LECTURE 5

BOGDAN ICHIM

Assessing the Accuracy of the Model

(via Dispersion Analysis)

Consider the sample mean $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. We want to *understand* $y_i - \overline{y}$.

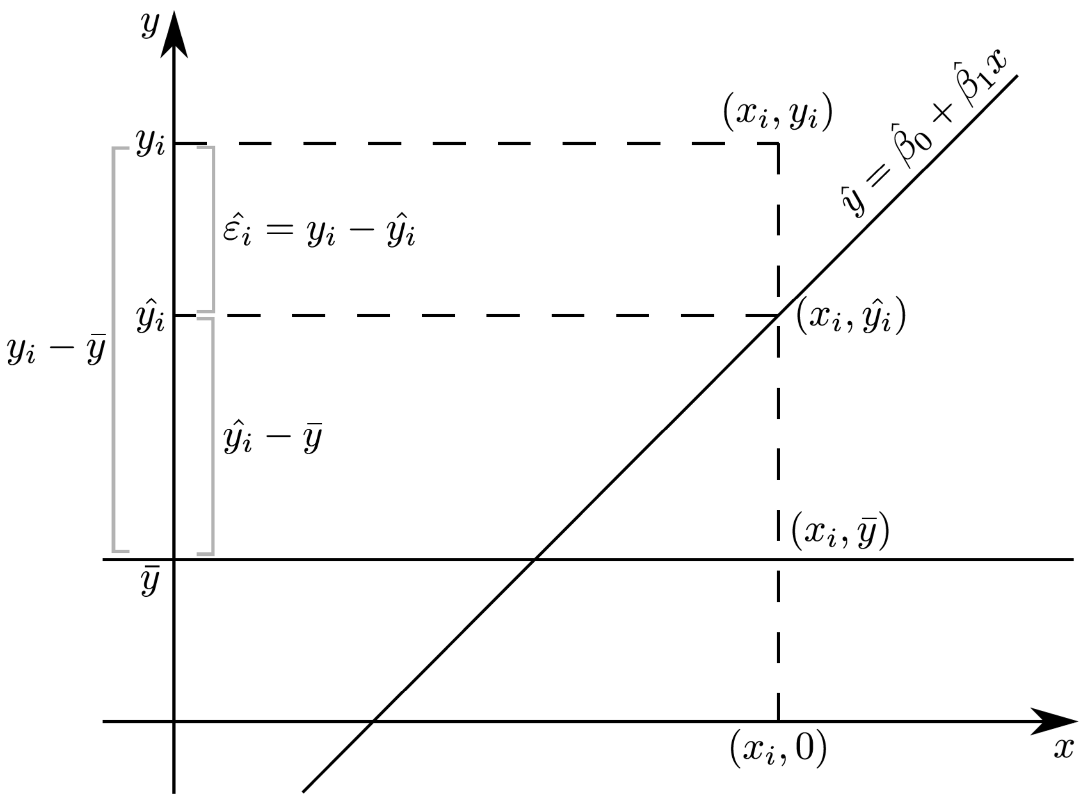

FIGURE 1

Recall from **Lecture 5**:

$$\begin{cases} \left.\begin{array}{l} \widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \\ \widehat{\epsilon}_i = y_i - \widehat{y}_i \end{array}\right\} \Longrightarrow y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{\epsilon}_i \qquad (1) \\ \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 x \qquad (2) \end{cases}$$

Remark 1. We have that

$$y_i - \overline{y} = (y_i - \widehat{y}_i) + (\widehat{y}_i - \overline{y}) \qquad (3)$$

We compute

$$\begin{aligned}
(1) + (2) &\Longrightarrow y_i = \overline{y} + \widehat{\beta}_1 x_i - \widehat{\beta}_1 \overline{x} + \widehat{\epsilon}_i \\
&\Longrightarrow y_i - \overline{y} = \widehat{\beta}_1 (x_i - \overline{x}) + \widehat{\epsilon}_i \qquad \Big| \sum_{i=1}^{n} \\
&\Longrightarrow 0 = \widehat{\beta}_1 \cdot 0 + \sum_{i=1}^{n} \widehat{\epsilon}_i \\
&\Longrightarrow \boxed{\sum_{i=1}^{n} \widehat{\epsilon}_i = 0.}
\end{aligned}$$

