Lecture 5: General Linear Model: Estimation and Testing

Reading Assignment:

- Muller and Fetterman Chapter 2
- Weisberg Chapter 3

We will now consider the general case in which we observe a single response and one or more covariates.

We write the general form of the linear model as

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\,\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1},$$

where

- $\mathbf{y}_{n\times 1} = \begin{bmatrix} y_1 \\ \vdots \\ \\ \vdots \\ y_n \end{bmatrix}$ contains the observed responses,
- $\mathbf{X}_{n\times p} = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{p-1}) = \begin{bmatrix} x_{1,0} & x_{1,1} & \cdots & x_{1,p-1} \\ x_{2,0} & x_{2,1} & \cdots & x_{2,p-1} \\ \vdots & & & \vdots \\ x_{n,0} & x_{n,1} & \cdots & x_{n,p-1} \end{bmatrix}$ is a matrix of fixed and known covariates,

- $\boldsymbol{\beta}_{p\times 1} = (\beta_0, \beta_1, \ldots, \beta_{p-1})' = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}$ is a vector of parameters to be estimated, and

- $\boldsymbol{\varepsilon}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ is a vector of unobserved random errors.

- The rows of $\mathbf{y}, \mathbf{X},$ and $\boldsymbol{\varepsilon}$ correspond to subjects or sampling units.
- Columns of $\mathbf{X}$ and the corresponding rows of $\boldsymbol{\beta}$ correspond to predictors. Often, the first column of $\mathbf{X}$, denoted $\mathbf{x}_0$, corresponds to an intercept variable and takes the value 1 for all subjects so that we have $\mathbf{x}_0 = \mathbf{J} = \mathbf{1}$. (We will assume this to be the case unless we state otherwise.)

We refer to the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ as the matrix representation of the linear model. All linear models may be written in this form.

Alternatively, the model is written in scalar notation as

$$y_i = \sum_{j=0}^{p-1} x_{ij}\beta_j + \varepsilon_i, \ i = 1, \ldots, n.$$

The row of $\mathbf{X}$ corresponding to subject i is denoted $\text{row}_i(\mathbf{X})$ or $\mathbf{x}_i'$ (which may lead to some confusion with the columns of $\mathbf{X}$ so that the precise meaning may depend on the context).

Example: Social Setting, Family Planning, and Birth Rate

Rodriguez (2002) considers factors related to the decline in the crude birth rate (CBR, the number of births per thousand population) between 1965 and 1975 for 20 Latin American and Caribbean countries. Covariates of interest include a social setting index (SOCIAL) and a family planning index (FAMPLAN). The social setting index is a function of literacy, school enrollment, life expectancy, infant mortality, percent of males aged 15-64 in the non-agricultural labor force, gross national product per capita, and percent of population living in urban areas. Higher social setting scores represent higher socio-economic levels. The family planning index is a function of availability of contraceptive methods, official government family planning policies, and structure of family planning programs in the country. Values of 20 or more indicate strong efforts in family planning, and values of 10-19 represent moderate efforts.

A few representative observations in the birth rate dataset are presented below.

Country	*SOCIAL*	*FAMPLAN*	*CBR*
Brazil	74	0	10
Costa Rica	84	21	29
Haiti	35	3	0
Mexico	83	4	9
Trinidad-Tobago	84	15	29

This model is written in the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ as follows:

$$\mathbf{y}_{20\times 1} = \mathbf{X}_{20\times 3}\,\boldsymbol{\beta}_{3\times 1} + \boldsymbol{\varepsilon}_{20\times 1}$$

$$\begin{bmatrix} 10 \\ 29 \\ 0 \\ 9 \\ 29 \\ \vdots \end{bmatrix}_{20\times 1} = \begin{bmatrix} 1 & 74 & 0 \\ 1 & 84 & 21 \\ 1 & 35 & 3 \\ 1 & 83 & 4 \\ 1 & 84 & 15 \\ \vdots & \vdots & \vdots \end{bmatrix}_{20\times 3} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}_{3\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \vdots \end{bmatrix}_{20\times 1}$$

- β_0 is called the intercept, which is the expected decline in birth rate when the social setting and family planning indices take the value zero.
  - Some investigators prefer to center all predictors so that the mean of each predictor is zero. Why?
- β_1 is the slope for social setting. It is interpreted as the expected increase in CBR decline for a one unit increase in the social setting index.
- β_2 is the slope for family planning. It is interpreted as the expected increase in CBR percent decline for a one unit increase in the family planning index.
- Each element of ε represents the distance between a country's observed percent CBR decline and the population regression line.

