

Salary Prediction Using Regression Model

Statistics for Business

Daftar Isi

- Pengantar
- Dataset
- Uji Statistik
- Pemodelan Regresi
- Kesimpulan dan Saran
- Referensi

Pengantar

Pengantar

Terdapat sebuah dataset yang menunjukkan gaji yang dimiliki seseorang berdasarkan usia, jenis kelamin, tingkat pendidikan, jabatan, dan lama pengalaman kerja. Dari dataset tersebut, penulis ingin mengetahui pengaruh dari faktor-faktor tersebut terhadap gaji dan melakukan prediksi gaji seseorang.

- Menguji pengaruh jenis kelamin terhadap gaji dengan uji statistik
- Memprediksi gaji dari lama pengalaman kerja seseorang dengan model regresi
- Memprediksi gaji dari variabel prediktor usia, jenis kelamin, tingkat pendidikan, dan lama pengalaman kerja dengan model regresi

Dataset

Dataset

- Dataset yang digunakan diambil dari kaggle.com. Dataset berisikan 375 baris data usia, jenis kelamin, tingkat pendidikan, jabatan, lama pengalaman kerja, dan besar gaji.
- Sebelum diolah lebih lanjut, dilakukan persiapan dengan menghapus missing value dan duplicated data sehingga didapatkan 324 baris data yang bisa digunakan.

	Age	Gender	EducationLevel	Job Title	YearsOfExperience	Salary
0	32.0	Male	Bachelor's	Software Engineer	5.0	90000.0
1	28.0	Female	Master's	Data Analyst	3.0	65000.0
2	45.0	Male	PhD	Senior Manager	15.0	150000.0
3	36.0	Female	Bachelor's	Sales Associate	7.0	60000.0
4	52.0	Male	Master's	Director	20.0	200000.0
...

Dataset

- Data Job Title tidak akan digunakan dalam pemodelan regresi karena terlalu bervariasi.
- Data jenis kelamin (Gender) akan diubah dari data kategorikal menjadi data numerik dengan Male = 0 dan Female = 1.
- Data tingkat pendidikan (Education Level) akan diubah dari data kategorikal menjadi data numerik dengan Bachelor's = 0, Master's = 1, dan PhD = 2.
- Link: <https://www.kaggle.com/datasets/rkiattisak/salaly-prediction-for-beginner>

	Age	Gender	EducationLevel	Job Title	YearsOfExperience	Salary
0	32.0	Male	Bachelor's	Software Engineer	5.0	90000.0
1	28.0	Female	Master's	Data Analyst	3.0	65000.0
2	45.0	Male	PhD	Senior Manager	15.0	150000.0

Dataset

Data numerik

	count	mean	std	min	25%	50%	75%	max
Age	324.0	37.382716	7.185844	23.0	31.0	36.5	44.0	53.0
YearsOfExperience	324.0	10.058642	6.650470	0.0	4.0	9.0	16.0	25.0
Salary	324.0	99985.648148	48652.271440	350.0	55000.0	95000.0	140000.0	250000.0

Korelasi antar data numerik

In [125]:

▶

df_salary[["Age", "YearsOfExperience", "Salary"]].corr()

Out[125]:

	Age	YearsOfExperience	Salary
Age	1.000000	0.979192	0.916543
YearsOfExperience	0.979192	1.000000	0.924455
Salary	0.916543	0.924455	1.000000

Korelasi antara usia, lama pengalaman kerja, dan gaji memiliki hasil positif dan berkorelasi kuat.

Dataset

Data kategorik

```
In [123]: df_salary["Gender"].value_counts()

Out[123]: Male      170
          Female    154
          Name: Gender, dtype: int64

In [124]: df_salary["EducationLevel"].value_counts()

Out[124]: Bachelor's    191
          Master's      91
          PhD           42
          Name: EducationLevel, dtype: int64
```

