CHAPTER 11

On Worthwhile Regressions, Big F's, and R^2

A message of this chapter is that a regression that is statistically significant is not necessarily a useful one, and that something more is needed. This "something more" is difficult to quantify. A "useful rule of thumb" is offered. Readers who are unhappy with this rule may still profit by reading the chapter and forming their own opinions of how they should feel about their fitted equations.

11.1. IS MY REGRESSION A USEFUL ONE?

Suppose, as in our steam data example, we had 25 observations with no pure error and fitted a straight line. The skeleton analysis of variance table would be as follows:

Source	df	MS	F
b_0	1		
$b_1 \mid b_0$	1	MS_{Reg}	Observed $F = MS_{Reg}/s^2$
Residual	23	s^2	
Total	25		

Assuming no defects were seen in the residuals, our observed F would be an $F(1, 23)$ variable under the null hypothesis of no slope. Suppose this were just significant at the $\alpha = 0.05$ test level, with the observed $F = 4.30 > 4.28 = F(1, 23, 0.95)$. Would we be happy? Momentarily, perhaps, but then we would look at the R^2 statistic. Since F and R^2 are directly connected by the relationship

$$R^2 = \nu_1 F/(\nu_1 F + \nu_2), \tag{11.1.1}$$

where (ν_1, ν_2) are the degrees of freedom of MS_{Reg} and s^2, respectively, we can evaluate $R^2 = 4.30/(4.30 + 23) = 0.1575$ or 15.75%. Not very high. So although our regression is "statistically significant," it is certainly not explaining much of the variation, and it is unlikely that it will predict future observations with much accuracy, or even provide a good summary of the behavior of the data.

One possible viewpoint of a useful or worthwhile model fit is that the range of values predicted by the fitted equation over the X-space should be considerably greater

than the size of the errors with which these predictions are made. This viewpoint was investigated by Box and Wetz in 1964. See Box and Wetz (1973). They concluded (details in Appendix 11A) that in order that an equation should be regarded as a satisfactory predictor (in the sense that the range of response values predicted by the equation is substantial compared with the standard error of the response), the observed F-ratio of (regression mean square)/(residual mean square) should exceed not merely the selected percentage point of the F-distribution, but several times the selected percentage point. How many times depends essentially on how great a ratio (prediction range)/(error of predictions) is specified. We offer the following conservative *Rule of Thumb*. Unless the observed F for overall regression exceeds the chosen test percentage point by at least a factor of four, and preferably more, the regression is unlikely to be of practical value for prediction purposes.

(We call this conservative because otherwise we would be tempted to replace "four" by "ten." See Appendix 11A for the reasoning.)

Let us see how this would affect the calculation above. If for $\nu_1 = 1$, $\nu_2 = 23$ we insisted on achieving four times the percentage point, the observed F would need to exceed $4(4.28) = 17.12$. The corresponding R^2 would then be, from (11.1.1), $R^2 = 17.12/(17.12 + 23) = 0.4267$ or 42.67%. Adoption of a "ten times" rule would lead to an F of at least 42.8, and a corresponding $R^2 = 42.8/(42.8 + 23) = 0.6505$ or 65.05%. Note that the steam data fit of Chapter 1 had $F = 57.54$, which is over 13 times as big as the 5% point 4.28.

A little thought about all this will lead the reader to ask: "Does this mean that if I am a 5% person—one willing to be wrong once in 20 times—I should instead act like a person with a much smaller percentage in general, to ensure that I get a useful model?" Essentially, yes, but it is easier to specify a factor than it is to specify how to change a 5% (or 2.5%, or 1%) test value down to a lower test value. (Also, this argument should be carried out only once, or one argues one's way iteratively down into the tail of the F-distribution!)

