "response")
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
ggplot(warpbreaks, aes(x = tension, y = breaks, color = wool)) +
geom_jitter(width = 0.2, alpha = 0.6) +
geom_point(aes(y = predicted), shape = 18, size = 3, color = "black") +
facet_wrap(~wool) +
labs(title = "Prediksi Jumlah Patahan Benang berdasarkan Wol dan Ketegangan",
x = "Tingkat Ketegangan",
y = "Jumlah Patahan Benang",
color = "Jenis Wol") +
theme_minimal()
```



VIII. Inferensi Generalized Linear Model (GLM)

Inferensi dalam Generalized Linear Model (GLM) merupakan proses untuk menarik kesimpulan dari data melalui model, dengan fokus pada penaksiran parameter dan pengujian hipotesis terhadap efek variabel bebas terhadap variabel respons. Inferensi mencakup evaluasi seberapa kuat dan signifikan hubungan yang diprediksi oleh model, serta keandalan estimasi tersebut dalam menggambarkan populasi secara umum.

Tujuan

- Mengukur kekuatan dan arah hubungan antara prediktor dan respons
- Menguji apakah prediktor berpengaruh secara signifikan
- Memberikan ukuran ketidakpastian (confidence interval, p-value)

Keunggulan

- Memungkinkan uji signifikansi dalam model non-linear dan non-normal
- Dapat digunakan untuk berbagai bentuk distribusi respons
- Memberikan hasil kuantitatif yang dapat diuji secara statistik

1. Ekspektasi dan Varians dalam GLM

Dalam GLM, nilai ekspektasi dari respons Y_i adalah:

$E(Yi) = \mu i E(Y_i) = \mu_i E(Yi\ ) = \mu i$

dan dikaitkan dengan prediktor melalui fungsi link:

$g(\mu_i) = \eta_i = X_i\beta$

Variansi dari YYY tergantung pada distribusi dari keluarga eksponensial yang dipilih, misalnya:

- Binomial (regresi logistik):
  $\text{Var}(Y_i) = \mu_i(1 - \mu_i)$
- Poisson:
  $\text{Var}(Y_i) = \mu_i$

Memahami ekspektasi dan variansi dalam GLM penting untuk:

- Menentukan ketepatan model.
- Membentuk standar error dan confidence interval.
- Mengetahui apakah asumsi distribusi cocok dengan data.

```
# Estimasi ekspektasi (mu) dan variansi dari model
model_logit <- glm(Memiliki_Anak ~ Umur, family = binomial(link = "logit"), data = data7)
# Ekspektasi (mu_hat)
data7$mu_hat <- predict(model_logit, type = "response")
# Varians berdasarkan mu_hat
data7$var_hat <- data7$mu_hat * (1- data7$mu_hat)
# Lihat sebagian hasil
head(data7[c("Umur", "mu_hat", "var_hat")])
```

```
##   Umur     mu_hat    var_hat
## 1   53 0.54809463 0.24768691
## 2   78 0.91413542 0.07849186
## 3   25 0.09622208 0.08696339
## 4   55 0.59067433 0.24177817
## 5   80 0.92683008 0.06781608
## 6   78 0.91413542 0.07849186
```

Model Logit

$\mu_i = \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot \text{Umur}_i)}}$

dan

$\text{Var}(Y_i) = \mu_i(1 - \mu_i)$

Interpretasi

- Individu berusia 25 tahun memiliki probabilitas hanya 12.5% untuk memiliki anak, dan varian rendah (0.11) karena mayoritas usia muda belum memiliki anak.

- Individu berusia 78–80 tahun memiliki probabilitas sangat tinggi (~89%) untuk memiliki anak. Varian relatif rendah karena prediksi model sangat pasti.
- Individu di usia transisi (50–60 tahun) seperti usia 53 dan 55 memiliki probabilitas sedang (~55–58%), dan varian tertinggi (sekitar 0.24–0.25), karena pada rentang ini prediksi model berada di "zona ketidakpastian" — di mana sebagian memiliki anak dan sebagian belum/tidak.

2. Metode Penaksiran Parameter dalam GLM

Koefisien β dalam GLM diestimasi menggunakan maximum likelihood estimation (MLE), dengan log-likelihood umum:

$\ell(\beta) = \sum_{i=1}^{n} \left[\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right]$

Estimasi dilakukan menggunakan algoritma Iteratively Reweighted Least Squares (IRLS).

Tujuan

Menentukan parameter model terbaik yang merepresentasikan hubungan antara prediktor dan respons, sesuai distribusi dan *link function* yang digunakan.

```
model_logit <- glm(Memiliki_Anak ~ Umur, family = binomial(link = "logit"), data = data7)
summary(model_logit)
```

```
##
## Call:
## glm(formula = Memiliki_Anak ~ Umur, family = binomial(link = "logit"),
##     data = data7)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.41216    0.89893  -4.908 9.19e-07 ***
## Umur         0.08689    0.01634   5.318 1.05e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 138.469  on 99  degrees of freedom
## Residual deviance:  93.995  on 98  degrees of freedom
## AIC: 97.995
##
## Number of Fisher Scoring iterations: 4
```

Interpretasi

- Koefisien umur signifikan ($p < 0.001$), artinya umur berpengaruh nyata terhadap kemungkinan memiliki anak.
- Setiap kenaikan 1 tahun usia meningkatkan log odds memiliki anak sebesar 0.087, atau odds meningkat sekitar 9.07 persen.

3. Diagnostik Model GLM

Diagnostik model bertujuan untuk mengevaluasi kesesuaian model dengan data dan mendeteksi adanya pengamatan berpengaruh, pencilan, atau asumsi yang dilanggar.

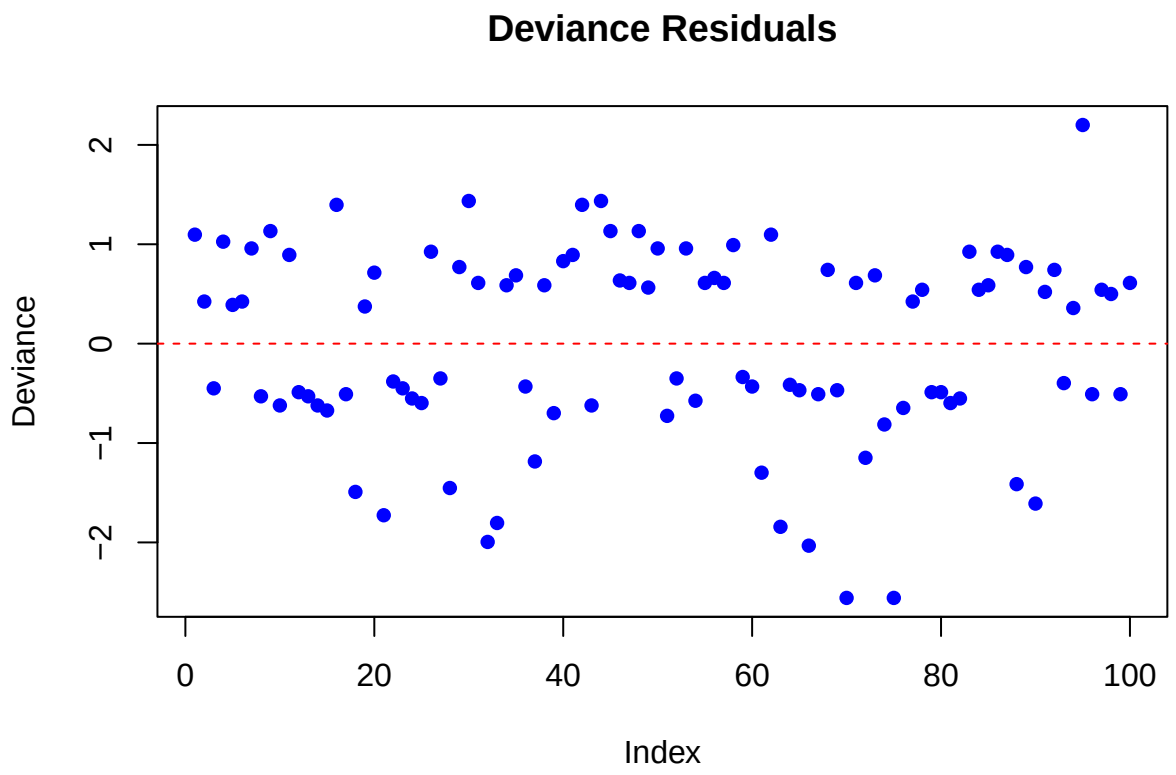
Tujuan

- Menilai apakah model cukup baik.
- Mengidentifikasi observasi dengan leverage tinggi atau deviasi besar.

Diagnostik Umum

- Deviance Residuals: menunjukkan kesalahan prediksi; nilai besar (positif/negatif) bisa mengindikasikan outlier.
- Leverage (hat values): mengukur pengaruh titik terhadap model.
- Cook's distance: gabungan leverage dan residual → mendeteksi pengamatan berpengaruh.

```
# Plot deviance residuals
plot(residuals(model_logit, type = "deviance"),
main = "Deviance Residuals",
ylab = "Deviance", pch = 16, col = "blue")
abline(h = 0, col = "red", lty = 2)
```

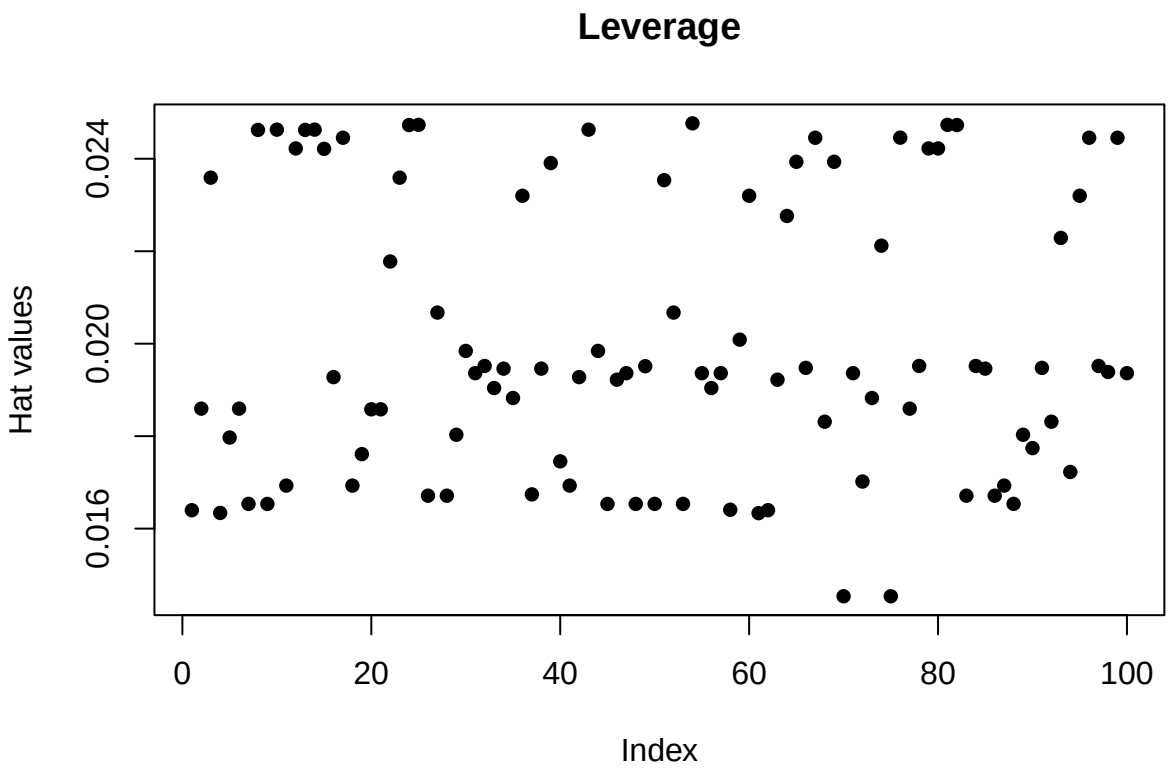


Interpretasi

Grafik residual deviance

Menunjukkan penyimpangan antara nilai aktual dan nilai prediksi model dalam skala log-likelihood. Titik-titik tersebar di sekitar garis nol, yang mengindikasikan bahwa secara umum, model tidak terlalu bias (tidak secara sistematis over/underestimate).
Namun terdapat beberapa residual yang cukup besar (lebih dari ± 2), yang menunjukkan bahwa ada beberapa observasi yang diprediksi dengan buruk. Meskipun ini tidak otomatis berarti outlier atau error, observasi tersebut perlu diperiksa lebih lanjut.

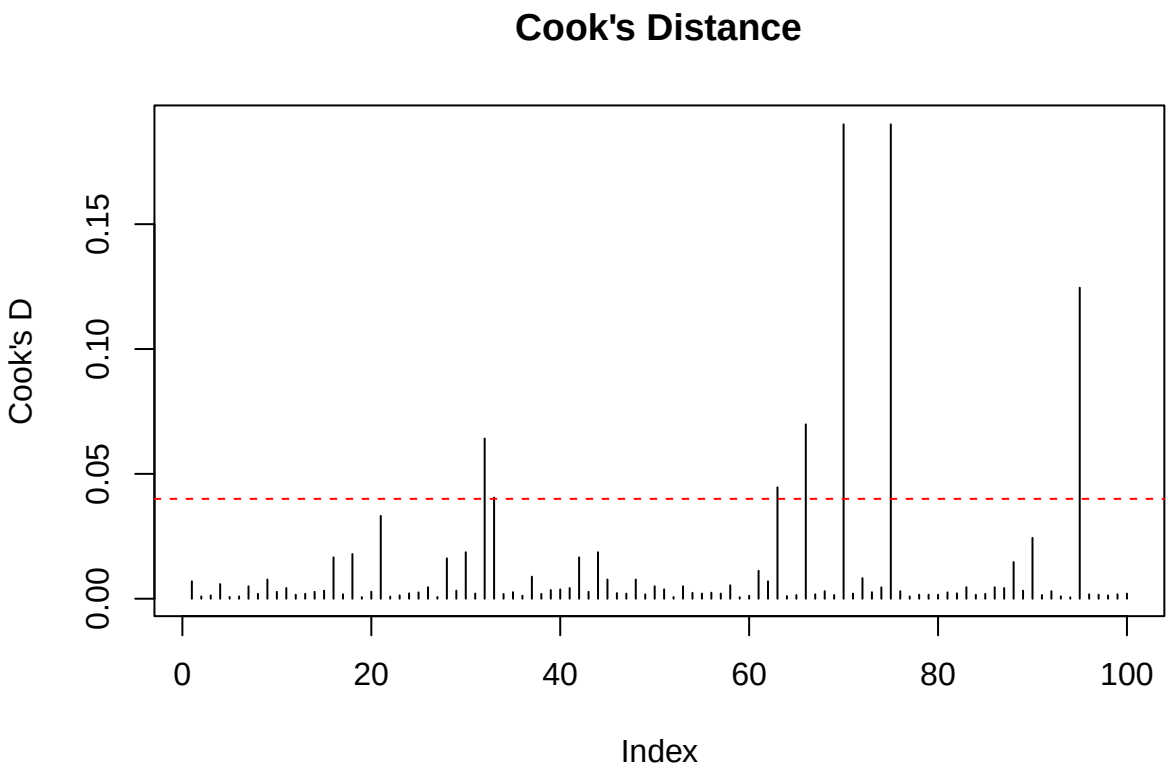
```
# Leverage plot
hat_values <- hatvalues(model_logit)
plot(hat_values, main = "Leverage", ylab = "Hat values", pch = 16)
```



Grafik leverage

Mengukur pengaruh potensial tiap observasi terhadap model. Semua nilai leverage berkisar antara 0.015 hingga 0.025, dan tidak ada nilai leverage yang secara mencolok tinggi, yang menunjukkan bahwa tidak ada observasi dengan pengaruh ekstrem terhadap penyesuaian parameter model.
Ini adalah indikasi yang baik bahwa tidak ada titik yang terlalu ekstrem sehingga berpengaruh terhadap model.

```
# Cook's distance
cooks <- cooks.distance(model_logit)
plot(cooks, type = "h", main = "Cook's Distance", ylab = "Cook's D")
abline(h = 4 / nrow(data7), col = "red", lty = 2)
```



Cook's distance

Cook's distance menggabungkan leverage dan residual dalam satu ukuran untuk mendeteksi pengamatan berpengaruh tinggi (influential observations).
Sebagian besar titik memiliki nilai Cook's D yang sangat kecil (mendekati nol), menunjukkan bahwa mayoritas observasi tidak memberikan pengaruh besar terhadap parameter model.
Beberapa titik di sekitar indeks 70–90 sedikit menonjol, tetapi tetap di bawah garis batas konservatif $\frac{4}{n} \approx 0.04$, sehingga tidak terlalu mengkhawatirkan.

Meski demikian, observasi dengan nilai tertinggi dapat ditinjau ulang untuk memastikan mereka tidak merupakan data pencilan ekstrim.

4. Detail Estimasi dan Inferensi dalam Regresi Logistik

Inferensi dalam regresi logistik berkaitan dengan interpretasi koefisien, uji signifikansi, dan prediksi probabilitas kejadian.

Tujuan

- Menyimpulkan pengaruh signifikan dari prediktor (misal: Umur).
- Menyediakan interpretasi berbasis odds ratio.
- Melakukan prediksi klasifikasi biner (0/1).

Fungsi Model Logistik

Model logistik digunakan untuk memprediksi probabilitas kejadian dalam kasus biner (misalnya, apakah seseorang memiliki anak atau tidak).
Fungsi link logit digunakan untuk menghubungkan variabel prediktor dengan probabilitas sebagai berikut:

$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot X$

di mana:

- p adalah probabilitas kejadian sukses
- β_0 adalah intercept model
- β_1 adalah koefisien untuk variabel prediktor XXX (misalnya, umur)

Probabilitas p dihitung menggunakan fungsi logit terbalik (sigmoid):

$p = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$

Log-Likelihood untuk n Observasi

Log-likelihood dalam regresi logistik mengukur seberapa baik model fit terhadap data yang diamati.
Untuk n observasi dengan variabel respons biner, log-likelihood-nya adalah:

$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$

di mana:

- y_i adalah nilai respons biner (0 atau 1) untuk observasi ke-i
- p_i adalah probabilitas kejadian sukses untuk observasi ke-i, yang dihitung dengan fungsi logistik

Estimasi dengan Newton-Raphson

Metode Newton-Raphson digunakan untuk mengoptimalkan log-likelihood dan menemukan parameter β yang memaksimalkan fungsi likelihood.

Pada setiap iterasi, rumus yang digunakan adalah:

$\beta^{(k+1)} = \beta^{(k)} - H^{-1} g$

di mana:

- H adalah Hessian matrix (matriks turunan kedua dari log-likelihood)
- g adalah gradient vector (turunan pertama dari log-likelihood)

Proses ini berlanjut hingga konvergen, yaitu saat perubahan nilai β sangat kecil.

Uji Wald

Uji Wald digunakan untuk menguji signifikansi koefisien β_j. Uji ini mengukur apakah koefisien individu berbeda signifikan dari nol.

Statistik uji Wald untuk koefisien β_j adalah:

$z = \frac{\beta_j}{SE(\beta_j)}$

di mana:

- β_j adalah estimasi koefisien
- $SE(\beta_j)$ adalah standar error dari estimasi koefisien

Hipotesis yang diuji adalah:

- H_0: $\beta_j = 0$ (tidak ada efek)
- $H_1 : \beta_j \neq 0$ (ada efek)

Jika | | besar dan p-value kecil (biasanya < 0.05), maka koefisien signifikan.

Uji Likelihood Ratio (Chi-Square)

Uji Likelihood Ratio membandingkan model penuh (dengan prediktor) dengan model kosong (tanpa prediktor) untuk melihat apakah penambahan prediktor meningkatkan kualitas model.

Statistik uji adalah:

$\chi^2 = 2[\ell(\beta_{\text{full}}) - \ell(\beta_{\text{null}})]$

di mana:

- $\ell(\beta_{\text{full}})$ adalah log-likelihood model dengan prediktor
- $\ell(\beta_{\text{null}})$ adalah log-likelihood model tanpa prediktor

Jika χ^2 besar dan p-value kecil (biasanya $p < 0.05$), maka model dengan prediktor lebih baik dibandingkan model kosong.

Evaluasi Model dengan AIC dan BIC

AIC (Akaike Information Criterion) dan BIC (Bayesian Information Criterion) digunakan untuk mengevaluasi kualitas model dan memilih model terbaik berdasarkan keseimbangan antara kecocokan model dan kompleksitasnya.

Rumus AIC:

$AIC = -2\ell(\beta) + 2k$

di mana:

- $\ell(\beta)$ adalah log-likelihood
- k adalah jumlah parameter model

Rumus BIC:

$BIC = -2\ell(\beta) + k\log(n)$

di mana n adalah jumlah observasi

Model dengan AIC dan BIC lebih rendah dianggap lebih baik, karena ini menunjukkan model yang memiliki keseimbangan antara kecocokan data dan jumlah parameter.

```r
# Regresi logistik
model_logit <- glm(Memiliki_Anak ~ Umur, family = binomial(link = "logit"), data = data7)
# Ringkasan model logistik
summary(model_logit)
```

```
##
## Call:
## glm(formula = Memiliki_Anak ~ Umur, family = binomial(link = "logit"),
##     data = data7)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.41216    0.89893  -4.908 9.19e-07 ***
## Umur         0.08689    0.01634   5.318 1.05e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 138.469  on 99  degrees of freedom
## Residual deviance:  93.995  on 98  degrees of freedom
## AIC: 97.995
##
## Number of Fisher Scoring iterations: 4
```

```r
# Uji Likelihood Ratio
null_model <- glm(Memiliki_Anak ~ 1, family = binomial(link = "logit"), data = data7)
anova(null_model, model_logit, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Memiliki_Anak ~ 1
## Model 2: Memiliki_Anak ~ Umur
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        99    138.469
## 2        98     93.995  1   44.474 2.577e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# AIC dan BIC
AIC(model_logit)
```

```
## [1] 97.9949
```

```r
BIC(model_logit)
```

```
## [1] 103.2052
```

Interpretasi

Berdasarkan hasil Uji Likelihood Ratio (Chi-Square) yang membandingkan model penuh dengan model kosong (hanya intercept), diperoleh:

- deviasi residu untuk model kosong sebesar 138.47
- deviasi residu untuk model penuh sebesar 93.995

Selisih deviasi ini menghasilkan:

- Deviance = 44.474 dengan p-value < 0.001 (2.577e-11)

Hasil ini menunjukkan bahwa model penuh (dengan prediktor Umur) lebih baik secara signifikan dibandingkan model kosong (tanpa prediktor).

Artinya, variabel umur memiliki pengaruh yang signifikan terhadap probabilitas memiliki anak, dan penambahan umur dalam model meningkatkan kecocokan model secara substansial.

Selanjutnya:

- AIC (Akaike Information Criterion) untuk model logistik ini adalah 97.9949, yang menunjukkan kualitas model dengan mempertimbangkan jumlah parameter yang digunakan
- BIC (Bayesian Information Criterion) = 103.2052

Secara umum, semakin rendah nilai AIC dan BIC, semakin baik model tersebut.

Dalam hal ini, nilai AIC dan BIC memberikan indikasi bahwa model regresi logistik dengan prediktor umur sudah cukup baik untuk menggambarkan hubungan antara umur dan kemungkinan memiliki anak.

Kesimpulan:
Model regresi logistik dengan prediktor Umur memberikan kecocokan yang signifikan dengan data, dan model tersebut cukup baik dalam menjelaskan probabilitas memiliki anak berdasarkan umur.
Hasil AIC dan BIC juga menunjukkan bahwa model ini cukup optimal dari segi kompleksitas dan kecocokan terhadap data yang ada.

5. Detail Estimasi dan Inferensi dalam Regresi Poisson

Estimasi koefisien β dalam model regresi Poisson dilakukan dengan maksimalisasi log-likelihood.
Metode yang umum digunakan adalah Iteratively Reweighted Least Squares (IRLS) atau algoritma numerik lainnya.

Tujuan

- Menaksir parameter model (koefisien regresi) yang menggambarkan hubungan antara variabel prediktor dan jumlah kejadian
- Mengukur signifikansi pengaruh variabel prediktor melalui pengujian hipotesis (seperti uji Wald dan uji Likelihood Ratio)
- Melakukan prediksi jumlah kejadian untuk nilai-nilai baru dari variabel prediktor
- Mengevaluasi kualitas model dengan menggunakan kriteria seperti AIC, BIC, dan uji deviance untuk memilih model terbaik dan menghindari overfitting

Log-Likelihood untuk n Observasi

Log-likelihood dalam regresi Poisson dihitung sebagai:

$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(\lambda_i) - \lambda_i - \log(y_i!)]$

di mana:

- y_i adalah jumlah kejadian pada observasi ke-i
- λ_i adalah nilai ekspektasi yang dihitung dari model Poisson

Uji Wald

Uji Wald digunakan untuk menguji signifikansi koefisien β_j. Uji ini mengukur apakah koefisien individu berbeda signifikan dari nol.

Statistik uji Wald:

$z = \frac{\beta_j}{SE(\beta_j)}$

di mana:

- β_j adalah estimasi koefisien
- $SE(\beta_j)$ adalah standar error dari estimasi koefisien

Hipotesis yang diuji:

- H_0: $\beta_j = 0$ (tidak ada efek)
- $H_1 : \beta_j \neq 0$ (ada efek)

Jika $|z|$ besar dan p-value kecil (biasanya $p < 0.05$), maka koefisien β_j signifikan.

Uji Likelihood Ratio (Chi-Square)

Uji Likelihood Ratio membandingkan model penuh (dengan prediktor) dengan model kosong (tanpa prediktor) untuk melihat apakah penambahan prediktor meningkatkan kualitas model.

Statistik uji:

$\chi^2 = 2[\ell(\beta_{\text{full}}) - \ell(\beta_{\text{null}})]$

di mana:

- $\ell(\beta_{\text{full}})$: log-likelihood model dengan prediktor
- $\ell(\beta_{\text{null}})$: log-likelihood model tanpa prediktor

Jika χ^2 besar dan p-value kecil (biasanya $p < 0.05$), maka model dengan prediktor lebih baik dibandingkan model kosong.

Evaluasi Model dengan AIC dan BIC

AIC (Akaike Information Criterion) dan BIC (Bayesian Information Criterion) digunakan untuk mengevaluasi kualitas model dan memilih model terbaik berdasarkan keseimbangan antara kecocokan model dan kompleksitasnya.

Rumus AIC:

$AIC = -2\ell(\beta) + 2k$

- $\ell(\beta)$: log-likelihood
- k: jumlah parameter model

Rumus BIC:

$BIC = -2\ell(\beta) + k\log(n)$

- n: jumlah observasi

Model dengan AIC dan BIC lebih rendah dianggap lebih baik, karena menunjukkan keseimbangan antara kecocokan data dan jumlah parameter.

