# Answer TD1 GLM

Xiaopeng ZHANG

October 24, 2025

**Exercise 1.1** (Proof of Cochran's theorem)**.** Let $Z$ be a Gaussian random vector in $\mathbb{R}^n$ with $Z \sim N(\mu, \sigma^2 I_n)$, where $\mu \in \mathbb{R}^n$ and $\sigma > 0$. Let $F_1, \dots, F_m$ be subspaces of dimension $d_i$, orthogonal to each other such that $\mathbb{R}^n = F_1 \oplus \cdots \oplus F_m$. For $i = 1, \dots, m$, let $P_{F_i}$ denote the orthogonal projection matrix onto $F_i$. Prove that

1. The random vectors $P_{F_1}Z, \dots, P_{F_m}Z$ have respective distributions
$$N(P_{F_1}\mu, \sigma^2 P_{F_1}), \dots, N(P_{F_m}\mu, \sigma^2 P_{F_m}) \tag{1}$$
2. The random vectors $P_{F_1}Z, \dots, P_{F_m}Z$ are pairwise independent.
3. The random variables
$$\frac{\|P_{F_1}(Z-\mu)\|^2}{\sigma^2}, \dots, \frac{\|P_{F_m}(Z-\mu)\|^2}{\sigma^2} \tag{2}$$
have respective distributions $\chi^2(d_1), \dots, \chi^2(d_m)$.
4. The random variables
$$\frac{\|P_{F_1}(Z-\mu)\|^2}{\sigma^2}, \dots, \frac{\|P_{F_m}(Z-\mu)\|^2}{\sigma^2} \tag{3}$$
are pairwise independent.

**1.** By the linearity of expectation and the properties of Gaussian distributions, we have:

$$\mathbb{E}[P_{F_i}Z] = P_{F_i}\mathbb{E}[Z] = P_{F_i}\boldsymbol{\mu},$$
$$\mathrm{Cov}(P_{F_i}Z) = P_{F_i}\mathrm{Cov}(Z)P_{F_i}^T = \sigma^2 P_{F_i}.$$

Now we need to show that $P_{F_i}Z$ also has a Gaussian distribution. We can't use the result that state linear transformations of Gaussian vectors are Gaussian, because we are proving this result. To do this, we can use the characteristic function of the Gaussian distribution. The characteristic function of a Gaussian random vector $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ is given by:

$$\phi_X(t) = \exp\left(it^T\boldsymbol{\mu} - \frac{1}{2}t^T\Sigma t\right).$$

For the random vector $P_{F_i}Z$, we have:

$$\begin{aligned}
\phi_{P_{F_i}Z}(t) &= \mathbb{E}\left[e^{it^T P_{F_i} Z}\right] \\
&= \mathbb{E}\left[e^{it^T P_{F_i}(\boldsymbol{\mu}+\sigma W)}\right] \quad \text{(where } W \sim \mathcal{N}(0, \mathbb{I}_n)\text{)} \\
&= e^{it^T P_{F_i}\boldsymbol{\mu}} \mathbb{E}\left[e^{it^T P_{F_i}\sigma W}\right] \\
&= e^{it^T P_{F_i}\boldsymbol{\mu}} \mathbb{E}\left[e^{i(P_{F_i}^T t)\sigma W}\right] \\
&= e^{it^T P_{F_i}\boldsymbol{\mu}} \cdot \exp\left(-\frac{1}{2}\|P_{F_i}^T t\|^2\sigma^2\right) \\
&= \exp\left(it^T P_{F_i}\boldsymbol{\mu} - \frac{1}{2}\|P_{F_i}^T t\|^2\sigma^2\right).
\end{aligned}$$

This shows that $P_{F_i}Z$ has a Gaussian distribution with mean $P_{F_i}\boldsymbol{\mu}$ and covariance $\sigma^2 P_{F_i}$.

**2.** To show that the random vectors $P_{F_i}Z$ and $P_{F_j}Z$ are independent for $i \neq j$, we can start by calculating the covariance, then we can prove that they are jointly Gaussian. Two jointly Gaussian random vectors are independent if and only if their covariance is zero. We have:

$$\begin{aligned}
\operatorname{Cov}(P_{F_i}Z, P_{F_j}Z) &= \mathbb{E}[(P_{F_i}Z - \mathbb{E}[P_{F_i}Z])(P_{F_j}Z - \mathbb{E}[P_{F_j}Z])^T] \\
&= \mathbb{E}[(P_{F_i}(Z - \boldsymbol{\mu}))(P_{F_j}(Z - \boldsymbol{\mu}))^T] \\
&= P_{F_i}\mathbb{E}[(Z - \boldsymbol{\mu})(Z - \boldsymbol{\mu})^T]P_{F_j}^T \\
&= P_{F_i}(\sigma^2 \mathbb{I}_n)P_{F_j}^T \\
&= \sigma^2 P_{F_i}P_{F_j}^T.
\end{aligned}$$

Since $F_i$ and $F_j$ are orthogonal subspaces, we have $P_{F_i}P_{F_j} = 0$.

