Univariate Ordinary Least Squares Estimator
Regression Constant Is Non-Zero

Gary Schurman MBE, CFA

May, 2011

There are many cases in finance where we need to estimate the unknown value of a dependent variable Y based upon the known value of an independent variable X. We can define the actual value of the dependent variable Y to be a function of the estimated value of Y plus an error term. We will define $\hat{Y}$ (Y hat) to be the estimated value of the dependent variable Y. The equation for the estimated value of Y is...

$$\hat{Y} = \alpha + \beta\, X \tag{1}$$

The actual value of Y is therefore the estimated value of Y plus an error term. Using Equation (1) above the equation for the actual value of the dependent variable Y where ϵ is the error term is...

$$\begin{aligned} Y &= \hat{Y} + \epsilon \\ &= \alpha + \beta\, X + \epsilon \end{aligned} \tag{2}$$

How do we go about estimating the values of α and β in equation (1) above? We will set these values such that the sum of squared estimation errors is minimized. To demonstrate these techniques we will work through the following problem...

The Problem

We are tasked with projecting next year's unemployment rate given an annualized rate of GDP growth. We will define the dependent variable Y to be the percentage unemployment rate and the independent variable X to be the percentage GDP growth rate. To help us establish this relationship we will use the table below that gives us historical data that represents a sample of the observed percentage unemployment rate, the observed percentage GDP growth rate and the statistics needed to calculate the means, variances, covariance and correlation...

Observ Num	X GDP	Y UnEmp	X Squared	Y Squared	XY Product
1	3.00	5.00	9.00	25.00	15.00
2	3.50	5.00	12.25	25.00	17.50
3	0.00	9.00	0.00	81.00	0.00
4	-2.00	11.00	4.00	121.00	-22.00
5	2.00	7.00	4.00	49.00	14.00
6	4.00	4.50	16.00	20.25	18.00
7	1.25	5.00	1.56	25.00	6.25
8	-4.00	10.50	16.00	110.25	-42.00
9	1.00	4.25	1.00	18.06	4.25
10	2.00	4.50	4.00	20.25	9.00
11	6.00	2.00	36.00	4.00	12.00
12	4.00	2.00	16.00	4.00	8.00
13	5.25	0.50	27.56	0.25	2.63
14	-1.00	7.00	1.00	49.00	-7.00
15	2.75	5.50	7.56	30.25	15.13
Total	27.75	82.75	155.94	582.31	50.75

Our goal is to (1) project the unemployment rate given a mild recession, which we will define as a negative one percent GDP growth rate, (2) calculate a confidence interval around that projected unemployment rate, and (3) calculating the goodness of fit.

The Equation for Total Sum of Squared Errors and it's Derivatives

The estimation error squared is the squared difference between the actual value of Y and the estimated value of Y. The equation for the total sum of squared errors (SSE) is therefore...

$$\begin{aligned}
SSE &= \sum_{i=1}^{N}\left[\left\{Y_i - (\alpha + \beta X_i)\right\}^2\right] \\
&= \sum_{i=1}^{N}\left[Y_i^2 - 2\alpha Y_i + 2\alpha\beta X_i - 2\beta Y_i X_i + \alpha^2 + \beta^2 X_i^2\right] \\
&= \sum_{i=1}^{N}\left[Y_i^2\right] - \sum_{i=1}^{N}\left[2\alpha Y_i\right] + \sum_{i=1}^{N}\left[2\alpha\beta X_i\right] - \sum_{i=1}^{N}\left[2\beta Y_i X_i\right] + \sum_{i=1}^{N}\left[\alpha^2\right] + \sum_{i=1}^{N}\left[\beta^2 X_i^2\right] \\
&= \sum_{i=1}^{N}\left[Y_i^2\right] - 2\alpha\sum_{i=1}^{N}\left[Y_i\right] + 2\alpha\beta\sum_{i=1}^{N}\left[X_i\right] - 2\beta\sum_{i=1}^{N}\left[Y_i X_i\right] + N\alpha^2 + \beta^2\sum_{i=1}^{N}\left[X_i^2\right] \qquad (3)
\end{aligned}$$

