Agribusiness Analysis and Forecasting

Multiple Regression Forecasts

Aleks Maisashvili

Texas A&M University

Multiple Regression Analysis

We would like to know following about regression analysis:

- How to use a regression to forecast a variable
- How to interpret the beta coefficients
- What the t ratio means
- What the p value is and what it means
- What the residuals are
- What the standard deviation is
- What is the F ratio and R^2 and what they are used for

Structural Variation

- Variables you want to forecast are often dependent on other variables.
  - Qt. Demand = f(Own Price, Substitute Price, Income, Population, Season, Tastes & Preferences, Trend, etc.).
  - $CropYield = a + b(Time, etc)$
- Structural models will explain most structural variation in a data series.
- Even when we build structural models, the forecast is not perfect.
- A residual remains as the unexplained portion.

Multiple Regression Forecasts

- Variables to include in a structural model are suggested by:
  - Economic theory
  - Knowledge of industry
  - Known relationships to other variables
- Examples of forecasting and uses:
  - Planted acres – needed by ag. input businesses
  - Demand for a product - sales and processors
  - Price of corn for cattle - feedlots, grain mills, etc.
  - Government payments - Congressional Budget Office
  - Exports or trade flows - international ag. business

Multiple Regression Forecasts

- Structural model

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + e$$

where X and Z are exogenous variables that explain the variation of Y over the historical period.

- Estimate parameters $(\beta_0, \beta_1, \beta_2, e)$ using OLS.
  - OLS minimizes the sum of squared residuals.
  - That is, we seek to explain as much of the variation in $\hat{Y}$ as possible, i.e., maximizing the precision of your probabilistic forecast.

Example of a Structural Forecast Model for a Crop

$$\text{Planted Acres}_t = f(Price_{t-1}, \text{Planted Acres}_{t-1}, IdleAcre_t, X_t)$$

$$HarvAc_t = f(PltAc_t)$$

$$Yield_t = f(Trend_t)$$

$$Production_t = Yield_t * HarvAc_t$$

$$Supply_t = Prod_t + EndStock_{t-1}$$

$$DomesticD_t = f(Price_t, Income/pop_t, Z_t)$$

$$Export_t = f(Price_t, Y_t)$$

$$EndingStock_t = f(Price_t, Production_t)$$

$$DomesticD_t + Export_t + Ending_Stock_t = Supply_t$$

Steps to build Multiple Regression Models

- Plot the Y variable in search of: trend, seasonal, cyclical, structural, and irregular variation.
- **Consider what relationships are suggested by economic theory and knowledge an industry**
- Plot Y vs. each X to evaluate the strength of hypothesized relationships; calculate correlation coefficients to Y.
- A Crop production forecast might be specified as follows:

$$PltAc_t = f(E(Price_t), PltAc_{t-1}, E(P^{th}Crop_t), Trend, Yield_{t-1})$$

$$HarvAc_t = \beta_0 + \beta_1 PltAc_t$$

$$Yield_t = \beta_0 + \beta_1 T$$

$$Prod_t = HarvAc_t * Yield_t$$

- Estimate with OLS.
- Make the deterministic forecast.
- Add stochastic error(s) for a probabilistic forecast.

US Planted Wheat Acreage Model

$$PltAc_t = f(E(Price_t), Yield_{t-1}, CRP_t, Year_t)$$

95%	Intercept	Years	Price t-1	CRP t	Yield t-1
Beta	**843.985**	**-0.412**	**9.237**	-0.073	0.459
S.E.	350.136	0.183	1.297	0.138	0.412
t-test	2.410	-2.245	7.124	-0.529	1.113
Prob(t)	0.024	0.034	0.000	0.601	0.276
Elasticity at Mean		-11.508	0.407	-0.008	0.229

- Statistically significant betas for the Trend ("Years") and Price.
- Leave *CRP* in model because its needed for policy analysis.
- Consider dropping $Yield_{t-1}$

Forecasting with Multiple Regression Models

- Specify (assume) alternative values for X's.
- Multiply Betas by their respective X's.
- Forecast Acres for alternative Prices and *CRP*.
- Lagged Yield and Year are constant in scenarios.