Least Squares Estimation

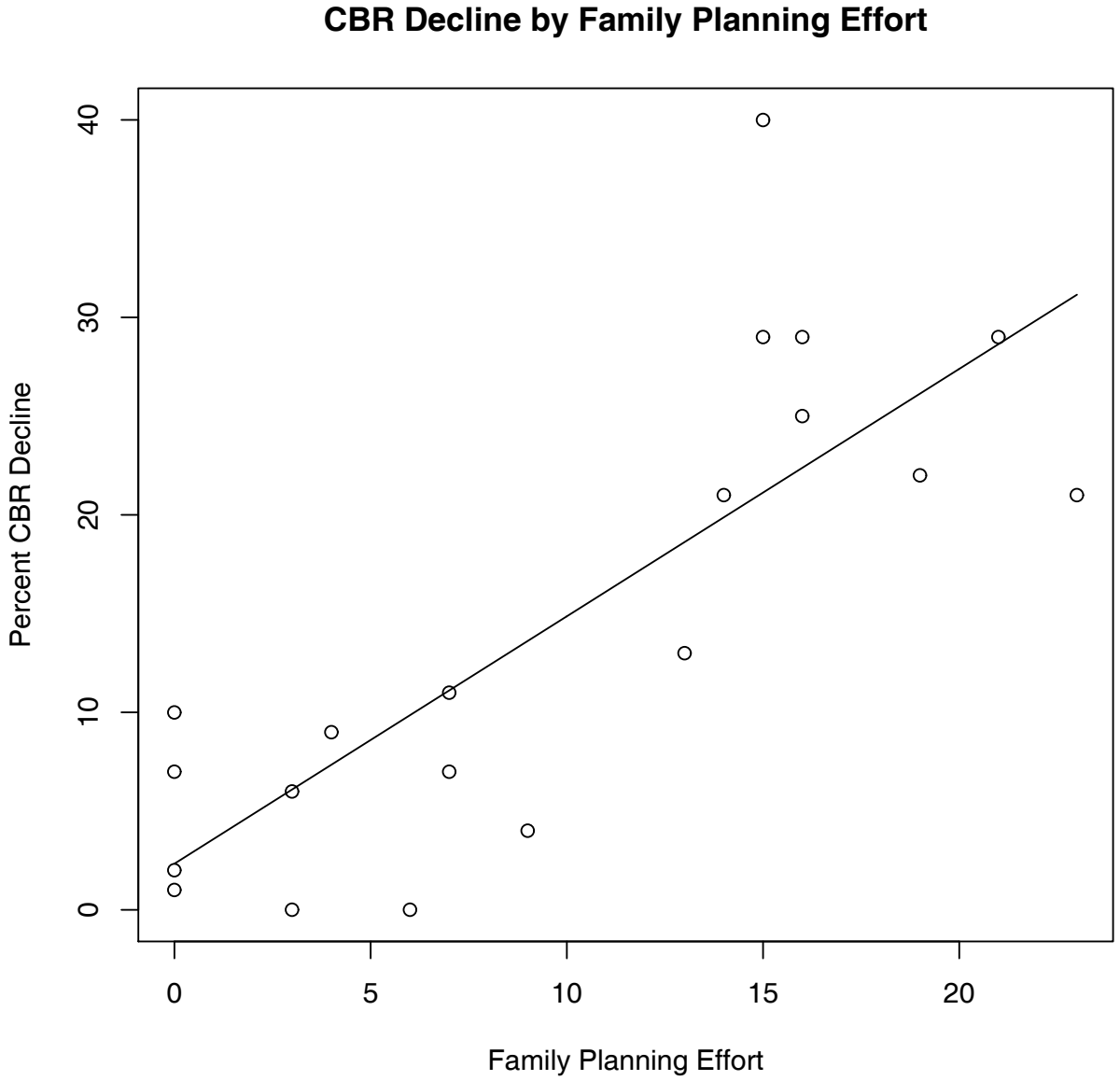
The **least squares** estimate of $\boldsymbol{\beta}$, denoted $\widehat{\boldsymbol{\beta}}$, satisfies

$$\begin{aligned}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) &= \min_{\beta}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \min_{\beta} \parallel \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 \\ &= \min_{\beta} \sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2,\end{aligned}$$

where $\mathbf{x}_i'$ indicates the i-th row of matrix $\mathbf{X}$. The least squares estimate of $\boldsymbol{\beta}$ minimizes the squared Euclidean distance between $\mathbf{y}$ and its mean $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$.

It can be shown that when we make the additional assumption that $y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$, the least squares estimates $\widehat{\boldsymbol{\beta}}$ are also *maximum likelihood* estimates. Maximum likelihood estimates have several desirable properties that you will learn more about in BIOS 661.

Below is a plot of the observed percent CBR decline versus family planning effort along with the least squares regression line. Consider two steps in fitting this line. First, we calculate the mean percent CBR decline, $\overline{y}$. Then, we pivot a line around $(\overline{x}, \overline{y}) = (9.55, 14.3)$, where $\overline{x}$ is the mean family planning effort, until we minimize the sum of squared deviations around the line.


CBR Decline by Family Planning Effort
Percent CBR Decline
0
10
20
30
40
0
5
10
15
20
Family Planning Effort

The least squares estimator, $\widehat{\boldsymbol{\beta}}$, has several good properties.

- First, if the linear model assumption holds, then the least squares estimator is *unbiased*; that is, $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
- Next, we will show later that if the observations are uncorrelated and have constant variance σ^2, then the variance-covariance matrix of the least squares estimator $\widehat{\boldsymbol{\beta}}$ is $\text{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. This estimator is *efficient* in the sense that it has the smallest variance in the class of all unbiased estimators that are linear functions of the data. We call this type of estimator a BLUE (Best Linear Unbiased Estimator).

- If we add the assumption that the error vector is normally distributed, then the least squares estimator is the "best" estimator among *all* unbiased estimators. (We define a "best" estimate to be an unbiased estimate with minimum variance. Other criteria for "best" estimates do exist but will not be addressed further here.)

- Later, we will also show that the sampling distribution of the least squares estimator $\widehat{\boldsymbol{\beta}}$ in *large* samples is approximately multivariate normal with the previously given covariance; that is, $\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$. In the intercept-only model, this means that $\widehat{\boldsymbol{\beta}} = \bar{\mathrm{y}} \sim N(\mu, \frac{\sigma^2}{n})$ in large samples. (This is true even if your errors are not exactly normal but are uncorrelated with constant variance. If they are exactly normal, then the sampling distribution of $\widehat{\boldsymbol{\beta}}$ is exactly normal, and you don't need to worry about attaining a certain sample size in order for asymptotic theory to hold.)

Least squares estimation requires a variety of assumptions:

Existence Assumption

Assume ε_i has finite first and second moments. That is, we observe values of random variables with finite variance. In practice, considering a finite number of subjects ensures that this assumption holds.

Linearity Assumption

We assume the expected values (means) of the response are linear functions of the parameters. That is, we assume

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta},$$

or equivalently, that

$$E(\mathbf{y}) - \mathbf{X}\boldsymbol{\beta} = \mathbf{0} \Leftrightarrow \mathbf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}.$$

The linearity assumption refers to the relationship between the response and parameters. Consider several examples.

