LAPORAN TUGAS BESAR
STATISTIKA MULTIVARIAT DENGAN R

Penentuan Faktor Penunjang Karir Pemain Badminton

Kelompok - 2
Michael Alvino Wijaya Darmawan - 6181801041
Yalvi Hidayat - 6181801044
Shannas Rizqi Raihan - 6181801046
Ame Fedora Ignacia Ginting - 6181801047
Mario Bagus Prasetya - 6181801071

UNIVERSITAS KATOLIK PARAHYANGAN
FAKULTAS TEKNIK INFORMATIKA DAN SAINS
JURUSAN TEKNIK INFORMATIKA

Abstract

Badminton merupakan salah satu cabang olahraga yang sangat populer dan banyak digemari tanpa melihat usia. Saat ini badminton memiliki asosiasi bernama *Badminton World Federation* (BWF) yang menjadi pengurus dari berbagai penyelenggaraan kegiatan dan administrasi acara badminton di tingkat dunia. Melihat BWF melakukan pencatatan pencapaian karir dan data-data pribadi dari setiap atlet yang mengikuti pertandingan, mengundang pertanyaan apakah data yang dicatat ada yang memiliki pengaruh terhadap total pencapaian prestasi kari dari atlet-atlet *ranking* papan atas dunia, maka dari itu penelitian ini akan melakukan penelitian terhadap data atlet yang dicatat oleh BWF untuk melihat apakah diantara data tersebut ada yang mempengaruhi pencapaian prestasi karir. Untuk mencari apakah terdapat data yang memiliki pengaruh terhadap pencapaian prestasi karir seorang atlet, digunakan metode regresi linear. Regresi linear merupakan salah satu alat statistik untuk memodelkan data numerik dan pada penelitian ini akan dicari sebuah model yang dimiliki. Hasil regresi yang sudah kami buat memiliki nilai AIC sebesar 1259.404 dengan seluruh asumsi yang telah terpenuhi.

Kata Kunci : *Badminton, Regresi Linear, Permodelan Box-Cox, Diagnosa Asumsi Model*

Daftar Isi

1 Pendahuluan

1.1 Latar Belakang

Badminton merupakan salah satu cabang olahraga yang sangat populer dan banyak digemari tanpa melihat usia. Saat ini badminton memiliki asosiasi bernama *Badminton World Federation* (BWF) yang menjadi pengurus dari berbagai penyelenggaraan kegiatan dan administrasi acara-acara badminton dunia dari tingkat *junior* maupun senior. BWF berdiri pada tahun 1934 dengan 9 anggota meliputi Kanada, Denmark, Inggris, Prancis, Irlandia, Belanda, Selandia Baru, Skotlandia, dan Wales. Namun kini anggotanya bertambah hingga 165 asosiasi bulu tangkis negara dari belahan dunia. Seluruh atlet yang berpartisipasi dalam pertandingan yang diselenggarakan resmi oleh BWF akan dicatat pencapaian karirnya dari pertandingan internasional tingkat rendah hingga event terbesar yang diselenggarakan oleh BWF, yaitu *Summer Olympics*. Selain melakukan pencatatan pencapaian karir, BWF juga mencatat data pribadi dari setiap atlet seperti jumlah *medal*, *ranking* dunia tertinggi dari karir, kategori dari bidang badminton yang paling dikenali dari setiap atlet, negara asal, dsb.

Melihat BWF melakukan pencatatan pencapaian karir dan data-data pribadi dari setiap atlet yang mengikuti pertandingan, mengundang pertanyaan apakah data yang dicatat ada yang memiliki pengaruh terhadap total pencapaian prestasi kari dari atlet-atlet *ranking* papan atas dunia? Maka dari itu penelitian ini akan melakukan penelitian terhadap data atlet yang dicatat oleh BWF untuk melihat apakah diantara data tersebut ada yang mempengaruhi pencapaian prestasi karir.

Untuk mencari apakah terdapat data yang memiliki pengaruh terhadap pencapaian prestasi karir seorang atlet, digunakan metode regresi linear. Regresi linear merupakan salah satu alat statistik untuk memodelkan data numerik. Pada penelitian ini akan dicari sebuah model yang dimiliki.

1.2 Tujuan Penelitian

Penelitian ini dilakukan untuk mencari hubungan antara atribut dan menentukan model mana yang terbaik untuk data `BWF_Player19` yang berisi data atlet badminton internasional dari berbagai negara yang tercatat masih aktif setidaknya sampai awal tahun 2019.

2 Metodologi

2.1 Catatan Teknis

2.1.1 Deskripsi Data

Dataset bernama `BWF_Players19.csv` berisi data 313 atlet badminton internasional dari berbagai negara yang tercatat masih aktif setidaknya sampai awal tahun 2019 (maksimum pensiun awal 2019). Data ini memiliki 12 kolom sebagai berikut:

No	Nama Kolom	Keterangan
1	*Name*	Nama lengkap dari atlet badminton
2	*Age*	Umur dari atlet badminton
3	*Height*	Tinggi badan atlet (dalam cm)
4	*Gender*	Jenis kelamin atlet M = *Male* F = *Female*
5	*Category*	Kategori dari bidang badminton yang paling dikenali dari para masing-masing atlet: MS = *Men's Singles* (Tunggal Putra) WS = *Women's Singles* (Tunggal Putri) MD = *Men's Doubles* (Ganda Putra) WD = *Women's Doubles* (Ganda Putri) XD = *Mixed Doubles* (Ganda Campuran)
6	*Country*	Negara asal atlet (29 negara)
7	*Continent*	Benua asal atlet *Asia*, *Europe* = Eropa, *Pan Am* = Amerika, *Oceania*
8	*Hand*	Tangan dominan yang dipakai atlet *Right* = Kanan *Left* = Kidal
9	*Medals*	Jumlah medali emas dari *major events* yang pernah diraih atlet dari tingkat Junior hingga Senior (contoh: Juara Dunia, Olimpiade, Sudirman Cup, *Thomas Uber* Cup, Piala Dunia, Kejuaraan Dunia Junior, *Suhandinata* Cup, *Asian Games*, *European Games*, *Pan Am Games*, *SEA Games*, dll)
10	*HRank*	*Ranking* dunia tertinggi dari karir masing-masing atlet
11	*Multiple*	Status apakah atlet dikenal bermain rangkap (lebih dari satu kategori), contoh: MD dan XD, WD dan XD, dsb 1 = *Yes* 0 = *No*
12	*Career*	Jumlah karir kemenangan dari semua babak pertandingan yang pernah dimenangkan oleh masing-masing atlet secara keseluruhan

2.1.2 Preparasi Data

- Pertama *import library* yang akan digunakan untuk mempermudah eksekusi data. *Library* yang digunakan adalah `ggplot`.
- Setelah memasukkan *library*, kami memasukkan *dataframe* yang akan digunakan untuk eksperimen.
- Untuk mengatasi data `NA`, kami mencari data yang ada di *website* BWF dan jika tidak ditemukan maka kami mengganti nilai `NA` dengan rata-rata pada kolom tersebut.
- Setelah itu, kami menemukan data-data yang nilainya tidak masuk akal seperti ada pemain dengan *Career/Height/Age* yang terlalu besar atau terlalu kecil. Cara kami mengatasinya adalah dengan mengganti data yang salah dengan data yang ada di *website* BWF

2.1.3 Analisis Data Eksplorasi

Analisis Data Eksploratif merupakan suatu alat menganalisis yang berguna untuk melihat pola suatu data sehingga mengetahui beberapa informasi awal terkait dengan data tersebut, yang nantinya akan

berguna untuk melakukan pengujian pada data tersebut dengan statistik inferensial[1]. Menurut Ronald K. Pearson pada bukunya *Exploratory Data Analysis Using R*[2], setidaknya terdapat tiga motivasi untuk melakukan analisis data yaitu :

1. Untuk memahami apa yang telah dan sedang terjadi.
2. Untuk memprediksi apa yang akan terjadi, baik di masa yang akan datang atau keadaan yang belum terlihat.
3. Untuk memberikan panduan dalam mengambil keputusan.

Namun pada bagian ini, analisis data eksploratif paling berguna dalam motivasi pertama yaitu memahami data. Untuk dapat memahami data hal yang dapat dilakukan menurut Diaconis[14] adalah melihat angka atau grafik dan mencoba menemukan pola, mengejar petunjuk yang disarankan oleh informasi latar belakang, imajinasi, pola yang dirasakan dan pengalaman analisis data lainnya [12]. Namun, meskipun terdapat data non numerik pada data yang akan dianalisis, analisis akan tetap dapat dilakukan dengan sebagian besar didasarkan pada karakteristik numerik yang dihitung dari nilai-nilai non numerik tersebut. Saat ini banyak alat eksplorasi yang telah dikembangkan, yang memungkinkan kita untuk dapat melakukan analisis pada variabel non numerik serta hubungannya dengan variabel lain, baik variabel kategorik maupun numerik.

Penyebutan ”grafik” pada ungkapan Diaconis[14] sangat penting karena manusia jauh lebih baik dalam melihat pola dalam grafik daripada melihat sekumpulan angka yang besar [12]. Terdapat berbagai macam tampilan grafik, bergantung pada kebutuhan dan data yang digunakan. Data merupakan kumpulan pada satu atau lebih variabel, dan variabel adalah karakteristik yang nilainya dapat berubah dari satu observasi ke observasi lainnya [12]. Kumpulan data yang terdiri dari pengamatan pada satu karakteristik disebut sebagai data univariat.

