Contents

1 Problem setup

True data The true, unobserved data is a standard regression setup, with the exception that one line of the data is reserved as 'validation data' with the rest called 'training data' — as is standard in predictive contexts. The response variable y is a noised linear combination of the covriates in X:

$$\tilde{X}_A = \begin{pmatrix} x_{11} & x_{12} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} \quad \text{and} \quad y_A = X_A\beta + \epsilon_A \quad \text{with} \quad \epsilon_A \sim \mathcal{N}(0, \sigma^2)$$

$$\tilde{X_V} = \begin{pmatrix} x_1^V & x_2^V \end{pmatrix} \quad \text{and} \quad y_V = X_V\beta + \epsilon_V \quad \text{with} \quad \epsilon_V \sim \mathcal{N}(0, \sigma^2)$$

The end goal is to learn an estimator on the training set that minimizes the expected loss on the validation set:

$$L(y_V, \hat{y}_V) = (y_V - \hat{y_V})^2$$

This is a well known problem with a simple solution. The regression estimator can be expressed as:

$$\tilde{\beta} = (\tilde{X}_A^T\tilde{X}_A)^{-1}\tilde{X}_A^T y_A$$

with and estimated response

$$\tilde{y}_V = \tilde{X}_V\tilde{\beta}$$

But there is one caveat: the data is actually not fully observed.

Observed data What we actually have access to is slightly different: one observation is missing in the training set, as is one of the entries in the testing set. We observe the full y^A, but the covariate matrices we actually have access to are:

$$X^A = \begin{pmatrix} ? & x_{12} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} \quad \text{and} \quad X^V = \begin{pmatrix} ? & x_2^V \end{pmatrix}$$

Imputed data and regression To perform a linear regression similarly to what we did previously, we first have to fill in the blanks: we impute the missing data by replacing them with chosen values ϕ and ψ (which we choose using the value of what we observe).

$$\hat{X}^A = \begin{pmatrix} \phi & x_{12} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix} \quad \text{and} \quad \hat{X}^V = \begin{pmatrix} \psi & x_2^V \end{pmatrix}$$

This in turn allows us to perform the regression just like before:

$$\hat{\beta} = (\hat{X}_A^T \hat{X}_A)^{-1} \hat{X}_A^T y_A \quad \text{and} \quad \hat{y}_V(\phi, \psi) = \hat{X}_V \hat{\beta}$$

This means that what we really want to minimize is the following risk function:

$$R(\phi, \psi) = \mathbb{E}[(y_V - \hat{y}_V(\phi, \psi))^2 | X_A, X_V]$$

where the only thing we choose is our decision rule for ϕ and ψ.

Distribution hypotheses Lastly, for this last expression to have any meaning, we need to make some assumption on the distribution of X.

We make a very simple hypotheses of a multivariate normal distribution for the covariates — the parameters are assumed to be known, in real life they would be estimates.

$$\tilde{X}_A \sim \mathcal{N}(\mu, \Sigma) \quad \tilde{X}_V \sim \mathcal{N}(\mu, \Sigma)$$

With these, we know the distribution of the missing observations x_{11} and X_1^V conditional of the observed ones and we can try to find the *phi* and *psi* values that give us the best expected loss.

2 Resolution

2.1 Loss

To be able to estimate the expected loss, we break it up into several components

$$\begin{aligned}
L(y_V, \hat{y}_V) &= (y_V - \hat{y_V})^2 \\
&= (\tilde{X}_V\beta + \epsilon_V - \hat{X}_V\hat{\beta})^2 \\
&= (\tilde{X}_V(\beta - \tilde{\beta}) + \tilde{X}_V(\tilde{\beta} - \hat{\beta}) + (\tilde{X}_V - \hat{X}_V)\hat{\beta} + \epsilon_V)^2 \\
&= (\tilde{X}_V(\beta - \tilde{\beta}))^2 && (1) \\
&\quad + (\tilde{X}_V(\tilde{\beta} - \hat{\beta}))^2 && (2) \\
&\quad + ((\tilde{X}_V - \hat{X}_V)\hat{\beta})^2 && (3) \\
&\quad + \tilde{X}_V(\beta - \tilde{\beta})\tilde{X}_V(\tilde{\beta} - \hat{\beta}) && (4) \\
&\quad + \tilde{X}_V(\beta - \tilde{\beta})(\tilde{X}_V - \hat{X}_V)\hat{\beta} && (5) \\
&\quad + \tilde{X}_V(\tilde{\beta} - \hat{\beta})(\tilde{X}_V - \hat{X}_V)\hat{\beta} && (6) \\
&+ \epsilon_V^2 \\
&+ \epsilon_V K
\end{aligned}$$

Where K is some term that will not matter (because when we take the expectation it will be zero). The risk we want to minimize is the expectation of this loss.

2.2 When the training set is actually known

Imputation The first thing we can easily do is to study the situation where the only missing data is in the validation set. In that case, $\tilde{X}_A = \hat{X}_A$ and so $\tilde{\beta} = \hat{\beta}$, and all we have to choose is ψ. In the previously computed loss, it means that terms (2), (4) and (6) are zero.

Furthermore, Cochran's theorem ensures that $(\beta - \tilde{\beta})$ and $\tilde{\beta}$ are independent so term (5) can be factorized and will have zero expectation (since $(\beta - \tilde{\beta})$ has zero expectation).

Term (1) depends only on the true values of X_V, independent of *psi*, so the choice of ψ is not impacted by this term.