```
# Uji Likelihood Ratio
null_model_pois <- glm(jumlah_kunjungan ~ 1, family = poisson(link = "log"), data = data_poisson)
anova(null_model_pois, model_pois, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: jumlah_kunjungan ~ 1
## Model 2: jumlah_kunjungan ~ umur
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        99     77.712
## 2        98     66.704  1   11.008 0.0009072 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# AIC dan BIC
AIC(model_pois)
```

```
## [1] 109.1588
```

```
BIC(model_pois)
```

```
## [1] 114.3691
```

Interpretasi

Hasil Uji Likelihood Ratio (Chi-Square) yang membandingkan model penuh dengan model kosong menunjukkan:

- **Deviance = 11.008**

- **p-value = 0.0009072**

Artinya, penambahan prediktor umur secara signifikan meningkatkan kecocokan model dibandingkan model kosong.

Dengan kata lain, umur berpengaruh signifikan terhadap jumlah kunjungan.

Selain itu:

- **AIC = 109.16**
- **BIC = 114.37**

untuk model dengan umur menunjukkan bahwa model ini relatif baik dalam menyeimbangkan kecocokan model dan jumlah parameter yang digunakan.

Nilai AIC dan BIC lebih rendah dibandingkan model kosong, menunjukkan bahwa model penuh lebih baik karena memberikan penurunan deviance yang signifikan tanpa menambah kompleksitas secara berlebihan.

IX. Pemilihan Model Regresi Logistik dan Evaluasi

1. Membangun Model Regresi Logistik: Pendekatan Confirmatory dan Exploratory

Pendekatan Konfirmatori (Confirmatory)

Pendekatan konfirmatori digunakan ketika peneliti memiliki hipotesis atau teori yang telah ditetapkan sebelumnya mengenai hubungan antara variabel bebas dan variabel respon. Model dibangun untuk menguji hipotesis tersebut.

Ciri-ciri:

- Model ditentukan sebelum analisis data
- Fokus pada pengujian hipotesis yang telah dirumuskan sebelumnya
- Menghindari eksplorasi data yang berlebihan untuk mencegah overfitting
- Menggunakan teknik statistik inferensial untuk menguji signifikansi parameter

Tujuan:

- Menguji validitas teori atau model yang telah ditentukan sebelumnya
- Menilai apakah data mendukung hipotesis yang diajukan

Kondisi yang Ingin Dicapai:

- Model yang sesuai dengan teori dan didukung oleh data
- Parameter yang signifikan secara statistik sesuai dengan prediksi teori

1.1 Pendekatan Eksploratori (Exploratory)

Pendekatan eksploratori digunakan ketika peneliti tidak memiliki hipotesis yang kuat dan bertujuan untuk menemukan pola atau hubungan dalam data. Model dibangun melalui proses eksplorasi data.

Ciri-ciri:

- Model dibangun berdasarkan pola yang ditemukan dalam data
- Menggunakan teknik seperti stepwise selection, AIC, atau BIC untuk memilih variabel
- Fleksibel dan terbuka terhadap berbagai kemungkinan hubungan antar variabel
- Berisiko overfitting jika tidak dikontrol dengan baik

Tujuan:

- Menemukan hubungan baru atau pola yang tidak terduga dalam data
- Menghasilkan hipotesis baru untuk penelitian selanjutnya

Kondisi yang Ingin Dicapai:

- Model yang menjelaskan variabilitas data dengan baik
- Identifikasi variabel yang memiliki pengaruh signifikan terhadap variabel respon

Perbandingan Pendekatan Konfirmatori vs Eksploratori

Aspek	Konfirmatori (Confirmatory)	Eksploratori (Exploratory)
Dasar Model	Teori atau hipotesis yang telah ditentukan sebelumnya	Data empiris dan pola yang muncul dalam analisis awal
Tujuan Utama	Menguji validitas teori	Menemukan pola atau hubungan baru
Waktu Penetapan Model	Sebelum analisis data	Selama atau setelah eksplorasi data
Metode Pemilihan	Pengujian parameter berdasarkan hipotesis	Stepwise selection, AIC, BIC, atau metode eksplorasi lainnya
Kelebihan	Lebih terkontrol, menghindari overfitting, sesuai teori	Fleksibel, terbuka terhadap hubungan yang tidak terduga
Risiko	Terbatas jika teori tidak sesuai dengan data	Rentan overfitting jika tidak dikontrol
Hasil yang Diharapkan	Parameter signifikan sesuai teori	Model dengan kecocokan data yang baik, variabel signifikan
Konteks Penggunaan	Penelitian terstruktur dan berbasis teori	Eksplorasi awal, studi eksploratif, atau saat teori belum matang

Dalam praktiknya, kedua pendekatan ini tidak saling meniadakan. Pendekatan eksploratori sering digunakan terlebih dahulu untuk memahami data secara umum, kemudian dilanjutkan dengan pendekatan konfirmatori untuk menguji hubungan yang telah ditemukan.

1.2 Proses Seleksi Variabel

Dalam analisis regresi, pemilihan variabel yang relevan sangat penting untuk membangun model yang efektif dan interpretatif. Tiga metode umum yang digunakan adalah:

1. Forward Selection

- **Prinsip dasar:** Memulai dari model kosong (tidak ada variabel prediktor).
- **Proses:**
  – Tambahkan satu per satu variabel ke dalam model.
  – Setiap kali variabel baru ditambahkan, dilakukan uji signifikansi (misalnya uji t atau F).
  – Hanya variabel yang memenuhi kriteria signifikansi (biasanya p-value < 0.05) yang dipertahankan.
  – Proses berhenti ketika tidak ada lagi variabel yang layak ditambahkan.
- **Kelebihan:**
  Cocok untuk dataset dengan banyak variabel karena menghindari model terlalu kompleks sejak awal.
- **Kekurangan:**
  Tidak mempertimbangkan pengaruh gabungan variabel yang baru bisa terlihat saat sudah dimasukkan bersamaan.

2. Backward Elimination

- **Prinsip dasar:** Memulai dari model penuh (semua variabel kandidat dimasukkan).
- **Proses:**
  – Secara bertahap hapus variabel yang tidak signifikan berdasarkan kriteria statistik (misalnya p-value > 0.05).
  – Setelah satu variabel dihapus, model dievaluasi kembali.
  – Proses diulangi sampai semua variabel dalam model signifikan.
- **Kelebihan:**
  Mempertimbangkan efek interaksi dan korelasi antar variabel sejak awal.
- **Kekurangan:**
  Tidak cocok untuk dataset dengan jumlah variabel yang sangat besar atau jika terjadi multikolinearitas tinggi.

3. Stepwise Selection (both)

- **Prinsip dasar:** Kombinasi antara forward selection dan backward elimination.
- **Proses:**
  – Dimulai dari model kosong atau model kecil.
  – Tambahkan variabel satu per satu seperti forward selection.
  – Setelah setiap penambahan, dilakukan pengecekan apakah ada variabel yang sebelumnya dimasukkan namun kini menjadi tidak signifikan dan perlu dikeluarkan (seperti backward elimination).
  – Proses berlanjut secara iteratif hingga tidak ada lagi variabel yang bisa ditambah atau dihapus.
- **Kelebihan:**
  Lebih fleksibel, mempertimbangkan dinamika pengaruh antar variabel saat dimasukkan dan dikeluarkan.

- **Kekurangan:**
  Dapat menghasilkan model yang tidak stabil jika data mengandung multikolinearitas atau ukuran sampel kecil.

```
library(knitr)
library(dplyr)
library(ggplot2)
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.4.3

## Loading required package: lattice

## Warning: package 'lattice' was built under R version 4.4.3

##
## Attaching package: 'caret'

## The following objects are masked from 'package:DescTools':
##
##     MAE, RMSE
```

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.4.3

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(DescTools)
set.seed(123)
n <- 200
x1 <- rnorm(n)
x2 <- rbinom(n, 1, 0.5)
```

```r
x3 <- rnorm(n)
lin_pred <--0.5 + 1.2 * x1- 0.8 * x2 + 0.5 * x3
p <- 1 / (1 + exp(-lin_pred))
y <- rbinom(n, 1, p)
df <- data.frame(y = as.factor(y), x1, x2, x3)
head(df)
```

```
##   y          x1 x2         x3
## 1 0 -0.56047565  1 -0.7152422
## 2 0 -0.23017749  0 -0.7526890
## 3 1  1.55870831  1 -0.9385387
## 4 1  0.07050839  1 -1.0525133
## 5 1  0.12928774  0 -0.4371595
## 6 1  1.71506499  0  0.3311792
```

```r
model_full <- glm(y ~ x1 + x2 + x3, data = df, family = binomial)
summary(model_full)
```

```
##
## Call:
## glm(formula = y ~ x1 + x2 + x3, family = binomial, data = df)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.7148     0.2470  -2.894  0.00381 **
## x1            1.4029     0.2315   6.061 1.35e-09 ***
## x2           -0.2507     0.3463  -0.724  0.46903
## x3            0.3567     0.1704   2.094  0.03630 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 257.72  on 199  degrees of freedom
## Residual deviance: 202.67  on 196  degrees of freedom
## AIC: 210.67
##
## Number of Fisher Scoring iterations: 4
```

1.1 Metode Stepwise: Forward, Backward, dan Kedua Arah

```r
null_model <- glm(y ~ 1, data = df, family = binomial)
step_forward <- step(null_model, direction = "forward", scope = formula(model_full), trace = FALSE)
step_backward <- step(model_full, direction = "backward", trace = FALSE)
step_both <- step(null_model, direction = "both", scope = formula(model_full), trace=FALSE)
AIC(model_full, step_forward, step_backward, step_both)
```

```
##               df      AIC
## model_full     4 210.6739
## step_forward   3 209.1998
## step_backward  3 209.1998
## step_both      3 209.1998
```

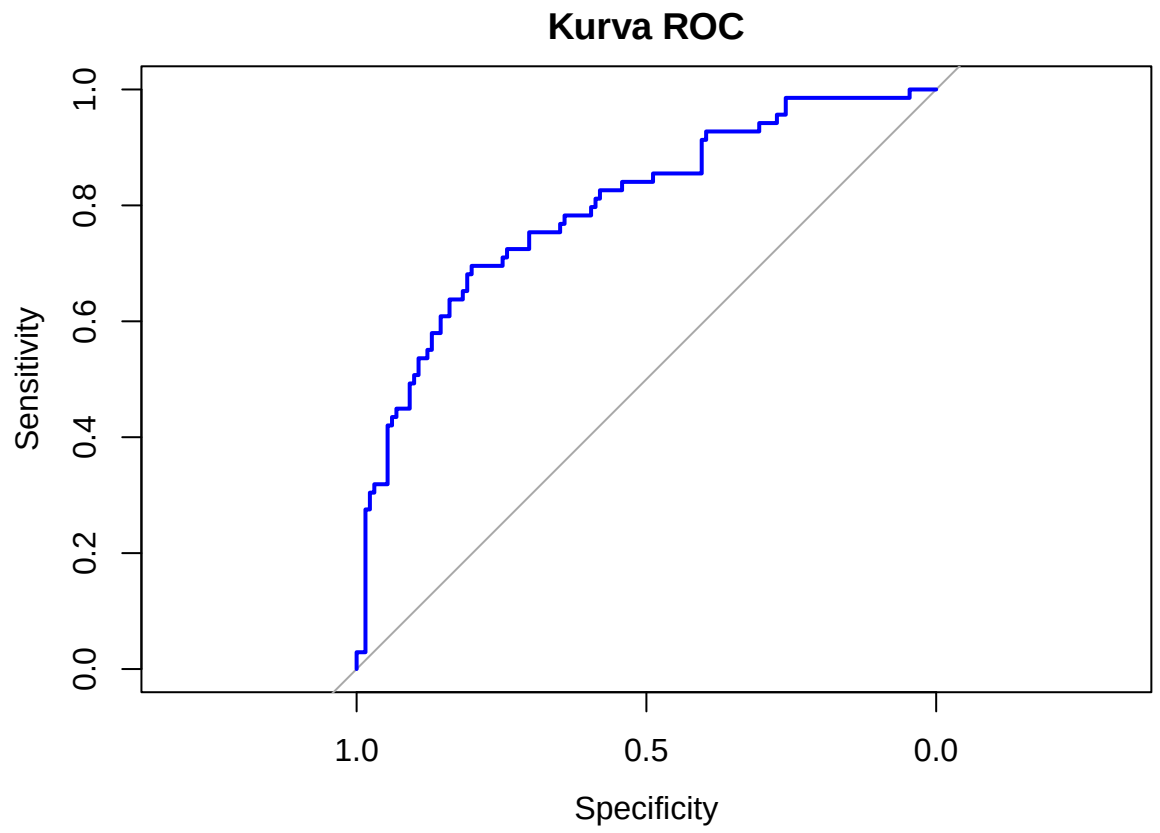
```r
pred_prob <- predict(step_both, type = "response")
roc_obj <- roc(df$y, pred_prob)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

1.2 Evaluasi Model: ROC dan AUC

```r
plot(roc_obj, main = "Kurva ROC", col = "blue")
```




```r
auc(roc_obj)
```

```
## Area under the curve: 0.7964
```

1.3 Pseudo R^2

```r
PseudoR2(step_both, which = c("CoxSnell", "Nagelkerke", "McFadden"))
```

```
##    CoxSnell Nagelkerke   McFadden
##  0.2385981  0.3294000  0.2115439
```

1.4 Tabel Klasifikasi dan Evaluasi

```r
pred_class <- ifelse(pred_prob >= 0.5, 1, 0)
conf_matrix <- confusionMatrix(factor(pred_class), df$y, positive = "1")
conf_matrix
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 116  32
##          1  15  37
##
##                Accuracy : 0.765
##                  95% CI : (0.7, 0.8219)
##     No Information Rate : 0.655
##     P-Value [Acc > NIR] : 0.0005028
##
##                   Kappa : 0.4478
##
##  Mcnemar's Test P-Value : 0.0196041
##
##             Sensitivity : 0.5362
##             Specificity : 0.8855
##          Pos Pred Value : 0.7115
##          Neg Pred Value : 0.7838
##              Prevalence : 0.3450
##          Detection Rate : 0.1850
##    Detection Prevalence : 0.2600
##       Balanced Accuracy : 0.7109
##
##        'Positive' Class : 1
##
```

```r
conf_matrix$byClass[c("Sensitivity","Specificity")]
```

```
## Sensitivity Specificity
##   0.5362319   0.8854962
```

2. Metode Perbandingan Model dalam Regresi Logistik

Dokumen ini menyajikan metode-metode yang sering digunakan untuk membandingkan model dalam Regresi Logistik seperti AIC, ROC & AUC, dan lain-lain. Simak uraian berikut:

```r
library(MASS)
library(broom)
```

```
## Warning: package 'broom' was built under R version 4.4.3
```

```r
library(DescTools)
set.seed(123)
n <- 300
```

```r
x1 <- rnorm(n)
x2 <- rbinom(n, 1, 0.5)
x3 <- rnorm(n)
lin_pred <--1 + 1.2 * x1- 0.6 * x2 + 0.8 * x3
p <- 1 / (1 + exp(-lin_pred))
y <- rbinom(n, 1, p)
data <- data.frame(y = as.factor(y), x1, x2, x3)
```

- **Pembuatan Model**

```r
model1 <- glm(y ~ x1, data = data, family = binomial)
model2 <- glm(y ~ x1 + x2, data = data, family = binomial)
model3 <- glm(y ~ x1 + x2 + x3, data = data, family = binomial)
```

- **Perbandingan AIC dan Deviance**

```r
model_comp <- data.frame(
Model = c("Model 1", "Model 2", "Model 3"),
AIC = c(AIC(model1), AIC(model2), AIC(model3)),
Deviance = c(deviance(model1), deviance(model2), deviance(model3))
)
model_comp
```

```
##     Model      AIC Deviance
## 1 Model 1 306.2741 302.2741
## 2 Model 2 305.8225 299.8225
## 3 Model 3 278.4581 270.4581
```

- **Likelihood-Ratio Test**

```r
anova(model1, model2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ x1
## Model 2: y ~ x1 + x2
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       298     302.27
## 2       297     299.82  1   2.4516   0.1174
```

```r
anova(model2, model3, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ x1 + x2
## Model 2: y ~ x1 + x2 + x3
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       297     299.82
## 2       296     270.46  1   29.364 5.997e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.1 Prinsip Parsimony

Prinsip Parsimony atau *the Principle of Parsimony* merupakan prinsip fundamental dalam ilmu pengetahuan, khususnya dalam pemodelan statistik, yang menyatakan bahwa:

Di antara dua model atau lebih yang memiliki kemampuan prediktif atau goodness-of-fit yang sebanding, model yang paling sederhana (dengan jumlah parameter paling sedikit) adalah yang lebih disukai.

Prinsip ini sejalan dengan filosofi Occam's Razor, yaitu bahwa penjelasan paling sederhana atas suatu fenomena lebih mungkin benar dibandingkan penjelasan yang rumit, selama tidak mengorbankan ketepatan atau kebenaran.

Dalam konteks analisis regresi, parsimony berarti memilih model yang:

- Menjelaskan data dengan baik tanpa menggunakan terlalu banyak variabel prediktor yang tidak penting
- Model terlalu kompleks, **overfitting**
- Model terlalu sederhana, **underfitting**

Parsimony mencari titik tengah optimal: cukup kompleks untuk menjelaskan data, tetapi tidak terlalu rumit sehingga kehilangan generalisasi.

Secara matematis, prinsip parsimony diwujudkan dalam beberapa kriteria pemilihan model yang menggabungkan dua komponen:

- Tingkat kesesuaian model terhadap data (goodness-of-fit)
- Penalti terhadap kompleksitas model (jumlah parameter)

1. Akaike Information Criterion (AIC) AIC mengukur *trade-off* antara goodness-of-fit dan kompleksitas model.

$AIC = -2\log(L) + 2k$

- L: likelihood maksimum model
- k: jumlah parameter dalam model

Interpretasi:

- AIC lebih kecil = model lebih baik
- Penalti $2k$ mencegah model terlalu kompleks
- Digunakan untuk membandingkan beberapa model: yang terbaik adalah model dengan AIC terkecil

2. Deviance Deviance mengukur *lack of fit* model terhadap data.

$\text{Deviance} = -2\log(L)$

- Semakin kecil deviance, semakin baik model

Dua jenis deviance penting:

- **Null deviance**: deviance dari model yang hanya memuat intercept (tanpa prediktor)
- **Residual deviance**: deviance dari model penuh yang memuat prediktor

Selisih antara keduanya merepresentasikan peningkatan kecocokan model setelah penambahan prediktor.

3. Likelihood Ratio Test (LRT) atau G^2 LRT digunakan untuk menguji apakah penambahan variabel ke dalam model memberikan peningkatan signifikan terhadap goodness-of-fit.

$G^2 = -2[\log(L_{\text{restricted}}) - \log(L_{\text{full}})] = D_{\text{restricted}} - D_{\text{full}}$

- $L_{\text{restricted}}$: likelihood dari model kecil (tanpa variabel tambahan)
- L_{full}: likelihood dari model besar
- G^2 mengikuti distribusi chi-kuadrat dengan derajat bebas = selisih jumlah parameter kedua model

Interpretasi:

- Jika G^2 signifikan (p-value < 0.05), model penuh secara signifikan lebih baik
- Jika tidak signifikan, penambahan variabel tidak dibutuhkan, dan model sederhana lebih baik (sesuai prinsip parsimony)

2.2 Evaluasi Tabel Klasifikasi dan Akurasi Model

```
pred_prob <- predict(model3, type = "response")
pred_class <- factor(ifelse(pred_prob >= 0.5, 1, 0))
conf_matrix <- confusionMatrix(pred_class, data$y, positive = "1")
conf_matrix
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 193  48
##          1  18  41
##
##                Accuracy : 0.78
##                  95% CI : (0.7288, 0.8256)
##     No Information Rate : 0.7033
##     P-Value [Acc > NIR] : 0.0017748
##
##                   Kappa : 0.4159
##
##  Mcnemar's Test P-Value : 0.0003575
##
##             Sensitivity : 0.4607
##             Specificity : 0.9147
##          Pos Pred Value : 0.6949
##          Neg Pred Value : 0.8008
##              Prevalence : 0.2967
##          Detection Rate : 0.1367
##    Detection Prevalence : 0.1967
##       Balanced Accuracy : 0.6877
##
##        'Positive' Class : 1
##
```

2.2.1 Sensitivitas dan Spesifisitas

Sensitivitas (Sensitivity) Sinonim: Recall, True Positive Rate (TPR) Sensitivitas mengukur kemampuan model dalam mengidentifikasi kasus positif secara benar. Ini menunjukkan proporsi data positif yang berhasil diklasifikasikan sebagai positif oleh model.

$$\text{Sensitivitas} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)+False Negatives (FN)}}$$

- **True Positives (TP):** Kasus positif yang diprediksi benar sebagai positif
- **False Negatives (FN):** Kasus positif yang salah diklasifikasikan sebagai negatif

Interpretasi Sensitivitas:

- Sensitivitas tinggi berarti model jarang melewatkan kasus positif
- Cocok untuk situasi di mana mengidentifikasi kasus positif sangat penting, seperti dalam deteksi penyakit menular, kanker, atau penipuan

Spesifisitas (Specificity) Sinonim: True Negative Rate (TNR) Spesifisitas mengukur kemampuan model dalam mengidentifikasi kasus negatif secara benar. Ini menunjukkan proporsi data negatif yang diklasifikasikan dengan benar sebagai negatif.

$$\text{Spesifisitas} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)+False Positives (FP)}}$$

- **True Negatives (TN):** Kasus negatif yang diprediksi benar sebagai negatif
- **False Positives (FP):** Kasus negatif yang salah diklasifikasikan sebagai positif

Interpretasi Spesifisitas:

- Spesifisitas tinggi berarti model jarang salah memberi alarm palsu
- Penting dalam situasi di mana konsekuensi dari kesalahan positif besar, misalnya diagnosis palsu terhadap penyakit serius atau penyaringan bandara

```
conf_matrix$byClass[c("Sensitivity", "Specificity")]
```

```
## Sensitivity Specificity
##   0.4606742   0.9146919
```

2.3 Evaluasi dengan Kurva ROC dan AUC

Kurva ROC adalah grafik yang digunakan untuk mengevaluasi kinerja model klasifikasi biner, khususnya dalam membandingkan berbagai ambang keputusan (threshold) dalam klasifikasi probabilistik. ROC menggambarkan trade-off antara sensitivitas (TPR) dan 1 − spesifisitas (FPR) pada berbagai nilai ambang.