Terdapat dua tipe data univariat yaitu, kategorikal(kualitatif) dan numerikal(kuantitatif). Sebuah data univariat dikatakan memiliki tipe kategorikal apabila observasi individu merupakan kategorik dan sebuah data univariat dikatakan memiliki tipe numerikal apabila setiap observasinya berupa angka. Sekumpulan data dikatakan bivariat apabila terdiri dari dua variabel pengukuran atau observasi [12]. Sekumpulan data disebut sebagai multivariat apabila terdapat dua variabel atau lebih yang masing-masingnya menghasilkan kategori atau nilai. Sehingga dapat disimpulkan bahwa data bivariat merupakan bagian dari data multivariat.

Grafik yang biasanya digunakan untuk melihat distribusi data *univariat* bertipe numerik adalah *histogram*, *density plot*, *dot chart*, dan *boxplot*. Sedangkan untuk melihat distribusi *univariat* bertipe kategorik yang biasanya adalah *barplot* dan *pie chart*. Grafik yang biasanya digunakan untuk melihat distribusi data *bivariat* bertipe kategorik dan kategorik adalah *stacked bar chart*, *grouped barplot*. Distribusi *bivariat* bertipe kategorik dan numerik biasanya menggunakan *barchart*

2.2 Pembuatan Model

Pembuatan model pada penelitian ini dilakukan menggunakan metodologi regresi linear.

2.2.1 Regresi Linear

Regresi linear merupakan metodologi statistik yang memanfaatkan hubungan antara dua variabel kuantitatif atau lebih sehingga variabel keluaran dapat diprediksi dari variabel lainnya.[13]

Regresi Linear Sederhana Regresi linear sederhana adalah hubungan secara linear antara satu variabel independen (X) dengan variabel dependen (Y). Analisis ini digunakan untuk mengetahui arah hubungan antara variabel independen dengan variabel dependen apakah positif atau negatif serta untuk memprediksi nilai dari variabel dependen apabila nilai variabel independen mengalami kenaikan

atau penurunan nilai. Data yang digunakan biasanya berskala interval atau rasio. Sebuah regresi linear sederhana memiliki persamaan sebagai berikut:

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

Dimana :

- y' adalah nilai respon
- β_0 adalah konstanta regresi
- β_1 adalah koefisien regresi; besaran respon yang ditimbulkan oleh prediktor
- x nilai dari variabel prediktor

Regresi Linear Berganda Regresi linear berganda (*multiple linear regression*) adalah hubungan secara linear antara dua atau lebih variabel independen ($X_1, X_2, ..., X_n$) dengan variabel dependen (Y). Analisis ini digunakan untuk mengetahui arah hubungan antara variabel independen dengan variabel dependen apakah masing-masing variabel independen berhubungan positif atau negatif dan untuk memprediksi nilai dari variabel dependen apabila nilai variabel independen mengalami kenaikan atau penurunan. Data yang digunakan biasanya berskala interval atau rasio. Sebuah regresi linear sederhana memiliki persamaan sebagai berikut:

$$y' = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

Dimana :

- y' adalah nilai respon
- β_0 adalah konstanta regresi
- β_1 adalah koefisien regresi; besaran respon yang ditimbulkan oleh prediktor
- x nilai dari variabel prediktor

2.3 Pemilihan Model

2.3.1 Algoritma *Stepwise* & *Backward*

Pada regresi *linear*, terdapat algoritma untuk menentukan model yang terbaik.

1. *Backward*
   Metode *backward*, adalah memasukkan semua prediktor kemudian mengeliminasi satu persatu hingga tersiswa prediktor yang signifikan saja. Eliminasi didasarkan pada prediktor yang memiliki nilai sig F yang diatas 0.1 [6]. Dalam R, algoritma *backward* ini melakukan eleminasi variabel yang kurang berpengaruh agar mendapatkan AIC terendah.

2. *Stepwise*
   Metode *stepwise* adalah memasukkan prediktor secara bertahap berdasarkan nilai F yang signifkan (sig F di bawah 0.05). Setelah dimasukkan lalu dikeluarkan lagi. Proses memasukkan dikombinasikan dengan mengeliminasi prediktor yang tidak signifikan (sig F di atas 0.01). [6]. Dalam R, algoritma *stepwise* ini melakukan eliminasi variabel yang kurang berpengaruh atau dapat memasukkan kembali variabel yang berpengaruh untuk mendapatkan AIC terendah.

Dalam penelitian ini, kami menggunakan algoritma *stepwise* ini untuk mencari model terbaik dengan nilai AIC terendah.

2.4 Diagnosa Model

Untuk mengetahui apakah model yang kita buat sudah baik atau tidak, kita dapat menggunakan uji asumsi klasik. Menurut Hayes(2013), tujuan uji asumsi bukan untuk mengetahui kita telah melanggar suatu asumsi, tetapi untuk mengetahui seberapa besar kemungkinan kita melakukan hal yang akan menyesatkan ketika kita akan menafsirkan hasil-hasil kita dan kesimpulan-kesimpulan yang kita buat darinya.Ia juga mengatakan dalam melakukan uji asumsi ini,kita tetap harus menghormati kompleksitas dan sifat data yang kita punya dan melakukan yang terbaik untuk menganalisisnya dengan metode yang paling cocok, tetapi jangan terobsesi dengan setiap pelanggaran asumsi yang kecil. [5] Oleh karena itu, kita membutuhkan pengujian asumsi agar model yang kita buat tidak menyesatkan atau bias, tetapi tetap mempertimbangkan untuk tidak sembarangan membuang data yang kita punya. Uji asumsi yang digunakan pada penelitian ini adalah uji linearitas, uji normalitas, uji homoskedastisitas, uji independensi.

2.4.1 Uji Asumsi Linearitas

Uji linearitas adalah sebuah uji untuk melihat hubungan yang *linear* antara variabel bebas dengan *output*. Menurut Ghozali, uji linearitas digunakan untuk melihat apakah spesifikasi model yang digunakan sudah benar atau tidak. Fungsi yang digunakan dalam suatu studi empiris sebaiknya berbentuk *linear*, kuadrat atau kubik [8]. Untuk mengetahui apakah semua variabel bebas berhubungan *linear* dengan variabel *output*, mengujinya dengan membuat plot antara *residual* dari model dengan nilai ekspentasi *output*.

2.4.2 Uji Asumsi Normalitas

Uji normalitas adalah sebuah uji yang dapat digunakan untuk mengetahui apakah variabel yang sedang diteliti itu merupakan terdistribusi normal atau tidak [7]. Menurut beberapa ahli statistik, data yang banyaknya lebih dari 30, sudah data dipastikan data tersebut sudah normal. Namun dalam hal ini, masih terdapat sebuah masalah yang belum terselesaikan. Bagaimana data tersebut itu terdistribusi? Apakah 30 data tersebut datanya memang terdistribusi normal atau angkanya sama-sama saja? Perlu ada pengujian lanjut untuk menguji data tersebut normal atau tidak. Jika tidak diuji dan ternyata data yang diolah ternyata tidak terdistribusi normal, data tersebut dapat dianggap sebagai data yang menyesatkan dan bias. Terdapat berbagai macam uji normalitas yang sudah digunakan. Antaranya adalah uji *Chi-Square*, *Kolomgorov Smirnov*, *Lillefors*, *Shapiro-Wilk*, dan *Jarque Bera* [9].

Dalam penelitian ini, kami menggunakan uji *Shapiro–Wilk* untuk menguji apakah variabel yang sedang diteliti normal atau tidak.

2.4.3 Uji Asumsi Homoskedasitisitas

Uji homoskedasititas merupakan uji yang melihat tingkat ke-konstanan suatu variansi yang bertujuan juga untuk melihat *error* dalam suatu model [7]. Jika model yang dibuat ternyata tidak konstan atau bersifat heteroskedastik, dapat menyebabkan *error* tipe I—sebuah *error* dimana terjadi *false positive* dimana salah dalam penarikan kesimpulan dari *null hypothesis* dengan menolak *null hypthesis* yang benar [10]. Dalam penelitian ini, kami menggunakan uji *Breusch–Pagan*.

2.4.4 Uji Asumsi Independensi

Uji independensi adalah suatu uji dimana untuk melihat setiap *residual* tidak dipengaruhi oleh *residual* data lainnya dalam suatu model regresi. Akan tetapi, uji asumsi ini jarang digunakan karena obstervasi yang sudah kita kumpulkan tentu nilainya tidak bergantung pada observasi lain.

2.5 Transformasi Model

Transformasi dilakukan jika asumsi-asumsi yang kita lakukan tidak terpenuhi. Kita dapat mentransformasikan variabel *output* [15] kita dengan mencari *lambda* dari fungsi *log-Likelihood* untuk mencari transformasi terbaik atau kita dapat mentransformasikan model dengan selang kepercayaan yang ada di fungsi *log-Likelihood*.

2.5.1 Tranformasi *Log*

Transformasi *log* mengaplikasikan nilai logaritma natural pada *output* dan dimodelkan, kemudian kita lakukan uji asumsi lagi.

2.5.2 Transformasi *Box-cox*

Transformasi *Box-Cox* adalah transformasi pangkat pada respon. [11] Jadi kita dapat mentranformasikan variabel *output* dari selang fungsi *log-Likelihood*.

2.6 Multikolinearitas

Multiklinearitas menyatakan bahwa setidaknya ada 3 atau lebih variabel yang saling terkorelasi satu sama lainnya terjadi jika variabel bebas dalam model regresi setidaknya ada 3 atau lebih variabel yang memiliki korelasi yang sangat kuat.

2.7 Penggunaan Perangkat Lunak

Perangkat lunak yang digunakan pada penelitian ini untuk melakukan analisis statistik adalah R studio, penulisan laporan penelitian menggunakan LaTeX.

