More on the Multiple Linear Model, and Bias and Standard Error of Ordinary Least Squares

Stats 203, Winter 2024

Lecture 4 [Last update: February 3, 2024]

1 Review and preview

- In our last lecture, we introduced the multiple linear model. We also talked about the ordinary least squares (OLS) estimates for the parameters of the multiple linear model, derived by the method of least squares:

$$\hat{\beta} := \operatorname*{argmin}_b \mathrm{MSE}_n(b) := \sum_{i=1}^{n} \Big(y_i - (b_0 + \sum_{j=1}^{p} b_j x_{ij}) \Big)^2$$

  We showed that the OLS estimates had the fairly simple closed-form solution

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top y,$$

  and that they could be used to make predictions at a new value of $x = (x_1, \ldots, x_p)$ using the least-squares line (aka line of best fit):

$$\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p.$$

  The lecture notes from last time discuss a quantity (R^2) for measuring the overall accuracy of the fitted model. I've decided to delay discussing this until a bit later in the quarter, when we talk about hypothesis tests for goodness-of-fit.

- Today we will turn to measuring the accuracy of each $\hat{\beta}_j$ as an estimate of the corresponding parameter β_j. Recall from Lecture 2 that the way we measured accuracy of an estimator was through its bias and standard error. So the bulk of today's class will be about working out the bias and standard error of the OLS estimator.

- As a (partial) spoiler: we will see that OLS is unbiased for β under the assumptions of the linear model. Thus we have our usual "rule of thumb" for unbiased estimators (see Lecture 2 notes), which goes something like this: "β_j is roughly equal to $\hat{\beta}_j$, give or take a standard error." This is an (informal) example of **inference**, that is, using the data to give a range of plausible values for unknown parameters. In a (not so distant) future lecture, we will make this rule of thumb precise using confidence intervals.

2 More on the linear model: what do the βs mean?

- Before we dive into the math, however, let's pause and think a little bit harder about what the βs in the linear model do and do not mean. (This discussion expands upon the corresponding section in the Lecture 3 notes, which we did not get to last Wednesday.)

- **Correlation does not imply causation.** An appealingly simple interpretation is to say that β_j represents the expected change in the response if the jth predictor were increased by one unit, with all other predictors held fixed. Unfortunately this is dead wrong. Consider the regression equation $GradeLevel = \beta_0 + \beta_1 ShoeSize + \epsilon$. It is intuitively clear that grade level and shoe size are positively

correlated – hence, β_1 should be positive – but if you went out and bought a larger pair of shoes, you would not expect your grade level to increase.

This mistake is so common that is has gained popular awareness through the expression "correlation doesn't imply causation." There is an entire sub-field of statistics, known as causal inference, that deals with what *does* imply causation. Many of the tools of causal inference build on (linear) regression, but they are not the same as (linear) regression. We will not have time to cover causal inference in this class.

- **Marginal correlation is different than conditional correlation.** Another common mistake is to confuse marginal and conditional correlation. This is best illustrated through the following hypothetical. Suppose that we were trying to use income level to predict likelihood of voting Republican, among voters in California and Texas. It is possible that in each of the two states; income is positively correlated with voting Republican, yet when the two states are considered together, the correlation flips! This seeming contradiction is an example of **Simpson's paradox**.

  When you first see it, Simpson' paradox can be surprising, but it is not actually a paradox. In the language of regression modeling, it simply means that it is possible for both of the following statements to be true:

$$\mathbb{E}[Y|X_1 = x_1] = \gamma_0 + \gamma_1 x_1, \quad \text{with} \quad \gamma_1 > 0$$
$$\mathbb{E}[Y|X_1 = x_1, X_2 = x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \quad \text{with} \quad \beta_1 < 0.$$

  In the previous example X_1 would represent income, X_2 would be a dummy variable representing state (so $X_2 = 0$ represents living in California, $X_2 = 1$ represents living in Texas), and Y would be another dummy variable representing vote (so $Y = 1$ represents voting Republican.)

  The general principle here is that each β_j in the multiple linear model represents the linear relationship between Y and X_j *after* conditioning on the other predictors. Whether or not this is the quantity you care about will depend on the context. NB: typically, language like "adjusting for..." or "accounting for..." is used instead of the more mathematical "conditional on...".