1. $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ is linear in parameters, predictors, and error
2. $y_i = \beta_0 + \beta_1 x_i^2 + \varepsilon_i$ is linear in parameters and error
3. $y_i = \beta_0 + x_i^{\beta_1} + \varepsilon_i$ is linear in error
4. $y_i = \beta_0 \exp(-\beta_1 x_i) + \varepsilon_i$ is nonlinear in parameters
5. $y_i = \beta_0 \left[\exp(-\beta_1 x_i)\right] \varepsilon_i \Leftrightarrow \log(y_i) = \log(\beta_0) - \beta_1 x_i + \ln(\varepsilon_i)$
   $\Leftrightarrow$ (5b)$y_i{}^* = \gamma_0 + \gamma_1 x_i + \delta_i$, which is linear in parameters, predictor, and error
6. $y_i = (\beta_0 + \beta_1 x_i)\varepsilon_i$ is nonlinear in parameters

Of these examples, only 1, 2, and 5b meet the linear model assumption of linearity. If linearity does not hold, then we should not attempt to fit linear models, and estimates from linear models will not have the nice properties discussed previously. Nonlinear models, such as the exponential growth or decay model given by $\mathbf{y} = \beta_1 e^{\beta_2 \mathbf{x}} + \varepsilon$, will not be covered in this class.

Independence Assumption

Each element of ε is statistically independent of every other. Equivalently, each element of $\mathbf{y}$ is statistically independent of every other, conditional on $\mathbf{X}$. This generally is fairly obvious to check.

For example, if a set of twins or a parent and child are included in data, the independence assumption will be violated. If positively correlated observations tend to have positively correlated predictor values, standard linear regression will yield anti-conservative tests of associations (i.e., your p-value will be too small). This is the type of error we expect with correlation in family data or longitudinal data.

We will discuss appropriate models for correlated data later in BIOS 663.

Homogeneity Assumption

We assume each element of $\boldsymbol{\varepsilon}$ has the same variance σ^2. Equivalently, $\text{Cov}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}$. It is possible to check this assumption, and we will spend a good deal of time later in the course discussing ways to check homogeneity of variances (which is sometimes called **homoscedasticity**). *Discussion of Homogeneity:*

- σ_i^2 is the variance of the error for subject i.
- $\sigma_i^2 = \sigma_i^2(y_i \mid x_{i1}, x_{i2}, \ldots, x_{ip})$ is the variance of the response, conditional on the value of the covariates $\mathbf{x}_i$ for subject i.
- Homogeneity of variances means that $\sigma_i^2 = \sigma^2$ for all subjects i.
- Homogeneity also means that the variance about the regression function $E(\mathbf{y} \mid \mathbf{X})$ is constant.

Error Covariance Matrix

Recall that we assume errors have mean zero. So $E(\boldsymbol{\varepsilon}) = \mathbf{0}$. Thus

$$\begin{aligned}
\text{Cov}(\boldsymbol{\varepsilon}) &= E\left[(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}))(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon}))'\right] \\
&= E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') - \boldsymbol{\varepsilon}E(\boldsymbol{\varepsilon})' - E(\boldsymbol{\varepsilon})\boldsymbol{\varepsilon}' + E(E(\boldsymbol{\varepsilon})E(\boldsymbol{\varepsilon})') \\
&= E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') - \mathbf{0} - \mathbf{0} + \mathbf{0} = \mathbf{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')_{\mathbf{n}\times\mathbf{n}} \\
&= \begin{bmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2)\ldots & E(\varepsilon_1\varepsilon_n) \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) & \vdots \\ \vdots & & \vdots \\ E(\varepsilon_n\varepsilon_1) & \ldots & \ldots \quad E(\varepsilon_n^2) \end{bmatrix}.
\end{aligned}$$

If the covariance $E(\varepsilon_i\varepsilon_j) = 0$, then the correlation

$$\frac{E(\varepsilon_i\varepsilon_j)}{\sqrt{E(\varepsilon_i^2)}\sqrt{E(\varepsilon_j^2)}} = 0,$$

and vice versa. Although independence of two random observations

always implies zero correlation (and zero covariance), the converse is not true. That is, unless we have Gaussian random variables, zero covariance does *not* imply independence.

The independence assumption means that the error covariance matrix is a diagonal matrix. In addition, because the covariance of a variable with itself equals its variance, we have

$$\begin{aligned} \text{Cov}(\boldsymbol{\varepsilon}) &= E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \\ &= \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & \dots & \dots & \sigma_n^2 \end{bmatrix}. \end{aligned}$$

The homogeneity assumption implies that $\sigma_i^2 = \sigma^2$ so that

$$\begin{aligned} \text{Cov}(\boldsymbol{\varepsilon}) &= E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') \\ &= \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & \dots & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}_n. \end{aligned}$$

For example, if the CBR decline is more variable for countries with low family planning effort than for countries with high family planning effort, conditional on other covariates of interest, the homogeneity of variances assumption would not hold.

Gaussian Errors Assumption

The Gaussian errors assumption is that $\varepsilon_i \sim N(0, \sigma_i^2)$. Assuming homogeneity of variances as well, we have that $\varepsilon_i \sim N(0, \sigma^2)$. This implies that $y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$ so that

$$f(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)^2\right).$$

The normality assumption is not needed for validity of least squares estimation. However, adding the assumption of Gaussian errors means that the least squares estimate of $\boldsymbol{\beta}$ is also a maximum likelihood estimate and a minimum variance unbiased estimate. The assumption of Gaussian errors also allows simple construction of exact small-sample hypothesis tests.

We combine all five assumptions as follows:

$$y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2),$$

with $y_i \perp y_j$ for $i \neq j$. Muller and Fetterman use the mnemonic *HILE Gauss* to describe the five assumptions.

Under the five assumptions (HILE Gauss), the $\{\varepsilon_i\}$ are i.i.d.(independent and identically distributed). However, the $\{y_i\}$ in general are not i.i.d.as $E(y_i) \neq E(y_j)$ because $\mathbf{x}_i' \neq \mathbf{x}_j'$ for all $i \neq j$. The $\{y_i\}$ are independent Gaussian random variables with equal variances, but they do not necessarily have equal expected values.

Defining the Normal Equations

For $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, there are two distinct sets of parameters:

- $\boldsymbol{\beta}$ with p elements, and
- σ^2 with 1 element.

We almost always assume $n \gg p$ so that the sample size $\gg$ the number of parameters.

Once the parameters $\boldsymbol{\beta}$ have been estimated, one may compute $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, the *predicted values* of the outcome. The *residuals* measure the accuracy of the prediction and are given by

$$\mathbf{y} - \widehat{\mathbf{y}} = \widehat{\boldsymbol{\varepsilon}} \neq \boldsymbol{\varepsilon}.$$

The error, ε_i, is unobserved, while the residual, $\widehat{\varepsilon}_i$, is observed (and an estimate of ε_i).

