This paper explains the research conducted by Minitab statisticians to develop the methods and data checks used in the Assistant in Minitab Statistical Software.

Simple Regression

Overview

The simple regression procedure in the Assistant fits linear and quadratic models with one continuous predictor (X) and one continuous response (Y) using least squares estimation. The user can select the model type or allow the Assistant to select the best fitting model. 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 and the coefficients. Finally, to ensure that the model is useful, it is important that the type of model selected accurately reflects the relationship between X and Y.

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
- Model fit

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 predictor and one continuous response and can fit two types of models:

- Linear: $F(x) = \beta_0 + \beta_1 X$
- Quadratic: $F(x) = \beta_0 + \beta_1 X + \beta_2 X^2$

The user can select the model before performing the analysis or can allow the Assistant to select the model. There are several methods that can be used to determine which model is most appropriate for the data. To ensure that the model is useful, it is important that the type of model selected accurately reflects the relationship between X and Y.

Objective

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

Method

We examined three methods that are typically used for model selection (Neter et al., 1996). The first method identifies the model in which the highest order term is significant. The second method selects the model with the highest R^2_{adj} value. The third method selects the model in which the overall F-test is significant. For more details, see Appendix A.

To determine the approach in the Assistant, we examined the methods and compared their calculations to one another. We also gathered feedback from experts in quality analysis.

Results

Based on our research, we decided to use the method that selects the model based on the statistical significance of the highest order term in the model. The Assistant first examines the quadratic model and tests whether the square term (β_2) in the model is statistically significant. If that term is not significant, then it drops the quadratic term from the model and tests the linear term (β_1). The model selected through this approach is presented in the Model Selection Report. Additionally, if the user selected a model that is different than the one selected by the Assistant, we report that in the Model Selection Report and the Report Card.

We chose this method in part because of feedback from quality professionals who said they generally prefer simpler models, which exclude terms that are not significant. Additionally, based on our comparison of the methods, using the statistical significance of the highest term in the model is more stringent than the method that selects the model based on the highest R^2_{adj} value. For more details, see Appendix A.

Although we use the statistical significance of highest model term to select the model, we also present the R^2_{adj} value and the overall F-test for the model in the Model Selection Report. To see the status indicators presented in the Report Card, see the Model fit data check section below.

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^2_{adj}, 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^2_{adj}, see Appendix A.

Method

To examine the power of the overall F-test, we performed power calculations for a range of R^2_{adj} values and sample sizes. To examine the precision of R^2_{adj}, we simulated the distribution of R^2_{adj} for different values of the population adjusted R^2 (ρ^2_{adj}) and different sample sizes. We examined the variability in R^2_{adj} values to determine how large the sample should be so that R^2_{adj} is close to ρ^2_{adj}. For more information on the calculations and simulations, see Appendix B.

Results

We found that for moderately large samples, regression has good power to detect relationships between X and Y, even if the relationships are not strong enough to be of practical interest. More specifically, we found that:

- With a sample size of 15 and a strong relationship between X and Y (ρ^2_{adj} = 0.65), the probability of finding a statistically significant linear relationship is 0.9969. Therefore, when the test fails to find a statistically significant relationship with 15 or more data points, it is likely that the true relationship is not very strong (ρ^2_{adj} value < 0.65).
- With a sample size of 40 and a moderately weak relationship between X and Y (ρ^2_{adj} = 0.25), the probability of finding a statistically significant linear relationship is 0.9398.Therefore, with 40 data points, the F-test is likely to find relationships between X and Y even when the relationship is moderately weak.

Regression can detect relationships between X and Y fairly easily. Therefore, if you find a statistically significant relationship, you should also evaluate the strength of the relationship using R^2_{adj}. We found that if the sample size is not large enough, R^2_{adj} is not very reliable and can vary widely from sample to sample. However, with a sample size of 40 or more, we found that R^2_{adj} values are more stable and reliable. With a sample size of 40, you can be 90% confident that observed value of R^2_{adj}will be within 0.20 of ρ^2_{adj} regardless of the actual value and the model type (linear or quadratic). For more detail on the results of the simulations, see Appendix B.

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

Status	Condition
	**Sample size < 40** Your 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 more precise estimate, larger samples (typically 40 or more) should be used. **Sample size > =40** Your sample is large enough to obtain a precise estimate of the strength of the relationship.

