Statistics For Business Pacmann – Analisa Gaji Dengan Model Regresi

CHARLES SUGIANTO

1 Background dan Tujuan Analisa

- Dalam project kali ini, peneliti akan mengolah dataset yang menggambarkan pendapatan individu berdasarkan berbagai faktor seperti usia, jenis kelamin, pendidikan, posisi pekerjaan, dan pengalaman kerja. Peneliti ingin memahami bagaimana faktor-faktor ini mempengaruhi pendapatan seseorang dan juga untuk melakukan prediksi mengenai pendapatan individu.

2 Dataset dan Tools

A. Dataset: https://www.kaggle.com/datasets/rkiattisak/salaly-prediction-for-beginer

B. Tools: Jupyter Notebook

C. Dataset terdiri dari 375 baris data yang mencakup informasi tentang usia, jenis kelamin, tingkat pendidikan, jabatan, lama pengalaman kerja, dan besaran gaji. Sebelum dilakukan pengolahan lebih lanjut, tahapan persiapan dilakukan dengan menghilangkan nilai yang hilang dan data yang duplikat.

	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
...	...	...	...	...	...	...
348	28.0	Female	Bachelor's	Junior Operations Manager	1.0	35000.0
349	36.0	Male	Bachelor's	Senior Business Development Manager	8.0	110000.0
350	44.0	Female	PhD	Senior Data Scientist	16.0	160000.0
351	31.0	Male	Bachelor's	Junior Marketing Coordinator	3.0	55000.0
371	43.0	Male	Master's	Director of Operations	19.0	170000.0

324 rows × 6 columns

2 Dataset dan Tools

- Dalam proses pengolahannya terdapat 3 point penting yang menjadi perhatian peneliti, sebagai berikut:
  - 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.

3.1 Deskripsi Data Numerik

- Penelitian ini mengungkapkan bahwa terdapat suatu korelasi yang signifikan antara variabel usia, lama pengalaman kerja, dan besaran gaji dalam konteks populasi yang diamati. Hasil analisis statistik menunjukkan adanya hubungan positif yang kuat antara usia dan lama pengalaman kerja dengan besaran gaji yang diterima oleh individu-individu dalam sampel studi ini. Korelasi positif ini menunjukkan bahwa semakin tinggi usia dan semakin lama pengalaman kerja, semakin tinggi pula besaran gaji yang cenderung diterima oleh individu-individu dalam studi ini.

```
1 #2 DESKRIPSI DATA
2 df_salary.describe().transpose()
```

	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

```
1 #2A Korelasi dalam variabel angka
2 df_salary[["Age","YearsOfExperience", "Salary"]].corr()
```

