# ØAMET2200
# Business Decision Making Using Data
## Lecture 8

Instructor: Fenella Carpena

October 25, 2019

# Announcements

- Problem Set # 3 has been posted
  - Due on Nov. 3 at 11:59PM
- Office Hours (OH) Schedule
  - Moved to Mondays, 4-5PM (Room PE737, Pilestredet 35)
  - No longer on Tuesdays and Fridays
- Schedule for remaining lectures
  - Lecture 9, Nov. 1 (Chapter 25, Categorical Explanatory Vars.)
  - Lecture 10, Nov. 8 (Chapter 27, Time Series)
  - Lecture 11, Nov. 22 (Practice Mini-Final Exam)
- Suggested practice problems from textbook posted on Canvas

# Part 4 of this Course: Building regression models

- You want to build a model to forecast outcomes from data you have
- What's the best model you can build?
- This class: When should you add variables to your model?
- Next week: Categorical explanatory variables

# Additional variables: What is the trade-off?

- Goal of this class: highlight the power of statistical models, but also illustrate potential risks
- Why not just add variables?
  - More variables mean a higher $R^2$
  - Explain more variation in $y$
  - So much data is available now, shouldn't we always just use whatever variables we have?

# Additional variables: What is the trade-off?

- Adding variables to the model is not always "better," there are potential problems too
- Does your model make sense? Can you explain it to customers or clients?
- Even with large datasets there is limited variation for your model to use to estimate a large number of slopes
  - Starbucks example from last lecture: what is the effect of the number of competitors on sales, holding fixed the median income in the area, the size of the store, the quality of the products, the quality of customer service, amount spent on advertising...
- Data mining and "multiple testing"

# Agenda for Today

- Building a multivariate regression model
- Collinearity
- Data mining: the good, the bad, and the ugly
- Chapter 24

