# Unit 2: Linear models: Inference

Eugene Katsevich

September 22, 2021

We now understand the least squares estimator $\widehat{\boldsymbol{\beta}}$ from geometric and algebraic points of view. In Unit 2, we will switch to a probabilistic perspective to derive inferential statements for linear models, in the form of hypothesis tests and confidence intervals. In order to facilitate this, we will assume that the error terms are normally distributed:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{where } \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n). \tag{1}$$

## 1 Building blocks for linear model inference

First we put in place some building blocks: The multivariate normal distribution (Section 1.1), the distributions of linear regression estimates and residuals (Section 1.2), and estimation of the noise variance $\sigma^2$ (Section 1.3).

### 1.1 The multivariate normal distribution

Recall that a random vector $\boldsymbol{w} \in \mathbb{R}^d$ has a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariate matrix $\boldsymbol{\Sigma}$ if it has probability density

$$p(\boldsymbol{w}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{w}-\boldsymbol{\mu})\right).$$

These random vectors have lots of special properties, including:

- (Linear transformation) If $\boldsymbol{w} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{A}\boldsymbol{w} + \boldsymbol{b} \sim N(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.
- (Independence) If $\begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \Sigma_{22} \end{pmatrix}\right)$, then $\boldsymbol{w}_1 \perp\!\!\!\perp \boldsymbol{w}_2$ if and only if $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$.

An important distribution related to the multivariate normal is the $\chi^2_d$ (chi-squared with $d$ degrees of freedom) distribution, defined as

$$\chi^2_d \equiv \sum_{j=1}^d w_j^2 \quad \text{for} \quad w_1, \ldots, w_d \overset{\text{i.i.d.}}{\sim} N(0,1).$$

### 1.2 The distributions of linear regression estimates and residuals

The most important distributional result in linear regression is that

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}). \tag{2}$$

Indeed, by the linear transformation property of the multivariate normal distribution,

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) \Longrightarrow \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \sim N((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}_n\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}) = N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}).$$

Next, let's consider the joint distribution of $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. We have

$$\begin{aligned}\begin{pmatrix}\widehat{\boldsymbol{\mu}} \\ \widehat{\boldsymbol{\epsilon}}\end{pmatrix} = \begin{pmatrix}\boldsymbol{H}\boldsymbol{y} \\ (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}\end{pmatrix} = \begin{pmatrix}\boldsymbol{H} \\ \boldsymbol{I}-\boldsymbol{H}\end{pmatrix}\boldsymbol{y} &\sim N\left(\begin{pmatrix}\boldsymbol{H} \\ \boldsymbol{I}-\boldsymbol{H}\end{pmatrix}\boldsymbol{X}\boldsymbol{\beta}, \begin{pmatrix}\boldsymbol{H} \\ \boldsymbol{I}-\boldsymbol{H}\end{pmatrix}\cdot\sigma^2\boldsymbol{I}\begin{pmatrix}\boldsymbol{H} & \boldsymbol{I}-\boldsymbol{H}\end{pmatrix}\right) \\ &= N\left(\begin{pmatrix}\boldsymbol{X}\boldsymbol{\beta} \\ \boldsymbol{0}\end{pmatrix}, \begin{pmatrix}\sigma^2\boldsymbol{H} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma^2(\boldsymbol{I}-\boldsymbol{H})\end{pmatrix}\right).\end{aligned} \tag{3}$$

In other words,

$$\widehat{\boldsymbol{\mu}} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{H}) \quad \text{and} \quad \widehat{\boldsymbol{\epsilon}} \sim N(\boldsymbol{0}, \sigma^2(\boldsymbol{I}-\boldsymbol{H})), \quad \text{with} \quad \widehat{\boldsymbol{\mu}} \perp\!\!\!\perp \widehat{\boldsymbol{\epsilon}}. \tag{4}$$

Since $\widehat{\boldsymbol{\beta}}$ is a deterministic function of $\widehat{\boldsymbol{\mu}}$ (in particular, $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{\mu}}$), it also follows that

$$\widehat{\boldsymbol{\beta}} \perp\!\!\!\perp \widehat{\boldsymbol{\epsilon}}. \tag{5}$$

### 1.3 Estimation of the noise variance $\sigma^2$

We can't quite do inference for $\boldsymbol{\beta}$ based on the distributional result (2) because the noise variance $\sigma^2$ is unknown to us. Intuitively, since $\sigma^2 = \mathbb{E}[\epsilon_i^2]$, we can get an estimate of $\sigma^2$ by looking at the quantity $\|\widehat{\boldsymbol{\epsilon}}\|^2$. To get the distribution of this quantity, we need the following lemma:

**Lemma 1.1.** *Let $\boldsymbol{w} \sim N(\boldsymbol{0}, \boldsymbol{P})$ for some projection matrix $\boldsymbol{P}$. Then, $\|\boldsymbol{w}\|^2 \sim \chi^2_d$, where $d = \text{trace}(\boldsymbol{P})$ is the dimension of the subspace onto which $\boldsymbol{P}$ projects.*

