# ST517 Note Outline 11: Model Fitting

## Lecture 11.1 Estimating regression coefficients via matrix algebra

**Recall:** The multiple regression equation $y=\beta_0+\beta_1 x_1+\beta_2 x_2 \ldots+\beta_k x_k+\varepsilon$

Focusing on the individuals: $y_i=\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i} \ldots+\beta_k x_{ki}+\varepsilon_i$

Using matrix notation: $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$

where

$$\mathbf{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\mathbf{X}=\begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ 1 & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}$$

$$\boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$\boldsymbol{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

**Estimating the regression coefficients**

- **Recall:** Least squares estimates $\hat{\boldsymbol{\beta}}$ minimize SSE
- In matrix notation, $SSE = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)' \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$
- Thus the least squares estimates can be written as: $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$
- The predicted values of y can be written as: $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$

**Example:** The data below show the number of bedrooms, number of bathrooms, and the sale prices for a random sample of eight homes a recently sold in a particular city. Use this data to derive the regression coefficients for the model using the number of bedrooms and bathrooms to predict sale price.

| **y = sale price (in thousands of dollars)** | **$x_1$ = number of bedrooms** | **$x_2$ = number of bathrooms** |
|---|---|---|
| 78.8 | 3 | 2 |
| 74.3 | 2 | 1 |
| 83.8 | 4 | 3 |
| 74.2 | 2 | 1 |
| 79.7 | 3 | 2 |
| 74.9 | 2 | 2 |
| 88.4 | 5 | 3 |
| 82.9 | 4 | 2 |

Writing the data table in matrix notation gives

$$\mathbf{y} = \begin{bmatrix} 78.8 \\ 74.3 \\ 83.8 \\ 74.2 \\ 79.7 \\ 74.9 \\ 88.4 \\ 82.9 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 3 & 2 \\ 1 & 2 & 1 \\ 1 & 4 & 3 \\ 1 & 2 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \\ 1 & 5 & 3 \\ 1 & 4 & 2 \end{bmatrix}$$

**Example (continued):** Solving for the quantities necessary to find the regression coefficients:

$$\mathbf{X'y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 2 & 4 & 2 & 3 & 2 & 5 & 4 \\ 2 & 1 & 3 & 1 & 2 & 2 & 3 & 2 \end{bmatrix} \begin{bmatrix} 78.8 \\ 74.3 \\ 83.8 \\ 74.2 \\ 79.7 \\ 74.9 \\ 88.4 \\ 82.9 \end{bmatrix} = \begin{bmatrix} 637 \\ 2031.1 \\ 1297.7 \end{bmatrix}$$

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 2 & 4 & 2 & 3 & 2 & 5 & 4 \\ 2 & 1 & 3 & 1 & 2 & 2 & 3 & 2 \end{bmatrix} \begin{bmatrix} 1 & 3 & 2 \\ 1 & 2 & 1 \\ 1 & 4 & 3 \\ 1 & 2 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \\ 1 & 5 & 3 \\ 1 & 4 & 2 \end{bmatrix} = \begin{bmatrix} 8 & 25 & 16 \\ 25 & 87 & 55 \\ 16 & 55 & 36 \end{bmatrix}$$

$$\left(\mathbf{X'X}\right)^{-1} = \frac{1}{84} \begin{bmatrix} 107 & -20 & -17 \\ -20 & 32 & -40 \\ -17 & -40 & 71 \end{bmatrix}$$

Finally, it can be shown that $\hat{\boldsymbol{\beta}} = \left(\mathbf{X'X}\right)^{-1}\mathbf{X'y} = \begin{bmatrix} 65.19 \\ 4.13 \\ 0.758 \end{bmatrix}$

Thus the estimated regression equation is:

**What should you be able to do?**

- Write the design matrix based on a data table
- Recognize dimensions for the design matrix, parameter vector, **X'X** etc.
- Write the estimated regression equation based on an estimated parameter vector
- Multiply, transpose, and invert small matrices

## Lecture: 11.2 Model Building

### Choosing the best model

- When choosing between two (or more) models using different subsets of predictors, there are several statistics that can be used to help pick the "best" model
    - Keep in mind:

