

## 2022-1 – Upstart – Research Scientist Phone Interview

### Question 1.

When doing the multiple linear regression problem, one column B is wrongly copied as column A. What is the consequence?

Table 1 has the results with correct parameters and Table 2 has the results with incorrect parameter, Administration = Marketing Spend.

Coefficients on Table 2 of columns A and B are very closed to each other, but not identical.

Notice the variance matrix has the smallest eigenvalue closed to zero, that means there exists strong multicollinearity.

Table 1.

```
runfile('/Users/zli/Desktop/Multiple-Linear-Regression/multiple_linear_regression.py',  
wdir='/Users/zli/Desktop/Multiple-Linear-Regression')
```

Intercept:

42554.16761773238

Coefficients:

[ 7.73467193e-01 3.28845975e-02 3.66100259e-02 -9.59284160e+02  
6.99369053e+02]

#### OLS Regression Results

```
=====
Dep. Variable:          Profit  R-squared:          0.950
Model:                  OLS    Adj. R-squared:      0.943
Method:                 Least Squares  F-statistic:    129.7
Date:                   Fri, 07 Jan 2022  Prob (F-statistic): 3.91e-21
Time:                   21:36:42  Log-Likelihood:    -421.10
No. Observations:       40      AIC:                854.2
Df Residuals:           34      BIC:                864.3
Df Model:                5
Covariance Type:        nonrobust
=====
```

```
=====
=====
              coef  std err      t  P>|t|  [0.025  0.975]
-----
const      4.255e+04  8358.538   5.091  0.000  2.56e+04  5.95e+04
R&D Spend    0.7735    0.055  14.025  0.000    0.661    0.886
Administration 0.0329    0.066   0.495  0.624   -0.102    0.168
Marketing Spend 0.0366    0.019   1.884  0.068   -0.003    0.076
Florida     -959.2842  4038.108  -0.238  0.814  -9165.706  7247.138
```

New York      699.3691   3661.563   0.191   0.850   -6741.822   8140.560

```
=====
Omnibus:          15.823   Durbin-Watson:          2.468
Prob(Omnibus):    0.000   Jarque-Bera (JB):        23.231
Skew:             -1.094   Prob(JB):           9.03e-06
Kurtosis:         6.025   Cond. No.          1.49e+06
=====
```

Warnings:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 1.49e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Table 2.

```
runfile('/Users/zli/Desktop/Multiple-Linear-Regression/multiple_linear_regression.py',
wdir='/Users/zli/Desktop/Multiple-Linear-Regression')
```

Intercept:

46329.06017854024

Coefficients:

[ 7.85141717e-01 1.69781929e-02 1.69782094e-02 -8.26468159e+02  
5.54657333e+02]

OLS Regression Results

```
=====
Dep. Variable:          Profit   R-squared:          0.950
Model:                  OLS   Adj. R-squared:        0.944
Method:                 Least Squares   F-statistic:    165.6
Date:                   Fri, 07 Jan 2022   Prob (F-statistic): 3.19e-22
Time:                   21:37:55   Log-Likelihood:    -421.24
No. Observations:      40   AIC:                   852.5
Df Residuals:          35   BIC:                   860.9
Df Model:               4
Covariance Type:       nonrobust
=====
```

```
=====
              coef   std err   t   P>|t|   [0.025   0.975]
-----
const   4.633e+04   3375.873   13.724   0.000   3.95e+04   5.32e+04
R&D Spend   0.7851   0.049   15.924   0.000   0.685   0.885
Administration   0.0170   0.009   1.839   0.074   -0.002   0.036
Marketing Spend   0.0170   0.009   1.838   0.075   -0.002   0.036
Florida   -826.4682   3985.466   -0.207   0.837   -8917.395   7264.459
```

New York      554.6573   3610.268   0.154   0.879   -6774.576   7883.891

```
=====
Omnibus:          14.873   Durbin-Watson:          2.511
Prob(Omnibus):    0.001   Jarque-Bera (JB):         21.150
Skew:             -1.038   Prob(JB):             2.56e-05
Kurtosis:         5.895   Cond. No.          1.11e+16
=====
```

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The smallest eigenvalue is 3.82e-20. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

Question 2.

