Regression Models

Taylor R. Brown

Last class

Last class we finished off modeling individual univariate time series.

It's also popular in finance to model one time series of returns in terms of others. For this we use **regression.**

These models aren't always used to forecast the future. Most of the time they try to explain the tradeoff between risk/return of a potential investment.

Simple Linear Regression

Say we are considering an investment in the technology etf `XLK`.

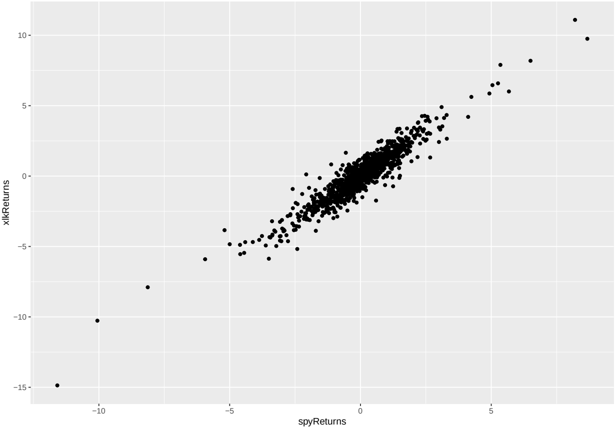
```
##            spyReturns xlkReturns
## 2018-06-20  0.1704472  0.2097876
## 2018-06-21 -0.6288188 -0.7713275
## 2018-06-22  0.1821467 -0.3243290
## 2018-06-25 -1.3706340 -2.0981394
## 2018-06-26  0.2211345  0.4030410
## 2018-06-27 -0.8318637 -1.3741332
```

Does it "beat" the overall market?

Simple Linear Regression

A rising tide lifts all boats...

```
## Warning: Removed 10 rows containing missing values (`geo
```



Simple Linear Regression

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$$

- Y_t: return at time t of XLK
- X_t: return at time t of SPY
- ϵ_t: unexplainable error at time t (typically normally distributed)
- β_0, β_1 coefficients

Simple Linear Regression

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$$

- Financial textbooks will often write the model as $Y_t = \alpha + \beta X_t + \epsilon_t$, and you'll hear people talking about the α and β of funds.
- single index model (Sharpe 1963): $r_{it} - r_f = \alpha + \beta(r_{m,t} - r_f) + \epsilon_{i,t}$
- capital asset pricing model (CAPM): $r_{it} - r_f = 0 + \beta(r_{m,t} - r_f) + \epsilon_{i,t}$
- with daily data, r_f is so small it almost doesn't matter
- whether or not $\alpha > 0$ is very important to know because this demonstrates a fund has an "edge" or some "secret sauce"
- β_1 is important to know because it's a measure of how "exposed" you are to the "risk" of the overall market

Simple Linear Regression

"Beta" (i.e. β_1) cuts both ways...

$$\begin{aligned} E[Y_t \mid X] &= E[\beta_0 + \beta_1 X + \epsilon \mid X_t] \\ &= \beta_0 + \beta_1 X + E[\epsilon \mid X_t] \\ &= \beta_0 + \beta_1 X_t \end{aligned}$$

- Typically $\beta_1 > 0$ so both move in the same direction and are positively correlated
- If $\beta_1 > 1$, then it magnifies both wins and losses
- If $0 < \beta_1 < 1$, then Y_t is in some sense "safer" (really less wiggly) than investing in the overall market X_t.
- $\beta_1 < 0$ is rare, but this might happen with an inverse ETF or short-seller fund.

Simple Linear Regression

Nonzero "alpha" (i.e. β_0) is fantastic but sometimes hard to find

$$\begin{aligned} E[Y_t \mid X] &= E[\beta_0 + \beta_1 X + \epsilon \mid X_t] \\ &= \beta_0 + \beta_1 X + E[\epsilon \mid X_t] \\ &= \beta_0 + \beta_1 X_t \end{aligned}$$

You get the same boost to the expected return of Y, no matter what X is

$$E[Y_t] = E[E(Y_t \mid X)] = \beta_0 + \beta_1 E[X_t]$$

Simple Linear Regression

X_t doesn't explain all the variability of Y_t...

$$\begin{aligned} V[Y_t \mid X_t] &= V[\beta_0 + \beta_1 X_t + \epsilon \mid X_t] \\ &= V[\epsilon_t \mid X_t] \\ &= \sigma^2 \end{aligned}$$

Simple Linear Regression

β_1 magnifies/attenuates the variance.

$$
\begin{aligned}
V[Y_t] &= E[V(Y_t|X_t)] + V[E(Y_t|X_t)] \\
&= \sigma^2 + V[\beta_0 + \beta_1 X_t] \\
&= \sigma^2 + \beta_1^2 V[X_t]
\end{aligned}
$$

$\sqrt{V[Y_t]}$ is often called **volatility**

Simple Linear Regression

β_1 also gives you a **hedge ratio**

$$Y_t - \beta_1 X_t = \beta_0 + \epsilon_t$$

$$V[Y_t - \beta_1 X_t] = \sigma^2$$

And σ^2 is the "risk" you can't diversify/hedge away.

TODO make this a homework problem that shows this minimizes variance.

Simple Linear Regression

```r
mod1 <- lm(xlkReturns ~ spyReturns, data = xlkSpy)
#summary(mod1)
```

Simple Linear Regression

Right now the 3 month annualized TBill rate is 4.5%, which equates to a daily rate of 0.0121%. So alternatively we could do

```
transformedData <- xlkSpy - 0.0121
# we scaled all returns by 100, otherwise it's .000121
mod2 <- lm(xlkReturns ~ spyReturns, data = transformedData)
#summary(mod2)
```

Simple Linear Regression

If you believe the intercept had to be zero, then

```
mod3 <- lm(xlkReturns ~ 0 + spyReturns, data = transformedI
#summary(mod3)
```

An even better risk-free rate adjustment would involve going back in time and finding all the time-varying risk-free rates.

Estimates (without subtracting risk free rate)

```
##
## Call:
## lm(formula = xlkReturns ~ spyReturns, data = xlkSpy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.48707 -0.34613  0.01014  0.33293  2.63773
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.01813    0.01750   1.036      0.3
## spyReturns   1.22362    0.01268  96.535   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
##
## Residual standard error: 0.6036 on 1188 degrees of free
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.8869, Adjusted R-squared:  0.886
```

Estimates (with subtracting risk free rate)

```
##
## Call:
## lm(formula = xlkReturns ~ spyReturns, data = transformed
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.48707 -0.34613  0.01014  0.33293  2.63773
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02084    0.01750   1.191    0.234
## spyReturns   1.22362    0.01268  96.535   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
##
## Residual standard error: 0.6036 on 1188 degrees of freed
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.8869, Adjusted R-squared:  0.886
```

Estimates (with subtracting risk free rate)

```
##
## Call:
## lm(formula = xlkReturns ~ 0 + spyReturns, data = transfo
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.46640 -0.32539  0.03081  0.35486  2.65913
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## spyReturns  1.22389    0.01268   96.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 0.6037 on 1189 degrees of freed
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.8869, Adjusted R-squared:  0.8868
## F-statistic:  9323 on 1 and 1189 DF,  p-value: < 2.2e-16
```

Sources:

Chapter 2.2,2.3,2.4 of Introduction to Time Series and Forecasting
Brockwell/Davis