Classic illustration: the Anscombe dataset

```r
> (anscombe <- read.csv(url("http://www.kenbenoit.net/courses/quant2/anscombe.csv")))
   x1  x2  x3  x4   y1   y2   y3   y4
 1  10  10  10   8  8.04  9.14  7.46  6.58
 2   8   8   8   8  6.95  8.14  6.77  5.76
 3  13  13  13   8  7.58  8.74 12.74  7.71
 4   9   9   9   8  8.81  8.77  7.11  8.84
 5  11  11  11   8  8.33  9.26  7.81  8.47
 6  14  14  14   8  9.96  8.10  8.84  7.04
 7   6   6   6   8  7.24  6.13  6.08  5.25
 8   4   4   4  19  4.26  3.10  5.39 12.50
 9  12  12  12   8 10.84  9.13  8.15  5.56
10  7   7   7   8  4.82  7.26  6.42  7.91
11  5   5   5   8  5.68  4.74  5.73  6.89
```
Classic illustration: the Anscombe dataset

```r
> round(apply(anscombe, 2, mean),2)  # same means for x, y
 x1  x2  x3  x4  y1  y2  y3  y4
9.0  9.0  9.0  9.0  7.5  7.5  7.5  7.5
> attach(anscombe)
> round(coef(lm(y1~x1)), 2)            # same b0, b1
 (Intercept)   x1
 3.0    0.5
> round(coef(lm(y2~x2)), 2)
 (Intercept)   x2
 3.0    0.5
> round(coef(lm(y3~x3)), 2)
 (Intercept)   x3
 3.0    0.5
> round(coef(lm(y4~x4)), 2)
 (Intercept)   x4
 3.0    0.5
> round(summary(lm(y1~x1))$r.squared, 2) # same R^2
[1] 0.67
> round(summary(lm(y2~x2))$r.squared, 2)
[1] 0.67
> round(summary(lm(y3~x3))$r.squared, 2)
[1] 0.67
> round(summary(lm(y4~x4))$r.squared, 2)
[1] 0.67
```
Anscombe dataset plotted
CLRM assumptions revisited

1. Specification:
   - $E(Y) = X\beta$ (linearity)
   - No extraneous variables in $X$
   - No omitted independent variables from $X$
   - Parameters ($\beta$) are constant

2. $E(\epsilon) = 0$

3. Error terms:
   - $\text{Var}(\epsilon) = \sigma^2$, or homoskedastic errors
   - $E(\epsilon_i, \epsilon_j) = 0$, or no auto-correlation

4. $X$ is non-stochastic
   - implies no *measurement error* in $X$
   - implies no serial correlation where a lagged value of $Y$ would be used as an independent variable
   - no *simultaneity* or *endogenous* $X$ variables

5. $\text{rank}(X) = k$

6. $\epsilon|X \sim N(0, \sigma^2)$
Omitting a relevant independent variable

- In general, $\beta^{OLS}$ of included coefficients will be biased, unless the excluded variable is uncorrelated with the included independent variables.

- If excluded variable is orthogonal to included variables, then $\beta^{OLS}$ unbiased but $\alpha^{OLS}$ (intercept) will be biased unless mean of excluded variable is zero.

- Variance-covariance matrix of $\beta^{OLS}$ will be smaller, meaning the MSE of $\beta^{OLS}$ can go up or down (depending on bias).

- Estimate of var-covariance matrix of $\beta^{OLS}$ is biased upward, because $\hat{\sigma^2}$ is biased upward, so inferences are inaccurate.
Omitting a relevant variable $Z$: graphical intuition

- Only blue and red areas reflect information used to estimate $\beta$ in $Y$ on $X$, but red also reflects variation in $Z$
- If $Z$ were included, only blue area would be used to estimate $\beta$
- Only yellow is used to estimate $\sigma^2$, except when $Z$ excluded, and then green area is also used
- If $X$ is orthogonal to $Z$, then no red area and bias disappears
Including an irrelevant independent variable

- $\beta^{OLS}$ and the estimator of its variance-covariance matrix will remain unbiased

- Generally the variance-covariance of $\beta^{OLS}$ will become larger, and therefore $\beta^{OLS}$ will be less efficient (increases MSE)

- Change in effect of $s_{b_1}$ of including irrelevant $x_2$:

$$s_{b_1} = \frac{\hat{\sigma}}{\sqrt{\sum(X_1 - \bar{X}_1)(1 - R^2)}}$$

so adding another variance will increase $R^2$ (unless $r_{x_1,x_2} = 0$)