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

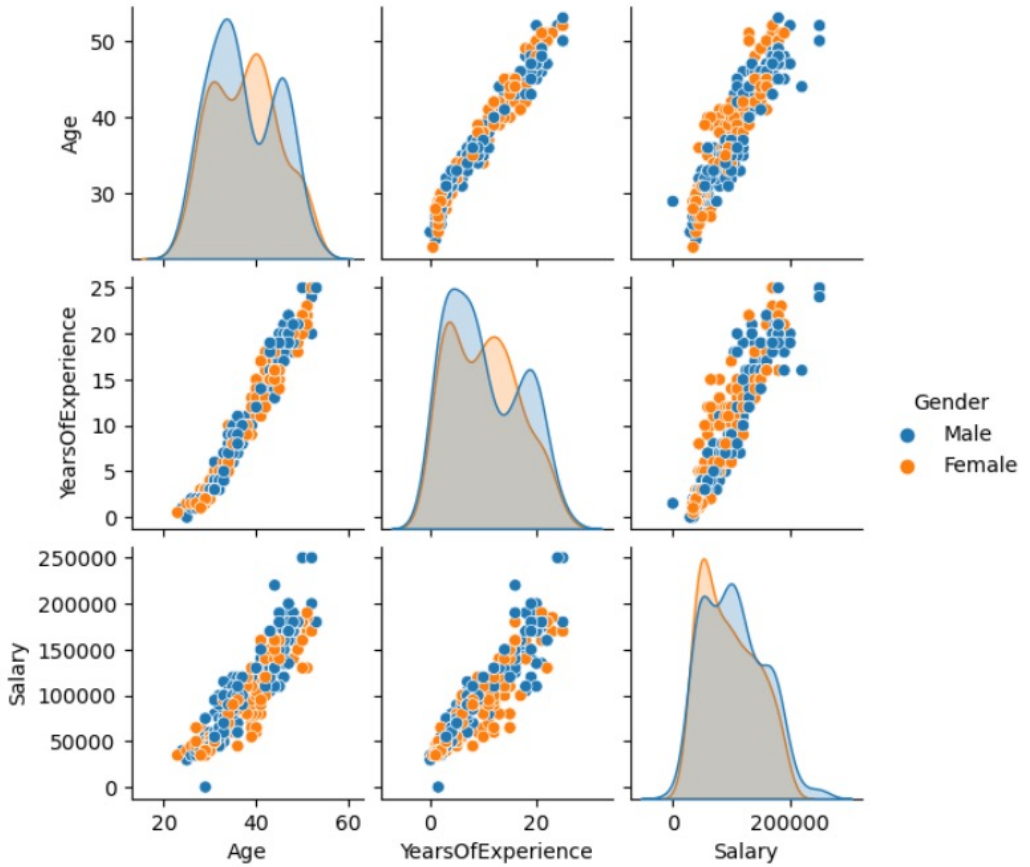
3.2 Deskripsi Data Kategorik

- Secara statistik, terdapat perbedaan yang signifikan dalam rata-rata gaji antara individu berjenis kelamin laki-laki dan perempuan, dengan rata-rata gaji laki-laki secara konsisten lebih tinggi dibandingkan dengan rata-rata gaji perempuan. Selain itu, terdapat pola yang menunjukkan bahwa rata-rata gaji meningkat seiring dengan peningkatan tingkat pendidikan, yang mengindikasikan adanya hubungan positif antara tingkat pendidikan yang lebih tinggi dan besaran rata-rata gaji dalam populasi yang diteliti.

```
In [67]:  1 df_salary["Gender"].value_counts()

Out[67]: Male      170
         Female    154
         Name: Gender, dtype: int64

In [68]:  1 df_salary["EducationLevel"].value_counts()

Out[68]: Bachelor's    191
         Master's       91
         PhD            42
         Name: EducationLevel, dtype: int64

In [72]:  1 #Gaji antar jenis kelamin
          2 df_salary.groupby("Gender")["Salary"].mean()

Out[72]: Gender
         Female     96136.363636
         Male      103472.647059
         Name: Salary, dtype: float64

In [73]:  1 #Gaji antar level pendidikan
          2 df_salary.groupby("EducationLevel")["Salary"].mean()

Out[73]: EducationLevel
         Bachelor's     73902.356021
         Master's      127912.087912
         PhD           158095.238095
         Name: Salary, dtype: float64
```

4 Visualisasi

- Dalam konteks penelitian ini, temuan menunjukkan bahwa terdapat suatu hubungan positif antara masa kerja individu dan besaran gaji yang diterimanya. Artinya, semakin lama seseorang telah bekerja, semakin tinggi pula besaran gaji yang cenderung diterimanya. Selain itu, temuan ini menunjukkan bahwa terdapat hubungan positif antara usia individu dan masa kerja yang dimilikinya. Ini mengindikasikan bahwa semakin tua seseorang, semakin banyak pengalaman kerja yang biasanya telah diakumulasinya.
- Namun, berdasarkan analisis yang dilakukan, jenis kelamin tidak memainkan peran yang signifikan dalam menentukan besaran gaji seseorang. Artinya, perbedaan gender tidak memiliki dampak yang cukup besar dalam memengaruhi besaran gaji individu dalam sampel yang diteliti.