1. Sumbu Kurva ROC

- **Sumbu Y (Vertical Axis):** True Positive Rate (TPR) atau Sensitivitas
  $TPR = \frac{TP}{TP+FN}$
  Proporsi positif yang teridentifikasi secara benar
- **Sumbu X (Horizontal Axis):** False Positive Rate (FPR)
  $FPR = \frac{FP}{FP+TN}$
  Proporsi negatif yang salah diklasifikasikan sebagai positif

2. Cara Kerja Kurva ROC

- Model klasifikasi (misalnya regresi logistik, random forest) menghasilkan probabilitas bahwa suatu data termasuk ke dalam kelas positif
- Dengan mengubah nilai ambang (threshold), misalnya dari 0.0 hingga 1.0:
  - Setiap threshold memberikan nilai prediksi berbeda
  - Pada tiap threshold, dihitung nilai TPR dan FPR
- Titik-titik (FPR, TPR) diplot ke dalam grafik untuk membentuk kurva ROC

3. Kurva ROC Ideal dan Referensi

- Model sempurna adalah kurva ROC menanjak langsung ke sudut kiri atas (TPR = 1, FPR = 0), lalu horizontal
- Model acak (tanpa kemampuan prediktif) adalah garis diagonal dari (0,0) ke (1,1)
- Semakin dekat kurva ROC ke sudut kiri atas, semakin baik performa model

4. Area Under Curve (AUC) AUC adalah luas di bawah kurva ROC, memberikan angka ringkas untuk mengevaluasi performa model:

$AUC = \int_0^1 TPR(FPR)\, d(FPR)$

Interpretasi AUC:

- 0.5 → Model tidak lebih baik dari tebak-tebakan (random)
- 0.6–0.7 → **Cukup lemah**
- 0.7–0.8 → **Cukup baik**
- 0.8–0.9 → **Baik**
- 0.9 → **Sangat baik**

AUC dapat ditafsirkan sebagai peluang bahwa model memberikan skor lebih tinggi untuk kasus positif daripada negatif secara acak.

5. Keunggulan Kurva ROC

- Threshold-invariant → mengevaluasi performa model pada berbagai threshold, bukan hanya satu titik seperti akurasi
- Class imbalance-friendly → tidak terlalu terpengaruh oleh ketidakseimbangan jumlah data antar kelas

6. Keterbatasan ROC

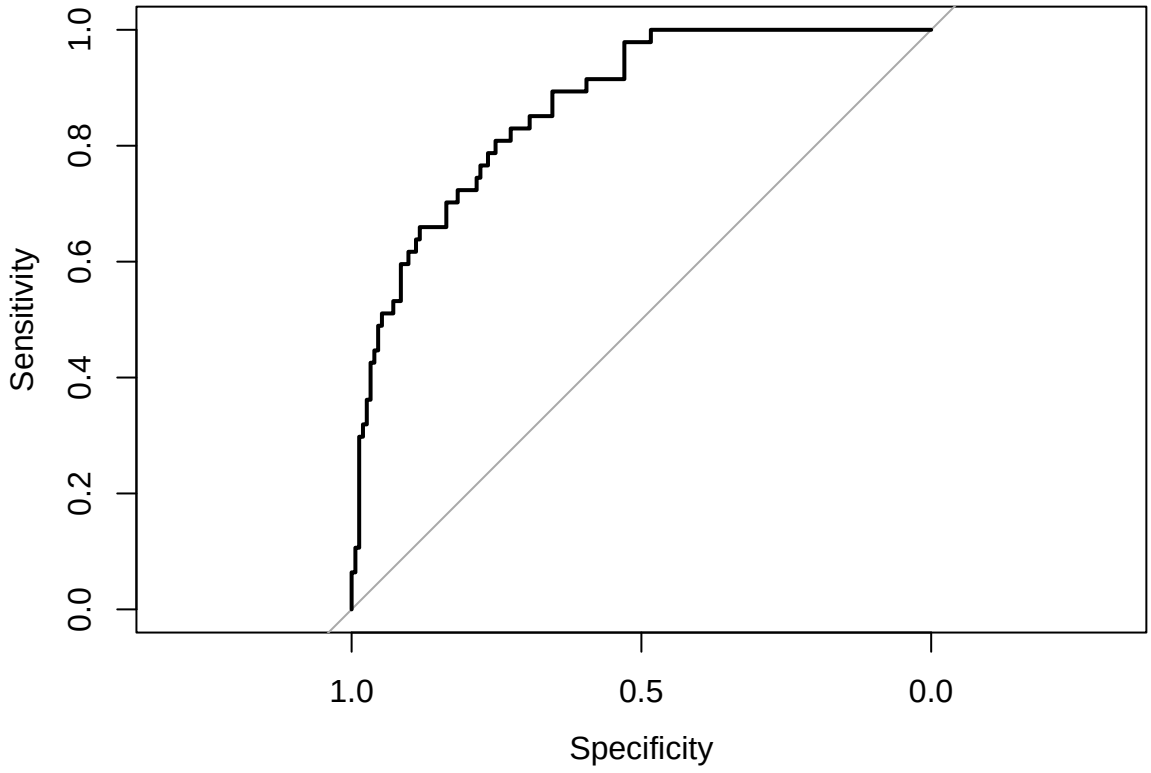
- Dalam kondisi kelas sangat tidak seimbang, ROC bisa menyesatkan
  Contoh: FPR bisa sangat kecil hanya karena jumlah data negatif sangat besar → Gunakan Precision-Recall Curve sebagai pelengkap

```r
library(pROC)
set.seed(123)
x1 <- rnorm(200)
x2 <- rbinom(200, 1, 0.5)
x3 <- rnorm(200)
lin_pred <--1 + 1.5 * x1- 0.7 * x2 + 0.6 * x3
p <- 1 / (1 + exp(-lin_pred))
y <- rbinom(200, 1, p)
data <- data.frame(y = as.factor(y), x1, x2, x3)
model <- glm(y ~ x1 + x2 + x3, data = data, family = binomial)
pred <- predict(model, type = "response")
roc_obj <- roc(data$y, pred)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```r
plot(roc_obj)
```




```
auc(roc_obj)
```

```
## Area under the curve: 0.8686
```

Pemilihan Threshold yang Optimal

Salah satu metode yang digunakan adalah mengevaluasi sensitivitas dan spesifisitas pada berbagai nilai cut-off.
Tujuan: mencari nilai ambang optimal yang memberikan keseimbangan terbaik antara True Positives dan True Negatives

```
thresholds <- seq(0.1, 0.9, by = 0.05)
results <- data.frame(Threshold = thresholds)
results$Sensitivity <- sapply(thresholds, function(t) {
pred_class <- ifelse(pred >= t, 1, 0)
cm <- table(Pred = pred_class, Obs = data$y)
TP <- cm["1", "1"]
FN <- cm["0", "1"]
TP / (TP + FN)
})
results$Specificity <- sapply(thresholds, function(t) {
pred_class <- ifelse(pred >= t, 1, 0)
cm <- table(Pred = pred_class, Obs = data$y)
TN <- cm["0", "0"]
FP <- cm["1", "0"]
```

```
TN / (TN + FP)
})
print(results)
```

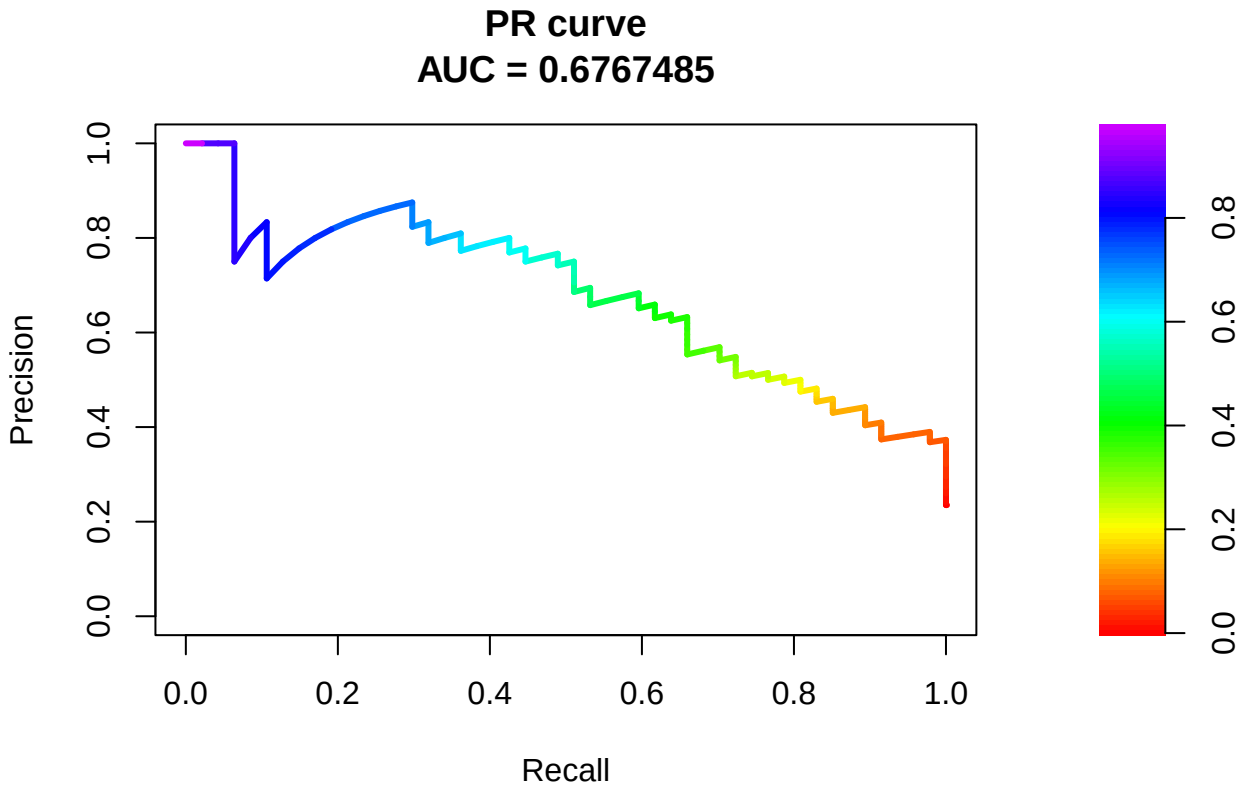
```
##    Threshold Sensitivity Specificity
## 1       0.10  0.91489362   0.5947712
## 2       0.15  0.85106383   0.6862745
## 3       0.20  0.80851064   0.7320261
## 4       0.25  0.76595745   0.7712418
## 5       0.30  0.72340426   0.8104575
## 6       0.35  0.68085106   0.8366013
## 7       0.40  0.61702128   0.8954248
## 8       0.45  0.59574468   0.9150327
## 9       0.50  0.51063830   0.9281046
## 10      0.55  0.51063830   0.9477124
## 11      0.60  0.42553191   0.9607843
## 12      0.65  0.36170213   0.9738562
## 13      0.70  0.29787234   0.9803922
## 14      0.75  0.19148936   0.9869281
## 15      0.80  0.12765957   0.9869281
## 16      0.85  0.06382979   1.0000000
## 17      0.90  0.02127660   1.0000000
```

```
library(PRROC)
```

```
## Warning: package 'PRROC' was built under R version 4.4.3
```

```
## Loading required package: rlang
```

```
set.seed(123)
x1 <- rnorm(200)
x2 <- rbinom(200, 1, 0.5)
x3 <- rnorm(200)
lin_pred <--1 + 1.5 * x1- 0.7 * x2 + 0.6 * x3
p <- 1 / (1 + exp(-lin_pred))
y <- rbinom(200, 1, p)
data <- data.frame(y = y, x1, x2, x3)
model <- glm(y ~ x1 + x2 + x3, data = data, family = binomial)
prob <- predict(model, type = "response")
pr <- pr.curve(scores.class0 = prob[data$y == 1],
scores.class1 = prob[data$y == 0],
curve = TRUE)
plot(pr)
```




```r
set.seed(123)
n <- 300
x1 <- rnorm(n)
x2 <- rbinom(n, 1, 0.5)
x3 <- rnorm(n)
lin_pred <--1 + 1.2 * x1- 0.6 * x2 + 0.8 * x3
p <- 1 / (1 + exp(-lin_pred))
y <- rbinom(n, 1, p)
data <- data.frame(y = as.factor(y), x1, x2, x3)
```

```r
model <- glm(y ~ x1 + x2 + x3, data = data, family = binomial)
model_null <- glm(y ~ 1, data = data, family = binomial)
logL0 <- logLik(model_null)
logLM <- logLik(model)
L0 <- exp(logL0)
LM <- exp(logLM)
n <- nobs(model)
cox_snell <- 1- (L0 / LM)^(2 / n)
mcfadden <- 1- (as.numeric(logLM) / as.numeric(logL0))
r2 <- data.frame(
R2_Cox_Snell = cox_snell,
R2_McFadden = mcfadden
)
r2
```

```
##   R2_Cox_Snell R2_McFadden
```

```
## 1    0.2698462   0.2586292
```

```
if (!require(pscl)) install.packages("pscl"); library(pscl)
```

```
## Loading required package: pscl
```

```
## Warning: package 'pscl' was built under R version 4.4.3
```

```
## Classes and Methods for R originally developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University (2002-2015),
## by and under the direction of Simon Jackman.
## hurdle and zeroinfl functions by Achim Zeileis.
```

```
pR2(model)
```

```
## fitting null model for pseudo-r2
```

```
##          llh      llhNull           G2     McFadden         r2ML         r2CU
## -135.2290328 -182.4040393   94.3500130    0.2586292    0.2698462    0.3835251
```

```
if (!require(rcompanion)) install.packages("rcompanion"); library(rcompanion)
```

```
## Loading required package: rcompanion
```

```
## Warning: package 'rcompanion' was built under R version 4.4.3
```

```
nagelkerke(model)
```

```
## $Models
##
## Model: "glm, y ~ x1 + x2 + x3, binomial, data"
## Null:  "glm, y ~ 1, binomial, data"
##
## $Pseudo.R.squared.for.model.vs.null
##                              Pseudo.R.squared
## McFadden                            0.258629
## Cox and Snell (ML)                  0.269846
## Nagelkerke (Cragg and Uhler)        0.383525
##
## $Likelihood.ratio.test
##  Df.diff LogLik.diff Chisq    p.value
##       -3     -47.175 94.35 2.5468e-20
##
## $Number.of.observations
##
## Model: 300
## Null:  300
##
```

```
## $Messages
## [1] "Note: For models fit with REML, these statistics are based on refitting with ML"
##
## $Warnings
## [1] "None"
```

```r
if (!require(DescTools)) install.packages("DescTools"); library(DescTools)
PseudoR2(model, which = "all")
```

```
##        McFadden     McFaddenAdj        CoxSnell      Nagelkerke   AldrichNelson
##       0.2586292       0.2366998       0.2698462       0.3835251       0.2392545
## VeallZimmermann           Efron McKelveyZavoina            Tjur             AIC
##       0.4360055       0.2893849       0.4315315       0.2936202     278.4580657
##             BIC          logLik         logLik0              G2
##     293.2731956    -135.2290328    -182.4040393      94.3500130
```

3. Evaluasi Model dengan AIC, ROC & AUC, Pseudo R^2, dan Tabel Klasifikasi

1. AIC (Akaike Information Criterion)

- Semakin rendah nilai AIC, semakin baik model dalam hal keseimbangan antara goodness-of-fit dan kompleksitas
- Bandingkan antar model:
  Pilih model dengan AIC terkecil, asal tidak mengorbankan interpretabilitas atau performa klasifikasi

2. ROC & AUC (Area Under the Curve)

- AUC mengukur kemampuan model membedakan antara kelas positif dan negatif

Nilai AUC:

- 0.5 = **tidak lebih baik dari tebak-tebakan**
- 0.7–0.8 = **cukup baik**
- 0.8–0.9 = **baik**
- 0.9 = **sangat baik**

Model dengan AUC tertinggi diutamakan untuk klasifikasi

3. Pseudo R^2 (misalnya McFadden's R^2)

- Mengukur kekuatan prediktif relatif terhadap model null
- Nilai mendekati 1 lebih baik, tetapi dalam praktik:
  R^2 sekitar 0.2–0.4 sudah cukup baik
- Gunakan sebagai pelengkap, bukan satu-satunya acuan

4. Tabel Klasifikasi (Confusion Matrix)

Lihat nilai:

- **Akurasi:** $\frac{TP+TN}{\text{Total}}$
  Sebaiknya tinggi, tapi hati-hati jika data imbalance
- **Sensitivitas (Recall / TPR):**
  Kemampuan model mendeteksi kelas 1 (positif)
- **Spesifisitas:**
  Kemampuan model mendeteksi kelas 0 (negatif)

Jika data imbalance, maka sensitivitas dan spesifisitas lebih informatif daripada akurasi

X. Multinomial and Ordinal Logistic Regression

10.1 Multinomial Logistic Regression

Multinomial Logistic Regression digunakan ketika variabel dependen (Y) bersifat kategori nominal dengan lebih dari dua kategori, dan tidak memiliki urutan logis atau hirarki. Ini merupakan perluasan dari binary logistic regression untuk kasus dengan k > 2 kategori dari variabel respons.

Contoh: Kategori preferensi makanan: {Pizza, Burger, Sushi}. Tidak ada urutan di antara kategori tersebut.

Model ini membandingkan peluang setiap kategori dengan kategori referensi (baseline), menggunakan logit (log odds ratio).

Tujuan

- Memprediksi probabilitas kemunculan kategori tertentu dari Y berdasarkan satu atau lebih prediktor (X).
- Mengukur pengaruh prediktor terhadap peluang relatif terhadap baseline category.

Rumus

Misalkan Y memiliki K kategori ($Y = 1, 2, ..., K$) dan kategori ke-K adalah kategori referensi (baseline), maka:

$\log\left(\frac{P(Y=j|\mathbf{x})}{P(Y=K|\mathbf{x})}\right) = \beta_{j0} + \beta_{j1}x_1 + \cdots + \beta_{jp}x_p, \quad \text{untuk } j = 1, ..., K-1$

Dimana:

- $\mathbf{x}$ adalah vektor prediktor.
- Setiap kategori memiliki satu set koefisien regresi (β_j) relatif terhadap baseline

Distribusi Multinomial

$P(x1 \ = n1 \ , ..., xk \ = nk \ ) = \frac{n!}{n1 \ !n2 \ ! \cdots nk \ !} \ \cdot p1^{n1} \ \ p2^{n2} \ \ \cdots pk^{nk}$

Dengan:

- $n = n_1 + n_2 + \cdots + n_k$

- $\sum_{i=1}^{k} p_i = 1$
- x_i adalah jumlah kejadian dalam kategori ke-i
- p_i adalah probabilitas kategori ke-i

Probabilitas untuk kategori ke-j:

$P(Y = j \mid \mathbf{x}) = \frac{\exp(\eta_j)}{1+\sum_{h=1}^{K-1} \exp(\eta_h)}$ dengan $\eta_j = \beta_{j0} + \sum_{k=1}^{p} \beta_{jk} x_k$

Untuk baseline category K:

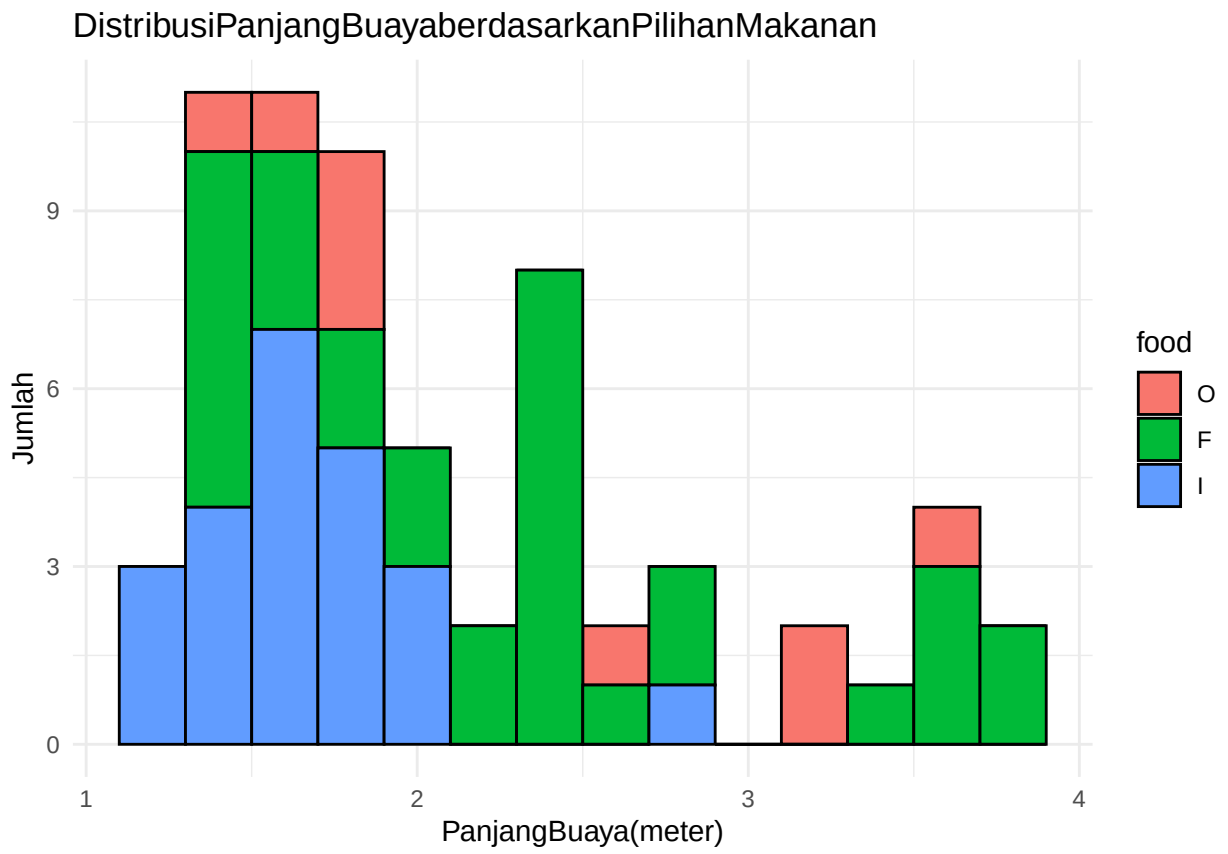
$P(Y = K \mid \mathbf{x}) = \frac{1}{1+\sum_{h=1}^{K-1} \exp(\eta_h)}$

Kasus 1

```
alligator_data <- data.frame(
length = c(1.24,1.30,1.30,1.32,1.32,1.40,1.42,1.42,1.45,1.45,
1.47,1.47,1.47,1.50,1.52,1.55,1.60,1.63,1.63,1.65, 1.65,1.65,1.65,1.68,1.70,1.73,1.73,1.78,1.78,1.78, 1
2.16,2.26,2.31,2.31,2.36,2.36,2.39,2.41,2.44,2.46,
2.56,2.67,2.72,2.79,2.84,3.25,3.28,3.33,3.56,3.58,
3.66,3.68,3.71,3.89),
food = factor(c("I","I","I","F","F","F","I","F","I","O", "I","F","F","I","I","I","I","I","I","O",
"I","F","F","F","I","O","O","I","I","O",
"I","F","F","I","I","I","I","I","F","F",
"F","F","F","F","F","F","F","F","F","F",
"O","F","I","F","F","O","O","F","F","F",
"F","O","F","F"),
levels = c("O", "F", "I")) # response order
)
summary(alligator_data)
```

```
##      length      food
##  Min.   :1.240   O: 9
##  1st Qu.:1.587   F:32
##  Median :1.825   I:23
##  Mean   :2.099
##  3rd Qu.:2.417
##  Max.   :3.890
```

```
ggplot(alligator_data, aes(x= length, fill= food))+
 geom_histogram(binwidth= 0.2, position= "stack",color= "black") +
 labs(title= "DistribusiPanjangBuayaberdasarkanPilihanMakanan",
 x= "PanjangBuaya(meter)", y= "Jumlah")+
theme_minimal()
```




```
#Modelmultinomial(default:baseline=firstlevel,yaituInvertebrates)
library(nnet)
model_mlr <-multinom(food~ length, data= alligator_data)
```

```
## # weights:  9 (4 variable)
## initial  value 70.311186
## iter  10 value 55.228714
## final  value 55.228598
## converged
```

```
summary(model_mlr)
```

```
## Call:
## multinom(formula = food ~ length, data = alligator_data)
##
## Coefficients:
##   (Intercept)      length
## F    1.329763 -0.02619614
## I    5.223790 -2.22400214
##
## Std. Errors:
##   (Intercept)    length
## F    1.227764 0.4988134
## I    1.654445 0.8318572
```

```
##
## Residual Deviance: 110.4572
## AIC: 118.4572
```

```r
z_values <- summary(model_mlr)$coefficients / summary(model_mlr)$standard.errors
p_values <- 2 * (1 - pnorm(abs(z_values)))

coef_table <- cbind(summary(model_mlr)$coefficients,
                    "z value" = round(z_values, 2),
                    "p value" = round(p_values, 4))
```

Pada variabel respons food, terdapat tiga kategori:

- **I = Invertebrata**
- **F = Fish**
- **O = Other**, yang digunakan sebagai kategori referensi (baseline)

Model multinomial logistik kemudian membandingkan masing-masing kategori terhadap baseline O melalui dua fungsi logit berikut:

Logit untuk I terhadap O:

$\log\left(\frac{P(I)}{P(O)}\right) = \beta_0^{(I)} + \beta_1^{(I)} \cdot \text{length}$

Logit untuk F terhadap O:

$\log\left(\frac{P(F)}{P(O)}\right) = \beta_0^{(F)} + \beta_1^{(F)} \cdot \text{length}l$

Untuk setiap kategori selain baseline, model mengestimasi:

- Nilai intercept
- Koefisien prediktor length
- Statistik uji Z dan p-value untuk masing-masing koefisie

Berikut adalah hasil estimasi model multinom(food ~ length):

Koefisien Model Multinomial Logistic Regression

Kategori	Intercept	Length	Z..Intercept.	Z..Length.	p..Intercept.	p..Length.
I	5.22	-2.220	3.16	-2.67	0.0016	0.0075
F	1.33	-0.027	1.08	-0.05	0.2788	0.9581

Interpretasi

Untuk kategori F (Fish), koefisien *length* sebesar -0.027 memiliki p-value sebesar 0.9581. Hal ini mengindikasikan bahwa panjang buaya tidak berpengaruh secara statistik terhadap kemungkinan memilih ikan dibandingkan kategori Other.