Therefore, $\operatorname{Cov}(P_{F_i}Z, P_{F_j}Z) = 0$, now it is left to be shown that $P_{F_i}Z$ and $P_{F_j}Z$ are jointly Gaussian.

To show that $P_{F_i}Z$ and $P_{F_j}Z$ are jointly Gaussian, we need to use their characteristic functions. The characteristic function of a Gaussian random vector is given by:

$$\begin{aligned}
\phi_{P_{F_i}Z}(t) &= \exp\left(it^T P_{F_i}\boldsymbol{\mu} - \frac{1}{2}\|P_{F_i}^T t\|^2\sigma^2\right), \\
\phi_{P_{F_j}Z}(t) &= \exp\left(it^T P_{F_j}\boldsymbol{\mu} - \frac{1}{2}\|P_{F_j}^T t\|^2\sigma^2\right).
\end{aligned}$$

Their joint characteristic function is given by:

$$\begin{aligned}
\phi_{P_{F_i}Z, P_{F_j}Z}(t_1, t_2) &= \mathbb{E}\left[e^{it_1^T P_{F_i}Z + it_2^T P_{F_j}Z}\right] \\
&= \mathbb{E}\left[e^{it_1^T P_{F_i}(\boldsymbol{\mu}+\sigma W) + it_2^T P_{F_j}(\boldsymbol{\mu}+\sigma W)}\right] \\
&= e^{it_1^T P_{F_i}\boldsymbol{\mu} + it_2^T P_{F_j}\boldsymbol{\mu}} \mathbb{E}\left[e^{i(t_1^T P_{F_i} + t_2^T P_{F_j})\sigma W}\right] \\
&= e^{it_1^T P_{F_i}\boldsymbol{\mu} + it_2^T P_{F_j}\boldsymbol{\mu}} \cdot \exp\left(-\frac{1}{2}\|(t_1^T P_{F_i} + t_2^T P_{F_j})\sigma W\|^2\right) \\
&= \exp\left(it_1^T P_{F_i}\boldsymbol{\mu} + it_2^T P_{F_j}\boldsymbol{\mu} - \frac{1}{2}\|(t_1^T P_{F_i} + t_2^T P_{F_j})\sigma W\|^2\right).
\end{aligned}$$

We thus have shown that $P_{F_i}Z$ and $P_{F_j}Z$ are jointly Gaussian. Since their covariance is zero, they are independent.

3. We know that $P_{F_i}Z \sim \mathcal{N}(P_{F_i}\boldsymbol{\mu}, \sigma^2 P_{F_i})$. Let $Y_i = P_{F_i}Z - P_{F_i}\boldsymbol{\mu}$. Then, $Y_i \sim \mathcal{N}(0, \sigma^2 P_{F_i})$. The matrix $P_{F_i}$ is a projection matrix onto a subspace of dimension $d_i$, so it has rank $d_i$. Therefore, we can write $P_{F_i} = U_i U_i^T$, where $U_i$ is an $n \times d_i$ matrix whose columns form an orthonormal basis for the subspace $F_i$. The $\chi^2$ distribution with $k$ degrees of freedom can be defined as the distribution of the sum of the squares of $k$ independent standard normal random variables. To show that $\frac{\|Y_i\|^2}{\sigma^2} \sim \chi^2(d_i)$, we can express $Y_i$ in terms of a standard normal vector. Let $W \sim \mathcal{N}(0, \mathbb{I}_n)$ be a standard normal vector in $\mathbb{R}^n$. Then, we can write:

$$\begin{aligned} Y_i &= P_{F_i}Z - P_{F_i}\boldsymbol{\mu} \\ &= P_{F_i}(\boldsymbol{\mu} + \sigma W) - P_{F_i}\boldsymbol{\mu} \\ &= P_{F_i}\sigma W \\ &= \sigma P_{F_i} W. \end{aligned}$$

Therefore, we have:

$$\begin{aligned} \frac{\|Y_i\|^2}{\sigma^2} &= \frac{\sigma^2 \|P_{F_i} W\|^2}{\sigma^2} \\ &= W^T P_{F_i}^T P_{F_i} W \\ &= W^T P_{F_i} W \\ &\sim \chi^2(d_i). \end{aligned}$$

Because $P_{F_i}$ is an orthogonal projection matrix onto a subspace of dimension $d_i$, $P_{F_i}^T = P_{F_i}$.