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

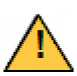
The standardized residual equals the value of a residual, e_i, divided by an estimate of its standard deviation. In general, an observation is considered unusual if the absolute value of the standardized residual is greater than 2. However, this guideline is somewhat conservative. You would expect approximately 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 type of 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. Unusual data points can have a strong influence on the results.
	There are at least one or more large standardized residuals or at least one or more high leverage values. You can hover over a point or use Minitab's brushing feature to identify the worksheet rows. Because unusual data can have a strong influence on the results, try to identify the cause for their unusual nature. Correct any data entry or measurement errors. Consider removing data that are associated with special causes and redoing the analysis.

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

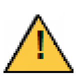
To estimate the Type I error rate, we performed multiple simulations with skewed, heavy-tailed, and light-tailed distributions that depart substantially from the normal distribution. We conducted simulations for the linear and quadratic models using a sample size of 15. We examined both the overall F-test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

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 0.05. 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 and for the test of the highest order term in the model, 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 0.038 and 0.0529, very close to the target significance level of 0.05.

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.

Model fit

You can select the linear or quadratic model before performing the regression analysis or you can choose for the Assistant to select the model. Several methods can be used to select an appropriate model.

Objective

We wanted to examine the different methods used to select a model type to determine which approach to use in the Assistant.

Method

We examined three methods that are typically used for model selection. The first method identifies the model in which the highest order term is significant. The second method selects the model with the highest R^2_{adj} value. The third method selects the model in which the overall F-test is significant. For more details, see Appendix A.

To determine the approach used in the Assistant, we examined the methods and how their calculations compared to one another. We also gathered feedback from experts in quality analysis.

Results

We decided to use the method that selects the model based on the statistical significance of the highest order term in the model. The Assistant first examines the quadratic model and tests whether the square term in the model (β_3) is statistically significant. If that term is not significant, then it tests the linear term (β_1) in the linear model. The model selected through this approach is presented in the Model Selection Report. Additionally, if the user selected a model that is different than the one selected by the Assistant, we report that in the Model Selection Report and the Report Card. For more information, see the Regression method section above.

Based on our findings, the Assistant Report Card displays the following status indicator:

Status	Condition
	**If the user's model matches the Assistant's best fitting model** You should evaluate the data and model fit in terms of your goals. Look at the fitted line plots to be sure that: • The sample adequately covers the range of X values. • The model properly fits any curvature in the data (avoid over-fitting). • The line fits well in any areas of special interest. **If the user's model does not match the Assistant's best fitting model** The Model Selection Report displays an alternative model that may be a better choice.

References

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

Appendix A: Model selection

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	$\beta_0 + \beta_1 X$
Quadratic	$\beta_0 + \beta_1 X + \beta_2 X^2$

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:

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

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.

To determine the method used in the Assistant to select the model type, we evaluated three options:

- Significance of the highest order term in the model
- The overall F-test of the model
- Adjusted R^2 value (R^2_{adj})

Significance of the highest order term in the model

In this approach, the Assistant starts with the quadratic model. The Assistant tests the hypotheses for the square term in the quadratic model:

$H_0: \beta_2 = 0$

$H_1: \beta_2 \neq 0$

If this null hypothesis is rejected, then the Assistant concludes that the square term coefficient is non-zero and selects the quadratic model. If not, the Assistant tests the hypotheses for the linear model:

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

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:

$H_0: f(X) is\ constant$

$H_1: f(X) is\ not\ constant$

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^2_{adj} 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^2_{adj} 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^2_{adj} decreases, indicating that the additional term is not useful. Therefore, the adjusted measure should be used to compare the linear and quadratic.

Relationship between model selection methods

We wanted to examine the relationship between the three model selection methods, how they are calculated, and how they affect one another.

First, we looked at the relationship between how the overall F-test and R^2_{adj} are calculated. 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^2_{adj}:

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

$$= 1 + \left(\frac{n-1}{p-1}\right)\frac{R^2_{adj}}{1-R^2_{adj}}.$$

The formulas above show that the F-statistic is an increasing function of R^2_{adj}. Thus, the test rejects H_0 if and only if R^2_{adj} exceeds a specific value determined by the significance level (α) of the test. To illustrate this, we calculated the minimum R^2_{adj} needed to obtain statistical significance of the quadratic model at $\alpha = 0.05$ for different sample sizes shown in Table 1 below. For example, with n = 15, the R^2_{adj} value for the model must be at least 0.291877 for the overall F-test to be statistically significant.