### $R^2$ criterion

- Basic idea:

- Works well for comparing simple linear regression models
    - Adjusted $R^2$ criterion better for multiple regression models

### Adjusted $R^2$ criterion

- Also known as MSE criterion since $R_a^2 = 1 - (n-1)\left[\frac{MSE}{SS_{yy}}\right]$

- Basic idea:

- Automatically printed with regression output, making for easy model comparison

### AIC and BIC

- Akaike information criterion: AIC = $2k - 2\ln(L)$
- Bayesian information criterion: BIC = $-2\ln(L) + k \cdot \ln(n)$
    - $L$ represents the maximum value of the likelihood function for the model

    - Terms involving $k$ represent

- Basic idea:

**Other considerations for model building**

- Multicollinearity
    - Two independent variables are **<u>orthogonal</u>** if their sample correlation coefficient is zero
    - If all pairs of independent variables are orthogonal, it is easy to solve the for the parameter estimates
    - If not, the matrix used to solve for the parameters is **<u>singular</u>**

    - Signs of multicollinearity:

    - Fixing Multicollinearity:

**Other considerations for model building (continued)**

- Estimating the regression coefficients (parameters) may not be possible if:

- Variable transformations
    - Consider if:

    - Common transformations:

**Final reminders**

- Always let context and knowledge be your guide!
- Start with the predictors that you believe are the most important
- Explore the data graphically to determine if curvilinear terms are appropriate
- Consider which interactions are most likely to be important or "real"
- Always check the fit of the model
- Go with the simplest model that fits the data well (**parsimony**)

## Lecture 11.3 Residual Analysis

**Recall:**

- Inference in regression context requires several assumptions about the error terms
  - Normally distributed
  - Centered at zero (average value = 0)
  - Constant variance
  - Independent

  ⇒

**Recall: Residual Plot**

- Plot of residuals (on the vertical axis) vs. predicted values (on the horizontal axis)
  - Shows if model is a good fit overall
- Plot of residuals (vertical axis) vs. values for one of the predictors (horizontal axis)
  - Shows if that predictor is well-modeled
- Should see:
  - Random scatter of residuals around zero
  - No systematic pattern
- What to do if there is a problem?
  - Evidence of non-linear relationship between x and y

  - Evidence of non-constant variance

**Checking Normality of the Residuals**

- Recall: QQ-plot checks is the data follow a normal distribution
  - If so, the plot shows a straight, upward sloping (nearly 45°) line
  - Any major departures from this indicate the normal distribution may not be a good fit for the data
- Could also look at a histogram of the residuals

**Studentized Residuals**

$$z_i^* = \frac{e_i}{s\sqrt{1-h_i}} = \frac{y_i - \hat{y}_i}{s\sqrt{1-h_i}}$$

where $h_i$ represents the amount of **<u>leverage</u>** a point has on the fitting of the model

- Useful for:
- Rule of thumb:

**Leverage**

- Recall the matrix formulation of the predicted values: $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$

  where $\boldsymbol{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ is the **<u>hat matrix</u>**

- For an individual observation *i*: $\hat{y}_i = h_1 y_1 + \cdots + h_i y_i + \cdots + h_n y_n$
- Leverage for observation *i*: $h_i =$ the diagonal entry in the hat matrix

- Useful for:
- Rule of thumb:

**Jackknife (or Deleted) Residuals**

$$d_i = y_i - \hat{y}_{(i)} = \frac{\hat{\varepsilon}_i}{1-h_i}$$

- **<u>Jackknife</u>**: Delete one observation at a time, fit the model using only the remaining $n-1$ observations; repeat for all observations

