Monday, April 8: Gauss-Markov estimation

**Theorem** (Weakly spherical normal projection theorem) Let \( \vec{Y} \) be an \( \mathbb{R}^n \)-valued random vector with distribution \( N(\vec{\mu}, \sigma^2 I) \), let \( \mathcal{W}_1, \ldots, \mathcal{W}_k \) be mutually orthogonal subspaces with dimensions \( n_1, \ldots, n_k \), and let \( \text{Proj}_i : \mathbb{R}^n \to \mathcal{W}_i \) denote orthogonal projection onto \( \mathcal{W}_i \). Then

1. The random vectors \( \text{Proj}_i(\vec{Y}) \) are normally distributed.
2. The random vectors \( \text{Proj}_i(\vec{Y}) \) are independent.
3. The random vectors \( \|\text{Proj}_i(\vec{Y})\|^2 / \sigma^2 \) are independent.
4. The random vectors \( \|\text{Proj}_i(\vec{Y})/\sigma\|^2 \) have a \( \chi^2(n_i, \|\text{Proj}_i(\vec{\mu})\|/\sigma) \) distribution.
Monday, April 8: Gauss-Markov estimation

The first item holds because a linear transformation of a normally distributed random variable is itself normally distributed.

The second holds because if random vectors are uncorrelated and normally distributed, then they are independent.

The third holds because of the second and the lemma.
Monday, April 8: Gauss-Markov estimation

For the fourth, let $\vec{b}_1, \ldots, \vec{b}_{n_i}$ be an orthonormal basis for $W_i$. Also, if $\text{Proj}_i(\vec{\mu}) \neq \vec{0}$, then define

$$\vec{b}_1 = \frac{\text{Proj}_i(\vec{\mu})}{\|\text{Proj}_i(\vec{\mu})\|}.$$ 

By the usual formula for orthogonal projection in terms of an orthonormal basis,

$$\frac{\text{Proj}_i(\vec{Y})}{\sigma} = \frac{\vec{Y} \cdot \vec{b}_1}{\sigma} \vec{b}_1 + \frac{\vec{Y} \cdot \vec{b}_2}{\sigma} \vec{b}_2 + \frac{\vec{Y} \cdot \vec{b}_{n_i}}{\sigma} \vec{b}_{n_i}$$

Also, by definition of multivariate normal, each coefficient is normally distributed.
Monday, April 8: Gauss-Markov estimation

In addition,

\[ E(\vec{Y} \cdot \vec{b}_j) = \vec{\mu} \cdot \vec{b}_j \]

\[ = (\text{Proj}_i(\vec{\mu}) + Q\text{proj}_i(\vec{\mu})) \cdot \vec{b}_j \]

\[ = \text{Proj}_i(\vec{\mu}) \cdot \vec{b}_j. \]

If \( \text{Proj}_i(\vec{\mu}) = \vec{0}, \) then \( \vec{\mu} \in (W_i)^\perp, \) so this equals 0, which also
equals \( \|\text{Proj}_i(\vec{\mu})\|. \)

If \( \text{Proj}_i(\vec{\mu}) \neq \vec{0}, \) then if \( j = 1 \) this equals

\[ \text{Proj}_i(\vec{\mu}) \cdot \left( \frac{\text{Proj}_i(\vec{\mu})}{\|\text{Proj}_i(\vec{\mu})\|} \right) = \|\text{Proj}_i(\vec{\mu})\|. \]

If \( j \neq 1, \) then this equals 0, since \( \text{Proj}_i(\vec{\mu}) \) is a scalar multiple of
\( \vec{b}_1, \) which is orthogonal to \( \vec{b}_j \) for \( j \neq 1. \)
Monday, April 8: Gauss-Markov estimation

In other words,

\[ E\left(\frac{\text{Proj}_i(\vec{Y}) \cdot \vec{b}_j}{\sigma}\right) = \begin{cases} \frac{\|\text{Proj}_i(\vec{\mu})\|}{\sigma} & \text{if } j = 1 \\ 0 & \text{otherwise.} \end{cases} \]
Monday, April 8: Gauss-Markov estimation