*Proof.* Let $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$ be an eigenvalue decomposition of $\boldsymbol{P}$, where $\boldsymbol{U}$ is orthogonal and $\boldsymbol{D}$ is a diagonal matrix with $D_{ii} \in \{0, 1\}$. We have $\boldsymbol{w} \overset{d}{=} \boldsymbol{U}\boldsymbol{D}\boldsymbol{z}$ for $\boldsymbol{z} \sim N(0, \boldsymbol{I}_n)$. Therefore,

$$\|\boldsymbol{w}\|^2 = \|\boldsymbol{D}\boldsymbol{z}\|^2 = \sum_{i:D_{ii}=1} z_i^2 \sim \chi^2_d, \quad \text{where } d = |\{i : D_{ii} = 1\}| = \text{trace}(D) = \text{trace}(\boldsymbol{P}).$$

□

Recall that $\boldsymbol{I} - \boldsymbol{H}$ is a projection onto the $(n-p)$-dimensional space $C(\boldsymbol{X})^\perp$, so by Lemma 1.1 and equation (4) we have

$$\|\widehat{\boldsymbol{\epsilon}}\|^2 \sim \sigma^2\chi^2_{n-p}. \tag{6}$$

From this result, it follows that $\mathbb{E}[\|\widehat{\boldsymbol{\epsilon}}\|^2] = n - p$, so

$$\widehat{\sigma}^2 \equiv \frac{1}{n-p}\|\widehat{\boldsymbol{\epsilon}}\|^2 \tag{7}$$

is an unbiased estimate for $\sigma^2$. Why does the denominator need to be $n - p$ rather than $n$ for the estimator above to be unbiased? The reason for this is that the residuals $\widehat{\boldsymbol{\epsilon}}$ are the projection of the true noise vector $\boldsymbol{\epsilon}$ onto the lower-dimensional subspace $C(\boldsymbol{X})^\perp$. To see this, note that

$$\widehat{\boldsymbol{\epsilon}} = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y} = (\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\epsilon}. \tag{8}$$

# 2 Hypothesis testing

Typically two types of null hypotheses are tested in a regression setting: Those involving one-dimensional parameters and those involving multi-dimensional parameters. For example, consider the null hypotheses $H_0 : \beta_j = 0$ and $H_0 : \boldsymbol{\beta}_S = \mathbf{0}$ for $S \subseteq \{0, 1, \ldots, p-1\}$, respectively. We discuss tests of these two kinds of hypothesis in Sections 2.1 and 2.2, and then discuss the power of these tests in Section 2.3.

## 2.1 Testing a one-dimensional parameter

**$t$-test for a single coefficient.** The most common question to ask in a linear regression context is: Is the $j$th predictor associated with the response, when controlling for the other predictors? In the language of hypothesis testing, this corresponds to the null hypothesis

$$H_0 : \beta_j = 0. \tag{9}$$

According to (2), we have $\widehat{\beta}_j \sim N(0, \sigma^2/s_j^2)$, where, as we learned in Unit 1,

$$s_j^2 \equiv [(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{jj}]^{-1} = \|\boldsymbol{x}_{*j}^{\perp}\|^2. \tag{10}$$

Therefore,

$$\frac{\widehat{\beta}_j}{\sigma/s_j} \sim N(0, 1), \tag{11}$$

and we are tempted to define a level $\alpha$ test of the null hypothesis (9) based on this normal distribution. While this is infeasible since we don't know $\sigma^2$, we can substitute in the unbiased estimate (7) derived in Section 1.3. Then,

$$\mathrm{SE}_j \equiv \frac{\widehat{\sigma}}{s_j} \quad \text{is the standard error of } \widehat{\beta}_j, \tag{12}$$

which is an approximation to the standard deviation of $\widehat{\beta}_j$. Dividing $\widehat{\beta}_j$ by its standard error gives us the $t$-statistic

$$t_j \equiv \frac{\widehat{\beta}_j}{\mathrm{SE}_j} = \frac{\widehat{\beta}_j}{\sqrt{\frac{1}{n-p}\|\widehat{\boldsymbol{\epsilon}}\|^2}/s_j}. \tag{13}$$

This statistic is *pivotal*, in the sense that it has the same distribution for any $\boldsymbol{\beta}$ such that $\beta_j = 0$. Indeed, we can rewrite it as

$$t_j = \frac{\frac{\widehat{\beta}}{\sigma/s_j}}{\sqrt{\frac{\sigma^{-2}\|\widehat{\boldsymbol{\epsilon}}\|^2}{n-p}}}. \tag{14}$$

Recalling the independence of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\epsilon}}$ (5), the scaled chi square distribution of $\|\widehat{\boldsymbol{\epsilon}}\|^2$ (6), the standard normal distribution of $\frac{\widehat{\beta}}{\sigma/s_j}$ (11), we find that

$$\text{under } H_0 : \beta_j = 0, \quad t_j \sim \frac{N(0,1)}{\sqrt{\frac{1}{n-p}\chi^2_{n-p}}}, \quad \text{with numerator and denominator independent.} \tag{15}$$

The latter distribution is called the *$t$ distribution with $n-p$ degrees of freedom* and denoted $t_{n-p}$. This paves the way for the two-sided $t$-test:

$$\phi_t(\boldsymbol{X}, \boldsymbol{y}) = \mathbb{1}(|t_j| > t_{n-p}(1-\alpha/2)), \tag{16}$$

where $t_{n-p}(1-\alpha/2)$ denotes the $1-\alpha/2$ quantile of $t_{n-p}$. Note that, by the law of large numbers,

$$\frac{1}{n-p}\chi^2_{n-p} \xrightarrow{P} 1 \quad \text{as} \quad n-p \to \infty, \tag{17}$$

so for large $n-p$ we have $t_j \sim t_{n-p} \approx N(0,1)$. Hence, the $t$-test is approximately equal to the following $z$-test:

$$\phi_t(\boldsymbol{X}, \boldsymbol{y}) \approx \phi_z(\boldsymbol{X}, \boldsymbol{y}) \equiv \mathbb{1}(|t_j| > z(1-\alpha/2)), \tag{18}$$

where $z(1-\alpha/2)$ is the $1-\alpha/2$ quantile of $N(0,1)$. The $t$-test can also be defined in a one-sided fashion, if power against one-sided alternatives is desired.

