Spring 2015 Statistics 151 (Linear Models) : Lecture Five

Aditya Guntuboyina

03 February 2015

1 Last Class

We looked at

1. **Fitted Values**: $\hat{Y} = X\hat{\beta} = HY$ where $H = X(X^TX)^{-1}X^T$. Remember that $\hat{Y}$ is the projection of Y on to the column space of X.
2. **Residuals**: $\hat{e} = Y - \hat{Y} = (I - H)Y$. Remember that $\hat{e}$ is orthogonal to **every vector** in the column space of X. The degrees of freedom of the residuals (also called residual degrees of freedom) equals $n - p - 1$ when X has full column rank. If X does not have full column rank, the residual degrees of freedom equals $n - rank(X)$.

2 The Residual Sum of Squares

The sum of squares of the residuals is called RSS:

$$RSS = \sum_{i=1}^{n} \hat{e}_i^2 = \hat{e}^T\hat{e} = Y^T(I - H)Y = e^T(I - H)e.$$

What is the residual sum of squares where there are no explanatory variables in the model (the model in this case only contains the intercept term)? Ans: $\sum_{i=1}^{n}(y_i - \bar{y})^2$ where $\bar{y} = (y_1 + \cdots + y_n)/n$. This quantity is called the TSS (Total Sum of Squares). The vector $(y_1 - \bar{y}, \ldots, y_n - \bar{y})$ has $n - 1$ degrees of freedom (because this is a vector of size n and it satisfies the linear constraint that sum is zero).

What is the residual sum of squares in simple linear regression (when there is exactly one explanatory variable)? Check that in simple linear regression:

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

where r is the sample correlation between $(y_1, \ldots, y_n)$ and $x = (x_1, \ldots, x_n)$. Because $1 - r^2 \leq 1$, the RSS in simple linear regression is smaller than the RSS in the linear model with no explanatory variables.

In general, RSS decreases (or remains the same) as we add more explanatory variables to the model.

3 The Coefficient of Determination

This is more commonly referred to as R-squared. It is defined as

$$R^2 := 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

What is the point of this definition? One of the goals of regression is to predict the value of the response variable for future subjects. For this purpose, we are given data $y_1, \dots, y_n$ and x_{ij} for $i = 1, \dots, n$ and $j = 1, \dots, p$.

Suppose we are told to predict the response of a future subject **without** using any of the data on the explanatory variables i.e., we are only supposed to use $y_1, \dots, y_n$. In this case, it is obvious that our prediction for the next subject would be $\bar{y}$. The error of this method of prediction on the ith subject is $y_i - \bar{y}$ and the total error is therefore:

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

If, on the other hand, we are allowed to use data on the explanatory variables, then the prediction will be given by

$$\hat{\beta}_0 + x_1\hat{\beta}_1 + \dots + x_p\hat{\beta}_p.$$

The error of this prediction on the ith subject is the residual $\hat{e}_i$ and the total error is the Residual Sum of Squares:

$$RSS = \sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Because using the explanatory variables is always better than not using them, RSS is always smaller than or equal to TSS (**this fact is crucially reliant on the fact that there is an intercept in our model**).

If RSS is very small compared to TSS, it means that the explanatory variables are really useful in predicting the response. On the other hand, if RSS is only a little bit smaller than TSS, it means that we are not really gaining much by using the explanatory variables. The quantity R^2 tries to quantify how useful the explanatory variables are in predicting the response. It always lies between 0 and 1

1. **If R^2 is high, it means that RSS is much smaller compared to TSS and hence the explanatory variables are really useful in predicting the response.**
2. **If R^2 is low, it means that RSS is only a little bit smaller than TSS and hence the explanatory variables are not useful in predicting the response.**

It must be noted that R^2 is an *in-sample* measure of prediction accuracy. In other words, the predictions are checked on the subjects already present in the sample (as opposed to checking them on new subjects). In particular, these are the same subjects on whom the model is fitted (or trained), so R^2 can be made to look very good by fitting models with lots of parameters.

Because RSS decreases when more parameters are added to the model, R^2 increases when more parameters are added to the model.

4 Expected Value of the RSS

What is the expected value of RSS?

$$\mathbb{E}(RSS) = \mathbb{E}e^T(I-H)e = \mathbb{E}\left(\sum_{i,j}(I-H)(i,j)e_ie_j\right) = \sum_{i,j}(I-H)(i,j)(\mathbb{E}e_ie_j)$$

Because $\mathbb{E}(e_ie_j)$ equals 0 when $i \neq j$ and σ^2 otherwise, we get

$$\mathbb{E}(RSS) = \sigma^2\sum_{i=1}^{n}(I-H)(i,i) = \sigma^2\left(n - \sum_{i=1}^{n}H(i,i)\right)$$

The sum of the diagonal entries of a square matrix is called its trace i.e, $tr(A) = \sum_i a_{ii}$. We can therefore write

$$\mathbb{E}(RSS) = \sigma^2\left(n - tr(H)\right).$$

A very important fact about trace is $tr(AB) = tr(BA)$. Thus

$$tr(H) = tr(X(X^TX)^{-1}X^T) = tr((X^TX)^{-1}X^TX) = tr(I_{p+1}) = p+1.$$

We proved

$$\mathbb{E}(RSS) = \sigma^2(n-p-1).$$

An *unbiased* estimator of σ^2 is therefore given by

$$\hat{\sigma^2} := \frac{RSS}{n-p-1}.$$

And σ is estimated by

$$\hat{\sigma} := \sqrt{\frac{RSS}{n-p-1}}.$$

This $\hat{\sigma}$ is called the **Residual Standard Error**.

5 Standard Errors of $\hat{\beta}$

We have seen that $\mathbb{E}\hat{\beta} = \beta$ and that $Cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}$. The standard error of $\hat{\beta}_i$ is therefore defined as $\hat{\sigma}$ multiplied by the square root of the ith diagonal entry of $(X^TX)^{-1}$. The standard error gives an idea of the accuracy of $\hat{\beta}_i$ as an estimator of β_i. These standard errors are part of the R output for the summary of the linear model.

6 Standardized or Studentized Residuals

The residuals $\hat{e}_1, \ldots, \hat{e}_n$ have different variances. Indeed, because $Cov(\hat{e}) = \sigma^2(I-H)$, we have

$$var(\hat{e}_i) = \sigma^2(1-h_{ii})$$

where h_{ii} denotes the ith diagonal entry of H. Because h_{ii} can be different for different i, the residuals have different variances.

The variance can be standardized to 1 if we divide the residuals by $\sigma\sqrt{1-h_{ii}}$. But because σ is unknown, one divides by $\hat{\sigma}\sqrt{1-h_{ii}}$ and we call the resulting quantities **Standardized Residuals** or **Studentized Residuals**:

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}.$$

The standardized residuals $r_1, \ldots, r_n$ are very important in regression diagnostics. Various assumptions on the unobserved errors $e_1, \ldots, e_n$ can be checked through them.