Bias and Variance Analysis of Multiple Linear Regression

Phil Schniter

September 11, 2020

1 Multiple Linear Regression

Given some observed features $\boldsymbol{x} \triangleq [x_1, \dots, x_d]^\top \in \mathbb{R}^d$, our goal is to predict a target $y \in \mathbb{R}$. To help us with this task, we are given access to n pairs of feature/label examples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \triangleq [x_{i1}, \dots, x_{id}]^\top$. We will use a linear predictor of the form

$$\widehat{y} = \widehat{\beta}_0 + \sum_{j=1}^d \widehat{\beta}_j x_j, \tag{1}$$

where the coefficients $\{\widehat{\beta}_j\}_{j=0}^d$ are trained to minimize the residual sum-of-squares (RSS) on the training data:

$$\text{RSS} \triangleq \sum_{i=1}^n (\widehat{y}_i - y_i)^2 \quad \text{with} \quad \widehat{y}_i = \widehat{\beta}_0 + \sum_{j=1}^d \widehat{\beta}_j x_{ij}. \tag{2}$$

Defining the quantities

$$\boldsymbol{y} \triangleq \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} \triangleq \begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \dots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{bmatrix}, \quad \boldsymbol{A} \triangleq \begin{bmatrix} \mathbf{1} & \boldsymbol{X} \end{bmatrix}, \quad \widehat{\boldsymbol{\beta}} \triangleq \begin{bmatrix} \widehat{\beta}_0 \\ \vdots \\ \widehat{\beta}_d \end{bmatrix}, \tag{3}$$

and assuming that $\boldsymbol{A}^\top \boldsymbol{A}$ is invertible, the RSS-minimizing coefficients take the form

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{y}, \tag{4}$$

as derived in the lecture. This is known as the "least-squares (LS) fit."

Note that, if $n < d + 1$, then $\boldsymbol{A}$ will be a wide matrix, in which case $\boldsymbol{A}^\top \boldsymbol{A}$ will *not* be invertible. In this case, the LS solution is not unique, i.e., there are an infinite number of $\widehat{\boldsymbol{\beta}}$ that yield the same training target predictions $\widehat{\boldsymbol{y}} = \boldsymbol{A}\widehat{\boldsymbol{\beta}}$. This suggests that we should always strive to have at least as many training examples, n, as adjustable model parameters, $d + 1$. In the sequel, we shall see that we actually want $n \gg d + 1$ for accurate predictions outside of the training data.

2 Error Analysis

We are interested in knowing how close the prediction $\widehat{y}$ will be to the true target y. Although we know the training RSS, its value may not be a good indicator of the squared "test" error

$(\widehat{y}-y)^2$. For example, the training RSS will equal zero whenever $\boldsymbol{A}$ is square and invertible, but that will not guarantee zero test error.

To understand the nature of the prediction error $\widehat{y}-y$, we analyze its mean and variance, which give us two views of what happens "on average." First, we establish conditions under which the mean is zero, and later we analyze the variance in the zero-mean case. We also connect our analysis to the terms "bias" and "variance" as typically used in machine learning.

For our analyses, we will assume that the target variable y is generated from some true weights $\{\beta_j\}$ according to

$$y = \beta_0 + \sum_{j=1}^{d_{\mathsf{true}}} \beta_j x_j + \epsilon, \tag{5}$$

where ϵ is random "noise" that is generated independently of the other model quantities, x_j and β_j. We model ϵ as zero mean, else there would an ambiguity between the mean of ϵ and the model quantity β_0. We assume that the variance of ϵ is finite and denote it by σ^2. No other assumptions on ϵ are needed for our analysis.

We will furthermore assume that the training data is consistent with the test data, in that

$$y_i = \beta_0 + \sum_{j=1}^{d_{\mathsf{true}}} \beta_j x_{ij} + \epsilon_i, \quad i = 1, \dots, n, \tag{6}$$

where random ϵ_i have mean zero and variance σ^2, and are drawn independently of β_j, the training features x_{ij}, the test quantities ϵ and $\boldsymbol{x}$. We will eventually model the training features as random (and independent of the test quantities $\boldsymbol{x}$ and ϵ), but we will postpone the details until we need them.

3 Bias

The first question we ask is: Under which conditions is the *average* value of the prediction error $\widehat{y}-y$ non-zero? This situation is known as "undermodeling." To be precise, the average value of the prediction error, also known as the "bias," is defined as

$$\text{bias}_{\widehat{y}}(\boldsymbol{x}) \triangleq \mathrm{E}\{\widehat{y} - y \mid \boldsymbol{x}\}. \tag{7}$$

In (7), the expectation is over the test noise ϵ and the trained weights $\widehat{\beta}_j$ (which themselves are affected by random effects in the training data) but not over the test features $\boldsymbol{x}$ nor true weights β_j, which we consider as fixed.