**Example: One-sample model.** Consider the intercept-only linear regression model $y = \beta_0 + \epsilon$, and let's apply the $t$-test derived above to test the null hypothesis $H_0 : \beta_0 = 0$. We have $\widehat{\beta}_0 = \bar{y}$. Furthermore, we have

$$\mathrm{SE}_0^2 = \frac{\widehat{\sigma}^2}{n}, \quad \text{where} \quad \widehat{\sigma}^2 = \frac{1}{n-1}\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2. \tag{19}$$

Hence, we obtain the $t$ statistic

$$t = \frac{\widehat{\beta}_0}{\mathrm{SE}_0} = \frac{\sqrt{n}\bar{y}}{\sqrt{\frac{1}{n-1}\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2}}. \tag{20}$$

According to the theory above, this test statistic has a null distribution of $t_{n-1}$.

**Example: Two-sample model.** Suppose we have $x_1 \in \{0,1\}$, in which case the linear regression $y = \beta_0 + \beta_1 x_1 + \epsilon$ becomes a two-sample model. We can rewrite this model as

$$y_i \sim \begin{cases} N(\beta_0, \sigma^2) & \text{for } x_i = 0; \\ N(\beta_0 + \beta_1, \sigma^2) & \text{for } x_i = 1. \end{cases} \tag{21}$$

It is often of interest to test the null hypothesis $H_0 : \beta_1 = 0$, i.e. that the two groups have equal means. Let's define

$$\bar{y}_0 \equiv \frac{1}{n_0}\sum_{i:x_i=0} y_i, \quad \bar{y}_1 \equiv \frac{1}{n_1}\sum_{i:x_i=1} y_i, \quad \text{where} \quad n_0 = |\{i : x_i = 0\}| \text{ and } n_1 = |\{i : x_i = 1\}|. \tag{22}$$

Then, we have seen before that $\widehat{\beta}_0 = \bar{y}_0$ and $\widehat{\beta}_1 = \bar{y}_1 - \bar{y}_0$. We can compute that

$$s_1^2 \equiv \|\boldsymbol{x}_{*1}^{\perp}\|^2 = \|\boldsymbol{x}_{*1} - \frac{n_1}{n}\mathbf{1}\|^2 = n_1\frac{n_0^2}{n^2} + n_0\frac{n_1^2}{n^2} = \frac{n_0 n_1}{n} = \frac{1}{\frac{1}{n_0} + \frac{1}{n_1}} \tag{23}$$

and

$$\widehat{\sigma}^2 = \frac{1}{n-2}\left(\sum_{i:x_i=0}(y_i - \bar{y}_0)^2 + \sum_{i:x_i=1}(y_i - \bar{y}_1)^2\right). \tag{24}$$

Therefore, we arrive at a $t$-statistic of

$$t = \frac{\sqrt{\frac{1}{\frac{1}{n_0}+\frac{1}{n_1}}}(\bar{y}_1 - \bar{y}_0)}{\sqrt{\frac{1}{n-2}\left(\sum_{i:x_i=0}(y_i - \bar{y}_0)^2 + \sum_{i:x_i=1}(y_i - \bar{y}_1)^2\right)}}. \tag{25}$$

Under the null hypothesis, this statistic has a distribution of $t_{n-2}$.

**$t$-test for a contrast among coefficients.** Given a vector $\boldsymbol{c} \in \mathbb{R}^p$, the quantity $\boldsymbol{c}^T\boldsymbol{\beta}$ is sometimes called a *contrast*. For example, suppose $\boldsymbol{c} = (1, -1, 0, \ldots, 0)$. Then, $\boldsymbol{c}^T\boldsymbol{\beta} = \beta_1 - \beta_2$ is the difference in effects of the first and second predictors. We are sometimes interested in testing whether such a contrast is equal to zero, i.e. $H_0 : \boldsymbol{c}^T\boldsymbol{\beta} = 0$. While this hypothesis can involve two or more of the predictors, the parameter $\boldsymbol{c}^T\boldsymbol{\beta}$ is still one-dimensional and therefore we can still apply a $t$-test. Going back to the distribution $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$, we find that

$$\boldsymbol{c}^T\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{c}^T\boldsymbol{\beta}, \sigma^2\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}).$$

Therefore, under the null hypothesis that $\boldsymbol{c}^T\boldsymbol{\beta} = 0$, we can derive that

$$\frac{\boldsymbol{c}^T\widehat{\boldsymbol{\beta}}}{\widehat{\sigma}\sqrt{\boldsymbol{c}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{c}}} \sim t_{n-p}, \tag{26}$$

giving us another $t$-test. Note that the $t$-tests described above can be recovered from this more general formulation by setting $\boldsymbol{c} = \boldsymbol{e}_j$, the indicator vector with $j$th coordinate equal to 1 and all others equal to zero.