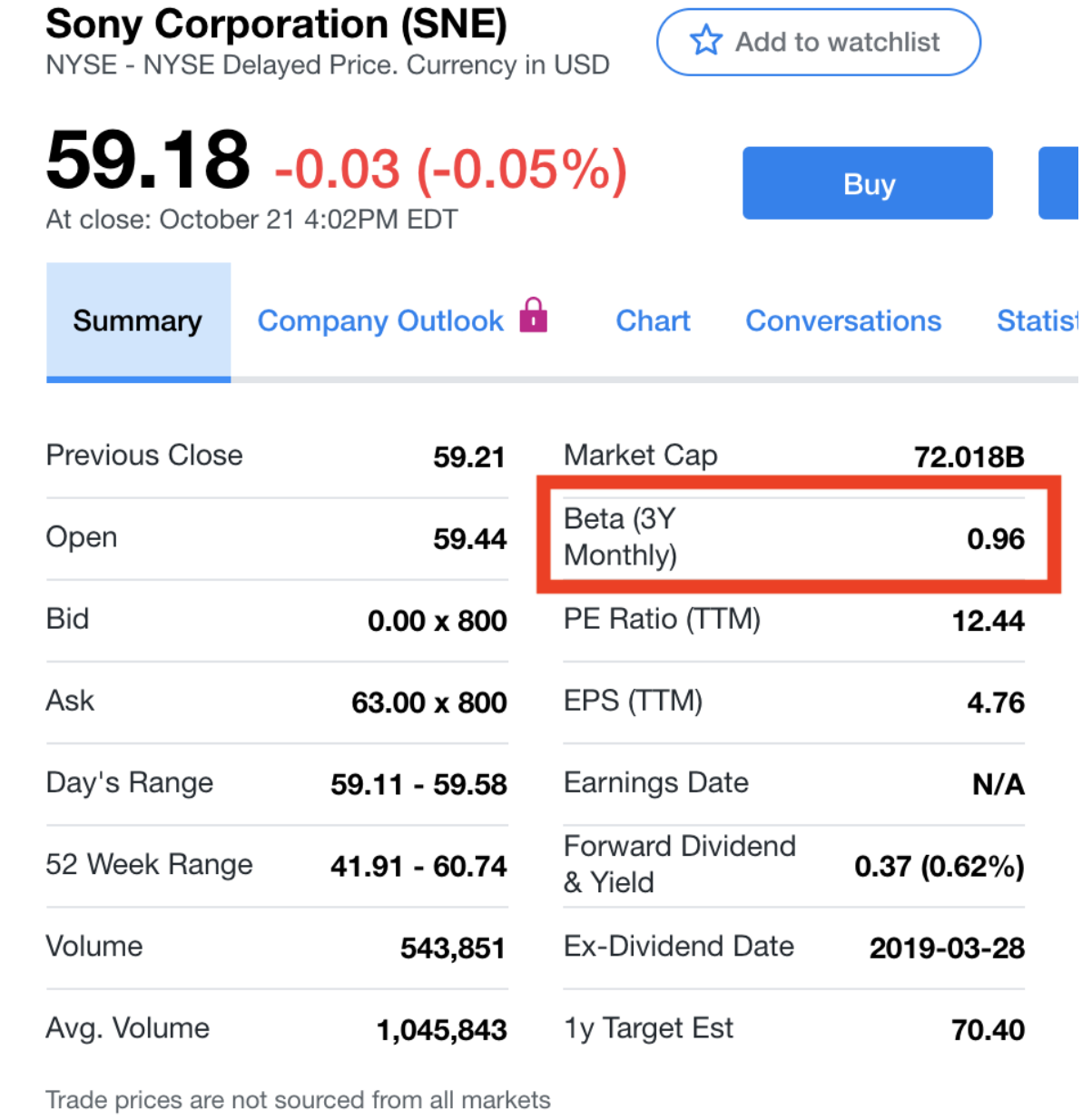
# Case Study for Today: Sony Stock

- We want to build a model that describes returns on stock in Sony Corporation
- A good place to start is the Capital Asset Pricing Model (CAPM), a model of financial markets
- The CAPM describes the relationship between returns on a **particular stock** and returns on the **whole stock market**
- According to the underlying theory, the market rewards investors for taking unavoidable risks, collectively known as **market risk**
- Other risks, collectively called **idiosyncratic risk**, are avoidable, and the CAPM model promises no compensation for these

# Case Study for Today: Sony Stock

- The CAPM can be represented as a simple regression (SRM)
- The response variable $y$: the percentage change in a Sony's value over some time period
- The explanatory variable $x_1$: the percentage change in the value of the whole stock market
- Finance analysts traditionally denote the intercept in this regression as **"alpha"** and the slope as **"beta"**
- "Beta" provides a measure of the degree of correlation between that stock and the rest of the market
- "Alpha" provides a measure of abnormal or excess return (e.g., ability to beat the market)

# Case Study for Today: Sony Stock



**Sony Corporation (SNE)**
NYSE - NYSE Delayed Price. Currency in USD

Add to watchlist

**59.18** -0.03 (-0.05%)
At close: October 21 4:02PM EDT

Buy

Summary | Company Outlook | Chart | Conversations | Statist

| | | | |
|---|---|---|---|
| Previous Close | 59.21 | Market Cap | 72.018B |
| Open | 59.44 | Beta (3Y Monthly) | 0.96 |
| Bid | 0.00 x 800 | PE Ratio (TTM) | 12.44 |
| Ask | 63.00 x 800 | EPS (TTM) | 4.76 |
| Day's Range | 59.11 - 59.58 | Earnings Date | N/A |
| 52 Week Range | 41.91 - 60.74 | Forward Dividend & Yield | 0.37 (0.62%) |
| Volume | 543,851 | Ex-Dividend Date | 2019-03-28 |
| Avg. Volume | 1,045,843 | 1y Target Est | 70.40 |