An Alternative and Simpler Check

A simple but effectively equivalent way of implementing the Box–Wetz type of check is to find the ratio

$$(\text{Max } \hat{Y}_i - \text{Min } \hat{Y}_i)/\{ps^2/n\}^{1/2}. \tag{11.1.2}$$

Provided this is sufficiently large (a rough rule of thumb of about four is suggested as *minimal*) and provided no other defect is seen in the fit, a worthwhile regression interpretation is likely to be possible. [Use of the ratio (11.1.2) in such a manner was suggested by G. E. P. Box. The "four[1] at least" rule of thumb is our suggestion.]

In (11.1.2), the numerator of the ratio is the range of $\hat{Y}$ values that arise *at data points*. The value ps^2/n estimates $p\sigma^2/n$, where p = number of parameters fitted, n = number of observations used, and $\sigma^2 = V(Y_i)$. This quantity $p\sigma^2/n$ is the average variance of the n fitted values, namely,

$$\overline{V}(\hat{Y}) = \sum_{i=1}^{n} V(\hat{Y}_i)/n = p\sigma^2/n. \tag{11.1.3}$$

(The last-mentioned result is proved below.)

[1] Some would say ten.

Example 1. For the steam data fit of Y on Y_8 in Chapter 1: Max $\hat{Y}_i = 11.38$, Min $\hat{Y}_i = 7.50$; $n = 25$, $p = 2$, $s^2 = 0.7923$. So the ratio is

$$(11.38 - 7.50)/\{2(0.7923)/25\}^{1/2} = 3.88/0.252 = 15.4.$$

Thus the range of predicted values is over 15 times the average standard error of prediction, clearly satisfactory. This is obviously consistent with the fact that the F-ratio of 57.54 is 13.4 times as great as the 95% point $F(1, 23, 0.95) = 4.28$.

Example 2. For the steam data fit of Y on X_8 and X_6 in Section 6.2, the ratio (11.1.2) is

$$(11.56 - 6.29)/\{3(0.4377)/25\}^{1/2} = 5.27/0.2292 = 23.0.$$

Alternatively, the F ratio of 61.90 is 17.99 times as great as the 95% point $F(2, 22, 0.95) = 3.44$. Both numbers indicate a useful predictive ability for this fit.

Proof of (11.1.3)

$$\begin{aligned} \mathbf{V}(\hat{\mathbf{Y}}) &= \mathbf{V}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) && \text{[using rule that} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I}\sigma^2)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' && \mathbf{V}(\mathbf{AY}) = \mathbf{A}\{\mathbf{V}(\mathbf{Y})\}\mathbf{A}'] \quad (11.1.4) \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2. \end{aligned}$$

The summation in (11.1.3) is now the sum of the diagonal terms of the above matrix, that is, its trace. In what follows we use result 6 of Appendix 5A, that trace **(AB)** = trace **(BA).** We have

$$\begin{aligned} n\overline{V}(\hat{Y}) = \Sigma V(\hat{Y}_i) &= \text{trace}\,\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\} \\ &= \sigma^2\,\text{trace}\,\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\} \\ &= \sigma^2\,\text{trace}\,\mathbf{I}_p \\ &= p\sigma^2. \end{aligned}$$

Comment

What we are really getting at, in this section, is that a test using the overall F-statistic in the usual manner is not a very good way to judge the future usefulness of the regression equation. Asking for a larger F-value (four times, or more times, the percentage point) is a crude way to attempt to recalibrate the F-test. Another way is the use of (11.1.2). Or one can decide on a suitable level of R^2. "How big should these statistics be?" you ask. In the end, like many things in life, one has to figure out how to judge regressions by acquiring experience. The statistics provide some guide but are not absolute (as we would wish them to be).