## 2.2 Testing a multi-dimensional parameter

**$F$-test for a group of coefficients.** Now we move on to the case of testing a multi-dimensional parameter: $H_0 : \boldsymbol{\beta}_S = \boldsymbol{0}$ for some $S \subseteq \{0, 1, \ldots, p-1\}$. In other words, we would like to test

$$H_0 : \boldsymbol{y} = \boldsymbol{X}_{*,\text{-}S}\boldsymbol{\beta}_{-S} + \boldsymbol{\epsilon} \quad \text{versus} \quad H_1 : \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{27}$$

To test this hypothesis, let us fit least squares coefficients $\widehat{\boldsymbol{\beta}}_{-S}$ and $\widehat{\boldsymbol{\beta}}$ for the partial model as well as the full model. If the partial model fits well, then the residuals $\boldsymbol{y} - \boldsymbol{X}_{*,\text{-}S}\widehat{\boldsymbol{\beta}}_{-\boldsymbol{S}}$ from this model will not be much larger than the residuals $\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ from the full model. To quantify this intuition, let us recall our analysis of variance decomposition from Unit 1:

$$\|\boldsymbol{y} - \boldsymbol{X}_{*,\text{-}S}\widehat{\boldsymbol{\beta}}_{-\boldsymbol{S}}\|^2 = \|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}_{*,\text{-}S}\widehat{\boldsymbol{\beta}}_{-\boldsymbol{S}}\|^2 + \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2. \tag{28}$$

Let's consider the ratio

$$\frac{\|\boldsymbol{y} - \boldsymbol{X}_{*,\text{-}S}\widehat{\boldsymbol{\beta}}_{-\boldsymbol{S}}\|^2 - \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2} = \frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}_{*,\text{-}S}\widehat{\boldsymbol{\beta}}_{-S}\|^2}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2}, \tag{29}$$

which is the relative increase in the residual sum of squares when going from the full model to the partial model. Let us rewrite this ratio in terms of projection matrices. Let $\boldsymbol{H}$ be the projection matrix for the full model, and let $\boldsymbol{H}_{\text{-}S}$ be the projection matrix for the partial model. Note that $\boldsymbol{H} - \boldsymbol{H}_{\text{-}S}$ is the projection matrix onto the $|S|$-dimensional space $C(\boldsymbol{X}) \cap C(\boldsymbol{X}_{\text{-}S})^T$. We have

$$\frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}_{*,\text{-}S}\widehat{\boldsymbol{\beta}}_{\text{-}S}\|^2}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2} = \frac{\|(\boldsymbol{H} - \boldsymbol{H}_{\text{-}S})\boldsymbol{y}\|^2}{\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}\|^2}. \tag{30}$$

Under the null hypothesis, we have

$$\frac{\|(\boldsymbol{H} - \boldsymbol{H}_{\text{-}S})\boldsymbol{y}\|^2}{\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}\|^2} = \frac{\|(\boldsymbol{H} - \boldsymbol{H}_{\text{-}S})\boldsymbol{\epsilon}\|^2}{\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}\|^2}. \tag{31}$$

Since the projection matrices in the numerator and denominator project onto orthogonal subspaces, we have $(\boldsymbol{H} - \boldsymbol{H}_{\text{-}S})\boldsymbol{\epsilon} \perp\!\!\!\perp (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}$, with $\|(\boldsymbol{H} - \boldsymbol{H}_{\text{-}S})\boldsymbol{\epsilon}\|^2 \sim \sigma^2\chi^2_{|S|}$ and $\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\epsilon}\|^2 \sim \sigma^2\chi^2_{n-p}$.

Renormalizing numerator and denominator to have expectation 1 under the null, we arrive at the $F$-statistic

$$F \equiv \frac{(\|\boldsymbol{y} - \boldsymbol{X}_{*,\text{-}S}\widehat{\boldsymbol{\beta}}_{\boldsymbol{-S}}\|^2 - \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2)/|S|}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/(n-p)}. \tag{32}$$

We have derived that under the null hypothesis,

$$F \sim \frac{\chi^2_{|S|}/|S|}{\chi^2_{n-p}/(n-p)}, \quad \text{with numerator and denominator independent.} \tag{33}$$

This distribution is called the $F$-distribution with $|S|$ and $n-p$ degrees of freedom, and denoted $F_{|S|,n-p}$. Denoting by $F_{|S|,n-p}(1-\alpha)$ the $1-\alpha$ quantile of this distribution, we arrive at the $F$-test

$$\phi_F(\boldsymbol{X}, \boldsymbol{y}) \equiv \mathbb{1}(F > F_{|S|,n-p}(1-\alpha)). \tag{34}$$

**Example: Testing for any significant coefficients except the intercept.** Suppose $\boldsymbol{x}_{*,0} = \mathbf{1}_n$ is an intercept term. Then, consider the null hypothesis $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$. In other words, the null hypothesis is the intercept-only model and the alternative hypothesis is the regression model with an intercept and $p-1$ additional predictors. In this case, $S = \{1, \ldots, p-1\}$ and -$S = \{0\}$. The corresponding $F$ statistic is

$$F \equiv \frac{(\|\boldsymbol{y} - \bar{y}\mathbf{1}\|^2 - \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2)/(p-1)}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/(n-p)}, \tag{35}$$

with null distribution $F_{p-1,n-p}$.