Then we have

$$\begin{aligned}
&\Longrightarrow \sum_{i=1}^{n} \widehat{\epsilon}_i = 0 \\
&\Longrightarrow \sum_{i=1}^{n} (y_i - \widehat{y}_i) = 0 \\
&\Longrightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \widehat{y}_i = 0 \\
&\Longleftrightarrow \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i \\
&\Longleftrightarrow \boxed{\overline{y} = \overline{\widehat{y}}.}
\end{aligned}$$

Recall that

$$\begin{aligned}
\frac{\partial \, \mathrm{RSS}}{\partial \widehat{\beta}_1} = 0 &\Longleftrightarrow \widehat{\beta}_0 \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \\
&\Longrightarrow \sum_{i=1}^{n} x_i (y_i - (\underbrace{\widehat{\beta}_0 + \widehat{\beta}_1 x_i}_{\widehat{y}_i})) = 0 \\
&\Longrightarrow \sum_{i=1}^{n} x_i (\underbrace{y_i - \widehat{y}_i}_{\widehat{\epsilon}_i}) = 0 \\
&\Longrightarrow \boxed{\sum_{i=1}^{n} x_i \widehat{\epsilon}_i = 0}
\end{aligned}$$

We also have that

$$
\begin{aligned}
& \widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \quad \Big| \cdot \widehat{\epsilon}_i \\
\Longrightarrow\ & \widehat{y}_i\widehat{\epsilon}_i = \widehat{\beta}_0\widehat{\epsilon}_i + \widehat{\beta}_1 x_i\widehat{\epsilon}_i \quad \Big| \sum_{i=1}^{n} \\
\Longrightarrow\ & \sum_{i=1}^{n} \widehat{y}_i\widehat{\epsilon}_i = \widehat{\beta}_0 \underbrace{\sum_{i=1}^{n} \widehat{\epsilon}_i}_{=0} + \widehat{\beta}_1 \underbrace{\sum_{i=1}^{n} x_i\widehat{\epsilon}_i}_{=0} \\
\Longrightarrow\ & \boxed{\sum_{i=1}^{n} \widehat{y}_i\widehat{\epsilon}_i = 0.}
\end{aligned}
$$

Then

$$
\sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) = \sum_{i=1}^{n}\widehat{\epsilon}_i(\widehat{y}_i - \overline{y}) = \underbrace{\sum_{i=1}^{n}\widehat{\epsilon}_i\widehat{y}_i}_{=0} - \overline{y}\underbrace{\sum_{i=1}^{n}\widehat{\epsilon}_i}_{=0} = 0.
$$

Now we ready to make the final computation. We square both sides of formula (3).

$$
\begin{aligned}
(3)^2 \Longrightarrow\ & (y_i - \overline{y})^2 = (y_i - \widehat{y}_i)^2 + (\widehat{y}_i - \overline{y})^2 + 2(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y}) \quad \Big| \sum_{i=1}^{n} \\
\Longrightarrow\ & \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \underbrace{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}_{\sum_{i=1}^{n}(\widehat{y}_i - \overline{\widehat{y}})^2} + 2\underbrace{\sum_{i=1}^{n}(y_i - \widehat{y}_i)(\widehat{y}_i - \overline{y})}_{=0} \\
\Longrightarrow\ & \underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{\text{variation in the y data}} = \underbrace{\sum_{i=1}^{n}(\widehat{y}_i - \overline{\widehat{y}})^2}_{\text{variation in the } \widehat{y} \text{ data}} + \underbrace{\sum_{i=1}^{n}(\underbrace{y_i - \widehat{y}_i}_{\widehat{\epsilon}_i} - \underbrace{\overline{(y - \widehat{y})}}_{\overline{\widehat{\epsilon}}=0})^2}_{\text{variation in the } \widehat{\epsilon} \text{ data}}
\end{aligned}
$$

