TMA4267 Linear Statistical Models
Part 2: Linear regresssion
Solutions to recommended exercise 3 - V2017

February 7, 2017

Problem 1: Simple linear regression

See the file `RecEx31.R`, accessible from the course website, for more details.

a) The `forbes` data set consists of $n = 17$ observations, and we are fitting a simple linear regression model which includes an intercept, so the design matrix $\boldsymbol{X}$ has 17 rows and 2 columns. Since the two columns are linearly independent, the rank of $\boldsymbol{X}$ is 2. We can confirm this using `R`, by calling the `rankMatrix()` function from the `Matrix` package.

```
library(Matrix) # Package needs to be installed, if it isn't already
rankMatrix(X)
```

b) The boiling point increases with increasing pressure. When the pressure is around 0.7 bar, the boiling point is approximately 90° C, and when the pressure is around 1 bar, the boiling point is approximately 100° C. The relationship between pressure and boiling point does look linear.

c) The estimated regression parameters are

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \approx \begin{bmatrix} 68.5 \\ 31.2 \end{bmatrix},$$

and the fitted regression function has the expression

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where x is pressure, and $\hat{y}$ is the expected boiling point in that pressure. From this expression it is clear that $\hat{\beta}_0$ is the boiling point the model would predict for $x = 0$ bar, i.e. in a vacuum, and $\hat{\beta}_1$ is the estimated rate of change of the boiling point with pressure; If the pressure increases by 1 bar, we expect the boiling point to increase by $\hat{\beta}_1$ degrees Celsius.

d) Two things are apparent when looking at the plot of the raw residuals $\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ versus the pressure:

- The plot has an inverted U-shape. Going from left to right, the residuals start off negative, then gradually become positive, before decreasing and becoming negative again.

- With the exception of one data point (observation number 12), the residuals all lie inside the interval $(-0.3, 0.3)$, and there is a roughly even split between positive and negative residuals.

e)
- Is the linear model appropriate?
  Yes. Although there may be a systematic relationship between the residuals and the covariate, it is clear that the relationship between the response and covariate is predominantly linear, and that the fitted model does a satisfactory job of representing that relationship.
- Are the error variances homoscedastic?
  Maybe. It's a little hard to tell, because of the remaining trend in the residuals. It looks like the residual variance may vary somewhat with pressure, but even if this is the case, there is only weak heteroscedasticity, and no serious violation of this model assumption.
- Are the error terms uncorrelated?
  No. The plot of residuals vs. pressure plainly shows that observations with similar pressures, have similar residuals. This is systematic variation not accounted for by the fitted model. If we want to improve the fit, we might try to use the logarithm of the pressure as a covariate, instead of the pressure.
- Does an additive model seem appropriate?
  Yes. The class of additive models subsumes linear models, and we have seen that a linear model is able to achieve a reasonably good fit in this case. There is nothing about this data set, or the analysis done so far, suggesting that a non-additive model would be preferable.

Normality of residuals can be ascertained by making a normal quantile-quantile plot of $\hat{\boldsymbol{\epsilon}}$, and comparing it to a straight line.

```
qqnorm(eps)
```

f) The parameter estimates, fitted response values and raw residuals all match the results from **b)** and **c)**. The residual plots (residuals vs. pressure and normal quantile-quantile) look the same as the ones in **d)** and **e)**.

Problem 2: Happiness

a) The regression coefficient for `work` is 0.4761. Assume that we look at two individuals that have scored the same values for `sex`, `love` and `money`. Further assume that one of the individuals has reported `work` to have value 1 (seeking other employment) and the other has 2 (inbetween seeking other employment and OK). Then, on average, we would expect that the happiness for the last individual is 0.4761 higher than for the first individual. Keeping the other variables fized, the effect of `work` on happiness is that happiness increases on average with 0.4761 units for every one unit increase in the `work` variable.

We can perform a t-test to see if `work` is significant in the full model (given that the other variables are present). Test statistic:

$$t = \frac{\hat{\beta}_4}{\widehat{\mathrm{Var}(\hat{\beta}_4)}} = \frac{0.4761}{0.1994} = 2.39$$

which under the null hypothesis that $\beta_4 = 0$ (vs. the two-sided alternative that $\beta_4 \neq 0$) is referred to a t-distribution with $n - k - 1 = 34$ degrees of freedom ($n = 39$ is the number of observations and $k = 4$ is the number of explanatory variables, and 1 for the intercept). We use significance level 0.05 and the critical value at level 0.025 (since two-sided test) in the t_{34} distribution is approximately 2.03 (found for 35 in the table), which means that we reject the null hypothesis.