Table 1 Minimum R^2_{adj} for a significant overall F-test for the quadratic model at $\alpha = 0.05$ at various sample sizes

Sample size	Minimum $\boldsymbol{R^2_{adj}}$
4	0.992500
5	0.900000
6	0.773799
7	0.664590
8	0.577608
9	0.508796
10	0.453712
11	0.408911
12	0.371895
13	0.340864
14	0.314512
15	0.291877
16	0.272238
17	0.255044

Sample size	Minimum R^2_{adj}
18	0.239872
19	0.226387
20	0.214326
21	0.203476
22	0.193666
23	0.184752
24	0.176619
25	0.169168
26	0.162318
27	0.155999
28	0.150152
29	0.144726
30	0.139677
31	0.134967
32	0.130564
33	0.126439
34	0.122565
35	0.118922
36	0.115488
37	0.112246
38	0.109182
39	0.106280
40	0.103528
41	0.100914
42	0.098429
43	0.096064

Sample size	Minimum R^2_{adj}
44	0.093809
45	0.091658
46	0.089603
47	0.087637
48	0.085757
49	0.083955
50	0.082227

Next, we examined the relationship between the hypothesis test of the highest order term in a model, and R^2_{adj}.The test for the highest order term, such as the square term in a quadratic model, can be expressed in terms of the sums of squares or of the R^2_{adj} of the full model (e.g. quadratic) and of the R^2_{adj} of the reduced model (e.g. linear):

$$F = \frac{SSE(Reduced) - SSE(Full)}{SSE(Full)/(n-p)}$$

$$= 1 + \frac{(n-p+1)\left(R^2_{adj}(Full) - R^2_{adj}(Reduced)\right)}{1 - R^2_{adj}(Full)}.$$

The formulas show that for a fixed value of $R^2_{adj}(Reduced)$, the F-statistic is an increasing function of $R^2_{adj}(Full)$. They also show how the test statistic depends on the difference between the two R^2_{adj}values. In particular, the value for the full model must be greater than the value for the reduced model to obtain an F-value large enough to be statistically significant. Thus, the method that uses the significance of the highest order term to select the best model is more stringent than the method that chooses the model with the highest R^2_{adj}. The highest order term method is also compatible with the preference of many users for a simpler model. Thus, we decided to use the statistical significance of the highest order term to select the model in the Assistant.

Some users are more inclined to choose the model that best fits the data; that is, the model with highest R^2_{adj}. The Assistant provides these values in the Model Selection Report and the Report Card.

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(·|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

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

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

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

Hence,

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

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:

$$\theta = \sum_{i=1}^{n} \left(f(X_i) - \bar{f}\right)^2/\sigma^2$$

$$= \frac{(n-1)\rho_{adj}^2}{1 - \rho_{adj}^2}$$

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

Using the formula above, we calculated the power of the overall F-tests for a range of ρ_{adj}^2 values when n = 15 for the linear and quadratic models. See Table 2 for the results.

Table 2 Power for linear and quadratic models with different ρ_{adj}^2 values with n=15

ρ_{adj}^2	θ	Power of F Linear	Power of F Quadratic
0.05	**0.737**	0.12523	0.09615
0.10	**1.556**	0.21175	0.15239
0.15	**2.471**	0.30766	0.21896
0.20	**3.500**	0.41024	0.29560
0.25	**4.667**	0.51590	0.38139
0.30	**6.000**	0.62033	0.47448
0.35	**7.538**	0.71868	0.57196
0.40	**9.333**	0.80606	0.66973
0.45	**11.455**	0.87819	0.76259
0.50	**14.000**	0.93237	0.84476
0.55	**17.111**	0.96823	0.91084
0.60	**21.000**	0.98820	0.95737
0.65	**26.000**	0.99688	0.98443
0.70	**32.667**	0.99951	0.99625
0.75	**42.000**	0.99997	0.99954
0.80	**56.000**	1.00000	0.99998
0.85	**79.333**	1.00000	1.00000
0.90	**126.000**	1.00000	1.00000
0.95	**266.000**	1.00000	1.00000

Overall, we found that the test has high power when the relationship between X and Y is strong and the sample size is at least 15. For example when ρ_{adj}^2 = 0.65, Table 2 shows that the probability of finding a statistically significant linear relationship at $\alpha = 0.05$ is 0.99688. The failure to detect such a strong relationship with the F-test would occur in less than 0.5% of samples. Even for a quadratic model, the failure to detect the relationship with the F-test would occur in less than 2% of samples. Thus, when the test fails to find a statistically significant relationship with 15 or more observations, it is a good indication that the true relationship, if there is one at all, has a ρ^2_{adj} value lower than 0.65. Note that ρ^2_{adj} does not have to be as large as 0.65 to be of practical interest.

