Topic 3
Linear Regression and Classification

Instructor: Farid Alizadeh

January 29, 2020

1 The Basic Linear Model: The Frequentist View

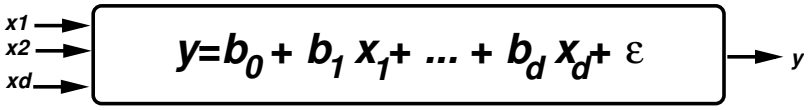
The linear regression model is very specific regression with the following assumptions:

1. The response variable y is numerical, and the feature variables $x_1, \ldots, x_d$ are either numerical or categorical.

2. If all feature variables x_i are numerical, then the relationship between y and x_i is *linear*:

$$y = b_1 x_1 + b_2 x_2 + \cdots + b_d x_d + \epsilon$$

   where $\epsilon \sim N(0, \sigma)$. This means that the relationship is essentially linear, but a random error ϵ is added to the underlying relationship. It is assumed that the random error follows the normal distribution with variance σ^2 which is fixed, that is it does not depend on values of features x_i.



3. If a variable x is categorical with levels say $\{a, b, \ldots, k\}$ then we can create dummy variables as follows: We choose one of the levels as the *base level*. Next we create *dummy variable*, z_a for level a, z_b for level b, and so on, except we *do not* make any dummy variables for the base level we chose. Each dummy variable z_i can only attain values one (if x is at level i) or zero (if x is not at level i). So, at most one of $z_i = 1$ and the rest are zero. If the x is at base level, then all $z_i = 0$. For example, if x represents "color" with levels blue, red and orange, and if we choose "blue" as the base level, then let z_{red} and z_{orange} are the dummy variables. A red item will have $z_{\text{red}} = 1$ and $z_{\text{orange}} = 0$. For a blue item, since blue is the base level, $z_{\text{red}} = z_{\text{orange}} = 0$.

4. We need to estimate the unknown parameters $b_1, b_2, \ldots, b_d$. In addition we would like to estimate the distribution of the estimated parameters and the distribution of the response variable y for a given feature vector $\mathbf{x} = (x_1, \ldots, x_d)$.

5. The linear model need not be a linear function of the original variables. It could also be a linear function depend on *functions* of feature variables. For example, if x_1 and x_2 are the original variables then $y = b_1 + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_2^2 + b_5 x_1 x_2$ is a legitimate linear model. In general as long as the model is linear with respect to the parameters, it is a linear model.

6. To estimate the parameters b_i we have a set of *training data* where N examples are given. For each example, both the y value and the $x_1, x_2, \ldots, x_d$ are given. So, like other methods, the training data set is usually organized in a table like this:

y	x_1	x_2	$\ldots$	x_d
y_1	x_{11}	x_{12}	$\ldots$	x_{1d}
y_2	x_{21}	x_{22}	$\ldots$	x_{2d}
$\vdots$	$\vdots$	$\vdots 2$	$\ddots$	$\vdots$
y_N	x_{N1}	x_{N2}	$\ldots$	x_{Nd}

As usual the data are assumed to be i.i.d. We represent the part of the of the data containing the feature vectors by a matrix X. Note that this matrix contains all the *features*, not just the original variables. So it contains columns corresponding to features that are actually used in the model. For instance, if there are two variables x_1, x_2, and the model is $y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_1 x_2$, then this matrix has a column of all ones, and one column for each of features, x_1, x_2, x_1^2 and $x_1 x_2$. So the matrix X, with these columns added to it is an $N \times d$ matrix, with d the number of features (including the constant or *bias variable*). Notice that if $\hat{\mathbf{b}} = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_d)$ is our estimate the vector of coefficients then $\hat{y} = X\hat{\mathbf{b}}$ is our prediction for y values of the training data. So $\|y - X\hat{\mathbf{b}}\|^2$ is the square of the (Euclidean) norm of the discrepancy between observed and predicted values in the training data.

1.1 The likelihood function for linear models

Let the vector $\mathbf{b} = (b_1, b_2, \ldots, b_d)$. For a new point $\mathbf{x}_{\text{new}} = (x_1, \ldots, x_n)$, since the error term ϵ is normal with mean zero, it follows that the predicted $\hat{y}_{\text{new}}$ value is a random variable that also follows the normal distribution with mean $\mu_y = b_1 x_1 + b_2 x_2 + \cdots + b_d x_d$, and the same standard deviation σ as ϵ. So the

pdf for the observed value of y_{new} is given by:

$$f(y|\mathbf{b},\sigma,\mathbf{x}_{new}) = \phi(b_1x_1 + b_2x_2 + \cdots + b_dx_d,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{y - b_1x_1 - b_2x_2 - \ldots - b_dx_d}{\sigma}\right)^2\right)$$

First, from the data we calculate the likelihood function. Unknown parameters here are the coefficients $\mathbf{b}$ and σ^2. Then if $\mathbf{y} = (y_1,\ldots,y_N)$ the likelihood function is given by:

$$\begin{aligned}\text{Lik}(\mathbf{b},\sigma^2|\mathbf{y},X) = \phi(y_1, b_1x_{11} + b_2x_{12} + \cdots + b_dx_{1d},\sigma)\times\\ \phi(y_2, b_1x12 + b_2x_{22} + \cdots + b_dx_{2d},\sigma)\times\\ \cdots \times \phi(y_N, b1x_{N1} + b_2x_{N2} + \cdots + b_dx_{Nd},\sigma)\end{aligned}$$

For this problem it is more convenient to consider the negative log likelihood:

$$\begin{aligned}-\log\text{Lik}(\mathbf{b},\sigma^2|\mathbf{y},X) &= N\log\sqrt{2\pi} + N\log\sigma + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - b_1x_1 - b_2x_2 - \cdots - b_dx_N)^2\\ &= \text{Const.} + N\log(\sigma) + \frac{1}{2\sigma^2}(\mathbf{y} - X\mathbf{b})^\top(\mathbf{y} - X\mathbf{b})\end{aligned}$$

The maximum likelihood solution (after dropping constant factors) is obtained by taking derivatives with respect to $\mathbf{b}$ and with respect to σ^2 and setting them equal to zero[1]:

$$\begin{aligned}\nabla_{\mathbf{b}}\log\text{Lik}(\mathbf{b}.\sigma^2|X,\mathbf{y}) &= 0\\ \frac{\partial}{\partial\sigma^2}\log\text{Lik}(\mathbf{b}.\sigma^2|X,\mathbf{y}) &= 0\end{aligned}$$