	C	D	E	F	G	H	I	J	K	L	M	N
92	Forecasting with alternative Values for Independent Variables											
93	Intercept	Years	Price t-1	CRP t	Yield t-1							
94	843.985	-0.412	9.237	-0.073	0.459	=M49						
95	Assumed Values for the Independent Variables					Forecast Values						
96	Year	Price t-1	CRP t	Yield t-1		Planted t						
97	2002	2.880	9.600	41.90		64.996	=$C$94+$D$94*C97+$E$94*D97+$F$94*E97+$G$94*F97					
98	2002	2.800	9.600	41.90		64.257	=$C$94+$D$94*C98+$E$94*D98+$F$94*E98+$G$94*F98					
99	2002	2.700	9.600	41.90		63.334	=$C$94+$D$94*C99+$E$94*D99+$F$94*E99+$G$94*F99					
100	2002	2.600	9.600	41.90		62.410	=$C$94+$D$94*C100+$E$94*D100+$F$94*E100+$G$94*F100					
101	2002	2.500	9.600	41.90		61.486	=$C$94+$D$94*C101+$E$94*D101+$F$94*E101+$G$94*F101					
102	2002	2.880	10.000	41.90		64.967	=$C$94+$D$94*C102+$E$94*D102+$F$94*E102+$G$94*F102					
103	2002	2.880	10.500	41.90		64.930	=$C$94+$D$94*C103+$E$94*D103+$F$94*E103+$G$94*F103					
104	2002	2.880	11.000	41.90		64.894	=$C$94+$D$94*C104+$E$94*D104+$F$94*E104+$G$94*F104					
105	2002	2.880	11.500	41.90		64.857	=$C$94+$D$94*C105+$E$94*D105+$F$94*E105+$G$94*F105					
106	2002	2.880	12.000	41.90		64.821	=$C$94+$D$94*C106+$E$94*D106+$F$94*E106+$G$94*F106					

Multiple Regression Forecast with Risk

We will begin probabilistic forecast using $\tilde{PA}_{t+i}$ and σ (Std. Dev) and assume a normal distribution for residuals.

$$\tilde{PA}_{t+i} = \hat{PA}_{t+i} + NORM(0, \sigma)$$

or

$$\tilde{PA}_{t+i} = \hat{PA}_{t+i} + \sigma * NORM(0, 1)$$

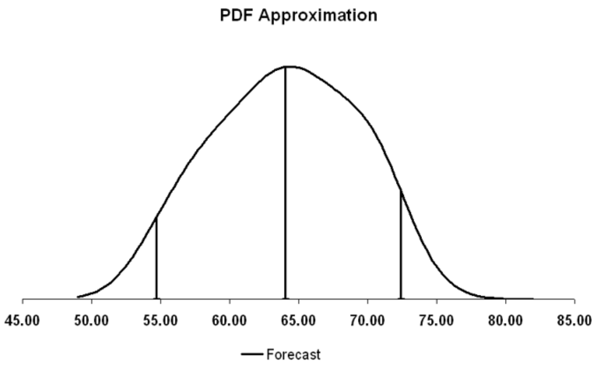
or

$$\tilde{PA}_{t+i} = NORM(\hat{PA}_{t+i}, \sigma)$$

	A	B	C	D	E	F	G
108	Intercept	Years	Price t-1	CRP t	Yield t-1		
109	843.985	-0.41164	9.23694	-0.07315	0.458552		
110	Assumed)	2003	2.88	9.6	41.1		
111	Forecasted		SE Predicted		Probabilistic		
112	Y-Hat for 2003		SEP for 2003		Forecast 2003		
113	64.218		5.141		57.044	=NORM(A113,C113)	
114	Formula to forecast determinisitc component						
115	=A109+B109*B110+C109*C110+D109*D110+E109*E110						

Multiple Regression Forecast with Risk

Present probabilistic forecast as a PDF with 95% Confidence Interval shown here as the bars about the mean for a probability density function (PDF).