5 Uji Statistik

- Dataset ini mencakup dua kategori jenis kelamin, yaitu *male* dan *female*. Peneliti ingin melakukan pengujian statistik untuk menentukan apakah terdapat perbedaan yang signifikan antara rata-rata gaji laki-laki dan rata-rata gaji Perempuan, dengan taraf signifikansi sebesar 10%. Selain itu, standar deviasi populasi tidak diketahui sehingga pengujian digunakan t-test.
- H0: μa = μb
- H1: μa > μb

- Uji Varians

```
#3B Analysis
# Gaji Male
df_male = df_salary[df_salary["Gender"]=="Male"]["Salary"].values
# Gaji Female
df_female = df_salary[df_salary["Gender"]=="Female"]["Salary"].values
# Variansi
np.var(df_male), np.var(df_female)
```

```
(2571353207.6989617, 2097896989.374262)
```

```
from scipy import stats
result = stats.ttest_ind(a = df_male, b = df_female, equal_var=False, alternative = "greater")
result.pvalue
```

```
0.08675461782037655
```

```
if result.pvalue < significance_level:
    print("Tolak hipotesis nol.")
else:
    print("Gagal menolak hipotesis nol.")
```

```
Tolak hipotesis nol.
```

Ada bukti yang cukup kuat menunjukkan bahwa terdapat perbedaan yang signifikan antara rata-rata gaji individu berjenis kelamin laki-laki dan perempuan. Rata-rata gaji laki-laki cenderung lebih tinggi daripada rata-rata gaji perempuan.

5 Uji Statistik

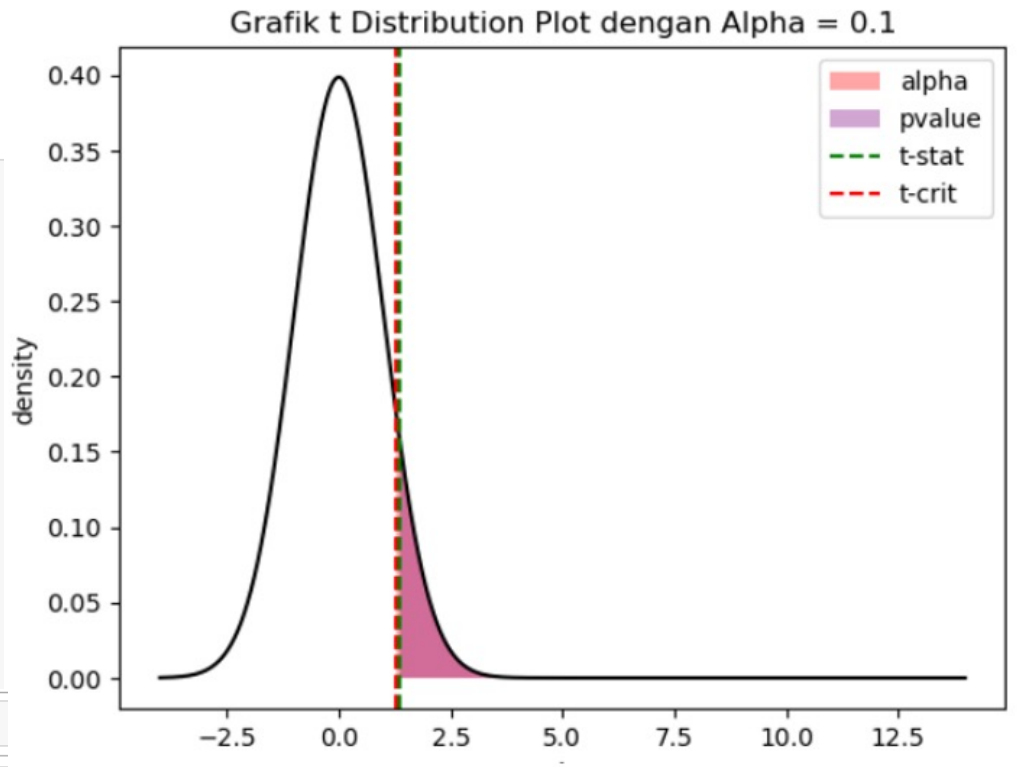
```python
# plot sample dist
x = np.arange(-4, 14, 0.001)
plt.plot(x, stats.t.pdf(x, df = df_data), color='black')
x_alpha = np.arange(stats.t.ppf(1-significance_level, df = df_data), 4, 0.01)
y_alpha = stats.t.pdf(x_alpha, df = df_data)
plt.fill_between(x = x_alpha, y1 = y_alpha, facecolor = 'red', alpha = 0.35, label = 'alpha')

# plot value
x_pvalue = np.arange(result.statistic, 4, 0.01)
y_pvalue = stats.t.pdf(x_pvalue, df = df_data)
plt.fill_between(x = x_pvalue, y1 = y_pvalue, facecolor = 'purple', alpha = 0.35, label = 'pvalue')
plt.axvline(np.round(result.statistic, 4), color ="green", linestyle = "--", label ="t-stat")
t_crit = np.round(stats.t.ppf(1-significance_level, df = df_data), 4)
plt.axvline(t_crit, color ="red", linestyle = "--", label ="t-crit")
plt.legend()
plt.xlabel("t")
plt.ylabel("density")
plt.title(f'Grafik t Distribution Plot dengan Alpha = {significance_level}');
plt.show()
```

```python
#3D Confidence Level
```

```python
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 adalah :", "[", lower, upper, "]")
```

```
Confidence Interval adalah : [ -1535.8717753119818 16208.438620231766 ]
```

