Statistical modelling

#2.i Diagnostic plots

Dr. Léo Belzile
HEC Montréal

Model assumptions

We postulate $\varepsilon_i \sim \mathbf{No}(0, \sigma^2)$ are independent errors.

- independence
- linearity
- homoscedasticity (equal variance)
- normality

Assumptions revisited

1. **Independence**: the errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent (thus, so are the observations)
2. **Linearity**: the expectation of the errors is $\mathsf{E}(\varepsilon_i) = 0$ for all $i = 1, \ldots, n$.
   - ✚ this implies that the mean model is correctly specified, so $\mathsf{E}(Y \mid \mathbf{X}) = \beta_0 + \beta_1 \mathrm{X}_1 + \cdots + \beta_p \mathrm{X}_p$
   - ✚ all the important explanatory variables have been included in the model
   - ✚ and their effects (presumed linear) have been properly captured by the model.
3. **Homoscedasticity**: the variance of the errors is **constant** $\mathsf{Va}(\varepsilon_i) = \sigma^2$ for $i = 1, \ldots, n$.
   - ✚ the variance of Y_i is constant and does not depend on $\mathbf{X}$.
4. **Normality**: the error terms $\boldsymbol{\varepsilon}$ follows a normal distribution.

Default graphics

- Use options `plots=diagnostics residuals(smooth)` to get default residual diagnostic plots and plots of residuals against continuous explanatories.

In **SAS**, we can save additional objects from the `glm` fit using the command `output`.

- In the code excerpt, we copy (names are user-specific)
  - the fitted values `fitted`
  - the ordinary residuals `ores`
  - the jackknife studentized residuals `jsr` in the temporary database `resid`.

- SAS code ✚ SAS output (1) ✚ SAS output (2)

```
ods graphics on;
proc glm data=infe.intention
    plots=diagnostics residuals;
class sex marital educ revenue;
model intention= fixation emotion marital
    sex age revenue educ / ss3 solution;
output out=resid predicted=fitted
    r=ores rstudent=jsr;
run;
```

Review of graphs (clockwise from top left)

- residual versus fitted values (linearity)
- Jackknife studentized residuals against fitted values (heteroscedasticity)
- Leverage plot (shows influence of observation on estimators)
- quantile-quantile plot of residual (normality)
- scatterplot of Y_i versus $\hat{Y}_i$ (linearity, but depends on R^2)
- Cook's distance plot (used to detect outliers)
- Density and histogram of ordinary residuals (normality)

In this example, the analysis of residual does not give us any reason to doubt the model assumptions. Therefore, we can be confident in the results of our analysis (hypothesis tests and confidence intervals).

Creating plots by hand: correlogram (ACF)

- Context ✚ Correlogram ✚ SAS code

✚ The `airpassengers` data contains monthly observations of the air traffic in the 1960s.

✚ We fit a linear model with month (categorical) and year (continuous) for log of the number of passengers.

✚ The autocorrelation function (ACF) shows there is residual dependence at different lags, both monthly and yearly dependence.

Linearity assumption

Many potential graphs of ordinary residuals...