In the case that $d < d_{\mathsf{true}}$, we can write

$$\begin{aligned} \text{bias}_{\widehat{y}}(\boldsymbol{x}) &= \mathrm{E}\left\{ (\widehat{\beta}_0 - \beta_0) + \sum_{j=1}^{d} (\widehat{\beta}_j - \beta_j) x_j - \sum_{j=d+1}^{d_{\mathsf{true}}} \beta_j x_j - \epsilon \;\middle|\; \boldsymbol{x} \right\} && (8) \\ &= \mathrm{E}\{\widehat{\beta}_0\} - \beta_0 + \sum_{j=1}^{d} \big(\mathrm{E}\{\widehat{\beta}_j\} - \beta_j \big) x_j - \sum_{j=d+1}^{d_{\mathsf{true}}} \beta_j x_j, && (9) \end{aligned}$$

since $\mathrm{E}\{\epsilon|\boldsymbol{x}\} = \mathrm{E}\{\epsilon\} = 0$ and since the trained weights $\widehat{\beta}_j$ are independent of the test features $\boldsymbol{x}$. Equation (9) shows that, even if the trained weights are correct on average (i.e., $\mathrm{E}\{\widehat{\beta}_j\} = \beta_j$ for

$j = 1 \ldots d$), the bias on $\widehat{y}$ will be nonzero due to the last term in (9). In summary, if $d < d_{\textsf{true}}$, then $\widehat{y}$ is "biased" and undermodeling occurs.

Next we examine the case that $d \geq d_{\textsf{true}}$. In this case, we can write (5) as

$$y = \beta_0 + \sum_{j=1}^{d} \beta_j x_j + \epsilon \tag{10}$$

by declaring that the true $\beta_j = 0$ for $j = d_{\textsf{true}}+1, \ldots, d$. To analyze the bias on $\widehat{y}$, it is convenient to first write the LS weights $\widehat{\boldsymbol{\beta}}$ in terms of the true weights $\boldsymbol{\beta} \triangleq [\beta_0, \ldots, \beta_d]^\top$. Using (4) and (5), we get

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{y} && (11)\\
&= (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top (\boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{\epsilon}) && (12)\\
&= \boldsymbol{\beta} + (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{\epsilon}, && (13)
\end{aligned}$$

where $\boldsymbol{\epsilon} \triangleq [\epsilon_1, \ldots, \epsilon_n]^\top$. For the last step, we used the fact that $(\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}$ due to the assumed invertibility of $\boldsymbol{A}^\top \boldsymbol{A}$. Thus, using $\boldsymbol{a}^\top \triangleq \begin{bmatrix} 1 \ \boldsymbol{x}^\top \end{bmatrix}$, we can write the prediction error as

$$\begin{aligned}
\widehat{y} - y &= \boldsymbol{a}^\top \widehat{\boldsymbol{\beta}} - (\boldsymbol{a}^\top \boldsymbol{\beta} + \epsilon) && (14)\\
&= \boldsymbol{a}^\top (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{\epsilon} - \epsilon. && (15)
\end{aligned}$$

Finally, the bias is simply the mean of the prediction error, which is

$$\begin{aligned}
\text{bias}_{\widehat{y}}(\boldsymbol{x}) &= \mathrm{E}\{\boldsymbol{a}^\top (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{\epsilon} - \epsilon \mid \boldsymbol{x}\} && (16)\\
&= \begin{bmatrix} 1 \ \boldsymbol{x}^\top \end{bmatrix} \mathrm{E}\{(\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top\} \, \mathrm{E}\{\boldsymbol{\epsilon} | \boldsymbol{x}\} - \mathrm{E}\{\epsilon | \boldsymbol{x}\} && (17)\\
&= 0, && (18)
\end{aligned}$$

since $\mathrm{E}\{\boldsymbol{\epsilon}|\boldsymbol{x}\} = \mathbf{0}$ and $\mathrm{E}\{\epsilon|\boldsymbol{x}\} = 0$. In summary, whenever $d \geq d_{\textsf{true}}$, the prediction $\widehat{y}$ is unbiased (i.e., the prediction error is zero on average for any fixed $\boldsymbol{x}$).

