Fall 2013 Statistics 151 (Linear Models) : Lecture Seven

Aditya Guntuboyina

19 September 2013

1 Last Class

We looked at

1. **Fitted Values**: $\hat{Y} = X\hat{\beta} = HY$ where $H = X(X^TX)^{-1}X^TY$. $\hat{Y}$ is the projection of Y onto the column space of X.
2. **Residuals**: $\hat{e} = Y - \hat{Y} = (I - H)Y$. $\hat{e}$ is orthogonal to **every vector** in the column space of X. The degrees of freedom of the residuals is $n - p - 1$.
3. **Residual Sum of Squares**: $RSS = \sum_{i=1}^n \hat{e}_i^2 = \hat{e}_i^T\hat{e}_i Y^T(I - H)Y$. RSS decreases when more explanatory variables are added to the model.
4. **Total Sum of Squares**: $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$. Can be thought of the RSS in a linear model with no explanatory variables (only the intercept term).
5. **Coefficient of Determination or Multiple** R^2: Defined as $1-(RSS/TSS)$. Always lies between 0 and 1. High value means that the explanatory variables are useful in explaining the response and low value means that the explanatory variables are not useful in explaining the response. R^2 increases when more explanatory variables are added to the model.

2 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**.

3 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.

4 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.

5 Normality of the Errors

Everything that we did so far was only under the assumption that the errors $e_1, \ldots, e_n$ were uncorrelated, had mean zero and variance σ^2. But if we want to test hypotheses about or if we want confidence intervals for linear combinations of β, we need distributional assumptions on the errors.

For example, consider the problem of testing the null hypothesis $H_0 : \beta_1 = 0$ against the alternative hypothesis $H_1 : \beta_1 \neq 0$. If H_0 were true, this would mean that the first explanatory variable has no role (in the presence of the other explanatory variables) in determining the expected value of the response. An obvious way to test this hypothesis is to look at the value of $\hat{\beta}_1$ and then to reject H_0 if $|\hat{\beta}_1|$ is large. But how large? To answer this question, we need to understand how $\hat{\beta}_1$ is distributed under the null hypothesis H_0. Such a study requires some distributional assumptions on the errors $e_1, \ldots, e_n$.

The mose standard assumption on the errors is that $e_1, \ldots, e_n$ are independently distributed according to the normal distribution with mean zero and variance σ^2. This is written in multivariate normal notation as $e \sim N(0, \sigma^2 I_n)$.

6 The Multivariate Normal Distribution

A random vector $U = (U_1, \ldots, U_p)^T$ is said to have the multivariate normal distribution with parameters μ and Σ if the joint density of $U_1, \ldots, U_p$ is given by

$$(2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(\frac{-1}{2}(u-\mu)^T\Sigma^{-1}(u-\mu)\right) \qquad \text{for } u \in \mathbb{R}^d.$$

Here $|\Sigma|$ denotes the determinant of Σ.

We use the notation $U \sim N_p(\mu, \Sigma)$ to express that U is multivariate normal with parameters μ and Σ.

Example 6.1. *An important example of the multivariate normal distribution occurs when $U_1, \ldots, U_p$ are independently distributed according to the normal distribution with mean 0 and variance σ^2. In this case, it is easy to show $U = (U_1, \ldots, U_p)^T \sim N_p(0, \sigma^2 I_p)$.*

The most important properties of the multivariate normal distribution are summarized below:

1. When $p = 1$, this is just the usual normal distribution.
2. **Mean and Variance-Covariance Matrix**: $\mathbb{E}U = \mu$ and $Cov(U) = \Sigma$.
3. **Independence of linear functions can be checked by multiplying matrices**: Two linear functions AU and BU are independent if and only if $A\Sigma B^T = 0$. In particular, this means that U_i and U_j are independent if and only if the (i, j)th entry of Σ equals 0.
4. Every linear function is also multivariate normal: $a + AU \sim N(a + A\mu, A\Sigma A^T)$.
5. Suppose $U \sim N_p(\mu, I)$ and A is a $p \times p$ symmetric and idempotent (symmetric means $A^T = A$ and idempotent means $A^2 = A$) matrix. Then $(U-\mu)^T A(U-\mu)$ has the chi-squared distribution with degrees of freedom equal to the rank of A. This is written as $(U-\mu)^T A(U-\mu) \sim \chi^2_{rank(A)}$.

7 Normal Regression Theory

We assume that $e \sim N_n(0, \sigma^2 I_n)$. Equivalently, $e_1, \ldots, e_n$ are independent normals with mean 0 and variance σ^2.

Under this assumption, we can calculate the distributions of many of the quantities studied so far.

7.1 Distribution of Y

Since $Y = X\beta + e$, we have $Y \sim N_n(X\beta, \sigma^2 I_n)$.

7.2 Distribution of $\hat{\beta}$

Because $\hat{\beta} = (X^T X)^{-1} X^T Y$ is a linear function of Y, it has a multivariate normal distribution. We already saw that $\mathbb{E}\hat{\beta} = \beta$ and $Cov(\hat{\beta}) = \sigma^2(X^T X)^{-1}$. Thus $\hat{\beta} \sim N_{p+1}(\beta, \sigma^2(X^T X)^{-1})$.

7.3 Distribution of Fitted Values

$\hat{Y} = HY$. Thus $\mathbb{E}\hat{Y} = H\mathbb{E}(Y) = HX\beta = X\beta$. Also $Cov(\hat{Y}) = Cov(HY) = \sigma^2 H$. Therefore $\hat{Y} \sim N_n(X\beta, \sigma^2 H)$.

7.4 Distribution of Residuals

$\hat{e} = (I - H)Y$. We saw that $\mathbb{E}\hat{e} = 0$ and $Cov(\hat{e}) = \sigma^2(I - H)$. Therefore $\hat{e} \sim N_n(0, \sigma^2(I - H))$.

7.5 Independence of residuals and $\hat{\beta}$

Recall that if $U \sim N_p(\mu, \Sigma)$, then AU and BU are independent if and only if $A\Sigma B^T$.

This can be used to verify that $\hat{\beta} = (X^T X)^{-1} X^T Y$ and $\hat{e} = (I - H)Y$ are independent. To see this, observe that both are linear functions of $Y \sim N_n(X\beta, \sigma^2 I)$. Thus if $A = (X^T X)^{-1} X^T Y$, $B = (I - H)$ and $\Sigma = \sigma^2 I$, then

$$A\Sigma B^T = \sigma^2 (X^T X)^{-1} X^T (I - H) = \sigma^2 (X^T X)^{-1} (X^T - X^T H)$$

Because $X^T H = (HX)^T = X^T$, we conclude that $\hat{\beta}$ and $\hat{e}$ are independent.

Also check that $\hat{Y}$ and $\hat{e}$ are independent.

7.6 Distribution of RSS

Recall

$$RSS = \hat{e}^T \hat{e} = Y^T (I - H) Y = e^T (I - H) e.$$

So

$$\frac{RSS}{\sigma^2} = \left(\frac{e}{\sigma}\right)^T (I - H) \left(\frac{e}{\sigma}\right).$$

Because $e/\sigma \sim N_n(0, I)$ and $I - H$ is symmetric and idempotent with rank $n - p - 1$, we have

$$\frac{RSS}{\sigma^2} \sim \chi^2_{n-p-1}.$$