We introduce several **notations**

$$
\begin{aligned}
\sum\nolimits_{i=1}^{n}(y_i - \overline{y})^2 &= \text{TSS} = \text{total sum of squares;} \\
&= \text{RSS}_0 = \text{residual sum of squares for a model with 0 predictors;} \\
&= \text{SST} = \text{sum of squares total;} \\
\sum\nolimits_{i=1}^{n}(\widehat{y}_i - \overline{\widehat{y}})^2 &= \text{SSR} = \text{sum of squares explained by regression;} \\
&= \text{ESS} = \text{explained sum of squares;} \\
\sum\nolimits_{i=1}^{n}(y_i - \widehat{y}_i)^2 &= \text{RSS} = \text{residual sum of squares;} \\
&= \text{RSS}_k = \text{residual sum of squares for a model with k predictors;} \\
&= \text{SSE} = \text{sum of squares of errors.}
\end{aligned}
$$

Above we have computed the **Regression Identity (Linear Model Identity)**

$$
\boxed{\text{TSS} = \text{SSR} + \text{RSS}\,.}
$$

The Coefficient of Determination

Note that RSS, MSE and RSE are absolute measurements of the performance of the regression model. But since they are measured in the units of Y, it is not always clear what is a good RSS, MSE or RSE.

Definition 2. The **coefficient of determination** R^2 (pronounced "R-squared") provides an alternative measure for the model's performance, which is independent of the scale of Y. It is defined by the following formula

$$\boxed{\mathrm{R}^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = \frac{\mathrm{SSR}}{\mathrm{TSS}}.}$$

In the case of the linear model $\mathrm{R}^2 \in [0, 1]$ and may be interpreted as the proportion of the total variation in the response variable Y which is explained by the model.

Remark 3. (1) $R^2 = 1$ indicates a perfect fit. We have

$$\begin{aligned} \mathrm{R}^2 = 1 &\Longleftrightarrow \\ \mathrm{RSS} = 0 &\Longleftrightarrow \\ \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = 0 &\Longleftrightarrow \\ & y_i = \widehat{y}_i \quad \forall i = \overline{1, n}. \end{aligned}$$

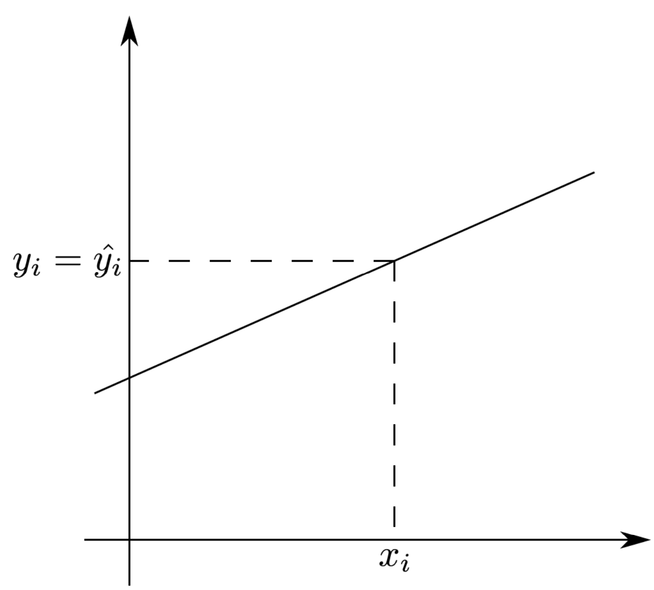

FIGURE 2

(2) If $\mathrm{R}^2 \approx 0$ then the regression did not explain much of the variability of the response.

$$\begin{aligned}\mathrm{R}^2 = 0 &\iff \\ \mathrm{SSR} = 0 &\iff \\ \sum_{i=1}^{n} (\widehat{y}_i - \overline{\widehat{y}})^2 = 0 &\iff \\ \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 = 0 &\iff \\ \widehat{y}_i = \overline{y} \quad \forall i = \overline{1,n}.\end{aligned}$$