We also wanted to examine the power of the overall F-test when the sample size was larger (n=40). We determined that the sample size n = 40 is an important threshold for the precision of the R^2_{adj} (see Strength of the relationship below) and we wanted to evaluate power values for the sample size. We calculated the power of the overall F-tests for a range of ρ^2_{adj} values when n = 40 for the linear and quadratic models. See Table 3 for the results.

Table 3 Power for linear and quadratic models with different ρ^2_{adj} values with n = 40

ρ^2_{adj}	θ	Power of F Linear	Power of F Quadratic
0.05	**2.0526**	0.28698	0.21541
0.10	**4.3333**	0.52752	0.41502
0.15	**6.8824**	0.72464	0.60957
0.20	**9.7500**	0.86053	0.76981
0.25	**13.0000**	0.93980	0.88237
0.30	**16.7143**	0.97846	0.94925
0.35	**21.0000**	0.99386	0.98217
0.40	**26.0000**	0.99868	0.99515
0.45	**31.9091**	0.99980	0.99905
0.50	**39.0000**	0.99998	0.99988
0.55	**47.6667**	1.00000	0.99999
0.60	**58.5000**	1.00000	1.00000
0.65	**72.4286**	1.00000	1.00000

We found that the power was high, even when the relationship between X and Y was moderately weak. For example, even when ρ^2_{adj} = 0.25, Table 3 shows that the probability of finding a statistically significant linear relationship at $\alpha = 0.05$ is 0.93980. With 40 observations, the F-test is unlikely to fail to detect a relationship between X and Y, even if that relationship is moderately weak.

Strength of the relationship

As we have already shown, 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^2_{adj} to tell them how strong the relationship actually is. If we consider R^2_{adj} as an estimate of ρ^2_{adj}, then we want to have confidence that the estimate is reasonably close to the true ρ^2_{adj} value.

To illustrate the relationship between R^2_{adj} and ρ^2_{adj}, we simulated the distribution of R^2_{adj} for different values of ρ^2_{adj} to see how variable R^2_{adj} is for different values of n. The graphs in Figures 1-4 below show histograms of 10,000 simulated values of R^2_{adj}. In each pair of histograms, the value of ρ^2_{adj} is the same so that we can compare the variability of R^2_{adj} for samples of size 15 to samples of size 40. We tested ρ^2_{adj} values of 0.0, 0.30, 0.60, and 0.90. All simulations were performed with the linear model.