3 Hasil

3.1 Analisis Data Eksploratif

"Analisis data eksploratif adalah pekerjaan detektif—pekerjaan detektif numerik—atau menghitung pekerjaan detektif—atau pekerjaan detektif *frafis*" [Tukey, 1977, page 1]. Sebuah filosofi analisa data dimana seorang peneliti melakukan pengujian terhadap data dalam rangka menemukan sesuatu yang dapat memberikan informasi mengenai penelitian yang dilakukan.

Dalam hal ini pertama kali kita melihat normalitas data dengan mengujinya dengan *shapiro-test* Kode yang akan kami pakai adalah `shapiro.test(<namaKolom>).` `<namaKolom>` dalam hal ini merupakan nama kolom dalam suatu tabel yang diujikan. Tipe data `<namaKolom>` diharuskan memiliki tipe data numerik[1]. Oleh karena itu, kami menggunakan kode program ini untuk kolom yang memiliki tipe data numerik.

Atribut	p-value	Keterangan
Age	$2.2 \times e^{-16}$	Alpha yang kami gunakan adalah 0.05. Karena p-value kurang dari 0.05 maka dapat dinyatakan bahwa data ini tidak berdistribusi normal.
Height	0.003544	Alpha yang kami gunakan adalah 0.05. Karena p-value kurang dari 0.05 maka dapat dinyatakan bahwa data ini tidak berdistribusi normal.
Medals	$2.2 \times e^{-16}$	Alpha yang kami gunakan adalah 0.05. Karena p-value kurang dari 0.05 maka dapat dinyatakan bahwa data ini tidak berdistribusi normal.
HRank	$1.015 \times e^{-15}$	Alpha yang kami gunakan adalah 0.05. Karena p-value kurang dari 0.05 maka dapat dinyatakan bahwa data ini tidak berdistribusi normal.
Career	$1.3 \times e^{-13}$	Alpha yang kami gunakan adalah 0.05. Karena p-value kurang dari 0.05 maka dapat dinyatakan bahwa data ini tidak berdistribusi normal.

Setelah melihat distribusi data dengan uji *Shapiro–Wilk*, kami membuat plot secara *univariat* untuk satu buah atribut dan plot secara *bivariat* untuk hubungan antara dua buah atribut dengan menggunakan keseluruhan atribut dalam data `BWF_Players19`, untuk melihat visualisasi persebaran data.

3.1.1 Plot *Univariat*

(a) (b)

Figure 1: Plot Univariat

Berdasarkan *piechart* yang telah dibuat di bagian 1a, Pemain bulu tangkis didominasi oleh pemain berjenis kelamin pria dibandingkan berjenis kelamin perempuan. Oleh karena itu, pemain bulu tangkis BWF didominasi oleh pemain laki-laki maupun perempuan. Untuk gambar bagian 1b, pemain dengan kategori XD—dimana merupakan kategori pemain campuran—lebih banyak mendominasi dibandingkan varian kategori lainnya.

(a) (b)

Figure 2: Plot Univariat

Berdasarkan *Barplot* 2a, diketahui bahwa pemain dengan *Right hand* memiliki jumlah lebih banyak dari pada pemain *Left hand*. Hasilnya cukup signifikan perbedaannya dimana sangat banyak pemain yang tercatat di situs BWF yang menggunakan tangan kanan dengan pemain yang menggunakan tangan kanan. Berdasarkan *histogram* pada gambar 2b, *Ranking* tertinggi para pemain berdistribusi terbanyak pada 0-5. Hal ini berarti kebanyakan pemain yang tercatat dalam situs BWF, setidaknya sudah banyak sekali pemain yang merasakan *ranking* 5 besar. Sangat sedikit sekali jumlah pemain yang berada di *ranking* 40 ke atas. Mungkin saja pemain yang memiliki *ranking* di atas 40–dalam kasus ini angka *ranking* besar, *ranking* jelek–merupakan pemain yang baru bergabung di dunia bulu tangkis.

3.1.2 Plot *Bivariat*

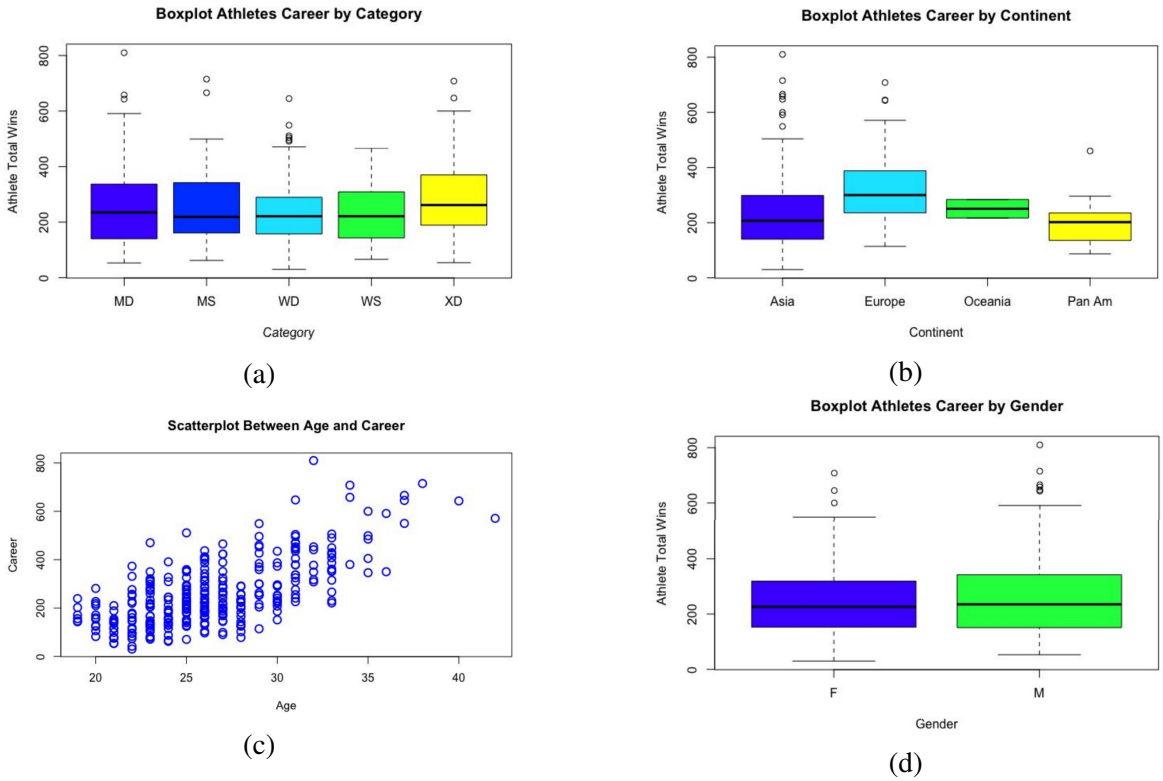

Figure 3: Plot *Bivariat*

Berdasarkan plot 3a diketahui bahwa rata-rata kategori terhadap *career* hampir sama. Meskipun nyaris sama jika dilihat dari distribusi *mean*, tetapi *range career* di tiap kategori memiliki *upper fence* yang berbeda-beda. Misalnya pada kategori MD, range *career* dari dianggap sebagai *outlier* di bawah *bottom fence* hingga *upper fence* cukup berbeda dengan kategori WS. Pada kategori MD, rentang *career* yang biasanya ditemui berkisar 0 - 600, sedangkan pada kategori WS, rentang *career* biasanya berada di rentang 50 - 500. Hal ini berarti pada kategori MD pemain dapat memiliki *career* antara 0 - 600 dibandingkan pemain WS hanya memiliki rentang dari 50 - 500

Pada plot 3b, rata-rata eropa paling tinggi diantara kontinen yang lainnya tetapi rentang di Asia lebih luas dibanding yang lainnya. Pemain Asia memiliki rentang dari 0 - 550, tetapi mengandung banyak sekali *outliers*. Pemain Eropa memiliki rentang sekitar 100 - 600, tetapi pemain Eropa tidak memiliki pemain yang karirnya di bawah 100. Hal ini menandakan, setidaknya pemain Eropa tidak ada yang tidak pernah menang atau pun jumlah kemenangannya di bawah 100.

Pada plot 3c terdapat pola yang terus naik pada umur terhadap *career*. Jika diperhatikan lebih baik lagi, *scatterplot* yang dihasilkan memiliki pola yang sangat terlihat bahwa faktor umur menjadi

faktor yang penting dalam menentukan tingkatan *career* seorang pemain bulu tangkis. Tidak ada titik dalam *scatterplot* tersebut yang berantakan atau terlalu jauh posisi titiknya dan terlihat terdapat pola persamaan garis yang terbentuk. Berdasarkan plot itu juga, nilai korelasi antara 2 variabel tersebut tergolong kuat dan positif. Kuat dalam hal ini dikarenakan kerapatan antar titik-titik dalam plot tersebut saling berdekatan dan positif karena arah kumpulan titik tersebut mengarah ke arah kanan atas.

Pada plot 3d rata-rata antara perempuan dan laki-laki hampir sama tetapi pada laki-laki rentang nya sedikit lebih luas. Dapat dikatakan juga, hasil *boxplot* tersebut tidak ada pengaruhnya sama sekali. Hal ini berarti nilai bagusnya sebuah *career* tidak ditentukan dari jenis *gender* pemain bulu tangkis yang terdaftar di *website* BWF.