11.2. A CONVERSATION ABOUT R^2

The R^2 statistic is used almost universally in judging regression equations. It is probably the first thing most of us look at, but it is not the only thing. It is an extremely useful indicator even if there are no absolute rules about how big it should be. It does have its drawbacks, however. Suppose we have n observations at m sites and there are n_e df for pure error. Suppose we fit a model with p parameters. Then a skeleton analysis of variance table takes the form:

Source	df	SS	MS
Regression$\|b_0$	$p - 1$	A	a
Lack of fit	$n - p - n_e$	B	b
Pure error	n_e	C	c
Total corrected	$n - 1$	D	d

Then

$$R^2 = A/D = 1 - (B + C)/D \tag{11.2.1}$$

and we can appreciate that:

1. If there is no pure error, $n_e = 0$, $C = 0$, and $R^2 = 1 - B/D$. B now has $(n - p)$ df. So as we enlarge the model, $n - p$ and B will be reduced, and R^2 may look deceptively good, simply because we are fitting to nearly all, or all, the degrees of freedom.
2. Suppose there is pure error, so that $C > 0$, $c > 0$. Consider the extreme case $B = 0$, $n - p - n_e > 0$, that is, there is no lack of fit but there are more sites of data than there are parameters. Then $R^2 = 1 - C/D$. Even though the model fits well, R^2 is limited by the fact that no model can explain the pure error so its maximum value is $1 - C/D$.
3. Outliers produce Y-values that are inflated or deflated. These either inflate C if they occur in a set of repeat runs, or B if not. In any case they reduce R^2 so R^2 is not resistant to outliers.
4. If the model has no intercept, $D = S_{YY} = \Sigma(Y_i - \overline{Y})^2$ does not occur naturally in the analysis and we recommend that calculation of R^2 should not be made. (Of course, we do not advocate omitting the intercept without careful consideration.)
5. Note that, if generalized least squares is used, the model becomes a no-intercept model, because the **1** vector of the first column is replaced by another vector.

Other problems with R^2 discussed in the literature are that:

6. R^2 is not comparable between models that contain different predictor variables. (If one model contains a subset of the parameters in another model, the comparison *is* valid, however.)
7. R^2 is not comparable between models that involve different transformations of the response vector.
8. Non-least-squares models need specially defined statistics.

What Should One Do for Linear Regression?

Probably the best advice for the moment is: continue to use R^2 (or adjusted R^2, if you prefer that) with caution appropriate to the points made above. For some of the debate, consult the references below.

References

Anderson-Sprecher (1994); Becker and Kennedy (1992); Ding and Bargmann (1991); Draper (1984); Healy (1984); Kvalseth (1985); Magee (1990); Nagelkerke (1991); Schemper (1990); Scott and Wild (1991); Shah (1991); Willett and Singer (1988).

APPENDIX 11A. HOW SIGNIFICANT SHOULD MY REGRESSION BE?

The application to regression situations of work by Box and Wetz is briefly explained. For a "useful" as distinct from a "significant" regression, the observed F-value for regression should exceed the usual percentage point by a multiple. The size of this multiple is arbitrary in the same way that significance levels are arbitrary, but guidelines are given for its choice.

The γ_m Criterion

In regression problems, provided no lack of fit is indicated, a test for regression is usually conducted by looking at the F-ratio of the "regression sum of squares given b_0" to "the residual mean square s^2." This value is then compared with an appropriate upper α-percentage point $F(\nu_m, \nu_r, 1 - \alpha)$, where ν_m and ν_r are, respectively, the degrees of freedom of the numerator and denominator of the F-statistic. When this F-ratio is statistically significant, that is, when $F > F(\nu_m, \nu_r, 1 - \alpha)$, it implies that a significantly large amount of the variation in the data about the mean has been taken up by the regression equation. This does not necessarily mean, however, that the fitted equation is a worthwhile predictor in the sense that the range of predicted values is "substantial" compared with the standard error of the response. The question then arises as to how we can distinguish statistically significant *and* worthwhile prediction equations from statistically significant prediction equations of limited practical value.

Some work that answers this question to a great extent appeared in a 1964 University of Wisconsin Ph.D. thesis "Criteria for judging adequacy of estimation by an approximating response function," by J. Wetz. (See Box and Wetz, 1973). The key finding may be summarized in the following way.