To calculate values for the parameters in equation (1) we will need the derivatives of equation (3) with respect to both α and β. The derivative of equation (3) with respect to α is...

$$\begin{aligned}
\frac{\delta SSE}{\delta\alpha} &= 2N\alpha + 2\beta\sum_{i=1}^{N}\left[X_i\right] - 2\sum_{i=1}^{N}\left[Y_i\right] \\
&= 2N\alpha + 2N\beta\sum_{i=1}^{N}\left[\frac{1}{N}X_i\right] - 2N\sum_{i=1}^{N}\left[\frac{1}{N}Y_i\right] \\
&= 2N\alpha + 2N\beta\bar{X} - 2N\bar{Y} \\
&= 2N\left[\alpha + \beta\bar{X} - \bar{Y}\right] \qquad (4)
\end{aligned}$$

Note that $\bar{X}$ (X bar) is the mean of the observed values of the independent variable X and $\bar{Y}$ (Y bar) is the mean of the observed values of the dependent variable Y.

The derivative of equation (3) with respect to β is...

$$\begin{aligned}
\frac{\delta SSE}{\delta\beta} &= 2\beta\sum_{i=1}^{N}\left[X_i^2\right] + 2\alpha\sum_{i=1}^{N}\left[X_i\right] - 2\sum_{i=1}^{N}\left[Y_i X_i\right] \\
&= 2N\beta\sum_{i=1}^{N}\left[\frac{1}{N}X_i^2\right] + 2N\alpha\sum_{i=1}^{N}\left[\frac{1}{N}X_i\right] - 2N\sum_{i=1}^{N}\left[\frac{1}{N}Y_i X_i\right] \\
&= 2N\left\{\beta\sum_{i=1}^{N}\left[\frac{1}{N}X_i^2\right] + \alpha\sum_{i=1}^{N}\left[\frac{1}{N}X_i\right] - \sum_{i=1}^{N}\left[\frac{1}{N}Y_i X_i\right]\right\} \\
&= 2N\left\{\beta\sum_{i=1}^{N}\left[\frac{1}{N}X_i^2\right] + \alpha\bar{X} - \sum_{i=1}^{N}\left[\frac{1}{N}Y_i X_i\right]\right\} \qquad (5)
\end{aligned}$$

Solving for Alpha and Beta

To solve for α and β in equation (1) above we will set derivative equations (4) and (5) both equal to zero and jointly solve for these two parameters. The value of α is...

$$\begin{aligned}
\frac{\delta SSE}{\delta\alpha} &= 0 \\
2N\left[\alpha + \beta\bar{X} - \bar{Y}\right] &= 0 \\
\alpha &= \bar{Y} - \beta\bar{X} \qquad (6)
\end{aligned}$$

To calculate the variance of the estimation error we will need an equation for α^2 which is...

$$\begin{aligned}\alpha^2 &= (\bar{Y} - \beta\bar{X})^2 \\ &= \bar{Y}^2 - 2\beta\bar{X}\bar{Y} + \beta^2\bar{X}^2 \end{aligned} \tag{7}$$

Using the value of α in equation (6) above, the value of β is...

$$\begin{aligned}\frac{\delta SSE}{\delta\beta} &= 0 \\ 2N\left\{\beta\sum_{i=1}^{N}\left[\frac{1}{N}X_i^2\right] + \alpha\bar{X} - \sum_{i=1}^{N}\left[\frac{1}{N}Y_iX_i\right]\right\} &= 0 \\ 2N\left\{\beta\sum_{i=1}^{N}\left[\frac{1}{N}X_i^2\right] + \left\{\bar{Y} - \beta\bar{X}\right\}\bar{X} - \sum_{i=1}^{N}\left[\frac{1}{N}Y_iX_i\right]\right\} &= 0 \\ \beta\left\{\sum_{i=1}^{N}\left[\frac{1}{N}X_i^2\right] - \bar{X}^2\right\} - \left\{\sum_{i=1}^{N}\left[\frac{1}{N}Y_iX_i\right] - \bar{Y}\bar{X}\right\} &= 0 \\ \beta\,\sigma_x^2 - Cov(xy) &= 0 \\ \beta &= \frac{Cov(xy)}{\sigma_x^2}\end{aligned} \tag{8}$$

