# Diagnostic tools

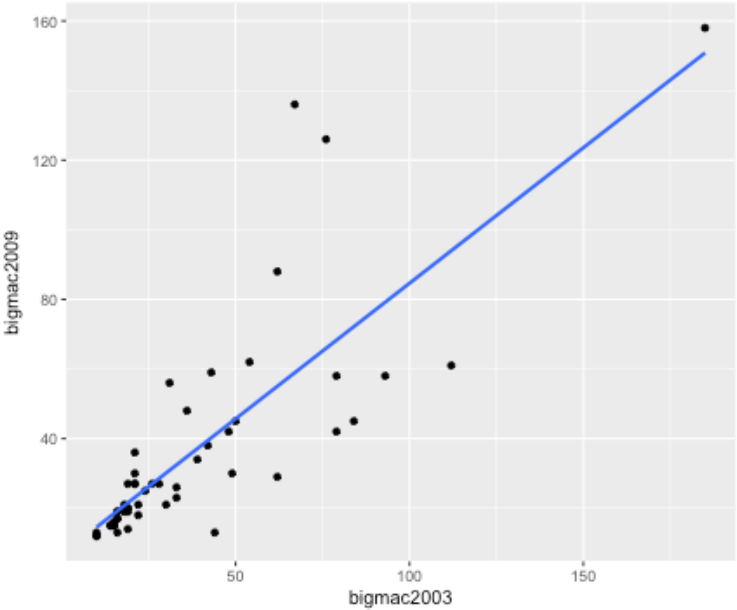
Math 430, Winter 2017

# Union Bank of Switzerland (UBS) reports

- Produces regular reports on prices & earnings in major cities throughout the world
- Measures (in minutes of labor required for "typical" worker to purchase the commodity):
  - prices of basic commodities (1 kg rice, 1 kg loaf of bread)
  - price of a Big Mac at McDonald's
- Data from 2003 (before recession) and 2009 (after recession) reports

# Some EDA results

```
qplot(x = bigmac2003, y = bigmac2009, data = UBSprices) +
  geom_smooth(method = "lm", se = FALSE)
```



# Fitting the SLR model

```
bigmac.lm <- lm(bigmac2009 ~ bigmac2003, data = UBSprices)
summary(bigmac.lm)
```

```
##
## Call:
## lm(formula = bigmac2009 ~ bigmac2003, data = UBSprices)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.968  -5.258  -2.159   0.187  77.081
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.73612    3.84985   1.750   0.0861 .
## bigmac2003   0.77886    0.07975   9.767 2.33e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.35 on 52 degrees of freedom
## Multiple R-squared:  0.6472,    Adjusted R-squared:  0.6404
## F-statistic: 95.39 on 1 and 52 DF,  p-value: 2.334e-13
```

# Review: SLR Model Assumptions

1. Linearity
2. Independence
3. Constant error variance
4. Normality

# Consequences of a violation

- **Non-linearity**: estimates are biased/meaningless
- **Non-independence**: estimates are unbiased (i.e. we fit the right line), but the SEs are a problem (typically too small)
- **Non-constant error variance**: estimates are unbiased but SEs are wrong (and we don't know how wrong)
- **Normality**: estimates are unbiased, SEs are correct BUT
    - confidence intervals are wrong for small sample sizes (we can't use t-distribution)
    - prediction intervals are wrong for all sample sizes

# Residuals

**Definition**: $e_i = y_i - \widehat{y}_i$

**Properties:**

- sum to zero (so, residuals can't be independent!)
- **uncorrelated** with $x$ and $\widehat{y}$
- normally distributed, but variance is not constant

$$e_i \sim \mathcal{N}\left(0, \sigma^2 \left[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right]\right)$$

# Standardized residuals

**Review:**

If $X$ has mean $\mu$ and standard deviation $\sigma$, then $\dfrac{X-\mu}{\sigma}$ has mean 0 and standard deviation 1

