THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Inference for simple regression 2

Lecture 4

STA 371G

In finance, the β of an asset indicates its volatility relative to the market. An asset with:

In finance, the β of an asset indicates its volatility relative to the market. An asset with:

- $\beta = 1$ rises and falls with the market as a whole.

In finance, the β of an asset indicates its volatility relative to the market. An asset with:

- $\beta = 1$ rises and falls with the market as a whole.
- $\beta > 1$ is **more** volatile than the market as a whole.

In finance, the β of an asset indicates its volatility relative to the market. An asset with:

- $\beta = 1$ rises and falls with the market as a whole.
- $\beta > 1$ is **more** volatile than the market as a whole.
- $\beta < 1$ is **less** volatile than the market as a whole.

In finance, the β of an asset indicates its volatility relative to the market. An asset with:

- $\beta = 1$ rises and falls with the market as a whole.
- $\beta > 1$ is **more** volatile than the market as a whole.
- $\beta < 1$ is **less** volatile than the market as a whole.

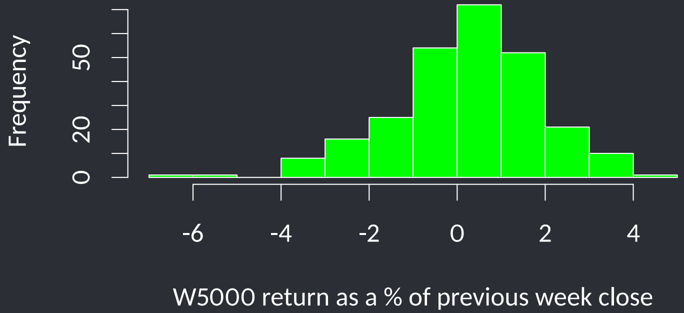
In finance, the β of an asset indicates its volatility relative to the market. An asset with:

- $\beta = 1$ rises and falls with the market as a whole.
- $\beta > 1$ is **more** volatile than the market as a whole.
- $\beta < 1$ is **less** volatile than the market as a whole.

β is just the slope of the regression line (i.e. $\hat{\beta}_1$) when we regress the asset's weekly returns against the weekly returns of a market index.

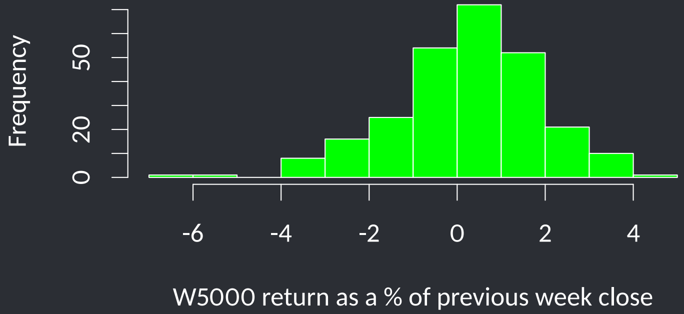
W5000 (Wilshire 5000, a broad market index)

```
> hist(stock.market$W5000, col='green',
+   main='', xlab='W5000 return as a % of previous week close')
```



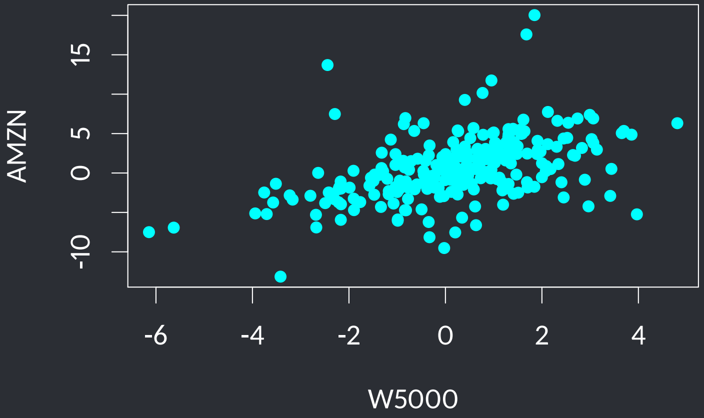
W5000 (Wilshire 5000, a broad market index)