1.2 The Maximum Likelihood Estimation $\mathbf{b}_{ML}$ of Coefficients

The vector of derivatives of the log likelihood function with respect to $\mathbf{b}$ is:

$$\nabla_{\mathbf{b}}\left(-\log\text{Lik}(\mathbf{b},\sigma|X,\mathbf{y})\right) = \frac{2}{\sigma^2}\nabla_{\mathbf{b}}\left((\mathbf{y} - X\mathbf{b})^\top(\mathbf{y} - X\mathbf{b})\right) = \frac{2}{\sigma^2}(X^\top\mathbf{y} - X^\top X\mathbf{b}).$$

Setting this quantity equal to zero we can solve for $\mathbf{b}$ and get:

$$X^\top\mathbf{y} - X^\top X\mathbf{b} = 0 \quad\Rightarrow\quad \mathbf{b} = \mathbf{b}_{ML} = (XX^\top)^{-1}X^\top\mathbf{y}$$

Before we move on let's make a few observations about the computation of the maximum likelihood coefficients $\mathbf{b}_{ML}$.

[1]The *gradient* of a function $f(\mathbf{b})$, that is $\nabla_{\mathbf{b}}f(\mathbf{b})$ is a row vector of derivatives of $f(\mathbf{b})$. So the i^{th} entry of this vector is $\frac{\partial}{\partial b_i}f(\mathbf{b})$.

1. When taking derivatives with respect to the coefficients $\mathbf{b}$ and set them equal to zero, the other parameter σ^2 is not involved in the resulting equation. This simplifies matters a lot, and the maximum likelihood $\mathbf{b}_{\text{ML}}$ does not depend on σ^2.

2. The logarithm of the log likelihood function is the *sum of squares of residuals*. Note that for each point in the training set: $(y_i, x_{i1}, \ldots, x_{id})$ the *residual* at that point is $\epsilon_i(\mathbf{b}) = y_i - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_d x_{id}$. This is the difference between the *predicted value* of the response variable $b_1 x_{i1} + b_2 x_{i2} + \cdots + b_d x_{id}$ and its observed value y_i. Notice also that $\boldsymbol{\epsilon}(\mathbf{b})$ is a function of the coefficients $\mathbf{b}$.

   The maximum likelihood estimate $\mathbf{b}_{\text{ML}}$ arises when minimizing the sum of squares of residuals:

$$\boldsymbol{\epsilon}(\mathbf{b})^\top \boldsymbol{\epsilon}(\mathbf{b}) = \sum_{i=1}^{N} \left(y_i - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_d x_{id}\right)^2 = (\mathbf{y} - X\mathbf{b})^\top (\mathbf{y} - X\mathbf{b}).$$

   So in this case the maximum likelihood solution coincides with the *least squares solution*. We represent the residuals $\hat{\epsilon}_i = \hat{\epsilon}_i(\mathbf{b}_{\text{ML}}) = \left(y_i - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_d x_{id}\right)^2$ and $\hat{\boldsymbol{\epsilon}}$ is the vector of $\hat{\epsilon}_i$. Notice that under the frequentist point of view $\hat{\boldsymbol{\epsilon}}$ are *realizations* of the random variable ϵ, the error added to the linear function $\epsilon = y - \mathbf{b}^\top \mathbf{x}$.

3. The system of equations that arises in determining $\mathbf{b}$ is a *linear system of equations*. As a result we can solve for $\mathbf{b}$ very easily and indeed the closed form formula $\mathbf{b}_{\text{ML}} = (X^\top X)^{-1} X^\top \mathbf{y}$.

4. Now that we have the maximum likelihood estimate of coefficients we can now use it to estimate the y_{new} value of a new item with feature vector $\mathbf{x}_{\text{new}}$:

$$\hat{y}_{\text{new}} = \mathbf{b}_{\text{ML}}^\top \mathbf{x}_{\text{new}}$$

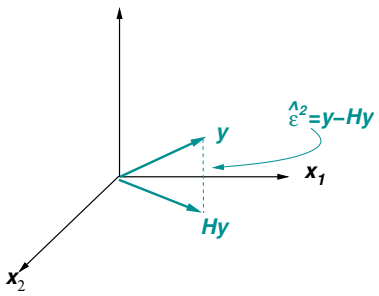
We are very fortunate here in that the linear regression model with normal error results in a *linear* system of equations in b_i alone, and so we can find a closed form linear formula for the maximum likelihood estimator $\mathbf{b}_{\text{ML}}$. Even for the variance, we will see that we will end up with a linear equation (once the $\mathbf{b}_{\text{ML}}$ is known.) This situation is an exception not the rule. In most other models, the maximum likelihood estimates are not this clean. In fact, in most cases the system of equations that arises from setting the partial derivatives of the likelihood function (or the log likelihood function) equal to zero, results in a *nonlinear* system of equations. In such cases the main method for solving for the unknown parameters is through *numerical methods* which produces successive approximations that ultimately converge to the solution. We will see this phenomenon for logistic regression and neural nets in future lectures.

The Hat matrix

Define the matrix

$$H = X(X^\top X)^{-1}X^\top.$$

H, as a linear transformation, *projects* a vector y to the space spanned by the columns X. It is a $N \times N$ square and symmetric matrix.



From this, and noting that $\hat{b} = (X^\top X)^{-1}X^\top y$ we have:

$$\hat{y} = X\hat{b} = Hy$$
$$\hat{\epsilon} = y - X\hat{b} = y - H\hat{y} = (I - H)y$$

where $\hat{y}$ are the *predicted* values of y_i for training data. H is called the *hat matrix* (since it puts the "hat" on the $\hat{y}$).

The hat matrix H and the matrix I−H have some important properties which makes analysis of distributions of various items much more straightforward.

