ETC5521: Exploratory Data Analysis

Sculpting data using models, checking assumptions, co-dependency and performing diagnostics

Lecturer: *Emi Tanaka*
Department of Econometrics and Business Statistics
ETC5521.Clayton-x@monash.edu
Week 8 - Session 1

Housekeeping

- Week 10 Monday October 19th, **tutorial will be 4-5.30PM** and **lecture will be from 5.30-7.30PM**.
- Dr. Mine Çetinkaya-Rundel will be guest lecturing from 6.30-7.30PM.
- FLUX quiz for week 8 is available now and due Friday 5PM.

Parametric regression

Parametric regression

- **Parametric** means that the researcher or analyst assumes in advance that the data fits some type of distribution (e.g. the normal distribution). E.g. one may assume that

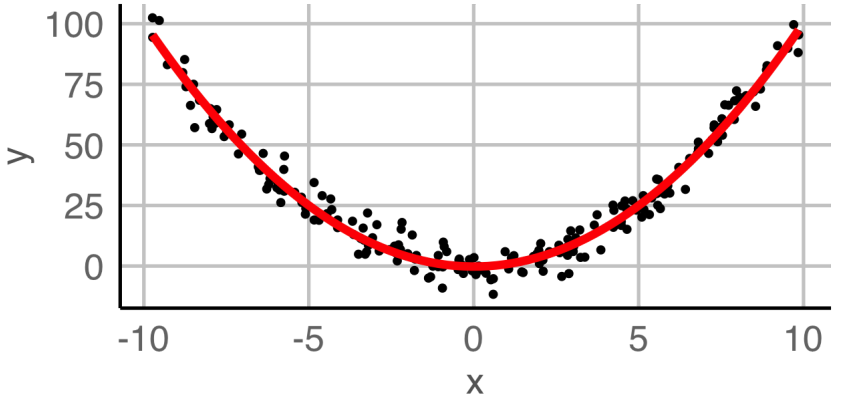
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$.

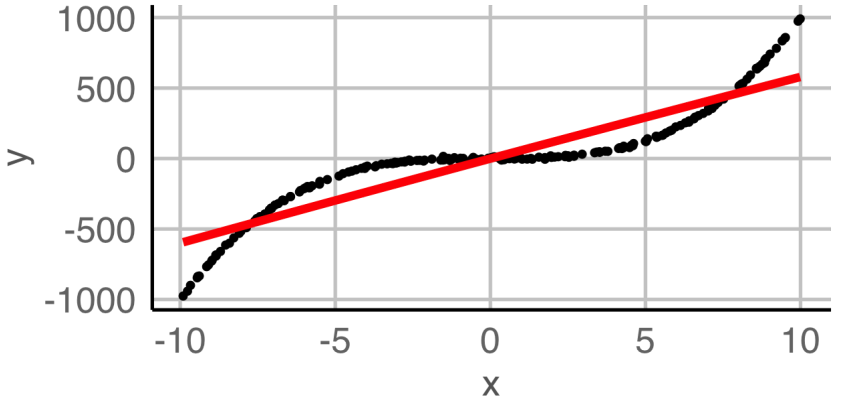
  - red = estimated
  - $blue$ = observed

- Because some type of distribution is assumed in advance, parametric fitting can lead to fitting a smooth curve that misrepresents the data.

Examples



Still assuming a quadratic fit:



Simulating data from parametric models

- In ETC5512 Wild-Caught Data, we talked about generating data from a simple linear model.
- If a model is say:

$$y = x^2 + e, \qquad e \sim N(0, 2^2)$$

we can simulate say 200 observations from this model for $x \in (-10, 10)$ by code as shown on the right.

```
set.seed(1)
df <- tibble(id = 1:200) %>%
        mutate(x = runif(n(), -10, 10),
                y = x^2 + rnorm(n(), 0, 2))
df
```

```
## # A tibble: 200 x 3
##       id     x      y
##    <int> <dbl>  <dbl>
##  1     1 -4.69 20.8
##  2     2 -2.56  6.63
##  3     3  1.46  0.301
##  4     4  8.16 67.0
##  5     5 -5.97 34.3
##  6     6  7.97 67.0
##  7     7  8.89 80.5
##  8     8  3.22 12.2
##  9     9  2.58  7.44
## 10    10 -8.76 80.2
## # … with 190 more rows
```