**Standardized residuals**

- Formula:

$$r_i = \frac{e_i}{s\sqrt{1-\frac{1}{n}-\frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}}}$$

# Residual Plots

- Plot residuals vs. predicted values
  - detect non-constant variance
  - detect non-linearity
  - detect outliers
- Plot residuals vs. $x$
  - in SLR, same as above
- Plot residuals vs. other possible predictors
  - detect important missing variable
- Plot residuals vs. lagged residuals
  - detect temporally correlated errors
  - sort errors in time order, plot $e_i$ vs. $e_{i-1}$

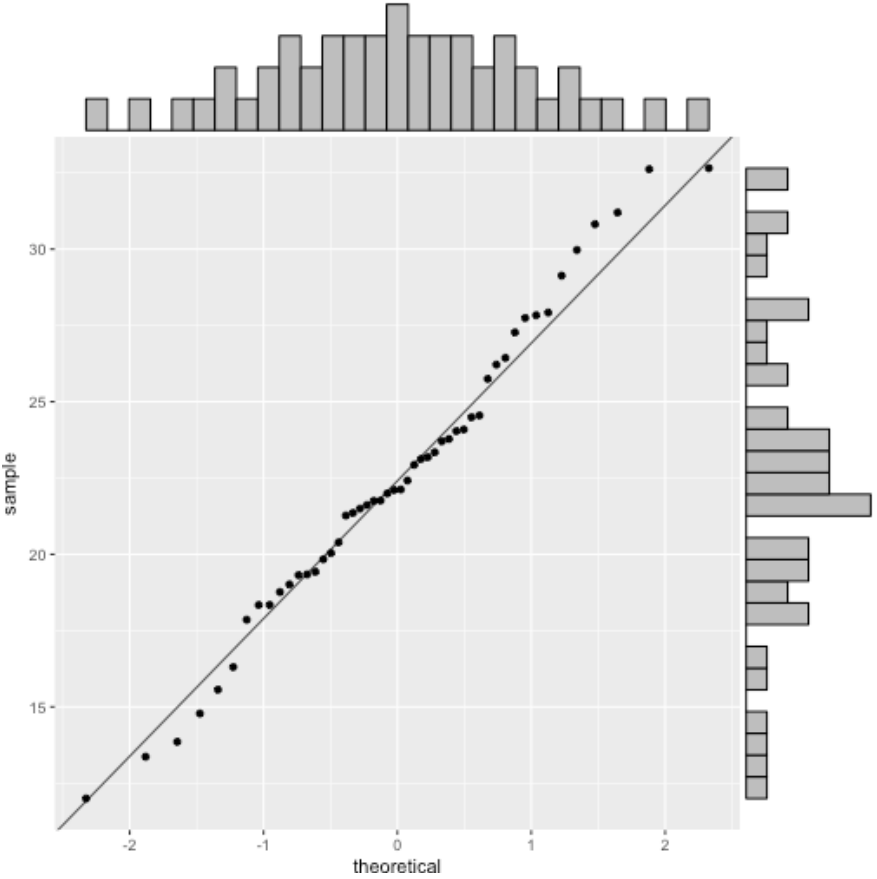
# Normal quantile-quantile (Q-Q) plots

**Use**: detect non-normality

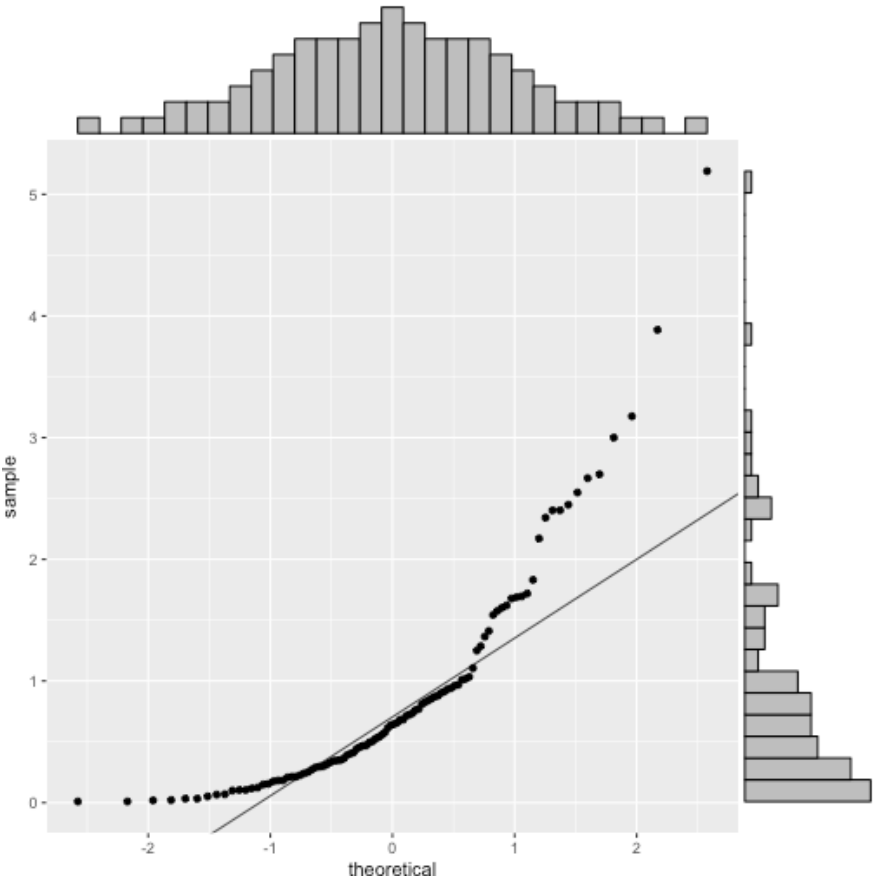
**Construction**:

- plots ordered residuals vs. quantiles from $\mathcal{N}(0, 1)$
- if in agreement, points will fall on diagonal line
- best to use standardized residuals