First, Note that

$$H^2 = X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top = X(X^\top X)^{-1}X^\top X = H$$

Now, H is a symmetric matrix so all its eigenvalues are real. Also, note that if λ is an eigenvalue of a matrix A, that is $Aq = \lambda q$, then λ^2 is an eigenvalue of A^2. This is so because $A^2q = A(Aq) = \lambda Aq = \lambda^2 q$. So if $H^2 = H$ then they have the same eigenvalues, that is all eigenvalues of H satisfy $\lambda^2 = \lambda$. So eigenvalues of H are all either 1 or 0. Also notice that

$$\text{Trace}(H) = \text{Trace}\Big(X(X^\top X)^{-1}X^\top\Big) = \text{Trace}\Big(X^\top X(X^\top X)^{-1}\Big) = \text{Trace}(I_d) = d$$

Recall[2] that trace of a matrix equals the sum of its eigenvalues. So since $\text{Trace}(H) = d$ and H is an $N \times N$ matrix, it follows that

$$H = q_1 q_1^\top + \cdots + q_d q_d^\top \quad \text{and} \quad \text{rank}(H) = d$$

where the q_i are the orthonormal eigenvectors of H corresponding to eigenvalue 1.

[2]We used the fact that $\text{Trace}(AB) = \text{Trace}(BA)$. Also I_d is the $d \times d$ identity matrix.

Also, the matrix $I-H$ has exactly the same eigenvectors as H. Furthermore,

$$(I-H)^2 = I-2H+H^2 = I-2H+H = I-H \quad \text{and} \quad \text{Trace}(I)-\text{Trace}(H) = n-d$$

So its eigenvalues are also either one or zero. And

$$I-H = q_{d+1}q_{d+1}^\top + \cdots + q_N q_N^t op \quad \text{and} \quad \text{rank}(I-H) = N-d$$

Note that the eigenvectors of $I-H$ are orthogonal to the eigenvectors of H, and $(I-H)H = 0$.

1.3 Maximum Likelihood Estimation of σ^2_{ML}

We now solve the second set of equations, taking derivatives with respect to σ^2 and setting it equal to zero. (Note that we are taking the derivatives with respect to σ^2 not σ.)

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\log \text{Lik}(\mathbf{b},\sigma^2|X,y) &= \frac{\partial}{\partial\sigma^2}\left(\frac{N}{2}\log(\sigma^2) + \frac{(y-X\mathbf{b}_{ML})^\top(y-X\mathbf{b}_{ML})}{2\sigma^2}\right)\\ &= \frac{N}{2\sigma^2} - \frac{(y-X\mathbf{b}_{ML})^\top(y-X\mathbf{b}_{ML})}{2\sigma^4}\\ &= \frac{N\sigma^2 - (y-X\mathbf{b}_{ML})^\top(y-X\mathbf{b}_{ML})}{2\sigma^4}\end{aligned}$$

Setting this partial derivative equal to zero we get

$$\sigma^2 = \sigma^2_{ML} = \frac{(y-X\mathbf{b}_{ML})^\top(y-X\mathbf{b}_{ML})}{N}$$

Note that we plug in the maximum likelihood estimate $\mathbf{b}_{ML}$ for $\mathbf{b}$ since we have already computed it.

The maximum likelihood estimation of variance σ^2_{ML} is a *biased* estimator in that its expected value does not equal the true variance of the error term. The *standard error* or mean square error (MSE) is defined as

$$\text{MSE} = S_e^2 = \frac{(y-X\mathbf{b}_{ML})^\top(y-X\mathbf{b}_{ML})}{N-d} = \frac{\hat{\epsilon}^\top\hat{\epsilon}}{N-d}$$

which is an *unbiased* estimate of the variance, in that its expected value is the actual variance σ^2.

The optimal least squares value, also known the *residual sum of squares* or *sum of square of error (SSE)*. We have

$$\hat{\epsilon}^\top\hat{\epsilon} = y^\top(I-H)^\top(I-H)y = y^\top(I-H)y$$

The last equality follows because $(I-H)^\top = I-H$, and $H^2 = H$ and $(I-H)^2 = I-H$.

1.4 Distribution of estimated coefficients

In general, suppose $\mathbf{X}$ is a random vector of length n that follows the multivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$ where Σ is the covariance matrix [3] of $\mathbf{X}$. Define the random vector $\mathbf{Y} = A\mathbf{X} + \mathbf{a}$, with A an $m \times n$ matrix, and $\mathbf{Y}$ and $\mathbf{a}$ vectors of length m. Then $\mathbf{Y}$ is also normal. We have

$$\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma) \quad \Rightarrow \quad \mathbf{Y} \sim N(A\boldsymbol{\mu} + \mathbf{a}, A\Sigma A^\top)$$

Keeping this in mind, let us now work to find the distribution of estimated $\mathbf{b}_{ML}$ and the predicted response y_{new}. In the common (ferquentist) approach, the feature matrix X is considered fixed. And the response $\mathbf{y} = X\mathbf{b} + \boldsymbol{\epsilon}$ is random since the vector of errors $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)$ are considered random with normal distribution. In fact each ϵ_i follows $N(0, \sigma)$ (the same σ for all ϵ_i.) Also since the data are i.i.d., the ϵ_i are also i.i.d. This implies that the vector $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma I)$ where $\mathbf{0}$ is the vector of all zeros of length N, and I is the identity matrix. With this point of view the random vector $\mathbf{Y} = X\mathbf{b} + \boldsymbol{\epsilon}$ follows the normal distribution $N(X\mathbf{b}, \sigma^2 X^\top X)$ (since $X^\top(\sigma^2 I)X = \sigma^2 X^\top X$).

Since $\hat{\mathbf{b}} = \mathbf{b}_{ML} = (X^\top X)^{-1} X^\top \mathbf{y}$, it follows that $\hat{\mathbf{b}}$ also follows the normal distribution, since it is a linear transformation of another normally distributed random vector $\mathbf{y}$. So its mean an variance are:

$$\begin{aligned}
\mathbb{E}(\hat{\mathbf{b}}) &= (X^\top X)^{-1} X^\top \mathbf{y} = (X^\top X)^{-1} X^\top \mathbf{b} = \mathbf{b} \\
\mathrm{Var}(\hat{\mathbf{b}}) &= \big((X^\top X)^{-1} X^\top\big)\big(\sigma^2 I\big)\big((X^\top X)^{-1} X^\top\big)^\top = \sigma^2 (X^\top X)^{-1}
\end{aligned}$$

