The ANOVA F-test

We've seen that the test of overall significance in the linear regression can be performed using the additional sum of squares principle. The F-statistic ($\frac{MSR}{MSE}$) in the ANOVA equivalently tests these hypothesis also.

To prove this, we consider the reduced model if $H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_p = 0$ is true, is:

$$y_i = \beta_0 + \varepsilon_i \longrightarrow y_i \sim N(\beta_0, \sigma^2)$$

To calculate SSE_A we need residuals and hence fitted values from the reduced model.

$$e_i = y_i - \hat{\mu}_i = y_i - \hat{\beta}_0$$

It can be shown (do it as an exercise) that the LSE of β_0 is

$$\hat{\beta}_0 = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\therefore \; SSE_A = \vec{e}_A^T \vec{e}_A = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0)^2$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$= SST$$

Thus the test statistic can be written as:

$$t = \frac{(SSE_A - SSE)/\ell}{SSE/(n-p-1)} = \frac{(SST - SSE)/p}{SSE/(n-p-1)} = \frac{SSR/p}{SSE/(n-p-1)} = \frac{MSR}{MSE}$$

(p annotated: $(p+1)-1$)

Example: Bike Rental Data

Suppose that we're interested in determining whether the expected number of bikes rented is the same in all seasons. We do this in the context of the following model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where $x_1 = \begin{cases} 1 & \text{summer} \\ 0 & \text{o.w.} \end{cases}$, $x_2 = \begin{cases} 1 & \text{fall} \\ 0 & \text{o.w.} \end{cases}$, $x_3 = \begin{cases} 1 & \text{winter} \\ 0 & \text{o.w.} \end{cases}$

The test of overall significance

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{vs.} \quad H_A: \beta_j \neq 0 \text{ for some } j$$

actually tests the equality of expected response across the four seasons. This is equivalent to

$$H_0: A\vec{\beta} = \vec{0} \text{ vs. } H_A: A\vec{\beta} \neq 0$$

where $\vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$ and $A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

The test statistic for this test is:

$$t = \frac{MSR}{MSE} = 236.9467$$

The p-value for the test is:

$$\text{p-value} = P(T \geq t) \text{ where } T \sim F_{(3, 10882)}$$
$$= 6.16 \times 10^{-149}$$

$\therefore$ we reject H_0 and conclude that the expected number of bikes rented is not the same in all seasons.

Residual Analysis

We've seen that the linear regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i$$

or, equivalently

$$\vec{y} = X\vec{\beta} + \vec{\varepsilon}$$

makes some assumptions about the ε's, and we should check them:

Specifically: $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

(i) Independence
(ii) Normality
(iii) Expectation is zero
(iv) Constant variance

We evaluate these assumptions using the residuals, which should follow a particular distribution and have a certain behavior if the error assumptions (i)-(iv) are true. If the residual behavior is incompatible with what is expected, it suggests that the error assumptions are invalid.

Recall, we've shown $\vec{e} \sim MVN(\vec{0}, \sigma^2(I-H))$ where I is the $n \times n$ identity matrix and $H = X(X^TX)^{-1}X^T$ is the "hat" matrix. Consequently, we have:

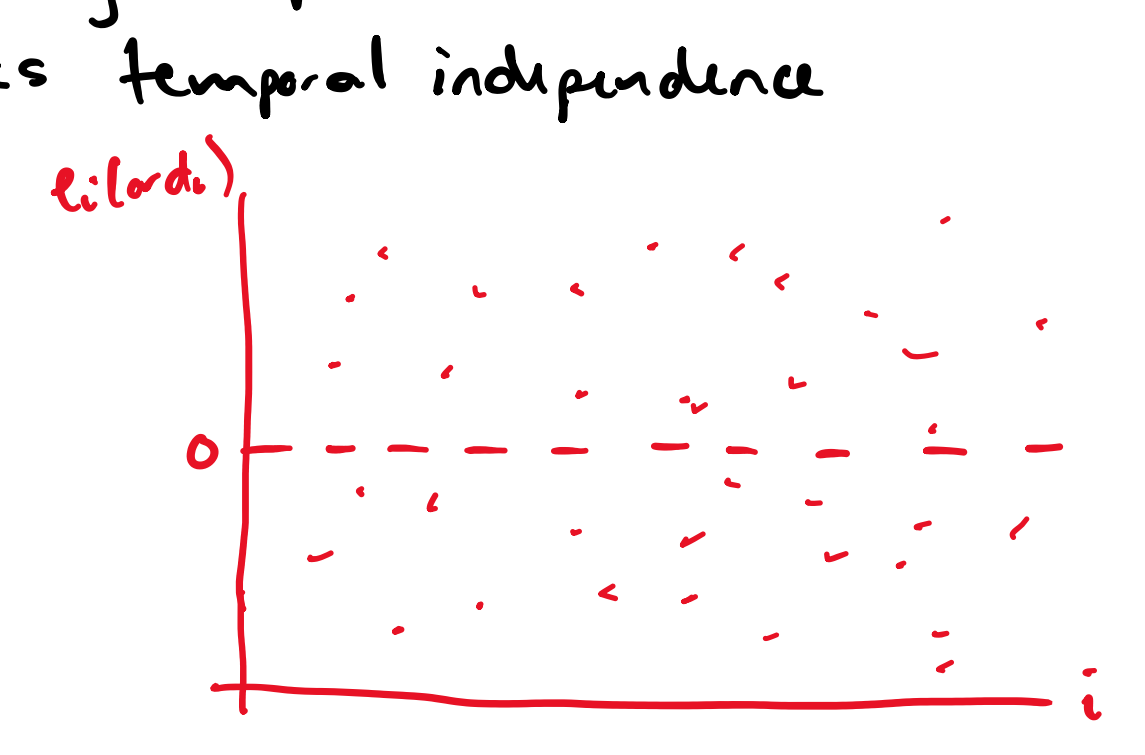
$$e_i \sim N(0, \sigma^2(1-h_{ii}))$$
$$\Updownarrow$$
$$\frac{e_i - 0}{\sigma\sqrt{1-h_{ii}}} \sim N(0,1)$$

We define the "studentized" residual as:

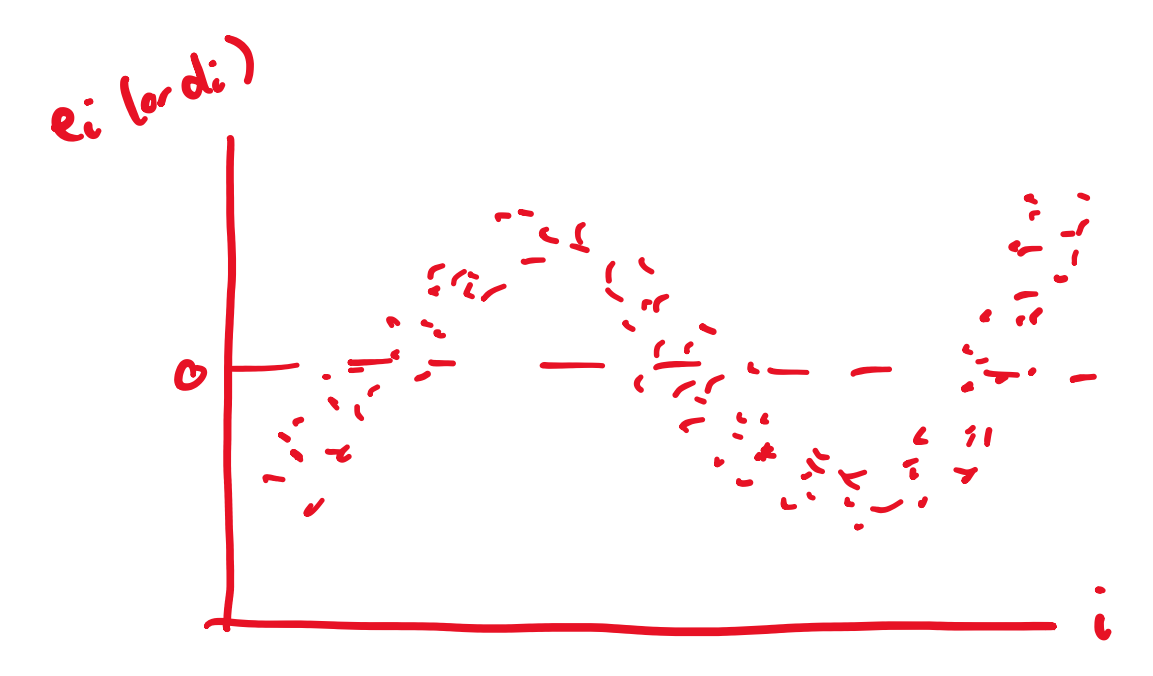
$$d_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

The assumptions (i)-(iv) are evaluated using plots of the residuals (e_i) and/or studentized residuals (d_i).