Sebaliknya, untuk kategori I (Invertebrata), koefisien *length* adalah -2.22, dengan p-value 0.0075. Ini menunjukkan bahwa peningkatan panjang buaya secara signifikan mengurangi peluang memilih invertebrata dibanding Other.

Interpretasi Odds Ratio

Kategori F (Fish vs Other)

Koefisien length = -0.027

OR = $e^{-0.027} \approx 0.973$

Penjelasan:
Penambahan 1 meter pada panjang buaya hanya menurunkan peluang memilih ikan dibanding Other sebesar sekitar 2.7%. Namun efek ini tidak signifikan secara statistik, sebagaimana tercermin dari p-value yang tinggi.

Kategori I (Invertebrata vs Other)

Koefisien length = -2.22

OR = $e^{-2.22} \approx 0.108$

Penjelasan:
Setiap pertambahan 1 meter pada panjang tubuh buaya mengurangi odds memilih invertebrata dibanding Other sebesar sekitar 89.2%. Efek ini signifikan secara statistik, menunjukkan bahwa panjang tubuh merupakan faktor penting dalam pemilihan invertebrata sebagai makanan.

Ringkasan Interpretasi Odds Ratio

Perbandingan	Koef..Length	Odds.Ratio	Interpretasi	Signifikansi
I vs O (Other)	-2.220	0.109	Semakin panjang buaya, semakin kecil kecenderungannya memilih I	Signifikan (p = 0.0075)
F vs O (Other)	-0.027	0.973	Panjang buaya tidak berpengaruh terhadap peluang memilih F	Tidak signifikan (p = 0.9581)

```
length_seq <- data.frame(length = seq(1.2, 4.0, by = 0.05))

predicted_probs <- predict(model_mlr, newdata = length_seq, type = "probs")

pred_df <- cbind(length_seq, predicted_probs)
pred_long <- tidyr::pivot_longer(pred_df, cols = -length, names_to = "food", values_to = "prob")

ggplot(pred_long, aes(x = length, y = prob, color = food)) +
  geom_line(size = 1.2) +
  labs(title = "Probabilitas Pilihan Makanan berdasarkan Panjang Buaya",
       x = "Panjang Buaya (meter)", y = "Probabilitas") +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

Grafik ini menunjukkan hubungan antara panjang buaya (dalam meter) dan probabilitas pemilihan jenis makanan, berdasarkan model regresi logistik multinomial. Terdapat tiga kategori makanan (food):

- **F (Fish)** – Garis merah
- **I (Invertebrata)** – Garis hijau
- **O (Other)** – Garis biru

Interpretasi Kurva Probabilitas

Fish (F)

- Probabilitas memilih ikan meningkat seiring bertambahnya panjang buaya.
- Untuk buaya pendek (~1.5 meter), probabilitas memilih ikan rendah (~20%).
- Namun, ketika panjang mencapai 3–4 meter, probabilitasnya naik hingga mendekati 80%.
- Ini menunjukkan buaya yang lebih besar cenderung memilih ikan sebagai sumber makanan utama.

Invertebrata (I)

- Probabilitas memilih invertebrata menurun tajam seiring peningkatan panjang.
- Pada buaya kecil, peluang memilih invertebrata sangat tinggi (>70%).
- Namun, semakin besar buaya, peluang ini turun drastis hingga hampir 0%.

- Ini konsisten dengan hasil regresi: koefisien negatif dan signifikan untuk panjang pada kategori I vs O.

Other (O)

- Probabilitas memilih kategori Other meningkat perlahan seiring panjang buaya, tapi tidak sekuat dua kategori lain.
- Peningkatannya kecil dan stabil dari sekitar 5% ke 20%.
- Hal ini menunjukkan bahwa makanan jenis "Other" bukan preferensi utama di semua rentang panjang, tetapi tetap relevan secara proporsional.

Kesimpulan

- Buaya kecil lebih memilih invertebrata, namun seiring pertumbuhan, pilihan makanannya beralih ke ikan.
- Kategori "Other" memiliki probabilitas relatif rendah, tetapi sedikit meningkat untuk buaya yang lebih besar.
- Panjang buaya berperan penting dalam pola preferensi makanannya, khususnya dalam peralihan dari invertebrata ke ikan.

Penghitungan G^2 (Likelihood Ratio Test)

Definisi

- G^2 (*Deviance*) dihitung sebagai:

$$G^2 = -2(logL_{null} - logL_{full})$$

- McFadden's Pseudo R^2 dihitung dengan rumus:

$$R^2 = 1 - \frac{logL_{full}}{logL_{null}}$$

Keterangan:

- L_{null}: log-likelihood dari model null (hanya intercept)
- L_{full}: log-likelihood dari model dengan prediktor
- G^2: mengukur peningkatan kecocokan model penuh dibanding model null
- R^2: memberikan ukuran proporsi informasi yang dijelaskan oleh model

Kita bandingkan model null vs model penuh:

- Model null: hanya intercept, tanpa prediktor
- Model penuh: menggunakan prediktor `length`

```
#Modelnull (tanpaprediktor)
model_null<-multinom(food ~ 1, data= alligator_data, trace= FALSE)
#Modelpenuh(denganprediktorlength)
model_full<-multinom(food ~ length, data= alligator_data, trace= FALSE)
```

```r
#Log-likelihoodmasing-masing
LL_null <-logLik(model_null)
LL_full <-logLik(model_full)
#G² =-2(LL_null-LL_full)
G2<--2* (as.numeric(LL_null)-as.numeric(LL_full)) #Derajatkebebasan=jumlahtambahanparameter
df<-attr(LL_full, "df")-attr(LL_null, "df")
#p-value
p_value <-1-pchisq(G2,df)
#Output
cat("NilaiG²=", round(G2, 4),"\n")
```

```
## NilaiG²= 16.29
```

```r
cat("Derajatkebebasan=",df, "\n")
```

```
## Derajatkebebasan= 2
```

```r
cat("p-value=", round(p_value, 4),"\n")
```

```
## p-value= 3e-04
```

Interpretasi

- Nilai $G^2 = 16.29$ menunjukkan seberapa besar peningkatan kecocokan model penuh (dengan prediktor length) dibandingkan model null (tanpa prediktor).
- Derajat kebebasan (df) = 2 karena ada dua kategori non-referensi (Fish dan Invertebrata) dalam model multinomial (masing-masing dengan satu parameter untuk prediktor length).
- p-value = 0.0003 sangat kecil (« 0.05), sehingga:

Kesimpulan Uji

Kita menolak H_0 yang menyatakan bahwa model penuh tidak lebih baik dari model null.
Artinya:
Panjang buaya (length) secara signifikan meningkatkan kemampuan model dalam memprediksi kategori makanan (food).

Kasus 2

```r
#BuattabelsesuaidenganTabel6.2 Agresti
belief_data <-tribble(
~Race,~Gender, ~Belief, ~Count,
"White", "Female", "Yes", 371,
"White", "Female", "Undecided",49,
"White", "Female", "No", 74,
"White", "Male", "Yes", 250,
"White", "Male", "Undecided",45,
"White", "Male", "No", 71,
"Black", "Female", "Yes", 64,
"Black", "Female", "Undecided",9,
"Black", "Female", "No", 15,
```

```
"Black", "Male", "Yes", 25,
"Black", "Male", "Undecided",5,
"Black", "Male", "No", 13
)
#Ubahkeformatfaktor
belief_data <-belief_data%>%
mutate(
Belief= factor(Belief, levels= c("Yes", "Undecided", "No")),
Gender= factor(Gender),
Race= factor(Race)
)
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.4.3
```

```
# Perluas data sesuai Count
expanded_data <- belief_data %>%
uncount(weights = Count)
head(expanded_data)
```

```
## # A tibble: 6 x 3
##   Race  Gender Belief
##   <fct> <fct>  <fct>
## 1 White Female Yes
## 2 White Female Yes
## 3 White Female Yes
## 4 White Female Yes
## 5 White Female Yes
## 6 White Female Yes
```

```
# Model multinomial: Belief sebagai response
model_mlr <- multinom(Belief ~ Race + Gender, data = expanded_data, trace = FALSE)
summary(model_mlr)
```

```
## Call:
## multinom(formula = Belief ~ Race + Gender, data = expanded_data,
##     trace = FALSE)
##
## Coefficients:
##           (Intercept)   RaceWhite GenderMale
## Undecided   -1.954553 -0.07078817  0.3134797
## No          -1.301607 -0.34176952  0.4185511
##
## Std. Errors:
##           (Intercept) RaceWhite GenderMale
## Undecided   0.2974392 0.3091936  0.2083285
## No          0.2264955 0.2370377  0.1712550
##
## Residual Deviance: 1547.453
## AIC: 1559.453
```

Hasil Estimasi Koefisien

Response	Intercept	RaceWhite	GenderMale
Undecided	-1.955	-0.071	0.313
No	-1.302	-0.342	0.419

Response	Std. Error (Intercept)	Std. Error (RaceWhite)	Std. Error (GenderMale)
Undecided	0.297	0.309	0.208
No	0.226	0.237	0.171

Interpretasi Koefisien

1. Kategori *Undecided* (vs *Yes*)

- Intercept = -1.955, Untuk *Black female* (kategori referensi), kemungkinan memilih *Undecided* dibanding *Yes* cukup rendah.
- RaceWhite = -0.071, Responden kulit putih sedikit kurang mungkin ragu-ragu (*Undecided*) dibandingkan responden kulit hitam, meskipun efeknya sangat kecil.
- GenderMale = 0.313, Laki-laki lebih besar kemungkinan menjawab *Undecided* dibanding *Yes* daripada perempuan.

2. Kategori *No* (vs *Yes*)

- Intercept = -1.302 → Untuk *Black female*, kemungkinan menjawab *No* (tidak percaya afterlife) dibanding *Yes* (percaya) juga relatif rendah.
- RaceWhite = -0.342 → Responden kulit putih sedikit kurang cenderung tidak percaya afterlife dibanding responden kulit hitam.
- GenderMale = 0.419 → Laki-laki cenderung lebih mungkin tidak percaya afterlife dibanding perempuan.

Model Fit

- Residual Deviance: 1547.453
- AIC: 1559.453

Nilai-nilai ini memberi ukuran seberapa baik model menjelaskan data. Dapat digunakan untuk membandingkan dengan model lain.

```
# Z dan p-value
z_values <- summary(model_mlr)$coefficients / summary(model_mlr)$standard.errors
p_values <- 2 * (1- pnorm(abs(z_values)))
coef_table <- cbind(summary(model_mlr)$coefficients,
"z value" = round(z_values, 2),
"p value" = round(p_values, 4))
knitr::kable(coef_table, caption = "Koefisien Model Multinomial Logistic Regression")
```

Table 30: Koefisien Model Multinomial Logistic Regression

	(Intercept)	RaceWhite	GenderMale	(Intercept)	RaceWhite	GenderMale	(Intercept)	RaceWhite	GenderMale
Undecided	-1.954553	-0.0707882	0.3134797	-6.57	-0.23	1.50	0	0.8189	0.1324
No	-1.301607	-0.3417695	0.4185511	-5.75	-1.44	2.44	0	0.1493	0.0145

Interpretasi Per Koefisien

Kategori *Undecided* (dibanding *Yes*)

- Intercept signifikan ($p < 0.0001$) → kelompok referensi (Black Female) memiliki kemungkinan yang jauh lebih rendah untuk menjawab *Undecided* daripada *Yes*.
- RaceWhite tidak signifikan ($p = 0.8189$) → tidak ada perbedaan yang berarti antara responden kulit putih dan kulit hitam dalam menjawab *Undecided*.
- GenderMale tidak signifikan ($p = 0.1324$) → tidak ada bukti kuat bahwa laki-laki lebih cenderung ragu (*Undecided*) dibanding perempuan.

Kategori *No* (dibanding *Yes*)

- Intercept signifikan ($p < 0.0001$) → kelompok referensi (Black Female) memiliki kemungkinan rendah menjawab *No* dibanding *Yes*.
- RaceWhite tidak signifikan ($p = 0.1493$) → perbedaan antara kulit putih dan kulit hitam dalam kemungkinan menjawab *No* tidak cukup kuat secara statistik.
- GenderMale signifikan ($p = 0.0145$) → laki-laki secara signifikan lebih mungkin untuk tidak percaya afterlife dibanding perempuan.

Ringkasan Temuan

Variabel	Undecided vs Yes	No vs Yes	Signifikansi
Intercept	Signifikan	Signifikan	Menunjukkan baseline (Black Female) rendah pada *Undecided* dan *No*
RaceWhite	Tidak signifikan	Tidak signifikan	Tidak ada pengaruh kuat dari ras
GenderMale	Tidak signifikan	Signifikan	Laki-laki lebih cenderung menjawab *No*

Kesimpulan

- Jenis kelamin memengaruhi kepercayaan pada afterlife: laki-laki lebih cenderung tidak percaya dibanding perempuan.
- Ras tidak memberikan efek signifikan terhadap kepercayaan afterlife dalam model ini.
- Kategori referensi (Black Female) adalah kelompok yang paling besar kemungkinan percaya pada afterlife.

```r
#Prediksi probabilitas
new_data<-expand.grid(
Race= levels(expanded_data$Race),
Gender= levels(expanded_data$Gender)
)
new_data$prob <-predict(model_mlr, newdata= new_data, type= "probs")
#Gabungkanhasil
cbind(new_data,predict(model_mlr, newdata= new_data, type= "probs"))
```

```
##    Race Gender   prob.Yes prob.Undecided    prob.No       Yes  Undecided
## 1 Black Female 0.70735270     0.10018080 0.19246651 0.7073527 0.10018080
## 2 White Female 0.75456040     0.09956337 0.14587623 0.7545604 0.09956337
## 3 Black   Male 0.62216559     0.12055825 0.25727616 0.6221656 0.12055825
## 4 White   Male 0.67827036     0.12244777 0.19928187 0.6782704 0.12244777
##          No
## 1 0.1924665
## 2 0.1458762
## 3 0.2572762
## 4 0.1992819
```

```r
model_mlr <-multinom(Belief ~ Race * Gender, data= expanded_data)
```

```
## # weights:  15 (8 variable)
## initial  value 1088.724778
## iter  10 value 774.603183
## final  value 773.299583
## converged
```

Tabel berikut menunjukkan probabilitas responden dari tiap kombinasi ras dan jenis kelamin untuk menjawab *Yes*, *Undecided*, atau *No* terhadap keyakinan akan afterlife:

Ras	Gender	Prob. Yes	Prob. Undecided	Prob. No
Black	Female	0.7074	0.1002	0.1925
White	Female	0.7546	0.0996	0.1459
Black	Male	0.6222	0.1205	0.2573
White	Male	0.6783	0.1225	0.1993

Interpretasi

- Probabilitas terbesar untuk semua kelompok adalah menjawab *Yes* — menunjukkan bahwa mayoritas percaya pada afterlife, terlepas dari ras dan gender.
- Perempuan (baik Black maupun White) cenderung memiliki probabilitas lebih tinggi untuk menjawab *Yes* dibanding laki-laki.
- Laki-laki Black memiliki probabilitas tertinggi untuk menjawab *No* (25.7%), disusul White Male (19.9%)

```r
#Modelnull:hanyaintercept
model_null<-multinom(Belief ~1, data= expanded_data, trace= FALSE)
#Log-likelihood
ll_null <-logLik(model_null)
ll_full <-logLik(model_mlr)
#G2 (deviance)
G2<--2* (as.numeric(ll_null)-as.numeric(ll_full)) #Derajatkebebasan
df<-attr(ll_full, "df")-attr(ll_null, "df")
```

```r
#p-value
pval <-1-pchisq(G2,df)
#McFadden's pseudoR2
pseudo_r2 <-1-(as.numeric(ll_full) /as.numeric(ll_null))
#Tampilkanhasil
cat("NilaiG²(Deviance):", round(G2, 4),"\n")
```

```
## NilaiG²(Deviance): 9.5975
```

```r
cat("DerajatKebebasan:",df, "\n")
```

```
## DerajatKebebasan: 6
```

```r
cat("p-value:", round(pval, 4),"\n")
```

```
## p-value: 0.1427
```

```r
cat("McFadden'sPseudoR²:", round(pseudo_r2, 4),"\n")
```

```
## McFadden'sPseudoR²: 0.0062
```

Hasil Uji Likelihood Ratio Test

- G^2 (Deviance) = 9.5975
- Derajat kebebasan (df) = 6
- p-value = 0.1427
- McFadden's Pseudo R^2 = 0.0062

Interpretasi

- G^2 (Deviance) = 9.5975 mengukur sejauh mana model penuh memberikan peningkatan kecocokan terhadap data dibandingkan model null.
- Derajat kebebasan (df) = 6 menunjukkan bahwa 6 parameter tambahan dipertimbangkan dalam model penuh dibanding model null.
- p-value = 0.1427 > 0.05, Artinya, tidak terdapat bukti yang cukup secara statistik untuk menyimpulkan bahwa model penuh secara signifikan lebih baik dibanding model null. Dengan kata lain, penambahan prediktor Race * Gender tidak memberikan peningkatan kecocokan model yang signifikan.

- McFadden's Pseudo $R^2 = 0.0062 \rightarrow$ Hanya sekitar 0.62% variasi dalam variabel Belief yang dapat dijelaskan oleh model penuh. Ini menunjukkan bahwa kemampuan prediksi model sangat renda**h**.

Kesimpulan

Model penuh yang memasukkan interaksi antara ras dan jenis kelamin (Race * Gender) tidak secara signifikan lebih baik dibandingkan model null (hanya intersep) dalam menjelaskan variasi keyakinan terhadap afterlife. Dengan demikian, tidak ada bukti kuat bahwa kombinasi ras dan jenis kelamin memengaruhi keyakinan seseorang pada afterlife secara statistik.

10.2 Ordinal Logistic Regression

Regresi Logistik Ordinal adalah metode regresi yang digunakan ketika variabel respons (Y) bersifat kategorik ordinal, yaitu memiliki urutan (ranking) tetapi jarak antar kategori tidak harus sama.

Contoh:

- Skala kepuasan: Sangat Tidak Puas, Tidak Puas, Netral, Puas, Sangat Puas
- Tingkat pendidikan: SD, SMP, SMA, S1, S2, S3

Dalam banyak kasus, prediksi atau pemodelan untuk variabel ordinal tidak tepat jika menggunakan regresi logistik multinomial, karena pendekatan tersebut mengabaikan struktur urutan kategori. Oleh karena itu, digunakanlah model logistik ordinal, yang mempertahankan struktur ordinal pada data.

Model Proporsional Odds (Cumulative Logit Model)

Model yang paling umum digunakan dalam regresi logistik ordinal adalah model proportional odds atau model kumulatif logit (cumulative logit model).

Bentuk Umum Model:

$\log\left(\frac{P(Y\leq j)}{P(Y>j)}\right) = \theta_j - \beta_1 X_1 - \beta_2 X_2 - \cdots - \beta_p X_p$

dengan $j = 1, 2, \ldots, J-1$, untuk J kategori ordinal.

Penjelasan Setiap Komponen:

Komponen	Penjelasan
Y	Variabel respons ordinal dengan J kategori berurutan.
$P(Y \leq j)$	Probabilitas kumulatif bahwa respon berada pada kategori ke-j atau lebih rendah.
$P(Y > j)$	Probabilitas bahwa respon berada di atas kategori ke-j.
θ_j	Intercept khusus kategori (cutpoint) untuk kategori ke-j. Menyatakan nilai ambang antara kategori j dan $j+1$.
β_k	Koefisien regresi untuk prediktor X_k. Sama untuk semua j dalam model proportional odds.
X_k	Prediktor/variabel penjelas.

Catatan Penting: Proportional Odds Assumption

Asumsi utama dari model ini adalah:

Efek prediktor terhadap logit kumulatif adalah konstan di seluruh ambang batas kategori.

Ini disebut proportional odds assumption, yaitu:

$\beta^{(1)} = \beta^{(2)} = \cdots = \beta^{(J-1)} = \beta$

Dengan kata lain, hanya intersep (cutpoint) yang berubah untuk setiap kategori, bukan koefisien regresor.

Interpretasi Model

Dari model:

$\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \theta_j - \beta_1 X_1 - \cdots - \beta_p X_p log(P(Y > j)$

diperoleh:

$\frac{P(Y \leq j)}{P(Y > j)} = \exp(\theta_j - \beta_1 X_1 - \cdots - \beta_p X_p)$

Artinya:

- Jika $\beta_k > 0$, maka semakin besar X_k, probabilitas berada pada kategori lebih rendah (lebih kecil dari atau sama dengan j) menjadi lebih kecil, sehingga kemungkinan berada pada kategori lebih tinggi meningkat.
- e^{β_k} adalah odds ratio kumulatif: perubahan odds kumulatif untuk setiap kenaikan satu unit dalam X_k dengan asumsi proporsionalitas.

Estimasi Parameter

Parameter θ_j dan β_k dalam model logistik ordinal diestimasi menggunakan maximum likelihood estimation (MLE). Fungsi likelihood untuk n observasi adalah:

$L(\beta, \theta) = \prod_{i=1}^{n} P(Y_i = j_i \mid X_i)$

yang dihitung dari perbedaan probabilitas kumulatif:

$P(Y_i = j) = P(Y_i \leq j) - P(Y_i \leq j - 1)$

Pengujian Model

a. Pengujian Signifikansi Koefisien:

- **Wald Test:** Menguji $H_0 : \beta_k = 0$
- **Likelihood Ratio Test:** Bandingkan model penuh dengan model tanpa prediktor

b. Uji Asumsi Proportional Odds:

- **Brant Test** digunakan untuk menguji apakah koefisien prediktor konsisten di seluruh kategori.

Alternatif jika Asumsi Proportional Odds Tidak Terpenuhi

Jika asumsi proportional odds tidak valid, alternatif model meliputi:

- Partial Proportional Odds Model: beberapa prediktor mengikuti asumsi, beberapa tidak.
- Non-Proportional Odds Model: membolehkan semua k _k k berubah untuk setiap kategori.

Kasus