Perbandingan gaji antar variabel kategorik

```
In [66]: df_salary.groupby("Gender")["Salary"].mean()

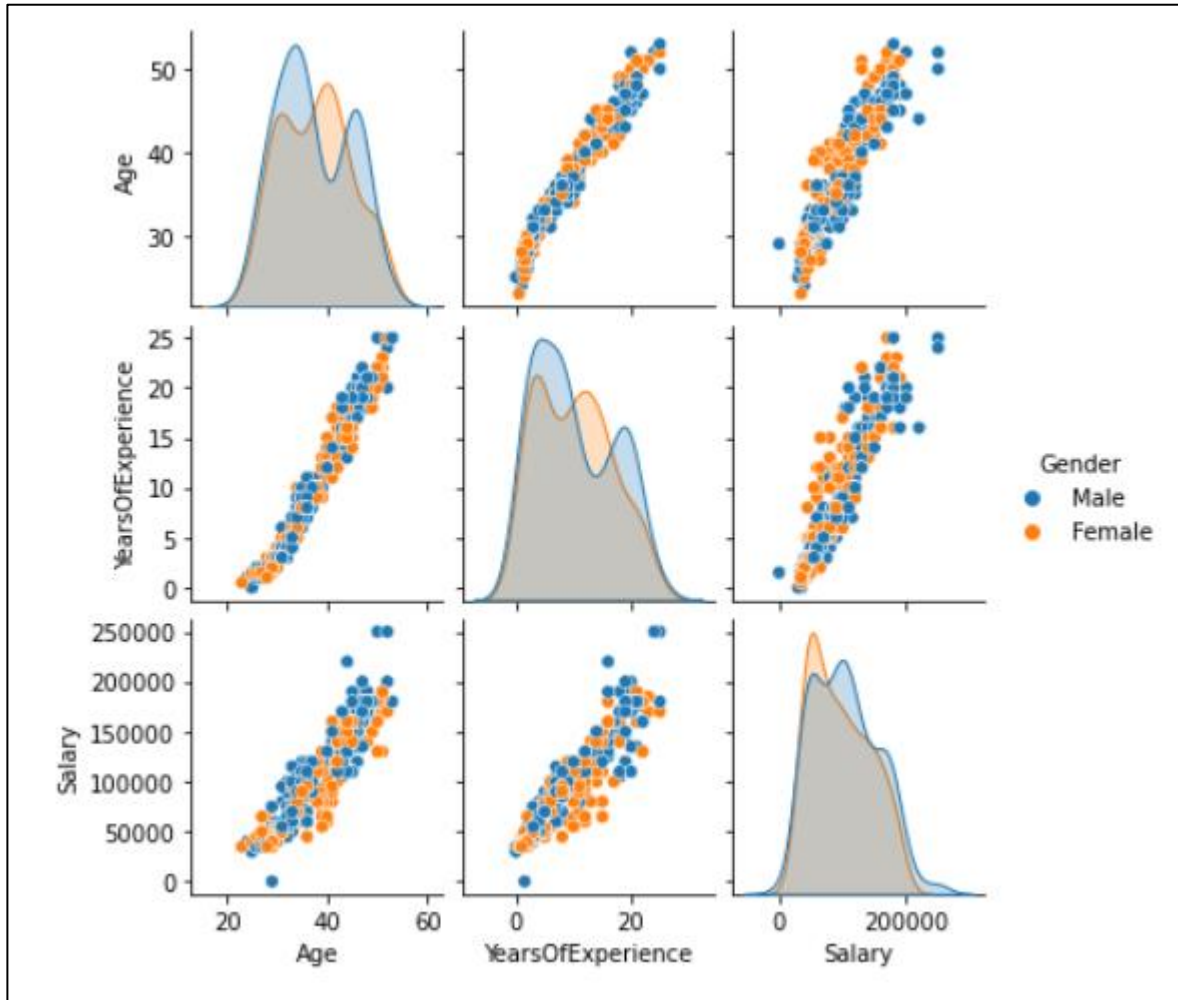
Out[66]: Gender
          Female    96136.363636
          Male     103472.647059
          Name: Salary, dtype: float64

In [67]: df_salary.groupby("EducationLevel")["Salary"].mean()

Out[67]: EducationLevel
          Bachelor's    73902.356021
          Master's     127912.087912
          PhD          158095.238095
          Name: Salary, dtype: float64
```