Also,

\[
\text{Cov}(\frac{\text{Proj}_i(\vec{Y}) \cdot \vec{b}_j}{\sigma}, \frac{\text{Proj}_i(\vec{Y}) \cdot \vec{b}_k}{\sigma}) = \frac{1}{\sigma^2} (\vec{b}_j \cdot V_{\text{Proj}_i(\vec{Y})}(\vec{b}_k)) \\
= \vec{b}_j \cdot \vec{b}_k \\
= \begin{cases} 
1 & \text{if } j = k \\
0 & \text{otherwise}
\end{cases}
\]

This means that the random variables

\[
\frac{\text{Proj}_i(\vec{Y}) \cdot \vec{b}_j}{\sigma}
\]

each have variance 1, and are uncorrelated.

Since they are uncorrelated and normally distributed, then they are also independent.
Overall, we have shown that the random variables

$$\frac{\text{Proj}_i(\bar{\mathbf{Y}}) \cdot \bar{b}_j}{\sigma}$$

are independent and normally distributed, and have variance 1. When $j = 1$, the expected value is $\|\text{Proj}_i(\bar{\mathbf{u}})\|/\sigma$; when $j \neq 1$, the expected value is 0.

By the definition of a chi-square distribution, this means that

$$\left\| \frac{\text{Proj}_i(\bar{\mathbf{Y}})}{\sigma} \right\|^2 = \left( \frac{\bar{\mathbf{Y}} \cdot \bar{b}_1}{\sigma} \right)^2 + \left( \frac{\bar{\mathbf{Y}} \cdot \bar{b}_2}{\sigma} \right)^2 + \cdots + \left( \frac{\bar{\mathbf{Y}} \cdot \bar{b}_{n_i}}{\sigma} \right)^2$$

has a $\chi^2(n_i, \|\text{Proj}_i(\bar{\mathbf{u}})\|/\sigma)$ distribution, as claimed.
The setup for multiple linear regression:

We have vectors \( \vec{y} \in \mathbb{R}^n \) (observations of the \textit{response} variable) and \( \vec{x}_1, \ldots, \vec{x}_p \in \mathbb{R}^n \) (observations of the explanatory variables), where

\[
\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad \vec{x}_j = \begin{bmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{n,j} \end{bmatrix}.
\]
Wednesday, April 10: Gauss-Markov estimation

For any $\alpha_0, \alpha_1, \ldots, \alpha_p \in \mathbb{R}$, we define

$$\tilde{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \alpha_0 \mathbf{1} + \alpha_1 \mathbf{x}_1 + \cdots + \alpha_p \mathbf{x}_p,$$

and we also define

$$\tilde{\mathbf{e}} = \tilde{\mathbf{y}} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \mathbf{y} - \tilde{\mathbf{y}}.$$
To fit a regression model means to find the unique $\alpha_0, \alpha_1, \ldots, \alpha_p \in \mathbb{R}$ that minimize

$$\| \vec{e} \|^2 = \sum_{i=1}^{n} e_i^2.$$

This is a problem that we have seen before...
Wednesday, April 10: Gauss-Markov estimation

We want to find the unique $\hat{y} \in \mathcal{W} = \text{Span}(\vec{1}, \vec{x}_1, \ldots, \vec{x}_p)$ that minimizes

$$\| \vec{e} \| = \| \vec{y} - \hat{y} \|.$$ 

The unique $\hat{y} \in \mathcal{W}$ that does this is

$$\hat{y} = \text{Proj}_W (\vec{y}).$$
Wednesday, April 10: Gauss-Markov estimation

But we want more: we want to find the coefficients 
\[ \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p \in \mathbb{R}^n \] (the minimizing values of \[ \alpha_0, \alpha_1, \ldots, \alpha_p \])

This means that we want to represent

\[ \tilde{y} = \text{Proj}_W(\vec{y}) \]

relative to the basis \( \{ \vec{1}, \vec{x}_1, \ldots, \vec{x}_p \} \) for \( W \).

