

Multiple Regression

Example problem: How can we predict a person's height from his/her parents' heights?

y = *person's height*, x_1 = *Mom's height*, x_2 = *Dad's height*, x_3 = *gender*,
 ε = *error*

Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

Introduction

- Linear model relates the value of an dependent variable y to the value of a single independent variable x .
- There are many situations, when a single independent variable is not enough.

Examples:

- Health status depends on genes, diet, lifestyle, etc.
- Salary depends on education, experience, position, etc.
- Exam grade depends on time used to study, study methods, abilities, etc.
- In situations like this, there are several independent variables, x_1, x_2, \dots, x_p , that are related to a dependent variable y .
- If the relationship between the dependent and independent variables is linear, the technique of multiple regression can be used.

The Multiple Regression Model

- A sample of n items and that on each item we have measured p dependent variables, x_1, x_2, \dots, x_p .
- The i th sampled item info: ordered set $(y_i, x_{1i}, \dots, x_{pi})$.
- We can then fit the **multiple regression model** $y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \varepsilon_i$.

Various Multiple Linear Regression Models

- **Polynomial regression model** (the independent variables are all powers of a single variable)

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 \dots + \beta_p x_i^p + \varepsilon_i.$$

- **Quadratic model** (polynomial regression of model of degree 2, and powers of several variables)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \beta_4 x_{i1}^2 + \beta_5 x_{i2}^2 + \varepsilon_i.$$



- A variable that is the product of two other variables is called an **interaction**.
- These models are considered **linear models**, even though they contain nonlinear terms in the independent variables. The reason is that they are **linear in the coefficients, β_i** .

Estimating the Coefficients

- In any multiple regression model, the estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ are computed by least-squares, just as in simple linear regression.

- The equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$$

is called the **least-squares equation** or **fitted regression equation**.

- Define \hat{y}_i to be the value of the least-squares function corresponding to the x values (x_{1i}, \dots, x_{pi}) .
- The residuals are the quantities $e_i = y_i - \hat{y}_i$ which are the differences between the observed y values and the \hat{y} values given by the equation.
- We compute $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ so as to minimize the sum of the squared residuals.

Sums of Squares

- Much of the analysis in multiple regression is based on three fundamental quantities.

- They are the **total sum of squares** $SST = \sum_{i=1}^n (y_i - \bar{y})^2$

the **error sum of squares** $SSE = \sum_{i=1}^n (y_i - \hat{y})^2$

and the **regression sum of squares** $SSR = \sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y})^2$

- The **analysis of variance identity** is $SST = SSR + SSE$.

Assumptions about the error terms

Assumptions for Errors in Multiple Linear Regression Models:

1. The errors $\varepsilon_1, \dots, \varepsilon_n$ are random and independent. In particular, the magnitude of any error ε_i does not influence the value of the next error ε_{i+1} .
2. The errors $\varepsilon_1, \dots, \varepsilon_n$ all have mean 0.
3. The errors $\varepsilon_1, \dots, \varepsilon_n$ all have the same variance, which we denote by σ^2 .
4. The errors $\varepsilon_1, \dots, \varepsilon_n$ are normally distributed.

There are the same assumptions as those for the Simple Linear Regression.

Mean and Variance of y_i

- The multiple linear regression model is $y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} + \varepsilon_i$.
- Under assumptions 1 through 4, the observations y_1, \dots, y_n are independent random variables that follow the normal distribution.
- The mean and variance of y_i are given by

$$\mu_{y_i} = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi}$$

$$\sigma_{y_i}^2 = \sigma^2$$

- Each coefficient represents the change in the mean of y associated with an increase of one unit in the value of x_i , when the other x variables are held constant.

Statistics

- The three statistics most often used in multiple regression are the estimated error variance s^2 , the coefficient of determination R^2 , and the F statistic.
- The estimated error variance

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p - 1} = \frac{SSE}{n - p - 1}$$

- The goodness of fit statistic in multiple regression denoted by R^2 and it is called the coefficient of determination (same as in SLPR).
- The value of R^2 is calculated in the same way as r^2 in simple linear regression. That is, $R^2 = SSR/SST$.

Distribution of β_i

- When assumptions 1 through 4 are satisfied, the quantity

$$\frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}}$$

has a Student's t distribution with $n - p + 1$ degrees of freedom.