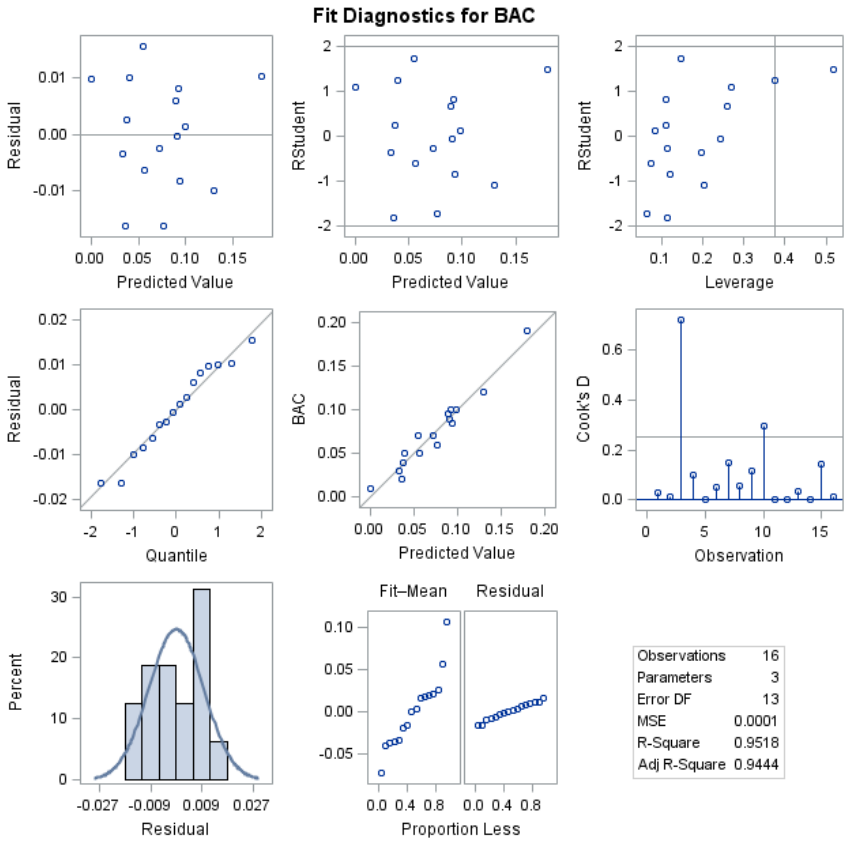
**Cook's Distances**

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(k+1) \cdot MSE}\left[\frac{h_i}{(1-h_i)^2}\right] = \frac{z_i^2}{p}\left[\frac{h_i}{1-h_i}\right]$$

- If the $i^{\text{th}}$ observation has a large studentized residual and high leverage, $D_i$ will be large
- Useful for:

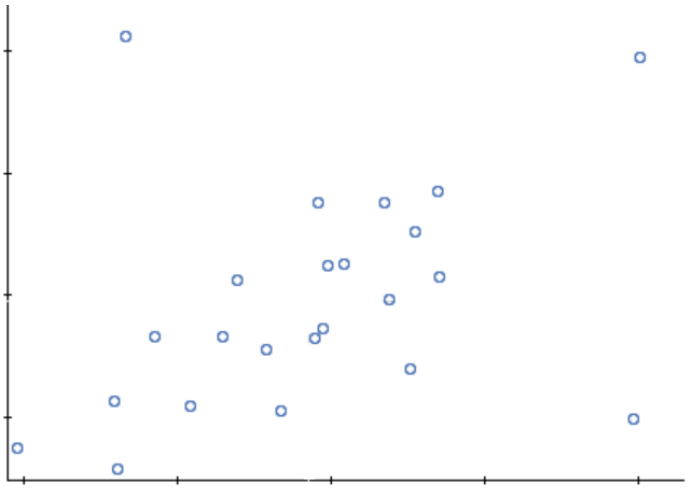
- Rule of thumb:

**Example:** Regression diagnostics for BAC model with beers, weight



## Lecture 11.4: Outliers, Leverage, and Influence

- Outlier: A point that doesn't fit overall pattern of the data
- Leverage point: Pulls the model close to it, due to a lack of neighboring points
- Influential observation: Has a large effect on the fitted model
-



**What if you find such a point?**

- Investigate that it is valid point
- Fit the model without that point and see if/how the fit changes
- Report it

## Lecture 11.5: Additional Examples

**Use the following to answer questions 1 to 5:** Recall (from the previous outline) the study conducted to determine whether infection surveillance and control program have reduced the rates of hospital-acquired infections. The data is a random sample of 28 US hospitals with variables: RISK ($y$), STAY ($x_1$), AGE ($x_2$), INS ($x_3$), SCHOOL ($x_4$), RC1 ($x_5$), RC2 ($x_6$), and RC3 ($x_7$).

