Linear Statistical Modeling Methods with SAS

Inferences in Simple Linear Regression Models

Xuemao Zhang
East Stroudsburg University

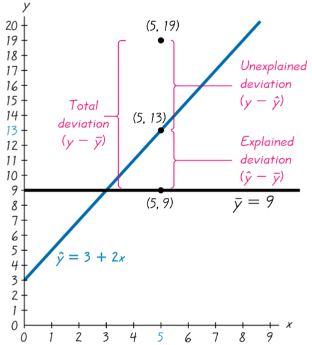
February 12 - 14, 2024

Outline

- ANOVA (Analysis of Variance)
- Statistical Inferences of Slope and Intercept
- Estimation of The Mean Response
- Estimation of an Individual Response

ANOVA (Analysis of Variance)

Explained and Unexplained Deviation



The figure shows (5,13) lies on the regression line, but (5,19) does not.

- Total Deviation (from $\overline{y} = 9$) of the point $(5, 19) = y - \overline{y} = 19 - 9 = 10$.
- Explained Deviation (from $\overline{y} = 9$) of the point $(5, 19) = \widehat{y} - \overline{y} = 13 - 9 = 4$.
- Unexplained Deviation (from $\overline{y} = 9$) of the point $(5, 19) = y - \widehat{y} = 19 - 13 = 6$.

ANOVA (Analysis of Variance)

$$\boxed{\begin{array}{ccc} \textit{total deviation} = & \textit{explained deviation} & + \textit{unexplained deviation} \\ y - \overline{y} = & \widehat{y} - \overline{y} & + y - \widehat{y} \end{array}}$$

It can be shown that

$$\boxed{\begin{array}{ccc} \textit{total variation} = & \textit{explained variation} & + \textit{unexplained variation} \\ \sum_{i=1}^{n}(y_i - \overline{y})^2 = & \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 & + \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \end{array}}$$

ANOVA (Analysis of Variance)

The **Total Sum of Squares (SST)**, is a measure of the variation in the response values ignoring the regression:

$$\text{SST} = \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

Now compare this with the **Error Sum of Squares (SSE)**, a measure of the variation remaining in the response values after predicting them using the fitted regression equation:

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2.$$

Their difference is called the **Regression Sum of Squares (SSR)**: $SST = SSR + SSE$.

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\widehat{Y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \widehat{Y}_i)^2.$$

SSR measures the amount of variation “explained by'' the model fit and also the reduction in uncertainty of predicting the response due to the model.

ANOVA (Analysis of Variance)

- The **coefficient of determination**, r^2, is given by

$$r^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

  The first expression shows it to be the proportion of the variation in the response explained by the regression.
- The second expression shows it to be the proportion by which the regression reduces the uncertainty in predicting the response.

- **Coefficient of Determination.** The coefficient of determination, r^2, is a measure of (take your pick):
  - How much of the variation in the response is "explained" by the regression.
  - How much of the variation in the response is reduced by predicting it using the regression.

Analysis of Variance

The ANOVA Table

- Total df $= n - 1$
- Regression df $= 1$
- Error df $= n - 1 - 1 = n - 2$

Mean Squares:

MSR=SSR/1

MSE = SSE/(n-2)

Table 1: ANOVA Table(more discussions in Chap 8)

Source	df	SS	MS(Mean Squares)	F
Regression	1	SSR	MSR=SSR/1	MSR/MSE
Error	$n-2$	SSE	MSE = SSE/(n-2) $=\widehat{\sigma^2}$	
Total	$n-1$	SST		

Estimation of σ^2

The Mean Square Error. The mean square error or MSE, is an **estimator** of σ^2 in the SLR model, the variance of the error terms ϵ, in the simple linear regression model. Its formula is

$$\text{MSE} = \widehat{\sigma^2} = \frac{1}{\mathbf{n-2}} \sum_{i=1}^{n} e_i^2.$$

It measures the "average squared prediction error" when using the regression.

Analysis of Variance

Example (Analysis of Variance)

Table 2: ANOVA Table

Source	df	SS	MS(Mean Squares)	F	Pr >F
Model	1	678.37	678.37	45.46	$< .0001$
Error	30	447.67	14.92		
C Total	31	1126.05			

```
proc reg data=mtcars;
model mpg = hp;
run;
```

Model Interpretation

Once an acceptable model has been obtained, the next step is to **interpret** the fitted model.