4. Since we have already shown that the random vectors $P_{F_i}Z$ and $P_{F_j}Z$ are independent for $i \neq j$, it follows that any functions of these independent random vectors are also independent. In particular, the random variables $\frac{\|P_{F_i}(Z-\boldsymbol{\mu})\|^2}{\sigma^2}$ and $\frac{\|P_{F_j}(Z-\boldsymbol{\mu})\|^2}{\sigma^2}$ are functions of the independent random vectors $P_{F_i}Z$ and $P_{F_j}Z$, respectively. Therefore, these random variables are also independent for $i \neq j$.

5. **Final note:** characteristic function of any distribution is:

$$\phi_Z(t) = \mathbb{E}(e^{i\langle t, Z\rangle})$$

**Properties of Fourier Transform:**

**General Formalization:** For a function $f : \mathbb{R} \to \mathbb{C}$, the Fourier transform is defined as:

$$\hat{f}(\xi) = \mathcal{F}[f](\xi) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i \xi t} dt$$

The inverse Fourier transform is:

$$f(t) = \mathcal{F}^{-1}[\hat{f}](t) = \int_{-\infty}^{\infty} \hat{f}(\xi) e^{2\pi i \xi t} d\xi$$

**Key Properties:**

(a) **Linearity:** $\mathcal{F}[af + bg] = a\mathcal{F}[f] + b\mathcal{F}[g]$

(b) **Time shifting:** $\mathcal{F}[f(t-a)](\xi) = e^{-2\pi i a\xi}\hat{f}(\xi)$

(c) **Frequency shifting:** $\mathcal{F}[e^{2\pi i a t}f(t)](\xi) = \hat{f}(\xi - a)$

(d) **Scaling:** $\mathcal{F}[f(at)](\xi) = \frac{1}{|a|}\hat{f}\left(\frac{\xi}{a}\right)$

(e) **Conjugation:** $\mathcal{F}[\overline{f(t)}](\xi) = \overline{\hat{f}(-\xi)}$

(f) **Time reversal:** $\mathcal{F}[f(-t)](\xi) = \hat{f}(-\xi)$

(g) **Differentiation:** $\mathcal{F}[f'(t)](\xi) = 2\pi i\xi\hat{f}(\xi)$

(h) **Integration:** $\mathcal{F}\left[\int_{-\infty}^{t} f(\tau)d\tau\right](\xi) = \frac{\hat{f}(\xi)}{2\pi i\xi}$

(i) **Convolution theorem:** $\mathcal{F}[(f * g)(t)](\xi) = \hat{f}(\xi)\hat{g}(\xi)$

(j) **Parseval's theorem:** $\int_{-\infty}^{\infty} |f(t)|^2 dt = \int_{-\infty}^{\infty} |\hat{f}(\xi)|^2 d\xi$

(k) **Plancherel's theorem:** $\langle f, g\rangle = \langle \hat{f}, \hat{g}\rangle$

**Relation to Fourier transform (answer to the question).** The characteristic function is the Fourier transform of the law (probability measure) of $Z$. If $Z$ has a density $f_Z$ on $\mathbb{R}^n$:

$$\phi_Z(t) = \int_{\mathbb{R}^n} e^{i\,t^\top x}\, f_Z(x)\, dx.$$

Using the "angular-frequency" convention $\mathcal{F}_\omega[f](\omega) = \int f(x)e^{-i\,\omega^\top x}dx$, one has

$$\phi_Z(t) = \mathcal{F}_\omega[f_Z](-t).$$

With the $2\pi$-normalized convention,

$$\hat{f}_Z(\xi) = \int f_Z(x)\, e^{-2\pi i\,\xi^\top x}\, dx, \qquad \Rightarrow \qquad \phi_Z(t) = \hat{f}_Z\big(-\tfrac{t}{2\pi}\big)\,.$$

When a density exists, an inversion formula is

$$f_Z(x) = \frac{1}{(2\pi)^n}\int_{\mathbb{R}^n} e^{-i\,t^\top x}\, \phi_Z(t)\, dt,$$

with the constant adjusted to the chosen Fourier convention.

**Exercise 1.2** (Proof of Proposition 1. of the chapter 1)**.** Let $X$ be the design matrix of size $n \times (p+1)$. We assume $X$ to be full rank ($\text{rank}(X) = p+1$). Let define the following linear model

$$Y = X\beta + \epsilon$$

with $\beta \in \mathbb{R}^{p+1}$. Let

$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^{p+1}} \|Y - X\beta\|^2$$

be the ordinary least square estimator (OLSE).