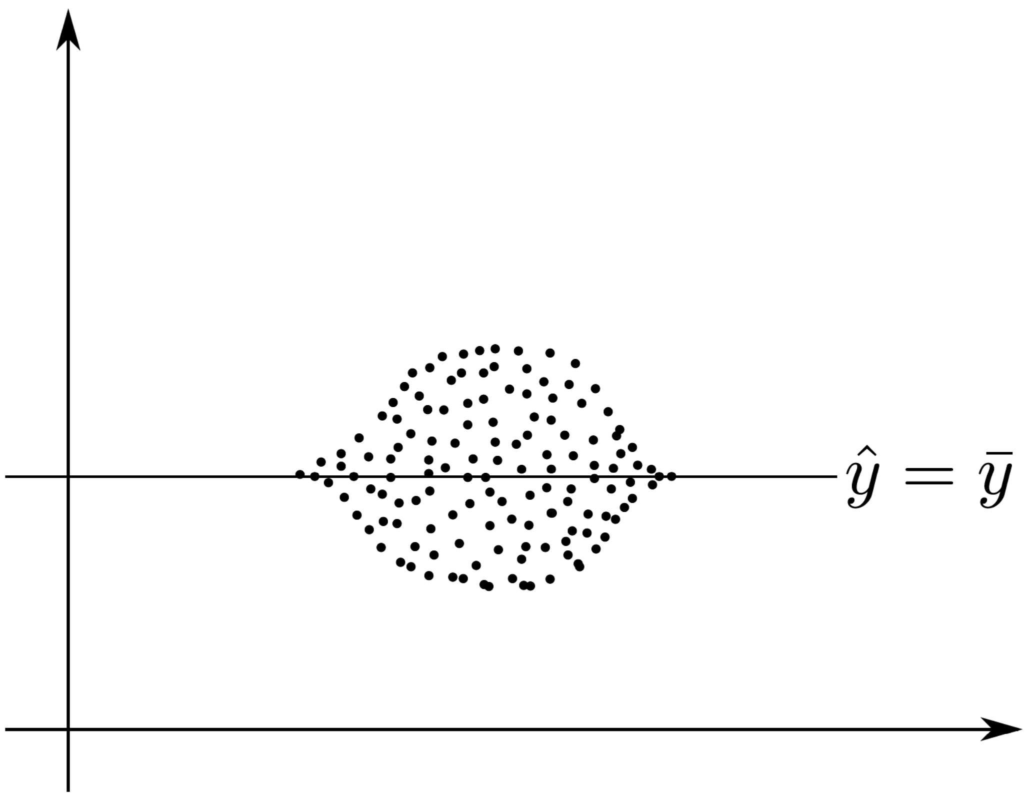

FIGURE 3

(3) R^2 will always increase when more variables are added to the model, no matter how useless those variables are for prediction.

(4) The formulas presented above may vary, depending on the implementation. For example in R if the intercept is removed, then the formula

$$\mathrm{R}_0^2 = 1 - \frac{\mathrm{RSS}}{\sum_{i=1}^{n} y_i^2}$$

is used.

This formula may be seen as anachronic since makes no sense to use it for other machine learning models.

Adjusted Coefficient of Determination

Definition 4. The **adjusted coefficient of determination** $\overline{\mathrm{R}^2}$ (pronounced "adjusted R-squared") is defined by the following formula

$$\boxed{\overline{\mathrm{R}^2} = 1 - \frac{\dfrac{\mathrm{RSS}}{n-k-1}}{\dfrac{\mathrm{TSS}}{n-1}},}$$

where k = the number of explanatory variables = the number of predictors.

Remark 5.
(1) $\overline{\mathrm{R}^2} \leq R^2$.
(2) It is possible that $\overline{\mathrm{R}^2} \leq 0$.
(3) $\overline{\mathrm{R}^2}$ may be used for choosing one model among several models with similar performance. In principle, one should choose the one with the highest $\overline{\mathrm{R}^2}$.
(4) Assume

$$\left.\begin{array}{l} \mathrm{RSS}_{k_1} = \mathrm{RSS}_{k_2} \\ k_1 < k_2 \end{array}\right\} \xRightarrow{\mathrm{TSS}_{k_1} = \mathrm{TSS}_{k_2} \text{ (always)}} \overline{\mathrm{R}^2_{k_1}} > \overline{\mathrm{R}^2_{k_2}}.$$

Occam's razor: One should prefer the model with k_1 predictors since it is the simpler model.