- The number of degrees of freedom is equal to the denominator used to compute the estimated error variance.
- This statistic is used to perform hypothesis tests (and compute confidence intervals), as in SLR.

Tests of Hypothesis

- The main test is multiple linear regression is

$$H_0 = \beta_1 = \beta_2 = \dots = \beta_p = 0 \text{ vs. } H_a: \text{At least one } \beta_i \neq 0.$$

- This is a very strong hypothesis. It says that none of the independent variables has any linear relationship with the dependent variable.
- **If H_0 is true**, then the regression is not significant.
- **If we reject H_0** , then we conclude that there is some significant relationship between the response and at least one of the predictors.

Tests of Hypothesis

- Hypotheses (**significance of regression**):


$$H_0 = \beta_1 = \beta_2 = \dots = \beta_p = 0 \text{ vs. } H_a: \text{At least one } \beta_i \neq 0.$$

- The test statistic for this hypothesis is

$$F = \frac{SSR/p}{SSE/(n-p-1)}$$

- This is an F statistic and its null distribution is $F_{p,n-p-1}$. Note that the denominator of the F statistic is s^2 . The subscripts p and $n-p-1$ are the **degrees of freedom** for the F statistic.
- The value of the F -statistic and the p -value for this test are always included in the output of statistical computing packages in the Analysis of Variance Table (ANOVA results).

More tests- nested F-tests

- In MLR we usually want to know if one model is better than another.
- Nested models:
- **Simple model:** : $ys = \beta_0 + \beta_1x_1 + \beta_2x_2 + ..+ \beta_kx_k + \varepsilon;$
- **Complex Model:** : $yc = \beta_0 + \beta_1x_1 + \beta_2x_2 + ...+ \beta_kx_k + \beta_{k+1}x_{k+1} + ...+ \beta_mx_m + \varepsilon$

simple model

Question: Is the complex model better than the simple model?

Method: nested F-test.

Test: $H_0 = \beta_{k+1} = \beta_{k+2} = \dots = \beta_m = 0$ vs. H_a : At least one of these $m-k$ $\beta_i \neq 0$.

Test statistic: $F = \frac{(SSEs - SSEc)/(m - k)}{SSEc/(n - (m + 1))}$.

Nested F -test

- **Question:** Is the complex model better than the simple model?
- **Method:** nested F-test.
- Test: $H_0 = \beta_{k+1} = \beta_{k+2} = \dots = \beta_m = 0$ vs. H_a : At least one of these $m-k$ $\beta_i \neq 0$.
- Test statistic:
$$F = \frac{(SSEs - SSEc)/(m-k)}{SSEc/(n-(m+1))}.$$

Under H_0 , the test stat has an F distribution with $m-k$ and $n-(m+1)$ d.f.

Decision rule: (1) Reject H_0 if computed $F > F_{\alpha, m-k, n-(m+1)}$ (=percentile cutting probability α to its right), or

(2) Reject H_0 if $p\text{-value} = P(F_{\alpha, m-k, n-(m+1)} > F) < \text{significance level}$.

Common Nested F-tests

1) “Model utility” or “significance of regression” test:

$$ys = \beta_0 + \varepsilon \text{ vs. } yc = \text{your model}$$

This means testing if having regression equation is better than no regression at all. P-value for this model is reported by all stat computing packages.

2) *Add one more variable to an existing model:*

$$ys = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k + \varepsilon;$$

$$yc = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k + \beta_{k+1}x_{k+1} + \varepsilon;$$

Here F-test evaluates if X_{k+1} adds explanatory power/value to the model with k variables.

WARNING! F-test depends on the actual variables used in the models, not only on the number of them.

Confidence Intervals for the model parameters

- Model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$
- $(1-\alpha)100\%$ confidence interval (CI) for β_i is given by:

$\widehat{\beta}_i = t_{\frac{\alpha}{2}, n-(k+1)} s_{\widehat{\beta}_i}$, where $s_{\widehat{\beta}_i}$ is the standard deviation of $\widehat{\beta}_i$ usually computed by a package.

