

Multiple Regression

Overview

The multiple regression procedure in the Assistant fits linear and quadratic models with up to five predictors (X) and one continuous response (Y) using least squares estimation. The user selects the model type and the Assistant selects model terms. In this paper, we explain the criteria the Assistant uses to select the regression model.

Additionally, we examine several factors that are important to obtain a valid regression model. First, the sample must be large enough to provide enough power for the test and to provide enough precision for the estimate of the strength of the relationship between X and Y . Next, it is important to identify unusual data that may affect the results of the analysis. We also consider the assumption that the error term follows a normal distribution and evaluate the impact of nonnormality on the hypothesis tests of the overall model.

Based on these factors, the Assistant automatically performs the following checks on your data and reports the findings in the Report Card:

- Amount of data
- Unusual data
- Normality

In this paper, we investigate how these factors relate to regression analysis in practice and we describe how we established the guidelines to check for these factors in the Assistant.

Regression methods

Model selection

Regression analysis in the Assistant fits a model with one continuous response and two to five predictors. One of the predictors may be categorical. There are two types of models to choose from:

- Linear: $F(x) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \dots + \beta_pX_p$
- Quadratic: $F(x) = \beta_0 + \sum_i \beta_iX_i + \sum_i \beta_{ii}X_i^2 + \sum_{i<j} \beta_{ij}X_iX_j$

The Assistant selects the model terms from the full linear or quadratic model.

Objective

We wanted to examine different methods that can be used for model selection to determine which one to use in the Assistant.

Method

We examined three different types of model selection: backward, forward, and stepwise. These model selection types include several options that we also examined, including:

- The criteria used to enter or remove terms from the model.
- Whether to force certain terms into the model or to include certain terms in the initial model.
- The hierarchy of the models.
- Standardizing the X variables in the model.

We reviewed these options, looked at their effect on the outcome of the procedure, and considered which methods were preferred by practitioners.

Results

The procedure we used to select the model terms in the Assistant is as follows:

- Stepwise model selection is used. Often a set of potential X variables are correlated, so that the effect of one term will depend on what other terms are also in the model. Stepwise selection is arguably the best approach under this condition because it allows terms to be entered at one step but to be removed later, depending on what other terms are included in the model.
- Hierarchy of the model is maintained at each step and multiple terms can enter the model in the same step. For example, if the most significant term is X_1^2 , then it is entered, along with X_1 , regardless of whether X_1 is significant. Hierarchy is desirable because it

allows the model to be translated from standardized to unstandardized units. And, because hierarchy allows multiple terms to enter the model at any step, it is possible to identify an important square or interaction term, even if the associated linear term is not strongly related to the response.

- Terms are entered or removed from the model based on $\alpha = 0.10$. Using $\alpha = 0.10$ makes the procedure more selective than the stepwise procedure in core Minitab, which uses $\alpha = 0.15$.
- For purposes of selecting the model terms, predictors are standardized by subtracting the mean and dividing by the standard deviation. The final model is displayed in units of the unstandardized X 's. Standardization of X 's removes most of the correlation between linear and square terms, which reduces the chance of adding higher order terms unnecessarily.

Data checks

Amount of data

Power is concerned with how likely a hypothesis test is to reject the null hypothesis, when it is false. For regression, the null hypothesis states that there is no relationship between X and Y. If the data set is too small, the power of the test may not be adequate to detect a relationship between X and Y that actually exists. Therefore, the data set should be large enough to detect a practically important relationship with high probability.

Objective

We wanted to determine how the amount of data affects the power of the overall F-test of the relationship between X and Y and the precision of R_{adj}^2 , the estimate of the strength of the relationship between X and Y. This information is critical to determine whether the data set is large enough to trust that the strength of the relationship observed in the data is a reliable indicator of the true underlying strength of the relationship. For more information on R_{adj}^2 , see Appendix A.

Method


We took a similar approach to determining the recommended sample size that we used for simple regression. We examined the variability in R_{adj}^2 values to determine how large the sample should be so that R_{adj}^2 is close to ρ_{adj}^2 . We also confirmed that the recommended sample size provided reasonable power even when the strength of the relationship between Y and the X variables is moderately weak. For more information on the calculations, see Appendix B.