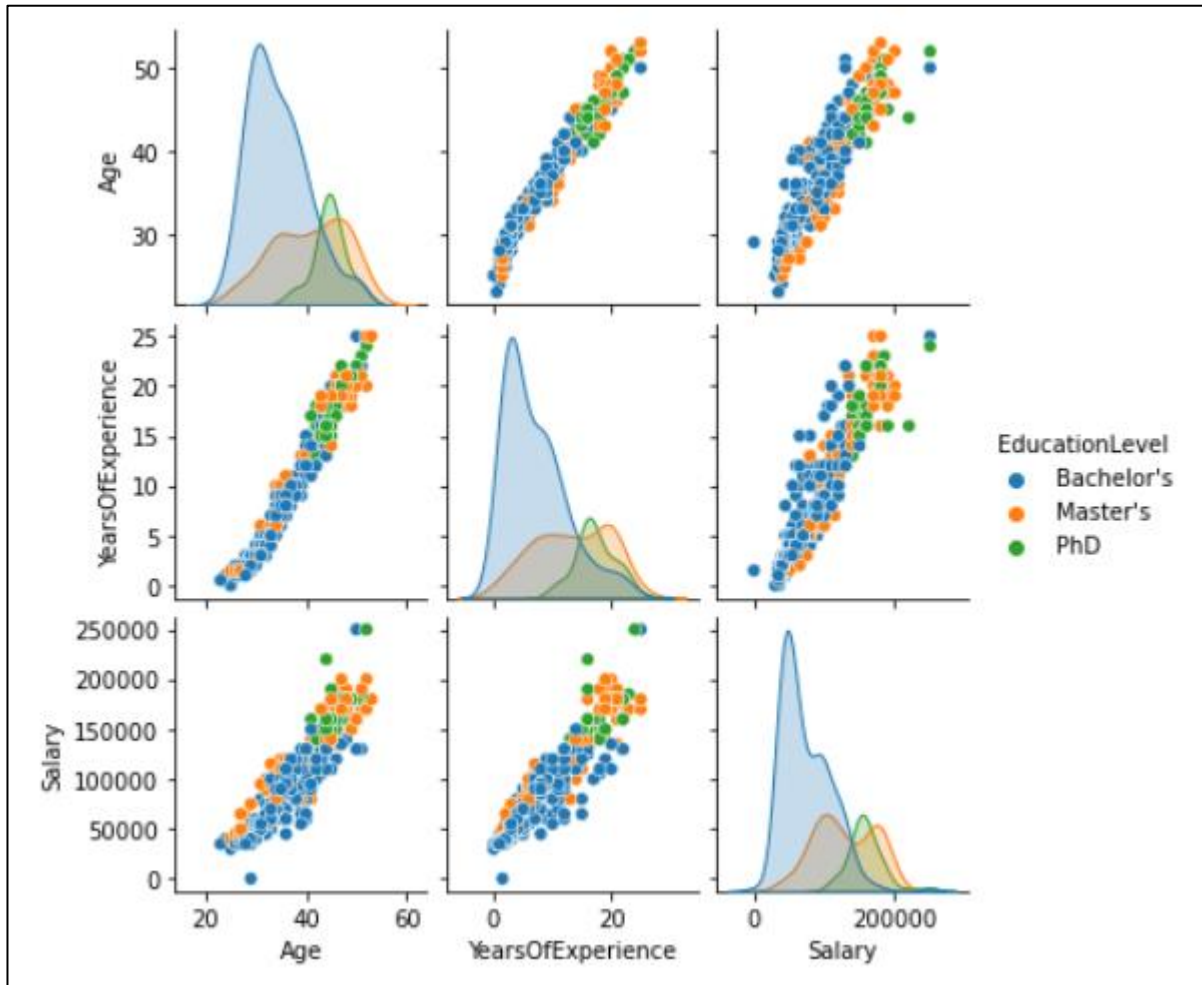
- Rata-rata gaji laki-laki lebih besar dari rata-rata gaji perempuan.
- Rata-rata gaji semakin besar seiring dengan level pendidikan yang lebih tinggi.

Visualisasi Data Numerik terhadap Jenis Kelamin



- Semakin lama pengalaman kerja seseorang, semakin tinggi gajinya.
- Semakin tua usia seseorang, semakin lama pula pengalaman kerja yang dimiliki.
- Jenis kelamin tidak terlalu berdampak signifikan pada gaji.

Visualisasi Data Numerik terhadap Tingkat Pendidikan



- Semakin tinggi tingkat pendidikan seseorang, semakin tinggi gajinya.
- Tingkat pendidikan yang tinggi cenderung dimiliki seseorang dengan usia yang lebih tua.
- Tingkat pendidikan yang tinggi cenderung dimiliki oleh seseorang dengan lama pengalaman kerja yang tinggi pula.

Uji Statistik

Uji Statistik

Penulis ingin mengetahui pengaruh jenis kelamin terhadap besarnya gaji seseorang. Dalam dataset terdapat 2 jenis kelamin yaitu male (laki-laki (a)) dan female (perempuan (b)). Penulis akan menguji apakah rata-rata gaji laki-laki lebih besar dari rata-rata gaji perempuan.

Taraf signifikansi = 10%

$$H_0: \mu_a = \mu_b$$

$$H_1: \mu_a > \mu_b$$

Uji Statistik

Karena standar deviasi populasi tidak diketahui, digunakan t-test. Sebelum menggunakan t-test, dilakukan uji variansi.

```
In [72]: ▶ # Gaji Laki-laki
df_male = df_salary[df_salary["Gender"]=="Male"]["Salary"].values

# Gaji Perempuan
df_female = df_salary[df_salary["Gender"]=="Female"]["Salary"].values

# Variansi
np.var(df_male), np.var(df_female)
```

```
Out[72]: (2571353207.6989617, 2097896989.374262)
```

Dari hasil tersebut dapat disimpulkan bahwa variansi tidak sama.

Uji Statistik

```
In [73]: ▶ from scipy import stats
          result = stats.ttest_ind(a = df_male,
                                   b = df_female,
                                   equal_var=False,
                                   alternative = "greater")
```

```
In [74]: ▶ result.pvalue
```

```
Out[74]: 0.08675461782037655
```

```
In [75]: ▶ result.statistic
```

```
Out[75]: 1.364034982496829
```

```
In [76]: ▶ # Menentukan aturan keputusan
          if result.pvalue < significance_level:
              print("Tolak hipotesis nol.")
          else:
              print("Gagal menolak hipotesis nol.")
```