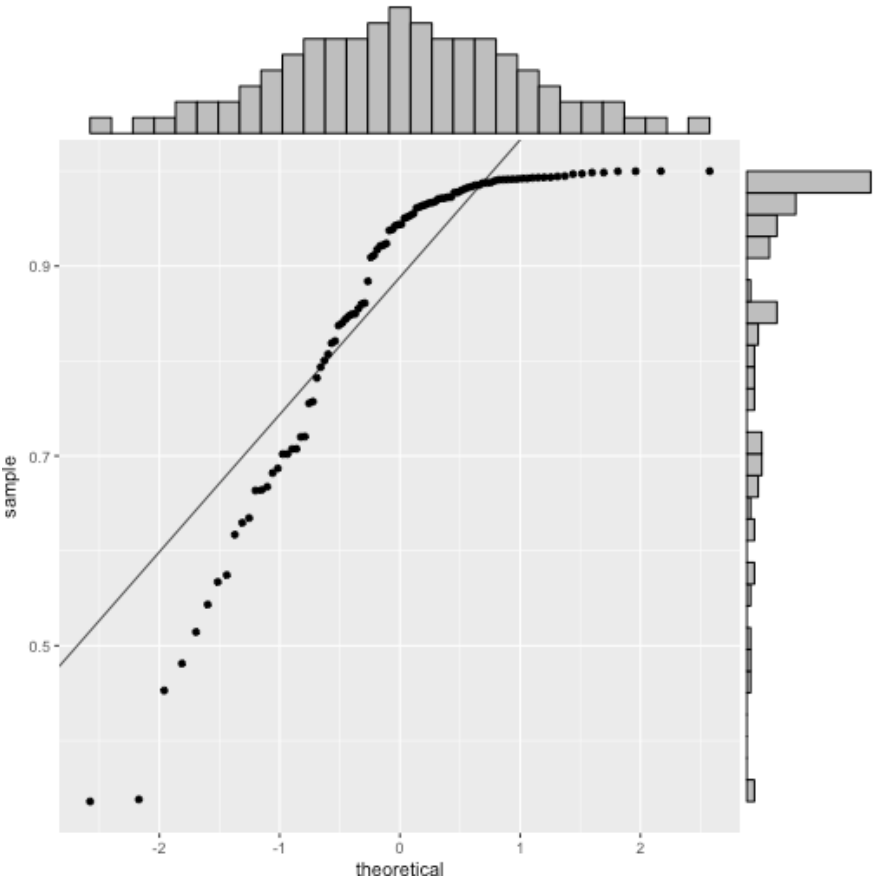
# Normal quantile-quantile (Q-Q) plots



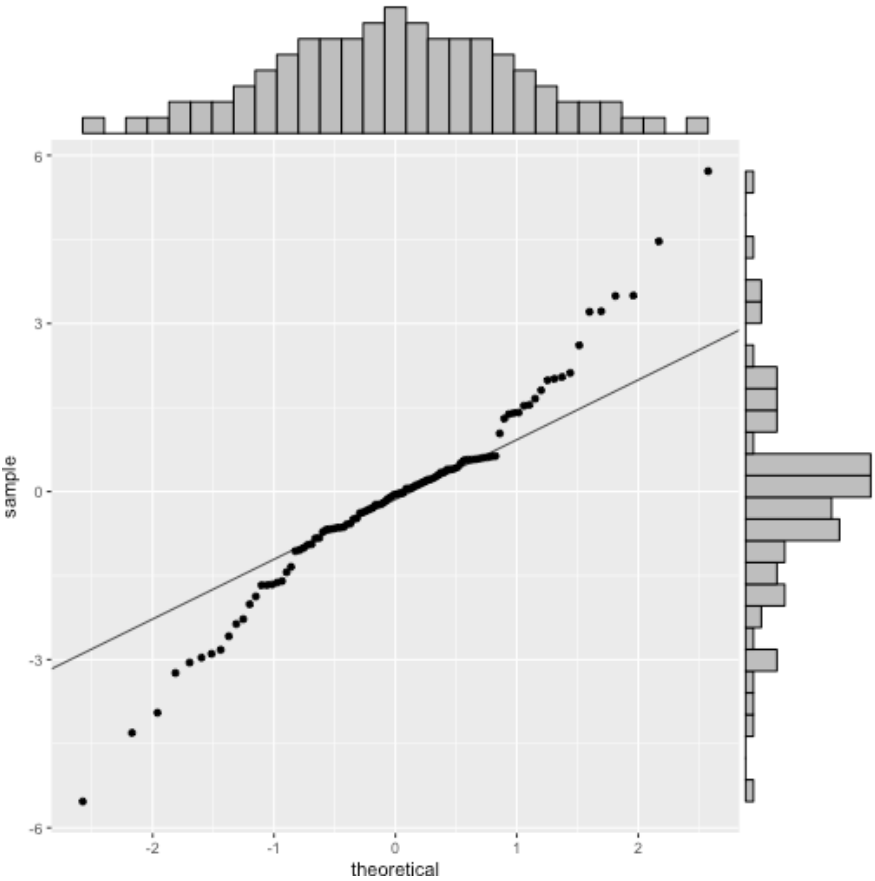
# Normal Q-Q plots



# Normal Q-Q plots



# Normal Q-Q plots

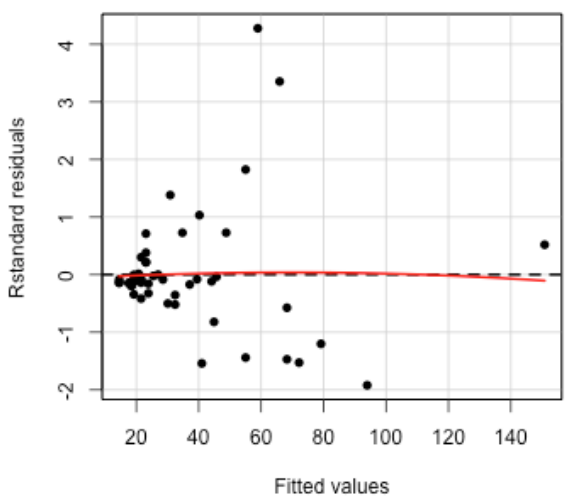


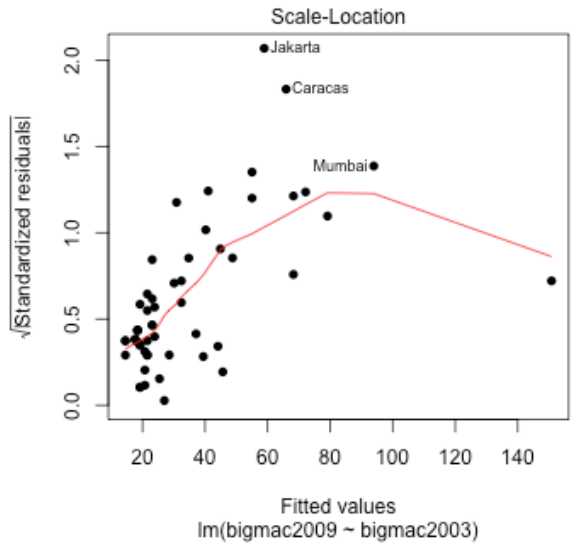
# Calibrating Our Perception

- https://gallery.shinyapps.io/slr_diag/
- App simulates residual plots under different models

# UBS Residual Plots

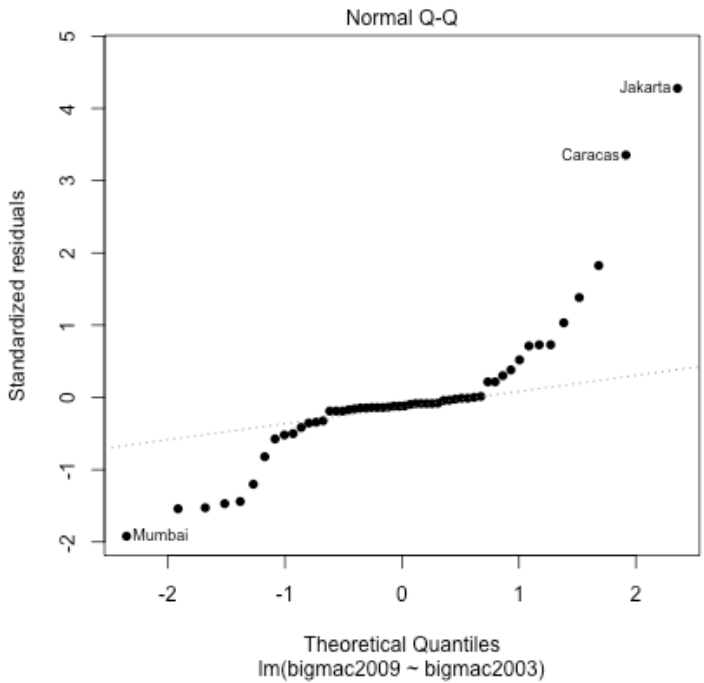
```
library(car) ## workhorse for residual plots
residualPlot(bigmac.lm, pch = 16, type = "rstandard")
plot(bigmac.lm, which = 2, pch = 16)
```



# UBS Residual Plots

```
plot(bigmac.lm, which = 2, pch = 16)
```



# Outliers and influential points

**Why we worry**

$$\widehat{\beta}_1 = r \cdot \frac{s_y}{s_x}$$

**Methods:**

1. **Graphical displays**: scatterplot, residual plot, boxplot of residuals, histogram of residuals