Elements of Hypothesis Testing

We test the **Null Hypothesis**

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

against the **Alternative Hypothesis**

$$H_a : \text{at least one } \ \beta_j \neq 0$$

in order to answer the following.

Question 6. Is at least one of the predictors useful in predicting the response?

This hypothesis test is performed by computing the **F-statistic**

$$F = \frac{\dfrac{\mathrm{RSS}_0 - \mathrm{RSS}_k}{k}}{\dfrac{\mathrm{RSS}_k}{n-k-1}}.$$

When H_0 is true and the errors ϵ_i have a normal distribution, the F-statistic follows an F-distribution (Fisher-Snedecor distribution).

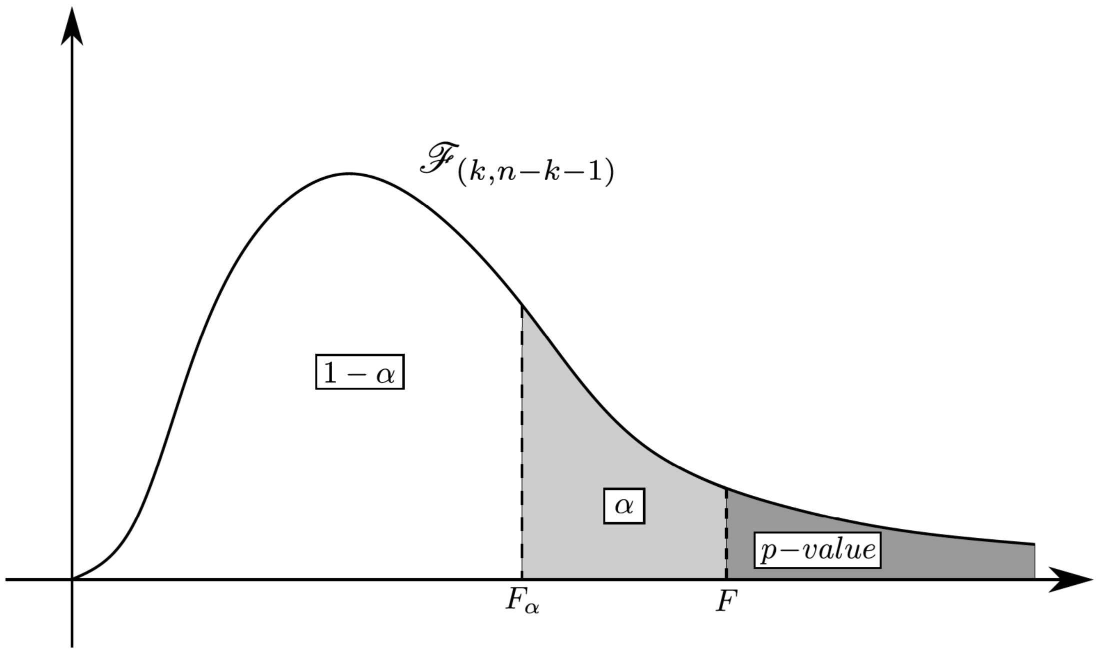

FIGURE 4. Fisher-Snedecor distribution $\mathcal{F}(k,\, n-k-1)$

We consider the following.

α = significance level $= \int_{F_\alpha}^{\infty} \mathcal{F}(k,\, n-k-1)$

$1-\alpha$ = confidence level $= \int_{0}^{F_\alpha} \mathcal{F}(k,\, n-k-1)$

p-value = significance $F = \int_{F}^{\infty} \mathcal{F}(k,\, n-k-1)$

Decisions based on p-values:

- p-value $\leq \alpha \iff F \geq F_\alpha \implies$ reject H_0 and then H_a is true;
- p-value $> \alpha \iff F < F_\alpha \implies H_0$ is true.