Therefore,

$$\mathbf{b}_{ML} = \hat{\mathbf{b}} \sim N\big(\mathbf{b}, \sigma^2 (X^\top X)^{-1}\big)$$

where $\mathbf{b}$ is the "true" value of coefficients, and σ^2 is the "true" variance of the error term ϵ. In particular, the individual ML coefficients $\hat{b}_i = \boldsymbol{e}_i^\top \hat{\mathbf{b}}_{ML}$, where $\boldsymbol{e}_i$ is the vector whose i^{th} entry is 1 and every other entry is zero. So the distribution of each ML b_i are also normal with mean $\boldsymbol{e}_i^\top \mathbf{b} = b_i$ and variance $\sigma^2 \boldsymbol{e}_i^\top (X^\top X)^{-1} \boldsymbol{e}_i = \sigma^2 \big((X^\top X)^{-1}\big)_{ii}$ the i^{th} diagonal entry of the matrix $(X^\top X)^{-1}$:

$$\hat{b}_i \sim N\Big(b_i, \sigma^2 (X^\top X)^{-1}\big)_{ii}\Big)$$

Note that i ranges from zero to d, the 0^{th} entry corresponding to the intercept (the column of X that is all ones).

Of course we do not know σ^2 in general, so we can instead use the S_e^2 as an approximation of σ^2. Also for each ML coefficients $\hat{b}_i$ define

$$s^2(\hat{b}_i) = S_e^2 \Big((X^\top X)^{-1} \Big)_{ii}$$

As a result it is now known that the standardized (or actually *Studentized*) values of the ML coefficients $\hat{b}_i$ follows the *t-distribution* with $N - d$ degrees of

[3] In multivariate cases it is more convenient to consider the variance σ^2 and the covariance matrix Σ as the parameter of the normal distribution rather than standard deviation σ.

freedom:

$$\frac{\hat{b}_i}{s_e(b_i)} \sim t(N-d) \quad \text{the } t \text{ distribution with degrees of freedom } df = N-d$$

We will prove this result in the Appendix below.

Hypothesis Testing for the coefficients $\hat{b}_i$, and finding confidence intervals

In the frequentist approach we can test whether the feature x_i is relevant in determining y or not. In the formula that relates the x_i to y, the relationship is assumed to be $y = b_1x_1 + b_2x_2 + \cdots + b_dx_d$. So if x_i is irrelevant to y its coefficient b_i must be zero. This is formulated by the following hypothesis testing:

Null Hyothesis: The coefficient $b_i = 0$

Alternative Hypothesis: The coefficient $b_i \neq 0$

The *p-value* of this test can be calculated by the t-distribution applied to $x = \hat{b}_i/s(\hat{b}_i)$. In R we can use `pt(x, df)` with `df` $= N-d$.

The confidence interval can also be calculated similarly with

$$\hat{b}_i \pm s(\hat{b}_i)\texttt{qt}\big(1-\alpha/2, \texttt{df} = N-d\big).$$

In R the `lm` produces the model in which all p-values of coefficients along with their 95% confidence intervals are provided.

1.5 The distribution of $\hat{\epsilon}$

We saw that

$$\hat{\epsilon} = y - X\hat{b} = y - Hy = (I-H)y$$
$$\hat{\epsilon}^\top\hat{\epsilon} = y^\top(I-H)y$$

On the other hand, we know that $y \sim N(Xb, \sigma^2 I)$. Therefore, $\hat{\epsilon} = (I-H)y$ also follows the normal distribution with mean

$$\mathbb{E}(\hat{\epsilon}) = (I-H)\mathbb{E}(y) = (I-H)Xb = Xb - HXb = Xb - Xb = 0$$

This because $HXb = \left[X^\top(X^\top X)^{-1}X^\top\right]Xb = Xb$.

Also the variance/covariance matrix $\hat{\epsilon}$ can be computed as

$$\text{Cov}(\hat{\epsilon}) = (I-H)^\top\big(\sigma^2 I\big)(I-H) = \sigma^2(I-H).$$

(Recall that $I-H$ is symmetric and $(I-H)^2 = I-H$.) Thus, we have:

$$\hat{\epsilon} \sim N\big(0, \sigma^2(I-H)\big)$$

One observation here is that even though the actual error ϵ_i are independent, the *estimated* residuals $\hat{\epsilon}_i$ are actually dependent (otherwise the covariance would have been $\sigma^2 I$ and not $\sigma^2(I-H)$.)

1.6 Degrees of freedom and an unbiased estimator for σ^2

We have:

$$X^\top \hat{\epsilon} = X(y - Xb_{ML}) = Xy - X^\top X b_{ML} = X^\top y - X^\top y = 0$$

What this says is that the vector $\hat{\epsilon}$ which has N entries, satisfies d equations (as $X^\top$ has d rows). What this means is that since $\hat{\epsilon}$ is a random vector, $N - d$ of its entries can be chosen randomly, but the remaining d entries must be chose in such a manner that $X^\top \hat{\epsilon} = 0$ is satisfied. We say that $\hat{\epsilon}$ has $n - d$ *degrees of freedom*:

$$\mathrm{df}(\hat{\epsilon}) = N - d$$

In fact, since the first column of X is all ones, it follows that $1^\top \hat{\epsilon} = \sum_i \hat{\epsilon}_i = 0$, and also the sample mean $\mathrm{avg}(\hat{\epsilon}) = 0$.

From here we can also calculate the expected value of MSE, that is $\frac{\hat{\epsilon}^\top \hat{\epsilon}}{N-d}$. In fact,

$$\sigma^2(I - H) = \mathrm{Cov}(\hat{\epsilon}) = \mathbb{E}\Big((\hat{\epsilon} - \bar{\epsilon}1)(\hat{\epsilon} - \bar{\epsilon}1)\Big)^\top = \mathbb{E}\Big(\hat{\epsilon}\hat{\epsilon}^\top\Big),$$

since $\bar{\epsilon} = 0$.