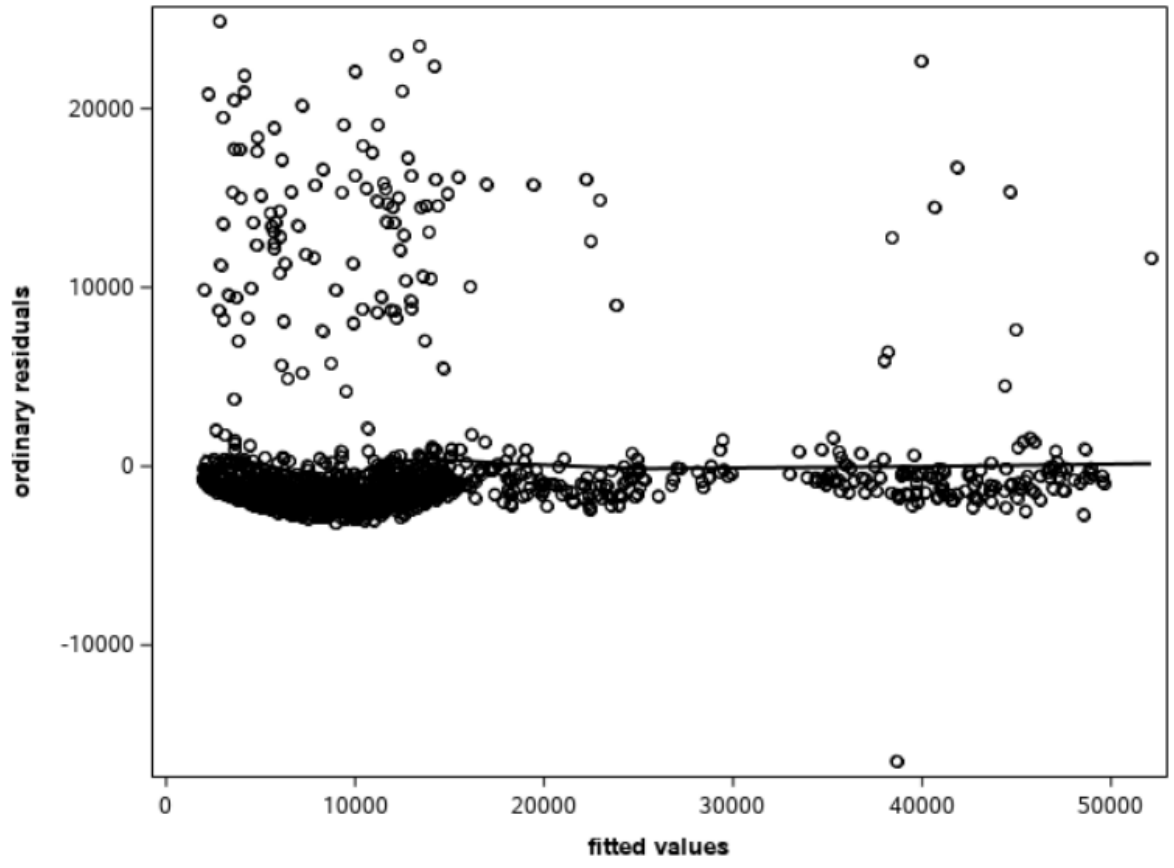
- against fitted values
- against explanatories
- against omitted covariates (not included in the mean model)
- added-variable plots

Insurance data

Consider a linear model with `age`, `sex`, `region` and the interaction between `smoker/obesity` and `bmi`.

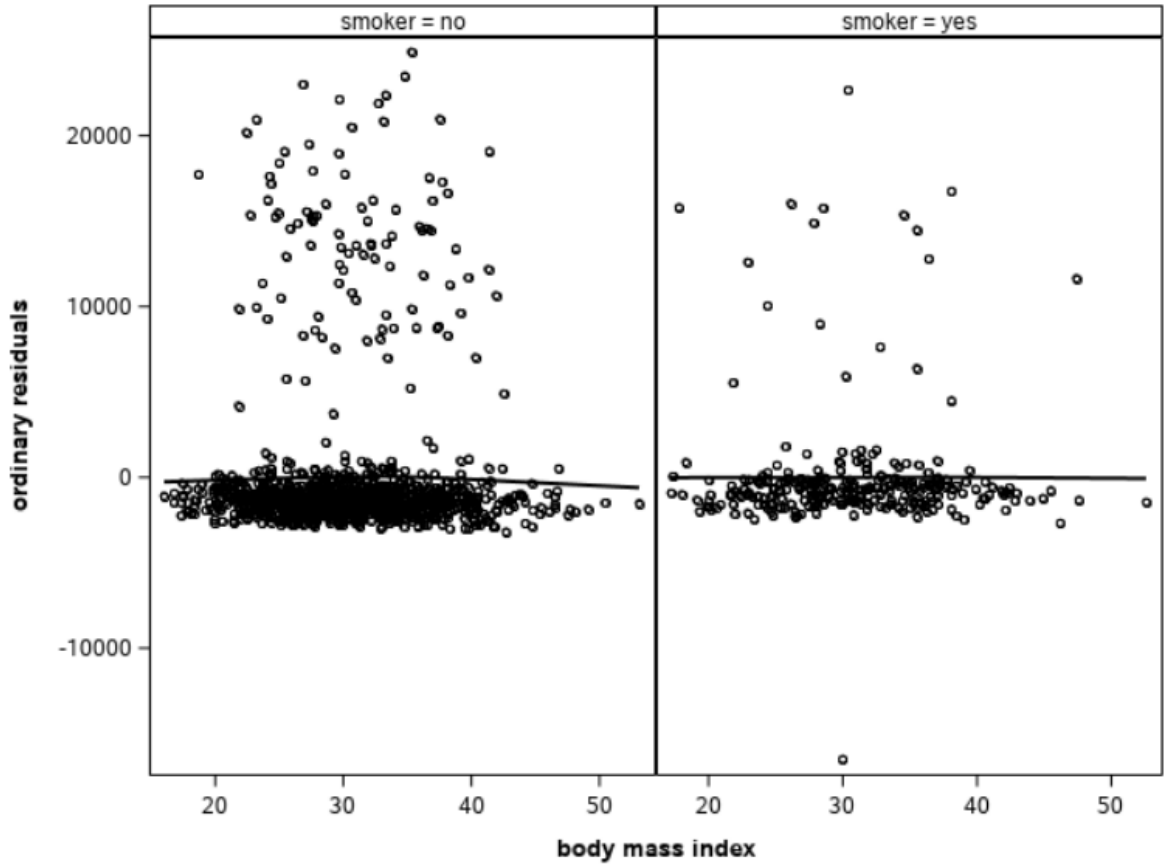
- The plots show that our model is inadequate, but this can lead to wrong diagnostics:
  - because of unexplained (abnormally high) charges, the line for e.g., non smoker is too high.
  - most data are well captured, but this impact quantile-quantile plot.
  - a log-transformation could reduce the impact of these abnormal values (smaller differences), or else robust regression