1. The first three observations (OBS) of the data are shown below. Use this to construct the first three rows of the design matrix that would be used to estimate the following models:

   a. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_6 + \hat{\beta}_7 x_7$

   b. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_6 + \hat{\beta}_7 x_7 + \hat{\beta}_8 x_1 x_2$

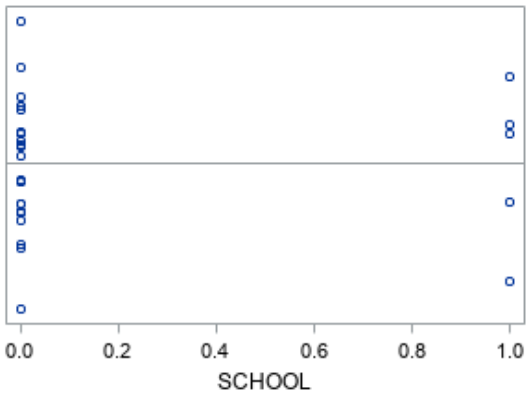
   c. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_6 + \hat{\beta}_7 x_7 + \hat{\beta}_8 x_3^2$

| OBS | RISK | STAY | AGE | INS | SCHOOL | RC1 | RC2 | RC3 |
|---|---|---|---|---|---|---|---|---|
| 1 | 4.1 | 7.13 | 55.7 | 9.0 | 0 | 0 | 0 | 1 |
| 2 | 1.6 | 8.82 | 58.2 | 3.8 | 0 | 1 | 0 | 0 |
| 3 | 2.7 | 8.34 | 56.9 | 8.1 | 0 | 0 | 1 | 0 |

2. Write the dimensions for each component of $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ ; consider the full design matrix (not just the first 3 rows) for each model in the previous question. (So your final answer should have dimensions for $\boldsymbol{y}$, $\boldsymbol{\varepsilon}$, and three different **X**s and $\boldsymbol{\beta}$s.)

On page 12 is SAS output for the estimated model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$ (call this model 1) Use this output to answer questions 3 to 5.

3. The plot of residuals vs. school is at the right. Why are there only two lines of residuals? (Do not overthink the answer to this question! It may help to review the variable descriptions from the last outline.)
4. Based on the residual plots, which of the conditions for the model to be valid appear to be met? For which are there reason to be concerned? Explain, referencing provided plot numbers.




5. One page 13 is SAS output for the estimated model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_3$ (call this model 2). Which model, 1 or 2, would you recommend using to predict RISK? Pick one model and explain your reasoning. Also explain any further changes to your chosen you would recommend investigating.