3.1.3 Plot Interaksi

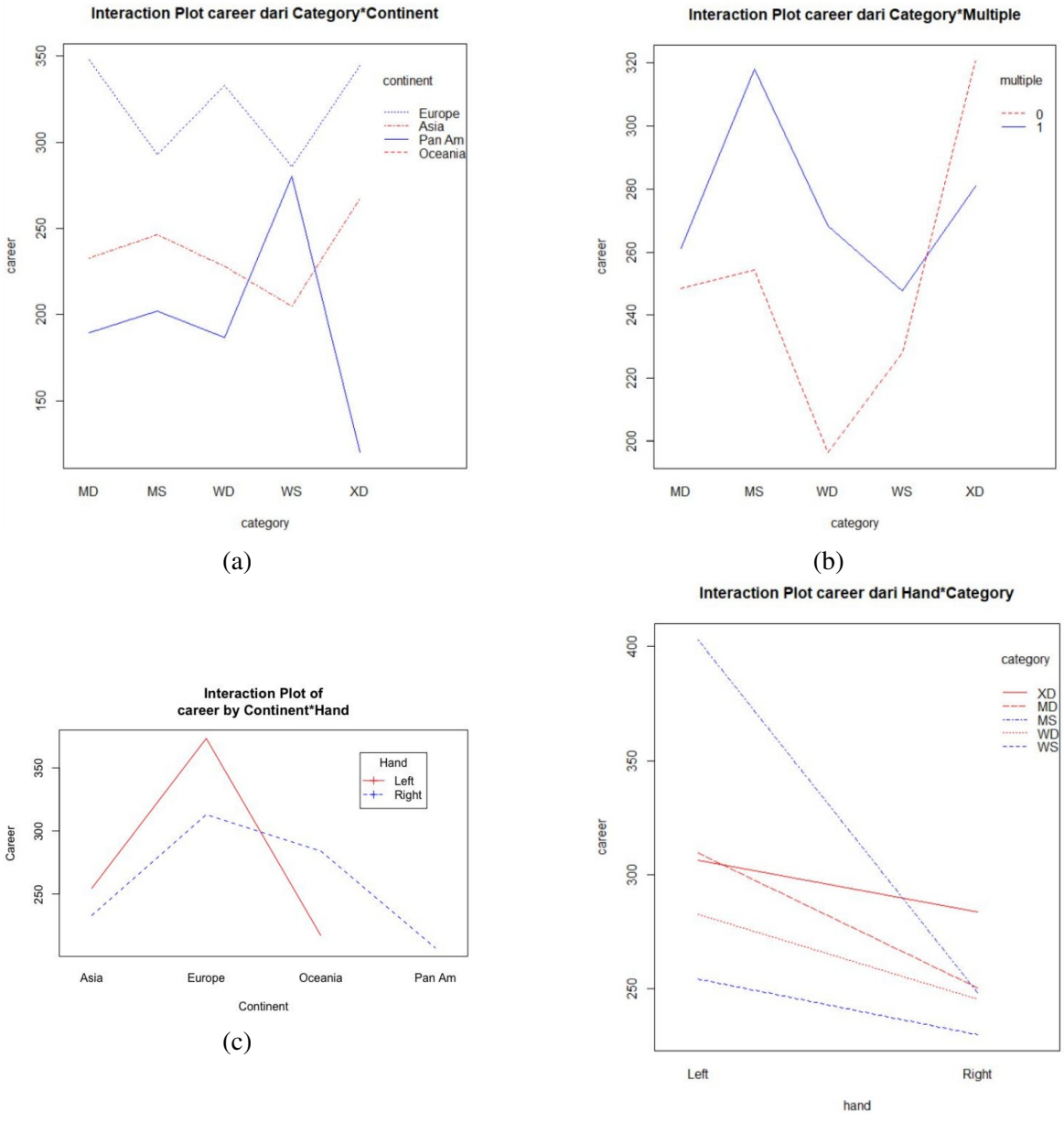

Figure 4: Plot Interaksi

Berdasarkan *interaction plot* 4a, 4b, 4c, 4d terdapat interaksi antara variabel-variabel terhadap *career* karena ada garis yang berpotongan.

4 Pembuatan Model

Dalam sub-bab ini kami menggunakan beberapa variabel untuk dijadikan sebagai model. Ada tiga model yang kami lakukan uji coba untuk menentukan mana model yang paling baik. Masing-masing dari model kami tentukan berdasarkan variabel yang dikira memberikan pengaruh terhadap variabel *career*.

4.1 Permodelan Regresi Linear

Pada pembuatan model pertama kali, kami menggunakan keseluruhan variabel.

No	Model	AIC
1	`Career = Age + Country + Hand + Medals + HRank+ Multiple + Category + Gender + Continent + Height`	3539

Kemudian dimasukkan ke dalam algoritma *stepwise* untuk mengambil variabel yang terbaik.

No	Model	AIC
1	`Career = Age + Country + Multiple + HRank + Medals`	3532

Setelah didapatkan variabel yang terbaik, kami menambahkan interaksi antar variabel yang kami anggap memiliki hubungan yang cukup berpengaruh untuk mengembangkan model yang kami buat. Kami melakukan pemodelan regresi untuk menentukan model yang terbaik dari ketiga model yang telah dibuat kemudian kami mencoba melihat nilai AIC dari setiap model dan didapatkan AIC pada model 1 sebesar 3532, model 2 sebesar 3439, dan model 3 sebesar 3428 yang berada pada Tabel.

No	Model	AIC
1	`Career = Age + relevel(Country, ref = \Indonesia") + Multiple + HRank + Medals`	3532
2	`Career = relevel(Country, ref = \Indonesia") + Multiple + Medals + HRank*Medals + HRank*Age`	3439
3	`Career = HRank*relevel(Country, ref = \Indonesia") + Multiple*Age + HRank*Medals + HRank*Age + Category`	3428

Setelah nilai AIC didapatkan diketahui model yang terbaik berdasarkan nilai AIC terendah adalah model 3, dan model yang terbaik ke dua adalah model 2 dengan AIC 3439. Lalu setelah didapatkan model model yang terbaik, kami mencoba melihat diagnosa apakah model tersebut layak dijadikan sebagai model terbaik yang kami pilih.

4.1.1 Asumsi Model

Pada asumsi model, disini kami mencoba berbagai diagnosa untuk menentukan apakah model yang kami buat memenuhi keseluruhan diagnosa.

4.1.2 Multikolinearitas

Adanya multikolinearitas dapat menyebabkan model pada regresi menjadi tidak stabil, oleh karena itu kami mencoba mengecek model-model yang telah kami buat apakah terdapat masalah multikolinearitas. Kami menggunakan `library(car)` untuk menghitung nilai *Variance Inflation Factor* atau VIF yang akan mengukur seberapa besar variansi dari variabel bebas model regresi kita membesar karena adanya multikolinearitas.

Variabel	GVIF	Df	GVIF ^(1/(2*Df))
`relevel(Country, ref="Indonesia")`	9.722958	28	1.041452
`Multiple`	1.331162	1	1.153760
`Medals`	2.778809	1	1.66976
`Age`	2.707258	1	1.645375
`HRank`	59.018359	1	7.682341
`Medals*HRank`	3.215528	1	1.793189
`Age*HRank`	56.892099	1	7.542685

Kami menggunakan model ke 2 untuk kami lanjutkan ke tahap diagnosa asumsi, karena nilai gvif yang dihasilkan cukup kecil dan hanya beberapa nilai besar namun masih di bawah 10 maka variabel tidak perlu dibuang. Sedangkan untuk model ke 3 kita tidak bisa mengecek masalah multikolinearitas karena nilai multikolinearitas yang sangat besar.

4.1.3 Diagnosa Asumsi Linearitas

Untuk melakukan diagnosa asumsi linearitas kami menggunakan `library(broom)` untuk melihat plot linearitas dari model yang kami buat.

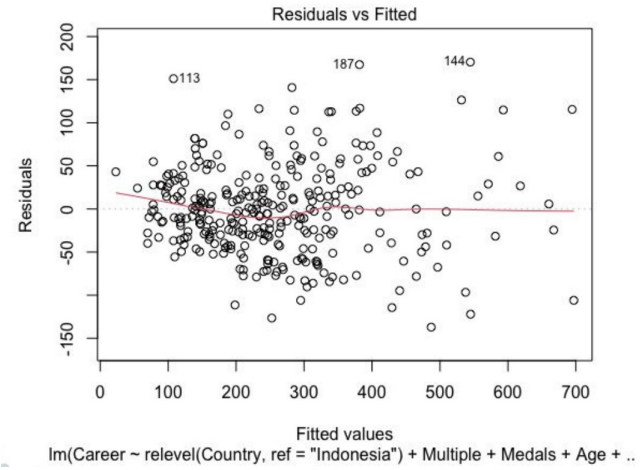

Figure 5: Linearitas

Seperti yang terlihat pada gambar 5, hasil plot menunjukkan bahwa asumsi linearitas masih kurang terpenuhi dikarenakan garis merah dalam plot kurang mendekati nilai 0. Maka dari itu, bisa kita katakan bahwa asumsi linearitas hampir terpenuhi, tetapi masih belum terlalu sejajar dengan garis nol.

4.1.4 Diagnosa Asumsi Normalitas pada Residual

Untuk melakukan diagnosa asumsi normalitas pada residual, kami menggunakan melihat *Q-Q plot* yang merupakan plot antara residual yang telah di standarisasi vs kuantil teoritik yang berdistribusi normal.