1. Show that OLSE exists and is unique such that

$$\hat{\beta} = \hat{\beta}(Y) = (X^\top X)^{-1}X^\top Y$$

2. Application for $p = 1$: Let $(x_1, y_1), \dots, (x_n, y_n)$ be $n$ pairs of real numbers. Determine the real $\hat{a}$ and $\hat{b}$ that minimize $\mathrm{RSS}(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2$. Interpret.

**1.** We want to minimize the function $\beta \mapsto \|Y - X\beta\|^2$. As $X$ is full rank, it has a smallest singular value $\sigma_{\min}(X) > 0$. We have

$$\|Y - X\beta\| \geq \|X\beta\| - \|Y\| \geq \sigma_{\min}(X)\|\beta\| - \|Y\|$$

As this shows that the function goes to infinity as $\|\beta\|$ goes to infinity, the minimum is attained at some point $\widehat{\beta}$. The function is differentiable and convex, so the minimum is attained at a point where the gradient is zero.

calculating the gradient, we have

$$\nabla_\beta \|Y - X\beta\|^2 = -2X^\top(Y - X\beta)$$

Setting this to zero, we have

$$X^\top Y = X^\top X\widehat{\beta}$$

As $X$ is full rank, $X^\top X$ is invertible, and we have

$$\widehat{\beta} = (X^\top X)^{-1} X^\top Y$$

**Exercise 1.3.** Let $X$ be a $n \times p$ matrix of rank $p$. Let $\hat{Y}$ be the orthogonal projection on the space $[X]$ generated by the column vectors of $X$ of a vector $Y$ of $\mathbb{R}^n$. Show that $\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = 0$ if one of the column vectors of $X$ is the vector $\mathbf{1}_n = (1, \dots, 1)$. Interpret.

**1.** Let $P_X$ be the orthogonal projection matrix on the space $[X]$. We have $\hat{Y} = P_X Y$.

$$Y - \hat{Y} = Y - P_X Y = (id - P_X)Y \in [X]^\perp$$

Since one of the column vectors of $X$ is $\mathbf{1}_n$, we have $\mathbf{1}_n \in [X]$, therefore:

$$(Y - \hat{Y})^T \mathbf{1}_n = ((id - P_X)Y)^T \mathbf{1}_n = 0$$

which is equivalent to

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i) = 0.$$

**2.** The residuals $Y - \hat{Y}$ sum to zero, which means that the average of the fitted values $\hat{Y}$ is equal to the average of the observed values $Y$. This is a desirable property in regression analysis, as it ensures that the model does not systematically overestimate or underestimate the response variable.

**Exercise 1.4.** We consider the following simple linear regression statistical model: $Y_i = \beta x_i + \varepsilon_i$, for $i = 1, \dots, n$ where the $\varepsilon_i$ are independent, centered, of constant variance. We define two estimators of $\beta \in \mathbb{R}$:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} \quad \text{and} \quad \beta^\star = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i}$$

1. What is the logic of construction of these estimators?

2. Show that they are unbiased estimators of $\beta$.

3. Compare the variances of these two estimators.

**1.** The estimator $\hat{\beta}$ is the ordinary least squares (OLS) estimator, which minimizes the sum of squared residuals between the observed values $Y_i$ and the predicted values $\beta x_i$. The estimator $\beta^\star$ is a simple average-based estimator that uses the total sum of $Y_i$ divided by the total sum of $x_i$.

**2.** To show that both estimators are unbiased, we compute their expected values:

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}\left[(X^TX)^{-1}X^TY\right] = (X^TX)^{-1}X^T\mathbb{E}[Y]$$

Since $Y = X\beta + \varepsilon$ and $\mathbb{E}[\varepsilon] = 0$, we have:

$$\mathbb{E}[Y] = X\beta$$

Thus,

$$\mathbb{E}[\hat{\beta}] = (X^TX)^{-1}X^TX\beta = \beta$$

Similarly, for $\beta^\star$:

$$\mathbb{E}[\beta^\star] = \mathbb{E}\left[\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}\right] = \frac{\sum_{i=1}^n \mathbb{E}[Y_i]}{\sum_{i=1}^n x_i} = \frac{\sum_{i=1}^n \beta x_i}{\sum_{i=1}^n x_i} = \beta$$

Therefore, both estimators are unbiased.