Logistic regression

Logistic regression

- Not all parametric models assume Normally distributed errors.
- Logistic regression models the relationship between a set of explanatory variables $(x_{i1}, \ldots, x_{ik})$ and a set of **binary outcomes** Y_i for $i = 1, \ldots, r$.
- We assume that $Y_i \sim B(n_i, p_i)$ and the model is given by

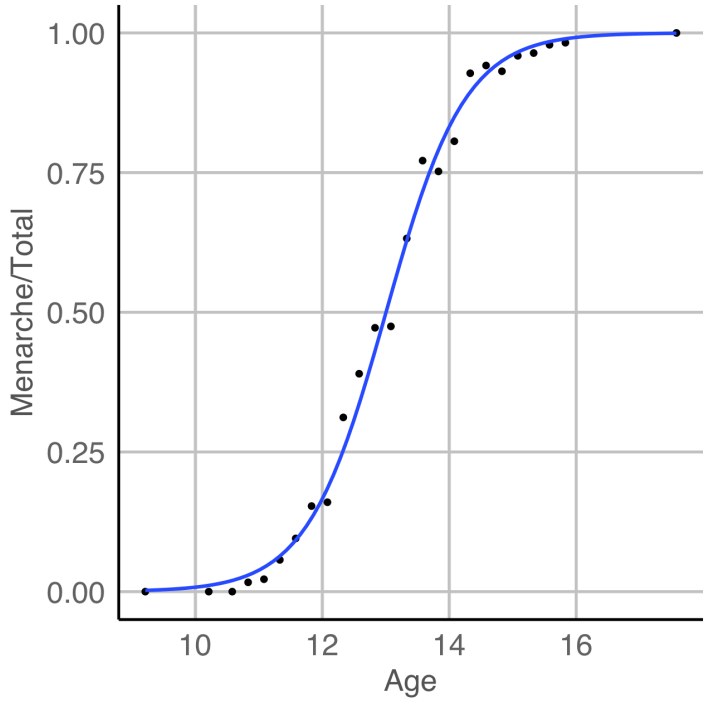
$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}.$$

- The function $f(p) = \ln\left(\frac{p}{1-p}\right)$ is called the **logit** function, continuous with range $(-\infty, \infty)$, and if p is the probablity of an event, $f(p)$ is the log of the odds.

Case study 1 Menarche

In 1965, the average age of 25 homogeneous groups of girls was recorded along with the number of girls who have reached menarche out of the total in each group.

data R



Milicer, H. and Szczotka, F. (1966) Age at Menarche in Warsaw girls in 1965. Human Biology 38, 199–203.

Simulating data from logistic regression

```
fit1 <- glm(Menarche/Total ~ Age,
            family = "binomial",
            data = menarche)
(beta <- coef(fit1))

## (Intercept)         Age
##  -20.911682    1.608169
```

- The fitted regression model is given as:

$$\text{logit}(\hat{p}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}.$$

- Taking the exponential of both sides and rearranging we get

$$\hat{p}_i = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_{i1})}}.$$

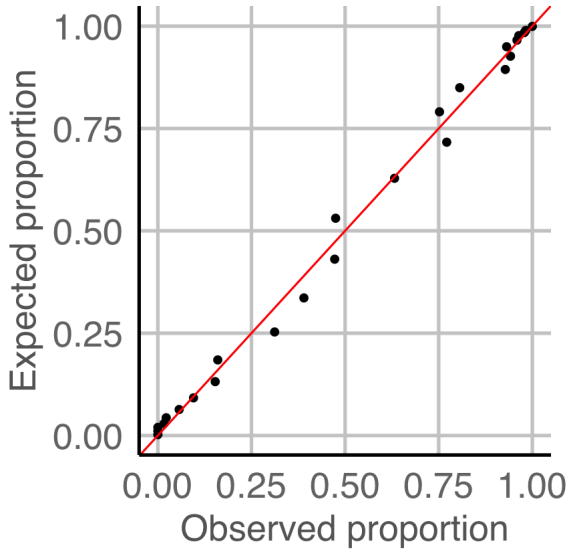
```
menarche %>%
  rowwise() %>% # simulating from first principles
  mutate(
    phat = 1/(1 + exp(-(beta[1] + beta[2] * Age))),
    simMenarche = rbinom(1, Total, phat))

## # A tibble: 25 x 5
## # Rowwise:
##      Age Total Menarche    phat simMenarche
##    <dbl> <dbl>    <dbl>   <dbl>       <int>
##  1  9.21   376        0 0.00224           2
##  2 10.2    200        0 0.0111            5
##  3 10.6     93        0 0.0199            0
##  4 10.8    120        2 0.0294            5
##  5 11.1     90        2 0.0434            8
##  6
##  7
##  8
##  9
## 10
## #
```

If simulating data from a model object, `simulate` function usually can do this for you!

Diagnostics for logistic regression models

- One diagnostic is to compare the observed and expected proportions under the logistic regression fit.

```
df1 <- menarche %>%
  mutate(
    pexp = 1/(1 + exp(-(beta[1] + beta[2] * Age))),
    pobs = Menarche / Total)
```



- Goodness-of-fit type test is used commonly to assess the fit as well.
- E.g. Hosmer–Lemeshow test, where test statistic is given as

$$H = \sum_{i=1}^{r} \left(\frac{(O_{1i} - E_{1g})^2}{E_{1i}} + \frac{(O_{0i} - E_{0g})^2}{E_{0i}} \right)$$

where O_{1i} (E_{1i}) and O_{0i} (E_{0i}) are observed (expected) frequencies for successful and non-successful events for group i, respectively.

```
vcdExtra::HLtest(fit1)

## Hosmer and Lemeshow Goodness-of-Fit Test
##
## Call:
## glm(formula = Menarche/Total ~ Age, family = "binom
##  ChiSquare df   P_value
##  0.1041887  8 0.9999997
```

Diagnostics for linear models

Assumptions for linear models

For $i \in \{1, \ldots, n\}$,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i,$$

where $\epsilon_i \sim NID(0, \sigma^2)$ or in matrix format,

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

where

- $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$,
- $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)^\top$,
- $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$, and
- $\mathbf{X} = \begin{bmatrix} \mathbf{1}_n & \boldsymbol{x}_1 & \ldots & \boldsymbol{x}_k \end{bmatrix}$, where
- $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})^\top$ for $j \in \{1, \ldots, k\}$

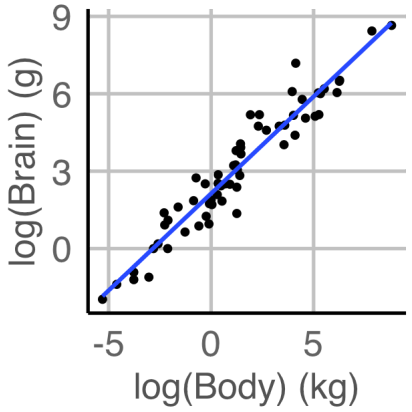
This means that we assume

1. $E(\epsilon_i) = 0$ for $i \in \{1, \ldots, n\}$.
2. $\epsilon_1, \ldots, \epsilon_n$ are independent.
3. $Var(\epsilon_i) = \sigma^2$ for $i \in \{1, \ldots, n\}$ (i.e. homogeneity).
4. $\epsilon_1, \ldots, \epsilon_n$ are normally distributed.

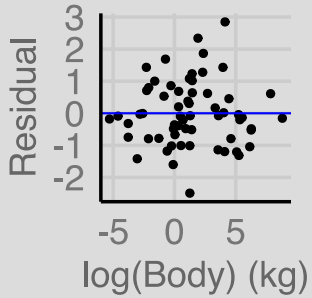
So how do we check it?