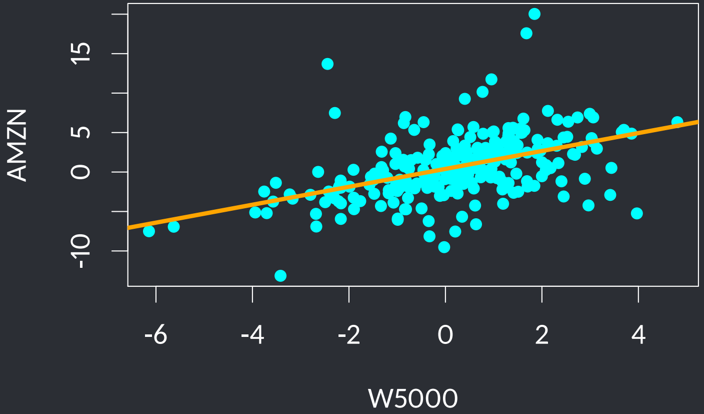
```
> hist(stock.market$W5000, col='green',
+   main='', xlab='W5000 return as a % of previous week close')
```



Amazon (AMZN)

```
> plot(AMZN ~ W5000, data=stock.market,
+   pch=16, col='cyan')
```





The regression line is

$$\widehat{\mathrm{AMZN}} = 0.4 + 1.13 \cdot \mathrm{W5000},$$

with $R^2 = 0.22$ and $p = 7.8 \times 10^{-16}$.

Interpreting the regression statistics

- $\hat{\beta}_1 = 1.13$ ("β") is the predicted increase in returns for AMZN when W5000 returns increase by 1 percentage point—since this is > 1 AMZN will swing more than the market as a whole

Interpreting the regression statistics

- $\hat{\beta}_1 = 1.13$ ("β") is the predicted increase in returns for AMZN when W5000 returns increase by 1 percentage point—since this is > 1 AMZN will swing more than the market as a whole
- $R^2 = 0.22$ indicates how closely AMZN tracks W5000 (the market as a whole)

Interpreting the regression statistics

- $\hat{\beta}_1 = 1.13$ ("β") is the predicted increase in returns for AMZN when W5000 returns increase by 1 percentage point—since this is > 1 AMZN will swing more than the market as a whole
- $R^2 = 0.22$ indicates how closely AMZN tracks W5000 (the market as a whole)
- $p = 7.8 \times 10^{-16}$ tells us whether we can reject the null hypothesis that AMZN does not move with the market at all

Interpreting the regression statistics

- $\hat{\beta}_1 = 1.13$ ("β") is the predicted increase in returns for AMZN when W5000 returns increase by 1 percentage point—since this is > 1 AMZN will swing more than the market as a whole
- $R^2 = 0.22$ indicates how closely AMZN tracks W5000 (the market as a whole)
- $p = 7.8 \times 10^{-16}$ tells us whether we can reject the null hypothesis that AMZN does not move with the market at all

Interpreting the regression statistics

- $\hat{\beta}_1 = 1.13$ ("β") is the predicted increase in returns for AMZN when W5000 returns increase by 1 percentage point—since this is > 1 AMZN will swing more than the market as a whole
- $R^2 = 0.22$ indicates how closely AMZN tracks W5000 (the market as a whole)
- $p = 7.8 \times 10^{-16}$ tells us whether we can reject the null hypothesis that AMZN does not move with the market at all (we can! since p is small)

Simple regression assumptions

We need four things to be true for statistical inference (i.e., hypothesis tests, *p*-values, confidence intervals) to work for regression:

Simple regression assumptions

We need four things to be true for statistical inference (i.e., hypothesis tests, p-values, confidence intervals) to work for regression:

1. The errors are independent.
2. Y is a linear function of X (except for the errors).
3. The errors are normally distributed.
4. The variance of Y is the same for any value of X ("homoscedasticity").

Assumption 1: Independence

Independence means that knowing the value of a variable for one case doesn't tell you anything about another case.

- From last time: Knowing how much Bob drinks doesn't give us any suggestion as to how much Lisa drinks.

Assumption 1: Independence

Independence means that knowing the value of a variable for one case doesn’t tell you anything about another case.

- From last time: Knowing how much Bob drinks doesn’t give us any suggestion as to how much Lisa drinks.
- From today: Knowing the return this week doesn’t tell us anything about the return next week, if we believe the efficient market hypothesis.

Assumption 1: Independence

Independence means that knowing the value of a variable for one case doesn't tell you anything about another case.

- From last time: Knowing how much Bob drinks doesn't give us any suggestion as to how much Lisa drinks.
- From today: Knowing the return this week doesn't tell us anything about the return next week, if we believe the efficient market hypothesis.
- But: Time-series data often violates the independence assumption!

Assumption 1: Independence

Independence means that knowing the value of a variable for one case doesn't tell you anything about another case.

- From last time: Knowing how much Bob drinks doesn't give us any suggestion as to how much Lisa drinks.
- From today: Knowing the return this week doesn't tell us anything about the return next week, if we believe the efficient market hypothesis.
- But: Time-series data often violates the independence assumption!
- We can only check this assumption by thinking about the situation conceptually.

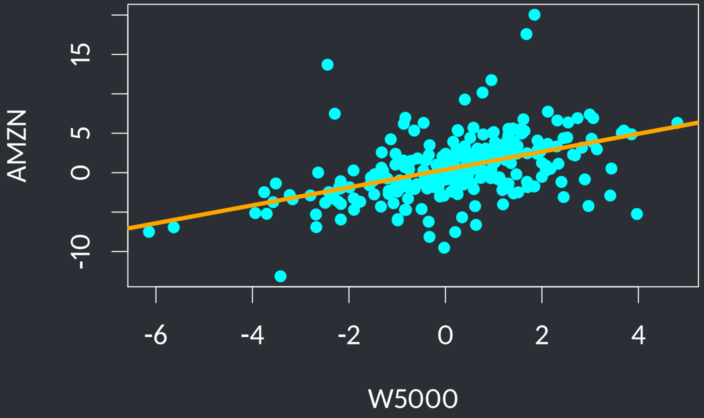
Simple regression assumptions

We need four things to be true for statistical inference (i.e., hypothesis tests, *p*-values, confidence intervals) to work for regression:

1. The errors are independent. $\checkmark$
2. Y is a linear function of X (except for the errors).
3. The errors are normally distributed.
4. The variance of Y is the same for any value of X ("homoscedasticity").

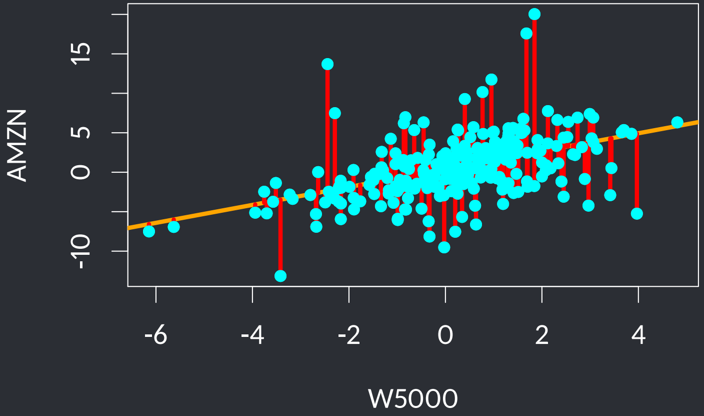
Assumption 2: Linearity

Step 1: Visually examine to ensure a line is a good fit for the data:



Assumption 2: Linearity

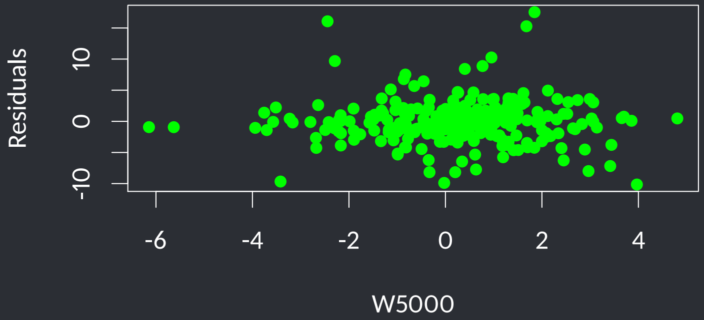
Each point has a **residual** ($Y - \hat{Y}$); this is the over/under-prediction of the model (red lines).



Assumption 2: Linearity

A **residual plot** (of residuals vs X) helps us ensure that there is not subtle nonlinearity. We want to see **no trend** in this plot:

```
> model <- lm(AMZN ~ W5000, data=stock.market)
> plot(stock.market$W5000, resid(model),
+   pch=16, col='green', xlab='W5000', ylab='Residuals')
```



Simple regression assumptions

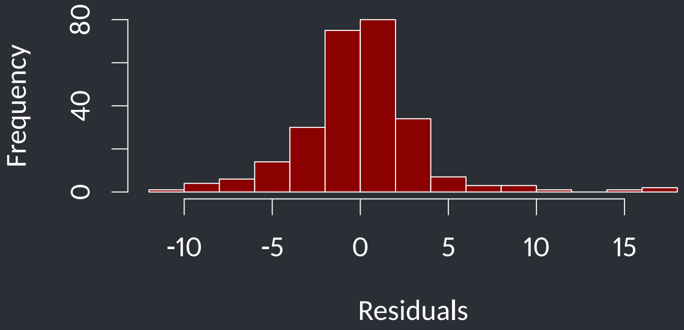
We need four things to be true for statistical inference (i.e., hypothesis tests, p-values, confidence intervals) to work for regression:

1. The errors are independent. $\checkmark$
2. Y is a linear function of X (except for the errors). $\checkmark$
3. The errors are normally distributed.
4. The variance of Y is the same for any value of X ("homoscedasticity").

Assumption 3: Errors are normally distributed

Step 1: Look at a histogram of the residuals and ensure they are approximately normally distributed:

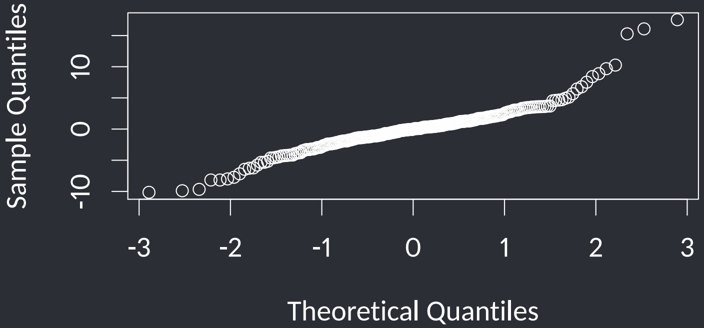
```
> hist(resid(model), col='darkred',
+   xlab='Residuals', main='')
```



Assumption 3: Errors are normally distributed

Step 2: Look at a Q-Q plot of the residuals and look for an approximately straight line:

```
> qqnorm(resid(model), main='')
```



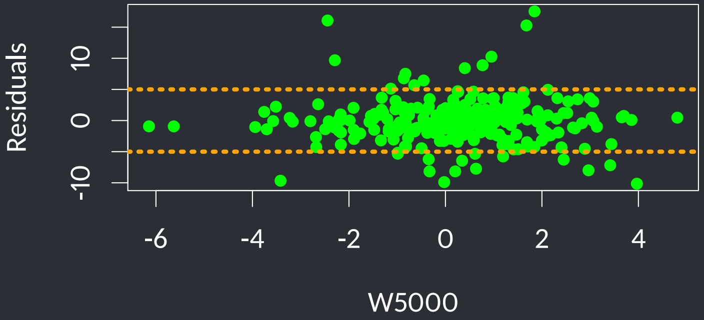
Simple regression assumptions

We need four things to be true for statistical inference (i.e., hypothesis tests, *p*-values, confidence intervals) to work for regression:

1. The errors are independent. $\checkmark$
2. Y is a linear function of X (except for the errors). $\checkmark$
3. The errors are normally distributed. $\checkmark$
4. The variance of Y is the same for any value of X ("homoscedasticity").

Assumption 4: The variance of Y is the same for any value of X

Look for the residual plot to have roughly equal vertical spread all the way across:



Simple regression assumptions

We need four things to be true for statistical inference (i.e., hypothesis tests, *p*-values, confidence intervals) to work for regression:

1. The errors are independent. ✓
2. Y is a linear function of X (except for the errors). ✓
3. The errors are normally distributed. ✓
4. The variance of Y is the same for any value of X ("homoscedasticity"). ✓

Simple regression assumptions

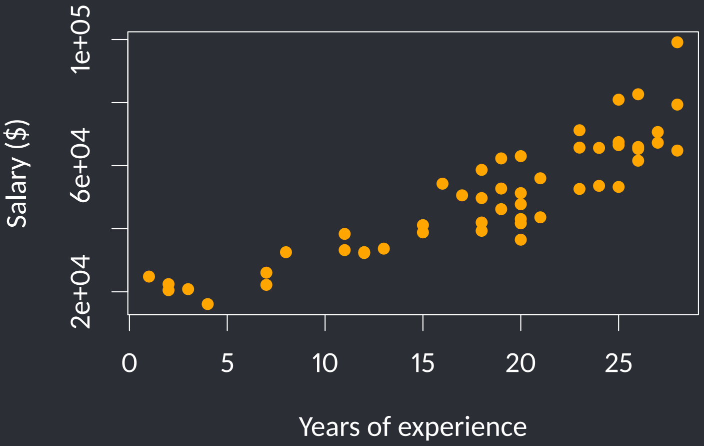
We need four things to be true for statistical inference (i.e., hypothesis tests, *p*-values, confidence intervals) to work for regression:

1. The errors are independent. ✓
2. Y is a linear function of X (except for the errors). ✓
3. The errors are normally distributed. ✓
4. The variance of Y is the same for any value of X ("homoscedasticity"). ✓

We always need to check these assumptions before interpreting *p*-values or confidence intervals!

An example where an assumption fails

This is a data set of social worker salaries based on years of experience. Which assumption might be violated here?



An example where an assumption fails