Noting that the correlation between X and Y is...

$$\begin{aligned}\rho_{xy} &= \frac{Cov(xy)}{\sigma_x\sigma_y} \\ \rho_{xy}\sigma_x\sigma_y &= Cov(xy)\end{aligned} \tag{9}$$

The value of β in equation (8) above can also be expressed as...

$$\beta = \frac{Cov(xy)}{\sigma_x^2} = \frac{\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2} = \rho_{xy}\frac{\sigma_y}{\sigma_x} \tag{10}$$

To calculate the variance of the estimation error we will need an equation for β^2 which is...

$$\begin{aligned}\beta^2 &= \left[\rho_{xy}\frac{\sigma_y}{\sigma_x}\right]^2 \\ &= \rho_{xy}^2\frac{\sigma_y^2}{\sigma_x^2}\end{aligned} \tag{11}$$

Residuals

A residual is defined as the differences between the estimated value of Y via Equation (1) above and the actual value of Y. Residuals are captured in the error term ϵ. We can rewrite Equation (2) as...

$$\epsilon = Y - \alpha - \beta\,X \tag{12}$$

The expected value of the error term ϵ in Equation (12) above is...

$$\begin{aligned}\mathbb{E}\left[\epsilon\right] &= \mathbb{E}\left[Y - \alpha - \beta\,X\right] \\ &= \mathbb{E}\left[Y\right] - \mathbb{E}\left[\alpha + \beta\,X\right] \\ &= Y - Y \\ &= 0\end{aligned} \tag{13}$$

The expected value of the square of the error term ϵ in Equation (12) above is...

$$\begin{aligned}\mathbb{E}\Big[\epsilon^2\Big] &= \mathbb{E}\Big[\Big(Y-\alpha-\beta\,X\Big)^2\Big] \\ &= \mathbb{E}\Big[Y^2-2\alpha Y-2\beta XY+2\alpha\beta X+\alpha^2+\beta^2X^2\Big] \end{aligned} \tag{14}$$

After substituting equations (6) and (7) above for α and α^2 in equation (14) above the equation for the expected value of the square of the error term becomes...

$$\begin{aligned}\mathbb{E}\Big[\epsilon^2\Big] &= \mathbb{E}\Big[Y^2-2(\bar{Y}-\beta\bar{X})Y-2\beta XY+2(\bar{Y}-\beta\bar{X})\beta X+(\bar{Y}^2-2\beta\bar{X}\bar{Y}+\beta^2\bar{X}^2)+\beta^2X^2\Big] \\ &= \mathbb{E}\Big[Y^2\Big]-\mathbb{E}\Big[2(\bar{Y}-\beta\bar{X})Y\Big]-\mathbb{E}\Big[2\beta XY\Big]+\mathbb{E}\Big[2(\bar{Y}-\beta\bar{X})\beta X\Big]+\mathbb{E}\Big[\bar{Y}^2-2\beta\bar{X}\bar{Y}+\beta^2\bar{X}^2)\Big]+\mathbb{E}\Big[\beta^2X^2\Big] \\ &= \mathbb{E}\Big[Y^2\Big]-2(\bar{Y}-\beta\bar{X})\mathbb{E}\Big[Y\Big]-2\beta\mathbb{E}\Big[XY\Big]+2\beta(\bar{Y}-\beta\bar{X})\mathbb{E}\Big[X\Big]+\bar{Y}^2-2\beta\bar{X}\bar{Y}+\beta^2\bar{X}^2)+\beta^2\mathbb{E}\Big[X^2\Big] \\ &= \mathbb{E}\Big[Y^2\Big]-2(\bar{Y}-\beta\bar{X})\bar{Y}-2\beta\mathbb{E}\Big[XY\Big]+2\beta(\bar{Y}-\beta\bar{X})\bar{X}+\bar{Y}^2-2\beta\bar{X}\bar{Y}+\beta^2\bar{X}^2+\beta^2\mathbb{E}\Big[X^2\Big] \\ &= \mathbb{E}\Big[Y^2\Big]-2\bar{Y}^2+2\beta\bar{X}\bar{Y}-2\beta\mathbb{E}\Big[XY\Big]+2\beta\bar{X}\bar{Y}-2\beta^2\bar{X}^2+\bar{Y}^2-2\beta\bar{X}\bar{Y}+\beta^2\bar{X}^2+\beta^2\mathbb{E}\Big[X^2\Big] \\ &= \Big\{\mathbb{E}\Big[Y^2\Big]-\bar{Y}^2\Big\}-2\beta\Big\{\mathbb{E}\Big[XY\Big]+\bar{X}\bar{Y}\Big\}+\beta^2\Big\{\mathbb{E}\Big[X^2\Big]-\bar{X}^2\Big\} \\ &= \sigma_y^2-2\beta Cov(xy)+\beta^2\sigma_x^2 \end{aligned} \tag{15}$$