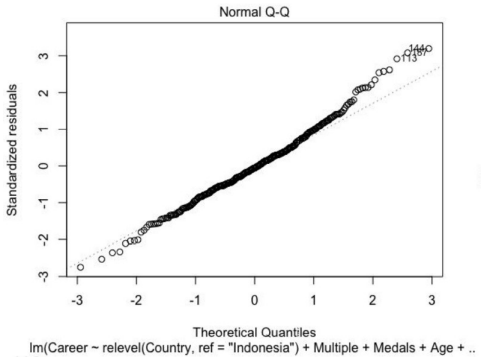

Figure 6: Normalitas

Seperti yang terlihat pada gambar 6, hasil plot menunjukkan bahwa asumsi normalitas belum terpenuhi karena data masih melenceng dari garis diagonal. Dapat diperjelas dengan menggunakan uji *shapiro-wilk* dan didapatkan nilai `p-value` kurang dari 0.05 pada gambar 7, maka dari itu normalitas dari data ini masih belum terpenuhi.

```
        Shapiro-Wilk normality test
data:  resid
W = 0.98977, p-value = 0.02739
```

Figure 7: Shapiro Wilk

4.1.5 Diagnosa Asumsi Homoskedastisitas pada Residual

Untuk melakukan diagnosa asumsi homoskedastisitas pada residual kami menggunakan *library(lmtest)* dan menggunakan *bptest*.

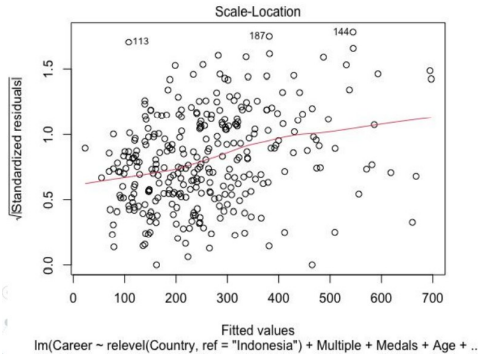

Figure 8: Homoskedasitistas pada Residual

Seperti yang terlihat pada gambar 8, hasil plot menunjukkan bahwa asumsi homoskedastisitas sudah cukup terpenuhi meskipun belum tersebar cukup merata. Maka dari itu, untuk membuktikan lebih detail kita gunakan uji *bptest* untuk melihat nilainya.

```
        studentized Breusch-Pagan test
data:  stepwise1
BP = 62.504, df = 34, p-value = 0.002054
```

Figure 9: Breusch Pagan

Dari hasil pada gambar 9, dapat dilihat bahwa nilai dalam *Breusch-pagan* masih kurang dari 0.05 oleh karena itu belum bisa dikatakan memenuhi asumsi homoskedastisitas.

4.1.6 Diagnosa Asumsi Independensi pada Residual

Untuk melakukan diagnosa asumsi independensi pada residual, asumsi ini menyatakan bahwa setiap residual tidak dipengaruhi oleh residual data lainnya dalam model regresi.

4.1.7 Deteksi Outlier

Dengan menggunakan `boxplot.stats()$out`, data *outlier* akan dikeluarkan beserta nilai residualnya, seperti tabel di bawah ini

No	Data *Outlier*	Nilai Residual
1	97	140.8121
2	113	151.1278
3	133	126.3589
4	144	170.3289
5	187	167.4622
6	261	-137.1064
7	263	-126.5164

Untuk mendeteksi apakah outlier berpengaruh atau tidak terhadap data, kami menggunakan plot residual vs leverage. Untuk *outlier* dapat dideteksi dengan melihat apakah ada nilai *standarized residuals* yang diluar rentang [-3,3] pada sumbu y, dan *high leverage* dideteksi dengan melihat apakah data memiliki nilai yang lebih besar dari $2(p+1)/n$ di sumbu x. Pada gambar 10 didapat bahwa ada beberapa data yang melewati *standarized residual* tetapi tidak begitu jauh, dan untuk *high leverage* tidak melebihi 0.6.

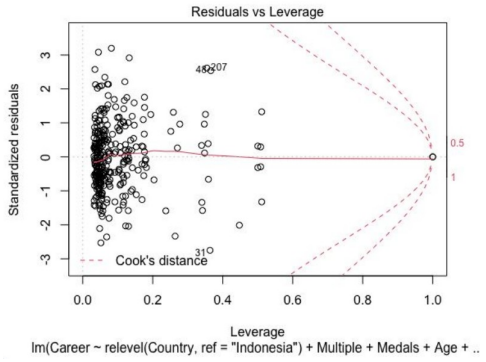

Figure 10: Outlier dan High Leverage

Kita bisa melihat lebih detail apakah *outlier* berpengaruh atau tidak dengan menggunakan `outliertest` pada model yang kita miliki, dan jika dilihat dari gambar 10 ditunjukkan bahwa nilai *bonferroni p* lebih dari 0.05 dan dapat diartikan jika *outlier* dari data ke 144 tidak signifikan berpengaruh pada data, maka dari itu data lebih baik tidak di buang.

Data Outlier	rstudent	Unadjusted p-value	Bonferroni p
144	140.8121	0.0013031	0.40397

4.1.8 Kesimpulan Diagnosa Asumsi

No	Diagnosa Asumsi	Keterangan
1	Linearitas	Hampir Terpenuhi
2	Normalitas pada Residual	Tidak Terpenuhi
3	Homoskedastisitas	Hampir Terpenuhi

4.2 log-Likelihood Function

Karena model yang kami buat masih belum memenuhi diagnosa asumsi, maka dari itu kami melakukan transformasi agar dapat memenuhi asumsi-asumsi tersebut.

Untuk menentukan transformasi mana yang cocok untuk model yang kami buat, disini kami melihat nilai λ dengan menggunakan `library(MASS)` untuk mengecek apakah nilai λ sama dengan 0 atau tidak sama dengan 0. Jika λ sama dengan 0 maka transformasi *log* cocok untuk digunakan, tetapi jika tidak sama dengan 0 maka transformasi *box-cox* lebih cocok untuk digunakan.

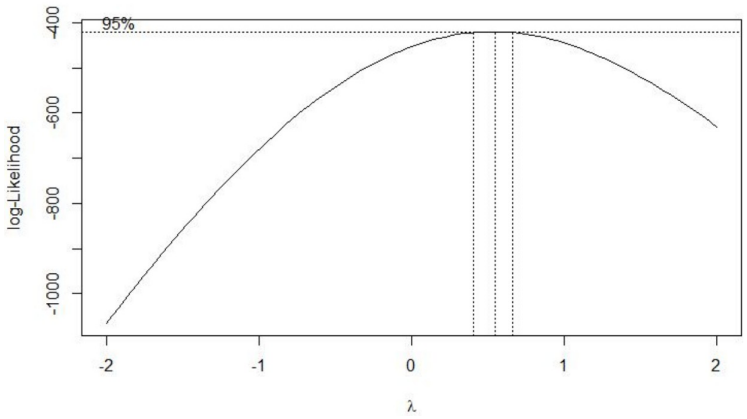

Figure 11: Lamda log-Likelihood

Dalam gambar 11, didapatkan bahwa λ berkisaran di antara 0.5 dan dapat dikatakan bahwa nilai λ = 0.5 adalah nilai terbaik untuk di transformasikan ke output model kita, karena nilai λ = 0.5 maka kita menggunakan transformasi dengan *box-cox*.

4.3 Pemodelan Box Cox

Model yang kita gunakan untuk transformasi *box-cox* sebelum di *stepwise* dan mendapatkan nilai AIC sebesar 1260.842

No	Model	AIC
1	`Career = relevel(Country, ref = \Indonesia") + Multiple + Medals + Age + HRank*Medals + HRank*Age`	1260.842

Model setelah di *stepwise* mendapatkan nilai AIC sebesar 1259.404

No	Model	AIC
1	`Career = relevel(Country, ref = \Indonesia") + Multiple + Medals + Age + HRank*Age`	1259.404

4.3.1 Asumsi Model

Pada asumsi model, disini kami mencoba berbagai diagnosa untuk menentukan apakah model yang kami buat memenuhi keseluruhan diagnosa.

4.3.2 Multikolinearitas

Adanya multikolinearitas dapat menyebabkan model pada regresi menjadi tidak stabil, oleh karena itu kami mencoba mengecek model-model yang telah kami buat apakah terdapat masalah multikolinearitas.

Kami menggunakan `library(car)` untuk menghitung nilai *Variance Inflation Factor* atau VIF yang akan mengukur seberapa besar variansi dari variabel bebas model regresi kita membesar karena adanya multikolinearitas.

Variabel	GVIF	Df	GVIF ^(1/(2*Df))
`relevel(Country, ref="Indonesia")`	5.535123	28	1.031027
`Multiple`	1.323087	1	1.150255
`Medals`	2.397572	1	1.548409
`Age`	2.690624	1	1.640312
`HRank`	58.013378	1	7.616651
`Age*HRank`	56.839931	1	7.539226

Karena nilai gvif yang dihasilkan cukup kecil dan hanya beberapa nilai besar namun masih di bawah 10 maka variabel tidak perlu dibuang.

4.3.3 Diagnosa Asumsi Linearitas

Untuk melakukan diagnosa asumsi linearitas kami menggunakan `library(broom)` untuk melihat plot linearitas dari model yang kami buat.

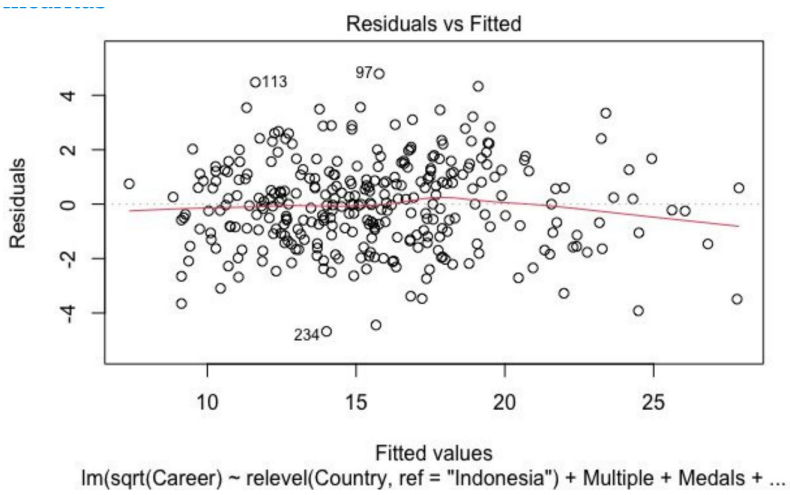

Figure 12: Linearitas Box Cox

Seperti yang terlihat pada gambar 12, hasil plot menunjukkan bahwa asumsi linearitas sudah cukup terpenuhi dikarenakan garis merah dalam plot mendekati nilai 0. Maka dari itu, bisa kita katakan bahwa asumsi linearitas cukup terpenuhi.