**3.** To compare the variances, we compute:

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Cov}((X^TX)^{-1}X^TY, (X^TX)^{-1}X^TY)$$

$$= \mathbb{E}\left[(X^TX)^{-1}X^TYY^TX(X^TX)^{-1}\right] - \beta\beta^T$$

Recalling that $Y = X\beta + \varepsilon$ thus $\mathbb{E}[YY^T] = \sigma^2 I + X\beta\beta^TX^T$, we have:

$$= (X^TX)^{-1}X^T\mathbb{E}[YY^T]X(X^TX)^{-1} - \beta\beta^T$$

$$= (X^TX)^{-1}X^T(\sigma^2 I + X\beta\beta^TX^T)X(X^TX)^{-1} - \beta\beta^T$$

$$= \sigma^2(X^TX)^{-1} + (X^TX)^{-1}X^TX\beta\beta^TX^TX(X^TX)^{-1} - \beta\beta^T$$

$$= \sigma^2(X^TX)^{-1} + \beta\beta^T - \beta\beta^T = \sigma^2(X^TX)^{-1}$$

For $\beta^\star$:

$$\mathrm{Var}(\beta^\star) = \mathrm{Cov}\left(\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}, \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i}\right)$$

$$= \frac{1}{(\sum_{i=1}^n x_i)^2}\mathrm{Cov}\left(\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i\right) = \frac{1}{(\sum_{i=1}^n x_i)^2}\sum_{i=1}^n \mathrm{Var}(Y_i) = \frac{n\sigma^2}{(\sum_{i=1}^n x_i)^2}$$

To compare the two variances, we note that $\mathrm{Var}(\hat{\beta}) = \sigma^2(X^TX)^{-1}$ and $\mathrm{Var}(\beta^\star) = \frac{n\sigma^2}{(\sum_{i=1}^n x_i)^2}$. The variance of $\hat{\beta}$ is generally smaller than that of $\beta^\star$, especially when the $x_i$ values are not all equal, making $\hat{\beta}$ the more efficient estimator.

**Exercise 1.5** (An important result)**.** We consider the Gaussian linear regression model:

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0_n, \sigma^2 I_n)$$

where $\beta \in \mathbb{R}^r$, $Y \in \mathbb{R}^n$ and $X$ matrix of size $n \times r$ of rank $r$.

1. Recall the matrix closed form of the OLSE and give an unbiased estimator of $\sigma^2 > 0$.
2. Compute the maximum likelihood estimators of $\beta$ and $\sigma^2$.
3. Conclude.

**1.** The matrix closed form of the ordinary least squares estimator (OLSE) is given by: $\hat{\beta} = (X^T X)^{-1} X^T Y$. An unbiased estimator of $\sigma^2$ can be constructed using the residuals from the regression. The residuals are given by $\hat{\epsilon} = Y - X\hat{\beta}$. The unbiased estimator of $\sigma^2$ is then given by:

$$\hat{\sigma}^2 = \frac{1}{n-r} \hat{\epsilon}^T \hat{\epsilon}$$

where $n$ is the number of observations and $r$ is the rank of $X$ (assumed to be full rank $1+p$). Now let's try to find this result.

$$\begin{aligned}
\hat{\epsilon} &= Y - X\hat{\beta} \\
&= Y - X(X^T X)^{-1} X^T Y \\
&= (id - P_X)Y \\
&= P_{X^\perp} Y
\end{aligned}$$

calculating the norm of the residuals:

$$\begin{aligned}
\hat{\epsilon}^T \hat{\epsilon} &= Y^T P_{X^\perp}^T P_{X^\perp} Y \\
&= Y^T P_{X^\perp} Y \\
&= (X\beta + \epsilon)^T P_{X^\perp} (X\beta + \epsilon) \\
&= \epsilon^T P_{X^\perp} \epsilon
\end{aligned}$$

Remind that $\epsilon^T P_{X^\perp} \epsilon = \|P_{X^\perp}(\epsilon)\|^2$, thus by the Cochran's theorem, we have:

$$\frac{\|P_{X^\perp}(\epsilon)\|^2}{\sigma^2} \sim \chi^2(n-r)$$

which implies that:

$$\mathbb{E}[\hat{\epsilon}^T \hat{\epsilon}] = \mathbb{E}[\|P_{X^\perp}(\epsilon)\|^2] = (n-r)\sigma^2$$

Remind that $\mathbb{E}[\chi^2(n-r)] = (n-r)$, thus we have:

$$\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}\left[\frac{\hat{\epsilon}^T \hat{\epsilon}}{n-r}\right] = \frac{1}{n-r} \mathbb{E}[\epsilon^T P_{X^\perp} \epsilon] = \sigma^2$$

which shows that $\hat{\sigma}^2$ is an unbiased estimator of $\sigma^2$.