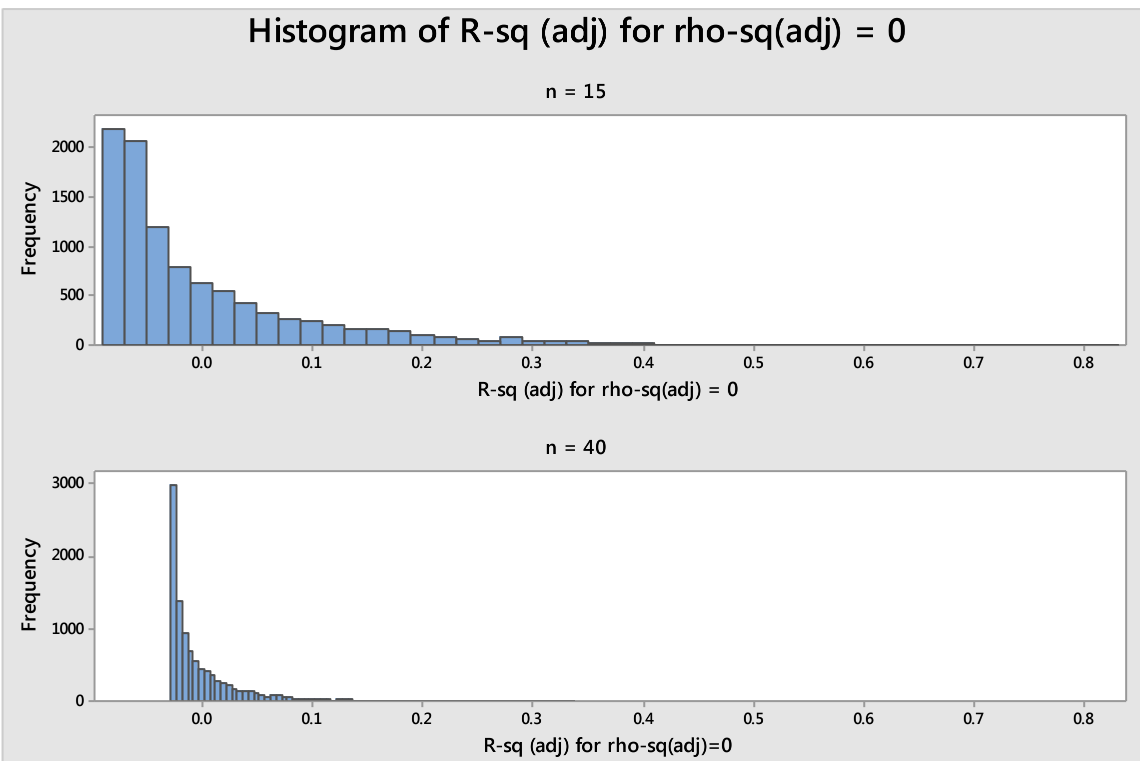


Figure 1 Simulated R^2_{adj} values for ρ^2_{adj} = 0.0 for n=15 and n=40

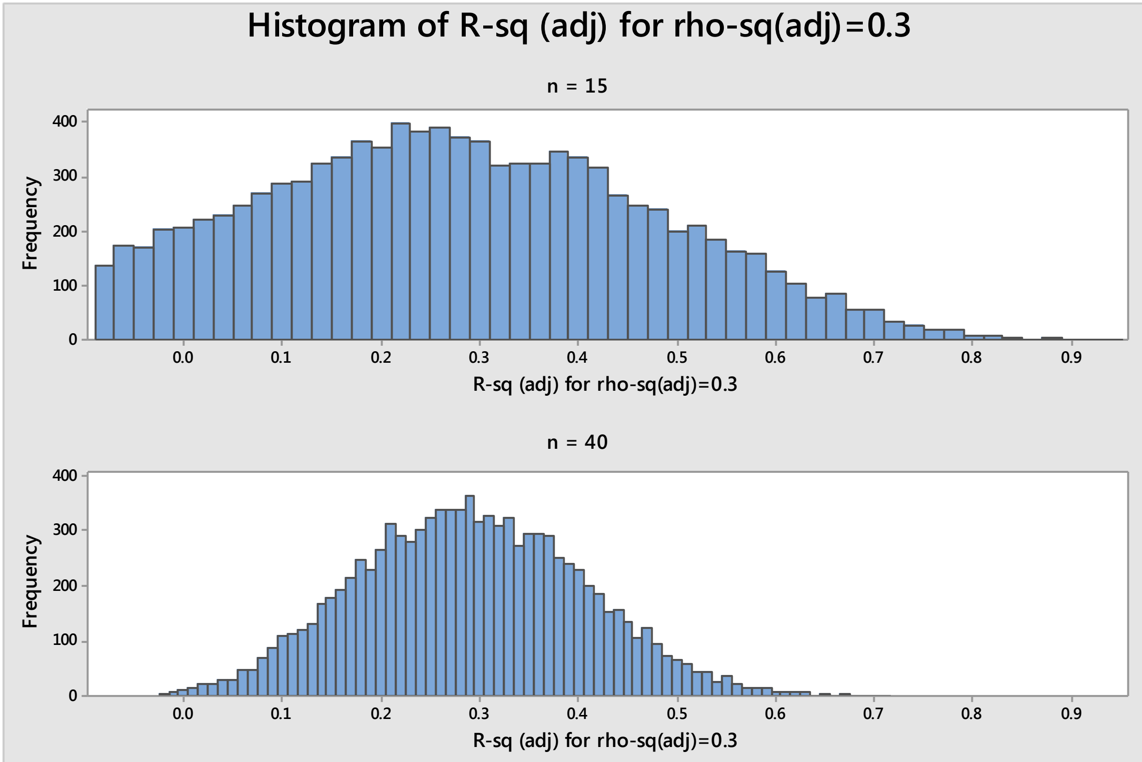


Figure 2 Simulated R^2_{adj} values for ρ^2_{adj} = 0.30 for n=15 and n=40

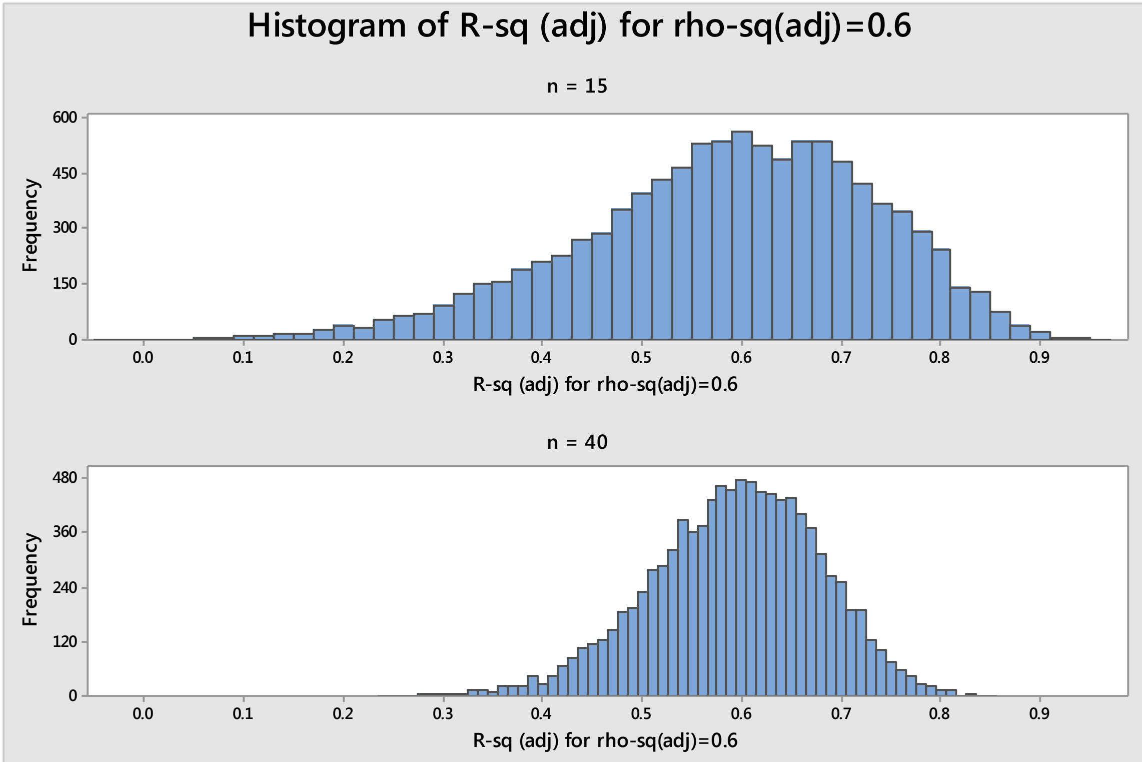


Figure 3 Simulated R^2_{adj} values for ρ^2_{adj} = 0.60 for n=15 and n=40

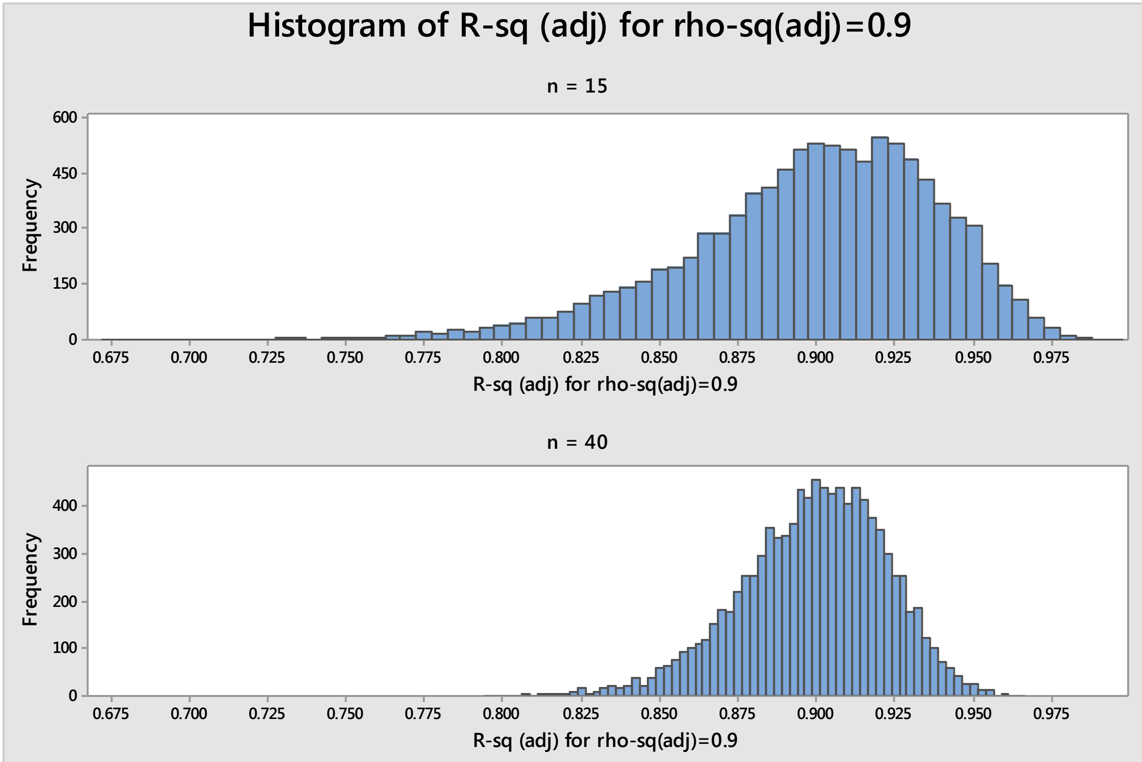


Figure 4 Simulated R^2_{adj} values for ρ^2_{adj} = 0.90 for n=15 and n=40

Overall, the simulations show that there can be a considerable difference between the actual strength of the relationship (ρ^2_{adj}) and the relationship observed in the data (R^2_{adj}). Increasing the sample size from 15 to 40 greatly reduces the likely magnitude of the difference. We determined that 40 observations is an appropriate threshold by identifying the minimum value of n for which absolute differences $|R^2_{adj} - \rho^2_{adj}|$ greater than 0.20 occur with no more than 10% probability. This is regardless of the true value of ρ^2_{adj} in any of the models considered. For the linear model, the most difficult case was ρ^2_{adj} = 0.31, which