- Derajat Kebebasan dan Confidence Level



- Berdasarkan hasil analisis, dapat disimpulkan bahwa dengan tingkat keyakinan sebesar 90%, terdapat bukti yang kuat bahwa rata-rata gaji laki-laki melebihi rata-rata gaji perempuan. Selain itu, hasil dari interval kepercayaan (confidence interval) menunjukkan bahwa dengan tingkat kepercayaan sebesar 90%, perkiraan interval untuk perbedaan rata-rata gaji adalah dari -1535 hingga 16208.

6.1 Model Regresi Single Predictor

- Terkait dengan prediksi gaji seseorang dari lama pengalaman kerjanya.
- Salary = 31960 + 6763 × Years of Experience

```
# Fit Linear Regression Using Horsepower Variable
# Create OLS model
model = smf.ols("Salary ~ YearsOfExperience", df_salary)
results_model_salary = model.fit()
results_salary = print_coef_std_err(results_model_salary)
results_salary
```

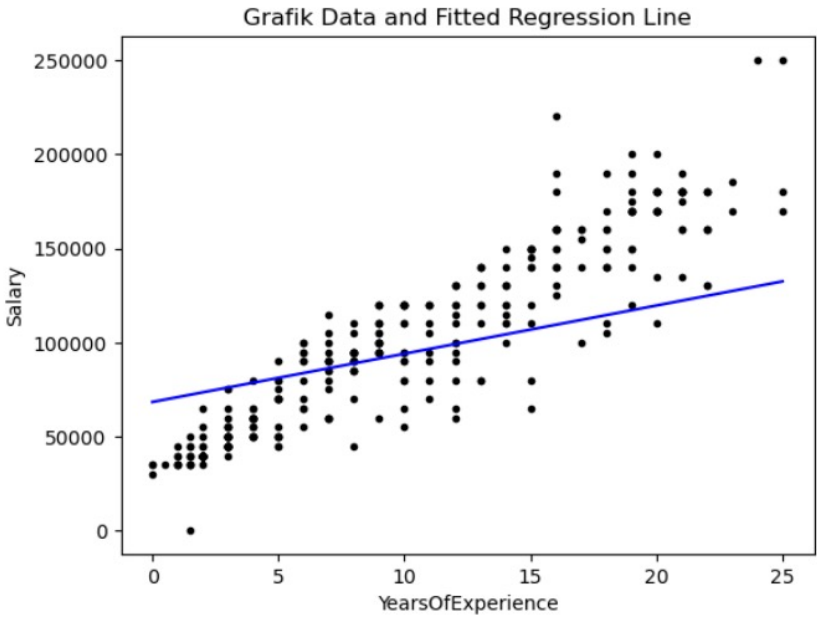
	coef	std err
Intercept	31959.508721	1873.552736
YearsOfExperience	6762.954641	155.446221

```
results_model_salary.rsquared
# Salary = 31960 + 6763 x Years of Experience
```