- SAS output + SAS code



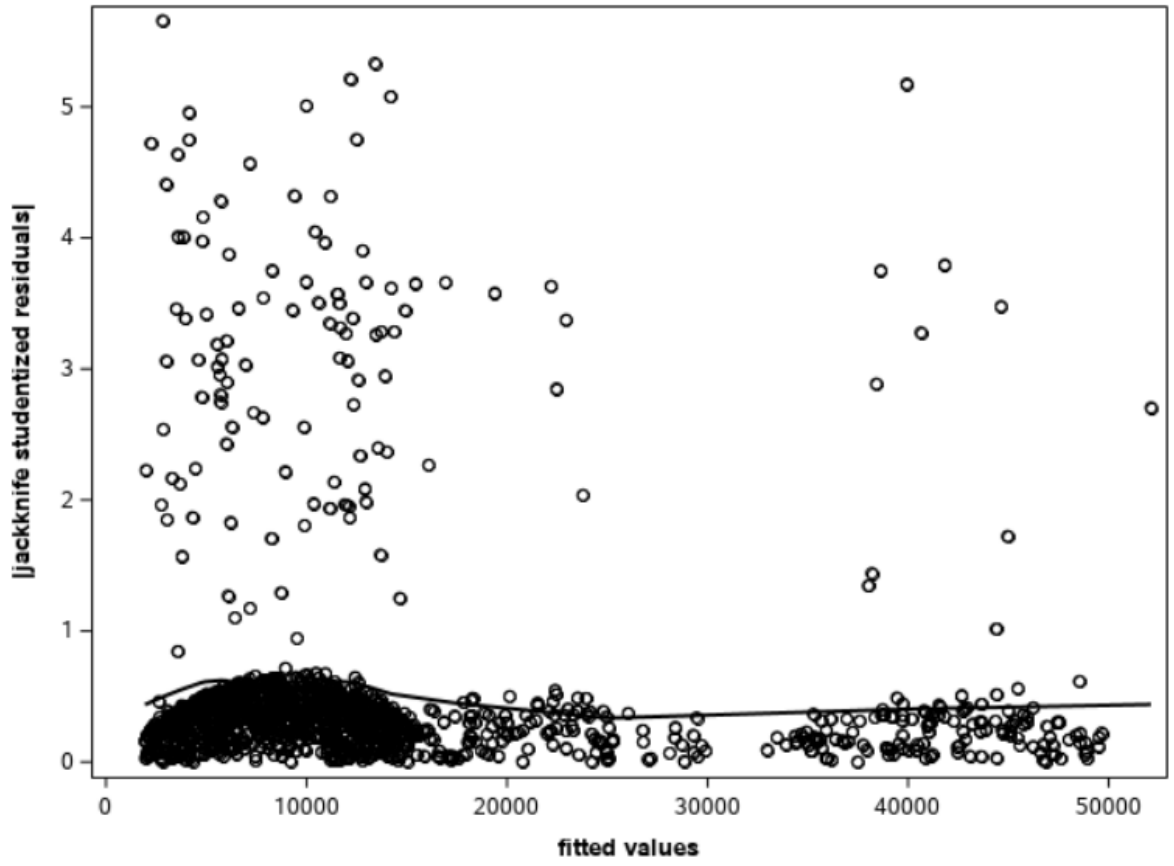
Linearity

- SAS output (1) + SAS output (2) + SAS code



Linearity (2)