Model diagnostics for linear models

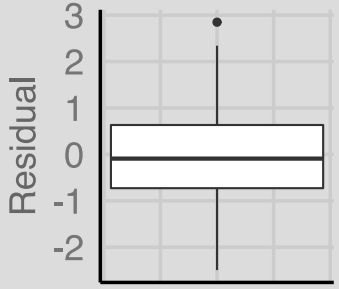
Plot Y_i vs x_i to see if there is $\approx$ a linear relationship between Y and x.



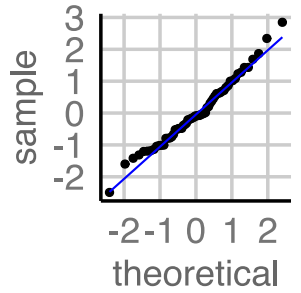
To check the homoscedasticity assumption, plot R_i vs x_i. There should be no obvious patterns.



A boxplot of the residuals R_i to check for symmetry.



A normal Q-Q plot, i.e. a plot of the ordered residuals vs $\Phi^{-1}(\frac{i}{n+1})$.



Assessing (A1) $E(\epsilon_i) = 0$ for $i = 1, \ldots, n$

- It is a property of the least squares method that

$$\sum_{i=1}^{n} R_i = 0, \quad \text{so} \quad \bar{R}_i = 0$$

  for $R_i = Y_i - \hat{Y}_i$, hence (A1) will always appear valid "overall".
- Trend in residual versus fitted values or covariate can indicate "local" failure of (A1).
- What do you conclude from the following plots?

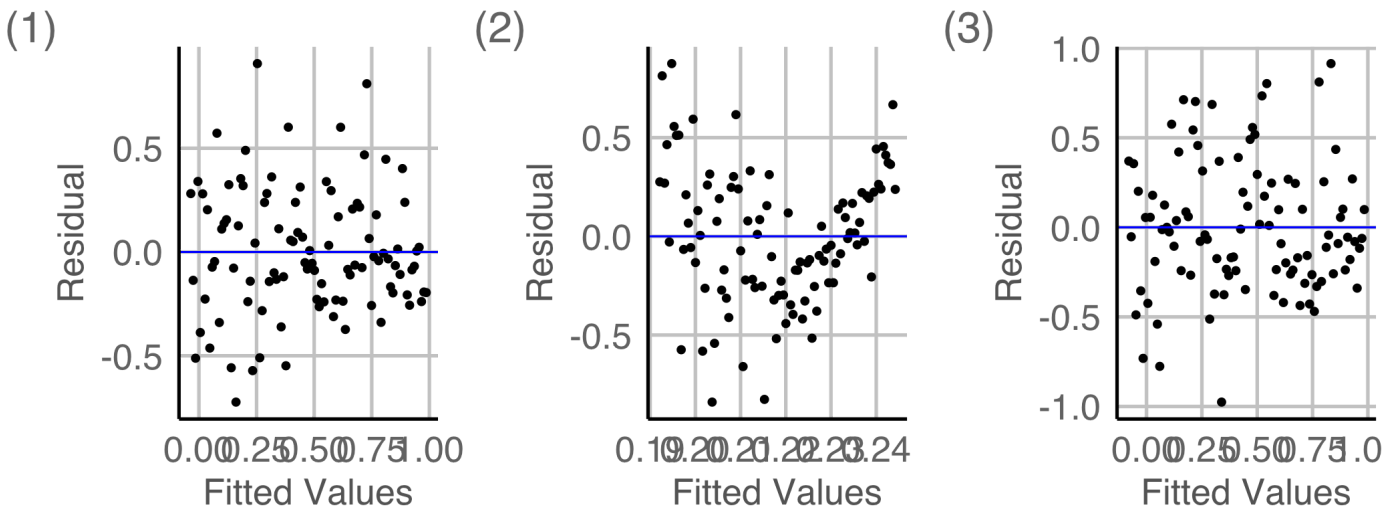
Assessing (A2)-(A3)

(A2) $\epsilon_1, \ldots, \epsilon_n$ are independent

- If (A2) is correct, then residuals should appear randomly scattered about zero if plotted against fitted values or covariate.
- Long sequences of positive residuals followed by sequences of negative residuals in R_i vs x_i plot suggests that the error terms are not independent.