We seek estimates $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}^2$ that are optimal in some sense. One criterion is least squares, which computes $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}^2$ that minimize the average squared distance from observed outcomes to predicted outcomes.

Total squared error of prediction, called the residual SS or the sum of squares for error (SSE), is

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2 \\ &= \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

To find $\widehat{\boldsymbol{\beta}}$ that minimizes SSE take derivatives:

$$\frac{\partial SSE}{\partial \boldsymbol{\beta}} = \mathbf{0} - 2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$$

Set $\partial SSE/\partial \boldsymbol{\beta} = \mathbf{0_{p \times 1}}$ to create a system of p simultaneous equations:

$$\mathbf{0} \ = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

Rearranging terms yields the *normal equations*:

$$(\mathbf{X}'\mathbf{X})_{p \times p} \ \widehat{\boldsymbol{\beta}}_{p \times 1} = \mathbf{X}'_{p \times n} \ \mathbf{y}_{n \times 1}.$$

A $\widehat{\boldsymbol{\beta}}$ that solves the normal equations yields predicted values $\widehat{\mathbf{y}}$ and residuals $\widehat{\boldsymbol{\varepsilon}}$ such that $\widehat{\boldsymbol{y}}'\widehat{\boldsymbol{\varepsilon}} = 0$; that is, the predicted values and residuals are orthogonal (normal to each other).

Solving Normal Equations if $\mathbf{X}$ is Full Rank

If $\boldsymbol{X}'\boldsymbol{X}$ is full rank (i.e.,rank$(\mathbf{X}) = p$ with $n > p$), then $\mathbf{X}'\mathbf{X}$ has a unique inverse, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Then we solve the normal equations:

$$\begin{aligned}
(\mathbf{X}'\mathbf{X})\widehat{\boldsymbol{\beta}} &= \mathbf{X}'\mathbf{y} \\
(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
\widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.
\end{aligned}$$

Facts about the Least Squares Estimate:

- $\widehat{\boldsymbol{\beta}}$ is the unique least squares estimate.
- $\widehat{\boldsymbol{\beta}}$ is best linear unbiased estimate (BLUE).
- With Gaussian errors, $\widehat{\boldsymbol{\beta}}$ is the MLE and minimum variance unbiased estimator.

Solving Normal Equations for X Less Than Full Rank

If $\boldsymbol{X}$ is less than full rank, say rank$(\boldsymbol{X}) = r < p$, then $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ does not exist.

Solutions:

1. Drop some covariates.
2. Generalized Inverse ($\hat{\boldsymbol{\beta}}$ not unique)
3. Penalized regression methods, such as Ridge regression, which minimizes

$$\min_{\beta} \parallel \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \parallel^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where λ is a tuning parameter that can be decided by cross-validation. In other words, we penalize the size of the regression coefficients The solution is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda I_{p\times p})^{-1}\mathbf{X}'\mathbf{y}$

Hat Matrix From now on, we assume $\mathbf{X}$ is full rank, unless otherwise specified.

The *hat matrix* is

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

The hat matrix has the special properties of idempotency ($\mathbf{HH} = \mathbf{H}$) and symmetry. In addition, rank$(\mathbf{H}) = r =$ rank$(\mathbf{X})$, and rank$(\mathbf{I} - \mathbf{H}) = n - r$. The prediction value can be written as

$$\widehat{\mathbf{y}} \quad = \quad \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y},$$

$\mathbf{H}$ is called the hat matrix because $\mathbf{H}\mathbf{y} = \widehat{\mathbf{y}}$. The residuals

$$\widehat{\boldsymbol{\varepsilon}} \quad = \quad \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Estimating σ^2

The MLE of σ^2 is $\frac{SSE}{n} = \frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{n}$.

$E\left[\frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{n}\right] = \sigma^2\left[\frac{n-r}{n}\right]$, so the MLE of σ^2 is biased.

Recall: Calculating Sample Variance

When we calculate a variance, first we calculate a mean, say $\overline{y}$, and then we find the distance of each point from the mean, $y_i - \overline{y}$, $i = 1, \ldots, n$. (Note: by definition of the mean, $\sum_{i=1}^{n}(y_i - \overline{y}) = 0$; for this reason, raw deviations are not a useful variance measure.) The *sum of squares* $\sum_{i=1}^{n}(y_i - \overline{y})^2$ is a useful measure of variability because it increases as the data are more dispersed about the mean. However, this measure also depends on n, and therefore it is not as useful in comparing groups. For comparison purposes, we convert the *sum of squares* to a *variance* by dividing by $n - 1$, where n is the number of subjects in a group.

Why not divide the sum of squares by n?

The reason we do not divide by n is that we do not have n *independent* pieces of information about the variance. First, we calculated a mean, and then we calculated deviations from the mean. If we calculate the first $n-1$ deviations, then we know the last $\sum_{i=1}^{n}(y_i - \overline{y}) = 0$. The independent pieces of information contributing to a statistic are called the *degrees of freedom*.

By similar logic, $\widehat{\boldsymbol{\sigma}}^2 = \frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{n-r} = \frac{SSE}{n-r}$ is an unbiased estimate of σ^2.

$\widehat{\sigma}^2$ is a *quadratic form* in $\boldsymbol{y}$:

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-r} = \frac{(\mathbf{y}-\widehat{\mathbf{y}})'(\mathbf{y}-\widehat{\mathbf{y}})}{n-r} = \frac{\mathbf{y}'\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{y}}{n-r} \\ &= \frac{\mathbf{y}'\left[\mathbf{I}-\mathbf{H}\right]\mathbf{y}}{n-r}\end{aligned}$$

Example: Ozone Exposure Assessment

One common problem in environmental epidemiology is determining personal exposures to environmental toxicants, such as ozone in the air. Adverse health effects associated with ozone exposure include increased incidence of cough, chest pain, and other respiratory symptoms. Although outdoor ozone concentrations are monitored by the Environmental Protection Agency (EPA), it is more difficult to determine indoor concentrations. Personal exposures, which vary based on the proportion of time spent outdoors, at home, in the workplace, and in other areas, are even more difficult to measure. Using outdoor ozone concentrations as a crude approximation of personal exposure can lead to substantial measurement error, which can in turn lead to biased parameter estimates.