```r
library(MASS) #Untukpolr
library(dplyr)
library(tidyr)
library(ggplot2)
# Data dari Tabel 6.3 Agresti
ideology_data <- tribble(
  ~Gender, ~Party, ~Ideology, ~Count,
  "Female", "Democrat",    "Very Liberal",       25,
  "Female", "Democrat",    "Slightly Liberal",  105,
  "Female", "Democrat",    "Moderate",           86,
  "Female", "Democrat",    "Slightly Conservative", 28,
  "Female", "Democrat",    "Very Conservative",   4,
  "Female", "Republican",  "Very Liberal",        0,
  "Female", "Republican",  "Slightly Liberal",    5,
  "Female", "Republican",  "Moderate",           15,
  "Female", "Republican",  "Slightly Conservative", 83,
  "Female", "Republican",  "Very Conservative",  32,
  "Male",   "Democrat",    "Very Liberal",       20,
  "Male",   "Democrat",    "Slightly Liberal",   73,
  "Male",   "Democrat",    "Moderate",           43,
  "Male",   "Democrat",    "Slightly Conservative", 20,
  "Male",   "Democrat",    "Very Conservative",   3,
  "Male",   "Republican",  "Very Liberal",        0,
  "Male",   "Republican",  "Slightly Liberal",    1,
  "Male",   "Republican",  "Moderate",           14,
  "Male",   "Republican",  "Slightly Conservative", 72,
  "Male",   "Republican",  "Very Conservative",  32
)


# Pastikan urutan faktor ordinal sesuai spektrum ideologi
ideology_data <- ideology_data %>%
  mutate(
    Ideology = factor(Ideology,
                      levels = c("Very Liberal", "Slightly Liberal", "Moderate",
                                 "Slightly Conservative", "Very Conservative"),
                      ordered = TRUE),
    Gender = factor(Gender),
    Party = factor(Party)
  )
```

```r
# Perluas berdasarkan Count
expanded_data <- uncount(ideology_data, weights = Count)
head(expanded_data)
```

```
## # A tibble: 6 x 3
##   Gender Party    Ideology
##   <fct>  <fct>    <ord>
## 1 Female Democrat Very Liberal
## 2 Female Democrat Very Liberal
## 3 Female Democrat Very Liberal
## 4 Female Democrat Very Liberal
## 5 Female Democrat Very Liberal
```

```
## 6 Female Democrat Very Liberal
```

```
library(MASS)
model_ordinal <- polr(Ideology ~ Gender + Party, data = expanded_data, Hess = TRUE)
summary(model_ordinal)
```

```
## Call:
## polr(formula = Ideology ~ Gender + Party, data = expanded_data,
##     Hess = TRUE)
##
## Coefficients:
##                     Value Std. Error t value
## GenderMale      -0.04731     0.1499 -0.3157
## PartyRepublican  3.63365     0.2175 16.7037
##
## Intercepts:
##                                          Value    Std. Error t value
## Very Liberal|Slightly Liberal            -2.1223   0.1692   -12.5461
## Slightly Liberal|Moderate                 0.1689   0.1143     1.4778
## Moderate|Slightly Conservative            1.8572   0.1511    12.2891
## Slightly Conservative|Very Conservative   4.6500   0.2343    19.8456
##
## Residual Deviance: 1555.786
## AIC: 1567.786
```

Koefisien Regresi

Variabel	Koefisien	Std. Error	z-value
GenderMale	-0.04731	0.1499	-0.316
PartyRepublican	3.63365	0.2175	16.704

Interpretasi Koefisien:

GenderMale = -0.04731

- Efek jenis kelamin tidak signifikan secara statistik (z = -0.316, jauh dari ± 1.96).
- Artinya, tidak ada cukup bukti bahwa laki-laki berbeda secara signifikan dari perempuan dalam kecenderungan ideologi politik setelah mengendalikan afiliasi partai.

PartyRepublican = 3.63365

- Sangat signifikan secara statistik (z = 16.704 > 1.96).
- Karena koefisien positif, ini menunjukkan bahwa menjadi Republican meningkatkan kecenderungan berada di kategori ideologi yang lebih konservatif dibandingkan dengan Democrat.
- **Interpretasi Odds Ratio:**

  $\exp(3.63365) \approx 37.84$

  Artinya, odds kumulatif seorang Republican untuk memiliki ideologi lebih konservatif (j) sekitar 38 kali lebih besar dibandingkan seorang Democrat, dengan gender tetap.

Intersep (Thresholds / Cutpoints)

Cutpoint	Estimate	Interpretation
VeryLiberal | SlightlyLiberal	-2.1223	Ambang batas antara kategori 1 dan 2
SlightlyLiberal | Moderate	0.1689	Ambang batas antara kategori 2 dan 3
Moderate | SlightlyConservative	1.8572	Ambang batas antara kategori 3 dan 4
SlightlyConservative | VeryConservative	4.6500	Ambang batas antara kategori 4 dan 5

Cutpoints ini berfungsi sebagai parameter threshold θ_j dalam model kumulatif logit.

Model yang digunakan adalah:

$\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \theta_j - \beta_1 X_1 - \beta_2 X_2$

Dengan $\theta\theta_j$ adalah cutpoint, dan β koefisien prediktor.

Statistik Model

- **Residual Deviance**: 1555.786
- **AIC**: 1567.786

Angka ini berguna untuk membandingkan model; semakin kecil AIC, semakin baik model.

```
# Hitung z dan p-value
coefs <- coef(summary(model_ordinal))
p_values <- pnorm(abs(coefs[, "t value"]), lower.tail = FALSE) * 2
cbind(coefs, "p value" = round(p_values, 4)) %>%
  knitr::kable(caption = "Koefisien dan p-value dari Model Regresi Logistik Ordinal")
```

Table 36: Koefisien dan p-value dari Model Regresi Logistik Ordinal

	Value	Std. Error	t value	p value
GenderMale	-0.0473126	0.1498822	-0.315665	0.7523
PartyRepublican	3.6336493	0.2175355	16.703705	0.0000
Very Liberal|Slightly Liberal	-2.1223326	0.1691632	-12.546067	0.0000
Slightly Liberal|Moderate	0.1689125	0.1143027	1.477764	0.1395
Moderate|Slightly Conservative	1.8571511	0.1511213	12.289140	0.0000
Slightly Conservative|Very Conservative	4.6500442	0.2343111	19.845599	0.0000

Interpretasi:

Efek Prediktor

PartyRepublican:

- Koefisien positif (3.63) dan p-value < 0.0001, signifikan secara statistik.

- Menjadi Republican secara drastis meningkatkan peluang berada dalam kategori ideologi yang lebih konservatif.
- **Odds Ratio**: exp(3.6336) ≈ 37.84
  Republican memiliki 38 kali lebih besar odds kumulatif untuk kategori ideologi lebih konservatif daripada Democrat.

GenderMale:

- Koefisien: -0.0473 (sangat kecil)
- p-value: 0.7523 (» 0.05)
- Tidak signifikan, tidak ada bukti statistik bahwa gender (male vs female) berdampak pada tingkat ideologi politik setelah memperhitungkan afiliasi partai.

Intersep / Thresholds (Cutpoints)

VeryLiberal | SlightlyLiberal: $p < 0.0001$

- **Sangat signifikan**, ada perbedaan tegas antara kategori "Very Liberal" dan "Slightly Liberal".

SlightlyLiberal | Moderate: $p = 0.1395$

- Tidak signifikan secara statistik, transisi dari "Slightly Liberal" ke "Moderate" tidak terlalu tajam dibanding batas lainnya.

Dua ambang lainnya (Moderate | SlightlyConservative, SlightlyConservative | VeryConservative) signifikan ($p < 0.0001$), menunjukkan transisi antar kategori cukup jelas secara statistik.

Kesimpulan

- **Variabel Partai (Republican)** adalah prediktor utama dalam menjelaskan kecenderungan ideologi politik.
- **Jenis kelamin** tidak signifikan, perbedaan ideologi antara laki-laki dan perempuan **tidak cukup kuat secara statistik**.
- **Model valid secara statistik**, tetapi ada satu ambang batas (Slightly Liberal → Moderate) yang perlu ditinjau lebih lanjut karena tidak signifikan.

```
new_data<-expand.grid(Gender= levels(expanded_data$Gender),
                      Party= levels(expanded_data$Party))
 #Prediksi probabilitas
predict_probs<-predict(model_ordinal, newdata = new_data, type= "probs")
cbind(new_data,predict_probs)
```

```
##   Gender      Party Very Liberal Slightly Liberal  Moderate
## 1 Female   Democrat  0.106945088       0.43518292 0.3228365
## 2   Male   Democrat  0.111548559       0.44229808 0.3165493
## 3 Female Republican  0.003153821       0.02717858 0.1144037
## 4   Male Republican  0.003306117       0.02844921 0.1189364
##   Slightly Conservative Very Conservative
## 1             0.1255648       0.009470629
## 2             0.1205672       0.009036939
## 3             0.5895337       0.265730233
## 4             0.5927066       0.256601598
```

Tabel Prediksi Probabilitas Ideologi Politik

Gender	Party	Very Liberal	Slightly Liberal	Moderate	Slightly Conservative	Very Conservative
Female	Democrat	0.107	0.435	0.323	0.126	0.009
Male	Democrat	0.112	0.442	0.317	0.121	0.009
Female	Republican	0.003	0.027	0.114	0.590	0.266
Male	Republican	0.003	0.028	0.119	0.593	0.257

Interpretasi Per Kelompok

Democrat (Female dan Male)

- Probabilitas tertinggi ada di kategori Slightly Liberal (43–44%), disusul Moderate (32%).
- Hanya sekitar 10–11% berada di kategori Very Liberal.
- Hanya ~1% memiliki probabilitas Very Conservative.
- Artinya, baik laki-laki maupun perempuan dari Partai Demokrat cenderung berada di kiri-tengah spektrum ideologi (Liberal ke Moderate).

Perbedaan Gender di Democrat:

- Negligible (sangat kecil): Gender tidak memberi perubahan berarti dalam distribusi probabilitas ideologi dalam kelompok Democrat (sesuai hasil model: Gender tidak signifikan).

Republican (Female dan Male)

- Probabilitas tertinggi ada di kategori Slightly Conservative (~59%), lalu diikuti Very Conservative (~26%).
- Sangat kecil kemungkinan menjadi Very Liberal atau Slightly Liberal (<3%).
- Artinya, anggota Partai Republican, baik pria maupun wanita, sangat cenderung konservatif.

Perbedaan Gender di Republican:

- Hampir tidak ada perbedaan: Probabilitas antar kategori ideologi **sangat mirip** antara perempuan dan laki-laki Republican.
- Ini mendukung koefisien Gender yang **tidak signifikan secara statistik** dalam model.

Kesimpulan

- **Afiliasi Partai** memiliki pengaruh sangat besar terhadap spektrum ideologi:
  - **Democrat**: mayoritas Liberal hingga Moderate.
  - **Republican**: mayoritas Slightly hingga Very Conservative
- **Jenis kelamin (Gender)**:

- `Hampir tidak berpengaruh dalam setiap kelompok partai.`

- `Hal ini konsisten dengan hasil model sebelumnya (p-value Gender = 0.7523, tidak signifikan).`

- **Model menghasilkan probabilitas** yang **secara substantif dan statistik** masuk akal dan mencerminkan pola umum dalam preferensi ideologi.

XI. Loglinear Model

Model log-linear merupakan metode statistik yang digunakan untuk menganalisis hubungan antar variabel kategorik dalam tabel kontingensi. Tidak seperti regresi linear yang memodelkan hubungan antara variabel numerik, model log-linear memodelkan logaritma dari frekuensi harapan (expected cell counts) sebagai fungsi linear dari efek utama dan interaksi antar variabel. Model ini mengasumsikan bahwa frekuensi dalam setiap sel tabel mengikuti distribusi Poisson.

Berbeda dengan regresi logistik yang memodelkan probabilitas kejadian berdasarkan variabel prediktor, model log-linear memperlakukan semua variabel secara setara tanpa membedakan antara variabel respon dan prediktor. Tujuan utamanya adalah mengevaluasi struktur asosiasi di antara variabel-variabel tersebut.

1. Model Log-Linear Dua Arah

Dalam tabel kontingensi dua dimensi, misalnya variabel X dengan I kategori dan Y dengan J kategori, frekuensi aktual dicatat sebagai n_{ij}, dan frekuensi harapan sebagai μ_{ij}. Model log-linear untuk dua variabel dapat diklasifikasikan sebagai berikut:

1. **Model Saturated**: Mencakup semua efek utama dan interaksi. Model ini selalu cocok sempurna dengan data karena tidak mengurangi kompleksitas, tetapi kurang berguna untuk interpretasi.
   $\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$
2. **Model Independence**: Hanya mencakup efek utama tanpa interaksi. Digunakan untuk menguji apakah variabel bebas satu sama lain.
   $\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$

Kasus 1

```
tabel <- matrix(c(30,20,10,40),nrow=2,byrow=TRUE)
colnames(tabel) <- c("Sakit","Sehat")
rownames(tabel) <- c("Ya","Tidak")
tabel
```

```
##       Sakit Sehat
## Ya       30    20
## Tidak    10    40
```

```
data <- as.data.frame(as.table(tabel))
colnames(data) <- c("Merokok", "Status", "Freq")
data
```

```
##   Merokok Status Freq
## 1      Ya  Sakit   30
## 2   Tidak  Sakit   10
## 3      Ya  Sehat   20
## 4   Tidak  Sehat   40
```

```
# Model tanpa interaksi
fit_no_inter <- glm(Freq ~ Merokok + Status, family = poisson, data = data)
summary(fit_no_inter)
```

```
##
## Call:
## glm(formula = Freq ~ Merokok + Status, family = poisson, data = data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.996e+00  1.871e-01  16.013   <2e-16 ***
## MerokokTidak 3.892e-10  2.000e-01   0.000    1.000
## StatusSehat  4.055e-01  2.041e-01   1.986    0.047 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 21.288  on 3  degrees of freedom
## Residual deviance: 17.261  on 1  degrees of freedom
## AIC: 43.036
##
## Number of Fisher Scoring iterations: 4
```

Model Log-Linear Tanpa Interaksi

Model:

$\log(\mu_{ij}) = \lambda + \lambda_i^{\text{Merokok}} + \lambda_j^{\text{Status}}$

Koefisien	Estimate	Std. Error	z-value	p-value
(Intercept)	2.996	0.1871	16.013	<0.0001
MerokokTidak	~0	0.2000	0.000	1.0000
StatusSehat	0.4055	0.2041	1.986	0.0470

Interpretasi:

- Model ini tidak mempertimbangkan adanya interaksi antara status merokok dan status kesehatan.
- Intercept (2.996): log dari expected count untuk kategori referensi (perokok-sakit).
- MerokokTidak (~0): menunjukkan bahwa status tidak merokok tidak signifikan berpengaruh terhadap frekuensi, setelah dikontrol oleh status.
- StatusSehat (0.4055): signifikan di 5% (p = 0.047), berarti individu sehat cenderung memiliki expected count lebih tinggi dari individu sakit, tanpa memperhitungkan status merokok.

Goodness-of-Fit:

- Residual deviance = 17.261 (df = 1) → cukup besar untuk 1 derajat bebas → menunjukkan kurang cocok (ada interaksi yang mungkin terabaikan).
- AIC = 43.036

```r
# Model dengan interaksi
fit_inter <- glm(Freq ~ Merokok * Status, family = poisson, data = data)
summary(fit_inter)
```

```
##
## Call:
## glm(formula = Freq ~ Merokok * Status, family = poisson, data = data)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                3.4012     0.1826  18.629  < 2e-16 ***
## MerokokTidak              -1.0986     0.3651  -3.009  0.00262 **
## StatusSehat               -0.4055     0.2887  -1.405  0.16015
## MerokokTidak:StatusSehat   1.7918     0.4564   3.926 8.65e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2.1288e+01  on 3  degrees of freedom
## Residual deviance: 3.9968e-15  on 0  degrees of freedom
## AIC: 27.775
##
## Number of Fisher Scoring iterations: 3
```

Model Log-Linear Dengan Interaksi

Model:

$\log(\mu_{ij}) = \lambda + \lambda_i^{\text{Merokok}} + \lambda_j^{\text{Status}} + \lambda_{ij}^{\text{Merokok}\times\text{Status}}$

Koefisien	Estimate	Std. Error	z-value	p-value
(Intercept)	3.4012	0.1826	18.629	<0.0001
MerokokTidak	-1.0986	0.3651	-3.009	0.0026
StatusSehat	-0.4055	0.2887	-1.405	0.1602
MerokokTidak:StatusSehat	1.7918	0.4564	3.926	<0.0001

Interpretasi:

- Sekarang kita memodelkan interaksi langsung antara merokok dan status.
- Interaksi signifikan: koefisien 1.7918 ($p < 0.0001$) → hubungan antara status merokok dan status kesehatan tidak independen, tetapi saling memengaruhi.
- Main effect StatusSehat tidak signifikan sendiri, tetapi karena interaksi signifikan, efek gabungannya tetap penting.
- Koefisien MerokokTidak = -1.0986 dan interaksi MerokokTidak:StatusSehat = 1.7918 → menunjukkan bahwa efek tidak merokok berubah tergantung apakah sehat atau tidak.

Perbandingan Model

Kriteria	Model Tanpa Interaksi	Model Dengan Interaksi
Residual Deviance	17.261 (df = 1)	0 (3.9968e-15) (df = 0)
AIC	43.036	27.775
Interaksi signifikan?	Tidak	Ya ($p < 0.001$)

Kesimpulan:

- Model dengan interaksi jauh lebih baik:
  - Deviansi residu sangat kecil dan df = 0 → cocok sempurna.
  - AIC jauh lebih kecil, menunjukkan kualitas model yang lebih baik.
- Model tanpa interaksi kurang mampu menangkap hubungan nyata antara variabel Merokok dan Status.
- Interaksi penting: artinya hubungan antara status kesehatan tergantung pada status merokok, dan sebaliknya.

Kasus 2

```r
# Membuat data frame dari tabel
tabel2x3 <- matrix(c(12, 20, 8, 18, 24, 10), nrow = 2, byrow = TRUE)
colnames(tabel2x3) <- c("Kurus", "Normal", "Gemuk")
rownames(tabel2x3) <- c("Laki-laki", "Perempuan")
tabel2x3
```

```
##           Kurus Normal Gemuk
## Laki-laki    12     20     8
## Perempuan    18     24    10
```

```r
# Ubah menjadi data.frame untuk glm
data2x3 <- as.data.frame(as.table(tabel2x3))
colnames(data2x3) <- c("JenisKelamin", "BMI", "Freq")
data2x3
```

```
##   JenisKelamin    BMI Freq
## 1    Laki-laki  Kurus   12
## 2    Perempuan  Kurus   18
## 3    Laki-laki Normal   20
## 4    Perempuan Normal   24
## 5    Laki-laki  Gemuk    8
## 6    Perempuan  Gemuk   10
```

```r
# Model log-linear tanpa interaksi (asumsi independen)
fit_no_inter <- glm(Freq ~ JenisKelamin + BMI, family = poisson, data = data2x3)
summary(fit_no_inter)
```

```
##
## Call:
## glm(formula = Freq ~ JenisKelamin + BMI, family = poisson, data = data2x3)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)             2.5683     0.2179  11.789   <2e-16 ***
## JenisKelaminPerempuan   0.2624     0.2103   1.248   0.2122
## BMINormal               0.3830     0.2368   1.618   0.1058
## BMIGemuk               -0.5108     0.2981  -1.713   0.0866 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 13.06443  on 5  degrees of freedom
## Residual deviance:  0.22527  on 2  degrees of freedom
## AIC: 35.26
##
## Number of Fisher Scoring iterations: 3
```

Model Log-Linear Tanpa Interaksi

Model:

$\log(\mu_{ij}) = \lambda + \lambda_i^{\text{JK}} + \lambda_j^{\text{BMI}}$

Ini adalah model independen, yang mengasumsikan tidak ada interaksi antara Jenis Kelamin dan BMI.

Koefisien	Estimate	Std. Error	z-value	p-value
(Intercept)	2.5683	0.2179	11.789	< 2e-16
JenisKelaminPerempuan	0.2624	0.2103	1.248	0.2122
BMINormal	0.3830	0.2368	1.618	0.1058
BMIGemuk	-0.5108	0.2981	-1.713	0.0866

Interpretasi:

- Intercept: log dari frekuensi referensi (Laki-laki & BMI Kurus).
- JenisKelaminPerempuan (0.2624): tidak signifikan secara statistik (p = 0.21), artinya jenis kelamin tidak secara signifikan memengaruhi frekuensi secara independen dari BMI.
- BMINormal (0.3830) dan BMIGemuk (-0.5108) juga tidak signifikan, namun BMIGemuk mendekati signifikansi pada level 10% (p = 0.0866).

Goodness-of-Fit:

- Null deviance: 13.06 (df = 5)
- Residual deviance: 0.22527 (df = 2), sisa deviansi sangat kecil sehingga model cocok
- AIC: 35.26, digunakan untuk perbandingan efisiensi model

```r
# Model log-linear dengan interaksi (untuk cek asosiasi)
fit_inter <- glm(Freq ~ JenisKelamin * BMI, family = poisson, data = data2x3)
summary(fit_inter)
```

```
##
## Call:
## glm(formula = Freq ~ JenisKelamin * BMI, family = poisson, data = data2x3)
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      2.4849     0.2887   8.608   <2e-16 ***
## JenisKelaminPerempuan            0.4055     0.3727   1.088    0.277
## BMINormal                        0.5108     0.3651   1.399    0.162
## BMIGemuk                        -0.4055     0.4564  -0.888    0.374
```