(A3) $Var(\epsilon_i) = \sigma^2$ for $i = 1, \ldots, n$

- If (A3) holds then the spread of the residuals should be roughly the same across the fitted values or covariate.



Assessing (A4) $\epsilon_1, \ldots, \epsilon_n$ are normally distributed

Q-Q Plots

- The function `qqnorm(x)` produces a Q-Q plot of the ordered vector `x` against the quantiles of the normal distribution.
- The n chosen normal quantiles $\Phi^{-1}(\frac{i}{n+1})$ are easy to calculate but more sophisticated ways exist:
  - $\frac{i}{n+1} \mapsto \frac{i-3/8}{n+1/4}$, default in `qqnorm`.
  - $\frac{i}{n+1} \mapsto \frac{i-1/3}{n+1/3}$, recommended by Hyndman and Fan (1996).

In R

```
fit <- lm(y ~ x)
```

By "hand"

```
plot(qnorm((1:n) / (n + 1)), sort(resid(fit)))
```

By `base`

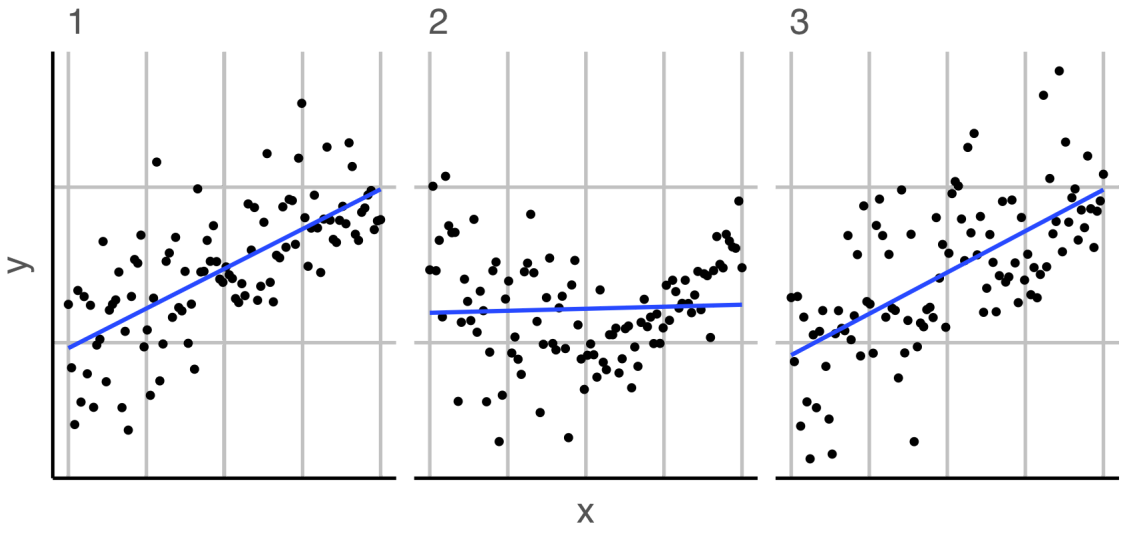
```
qqnorm(resid(fit))
qqline(resid(fit))
```

By `ggplot2`

```
data.frame(residual = resid(fit)) %>%
  ggplot(aes(sample = residual)) +
  stat_qq() + stat_qq_line(color="blue")
```

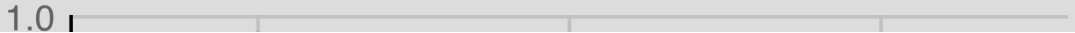

Reference: Hyndman and Fan (1996). *Sample quantiles in statistical packages*, American Statistician, 50, 361--365.