We consider data from a study conducted in State College, Pennsylvania, in which children wore small (2 cm $\times$ 3 cm) personal

ozone samplers. Investigators wish to model personal ozone exposures ($O_{PERSONAL}$) measured by the samplers as a function of outdoor ($O_{OUTDOOR}$) ozone concentrations (measured at a central State College site), home indoor ozone concentrations (O_{HOME}) for each child, and the proportion of time each child spent outdoors ($TIME_{OUTDOORS}$).

The data we consider include 64 measurements of personal ozone exposure (in parts per billion or ppb) along with the corresponding measurements of outdoor ozone concentrations, home indoor ozone concentrations, and the proportion of time spent outdoors.

This model is written in the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ as follows:

$$\mathbf{y}_{64\times 1} = \mathbf{X}_{64\times 4}\,\boldsymbol{\beta}_{4\times 1} + \boldsymbol{\varepsilon}_{64\times 1}$$

$$\begin{bmatrix} 26.29 \\ 3.30 \\ 29.28 \\ 28.55 \\ 38.28 \\ \vdots \end{bmatrix}_{64\times 1} = \begin{bmatrix} 1 & 35.88 & 22.29 & 0.57 \\ 1 & 34.37 & 22.27 & 0.17 \\ 1 & 45.96 & 23.40 & 0.00 \\ 1 & 92.56 & 7.14 & 0.26 \\ 1 & 30.44 & 35.38 & 0.69 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}_{64\times 4} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \vdots \end{bmatrix}_{64\times 1}$$

- β_0 is called the intercept, which is the expected value of $O_{PERSONAL}$ when all other predictors

$$(O_{OUTDOOR}, O_{HOME}, TIME_{OUTDOORS})$$

take the value zero.

- β_1 is the slope for outdoor ozone. It is interpreted as the expected ppb increase in personal exposure for a one ppb increase in outdoor ozone concentration.
- β_2 is the slope for home indoor ozone. It is interpreted as the expected ppb increase in personal exposure for a one ppb increase in home indoor ozone concentration.
- β_3 is the slope for the proportion of time spent outdoors. It is interpreted as the expected ppb increase in personal exposure for an additional one percent of time spent outdoors.

The following R code may be used to obtain parameter estimates for the ozone data.

```
> ozone = read.table("ozone.txt", header = T, sep = " ") # space delimited
> n = nrow(ozone) # number of rows of ozone (obs)
> X = model.matrix(~ outdoor + home + time_out, data = ozone) # predictor matrix
> y = ozone$personal # response
> head(X)

  (Intercept)  outdoor  home time_out
1           1 35.87771 22.29     0.57
2           1 43.79189 13.97     0.90
3           1 49.81255 18.96     0.55
4           1 34.37366 22.27     0.17
5           1 45.95496 23.40     0.00
6           1 64.76558 39.62     0.30

> bhat = solve(t(X) %*% X) %*% t(X) %*% y # from notes
> yhat=X%*%bhat; # predicted values
> ehat=y-yhat; # residuals
> p=ncol(X); # num columns of X
> df=n-p; # df
> sse = t(y) %*% y - t(bhat) %*% t(X) %*% y # SSE
> mse=sse/df; # MSE
```

The output from the R is given below.

```
> print(bhat)
```

```
                    [,1]
(Intercept)  3.78348593
outdoor      0.09142005
home         0.59543659
time_out    13.64453832
```

```
> print(mse)
```

```
         [,1]
[1,] 169.1365
```