After noting that per Equation (8) above...

$$\begin{aligned}\beta &= \frac{Cov(xy)}{\sigma_x^2} \\ \beta\sigma_x^2 &= Cov(xy) \end{aligned} \tag{16}$$

Equation (15) becomes...

$$\begin{aligned}\mathbb{E}\Big[\epsilon^2\Big] &= \sigma_y^2-2\beta^2\sigma_x^2+\beta^2\sigma_x^2 \\ &= \sigma_y^2-\beta^2\sigma_x^2 \end{aligned} \tag{17}$$

After substituting equation (11) for β^2 in equation (17), the equation for the expected value of the square of the error term becomes...

$$\begin{aligned}\mathbb{E}\Big[\epsilon^2\Big] &= \sigma_y^2-\rho_{xy}^2\frac{\sigma_y^2}{\sigma_x^2}\sigma_x^2 \\ &= \sigma_y^2-\rho_{xy}^2\sigma_y^2 \\ &= (1-\rho_{xy}^2)\sigma_y^2 \end{aligned} \tag{18}$$

We now have enough information to calculate the mean and variance of the error term ϵ. Using Equation (13) above the mean of the error term ϵ is...

$$\text{mean of } \epsilon = \mathbb{E}\Big[\epsilon\Big] = 0 \tag{19}$$

Using Equations (13) and (18) above the variance of the error term ϵ is...

$$\text{variance of } \epsilon = \mathbb{E}\Big[\epsilon^2\Big]-\Big(\mathbb{E}\Big[\epsilon\Big]\Big)^2 = (1-\rho_{xy}^2)\sigma_y^2-0 = (1-\rho_{xy}^2)\sigma_y^2 \tag{20}$$

Goodness of Fit

A well-fitting regression model results in predicted values close to the observed (i.e. actual) values. The mean model, which uses the mean of the data series for every predicted value, would be used if there were no informative predictor variables. The fit of a proposed regression model should therefore be better than the fit of the mean model.

We defined SSE to be the sum of squared errors using our proposed regression model (Equation (3) above). We will define SST to be the sum of squared errors using the mean model. The equations for SSE and SST are...

$$SSE = \sum_{i=1}^{N} \left[\left\{ Y_i - (\alpha + \beta X_i) \right\}^2 \right] \text{ ...and... } SST = \sum_{i=1}^{N} \left[\left\{ Y_i - \bar{Y} \right\}^2 \right] \text{ ...where... } \bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i \tag{21}$$

The difference between SST and SSE is the improvement in predictive power from using the regression model as compared to the mean model. The equation for R-squared is..

$$\text{R-squared} = \frac{SST - SSE}{SST} \tag{22}$$

The value of R-squared from zero to one. Zero means that the proposed model does not improve prediction over the mean model. One indicates perfect prediction. The closer to one that the value of R-squared gets the better the predictive power of the regression model.

The Solution To Our Problem

Step 1 - Calculate the means, variances, covariance and correlation of X and Y...