Suppose we fit, by least squares, the model

$$\mathbf{Y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\psi} + \boldsymbol{\varepsilon}, \tag{11A.1}$$

where $\mathbf{X}\boldsymbol{\beta}$ is the part to be tested in a "test for regression" and $\mathbf{Z}\boldsymbol{\psi}$ represents effects such as the mean, block variables, time trends, and so on, that we wish to eliminate from the variation in the data but in which we otherwise have no interest. We assume $\mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $V(\boldsymbol{\varepsilon}) = \mathbf{I}\sigma^2$. The changes in response values η_i over the n experimental points can conveniently be measured by the quantity

$$\sum_{i=1}^{n} (\eta_i - \tilde{\eta}_i)^2/n, \tag{11A.2}$$

where η_i is the true response of the ith observation, and $\tilde{\eta}_i$ is the ith element of the vector $\tilde{\boldsymbol{\eta}} = \mathbf{Z}\boldsymbol{\psi}$. When only the intercept term β_0 is eliminated, $\tilde{\eta}_i = \overline{\eta}$, the mean of the η_i.

We can compare the quantity in Eq. (11A.2) with the size of the errors we commit in estimating the differences $\eta_i - \tilde{\eta}_i$. The least squares estimator of $\eta_i - \tilde{\eta}_i$ is $\hat{Y}_i - \tilde{Y}_i$, the ith element of the vector

$$\hat{\mathbf{Y}} - \tilde{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}, \tag{11A.3}$$

say, and the variance–covariance matrix of this is given by

$$E\{\mathbf{H}\mathbf{Y} - E(\mathbf{H}\mathbf{Y})\}\{\mathbf{H}\mathbf{Y} - E(\mathbf{H}\mathbf{Y})\}' = \mathbf{H}\sigma^2\mathbf{H}' = \mathbf{H}\sigma^2, \tag{11A.4}$$

where $\mathbf{V}(\mathbf{Y}) = \mathbf{I}\sigma^2$, due to the fact that $\mathbf{H}$ is symmetric and idempotent. Thus

$V(\hat{Y}_i - \tilde{Y}_i)$ is the ith diagonal element of $\mathbf{H}\sigma^2$, and the average of these variances, which is an overall measure of how well we estimate the quantities $\eta_i - \tilde{\eta}_i$ is given by

$$\begin{aligned} \sigma^2_{\hat{\mathbf{Y}}-\tilde{\mathbf{Y}}} &= \text{trace}(\mathbf{H}\sigma^2)/n \\ &= \nu_m\sigma^2/n \end{aligned} \tag{11A.5}$$

due to the fact that trace $\mathbf{H}$ = trace $\mathbf{X}\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}$ = trace$\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\mathbf{X}$ = trace$(\mathbf{I}_{\nu_m})$, where ν_m is the number of parameters in $\boldsymbol{\beta}$. It follows that a sensible comparison of the sizes of changes in the $\eta_i - \tilde{\eta}_i$ with their errors of estimate is given by the square root of the ratio of Eq. (11A.2) to Eq. (11A.5), namely,

$$\gamma_m = \left\{\sum_{i=1}^{n} (\eta_i - \tilde{\eta}_i)^2/(\nu_m\sigma^2)\right\}^{1/2}. \tag{11A.6}$$

Figure 11A.1 shows the situation for a single predictor variable X; γ_m is comparing the spread of the heavy lines (the $\eta_i - \tilde{\eta}_i$) with the average spread of their estimates, whose distributions are shown at the various X_i. How big should γ_m be for a fitted equation to be practically useful as distinct from just statistically significant? This is, to a great extent, arbitrary, in the same sense that a selected statistical significance level is arbitrary. (However, to help fix ideas, we shall soon think in terms of values γ_m = 2, 3, and 4 so that we can examine the consequences and choose accordingly.) Suppose that γ_0 is the minimally acceptable level of γ_m. Then Box and Wetz show that we shall need to determine a certain value F_0, dependent on γ_0, and, if the usual regression F-ratio exceeds this value F_0, then we shall accept that γ_m is sufficiently large for the regression fit to be a practically useful one. Box and Wetz show further that, approximately, this critical value F_0 is