4.3.4 Diagnosa Asumsi Normalitas pada Residual

Seperti yang terlihat pada gambar 13, hasil plot menunjukkan bahwa asumsi normalitas sudah terpenuhi karena hampir semua data berada di garis diagonal. Dapat diperjelas dengan menggunakan uji *shapiro-wilk* dan didapatkan nilai `p-value` lebih dari 0.05 pada gambar 14, maka dari itu normalitas dari sudah terpenuhi.

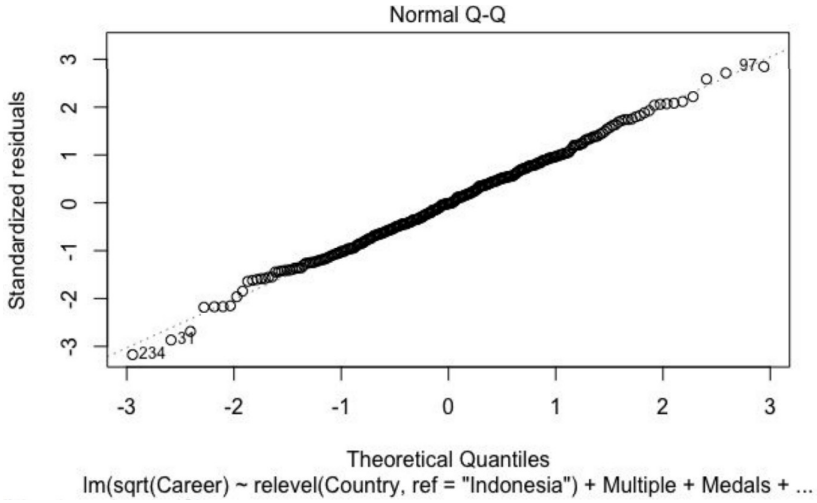

Figure 13: Normalitas Box Cox

```
        Shapiro-Wilk normality test

data:  resid2
W = 0.99791, p-value = 0.9656
```

Figure 14: *Shapiro-Wilk Box-Cox*

4.3.5 Diagnosa Asumsi Homoskedasitistas pada Residual

Seperti yang terlihat pada gambar 15, hasil plot menunjukkan bahwa asumsi homoskedastisitas sudah cukup terpenuhi dan nilai tersebar cukup merata. Maka dari itu, untuk membuktikan lebih detail kita gunakan uji `bptest` untuk melihat nilainya.

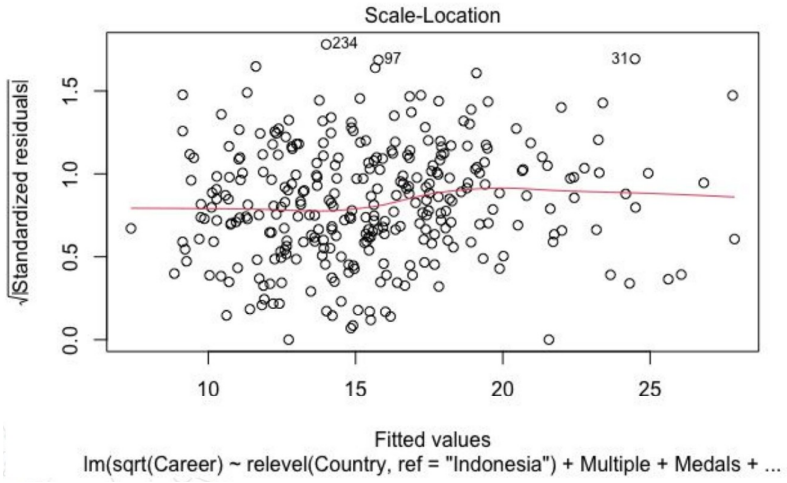

Figure 15: Homoskedasitistas Box Cox

Dari hasil pada gambar 16, dapat dilihat bahwa nilai dalam *Breusch-pagan* sudah lebih dari 0.05 oleh karena itu dapat dikatakan bahwa model memenuhi asumsi homoskedastisitas.

```
        studentized Breusch-Pagan test

data:  stepwiseLog5
BP = 37.843, df = 33, p-value = 0.2577
```

Figure 16: Breusch Pagan Box Cox

4.3.6 Diagnosa Asumsi Independensi pada Residual

Untuk melakukan diagnosa asumsi independensi pada residual, asumsi ini menyatakan bahwa setiap residual tidak dipengaruhi oleh residual data lainnya dalam model regresi.

4.3.7 Deteksi *Outlier*

Dengan menggunakan `boxplot.stats()$out`, data outlier akan dikeluarkan beserta nilai residualnya, seperti tabel dibawah ini

No	Data *Outlier*	Nilai Residual
1	97	4.792831
2	113	4.477672
3	234	-4.678020
4	263	-4.443065

Untuk mendeteksi apakah *outlier* berpengaruh atau tidak terhadap data, kami menggunakan plot *residual vs leverage*. Untuk *outlier* dapat dideteksi dengan melihat apakah ada nilai *standarized residuals* yang diluar rentang [-3,3] pada sumbu y, dan *high leverage* dideteksi dengan melihat apakah data memiliki nilai yang lebih besar dari $2(p+1)/n$ di sumbu x. Pada gambar 17 didapat bahwa ada beberapa data yang melewati *standarized residual* tetapi tidak begitu jauh, dan untuk *high leverage* tidak melebihi 0.5.

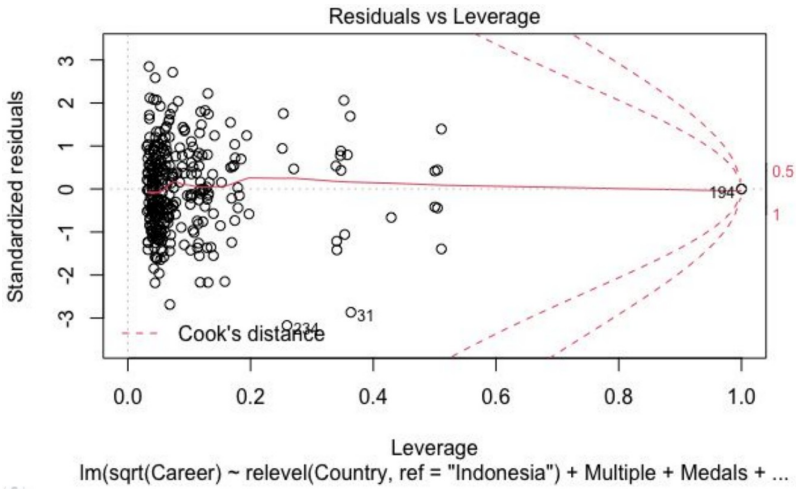

Figure 17: Outlier dan High Leverage Box Cox

Kita bisa melihat lebih detail apakah *outlier* berpengaruh atau tidak dengan menggunakan `outliertest` pada model yang kita miliki, dan jika dilihat dari gambar 10 ditunjukkan bahwa nilai *bonferroni p* lebih dari 0.05 dan dapat diartikan jika *outlier* dari data ke 234 tidak signifikan berpengaruh pada data, maka dari itu data lebih baik tidak di buang.

Data Outlier	rstudent	Unadjusted p-value	Bonferroni p
234	-3.22518	0.0014093	0.43688

4.3.8 Kesimpulan Diagnosa Asumsi

No	Diagnosa Asumsi	Keterangan
1	Linearitas	Hampir Terpenuhi
2	Normalitas pada Residual	Terpenuhi
3	Homoskedastisitas	Terpenuhi

5 Kesimpulan

Model ini adalah model akhir yang kami buat dengan menggunakan transformasi *box-cox* yang memenuhi semua diagnosa asumsi dengan nilai AIC sebesar 1259.404. Berikut model yang kami buat:
`sqrt(Career) = relevel(Country, ref = \Indonesia") + Multiple + Medals + Age + HRank*Age`

$$Career = \beta_0 - \beta_1 \times CountryAustralia + \beta_2 \times CountryBelgium + \beta_3 \times CountryBulgaria - \beta_4 \times CountryCanada - \beta_5 \times CountryChina - \beta_6 \times CountryDenmark + \beta_7 \times CountryEngland + \beta_8 \times CountryEstonia + \beta_9 \times CountryFrance + \beta_{10} \times CountryGermany + \beta_{11} \times CountryHongKong + \beta_{12} \times CountryIndia + \beta_{13} \times CountryIreland + \beta_{14} \times CountryJapan + \beta_{15} \times CountryKorea + \beta_{16} \times CountryMalaysia + \beta_{17} \times CountryNetherlands + \beta_{18} \times CountryRussia + \beta_{19} \times CountryScotland + \beta_{20} \times CountrySingapore + \beta_{21} \times CountrySpain + \beta_{22} \times CountrySweden + \beta_{23} \times CountrySwizterland + \beta_{24} \times CountryTaiwan + \beta_{25} \times CountryThailand + \beta_{26} \times CountryTurkey + \beta_{27} \times CountryUSA + \beta_{28} \times CountryVietnam + \beta_{29} \times Multiple1 + \beta_{30} \times Medals + \beta_{31} \times Age + \beta_{32} \times HRank - \beta_{33} \times HRank : Age + \epsilon_i$$