4 Prediction-Error Variance

From now on, we will focus on the case that $d \geq d_{\textsf{true}}$. We know that, when $d \geq d_{\textsf{true}}$, the prediction error $\widehat{y} - y$ is zero *on average* given any test features $\boldsymbol{x}$ (which determine $\boldsymbol{a}$). Here, the averaging is over the test noise and randomness in the weights $\widehat{\boldsymbol{\beta}}$. But the prediction error is *not* expected to be zero for any specific realization of those random quantities.

A natural question would then be: What is the *variance* of the prediction error $\widehat{y} - y$ given $\boldsymbol{x}$? Recall that

$$\text{var}\{\widehat{y} - y \mid \boldsymbol{x}\} = \mathrm{E}\{(\widehat{y} - y)^2 \mid \boldsymbol{x}\} - \mathrm{E}\{\widehat{y} - y \mid \boldsymbol{x}\}^2 = \mathrm{E}\{(\widehat{y} - y)^2 \mid \boldsymbol{x}\} \tag{19}$$

since $\mathrm{E}\{\widehat{y} - y \mid \boldsymbol{x}\} = 0$. So, the variance of the prediction error is equal to the mean-squared prediction error due to the unbiased nature of the prediction. For the mean-squared prediction error, we have

$$
\begin{aligned}
&\mathrm{E}\left\{(\widehat{y}-y)^2 \;\middle|\; \boldsymbol{x}\right\} \\
&= \mathrm{E}\left\{(\widehat{y}-y)(\widehat{y}-y)^\top \;\middle|\; \boldsymbol{x}\right\} && (20)\\
&= \mathrm{E}\left\{[\boldsymbol{a}^\top(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{\epsilon}-\epsilon][\boldsymbol{a}^\top(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{\epsilon}-\epsilon]^\top \;\middle|\; \boldsymbol{x}\right\} && (21)\\
&= \mathrm{E}\left\{\boldsymbol{a}^\top(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{a} \;\middle|\; \boldsymbol{x}\right\} - \mathrm{E}\left\{\boldsymbol{a}^\top(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{\epsilon} \;\middle|\; \boldsymbol{x}\right\}\mathrm{E}\{\epsilon \mid \boldsymbol{x}\}\\
&\quad - \mathrm{E}\{\epsilon \mid \boldsymbol{x}\}\,\mathrm{E}\left\{\boldsymbol{\epsilon}^\top\boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{a} \;\middle|\; \boldsymbol{x}\right\} + \mathrm{E}\{\epsilon^2 \mid \boldsymbol{x}\} && (22)\\
&= \mathrm{E}\left\{\boldsymbol{a}^\top(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top\boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{a} \;\middle|\; \boldsymbol{x}\right\} + \sigma^2, && (23)
\end{aligned}
$$

where we leveraged the fact that ϵ is independent of the other random variables to decouple the expectation, and then we used the fact that $\mathrm{E}\{\epsilon \mid \boldsymbol{x}\} = \mathrm{E}\{\epsilon\} = 0$ to drop two terms. Next, we separate the expectation over $\boldsymbol{\epsilon}$ from that over the other quantities and move it to the inside, giving

$$
\begin{aligned}
\mathrm{E}\left\{(\widehat{y}-y)^2 \;\middle|\; \boldsymbol{x}\right\} &= \mathrm{E}\left\{\boldsymbol{a}^\top(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\,\mathrm{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top|\boldsymbol{x},\boldsymbol{A}\}\boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{a} \;\middle|\; \boldsymbol{x}\right\} + \sigma^2 && (24)\\
&= \sigma^2\,\mathrm{E}\left\{\boldsymbol{a}^\top(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{a} \;\middle|\; \boldsymbol{x}\right\} + \sigma^2, && (25)
\end{aligned}
$$

where the second step invoked the independence of $\{\epsilon_i\}$ to write $\mathrm{E}\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \mid \boldsymbol{x}, \boldsymbol{A}\} = \sigma^2\boldsymbol{I}$. Exploiting the invertibility of $\boldsymbol{A}^\top\boldsymbol{A}$, we find

$$
\begin{aligned}
\mathrm{E}\left\{(\widehat{y}-y)^2 \;\middle|\; \boldsymbol{x}\right\} &= \sigma^2\,\mathrm{E}\left\{\boldsymbol{a}^\top(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{a} \;\middle|\; \boldsymbol{x}\right\} + \sigma^2 && (26)\\
&= \sigma^2\,\mathrm{E}\left\{\,\mathrm{tr}\left[\boldsymbol{a}\boldsymbol{a}^\top(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\right] \;\middle|\; \boldsymbol{x}\right\} + \sigma^2, && (27)
\end{aligned}
$$

