Regression and Simulation Methods

Week 3: Linear Models Continued and Non-Linear Regression

Chris Oldnall, 24th October 2023

4 Steps to (Normal) Linear Modelling

- Formation
- Estimation
- Checking
- Analysis

Analysis

Residual sum of squares

Confidence intervals

Hypothesis testing

F-Tests

How to quantify the RSS?

$$\text{RSS} = \sum_{i=1}^{n} \{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \ldots + \hat{\beta}_{p-1} x_{p-1,i})\}^2$$

$$\mathbb{E}[\text{RSS}] = (n-p)\sigma^2$$

$$\hat{\sigma}^2 = \text{RSS}/(n-p)$$

How do I define a confidence interval?

Definition 23.1. Suppose a dataset $x_1, \ldots, x_n$ is modelled by random variables $X_1, \ldots, X_n$. Let θ be the model parameter and let $\gamma \in [0, 1]$. Let $L = g(X_1, \ldots, X_n)$ and $U = h(X_1, \ldots, X_n)$ be such that

$$P(L \leq \theta \leq U) = \gamma$$

for any value of θ. Then the interval

$$[l, u]$$

is a $100\gamma\%$ **confidence interval** for θ, where $l = g(x_1, \ldots, x_n)$ and $u = h(x_1, \ldots, x_n)$. γ is the **confidence level**.

How do I define a confidence interval?

Proposition 23.2. *Suppose a dataset* $x_1, \ldots, x_n$ *is modelled as an i.i.d. sample* $X_1, \ldots, X_n$ *from an* $\mathrm{N}(\mu, \sigma^2)$ *distribution. Then the interval*

$$\left[\bar{x}_n - z(\alpha/2)\frac{\sigma}{\sqrt{n}} \quad , \ \bar{x}_n + z(\alpha/2)\frac{\sigma}{\sqrt{n}}\right]$$

is a $100(1-\alpha)\%$ *confidence interval for* μ*.*

Proof. We know that the sample mean $\bar{X}_n = (X_1 + \cdots + X_n)/n$ is normally distributed, $\bar{X}_n \sim \mathrm{N}(\mu, \sigma^2/n)$. Therefore

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

is a standard normal random variable, $Z \sim \mathrm{N}(0,1)$. Therefore

$$\begin{aligned}
1 - \alpha &= P(-z(\alpha/2) < Z < z(\alpha/2)) \\
&= P\left(-z\left(\frac{\alpha}{2}\right) < \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} < z\left(\frac{\alpha}{2}\right)\right) \\
&= P\left(-\bar{X}_n - z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X}_n + z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}\right) \\
&= P\left(\bar{X}_n - z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}\right) \\
&= P(L \le \mu \le U),
\end{aligned}$$

for

$$[L, U] = \left[\bar{X}_n - z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}, \bar{X}_n + z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}\right].$$

According to Definition 23.1 we obtain the $100(1-\alpha)\%$ confidence interval by evaluating the random interval $[L, U]$ on the data, giving

$$[l, u] = \left[\bar{x}_n - z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z\left(\frac{\alpha}{2}\right)\frac{\sigma}{\sqrt{n}}\right].$$

$\square$

<u>Are there different CIs?</u>

If we want a CI on the <u>**mean**</u> and ***know*** the standard deviation…

Z-Test

If we want a CI on the <u>**mean**</u> and ***don't know*** the standard deviation…

T-Test

If we want a CI on the <u>**standard deviation**</u>…

Chi-Squared Test

How do I define a hypothesis test?

- Most often the hypotheses relate to a parameter θ in a model.
  - $H_0 : \theta \in \Omega_0$
  - $H_1 : \theta \in \Omega_1$
- If Ω_0 contains a single element, then H_0 is called a *simple hypothesis*. If it contains multiple elements, it is called a *composite hypothesis*.

What is a p-value?

- The p-value is the probability of obtaining the observed result or a more extreme result if the null hypothesis, H0, is true.

- Conventionally we are looking for a p-value < 0.05 for a test which is at a 95% significance value.

- For a normal distribution we have…

$$\text{p-value} \propto \frac{1}{k} \times \text{C.I.}$$

What abut errors?

- Type I error corresponds to false positive.
- Type II error corresponds to false negative.

Decision based on statistical hypothesis test applied to sample	Truth about population	
	H_0 is true	H_0 is false
Do not reject H_0	No error	**Type II error**
Reject H_0	**Type I error**	No error

Conventionally, the probability of making a Type II error is denoted by β.

The level of statistical significance, α, is equal to the probability of making a Type I error.

BIG STATS POINT…

REJECT THE NULL

or

NOTHING

What is an F-Test?

$$F = \frac{\left(RSS_0 - RSS_1\right) / \left(df_1 - df_0\right)}{RSS_1 / \left(n - df_1\right)} \sim F_{df_1 - df_0, n - df_1}$$

- Used to compare if one model is better than the other!

Are there other checks?

$$C_p = RSS + 2\hat{\sigma}^2 p$$

$$AIC = n \log(RSS/n) + 2p$$

- These quantities and similar (for example BIC) are starting to become increasingly challenged in the literature.