The mean of the percentage GDP growth rate is...

$$\bar{X} = \sum_{i=1}^{N} \left[\frac{1}{N} X_i \right] = \frac{1}{N} \sum_{i=1}^{N} \left[X_i \right] = \frac{27.75}{15} = 1.85$$

The mean of the percentage unemployment rate is...

$$\bar{Y} = \sum_{i=1}^{N} \left[\frac{1}{N} Y_i \right] = \frac{1}{N} \sum_{i=1}^{N} \left[Y_i \right] = \frac{82.75}{15} = 5.52$$

The variance of the percentage GDP growth rate is...

$$\sigma_x^2 = \sum_{i=1}^{N} \left[\frac{1}{N} X_i^2 \right] - \bar{X}^2 = \frac{1}{N} \sum_{i=1}^{N} \left[X_i^2 \right] - \bar{X}^2 = \frac{155.94}{15} - 1.85^2 = 6.97$$

The variance of the percentage unemployment rate is...

$$\sigma_y^2 = \sum_{i=1}^{N} \left[\frac{1}{N} Y_i^2 \right] - \bar{Y}^2 = \frac{1}{N} \sum_{i=1}^{N} \left[Y_i^2 \right] - \bar{Y}^2 = \frac{582.31}{15} - 5.52^2 = 8.39$$

The covariance of the percentage GDP growth rate and percentage unemployment rate is...

$$Cov(x, y) = \sum_{i=1}^{N} \left[\frac{1}{N} X_i Y_i \right] - \bar{X}\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} \left[X_i Y_i \right] - \bar{X}\bar{Y} = \frac{50.75}{15} - (1.85)(5.52) = -6.82$$

The correlation of the percentage GDP growth rate and percentage unemployment rate is...

$$\rho_{xy} = \frac{Cov(x, y)}{\sigma_x \sigma_y} = \frac{-6.82}{\sqrt{6.97}\sqrt{8.39}} = -0.89$$

Step 2 - Calculate the values of alpha and beta...

The value of beta is...

$$\beta = \frac{Cov(xy)}{\sigma_x^2} = \frac{-6.82}{6.97} = -0.9784$$

The value of alpha is...

$$\alpha = \bar{Y} - \beta \bar{X} = 5.52 - (-0.9784 \times 1.85) = 7.3267$$

Step 3 - Calculate the projected unemployment rate and a 95% confidence interval...

The projected unemployment...
$\hat{Y} = \alpha + \beta X = 7.3267 + (-0.9784)(-1.00) = 8.31$

The variance of projected unemployment...
variance of $\epsilon = (1 - \rho_{xy}^2)\sigma_y^2 = (1 - (-0.89)^2) \times 8.39 = 1.71$

Confidence interval...
$CI = 8.31 \pm 1.96 \times \sqrt{1.71}$

Step 4 - Calculate the goodness of fit...

Observ	GDP	Actual	Regression Model		Mean Model	
Num	X	Y	Estimate	Sq Err	Estimate	Sq Err
1	3.00	5.00	4.39	0.37	5.52	0.27
2	3.50	5.00	3.90	1.20	5.52	0.27
3	0.00	9.00	7.33	2.80	5.52	12.13
4	-2.00	11.00	9.28	2.95	5.52	30.07
5	2.00	7.00	5.37	2.66	5.52	2.20
6	4.00	4.50	3.41	1.18	5.52	1.03
7	1.25	5.00	6.10	1.22	5.52	0.27
8	-4.00	10.50	11.24	0.55	5.52	24.83
9	1.00	4.25	6.35	4.40	5.52	1.60
10	2.00	4.50	5.37	0.76	5.52	1.03
11	6.00	2.00	1.46	0.30	5.52	12.37
12	4.00	2.00	3.41	2.00	5.52	12.37
13	5.25	0.50	2.19	2.86	5.52	25.17
14	-1.00	7.00	8.31	1.70	5.52	2.20
15	2.75	5.50	4.64	0.75	5.52	0.00
Total	27.75	82.75	—	25.68	—	125.81

Using Equation (23) and the table above the goodness of fit is...

$$\text{R-squared} = \frac{SST - SSE}{SST} = \frac{125.81 - 25.68}{125.81} = 0.80 \tag{23}$$