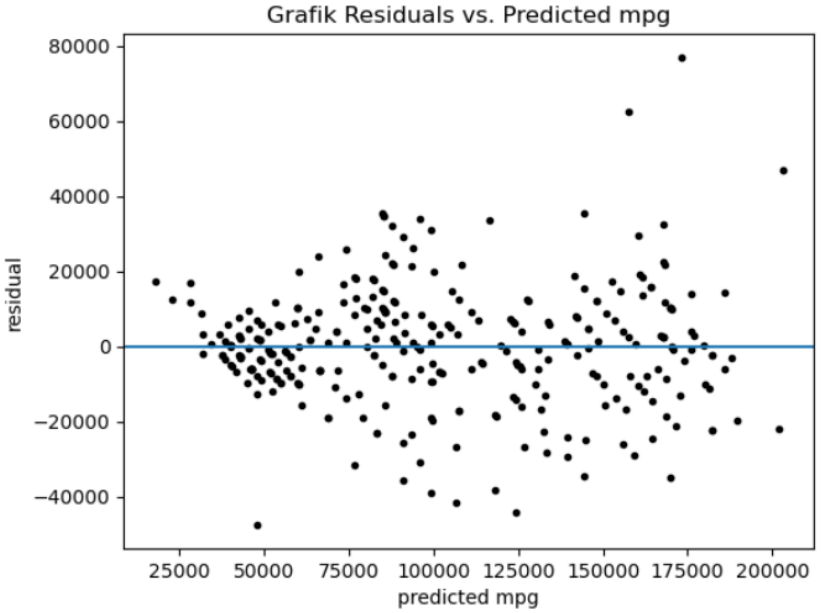
```
0.8546166681460778
```

Hasil analisis menghasilkan model regresi yang memiliki tingkat R-squared yang cukup tinggi sebesar 85,46%. Dalam konteks perbandingan dua individu dengan perbedaan pengalaman kerja selama satu tahun, model ini memprediksi bahwa individu dengan pengalaman kerja lebih lama 1 tahun akan memiliki perbedaan pendapatan sekitar 6763.