Dari hasil di gambar 18, maka kita dapat menulis persamaan regresi linear untuk model ini menjadi:

$$\widehat{Career} = 0.96 - 1.55 \times CountryAustralia + 7.34 \times CountryBelgium + 6.71 \times CountryBulgaria - 1.76 \times CountryCanada - 2.23 \times CountryChina - 0.33 \times CountryDenmark + 4.4 \times CountryEngland + 9.42 \times CountryEstonia + 5.02 \times CountryFrance + 4.29 \times CountryGermany + 0.54 \times CountryHongKongChina + 1.39 \times CountryIndia + 0.12 \times CountryJapan + 1.83 \times CountryKorea + 0.21 \times CountryMalaysia + 5.02 \times CountryNetherlands + 6.47 \times CountryRussia + 5.22 \times CountryScotland + 1.94 \times CountrySingapore + 6.21 \times CountrySpain + 6.74 \times CountrySweden + 8.98 \times CountrySwitzerland + 1.81 \times CountryTaiwan + 1.46 \times CountryThailand + 9.23 \times CountryTurkey + 2.4 \times CountryUSA + 3.59 \times CountryVietnam + 0.59 \times Multiple1 + 0.43 \times Medals + 0.5 \times Age + 0.33 \times HRank - 0.01 \times Age : HRank$$

```
Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 0.959577   1.074416   0.893 0.372565
relevel(Country, ref = "Indonesia")Australia -1.557922   1.309346  -1.190 0.235118
relevel(Country, ref = "Indonesia")Belgium   7.346887   1.823167   4.030 7.21e-05 ***
relevel(Country, ref = "Indonesia")Bulgaria  6.712904   1.041217   6.447 5.00e-10 ***
relevel(Country, ref = "Indonesia")Canada   -1.761163   0.732719  -2.404 0.016887 *
relevel(Country, ref = "Indonesia")China    -2.230331   0.443513  -5.029 8.85e-07 ***
relevel(Country, ref = "Indonesia")Denmark  -0.339739   0.475233  -0.715 0.475276
relevel(Country, ref = "Indonesia")England   4.463381   0.639003   6.985 2.09e-11 ***
relevel(Country, ref = "Indonesia")Estonia   9.420272   1.775403   5.306 2.29e-07 ***
relevel(Country, ref = "Indonesia")France    5.025333   0.638331   7.873 7.74e-14 ***
relevel(Country, ref = "Indonesia")Germany   4.296728   0.715490   6.005 5.94e-09 ***
relevel(Country, ref = "Indonesia")Hong Kong China  0.549149   0.647619   0.848 0.397192
relevel(Country, ref = "Indonesia")India     1.398576   0.638962   2.189 0.029438 *
relevel(Country, ref = "Indonesia")Ireland   6.408582   1.073008   5.973 7.09e-09 ***
relevel(Country, ref = "Indonesia")Japan     0.126278   0.429624   0.294 0.769032
relevel(Country, ref = "Indonesia")Korea     1.839457   0.462922   3.974 9.02e-05 ***
relevel(Country, ref = "Indonesia")Malaysia  0.215224   0.437251   0.492 0.622949
relevel(Country, ref = "Indonesia")Netherlands  5.026525   0.781986   6.428 5.58e-10 ***
relevel(Country, ref = "Indonesia")Russia    6.473119   0.781678   8.281 5.13e-15 ***
relevel(Country, ref = "Indonesia")Scotland  5.224877   1.042276   5.013 9.54e-07 ***
relevel(Country, ref = "Indonesia")Singapore 1.949772   1.249319   1.561 0.119736
relevel(Country, ref = "Indonesia")Spain     6.216898   1.056561   5.884 1.14e-08 ***
relevel(Country, ref = "Indonesia")Sweden    6.741468   1.807500   3.730 0.000232 ***
relevel(Country, ref = "Indonesia")Switzerland  8.980516   1.840005   4.881 1.78e-06 ***
relevel(Country, ref = "Indonesia")Taiwan    1.816151   0.549501   3.305 0.001074 **
relevel(Country, ref = "Indonesia")Thailand  1.463062   0.472689   3.095 0.002167 **
relevel(Country, ref = "Indonesia")Turkey    9.233392   1.767240   5.225 3.42e-07 ***
relevel(Country, ref = "Indonesia")USA       2.404941   0.920718   2.612 0.009488 **
relevel(Country, ref = "Indonesia")Vietnam   3.592696   1.288635   2.788 0.005668 **
Multiple1                                   0.596176   0.224567   2.655 0.008392 **
Medals                                      0.430383   0.033196  12.965  < 2e-16 ***
Age                                         0.503643   0.038084  13.224  < 2e-16 ***
HRank                                       0.330670   0.053197   6.216 1.85e-09 ***
Age:HRank                                  -0.018259   0.002233  -8.175 1.05e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.714 on 279 degrees of freedom
Multiple R-squared:  0.8449,    Adjusted R-squared:  0.8266
F-statistic: 46.07 on 33 and 279 DF,  p-value: < 2.2e-16
```

Figure 18

Model ini mengatakan beberapa hal:

- Untuk *Multiple1* maka akan ada kenaikan 0.59 untuk *output sqrt Career* secara signifikan
- Untuk Kenaikan 1 *Medal* maka akan ada kenaikan 0.5 untuk *output sqrt Career* secara signifikan
- Untuk Kenaikan 1 *Age* maka akan ada kenaikan dengan 0.33 untuk *output sqrt Career* secara signifikan
- Untuk Negara Belgium maka akan ada kenaikan 7.34 untuk *output sqrt Career* secara signifikan
- Untuk Kenaikan 1 *HRank* maka akan ada penurunan - 0.01 *output sqrt Career* secara signifikan

References

[1] Kurniawan, Robert. *Cara Mudah Belajar Statistik Analisis Data & Eksplorasi*. Prenada Media, 2019.

[2] Pearson, Ronald K. *Exploratory data analysis using R*. CRC Press, 2018.

[3] Larry E. Beutler. The dodo bird is extinct Clinical Psychology: Science and Practice 9, no. 1 (2002): 30-34.

[4] Informatika UNPAR. `https://informatika.unpar.ac.id`

[5] Hayes, A.F., 2017. Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. Guilford publications.

[6] Berkenalan dengan Metode Metode Analisis Regresi Melalui SPSS Oleh Wahyu Widhiarso

[7] Meuleman, B., Loosveldt, G. and Emonds, V., 2014. Regression analysis: Assumptions and diagnostics. The SAGE handbook of regression analysis and causal inference, pp.83-110.

[8] Ghozali, I., 2018. Aplikasi analisis multivariate dengan program IBM SPSS 25.

[9] Uji Normalitas, https://www.statistikian.com/2013/01/uji-normalitas.html

[10] Uji Homoskedastik, Bilson Simamora, https://www.bilsonsimamora.com/blog/2017/08/26/uji-homoskedastisitas/

[11] Fransiska, W., Nugroho, S. and Faisal, F., 2012. Transformasi Box Cox dalam Analisis Regresi Linier Sederhana. E-jurnal FMIPA Universitas Bengkulu.

[12] Peck, R., Olsen, C. and Devore, J.L., 2015. Introduction to statistics and data analysis. Cengage Learning.

[13] , Kutner, Michael H and Nachtsheim, Christopher J and Wasserman, William, 1996. Applied linear statistical models. McGraw-Hill/Irwin.

[14] , Diaconis, Persi, 2006. Theories of data analysis: From magical thinking through classical statistics. Exploring data tables, trends, and shapes. Wiley Online Library

[15] Dwi Ispriyanti, D.I., 2004. Pemodelan Statistika dengan Transformasi Box Cox. Jurnal Matematika, 7(3), pp.8-17.