Results

As with simple regression, we recommend a sample large enough that you can be 90% confident that the observed value of R_{adj}^2 will be within 0.20 of ρ_{adj}^2 . We found that the required sample size increases as you add more terms to the model. Therefore, we calculated the sample size needed for each model size. The recommended size is rounded up to the nearest multiple of 5. For example, if the model has eight coefficients in addition to the constant, such as four linear terms, three interaction terms, and one square term, then the minimum sample size required to meet the criterion is $n = 49$. The Assistant rounds this up to a recommended sample size of $n = 50$. For more information on specific sample size recommendations based on the number of terms, see Appendix B.

We also verified that the recommended sample sizes provide good enough power. We found that, for moderately weak relationships, $\rho_{adj}^2 = 0.25$, the power is typically about 80% or more. Therefore, following the Assistant's recommendations for sample size ensures that you will have reasonably good power and good precision in estimating the strength of the relationship.

Based on these results, the Assistant displays the following information in the Report Card when checking the amount of data:

| Status | Condition |
|---|--|
|  | <p>Sample size < recommended</p> <p>The sample size is not large enough to provide a very precise estimate of the strength of the relationship. Measures of the strength of the relationship, such as R-Squared and R-Squared (adjusted), can vary a great deal. To obtain a precise estimate, larger samples should be used for a model of this size.</p> |
| | <p>Sample size >= recommended</p> <p>The sample is large enough to obtain a precise estimate of the strength of the relationship.</p> |

Unusual data

In the Assistant Regression procedure, we define unusual data as observations with large standardized residuals or large leverage values. These measures are typically used to identify unusual data in regression analysis (Neter et al., 1996). Because unusual data can have a strong influence on the results, you may need to correct the data to make the analysis valid. However, unusual data can also result from the natural variation in the process. Therefore, it is important to identify the cause of the unusual behavior to determine how to handle such data points.

Objective

We wanted to determine how large the standardized residuals and leverage values need to be to signal that a data point is unusual.

Method

We developed our guidelines for identifying unusual observations based on the standard Regression procedure in Minitab (**Stat > Regression > Regression**).

Results

STANDARDIZED RESIDUAL



The standardized residual equals the value of a residual, e_i , divided by an estimate of its standard deviation. In general, an observation is considered unusual if the absolute value of the standardized residual is greater than 2. However, this guideline is somewhat conservative. You would expect approximately 5% of all observations to meet this criterion by chance (if the errors are normally distributed). Therefore, it is important to investigate the cause of the unusual behavior to determine if an observation truly is unusual.

LEVERAGE VALUE

Leverage values are related only to the X value of an observation and do not depend on the Y value. An observation is determined to be unusual if the leverage value is more than 3 times the number of model coefficients (p) divided by the number of observations (n). Again, this is a commonly used cut-off value, although some textbooks use $\frac{2 \times p}{n}$ (Neter et al., 1996).

If your data include any high leverage points, consider whether they have undue influence over the model selected to fit the data. For example, a single extreme X value could result in the selection of a quadratic model instead of a linear model. You should consider whether the observed curvature in the quadratic model is consistent with your understanding of the process. If it is not, fit a simpler model to the data or gather additional data to more thoroughly investigate the process.

When checking for unusual data, the Assistant Report Card displays the following status indicators:

| Status | Condition |
|--|---|
|  | There are no unusual data points. |
|  | There are at least one or more large standardized residuals or at least one or more high leverage points. |

Normality

A typical assumption in regression is that the random errors (ε) are normally distributed. The normality assumption is important when conducting hypothesis tests of the estimates of the coefficients (β). Fortunately, even when the random errors are not normally distributed, the test results are usually reliable when the sample is large enough.

Objective

We wanted to determine how large the sample needs to be to provide reliable results based on the normal distribution. We wanted to determine how closely the actual test results matched the target level of significance (alpha, or Type I error rate) for the test; that is, whether the test incorrectly rejected the null hypothesis more often or less often than expected for different nonnormal distributions.