Observe that

$$\hat{\epsilon}^\top \hat{\epsilon} = \mathrm{Trace}\Big(\hat{\epsilon}^\top \hat{\epsilon}\Big),$$

where $\mathrm{Trace}(A)$, the trace of matrix A, is the sum of its diagonal entries. Therefore,

$$\mathbb{E}\big(\hat{\epsilon}^\top \hat{\epsilon}\big) = \mathbb{E}\,\mathrm{Trace}\Big(\hat{\epsilon}^\top \hat{\epsilon}\Big) = \mathrm{Trace}\,\mathbb{E}\Big(\hat{\epsilon}^\top \hat{\epsilon}\Big) = \sigma^2\,\mathrm{Trace}(I - H)$$

Finding $\mathrm{Trace}(I - H)$ is easy since, noting that $I - H$ is a $N \times N$ matrix, we get

$$\begin{aligned}
\mathrm{Trace}(I - H) = \mathrm{Trace}(I) - \mathrm{Trace}(H) &= N - \mathrm{Trace}\Big(X(X^\top X)^{-1}X^\top\Big) \\
&= N - \mathrm{Trace}\Big(X^\top X(X^\top X)^{-1}\Big) \\
&= N - \mathrm{Trace}(I_d) \\
&= N - d
\end{aligned}$$

It follows that $\mathbb{E}\big(\hat{\epsilon}^\top \hat{\epsilon}\big) = \sigma^2(N - d)$. The punch line now is that

$$\mathbb{E}(\mathrm{MSE}) = \mathbb{E}\left(\frac{\hat{\epsilon}^\top \hat{\epsilon}}{N - d}\right) = \sigma^2.$$

In other words, SSE is an *unbiased estimator of* σ^2. This is in contrast with the ML estiamtor of σ^2, that is $\frac{\hat{\epsilon}^\top \hat{\epsilon}}{N}$ which is biased estimator, since its expected value is not σ^2.

1.7 R^2 and the explanatory power of the model

One way to see how effective the linear model is for a particular set of data is to see how much of the variation in the response y is explained n=by the model, and how much os it is not explained by it.

Define *total variation* or *total sum of squares (SST)* by

$$= \text{SST} = \sum_{i=1}^{N} (y_i - \bar{y})^2 = y^\top \left(I - \frac{1}{N} J\right) y$$

Here $J = 11^\top$, that is the $N \times N$ matrix of all ones.

We also know that the variation as a result of error, the *unexplained variation* is the SSE. This quantity can also be written in matrix notation as:

$$\text{SSE} = \hat{\epsilon}^\top \hat{\epsilon} = y^\top (I - H) y$$

Define the *regression sum of squares* or *explained variation* as

$$\text{SSR} = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{N} (\hat{b_1} x_{xi1} + \cdots + \hat{b_d} x_{id} - \bar{y})^2.$$

This quantity shows measures how the *predicted value* $\hat{y}_i$ varies as a result of variation on features. It is the "explained variation" since the model completely determines it.

In matrix notation we can show that

$$\text{SSR} = y^\top \left(H - \frac{1}{N} J\right) y$$

Now it is easily seen that

$$\begin{aligned} \text{SSE} + \text{SSR} &= y^\top \left(H - \frac{1}{N} J\right) y + y^\top (I - H) y \\ &= y^\top \left(H - \frac{1}{N} J + I - H\right) y \\ &= y^\top \left(I - \frac{1}{n} J\right) y \\ &= \text{SST} \end{aligned}$$

So the equation SST=SSR+SSE. Also all of these quantities are nonnegative. It follows that the total variation of SST can be decomposed into into the explained variation SSR and unexplained variation SSE. Define

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

Then clearly $0 \leq R^2 \leq 1$. The closer R^2 is to 1, the better the model explains the changes in y. The closer it is to 0, the more of its variation is unexplained,

and its is possible that more features not included in the model may have to be added. So in short, R^2 is a measure of model's fit to the data.

It is also possible to show that $\mathrm{df}(\mathrm{SSR}) = d-1$, and $\mathrm{df}(\mathrm{SST}) = N-1$. This means that there is a $(d-1)\times N$ matrix B such that $B\left(H-\frac{1}{N}y\right) = 0$ and $d-1$ is the largest such dimension. Also, there is a vector a such that $a^\top\left(I-\frac{1}{N}J\right) = 0$.

Notice that $\mathrm{df}(\mathrm{SSE}) + \mathrm{df}(\mathrm{SSR}) = \mathrm{df}(\mathrm{SST})$.

1.8 Analysis of Variance (ANOVA)

We can also test the relevance of a subset of x_i and see, together if they are relevant or not. For instance suppose we are considering two models: one with features $x_1, \ldots, x_d$ called the *reduced model*, and another with these features *plus* the additional $x_{d+1}, \ldots, x_{d+k}$, called the *complete model*. Suppose X_d is the matrix of the reduced model, and X_{d+k} is the matrix of the complete model: Also, X_k be the matrix made up of all only features $x_{d+1}, \ldots, x_{d+k}$.

$$X_{d+k} = [X_d, X_k]$$

Finally, let $\hat{b}_{d+k}$ be the ML coefficients of the complete model, and $\hat{\hat{b}}_d$ the coefficients of the reduced model. Note that $\hat{\hat{b}}_d$ are different from the from the first d components of $\hat{b}_d$. If the additional variables $x_{d+1}, \ldots, x_{d+k}$ are not significant, then $\hat{b}_{d+1} = \cdots = \hat{b}_{d+k} = 0$. If $\hat{b_r}$ is the estimated coefficients for the reduced model and $\hat{b_c}$ for the complete model, we have

$$\log \mathrm{Lik}(\text{reduced model}) = \frac{\left(y - X_d\hat{b_r}\right)^\top\left(y - X_d\hat{b_r}\right)}{\sigma^2}$$

$$\log \mathrm{Lik}(\text{complete model}) = \frac{\left(y - X_{d+k}\hat{b_c}\right)^\top\left(y - X_{d+k}\hat{b_c}\right)}{\sigma^2}$$

We learned in the previous lecture that we could use the likelihood ratio test to compare two models. In this case, when the likelihood ratio test is specialized to the linear regression with normal error, the likelihood ratio test is essentially the same as the *Analysis of Variance (ANOVA)* test. To see how ANOVA works let's first recall the χ^2 and F distributions.