- **The Fitted Slope.** The fitted slope may be interpreted in a couple of ways:
  - **As the estimated change in the mean response per unit increase in the regressor.** This is another way of saying it is the derivative of the predicted response with respect to the regressor:

$$\frac{d\hat{Y}}{dx} = \frac{d}{dX}(\hat{\beta}_0 + \hat{\beta}_1 X) = \hat{\beta}_1.$$

  - **In terms of the estimated change in the mean response per unit increase in the predictor.** In this formulation, if the regressor X, is a differentiable function of the predictor, Z,

$$\frac{d\hat{Y}}{dz} = \frac{d}{dz}(\hat{\beta}_0 + \hat{\beta}_1 X) = \hat{\beta}_1 \frac{dX}{dz},$$

    which means

$$\hat{\beta}_1 = \frac{d\hat{Y}}{dz} \bigg/ \frac{dX}{dz}$$

- **The Fitted Intercept.** The fitted intercept is the estimate of the response when the regressor equals 0, provided this makes sense.

Model Interpretation

Example.

The fitted model for the mtcars data is

$$\hat{Y} = 30.099 - 0.0682X,$$

where Y is mpg and X is hp (horsepower).

- **The Fitted Slope.** The fitted slope, -0.0682, is interpreted as the estimated change in mean mpg for each additional hp.
- **The Fitted Intercept.** The fitted intercept is 30.099. What might be its interpretation?
- **The Mean Square Error.** The MSE, 14.922, estimates the variance of the random errors.

Sampling Distribution of $\widehat{\beta}_1$ and $\widehat{\beta}_0$

To derive the sampling distribution of $\widehat{\beta}_1$ and $\widehat{\beta}_0$, we need the following results.

THEOREM 1. Let $Y_1, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then the sample mean

$$\overline{Y} = \frac{Y_1 + \cdots + Y_n}{n} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

is normally distributed with mean $\mu_{\overline{Y}} = \mu$ and variance $\sigma^2_{\overline{Y}} = \sigma^2/n$.

THEOREM 2. Suppose $X_i \sim N(\mu_i, \sigma = \sigma_i), i = 1, 2, \ldots, n$ are independent. Let a_j and b_j, $j = 1, \ldots, n$ be constants. Define

$$U = \sum_{j=1}^{n} a_j X_j \quad \text{and } V = \sum_{j=1}^{n} b_j X_j$$

Then,

U and V are independent if and only if $cov(U, V) = 0$.

Sampling Distribution of $\widehat{\beta}_1$

Recall that

$$\begin{aligned}\hat{\beta}_1 =& \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}} \\ =& \frac{\sum_{i=1}^n (x_i - \overline{x}) y_i}{S_{xx}} \\ =& \sum_{i=1}^n c_i y_i,\end{aligned}$$

where $c_i = \frac{x_i - \overline{x}}{S_{xx}}$. If we assume that Y_i are n independent $N(\beta_0 + \beta_1 x_i, \sigma^2)$ normal random variables, $i = 1, \ldots, n$, then by **Theorem 1**, it can be shown that

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right),$$

where $S_{xx} = \sum_{i=1}^n (x_i - \overline{x})^2$. And

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE}/\sqrt{S_{xx}}} \sim t_{n-2}.$$

Sampling Distribution of $\widehat{\beta}_1$

Both ρ and β_1 measure the direction and strength of the linear correlation between X and Y. Therefore, testing $H_0 : \rho = 0$ should be equivalent to testing $H_0 : \beta_1 = 0$.

Recall (see lecture 2) that under $H_0 : \rho = 0$, the test statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}}.$$

It can be shown that

$$\frac{\hat{\beta}_1 - 0}{\sqrt{MSE}/\sqrt{S_{xx}}} = r\sqrt{\frac{n-2}{1-r^2}}.$$

Sampling Distribution of $\widehat{\beta}_0$

- Recall that $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$.
- It can be shown that $\overline{Y}$ and $\hat{\beta}_1$ are independent using **Theorem 2**.

Again, if we assume that Y_i are independent $N(0, \sigma^2)$ normal random variables, $i = 1, \ldots, n$, then it can be shown using **Theorem 1** that

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}\right]\right),$$

where $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$. And

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSE} \cdot \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}} \sim t_{n-2}.$$

Remark. The degrees of freedom (df) for the t-distribution are $\mathbf{n-2}$.