To test if the regression is significant, that is, not all coefficients are zero, we look at the $F = 20.83$ value. Using a significance level of 0.05 this is referred to a critical value for the Fisher distribution with 4 and 34 degrees of freedom: approximately 2.64 (with 35 from the table). This means that the regression is highly significant.

Residual plots: We see no clear trend in the plot of residuals vs. fitted values, which is good. The quantile-quantile plot shows no clear deviation from normality, but at least one outlier is identified.

b) In a multiple regression the least squares estimates for the regression coefficients are found by solving a set of equations. When the explanatory variables are not orthogonally selected the value for one explanatory variable will influence the estimate of the regression coefficient for the other. In a design of experiments where explanatory variables are chosen so that they are independent of each other (orthogonal columns) the normal equations will become uncoupled and the regression coefficient estimate for the explanatory variables will not influence each other.

Problem 3: Results on $\hat{\boldsymbol{\beta}}$ and SSE in multiple linear regression

(Exam K2014, Problem 4) Define the matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

a) $\boldsymbol{H}$ is a orthogonal projection matrix, since it is symmetric and idempotent, $\boldsymbol{H}^T = \boldsymbol{H}$ and $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$. For a symmetric and idempotent matrix the rank is equal to the trace. The rank of $\boldsymbol{H}$ is p. See proof below.

$$
\begin{aligned}
\boldsymbol{H} &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \\
\boldsymbol{H}^T &= (\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)^T = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H} \\
\boldsymbol{H}^2 &= (\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H} \\
\mathrm{tr}(\boldsymbol{H}) &= \mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) = \mathrm{tr}(\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}) = \mathrm{tr}(\boldsymbol{I}_{p\times p}) = p
\end{aligned}
$$

Graphically: The vector $\boldsymbol{H}\boldsymbol{Y}$ is a projection of the vector $\boldsymbol{Y}$ onto the space spanned by the columns of $\boldsymbol{X}$.

The matrix $\boldsymbol{I} - \boldsymbol{H}$ is also symmetric and idempotent, and thus a symmetric projection matrix. The rank of $\boldsymbol{I} - \boldsymbol{H}$ is $n - p$. See proof below.

$$
\begin{aligned}
(\boldsymbol{I} - \boldsymbol{H})^T &= \boldsymbol{I} - \boldsymbol{H}^T = \boldsymbol{I} - \boldsymbol{H} \\
(\boldsymbol{I} - \boldsymbol{H})^2 &= (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{I} - 2\boldsymbol{H} + \boldsymbol{H}^2 = \boldsymbol{I} - \boldsymbol{H} \\
\mathrm{tr}(\boldsymbol{I} - \boldsymbol{H}) &= \mathrm{tr}(\boldsymbol{I}_{n\times n}) - \mathrm{tr}(\boldsymbol{H}) = n - p
\end{aligned}
$$

Graphically: The vector $(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}$ is a projection of the vector $\boldsymbol{Y}$ onto the space orthogonal to the space spanned by the columns of $\boldsymbol{X}$.

b) Let $\text{SSE} = \boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}$. Derive the distribution of SSE.

One of the key theorems of this course (point 2 of Theorem B.8 on p. 651 of Fahrmeir et al.) states that if $\boldsymbol{D}$ is a symmetric and idempotent matrix with rank r and $\boldsymbol{Z} \sim N_n(\boldsymbol{0}, \boldsymbol{I})$, then

$$\boldsymbol{Z}^T\boldsymbol{D}\boldsymbol{Z} \sim \chi^2_r,$$

which means that if $\boldsymbol{Z} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, then

$$\frac{\boldsymbol{Z}^T\boldsymbol{D}\boldsymbol{Z}}{\sigma^2} \sim \chi^2_r.$$

In this case we have $\boldsymbol{D} = (\boldsymbol{I}-\boldsymbol{H})$ symmetric and idempotent with rank $n-p$, and $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. To use the theorem we need to look at $\boldsymbol{Y}^* = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$.

$$\begin{aligned}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}^* &= (\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) \\ &= (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} - (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta} = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} - (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta}) = \\ &= (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} - (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}\end{aligned}$$

since $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{X}$. Projecting $\boldsymbol{X}$ onto the space spanned by the columns of $\boldsymbol{X}$ gives $\boldsymbol{X}$.