- **Test for $H_0: \beta_i = \beta_{i0}$ vs.**
- **Decision rule:** Level α test rejects H_0 if

$$H_a: \beta_i \neq \beta_{i0} \quad |t| > t_{\frac{\alpha}{2}, n-(k+1)}$$

$$H_a: \beta_i > \beta_{i0} \quad |t| > t_{\alpha, n-(k+1)}$$

$$H_a: \beta_i < \beta_{i0} \quad |t| < -t_{\alpha, n-(k+1)}$$

- **NOTE:** This test works when we assume all other variables are staying in the model.
- These are called “partial t-tests” because they deal with “partial info about β_i given a model with other variables.
- Usually p-values for the partial t-tests are reported along with the value of the test statistic by computing package.

Prediction of the dependent variable

- Given values of explanatory variables in a model and the estimated model we can look for
 - (1) $(1-\alpha)100\%$ confidence interval for the mean of the response value for these explanatory variables, or
 - (2) $(1-\alpha)100\%$ prediction interval for A value of the response for these explanatory variables.

Both CI and PI are easily computed by MINITAB.

R² and Adjusted R²

- **R² always increase when a regressor is added to the model, regardless of the value of the contribution of that variable.**
- **An adjusted R²:**

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-p}}{\frac{SST}{n-1}} = 1 - \frac{MSE}{\frac{SST}{n-1}} = 1 - \frac{n-1}{n-p} \frac{SSE}{SST}$$

- **The adjusted R² will only increase on adding a variable to the model if the addition of a variable reduces the error mean squares MSE=SSE/(n-p).**
- **SST is independent of the model, it is only dependent on the spread of Y.**
- **If we add a predictor, then (n-1)/(n-p) increases, and SSE decreases.**
- **If a predictor has little explanatory power, then SSE decreases a little, and that decrease in SSE can be offset by increase in (n-1)/(n-p), so adjusted R² does not change.**
- **It will only change is the decrease in SSE weighs more than the adjustment (n-1)/(n-p).**

Measure of Goodness of Fit- Multiple Regression: R^2 and adjusted R^2

- The quantity R^2

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

- OR written as

$$R^2 = \frac{SST - SSE}{SST} = \text{proportion of total variation in Y explained by regression.}$$

- In multiple regression R is the Pearson correlation between observed Y s and predicted \hat{Y} s and we call it the coefficient of determination.

Diagnostics- plots

To check independence of the residuals and their normal distribution we usually make and evaluate the following plots:

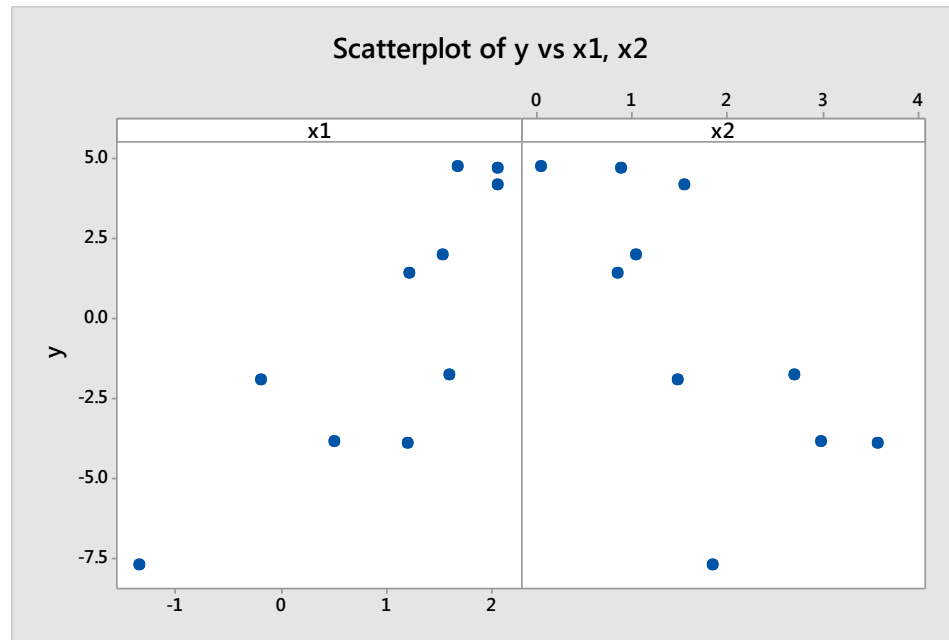
- Residuals vs. fits;
- Residuals vs. explanatory variables,
- Residuals vs. order of observations,
- Normal probability plot of residuals.

