투자론

- R과 Excel을 통한 금융데이터 분석 -

6주차
회귀 분석 및 APT and Multifactor Model

충남대학교
장호규 교수

Unit 01

Brief Introduction to Regression

Regression

- Technique used for the modeling and analysis of numerical data
- Exploits the relationship between two or more variables so that we can gain information about one of them through knowing values of the other
- Regression can be used for prediction, estimation, hypothesis testing, and modeling causal relationships

Regression Lingo

$$Y = \underbrace{X_1 + X_2 + X_3}$$

Dependent Variable
Outcome Variable
Response Variable

Independent Variable
Predictor Variable
Explanatory Variable

Why Linear Regression?

- Suppose we want to model the dependent variable Y in terms of three predictors, X_1, X_2, X_3:

$$Y = f(X_1, X_2, X_3)$$

- Typically will not have enough data to try and directly estimate f

- Therefore, we usually have to assume that it has some restricted form, such as linear:

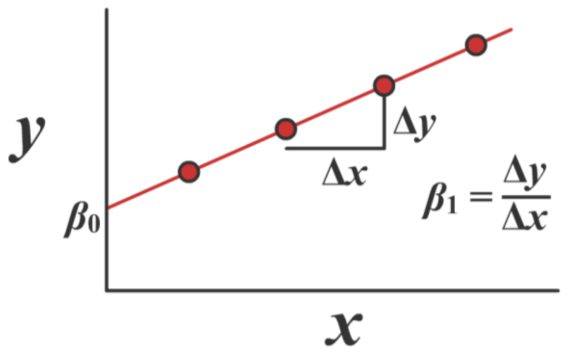
$$Y = X_1 + X_2 + X_3$$

Linear Regression

A Probabilistic Model

- Much of mathematics studies variables that are deterministically related to one another

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

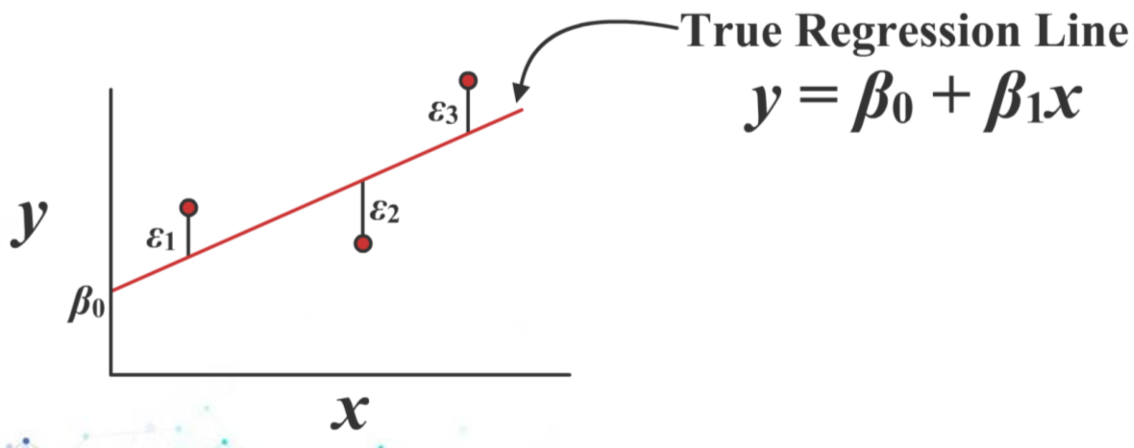


- But a non-deterministic relation between variables is more interesting and more realistic

Linear Regression

A Probabilistic Model

- Definition: There exists parameters β_0, β_1, and σ^2 such that for any fixed value of the independent variable, x, the dependent variable is related to x through the model equation: $y = \beta_0 + \beta_1 x + \in$
- $\in$ is a random variable that follows $N(0, \sigma^2)$