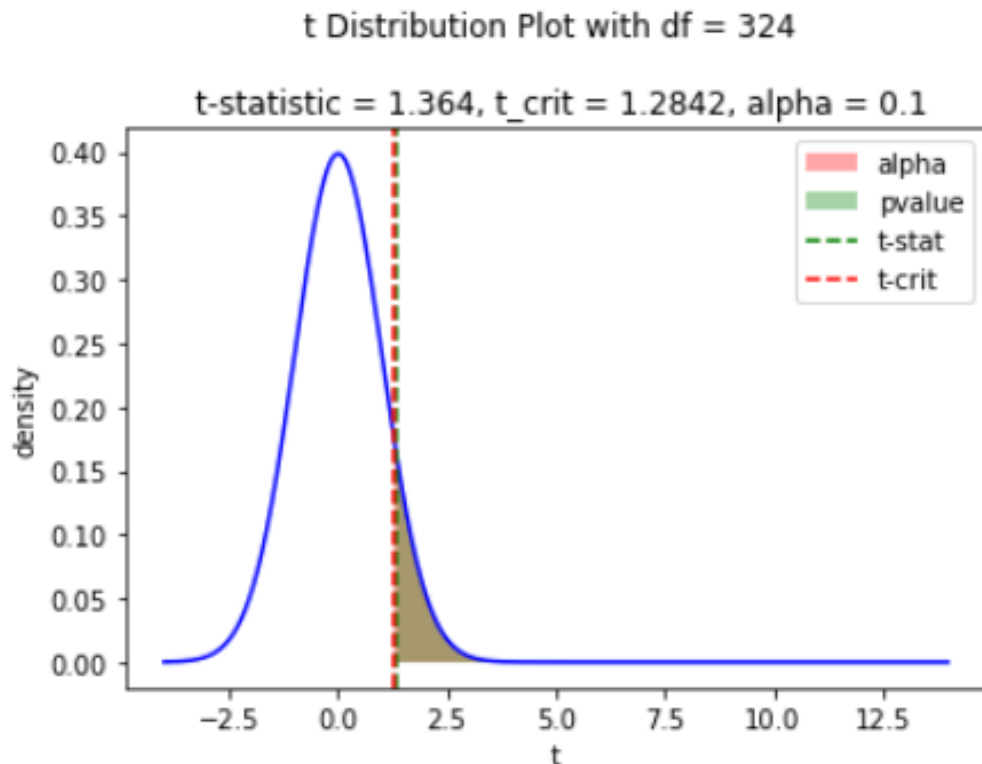
```
Tolak hipotesis nol.
```

Terdapat cukup bukti bahwa rata-rata gaji laki-laki dan perempuan tidak sama.

Rata-rata gaji laki-laki lebih tinggi dari rata-rata gaji perempuan.

Uji Statistik

Derajat kebebasan



Confidence level

```
In [79]: from statsmodels.stats.weightstats import DescrStatsW, CompareMeans

cm = CompareMeans(d1 = DescrStatsW(data=df_male),
                  d2 = DescrStatsW(data=df_female))

lower, upper = cm.tconfint_diff(alpha=significance_level,
                               alternative='two-sided',
                               usevar='unequal')

print("Confidence Interval", ":", "[", lower, upper, "]")

Confidence Interval : [ -1535.8717753119818 16208.438620231766 ]
```

- Dari hasil yang didapat, disimpulkan bahwa kita 90% yakin bahwa rata-rata gaji laki-laki lebih dari rata-rata gaji perempuan.
- Dari confidence interval yang didapat, disimpulkan kita 90% yakin bahwa rata-rata perbedaan gaji memiliki interval di -1535 sampai dengan 16208.

Regression Model

Regression: Single Predictor

Dilakukan pemodelan regresi untuk memprediksi gaji seseorang dari lama pengalaman kerjanya.

```
# Create OLS model object
model = smf.ols("Salary ~ YearsOfExperience", df_salary)

# Fit the model
results_model_salary = model.fit()

# Extract the results (Coefficient and Standard Error) to DataFrame
results_salary = print_coef_std_err(results_model_salary)
results_salary
```

	coef	std err
Intercept	31959.508721	1873.552736
YearsOfExperience	6762.954641	155.446221

```
In [176]: results_model_salary.rsquared
Out[176]: 0.8546166681460778
```