**Example: Testing for equality of group means in $C$-groups model.** As a further special case, consider the $C$-groups model from Unit 1. Recall the ANOVA decomposition

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\bar{y}_{c(i)} - \bar{y})^2 + \sum_{i=1}^n (y_i - \bar{y}_{c(i)})^2 = \text{SSB} + \text{SSW}. \tag{36}$$

The $F$-statistic in this case becomes

$$F = \frac{\sum_{i=1}^n (\bar{y}_{c(i)} - \bar{y})^2/(C-1)}{\sum_{i=1}^n (y_i - \bar{y}_{c(i)})^2/(n-C)} = \frac{\text{SSB}/(C-1)}{\text{SSW}/(n-C)}, \tag{37}$$

with null distribution $F_{C-1,n-C}$.

## 2.3 Power

So far we've been focused on finding the null distributions of various test statistics in order to construct tests with Type-I error control. Now let's shift our attention to examining the power of these tests.

**The power of a $t$-test.** Consider the $t$-test of the null hypothesis $H_0 : \beta_j = 0$. Suppose that, in reality, $\beta_j \neq 0$. What is the probability the $t$-test will reject the null hypothesis? To answer this question, recall that $\widehat{\beta}_j \sim N(\beta_j, \sigma^2/s_j^2)$. Therefore,

$$t = \frac{\widehat{\beta}_j}{\text{SE}_j} = \frac{\beta_j}{\text{SE}_j} + \frac{\widehat{\beta}_j - \beta_j}{\text{SE}_j} \dot{\sim} N\left(\frac{\beta_j s_j}{\sigma}, 1\right). \tag{38}$$

Here we have made the approximation $\text{SE}_j \approx \frac{\sigma}{s_j}$, which is pretty good when $n-p$ is large. Therefore, the power of the two-sided $t$-test is

$$\mathbb{E}[\phi_t] = \mathbb{P}[\phi_t = 1] \approx \mathbb{P}[|t| > z_{1-\alpha/2}] \approx \mathbb{P}\left[\left|N\left(\frac{\beta_j s_j}{\sigma}, 1\right)\right| > z_{1-\alpha/2}\right]. \tag{39}$$

Therefore, the quantity $\frac{\beta_j s_j}{\sigma}$ determines the power of the $t$-test. To understand $s_j$ a little better, let's assume that the rows $\boldsymbol{x}_{i*}$ of the model matrix are drawn i.i.d. from some distribution $(x_0, \ldots, x_{p-1})$. Then we have roughly

$$\boldsymbol{x}_{*j}^{\perp} \approx \boldsymbol{x}_{*j} - \mathbb{E}[\boldsymbol{x}_{*j} | \boldsymbol{X}_{*,\text{-}j}], \tag{40}$$

so $x_{ij}^{\perp} \approx x_{ij} - \mathbb{E}[x_{ij} | \boldsymbol{x}_{i,\text{-}j}]$. Hence,

$$s_j^2 \equiv \|\boldsymbol{x}_{*j}^{\perp}\|^2 \approx n\mathbb{E}[(x_j - \mathbb{E}[x_j|\boldsymbol{x}_{\text{-}j}])^2] = n\mathbb{E}[\text{Var}[x_j|\boldsymbol{x}_{\text{-}j}]]. \tag{41}$$

Hence, we can rewrite the alternative distribution (38) as

$$t \dot{\sim} N\left(\frac{\beta_j \cdot \sqrt{n} \cdot \sqrt{\mathbb{E}[\text{Var}[x_j|\boldsymbol{x}_{\text{-}j}]]}}{\sigma}, 1\right). \tag{42}$$

We can see clearly now how the power of the $t$-test varies with the effect size $\beta_j$, the sample size $n$, the degree of collinearity $\mathbb{E}[\text{Var}[x_j|\boldsymbol{x}_{\text{-}j}]]$, and the noise standard deviation $\sigma$.

**The power of an $F$-test.** Now let's turn our attention to computing the power of the $F$-test. We have

$$F = \frac{\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}_{*,\text{-}S}\widehat{\boldsymbol{\beta}}_{\text{-}S}\|^2/|S|}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/|n-p|} = \frac{\|(\boldsymbol{H} - \boldsymbol{H}_{\text{-}S})\boldsymbol{y}\|^2/|S|}{\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}\|^2/|n-p|} \approx \frac{\|(\boldsymbol{H} - \boldsymbol{H}_{\text{-}S})\boldsymbol{y}\|^2/|S|}{\sigma^2}. \tag{43}$$

To calculate the distribution of the numerator, we need to introduce the notion of a non-central chi-squared random variable.