  (PS: The above hypothetical is related to, though distinct from, a well-known fact among political scientists, which is that lower-income voters tend to lean Democratic, but lower-income states are more likely to vote Republican. Some political commentators in the early aughts used the latter correlation to declare that the Republican Party was the party of the working class. Analyzing data after aggregating unit-level information is called **ecological fallacy**, though it is no more a fallacy than Simpson's paradox is a paradox. For more information see Gelman et al. (2005).)

- **A correct interpretation.** The correct way to interpret the β_js comes from the mathematical fact that if (X, Y) follow the linear model, then

$$\beta_j = \mathbb{E}[Y|X_1 = x_1, \ldots, X_j = x_j + 1, X_{j+1} = x_{j+1}, \ldots] - \mathbb{E}[Y|X_1 = x_1, \ldots, X_j = x_j, X_{j+1} = x_{j+1}, \ldots].$$

  Translating math to English, we would say that "β_j is the expected difference in the response, between individuals in the population whose value of X_j differs by 1, and whose values of X_k are equal for all $k \neq j$." Perhaps the only redeeming quality of this dense, awkward phrasing is that it is correct.

3 More on the linear model: why do we care about inferring the βs?

- Now that we know what the βs mean, the next question is: why do we care about inferring them? Said differently, why don't we simply content ourselves with calculating the $\hat{\beta}$s? After all the $\hat{\beta}$s suffice for making predictions. A common answer is that this is just what we *do* in statistics – that is, we write down a model, fit the model using some data, and then try to draw inferences about true values of unknown parameters – but of course that completely begs the question. Let me try to give a better answer through a motivating example.

- **Motivating example: smoking and lung cancer.** All of you (hopefully!) know that smoking cigarettes increases your chances of getting lung cancer. But in the first half of the 20th century, things were not quite so clear. In fact it took an incredible effort of data collection and analysis to convince the general public and public officials (in the US and elsewhere) to regulate cigarette smoking. Many of these analyses were, in one form or another, multiple linear regressions.
- The basic data: dramatic increases in the rates of cigarette smoking, and concurrently, of lung cancer, in the years 1900 - 1959. In the language of simple linear regression, rates of cigarette smoking (the predictor, x) and lung cancer (the response, y) were seen to be highly correlated.
- Nevertheless, there were skeptics. Two common objections of the skeptics were:
  - How do we know the correlation is "a fact, and not a spurious result..."?
  - How do we know the correlation is not due to "numerous other possible variables..." such as age, sex, urban-rural differences, etc.?[1]
- The purpose of linear modeling and inference is to rule out (some of) these kinds of objections, by allowing us to
  - **Generalize**: By making statements about parameters instead of estimates, we can draw conclusions about the entire population rather than the particular sample we're looking at.
  - **Represent confidence**: Reporting ranges of plausible values allows us to have a certain level of confidence in our answers, instead of just reporting a single number which will inevitably be wrong. (I am not yet giving you a definition of what "confidence" means. That will come later, when we talk about confidence intervals.)
  - **Account for other variables**: Including multiple variables in the model allows us to estimate the relationship between a given predictor and response after accounting for other factors.
- In the context of the smoking and lung cancer example, appropriate linear modeling and inference justifies making claims like "we are 90% confident that there is a true association between cigarette smoking and lung cancer among US adults, after accounting for age and sex," by collecting data on rates of lung cancer and smoking, along with other variables such as age, sex, and then reporting inferences from the linear regression. (In fact, it turns out that the data as of 1959 justified making claims like "we are 99.999% confident that there is a true association between cigarette smoking and lung cancer, after accounting for age, sex, and many other variables" which is what it took to override the objections of the tobacco industry.)
- To make some (pretty obvious) observations:
  - Simply writing down a linear model doesn't mean its assumptions are correct. Much of the hard work of statistics involves collecting data in accordance with the assumptions of the linear model – for instance, by drawing a simple random sample from a population, or taking repeated measurements of a quantity over multiple days – and then checking and rechecking the assumptions after the data analysis is performed.
  - Even if the model is plausible, and inferences are valid, we still need to be careful about interpretation; see the previous section. For instance, just doing a regression is not enough to establish that smoking causes cancer.
- PS: the previous discussion is drawn from the paper Cornfield et al. (1959). I have skipped over many of the relevant details. The example turns out to be a proud moment for statistics but not for statisticians – many statisticians of the first rank continued disputing the link between cigarette smoking and lung cancer long after there was any serious scientific basis for doing so. An (incredibly detailed and damning) historiography is the book "Golden Holocaust" by Robert Proctor.

[1] Quotations are taken from the paper Cornfield et al. (1959).