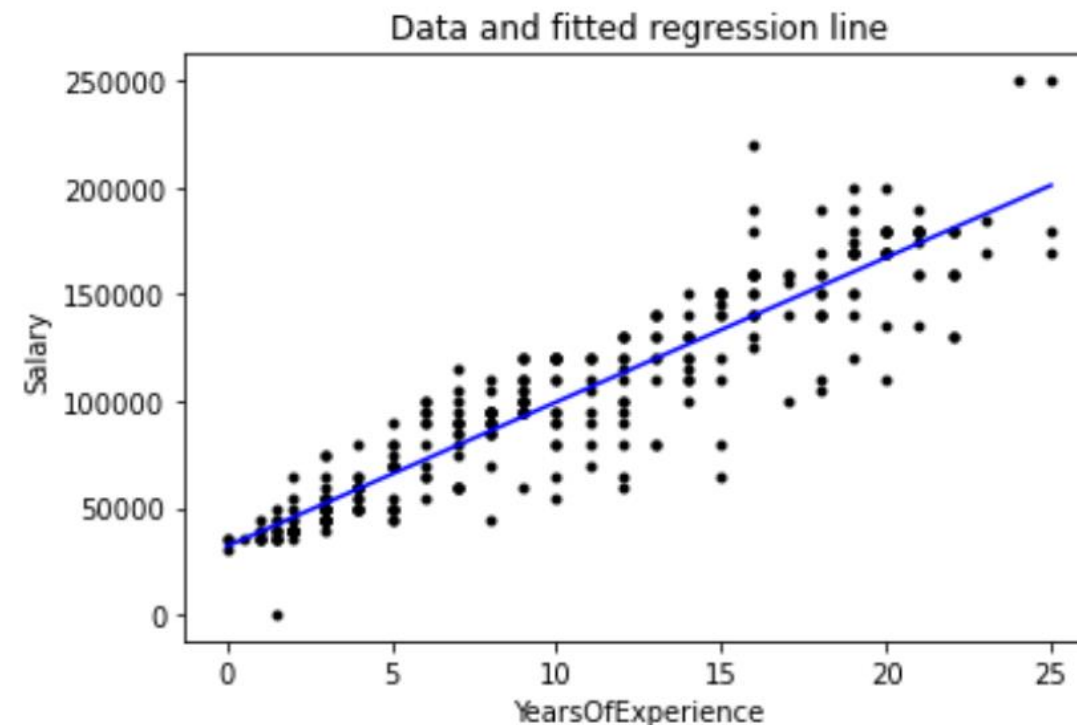
Dari hasil tersebut didapatkan persamaan regresi berikut dengan R-squared yang cukup baik yaitu 0,85.

$$\text{Salary} = 31960 + 6763 \times \text{Years of Experience}$$

Regression: Single Predictor

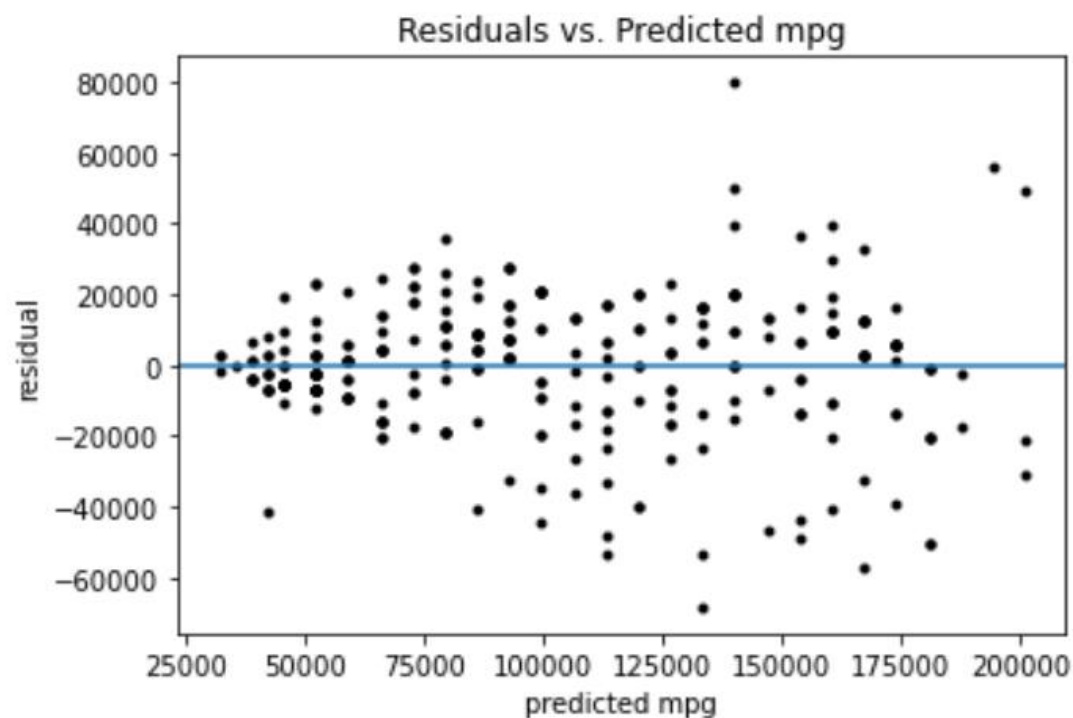
$$\text{Salary} = 31960 + 6763 \times \text{Years of Experience}$$

- Jika membandingkan dua orang yang memiliki 1 tahun perbedaan pada lama pengalaman kerja, diperkirakan orang yang memiliki pengalaman kerja lebih lama memiliki gaji yang lebih besar dengan selisih 6763.
- Untuk seseorang yang memiliki lama pengalaman kerja 0 tahun, perkiraan rata-rata gaji yang didapatkan adalah sebesar 31960.