**Definition 2.1.** *For some vector $\boldsymbol{\mu} \in \mathbb{R}^d$, suppose $\boldsymbol{z} \sim N(\boldsymbol{\mu}, \boldsymbol{I}_d)$. Then, we define the distribution of $\|\boldsymbol{z}\|^2$ as the non-central chi-square random variable with $d$ degrees of freedom and noncentrality parameter $\|\boldsymbol{\mu}\|^2$ and denote this distribution by $\chi_d^2(\|\boldsymbol{\mu}\|^2)$.*

It can be shown that if $\boldsymbol{P}$ is a projection matrix and $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, then $\frac{1}{\sigma^2}\|\boldsymbol{P}\boldsymbol{y}\|^2 \sim \chi^2_{\text{tr}(\boldsymbol{P})}(\frac{1}{\sigma^2}\|\boldsymbol{P}\boldsymbol{\mu}\|^2)$. It therefore follows that

$$F \approx \frac{\|(\boldsymbol{H} - \boldsymbol{H}_{\text{-}S})\boldsymbol{y}\|^2/|S|}{\sigma^2} \sim \frac{1}{|S|}\chi^2_{|S|}(\|(\boldsymbol{H} - \boldsymbol{H}_{\text{-}S})\boldsymbol{X}\boldsymbol{\beta}\|^2) = \frac{1}{|S|}\chi^2_{|S|}\left(\frac{1}{\sigma^2}\|\boldsymbol{X}_{*,S}^{\perp}\boldsymbol{\beta}_S\|^2\right). \tag{44}$$

Assuming as before that the rows of $\boldsymbol{X}$ are samples from a joint distribution, we can write

$$\|\boldsymbol{X}_{*,S}^{\perp}\boldsymbol{\beta}_S\|^2 \approx n\boldsymbol{\beta}_S^T\mathbb{E}[\text{Var}[\boldsymbol{x}_S|\boldsymbol{x}_{\text{-}S}]]\boldsymbol{\beta}_S. \tag{45}$$

Therefore,

$$F \dot{\sim} \frac{1}{|S|}\chi^2_{|S|}\left(\frac{n\boldsymbol{\beta}_S^T\mathbb{E}[\text{Var}[\boldsymbol{x}_S|\boldsymbol{x}_{\text{-}S}]]\boldsymbol{\beta}_S}{\sigma^2}\right), \tag{46}$$

which is similar in spirit to equation (42).

**Power when predictors are added to the model.** As we know, the outcome of a regression is a function of the predictors that are used. What happens to the $t$-test $p$-value for $H_0 : \beta_j = 0$ when a predictor is added to the model? To keep things simple, let's consider the

$$\text{true underlying model: } y = \beta_0 x_0 + \beta_1 x_1 + \epsilon. \tag{47}$$

Let's consider the power of testing $H_0 : \beta_0 = 0$ in the regression models

$$\text{model 0: } y = \beta_0 x_0 + \epsilon \quad \text{versus} \quad \text{model 1: } y = \beta_0 x_0 + \beta_1 x_1 + \epsilon. \tag{48}$$

There are four cases based on $\text{cor}[\boldsymbol{x}_{*0}, \boldsymbol{x}_{*1}]$ and the value of $\beta_1$ in the true model:

1. $\text{cor}[\boldsymbol{x}_{*0}, \boldsymbol{x}_{*1}] \neq 0$ and $\beta_1 \neq 0$. In this case, in model 0 we have omitted an important variable that is correlated with $\boldsymbol{x}_{*0}$. Therefore, the meaning of $\beta_0$ differs between model 0 and model 1, so it may not be meaningful to compare the $p$-values arising from these two models.
2. $\text{cor}[\boldsymbol{x}_{*0}, \boldsymbol{x}_{*1}] \neq 0$ and $\beta_1 = 0$. In this case, we are adding a null predictor that is correlated with $x_{*0}$. Recall that the power of the $t$-test hinges on the quantity $\frac{\beta_j \cdot \sqrt{n} \cdot \sqrt{\mathbb{E}[\text{Var}[x_j | \boldsymbol{x}_{\text{-}j}]]}}{\sigma}$. Adding the predictor $x_1$ has the effect of reducing the conditional predictor variance $\mathbb{E}[\text{Var}[x_j | \boldsymbol{x}_{\text{-}j}]]$, therefore reducing the power. This is a case of *predictor competition*.
3. $\text{cor}[\boldsymbol{x}_{*0}, \boldsymbol{x}_{*1}] = 0$ and $\beta_1 \neq 0$. In this case, we are adding a non-null predictor that is orthogonal to $\boldsymbol{x}_{*0}$. While the conditional predictor variance $\mathbb{E}[\text{Var}[x_j | \boldsymbol{x}_{\text{-}j}]]$ remains the same due to orthogonality, the residual variance $\sigma^2$ is reduced when going from model 0 to model 1. Therefore, in this case adding $x_1$ to the model increases the power for testing $H_0 : \beta_0 = 0$. This is a case of *predictor collaboration*.
4. $\text{cor}[\boldsymbol{x}_{*0}, \boldsymbol{x}_{*1}] = 0$ and $\beta_1 = 0$. In this case, we are adding an orthogonal null variable, which does not change the conditional predictor variance or the residual variance, and therefore keeps the power of the test the same.

In conclusion, adding a predictor can either increase or decrease the power of a $t$-test. Similar reasoning can be applied to the $F$-test.

# 3 Confidence and prediction intervals

In addition to hypothesis testing, we often want to construct confidence intervals for the coefficients.

**Confidence interval for a coefficient.** Under $H_0 : \beta_j = 0$, we showed that $\frac{\widehat{\beta}_j}{\widehat{\sigma}/s_j} \sim t_{n-p}$. The same argument shows that for arbitrary $\beta_j$, we have

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}/s_j} \sim t_{n-p}. \tag{49}$$

We can use this relationship to construct a confidence interval for $\beta_j$ as follows:

$$\begin{aligned} 1 - \alpha = \mathbb{P}[|t_{n-p}| \leq t_{n-p}(1-\alpha/2)] &= \mathbb{P}\left[\left|\frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}/s_j}\right| \leq t_{n-p}(1-\alpha/2)\right] \\ &= \mathbb{P}\left[\beta_j \in \left[\widehat{\beta}_j - \frac{\widehat{\sigma}}{s_j} t_{n-p}(1-\alpha/2), \widehat{\beta}_j + \frac{\widehat{\sigma}}{s_j} t_{n-p}(1-\alpha/2)\right]\right] \\ &\equiv \mathbb{P}[\beta_j \in I_j]. \end{aligned} \tag{50}$$

The confidence interval $I_j$ defined above therefore has $1 - \alpha$ coverage.