The same estimates may be obtained from SAS PROC REG.

```
proc reg data=ozone;
model personal=outdoor home time_out;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: personal Personal Ozone Exposure (ppb)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5034.90667	1678.30222	9.92	<.0001
Error	60	10148	169.13652		
Corrected Total	63	15183			

Root MSE	13.00525	R-Square	0.3316
Dependent Mean	23.54578	Adj R-Sq	0.2982
Coeff Var	55.23389		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value
Intercept	Intercept	1	3.78349	4.34206	0.87
outdoor	Outdoor Ozone Concentration (ppb)	1	0.09142	0.09042	1.01
home	Home Indoor Ozone Concentration (ppb)	1	0.59544	0.16478	3.61
time_out	Proportion of Time Spent Outdoors	1	13.64454	7.70973	1.77

Parameter Estimates

Variable	Label	DF	Pr > |t|
Intercept	Intercept	1	0.3870
outdoor	Outdoor Ozone Concentration (ppb)	1	0.3160
home	Home Indoor Ozone Concentration (ppb)	1	0.0006
time_out	Proportion of Time Spent Outdoors	1	0.0818

Hypothesis Testing

The General Linear (Univariate) Hypothesis, GLH

For testing, we assume i.i.d.Gaussian errors. $\boldsymbol{\beta}$ is the matrix of primary parameters, and $\boldsymbol{\theta}_{a\times 1} = \mathbf{C}_{a\times p}\ \boldsymbol{\beta}_{p\times 1}$ is a matrix of secondary parameters, defined by $\mathbf{C}$, the *contrast matrix*. Each row of $\mathbf{C}$ defines a new scalar parameter in terms of the $\boldsymbol{\beta}$'s, e.g.,$\beta_1 - \beta_2$.

Let $\boldsymbol{\theta}_0$ be a matrix of known constants (the hypothesized values). Most often $\boldsymbol{\theta}_0$ is taken to be the zero matrix. The (univariate) general linear hypothesis is

$$\begin{aligned} H_0 : \boldsymbol{\theta}_{a\times 1} &= \boldsymbol{\theta}_0 \\ H_A : \boldsymbol{\theta}_{a\times 1} &\neq \boldsymbol{\theta}_0. \end{aligned}$$

Example: Choosing Contrast Matrix and Secondary Parameter Matrix

For the ozone data, consider the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $O_{PERSONAL}$ as the response and $O_{OUTDOOR}$, O_{HOME}, and $TIME_{OUTDOORS}$ as predictors. Suppose we hypothesize that the outdoor and home exposures have the same effect on personal exposure levels. Thus we have $H_0 : \beta_1 - \beta_2 = 0$. This implies the secondary parameter $\theta = (\beta_1 - \beta_2)$ with corresponding contrast matrix $\mathbf{C} = \begin{bmatrix} 0 & 1 & -1 & 0 \end{bmatrix}$ and $\theta_0 = 0$.

Suppose our hypothesis is that all the slopes are the same. What are the values of $\boldsymbol{\theta}$, $\mathbf{C}$, and $\boldsymbol{\theta}_0$?

Estimability of a Parameter

Choosing appropriate $\boldsymbol{C}$ and $\boldsymbol{\theta}_0$, coupled with fitting appropriate models, allows testing hypotheses about all *estimable* parameters.

(Searle, 1971, p. 180): "A (linear) function of the parameters is defined to be estimable if it is identically equal to some linear function of the expected value of the vector of observations, $\boldsymbol{y}$."

Thus a scalar parameter, $\theta_i = \mathbf{C}_{1\times p}\,\boldsymbol{\beta}_{p\times 1}$, is estimable

$$\Leftrightarrow \mathbf{C}_{1\times p}\,\boldsymbol{\beta}_{p\times 1} = \mathbf{t}'_{1\times n} E(\mathbf{y}_{n\times 1}),$$

for $\boldsymbol{t}$ a vector of constants.

More generally, for a vector we need $\boldsymbol{\theta_{a\times 1}} = \boldsymbol{T_{a\times n}} E\left(\boldsymbol{y}_{n\times 1}\right)$

There always exist r =rank$(\boldsymbol{X})$ distinct and estimable parameters (which are not necessarily elements of $\boldsymbol{\beta}$ but may be linear combinations of elements).

If rank$(\boldsymbol{X}) = r = p$, then $\widehat{\boldsymbol{\beta}}$ exists (uniquely), $\boldsymbol{\beta}$ is estimable, and any (nonzero) $\boldsymbol{C}$ gives estimable $\boldsymbol{\theta}$. This is usually the case with continuous predictors unless some predictors are collinear.

If rank$(\boldsymbol{X}) = r < p$, $\boldsymbol{\beta}$ is not estimable (although as many as r elements may be), and for $\widehat{\boldsymbol{\theta}} = \boldsymbol{C}\boldsymbol{\beta}$, we must check estimability.

To show set of parameters

$$\boldsymbol{\theta}_{a\times 1} = \mathbf{C}_{a\times p}\, \boldsymbol{\beta}_{p\times 1} = \mathbf{T}_{a\times n} E(\mathbf{y}_{n\times 1})$$

is estimable, it suffices to show that $\mathbf{C}_{a\times p} = \mathbf{T}_{a\times n}\mathbf{X}_{n\times p}$.

Estimable $\widehat{\boldsymbol{\theta}}$ shares the optimality of $\widehat{\boldsymbol{\beta}}$ (whatever r is): BLUE for least squares and MLE with Gaussian errors.

Show that $H_0 : \beta_1 = \beta_2$ is estimable for the ozone data.

Testability of a Hypothesis

Consider the likelihood ratio (LR) test. Let $\mathbf{M}_{a\times a} = \mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'$.
Define GLH testability as the (unique) existence of the LR test.

$\boldsymbol{\theta}$ is testable $\Leftrightarrow$

- $\boldsymbol{C}$ is full rank a (no redundancies), and
- $\boldsymbol{\theta}$ is estimable,

OR, equivalently,

- $\boldsymbol{M}$ is full rank a, and
- $\boldsymbol{\theta}$ is estimable.

If $\boldsymbol{X}$ is full rank, then $\boldsymbol{\theta}$ is testable $\Leftrightarrow$
$\boldsymbol{C}$ is full rank a OR $\boldsymbol{M}$ is full rank a (because any $\boldsymbol{\theta}$ is estimable)

Show that $H_0 : \beta_1 = \beta_2$ is testable for the ozone data.

Computation of Test Statistic and p-Value

Define the sums of squares hypothesis as
$SSH_{1\times 1} = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \boldsymbol{M}^{-1} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$.

With HILE Gauss, the likelihood ratio statistic equals

$$\begin{aligned} F_{obs} &= \frac{SSH/a}{SSE/(n-r)} = \frac{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \boldsymbol{M}^{-1} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/a}{\widehat{\sigma}^2} \\ &= \frac{MSH}{MSE} \end{aligned}$$

Under H_0: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, SSH and SSE are scaled χ^2 random variables, with $SSH/\sigma^2 \sim \chi^2(a)$, independently of $SSE/\sigma^2 \sim \chi^2(n-r)$. It can be shown that if $z_1 \sim \chi^2_{d_1}$, $z_2 \sim \chi^2_{d_2}$, and $z_1 \perp z_2$, then $\frac{z_1/d_1}{z_2/d_2}$ follows an $F_{d1,d2}$ distribution. Thus

$$F_{obs} = \frac{[SSH/\sigma^2]/a}{[SSE/\sigma^2]/(n-r)} = \frac{SSH/a}{SSE/(n-r)} \sim F(a, n-r).$$

The p-value equals the probability of observed or more extreme data arising under the null, that is,

$$\text{p-val} = \Pr\{F(a, n-r) \geq F_{obs}\} = 1 - \Pr\{F(a, n-r) < F_{obs}\}.$$

Reject H_0 if $F_{obs} > f_{crit} = F^{-1}(1-\alpha, a, n-r)$.

Obtain f_{crit} in SAS as FINV(prob, df_1, df_2), the value of an F statistic with df_1 numerator and df_2 denominator degrees of freedom, such that $\Pr\{F \leq f_{crit}\}$ =prob. In R, use qf(prob, df_1, df_2). (To get p-values in R, use 1 - pf(crit, df_1, df_2).

All linear model GLH tests correspond to comparing two models, the "full" model, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and a reduced model defined by constraints. This concept lies at the heart of any LR test and is critical in understanding any particular GLH test.

Example: Computing a GLH Test

For the ozone data, again consider the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $O_{PERSONAL}$ as the response and $O_{OUTDOOR}$, O_{HOME}, and $TIME_{OUTDOORS}$ as predictors. Suppose we wish to test that all slopes are equal. Then $H_0 : \boldsymbol{\theta} = \begin{bmatrix} \beta_1 - \beta_2 \\ \beta_2 - \beta_3 \end{bmatrix} = \mathbf{0}$, with corresponding contrast matrix $\mathbf{C} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$.

The additional code needed to fit the model, along with PROC REG code, is given below.