$$F_0 \simeq (1 + \gamma_0^2)F(\nu_0, \nu_r, 1 - \alpha), \tag{11A.7}$$

where ν_r is the number of residual degrees of freedom and where

$$\nu_0 = \nu_m(1 + \gamma_0^2)^2/(1 + 2\gamma_0^2). \tag{11A.8}$$

In other words, in order for the fit to be practically worthwhile, $F > F_0$. It is, of course, easy to work out F_0 in any specific case, but it is also informative to look at the ratios

$$F_0/F(\nu_m, \nu_r, 1 - \alpha) \tag{11A.9}$$

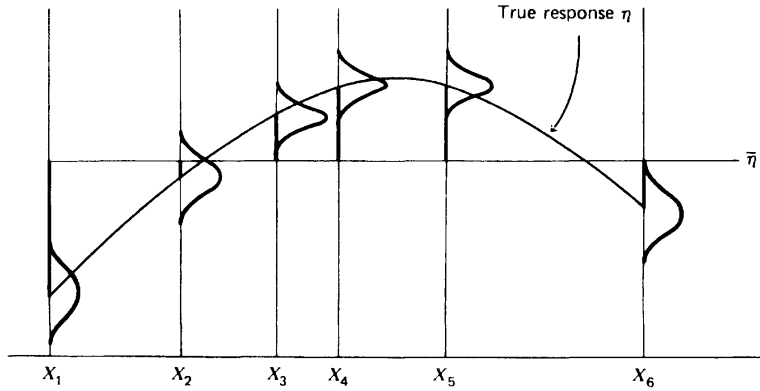

Figure 11A.1. Deviations of true η values about their mean compared with the spreads of the estimates $\hat{Y}_i - \tilde{Y}_i$ for a single X-variable.

TABLE 11A.1. Ratios $F_0/F(\nu_m, \nu_r, 0.95)$ for $\gamma_0 = 2$

Residual Degrees of Freedom, ν_r	Regression Degrees of Freedom, ν_m								
	1	2	3	4	5	6	10	15	21
3	5	5	5	5	5	5	5	5	5
4	4	4	5	5	5	5	5	5	5
5	4	4	4	5	5	5	5	5	5
10	4	4	4	4	4	4	5	5	5
15	4	4	4	4	4	4	4	5	5
20	4	4	4	4	4	4	4	5	5
30	4	4	4	4	4	4	4	4	5
40 to ∞	3	4	4	4	4	4	4	4	4

for given values of γ_0, and various values of ν_m, ν_r, and α. Tables 11A.1, 11A.2, and 11A.3 show these ratios rounded to the nearest unit for $\gamma_0 = 2$, 3, and 4, respectively, and for $\alpha = 0.05$. We see that, at this probability level, if we are prepared to accept a value of $\gamma_0 = 2$ as being sufficiently informative for our purposes, we need to have an observed F of at least four or five times as large as the usual percentage point to declare the regression practically useful to us. Or, if we are prepared to accept a value of $\gamma_0 = 3$, the F must be at least six to ten times as large as the usual percentage point. Moving to Table 11A.3, we see that, as our selected γ_0 increases, the ratios both increase and vary more, depending on the degrees of freedom involved. (For $\alpha = 0.01$, the general picture is much the same, with ratio values being either the same or one or two units lower in most cases.)

In summary, it is clear that an observed F-ratio must be *at least* four or five times the usual percentage point for the minimum level of proper representation, as Table 11A.1 indicates. In practice, the multiples in Table 11A.2 are probably the ones that should be attained or exceeded in most practical cases, guaranteeing, as they do, a γ_m ratio of 3 or more. However, like the choice of a confidence level, much depends on individual preferences, and the tables are offered as guidelines for personal choice.

