MA50260 Statistical Modelling

Lecture 5: Linear Combinations and Collinearity in Linear Regression

Ilaria Bussoli

February 20, 2024

Summary of last week & new topics

For the linear regression model

$$\mathbf{Y} = \mathbf{X}\underline{\beta} + \underline{\epsilon},$$

with $\dim(\mathbf{Y}) = \dim(\underline{\epsilon}) = n \times 1$, $\dim(\mathbf{X}) = n \times p$ and $\dim(\underline{\beta}) = p \times 1$, the distribution of the least square estimator is

$$\hat{\underline{\beta}}(\mathbf{Y}) \sim \mathrm{MVN}_p \left(\underline{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right).$$

We used this result to

- Test $H_0 : \beta_j = b$ against $H_1 : \beta_j \neq b$
- Derive $(1-\alpha) \times 100\%$ confidence interval for β_j ($j = 1, \ldots, p$),

$$\hat{\beta}_j \pm t_{n-p}(1-\alpha/2) \times \sqrt{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{j,j}}.$$

Summary of last week & new topics

For the linear regression model

$$\mathbf{Y} = \mathbf{X}\underline{\beta} + \underline{\epsilon},$$

with $\dim(\mathbf{Y}) = \dim(\underline{\epsilon}) = n \times 1$, $\dim(\mathbf{X}) = n \times p$ and $\dim(\underline{\beta}) = p \times 1$, the distribution of the least square estimator is

$$\underline{\hat{\beta}}(\mathbf{Y}) \sim \mathrm{MVN}_p\left(\underline{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right).$$

We used this result to

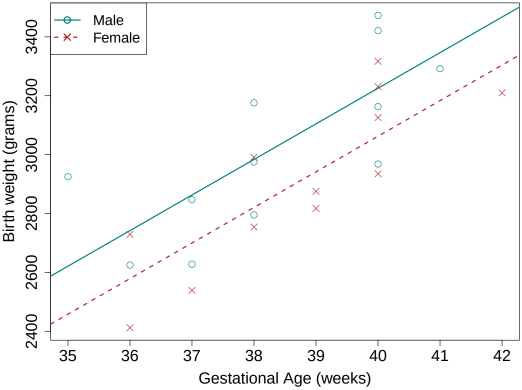
- Test $H_0 : \beta_j = b$ against $H_1 : \beta_j \neq b$
- Derive $(1 - \alpha) \times 100\%$ confidence interval for β_j $(j = 1, \ldots, p)$,

$$\hat{\beta}_j \pm t_{n-p}(1 - \alpha/2) \times \sqrt{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{j,j}}.$$

What if we want to test more than one regression coefficient? Or combinations of them?

Linear Combinations of Regression Coefficients (I)

Recall the birth weight example with separate intercepts for males (β_1) and females (β_2).



Is there a difference between males and females?

Linear Combinations of Regression Coefficients (II)

We wish to test

$$H_0 : \beta_1 - \beta_2 = 0 \qquad \text{vs.} \qquad H_1 : \beta_1 - \beta_2 \neq 0.$$

We require the distribution of $\mathbf{a}^T \underline{\hat{\beta}}(\mathbf{Y})$, with $\mathbf{a} = (1, -1, 0)^T$:

$$\mathbb{E}\left[\mathbf{a}^T \underline{\hat{\beta}}(\mathbf{Y})\right] = \mathbf{a}^T \; \mathbb{E}\left[\underline{\hat{\beta}}(\mathbf{Y})\right] = \mathbf{a}^T \underline{\beta};$$

and

$$\begin{aligned} \mathrm{Var}\left[\mathbf{a}^T \underline{\hat{\beta}}(\mathbf{Y})\right] &= \mathbf{a}^T \mathrm{Var}\left[\underline{\hat{\beta}}(\mathbf{Y})\right] \mathbf{a} \\ &= \sigma^2 \mathbf{a}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{a}. \end{aligned}$$

Since $\underline{\hat{\beta}}(\mathbf{Y})$ follows a multivariate normal distribution,

$$\mathbf{a}^T \underline{\hat{\beta}}(\mathbf{Y}) \sim \mathrm{Normal}(\mathbf{a}^T \underline{\beta}\,, \, \sigma^2 \mathbf{a}^T \left(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{a}\right).$$

Hypothesis Testing

To test

$$H_0 : \mathbf{a}^T \underline{\beta} = b \qquad \text{vs.} \qquad H_1 : \mathbf{a}^T \underline{\beta} \neq b,$$

we consider the observed test statistic

$$t = \frac{\mathbf{a}^T \hat{\underline{\beta}} - b}{\sqrt{\hat{\sigma}^2 \mathbf{a}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{a}}}$$

from the test statistic

$$T(\mathbf{Y}) = \frac{\mathbf{a}^T \hat{\underline{\beta}}(\mathbf{Y}) - b}{\sqrt{\hat{\sigma}^2(\mathbf{Y})\ \mathbf{a}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{a}}}$$

which, under the null hypothesis, is t_{n-p} distributed.