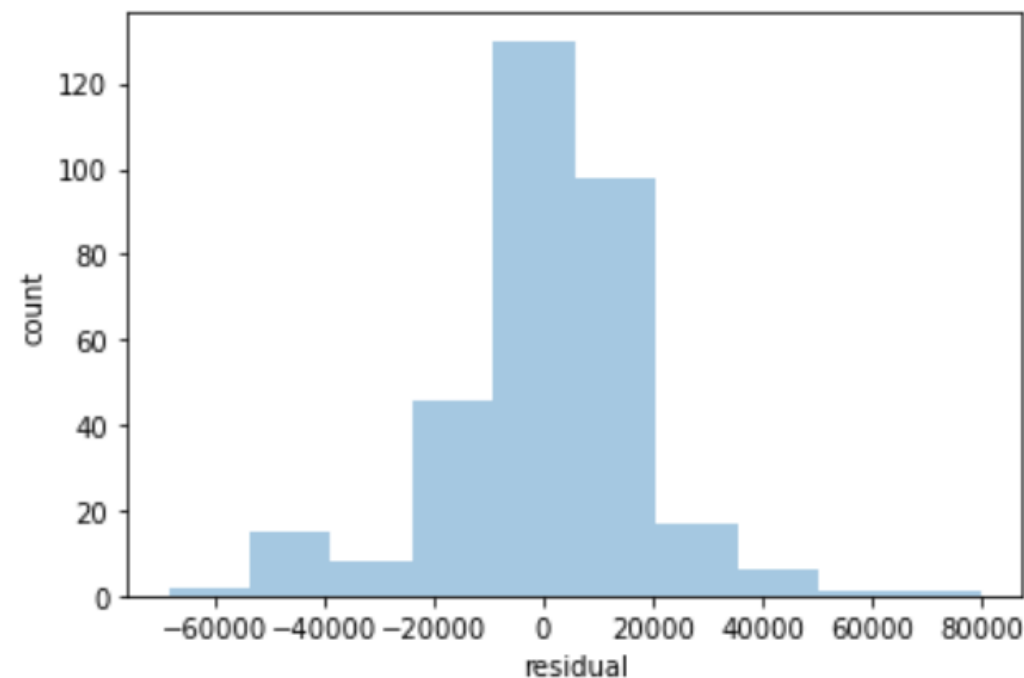


Regression: Single Predictor

Residual Plot




Normality of Error Assumption



Regression: Single Predictor with Log Transformation

Dilakukan pemodelan regresi untuk memprediksi gaji seseorang dari lama pengalaman kerjanya namun dilakukan transformasi logaritmik pada variabel prediktor.

```
In [170]:  # Create OLS model object  
model = smf.ols("Salary ~ logYOE", df_salary)  
  
# Fit the model  
results_logtransform = model.fit()  
  
# Extract the results (Coefficient and Standard Error) to DataFrame  
results_salary_log = print_coef_std_err(results_logtransform)  
results_logtransform.rsquared
```

Out[170]: 0.7656239539695425

Didapatkan hasil R-squared sebesar 0,76. Hasil ini lebih rendah dari hasil R-squared 0,85 pada pemodelan tanpa transformasi logaritmik. Sehingga untuk pemodelan regresi dengan satu variabel prediktor, digunakan model regresi tanpa transformasi.

Regression: Multiple Predictors with One Interaction

Dalam pemodelan ini, digunakan semua variabel prediktor yaitu usia, jenis kelamin, tingkat pendidikan, dan lama pengalaman kerja. Ditambahkan juga satu interaksi antar variabel prediktor yaitu usia dan lama pengalaman kerja. Untuk variabel tingkat pendidikan (Education Level) diperlakukan sebagai variabel kategorikal.

Evaluasi model dengan K-Fold cross validation

```
Out[44]:
```

	test_rsquared	folds
0	0.892141	Folds 1
1	0.902729	Folds 2
2	0.912515	Folds 3
3	0.825113	Folds 4
4	0.897267	Folds 5

```
In [45]: scores_ols_all_pred["test_rsquared"].mean()
```

```
Out[45]: 0.8859529642576728
```

Didapatkan R-squared rata-rata sebesar 0,88 yang berarti model ini baik dan dapat menjelaskan 88% variansi gaji.

Regression: Multiple Predictors with One Interaction

Fitting model

	coef	std err
Intercept	-44159.185552	16580.736611
C(EducationLevel)[T.1]	19574.074815	2257.344892
C(EducationLevel)[T.2]	26339.473807	3160.610738
Age	3042.039143	611.919060
Gender	-9310.571777	1766.475849
YearsOfExperience	2433.641886	1211.995905
Age:YearsOfExperience	3.452762	21.044653

Didapatkan hasil koefisien persamaan regresi di samping.

Hasil intercept negatif kurang baik karena kurang dapat menghasilkan interpretasi yang baik (gaji tidak mungkin negatif) dan usia kerja seseorang biasanya tidak dimulai dari nol.

Oleh karena itu dilakukan centering variabel usia (age).

