LO 1

Quadratic Terms

In the women's downhill ski racing event at the Winter Olympic Games in Salt Lake City, Picabo Street of the U.S. team was disappointed with her 16th-place finish after she'd posted the fastest practice time. Changing snow conditions can affect finish times, and in fact the top seeds can choose their starting positions and try to guess when the conditions will be best. But how much impact was there? On the day of the women's downhill race, it was unusually sunny. Skiers expect conditions to improve and then, as the day wears on, to deteriorate, so they try to pick the optimum time. But their calculations were upset by a two-hour delay. Picabo Street chose to race in 26th position. By then conditions had turned around, and the slopes had begun to deteriorate. Was that the reason for her disappointing finish?

The regression in Table 1 seems to support her point. Times did get slower as the day wore on.

Dependent variable is: Time
R squared $= 37.9\%$ R squared (adjusted) $= 36.0\%$
s $= 1.577$ with $35 - 2 = 33$ degrees of freedom

Variable	Coeff	SE(Coeff)	*t*-ratio	P-Value
Intercept	100.069	0.5597	179	<0.0001
StartOrder	0.108563	0.0242	4.49	<0.0001

Table 1 Time to ski the women's downhill event at the 2002 Winter Olympics depended on starting position.

But a plot of the residuals (Figure 1) warns us that the Linearity Assumption isn't met.

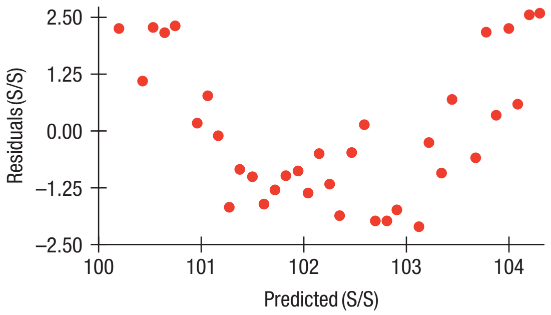


Figure 1 The residuals reveal a bend.

If we return to plot the data, we can see that re-expression can't help us because the times first trend down and then turn around and increase (Figure 2).

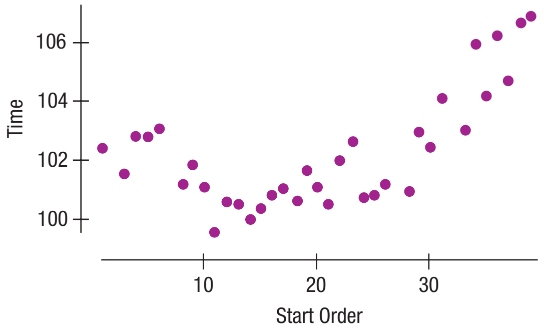


Figure 2 The original data trend down and then up. That kind of bend can't be improved with re-expression.

How can we use regression here? We can introduce a squared term to the model:

$$\hat{y} = b_0 + b_1\,startorder + b_2\,startorder^2$$

The fitted function is a *quadratic*, which can follow bends like the one in these data. Here is the regression table (Table 2).

Dependent variable is: Time
R squared = 83.3% R squared (adjusted) = 82.3%
s = 0.8300 with 35 − 3 = 32 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	110.139	2	55.0694	79.9
Residual	22.0439	32	0.688871	

Variable	Coeff	SE(Coeff)	*t*-ratio	P-Value
Intercept	103.547	0.4749	218	<0.0001
StartOrder	−0.367408	0.0525	−6.99	<0.0001
StartOrder^2	0.011592	0.0012	9.34	<0.0001

Table 2 A regression model with a quadratic term fits these data better.

This model fits the data better. Adjusted R^2 is 82.3%, up from 36.0% for the linear version. And the residuals look generally unstructured (Figure 3).