Interval Estimation of Slope and Intercept

Level $1-\alpha$ confidence intervals for β_0 and β_1 are

$$(\hat{\beta}_0 - \hat{\sigma}(\hat{\beta}_0)t_{n-2,\frac{\alpha}{2}}, \hat{\beta}_0 + \hat{\sigma}(\hat{\beta}_0)t_{n-2,\frac{\alpha}{2}}),$$

and

$$(\hat{\beta}_1 - \hat{\sigma}(\hat{\beta}_1)t_{n-2,\frac{\alpha}{2}}, \hat{\beta}_1 + \hat{\sigma}(\hat{\beta}_1)t_{n-2,\frac{\alpha}{2}}),$$

respectively, where

$$\hat{\sigma}(\hat{\beta}_0) = \sqrt{\mathrm{MSE}\left[\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}\right]},$$

$$\hat{\sigma}(\hat{\beta}_1) = \sqrt{\mathrm{MSE}/S_{xx}}$$

and

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Interval Estimation of Slope and Intercept

NOTE: Whether the interval for β_1 contains 0 is of particular interest(we are testing $H_0 : \beta_1 = 0$). If it does, it means that we cannot statistically distinguish β_1 from 0. This means we have to consider plausible the model for which $\beta_1 = 0$:

$$Y = \beta_0 + \epsilon$$

This model implies that there is no linear association between Y and X. That is, the linear regression model is useless.

Besides confidence intervals, we can use Hypothesis test to check the usefulness of the model.

Testing the Usefulness of the SLR Model - T test

- **The Statistical Hypotheses:**

$$H_0: \quad \beta_1 \quad = \quad 0$$

versus one of the alternative hypotheses

$$\begin{array}{lcc} H_{a^+}: & \beta_1 & > & 0 \\ H_{a_-}: & \beta_1 & < & 0 \\ H_{a\pm}: & \beta_1 & \neq & 0 \end{array}$$

- **Observed Value of Standardized Test Statistic:**

$$t^* = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}(\hat{\beta}_1)}$$

Testing the Usefulness of the SLR Model - T test

The Test is using t distribution with $df = \mathbf{n-2}$:

- p**-value method:** The p-values for the tests are
  - For the test of H_0 versus H_{a^+}, $p^+ = P(t \geq t^*)$,
  - For the test of H_0 versus H_{a_-}, $p_- = P(t \leq t^*)$,
  - For the test of H_0 versus $H_{a\pm}$, $p\pm = 2P(t \geq |t^*|) = 2\min(p_-, p^+)$,

- **Critical value method:** Suppose the test significance level α is given, then
  - For H_{a^+}: H_0 is rejected only if $t^* \geq t_{n-2,\alpha}$.
  - For H_{a_-}: H_0 is rejected only if $t^* \leq -t_{n-2,\alpha}$.
  - For $H_{a\pm}$: H_0 is rejected only if $t^* \leq -t_{n-2,\alpha/2}$ or $t^* \geq t_{n-2,\alpha/2}$ $(|t^*| > t_{n-2,\alpha/2})$.

Testing the Usefulness of the SLR Model - F test

We can test the overall usefulness of the model using an F test. If the model is useful, MSR will be large compared to the unexplained variation, MSE. The Test is using F distribution with $(\mathbf{df_1 = 1, df_2 = n - 2})$:

Test of $H_0 : \beta_1 = 0$ versus $H_{a\pm} : \beta_1 \neq 0$ at significance level α: The test statistic is

$$F^* = \frac{MSR}{MSE}$$

H_0 is rejected only if $F^* \geq F_\alpha$ with $df_1 = 1, df_2 = n - 2$.

Remark. This test for $H_0 : \beta_1 = 0$ versus $H_{a\pm} : \beta_1 \neq 0$ is exactly equivalent to the t-test, with

$$(t^*)^2 = F^*.$$

Testing the Usefulness of the SLR Model - F test

The analysis of variance test of $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ is an example of the general test for a linear statistical model. To test

$$\begin{aligned} H_0 : &\beta_1 = 0 \\ H_a : &\beta_1 \neq 0, \end{aligned}$$

We consider the full model

$$E(Y) = \beta_0 + \beta_1 x_1, \qquad \text{full model}$$

and the reduced model

$$E(Y) = \beta_0, \qquad \text{reduced model}$$

We fit the full model and denote the obtained error sum of squares by $SSE(F)$; We fit the reduced model and obtain the error sum of squares

$$SSE(R) = \sum (Y_i - \widehat{\beta}_0)^2 = \sum (Y_i - \overline{Y})^2 = SST.$$

The logic now is to compare the two error sums of squares $SSE(F)$ and $SSE(R)$. It can be shown that

$$SSE(F) \leq SSE(R).$$

And a small difference of $SSE(R) - SSE(F)$ suggests that H_0 holds. On the other hand, a large difference suggests that H_a holds because the additional parameters in the model do help to reduce substantially the variation of the observations Y_i around the fitted regression function

Testing the Usefulness of the SLR Model - F test

The actual test statistic is a function of $SSE(R) - SSE(F)$, namely:

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \Big/ \frac{SSE(F)}{df_F}$$

which follows the F distribution when H_0 is true. The degrees of freedom df_R and df_F are those associated with the reduced and full model error sums of squares, respectively. That is $df_R = n - 1$ and $df_F = n - 2$. At significance level α, H_0 is rejected only if

$$F^* \geq F_{\alpha, df_R - df_F, df_F}$$

or

$$F^* \geq F_{\alpha,1,n-2}.$$

Noticed that

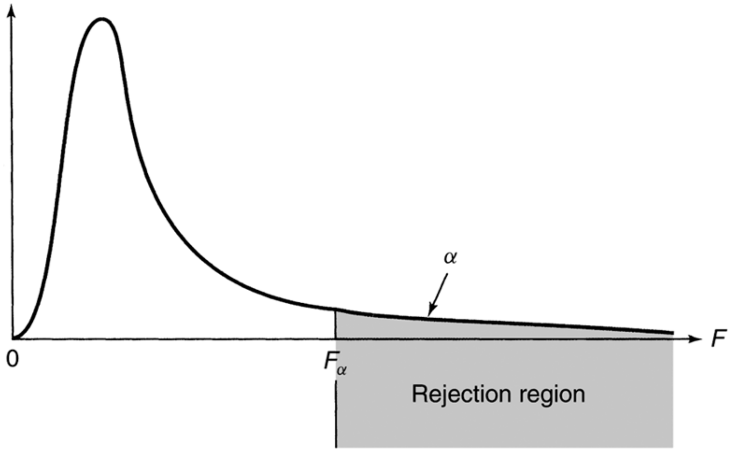
$$SSE(R) = SST, \quad SSE(F) = SSE.$$

So

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \Big/ \frac{SSE(F)}{df_F} = \frac{SSR}{1} \Big/ \frac{SSE}{n-2} = \frac{MSR}{MSE}$$

which is identical to the analysis of variance test statistic.

Testing the Usefulness of the SLR Model - F test



Testing the Usefulness of the SLR Model - F test

Analysis of SAS DATA `mtcars` gives the fitted model

$$\hat{Y} = 30.09886 - 0.06823X,$$

and the ANOVA table:

Table 3: ANOVA Table

Source	df	SS	MS(Mean Squares)	F	Pr >F
Model	1	678.37287	678.372878	45.46	$<$ .0001
Error	30	447.67431	14.92248		
Total	31	1126.04719			

Note: $F^* = 45.46$ vs. $t^* = \frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)} = \frac{-0.06823}{0.01012} \approx -6.74$, where 0.01012 is the standard error of $\hat{\beta}_1 = -0.06823$.

Testing Hypotheses Concerning Intercept - T test

- **The Statistical Hypotheses:**

$$H_0 : \quad \beta_0 \quad = \quad b_0$$

versus one of the alternative hypotheses

$$\begin{aligned} H_{a^+} : &\quad \beta_0 \quad > \quad b_0 \\ H_{a_-} : &\quad \beta_0 \quad < \quad b_0 \\ H_{a\pm} : &\quad \beta_0 \quad \neq \quad b_0 \end{aligned}$$

- **Observed Value of Standardized Test Statistic:**

$$t^* = \frac{\hat{\beta}_0 - b_0}{\hat{\sigma}(\hat{\beta}_0)}.$$

- **The Test:** The p-values for the tests are
  - For the test of H_0 versus H_{a^+}, $p^+ = P(t \geq t^*)$,
  - For the test of H_0 versus H_{a_-}, $p_- = P(t \leq t^*)$,
  - For the test of H_0 versus $H_{a\pm}$, $p\pm = 2P(t \geq |t^*|) = 2\min(p_-, p^+)$,

  where $t \sim t_{\mathbf{n-2}}$.

Note that the test can be conducted using critical value method as well.

Estimation of The Mean Response

Let the mean response at $X = x_0$ denoted by

$$\mu_0 = \beta_0 + \beta_1 x_0.$$

The point estimator of μ_0 is

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x_0 = \overline{y} + \hat{\beta}_1 (x_0 - \overline{x}).$$

It is know that $\overline{Y}$ and $\hat{\beta}_1$ are independent. Therefore, it can be derived that

$$\hat{Y}_0 \sim N\left(\mu_0, \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]\sigma^2\right).$$

And

$$\frac{\hat{Y}_0 - \mu_0}{\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}} \sim t_{n-2}.$$