Thus, we have shown that $\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y} = \boldsymbol{Y}^{*T}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}^*$, and we may use the theorem to conclude that $\text{SSE}/\sigma^2 \sim \chi^2_{n-p}$.

The mean of a χ^2-distributed variable equals the number of degrees of freedom, so

$$\begin{aligned}\text{E}\left(\frac{\text{SSE}}{\sigma^2}\right) &= n-p \\ \text{E}\left(\text{SSE}\right) &= (n-p)\sigma^2 \\ \text{E}\left(\frac{\text{SSE}}{n-p}\right) &= \sigma^2\end{aligned}$$

Thus, $\hat{\sigma}^2 = \frac{\text{SSE}}{n-p}$ will be an unbiased estimator for σ^2. The variance of a χ^2-distributed random variable is twice the number of degrees of freedom, so the variance of $\hat{\sigma}^2$ is

$$\begin{aligned}\text{Var}(\hat{\sigma}^2) &= \text{Var}\left(\frac{\text{SSE}}{n-p}\right) = \frac{1}{(n-p)^2}\text{Var}(\text{SSE}) \\ &= \frac{1}{(n-p)^2}\text{Var}\left(\sigma^2\frac{\text{SSE}}{\sigma^2}\right) = \frac{\sigma^4}{(n-p)^2}\text{Var}\left(\frac{\text{SSE}}{\sigma^2}\right) \\ &= \frac{1}{(n-p)^2}2(n-p)\sigma^4 = \frac{2\sigma^4}{n-p}.\end{aligned}$$

c) We consider the two matrices

$$\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \quad \text{and} \quad \boldsymbol{B} = \boldsymbol{I}-\boldsymbol{H} = \boldsymbol{I}-\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$$

and the corresponding transformed random variables $\boldsymbol{Z}_1 = \boldsymbol{AY}$ and $\boldsymbol{Z}_2 = \boldsymbol{BY}$. Here $\boldsymbol{A}$ is a $p \times n$ matrix (since $\boldsymbol{X}$ is $n \times p$), and $\boldsymbol{B}$ is the same dimension as $\boldsymbol{H}$, that is, $n \times n$, and is symmetric and idempotent (found previously).

Since $\boldsymbol{Y} \sim N_n(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$ then $\boldsymbol{AY}$ and $\boldsymbol{BY}$ are independent random variables if $\sigma^2\boldsymbol{AB}^T = \boldsymbol{0}$.

$$\begin{aligned}\boldsymbol{AB}^T &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = 0\end{aligned}$$

We have proven that $\boldsymbol{Z}_1 = \boldsymbol{AY}$ and $\boldsymbol{Z}_2 = \boldsymbol{BY}$ are independent random variables. Then it follows that $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2^T\boldsymbol{Z}_2$ are also independent random variables. Since $\boldsymbol{Z}_1 = \hat{\boldsymbol{\beta}}$ and $\boldsymbol{Z}_2^T\boldsymbol{Z}_2 = \boldsymbol{Y}^T\boldsymbol{BY} = \text{SSE}$, we have proven that $\hat{\boldsymbol{\beta}}$ and SSE are independent random variables.

Application to multiple linear regression: The independence of $\hat{\boldsymbol{\beta}}$ and SSE is used in the construction of a t-test for hypotheses about $\boldsymbol{\beta}$.

Problem 4: Weighted linear regression

(Exam V2014 Problem 4)

a) Let $(\lambda_i, \boldsymbol{e}_i)$, $i = 1, ..., p$ be the eigenvalues and eigenvectors of $\boldsymbol{V}$. Let $\boldsymbol{P}$ be the $(p \times p)$ matrix of eigenvectors,

$$\boldsymbol{P} = [\boldsymbol{e}_1\boldsymbol{e}_2 \cdots \boldsymbol{e}_p]$$

and $\boldsymbol{\Lambda}$ be a diagonal matrix with the eigenvalues $\lambda_1, \lambda_1, ..., \lambda_p$ on the diagonal. Then $\boldsymbol{V}^{-\frac{1}{2}}$ is defined as

$$\boldsymbol{V}^{-\frac{1}{2}} \quad = \quad \boldsymbol{P}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{P}^T$$

Observe that $\boldsymbol{V}^{-\frac{1}{2}}$ is symmetric, and that $\boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{V}^{-\frac{1}{2}} = \boldsymbol{V}^{-1}$.