Method



To estimate the Type I error rate, we performed multiple simulations with skewed, heavy-tailed, and light-tailed distributions that depart substantially from the normal distribution. We conducted simulations using a sample size of 15. We examined the overall F-test for several models.

For each condition, we performed 10,000 tests. We generated random data so that for each test, the null hypothesis is true. Then, we performed the tests using a target significance level of 0.10. We counted the number of times out of 10,000 that the tests actually rejected the null hypothesis, and compared this proportion to the target significance level. If the test performs well, the Type I error rates should be very close to the target significance level. See Appendix C for more information on the simulations.

Results

For both the overall F-test, the probability of finding statistically significant results does not differ substantially for any of the nonnormal distributions. The Type I error rates are all between 0.08820 and 0.11850, reasonably close to the target significance level of 0.10.

Because the tests perform well with relatively small samples, the Assistant does not test the data for normality. Instead, the Assistant checks the size of the sample and indicates when the sample is less than 15. The Assistant displays the following status indicators in the Report Card for Regression:

| Status | Condition |
|---|--|
|  | The sample size is at least 15, so normality is not an issue. |
|  | Because the sample size is less than 15, normality may be an issue. You should use caution when interpreting the p-value. With small samples, the accuracy of the p-value is sensitive to nonnormal residual errors. |

References

Neter, J., Kutner, M.H., Nachtsheim, C.J., & Wasserman, W. (1996). *Applied linear statistical models*. Chicago: Irwin.

Appendix A: Model and statistics

A regression model relating a predictor X to a response Y is of the form:

$$Y = f(X) + \varepsilon$$

where the function $f(X)$ represents the expected value (mean) of Y given X .

In the Assistant, there are two choices for the form of the function $f(X)$:

| Model type | $f(X)$ |
|------------|--|
| Linear | $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$ |
| Quadratic | $\beta_0 + \sum_i \beta_i X_i + \sum_i \beta_{ii} X_i^2 + \sum_{i < j} \beta_{ij} X_i X_j$ |

The values of the coefficients β are unknown and must be estimated from the data. The method of estimation is least squares, which minimizes the sum of squared residuals in the sample:

$$\min \sum_{i=1}^n (Y_i - \hat{f}(X_i))^2.$$

A residual is the difference between the observed response Y_i and the fitted value $\hat{f}(X_i)$ based on the estimated coefficients. The minimized value of this sum of squares is the SSE (error sum of squares) for a given model.

Overall F-test

This method is a test of the overall model (linear or quadratic). For the selected form of the regression function $f(X)$, it tests:

$$H_0: f(X) \text{ is constant}$$

$$H_1: f(X) \text{ is not constant}$$

Adjusted R^2

Adjusted R^2 (R^2_{adj}) measures how much of the variability in the response is attributed to X by the model. There are two common ways of measuring the strength of the observed relationship between X and Y :

$$R^2 = 1 - \frac{SSE}{SSTO}$$

And

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$$

Where

$$SSTO = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

SSTO is the total sum of squares, which measures the variation of the responses about their overall average \bar{Y} . SSE measures their variation about the regression function $f(X)$. The adjustment in R_{adj}^2 is for the number of coefficients (p) in the full model, which leaves $n - p$ degrees of freedom to estimate the variance of ε . R^2 never decreases when more coefficients are added to the model. However, because of the adjustment, R_{adj}^2 can decrease when additional coefficients do not improve the model. Thus, if adding another term to the model does not explain any additional variance in the response, R_{adj}^2 decreases, indicating that the additional term is not useful. Therefore, the adjusted measure should be used to compare models of different sizes.

Relationship between the F-test and R_{adj}^2

The F- statistic for the test of the overall model can be expressed in terms of SSE and SSTO which are also used in the calculation of R_{adj}^2 :

$$F = \frac{(SSTO - SSE)/(p-1)}{SSE/(n-p)}$$
$$= 1 + \left(\frac{n-1}{p-1}\right) \frac{R_{adj}^2}{1-R_{adj}^2}$$

The formulas above show that the F-statistic is an increasing function of R_{adj}^2 . Thus, the test rejects H_0 if and only if R_{adj}^2 exceeds a specific value determined by the significance level (α) of the test.