Also, a personal opinion: classes such as this one inevitably spend a long time dwelling on the assumptions of the linear model, how they are often unrealistic, how violations invalidate inferential conclusions, how parameters don't what you wan't them mean, etc. This gives the (not totally incorrect) impression that statistics has a pessimistic and even skeptical bent. In many cases, this skepticism has served the broader scientific community well, by guarding against pseudo-scientific, unreplicable, and otherwise unwarranted claims. This example turns out to be the (rare?) case where the skepticism of statisticians probably hurt more than it helped.

4 Probability review: random vectors

- Ok, now back to the math. It turns out to be much easier to derive bias and standard error by working with the entire vector $\hat{\beta} \in \mathbb{R}^{p+1}$, rather than dealing with each $\hat{\beta}_j$ in turn. To do this, we need to remember what the definitions of expectation and variance of a random vector are.
- A random vector $U \in \mathbb{R}^p$ is a p-tuple of random variables:

$$U = \begin{bmatrix} U_1 \\ \vdots \\ U_p \end{bmatrix}$$

The expectation of U is a length-p vector:

$$\mathbb{E}[U] = \begin{bmatrix} \mathbb{E}[U_1] \\ \vdots \\ \mathbb{E}[U_p] \end{bmatrix}$$

The covariance matrix of U is a $p \times p$ matrix:

$$\text{Cov}[U] = \begin{bmatrix} \text{Cov}[U_1, U_1] & \text{Cov}[U_1, U_2] & \dots & \text{Cov}[U_1, U_p] \\ \text{Cov}[U_2, U_1] & \text{Cov}[U_2, U_2] & \dots & \text{Cov}[U_2, U_p] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[U_p, U_1] & \text{Cov}[U_p, U_2] & \dots & \text{Cov}[U_p, U_p] \end{bmatrix}$$

Notice that $\text{Cov}[U]$ is symmetric.

- *Properties of random vectors.* Let U and V be random vectors, a and b be fixed vectors, and $\mathbf{A}$ be a matrix. The following properties are all more or less straightforward generalizations of properties of random variables.
  - **Linearity of expectation**. $\mathbb{E}[a^\top U + b^\top V] = a^\top \mathbb{E}[U] + b^\top \mathbb{E}[V]$.
  - **Linearity of conditional expectation** .$\mathbb{E}[f(X) \cdot U + g(X) \cdot V | X] = f(X)\mathbb{E}[U|X] + g(X)\mathbb{E}[V|X]$.
  - **Variance is quadratic**. $\text{Cov}[v^\top U] = v^\top \text{Cov}[U] v$. $\text{Cov}[\mathbf{A}U] = \mathbf{A}\text{Cov}[U]\mathbf{A}^\top$.
  - **Independence implies zero-covariance**. U and V are independent if

$$\mathbb{P}(U \in A, V \in B) = \mathbb{P}(U \in A)\mathbb{P}(V \in B).$$

    If U and V are independent random vectors, then $\text{Cov}(f(U), g(V)) = 0$ for all functions $f, g \in \mathbb{R}^p \to \mathbb{R}$, and $\mathbb{E}[g(V)|U] = \mathbb{E}[g(V)]$.

5 Bias and standard error of OLS

- Recall the multiple linear model from last class: random variables $(X_1, Y_1), \dots, (X_n, Y_n)$ are distributed independently, with $X_i \in \mathbb{R}^p$ coming from an arbitrary distribution P_i,

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i, i = 1, \dots, n \tag{1}$$

and $\epsilon_i \perp\!\!\!\perp X_i, \mathbb{E}[\epsilon_i] = 0$ and $\mathrm{Var}[\epsilon_i] = \sigma^2$. This can be more concisely written using matrices and vectors. Let

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Then (1) can be rewritten as

$$Y = \mathbf{X}\beta + \epsilon, \tag{2}$$

where $\epsilon \perp\!\!\!\perp \mathbf{X}, \mathbb{E}[\epsilon] = 0, \mathrm{Cov}[\epsilon] = \sigma^2 \mathbf{I}_n$. (A good exercise in checking your understanding is to work out why the linear model assumptions imply that $\mathrm{Cov}[\epsilon] = \sigma^2 \mathbf{I}_n$. [Update: corrected the subscript of $\mathbf{I}_n$ to be n not p.])