2. **Measures of influence**: leverage, Cook's distance

# Leverage

- **Idea**: Points farther from $\bar{x}$ have greater potential to influence the slope
- **Metric**: Leverage

$$h_i = \frac{1}{n} + \frac{\left(x_i - \bar{x}\right)^2}{\sum_{j=1}^{n} \left(x_j - \bar{x}\right)^2}$$

- $\sum h_i = 2$ for SLR, so "typical" leverage is about $2/n$
- red flag if $h_i > 4/n$
- **Caution:** leverage only refers to $x$ coordinate, does not take into account the $y$ coordinate

# UBS Leverage

```
bigmac.lev <- hatvalues(bigmac.lm)
which(bigmac.lev > 4/nrow(BigMac2003))
```

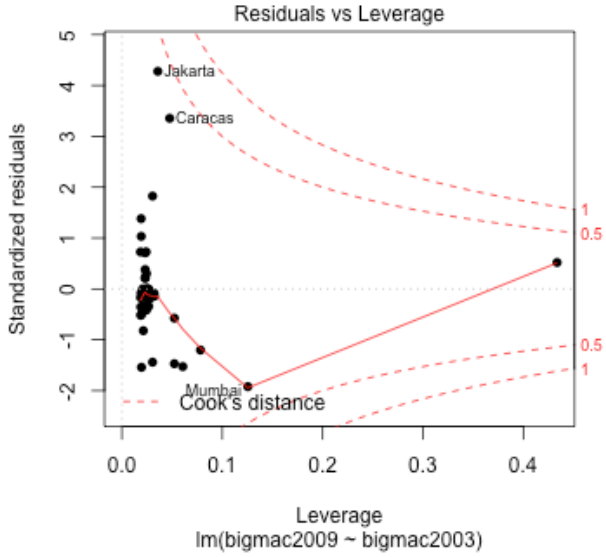
```
##  Bogota    Kiev  Mumbai Nairobi
##       7      23      36      37
```

```
which(bigmac.lev > 6/nrow(BigMac2003))
```

```
##  Mumbai Nairobi
##      36      37
```

# UBS Leverage

```
plot(bigmac.lm, which = 5, pch = 16)
```



# Cook's distance

Measures amount of influence observation has on the estimated regression equation

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_{j(i)} - \hat{y}_j\right)^2}{2s^2} = \frac{r_i}{2} \cdot \frac{h_{ii}}{1 - h_{ii}}$$

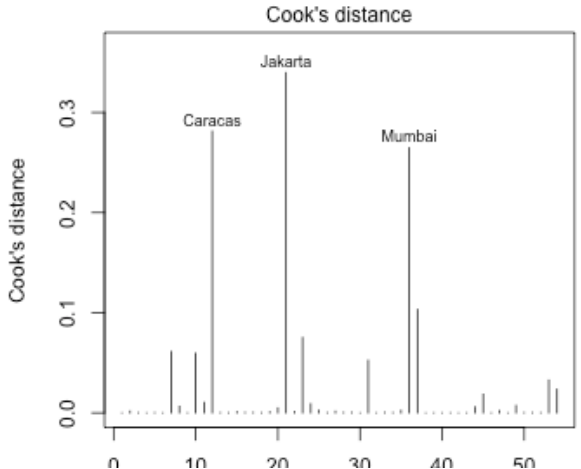
- Rule of thumb: Investigate points with $D_i > 4/(n-2)$
- Better idea: look for gaps in a plot of $D_i$

# UBS Cook's Distance

```
bigmac.cooksd <- cooks.distance(bigmac.lm)
head(bigmac.cooksd, 5)
```

```
##    Amsterdam       Athens     Auckland      Bangkok    Barcelona
## 1.633498e-06 1.722379e-03 2.451051e-04 1.561796e-05 2.897748e-04
```

```
plot(bigmac.lm, which = 4, pch = 16)
```



# What do we do if our assumptions are violated?

1. Change your assumptions (harder, need more theory)
2. Transform $y$, $x$, or both