using the fact that $\boldsymbol{b}^\top\boldsymbol{C}\boldsymbol{b} = \mathrm{tr}[\boldsymbol{b}^\top\boldsymbol{C}\boldsymbol{b}] = \mathrm{tr}[\boldsymbol{b}\boldsymbol{b}^\top\boldsymbol{C}]$ for any vector $\boldsymbol{b}$ and matrix $\boldsymbol{C}$ of compatible dimensions. Since the trace is a linear operation, we can move the expectation inside to write

$$
\mathrm{E}\left\{(\widehat{y}-y)^2 \;\middle|\; \boldsymbol{x}\right\} = \sigma^2\,\mathrm{tr}\left[\boldsymbol{a}\boldsymbol{a}^\top\,\mathrm{E}\{(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\}\right] + \sigma^2, \tag{28}
$$

where we also used the fact that $\boldsymbol{a}$ is a deterministic function of $\boldsymbol{x}$, and that the training features in $\boldsymbol{A}$ are generated independently of the test features $\boldsymbol{x}$.

To proceed further, we need to make some additional assumptions. First, we will assume that the training feature vectors $\{\boldsymbol{x}_i\}$ are independent and identically distributed, and also independent of the quantities $\boldsymbol{x}$, ϵ, and $\{\epsilon_i\}$. Furthermore, we will assume that the correlation matrix $\boldsymbol{R}_{aa} \triangleq \mathrm{E}\{\boldsymbol{a}_i\boldsymbol{a}_i^\top\}$ of the augmented training feature vectors $\boldsymbol{a}_i \triangleq \begin{bmatrix}1 & \boldsymbol{x}_i^\top\end{bmatrix}$ is invertible. This prevents, for example, any features from being perfectly correlated or deterministic. In addition, we shall assume that n is large, so that

$$
\frac{1}{n}\boldsymbol{A}^\top\boldsymbol{A} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{a}_i\boldsymbol{a}_i^\top \approx \boldsymbol{R}_{aa}. \tag{29}
$$

This approximation is suggested by the law of large numbers, which states that

$$
\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{a}_i\boldsymbol{a}_i^\top = \boldsymbol{R}_{aa}. \tag{30}
$$

With these additional assumptions, we can write the mean-squared error for fixed $\boldsymbol{x}$ as

$$\begin{aligned}
\mathrm{E}\left\{(\widehat{y}-y)^2 \mid \boldsymbol{x}\right\} &= \sigma^2 \operatorname{tr}\left[\boldsymbol{a}\boldsymbol{a}^\top \mathrm{E}\{(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\}\right] + \sigma^2 && (31)\\
&= \tfrac{\sigma^2}{n} \operatorname{tr}\left[\boldsymbol{a}\boldsymbol{a}^\top \mathrm{E}\{(\tfrac{1}{n}\boldsymbol{A}^\top\boldsymbol{A})^{-1}\}\right] + \sigma^2 && (32)\\
&\approx \tfrac{\sigma^2}{n} \operatorname{tr}\left[\boldsymbol{a}\boldsymbol{a}^\top \mathrm{E}\{\boldsymbol{R}_{aa}^{-1}\}\right] + \sigma^2 && (33)\\
&= \tfrac{\sigma^2}{n} \operatorname{tr}\left[\boldsymbol{a}\boldsymbol{a}^\top \boldsymbol{R}_{aa}^{-1}\right] + \sigma^2 && (34)\\
&= \tfrac{\sigma^2}{n} \boldsymbol{a}^\top \boldsymbol{R}_{aa}^{-1}\boldsymbol{a} + \sigma^2, && (35)
\end{aligned}$$

where assumed large n to apply the approximation in (29).

The result in (35) is not an intuitive as we might like. This because it is stated for a fixed $\boldsymbol{x}$ (i.e., fixed $\boldsymbol{a}$). A much more intuitive expression results if we also average over $\boldsymbol{x}$, i.e., if we compute $\mathrm{E}\{(\widehat{y}-y)^2\}$. To do this, we must impose a statistical model on the test features $\boldsymbol{x}$. We won't assume too much about $\boldsymbol{x}$, just that it is independent of ϵ and the training quantities $\{\boldsymbol{x}_i\}$ and $\boldsymbol{\epsilon}$, and that it is distributed identically to $\{\boldsymbol{x}_i\}$, so that $\mathrm{E}\{\boldsymbol{a}\boldsymbol{a}^\top\} = \boldsymbol{R}_{aa}$. Then, starting from (34), we get

