Unit 4: Regression with multiple predictors

Ethan Levien

November 19, 2025

Introduction

Now we are ready to study linear regression with multiple predictors. Much of the concepts carry over from the single predictor case and the Python code is nearly the same. There are some new aspects though when it comes to how we interpret the predictors. In particular, we have to remember that these represent average differences in the response variable when all other predictors are fixed. This is the idea of controlling for another variable. We will also understand what happens when we add regression coefficients to the model.

4.1 Multiple predictor linear regression

The real power of regression comes when we work with models of the form

$$Y = \beta_0 + \sum_{i=1}^{K} \beta_i X_i + \epsilon \tag{1}$$

$$\epsilon \sim \text{Normal}(0, \sigma^2) \tag{2}$$

where X_i is a set of K predictor variables. Alternatively, we can write

$$Y|(X_1 = x_1, \ldots, X_K = x_K) \sim \text{Normal}\left(\beta_0 + \sum_{i=1}^{K} \beta_i X_i, \sigma^2\right) \tag{3}$$

You might see the shorthand,

$$Y \sim \text{LR}(X, \beta, \sigma^2). \tag{4}$$

In these notes, our goal is to answer the following questions

1. What are estimators of the parameters in this model?
2. How do we interpret the regression coefficients β_i?
3. What is the sample distribution of the regression coefficients?
4. How do the correlations between predictors influence the answers to these questions?

Example 1 (Two-predictor linear regression in `statsmodels`)**.** Question: Generate data from the two-predictor model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

and fit the linear regression $Y \sim X_1 + X_2$ using `statsmodels`.

Solution:

```
import numpy as np
```

```
import statsmodels.api as sm

# --- simulate data ---
n = 300
rng = np.random.default_rng(0)

beta0, beta1, beta2, sigma = 1.0, 2.0, -1.0, 0.5
X1 = rng.normal(size=n)
X2 = rng.normal(size=n)
eps = rng.normal(scale=sigma, size=n)
Y = beta0 + beta1*X1 + beta2*X2 + eps

# --- fit OLS: Y ~ X1 + X2 ---
X = sm.add_constant(np.column_stack([X1, X2]))
model = sm.OLS(Y, X)
results = model.fit()
```

The output from the regression with multiple predictors is basically the same as for single-predictor, except now we have multiple rows for the difference regression coefficients. In each case, the interpretation of the p-value and confidence intervals are nearly the same as they were for the single predictor case. However, for the p-value, we need to remember that this is the p-value testing the hypothesis that a particular predictor is zero. The F-statistic is used to test the hypothesis that all predictors are zero, although I won't go into much more detail because I don't place a big emphasis on hypothesis testing in this course.

The interpretation of R^2 is the same as before, except that now we are considering the ratio of the variance conditioned on ALL predictors to the overall variation in Y; that is,

$$R^2 = 1 - \frac{\sum_i r_i^2}{\sum_i (y_i - \bar{y})^2} \approx 1 - \frac{\text{var}(Y|X_1, X_2)}{\text{var}(Y)}$$

where in the multi-predictor case

$$r_i = Y_i - \left(\hat{\beta}_0 + \sum_k^m \hat{\beta}_k X_{i,k} \right). \tag{5}$$

4.1.1 Basic interpretation and estimation of the parameters

References: [1, Ch. 10]

In order to interpret the parameters, it's easiest to work with just two predictors like we have in the example above. The formula for the conditional expectation of Y is

$$E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \tag{6}$$

where I'm using the shorthand

$$E[Y|X] = E[Y|(X_1, X_2)]$$

to mean the expected value of Y conditioned on **both** predictors.

Equation 6 is the equation for a flat surface in two dimensions:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{7}$$

A drawing of y is shown in 2.

If we make a slice through the surface in the x_1 direction and look it at from the side, we see a line with slope β_1 (and similarly for x_2). This leads to the following interpretation of β_i:

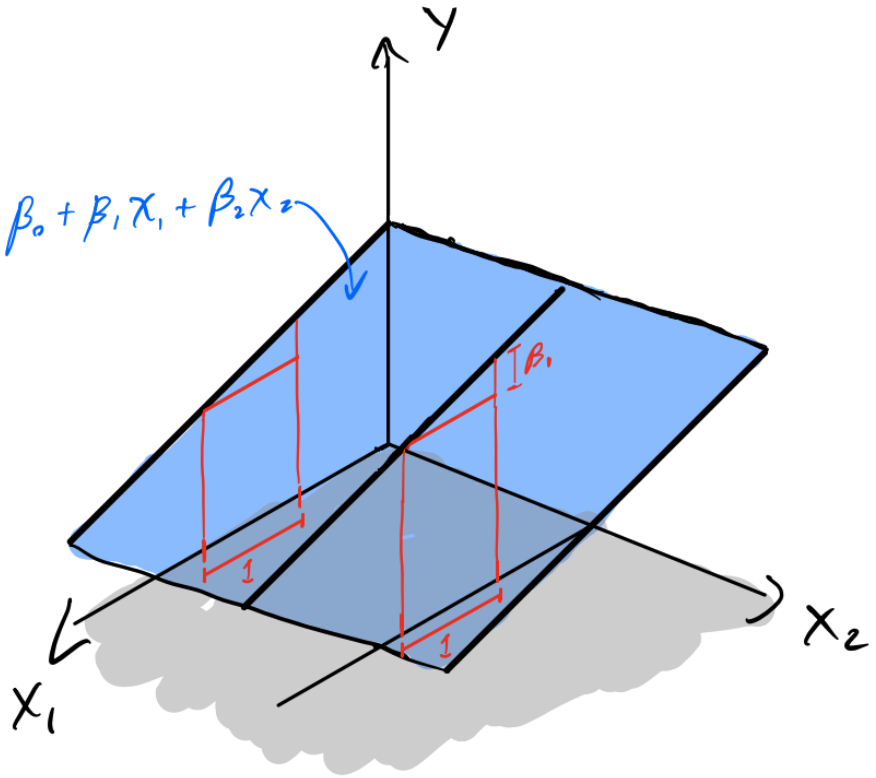

Figure 1: The function $y(x_1, x_2)$

β_1 is the slope of $E[Y|X]$ vs. X_1 for fixed X_2.