**2.** The likelihood function of the Gaussian linear regression model is given by:

$$L(\beta, \sigma^2; Y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)\right)$$

To find the maximum likelihood estimators (MLEs) of $\beta$ and $\sigma^2$, we take the logarithm of the likelihood function:

$$\ell(\beta, \sigma^2; Y) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)$$

We then take the partial derivatives of $\ell$ with respect to $\beta$ and $\sigma^2$, set them to zero, and solve for the parameters.

**Exercise 1.6** (Unbiased estimator of $\sigma^2$ in the non-Gaussian model)**.** Consider the following non-Gaussian linear model:

$$Y = X\beta + \epsilon$$

with $\beta \in \mathbb{R}^p$, $X$ of full rank, and the $\epsilon_i$ independent, centered and of variance $\sigma^2$. We pose:

$$\hat{\sigma}^2 = \frac{1}{n-p}\|Y - X\hat{\beta}\|^2$$

We note $\mathrm{Tr}(\cdot)$ the trace of a matrix.

1. Show that $(n-p)\hat{\sigma}^2 = \mathrm{Tr}(\epsilon^\top P_{X^\perp}\epsilon)$
2. Using the fact that $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ for $A$ and $B$ of respective size $(m \times n)$ and $(n \times m)$, show that
$$(n-p)E_\beta[\hat{\sigma}^2] = \sigma^2\mathrm{Tr}(P_{X^\perp})$$
3. Deduce that $E_\beta[\hat{\sigma}^2] = \sigma^2$.

**1.** We have:

$$\begin{aligned}
(n-p)\hat{\sigma}^2 &= \|Y - X\hat{\beta}\|^2 \\
&= (Y - X\hat{\beta})^\top(Y - X\hat{\beta}) \\
&= (Y - X(X^\top X)^{-1}X^\top Y)^\top(Y - X(X^\top X)^{-1}X^\top Y) \\
&= (id - P_X)Y^\top(id - P_X)Y \\
&= Y^\top(id - P_X)^\top(id - P_X)Y \\
&= Y^\top P_{X^\perp}Y \\
&= (X\beta + \epsilon)^\top P_{X^\perp}(X\beta + \epsilon) \\
&= \epsilon^\top P_{X^\perp}\epsilon
\end{aligned}$$

**2.** Using the linearity of the expectation and the property of the trace, we have:

$$\begin{aligned}
(n-p)E_\beta[\hat{\sigma}^2] &= E_\beta[\epsilon^\top P_{X^\perp}\epsilon] \\
&= E_\beta[\mathrm{Tr}(\epsilon^\top P_{X^\perp}\epsilon)] \\
&= E_\beta[\mathrm{Tr}(P_{X^\perp}\epsilon\epsilon^\top)] \\
&= \mathrm{Tr}(P_{X^\perp}E_\beta[\epsilon\epsilon^\top]) \\
&= \mathrm{Tr}(P_{X^\perp}\sigma^2 I_n) \\
&= \sigma^2\mathrm{Tr}(P_{X^\perp})
\end{aligned}$$

**3.** Since $P_{X^\perp}$ is a projection matrix onto a subspace of dimension $n-p$, we have $\text{Tr}(P_{X^\perp}) = n-p$. Therefore:

$$(n-p)E_\beta[\hat{\sigma}^2] = \sigma^2(n-p)$$

Dividing both sides by $n-p$, we obtain:

$$E_\beta[\hat{\sigma}^2] = \sigma^2$$

which shows that $\hat{\sigma}^2$ is an unbiased estimator of $\sigma^2$.

**Exercise 1.7** (Proof of theorem 4 chapter 4)**.** Consider the following Gaussian linear model $Y = X\beta + \epsilon$ where $\beta \in \mathbb{R}^r$, $X$ is a full rank matrix of size $n \times r$ $(n > r)$. Let $C \in M_{q,r}(\mathbb{R})$. We want to test

$$H_0 : C\beta = 0_q \quad \text{versus} \quad H_1 : C\beta \neq 0_q$$

We assume that $\text{rg}(C) = q \leq r$. Therefore, you will note that $\text{rg}(C^\top) = q$ where $C^\top$ is the transpose of $C$.

1. Show that if $Z \sim N_q(0_q, \Sigma)$ then $Z^\top \Sigma^{-1} Z \sim \chi^2_q$.
2. Show that $C(X^\top X)^{-1}C^\top$ is a symmetric and invertible matrix.
3. Recall the ordinary least squares expression $\hat{\beta}$.
4. What is the law of $\hat{\beta}$?
5. Deduce the law of $C\hat{\beta}$ under the hypothesis $H_0$.
6. Deduce that, under $H_0$,

$$R = \frac{(C\hat{\beta})^\top (C(X^\top X)^{-1}C^\top)^{-1}(C\hat{\beta})}{\sigma^2} \sim \chi^2_q$$

7. Conclude that, under $H_0$,

$$F = \frac{\hat{\beta}^\top C^\top (C(X^\top X)^{-1}C^\top)^{-1} C\hat{\beta}}{q\hat{\sigma}^2}$$

   is distributed according to a Fisher distribution with $(q, n-r)$ degrees of freedom. Each step of the reasoning must be carefully justified.