6 Code

```
BWF_players = BWF_players = read.csv("BWF_players19.csv")
 is .na(BWF_players)
na_dataNew = BWF_players[!complete.cases(BWF_players),]
#NA Height Change
#Asia
BWF_players$Height[BWF_players$Name == "Adnan Maulana"] <- 175
BWF_players$Height[BWF_players$Name == "Bagas Maulana"] <- 175
BWF_players$Height[BWF_players$Name == "Cheah Yee See"] <- 164
BWF_players$Height[BWF_players$Name == "Di Zi Jian"] <- 183
BWF_players$Height[BWF_players$Name == "Lu Guang Zu"] <- 183
BWF_players$Height[BWF_players$Name == "Muhammad Shohibul Fikri"] <- 175
BWF_players$Height[BWF_players$Name == "Nguyen Thuy Linh"] <- 165
BWF_players$Height[BWF_players$Name == "Pearly Tan"] <- 164
BWF_players$Height[BWF_players$Name == "Pitha Haningtyas Mentari"] <- 165
BWF_players$Height[BWF_players$Name == "Sim Yu-jin"] <- 170
BWF_players$Height[BWF_players$Name == "Siti Fadia Silva Ramadhanti"] <- 165
BWF_players$Height[BWF_players$Name == "Taichi Saito"] <- 170
BWF_players$Height[BWF_players$Name == "Wang Chang"] <- 183
BWF_players$Height[BWF_players$Name == "Yeremia Erich Yoche Yacob Rambitan"] <- 175
#Europe
BWF_players$Height[BWF_players$Name == "Nhat Nguyen"] <- 183
BWF_players$Height[BWF_players$Name == "Qi Xuefei"] <- 186
#Pan Am
BWF_players$Height[BWF_players$Name == "Ryan Chew"] <- 172
#NA Hand Change
BWF_players$Hand[BWF_players$Name == "Nguyen Thuy Linh"] <- "Right"
BWF_players$Hand[BWF_players$Name == "Ryan Chew"] <- "Right"
BWF_players$Hand[BWF_players$Name == "Sim Yu-jin"] <- "Right"
BWF_players$Hand[BWF_players$Name == "Wang ZhiYi"] <- "Right"
#NA Medal Change
BWF_players$Medals[BWF_players$Name == "Chang Tak Ching"] <- 4
BWF_players$Medals[BWF_players$Name == "Ekaterina Malkova"] <- 4
BWF_players$Medals[BWF_players$Name == "Kang Min-hyuk"] <- 4
BWF_players$Medals[BWF_players$Name == "Kim Jae-hwan"] <- 4
BWF_players$Medals[BWF_players$Name == "Ng Wing Yung"] <- 4
BWF_players$Medals[BWF_players$Name == "Ryan Chew"] <- 4
#change multiple
BWF_players$Multiple[BWF_players$Name == "Di Zi Jian"] <- 1
#change Age
BWF_players$Age[BWF_players$Name == "Vu Thi Trang"] <- 29
BWF_players$Age[BWF_players$Name == "Ayako Sakuramoto"] <- 25
BWF_players$Age[BWF_players$Name == "Cheung Ngan Yi"] <- 28
BWF_players$Age[BWF_players$Name == "Wakana Nagahara"] <- 25
#change Height
BWF_players$Height[BWF_players$Name == "Kang Min-hyuk"] <- 183
#change career
BWF_players$Career[BWF_players$Name == "Cheah Yee See"] <- 124
BWF_players$Career[BWF_players$Name == "Huang Yu Xiang"] <- 102
BWF_players$Career[BWF_players$Name == "Akira Koga"] <- 90
BWF_players$Career[BWF_players$Name == "Qi Xuefei"] <- 114
BWF_players$Career[BWF_players$Name == "Lee Yong-dae"] <- 810
BWF_players$Career[BWF_players$Name == "Fitriani Fitriani"] <- 112
BWF_players$Career[BWF_players$Name == "Zhang Yi Man"] <- 66
#change medals
BWF_players$Medals[BWF_players$Name == "Lin Dan"] <- 21
# asfactor
BWF_players$Gender = as.factor(BWF_players$Gender)
```

```
BWF_players$Hand = as.factor(BWF_players$Hand)
BWF_players$Category = as. factor (BWF_players$Category)
BWF_players$Continent = as. factor (BWF_players$Continent)
BWF_players$Multiple = as. factor (BWF_players$Multiple)
BWF_players$Country = as.factor (BWF_players$Country)
BWF_players$Age = as.integer(BWF_players$Age)
BWF_players$Medals= as.integer(BWF_players$Medals)
BWF_players$Height = as. integer (BWF_players$Height)
BWF_players$Medals = as.integer(BWF_players$Medals)
BWF_players$Career = as. integer (BWF_players$Career)
# grafik  univariat
ggplot(BWF_players, aes(x = Hand,  fill =factor (Hand)))+
  geom_bar()+
  xlab("Jenis  Hand")+
  ylab("Frequency")+
   ggtitle ("Barplot  Hand dalam Permainan Bulu Tangkis")+
  guides( fill =guide_legend("Hand"))+
  scale _ fill _brewer( palette ="Set2")+
  theme(panel.background = element_blank())
#Gender
ggplot(BWF_players, aes(x = "", y = Gender, fill =factor (Gender)))+
  geom_bar( stat  = " identity ", width = 1)+
   ggtitle ("Piechart  Gender Pemain Bulu Tangkis")+
  guides( fill =guide_legend("Gender"))+
  coord_polar ("y", start =0)+
  theme_void()+
  scale _ fill _brewer( palette ="Set2")
#Category
ggplot(BWF_players, aes(x = "", y = Category, fill =factor (Category)))+
  geom_bar( stat  = " identity ", width = 1)+
   ggtitle ("Piechart  Kategori  Pemain Bulu Tangkis")+
  guides( fill =guide_legend("Kategori"))+
  coord_polar ("y", start =0)+
  theme_void()+
  scale _ fill _brewer( palette ="Set2")
#Continent
ggplot(BWF_players, aes(x = Continent, fill =factor (Continent)))+
  geom_bar()+
   ggtitle ("Barplot  Asal  Kontinent  Pemain Bulu Tangkis")+
  guides( fill =guide_legend("Country"))+
  theme_void()
#Medal
par(mfrow = c(2,2))
h1 <- hist(BWF_players$Medals,
           main = "Peraihan  Medali Emas oleh Pemain",
           xlab  = "Jumlah Medali")
xfit <-seq(min(BWF_players$Medals),max(BWF_players$Medals),length=40)
yfit <-dnorm(xfit,mean=mean(BWF_players$Medals),sd=sd(BWF_players$Medals))
yfit <-yfit* diff (h1$mids[1:2])*length (BWF_players$Medals)
lines ( xfit ,  yfit ,  col="red",  lwd=2)
# Bivariat  Antara  Age dan Height
ggplot(BWF_players, aes(x = Height,  y = Age))+
  geom_point()+
  xlab("Height")+
  ylab("Umur")+
   ggtitle (" Scatterplot  Hubungan Height dan Age")+
  theme(panel.background = element_blank())
# Bivariat  Antara  Age dan Gender
ggplot(BWF_players, aes(x = Gender,y=Age))+
  geom_boxplot()+
```

```
    xlab("Gender")+
    ylab("Umur")+
     ggtitle ("Boxplot Antara Age dan Gender dalam Permainan Bulu Tangkis")+
    theme(panel.background = element_blank())
# Bivariat  Antara Age dan Category
ggplot(BWF_players, aes(x = Category,y=Age))+
    geom_boxplot()+
    xlab("Category")+
    ylab("Umur")+
     ggtitle ("Boxplot Antara Age dan Category dalam Permainan Bulu Tangkis")+
    theme(panel.background = element_blank())
# grafik  interaksi
 interaction.plot(x.factor = continent,
                  trace.factor = hand,
                  response = career,
                  main="Interaction Plot career dari Hand*Contintent",
                  xlab = "continent", ylab = "career",
                  col=c("red","blue"))
 interaction.plot(x.factor = hand,
                  trace.factor = category,
                  response = career,
                  main="Interaction Plot career dari Hand*Category",
                  xlab = "hand", ylab = "career",
                  col=c("red","blue"))
 interaction.plot(x.factor = category,
                  trace.factor = continent,
                  response = career,
                  main="Interaction Plot career dari Category*Continent",
                  xlab = "category", ylab = "career",
                  col=c("red","blue"))
 interaction.plot(x.factor = category,
                  trace.factor = multiple,
                  response = career,
                  main="Interaction Plot career dari Category*Multiple",
                  xlab = "category", ylab = "career",
                  col=c("red","blue"))
#Regresi model
 str(BWF_players)
# regresi banyak variable
 ols <- lm(Career~Age + Height + Country + Hand + Medals + HRank + Multiple + Category + Gender +
     Continent,data=BWF_players)
summary(ols)
AIC(ols)
 stepwise <- step(ols, direction ="both")
 ols <- lm(Career~relevel(Country,ref="Indonesia")+ Multiple + Medals + Age + HRank*Medals+
     HRank*Age,data=BWF_players)
AIC(ols)
summary(ols)
 stepwise1 <- step(ols, direction ="both")
 library(broom)
model.diag.metrics <- augment(stepwise1)
head(model.diag.metrics,10)
head(stepwise1$residuals, 10)
head(stepwise1$fitted.values, 10)
 plot(stepwise1,1)
 plot(stepwise1,2)
 resid <- stepwise1$residuals
 shapiro.test(resid)
###Breusch-Pagan Test heteroskedastik (p-value < 0,05)
 library(lmtest)
```

```
bptest(stepwise1)
library(car)
vif(stepwise1)
library(corrplot)
career.num <- data.frame(BWF_players$Age, BWF_players$Height, BWF_players$Medals, BWF_players$HRank,
    BWF_players$Career)
mat <- cor(career.num) #matriks korelasi
corrplot(mat, method="square")
resid <- stepwise1$residuals
bon <- rstudent(stepwise1)
boxplot(resid)
boxplot.stats(resid)$out
library(car)
outlierTest(stepwise)
#Lihat Lamda untuk menentukan transformasi
model1 <- lm(Career~ relevel(Country, ref="Indonesia") + Multiple + Age + Medals + HRank*Medals +
    HRank*Age, data=BWF_players)
stepwise1 <- step(model1, direction ="both")
library(MASS)
box <- boxcox(stepwise1)
(lambda <- box$x[which.max(box$y)]) #0.54 --- boxcox
#Transformasi Box-Cox dengan full data (BWF_players)
model.sq <- lm(sqrt(Career)~ relevel(Country, ref="Indonesia") + Multiple + Age + Medals + HRank*Medals +
    HRank*Age, data=BWF_players)
stepwiseLog2 <- step(model.sq, direction ="both")
summary(stepwiseLog2)
AIC(model.sq) #1238 #1259
model.sq <- lm(sqrt(Career) ~ relevel(Country, ref = "Indonesia") + Multiple +
                    Medals + Age + HRank + Age*HRank, data=BWF_players)
stepwiseLog2 <- step(model.sq, direction ="both")
summary(stepwiseLog2)
AIC(model.sq)
par(mfrow=c(2,2))
plot(stepwiseLog2)
library(car)
vif(stepwiseLog2)
library(lmtest)
resid2 <- stepwiseLog2$residuals
shapiro.test(resid2)
bptest(stepwiseLog2)
outlierTest(stepwiseLog2)
```