A level $1 - \alpha$ confidence interval for μ_0 is

$$(\hat{Y}_0 - \hat{\sigma}(\hat{Y}_0) t_{n-2, \frac{\alpha}{2}}, \hat{Y}_0 + \hat{\sigma}(\hat{Y}_0) t_{n-2, \frac{\alpha}{2}}),$$

where

$$\hat{\sigma}(\hat{Y}_0) = \sqrt{\text{MSE}\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]}, S_{xx} = \sum (x_i - \overline{x})^2.$$

Estimation of an Individual Response

We consider now the prediction of an individual new observation Y corresponding to a given future observation $X = x_0$,

$$Y|_{X=x_0} = \beta_0 + \beta_1 x_0 + \epsilon.$$

- We estimate $Y|_{X=x_0}$ by $\hat{Y}|_{X=x_0} = \hat{\mu}|_{X=x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$.
- $Y|_{X=x_0} \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$.
- $\hat{Y}|_{X=x_0} \sim N\left(\beta_0 + \beta_1 x_0, \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]\sigma^2\right)$.
- $Y|_{X=x_0}$ and $\hat{Y}|_{X=x_0}$ are independent since we are predicting a future value $Y|_{X=x_0}$ that is not used in the computation of $\hat{Y}|_{X=x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Therefore, it can be shown that

$$Y|_{X=x_0} - \hat{Y}|_{X=x_0} \sim N\left(0, \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]\sigma^2\right)$$

and

$$\frac{Y|_{X=x_0} - \hat{Y}|_{X=x_0}}{\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}} \sim t_{n-2}.$$

Estimation of an Individual Response

A level $1-\alpha$ prediction interval for a **future** observation at $\mathbf{X} = \mathbf{x_0}$ is

$$(\hat{Y}_{new} - \hat{\sigma}(Y_{new} - \hat{Y}_{new})t_{n-2,\frac{\alpha}{2}}, \hat{Y}_{new} + \hat{\sigma}(Y_{new} - \hat{Y}_{new})t_{n-2,\frac{\alpha}{2}}),$$

where

$$\hat{Y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x_0},$$

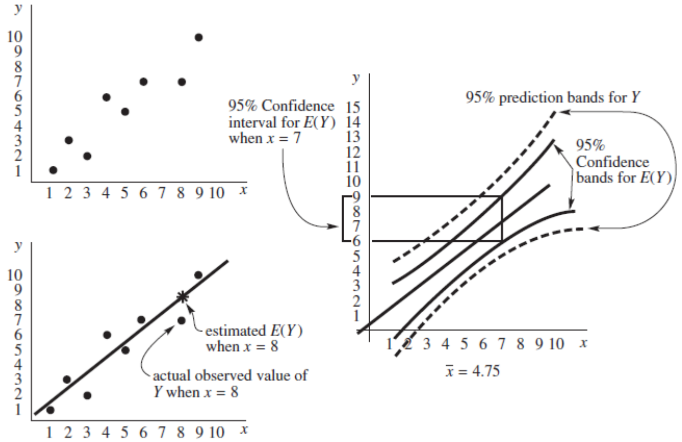
and

$$\hat{\sigma}(Y_{new} - \hat{Y}_{new}) = \sqrt{\text{MSE}\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum(x_i - \overline{x})^2}\right]}.$$

Remark. Prediction intervals for the actual value of Y are longer than confidence intervals for $E(Y)$ if both confidence levels are the same and both are determined for the same value of $x = x_0$.

Estimation of an Individual Response

Some hypothetical data and associated confidence and prediction bands



Example

```
proc reg data=mtcars  ALPHA=0.05; /* significance level = 1-alpha*/
model mpg = hp /clb;
run;
```

- CLB computes confidence limits for the parameter estimates;
- CLM computes confidence limits for the expected value of the dependent variable;
- CLI computes confidence limits for for an individual predicted value

Example

- It is possible to let SAS do the predicting of new observations and/or estimating of mean responses. The way to do this is to enter the values of the independent variables you are interested in during the data input step,but put a period (.) for the unknown y value. That is,

```
data newobs;
input hp mpg@@;
datalines;
160 .
;
run;

data mtcars;
set mtcars newobs;
run;

proc print data=mtcars;
run;
```

Example

```
proc reg data=mtcars;
model mpg = hp / r cli clm; /* r produces analysis of residuals */
run;
```

- We can use proc iml to remove the new observation from the data set.

```
proc iml;
edit mtcars;
delete point 33;
run;
quit;
```

Example

- We can use sas delete observation 33 directly
  - `_n_`: Represents the observation number.
  - `ne`: Stands for “not equal.”

```
data mtcars;
set mtcars newobs;
run;

data mtcars;
    set mtcars;
    if _n_ ne 33; /* Exclude observation 33 */
run;
```

License