**Confidence interval for $\mathbb{E}[y|\boldsymbol{x}_0]$.** Suppose now that we have a new predictor vector $\boldsymbol{x}_0 \in \mathbb{R}^p$. The mean of the response for this predictor vector is $\mathbb{E}[y|\boldsymbol{x}_0] = \boldsymbol{x}_0^T\boldsymbol{\beta}$. Plugging in $\boldsymbol{x}_0$ for $\boldsymbol{c}$ in the relation (26), we obtain

$$\frac{\boldsymbol{x}_0^T\widehat{\boldsymbol{\beta}} - \boldsymbol{x}_0^T\boldsymbol{\beta}}{\widehat{\sigma}\sqrt{\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}} \sim t_{n-p}.$$

From this we can derive that

$$\boldsymbol{x}_0^T\widehat{\beta}_j \pm \widehat{\sigma}\sqrt{\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0} \cdot t_{n-p}(1-\alpha/2) \tag{51}$$

is a $1-\alpha$ confidence interval for $\boldsymbol{x}_0^T\boldsymbol{\beta}$.

**Prediction interval for $y|\boldsymbol{x}_0$.** Instead of creating a confidence interval for a point on the regression line, we may want to create a confidence interval for a new draw $y_0$ of $y$ for $\boldsymbol{x} = \boldsymbol{x}_0$, i.e. a *prediction interval*. Note that

$$y_0 - \boldsymbol{x}_0^T\widehat{\beta} = \boldsymbol{x}_0^T\beta + \epsilon_0 - \boldsymbol{x}_0^T\widehat{\beta} = \epsilon_0 + \boldsymbol{x}_0^T(\beta - \widehat{\beta}) \sim N(0, \sigma^2 + \sigma^2\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0). \tag{52}$$

Therefore, we have

$$\frac{y_0 - \boldsymbol{x}_0^T\widehat{\beta}}{\widehat{\sigma}\sqrt{1 + \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0}} \sim t_{n-p}, \tag{53}$$

which leads to the $1-\alpha$ prediction interval

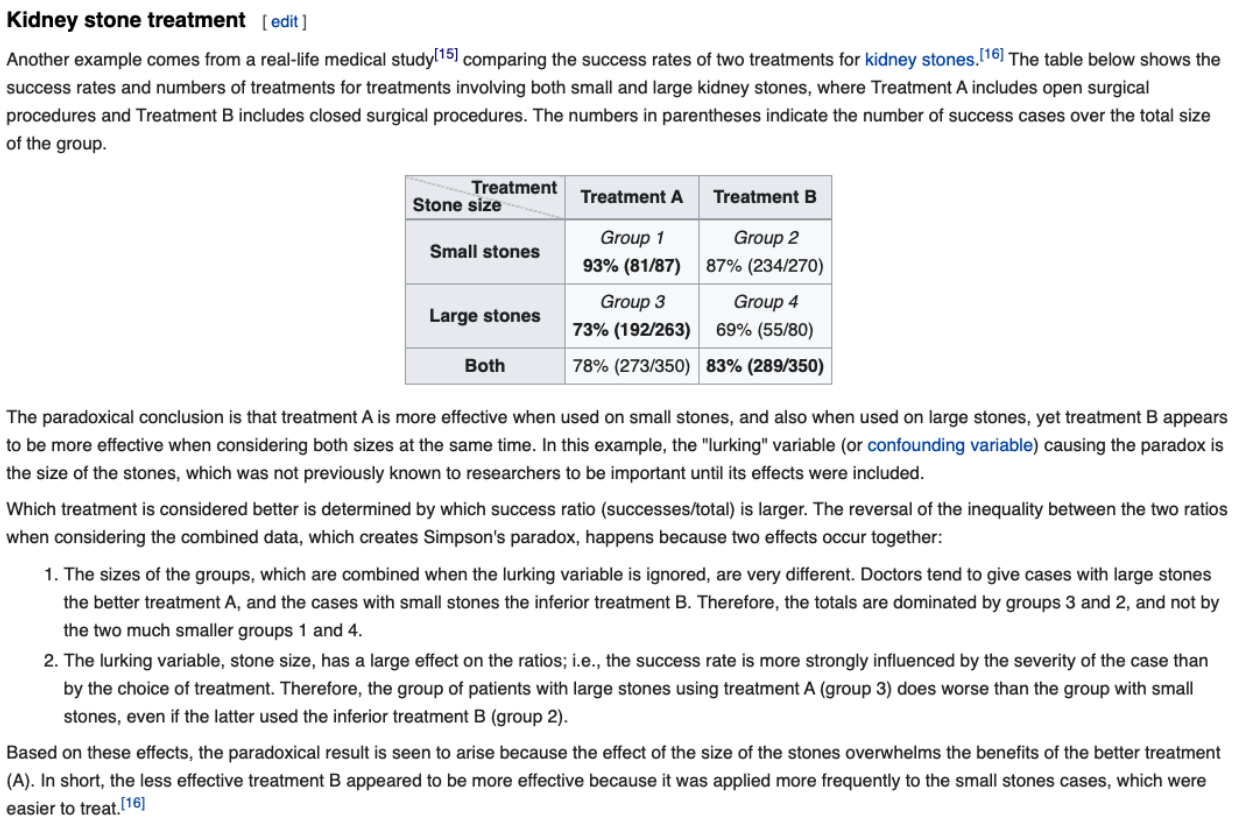
$$\boldsymbol{x}_0^T\widehat{\beta}_j \pm \widehat{\sigma}\sqrt{1 + \boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0} \cdot t_{n-p}(1-\alpha/2). \tag{54}$$