This leaves us with only term (3), with expectation:

$$\begin{aligned}
\mathbb{E}[((\tilde{X}_V - \hat{X}_V)\tilde{\beta})^2|x_2^V, X_A] &= \mathbb{E}[(x_1^V - \psi)^2\tilde{\beta}_1^2|x_2^V, X_A] \\
&= \mathbb{E}[\tilde{\beta}_1^2|x_2^V, X_A](\mathbb{E}[(x_1^V)^2|x_2^V, X_A] - 2\psi\mathbb{E}[x_1^V|x_2^V, X_A] + \psi^2)
\end{aligned}$$

Once we are there, we can differentiate this expression to easily derive the optimal expression for ψ: $\hat{\psi} = \mathbb{E}[x_1^V|x_2^V]$.

Incidentally, this does not use any assumption on the distribution of X: this would be true for any joint distribution we choose for the covariates.

Expected loss First note:

$$\begin{aligned}\eta = \beta - \tilde{\beta} &= \beta - (\tilde{X}_A^T \tilde{X}_A)^{-1} \tilde{X}_A^T y_A \\ &= (\tilde{X}_A^T \tilde{X}_A)^{-1} \tilde{X}_A^T (\tilde{X}_A \beta + \epsilon_A) \\ &= (\tilde{X}_A^T \tilde{X}_A)^{-1} \tilde{X}_A^T \epsilon_A\end{aligned}$$

That is, the difference between the estimated and real parameter is distributed following some centred normal distribution. Let us denote its covariance matrix by $S = \sigma^2 (\tilde{X}_A^T \tilde{X}_A)^{-1}$.

Now, term (1) can be expressed as :

$$\begin{aligned}\mathbb{E}[(\tilde{X}_V \eta)^2 | x_2^V] &= \mathbb{E}[(x_1^V \eta_1 + x_2^V \eta_2)^2 | x_2^V] \\ &= S_{11} \mathbb{E}[(x_1^V)^2 | x_2^V] + 2 x_2^V S_{12} \mathbb{E}[x_1^V | x_2^V] + S_{22} (x_2^V)^2\end{aligned}$$

Term (3) can be expressed as:

$$\begin{aligned}\mathbb{E}[((\tilde{X}_V - \hat{X}_V)\hat{\beta})^2 | x_2^V] &= \mathbb{E}[((x_1^V - \hat{\psi})\tilde{\beta}_1)^2 | x_2^V] \\ &= \mathbb{E}[((x_1^V - \hat{\psi})(\beta_1 + \eta_1))^2 | x_2^V] \\ &= \beta_1^2 \mathrm{Var}[x_1^V | x_2^V] + S_{11}^2 \mathrm{Var}[x_1^V | x_2^V]\end{aligned}$$

Finaly, adding the variance of the noise ϵ_V, we see that the expected loss is :

$$S_{11} \mathbb{E}[(x_1^V)^2 | x_2^V] + 2 x_2^V S_{12} \mathbb{E}[x_1^V | x_2^V] + S_{22} (x_2^V)^2 + \beta_1^2 \mathrm{Var}[x_1^V | x_2^V] + S_{11}^2 \mathrm{Var}[x_1^V | x_2^V] + \sigma^2$$

The terms from (1) would be more or less the same if the test data were fully observed. They represent the impact on the prediction of the error made when estimating β.

On the other hand, those from (3) are both positive and unique to the incomplete case. They do not depend on the test data at all. The first term reflects how an error in the imputation of X_V is amplified by the regression coefficient when predicting y. The second one shows how errors in the estimation of β and of X_V combine when predicting y.

Most importantly, the least squares estimator is a strongly consistent one, which implies that the only error terms that matter with large n are those that remain when η is set to zero:

$$\beta_1^2 \mathrm{Var}[x_1^V | x_2^V] + \sigma^2$$

The missing data in the validation set adds a term that does not vanish for large n.

When there are more than two covariates An important point is that in a context with more than two covariates, the conditional variance of an observation increases when the number of unobserved variables increases: this means that when the dimension increases, not only do new error terms appear, but the existing ones also increase.

For X_V with $p > 2$ covariates, we denote X_V^{miss} and X_V^{obs} the missing and observed values in X_V. We can now write (3) again:

$$\begin{aligned}\mathbb{E}[((\tilde{X}_V - \hat{X}_V)\beta)^2|X_V^{\text{obs}}] &= \mathbb{E}[(\sum_{\substack{i=1\\ i\in X_V^{\text{miss}}}}^{p} (x_i^V - \hat{x}_i^V)\beta_i)^2|X_V^{\text{obs}}]\\ &= \mathbb{E}[\sum_{\substack{i=1\\ i\in X_V^{\text{miss}}}}^{p} \sum_{\substack{j=1\\ j\in X_V^{\text{miss}}}}^{p} (x_i^V - \hat{x}_i^V)(x_j^V - \hat{x}_j^V)\beta_i\beta_j|X_V^{\text{obs}}]\end{aligned}$$

If, as previously, we choose $\hat{x}_i^V = \mathbb{E}[x_i^V|X_V^{\text{obs}}]$, we end up with:

$$\mathbb{E}[((\tilde{X}_V - \hat{X}_V)\beta)^2|X_V^{\text{obs}}] = \sum_{\substack{i,j=1\\ i,j\in X_V^{\text{miss}}}}^{p} \beta_i\beta_j Cov(x_i^V, x_j^V|X_V^{\text{obs}})$$

2.3 When n is large and the validation data is known

We can suppose that n is large and that we know $\tilde{X_V}$. In that case, we can take the approximation that $\beta = \tilde{\beta}$, that is, the only error in the estimation of β comes from the missing data in the training set. Then, the only error term that is nonzero is (2):

$$(\tilde{X}_V(\beta - \tilde{\beta}))^2$$