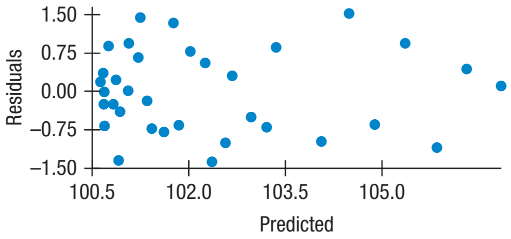


Figure 3 The residuals from the quadratic model show no structure.

However, one problem remains. In the new model, the coefficient of *Start Order* has changed from significant and positive to significant and negative. As we've just seen, that's a signal of possible collinearity. Quadratic models often have collinearity problems because for many variables, x and x^2 are highly correlated. In these data, *Start Order* and *Start Order*^{2} have a correlation of 0.97.

There's a simple fix for this problem. Instead of using *Start Order*^{2}, we can use $(Start\ Order - \overline{Start}\ Order)^2$. The form with the mean subtracted has a zero correlation with the linear term. Here's the resulting regression (Table 3).

Dependent variable is: Time
R squared = 83.3% R squared (adjusted) = 82.3%
s = 0.8300 with 35 − 3 = 32 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	110.139	2	55.0694	79.9
Residual	22.0439	32	0.688871	

Variable	Coeff	SE(Coeff)	*t*-ratio	P-Value
Intercept	98.7493	0.3267	302	<0.0001
StartOrder	0.104239	0.0127	8.18	<0.0001
(SO-mean)^2	0.011592	0.0012	9.34	<0.0001

Table 3 Using a centred quadratic term alleviates the collinearity.

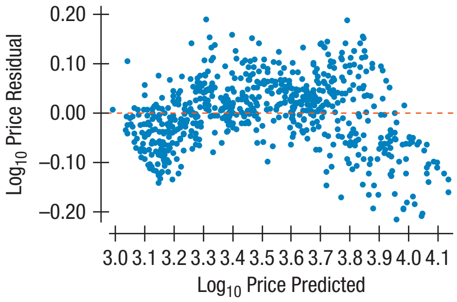
The predicted values and residuals are the same for these two models (this can be shown algebraically for any quadratic model with a centred squared term), but the coefficients of the second one are easier to interpret.

So did Picabo Street have a valid complaint? Well, times did increase with start order, but (from the quadratic term) they decreased before they turned around and increased. Picabo's start order of 26 has a predicted time of 101. 83 seconds. Her performance at 101.17 was better than predicted, but her residual of −0.66 is not large in magnitude compared with that of some of the other skiers. Skiers who had much later starting positions *were* disadvantaged, but Picabo's start position was only slightly later than the best possible one (about 16th according to this model), and her performance was not extraordinary by Olympic standards.

One final note: Quadratic models can do an excellent job of fitting curved patterns such as this one. But they are particularly dangerous to extrapolate beyond the range of the *x*-values. So you should use them with care.

FOR EXAMPLE Quadratic terms for diamond prices

We've fit a model *of Log*$_{10}$*Price* to *Carat Weight*, *Colour*, *Cut*, and *Clarity* (see For Example: "Stepwise regression for diamond prices".) The R^2 is an impressive 94.46%. The 749 diamonds are a random sample of diamonds of three *Colour* levels and of *Weight* between 0.3 and 1.5 carats. We transformed *Price* using the log (base 10) to linearize the relationship with *Carat Weight*. However, a plot of residuals vs. predicted values reveals:



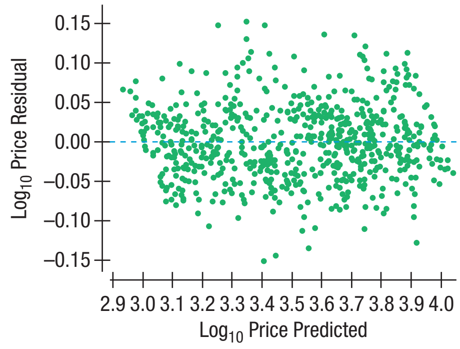
There appears to be a curved relationship between residuals and predicted values. What happens if we add *(Carat Weight)*^{2} as a predictor?