- SAS output (1) + SAS output (2) + SAS code



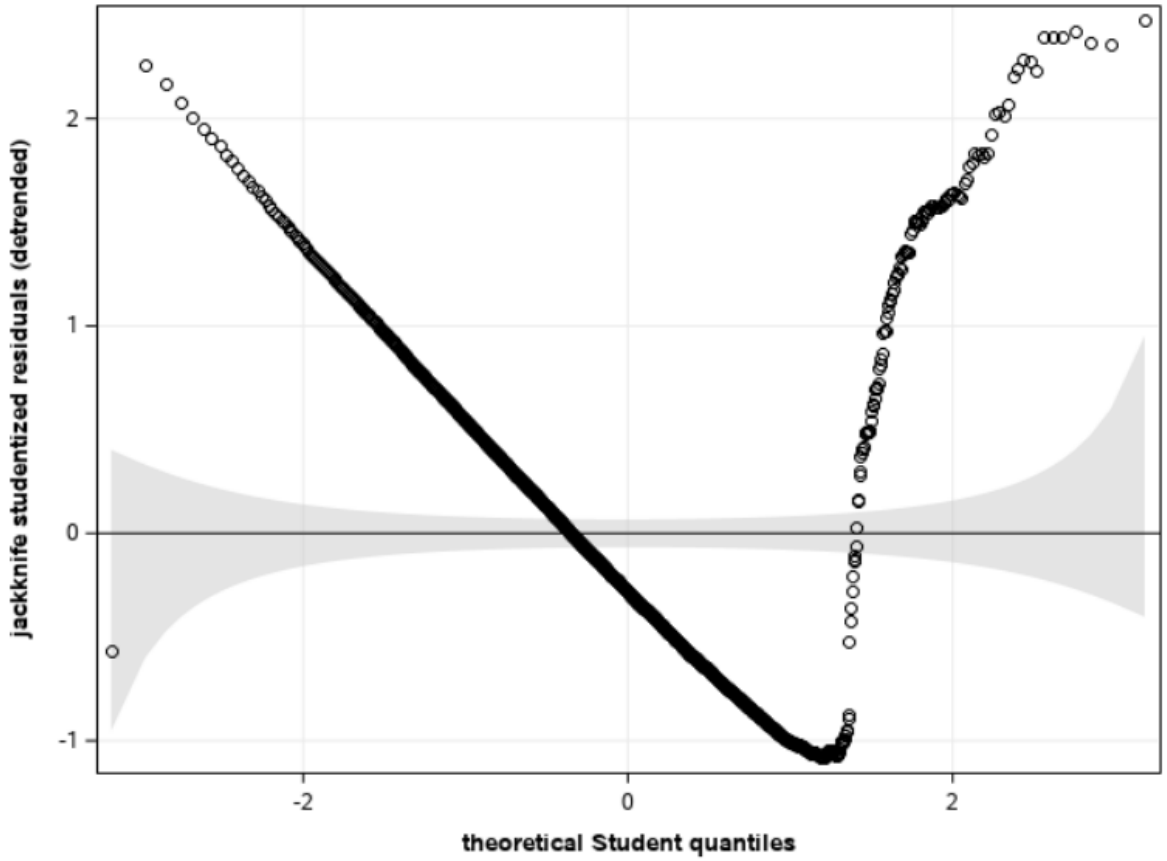
Quantile-quantile plots

To create a quantile-quantile plot manually

- sort the data (jackknife studentized residuals)
- compute plotting positions $i/(n+1)$, $i = 1, \ldots, n$
- calculate inverse transform $F^{-1}\{i/(n+1)\}$, where F^{-1} is the quantile function of the postulated distribution.
- add approximate pointwise confidence bands (computed using order statistics)
  - $U_{(j)} \sim \mathsf{Be}(j, n+1-j)$
  - therefore pick 0.025 and 0.975 quantiles of $\mathsf{Be}(j, n+1-j)$
  - back-transform to Student
  - detrend

Normality

- SAS output + SAS code (1) + SAS code (2)



Quantile-quantile plots

- `proc univariate` also supports a limited number of distributions, including the normal distribution.
- You could use the normal approximation to the Student-t distribution provided the degrees of freedom parameter $n - p - 2$ are large (greater than 20).

```
/* Histogram of jackknife studentized residuals
    with density estimate */
proc sgplot data=resid;
histogram jsr;
density jsr / type=kernel;
keylegend / position=bottom;
run;

proc univariate data=resid noprint;
qqplot jsr / normal(mu=est sigma=est l=2)
square;
run;
```