What about non-linear models?

- Some form of non-linear transformation of the parameters.
- Don’t get too confused with what comes later in the course…
- We will see non-normal models but all will be linear.

A more formal definition

$$\underline{Y} \sim f(\underline{x}, \underline{\theta}) + \underline{\epsilon}$$

- Where the function is such that it can not be expressed as a linear combination of the components of theta.

- Assumption that we can approximate with a linear function (use Taylors' theorem)

$$f\left(x_i, \boldsymbol{\theta}\right) \approx f\left(x_i, \boldsymbol{\theta}^k\right) + \sum_j \frac{\partial f\left(x_i, \boldsymbol{\theta}^k\right)}{\partial \theta_j}\left(\theta_j - \theta_j^k\right)$$

An example…

- Holliday model

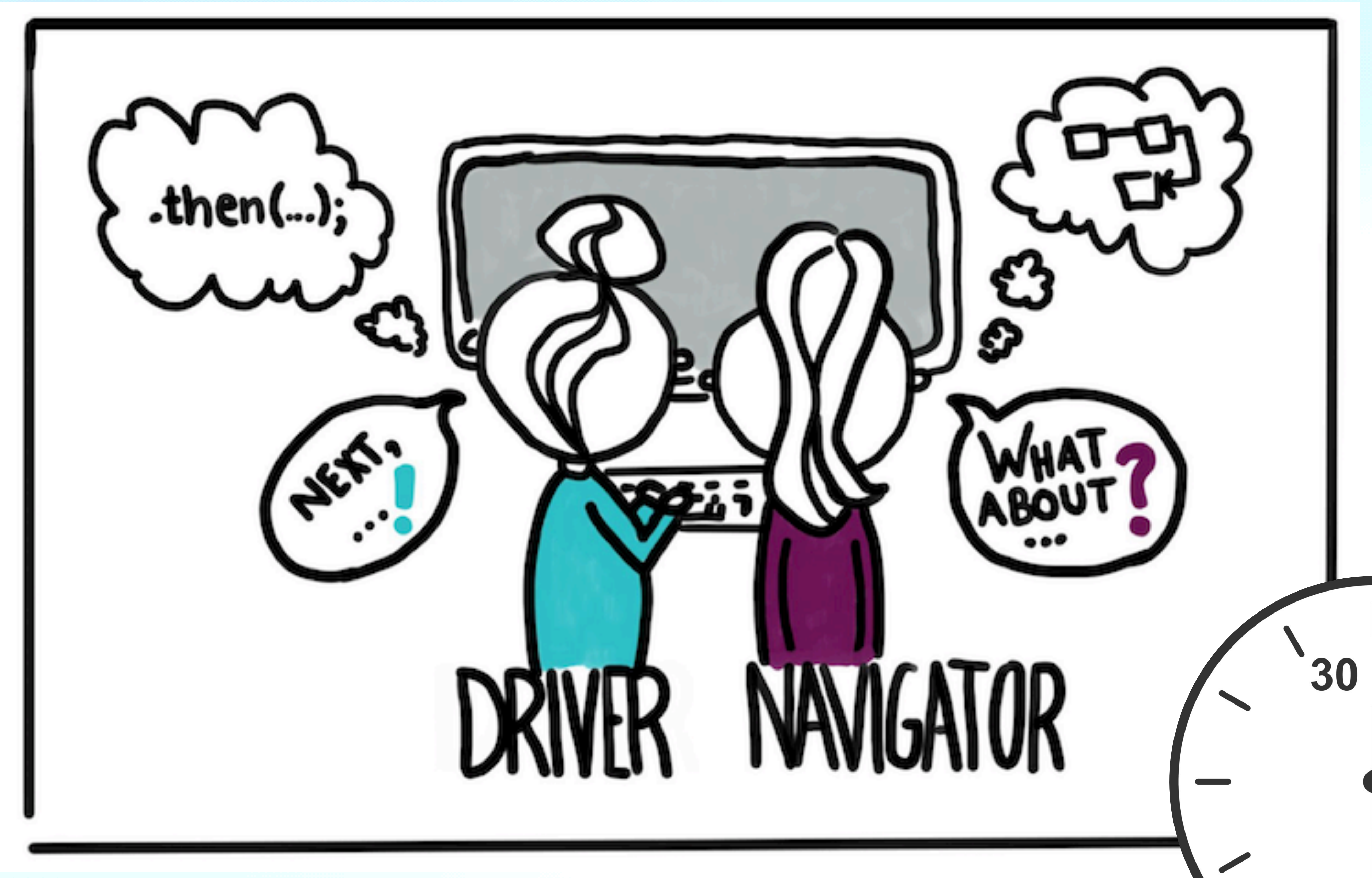
$$y = -\log\left(\beta_0 + \beta_1 x + \beta_2 x^2\right) + \varepsilon$$