Fitting regression model - MINITAB

- **Data set: classexample.MTW**
- For a model with X1 and x2 as explanatory variables find/compute/test
- Model equation,
- Discuss diagnostic plots,
- Are the slopes significantly different from zero on 5% significance level?
- Estimate of variance of the error term σ^2 , that is s^2 .
- What is the Pearson correlation coefficient between observed and predicted responses?
- Predict the value of Y for $x_1=1$, and $x_2=2$,
- Find a 99% confidence interval for the mean value of Y when $x_1=1$, and $x_2=2$,
- Find a 99% prediction interval for the mean value of Y when $x_1=1$, and $x_2=2$,
- Is the model with x_1 and x_2 significantly better than a model with intercept only?
That is, do the explanatory variables add significant information about Y?
- Is a model with X1 and X2 significantly better than a model with X1 only?

Multiple regression in MINITAB: data set classexample.MPJ

- Explanatory variables: x_1 and x_2 ; response variable: y
- Start with scatter plots of the response vs. explanatory variables to see if there is evidence of association:



- It seems that both x_1 and x_2 have some association with y .

Multiple regression in MINITAB

Regression Analysis: y versus x1, x2

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	157.381	78.6904	99.59	0.000
x1	1	76.931	76.9311	97.37	0.000
x2	1	45.747	45.7471	57.90	0.000
Error	7	5.531	0.7901		
Total	9	162.912			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.888893	96.60%	95.63%	92.93%

Sequential sums of squares- info for nested models

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Seq MS	F-Value	P-Value
Regression	2	157.381	96.60%	157.381	78.690	99.59	0.000
x1	1	111.634	68.52%	76.931	111.634	141.29	0.000
x2	1	45.747	28.08%	45.747	45.747	57.90	0.000
Error	7	5.531	3.40%	5.531	0.790		
Total	9	162.912	100.00%				

Tests use the sequential sums of squares

Multiple Regression in MINITAB contd.

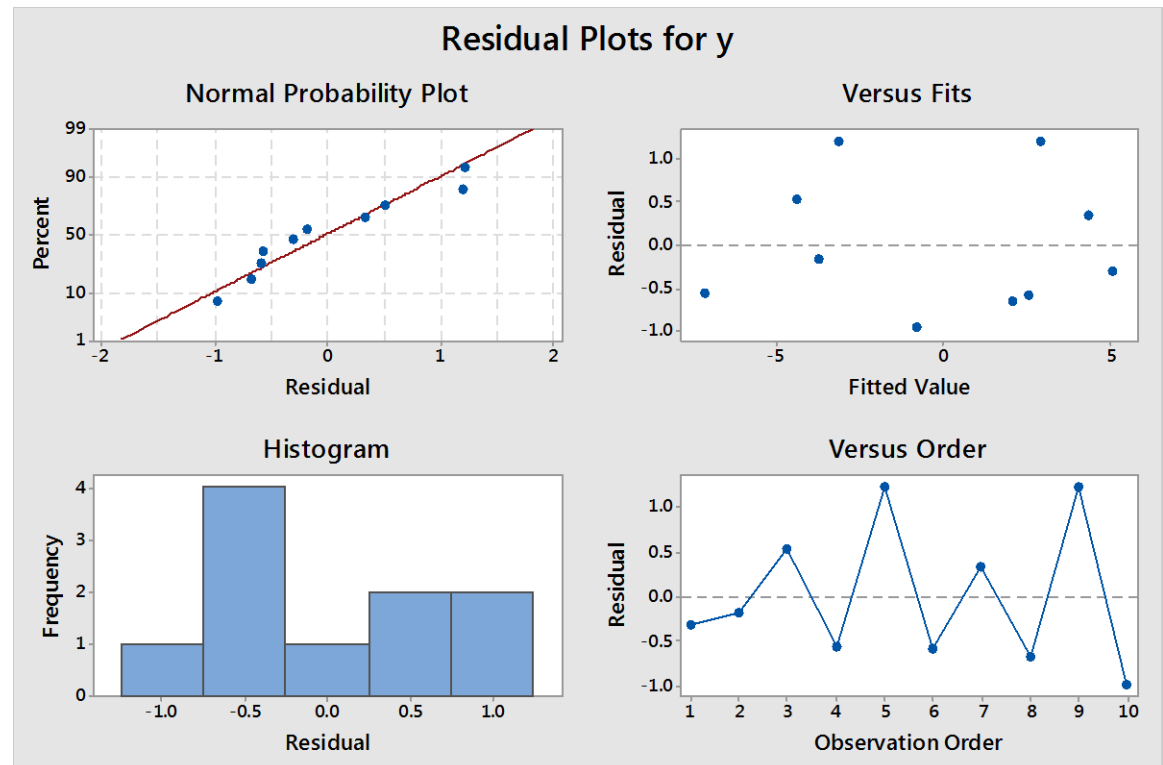
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.530	0.671	0.79	0.456	
x1	2.766	0.280	9.87	0.000	1.05
x2	-2.117	0.278	-7.61	0.000	1.05