- At this point you might notice that there's some notational overload – does $\mathbf{X}$ refer to the matrix of observed predictors, or predictors as random variables? It turns out not to matter too much, since as in the case of the simple linear model, we are going to condition on the X_{ij}s when computing bias and standard error. But, as always, be aware of whether you're talking about data or random variables, estimators or estimates, etc.

- Now to compute the bias and standard error, starting with the bias which, as usual, is the easier calculation:

$$\begin{aligned} \mathbb{E}[\hat{\beta}|X_1, \dots, X_n] &= \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top Y | X_1, \dots, X_n] \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[Y|X_1, \dots, X_n] \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[\mathbf{X}\beta + \epsilon | X_1, \dots, X_n] \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}\beta \\ &= \beta. \end{aligned}$$

  Going line by line: the first equality follows by definition of $\hat{\beta}$; the second equality used linearity of conditional expectation, along with the fact that the matrix[2] $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$ is fixed given $X_1, \dots, X_n$; the third equality just plugs in the regression equation (2) for Y; the fourth equality uses the fact that $\mathbf{X}\beta$ is fixed given $X_1, \dots, X_n$, while $\mathbb{E}[\epsilon | X_1, \dots, X_n] = \mathbb{E}[\epsilon] = 0$; and finally, the fifth line follows from the definition of an inverse.

  We conclude that $\mathbb{E}[\hat{\beta}_j | X_1, \dots, X_n] = \beta_j, j = 0, \dots, p$, and thus OLS is **unbiased**[3] for the true parameters under the linear model.

  Now for the covariance matrix, which is (thankfully) easier than you might have guessed. To start, remember the linear algebra facts $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$ and $(\mathbf{A}^{-1})^\top = (\mathbf{A}^\top)^{-1}$. (The latter, of course, assumes that $\mathbf{A}$ is square and invertible). From these facts we can conclude that

$$\left((\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top\right)^\top = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}.$$

[2] This matrix is called the Moore-Penrose pseudoinverse of $\mathbf{X}$.
[3] Both conditional on $X_1, \dots, X_n$ and unconditionally, by the law of total expectation.

Along with the properties of covariance of random vectors mentioned above, this gives us what we need:

$$\begin{aligned}
\text{Cov}[\hat{\beta}|X_1,\ldots,X_n] &= \text{Cov}[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y|X_1,\ldots,X_n] \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\text{Cov}[Y|X_1,\ldots,X_n]\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\text{Cov}[\mathbf{X}\beta+\epsilon|X_1,\ldots,X_n]\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\text{Cov}[\epsilon|X_1,\ldots,X_n]\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\sigma^2\mathbf{I}_p)\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= \frac{\sigma^2}{n}(\frac{1}{n}\mathbf{X}^\top\mathbf{X})^{-1}.
\end{aligned}$$

Going line by line: the first equality follows by definition of $\hat{\beta}$; the second equality uses that variance is quadratic; the third equality plugs in the regression equation (2) for Y; the fourth equality uses the fact that $\mathbf{X}\beta$ is fixed given $X_1,\ldots,X_n$; the fifth equality uses that $\text{Cov}[\epsilon|X_1,\ldots,X_n] = \text{Cov}[\epsilon] = \sigma^2\mathbf{I}_p$; and finally, the sixth equality follows from the definition of an inverse. In the last line, we simply multiply and divide by the sample size n to put $\mathbf{X}^\top\mathbf{X}$ onto the right scale; this will be made clearer momentarily.

- From here we can read off $\text{SE}[\hat{\beta}_j|X_1,\ldots,X_n]$:

$$\text{SE}[\hat{\beta}_j|X_1,\ldots,X_n] = \sqrt{\text{Var}[\hat{\beta}_j|X_1,\ldots,X_n]} = \sqrt{\frac{\sigma^2}{n}[(\frac{1}{n}\mathbf{X}^\top\mathbf{X})^{-1}]_{jj}} = \frac{\sigma}{\sqrt{n}}\sqrt{[(\frac{1}{n}\mathbf{X}^\top\mathbf{X})^{-1}]_{(j+1),(j+1)}}.$$

There was a mistake in an earlier version of these notes. Notice that the $j+1$st diagonal element gives the standard error for the $\hat{\beta}_j$. This is because we index β starting from 0, and the elements of a matrix starting from 1.