$$\begin{aligned}
\mathrm{E}\left\{(\widehat{y}-y)^2\right\} &\approx \tfrac{\sigma^2}{n} \operatorname{tr}\left[\mathrm{E}\{\boldsymbol{a}\boldsymbol{a}^\top\}\boldsymbol{R}_{aa}^{-1}\right] + \sigma^2 && (36)\\
&= \tfrac{\sigma^2}{n} \operatorname{tr}\left[\boldsymbol{R}_{aa}\boldsymbol{R}_{aa}^{-1}\right] + \sigma^2 && (37)\\
&= \tfrac{\sigma^2}{n} \operatorname{tr}[\boldsymbol{I}_{d+1}] + \sigma^2 && (38)\\
&= \frac{d+1}{n}\sigma^2 + \sigma^2, && (39)
\end{aligned}$$

which is very insightful. It says that, as the number of training samples n grows (for a fixed d), the first term vanishes, leaving a mean-squared error of σ^2. But notice that a mean-squared error of σ^2 is what we would get with the ideal linear estimator, i.e., with $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$. For this reason, σ^2 is often referred to as the "irreducible error."

5 Predictor Variance

One additional clarification is in order. As discussed in the lecture, it is common in machine-learning to write the mean-squared prediction error $\mathrm{E}\{(\widehat{y}-y)^2 \mid \boldsymbol{x}\}$ as the sum of three terms: the squared "bias", the "variance", and the "irreducible error". The "bias" and the "irreducible error" terms are the same ones that we described earlier. But the "variance" is not the variance of the prediction *error*, as in (19), but rather the variance of the predictor itself, i.e.,

$$\operatorname{var}_{\widehat{y}}(\boldsymbol{x}) \triangleq \mathrm{E}\left\{\left(\widehat{y} - \mathrm{E}\{\widehat{y}|\boldsymbol{x}\}\right)^2 \,\middle|\, \boldsymbol{x}\right\} \qquad (40)$$

Because

$$\begin{aligned}
\mathrm{E}\{\widehat{y}|\boldsymbol{x}\} &= \mathrm{E}\left\{\boldsymbol{a}^\top\widehat{\boldsymbol{\beta}} \mid \boldsymbol{x}\right\} && (41)\\
&= \boldsymbol{a}^\top \mathrm{E}\left\{\widehat{\boldsymbol{\beta}} \mid \boldsymbol{x}\right\} && (42)\\
&= \boldsymbol{a}^\top \mathrm{E}\left\{\left(\boldsymbol{\beta} + (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{\epsilon}\right) \mid \boldsymbol{x}\right\} && (43)\\
&= \boldsymbol{a}^\top \Big(\boldsymbol{\beta} + \mathrm{E}\left\{(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\right\}\mathrm{E}\left\{\boldsymbol{\epsilon} \mid \boldsymbol{x}\right\}\Big) && (44)\\
&= \boldsymbol{a}^\top\boldsymbol{\beta} && (45)
\end{aligned}$$

we have

$$\begin{aligned}
\mathrm{var}_{\widehat{y}}(\boldsymbol{x}) &= \mathrm{E}\left\{\left(\widehat{y} - \boldsymbol{a}^\top \boldsymbol{\beta}\right)^2 \;\middle|\; \boldsymbol{x}\right\} && (46)\\
&= \mathrm{E}\left\{\left(\boldsymbol{a}^\top (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)^2 \;\middle|\; \boldsymbol{x}\right\} && (47)\\
&= \mathrm{E}\left\{\left(\boldsymbol{a}^\top (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{\epsilon}\right)^2 \;\middle|\; \boldsymbol{x}\right\} && (48)\\
&= \mathrm{E}\left\{\boldsymbol{a}^\top (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top \boldsymbol{A} (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{a} \;\middle|\; \boldsymbol{x}\right\}, && (49)
\end{aligned}$$

which is identical to (23) except that it is missing the σ^2 term. Thus, if we apply the same analysis as before, we will eventually arrive at

$$\mathrm{var}_{\widehat{y}} \triangleq \mathrm{E}\{\mathrm{var}_{\widehat{y}}(\boldsymbol{x})\} \approx \frac{d+1}{n}\sigma^2. \tag{50}$$

This last expression is the "variance" of LS linear regression, and it differs from the (unbiased) error-variance $\mathrm{var}\{\widehat{y} - y\} = \mathrm{E}\{(\widehat{y} - y)^2\}$ from (39) in that it does not include the irreducible error term, σ^2.