Trade prices are not sourced from all markets

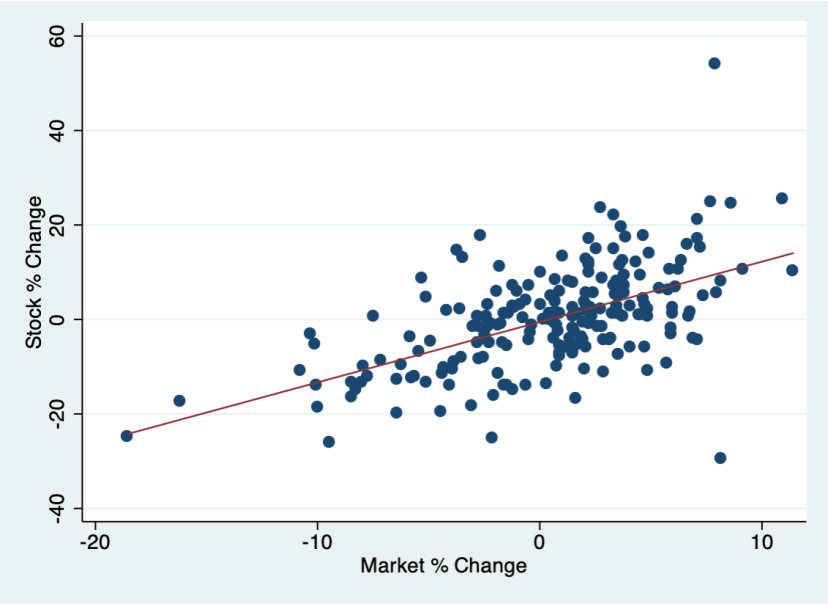
# Case Study for Today: Sony Stock

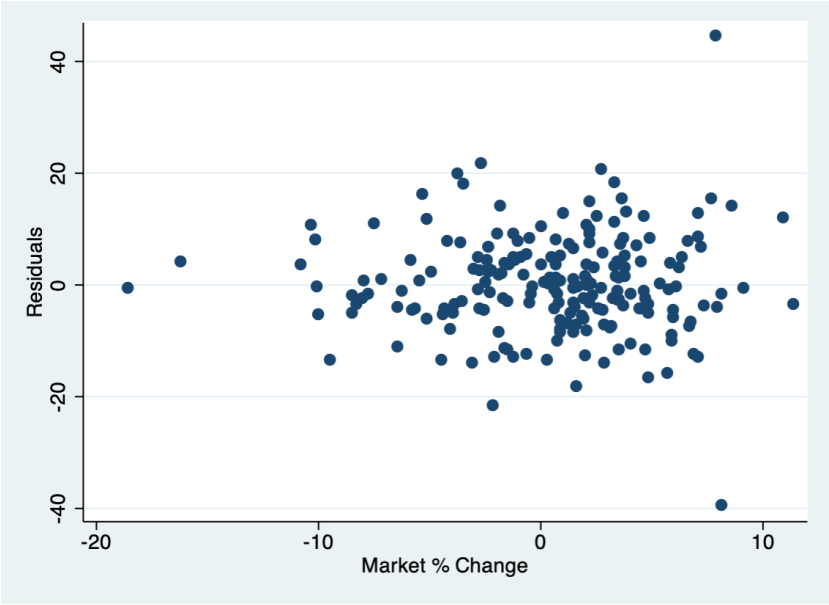
- Questions we'll analyze for our case study
  - What is the relationship between returns on Sony stock and returns in the broader market?
  - What is the "beta" of Sony stock?
- Why do we care?
  - Hedging strategy: Does Sony move 1-for-1 with the market?
  - Evaluating Sony's performance: How is Sony's share price performing relative to the market as a whole?

# Sony Stock: The Dataset

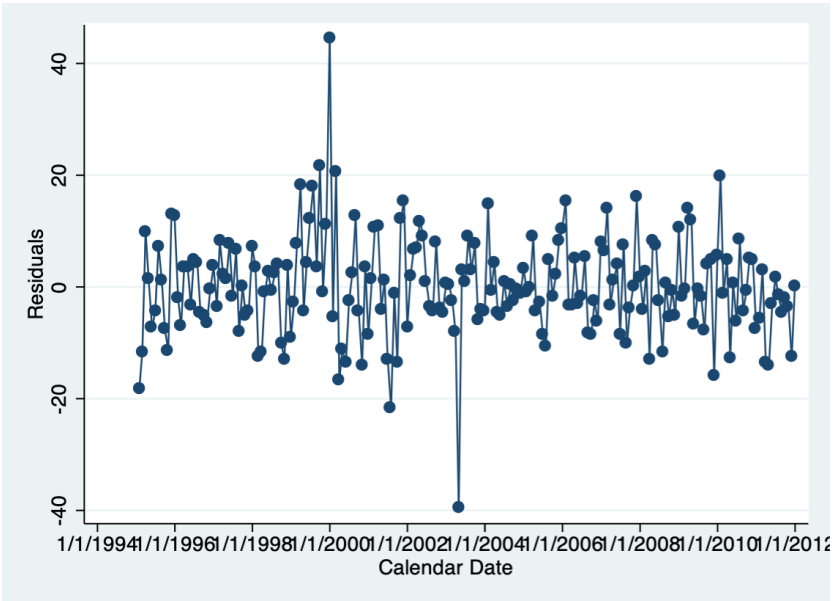
- You have 17 years of monthly data ($n = 204$) from Sony and from the rest of the U.S. stock market
- Response variable: Monthly percentage change in value of Sony stock (`SonyChange`)
- Explanatory variables:
  - Monthly percentage change in U.S. stock market (`MarketChange`)
  - Monthly percentage change in Dow Jones Industrial Average (30 largest companies) (`DowChange`)
  - Monthly percentage difference in returns between small and large companies (`SmallBig`)
  - Monthly percentage difference in returns between growth and value stocks (`HighLow`)