Examining the simulated data further



Simulation scheme

```
n <- 100
x <- seq(0, 1, length.out = n)
y1 <- x + rnorm(n) / 3                          #  Linear
y2 <- 3 * (x - 0.5) ^ 2 +
  c(rnorm(n / 2)/3, rnorm(n / 2)/6)             #  Quadratic
y3 <- -0.25 * sin(20 * x - 0.2) +
  x + rnorm(n) / 3                              #  Non-linear

M1 <- lm(y1 ~ x); M2 <- lm(y2 ~ x); M3 <- lm(y3 ~ x)
```



Revisiting outliers

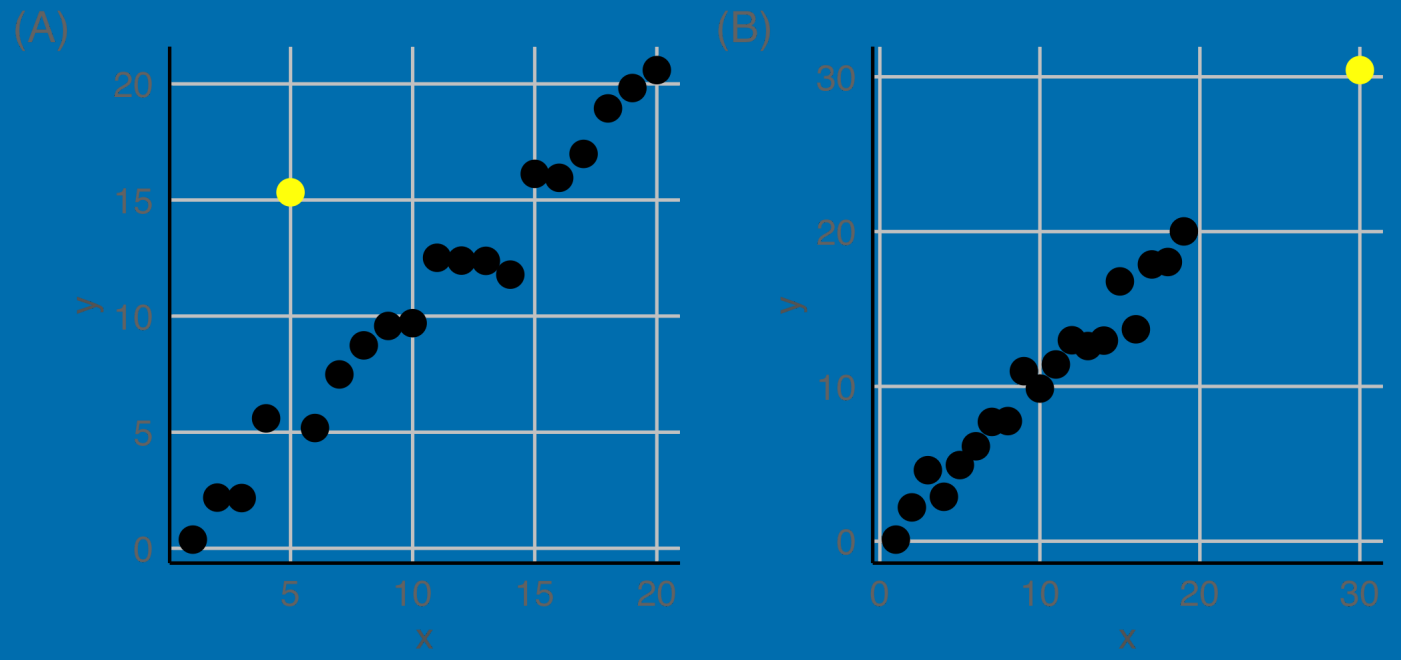
- We defined outliers in week 4 as "observations that are significantly different from the majority" when studying univariate variables.
- There is actually no hard and fast definition.

> We can also define an outlier as a data point that emanates from a different model than do the rest of the data.

- Notice that this makes this definition *dependent on the model* in question.