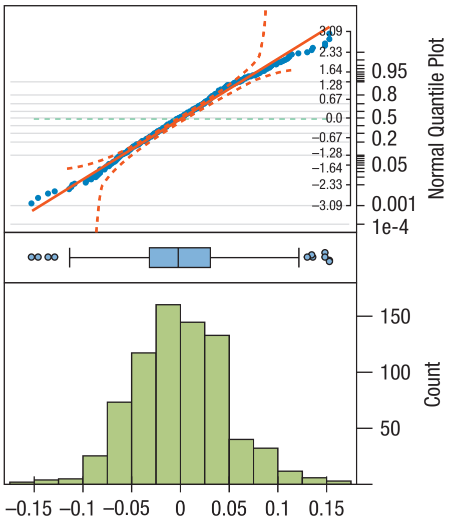
Here's the regression output:

Response Variable: Log_{10} Price
$R^2 - 97.31\%$ Adjusted $R^2 - 97.27\%$
$s = 0.04741$ with $749 - 14 = 735$ degrees of freedom

Variable	Coeff	SE(Coeff)	*t*-ratio	P-Value
Intercept	2.040547	0.017570	116.138	<0.0001
Carat.Weight	2.365036	0.042400	55.780	<0.0001
Carat.Weight^2	−0.699846	0.025028	−27.962	<0.0001
ColourD	0.347479	0.006346	54.755	<0.0001
ColourG	0.252336	0.005683	44.398	<0.0001
CutGood	−0.038896	0.007763	−5.010	<0.0001
CutIdea1	0.016165	0.006982	2.315	0.020877
CutVery Good	−0.014988	0.003834	−3.910	0.000101
ClaritySI1	−0.286641	0.008359	−34.292	<0.0001
ClaritySI2	−0.353371	0.008800	−40.155	<0.0001
ClarityVSI	−0.161288	0.008513	−18.947	<0.0001
ClarityVS2	−0.215559	0.008246	−26.141	<0.0001
ClarityVVSI	−0.077703	0.008682	−8.950	<0.0001
ClarityVVS2	−0.103078	0.008327	−12.378	<0.0001

The residuals—together with their Normal Probability plot, boxplot, and histogram—look like this:





(continued)

QUESTION Summarize this regression and comment on its appropriateness for predicting the price of diamonds (of this *Carat Weight*).

ANSWER The model for $Log_{10}Price$ is based on *Carat Weight*, *Carat Weight*^{2}, *Colour (three levels)*, *Cut (four levels)*, and *Clarity (seven levels)*. The R^2 for this model is 97.31% (97.27% adjusted), and the residual standard deviation is only 0.047 (in $Log_{10}Price$). Every term included in the model is statistically significant. The assumptions and conditions of multiple regression are all met. The residuals appear to be symmetric and roughly Normal. The inclusion of the squared term for *Carat Weight* has eliminated the pattern in the plot of residuals vs. predicted values. This model seems appropriate to use for other diamonds of this *Carat Weight*.

Regression Roles

We build regression models for a number of reasons. One reason is to model how variables are related to each other in the hope of understanding the relationships. Another is to build a model that might be used to predict values for a response variable when given values for the predictor variables. When we hope to understand, we're often particularly interested in simple, straightforward models in which predictors are as unrelated to each other as possible. We're especially happy when the t-statistics are large, indicating that the predictors each contribute to the model.

By contrast, when prediction is our goal, we're more likely to care about the overall R^2. Good prediction occurs when much of the variability in y is accounted for by the model. We might be willing to keep variables in our model that have relatively small t-statistics simply for the stability that having several predictors can provide. We care less whether the predictors are related to each other because we don't intend to interpret the coefficients anyway, so collinearity is less of a concern.

In both roles, we may include some predictors to "get them out of the way." Regression offers a way to approximately control for factors when we have observational data because each coefficient estimates a relationship *after removing the effects* of the other predictors. Of course, it would be better to control for factors in a randomized experiment, but in the real world of business that's often just not possible.