Notice that in the statement above, even though we are conditioning on both variables, the slope β_1 is independent of which value of X_2 we condition on. We can obtain the interpretation of β_2 by flipping the role of X_1 and X_2. The fact that is doesn't matter which value of X_2 (respectively X_1) we have conditioned on is a consequence of linearity, and thus one of the core model assumption of linear regression with multiple predictors, which we do not encounter in the single predictor case. Another way of articulating it is to say: the "effect" of X_1 and X_2 are not dependent on the other predictor's value.

Example 2 (Test scores)**.** We will now work with a new example of Children's test scores. To motivate this, we can imagine we are interested in studying what factors determine children's success in school in order to effective design interventions which help students that are struggling. The predictors are mother IQ and high school education. In this case, the model assumptions are saying that the association between the mother's high school education and test scores is not influenced by the mother's IQ. that is, If we compare two random children whose mothers have the same IQ, differ in whether they attended high school, then the average *difference* between their test scores will not depend on the IQ of their mothers, although the average magnitude of their test scores will depend on the mother's IQ.
Question: Fit the data to a linear regression model with two predictors and answer the questions

(a) What are the regression coefficients and the interpretations?

(b) Based on this regression analysis, which factor, IQ or high school education do we believe is more predictive of test scores?

(c) Overall, how well do high school education and IQ as methods do at predicting the test scores of children?

(d) What is the chance a student whose mother has an IQ of 90 and did not go to high school does better than a student whose mother has an IQ of 110 and did go to high school?

Solutions: We get the following output from `statsmodels` in the colab notebook:

```
>                            OLS Regression Results
> ==============================================================================
> Dep. Variable:                      y   R-squared:                       0.214
> ==============================================================================
>                  coef    std err          t      P>|t|      [0.025      0.975]
> ------------------------------------------------------------------------------
> const         25.7315      5.875      4.380      0.000      14.184      37.279
> mom_hs         5.9501      2.212      2.690      0.007       1.603      10.297
> mom_iq         0.5639      0.061      9.309      0.000       0.445       0.683
>
>
```

(a) For the regression coefficients we find the follow:

- $\beta_{\text{hs}} \approx 5.95$. This means that among students whose mothers **have the same IQ**, a student whose mother attended high school will, on average, have a score that is 5.95 points higher than a student whose mother did not.
- $\beta_{\text{iq}} \approx 0.56$. This means that among students whose mother's **have the same high school education** (either they all attended or did not attend high school), the difference between scores of students whose mothers IQ differs by one point is, on average, 0.56 points.
- $\hat{\beta}_0 \approx 26$. Mathematically, this tells us the average score of students whose mother did not attend high school and have zero IQ, but this is not a meaningful quantity since noone has zero iq. We can therefore ignore it when it comes to interpreting the output.

(b) Clearly β_{hs} is smaller, but we need to remember that are comparing quantities that have different units. X_{iq} takes values from around 70 to 130, while X_{hq} is either zero or 1. What is actually more useful is to compare how much a difference in one standard deviation of the predictor makes. For example, $\beta_{\text{iq}}\sigma_{\text{iq}}$ is the average difference in test scores between students whose mothers have the same high school education, but whose mother's IQ differ by one standard deviation. To this end, we can compute the following measures of effects

$$\hat{\beta}_{\text{hs}}\hat{\sigma}_{\text{hs}} \approx 2.44$$
$$\hat{\beta}_{\text{iq}}\hat{\sigma}_{\text{iq}} \approx 8.44.$$

The association between IQ and scores is actually larger. Note that the comparison is not perfect, since X_{hs} is a binary variable, but it still gives us a generally idea of the effects.

(c) The R^2 value is 0.214, so about 20% of the variation in test scores is explained by the variation in high school education and IQ of mothers.

(d) In the colab notebook we calculate this to be about 25%.

The interpretation can of course be generalized to the situation where we have many predictors. The general formula for the regression coefficient in terms of expectation is

$$\begin{aligned}\beta_i = E[Y|X_1, \ldots, X_{i-1}, X_i = x_i + 1, X_{i+1}, \ldots, X_K] \\ - E[Y|X_1, \ldots, X_{i-1}, X_i = x_i, X_{i+1}, \ldots, X_K]\end{aligned}$$

Note how this is a very natural extension of Equation 8. We get a more complex expression for the coefficients but the idea is the same.

4.1.2 Relationship between single and two predictor regression coefficients

We can express the regression coefficients explicitly in terms of conditional averages as

$$\beta_1 = E[Y|X_1 = (x + 1), X_2] - E[Y|X_1 = x, X_2]. \tag{8}$$

Now let's think about how the regression coefficients are related to covariance. One guess would be that, just as in the single-predictor case, β_1 is given by $\text{cov}(Y, X_1)/\sigma^2_{X_1}$. After all, if we look a slice of the 2D planer function $y(x_1, x_2)$ along the x_1 direction, we get the same slope for all x_2. It stands to reason that if we look at only the points in the x_1-y plane our regression slope would be β_1. However, **this argument assumes that when we change x_1, x_2 does not also change**. This is best understood with an example.

Example 3 (Test scores with multiple vs. single predictors)**.** Here we will consider once again the example of children's test scores and compare using both predictors in the sample above to the results we obtain we using only one predictor (high school education).

Question: What is the difference between the coefficient of X_{hs} when this is the only predictor and the coefficient when X_{iq} is also used? How is the coefficient in the multiple predictor case related to coefficient in the single predictor case?

Solution: When we performed the regression using only the mother's high school education as a predictor, we obtained a coefficients of about $\hat{\beta}'_{\text{hs}} \approx 12$ and $\hat{\beta}'_0 \approx 78$ (i'll use β' indicate coefficients in the single predictor model, as opposed to the multiple predictor model). The fitted model is

$$\hat{y} = 12X_{\text{hs}} + 78$$

while when also using X_{iq} as a predictor, the coefficient is about half that.
In the model with one predictor, the regression coefficient of 12 means that on average a student whose mother went to high school will do 12 points better than one whose mother did not. That is, we are predicting

$$E[Y|X_{\text{hs}} = 1] - E[Y|X_{\text{hs}} = 0] = \beta'_{\text{hs}} \approx 12$$

Let's compare this to what we would predict in the model with two predictors. In that case, the average test score of student whose mother went to high school is

$$\begin{aligned}\hat{y}_{\text{hs}} &\approx E[Y|X_{\text{hs}} = 1] \\ &= E[\beta_0 + \beta_{\text{hs}} + \beta_{\text{iq}}X_{\text{iq}}|X_{\text{hs}} = 1] \\ &= \beta_0 + \beta_{\text{hs}} + \beta_{\text{iq}}E[X_{\text{iq}}|X_{\text{hs}} = 1] \\ &\approx 6 \times 1 + 26 + 0.6\overline{X}_{\text{iq|hs}}\end{aligned}$$

where

$$\overline{X}_{\text{iq|hs}} = \text{sample average IQ of mother who attended high school} \approx E[X_{\text{iq}}|X_{\text{hs}} = 1]$$

On the other hand

$$\hat{y}_{\text{no-hs}} = 6 \times 0 + 26 + 0.6\overline{X}_{\text{iq,no-hs}}$$

where

$$\overline{X}_{\text{iq|no-hs}} = \text{sample average IQ of mother who DID NOT attend high school} \approx E[X_{\text{iq}}|X_{\text{hs}} = 0]$$

Thus, according to the model with two predictors, the average difference in test scores between the `hs` and `no-hs` groups is

$$\Delta\hat{y}_{\text{hs}} = 6 + 0.6(\overline{X}_{\text{iq|hs}} - \overline{X}_{\text{iq|no-hs}})$$

or written in terms of more probabilistic notation

$$E[Y|X_{\text{hs}} = 1] - E[Y|X_{\text{hs}} = 0] = \beta_{\text{hs}} + \beta_{\text{iq}}(E[X_{\text{iq}}|X_{\text{hs}} = 1] - E[X_{\text{iq}}|X_{\text{hs}} = 0])$$

We can compute $\overline{X}_{\text{iq|hs}} - \overline{X}_{\text{iq|no-hs}} \approx 10.3$, which gives $\Delta\hat{y}_{\text{hs}} \approx 12$. Thus, we have calculated the single-predictor regression coefficient from the multiple predictor case.

The important thing is that the two predictors are not independent. If they were, then $\overline{X}_{\text{iq|hs}} - \overline{X}_{\text{iq|no-hs}}$ would be zero, and it would have to be that the coefficient of X_{hs} is the same in both cases. We can generalize this to any model where X_1 is a binary predictor to obtain a relationship between the regression coefficient for β_1 with and without the second predictor; that is,

$$\beta_1' = \beta_1 + \beta_2(E[X_2|X_1 = 1] - E[X_2|X_1 = 0]) \tag{9}$$

where β_1' is the regression coefficient without using X_2 as a predictor in our model.

4.1.3 Effect of adding predictors on R^2

Adding a new predictor to a regression model can never decrease the coefficient of determination R^2. Let $R^2_{1\,\text{pred}}$ denote the R^2 from the single-predictor model

$$Y = \beta_1 X_1 + \epsilon',$$

and $R^2_{2\,\text{pred}}$ the R^2 from the two-predictor model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

By definition, $R^2 = 1 - \text{Var(residual)}/\text{Var}(Y)$, so R^2 increases whenever the residual variance decreases.

The key point is that the new error term ϵ' in the reduced model is *not* the same as $\beta_2 X_2 + \epsilon$. Instead, ϵ' absorbs the variation in Y that is correlated with X_2 but unaccounted for by X_1. Because X_1 and X_2 are generally correlated, part of the systematic variation explained by X_2 is treated as noise in the single-predictor model. Rather we have

$$\text{var}(\epsilon') = \text{var}(Y|X_1) = \text{var}(\beta_1 X_1 + \beta_2 X_2 + \epsilon|X_1) = \beta_1^2 \text{var}(X_2|X_1) + \sigma_\epsilon^2 > \sigma_{\epsilon^2} \tag{10}$$

hence

$$R^2_{2\,\text{pred}} \;\geq\; R^2_{1\,\text{pred}},$$

with equality only if X_2 is orthogonal to both X_1 and Y (i.e. it provides no additional explanatory power). Thus, adding predictors either improves the fit or leaves it unchanged, but never worsens it.

4.2 Covariance matrix and estimation of regression coefficients

4.2.1 Regression coefficients in terms of the covariance matrix

Here we will derive formulas for the regression coefficients in terms of covariances between the predictors, and covariance between the predictors and the response variable. This will (1) allow us to better understand

the relationship between single and multiple predictor regression coefficients and (2) lead to an estimator the regression coefficients in the multiple predictor case. In particular, we obtain formulas that generalize the relationship $\text{cov}(X, Y) = \beta_1 \sigma_x^2$, which we discovered to hold in the single predictors case.

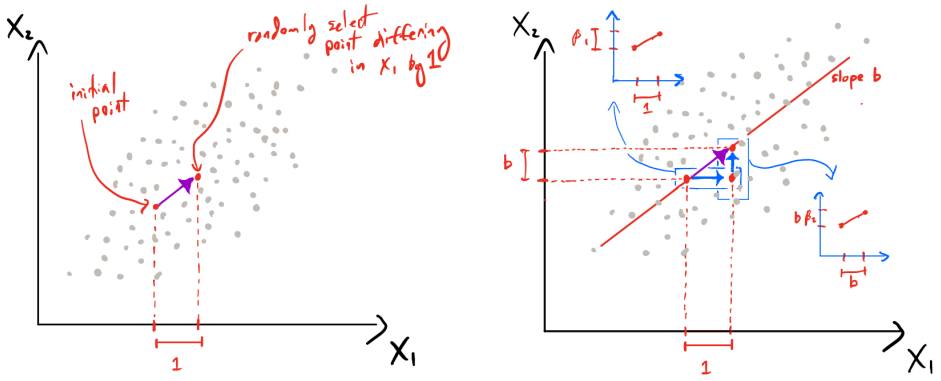

Figure 2: Here I'm illustrated the difference between the marginal regression slope (the slope of $E[Y|X_1]$ vs. X_1) and the regression coefficient β_1 in the two predictor model. Here I use b to denote the regression coefficient of X_1 with X_2 as a predictor. I use the notation of Example 4, although the idea applies more generally. When we increase x_1 by 1 without fixing X_2, then on average X_2 changes by b (which is the slope between x_1 and x_2 here, not the intercept.) Therefore, in order to relate this marginal slope to the regression slope β_1, subtract the increase in Y that is caused by the increase in X_2 (corresponding to the vertical blue arrow).

Consider a linear regression model with two predictors. The formulas we derive will generalize to many predictors. We will set $\beta_0 = E[X_1] = E[X_2] = 0$ for simplicity, since these cancels out in the end. An practice, I recommend checking that everything works out when this are not zero! We start by computing $\text{cov}(X_1, Y)$, which is simply $E[X_1 Y]$ since $E[X_1] = E[Y] = 0$. Just as we did for the single-predictor case in Unit 2, we write

$$\begin{aligned}\text{cov}(X_1, Y) &= E[X_1 Y] = E[X_1 E[Y|X_1]] \\ &= E[X_1(\beta_1 X_1 + \beta_2 X_2)] = \beta_1 E[X_1^2] + \beta_2 E[X_1 X_2] \\ &= \beta_1 \sigma_{x_1}^2 + \beta_2 \text{cov}(X_1, X_2)\end{aligned}$$

where we have used that, since $E[X_1] = E[X_2] = 0$, $\text{var}(X_1) = E[X_1^2]$ and $\text{cov}(X_1, X_2) = E[X_1 X_2]$. If we do the same for X_2, we get two equations

$$\begin{aligned}\text{cov}(X_1, Y) &= \beta_1 \sigma_{x_1}^2 + \beta_2 \text{cov}(X_1, X_2) \\ \text{cov}(X_2, Y) &= \beta_2 \sigma_{x_2}^2 + \beta_1 \text{cov}(X_1, X_2)\end{aligned}$$

Notice that as with the single-predictor case, it is useful to represent β_1 and β_2 as expectations which can be computed as averages over our data points. In addition to providing some insight into the meaning of the regression coefficients, this will yield candidates for our estimators of these quantities. This formulas can be rewritten in terms of a matrix, known as the covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_{x_1}^2 & \text{cov}(X_1, X_2) \\ \text{cov}(X_1, X_2) & \sigma_{x_2}^2 \end{bmatrix}. \tag{11}$$

We then have

$$\begin{bmatrix} \text{cov}(X_1, Y) \\ \text{cov}(X_2, Y) \end{bmatrix} = \Sigma \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \tag{12}$$

This is consistent with Eq. 9. In particular, by taking the first equation in this system and dividing by $\text{var}(X_1)$, we can rewrite Eq. 9 as

$$\beta_1' = \beta_1 + \beta_2\beta_{1,2} \tag{13}$$

where $\beta_{1,2} = \text{cov}(X_1, X_2)/\text{var}(X_1)$ is the (single-predictor) regression coefficient of X_1 with X_2 as the response variable. Eq. 12 can easily be generalized to many predictors. In the general case, the covariance matrix Σ is a $K \times K$ matrix with entries $\Sigma_{i,j} = \text{cov}(X_i, X_j)$.

Letting Σ^{-1} be the inverse of Σ, meaning

$$\Sigma^{-1}\Sigma = I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{14}$$

we then have

$$\begin{aligned} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} &= \Sigma^{-1} \begin{bmatrix} \text{cov}(X_1, Y) \\ \text{cov}(X_2, Y) \end{bmatrix} \\ &= \frac{1}{\sigma^2_{x_2}\sigma^2_{x_1} - \text{cov}(X_1, X_2)^2} \begin{bmatrix} \sigma^2_{x_2} & -\text{cov}(X_1, X_2) \\ -\text{cov}(X_1, X_2) & \sigma^2_{x_1} \end{bmatrix} \begin{bmatrix} \text{cov}(X_1, Y) \\ \text{cov}(X_2, Y) \end{bmatrix} \end{aligned}$$

You should notice the similarity between this formula and the single predictor formula $\beta_1' = \text{cov}(X, Y)/\text{var}(X)$. Here, Σ^{-1} is playing the role of $1/\text{var}(X)$, while the vector of βs and Y-X covariances play the roles of β_1 and $\text{cov}(X, Y)$ respectively. Notice that if $\Sigma = I$, then we can solve for each β_i separately and we obtain the same formulas as the single predictor case.

We don't usually worry about the closed form solutions for the β because they become impossibly complicated as the number of predictors grows. However, for two predictors we can derive a formula for β_1, which is

$$\beta_1 = \frac{\text{cov}(X_1, Y)\sigma^2_{x_2} - \text{cov}(X_2, Y)\text{cov}(X_1, X_2)}{\sigma^2_{x_2}\sigma^2_{x_1} - \text{cov}(X_1, X_2)^2} \tag{15}$$

The formula is particularly revealing if all the variances are set to one

$$\beta_1 = \frac{1}{1 - \rho^2_{1,2}}(\rho_1 - \rho_{1,2}\rho_2)$$

where $\rho_{1,2}$ is the correlation coefficient between X_1 and X_2. Notice that if X_1 and X_2 are uncorrelated ($\rho_{1,2} = 0$), we obtain the usual connection between the regression coefficient and the correlation coefficient between X_1 and X_2.

Example 4 (Correlated predictors)**.** Consider the model

$$\begin{aligned} X_1 &\sim \text{Normal}(0, 1) \\ X_2|X_1 &\sim \text{Normal}(bX_1, 1 - b^2) \\ Y|(X_1, X_2) &\sim \text{Normal}(\beta_1 X_1 + \beta_2 X_2, \sigma^2). \end{aligned}$$

where $b \in [0, 1]$. Note that b is just $\beta_{1,2}$ – the regression coefficient of X_1 on X_2, which in this case is the same as the correlation coefficient (see below).

Question:

(a) Show that $\text{var}(X_1) = \text{var}(X_2) = 1$ and $\text{cov}(X_1, X_2) = b$

(b) Expression β_1 as a function of b.

(c) Obtain the single predictor regression coefficient β_1' when only X_1 is the predictor.

Solution:

(a) By definition of the model $\text{var}(X_1) = 1$ and

$$\text{var}(X_2) = b^2\text{var}(X_1) + 1 - b^2 = b^2 + 1 - b^2 = 1$$
$$\text{cov}(X_1, X_2) = b\text{var}(X_1) = b$$

(b) We can write Equation 15 as

$$\beta_1 = \frac{\text{cov}(X_1, Y) - \text{cov}(X_2, Y)b}{1 - b^2}$$

(c) The single-predictor slope of Y on X_1 is $\beta_1 + b\beta_2$.

4.2.2 Simpson's paradox

In the example above we can see that β_1' and β_1 can have different signs depending on b. This effect is called Simpson's "paradox". This is not really a paradox, but a simple consequence of the fact that the correlation between predictors influences the single-predictor regression coefficient as illustrated in Figure 2. Let's look at a another example.

Example 5. Consider two binary predictors $X_1, X_2 \in \{0, 1\}$ with joint distribution given by the table of probabilities:

$P(X_1, X_2)$	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	0.4	0.1
$X_1 = 1$	0.1	0.4

Suppose the outcome Y satisfies the linear model

$$Y \mid (X_1, X_2) = X_1 - 2X_2 + \epsilon$$

Question: Compute the single-predictor regression coefficient β_1' of X_1 on Y.
Solution: First compute the difference in conditional means of X_2 across X_1:

$$\mathbb{P}(X_2 = 1 \mid X_1 = 1) = \frac{0.4}{0.1 + 0.4} = 0.8, \qquad \mathbb{P}(X_2 = 1 \mid X_1 = 0) = \frac{0.1}{0.4 + 0.1} = 0.2,$$
$$\beta_{1,2} \equiv \mathbb{E}[X_2 \mid X_1 = 1] - \mathbb{E}[X_2 \mid X_1 = 0] = 0.8 - 0.2 = 0.6$$

With the true model $Y = X_1 + (-2)X_2 + \varepsilon$ (so $\beta_1 = 1,\ \beta_2 = -2$), the single-predictor slope satisfies

$$\beta_1' \ = \ \beta_1 + \beta_2\beta_{1,2}.$$

Substituting $\Delta_{2|1} = 0.6$ yields

$$\beta_1' \ = \ 1 + (-2) \cdot 0.6 \ = \ 1 - 1.2 \ = \ -0.2.$$

4.2.3 Estimating regression coefficients

Now suppose we have data points $\{(X_{i,1}, \ldots, X_{i,K}, Y_i)\}_{i=1}^N$. We form the design matrix with *samples in rows* and *predictors in columns*:

$$X = \begin{bmatrix} X_{1,1} & \cdots & X_{1,K} \\ \vdots & \ddots & \vdots \\ X_{N,1} & \cdots & X_{N,K} \end{bmatrix} \in \mathbb{R}^{N \times K}, \qquad Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix} \in \mathbb{R}^N. \tag{16}$$

Assume X and Y are centered (no intercept). The empirical covariance quantities are

$$\hat{\Sigma} = \frac{1}{N} X^\top X \in \mathbb{R}^{K \times K}, \qquad \hat{c} = \frac{1}{N} X^\top Y \in \mathbb{R}^K. \tag{17}$$

When $\hat{\Sigma}$ is invertible,

$$\hat{\beta} = \hat{\Sigma}^{-1} \hat{c} = (X^\top X)^{-1} X^\top Y. \tag{18}$$

Moore–Penrose pseudoinverse. If $\hat{\Sigma}$ is singular (e.g. collinearity or $K > N$), use the Moore–Penrose pseudoinverse:

$$\hat{\beta} = X^+ Y, \qquad X^+ = V \Sigma^+ U^\top \quad \text{for } X = U \Sigma V^\top \text{ (SVD)}, \tag{19}$$

where Σ^+ inverts nonzero singular values and leaves zeros unchanged.

4.3 Sample distribution and collinearity

Just as before, we want to understand what the sample distribution of the coefficients looks like. In the multiple predictor case, we need to think about the joint distribution of $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K)$. We will start by focusing on the two predictor case.

Example 6 (Predictor sample distribution)**.** Consider the model in 4. Let's look at the sample distribution by fitting many simulated replicates.

Question: Write a function to generate a dataframe containing samples from the sample distribution of $(\hat{\beta}_1, \hat{\beta}_2)$. Make a scatter plot and explore the structure of the sample distribution, in particular it dependence on b, which controls the correlations between X_1 and X_2.

Solution: See colab notebook

4.3.1 The sample distribution

To understand more intuitively what is going on, imagine X_1 and X_2 are very highly correlated (if they are perfectly correlated we say they are colinear). We can then write

$$\begin{aligned} Y &= \beta_1 X_1 + \beta_2 X_2 + \epsilon \approx \beta_1 X_1 + \beta_2 X_1 + \epsilon \\ &\approx (\beta_1 + \beta_2) X_1 + \epsilon \end{aligned}$$

There are many ways to select β_1 and β_2 so that the surface $\beta_1 x_1 + a_2 \beta_2$ is close to the lines, since a change in β_1 can be compensated by a change in β_2. This means that **if we estimate β_1 and β_2 and then generated new data, it would be possible to get a VERY different value of $\hat{\beta}_1$ and $\hat{\beta}_2$, so long as $\hat{\beta}_1 + \hat{\beta}_2$ is close to what we got before**. This is illustrated in Figure 3 and Figure 4.

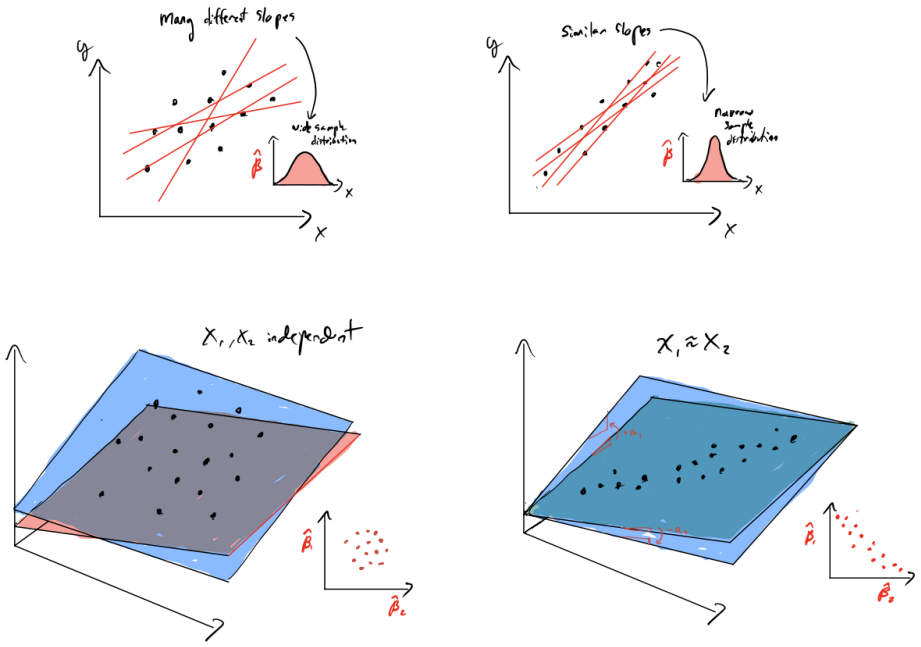

Figure 3: (top) In the single-predictor case, the width of the sample distribution measures how confident we are of a particular slope. It will be narrow if a replicate of our data is likely to produce a very similar slope. These means we get a rough idea of the width of sample distribution by seeing much we can change our regression line and still obtain something that appears to pass through our data. (bottom) In the two predictor case, we have a regression plane and changing β_1 and β_2 will "wiggle" the plane by tilting it in the x_1 and x_2 directions (there is also the intercept which can shift the plane up and down, but I'm not illustrating that). If X_1 and X_2 are uncorrelated, it doesn't matter which way we wiggle it, the fit will be similar, but if X_1 and X_2 are strongly correlated, wiggling the plane in the direction perpendicular to the points has a much smaller effect that parallel to them.

4.3.2 Sample distribution formula

It can be shown that (assuming the covariance matrix Σ is know),

$$\hat{\beta} \sim \text{MvNormal}(\sigma_\epsilon^2 \Sigma^{-1}/N) \tag{20}$$

In particular, if $\Sigma = I$ we obtain the sample distribution for the single-predictor case.

4.4 Dealing with categorical data

One situation in which models with multiple predictors frequently arrises is when trying to predict a Y variable based on categorical predictors, such as race. In this case, we need to transform the categories into numerical values. For example, if there are two categories (e.g. YES and NO) we map our variable to 0 or 1. If we have 3 categories (e.g. White, Black, Other), we might first think to map them to 0, 1 and 2. This has a problem though: A chancge from 1 to 2 should not necessarily correspond to a change from 0 to 1. In other words, **there is no clear ordering of the x values**. Sometimes we refer to such predictors and qualitative rather than quantitive, since they express a quality of our data points instead of a numerical quantity. To address this issue, we create dummy variables. In particular, In order to take a categorical variable and transform it into a set of indicator variables in python, we use the python function `get_dummies`. The usage of this is illustrated in the following example.

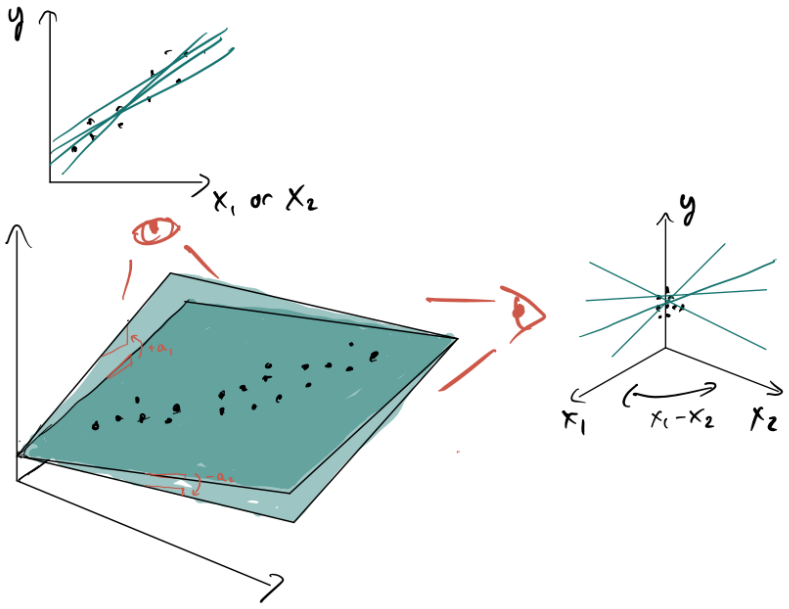

Figure 4: Different views of the data in the case when X_1 and X_2 are correlated. If we look at the data from the side, or along the $X_1 = X_2$ direction, then all our regression planes appear similar; however, when looked at from the "front" as shown in the right panel, we see that the places actually have very different slopes in the other direction.

Example 7 (Racial disparities in earnings)**.** Here we will fit the earnings data to a model with race as a predictor. In particular, we want to know: What is the association between race and earnings among adults in the US? We will start with a model using only race as a predictor. One way to approach this would be to simply use a binary predictor and consider only 2 race categories (e.g. White and non-White). This is limiting though. Instead, we can create a variable for each rate category we are interested in. In the dataset there are 4 race categories (not sure why these 4, but that's what we'll work with)

$$\{\text{Black, White, Hispanic, Other}\}$$

In principle, we could create a binary variable for each one (these are what we call dummy variable), to obtain a model like

$$Y = \beta_0 + \beta_{\text{black}} X_{\text{black}} + \beta_{\text{hispanic}} X_{\text{hispanic}} + \beta_{\text{other}} X_{\text{other}} + \beta_{\text{white}} X_{\text{white}} + \epsilon$$

This is problematic though, since at least one of the predictors above MUST be 1. This means that the first 3 of the predictors are perfectly correlated with the other one. By the default, python will drop the first predictor (in alphabetical order), leaving us with the model

$$Y = \beta_0 + \beta_{\text{hispanic}} X_{\text{hispanic}} + \beta_{\text{other}} X_{\text{other}} + \beta_{\text{white}} X_{\text{white}} + \epsilon.$$

Question: Fit the data to the model above. What is the expected disparity in earnings between someone who is white and someone who is hispanic.

Solution: See colab notebook. To answer the question posed above, we begin with the interpretations of the regression coefficients. In terms of conditional expectation, these are

$$\begin{aligned}
\beta_{\text{white}} &= E[Y|X_{\text{white}} = 1, X_{\text{hispanic}} = X_{\text{other}} = 0] - E[Y|X_{\text{white}} = 0, X_{\text{hispanic}} = X_{\text{other}} = 0] \\
&= E[Y|\text{someone is white}] - E[Y|\text{someone is black}] \approx 4.9 \\
\beta_{\text{hispanic}} &= E[Y|X_{\text{hispanic}} = 1, X_{\text{white}} = X_{\text{other}} = 0] - E[Y|X_{\text{hispanic}} = 0, X_{\text{white}} = X_{\text{other}} = 0] \\
&= E[Y|\text{someone is hispanic}] - E[Y|\text{someone is black}] \approx -0.7
\end{aligned}$$

Our goal however is to compute

$$\begin{aligned}
&E[Y|\text{someone is white}] - E[Y|\text{someone is hispanic}] \\
&= E[Y|X_{\text{white}} = 1, X_{\text{hispanic}} = 0, X_{\text{other}} = 0] - E[Y|X_{\text{white}} = 0, X_{\text{hispanic}} = 1, X_{\text{other}} = 0] \\
&= \beta_0 + \beta_{\text{white}} - \beta_0 - \beta_{\text{hispanic}} \\
&= \beta_{\text{white}} - \beta_{\text{hispanic}}
\end{aligned}$$

Exercises

Exercise 1 (A binary and normal predictor ❑)**:** Consider the a linear regression model

$$Y|(X_1, X_2) \sim \text{Normal}(\beta_0 + \beta_1 X_1 + \beta_2 X_2, \sigma^2)$$

where the two predictors obey

$$X_1 \sim \text{Bernoulli}(q)$$
$$X_2|X_1 \sim \text{Normal}(bX_1, s^2)$$

(a) Can you think are at least two examples where this would be a reasonable model of the relationship between 3 variables X_1, X_2 and Y?

(b) What are the formulas for $\text{cov}(X_1, X_2)$ and $\text{var}(X_2)$ in terms of the model parameters $q, b, s, \beta_0, \beta_1, \beta_2, \sigma^2$? You should be able to derive these formulas, but you may also reference formulas in previous exercises and class notes. You may assume $\beta_0 = 0$ for this part and the reminder of the exercise, as this simplifies some calculations and doesn't change the results.

(c) Derive a formula for $\text{cov}(Y, X_1)$ in terms of β_1, q, β_2 and b.

(d) Explain how the formula you derived in part (b) is related to the equation for $\text{cov}(Y, X_1)$ in the single predictor regression model (page 4 on week 3 notes). In particular, for what parameter values do the two formulas coincide? Your conclusion will be a particular case of what we saw to be true more generally in class concerning the relationship between β_1 and the covariances in a regression model with two predictions.

(e) Now derive the formula

$$\text{var}(Y) = q(1-q)\left(\beta_1^2 + \beta_2^2 b^2 + 2\beta_1\beta_2 b\right) + \beta_2^2 s^2 + \sigma^2$$

You will need to use the formula for the variance of the sum of two (not-necessarily independent) random variables, which is given on the midterm practice problems. This is also in the "addition and multiplication section" on the wikipedia page.

(f) The calculations in part (c) allows us to solve an exercise in Chapter 8 in Demidenko's textbook [2], albeit in the more restrictive context of a binary and normal predictor: Is it possible that β_1 and β_2 are **both negative**, yet the (marginal) slope of Y vs. X_1 is **positive**? If so, generate simulated data where this is the case.

Exercise 2 (Earnings data)**:** Consider the earnings data. This can be loaded with

```
> df = pd.read_csv("https://raw.githubusercontent.com/avehtari
> /ROS-Examples/master/Earnings/data/earnings.csv")
```

As in the previous exercise set, you will study the association between earnings and gender, but now using regression with multiple predictors.

(a) Perform a linear regression using `statsmodels` with gender and height as predictors.

(b) Provide interpretations for each regression coefficient (like we did in class for the test score example).

(c) Which factor, height or gender is more important based on your analysis?

(d) Based one the fitted model, predict the chance that someone who is not male and is 5.8ft earns more than a male who is the same height? To get a sense for the importance (or lack-thereof) of the height predictor, compare this to the chance that a male earns more than a non-male (regardless of height).

Exercise 3 (Sample distribution ❑)**:** In the notebook from class, we wrote code to generate samples from the sample distribution of $(\hat{\beta}_1, \hat{\beta}_2)$ in the model

$$\begin{aligned} X_1 &\sim \text{Normal}(0, 1). \\ X_2|X_1 &\sim \text{Normal}(bX_1, 1 - b^2) \\ Y|(X_1, X_2) &\sim \text{Normal}(\beta_1 X_1 + \beta_2 X_2, \sigma^2) \end{aligned}$$

Specifically, we had a function which takes β_1, β_2 and β_0 as inputs and returns a dataframe where the columns are the samples of $\hat{\beta}_1$ and $\hat{\beta}_2$ respectively. When we plotted the correlation coefficient as a function of b values and estimates the correlation coefficient between $\hat{\beta}_1$ and $\hat{\beta}_2$, it was a decreasing line.

(a) What would happen if instead of plotting the correlation coefficient, we plotted se$(\hat{\beta}_1)$ as a function of b? Would it increase? decrease? neither? Note that both X_1 and X_2 are standardized, so the distribution of X_1 values is not changed when we adjust b. In answering this question, you can either give a geometric intuition, or do a calculation. You should check your answer with simulations, but you still need to provide a detailed explanation.

(b) Is it possible to have large standard errors on all the $\hat{\beta}_i$ values (measured relative to the true values of course), but still have a large (meaning close to one) value of R^2? If so, for what parameter values does this happen? Run simulation(s) to support your answer.

Exercise 4 (❑)**:** Suppose we have a large amount of data from the model

$$X_1 \sim \text{Normal}(0, 1) \tag{21}$$
$$X_2|X_1 \sim \text{Normal}(3X_1, 1) \tag{22}$$
$$Y|(X_1, X_2) \sim \text{Normal}(X_1 - 2X_2, 1) \tag{23}$$

if a single-predictor linear regression is performed with Y as the response variable and ONLY X_2 as the predictor, what is the regression coefficient β_2'?

Exercise 5 (❑)**:** Consider two binary predictors $X_1, X_2 \in \{0, 1\}$ with joint distribution given by the table of probabilities:

$P(X_1, X_2)$	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	0.4	0.1
$X_1 = 1$	0.1	0.4

Suppose the outcome Y satisfies the linear model

$$Y \mid (X_1, X_2) = X_1 + cX_2 + \epsilon$$

(a) Compute the single-predictor regression coefficient β_1' (in terms of c)

(b) Compute the single-predictor regression coefficient β_2' (in terms of c)

(c) For what values of c does the model exhibit Simpson's paradox?

Exercise 6 (❑)**:** Let $X = (X_1, X_2)^T$ be a bivariate normal vector with mean zero and covariance matrix

$$\Sigma = \begin{bmatrix} 10 & 2 \\ 2 & 10 \end{bmatrix}.$$

Suppose the response satisfies

$$Y = 0.5X_1 - 1.5X_2 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1).$$

(a) Compute the single-predictor regression coefficient β_1' for Y on X_1.

(c) Comment on whether Simpson's paradox occurs in this setup.

Exercise 7 (❑)**:** A sports scientist is studying the relationship between athletes' endurance performance and two predictors: **weekly training hours** (X_1) and **VO2 max** (X_2, in ml/kg/min), measuring their effect on a 10-km run time Y (in minutes). Data from 100 athletes are collected.

A single predictor linear regression model of 10-km time Y on weekly training hours X_1 alone yields:

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.8884	0.285	3.114	0.002	0.322	1.454
x1	1.8436	0.284	6.487	0.000	1.280	2.408

A second predictor, VO2 max (X_2), is included in the model, giving:

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.0170	0.011	94.148	0.000	0.996	1.038
x1	1.0027	0.011	89.354	0.000	0.980	1.025
x2	3.0020	0.011	261.492	0.000	2.979	3.025

What is the regression coefficient $\beta_{1,2}$ of X_1 with X_2 as the response variable?

Exercise 8 (❑)**:** A researcher creates a linear model with two predictors, X_1 and X_2, but the dataset available only contains X_1.

- The **true** (long) model describing the data is:

$$Y = 1.0 + 2.0X_1 + 4.0X_2 + \epsilon_1$$

- The (unobserved) relationship between the predictors is given by:

$$X_2 = 0.5 + 0.5X_1 + \epsilon_2$$

Assume $E[\epsilon_1] = 0$ and $E[\epsilon_2] = 0$.

The researcher runs a regression using only X_1:

$$Y = \beta_0^+ \hat{\beta}_1 X_1 + \epsilon_1'$$

Question: What is the expected value (approximately) for the coefficient $\hat{\beta}_1$ that the researcher will estimate when they run this regression?

References

[1] Andrew Gelman, Jennifer Hill, and Aki Vehtari. *Regression and other stories*. Cambridge University Press, 2020.

[2] Eugene Demidenko. *Advanced statistics with applications in R*, volume 392. John Wiley & Sons, 2019.