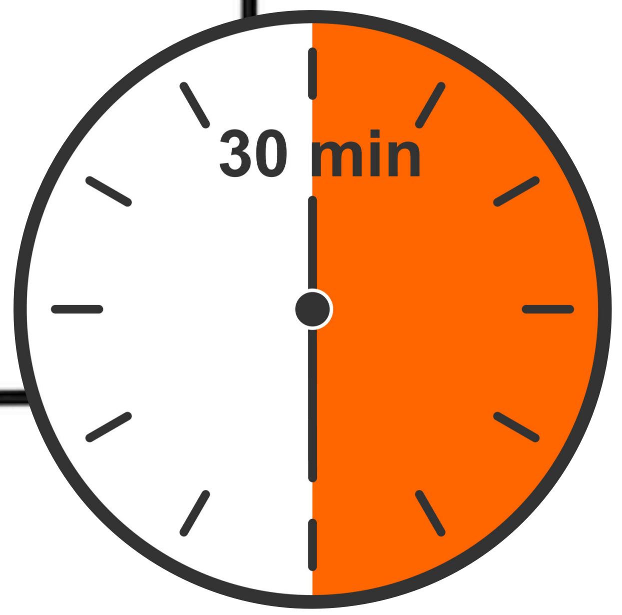
- Minimise a different sum of squares function

$$\sum_{j=1}^{n_i} \left\{ y_{ij} + \log\left(\beta_0 + \beta_1 x + \beta_2 x^2\right) \right\}^2$$

Rest of the tutorial...

- In pairs work on the third notebook.
  Paired programming will continue!





Rest of the week...

3-1: Using the poisons data (available on the SMSTC website), find a transformation for which the model assumptions, when checked by the standard residuals plots, are reasonable.

3-4: A set of data on brown onions is available (on the SMSTC website). Fit a similar type of model as that of the white onions presented within the SMSTC notes. Comment on your findings.

1-1: Return to the data-frame cats in the MASS package from Tutorial 1. Plot and model the relationship between bwt and hwt using a non-linear function?

If you want to consider more theory questions on the CI

1 (of 5). You are asked to investigate inequality in a country's population as measured by the variance, σ^2, of household incomes. Summary statistics from a preliminary survey of $n = 13$ households are given by

$$\bar{x} = n^{-1}\sum_{i=1}^{n} x_i = 24555, \qquad S^2 = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 30365.$$

	p=0	p=0.05	p=0.1	p=0.9	p=0.95	p=1
k=12	0	5.23	6.30	18.55	21.03	∞
k=13	0	5.89	7.04	19.81	22.36	∞
k=14	0	6.57	7.79	21.06	23.68	∞

Table 1: Selected quantiles for the χ^2_k distribution.

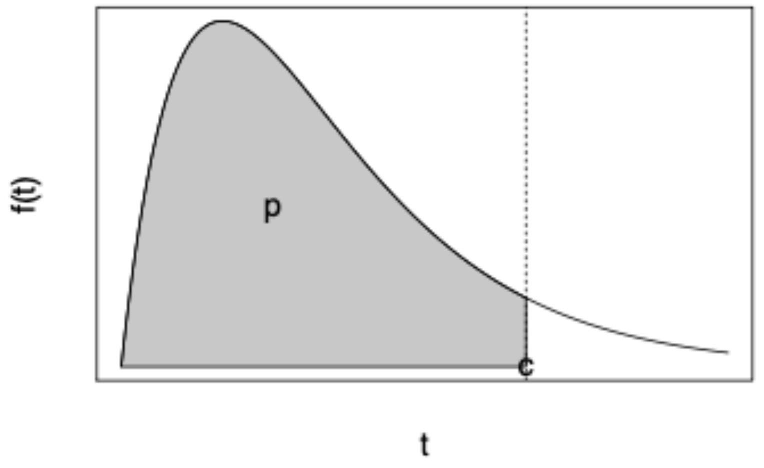


Figure 1: A sketch identifying the probabilities referred to in Table 1.

(a) Provide a set of assumptions under which you can compute a confidence interval for the variance. **[5]**

(b) (i) Given that your assumptions hold, calculate a central 90% confidence interval for the statistic $t(X) = (n-1)S^2/\sigma^2$. **[5]**

(a) If we **assume that the sampled incomes of households are well described as a set of iid normal random variables**, then the statistic

$$t(X) = \frac{(n-1)S^2}{\sigma^2}$$

is distributed as a χ^2_{n-1} random variable. Crucially, this means that probability statements about a χ^2_{n-1} random variable apply to $t(X)$ (before it is actually observed and can also be considered to be a random variable). 5 Marks

(b) (i) Given the distributional assumptions from part a), **the statistic** $(n-1)S^2/\sigma^2$ **is distributed according to a chi-squared distribution with** $n-1$ **degrees of freedom**. Computing the required interval thus requires finding an interval in which a χ^2_{n-1} random variable will fall with probability $\alpha = 0.9 = 90\%/100\%$, i.e.

$$0.9 = P\left(c_1 \leq \chi^2_{n-1} \leq c_2\right),$$

with the added condition that the probabilities of falling beyond either boundary of the interval is the same. This is achieved by **setting** c_1 **and** c_2 **to be the** 0.05^{th} **and** 0.95^{th} **quantiles of the** χ^2_{12} **distribution, which are** 5.23 **and** 21.03 **respectively**. 5 Marks