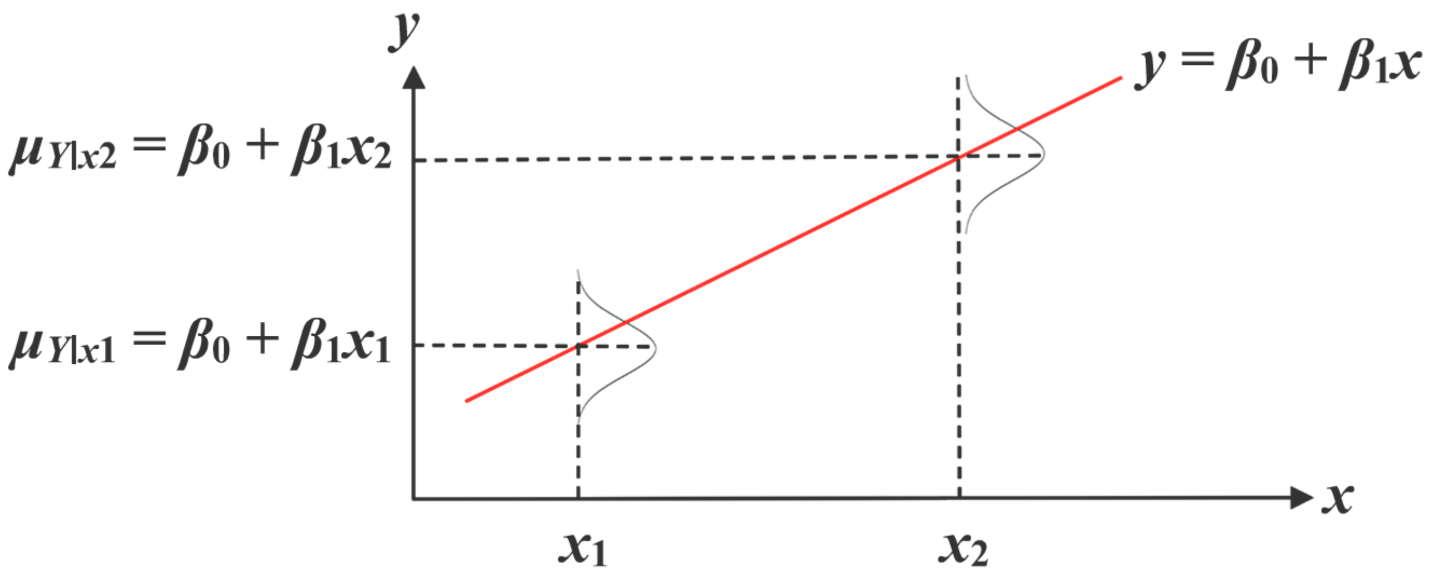


Linear Regression

A Probabilistic Model

- True regression line – meaning:
- $E(Y|X) = \beta_0 + \beta_1 X$
- The expected value of Y is a linear function of X, but for fixed x, the variable Y differs from its expected value by a random amount
- Formally, let x^* denote a particular value of the independent variable, x, then the linear probabilistic model says
- $E(Y|x^*) = \mu_{Y|x^*}$ = mean value of Y when x is x^*
- $Var(Y|x^*) = \sigma^2_{Y|x^*}$ = variance of Y when x is x^*

Graphical Interpretation

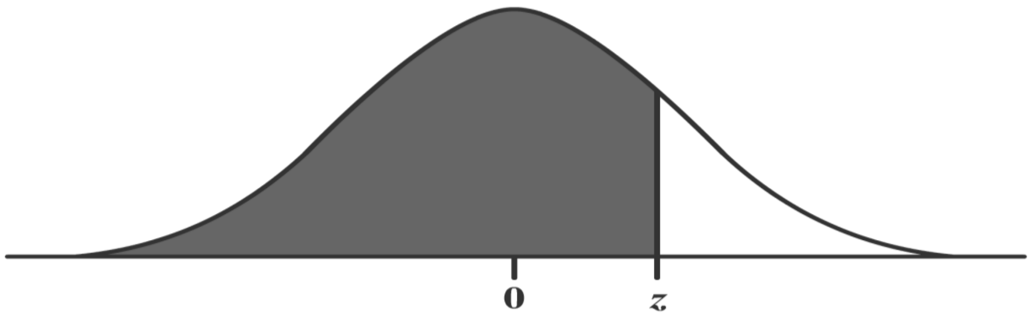


- For example, if x = height and y = weight, then $\mu_{Y|X=60}$ is the average weight for all individual 60 inches tall in the population

Example

- Suppose the relationship between the independent variable height (x) and dependent variable weight (y) is described by a simple linear regression model with true regression line y=7.5+0.5x and σ=3
  - Q1: What is the interpretation of $\beta_1 = 0.5$?
  - Q2 : If $x = 20$, what is the expected value of Y?
  - Q3 : If $x = 20$, what is $P(Y > 22)$?