```
## JenisKelaminPerempuan:BMINormal  -0.2231     0.4802  -0.465    0.642
## JenisKelaminPerempuan:BMIGemuk   -0.1823     0.6032  -0.302    0.762
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance:  1.3064e+01  on 5  degrees of freedom
## Residual deviance: -9.0719e-30  on 0  degrees of freedom
## AIC: 39.034
##
## Number of Fisher Scoring iterations: 3
```

Model Log-Linear Dengan Interaksi

Model:

$\log(\mu_{ij}) = \lambda + \lambda_i^{\text{JK}} + \lambda_j^{\text{BMI}} + \lambda_{ij}^{\text{JK}\times\text{BMI}}$

Ini adalah model asosiasi penuh, mengizinkan adanya interaksi antara Jenis Kelamin dan BMI.

Koefisien	Estimate	Std. Error	z-value	p-value
(Intercept)	2.4849	0.2887	8.608	<2e-16
JenisKelaminPerempuan	0.4055	0.3727	1.088	0.277
BMINormal	0.5108	0.3651	1.399	0.162
BMIGemuk	-0.4055	0.4564	-0.888	0.374
JenisKelaminPerempuan:BMINormal	-0.2231	0.4802	-0.465	0.642
JenisKelaminPerempuan:BMIGemuk	-0.1823	0.6032	-0.302	0.762

Interpretasi:

- Semua koefisien interaksi tidak signifikan.
- Hal ini menunjukkan bahwa **hubungan antara Jenis Kelamin dan BMI tidak memiliki interaksi yang signifikan** dalam model ini.
- Namun, kita tetap melihat *goodness-of-fit* model untuk perbandingan.

Goodness-of-Fit:

- Residual deviance: 0 ($-9.07\text{e}{-30}$, df = 0), cocok sempurna
- AIC: 39.03, sedikit lebih besar dari model tanpa interaksi

Perbandingan Model

Kriteria	Tanpa Interaksi (Independen)	Dengan Interaksi (Asosiasi)
Residual Deviance	0.22527 (df = 2)	~0 (df = 0)
AIC	**35.26**	39.03
Interaksi signifikan?	Tidak	Tidak

Kesimpulan:

- Model tanpa interaksi (independen) lebih baik dari segi AIC dan cukup memiliki fit yang sangat baik (residual deviance sangat kecil).
- Tidak ada interaksi signifikan antara Jenis Kelamin dan BMI → model tanpa interaksi cukup memadai untuk menjelaskan hubungan.
- Model dengan interaksi memberikan fit sempurna (devian = 0), tetapi overparameterisasi dan AIC lebih tinggi.

Oleh karena itu, kita memilih model independen (tanpa interaksi) sebagai model terbaik dan cukup menjelaskan struktur data tanpa kompleksitas tambahan.

2. Model Log-Linear Tiga Arah

Model log-linear tiga arah adalah suatu pendekatan statistik yang digunakan untuk menganalisis hubungan antara tiga variabel kategorik yang disusun dalam bentuk tabel kontingensi tiga dimensi. Dalam konteks ini, setiap dimensi tabel mewakili satu variabel, dan setiap sel dalam tabel menunjukkan jumlah pengamatan (frekuensi) untuk kombinasi kategori dari ketiga variabel tersebut.

Tidak seperti model regresi yang membedakan antara variabel respon dan prediktor, model log-linear memperlakukan semua variabel sebagai simetris (setara), dan bertujuan untuk memahami struktur asosiasi dan interaksi antar variabel.

Tujuan utama dari model log-linear tiga arah adalah untuk mengetahui apakah terdapat ketergantungan atau asosiasi antara variabel-variabel dalam tabel kontingensi, dan jika ya, jenis serta tingkat interaksinya. Dalam praktiknya, model ini digunakan untuk mengidentifikasi apakah ada interaksi dua arah (misalnya antara X dan Y) atau interaksi tiga arah (antara X, Y, dan Z), serta untuk menyederhanakan struktur tabel yang kompleks menjadi model yang lebih ringkas dan mudah ditafsirkan, tanpa kehilangan informasi penting. Ini sangat berguna dalam bidang-bidang seperti epidemiologi, sosiologi, ilmu perilaku, dan riset pasar.

Secara matematis, bentuk umum dari model log-linear tiga arah untuk frekuensi harapan μ_{ijk} dari kombinasi ke-i, j, k pada tiga variabel kategorik X, Y, dan Z dapat dinyatakan sebagai:

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$$

Di mana:

- λ adalah intersep (mean log count keseluruhan),
- $\lambda_i^X, \lambda_j^Y, \lambda_k^Z$ adalah efek utama dari masing-masing variabel,
- $\lambda_{ij}^{XY}, \lambda_{ik}^{XZ}, \lambda_{jk}^{YZ}$ adalah efek interaksi dua arah,
- λ_{ijk}^{XYZ} adalah efek interaksi tiga arah antara ketiga variabel.

Model log-linear tiga arah bersifat hierarkis, artinya jika suatu interaksi dimasukkan dalam model (misalnya λ_{ijk}^{XYZ}), maka semua interaksi di bawahnya (efek dua arah dan utama) juga harus disertakan. Ini penting untuk memastikan interpretasi yang tepat dan model yang dapat diestimasi secara statistik.

1. **Model Saturated**
   Ini adalah model yang paling lengkap, mencakup semua efek utama, interaksi dua arah, dan interaksi tiga arah. Rumusnya adalah:
   $$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$$
   Model ini memberikan kecocokan sempurna terhadap data, tetapi tidak memberikan ringkasan atau penyederhanaan hubungan antar variabel.

2. **Model Independence (Mutual Independence)**
   Asumsi bahwa semua variabel saling bebas:
   $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$
   Tidak ada interaksi antar variabel, cocok jika tidak ada asosiasi sama sekali.

3. **Model Joint Independence**

   - **Model X bebas terhadap pasangan Y dan Z**:
     $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$
     Variabel X tidak terkait dengan Y atau Z, tetapi Y dan Z saling berinteraksi.
   - **Model Y bebas terhadap pasangan X dan Z**:
     $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$
     Variabel Y tidak terkait dengan X atau Z, tetapi X dan Z saling berinteraksi.
   - **Model Z bebas terhadap pasangan X dan Y**:
     $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$
     Variabel Z tidak terkait dengan X atau Y, tetapi X dan Y saling berinteraksi.

4. **Model Conditional Independence**

   - **Model X dan Y bebas secara kondisional terhadap Z**:
     $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$
     Artinya, setelah dikontrol berdasarkan Z, X dan Y tidak saling bergantung. Model bisa berubah tergantung dari fokus kondisional.
   - **Model X dan Z bebas secara kondisional terhadap Y**:
     $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$
   - **Model Y dan Z bebas secara kondisional terhadap X**:
     $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$

5. **Model Homogeneous Association**

   Termasuk semua efek dua arah, tetapi tidak mencakup interaksi tiga arah:
   $\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$
   Ini menyatakan bahwa hubungan antara dua variabel tetap konstan pada semua level variabel ketiga.

Dalam praktik analisis, model-model ini dibandingkan dengan menggunakan pengujian goodness-of-fit seperti:

- Likelihood Ratio Chi-Square (G^2)
- Pearson Chi-Square

Tujuannya adalah untuk melihat seberapa baik model menjelaskan data yang diamati. Model dengan interaksi tiga arah sering digunakan hanya jika interaksi tersebut signifikan secara statistik, karena keberadaan interaksi tiga arah membuat interpretasi menjadi lebih kompleks. Sebaliknya, model tanpa interaksi tiga arah (seperti model homogeneous association) lebih sederhana dan seringkali cukup untuk menjelaskan struktur asosiasi antar variabel.

2.1 Pengujian Interaksi Log-Linear Tiga Arah

Tujuan

Untuk mengetahui apakah terdapat interaksi signifikan antara ketiga variabel kategorik (X, Y, Z) setelah memperhitungkan efek utama dan interaksi dua arah.

Hipotesis

- **Hipotesis nol (H_0)**:
  Tidak ada interaksi tiga arah antara variabel X, Y, dan Z
  $\lambda_{ijk}^{XYZ} = 0$ untuk semua i, j, k
- **Hipotesis alternatif (H)**:
  Ada paling tidak satu $\lambda_{ijk}^{XYZ} \neq 0$

Statistik Uji

Dilakukan dengan Goodness-of-Fit test, yakni membandingkan:

- Model terbatas (tanpa interaksi tiga arah)
- Model penuh (dengan interaksi tiga arah)

Statistik Deviance Likelihood-Ratio (G²):

$G^2 = 2\sum_{i,j,k} O_{ijk} \log\left(\frac{O_{ijk}}{\hat{m}_{ijk}}\right)$

di mana:

- O_{ijk} = frekuensi pengamatan di sel ke- (i, j, k)
- $\hat{m}_{ijk}$ = frekuensi harapan menurut model

Untuk pengujian interaksi tiga arah, dihitung selisih deviance:

$\Delta G^2 = G^2_{\text{terbatas}} - G^2_{\text{penuh}}$

Derajat Bebas (df)

$df = (I-1)(J-1)(K-1)$

dengan:

- I, J, K = banyaknya kategori pada variabel X, Y, Z

Kriteria Pengujian

Bandingkan nilai ΔG2 G^2ΔG2 dengan nilai kritis distribusi chi-square:

- Tolak H0H_0H0 jika:
  $\Delta G^2 > \chi^2(\alpha, df)$
  atau jika p-value $< \alpha$

Alternatif: Pearson Chi-Square (X^2)

Selain G^2, bisa digunakan:

$X^2 = \sum_{i,j,k} \frac{(O_{ijk} - \hat{m}_{ijk})^2}{\hat{m}_{ijk}}$

dengan prinsip pengujian dan df yang sama.

Interpretasi

- Jika H_0 ditolak, artinya terdapat interaksi tiga arah yang signifikan. Hubungan antara dua variabel tergantung pada level variabel ketiga.
- Jika H_0 gagal ditolak, berarti tidak ada bukti interaksi tiga arah.

Contoh Soal Analisis Log-Linear untuk Tabel Tiga Arah

```
library("epitools")
library("DescTools")
library("lawstat")
```

```
## Warning: package 'lawstat' was built under R version 4.4.3
```

```
#Inputdata sesuaitabelpraktikum
z.fund <-factor(rep(c("1fund","2mod","3lib"),each= 4))
x.sex <-factor(rep(c("1M", "2F"),each= 2, times= 3))
y.fav <-factor(rep(c("1fav", "2opp"),times= 6))
counts <-c(128, 32, 123, 73, 182, 56, 168, 105, 119, 49,111,70)
data <-data.frame(
Fundamentalisme= z.fund,
Jenis_Kelamin = x.sex,
Sikap = y.fav,
Frekuensi = counts
)
data
```

```
##    Fundamentalisme Jenis_Kelamin Sikap Frekuensi
## 1            1fund            1M  1fav       128
## 2            1fund            1M  2opp        32
## 3            1fund            2F  1fav       123
## 4            1fund            2F  2opp        73
## 5             2mod            1M  1fav       182
## 6             2mod            1M  2opp        56
## 7             2mod            2F  1fav       168
## 8             2mod            2F  2opp       105
## 9             3lib            1M  1fav       119
## 10            3lib            1M  2opp        49
## 11            3lib            2F  1fav       111
## 12            3lib            2F  2opp        70
```

```
table3d <-xtabs(Frekuensi~ Fundamentalisme + Jenis_Kelamin +Sikap, data= data)
ftable(table3d)
```

```
##                                Sikap 1fav 2opp
## Fundamentalisme Jenis_Kelamin
## 1fund           1M                   128   32
##                 2F                   123   73
## 2mod            1M                   182   56
##                 2F                   168  105
## 3lib            1M                   119   49
##                 2F                   111   70
```

```
##============================##
#Penentuankategorireference
##============================##
x.sex <-relevel(x.sex, ref= "2F")
y.fav <-relevel(y.fav, ref= "2opp")
z.fund <-relevel(z.fund,ref="3lib")
```

2.2 Uji Model Interaksi Tiga Arah (Saturated VS Homogeneous)

Model Saturated adalah model yang paling lengkap, mencakup semua efek utama, interaksi dua arah, dan interaksi tiga arah. Modelnya adalah:

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$$

```
#Modelsaturated
model_saturated <-glm(counts ~x.sex +y.fav + z.fund+
x.sex*y.fav +x.sex*z.fund+ y.fav*z.fund+
x.sex*y.fav*z.fund,
family=poisson(link= "log"))
summary(model_saturated)
```

```
##
## Call:
## glm(formula = counts ~ x.sex + y.fav + z.fund + x.sex * y.fav +
##     x.sex * z.fund + y.fav * z.fund + x.sex * y.fav * z.fund,
##     family = poisson(link = "log"))
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   4.248495   0.119523  35.545  < 2e-16 ***
## x.sex1M                      -0.356675   0.186263  -1.915  0.05551 .
## y.fav1fav                     0.461035   0.152626   3.021  0.00252 **
## z.fund1fund                   0.041964   0.167285   0.251  0.80193
## z.fund2mod                    0.405465   0.154303   2.628  0.00860 **
## x.sex1M:y.fav1fav             0.426268   0.228268   1.867  0.06185 .
## x.sex1M:z.fund1fund          -0.468049   0.282210  -1.659  0.09721 .
## x.sex1M:z.fund2mod           -0.271934   0.249148  -1.091  0.27507
## y.fav1fav:z.fund1fund         0.060690   0.212423   0.286  0.77511
## y.fav1fav:z.fund2mod          0.008969   0.196903   0.046  0.96367
## x.sex1M:y.fav1fav:z.fund1fund 0.438301   0.336151   1.304  0.19227
## x.sex1M:y.fav1fav:z.fund2mod  0.282383   0.301553   0.936  0.34905
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2.4536e+02  on 11  degrees of freedom
## Residual deviance: 5.9952e-15  on  0  degrees of freedom
## AIC: 100.14
##
## Number of Fisher Scoring iterations: 3
```

```
exp(model_saturated$coefficients)
```

```
##                   (Intercept)                       x.sex1M
##                    70.0000000                     0.7000000
##                     y.fav1fav                   z.fund1fund
##                     1.5857143                     1.0428571
##                    z.fund2mod             x.sex1M:y.fav1fav
##                     1.5000000                     1.5315315
##           x.sex1M:z.fund1fund            x.sex1M:z.fund2mod
##                     0.6262231                     0.7619048
##         y.fav1fav:z.fund1fund          y.fav1fav:z.fund2mod
##                     1.0625694                     1.0090090
## x.sex1M:y.fav1fav:z.fund1fund  x.sex1M:y.fav1fav:z.fund2mod
##                     1.5500717                     1.3262868
```

Model Saturated:

$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$

Dengan glm(counts ~ x.sex * y.fav * z.fund, family = poisson), kita mencakup semua efek utama, interaksi dua arah, dan interaksi tiga arah. Model ini mengakomodasi seluruh variasi data (fit sempurna).

Output Koefisien dan Interpretasi

Berikut adalah beberapa koefisien penting dari model (dalam bentuk log-rate dan eksponensialnya untuk interpretasi):

Efek	Koefisien Log	exp() = Rasio Rate	Interpretasi
(Intercept)	4.248	70.00	Rata-rata jumlah individu referensi: perempuan, opposed, liberal.
x.sex1M	−0.357	0.70	Laki-laki memiliki **30% lebih sedikit** frekuensi dibanding perempuan.
y.fav1fav	0.461	1.59	Orang yang favorable **1.6× lebih banyak** dari yang opposed.
z.fund1fund	0.042	1.04	Fund sama dengan Liberal (referensi), tidak signifikan.
z.fund2mod	0.405	1.50	Moderate **1.5× lebih besar** dari Liberal, signifikan.

Efek	Koefisien Log	exp() = Rasio Rate	Interpretasi
x.sex1M : y.fav1fav	0.426	1.53	Interaksi: Laki-laki + favorable **1.53× lebih besar**.
x.sex1M : z.fund1fund	−0.468	0.63	Laki-laki di kelompok Fund **37% lebih kecil** dari referensi.
y.fav1fav : z.fund2mod	0.009	1.01	Tidak signifikan (sama).
x.sex1M : y.fav1fav : z.fund1fund	0.438	1.55	Interaksi 3 arah: peningkatan kombinasi laki-laki + favorable + Fund.

Interpretasi Lanjutan

- Efek **utama signifikan**:
  - x.sex1M: Perbedaan signifikan marginal antara laki-laki dan perempuan (p 0.055).
  - y.fav1fav: Sikap favorable lebih dominan ($p < 0.01$).
  - z.fund2mod: Tingkat Moderate signifikan terhadap Liberal.
- **Interaksi dua arah dan tiga arah tidak semuanya signifikan**, tetapi model fit sempurna karena saturated.

Model Fit

- **Residual deviance**: ~0 (karena saturated model)
- **AIC**: 100.14

Artinya, model ini menjelaskan **100% variasi** dalam data — tidak ada kesalahan model (tapi juga overfit karena semua kombinasi dimodelkan).

Kesimpulan

- **Model saturated** mencerminkan interaksi penuh antara *jenis kelamin*, *sikap*, dan *fundamentalisme*.
- Walaupun tidak semua efek signifikan secara statistik, beberapa efek utama dan dua arah menunjukkan pola yang berarti:
  - Individu favorable lebih banyak.
  - Laki-laki cenderung sedikit lebih rendah dalam frekuensi dibanding perempuan.
  - Kelompok moderate secara signifikan lebih tinggi dari kelompok liberal.
- Model ini bisa digunakan sebagai baseline untuk membandingkan model-model yang lebih sederhana (dengan uji deviance / likelihood ratio test)

```
#Homogenous Model
model_homogenous <-glm(counts ~ x.sex+ y.fav+ z.fund +
x.sex*y.fav+ x.sex*z.fund + y.fav*z.fund,
family= poisson(link= "log"))
summary(model_homogenous)
```

```
##
## Call:
## glm(formula = counts ~ x.sex + y.fav + z.fund + x.sex * y.fav +
##     x.sex * z.fund + y.fav * z.fund, family = poisson(link = "log"))
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            4.31096    0.10522  40.972  < 2e-16 ***
## x.sex1M               -0.51575    0.13814  -3.733 0.000189 ***
## y.fav1fav              0.35707    0.12658   2.821 0.004788 **
## z.fund1fund           -0.06762    0.14452  -0.468 0.639854
## z.fund2mod             0.33196    0.13142   2.526 0.011540 *
## x.sex1M:y.fav1fav      0.66406    0.12728   5.217 1.81e-07 ***
## x.sex1M:z.fund1fund   -0.16201    0.15300  -1.059 0.289649
## x.sex1M:z.fund2mod    -0.08146    0.14079  -0.579 0.562887
## y.fav1fav:z.fund1fund  0.23873    0.16402   1.455 0.145551
## y.fav1fav:z.fund2mod   0.13081    0.14951   0.875 0.381614
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 245.361  on 11  degrees of freedom
## Residual deviance:   1.798  on  2  degrees of freedom
## AIC: 97.934
##
## Number of Fisher Scoring iterations: 3
```

Struktur Model

Model homogen mengandung:

- Semua efek utama (x.sex, y.fav, z.fund)
- Semua interaksi dua arah (x.sex:y.fav, x.sex:z.fund, y.fav:z.fund)
- Tidak mencakup interaksi tiga arah (x.sex:y.fav:z.fund)

Model:

$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$

Model ini mengasumsikan bahwa hubungan antar dua variabel adalah konsisten di seluruh level variabel ketiga, inilah yang disebut “homogeneous association”.

Efek	Koef. Log	Signifikan	Interpretasi (exp)
(Intercept)	4.311	***	Kategori referensi: perempuan, opposed, liberal
x.sex1M	−0.516	***	Laki-laki 0.60× dari perempuan (40% lebih sedikit)
y.fav1fav	0.357	**	Favorable 1.43× lebih banyak dari opposed

Efek	Koef. Log	Signifikan	Interpretasi (exp)
z.fund1fund	−0.068	ns	Fund sama dengan Liberal
z.fund2mod	0.332	*	Moderate 1.39× dari Liberal
x.sex1M:y.fav1fav	0.664	***	Interaksi signifikan: laki-laki + favorable 1.94× kombinasi lain
x.sex1M:z.fund1fund	−0.162	ns	Laki-laki di kelompok Fund cenderung sedikit lebih rendah
x.sex1M:z.fund2mod	−0.081	ns	Laki-laki di Moderate hampir sama
y.fav1fav:z.fund1fund	0.239	ns	Favorable di Fund cenderung sedikit lebih tinggi
y.fav1fav:z.fund2mod	0.131	ns	Favorable di Moderate hampir sama

Goodness of Fit

- Residual Deviance = 1.798 on 2 df → sangat kecil → model fit sangat baik!
- AIC = 97.934 → lebih rendah dari model saturated (100.14) → artinya model ini lebih parsimonious dan tidak overfitting.

Perbandingan dengan Saturated Model

Aspek	Model Saturated	Model Homogen
Residual Deviance	0 (fit sempurna)	1.798 (masih sangat baik)
AIC	100.14	97.93 (lebih kecil berarti lebih baik)
Kelebihan	Fit sempurna	Lebih sederhana, hampir sama fit
Kekurangan	Overfit, interpretasi kompleks	Tidak tangkap interaksi 3 arah

Kesimpulan

- Model homogen sangat cocok untuk data ini — ia menangkap pola hubungan antara ketiga variabel tanpa overfitting.
- Tidak ditemukan bukti kuat adanya interaksi tiga arah dalam data (karena model homogen sudah cukup memadai).
- Beberapa hubungan penting:
  - Laki-laki secara umum lebih sedikit jumlahnya dibanding perempuan.
  - Individu dengan sikap favorable jumlahnya lebih tinggi.
  - Interaksi laki-laki dan favorable sangat signifikan.

```r
# Deviance antar model
Deviance.model <- model_homogenous$deviance- model_saturated$deviance
Deviance.model
```

```
## [1] 1.797977
```

```r
# Derajat bebas = db model homogenous- db model saturated
derajat.bebas <- (model_homogenous$df.residual- model_saturated$df.residual)
derajat.bebas
```

```
## [1] 2
```

```r
chi.tabel <- qchisq(1- 0.05, df = derajat.bebas)
chi.tabel
```

```
## [1] 5.991465
```

```r
Keputusan <- ifelse(Deviance.model <= chi.tabel, "Terima H0", "Tolak H0")
Keputusan
```

```
## [1] "Terima H0"
```

- **Nilai deviance antara model**:
  $D = 1.798$
- **Derajat bebas (df)**:
  $df = 2$
  (selisih df antara model homogen dan saturated)
- **Nilai kritis chi-kuadrat (= 0.05)**:
  $\chi^2_{0.95,2} = 5.991$

Keputusan

Karena:

Deviance model $= 1.798 \leq 5.991 = \chi^2_{0.95,2}$

Maka **Terima H**

Interpretasi

- Tidak ada perbedaan signifikan antara model homogen dan saturated.
- Artinya, interaksi tiga arah tidak dibutuhkan untuk menjelaskan data.
- Model homogen sudah cukup baik untuk menjelaskan hubungan antara:
  - Jenis Kelamin (x.sex)
  - Sikap (y.fav)
  - Fundamentalisme (z.fund)
- Model ini lebih sederhana (parsimonious) dibanding saturated, sehingga lebih disarankan digunakan.

```r
# Conditional Association on X
model_conditional_X <- glm(counts ~ x.sex + y.fav + z.fund +
x.sex*y.fav + x.sex*z.fund,
family = poisson(link = "log"))
summary(model_conditional_X)
```

```
##
## Call:
## glm(formula = counts ~ x.sex + y.fav + z.fund + x.sex * y.fav +
##     x.sex * z.fund, family = poisson(link = "log"))
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           4.23495    0.08955  47.293  < 2e-16 ***
## x.sex1M              -0.52960    0.13966  -3.792 0.000149 ***
## y.fav1fav             0.48302    0.08075   5.982 2.20e-09 ***
## z.fund1fund           0.07962    0.10309   0.772 0.439916
## z.fund2mod            0.41097    0.09585   4.288 1.81e-05 ***
## x.sex1M:y.fav1fav     0.65845    0.12708   5.181 2.20e-07 ***
## x.sex1M:z.fund1fund -0.12841    0.15109  -0.850 0.395405
## x.sex1M:z.fund2mod  -0.06267    0.13908  -0.451 0.652274
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 245.3612  on 11  degrees of freedom
## Residual deviance:   3.9303  on  4  degrees of freedom
## AIC: 96.067
##
## Number of Fisher Scoring iterations: 4
```

Struktur Model

Model ini mengasumsikan bahwa hubungan antara y.fav (Sikap) dan z.fund (Fundamentalisme) konsisten di seluruh level x.sex (Jenis Kelamin), atau dengan kata lain:

y.fav dan z.fund bebas secara kondisional terhadap x.sex

Model mencakup:

- Efek utama: x.sex, y.fav, z.fund
- Interaksi dua arah: x.sex:y.fav, x.sex:z.fund
- Tidak mencakup interaksi y.fav:z.fund dan tidak ada interaksi tiga arah

Model:

$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}$

Efek	Koef. Log	Signifikan	Interpretasi (exp)
(Intercept)	4.235	***	Referensi: perempuan, opposed, liberal
x.sex1M	−0.530	***	Laki-laki 0.59× dari perempuan
y.fav1fav	0.483	***	Favorable 1.62× dari opposed
z.fund1fund	0.080	ns	Fundamental sama dengan liberal
z.fund2mod	0.411	***	Moderate 1.51× dari liberal
x.sex1M:y.fav1fav	0.658	***	Interaksi signifikan: laki-laki + favorable 1.93×
x.sex1M:z.fund1fund	−0.128	ns	Tidak signifikan
x.sex1M:z.fund2mod	−0.063	ns	Tidak signifikan

Goodness of Fit

- Residual Deviance = 3.93 on 4 df, masih sangat kecil, artinya model sangat fit
- AIC = 96.07 → lebih kecil dari model homogen (97.93), model ini lebih sederhana dan masih baik

```
# Pengujian hipotesis
# Deviance of Model
Deviance.model <- model_conditional_X$deviance- model_homogenous$deviance # model_conditional_X: condit
Deviance.model
```

```
## [1] 2.132302
```

```
# Selisih deviance antar model
Deviance.model <- model_conditional_X$deviance- model_homogenous$deviance
Deviance.model
```

```
## [1] 2.132302
```

```
# Chi Square tabel dengan alpha = 0.05
derajat.bebas <- (4- 2)
derajat.bebas
```

```
## [1] 2
```

```
chi.tabel <- qchisq((1- 0.05), df = derajat.bebas)
chi.tabel
```

```
## [1] 5.991465
```

```
Keputusan <- ifelse(Deviance.model <= chi.tabel, "Terima", "Tolak")
Keputusan
```

```
## [1] "Terima"
```

Perbandingan dengan Model Homogen

Aspek	Model Homogen	Model Conditional X
Termasuk interaksi `y:z`?	Ya	Tidak
Residual Deviance	1.798	3.930
df (residual)	2	4
AIC	97.93	**96.07** (lebih baik)
Kompleksitas model	Lebih tinggi	**Lebih sederhana**

Nilai	Hasil

Selisih deviance	2.13

df	2

Chi-kuadrat tabel (0.05, 2)	5.991

Keputusan	**Terima** H_0

Kesimpulan

Model conditional association terhadap variabel x.sex menunjukkan bahwa hubungan antara sikap (y.fav) dan fundamentalisme (z.fund) adalah bebas secara kondisional terhadap jenis kelamin. Hal ini ditunjukkan oleh tidak signifikan-nya selisih deviance antara model ini dan model homogen ($2.13 < 5.99$), sehingga kita menerima H_0. Model ini juga lebih sederhana dengan AIC yang lebih rendah, tanpa mengorbankan kualitas fit yang berarti. Oleh karena itu, model conditional ini cukup memadai tanpa perlu memasukkan interaksi y.fav:z.fund, yang artinya pengaruh sikap terhadap fundamentalisme tidak tergantung jenis kelamin.