Example - Birth Weights

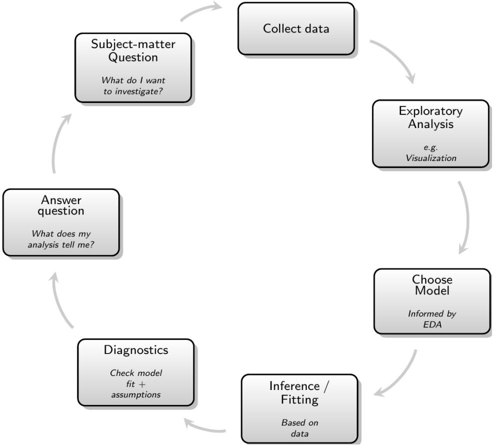
For the birth weight example with $\mathbf{a}^T = (1, -1, 0)$ and $b = 0$,

$$
\begin{aligned}
t &= \frac{\mathbf{a}^T \underline{\hat{\beta}} - b}{\sqrt{\hat{\sigma}^2 \mathbf{a}^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}}} \\
&= \frac{(\hat{\beta}_1 - \hat{\beta}_2) - 0}{\sqrt{\hat{\sigma}^2 \times \left[(\mathbf{X}^T\mathbf{X})^{-1}_{1,1} + (\mathbf{X}^T\mathbf{X})^{-1}_{2,2} - (\mathbf{X}^T\mathbf{X})^{-1}_{2,1} - (\mathbf{X}^T\mathbf{X})^{-1}_{1,2}\right]}} \\
&= \frac{163}{31370 \times 0.169} = 2.24.
\end{aligned}
$$

Compare to 97.5% quantile of a t_{24-3}-distribution, which is 2.08.

Since $2.24 > 2.08$, we reject H_0 at the 5% level and conclude that there is a significant difference between males and females.

What have we achieved so far?



Collinearity

Collinearity refers to linear dependence (strong correlation) between explanatory variables.

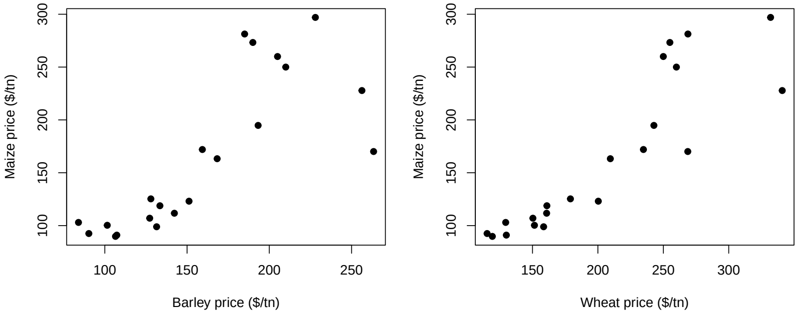
We say that two explanatory variables x_i and x_j are

- **Orthogonal** if $\mathrm{Corr}(x_i, x_j)$ is close to zero;
- **Collinear** if $\mathrm{Corr}(x_i, x_j)$ is close to one.

Collinearity can make results difficult to interpret and may cause numerical issues when deriving $(\mathbf{X}^T\mathbf{X})^{-1}$.

Example - Cereal Prices (I)

We investigate global commodity price forecasts for maize, barley and wheat from 1995–2015.



Example - Cereal Prices (II)

We relate annual maize prices, Y_i, to annual prices of barley, $x_{i,1}$, and wheat, $x_{i,2}$.

Consider the three models:

$$\begin{aligned}
\text{Model 1} &\rightarrow \mathbb{E}(Y_i) = \beta_1 + \beta_2 x_{i,1}, \\
\text{Model 2} &\rightarrow \mathbb{E}(Y_i) = \beta_1 + \beta_2 x_{i,2}, \\
\text{Model 3} &\rightarrow \mathbb{E}(Y_i) = \beta_1 + \beta_2 x_{i,1} + \beta_3 x_{i,2}.
\end{aligned}$$

The estimated regression coefficients are

$$\begin{aligned}
\text{Model 1} &\rightarrow \hat{\beta}_1 = -9.48,\ \hat{\beta}_2 = 1.09 \\
\text{Model 2} &\rightarrow \hat{\beta}_1 = -30.83,\ \hat{\beta}_2 = 0.95 \\
\text{Model 3} &\rightarrow \hat{\beta}_1 = -25.66,\ \hat{\beta}_2 = -0.51,\ \hat{\beta}_3 = 1.32
\end{aligned}$$

Example - Cereal Prices (III)

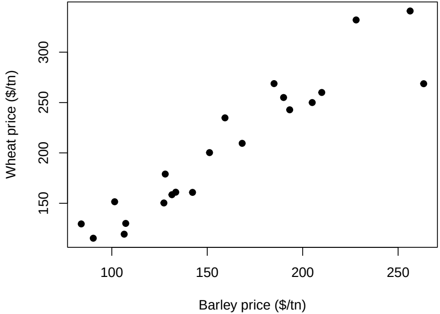
Let's derive the 95% confidence levels:

$$
\begin{aligned}
\text{Model } 1 \quad &\rightarrow \quad \hat{\beta}_2 \pm t_{21-2}(0.975) \times \mathrm{se}\left(\hat{\beta}_2\right) = (0.684, 1.487). \\
\text{Model } 2 \quad &\rightarrow \quad \hat{\beta}_2 \pm t_{21-2}(0.975) \times \mathrm{se}\left(\hat{\beta}_2\right) = (0.717, 1.18). \\
\text{Model } 3 \quad &\rightarrow \quad \hat{\beta}_2 \pm t_{21-3}(0.975) \times \mathrm{se}(\hat{\beta}_2) = (-1.366, 0.347), \\
& \quad\; \hat{\beta}_3 \pm t_{21-3}(0.975) \times \mathrm{se}(\hat{\beta}_3) = (0.655, 1.99).
\end{aligned}
$$

So we would conclude that β_2 is not significant in Model 3.

Example - Cereal Prices (IV)

Let's investigate the relationship between barley (β_2) and wheat (β_3) prices:



Corr(Barley, Wheat) = 0.94

We see that the two explanatory variables are strongly dependent.

$\Rightarrow$ We should fit either Model 1 or Model 2.

$\Rightarrow$ We have to check the dependence amongst explanatory variables.