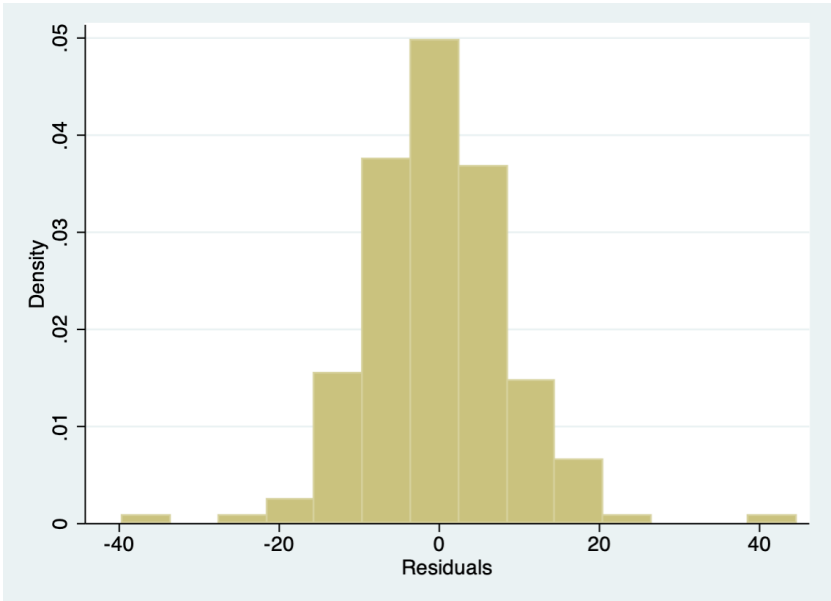
# Sony Stock: SRM using CAPM as starting point



# Sony Stock: SRM using CAPM as starting point



# Sony Stock: SRM using CAPM as starting point

```
. reg SonyChange MarketChange ;
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 7600.26565 | 1 | 7600.26565 |
| Residual | 16350.3952 | 202 | 80.9425503 |
| Total | 23950.6608 | 203 | 117.983551 |

| | |
|---|---|
| Number of obs | 204 |
| F( 1, 202) | 93.90 |
| Prob > F | 0.0000 |
| R-squared | 0.3173 |
| Adj R-squared | 0.3140 |
| Root MSE | 8.9968 |

| SonyChange | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| MarketChange | 1.277969 | .1318847 | 9.69 | 0.000 | 1.017921 | 1.538016 |
| _cons | -.5316643 | .633765 | -0.84 | 0.403 | -1.781308 | .7179791 |

# Sony Stock: SRM using CAPM as starting point

- How do you interpret $b_0$ (Sony stock's alpha)?
- How do you interpret $b_1$ (Sony stock's beta)?
- Is the estimate $b_1$ statistically significant at the 1% level?
- Is the two-sided test $H_0 : \beta_1 = 0$, $H_1 = \beta_1 \neq 0$ very interesting?

# Sony Stock: SRM using CAPM as starting point

- What is the $p$-value of the hypothesis test $H_0 : \beta_1 \leq 1,\ H_1 : \beta_1 > 1$?

# Sony Stock: Expanding the regression model

- We want to improve our estimate of $\beta_1$ by adding more $x$'s
- How to pick additional $x$'s?
  - Naive approach: add "more of the same" variables
  - Exploring the data and learning more about the problem
  - Theoretical knowledge
- Three additional variables we'll consider for our case study
  - `DowChange`
  - `SmallBig`
  - `HighLow`

# Sony Stock: Expanding the regression model

```
. correlate SonyChange MarketChange Dow SmallBig HighLow ;
(obs=204)
```

| | SonyCh~e | MarketC~ | DowCha~e | SmallBig | HighLow |
|---|---|---|---|---|---|
| SonyChange | 1.0000 | | | | |
| MarketChange | 0.5633 | 1.0000 | | | |
| DowChange | 0.4600 | 0.9068 | 1.0000 | | |
| SmallBig | 0.3665 | 0.2557 | 0.0022 | 1.0000 | |
| HighLow | -0.2407 | -0.2334 | -0.0492 | -0.3622 | 1.0000 |

# Sony Stock: Multiple Regression

```
. regress SonyChange MarketChange Dow SmallBig HighLow ;
```

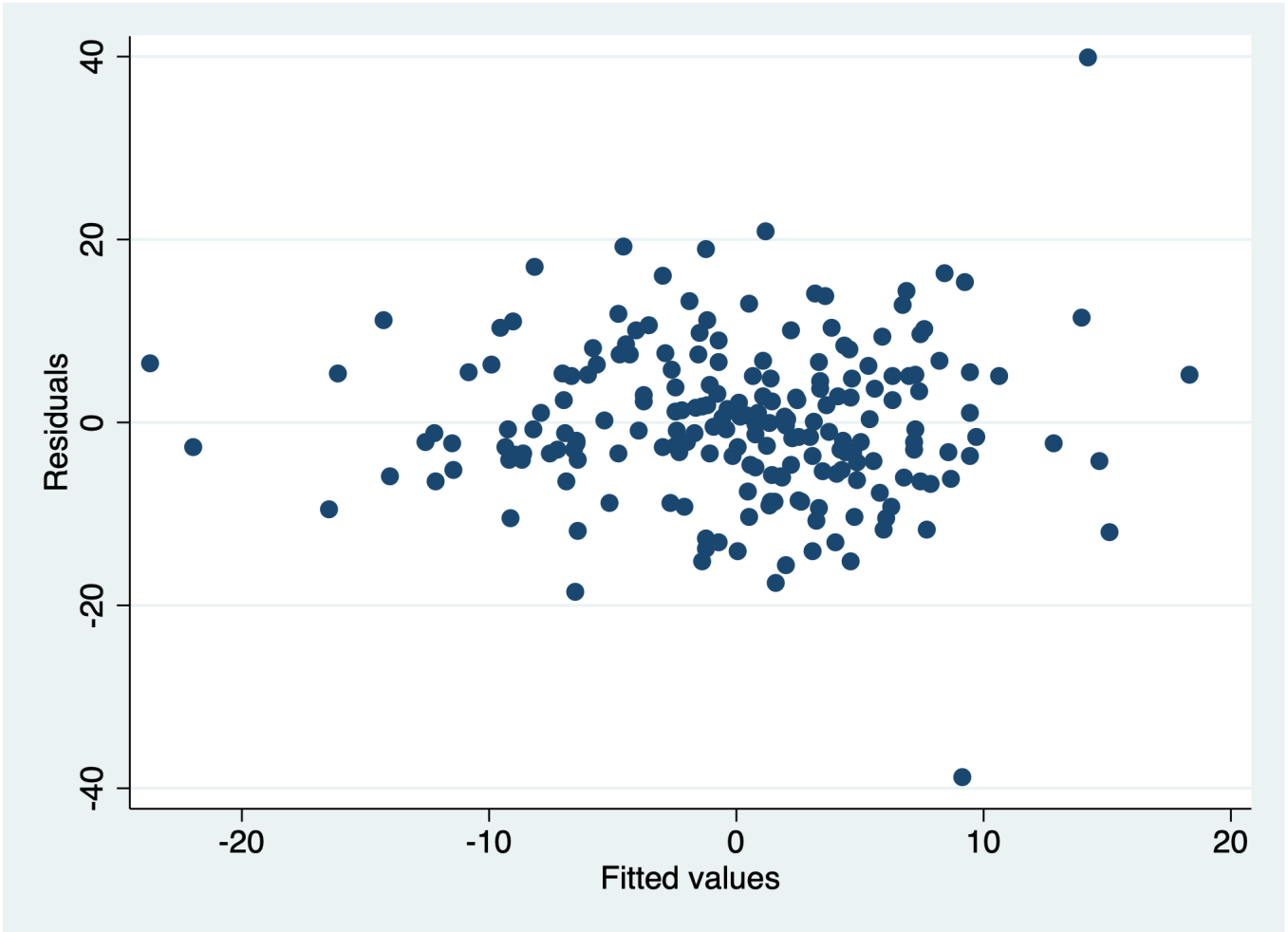
| Source | SS | df | MS |
|---|---|---|---|
| Model | 8920.48983 | 4 | 2230.12246 |
| Residual | 15030.171 | 199 | 75.5284974 |
| Total | 23950.6608 | 203 | 117.983551 |

| | |
|---|---|
| Number of obs | = 204 |
| F( 4, 199) | = 29.53 |
| Prob > F | = 0.0000 |
| R-squared | = 0.3725 |
| Adj R-squared | = 0.3598 |
| Root MSE | = 8.6907 |

| SonyChange | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| MarketChange | .970782 | .3979745 | 2.44 | 0.016 | .1859936 | 1.75557 |
| DowChange | .1667321 | .4057335 | 0.41 | 0.682 | -.6333566 | .9668209 |
| SmallBig | .7028111 | .209041 | 3.36 | 0.001 | .2905914 | 1.115031 |
| HighLow | -.159028 | .1974277 | -0.81 | 0.421 | -.5483469 | .2302909 |
| _cons | -.5625448 | .6170229 | -0.91 | 0.363 | -1.779287 | .6541974 |

# Sony Stock: Check MRM conditions

Scatter plot of residuals vs. predicted values



# Sony Stock: Comparing MRM vs. SRM using CAPM

- Does the MRM model have explanatory power?
- Does the MRM explain more variation in Sony stock than the CAPM regression?
- Which variables statistically significantly improve the fit of the MRM?
- How did the standard error for the coefficient on `MarketChange` change? Why?

# Collinearity: Definition and Implications

- **Collinearity** is the situation in which two or more of the explanatory variables are highly correlated
- Some correlation between explanatory variables is inevitable
- However, when your explanatory variables are **highly correlated** this **decreases the precision of the estimates**
  - In particular, standard errors tend to increase, and there may be large changes in the size of the estimates
  - Intuition: MRM estimates the effect of `MarketChange`, holding `DowChange` and other variables constant
- Collinearity **does not** violate MRM assumptions, statistical inference/hypothesis tests are still valid
  - But makes it harder to interpret the regression

# Collinearity: Example of **Perfect** Collinearity

```
. regress SonyChange MarketChange Dow SmallBig BigSmall HighLow ;
note: BigSmall omitted because of collinearity
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 8920.48983 | 4 | 2230.12246 |
| Residual | 15030.171 | 199 | 75.5284974 |
| Total | 23950.6608 | 203 | 117.983551 |

| | |
|---|---|
| Number of obs | 204 |
| F( 4, 199) | 29.53 |
| Prob > F | 0.0000 |
| R-squared | 0.3725 |
| Adj R-squared | 0.3598 |
| Root MSE | 8.6907 |

| SonyChange | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| MarketChange | .970782 | .3979745 | 2.44 | 0.016 | .1859936 | 1.75557 |
| DowChange | .1667321 | .4057335 | 0.41 | 0.682 | -.6333566 | .9668209 |
| SmallBig | .7028111 | .209041 | 3.36 | 0.001 | .2905914 | 1.115031 |
| BigSmall | 0 | (omitted) | | | | |
| HighLow | -.159028 | .1974277 | -0.81 | 0.421 | -.5483469 | .2302909 |
| _cons | -.5625448 | .6170229 | -0.91 | 0.363 | -1.779287 | .6541974 |

# Collinearity: Dealing with Collinearity

**Signs of Collinearity**

- High correlation between the explanatory variables
- $R^2$ doesn't change much as you add a variable
- Standard errors increase when you add a variable
- Coefficient estimates change dramatically when you add a variable
- Large $F$-statistic but many small $t$-statistics

**What to do if there is collinearity?**

- Do you have a good reason to keep the $x$'s in the model?
- If not, consider dropping them from the model, one at a time

# Collinearity: What is the "best" model?

Typically combines intuition, theory, and model testing.

1. What are candidate hypothesis?
2. Does the model have predictive power with all of these explanatory variables? ($F$-stat)
3. Does each $x$ variable have statistical significance? ($t$-stat)
   - If yes, great!
   - If no, think whether to include or exclude the variable from the regression (judgement call).

     By excluding, you are forcing the slope to be 0.

     By including, you may be adding "noise" to the model.
4. If excluding a variable, do so one at a time. Then start back at Step 1.

# Best Practices

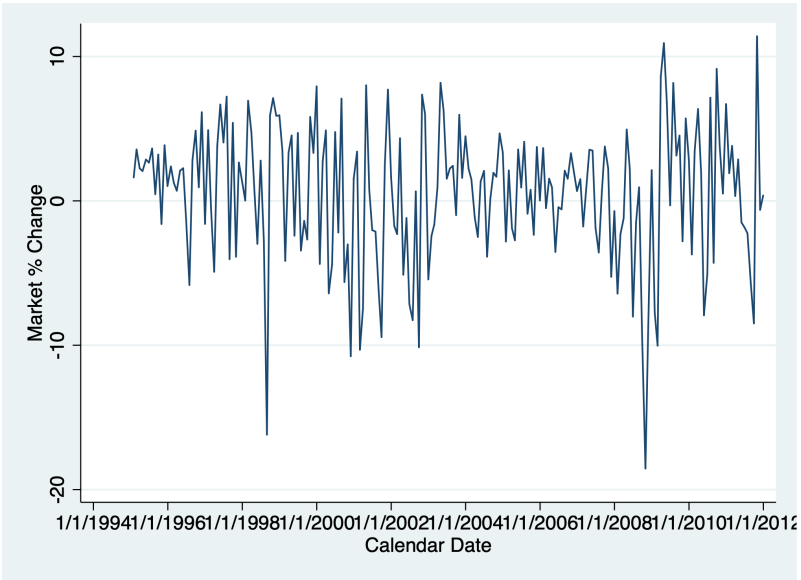
- Begin a regression analysis by looking at plots
- Remember to check the F-statistic before interpreting regression results
- Learn to recognize the presence of collinearity
- Do not remove explanatory variables at the first sign of collinearity
- Remove explanatory variables one at a time
- Remember: Regression modeling is iterative
  - Start with a model that makes sense theoretically and intuitively
  - Then refine and improve the fit of the model

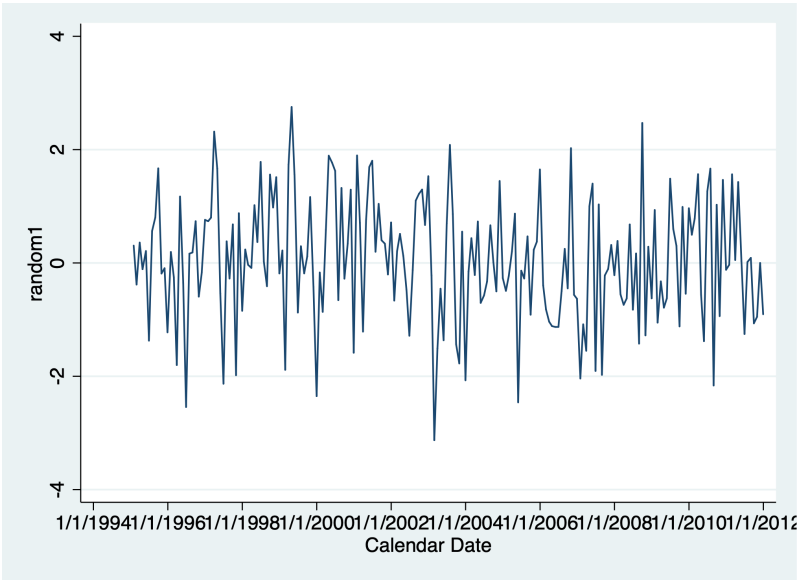
# Automated Model Selection

- Tools for automating model selection exists (beyond the scope of this course)
  - Stepwise regression
  - Ridge regression
  - LASSO (Least Absolute Shrinkage and Selection Operator)
  - Principal Component Analysis
- These methods are not without problems
  - Data Mining (looking for a "significant" explanatory variable)
  - Overfitting
  - High "False Discovery Rates"
  - They cannot account for common sense
- And some of these problems also have solutions
  - Training, Cross-Validation, Testing

# Example of Data Mining

Can you "explain" changes in the stock market with random numbers?