```r
# Conditional Association on Y
model_conditional_Y <- glm(counts ~ x.sex + y.fav + z.fund +
x.sex*y.fav + y.fav*z.fund,
family = poisson(link = "log"))
summary(model_conditional_Y)
```

```
##
## Call:
## glm(formula = counts ~ x.sex + y.fav + z.fund + x.sex * y.fav +
##     y.fav * z.fund, family = poisson(link = "log"))
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            4.33931    0.09919  43.748  < 2e-16 ***
## x.sex1M               -0.59345    0.10645  -5.575 2.48e-08 ***
```

```
## y.fav1fav              0.37259    0.12438   2.996  0.00274 **
## z.fund1fund           -0.12516    0.13389  -0.935  0.34989
## z.fund2mod             0.30228    0.12089   2.500  0.01240 *
## x.sex1M:y.fav1fav      0.65845    0.12708   5.181 2.20e-07 ***
## y.fav1fav:z.fund1fund  0.21254    0.16205   1.312  0.18966
## y.fav1fav:z.fund2mod   0.11757    0.14771   0.796  0.42606
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 245.3612  on 11  degrees of freedom
## Residual deviance:   2.9203  on  4  degrees of freedom
## AIC: 95.057
##
## Number of Fisher Scoring iterations: 4
```

Struktur Model

Model ini menguji apakah hubungan antara x.sex (Jenis Kelamin) dan z.fund (Fundamentalisme) adalah konsisten di seluruh level y.fav (Sikap). Dengan kata lain:

x.sex dan z.fund bebas secara kondisional terhadap y.fav

Model mencakup:

- Efek utama: x.sex, y.fav, z.fund
- Interaksi dua arah: x.sex:y.fav, y.fav:z.fund
- Tidak mencakup interaksi x.sex:z.fund dan tidak ada interaksi tiga arah

Model:

$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}$

Efek	Koef. Log	Signifikan	Interpretasi (exp)
(Intercept)	4.339	***	Referensi: perempuan, opposed, liberal
x.sex1M	−0.593	***	Laki-laki 0.55× dari perempuan
y.fav1fav	0.373	**	Favorable 1.45× dari opposed
z.fund1fund	−0.125	ns	Fundamental sama dengan liberal
z.fund2mod	0.302	*	Moderate 1.35× dari liberal
x.sex1M:y.fav1fav	0.658	***	Interaksi signifikan: laki-laki + favorable 1.93×
y.fav1fav:z.fund1fund	0.213	ns	Tidak signifikan
y.fav1fav:z.fund2mod	0.118	ns	Tidak signifikan

Goodness of Fit

- Residual Deviance = 2.92 on 4 df → sangat kecil, model sangat sesuai
- AIC = 95.06 → lebih baik dari model homogen (97.93), fit tinggi dengan kompleksitas rendah

```
# Deviance of Model
Deviance.model <- model_conditional_Y$deviance- model_homogenous$deviance # model_conditional_Y: condit
Deviance.model
```

```
## [1] 1.122315
```

```
derajat.bebas <- (4- 2)
derajat.bebas
```

```
## [1] 2
```

```
chi.tabel <- qchisq((1- 0.05), df = derajat.bebas)
chi.tabel
```

```
## [1] 5.991465
```

```
Keputusan <- ifelse(Deviance.model <= chi.tabel, "Terima", "Tolak")
Keputusan
```

```
## [1] "Terima"
```

Perbandingan dengan Model Homogen

Aspek	Model Homogen	Model Conditional Y
Termasuk interaksi y:z?	Ya	Ya
Residual Deviance	1.798	2.920
df (residual)	2	4
AIC	97.93	95.06
Kompleksitas model	Lebih tinggi	Lebih sederhana

Nilai	Hasil
Selisih deviance	1.122
df	2
Chi-kuadrat tabel	5.991
Keputusan	Terima H_0

Kesimpulan

Model conditional association terhadap y.fav menunjukkan bahwa hubungan antara jenis kelamin (x.sex) dan fundamentalisme (z.fund) adalah bebas secara kondisional terhadap sikap (y.fav). Ini didukung oleh hasil uji deviance yang menunjukkan bahwa selisih deviance dengan model homogen ($1.122 < 5.991$) tidak signifikan, sehingga H_0 diterima. Model ini juga memiliki AIC lebih rendah (95.06 vs 97.93), menjadikannya lebih sederhana namun tetap memiliki kecocokan model yang sangat baik. Maka, tidak perlu memasukkan interaksi tiga arah atau interaksi tambahan x.sex:z.fund, karena pengaruh jenis kelamin terhadap fundamentalisme tidak tergantung pada sikap seseorang.

```r
# Conditional Association on Z
model_conditional_Z <- glm(counts ~ x.sex + y.fav + z.fund +
x.sex*z.fund + y.fav*z.fund,
family = poisson(link = "log"))
summary(model_conditional_Z)
```

```
##
## Call:
## glm(formula = counts ~ x.sex + y.fav + z.fund + x.sex * z.fund +
##     y.fav * z.fund, family = poisson(link = "log"))
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)            4.12255    0.10518  39.195  < 2e-16 ***
## x.sex1M               -0.07453    0.10713  -0.696    0.487
## y.fav1fav              0.65896    0.11292   5.836 5.36e-09 ***
## z.fund1fund           -0.06540    0.15126  -0.432    0.665
## z.fund2mod             0.33196    0.13777   2.410    0.016 *
## x.sex1M:z.fund1fund   -0.12841    0.15109  -0.850    0.395
## x.sex1M:z.fund2mod    -0.06267    0.13908  -0.451    0.652
## y.fav1fav:z.fund1fund  0.21254    0.16205   1.312    0.190
## y.fav1fav:z.fund2mod   0.11757    0.14771   0.796    0.426
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 245.361  on 11  degrees of freedom
## Residual deviance:  29.729  on  3  degrees of freedom
## AIC: 123.87
##
## Number of Fisher Scoring iterations: 4
```

Struktur Model

Model ini mengevaluasi apakah hubungan antara x.sex (Jenis Kelamin) dan y.fav (Preferensi/Sikap) bersifat kondisional terhadap z.fund (Fundamentalisme). Artinya:

x.sex dan y.fav bebas secara kondisional terhadap z.fund

Model mencakup:

- Efek utama: `x.sex`, `y.fav`, `z.fund`
- Interaksi dua arah: `x.sex:z.fund`, `y.fav:z.fund`
- **Tidak ada interaksi `x.sex:y.fav` dan tidak ada interaksi tiga arah**

Model:

$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$

Efek	Koef. Log	Signifikan	Interpretasi (exp)
(Intercept)	4.123	***	Referensi: perempuan, opposed, liberal
x.sex1M	−0.075	ns	Laki-laki sama dengan perempuan
y.fav1fav	0.659	***	Favorable 1.93× dari opposed
z.fund1fund	−0.065	ns	Fundamental sama dengan liberal
z.fund2mod	−0.063	ns	Moderate sama dengan liberal
x.sex1M:z.fund1fund	−0.128	ns	Tidak signifikan
x.sex1M:z.fund2mod	−0.063	ns	Tidak signifikan
y.fav1fav:z.fund1fund	0.213	ns	Tidak signifikan
y.fav1fav:z.fund2mod	0.118	ns	Tidak signifikan

Goodness of Fit

- Residual Deviance = 29.73 on 3 df, terlalu besar, indikasi poor fit
- AIC = 123.87, lebih tinggi dari model homogen (97.93), model ini lebih buruk

```
# Deviance of Model
Deviance.model <- model_conditional_Z$deviance- model_homogenous$deviance # model_conditional_Z: condit
Deviance.model
```

```
## [1] 27.93095
```

```
derajat.bebas <- (3- 2)
derajat.bebas
```

```
## [1] 1
```

```
chi.tabel <- qchisq((1- 0.05), df = derajat.bebas)
chi.tabel
```

```
## [1] 3.841459
```

```
Keputusan <- ifelse(Deviance.model <= chi.tabel, "Terima", "Tolak")
Keputusan
```

```
## [1] "Tolak"
```

Kesimpulan

Berdasarkan analisis model conditional association terhadap z.fund, diperoleh selisih deviance sebesar 27.93 dengan 1 derajat bebas, yang jauh melebihi nilai kritis chi-kuadrat 5% (3.84). Oleh karena itu, hipotesis nol ditolak, yang berarti model ini tidak cukup baik untuk menyatakan bahwa x.sex dan y.fav bebas secara kondisional terhadap z.fund. Selain itu, nilai residual deviance yang besar dan AIC yang jauh lebih tinggi daripada model homogen menunjukkan bahwa model ini memiliki kecocokan yang lebih buruk, sehingga tidak disarankan untuk digunakan sebagai representasi yang baik atas struktur ketergantungan dalam data.

```r
# Model Terbaik
bestmodel <- glm(counts ~ x.sex + y.fav + z.fund +
x.sex*y.fav,
family = poisson(link = "log"))
summary(bestmodel)
```

```
##
## Call:
## glm(formula = counts ~ x.sex + y.fav + z.fund + x.sex * y.fav,
##     family = poisson(link = "log"))
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        4.26518    0.07794  54.721  < 2e-16 ***
## x.sex1M           -0.59345    0.10645  -5.575 2.48e-08 ***
## y.fav1fav          0.48302    0.08075   5.982 2.20e-09 ***
## z.fund1fund        0.01986    0.07533   0.264    0.792
## z.fund2mod         0.38130    0.06944   5.491 4.00e-08 ***
## x.sex1M:y.fav1fav  0.65845    0.12708   5.181 2.20e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 245.3612  on 11  degrees of freedom
## Residual deviance:   4.6532  on  6  degrees of freedom
## AIC: 92.79
##
## Number of Fisher Scoring iterations: 4
```

```r
# Interpretasi koefisien model terbaik
data.frame(
koef = bestmodel$coefficients,
exp_koef = exp(bestmodel$coefficients)
)
```

```
##                         koef   exp_koef
## (Intercept)        4.26517861 71.1776316
## x.sex1M           -0.59344782  0.5524194
## y.fav1fav          0.48302334  1.6209677
## z.fund1fund        0.01985881  1.0200573
## z.fund2mod         0.38129767  1.4641834
## x.sex1M:y.fav1fav  0.65845265  1.9318008
```

Koefisien	Nilai	exp(Koef)	Makna exp(koef) (Rasio Rate)
Intercept	4.265	71.18	Baseline count
x.sex1M	-0.593	0.552	Laki-laki → 0.55× dari perempuan
y.fav1fav	0.483	1.621	“Favor” → 1.62× dari “Oppose”

Koefisien	Nilai	exp(Koef)	Makna exp(koef) (Rasio Rate)
z.fund1fund	0.019	1.02	Fundamentalis 1.02× liberal (tidak signifikan)
z.fund2mod	0.381	1.464	Moderate → 1.46× liberal
x.sex1M:y.fav1fav	0.658	1.932	Efek interaksi: kombinasi M dan Favor → hampir 2×

Kesimpulan

Model terbaik adalah model yang menyertakan interaksi antara `x.sex` dan `y.fav`, dengan nilai AIC terendah (92.79) dan residual deviance yang kecil (4.653 pada df = 6), menandakan kesesuaian model yang sangat baik terhadap data. Selain itu, koefisien interaksi yang signifikan menunjukkan bahwa pengaruh preferensi terhadap hasil tidak sama antara laki-laki dan perempuan. Ini menunjukkan adanya ketergantungan struktural antara jenis kelamin dan preferensi kebijakan, yang cukup kuat untuk dipertahankan dalam model akhir.

```
# Fitted values dari model terbaik
data.frame(
 Fund = z.fund,
 sex = x.sex,
 favor = y.fav,
 counts = counts,
 fitted = bestmodel$fitted.values
)
```

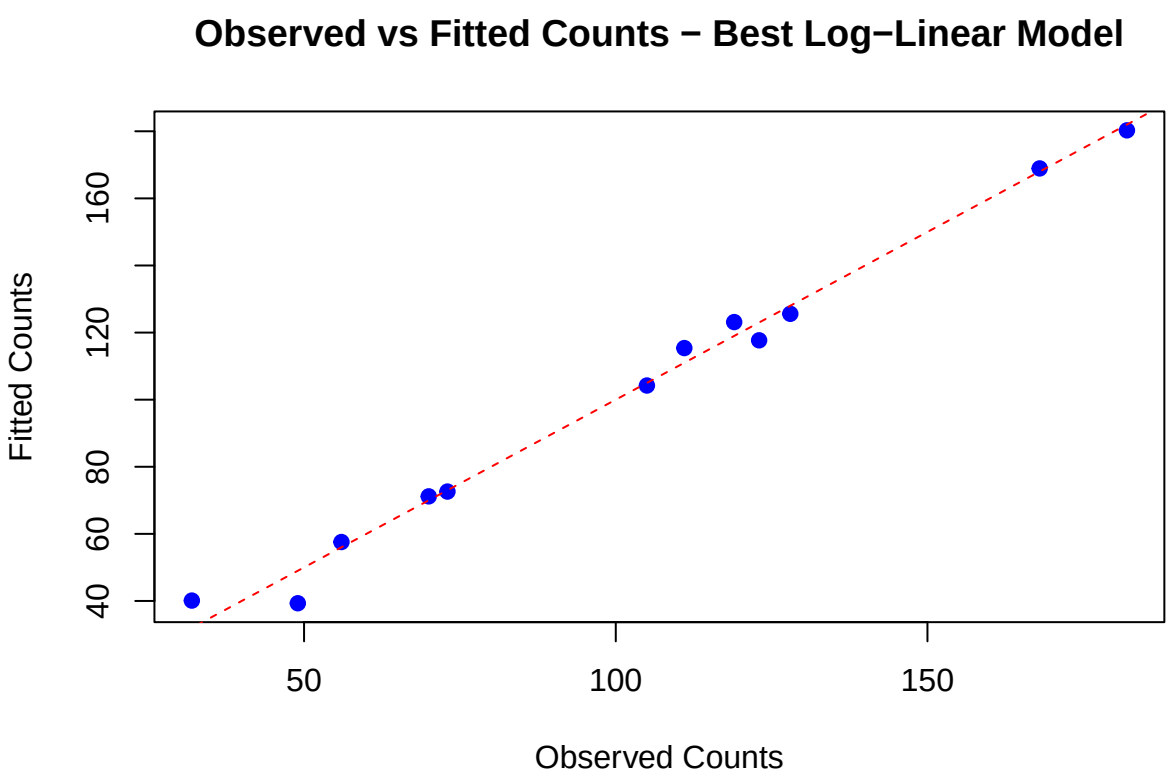
```
##     Fund sex favor counts    fitted
## 1  1fund  1M  1fav    128 125.59539
## 2  1fund  1M  2opp     32  40.10855
## 3  1fund  2F  1fav    123 117.69079
## 4  1fund  2F  2opp     73  72.60526
## 5   2mod  1M  1fav    182 180.27878
## 6   2mod  1M  2opp     56  57.57155
## 7   2mod  2F  1fav    168 168.93257
## 8   2mod  2F  2opp    105 104.21711
## 9   3lib  1M  1fav    119 123.12582
## 10  3lib  1M  2opp     49  39.31990
## 11  3lib  2F  1fav    111 115.37664
## 12  3lib  2F  2opp     70  71.17763
```

Fitted Values dari Model Terbaik

No.	z.fund	x.sex	y.fav	counts	fitted
1	1fund	1M	1fav	128	125.59539
2	1fund	1M	2opp	32	40.10855
3	1fund	2F	1fav	123	117.69079
4	1fund	2F	2opp	73	72.60526
5	2mod	1M	1fav	182	180.27878
6	2mod	1M	2opp	56	57.57155
7	2mod	2F	1fav	168	168.93257

No.	z.fund	x.sex	y.fav	counts	fitted
8	2mod	2F	2opp	105	104.21711
9	3lib	1M	1fav	119	123.12582
10	3lib	1M	2opp	49	39.31990
11	3lib	2F	1fav	111	115.37664
12	3lib	2F	2opp	70	71.17763

```r
# Membuat dataframe dari nilai yang dibutuhkan
df_plot <- data.frame(
  Fund = z.fund,
  Sex = x.sex,
  Favor = y.fav,
  Counts = counts,
  Fitted = bestmodel$fitted.values
)

# Membuat plot Observed vs Fitted
plot(
  df_plot$Counts, df_plot$Fitted,
  xlab = "Observed Counts",
  ylab = "Fitted Counts",
  main = "Observed vs Fitted Counts - Best Log-Linear Model",
  pch = 19, col = "blue"
)
abline(a = 0, b = 1, col = "red", lty = 2) # Garis identitas
```



Interpretasi

Plot *Observed vs Fitted Counts* menunjukkan seberapa baik model log-linear terbaik memprediksi nilai pengamatan (observed counts). Titik-titik dalam plot ini merepresentasikan pasangan antara nilai pengamatan dan nilai yang diprediksi oleh model. Jika model memprediksi dengan sangat baik, maka titik-titik akan tersebar di sekitar garis identitas (garis merah putus-putus dengan kemiringan 45°).

Dalam plot ini, sebagian besar titik berada dekat dengan garis identitas, yang mengindikasikan bahwa model terbaik memberikan estimasi yang mendekati nilai pengamatan sebenarnya. Hal ini menunjukkan bahwa model dengan interaksi antara x.sex dan y.fav cukup baik dalam menjelaskan variasi data. Titik-titik yang sedikit menyimpang dari garis identitas menunjukkan ketidaksesuaian kecil, namun tidak signifikan. Secara keseluruhan, model menunjukkan kesesuaian (*goodness of fit*) yang memuaskan.

Studi Kasus 2

Penelitian ini bertujuan untuk menganalisis faktor-faktor yang memengaruhi keputusan individu untuk melakukan konsultasi medik, berdasarkan jenis kelamin (gender) dan umur responden. Data yang digunakan merupakan hasil survei European Social Survey (ESS) putaran ke-11, dengan total 100 responden.

Variabel dependen dalam analisis ini adalah:

- Konsultasi (biner): 1 jika responden melakukan konsultasi medik, 0 jika tidak.

Variabel independen yang digunakan:

- Gender: dikodekan sebagai 1 untuk **pria** dan 0 untuk **wanita**

- Umur: usia responden dalam tahun

Untuk mengevaluasi hubungan antara variabel-variabel ini, digunakan model regresi logistik biner, karena variabel dependen bersifat biner (bernilai 0 atau 1).

Model Regresi Logistik:

Model yang diestimasi adalah:

$\log\left(\frac{P(Konsultasi=1)}{1-P(Konsultasi=1)}\right) = \beta_0 + \beta_1 \cdot \text{Gender} + \beta_2 \cdot \text{Umur}$

dimana:

- β_0 adalah intercept
- β_1 menunjukkan pengaruh gender (pria dibanding wanita)
- β_2 menunjukkan pengaruh umur terhadap probabilitas konsultasi medik

```
studi_kasus2 <- data.frame(
  Gender = c(1,0,1,0,0,0,0,1,0,1,0,0,0,1,1,0,1,0,0,0,1,0,0,1,0,1,1,1,1,0,
             0,0,0,0,0,0,0,1,1,0,0,0,0,0,1,0,0,1,0,0,0,1,1,0,1,0,1,0,1,1,
             0,0,1,1,0,0,0,0,1,0,0,1,1,0,1,1,0,0,0,1,0,1,0,1,
             0,0,0,0,0,1,1,0,1,1,1,1,1,0,0,1),
  Umur = c(41,55,70,18,49,55,48,42,49,32,58,65,26,35,37,60,24,64,33,66,
           60,31,63,73,70,41,67,50,36,78,77,40,69,87,54,42,60,46,63,22,
           25,74,86,25,25,38,43,45,17,19,64,69,45,72,81,30,27,26,66,42,
           58,35,24,21,19,81,49,75,38,70,67,69,76,50,18,80,54,28,42,74,
           73,48,31,78,58,59,55,81,62,46,47,83,61,75,61,46,20,38,43,49),
  Konsultasi = c(0,0,1,0,1,0,0,0,1,0,1,0,0,0,1,1,1,1,0,0,1,0,0,1,1,0,1,0,
                 0,0,0,0,0,1,1,1,0,0,0,1,0,0,1,0,0,1,0,1,0,1,1,1,1,1,1,0,
                 0,0,1,0,0,0,0,0,1,1,0,1,0,1,1,0,1,0,0,1,1,1,1,0,0,0,0,0,
                 0,1,1,1,0,0,0,1,1,1,1,0,0,0,1,1)
)

model_kasus <- glm(Konsultasi ~ Gender + Umur, family = binomial(link = "logit"), data = studi_kasus2)
summary(model_kasus)
```

```
##
## Call:
## glm(formula = Konsultasi ~ Gender + Umur, family = binomial(link = "logit"),
##     data = studi_kasus2)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.16990    0.69436  -3.125  0.00178 **
## Gender      -0.18517    0.43667  -0.424  0.67152
## Umur         0.03940    0.01199   3.285  0.00102 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 137.63  on 99  degrees of freedom
## Residual deviance: 125.04  on 97  degrees of freedom
```