Appendix B: Amount of data

In this section we consider how n , the number of observations, affects the power of the overall model test and the precision of R_{adj}^2 , the estimate of the strength of the model.

To quantify the strength of the relationship, we introduce a new quantity, ρ_{adj}^2 , as the population counterpart of the sample statistic R_{adj}^2 . Recall that

$$R_{adj}^2 = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$$

Therefore, we define

$$\rho_{adj}^2 = 1 - \frac{E(SSE|X)/(n-p)}{E(SSTO|X)/(n-1)}$$

The operator $E(\cdot|X)$ denotes the expected value, or the mean of a random variable given the value of X . Assuming the correct model is $Y = f(X) + \varepsilon$ with independent identically distributed ε , we have

$$\frac{E(SSE|X)}{n-p} = \sigma^2 = \text{Var}(\varepsilon)$$

$$\frac{E(SSTO|X)}{n-1} = \sum_{i=1}^n \frac{(f(X_i) - \bar{f})^2}{(n-1)} + \sigma^2$$

where $\bar{f} = \frac{1}{n} \sum_{i=1}^n f(X_i)$.

Hence,

$$\rho_{adj}^2 = \frac{\sum_{i=1}^n (f(X_i) - \bar{f})^2 / (n-1)}{\sum_{i=1}^n (f(X_i) - \bar{f})^2 / (n-1) + \sigma^2}$$

Overall model significance

When testing the statistical significance of the overall model, we assume that the random errors ε are independent and normally distributed. Then, under the null hypothesis that the mean of Y is constant ($f(X) = \beta_0$), the F-test statistic has an $F(p-1, n-p)$ distribution. Under the alternative hypothesis, the F-statistic has a noncentral $F(p-1, n-p, \theta)$ distribution with noncentrality parameter:

$$\begin{aligned} \theta &= \sum_{i=1}^n (f(X_i) - \bar{f})^2 / \sigma^2 \\ &= \frac{(n-1)\rho_{adj}^2}{1 - \rho_{adj}^2} \end{aligned}$$

The probability of rejecting H_0 increases with the noncentrality parameter, which is increasing in both n and ρ_{adj}^2 .

Strength of the relationship

As we showed for simple regression, a statistically significant relationship in the data does not necessarily indicate a strong underlying relationship between X and Y . This is why many users look to indicators such as R_{adj}^2 to tell them how strong the relationship actually is. If we consider R_{adj}^2 as an estimate of ρ_{adj}^2 , then we want to have confidence that the estimate is reasonably close to the true ρ_{adj}^2 value.

For each possible model size, we determined an appropriate threshold for acceptable sample size by identifying the minimum value of n for which absolute differences $|R_{adj}^2 - \rho_{adj}^2|$ greater than 0.20 occur with no more than 10% probability. This is regardless of the true value of ρ_{adj}^2 . The recommended sample sizes $n(T)$ are summarized in the table below where T is the number of coefficients in the model other than the constant coefficient.

| T | n(T) |
|-------|------|
| 1-3 | 40 |
| 4-6 | 45 |
| 7-8 | 50 |
| 9-11 | 55 |
| 12-14 | 60 |
| 15-18 | 65 |
| 19-21 | 70 |
| 22-24 | 75 |
| 25-27 | 80 |
| 28-31 | 85 |
| 32-34 | 90 |
| 35-38 | 95 |
| 39-41 | 100 |
| 42-45 | 105 |
| 46-48 | 110 |

| T | n(T) |
|-------|------|
| 49-52 | 115 |
| 53-56 | 120 |
| 57-59 | 125 |
| 60-63 | 130 |
| 64-67 | 135 |
| 68-70 | 140 |
| 71-73 | 145 |