8. Justify and construct a test of $H_0$ against $H_1$ of level $\alpha$.

**1.** Remind that the Cholesky decomposition states that any symmetric positive definite matrix $\Sigma$ can be written as $\Sigma = LL^\top$ where $L$ is a lower triangular matrix with strictly positive diagonal entries.

If $Z \sim N_q(0_q, \Sigma)$, according to Cholesky decomposition, there exists a matrix $L$ such that $\Sigma = LL^\top$. We have $Z = LW$ where $W \sim N_q(0_q, I_q)$. Thus,

$$Z^\top \Sigma^{-1} Z = W^\top L^\top (LL^\top)^{-1} LW = W^\top I_q W = W^\top W \sim \chi^2_q$$

**2.** The matrix $C(X^\top X)^{-1}C^\top$ is symmetric because

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

. Moreover, for any non-zero vector $u \in \mathbb{R}^q$,

$$u^\top C(X^\top X)^{-1}C^\top u = (C^\top u)^\top (X^\top X)^{-1}(C^\top u) > 0$$

$\mathrm{rg}(C) = q$, in other words, $C \in M_{q,r}(\mathbb{R})$ is of full row rank, which implies that $C^\top \in M_{r,q}(\mathbb{R})$ is of full rank. Thus, for any non-zero vector $u \in \mathbb{R}^q$, $C^\top u \neq 0$. On the other hand, since $X$ is of full rank, $X^\top X$ is invertible. Therefore, $C(X^\top X)^{-1}C^\top$ is invertible.

**3.** The ordinary least squares estimator is given by

$$\hat{\beta} = (X^\top X)^{-1}X^\top Y$$

**4.** Since $Y = X\beta + \epsilon$ where $\epsilon \sim N_n(0_n, \sigma^2 I_n)$, we have

$$\hat{\beta} = (X^\top X)^{-1}X^\top(X\beta + \epsilon) = \beta + (X^\top X)^{-1}X^\top \epsilon$$

Thus, $\hat{\beta} \sim N_r(\beta, \sigma^2(X^\top X)^{-1})$.

**5.** Under $H_0$, $C\beta = 0_q$. Therefore,

$$C\hat{\beta} = C\beta + C(X^\top X)^{-1}X^\top\epsilon = C(X^\top X)^{-1}X^\top\epsilon$$

Thus, $C\hat{\beta} \sim N_q(0_q, \sigma^2 C(X^\top X)^{-1}C^\top)$.

**6.** Under $H_0$, we have

$$R = \frac{(C\hat{\beta})^\top (C(X^\top X)^{-1}C^\top)^{-1}(C\hat{\beta})}{\sigma^2}$$

Since $C\hat{\beta} \sim N_q(0_q, \sigma^2 C(X^\top X)^{-1}C^\top)$, according to the result of question 1, we have $R \sim \chi^2_q$.

**7.** We know that $\frac{(n-r)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-r}$ and if $R$ and $\hat{\sigma}^2$ are independent, then

$$F = \frac{\sigma^{-2}R/q}{(\sigma^{-2}(n-r)\hat{\sigma}^2)/(n-r)}$$

is distributed according to a Fisher distribution with $(q, n-r)$ degrees of freedom.

It remains to show that $R$ and $\hat{\sigma}^2$ are independent. Note that

$$\hat{\sigma}^2 = \frac{1}{n-r}(Y - X\hat{\beta})^\top (Y - X\hat{\beta}) = \frac{1}{n-r}\epsilon^\top (I_n - P_X)\epsilon$$

where $P_X = X(X^\top X)^{-1}X^\top$ is the projection matrix. On the other hand,

$$R = \frac{(C\hat{\beta})^\top (C(X^\top X)^{-1}C^\top)^{-1}(C\hat{\beta})}{\sigma^2} = \frac{\epsilon^\top A \epsilon}{\sigma^2}$$

where $A = X(X^\top X)^{-1}C^\top(C(X^\top X)^{-1}C^\top)^{-1}C(X^\top X)^{-1}X^\top$. Thus, to show the independence between $R$ and $\hat{\sigma}^2$, it suffices to show that $A(I_n - P_X) = 0$. Indeed,

$$A(I_n - P_X) = X(X^\top X)^{-1}C^\top(C(X^\top X)^{-1}C^\top)^{-1}C(X^\top X)^{-1}X^\top(I_n - P_X) = 0$$

since $X^\top(I_n - P_X) = 0$. Therefore, $R$ and $\hat{\sigma}^2$ are independent.