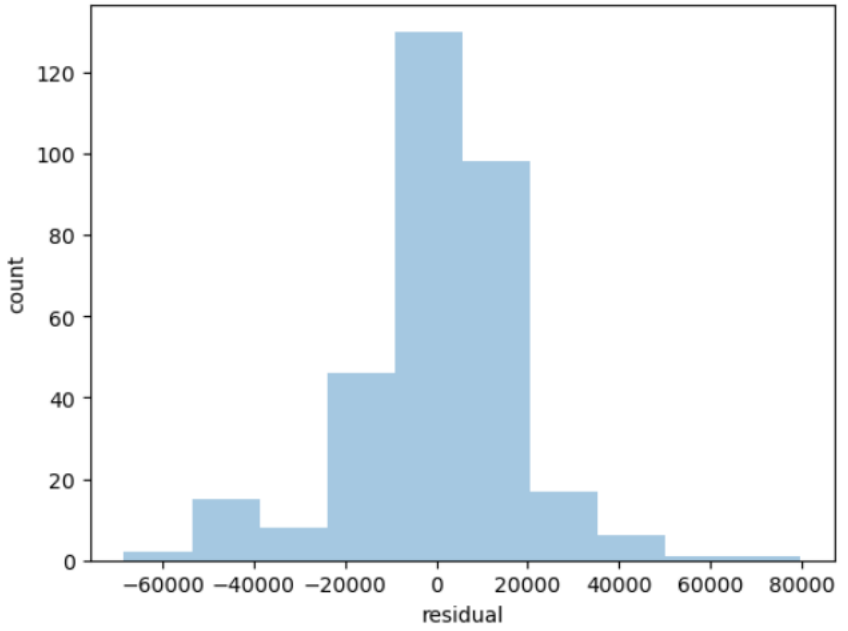
6.1 Model Regresi Single Predictor



- Residual Plot



- Normality of Error Assumption



6.2 Single Predictor with Log Transformation

- Terkait dengan gaji seseorang dari lama pengalaman kerjanya, serta transformasi logaritmik pada variabel prediktor.

```
# Create OLS
model = smf.ols("Salary ~ logYOE", df_salary)
results_logtransform = model.fit()
results_salary_log = print_coef_std_err(results_logtransform)
results_logtransform.rsquared
```

```
0.7656239539695425
```

Dalam hasil analisis, tercatat bahwa nilai R-squared yang diperoleh adalah sekitar 0,76. Nilai ini menunjukkan bahwa kemampuan model untuk menjelaskan variasi dalam data tidak sekuat pada model sebelumnya yang memiliki R-squared sebesar 0,85, yang tidak menggunakan transformasi logaritmik. Oleh karena itu, dalam konteks pemodelan regresi dengan satu variabel prediktor, lebih disarankan untuk menggunakan model regresi tanpa transformasi logaritmik.

6.3 Multiple Predictors with One Interaction

- Dalam analisis, peneliti akan menggunakan semua variabel prediktor, termasuk interaksi usia dan lama pengalaman kerja, dengan tingkat pendidikan sebagai variabel kategorikal.

6.3.1 Evaluasi model dengan K-Fold cross validation

- Nilai rata-rata R-squared yang ditemukan adalah sekitar 0,88, mengindikasikan bahwa model ini memiliki kualitas yang baik dan mampu menjelaskan sekitar 88% variasi dalam besaran gaji.

```
#Evaluate a model using K-fold cross validation
# Create a class model
ols_all_pred = StatsmodelsRegressor(
    smf.ols, "Salary ~ Age + Gender + C(EducationLevel) + YearsOfExperience + Age:YearsOfExperience")

# Create k-fold splitter object
kfold = KFold(n_splits=5, shuffle = True, random_state=123)
scores_ols_all_pred = cross_val_score(
    estimator = ols_all_pred, X = df_salary, y = df_salary["Salary"], cv = kfold,scoring = "r2")
scores_ols_all_pred = pd.DataFrame(data = scores_ols_all_pred, columns=["test_rsquared"])
scores_ols_all_pred["folds"] = [f"Folds {i+1}" for i in range(5)]
scores_ols_all_pred
```

	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

```
scores_ols_all_pred["test_rsquared"].mean()
```

```
0.8859529642576718
```

6.3.2 Fitting Model

- Koefisien persamaan regresi di atas menghasilkan intercept yang kurang optimal, karena interpretasinya tidak sesuai (gaji tidak dapat bernilai negatif) dan umumnya usia kerja tidak dimulai dari nol, sehingga variabel usia dicentering.

```
# Fit Linear Regression Using All Predictors
# Create OLS model
model = smf.ols("Salary ~ Age + Gender + C(EducationLevel) + YearsOfExperience + Age:YearsOfExperience", df_salary)
results_model_salary = model.fit()
results_salary = print_coef_std_err(results_model_salary)
results_salary
```

	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

6.3.3 Centering Variabel Usia

```
|:
1 # Centering Predictor Age
2 mean_age = df_salary["Age"].mean()
3 mean_age = np.round(mean_age,0)
4 mean_age
5
6 df_salary["Age"] = df_salary["Age"]-mean_age
7 df_salary.rename(columns = {"Age":"AgeCentered"}, inplace=True)
8 df_salary.head()
```

|:

	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

- Average = 37 tahun, digunakan sebagai dasar perhitungan

6.3.4 K Fold Cross Validation

```python
# Create a class model
ols_all_pred = StatsmodelsRegressor(
    smf.ols, "Salary ~ AgeCentered + Gender + C(EducationLevel) + YearsOfExperience + AgeCentered:YearsOfExperience")

# Create k-fold splitter
kfold = KFold(n_splits=5, shuffle = True, random_state=12)
scores_ols_all_pred = cross_val_score(
    estimator = ols_all_pred, X = df_salary, y = df_salary["Salary"], cv = kfold, scoring = "r2")
scores_ols_all_pred = pd.DataFrame(data = scores_ols_all_pred, columns=["test_rsquared"])
scores_ols_all_pred["folds"] = [f"Folds {i+1}" for i in range(5)]
scores_ols_all_pred
```

	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

```python
scores_ols_all_pred["test_rsquared"].mean()
#Model yang digunakan semua media memiliki kecocokan yang baik
#Dapat menjelaskan 89% varians gaji.
```

```
0.8901007028969221
```

- Tercapai nilai R-squared rata-rata sekitar 0,89, yang mengindikasikan bahwa model ini memiliki kualitas yang baik dan mampu menjelaskan sekitar 89% variasi dalam besaran gaji.

6.3.4 K Fold Cross Validation

```
:
1  # Create OLS model
2  model = smf.ols(
3      'Salary ~ AgeCentered + Gender + C(EducationLevel) + YearsOfExperience + AgeCentered:YearsOfExperience', df_salary)
4  results = model.fit()
5  results_salary = print_coef_std_err(results)
6  results_salary
```

:

	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

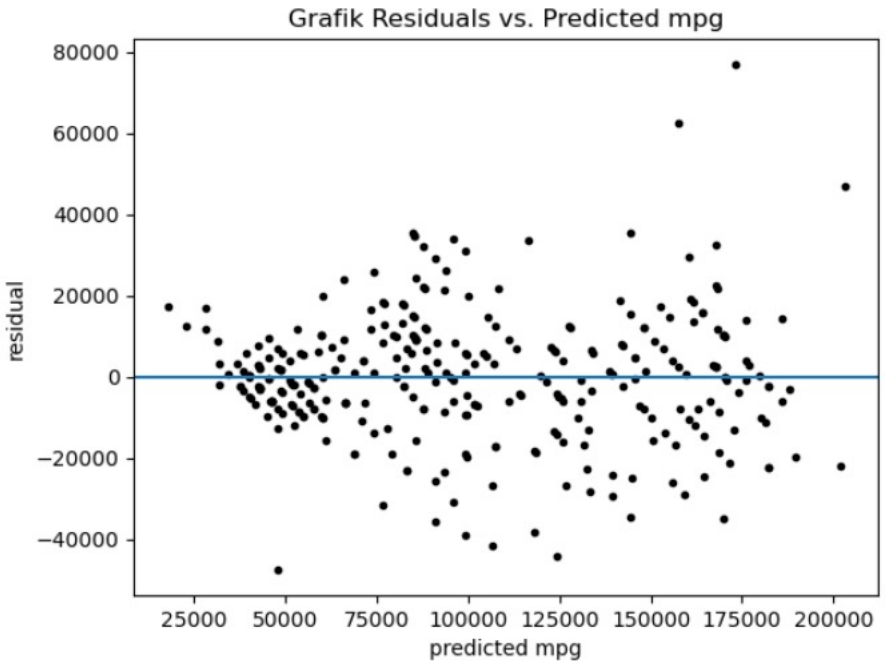
- # Gaji Bachelor = 68396+3042 ×(Age - 37)−9311×Gender+2561×YearsOfExperience+3×(Age - 37)×YearsOfExperience
- # Gaji Master = 68396+19574+3042×(Age - 37)−9311×Gender+2561×YearsOfExperience+3×(Age - 37)×YearsOfExperience
- # Gaji PhD = 68396+26339+3042×(Age - 37)−9311×Gender+2561×YearsOfExperience+3×(Age - 37)×YearsOfExperience