Brief Introduction to Regression



$$x \sim N(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

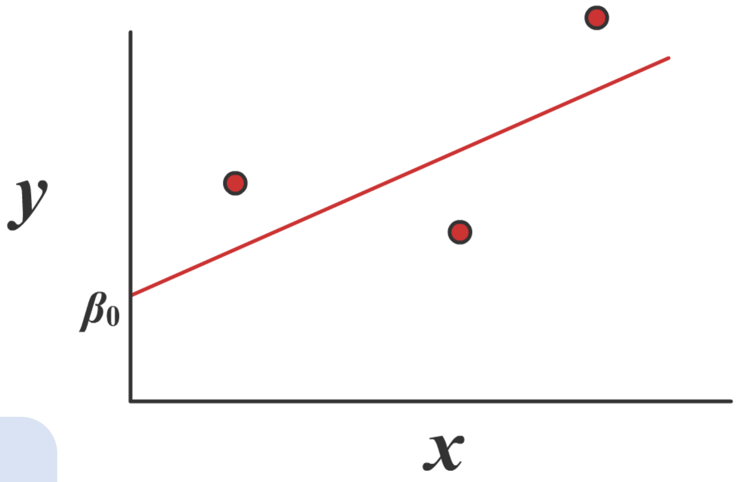
Normal Deviate z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-4.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.8	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.7	.0001	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483

출처 : 교수자 제공

Estimating Model Parameters

- Point estimates of $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are obtained by the principle of least squares:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)^2]$$



$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{s_{xy}}{s_x^2} = \rho_{x,y}\frac{s_y}{s_x}$$

$$\widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x}$$

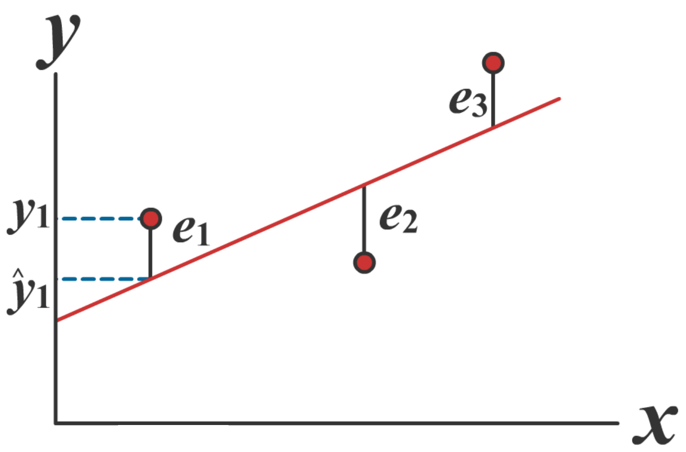
Estimating Model Parameters

- Predicted or fitted, values are values of y predicted by the least squares regression line obtained by plugging in $x_1, \cdots, x_n$ into the estimated regression line

$$\widehat{y_1} = \widehat{\beta_0} + \widehat{\beta_1} x_1, \; \widehat{y_2} = \widehat{\beta_0} + \widehat{\beta_1} x_2, \cdots$$

- Residuals are the deviations of observed and predicted values

$$\widehat{e_1} = y_1 - \widehat{y_1}, \; \widehat{e_2} = y_2 - \widehat{y_2}, \cdots$$



Decomposition of Sum of Squares

$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}$$

$$\Rightarrow (y_i - \bar{y})^2 = (y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

- Taking summation on both sides yields

$$\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\Leftrightarrow TSS = ESS + RSS \Longrightarrow 1 = \frac{ESS}{TSS} + \frac{RSS}{TSS} = R^2 + \frac{RSS}{TSS}$$

Statistical Test

- H_0 vs H_1: usually H_0 implies $\beta_1 = 0$ (no effect, most conservative)
  - (We do not cover details of the statistical inference.) Under the assumption that error terms are normally distributed, $\epsilon_i \sim N(0, \sigma^2)$ followings are known:

$$\hat{\beta}_0 \sim N\left(\beta_0, \left\{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right\}\sigma^2\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)$$

Standard error = S.E.

  - Usually σ^2 is unknown. In this case, we use the sample variance, s^2, instead of σ^2 in the above expression. But, the probability distribution changes from normal distribution to Student's **t-distribution**

Statistical Test

t-distribution

- $H_0 : \beta_1 = b_1, \beta_0 = b_0$, then

$$\frac{\frac{\widehat{\beta_1}-b_1}{s}}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}} \sim t(n-2) \quad , \frac{\widehat{\beta_0}-b_0}{s\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}} \sim t(n-2)$$