required n = 36. For the quadratic model, the most difficult case was ρ^2_{adj} = 0.30, which required n = 38. With 40 observations, you can be 90% confident that observed value of R^2_{adj} will be within 0.20 of ρ^2_{adj}, regardless of what that value is and whether you use the linear or quadratic model.

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 4 below shows the proportion of 10,000 simulations in which the overall F-test was significant at $\alpha = 0.05$ for various distributions of ε for the linear and quadratic models. In these simulations, the null hypothesis, which states that there is no relationship between X and Y, was true. The X values were evenly spaced over an interval. We used a sample size of n=15 for all tests.

Table 4 Type I error rates for overall F-tests for linear and quadratic models with n=15 for nonnormal distributions

Distribution	Linear significant	Quadratic significant
Normal	0.04770	0.05060
t(3)	0.04670	0.05150
t(5)	0.04980	0.04540
Laplace	0.04800	0.04720
Uniform	0.05140	0.04450
Beta(3, 3)	0.05100	0.05090
Exponential	0.04380	0.04880
Chi(3)	0.04860	0.05210
Chi(5)	0.04900	0.05260
Chi(10)	0.04970	0.05000
Beta(8, 1)	0.04780	0.04710

Next, we examined the test of the highest order term used to select the best model. For each simulation, we considered whether the square term was significant. For cases where the square term was not significant, we considered whether the linear term was significant. In these simulations, the null hypothesis was true, target $\alpha = 0.05$ and n=15.

Table 5 Type I error rates for tests of highest order term for linear or quadratic models with n=15 for nonnormal distributions

Distribution	Square	Linear
Normal	0.05050	0.04630
t(3)	0.05120	0.04300
t(5)	0.04710	0.04820
Laplace	0.04770	0.04660
Uniform	0.04670	0.04900
Beta(3, 3)	0.05000	0.04860
Exponential	0.04600	0.03800
Chi(3)	0.05110	0.04290
Chi(5)	0.05290	0.04490
Chi(10)	0.04970	0.04610
Beta(8, 1)	0.04770	0.04380

The simulation results show, that for both the overall F-test and for the test of the highest order term in the model, the probability of finding statistically significant results does not differ substantially for any of the error distributions. The Type I error rates are all between 0.038 and 0.0529.