```
> C = matrix(c(0,1,-1,0,0,0,1,-1), nrow = 2, byrow = T) # contrast matrix
> print(C)

     [,1] [,2] [,3] [,4]
[1,]    0    1   -1    0
[2,]    0    0    1   -1

> M=C %*% solve(t(X)%*%X)%*%t(C)
> thetahat=C%*%bhat
> ssh=t(thetahat)%*%solve(M)%*%thetahat
> f_obs=(ssh/nrow(thetahat))/mse
> p=1-pf(f_obs,2,60)
```

The results from R are below.

```
> print(ssh)

          [,1]
[1,] 1237.324

> print(f_obs)

          [,1]
[1,] 3.657767

> print(p)

             [,1]
[1,] 0.03170141
```

We reject the null hypothesis and conclude that not all slopes are identical. At least one slope is not equal to the others.

This output corresponds to PROC REG below.

```
proc reg data=ozone;
model personal=outdoor home time_out;
test outdoor-home=0, home-time_out=0;
run;
```

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable personal

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	618.66202	3.66	0.0317
Denominator	60	169.13652		

Wald Tests

For a single coefficient β_j, we can test $H_0 : \beta_j = 0$ if β_j is estimable. SAS automatically reports *Wald* tests for parameters in a regression model. These are tests of the hypothesis $H_0 : \beta_j = 0$ for each j.

We know that in large samples (or in small samples if our errors are exactly normal), $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$. The variance of $\widehat{\beta}_j$ is the j^{th} diagonal element of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Using properties of the standard normal distribution, we can base our test on the ratio

$$t = \frac{\widehat{\beta}_j - 0}{\sqrt{\mathsf{var}(\widehat{\beta}_j)}}.$$

Usually, we do not know σ^2 exactly and obtain an estimate as $\widehat{\sigma}^2 = \frac{SSE}{dfE}$. If we do know σ^2 exactly, then $t \sim N(0,1)$. If we estimate σ^2 from the data, then $t \sim t_{dfE}$, a Student's t distribution with dfE degrees of freedom.