Regression: Multiple Predictors with One Interaction

Centering Variabel Usia (Age)

Digunakan rata-rata usia pada dataset (37 tahun) sebagai acuan. Sehingga data usia akan dihitung dari jaraknya terhadap usia 37 tahun.

```
In [48]: ▶ df_salary["Age"] = df_salary["Age"] - mean_age  
df_salary.rename(columns = {"Age": "AgeCentered"}, inplace=True)  
df_salary.head()
```

```
Out[48]:
```

	AgeCentered	Gender	EducationLevel	YearsOfExperience	Salary
0	-5.0	0	0	5.0	90000.0
1	-9.0	1	1	3.0	65000.0
2	8.0	0	2	15.0	150000.0
3	-1.0	1	0	7.0	60000.0
4	15.0	0	1	20.0	200000.0

Regression: Multiple Predictors with One Interaction

Evaluasi model dengan K-Fold cross validation

Out[49]:

	test_rsquared	folds
0	0.849681	Folds 1
1	0.907836	Folds 2
2	0.873470	Folds 3
3	0.938117	Folds 4
4	0.881399	Folds 5

Didapatkan R-squared rata-rata sebesar 0,89 yang berarti model ini baik dan dapat menjelaskan 89% variansi gaji.

```
In [50]: > scores_ols_all_pred["test_rsquared"].mean()
```

Out[50]: 0.8901007028969221

Regression: Multiple Predictors with One Interaction

Fitting model

	coef	std err
Intercept	68396.262743	6722.803498
C(EducationLevel)[T.1]	19574.074815	2257.344892
C(EducationLevel)[T.2]	26339.473807	3160.610738
AgeCentered	3042.039143	611.919060
Gender	-9310.571777	1766.475849
YearsOfExperience	2561.394070	714.405923
AgeCentered:YearsOfExperience	3.452762	21.044653

Interpretasi tingkat pendidikan

Jika membandingkan dua orang yang memiliki usia, jenis kelamin, lama pengalaman kerja yang sama, gaji seseorang dengan tingkat pendidikan Master's diperkirakan lebih tinggi 19574 dollar daripada gaji seseorang dengan tingkat pendidikan Bachelor's.

Interpretasi jenis kelamin

Jika membandingkan dua orang yang memiliki usia, lama pengalaman kerja, dan tingkat pendidikan yang sama, perempuan diperkirakan memiliki gaji lebih sedikit 9311 dollar dibandingkan laki-laki.

$$\text{Salary for Bachelor's} = 68396 + 3042 \times (\text{Age} - 37) - 9311 \times \text{Gender} + 2561 \times \text{YearsOfExperience} + 3 \times (\text{Age} - 37) \times \text{YearsOfExperience}$$

$$\text{Salary for Master's} = 68396 + 19574 + 3042 \times (\text{Age} - 37) - 9311 \times \text{Gender} + 2561 \times \text{YearsOfExperience} + 3 \times (\text{Age} - 37) \times \text{YearsOfExperience}$$

$$\text{Salary for PhD} = 68396 + 26339 + 3042 \times (\text{Age} - 37) - 9311 \times \text{Gender} + 2561 \times \text{YearsOfExperience} + 3 \times (\text{Age} - 37) \times \text{YearsOfExperience}$$

Regression: Multiple Predictors with One Interaction

$$\text{Salary for Bachelor's} = 68396 + 3042 \times (\text{Age} - 37) - 9311 \times \text{Gender} + 2561 \times \text{YearsOfExperience} + 3 \times (\text{Age} - 37) \times \text{YearsOfExperience}$$
$$\begin{aligned} \text{Salary for Master's} = & 68396 + 19574 + 3042 \times (\text{Age} - 37) - 9311 \times \text{Gender} + 2561 \times \text{YearsOfExperience} + 3 \times (\text{Age} - 37) \\ & \times \text{YearsOfExperience} \end{aligned}$$
$$\text{Salary for PhD} = 68396 + 26339 + 3042 \times (\text{Age} - 37) - 9311 \times \text{Gender} + 2561 \times \text{YearsOfExperience} + 3 \times (\text{Age} - 37) \times \text{YearsOfExperience}$$

Interpretasi usia

Jika membandingkan dua orang yang memiliki jenis kelamin dan tingkat pendidikan yang sama, serta pengalaman kerja 0 tahun, seseorang yang usianya 1 tahun lebih tua dari 37 tahun diperkirakan memiliki gaji lebih tinggi 3042 dollar daripada seseorang berusia 37 tahun.