# 4 Practical considerations

**Practical versus statistical significance.** You can have a statistically significant effect that is not practically significant. The hypothesis testing framework is most useful in the case when the signal to noise ratio is relatively small. Otherwise, constructing a confidence interval for the effect size is a more meaningful approach.

**Correlation versus causation, and Simpson's paradox.** We need to be very careful when interpreting linear regression coefficients, which can be sensitive to the choice of other predictors to include. You can get misleading conclusions if you omit important variables from the regression. A special case of this is *Simpson's paradox*, where an important discrete variable is omitted. Consider the example in Figure 1.

**Dealing with correlated predictors.** It depends on the goal. If we're trying to tease apart effects of correlated predictors, then we have no choice but to proceed as usual despite lower power. Otherwise, we can test predictors in groups via the $F$-test to get higher power at the cost of lower "resolution."



**Kidney stone treatment** [ edit ]

Another example comes from a real-life medical study[15] comparing the success rates of two treatments for kidney stones.[16] The table below shows the success rates and numbers of treatments for treatments involving both small and large kidney stones, where Treatment A includes open surgical procedures and Treatment B includes closed surgical procedures. The numbers in parentheses indicate the number of success cases over the total size of the group.

| Stone size \ Treatment | Treatment A | Treatment B |
|---|---|---|
| **Small stones** | *Group 1* **93% (81/87)** | *Group 2* 87% (234/270) |
| **Large stones** | *Group 3* **73% (192/263)** | *Group 4* 69% (55/80) |
| **Both** | 78% (273/350) | **83% (289/350)** |

The paradoxical conclusion is that treatment A is more effective when used on small stones, and also when used on large stones, yet treatment B appears to be more effective when considering both sizes at the same time. In this example, the "lurking" variable (or confounding variable) causing the paradox is the size of the stones, which was not previously known to researchers to be important until its effects were included.

Which treatment is considered better is determined by which success ratio (successes/total) is larger. The reversal of the inequality between the two ratios when considering the combined data, which creates Simpson's paradox, happens because two effects occur together:

1. The sizes of the groups, which are combined when the lurking variable is ignored, are very different. Doctors tend to give cases with large stones the better treatment A, and the cases with small stones the inferior treatment B. Therefore, the totals are dominated by groups 3 and 2, and not by the two much smaller groups 1 and 4.
2. The lurking variable, stone size, has a large effect on the ratios; i.e., the success rate is more strongly influenced by the severity of the case than by the choice of treatment. Therefore, the group of patients with large stones using treatment A (group 3) does worse than the group with small stones, even if the latter used the inferior treatment B (group 2).

Based on these effects, the paradoxical result is seen to arise because the effect of the size of the stones overwhelms the benefits of the better treatment (A). In short, the less effective treatment B appeared to be more effective because it was applied more frequently to the small stones cases, which were easier to treat.[16]

Figure 1: An example of Simpson's paradox (source: Wikipedia).

**Model selection.** We need to ask ourselves: Why do we want to do model selection? It can either be for prediction purposes or for inferential purposes. If it is for prediction purposes, then we can apply cross-validation to select a model and we don't need to think very hard about statistical significance. If it is for inference, then we need to be more careful. There are various classical model selection criteria (e.g. AIC, BIC), but it is not entirely clear what statistical guarantee we are getting for the resulting models. A simpler approach is to apply a $t$-test for each variable in the model, apply a multiple testing correction to the resulting $p$-values, and report the set of significant variables and the associated guarantee. Re-fitting the linear regression after model selection leads us into some dicey inferential territory due to selection bias. This is the subject of ongoing research and the jury is still out on the best way of doing this.

# 5 R demo

Let's put into practice what we've learned in Units 1 and 2.

```
houses_data = read_tsv("../data/Houses.dat")
houses_data
```

```
## # A tibble: 100 x 7
##     case taxes  beds baths   new price  size
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1     1  3104     4     2     0  280.  2048
```

```
##  2     2  1173     2     1     0  146.   912
##  3     3  3076     4     2     0  238.  1654
##  4     4  1608     3     2     0  200   2068
##  5     5  1454     3     3     0  160.  1477
##  6     6  2997     3     2     1  500.  3153
##  7     7  4054     3     2     0  266.  1355
##  8     8  3002     3     2     1  290.  2075
##  9     9  6627     5     4     0  587   3990
## 10    10   320     3     2     0   70   1160
## # ... with 90 more rows

# explore the variables

# should we model beds/baths as categorical or continuous?

# running a regression and interpreting the summary

# hypothesis tests, confidence intervals

# interactions

# confidence bands
```