- 95% Confidence Interval (2-sided test):

$$[\widehat{\beta_1} - t_{0.025}(n-2) \times S.E.(\widehat{\beta_1}), \widehat{\beta_1} + t_{0.025}(n-2) \times S.E.(\widehat{\beta_1})]$$

$$[\widehat{\beta_0} - t_{0.025}(n-2) \times S.E.(\widehat{\beta_0}), \widehat{\beta_0} + t_{0.025}(n-2) \times S.E.(\widehat{\beta_0})]$$

Hypothesis Test

- $H_0 : \beta_1 = b_1, \beta_0 = b_0$
  - We would reject these hypotheses if

$$\left|\frac{\widehat{\beta_1} - b_1}{S.E.(\widehat{\beta_1})}\right| > t_{0.025}(n-2)$$

$$\left|\frac{\widehat{\beta_0} - b_0}{S.E.(\widehat{\beta_0})}\right| > t_{0.025}(n-2)$$

Hypothesis Test

Criticla Values for Two-Sided and One-sided Tests Using the Student t Distribution

Degree of Freedom	Signigicance Level 20%(2-sided) 10%(1-sided)	10%(2-sided) 5%(4-sided)	5%(2-sided) 2.5%(1-sided)	2%(2-sided) 1%(1-sided)	1%(2-sided) 0.5%(1-sided)
1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79
26	1.32	1.71	2.06	2.48	2.78
27	1.31	1.70	2.05	2.47	2.77
28	1.31	1.70	2.05	2.47	2.76
29	1.31	1.70	2.05	2.46	2.76
30	1.31	1.70	2.04	2.46	2.75
60	1.30	1.67	2.00	2.39	2.66
90	1.29	1.66	1.99	2.37	2.63
120	1.29	1.66	1.98	2.36	2.62
∞	1.28	1.64	1.96	2.33	2.58

- Valuse are shown for the critical values for two-sided($\neq$)and one-side($>$) alternative hypotheses.
  The critical value for the one-sided($<$) test is the negative of the one-sided($>$) critical value shown in the table.
  For example, 2.13 is the critical value for a two-sided test with a significance level of 5% using the Student t distribution with 15 degrees of freedom.

출처 : 교수자 제공

Multiple Linear Regression

- Extension of the simple linear regression model to two or more independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

  - Expression = Baseline + Age + Tissue + Sex + Error

- Partial Regression Coefficients: $\boldsymbol{\beta}_i \equiv$ effect on the dependent variable when increasing the i^{th} independent variable by 1 unit, holding all other predictors constant

Example with Real Data

- We run simple linear regressions to obtain betas.

$$r_i - r_f = \alpha_i + \beta_i \left(r_m - r_f\right) + \varepsilon_i$$

For simplicity, we assume zero riskless rate.

Use the following data (use the excel sheet attached)

	A	B	C	D	E	F	G	H	I	J	K	L
1	PRICE DATA: 10 STOCKS AND SP500, 2015-2020 SP500 represented by Vanguard's Index 500 fund (includes dividends)											
2		1	2	3	4	5	6	7	8	9	10	11
3		Apple	Google	Amazon	Seagate	Comcast	Merck	Johnson-Johnson	General Electric	Hewlett Packard	Goldman Sachs	SP500
4												
5	Date	AAPL	GOOG	AMZN	STX	CMCSA	MRK	JNJ	GE	HPQ	GS	GSPC
6	01-Jan-20	308.78	1,434.23	2,008.72	56.08	42.99	84.74	147.93	12.44	21.12	236.31	3,225.52
7	01-Dec-19	292.95	1,337.02	1,847.84	57.92	44.76	89.59	144.95	11.14	20.18	228.53	3,230.78
8	01-Nov-19	265.82	1,304.96	1,800.80	58.10	43.94	85.88	135.68	11.25	19.72	218.77	3,140.98
9	01-Oct-19	247.43	1,260.11	1,776.66	56.49	44.40	85.36	130.30	9.96	17.06	210.89	3,037.56
10	01-Sep-19	222.77	1,219.00	1,735.91	51.74	44.66	82.37	127.68	8.91	18.43	204.82	2,976.74
11	01-Aug-19	206.84	1,188.10	1,776.29	48.29	43.85	84.62	125.73	8.23	17.81	200.28	2,926.46
12	01-Jul-19	211.10	1,216.68	1,866.78	44.54	42.56	81.21	127.55	10.42	20.49	216.21	2,980.38
13	01-Jun-19	196.12	1,080.91	1,893.63	44.68	41.68	81.51	136.42	10.46	20.08	193.50	2,941.76
14	01-May-19	172.81	1,103.63	1,775.07	39.68	40.42	77.00	127.59	9.40	18.05	178.43	2,752.06