We evaluated the power of the overall F test of the model for a moderately weak value of $\rho_{adj}^2 = 0.25$, to confirm that there is sufficient power at the recommended sample sizes. The model sizes in the table below represent the worst-case for each value of n(T). Smaller models with the same n(T) will have greater power.

| T | n(T) | Power at $\rho_{adj}^2 = 0.25$ |
|----|------|-----------------------------------|
| 3 | 40 | 0.902791 |
| 6 | 45 | 0.854611 |
| 8 | 50 | 0.850675 |
| 11 | 55 | 0.831818 |
| 14 | 60 | 0.820592 |
| 18 | 65 | 0.798003 |
| 21 | 70 | 0.796425 |
| 24 | 75 | 0.796911 |
| 27 | 80 | 0.798856 |
| 31 | 85 | 0.789861 |
| 34 | 90 | 0.794367 |
| 38 | 95 | 0.788625 |
| 41 | 100 | 0.794511 |
| 45 | 105 | 0.790864 |

| T | n(T) | Power at $\rho_{adj}^2 = 0.25$ |
|----|------|-----------------------------------|
| 48 | 110 | 0.797487 |
| 52 | 115 | 0.795250 |
| 56 | 120 | 0.793698 |
| 59 | 125 | 0.800982 |
| 63 | 130 | 0.800230 |
| 67 | 135 | 0.799906 |
| 69 | 140 | 0.814664 |

Appendix C: Normality

The regression models used in the Assistant are all of the form:

$$Y = f(X) + \varepsilon$$

The typical assumption about the random terms ε is that they are independent and identically distributed normal random variables with mean zero and common variance σ^2 . The least squares estimates of the β parameters are still the best linear unbiased estimates, even if we forgo the assumption that the ε are normally distributed. The normality assumption only becomes important when we try to attach probabilities to these estimates, as we do in the hypothesis tests about $f(X)$.

We wanted to determine how large n needs to be so that we can trust the results of a regression analysis based on the normality assumption. We performed simulations to explore the Type I error rates of the hypothesis tests under a variety of nonnormal error distributions.

Table 1 below shows the proportion of 10,000 simulations in which the overall F-test was significant at $\alpha = 0.10$ for various distributions of ε for three different models. In these simulations, the null hypothesis, which states that there is no relationship between X and Y , was true. The X values were generated as multivariate normal variables by Minitab's RANDOM command. We used a sample size of $n=15$ for all tests. All the models involved five continuous predictors. The first model was the linear model with all five X variables. The second model had all linear and square terms. The third model had all linear terms and seven of the 2-way interactions.

Table 1 Type I error rates for overall F-tests with $n=15$ for nonnormal distributions

| Distribution | Linear | Linear + square | Linear + 7 interactions |
|--------------|---------|-----------------|-------------------------|
| Normal | 0.09910 | 0.10270 | 0.10060 |
| t(3) | 0.09840 | 0.11850 | 0.11800 |
| t(5) | 0.09980 | 0.10010 | 0.10430 |
| Laplace | 0.09260 | 0.09400 | 0.09650 |
| Uniform | 0.10630 | 0.10080 | 0.09480 |
| Beta(3, 3) | 0.09980 | 0.10120 | 0.10020 |
| Exponential | 0.08820 | 0.09500 | 0.09960 |
| Chi(3) | 0.09890 | 0.11400 | 0.10970 |
| Chi(5) | 0.09730 | 0.10590 | 0.10330 |

| Distribution | Linear | Linear + square | Linear + 7 interactions |
|--------------|---------|-----------------|-------------------------|
| Chi(10) | 0.10150 | 0.09930 | 0.10360 |
| Beta(8, 1) | 0.09870 | 0.10230 | 0.10490 |

The simulation results show, that the probability of finding statistically significant results does not differ substantially from the nominal value of 0.10 for any of the error distributions. The Type I error rates observed are all between 0.08820 and 0.11850.

© 2015, 2017 Minitab Inc. All rights reserved.

Minitab®, Quality. Analysis. Results.® and the Minitab® logo are all registered trademarks of Minitab, Inc., in the United States and other countries. See minitab.com/legal/trademarks for more information.