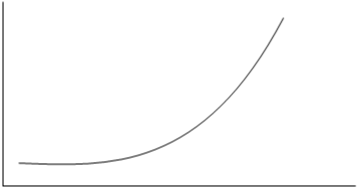


PDF Approximation	
	Forecast 2
Start	49
End	82
Band Widt	2.157497
Kernel	Gaussian
Confidenc	95.0%
Lower Qua	54.685
Average	64.045
Upper Qua	72.378

Regression Model for Growth

- Some data display a growth pattern.
- Easy to forecast with multiple regression
- Add T^2 variable to capture the growth or decay of Y variable.
- Growth function

$$Y = a + b_1 T + b_2 T^2$$



- Growth at decreasing rate

$$Y = a + b_1 Log(T)$$

Data & Results for Log Models

Single Log Form

$$Log(Y_t) = b_0 + b_1 T$$

Try the Single Log Form			95%	Intercept	Years
KOV	Log Y	Years	**Beta**	**184.039**	**-0.091**
110.0	4.70048	1970	**S.E.**	4.393	0.002
93.0	4.532599	1971	**t-test**	41.895	-41.176
80.0	4.382027	1972	**Prob(t)**	0.000	0.000
81.0	4.394449	1973	**Elasticity at Mean**		-57.184
68.0	4.219508	1974	**Variance Inflation Factor**		NA
55.0	4.007333	1975	**Partial Correlation**		NA
52.0	3.951244	1976	**Semipartial Correlation**		NA

Double Log Form

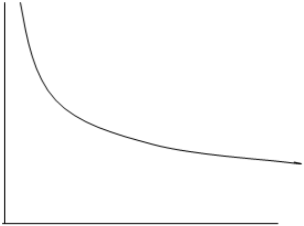
$$Log(Y_t) = b_0 + b_1 Log(T)$$

Double Log Transformation of the Data				95%	Intercept	Log X
Actual KO'	Years	Log Y	Log X	**Beta**	**1376.746**	**-180.887**
110.0	1970	4.70048	7.585788822	**S.E.**	33.222	4.375
93.0	1971	4.532599	7.586296307	**t-test**	41.441	-41.346
80.0	1972	4.382027	7.586803535	**Prob(t)**	0.000	0.000
81.0	1973	4.394449	7.587310506	**Elasticity at Mean**		-434.258

Regression Model For Decay Functions

- Some data display a decay pattern.
- There are various possible models for this situation
- One example: use an exogenous variable of the form $\frac{1}{T}$
- Decay function

$$Y = \beta_0 + \beta_1(\frac{1}{T}) + \beta_2(\frac{1}{T^2})$$



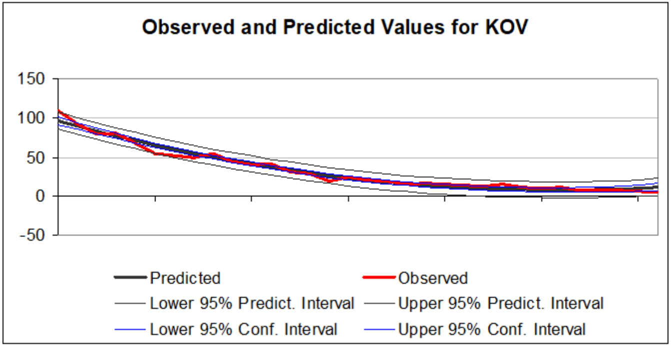
Forecasting Growth or Decay Patterns

Here is the regression result for estimating a decay function.