- To both sanity check our calculations, and get some intuition about what the matrix $(\mathbf{X}^\top\mathbf{X})^{-1}$ represents, let's look at the case where $p = 1$, which is just simple linear regression. In this case

$$\mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \Longrightarrow \frac{1}{n}\mathbf{X}^\top\mathbf{X} = \begin{bmatrix} 1 & \bar{X} \\ \bar{X} & \overline{X^2} \end{bmatrix} = \begin{bmatrix} 1 & \bar{X} \\ \bar{X} & s_X^2 + \bar{X}^2 \end{bmatrix},$$

where $\overline{X^2} = \frac{1}{n}\sum_{i=1}^n X_i^2$ is the sample mean of the X_i^2s, and the last equality just says that sample variance = sample mean of X_i^2 − sample mean of X_i, and can be confirmed by a bit of algebra (which you should do if you don't believe me).

(Now it should be clearer why we divided by n: if each $X_i \sim P$ comes from the same distribution, then the entries of $\frac{1}{n}\mathbf{X}^\top\mathbf{X}$ will converge to fixed numbers, $\bar{X} \to \mathbb{E}[X], s_X^2 \to \text{Var}[X]$ as $n \to \infty$, by the law of large numbers.)

Now, using the definition of the inverse of a 2×2 matrix, we have that:

$$\text{Cov}[\hat{\beta}|X_1,\ldots,X_n] = \frac{\sigma^2}{n}\Big(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\Big)^{-1} = \frac{\sigma^2}{s_X^2 n}\begin{bmatrix} s_X^2 + \bar{X}^2 & -\bar{X} \\ -\bar{X} & 1 \end{bmatrix}$$

By looking at the diagonal entries of this matrix, we conclude that

$$\text{SE}[\hat{\beta}_0|X_1,\ldots,X_n] = \frac{\sigma}{\sqrt{n}}\sqrt{1 + \frac{\bar{X}^2}{s_X^2}}, \quad \text{SE}[\hat{\beta}_1|X_1,\ldots,X_n] = \frac{\sigma}{\sqrt{n}s_X},$$

which agrees with the calculations from Lecture 2 and (hopefully) your answer to Question 2a in Homework 1.

- For general p, when an intercept is included, there turns out to be a (pretty nice!) interpretation of the entries of $(\mathbf{X}^\top \mathbf{X})^{-1}$. However it requires a bit of matrix algebra to work out, so I think I'll leave it for a homework question.

References

Jerome Cornfield, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1):173–203, 1959.

Andrew Gelman, Boris Shor, Joseph Bafumi, and David Park. Rich state, poor state, red state, blue state: What's the matter with connecticut? *Poor State, Red State, Blue State: What's the Matter with Connecticut*, 2005.

OLS standard errors, demonstrated by simulation

Let's look at the OLS standard errors in a simulation example.

Here is a function for generating n independent random variables $(X_i, Y_i), i = 1, \ldots, n$, according to the model

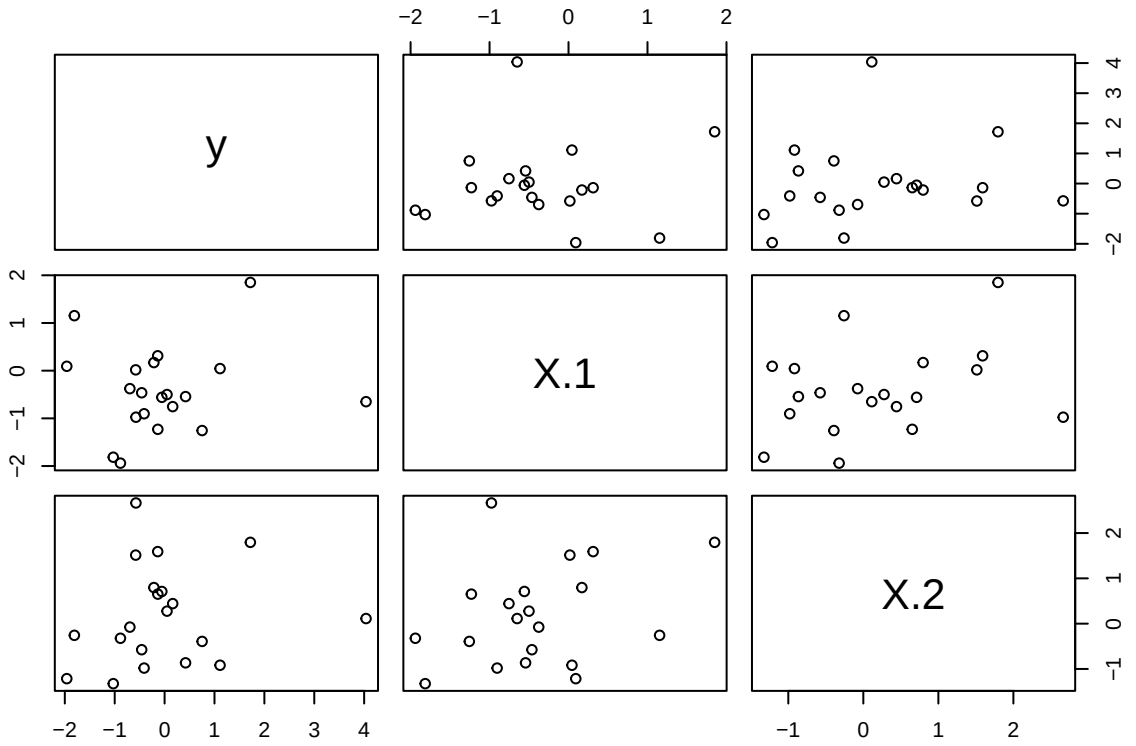
$$X_i = (X_{i1}, X_{i2}) \sim N_2(0, \Sigma), \Sigma = \begin{bmatrix} 1 & .8 \\ .8 & 2 \end{bmatrix}$$
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i,$$