Regression Equation

$$y = 0.530 + 2.766 x1 - 2.117 x2$$

Residual Plots for y



Multiple Regression in MINITAB -PREDICTION

Regression Equation

$$y = 0.530 + 2.766 x_1 - 2.117 x_2$$

Variable Setting

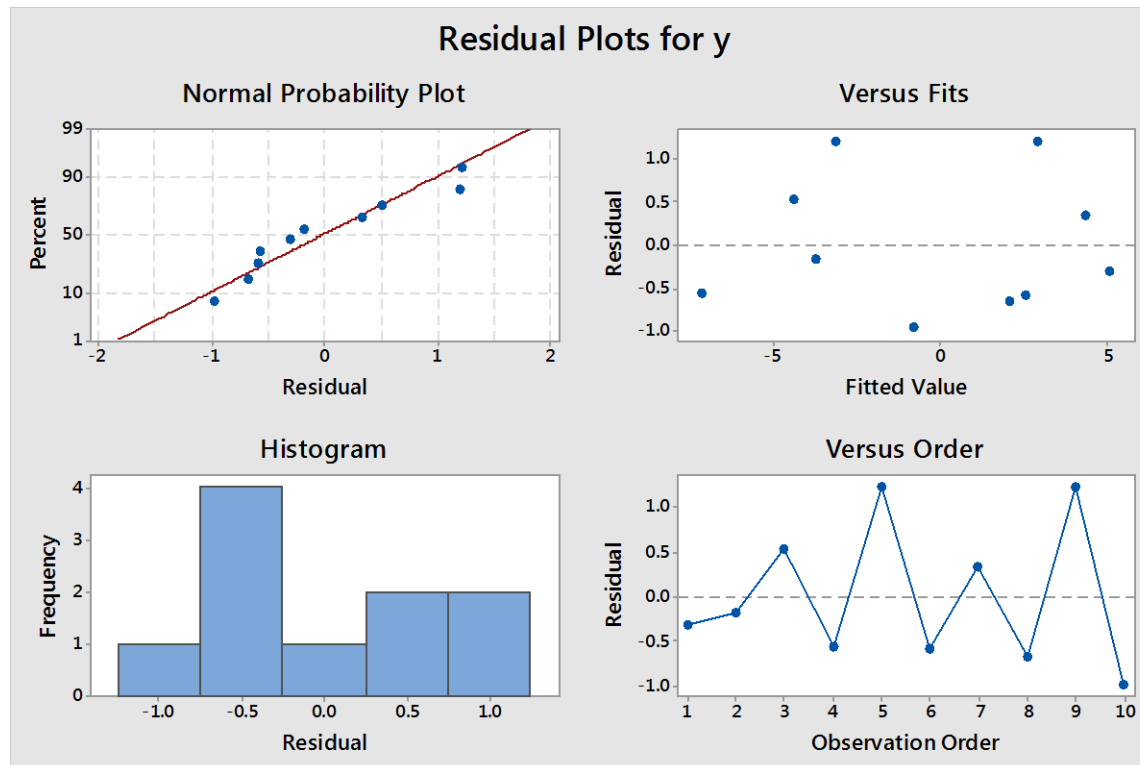
x1 1

x2 2

Fit	SE Fit	95% CI	95% PI
-0.938632	0.292494	(-1.63027, -0.246995)	(-3.15140, 1.27413)

ANSWERS

- Data set: `classexample.MTW`
- For a model with X1 and x2 as explanatory variables find/compute/test
- Model equation: $y = 0.530 + 2.766 x_1 - 2.117 x_2$
- Discuss diagnostic plots: **all good.**



- Are the slopes significantly different from zero on 5% significance level?