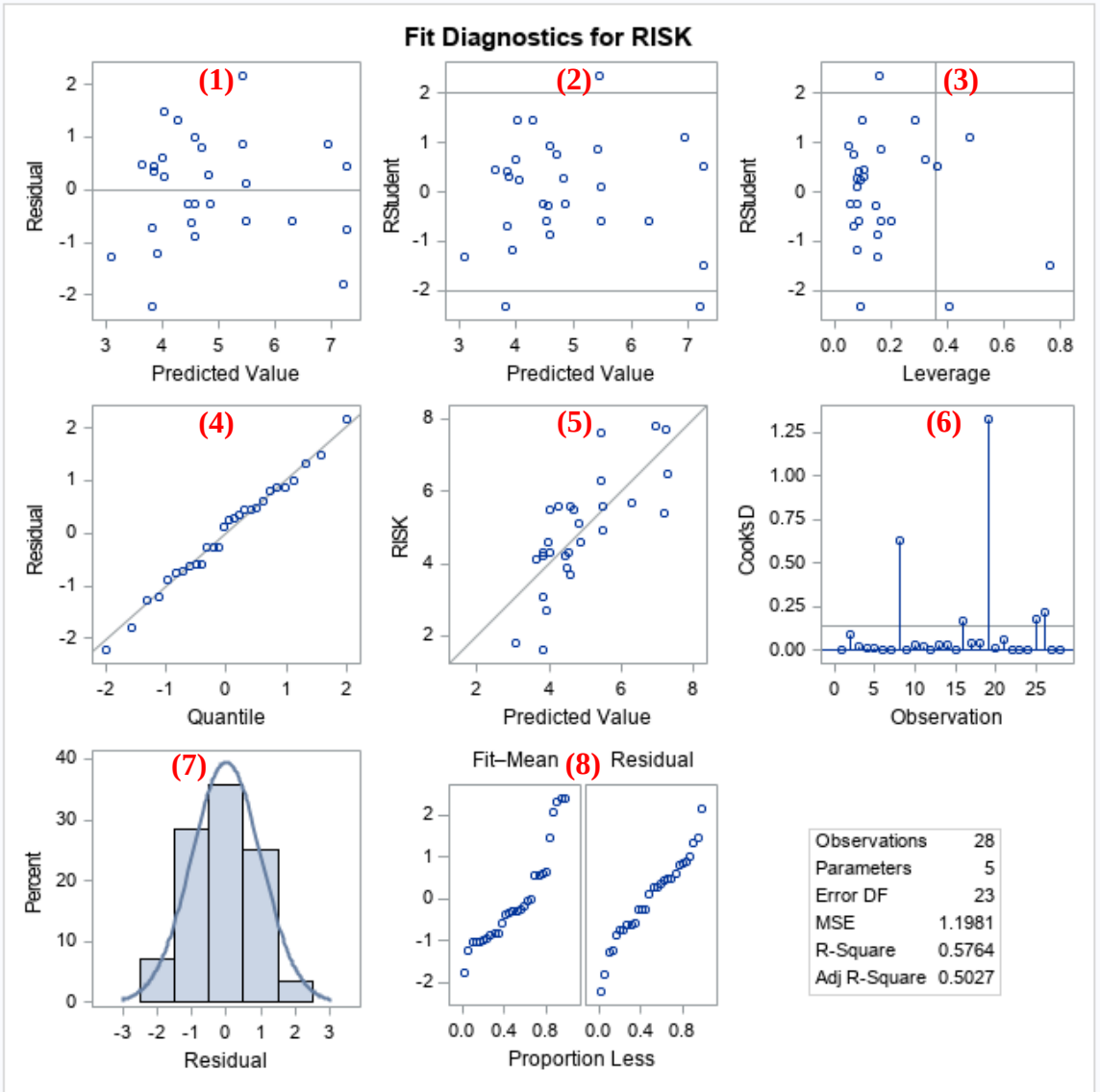
**SAS Output for Model 1:**

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 4 | 37.48854 | 9.37214 | 7.82 | 0.0004 |
| Error | 23 | 27.55574 | 1.19808 | | |
| Corrected Total | 27 | 65.04429 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 1.09457 | R-Square | 0.5764 |
| Dependent Mean | 4.86429 | Adj R-Sq | 0.5027 |
| Coeff Var | 22.50210 | | |

| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | Intercept | 1 | -1.88866 | 3.41449 | -0.55 | 0.5855 |
| STAY | STAY | 1 | 0.22821 | 0.10493 | 2.17 | 0.0402 |
| AGE | AGE | 1 | 0.05918 | 0.06476 | 0.91 | 0.3703 |
| INS | INS | 1 | 0.06653 | 0.01999 | 3.33 | 0.0029 |
| SCHOOL | SCHOOL | 1 | -0.19854 | 0.59622 | -0.33 | 0.7421 |



**SAS Output for Model 2:**
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_3$

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 2 | 36.27961 | 18.13981 | 15.77 | <.0001 |
| Error | 25 | 28.76467 | 1.15059 | | |
| Corrected Total | 27 | 65.04429 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 1.07265 | R-Square | 0.5578 |
| Dependent Mean | 4.86429 | Adj R-Sq | 0.5224 |
| Coeff Var | 22.05163 | | |

| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
| Intercept | Intercept | 1 | 1.15124 | 0.89658 | 1.28 | 0.2109 |
| STAY | STAY | 1 | 0.26598 | 0.09288 | 2.86 | 0.0084 |
| INS | INS | 1 | 0.05416 | 0.01538 | 3.52 | 0.0017 |