- Keep in mind that “relevant” is a very substantive matter
Adding an irrelevant variable $Z$: graphical intuition

- Blue area reflects variation in $Y$ due entirely to $X$, so $\beta$ unbiased
- Since blue area $< (\text{blue}+\text{red})$ area, $\text{var}(\hat{\beta})$ increases
- Yellow area used to estimate $\sigma$ unbiased so var-cov matrix of $\hat{\beta}$ remains unbiased
- If $Z$ is orthogonal to $X$ then no red area and then no efficiency loss
Non-linearity

- Some non-linear forms simply cannot be used with OLS
- But others can be, if the transformation of one or more variables results in a linear function in the transformed variables
- Two types of transformations, depending on whether the whole equation or only independent variables are transformed
- Transforming only the independent variables example:

\[ y = \alpha + \beta_1 x + \beta_2 x^2 + \epsilon \]

\[ y = \alpha + \beta_1 x + \beta_2 z + \epsilon \]

where a new variable \( z = x^2 \) is created from squaring \( x \)
- The equation with \( z \) is linear in the parameters but not in the variables
Non-linearity

Transforming the entire equation means applying a transformation to both sides, not just the independent variables.

Example: the Cobb-Douglas production function:

\[
Y = AK^\alpha L^\gamma
\]

\[
\ln Y = \ln A + \alpha \ln K + \gamma \ln L + \ln \epsilon
\]

\[
Y^* = A^* + \alpha K^* + \gamma L^* + \epsilon^*
\]

is now linear in the transformed variables \( Y^*, K^* \) and \( L^* \).
Functional forms for additional non-linear transformations

log-linear as with the Cobb-Douglas production function example

semi-log has two forms:
- \( Y = \alpha + \beta \ln X \) (where \( \beta \) is \( \Delta Y \) due to \( \%\Delta X \))
- \( \ln Y = \alpha + \beta X \) (where \( \beta \) is \( \%\Delta Y \) due to \( \Delta X \))

inverse or reciprocal: \( Y = \alpha + \beta (1/X) \)

polynomial \( Y = \alpha + \beta X + \gamma X^2 \)

logit \( y = \frac{e^{\alpha + \beta X}}{1 + e^{\alpha + \beta X}} \) constrains \( y \) to lie in \([0, 1]\). Estimation is done by transforming \( y \) into log-odds ratio \( \ln[y/(1-y)] = \alpha + \beta x \)
Changing parameter values

- No real OLS solutions to this problem in the manner of previous solutions (through transformation)
- For simple “switching regimes” it is possible to divide a dataset into discrete sections, and regress using dummy variables
- A test is available for this, known as the Chow test
- For more complicated and more general models, we must use maximum-likelihood or (even better) Bayesian models
- Example:

\[
\begin{align*}
  y &= \beta_1 + \beta_2 x + \epsilon \\
  \text{where: } \beta_2 &= \alpha_1 + \alpha_2 z + \nu \\
  \text{combine to get: } y &= \beta_1 + \alpha_1 x + \alpha_2 (xz) + (\epsilon + x\nu)
\end{align*}
\]
A very easy set of diagnostic plots can be accessed by plotting a `lm` object, using `plot.lm()`.

This produces, in order:

1. residuals against fitted values
2. Normal Q-Q plot
3. scale-location plot of $\sqrt{|e_i|}$ against fitted values
4. Cook’s distances versus row labels
5. residuals against leverages
6. Cook’s distances against leverage/(1-leverage)

Note that by default, `plot.lm()` only gives you 1,2,3,5.
Residuals v. fitted plots

If constant variance assumption holds, then residuals would not show a pattern against fitted values — this pattern suggests a transformation is needed.

- plot(lm(votes1st ~ spend_total*incumb, data=dail), which=1)
plot(lm(votes1st~log(spend_total)*incumb, data=dail), which=1)
Residuals v. fitted plots: log(votes)

plot(lm(log(votes1st)~spend_total*incumb, data=dail), which=1)
Residuals vs. fitted plots: log(votes) and log(spending)

plot(lm(log(votes1st) ~ log(spend_total) * incumb, data = dail), which = 1)
plot(lm(votes1st~spend_total*incumb, data=dail), which=2)
Normal Q-Q plot: logged(votes)

plot(lm(log(votes1st) ~ spend_total*incumb, data=dail), which=2)
Examine the outliers!