One and Two-Sided Tests

One-sided tests exist only for scalar hypotheses, not for vector hypotheses. Let $a =\#$ rows of $\boldsymbol{C}$. If $a = 1$, then $\boldsymbol{\theta}$ is a scalar (1×1) and $F_{obs}(1, n-r) = t^2(n-r)$, where $t(n-r)$ is t-statistic with d.f. of $n - r$.

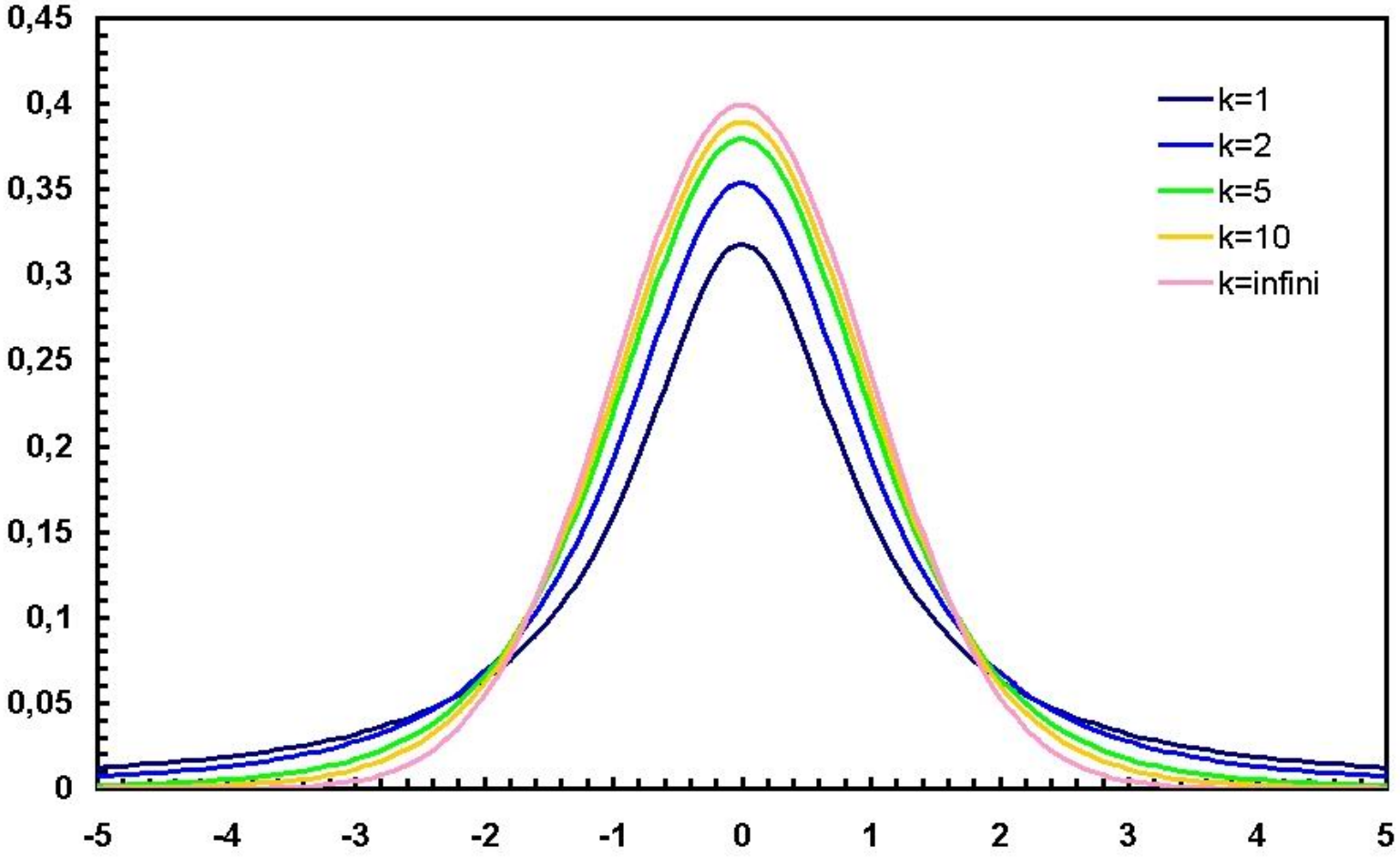
- Two-sided test: $H_0 : \theta = \theta_0$ vs. $H_A : \theta \neq \theta_0$
- One-sided tests: $\begin{cases} H_0\text{:} \;\; \theta = \theta_0 \;\; \text{vs.} \;\; H_\text{A}\text{:} \;\; \theta > \theta_0 \\ H_0\text{:} \;\; \theta = \theta_0 \;\; \text{vs.} \;\; H_\text{A}\text{:} \;\; \theta < \theta_0 \end{cases}$

Conducting a test of size α by t-test:

- A two-sided t-test, $H_\text{A}\text{:} \;\; \theta \neq \theta_0$, uses the $\alpha/2$ and $(1 - \alpha/2)$ critical values.
- A one-sided t-test,, $H_\text{A}\text{:} \;\; \theta < \theta_0$, uses the α critical value.
- A one-sided t-test,, $H_\text{A}\text{:} \;\; \theta > \theta_0$, uses the $(1 - \alpha)$ critical value.

In all cases, one rejects H_0 if the test statistic is farther from zero than

the critical value.

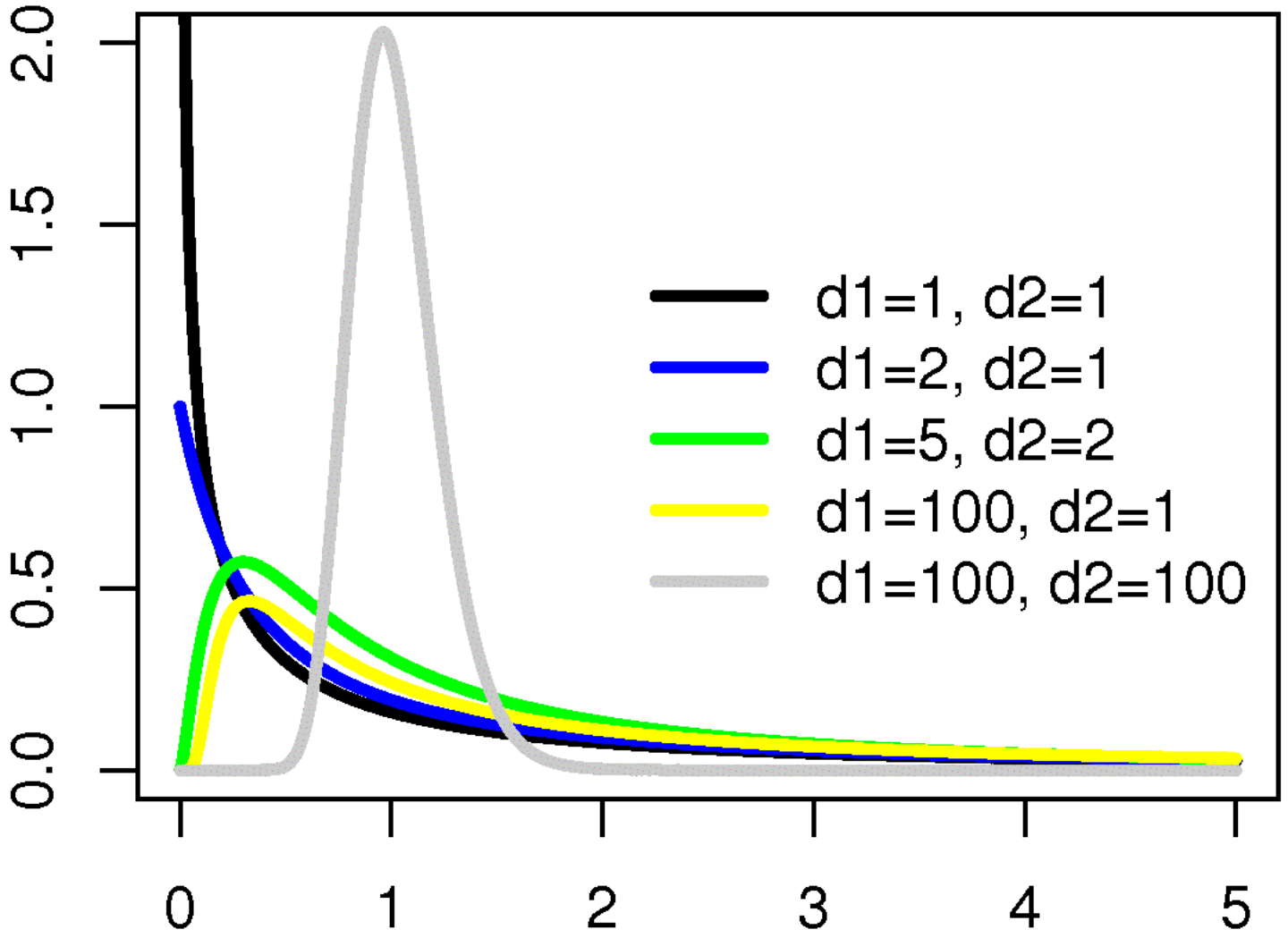


F-tests

- A two-sided F test, H_A: $\theta \neq \theta_0$, uses the α critical value.
- A one-sided F test, H_A: $\theta < \theta_0$, uses the 2α critical value and also requires appropriate sign of the difference.

Recall the definition of a CDF: $F_X(x) = \Pr\{X \leq x\}$.

$$F_F(f) = \Pr\{F \leq f\} = \int_0^f f_F(u)du$$


0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5
d1=1, d2=1
d1=2, d2=1
d1=5, d2=2
d1=100, d2=1
d1=100, d2=100

Next: Distributional Results

Reading Assignment:

- Muller and Fetterman Chapter 3: "Some Distributions for the GLM" (Required)