THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Inference for Simple Regression 2

Lecture 12

STA 371G

TX Votes presentation

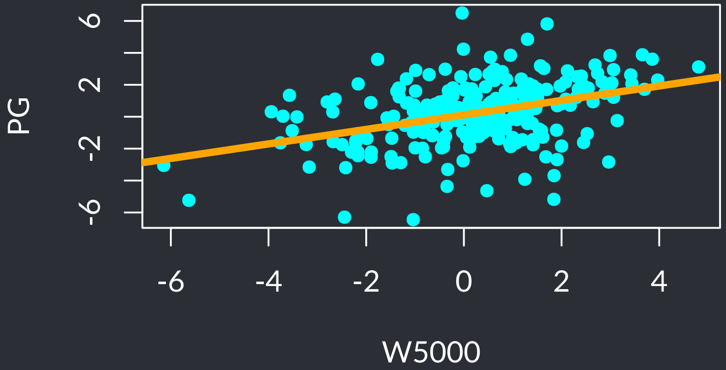
A few notes about the project

- You'll build up your paper over time—for the Preliminary Proposal all you need are the title, authors, introduction, and data description
- Also submit your R script and codebook
- Your codebook should include all of the information from the original data set's codebook for each of the variables you include
- Data sets and codebooks are posted in Canvas under Files
- Think about what you want to predict/model, and then come up with an (inital) list of 8-20 explanatory variables that might be helpful in predicting it

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.



The regression line is

$$\widehat{\mathrm{PG}} = 0.11 + 0.45 \cdot \mathrm{W5000},$$

with $R^2 = 0.15$ and $p = 4.81 \times 10^{-11}$.

Interpreting the regression statistics

- $\hat{\beta}_1 = 0.45$ ("β") is the predicted increase in returns for PG when W5000 returns increase by 1 percentage point—since this is < 1 PG will swing less than the market as a whole

Interpreting the regression statistics

- $\hat{\beta}_1 = 0.45$ (“β”) is the predicted increase in returns for PG when W5000 returns increase by 1 percentage point—since this is < 1 PG will swing less than the market as a whole
- $R^2 = 0.15$ indicates how closely PG tracks W5000 (the market as a whole)

Interpreting the regression statistics

- $\hat{\beta}_1 = 0.45$ ("β") is the predicted increase in returns for PG when W5000 returns increase by 1 percentage point—since this is < 1 PG will swing less than the market as a whole
- $R^2 = 0.15$ indicates how closely PG tracks W5000 (the market as a whole)
- $p = 4.81 \times 10^{-11}$ tells us whether we can reject the null hypothesis that PG does not move with the market at all

Interpreting the regression statistics

- $\hat{\beta}_1 = 0.45$ ("β") is the predicted increase in returns for PG when W5000 returns increase by 1 percentage point—since this is < 1 PG will swing less than the market as a whole
- $R^2 = 0.15$ indicates how closely PG tracks W5000 (the market as a whole)
- $p = 4.81 \times 10^{-11}$ tells us whether we can reject the null hypothesis that PG does not move with the market at all

Interpreting the regression statistics

- $\hat{\beta}_1 = 0.45$ ("β") is the predicted increase in returns for PG when W5000 returns increase by 1 percentage point—since this is < 1 PG will swing less than the market as a whole
- $R^2 = 0.15$ indicates how closely PG tracks W5000 (the market as a whole)
- $p = 4.81 \times 10^{-11}$ tells us whether we can reject the null hypothesis that PG does not move with the market at all (we can! since p is small)

Simple regression assumptions for inference

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 for inference

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 of errors

Independence means that knowing the error (over-/under-prediction by the regression line) for one case doesn't tell you anything about the error for another case.

Assumption 1: Independence of errors

- From last time: Knowing how much more Bob drinks than expected (based on the age he started drinking) doesn’t give us any suggestion as to how much more Lisa drinks than expected.

Assumption 1: Independence of errors

- From last time: Knowing how much more Bob drinks than expected (based on the age he started drinking) doesn't give us any suggestion as to how much more Lisa drinks than expected.
- From today: Knowing how much better or worse PG performs relative to the market *last* week doesn't tell us anything about how much better or worse PG performs relative to the market *this* week, if we believe the efficient market hypothesis.

Assumption 1: Independence of errors

- From last time: Knowing how much more Bob drinks than expected (based on the age he started drinking) doesn't give us any suggestion as to how much more Lisa drinks than expected.
- From today: Knowing how much better or worse PG performs relative to the market *last* week doesn't tell us anything about how much better or worse PG performs relative to the market *this* week, if we believe the efficient market hypothesis.
- But: Time-series data often violates the independence assumption!

Assumption 1: Independence of errors

- From last time: Knowing how much more Bob drinks than expected (based on the age he started drinking) doesn't give us any suggestion as to how much more Lisa drinks than expected.
- From today: Knowing how much better or worse PG performs relative to the market *last* week doesn't tell us anything about how much better or worse PG performs relative to the market *this* week, if we believe the efficient market hypothesis.
- But: Time-series data often violates the independence assumption!
- We can often only verify this assumption by thinking about the situation conceptually.

Simple regression assumptions for inference

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 variance of Y is the same for any value of X ("homoscedasticity").
4. The errors are normally distributed.

Simple regression assumptions for inference

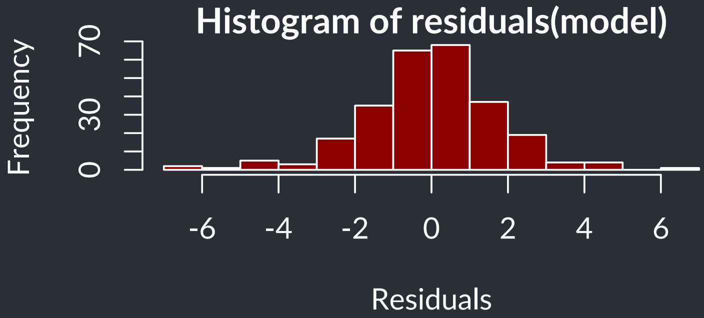
We need four things to be true for statistical inference (i.e., hypothesis tests, *p*-values, confidence intervals) to work for regression:

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

Assumption 4: Errors are normally distributed

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

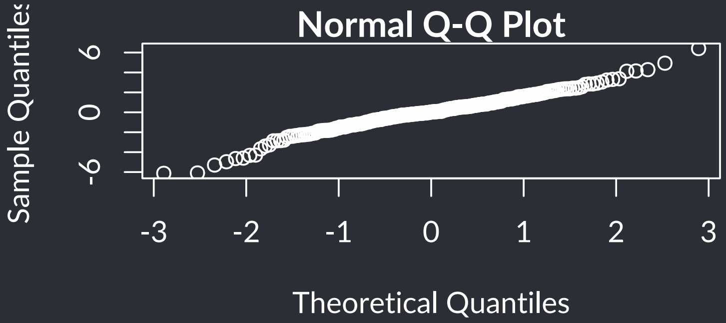
```
hist(residuals(model), col='darkred', xlab='Residuals')
```



Assumption 4: Errors are normally distributed

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

```
qqnorm(residuals(model))
```



Simple regression assumptions for inference

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 variance of Y is the same for any value of X ("homoscedasticity"). ✓
4. The errors are normally distributed. ✓

Simple regression assumptions for inference

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 variance of Y is the same for any value of X ("homoscedasticity"). ✓
4. The errors are normally distributed. ✓

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