Ho: $\beta_1 = 0$ vs. $\beta_1 \neq 0$

Test statistic = 9.87, P-value=0.00, reject Ho, $\beta_1 \neq 0$

Ho: $\beta_2 = 0$ vs. $\beta_2 \neq 0$

Test statistic = -7.61, P-value=0.00, reject Ho, $\beta_2 \neq 0$

- **Coefficients**

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.530	0.671	0.79	0.456	
x1	2.766	0.280	9.87	0.000	1.05
x2	-2.117	0.278	-7.61	0.000	1.05

- Estimate of variance of the error term σ^2 , that is s^2 .

Model Summary

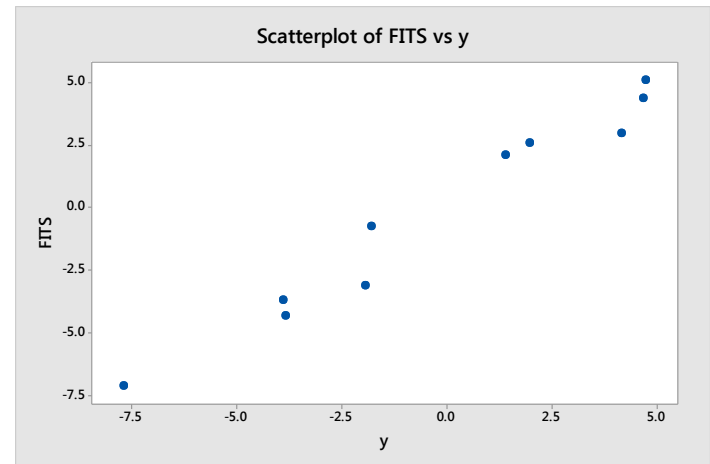
S	R-sq	R-sq(adj)	R-sq(pred)
0.888893	96.60%	95.63%	92.93%

$$s^2 = (0.888893)^2$$

- What is the Pearson correlation coefficient between observed and predicted responses?

$$R = \sqrt{0.966} = 0.983.$$

It seems that we have a pretty good fit of the model to the data



- Predict the value of Y for $x_1=1$, and $x_2=2$,
- Find a 99% **confidence interval** for the mean value of Y when $x_1=1$, and $x_2=2$,
- Find a 99% **prediction interval** for the mean value of Y when $x_1=1$, and $x_2=2$,

Variable Setting

x1 1
x2 2

Fit	SE Fit	95% CI	95% PI
-0.938632	0.292494	(-1.63027, -0.246995)	(-3.15140, 1.27413)

- Is the model with x1 and x2 significantly better than a model with intercept only? That is, do the explanatory variables add significant information about Y?

Here we need to test “significance of regression”.

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Seq MS	F-Value	P-Value
Regression	2	157.381	96.60%	157.381	78.690	99.59	0.000
x1	1	111.634	68.52%	76.931	111.634	141.29	0.000
x2	1	45.747	28.08%	45.747	45.747	57.90	0.000
Error	7	5.531	3.40%	5.531	0.790		
Total	9	162.912	100.00%				

Ho: $y = \beta_0 + \varepsilon$ vs. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

Test statistic is $F=99.95$, $p\text{-value}=0$, reject Ho, meaning that model with x1 and x2 is significantly better than a model with intercept only.

$$F = \frac{(SSEs - SSEc)/(m - k)}{SSEc/(n - (m + 1))} = \frac{45.747/1}{5.531/7} = 57.9.$$

- Is a model with X1 and X2 significantly better than a model with X1 only?

Test Ho: $y = \beta_0 + \beta_1 x_1 + \varepsilon$ vs. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

P-value=0, reject Ho, conclude that model with x1 and x2 is significantly better than model with x1 only.

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Seq MS	F-Value	P-Value
Regression	2	157.381	96.60%	157.381	78.690	99.59	0.000
x1	1	111.634	68.52%	76.931	111.634	141.29	0.000
x2	1	45.747	28.08%	45.747	45.747	57.90	0.000
Error	7	5.531	3.40%	5.531	0.790		
Total	9	162.912	100.00%				