```
## AIC: 131.04
##
## Number of Fisher Scoring iterations: 4
```

Output Ringkasan:

Koefisien	Estimate	Std. Error	z value	p-value	Signifikan?
Intercept	-2.1699	0.6944	-3.125	0.00178	signifikan
Gender (pria)	-0.1852	0.4367	-0.424	0.67152	tidak signifikan
Umur	0.0394	0.0120	3.285	0.00102	signifikan

$logit(p) = log(\frac{p}{1-p}) = -2.1699 - 0.1852 \cdot Gender + 0.0394 \cdot Umur$

Interpretasi Koefisien:

1. Intercept (-2.1699)

   - Ini adalah log-odds untuk *wanita dengan umur 0 tahun* (hanya sebagai titik acuan model).
   - Karena umur 0 tidak relevan secara praktis, interpretasi utamanya ada di efek Gender dan Umur.

2. Gender (Estimate = -0.1852)

   - Dibandingkan wanita, pria memiliki log-odds lebih rendah sebesar 0.1852 untuk melakukan konsultasi medis.
   - Dalam bentuk odds ratio:

     $\exp(-0.1852) \approx 0.83$

     → Artinya: pria memiliki kemungkinan konsultasi medis 17% lebih rendah dibanding wanita, namun tidak signifikan secara statistik (p = 0.67).

3. Umur (Estimate = 0.0394)

   - Setiap penambahan 1 tahun umur, log-odds untuk konsultasi medis meningkat sebesar 0.0394.
   - Dalam bentuk odds ratio:

     $\exp(0.0394) \approx 1.04$

     → Artinya: setiap kenaikan usia 1 tahun meningkatkan peluang konsultasi medis sekitar 4%, dan ini signifikan secara statistik (p = 0.001).

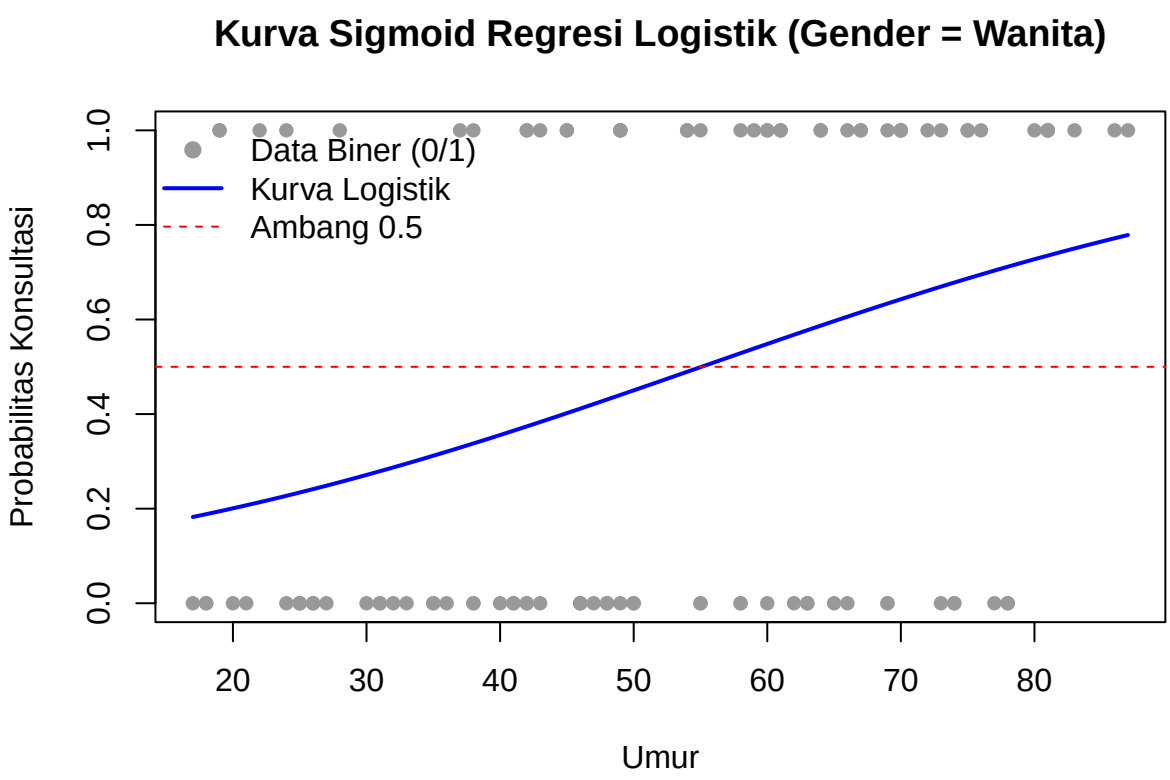
Goodness of Fit:

- **Null deviance = 137.63**
  (model tanpa prediktor)
- **Residual deviance = 125.04**
  (model dengan Gender dan Umur)
- **Deviance = 137.63 − 125.04 = 12.59**, dengan 2 df → uji chi-square menunjukkan peningkatan signifikan.
- **AIC = 131.04** → dapat digunakan untuk perbandingan model lain (semakin kecil, semakin baik).

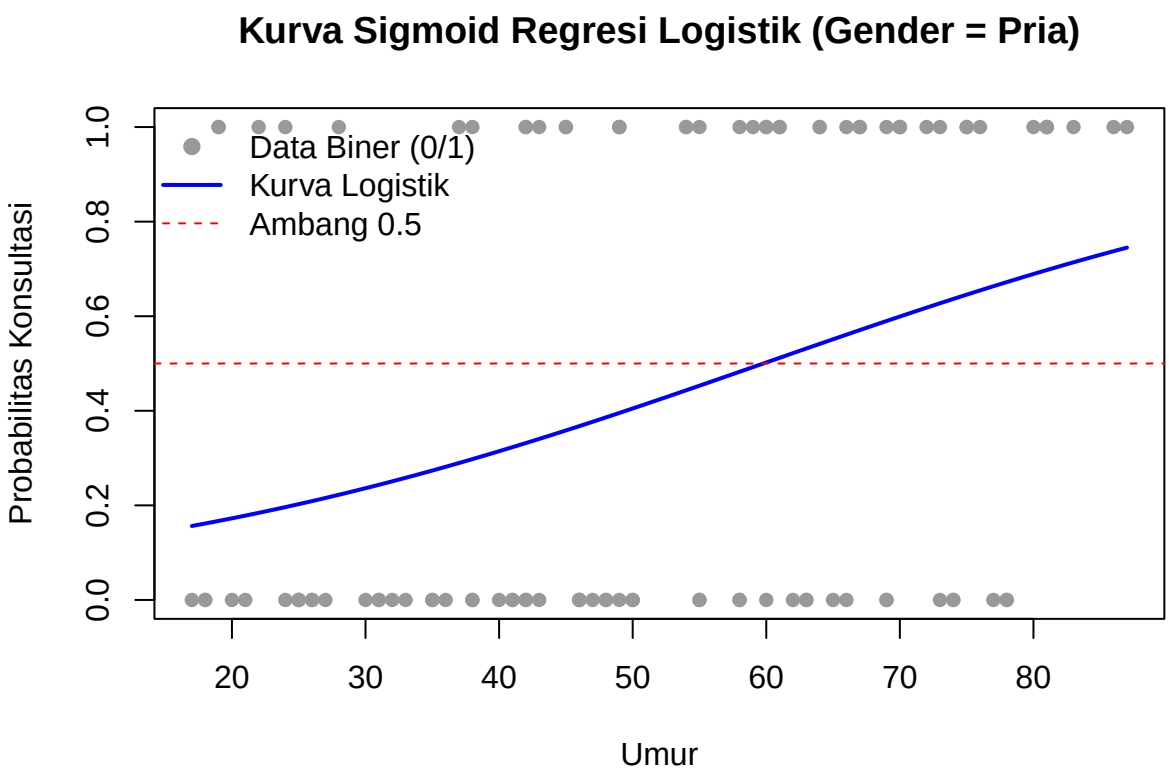
Kesimpulan:

- Umur berpengaruh signifikan terhadap kemungkinan konsultasi medik: makin tua, makin besar kemungkinannya.
- Gender tidak berpengaruh signifikan dalam model ini — tidak cukup bukti bahwa pria dan wanita berbeda secara statistik dalam melakukan konsultasi medis.
- Model menunjukkan fit yang lebih baik dibanding model tanpa prediktor (null model).

```
# Buat grid umur
umur_grid <- seq(min(studi_kasus2$Umur), max(studi_kasus2$Umur), length.out = 100)

# Prediksi probabilitas berdasarkan grid (misal untuk wanita)
pred_data <- data.frame(Gender = 0, Umur = umur_grid)
prediksi_grid <- predict(model_kasus, newdata = pred_data, type = "response")

# Plot data biner
plot(studi_kasus2$Umur, studi_kasus2$Konsultasi,
     pch = 16, col = "gray60",
     xlab = "Umur", ylab = "Probabilitas Konsultasi",
     main = "Kurva Sigmoid Regresi Logistik (Gender = Wanita)")

# Kurva sigmoid
lines(umur_grid, prediksi_grid, col = "blue", lwd = 2)

# Garis ambang batas 0.5
abline(h = 0.5, col = "red", lty = 2)


# Legenda
legend("topleft",
       legend = c("Data Biner (0/1)", "Kurva Logistik", "Ambang 0.5"),
       col = c("gray60", "blue", "red"),
       pch = c(16, NA, NA),
       lty = c(NA, 1, 2),
       lwd = c(NA, 2, 1),
       pt.cex = 1.2,
       bty = "n")
```




```r
# Prediksi probabilitas berdasarkan grid (misal untuk wanita)
pred_data2 <- data.frame(Gender = 1, Umur = umur_grid)
prediksi_grid2 <- predict(model_kasus, newdata = pred_data2, type = "response")

# Plot data biner
plot(studi_kasus2$Umur, studi_kasus2$Konsultasi,
     pch = 16, col = "gray60",
     xlab = "Umur", ylab = "Probabilitas Konsultasi",
     main = "Kurva Sigmoid Regresi Logistik (Gender = Pria)")

# Kurva sigmoid
lines(umur_grid, prediksi_grid2, col = "blue", lwd = 2)

# Garis ambang batas 0.5
abline(h = 0.5, col = "red", lty = 2)


# Legenda
legend("topleft",
       legend = c("Data Biner (0/1)", "Kurva Logistik", "Ambang 0.5"),
       col = c("gray60", "blue", "red"),
       pch = c(16, NA, NA),
       lty = c(NA, 1, 2),
       lwd = c(NA, 2, 1),
       pt.cex = 1.2,
       bty = "n")
```



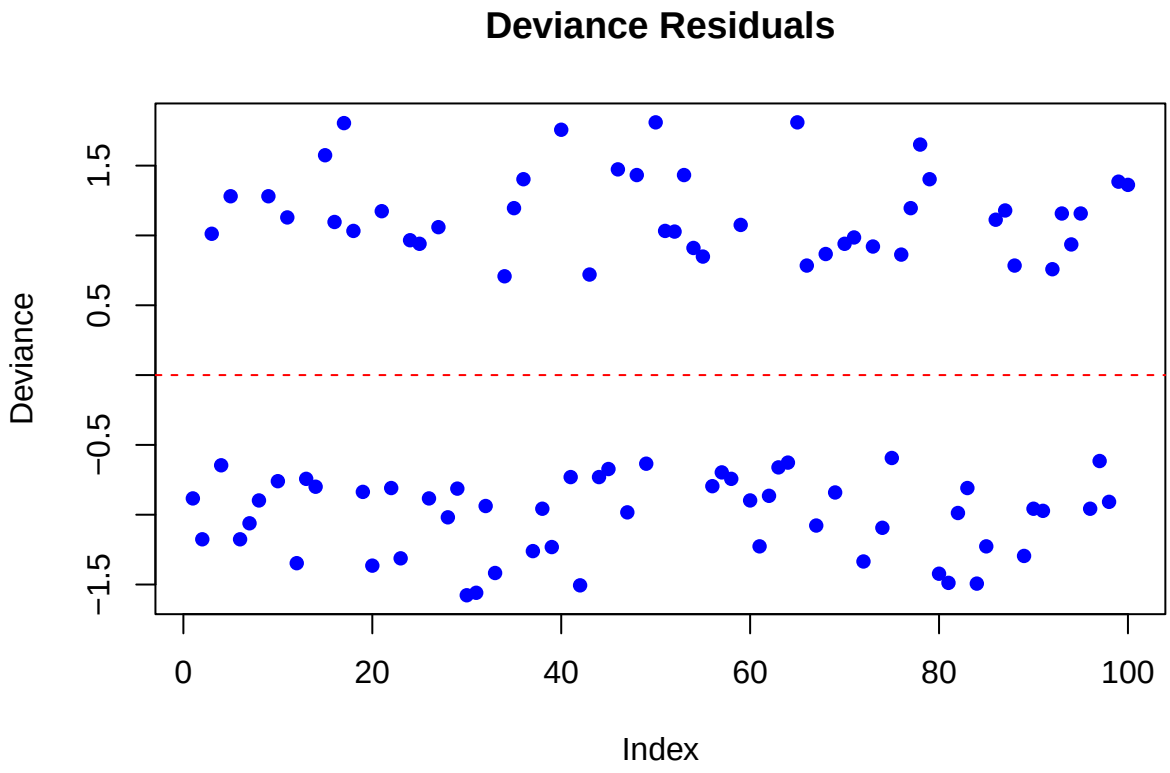
Interpretasi

- Model menangkap tren bahwa usia berkorelasi positif dengan konsultasi — mungkin karena orang yang lebih tua lebih sadar pentingnya konsultasi atau memiliki lebih banyak masalah yang perlu dikonsultasikan.
- Gender bukan faktor penting dalam prediksi: tidak cukup bukti statistik bahwa pria atau wanita berbeda secara signifikan dalam perilaku konsultasi.

```
# Ekspektasi (mu_hat) dari model
studi_kasus2$mu_hat <- predict(model_kasus, type = "response")

# Variansi dari distribusi binomial: mu*(1 - mu)
studi_kasus2$var_hat <- studi_kasus2$mu_hat * (1 - studi_kasus2$mu_hat)

# Lihat sebagian data
head(studi_kasus2[c("Umur", "Gender", "mu_hat", "var_hat")])
```

```
##   Umur Gender    mu_hat   var_hat
## 1   41      1 0.3230391 0.2186848
## 2   55      0 0.4992178 0.2499994
## 3   70      1 0.5993210 0.2401353
## 4   18      0 0.1883493 0.1528738
## 5   49      0 0.4404087 0.2464489
## 6   55      0 0.4992178 0.2499994
```

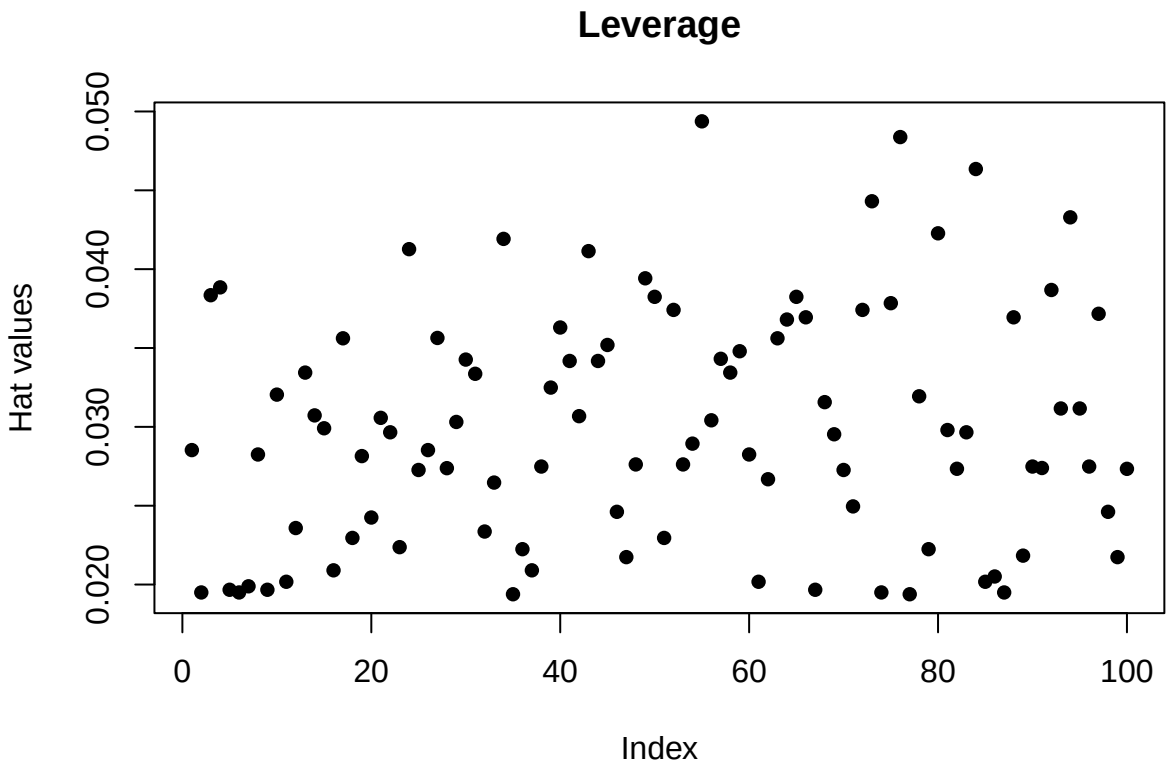
```r
# Deviance residuals
plot(residuals(model_kasus, type = "deviance"),
     main = "Deviance Residuals",
     ylab = "Deviance", pch = 16, col = "blue")
abline(h = 0, col = "red", lty = 2)
```



Interpretasi Gambar:

- Titik biru tersebar cukup merata di sekitar garis horizontal nol:
  - Ini menunjukkan tidak ada pola sistematis yang kuat dalam residual.
  - Artinya model cukup baik dalam menangkap tren data.
- Sebagian besar residu berada antara -1.5 dan 1.5, tidak ada outlier mencolok:
  - Tidak ada indikasi pengamatan yang sangat tidak sesuai dengan model.
  - Ini pertanda tidak ada masalah ekstrem seperti leverage tinggi atau outlier kasar.
- Distribusi acak dari titik-titik:
  - Menunjukkan asumsi model terpenuhi dengan cukup baik, terutama tidak adanya pola yang menyarankan hubungan non-linear antara prediktor dan logit(p).

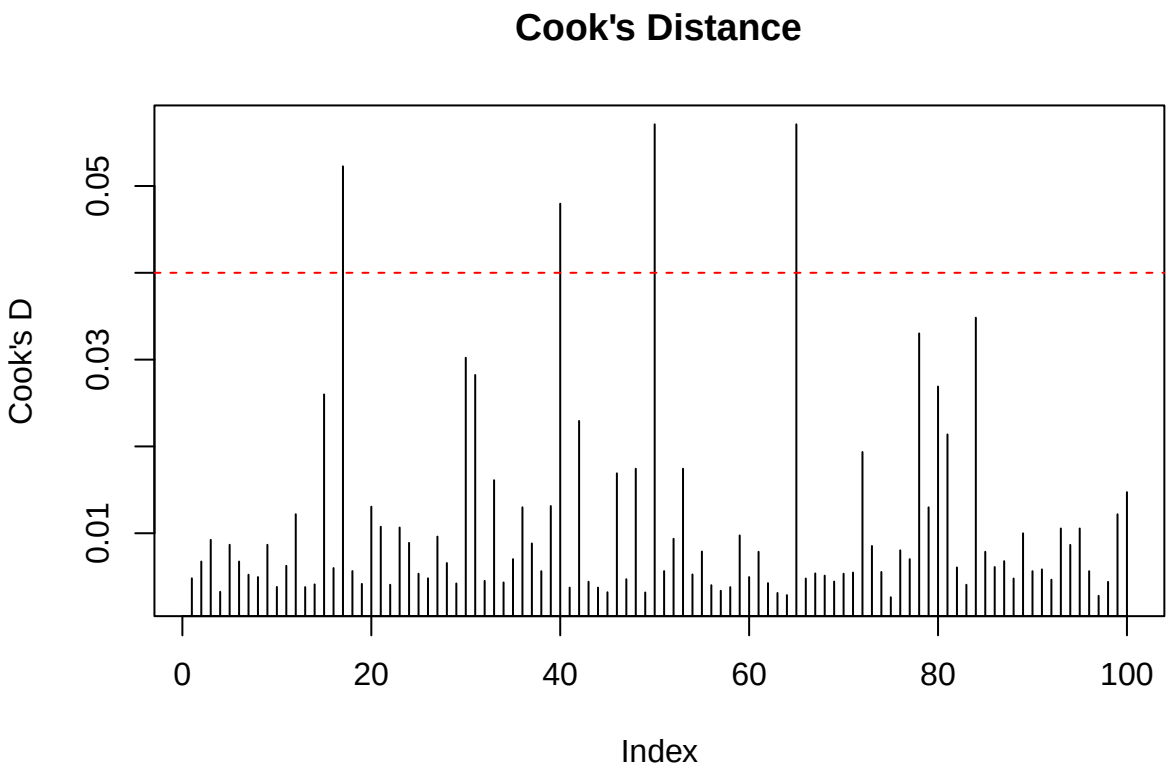
```r
# Leverage (hat values)
hat_values <- hatvalues(model_kasus)
plot(hat_values, main = "Leverage", ylab = "Hat values", pch = 16)
```



Interpretasi

- Tidak ada pengamatan yang menunjukkan leverage mencurigakan.
- Tidak ada kebutuhan mendesak untuk menghapus atau memberi perhatian khusus terhadap observasi tertentu.
- Ini mendukung bahwa model regresi logistik yang dibangun stabil dan tidak bergantung pada data ekstrim.

```
# Cook's Distance
cooks <- cooks.distance(model_kasus)
plot(cooks, type = "h", main = "Cook's Distance", ylab = "Cook's D")
abline(h = 4 / nrow(studi_kasus2), col = "red", lty = 2)
```



Interpretasi

- Ada beberapa batang (pengamatan) yang melewati batas merah.
- Namun tidak ada nilai ekstrem (misal Cook's D > 1), yang biasanya benar-benar menunjukkan influential point berat.

```r
# Model null (tanpa prediktor)
null_model_kasus <- glm(Konsultasi ~ 1, family = binomial, data = studi_kasus2)

# Model penuh (dengan Gender dan Umur)
model_kasus <- glm(Konsultasi ~ Gender + Umur, family = binomial, data = studi_kasus2)

# Bandingkan model menggunakan uji Likelihood Ratio (Chi-square)
anova(null_model_kasus, model_kasus, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Konsultasi ~ 1
## Model 2: Konsultasi ~ Gender + Umur
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        99     137.63
## 2        97     125.04  2   12.591 0.001844 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hasil Tabel Analysis of Deviance:

Model	Residual Df	Residual Deviance	Df	Deviance	Pr(>Chi)
Model 1	99	137.63			
Model 2	97	125.04	2	12.591	0.001844

Interpretasi:

- **Penurunan deviance**:
  Total deviance berkurang dari 137.63 → 125.04, yaitu 12.591 poin dengan penambahan 2 variabel.
- Nilai p = 0.001844, yang sangat signifikan (p < 0.01) → ditandai dengan **.

Kesimpulan:

- Model 2 secara signifikan lebih baik daripada Model 1.
- Artinya, Gender dan Umur secara bersama-sama memberikan kontribusi yang signifikan dalam menjelaskan variasi dalam variabel respon Konsultasi.
- Dengan kata lain, ada bukti kuat bahwa minimal salah satu dari Gender atau Umur berpengaruh terhadap probabilitas konsultasi.

```
AIC(model_kasus)
```

```
## [1] 131.0363
```

```
AIC(null_model_kasus)
```

```
## [1] 139.6278
```

```
BIC(model_kasus)
```

```
## [1] 138.8518
```

```
BIC(null_model_kasus)
```

```
## [1] 142.2329
```

Nilai AIC dan BIC:

Model	AIC	BIC
Model Null	139.63	142.23
Model Kasus (Konsultasi ~ Gender + Umur)	131.04	138.85

Interpretasi AIC (Akaike Information Criterion):

- Semakin rendah AIC, semakin baik model dalam hal keseimbangan antara kompleksitas dan goodness-of-fit.

- AIC Model dengan variabel (131.04) lebih rendah dari AIC Model Null (139.63).
- Model dengan Gender dan Umur lebih baik secara keseluruhan dibandingkan model yang hanya memiliki intercept.

Interpretasi BIC (Bayesian Information Criterion):

- Semakin rendah BIC, semakin baik model dalam konteks informasi Bayesian (lebih penal terhadap kompleksitas model dibanding AIC).
- BIC Model dengan variabel (138.85) < BIC Model Null (142.23).
- Model dengan variabel tetap lebih unggul, meskipun penalti BIC terhadap jumlah parameter lebih besar.

Kesimpulan:

- Baik AIC maupun BIC mendukung penggunaan model dengan prediktor Gender dan Umur.
- Ini memperkuat hasil dari Analysis of Deviance sebelumnya, bahwa penambahan Gender dan Umur memberikan model yang secara signifikan lebih baik.

XII. Referensi

- ESS Datafile. (2025). *European Social Survey*. Retrieved from https://ess.sikt.no/en/datafile/242aaa393bbb-40f5-98bf-bfb1ce53d8ef
- Wiley Series. (2012). *An Introduction to Categorical Analysis* (2nd ed.). Wiley Series.
- I Gede Nyoman Mindra Jaya (2025). *Analisis Data Kategori*. E-Book.
- Agresti, A. (2007). *An Introduction to Categorical Data Analysis* (2nd ed.). Wiley.
- Agresti, A. (2013). *Categorical Data Analysis* (3rd ed.). Wiley.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.
- Kleinbaum, D. G., & Klein, M. (2010). *Logistic Regression: A Self-Learning Text* (3rd ed.). Springer.
- Powers, D. M. W. (2011). *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation*. Journal of Machine Learning Technologies.
- Lemeshow, S., & Hosmer, D. W. (1982). *A Review of Goodness of Fit Statistics for Use in the Development of Logistic Regression Models*. American Journal of Epidemiology.
- Wright, R. E. (1995). *Logistic Regression*. In *Reading and Understanding Multivariate Statistics*. American Psychological Association.
- Mantel, N., & Haenszel, W. (1959). *Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease*. Journal of the National Cancer Institute.
- Friendly, M. (2000). *Visualizing Categorical Data*. SAS Institute.