Interpretasi lama pengalaman kerja

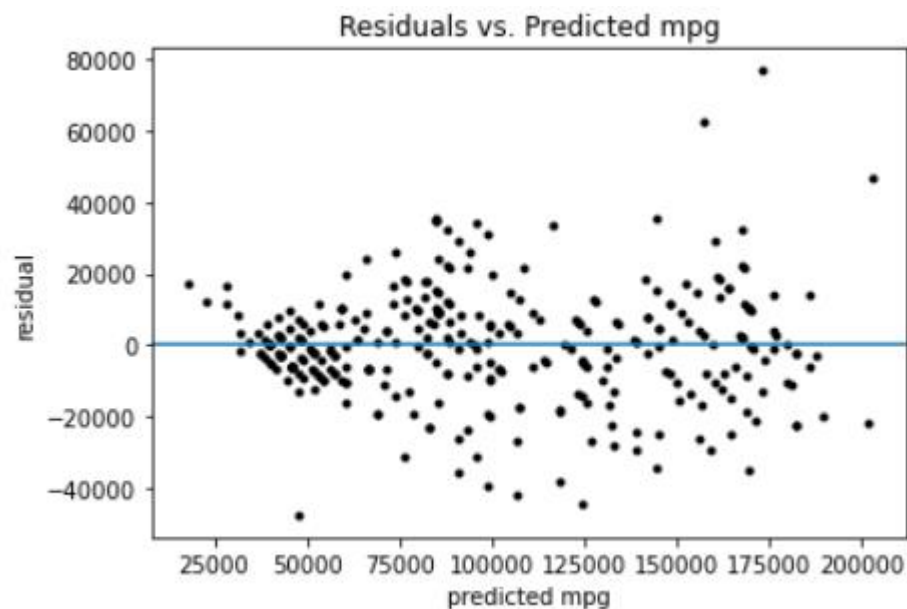
Jika membandingkan dua orang berusia 37 tahun yang memiliki jenis kelamin dan tingkat pendidikan yang sama, seseorang dengan lama pengalaman kerja lebih lama 1 tahun diperkirakan memiliki gaji lebih tinggi 2561 dollar.

Interpretasi intercept

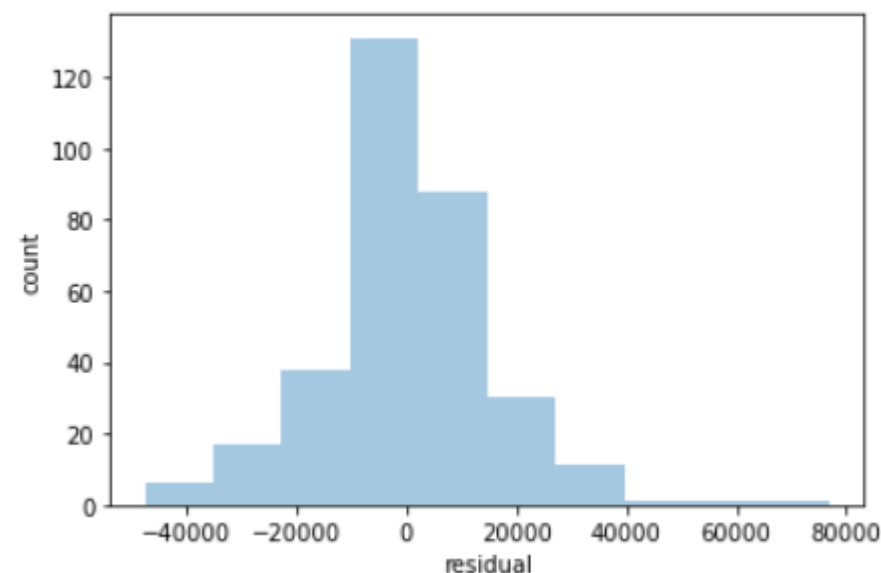
Seorang laki-laki yang berusia 37 tahun dengan tingkat pendidikan Bachelor's tanpa pengalaman kerja diperkirakan memiliki gaji sebesar 68396 dollar.

Regression: Multiple Predictors with One Interaction

Residual Plot



Normality of Error Assumption



Kesimpulan dan Saran

Kesimpulan

- Dapat disimpulkan bahwa usia, jenis kelamin, lama pengalaman kerja, dan tingkat pendidikan berpengaruh terhadap besaran gaji seseorang sehingga bisa digunakan untuk memprediksi besaran gaji tersebut.
- Model regresi yang dibangun dengan single predictor yaitu lama pengalaman kerja menghasilkan performa yang cukup bagus dengan R-squared 0,85. Transformasi logaritmik pada model ini tidak menghasilkan model yang lebih baik karena memiliki skor R-squared lebih rendah yaitu 0,76.
- Model regresi yang dibangun dengan semua predictor disertai interaksi antara usia dan lama pengalaman kerja, menghasilkan performa yang lebih baik dengan R-squared 0,89. Model tersebut juga menghasilkan interpretasi yang baik dengan dilakukannya centering pada variabel usia (age).

Saran

- Untuk pengembangan selanjutnya dapat dilakukan percobaan untuk berbagai variasi jumlah predictor yang digunakan. Dapat juga dilakukan pengelompokan data gaji berdasarkan industrinya sehingga didapatkan model regresi yang akurat untuk masing-masing industri.

Referensi

- Statistics for Business : Decision Making and Analysis — Robert Stine and Dean Foster
- Regression and Other Stories. — Andrew Gelman, Jennifer Hill, and Aki Vehtari
- The Effect: An Introduction to Research Design and Causality. Chapter 13
Huntington-Klein, N. 2021

Terima Kasih