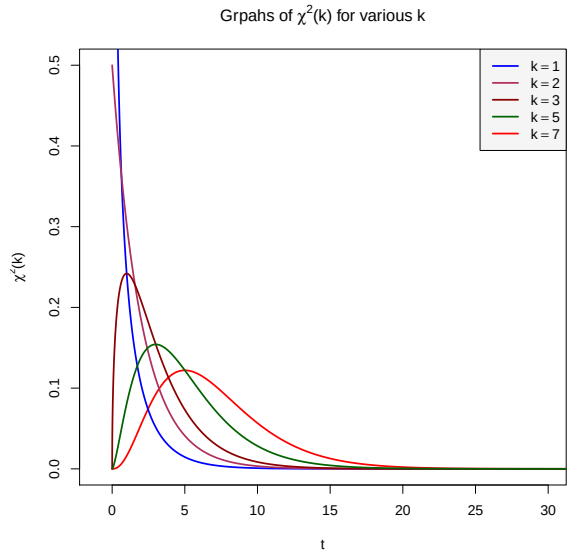
The $\chi^2(k)$ and $F(m,n)$ distributions

Before showing how to evaluate significance of each variable x_i or subset of these variables, let's first review the two basic distributions.

Suppose $Z_1, Z_2, \ldots, Z_k$ are i.i.d. *standard normal* random variables, that is each $Z_i \sim N(0,1)$, and they are all independent. Then the random variable

$$Z_1^2 + Z_2^2 + \cdots + Z_k^2 \sim \chi^2(k)$$

The $\chi^2(k)$ distribution for various values of k are drawn in the following graph:



The *Fisher's* F *distribution* arises as the ratio of two independent χ^2 random variables. If U and V are random variables with $U \sim \chi^2(i)$ and $V \sim \chi^2(j)$, then the ratio

$$\frac{U/i}{V/j} \sim F(i, j)$$

where $F(i, j)$ is the F-distribution with two parameters $df = i$ and $df = j$. This distribution arise in determining whether a *subset* of features are relevant in the response variable y.

Now consider the following quantities:

1. $X = X_{d+k}$ is the design matrix of the complete model, and $H = X(X^\top X)^{-1}X^\top$ is the hat matrix for the complete model.

2. $SST = (y - \bar{y}\mathbf{1})^\top (y - \bar{y}\mathbf{1})$ is the variation in observed y_i. It can be shown that
$$\frac{SST}{\sigma^2} \sim \chi^2(N - 1).$$
See Appendix for details. Note that since $\mathbf{1}^\top(y - \bar{y}\mathbf{1}) = 0$, it has $N - 1$ degrees of freedom.

3. X_d is the first d columns of the design matrix. It is in fact the design matrix of the reduced model. Similarly, the hat matrix for the reduced matrix is $H_d = X_d\left(X_d^\top X_d\right)^{-1}X_d^\top$.

4. X_k is the columns of the design matrix which are in the complete model but not in the reduced one. So $X = [X_d, X_k]$.

5. For the full model we have

$$\begin{aligned}\text{SSE}(d+k) = \sum_{i=1}^{N} &= \left(y_i - \hat{b}^\top x_i\right)^\top \left(y_i - \hat{b}^\top x_i\right) \\ &= (y - Xb)^\top (y - Xb) \\ &= (y - Hy)^\top (y - Hy) \\ &= y^\top (I - H)^\top (I - H) y\end{aligned}$$

   It can be shown that (See Appendix)

$$\frac{\text{SSE}(d+k)}{\sigma^2} \sim \chi^2(N - d - k)$$

   So $\text{SSE}(d+k)$ has $N - d - k$ degrees of freedom.

6. Similarly, for the reduced model we have $\text{SSE}(d) = y^\top (I - H_d)^\top (I - H_d) y \sim \chi^2(N - d)$, that is $\text{SSE}(d)$ has $N - d$ degrees of freedom.
   Finally, if the new k features $x_{d+1}, \ldots, x_{d+k}$ are not significant, and $b_{d+1} = \cdots = b_{d+k} = 0$, then it can be shown

$$\frac{\text{SSE}(d) - \text{SSE}(d+k)}{\sigma^2 k} \sim \chi^2(k)$$

   Note that this is true only if the extra features are insignificant.

The conclusion from items 6 and 2 is that

$$\text{If } b_{d+1} = \cdots = b_{k+d} = 0 \quad \text{then} \quad \frac{\big(\text{SSE}(d) - \text{SSE}(d+k)\big)/k}{\text{MSE}(d+k)} \sim F(k, d+k)$$

where $\text{MSE}(d+k) = \text{SSE}(d+k)/(d+k)$. The *analysis of variance (ANOVA)* is the following statistical test:

Null hypogthesis: The variables $X_{d+1}, \ldots, X_{d+k}$ are not significant and thus, $b_{d+1} = \cdots = b_{d+k} = 0$,
Alternative hypothesis: At least one of the coefficients $b_{d+1}, \ldots, b_{d+k}$ is non zero.

The p-value of this test is computed by calculating the ratio in item 6 over 2 which should follow the F distribution if the null hypothesis is true.

In R this can be achieved very easily as follows. Suppose the complete model is $y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$ and the reduced model is $y = b_0 + b_1 x_1 + b_2 x_2$. Then, using the `lm` package, do

```
completeModel <- lm(Y~X1+X2+X3+X4, data=someDat)
reducedModel <- lm(Y~X1+X2, data=someDat)
anova(reducedModel,completeModel)
```

will give the p-value of the two additional variables `X3,X4`.

Appendix: Details of ANOVA

We saw that the estimated $\hat{\mathbf{b}} = (X^\top X)^{-1} X^\top \mathbf{y}$ follows $N\big(\mathbf{b}, \sigma^2 (X^\top X)^{-1}\big)$. Also, the residuals $\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - X\mathbf{b} = \mathbf{y} - H\mathbf{y}$ are normal $N\big(\mathbf{0}, \sigma^2 (I - H)\big)$. We first prove a useful fact. Recall that in general for the random vectors X and Y of lengths n and m, respectively, we have

$$\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}\left(\big(\mathbf{X} - \mathbb{E}(\mathbf{Y})\big)\big(\mathbf{Y} - \mathbb{E}(\mathbf{Y})\big)^\top \right) \quad \text{and} \quad \mathrm{Cov}\big(A\mathbf{X}, B\mathbf{Y}\big) = A\,\mathrm{Cov}(\mathbf{X}, \mathbf{Y})B^\top.$$

In particular, if $\mathbf{X}$ and $\mathbf{Y}$ are un-correlated, that is $\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = 0$, then any linear transformation of them is also un-correlated: $\mathrm{Cov}(A\mathbf{X}, B\mathbf{Y}) = A\,\mathrm{Cov}(\mathbf{X}, \mathbf{Y})B^\top = 0$. We can now prove a useful fact:

Lemma 1 *The normal variables $\hat{\boldsymbol{\epsilon}}$ and $\hat{\mathbf{b}}$ are un-correlated and thus independent.*

Proof:

$$\begin{aligned}
\mathrm{Cov}\big(\hat{\boldsymbol{\epsilon}}, \hat{\mathbf{b}}\big) &= \mathrm{Cov}\big(\mathbf{y} - \hat{\mathbf{y}}, \hat{\mathbf{b}}\big) \\
&= \mathrm{Cov}\Big(\mathbf{y} - H\mathbf{y}, \big(X^\top X\big)^{-1} X^\top \mathbf{y}\Big) \\
&= \mathrm{Cov}\Big((I - H)\mathbf{y}, \big(X^\top X\big)^{-1} X^\top \mathbf{y}\Big) \\
&= (I - H)\,\mathrm{Cov}(\mathbf{y}, \mathbf{y})\big(X^\top X\big)^{-1} X^\top \quad \text{noting that} \quad \mathrm{Cov}(\mathbf{y}, \mathbf{y}) = \sigma^2 I \\
&= \sigma^2 \Big(I - X\big(X^\top X\big)^{-1} X^\top\Big) X \big(X^\top X\big)^{-1} \\
&= \sigma^2 \Big(X\big(X^\top X\big)^{-1} - X\big(X^\top X\big)^{-1}\Big) \\
&= 0
\end{aligned}$$

■

Next recall that $\hat{\boldsymbol{\epsilon}} = (I - H)\mathbf{y}$ and the sum of squares of residuals is

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}}^\top \hat{\boldsymbol{\epsilon}} &= \mathbf{y}^\top (I - H)^2 \mathbf{y} \\
&= \mathbf{y}^\top (I - H)\mathbf{y} \\
&= \mathbf{y}^\top \Big(\mathbf{q}_1 \mathbf{q}_1^\top + \cdots + \mathbf{q}_{N-d} \mathbf{q}_{N-d}^\top\Big)\mathbf{y} \\
&= \big(\mathbf{y}^\top \mathbf{q}_1\big)^2 + \cdots + \big(\mathbf{y}^\top \mathbf{q}_{N-d}\big)^2
\end{aligned}$$

where $\mathbf{q}_i$ are the set of orthonormal eigenvectors of $I - H$ corresponding to eigenvalue 1. Now,

$$\mathrm{Cov}(\mathbf{q}_i^\top \mathbf{y}, \mathbf{q}_j^\top \mathbf{y}) = \mathbf{q}_i^\top\, \mathrm{Cov}(\mathbf{y}, \mathbf{y})\mathbf{q}_j = \sigma^2 \mathbf{q}_i^\top \mathbf{q}_j = 0.$$

And

$$\mathrm{Var}(\mathbf{q}_i^\top \mathbf{y}) = \mathrm{Cov}(\mathbf{q}_i^\top \mathbf{y}, \mathbf{q}_i^\top \mathbf{y}) = \mathbf{q}_i^\top\, \mathrm{Cov}(\mathbf{y}, \mathbf{y})\mathbf{q}_i = \sigma^2 \mathbf{q}_i^\top \mathbf{q}_i = \sigma^2.$$

So for different q_i, q_j, the quantities $q_i^\top y$ are independent. Also,

$$\begin{aligned}\mathbb{E}(q_i^\top y) &= \mathbb{E}(q_i^\top X b)\\ &= \mathbb{E}\Big(q_i^\top X(\hat{b} - (\hat{b} - b))\Big)\\ &= \mathbb{E}(q_i^\top X\hat{b}) - \mathbb{E}(q_i^\top X(\hat{b} - b))\\ &= \mathbb{E}(q_i^\top Hy) - q_i^\top X\mathbb{E}(\hat{b} - b) = 0 - 0 = 0 \quad (\text{since } q_i^\top H = 0)\end{aligned}$$

In summary the random variables $q_i^\top y \sim N(0, \sigma^2)$ each, and they are independent. Therefore, dividing each by σ we get a a set of i.i.d standard normal $N(0, 1)$ random variable, and so their sum of squares is a chi-square random variable:

$$\frac{\hat{\epsilon}^\top \hat{\epsilon}}{\sigma^2} = \sum_{i=1}^{N-d} \frac{(q_i^\top y)^2}{\sigma^2} \sim \chi^2(N - d)$$

Next let's consider the vector $v = \left(X^\top X\right)^{1/2}(\hat{b} - b)$.[4] This vector is a normal random vector since it is a linear transformation of of the vector $\hat{b} - b$, which itself is a normal random vector. So v/σ is a standardized version of $\hat{b}$. In fact, we have,

$$\mathbb{E}(v) = \left(X^\top X\right)^{1/2}\left(\mathbb{E}(\hat{b}) - b\right) = \left(X^\top X\right)^{1/2}(b - b) = 0.$$

And

$$\mathrm{Cov}(v) = \mathrm{Cov}\Big((X^\top X)^{1/2}\hat{b}\Big) = \sigma^2\left(X^\top X\right)^{1/2}\left(X^\top X\right)^{-1}\left(X^\top X\right)^{1/2} = \sigma^2 I$$

So $v \sim N(0, \sigma^2 I)$. This also implies $v^\top v/\sigma^2 \sim \chi^2(d)$.

Also, since $\mathrm{Cov}(\hat{\epsilon}, \hat{b}) = 0$, and v is a linear transformation of $\hat{b} - b$ it follows that $\mathrm{Cov}(\hat{\epsilon}, v) = 0$.

So since v and $\hat{\epsilon}$ are independent, $v^\top v/\sigma^2 \sim \chi^2(d)$, and $\hat{\epsilon}^\top \hat{\epsilon} \sim \chi^2(N - d)$, it follows that

$$\frac{v^\top v/d}{\hat{\epsilon}^\top \hat{\epsilon}/(N - d)} = \frac{\left(\hat{b} - b\right)^\top (X^\top X)\left(\hat{b} - b\right)/d}{\hat{\epsilon}^\top \hat{\epsilon}/(N - d)} \sim F(d, N - d)$$

In particular, if the entire model is insignificant then $b = 0$ and so

$$F - \text{ratio} = \frac{\hat{b}^\top (X^\top X)\hat{b}/d}{\hat{\epsilon}^\top \hat{\epsilon}/(N - d)} \sim F(d, N - d)$$

So we can do the following hypothesis test:

Null hypothesis: All variables in the linear model are insignificant in determining y, (that is $b = 0$).