The stated results have been given in terms of the F-statistic for overall regression. However, similar results apply for subsets of coefficients in the fitted model, and so the rule may be applied to the F-values arising from such subsets, also, if desired. (See Ellerton, 1978.)

TABLE 11A.2. Ratios $F_0/F(\nu_m, \nu_r, 0.95)$ for $\gamma_0 = 3$

Residual Degrees of Freedom, ν_r	Regression Degrees of Freedom, ν_m								
	1	2	3	4	5	6	10	15	21
3	9	9	9	9	10	10	10	10	10
4	8	9	9	9	9	9	10	10	10
5	8	8	9	9	9	9	9	10	10
10	7	7	8	8	8	8	9	9	9
15	6	7	7	8	8	8	9	9	9
20	6	6	7	7	8	8	8	9	9
30	6	6	7	7	7	8	8	9	9
40	6	6	7	7	7	7	8	8	9
60	6	6	7	7	7	7	8	8	8
120	6	6	7	7	7	7	8	8	8
∞	6	6	6	7	7	7	7	8	8

TABLE 11A.3. Ratios $F_0/F(\nu_m, \nu_r, 0.95)$ for $\gamma_0 = 4$

Residual Degrees of Freedom, ν_r	Regression Degrees of Freedom, ν_m								
	1	2	3	4	5	6	10	15	21
3	15	15	16	16	16	16	17	17	17
4	13	14	15	15	16	16	16	16	17
5	12	13	14	15	15	15	16	16	16
10	10	12	13	13	14	14	15	15	16
15	10	11	12	12	13	13	14	15	15
20	9	11	11	12	12	13	14	15	15
30	9	10	11	11	12	12	14	14	15
40	9	10	11	11	12	12	13	14	14
60	9	10	10	11	11	12	13	14	14
120	9	9	10	11	11	11	13	13	14
∞	8	9	10	10	11	11	12	12	12

EXERCISES FOR CHAPTER 11

Note: You will need the formula

$$R^2 = \nu_1 F/(\nu_1 F + \nu_2) \tag{11E}$$

where ν_1 = number of parameters fitted − 1 (for β_0),
ν_2 = residual degrees of freedom.

A. Substitute specific values of ν_1, ν_2 and $F(\nu_1, \nu_2, 1 - \alpha)$ for chosen α (e.g., $\alpha = 0.05$) into the right-hand side of Eq. (11E). The R^2 values you get are those that correspond to "just significant at $100\alpha\%$" F-values. You will be surprised at how low these R^2 values can be, and this will provide additional motivation for reading Appendix 11A. Application of Tables 11A.1, 11A.2, and 11A.3 (in which $\nu_m = \nu_1$, $\nu_r = \nu_2$) will ensure higher R^2 values; you may wish to do a few sample calculations to see what the effects can be.

B. Your friend says he has fitted a plane to $n = 33$ observations on (X_1, X_2, Y) and that his overall regression (given b_0) is just significant at the $\alpha = 0.05$ level. You ask him for his R^2 value but he doesn't know. You work it out for him on the basis of what he has told you.

C. You perform a regression for a colleague. She gives you 46 data points, which include 5 sets of points each with 6 repeat runs, and the model contains 6 parameters including β_0. She tells you: "I hope the R^2 value will exceed 90%." For this to happen, how many times bigger than the 5% tabled F-value would F (for all parameters except b_0) have to be? (Assume that there is no lack of fit.)

D. You are given a regression printout that shows a planar fit to $X_1, X_2, \ldots, X_5$, plus intercept of course, obtained from a set of 50 observations.

1. The overall F for regression $F(b_1, b_2, \ldots, b_5|b_0)$ is ten times as big as the 5% upper-tail F percentage point. How big is R^2?

2. Looking at the data, you observe that they consist of six groups of repeats with 8, 8, 8, 8, 8, and 10 observations in them. What would you do now, and why?