$$Y_t = \beta_0 + \beta_1 \left(\frac{1}{T}\right)$$

or

$$Y_t = \beta_0 + \beta_1 \frac{1}{T} + \beta_2 \frac{1}{T^2}$$



Multiple Regression Forecasts

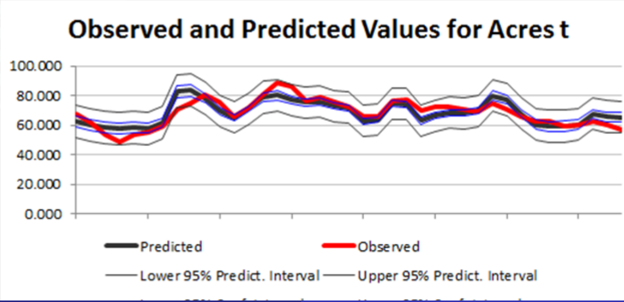
Example of a structural regression model that contains both a *Trend* and an independent X variable

$$Y = \beta_0 + \beta_1 T + \beta_2 X_t$$

This equation does not explain all of the variability, a seasonal or cyclical variability may be present, if so, you need to “remove” its effect.

OLS Regression Statistics for Acres t,

F-test	43.620	Prob(F)	0.000
MSE$^{1/2}$	5.099	CV Regr	7.446
R^2	0.708	Durbin-W	0.659
RBar^2	0.692	Rho	0.671
Akaike Inf	3.281	Goldfeld-(	4.614
Schwarz I	3.366		
95%	Intercept	EPrice t-1	Trend
Beta	805.020	9.758	-0.386
S.E.	162.842	1.045	0.083
t-test	4.944	9.339	-4.659
Prob(t)	0.000	0.000	0.000
Elasticity at Mean		0.424	-11.180
Variance Inflation Factor			
Partial Correlation			



Regression evaluation

- Make sure parameter signs are based on sound economic theory for all variables.
- Student t ratios greater than 1.96 (P values for betas < 0.05).
- F ratio larger than 20.0 and its P value < 0.05.
- Add explanatory elements that increase $\bar{R}^2$ by a non-trivial amount.

Goodness of Fit Measures

Models with high R^2 may not forecast well. *Every* time we add a new X, we will get some increase in R^2.

$$R^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2}{\sum_{t=1}^{T}(Y_t - \bar{Y})^2}$$

$\bar{R}^2$ is preferred as it is not affected by no. of Xs.

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k}$$

We might also use mean absolute percentage error (MAPE) for model selection. MAPE is used to measure the accuracy of a forecasting measure and usually the accuracy is expressed in terms of percentage.

$$MAPE = \frac{100\%}{n} \sum_{t=1}^{T} \left| \frac{Y_t - \hat{Y}_t}{Y_t} \right|$$

Information Criteria

For a residual sum of squares RSS, a number of right-hand-side variables k, and a number of observations n, Akaike Information Criterion and Bayesian Information Criterion are calculated as follows:

- Akaike Information Criterion (AIC)

$$AIC = \ln\left(\frac{RSS}{n}\right) + \frac{2(k+1)}{n}$$

- Bayesian Information Criterion (BIC), sometimes called Schwartz Information Criterion (SIC) or Schwarz Bayesian Criterion (SBC)

$$BIC = \ln\left(\frac{RSS}{n}\right) + \frac{k}{n} ln(n)$$

- When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. The AIC & BIC resolve this problem by introducing a penalty term for the number of parameters in the model. The penalty term is larger in BIC than in AIC.

Summary of Goodness of Fit Measures

- *MAPE* works best to determine model for "in sample" forecasting.
- R^2 does not penalize for adding Xs.
- $\bar{R}^2$ provides some penalty as k increases.
- *AIC* is better than R^2 but *BIC* results in the most parsimonious models (fewest Xs).

Black Swans (BSs)

- Black Swans are low probability events.
  - An outlier, which is "outside realm of reasonable expectations."
  - Carries an extreme impact.
  - Human nature causes us to concoct explanations.
- Black swans are an example of uncertainty
  - Uncertainty is generated by unknown probability distributions.
  - Risk is generated by "known" distributions.
- 1917 influenza pandemic would be a black swan
- COVID-19 pandemic
- War outbreak
- Sudden policy change
- We will discuss more about incorporating uncertainty in the model when we start simulating non-parametric distributions.