Alternative hypothesis: At least one of the variables in the linear model is significant (that is at least one of $b_i \neq 0$).

[4] The matrix $X^\top X$ is symmetric and has positive eigenvalues λ_i. So, if $X^\top X = \lambda_1 q_1 q^\top + \cdots + \lambda_n q_n q_n^\top$, then $\left(X^\top X\right)^{1/2} = \sqrt{\lambda_1} q_1 q_1^\top + \cdots + \sqrt{\lambda_n} q_n q_n^\top$.

To test this the null hypothesis, we form the F-ratio above. If the Null hypothesis is true the F-ratio should hover around its mean which for $F(m,n)$ equals $\frac{n}{n+2}$. Of on the other hand the Null hypothesis is false, then the F-ratio will be much larger than its expected value. So to get the p-value, we need to calculate the $Pr[X > F-\text{ratio}]$ where $X \sim F(m,n)$. This can be done in R for instance like this:

```
pVal = fp(Fatio,df1=d,df2=N-d, tail=TRUE)
```

Also the command

```
> linModel<-lm(y~x1+x2, data=dat)
> summary(linModel)

Call:
lm(formula = y ~ x1 + x2, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-14.3237  -4.0080  -0.3959   4.2126  19.7643

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.1718     4.9098   3.905 0.000174 ***
x1            2.5622     0.2995   8.555 1.75e-13 ***
x2            1.8570     0.0768  24.179  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.216 on 97 degrees of freedom
Multiple R-squared:  0.8696,    Adjusted R-squared:  0.8669
F-statistic: 323.4 on 2 and 97 DF,  p-value: < 2.2e-16
```

As you can see the summary gives the estimated coefficients of x_1, x_2 and the constant term (Intercept). It also gives the both the F-ratio (323.4) and its p-value (2.2×10^{-16}). In this case since the p-value is extremely small, we must reject the Null hypothesis and declare that at least one of x_1 or x_2 are significant.

Appendix: Partial F tests

We now continue by checking the significance of a *subset of variables*. So as before, suppose the complete and reduced models are given by:

Reduced model: $y = b_1x_1 + \cdots + b_dx_d$

Complete model: $y = b_1x_1 + \cdots + b_dx_d + b_{d+1}x_{d+1} + \cdots + b_{d+k}x_{d+k}$

We will use the subscript d for the reduced model and subscript $d+k$ for the complete one. So the design matrix for the reduced models,X_d, is the first d columns of the design matrix for the complete model, X_{d+k}. We have

$$\mathrm{SSE}_d = \mathbf{y}^\top (I - H_d) \quad \mathrm{SSE}_{k+d} = \mathbf{y}^\top (I - H_{d+k})\mathbf{y}$$

where H_d and H_{d+k} are, respectively, the hat matrices of the reduce and complete models. We now show that $\mathrm{SSE}_d - \mathrm{SSE}_{d+k}$ itself is a sum of squares and has a $\chi^2(k)$ distribution. Note that

$$\mathrm{SSE}_d - \mathrm{SSE}_{d+k} = \mathbf{y}^\top (I - H_d)\mathbf{y} - \mathbf{y}^\top (I - H_{d+k})\mathbf{y} = \mathbf{y}^\top \big(H_{d+k} - H_d\big)\mathbf{y}$$

It turns out the matrix $(H_{d+k} - H_d)^2 = H_{d+k} - H_d$. First, observe that $H_{d+k}H_d = H_d H_{d+k} = H_d$. This is is so because in general, for a vector $\boldsymbol{z}$ the vector $\boldsymbol{z}_1 = H_d\boldsymbol{z}$ is the orthogonal projection of $\boldsymbol{z}$ to the space spanned by the first d column so the design matrix. And $H_{d+k}H_d\boldsymbol{z} = H_{d+k}\boldsymbol{z}_1 = \boldsymbol{z}_1$ since $\boldsymbol{z}_1$ is already in the space spanned by the $d+k$ columns of the design matrix. So $H_{d+k}H_d = H_d$. Similarly, $H_d H_{d+k}\boldsymbol{z} = H_d\boldsymbol{z}$, because orthogonal projection of $\boldsymbol{z}$ first to the space spanned $x_1, \ldots, x_{d+k}$ and subsequently projecting the result to the subspace spanned by $x_1, \ldots, x_d$ gives us the same point as if we had directly projected orthogonally to the space spanned by $x_1, \ldots, x_d$.

So, we $\big(H_{d+k} - H_d\big)^2 = H_{d+k}^2 - H_{d+k}H_d - H_d H_{d+k} + H_d^2 = H_{d+k} - H_d$. So the eigenvalues of $H_{d+k} - H_d$ are either zero or one. Also $\mathrm{rank}(H_{d+k} - H_d) = \mathrm{rank}(H_{d+k} - \mathrm{rank}(H_d) = d+k-d = k$. So $\mathbf{y}^\top(H_{d+k} - H_d)\mathbf{y}$ is a sum of squares k independent normal $N(0, \sigma^2)$ variables and so its distribution is $\chi^2(k)$. In summary

$$\mathrm{SSE}_d - \mathrm{SSE}_{d+k} = \mathbf{y}^\top \big(H_{d+k} - H_d\big)^2 \mathbf{y} \sim \chi^2(k)$$

Also

$$\begin{aligned}
\mathrm{Cov}\Big(\big(I - H_{d+k}(\mathbf{y}, \big(H_{d+k} - H_d\big)\mathbf{y}\Big) &= \sigma^2 \big(I - H_{d+k}\big)\big(H_{d+k} - H_d\big) \\
&= \sigma^2 \big(H_{d+k} - H_d - H - H_{d+k}^2 + d + kH_d\big) \\
&= \sigma^2 \big(H_{d+k} - H_d - H_{d+k} + H_d\big) = 0
\end{aligned}$$

It follows that $\big(I - H_{d+k}\big)\mathbf{y}$ and $\big(H_{d+k} - H_d\big)\mathbf{y}$ are independent. So we can form an F ratio from their sum of squares as follows: