Lecture 7: MATH 342W: Introduction to Data Science and Machine Learning

Sergio E. Garcia Tapia*

February 20, 2025 (last updated May 3, 2025)

Estimating Covariance

Last time, we introduced ordinary least squares regression for $p = 1$ and derived the following expressions for the least squares estimates:

$$b_0 = \bar{y} - b_1\bar{x} \tag{1}$$

$$b_1 = r\frac{S_y}{S_x} \tag{2}$$

$$r = \frac{S_{xy}}{S_x S_y}$$

where

$$S_x^2 := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S_y^2 := \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$S_{xy} := \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{3}$$

The quantities S_x^2 and S_y^2 estimate the variance of σ_x^2 and σ_y^2, respectively (the variances of x and y). This is similar to how $\bar{x}$ seeks to estimate μ_X and $\bar{y}$ seeks to estimate μ_Y. The quantity S_{xy} estimates the *covariance* of x and y, given by

$$Cov[X, Y] := \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

where X and Y are random variables, μ_X and μ_Y are their respective means, and E is used to compute the expectation. Though not a core part of this course, we will briefly talk about the meaning of $Cov[X, Y]$, and we'll use the estimate in S_{xy} from Equation 3.

Case 1: $S_{xy} > 0$

Consider the data set depicted in Figure 1. We've plotted a vertical and horizontal line through $\bar{x}$ and $\bar{y}$. This causes the plane to be partitioned into four regions. We focus on two cases:

*Based on lectures of Dr. Adam Kapelner at Queens College. See also the course GitHub page.

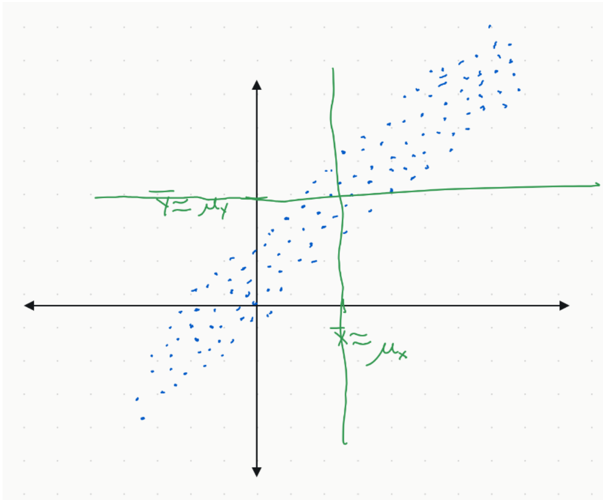

Figure 1: A data set where $S_{xy} > 0$

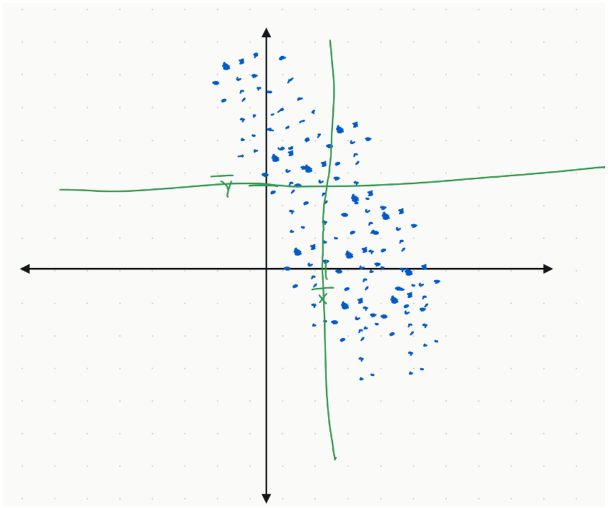

Figure 2: A data set where $S_{xy} < 0$

(i) If $x_i \geq \bar{x}$, we see that y_i tends to be above the mean. Now $(x_i - \bar{x}) \geq 0$, and $(y_i - \bar{y})$ is mostly positive also. Therefore, the product is mostly positive.

(ii) If $x_i < \bar{x}$, we see that y_i tends to be below the mean. Now $(x_i - \bar{x}) < 0$, and $(y_i - \bar{y})$ is mostly negative also. Once again, the product tends to be positive.

Thus, in this case we have $S_{xy} > 0$.

Case 2: $S_{xy} < 0$

In Figure 2, we've plotted a different data set. Once again, we've plotted the vertical and horizontal lines through $\bar{x}$ and $\bar{y}$, respectively. This causes the plane to be partitioned into four regions like before. We focus on two similar cases:

(i) If $x_i \geq \bar{x}$, we see that y_i tends to be below the mean. Now $(x_i - \bar{x}) \geq 0$, and $(y_i - \bar{y})$ is mostly negative also. Therefore, the product is mostly negative.

(ii) If $x_i < \bar{x}$, we see that y_i tends to be above the mean. Now $(x_i - \bar{x}) < 0$, and $(y_i - \bar{y})$ is mostly positive also. Once again, the product tends to be negative.

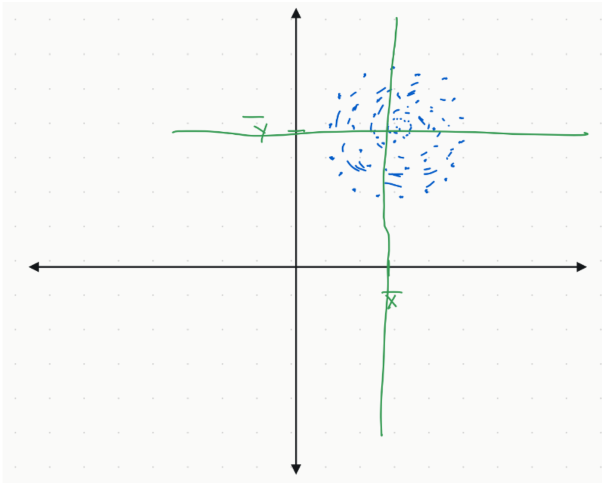

Figure 3: A data set where $S_{xy} \approx 0$

Thus, in this case we have $S_{xy} < 0$.

Case 3: $S_{xy} = 0$

For this last case, we've plotted points in Figure 3 such that they appear to be distributed like a disk roughly around the point $(\bar{x}, \bar{y})$.

(i) If $x_i \geq \bar{x}$, we see that y_i tends to be above the mean just as often as it is below the mean, so on average, we see $(y_i - \bar{y}) \approx 0$. Therefore, the product is mostly near zero.

(ii) If $x_i < \bar{x}$, we see the same behavior.

Therefore, in this case we have $S_{xy} \approx 0$.

Interpretation

Having observed the previous three cases and considered the estimates S_{xy}, what can we say about $Cov[X, Y]$? It measures the degree and direction of the linear relationship between X and Y in units of x times y. For example, if x is measured in kilograms (kg) and y is measured in meters (m), then the units of S_{xy} are kilograms-meters (kg-m). Meanwhile, the quantity $r = S_{xy}/S_x S_y$ is unitless.

Least Squares Facts

Theorem 1. Suppose $\mathbb{D} = \{(x_i, y_i)\}_{i=1}^n$ is a data set where each $x_i, y_i \in \mathbb{R}$. If b_0 and b_1 are the ordinary least square estimates, $\hat{y}_i$ is the prediction corresponding to x_i using the least squares estimates, and $e_i = y_i - \hat{y}_i$ is the residual for the ith response, then the sum of the residuals is zero:

$$\sum_{i=1}^{n} e_i = 0$$

Proof. Given the least squares estimates in Equations 1 and 2, the linear model is

$g(x) = b_0 + b_1 x$. In particular, $\hat{y}_i = g(x_i) + b_0 + b_1 x_i$, so

$$\begin{aligned}
\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n}(y_i - \hat{y}_i) \\
&= \sum_{i=1}^{n}(y_i - g(x_i)) \\
&= \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) \\
&= \sum_{i=1}^{n} y_i - b_0 \sum_{i=1}^{n}(1) - b_1 \sum_{i=1}^{n} x_i \\
&= n\bar{y} - nb_0 - b_1 n\bar{x} \\
&= n(\underbrace{\bar{y} - b_1\bar{x}}_{b_0} - b_0) \\
&= n(b_0 - b_0) \\
&= 0.
\end{aligned}$$

□

Corollary 1. The mean of the residuals in OLS is zero: $\bar{e} = 0$.

Theorem 2. Let $\mathbb{D} = \{(x_i, y_i)\}_{i=1}^{n}$ be a data set with $x_i, y_i \in \mathbb{R}$, and let b_0, b_1 be the least squares estimates. If $\bar{x}$ and $\bar{y}$ denote the averages of the x_i's and y_i's respectively, then the point $(\bar{x}, \bar{y})$ lies on the least squares line defined by the estimates.

Proof. Since $g(x) = b_0 + b_1 x$, we can once again use Equation 1 to conclude that

$$g(\bar{x}) = b_0 + b_1\bar{x} = (\bar{y} - b_1\bar{x}) + b_1\bar{x} = \bar{y}.$$

□

The Performance of the Linear Least Squares Model

Recall that OLS was introduced when moving to a numeric response space, i.e., $\mathcal{Y} = \mathbb{R}$. We began by considering the simple case where $p = 1$. In this setting, we saw that the null model $g_0 = \bar{y}$. That is, the null model always predicts $\bar{y}$. This too is a linear model, with a slope of 0 and an intercept of $\bar{y}$. Figure 4 illustrates the null model when $p = 1$. Notice how the residuals $e_i = y_i - \bar{y}_i$ are positive for points above the graph of g_0 and negative for points below the graph of g_0. One way to think about the performance of the null model is to view the distribution of the residuals, as illustrated in Figure 5. Notice how this distribution is centered around 0, which is consistent with the fact that $\bar{e} = 0$, as in Corollary 1.

We seek to compare the performance of the null model with the least squares estimates

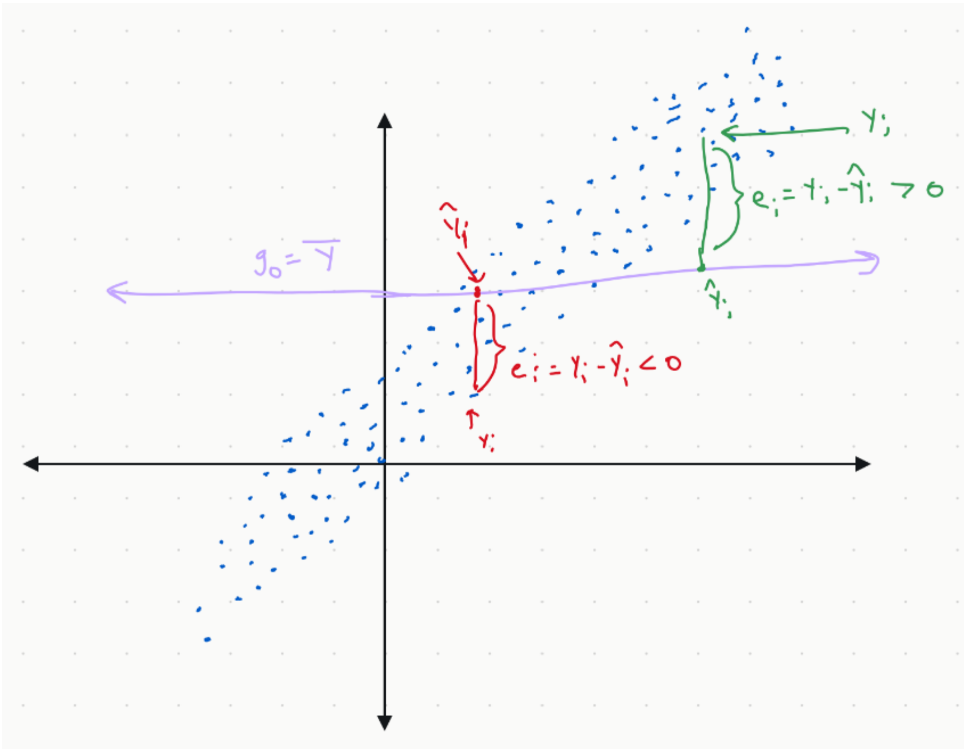

Figure 4: The null model in OLS with $p = 1$.

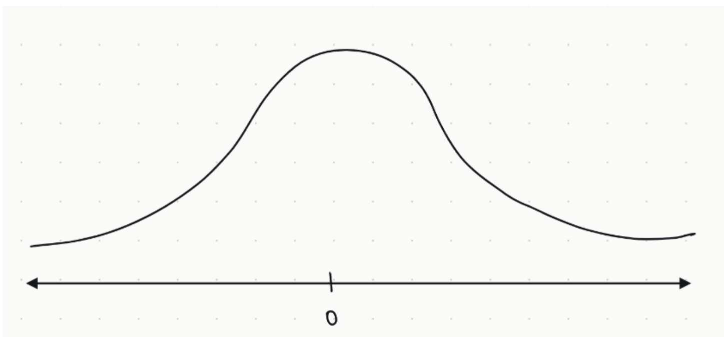

Figure 5: The distribution of the residuals in the null model in Figure 4.

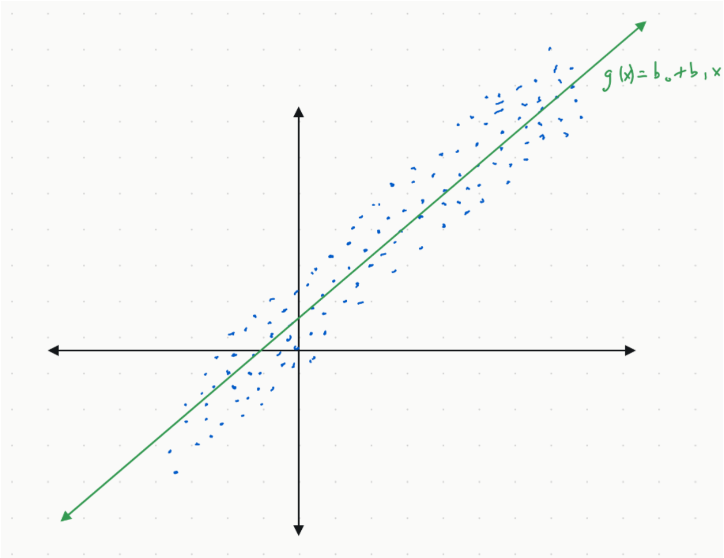

Figure 6: The least squares model for the same data set as in Figure 4.

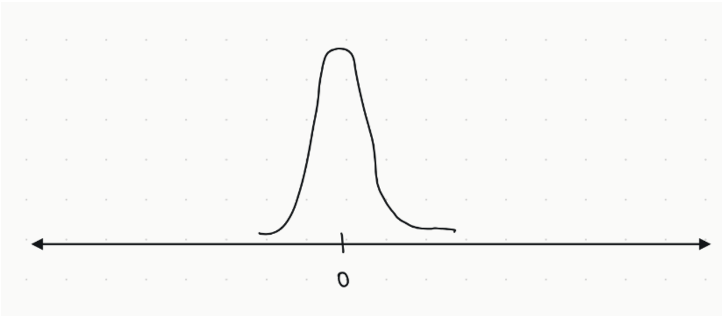

Figure 7: Distribution of the residuals in the model in Figure 6.

by comparing the distribution of the residuals. Figure 6 shows the linear model defined by the least squares estimates (Equations 1 and 2). In Figure 7, we see the corresponding distribution of the residuals. Notice that similar to g_0, the distribution of the residuals is centered at 0, but it has a smaller variance compared to g_0. In fact, we will later show that this is always the case for OLS[1].

The Variance in the Residuals

Let E_0 be a random variable for the residuals of the null model g_0, and consider the variance σ_0^2 of E_0:

$$\sigma_0^2 := \text{Var}[E_0] := \mathbb{E}[(E_0 - \mu_{E_0}^2)]$$

where $\mathbb{E}[\cdot]$ stands for expectation, and μ_{E_0} stands for the mean of E_0. As we've seen, the variance is estimated by $S_{e_0}^2$ (where now we use lowercase e_0 for the *realization* of the

[1]Based on our discussion on covariance, we can also better understand the effect of r in Equation 2. The quantities S_x and S_y are estimations of the standard deviation of x and y, respectively, so they are positive quantities. Meanwhile, as we saw, $r = S_{xy}/S_xS_y$, and its sign is dictated by S_{xy}. Hence, the slope of the least squares regression line is b_1, whose sign is determined by r, which in turn has its sign determined by S_{xy}.

random variable E_0). By Corollary 1, we know $\overline{e} = 0$. Recalling that the model always predicts $\overline{y}$, we can show that

$$\begin{aligned}
S_{e_0}^2 &:= \frac{1}{n-2}\sum_{i=1}^{n}(e_i - \bar{e})^2 \\
&= \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 \\
&= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \bar{y})^2 && (\hat{y}_i = \bar{y} \text{ for } g_0) \\
&:= \frac{1}{n-2} SSE_0,
\end{aligned}$$

where we use the symbol SSE_0 to denote the sum of squared errors for the null model. The sum of squared errors for the null model is also called the *sum of squared totals*, or SST.

Definition. Given a finite sequence of values $(y_i)_{i=1}^n$ with average $\overline{y}$, the **sum of squared totals**, or SST, is defined by

$$SST := SSE_0 = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

The SST is the SSE of the null model g_0.

Now let E_g be the random variable for the residuals of the least square model g. Once again, we consider the variance in E_g

$$\sigma_{E_g}^2 := \mathrm{Var}[E_g] := \mathbb{E}[(E_g - \mu_{E_g}^2)]$$

which is estimated by $S_{e_g}^2$ (noticed again lowercase e_g for the realization of E_g)

$$\begin{aligned}
S_{e_g}^2 &:= \frac{1}{n-2}\sum_{i=1}^{n}(e_{g,i} - \bar{e}_g)^2 \\
&= \frac{1}{n-2}\sum_{i=1}^{n} e_{g,i}^2 && (\bar{e}_g = 0 \text{ by Corollary 1}) \\
&= \frac{1}{n-2}\sum_{i=1}^{n} SSE_g \\
&:= MSE_g
\end{aligned}$$

where MSE_g is the mean-squared error of g and SSE_g is the sum of squared errors of g. The relations shown reveal why SSE_g is an important performance metric: it is related to the variance in the errors of the residuals.

The R^2 Performance Metric

In literature, $S_{e_0}^2$ is referred to as the estimate of the variance of the error that you "begin with", while $S_{e_g}^2$ is the estimate of the variance of the error that you "end with". One intuitive way to understand this is to recall that the null model is our naive approach to making predictions and it does not use take into account any of the features. The least squares estimates, on the other hand, does consider the feature inputs and tries to minimize the errors in the predictions. Thus, it is reasonable to say that as you obtain more data, you veer away from $S_{e_0}^2$ and towards $S_{e_g}^2$ (henceforth referred to as S_e^2).

Another important quantity is $S_{e_0}^2 - S_e^2$, which is an estimate of the variance of the error "explained". What does this mean? Well, the purpose of this class is to model phenomena with some response y, to learn from the data, and to make predictions. An increase in $S_{e_0}^2 - S_e^2$ indicates that we are better able to "explain" the the response, or the variance in it. In fact, this is related to the **R-squared (R^2) performance metric** for regression analysis:

$$R^2 := \frac{S_{e_0}^2 - S_e^2}{S_{e_0}^2} \tag{4}$$

This represents the *estimate of the proportion of variance explained.* This will serve as our fourth performance metric. Though defined this way, it is often presented in different ways. We will cast this in an alternate form:

$$\begin{aligned}
R^2 &:= \frac{S_{e_0}^2 - S_e^2}{S_{e_0}^2} \\
&= \frac{\frac{1}{n-2}SST - \frac{1}{n-2}SSE}{\frac{1}{n-2}SST} \\
&= \frac{SST - SSE}{SST} \\
&= 1 - \frac{SSE}{SST}
\end{aligned}$$

Hence, we can alternatively define the R^2 performance metric as

$$R^2 := 1 - \frac{SSE}{SST} \tag{5}$$

As we will show, $R^2 \in [0, 1]$ for the linear least squares model. For now, we make the following notes:

- From Equation 5, we see that $R^2 = 1$ is equivalent to having $SSE = 0$, which in turn is equivalent to $e_i = 0$ for all i. We say in that case that the least squares estimate is a *perfect fit.*

- $R^2 = 0$ is equivalent to $SSE = SST$, which means the least squares estimate corresponds to the null model g_0.

How do R^2 and $RMSE$ (root mean squared error) compare? Since $RMSE = \sqrt{\frac{1}{n-2}SSE}$, we see that $RMSE$ and SSE increase (and decrease) together. Thus, from Equation 5 it is evident that a decrease in $RMSE$ corresponds to a decrease in SSE, and hence an increase in R^2. Similarly, an increase in $RMSE$ leads to a decrease in R^2. Since we have seen that a low $RMSE$ is desirable, this means that a high R^2 is preferred.

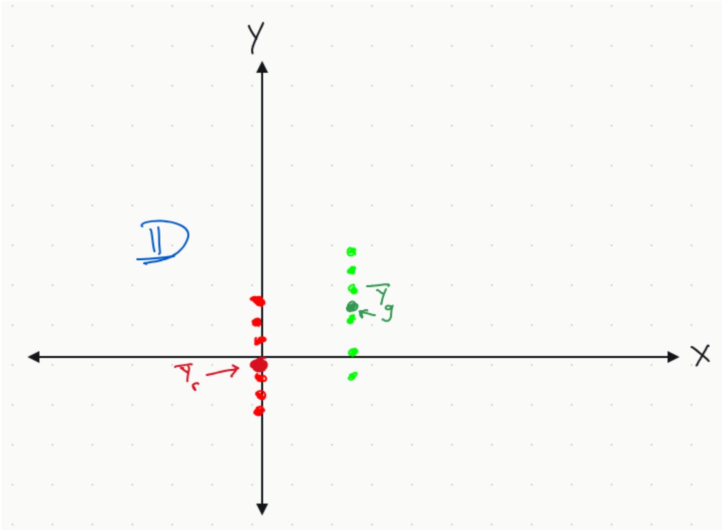

Figure 8: A data set $\mathbb{D}$ with $\mathcal{X} = \{\text{red}, \text{green}\}$ and $\mathbb{Y} = \mathbb{R}$.

R^2 vs. $RMSE$

Which performance metric is more informative: R^2 or $RMSE$? It turns out $RMSE$ tends to be more informative. For example, it's possible for R^2 to be very close to 1, but we are in a context where due to the large scale of values, the $RMSE$ is still large in an absolute sense (even though it is small in a relative sense). In fact, we can make the following statement about $RMSE$. Given a prediction $\hat{y}$, the *95% confidence interval* has radius $4 \cdot RMSE$:

$$CI_{y,95\%} = [\hat{y} - 2 \cdot RMSE, \hat{y} + 2 \cdot RMSE]$$

That is, given the prediction $\hat{y}$, we can say the interval above is a reasonable range for the actual value of y. This result is out of the scope of this course, but we mention it because we will nevertheless use it from time to time.

ANOVA

Suppose we are using linear modeling in a setting where $\mathcal{Y} = \mathbb{R}$, but $\mathcal{X} = \{0, 1\}$. That is, the response space is numeric while the feature space is binary (equivalently, the feature is categorical). This is a special case of linear least squares modeling before we move on to the multivariate case. Take, for example, the data set $\mathbb{D}$ in Figure 8. The inputs are "red" and "green", which is a categorical variable that we can encode as 0 and 1, respectively. For these inputs, we have several points each with a different real-valued response. Figure 8 also highlights the values $\bar{y}_\text{r}$ and $\bar{y}_\text{g}$ (that's g for green, not the g function), the average of the responses when limiting ourselves to the red and green inputs, respectively. A reasonable, yet naive model g would predict that

$$g(\text{red}) = \bar{y}_\text{r}, \tag{6}$$
$$g(\text{green}) = \bar{y}_\text{g}. \tag{7}$$

We will show that these are in fact the least squares estimates. This model is historically important, and it is called **ANOVA (ANalaysis Of VAriance)**. It is a linear "regression" model for categorical variables. We proceed to prove our assertion about Equations 6 and 7.

First, we will also introduce some notation. Let n_r denote the number of red inputs and n_g denote the number of green inputs. Also, let p_r denote the proportion of red inputs and let p_g denote the proportion of green inputs. Then we have the following relations:

$$n = n_r + n_g, \quad p_r = \frac{n_r}{n}, \quad p_g = \frac{n_g}{n}, \quad p_r = (1 - p_g)$$

Since we are encoding "red" as 0 and "green" as 1, we can use this notation to write

$$\begin{aligned}
\bar{y}_r &= \frac{1}{n_r} \sum_{\{\,i|\ x_i=0\,\}} y_i \\
\bar{y}_g &= \frac{1}{n_g} \sum_{\{\,i|\ x_i=1\,\}} y_i
\end{aligned}$$

Next, we note that since $x_i \in \{0, 1\}$, it follows that $x_i = x_i^2$, so we have the following identities:

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \\
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{\{\,i|\ x_i=1\,\}} x_i = \frac{n_g}{n} = p_g \qquad \iff \qquad n\bar{x} = np_g
\end{aligned}$$

We now invoke the equivalent of Equation 2 that we derived last time:

$$\begin{aligned}
b_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\
&= \frac{\sum\limits_{\{\,i|\ x_i=1\,\}} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i - n\bar{x}^2} && \text{(Zeros do not contribute to sum, and } x_i = x_i^2\text{)} \\
&= \frac{n_g \cdot \frac{1}{n_g}\sum\limits_{\{\,i|\ x_i=1\,\}} y_i - np_g\bar{y}}{np_g - np_g^2} && (x_i = 1 \implies x_i \cdot y_i = y_i) \\
&= \frac{np_g\bar{y}_g - np_g\bar{y}}{np_g - np_g^2} && (n_g = np_g) \\
&= \frac{\bar{y}_g - \bar{y}}{1 - p_g} \\
&= \frac{\bar{y}_g - [p_g\bar{y}_g + (1 - p_g)\bar{y}_r]}{1 - p_g} \\
&= \bar{y}_g - \bar{y}_r
\end{aligned}$$

Now using Equation 1:

$$\begin{aligned}
b_0 &= \bar{y} - b_1\bar{x} \\
&= (p_g\bar{y}_g + (1 - p_g)\bar{y}_r) - (\bar{y}_g - \bar{y}_r)p_g \\
&= p_g\bar{y}_g - p_g\bar{y}_g + \bar{y}_r - p_g\bar{y}_r + \bar{y}_r p_g \\
&= \bar{y}_r
\end{aligned}$$

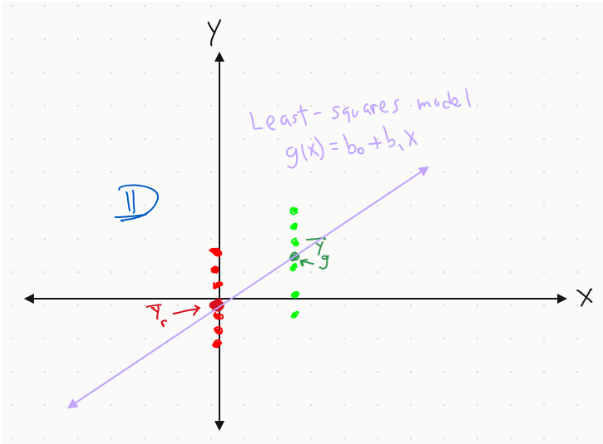

Figure 9: A linear least squares fit for the data set in Figure 8.

In summary, the least squares estimate are given by

$$b_0 = \bar{y}_r \tag{8}$$

$$b_1 = \bar{y}_g - \bar{y}_r \tag{9}$$

Therefore, the linear least squares model is $g(x) = b_0 + b_1 x$. Now we can prove our assertion from Equations 6 and 7:

$$\begin{aligned} g(0) &= b_0 + b_1(0) = b_0 = \bar{y}_r, \\ g(1) &= b_0 + b_1(1) = \bar{y}_r + (\bar{y}_g - \bar{y}_r) = \bar{y}_g, \end{aligned}$$

as we set out to show. This line is shown in Figure 9. Note that the intercept corresponds to the "red" inputs, causing "red" to take on the role of the reference variable. In particular, b_0 is $\bar{y}_r$, and b_1 is not $\bar{y}_g$, but rather, an offset from $\bar{y}_r$. This follows because we derived least squares by prepending to our design matrix X a column of 1's, which we have been denoting as $\vec{\mathbf{1}}_n$. If, however, we omitted $\vec{\mathbf{1}}_n$ and instead used one binary column for each category (level), then each coefficient b_0, b_1 would correspond to the mean response for that level.

Multivariate Linear Least Squares Regression

We will now work towards extending the least square models to the multivariate case where $p > 1$. Our set of candidate functions is the set of hyperplanes:

$$\mathcal{H} = \{\mathbf{x} \cdot \mathbf{w} = 0 : \ \mathbf{w} \in \mathbb{R}^{p+1}\}$$

As a reminder, though we have p features, we extend each input to length $p + 1$ by prepending a 1 entry to each input vector. This is simply for convenience, so that instead of writing $\mathbf{x} \cdot \mathbf{w} + b_0$, we can let $w_0 = b_0$, $x_0 = 1$, and write $\mathbf{x} \cdot \mathbf{w} = 0$. Thus, our matrix X of inputs from $\mathbb{D}$ is of dimension $n \times (p + 1)$, and looks like

$$X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}$$

As in the case for $p = 1$, the least squares algorithm seeks to find the vector $\mathbf{b}$ defining a hyperplane that minimizes the SSE (the sum of squared errors):

$$\mathcal{A}(\mathbb{D}, \mathcal{H}) = \mathbf{b} = \underset{\mathbf{w} \in \mathbb{R}^{p+1}}{\text{argmin}}\{SSE\}$$

where

$$SSE := \sum_{i=1}^{n} e_i^2$$

We will use vector notation for our computations. Let

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_p \end{bmatrix}$$

Since $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, we can recast the equation for SSE:

$$\begin{aligned} SSE &= \sum_{i=1}^{n} e_i^2 \\ &= \mathbf{e} \cdot \mathbf{e} \\ &= \mathbf{e}^\top \mathbf{e} \\ &= (\mathbf{y} - \hat{\mathbf{y}})^\top (\mathbf{y} - \hat{\mathbf{y}}) \\ &= (\mathbf{y}^\top - \hat{\mathbf{y}}^\top)(\mathbf{y} - \hat{\mathbf{y}}) \\ &= \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \hat{\mathbf{y}} - \hat{\mathbf{y}}^\top \mathbf{y} + \hat{\mathbf{y}}^\top \hat{\mathbf{y}} \\ &= \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \hat{\mathbf{y}} - (\mathbf{y}^\top \hat{\mathbf{y}})^\top + \hat{\mathbf{y}}^\top \hat{\mathbf{y}} \end{aligned}$$

Note we made use of the fact that $(A + B)^\top = A^\top + B^\top$, and that $(AB)^\top = B^\top A^\top$. Moreover, in general $A \neq A^\top$, but since $\mathbf{y}^\top \hat{\mathbf{y}}$ is a scalar, it does equal $(\mathbf{y}^\top \hat{\mathbf{y}})^\top$. Another relation that we will make use of is the fact that the prediction is given by matrix multiplication:

$$\hat{\mathbf{y}} = X\mathbf{w}$$

Be sure to convince yourself of this. Putting these facts together, we can write

$$\begin{aligned} SSE &= \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \hat{\mathbf{y}} - (\mathbf{y}^\top \hat{\mathbf{y}})^\top + \hat{\mathbf{y}}^\top \hat{\mathbf{y}} \\ &= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \hat{\mathbf{y}} + \hat{\mathbf{y}}^\top \hat{\mathbf{y}} \\ &= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top (X\mathbf{w}) + (X\mathbf{w})^\top (X\mathbf{w}) \\ &= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top X\mathbf{w} + \mathbf{w}^\top X^\top X\mathbf{w} \end{aligned}$$

To minimize the SSE, we now take $p + 1$ partial derivatives, one with respect to each of $w_0, w_1, w_2, \ldots, w_p$, and we set them all to zero:

$$\frac{\partial}{\partial \mathbf{w}}[SSE] = \begin{bmatrix} \frac{\partial}{\partial w_0}[SSE] \\ \frac{\partial}{\partial w_1}[SSE] \\ \frac{\partial}{\partial w_2}[SSE] \\ \vdots \\ \frac{\partial}{\partial w_p}[SSE] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}_{p+1}$$

In the case when $p = 1$, we laboriously solved a system of equations to obtain the estimates b_0 and b_1. For the general case where $p > 1$, we will resort to useful results from differential calculus and linear algebra to solve this system in one swoop. Next, we digress to mention these key results.

Digression on Differential Calculus and Linear Algebra

Proposition 1. *Let $\mathbf{x} \in \mathbb{R}^n$, and let $a \in \mathbb{R}$ be a constant with respect to all entries of $\mathbf{x}$. Then*

$$\frac{\partial}{\partial \mathbf{x}}[a] := \begin{bmatrix} \frac{\partial}{\partial x_1}[a] \\ \frac{\partial}{\partial x_2}[a] \\ \vdots \\ \frac{\partial}{\partial x_n}[a] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}_n$$

Proposition 2. *Let $\mathbf{a} \in \mathbb{R}^n$, where all entries are constant with respect to all entries of $\mathbf{x}$. Then*

$$\frac{\partial}{\partial \mathbf{x}}[\mathbf{a} \cdot \mathbf{x}] = \frac{\partial}{\partial \mathbf{x}}[\mathbf{a}^\top \mathbf{x}] = \mathbf{a}$$

Next time we will show other key facts that we will use.