**8.** To construct a test of level $\alpha$, we reject $H_0$ when

$$F > F_{q,n-r,1-\alpha}$$

where $F_{q,n-r,1-\alpha}$ is the $(1-\alpha)$-quantile of the Fisher distribution with $(q, n-r)$ degrees of freedom.

**Exercise 1.8** (MCQ)**.** We have observations $(x_i, y_i) \in \mathbb{R}^2$, $\forall i = 1, \dots, n$. We consider the following classical Gaussian linear model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \forall i = 1, \dots, n$$

where $(\beta_0, \beta_1) \in \mathbb{R}^2$ and $\varepsilon_i \sim N(0, \sigma^2)$ are i.i.d.

Let $X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$. Assume $X$ is a full rank matrix and note $\hat{\beta}_0$ and $\hat{\beta}_1$ the least squares estimators of $\beta_0$ and $\beta_1$.
For each of the following questions, give the answer.

1. Are the variables $Y_i$ independent and identically distributed?

   a) Yes b) No c) not always

2. Does the regression line calculated on the observations pass through the mean point $(\bar{x}, \bar{y})$?

   a) Yes b) No c) Only if I am lucky

3. Is it possible to find estimators of $\beta_0$ and $\beta_1$ with smaller variance than the ordinary least squares estimators?

   a) Yes b) No c) Maybe.

4. Are $\hat{\beta}_0$ and $\hat{\beta}_1$ independent?

   a) Yes b) No c) It depends on the matrix $X$

5. If the coefficient of determination $R^2$ calculated on the observations is equal to 1, are the points $(x_i, y_i)_{i=1,\dots,n}$ aligned?

   a) Yes b) No c) Not necessarily

6. Are $\hat{Y}$ and $Y - \hat{Y}$ independent?

   a) Yes b) No c) It depends on the matrix $X$

7. Are $\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n}$ and $Y - \hat{Y}$ independent?

   a) Yes b) No c) It depends on the matrix $X$

8. Is the maximum likelihood estimator of $\sigma^2$ unbiased?

   a) Yes b) No c) We don't know

**Exercise 1.9** (This exercise will be solved without the tools of linear algebra)**.** Let $(x_1, y_1), \dots, (x_n, y_n)$ be $n$ pairs of real numbers. We suppose that $y_i$ are the realization of $Y_i$ whose law is given by the following equation:

$$Y_i = a + bx_i + \varepsilon_i, \quad \varepsilon_i \sim_{i.i.d.} N(0, \sigma^2)$$

1. Determine $\hat{A}$ and $\hat{B}$ the maximum likelihood estimators of $a$ and $b$. Interpret the estimators.

2. Show that these estimators are unbiased.

3. Calculate the variance of the estimators $\mathrm{Var}_\beta(\hat{A})$ and $\mathrm{Var}_\beta(\hat{B})$. How do these variances vary as a function of $\sigma^2$ and the experimental design $x_1, \dots, x_n$?

4. Compute the covariance of $\hat{A}$ and $\hat{B}$. Comment.

5. Let $\hat{Y}_i = \hat{A} + \hat{B}x_i$ and $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$. Show that $\sum_{i=1}^n \hat{\varepsilon}_i = 0$.

6. Show that $\frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-2}$ is an unbiased estimator of $\sigma^2$.

7. Let $x_{n+1}$ be another value. We define $\hat{Y}_{n+1} = \hat{A} + \hat{B}x_{n+1}$. Compute the variance of this prediction.

8. Furthermore, let $Y_{n+1} = A + Bx_{n+1} + \varepsilon_{n+1}$. Calculate the variance of $\hat{\varepsilon}_{n+1} = Y_{n+1} - \hat{Y}_{n+1}$. Compare it to the variance of $\varepsilon_i$ (for $i = 1, \dots, n$).

9. Gauss-Markov Theorem:

   (a) Show that $\hat{B}$ is written as a linear combination of the observations (we will explain the weights).

   (b) Consider $\tilde{B} = \sum_{i=1}^n \lambda_i Y_i$ another unbiased estimator of $B$, written as a linear combination of $Y_i$. Show that $\sum_{i=1}^n \lambda_i = 0$ and $\sum_{i=1}^n \lambda_i x_i = 1$.

   (c) Deduce that $\mathrm{Var}_\beta(\tilde{B}) \geq \mathrm{Var}_\beta(\hat{B})$