where as usual ϵ_i is independent of X_i, and $\mathbb{E}][\epsilon_i] = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma^2$.

```
# A function to generate data from the multiple linear model, with p = 2
# normal predictors.
simulate_linear_model_with_normal_predictors = function(n,beta,sigma){
  # Multivariate normal predictors, with matrix
  Z = matrix(rnorm(n*2),n,2)  # Matrix with rows Z
  Sigma = matrix(c(1,.8,
                   .8,2),ncol = 2,nrow = 2)
  S = chol(Sigma) # matrix square root of Sigma
  X = Z %*% S
  # (Exercise: why is the covariance matirx of each row of X equal to Sigma?)

  # Normal errors and response
  epsilon = rnorm(n,mean = 0,sd = sigma)
  y = X %*% beta + epsilon

  data = data.frame(y = y,X = X)
  return(data)
}
```

Here we draw the data once, plot it, and fit a linear model. We

```
# Setting a seed ensures that the random numbers come out the same way each time.
set.seed(1993)

# Simulating settings
n = 20          # number of observations
beta = c(-.5,.2) # true coefficients
sigma = 1.5      # true error standard deviation

# Simulate the data
observed_data = simulate_linear_model_with_normal_predictors(n,beta = beta,sigma = sigma)
plot(observed_data)
```




```
# Fit the model
obj = lm(y ~ X.1 + X.2 + 1,data = observed_data)
coefs = obj$coefficients
coefs
```

```
## (Intercept)         X.1         X.2
## -0.04623496  0.05000644  0.17870309
```

Now we calculate the standard error.

```
# Calculate standard errors
X = as.matrix(cbind(rep(1,n),observed_data[,-1])) # notice we've added the all-ones vector
cov_ols = sigma^2 * solve(t(X) %*% X)             # covariance of \hat{beta}|X_1,...,X_n
standard_errors = sqrt(diag(cov_ols))

standard_errors
```

```
## rep(1, n)       X.1       X.2
## 0.3875229 0.4004753 0.3308395
```

In this case, our rule of thumb would tell us that

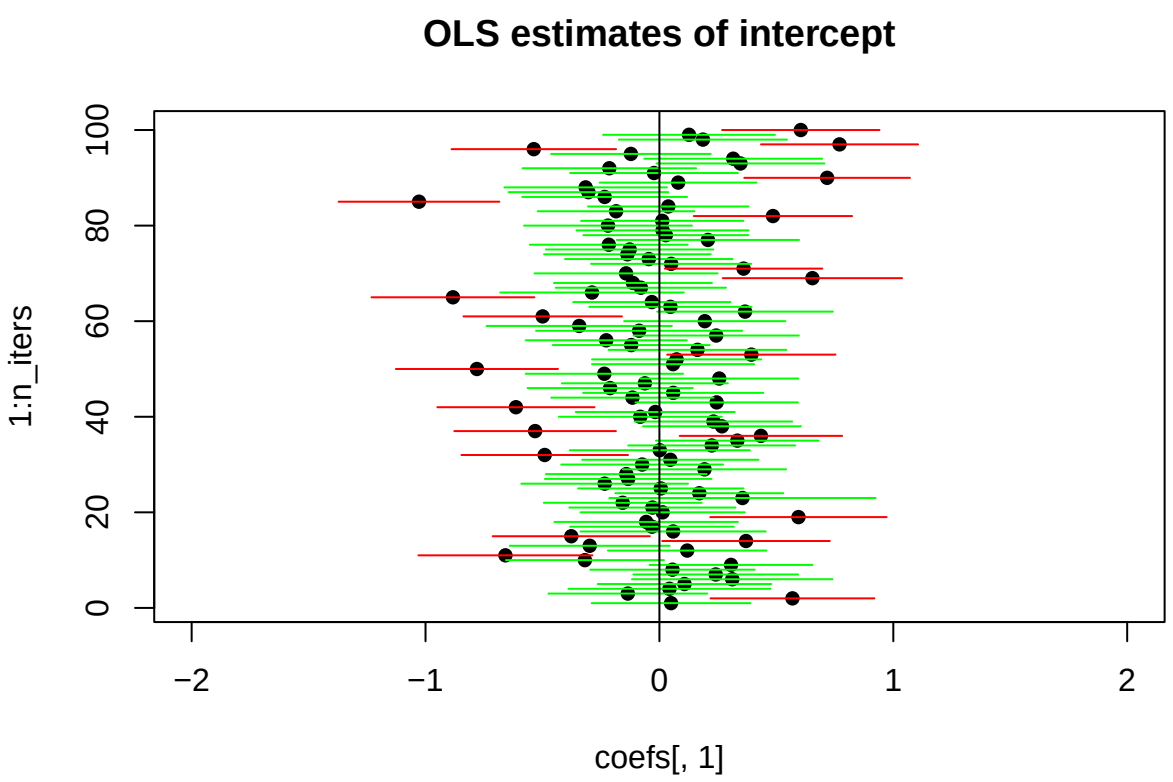
- $\beta_0 \approx$ -0.046, plus or minus 0.388
- $\beta_1 \approx$ 0.05, plus or minus 0.4
- $\beta_2 \approx$ 0.179, plus or minus 0.331.

Notice that β_0 and β_2 are within one standard error of $\hat{\beta}_0$ and $\hat{\beta}_2$, but not β_1.

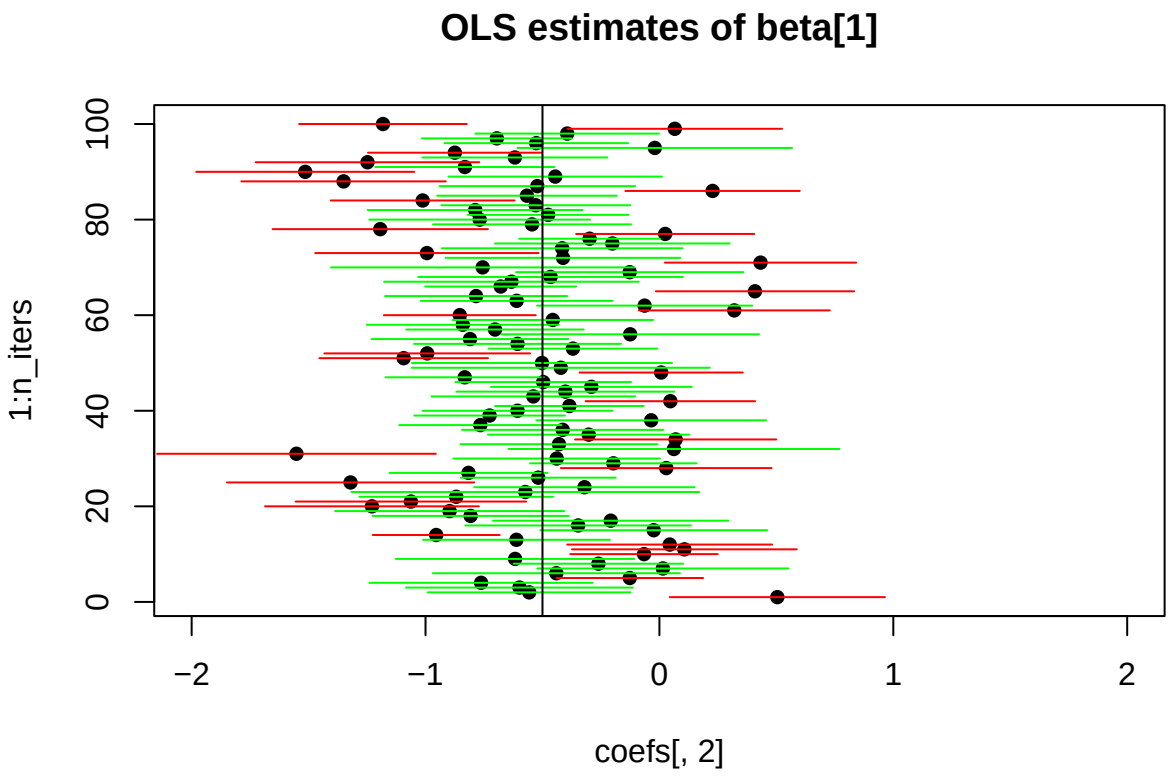
The rule of thumb over 100 different datasets