When doing the linear regression, if the dataset is wrongly copied twice. What is the consequence?

The estimated parameter won't change but the confidence interval (or the c.i. range) could shrink by approx  $\sqrt{2}$ . In addition, the R squared won't change but the adjusted R squared changes.

Table 1.

```
runfile('/Users/zli/Desktop/Multiple-Linear-Regression/multiple_linear_regression.py',
wdir='/Users/zli/Desktop/Multiple-Linear-Regression')
```

Intercept:  
45299.49140836343

Coefficients:  
[0.51986565]

OLS Regression Results

```
=====
Dep. Variable:          Profit   R-squared:          0.111
Model:                  OLS   Adj. R-squared:        0.087
Method:                 Least Squares   F-statistic:    4.726
Date:                   Fri, 07 Jan 2022   Prob (F-statistic): 0.0360
Time:                   21:50:10   Log-Likelihood:     -478.74
No. Observations:       40   AIC:                   961.5
Df Residuals:           38   BIC:                   964.9
Df Model:               1
Covariance Type:        nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
const	4.53e+04	3.02e+04	1.502	0.141	-1.57e+04	1.06e+05
Administration	0.5199	0.239	2.174	0.036	0.036	1.004
Omnibus:	0.124	Durbin-Watson:	1.946			
Prob(Omnibus):	0.940	Jarque-Bera (JB):	0.070			
Skew:	-0.081	Prob(JB):	0.966			
Kurtosis:	2.874	Cond. No.	6.14e+05			

#### Warnings:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 6.14e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Table 2.

```
runfile('/Users/zli/Desktop/Multiple-Linear-Regression/multiple_linear_regression.py',
wdir='/Users/zli/Desktop/Multiple-Linear-Regression')
```

Intercept:

45299.491408363414

Coefficients:

[0.51986565]

#### OLS Regression Results

Dep. Variable:	Profit	R-squared:	0.111
Model:	OLS	Adj. R-squared:	0.099
Method:	Least Squares	F-statistic:	9.700
Date:	Fri, 07 Jan 2022	Prob (F-statistic):	0.00258
Time:	21:51:41	Log-Likelihood:	-957.48
No. Observations:	80	AIC:	1919.
Df Residuals:	78	BIC:	1924.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
const	4.53e+04	2.1e+04	2.153	0.034	3403.607	8.72e+04

Administration	0.5199	0.167	3.115	0.003	0.188	0.852
----------------	--------	-------	-------	-------	-------	-------

```
=====
Omnibus:          0.109  Durbin-Watson:          1.976
Prob(Omnibus):    0.947  Jarque-Bera (JB):          0.140
Skew:            -0.081  Prob(JB):          0.932
Kurtosis:         2.874  Cond. No.          6.14e+05
=====
```

Warnings:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 6.14e+05. This might indicate that there are strong multicollinearity or other numerical problems.

## 2022-3-17 Meta 买它 infra research data scientist 电面

ML/STATISTICS: credit fraud,

Q1: given amount and distance as features, what algorithm you will use?

Answer: Build a classification model to predict probability of fraud.

Q2: what other algorithms you can think of and what are the pro and cons compared to the one you proposed in Q1.

Answer:

2 features -> decision tree/boosting/deep learning is not adequate.

Decision Tree:

- \* Not be efficient because lots of data but very few features

KNN:

- \* Frauds change over time, not a good patterns as new tech used in the new fraud cases
- \* Save all the data but not training needed

Anomaly Detection (to be reviewed):

- \* Distribution of individual features

Logistic regression:

- \* Good interpretability
- \* Score fast
- \* Training is relatively slow
- \* Its relative simplicity makes it a high-bias and low-variance model, so it may not perform well when the decision boundary is not linear.

Q3: coefficient of amount to fraudulence is 0.10 with standard error 0.02, what's the relationship between amount and fraudulence? Is it statistically significant? How do you prove it?

Answer: (See ESL Page 124) Each unit increase in the distance accounts for *an increase in the odds of fraudulence of  $\exp(0.10) \approx 1.105$  or 10.5% (alternatively the increase in the log-odds of*

*fraud of 0.1 or 10%*). The Z score is  $0.10/0.02=5$  which means the coefficient is significant. The is proved by the CLT.