$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{Y} &= \boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{X\beta} + \boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{\varepsilon} \\ \boldsymbol{Y}^* &= \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*\end{aligned}$$

where $\boldsymbol{\varepsilon}^* \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. To see this calculate $\text{Cov}(\boldsymbol{\varepsilon}^*) = \boldsymbol{V}^{-\frac{1}{2}}\,\text{Cov}(\boldsymbol{\varepsilon})\boldsymbol{V}^{-\frac{1}{2}} = \boldsymbol{V}^{-\frac{1}{2}}\sigma^2\boldsymbol{V}\boldsymbol{V}^{-\frac{1}{2}} = \sigma^2\boldsymbol{I}$.

We have now the ordinary least squares problem in the new quantities $\boldsymbol{Y}^*$, $\boldsymbol{X}^*$ and $\boldsymbol{\varepsilon}^*$, and know that the least squares solution is

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= (\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{Y}^* \\ &= (\boldsymbol{X}^T\boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{Y} \\ &= (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}\end{aligned}$$

Mean:

$$\begin{aligned}\mathrm{E}(\tilde{\boldsymbol{\beta}}) &= \mathrm{E}((\boldsymbol{X}^{*T}\boldsymbol{X}^{*})^{-1}\boldsymbol{X}^{*T}\boldsymbol{Y}^{*}) = (\boldsymbol{X}^{*T}\boldsymbol{X}^{*})^{-1}\boldsymbol{X}^{*T}\,\mathrm{E}(\boldsymbol{Y}^{*}) \\ &= (\boldsymbol{X}^{*T}\boldsymbol{X}^{*})^{-1}\boldsymbol{X}^{*T}\boldsymbol{X}^{*}\boldsymbol{\beta} = \boldsymbol{\beta}\end{aligned}$$

since $\mathrm{E}(\boldsymbol{Y}^{*}) = \boldsymbol{X}^{*}\boldsymbol{\beta}$.

The ordinary least square estimator $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ is unbiased in this model since the mean of $\boldsymbol{Y}$ doesn't depend on $\boldsymbol{V}$.

$$\begin{aligned}\mathrm{E}(\hat{\beta}) &= \mathrm{E}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathrm{E}(\boldsymbol{Y}) \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}\end{aligned}$$

If we just look at unbiasedness it may appear that the two estimators are equally good. However, since $\tilde{\boldsymbol{\beta}}$ is the least squares estimator (from looking at transformed quantities) we may conclude using the Gauss-Markov Theorem (p 181 in Fahrmeir et al (2013)) that $\tilde{\boldsymbol{\beta}}$ has the minimum variance in each component among all the unbiased estimators, BLUE. If we had calculated the covariance matrices of $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$, we should see this. Thus, $\tilde{\boldsymbol{\beta}}$ should be preferred. Another issue is the fact that $\boldsymbol{V}$ seldom is known, and need to be estimated. The concept of BLUE is handled in detail in our Statistical Inference course.

b) Find the expected value and covariance matrix of $\hat{\boldsymbol{\alpha}_1}$ under the true model

$$\begin{aligned}\mathrm{E}(\hat{\boldsymbol{\alpha}_1}) &= \mathrm{E}((\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}) \\ &= (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\,\mathrm{E}(\boldsymbol{Y}) == (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T(\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2) \\ &= \boldsymbol{\beta}_1 + (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{\beta}_2\end{aligned}$$

Thus, $\hat{\boldsymbol{\alpha}}_1$ is a biased estimator for $\boldsymbol{\beta}_1$.

$$\begin{aligned}\mathrm{Cov}(\hat{\boldsymbol{\alpha}_1}) &= \mathrm{Cov}((\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}) \\ &= (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\,\mathrm{Cov}(\boldsymbol{Y})\boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1} \\ &= (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\sigma^2\boldsymbol{I}\boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1} \\ &= \sigma^2(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\end{aligned}$$

Observe, $\mathrm{Cov}(\hat{\boldsymbol{\alpha}_1})$ is not dependent on $\boldsymbol{\beta}_2$.

We see that the bias term for $\hat{\boldsymbol{\alpha}}_1$ is $(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{\beta}_2$. When is the bias term equal to zero?

When $\boldsymbol{\beta}_2 = \boldsymbol{0}$ there is no bias (but that is not so exciting). The bias is also zero when $\boldsymbol{X}_1^T\boldsymbol{X}_2 = \boldsymbol{0}$. This will happen if the two matrices are orthogonal. In Part 4: DOE this will be useful to know.