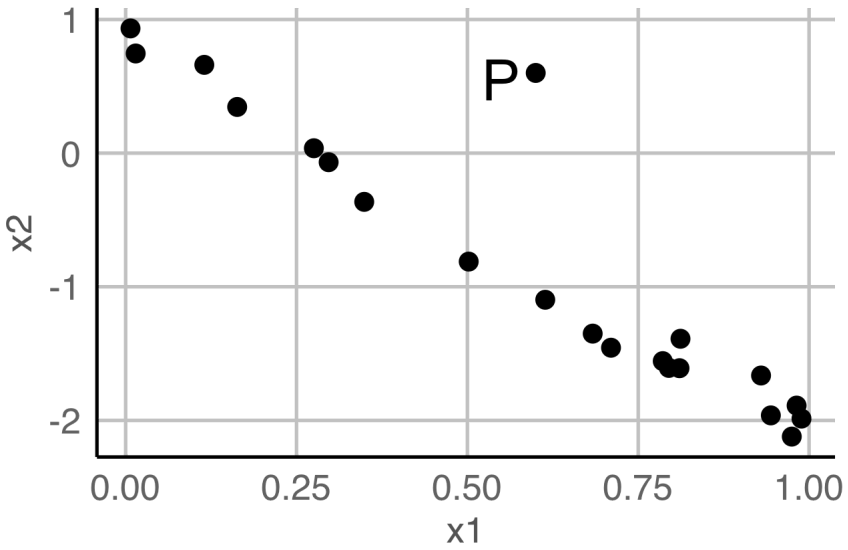
Pop Quiz

Would you consider the yellow points below as outliers?



Outlying values

- As with simple linear regression the fitted model should not be used to predict Y values for $\boldsymbol{x}$ combinations that are well away from the set of observed $\boldsymbol{x}_i$ values.
- This is not always easy to detect!



- Here, a point labelled P has x_1 and x_2 coordinates well within their respective ranges but P is not close to the observed sample values in 2-dimensional space.
- In higher dimensions this type of behaviour is even harder to detect but we need to be on guard against extrapolating to extreme values.

Leverage

- The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is referred to as the **hat matrix**.
- The i-th diagonal element of $\mathbf{H}$, h_{ii}, is called the **leverage** of the i-th observation.
- Leverages are always between zero and one,

$$0 \leq h_{ii} \leq 1.$$

- Notice that leverages are not dependent on the response!
- Points with high leverage can exert a lot of influence on the parameter estimates

Studentized residuals

In order to obtain residuals with equal variance, many texts recommend using the **studentised residuals**

$$R_i^* = \frac{R_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

for diagnostic checks.

Cook's distance

- Cook's distance, D, is another measure of influence:

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[-i]})^\top Var(\hat{\boldsymbol{\beta}})^{-1}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{[-i]})}{p}$$

$$= \frac{R_i^2 h_{ii}}{(1-h_{ii})^2 p\hat{\sigma}^2},$$

where p is the number of elements in $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{[-i]}$ and $\hat{Y}_{j[-i]}$ are least squares estimates and the fitted value obtained by fitting the model ignoring the i-th data point $(\boldsymbol{x}_i, Y_i)$, respectively.

Case study 2 Social media marketing

Data collected from advertising experiment to study the impact of three advertising medias (youtube, facebook and newspaper) on sales.

`data R`

Extracting values from models in R

- The leverage value, studentised residual and Cook's distance can be easily extracted from a model object using `broom::augment`.
  - `.hat` is the leverage value
  - `.std.resid` is the studentised residual
  - `.cooksd` is the Cook's distance

```
fit <- lm(sales ~ youtube * facebook, data = marketing)
broom::augment(fit)
```

```
## # A tibble: 200 x 9
##    sales youtube facebook .fitted  .resid .std.resid    .hat .sigma   .cooksd
##    <dbl>   <dbl>    <dbl>   <dbl>   <dbl>      <dbl>   <dbl>  <dbl>     <dbl>
##  1 26.5    276.     45.4    26.0   0.496      0.442  0.0174    1.13 0.000864
##  2 12.5     53.4    47.2    12.8  -0.281     -0.252  0.0264    1.13 0.000431
##  3 11.2     20.6    55.1    11.1   0.0465     0.0423 0.0543    1.14 0.0000256
##  4 22.2    182.     49.6    21.2   1.04       0.923  0.0124    1.13 0.00268
```

That's it, for this lecture!

Lecturer: Emi Tanaka

Department of Econometrics and Business Statistics

ETC5521.Clayton-x@monash.edu