# Example of Data Mining

## Regression of `MarketChange` on 25 variables with random numbers

```
. regress MarketChange random1-random25 ;
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 537.689657 | 25 | 21.5075863 |
| Residual | 4115.90311 | 178 | 23.1230512 |
| Total | 4653.59277 | 203 | 22.9241023 |

| | |
|---|---|
| Number of obs | = 204 |
| F( 25, 178) | = 0.93 |
| Prob > F | = 0.5638 |
| R-squared | = 0.1155 |
| Adj R-squared | = -0.0087 |
| Root MSE | = 4.8086 |

| MarketChange | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| random1 | -.0555152 | .326365 | -0.17 | 0.865 | -.6995577 | .5885273 |
| random2 | .2101161 | .3542982 | 0.59 | 0.554 | -.4890492 | .9092814 |
| random3 | .4173678 | .3420193 | 1.22 | 0.224 | -.2575665 | 1.092302 |
| random4 | -.2239637 | .3462961 | -0.65 | 0.519 | -.9073378 | .4594104 |
| random5 | -.5583642 | .3283165 | -1.70 | 0.091 | -1.206258 | .0895293 |
| random6 | .0666732 | .3685756 | 0.18 | 0.857 | -.6606668 | .7940132 |
| random7 | -.8351073 | .3992565 | -2.09 | 0.038 | -1.622993 | -.0472221 |
| random8 | -.2658548 | .3903096 | -0.68 | 0.497 | -1.036084 | .5043748 |
| random9 | -.5863851 | .4023789 | -1.46 | 0.147 | -1.380432 | .2076618 |
| random10 | -.1411426 | .3698135 | -0.38 | 0.703 | -.8709254 | .5886402 |
| random11 | .1065831 | .3528114 | 0.30 | 0.763 | -.5896483 | .8028144 |
| random12 | -.3951103 | .3678266 | -1.07 | 0.284 | -1.120972 | .3307516 |
| random13 | -.370126 | .3706793 | -1.00 | 0.319 | -1.101618 | .3613655 |
| random14 | -.4033955 | .3594233 | -1.12 | 0.263 | -1.112674 | .3058835 |
| random15 | .1361262 | .3745083 | 0.36 | 0.717 | -.6029214 | .8751738 |
| random16 | -.0771639 | .3169369 | -0.24 | 0.808 | -.7026012 | .5482734 |
| random17 | .2676346 | .3551067 | 0.75 | 0.452 | -.4331262 | .9683953 |
| random18 | .294177 | .3517137 | 0.84 | 0.404 | -.3998882 | .9882421 |
| random19 | -.0046322 | .3564877 | -0.01 | 0.990 | -.7081183 | .6988539 |
| random20 | .1232889 | .3377879 | 0.36 | 0.716 | -.5432953 | .7898731 |
| random21 | .909352 | .3887375 | 2.34 | 0.020 | .1422248 | 1.676479 |
| random22 | -.002114 | .3866711 | -0.01 | 0.996 | -.7651633 | .7609353 |
| random23 | -.3717811 | .3745569 | -0.99 | 0.322 | -1.110925 | .3673623 |
| random24 | .3257152 | .3466641 | 0.94 | 0.349 | -.3583851 | 1.009816 |
| random25 | -.0127162 | .355392 | -0.04 | 0.971 | -.7140401 | .6886076 |
| _cons | .5222358 | .3487762 | 1.50 | 0.136 | -.1660325 | 1.210504 |

# Example of Data Mining

Regression of `MarketChange` on `random21`

```
. regress MarketChange random21 ;
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 76.9151382 | 1 | 76.9151382 |
| Residual | 4576.67763 | 202 | 22.65682 |
| Total | 4653.59277 | 203 | 22.9241023 |

| | |
|---|---|
| Number of obs | 204 |
| F( 1, 202) | 3.39 |
| Prob > F | 0.0669 |
| R-squared | 0.0165 |
| Adj R-squared | 0.0117 |
| Root MSE | 4.7599 |

| MarketChange | Coef. | Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| random21 | .6517086 | .3537098 | 1.84 | 0.067 | -.0457283 | 1.349146 |
| _cons | .4613853 | .3353188 | 1.38 | 0.170 | -.1997886 | 1.122559 |

- High $F$-stat! High $t$-stat! Great model?
- Is this likely to predict changes in the stock market in the future? Why or why not?

# Preventing "bad" data mining

- Always check the F-statistic before interpreting regression results
- Be suspicious of models where explanatory variables are too "hand picked" or arbitrary
- Your results need to make sense: regardless of how large your $t$-statistic is, you need a plausible theory
  - Use your intuition and expert knowledge about economics