Well, that’s assuming that this set is linearly independent. If it isn’t, the data is called multicollinear. In this case, there is still a unique nearest point, but there is not a unique way to represent that point relative to this set.

Exact multicollinearity is not usually an issue, but approximately multicollinearity leads to unstable estimates and other problems.
Friday, April 12: Gauss-Markov estimation

Note that if 
\[
\{\vec{1}, \vec{x}_1, \ldots, \vec{x}_p\}
\]
is a linearly independent set, then it is a basis for \(W\).

This means that \(\vec{y} \in W\) can be written as
\[
\vec{y} = \hat{\beta}_0 \vec{1} + \hat{\beta}_1 \vec{x}_1 + \ldots + \hat{\beta}_p \vec{x}_p
\]
for some unique \(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p \in \mathbb{R}\).

How can we compute these coefficients?
Friday, April 12: Gauss-Markov estimation

First, let’s rewrite the above linear combination in terms of the $X$ matrix

$$X = \begin{bmatrix} \vec{1} & \vec{x}_1 & \cdots & \vec{x}_p \end{bmatrix}.$$ 

A slightly different way of looking at matrix multiplication gives

$$X \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{bmatrix} = a_0 \vec{1} + a_1 \vec{x}_1 + \cdots + a_p \vec{x}_p.$$ 

Note that the linear transformation $X : \mathbb{R}^{p+1} \to \mathbb{R}^n$ that it represents (relative to the standard basis) has

$$\text{Ker}(X) = \{ \vec{0} \}$$

because $\{ \vec{1}, \vec{x}_1, \ldots, \vec{x}_p \}$ is a linearly independent set.
If we define

\[ \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}, \]

then with this notation,

\[ \tilde{y} = \hat{\beta}_0 \tilde{1} + \hat{\beta}_1 \tilde{x}_1 + \ldots + \hat{\beta}_p \tilde{x}_p = X \hat{\beta}. \]

Or, from a linear transformation point of view:

\[ \tilde{y} = X(\hat{\beta}). \]

We would like to solve for \( \hat{\beta} \) (represented relative to the standard basis).
Friday, April 12: Gauss-Markov estimation

The trick here is to recognize that it would be useful to apply the adjoint $X' : \mathbb{R}^n \rightarrow \mathbb{R}^{p+1}$ to both sides of this equation. Once we do that and apply some of the upcoming homework problems, we’ll have it.

We have:

$$X'(\tilde{y}) = X'X(\tilde{\beta}).$$

Now note that $\tilde{y} = \text{Proj}_W(\tilde{y}) \in W$.

By a homework problem,

$$\text{Ker}(X') = \text{Range}(X)^\perp = W^\perp,$$

so $Q\text{Proj}_W(\tilde{y}) \in \text{Ker}(X')$. 
Friday, April 12: Gauss-Markov estimation

This means that

\[ X'(\tilde{y}) = X'(\text{Proj}_W(\tilde{y})) \]
\[ = X'(\text{Proj}_W(\tilde{y}) + Q\text{Proj}_W(\tilde{y})) \]
\[ = X'(\tilde{y}). \]

Putting this back into our previous equation, this implies that

\[ X'(\tilde{y}) = X'X(\hat{\beta}). \]

By another homework problem, since Ker(X) = \{\tilde{0}\} then

\[ X'X : \mathbb{R}^{p+1} \rightarrow \mathbb{R}^{p+1} \]

is invertible, so

\[ \hat{\beta} = (X'X)^{-1}X'(\tilde{y}), \]

which is a formula for the entries in the representation of

\[ \text{Proj}_W(\tilde{y}) \]

relative to the given basis \{\tilde{1}, \tilde{x}_1, \ldots, \tilde{x}_p\} of \mathbb{W}. 
If we want to compute this, we represent everything relative to the standard basis, which gives us

\[
\begin{bmatrix}
\hat{\beta}_0 \\
\hat{\beta}_1 \\
\vdots \\
\hat{\beta}_p \\
\end{bmatrix}
= (X^T X)^{-1} X^T 
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n \\
\end{bmatrix},
\]

where $X^T$ denotes the transpose of the matrix $X$. 