- Assumption (i) is evaluated with a scatter plot of e_i (or d_i) vs i. The rationale is that temporal dependence may exist, but a scatter plot with a random scattering of points (ie., absence of an obvious relationship) indicates temporal independence

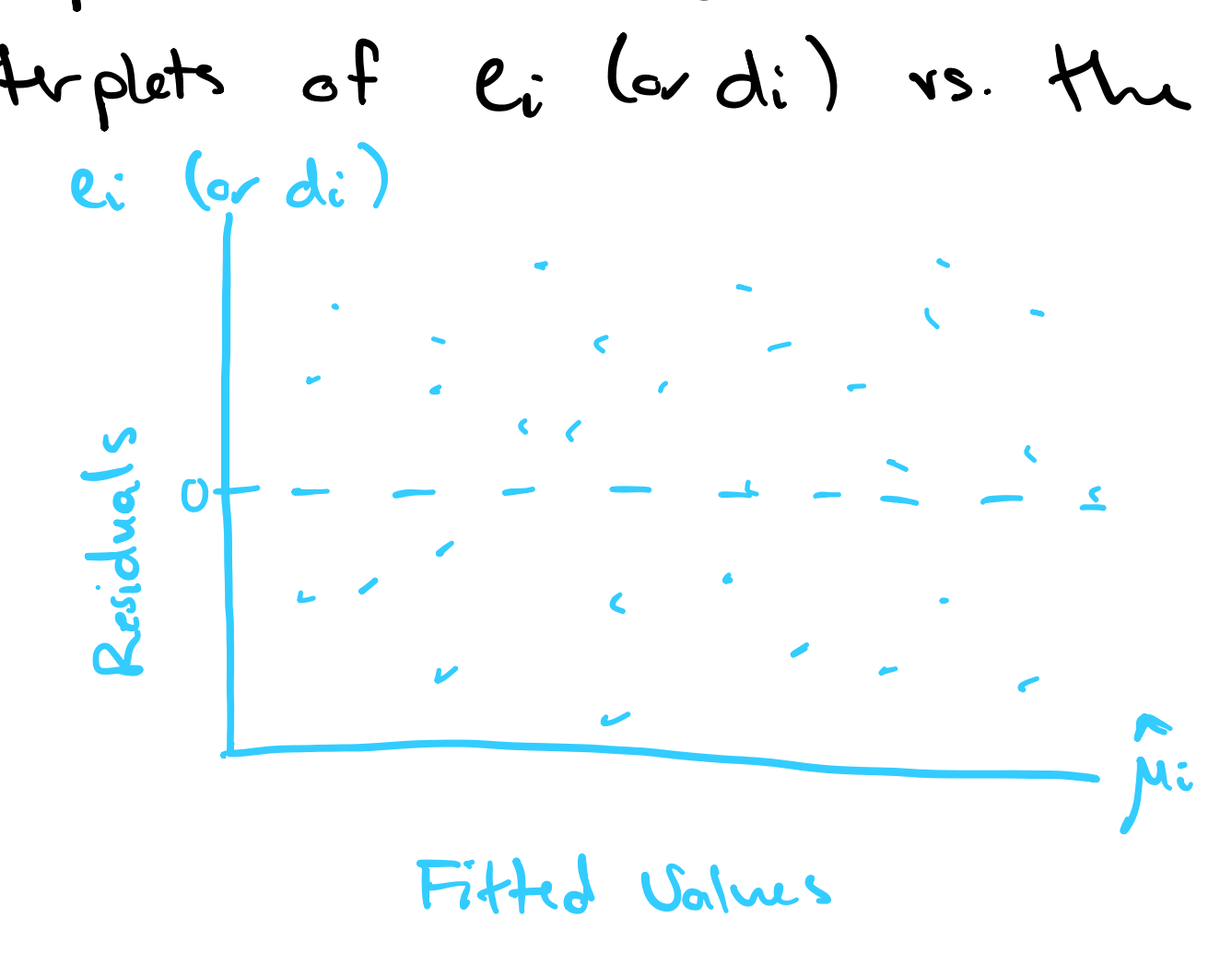


← Good - there are no obvious relationships

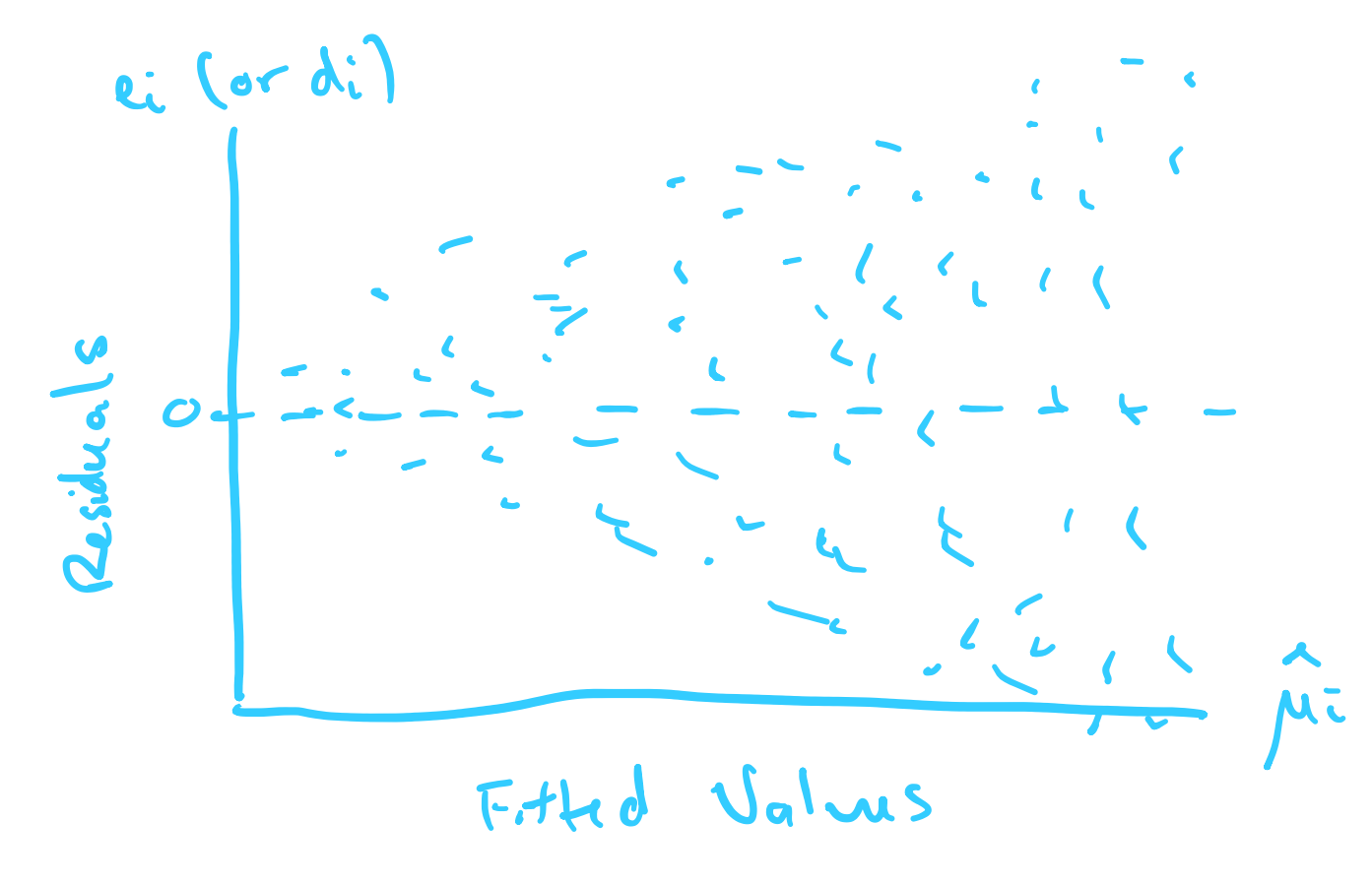


← Bad - these residuals are not independent

- Assumptions (i) and (iv) can be evaluated by looking at scatterplots of e_i (or d_i) vs. the fitted values ($\hat{\mu}_i$).



← Good - no obvious patterns



← Bad - non constant variation

* we could also diagnose these problems with scatterplots of the residuals vs. the explanatory variables (ie, e_i vs. x_{ij} or d_i vs. x_{ij}) but this is less efficient.