7 Penjelasan

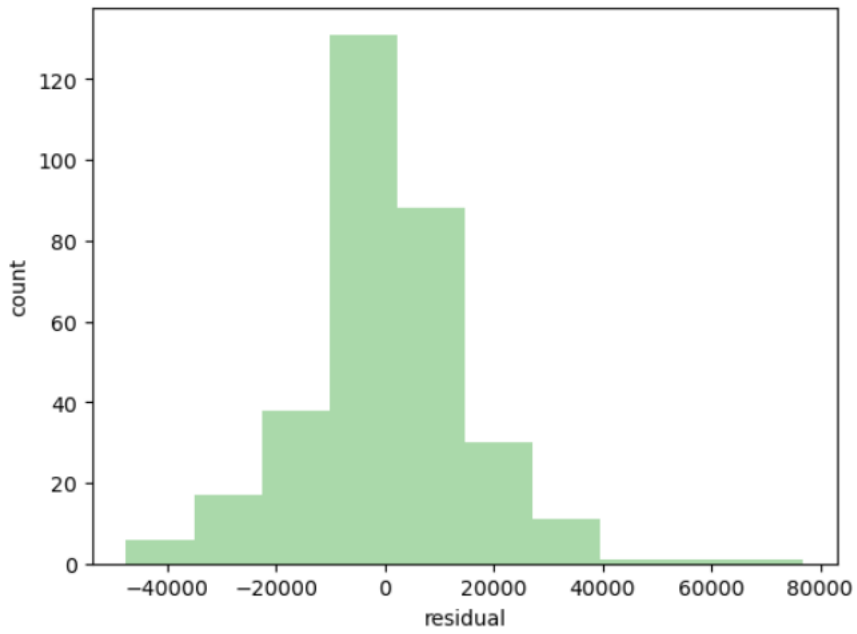
- Jenis kelamin: Perempuan diperkirakan memiliki gaji 9,311 dollar lebih rendah daripada laki-laki dengan kondisi yang sama.
- Lama pengalaman kerja: Lama pengalaman kerja tambahan 1 tahun diperkirakan meningkatkan gaji sebesar 2,561 dollar.
- Intercept: Seorang laki-laki berusia 37 tahun dengan gelar Bachelor’s tanpa pengalaman kerja diperkirakan memiliki gaji sebesar 68,396 dollar.
- Tingkat pendidikan: Seseorang dengan gelar Master’s diperkirakan memiliki gaji 19,574 dollar lebih tinggi daripada yang memiliki gelar Bachelor’s.
- Usia: Usia 1 tahun lebih tua dari 37 tahun diperkirakan berarti gaji lebih tinggi sebesar 3,042 dollar.

7 Penjelasan

- Residual Plot



- Normality of Error Assumption



8 Kesimpulan

- Berdasarkan hasil Analisa, maka dapat disimpulkan usia, jenis kelamin, lama pengalaman kerja, dan tingkat pendidikan memengaruhi gaji dan dapat digunakan dalam prediksi. Model regresi dengan lama pengalaman kerja sebagai satu-satunya predictor memiliki R-squared 0,85, sementara model dengan transformasi logaritmik memiliki R-squared 0,76. Model regresi dengan semua predictor dan interaksi usia dan lama pengalaman kerja memiliki R-squared 0,89, dengan centering pada variabel usia.

9 Referensi

- Anderson, D.R., Sweeney, D.J., Williams, T.A., Camm, J.D. and Cochran, J.J., 2016. *Statistics for business & economics*. Cengage Learning.
- Ellis, P.D., 2010. *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge university press.