- We can examine the points with row labels \(264, 269, 404\).
- Note: these are not the row numbers any longer, since we removed some with missing values.
- Let’s see what is strange about these cases:

```r
> dail[c("264","269","404"), c("district", "wholename", "party", "votes1st", "incumb", "spend_total")]

district    wholename party votes1st incumb spend_total
264  Cavan Monaghan  Vincent Martin  ind    1943      0   34542.73
269  Cavan Monaghan  Gerry McCaughey  pd     1131      0   30573.12
404   Limerick East     Aidan Ryan  ind      19      0   10890.19
```
Looks at the square root of the absolute (standardized) residuals instead of just residuals, since $\sqrt{|e|}$ is less skewed

Note the use of standardized or studentized residuals

```
plot(lm(votes1st~spend_total*incumb, data=dail), which=3)
```
Studentized and standardized residuals

▶ Define $h_i$ as the diagonal $H_{ii}$ from the hat matrix. These are also known as the leverage of each observation, and is computed as:

$$h_i = x_i (X'X)^{-1} x'_j$$

▶ The standardized residual is then:

$$\hat{e}_i = \frac{e_i}{s \sqrt{1 - h_i}}$$

▶ The studentized residual is the root mean squared error of the regression with the $i$th observation removed:

$$r_i = \frac{e_i}{s(i) \sqrt{1 - h_i}}$$

▶ Both standardized and studentized residuals are attempts to adjust residuals by their standard errors, where the $\text{Var}(e_i) = \sigma^2(1 - h_i)$

▶ Note that the calculated $e_i = Y_i - \hat{Y}_i$ all have the same variance (the homoskedasticity assumption), but the calculated $e_i$ do not
Cook’s Distance plot

Cook’s Distance is a measure of influence

plot(lm(votes1st ~ spend_total * incumb, data=dail), which=4)
Cook's Distance

- Cook's distance for observation $i$ measures the effect of deleting that observation
- Defined as:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{p\hat{\sigma}}$$

$$= \frac{e_i^2}{p\hat{\sigma}} \left[ \frac{h_{ii}}{(1 - h_{ii})^2} \right]$$

- Guideline: points for which $D_i > 1.0$ usually need closer examination
- Closely connected to leverage, which is $h_{ii}$ or the diagonal of the hat matrix $H$

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{s_{xx}}$$

- In R, `hii <- influence()$hat`
Leverage is defined as $h_{ii}$ for the $i$th observation.

```
plot(lm(votes1st~spend_total*incumb, data=dail), which=5)
```
Cook’s Distance v. leverage

plot(lm(votes1st ~ spend_total*incumb, data=dail), which=6)
What to do with Outliers?

1. Ignore the problem!
2. Investigate why the data are outliers — what makes them unusual?
3. Consider respecifying the model, either by transforming a variable or by including an additional variable (but beware of overfitting)
4. Consider a variant of “robust regression” that downweights outliers
5. Note: Fox contains a number of good rules of thumb for detecting what constitutes an “outlier” based on thresholds applied to diagnostic statistics (but best method is usually graphical)
Robust regression methods

- Robust regression methods are methods to find central tendencies without giving undue influence to outliers
- For simple case like $\bar{X}$, assume we have extreme values
- Simple solution: Trim off extreme values, such as the outer deciles: $\bar{X}_{.10}$ is the mean of $X$ value with outer 10% trimmed away
- Or: use $\bar{X}_b$ as “biweighted” mean, gradually weighting observations

\[
w = \begin{cases} 
(1 - z^2)^2 & \text{if } |z| \leq 1 \\
0 & \text{if } |z| > 1 
\end{cases}
\]

where \[z = \frac{X - \text{median}(X)}{3(\text{IQR})}\]

and \[\bar{X}_b = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i}\]

- The “bi-weighted iterative” mean takes this a step further:

\[z = \frac{X - \bar{X}_b}{3(\text{IQR})}\]
Biweighted least squares

\[ e_i = Y_i - \hat{Y}_i \]
\[ z_i = \frac{Y_i - \hat{Y}_i}{3s} \]
\[ w = (1 - z^2)^2 \quad \text{if } |z| \leq 1 \]
\[ w = 0 \quad \text{if } |z| > 1 \]
\[ b = \frac{\sum wXY}{\sum wX^2} \]
\[ a = \bar{Y} - b\bar{X} \]

1. Start with OLS (same as all \( w_i = 1 \))
2. Measure \( Y_i - \hat{Y}_i \) — this determines the weights \( w_i \)
3. Fit the new line and repeat steps 2–3 until improvement stops

In R, this is \texttt{rlm()} from the \texttt{MASS} library