Of course, if we draw more random variables from the linear model, the data and the reuslting estimates will come out differently. Let's look at what happens when we redo the previous simulation over 100 different datasets.

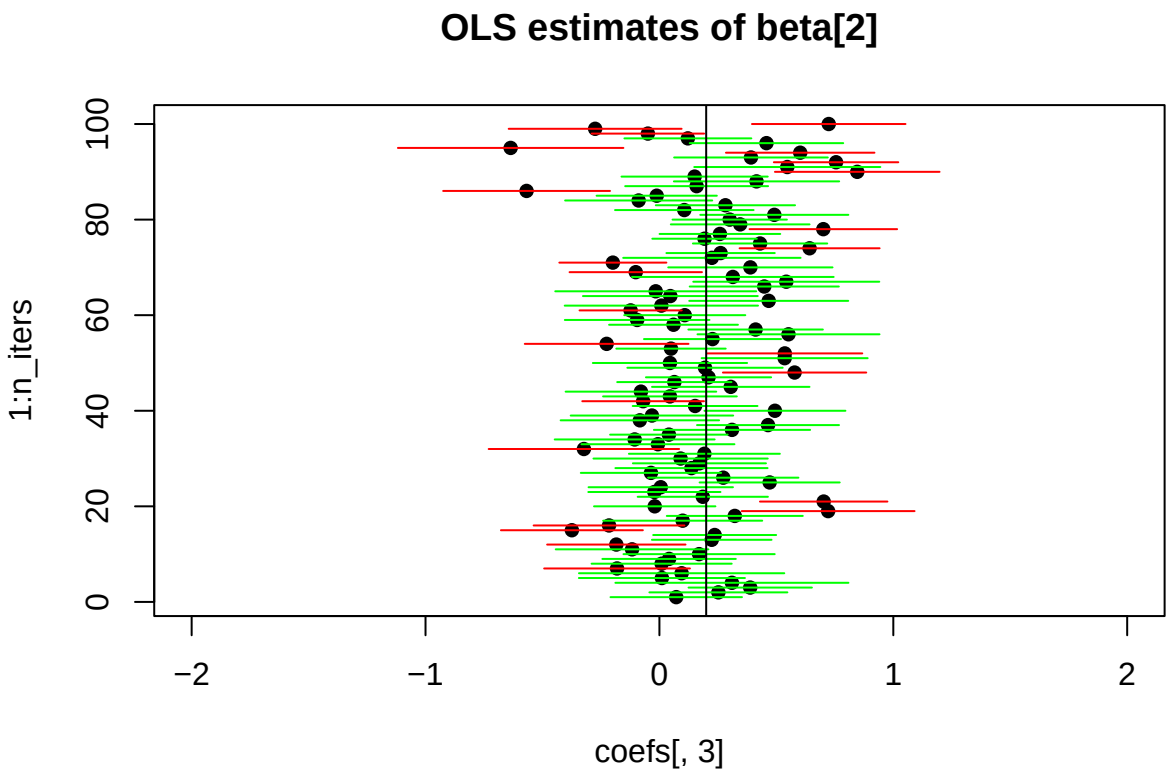
```
# Run a simulation calculating coefficients and standard errors
# for 100 different hypothetical datasets, each following a multiple linear model.

n_iters = 100  # number of hypothetical datasets

coefs = standard_errors = matrix(nrow = n_iters,ncol = 3)
for(ii in 1:n_iters)
{
  # Simulate the data
  observed_data = simulate_linear_model_with_normal_predictors(n,beta = beta,sigma = sigma)

  # Fit the model
  obj = lm(y ~ X.1 + X.2 + 1,data = observed_data)
  coefs[ii,] = obj$coefficients

  # Calculate standard errors
  X = as.matrix(cbind(rep(1,n),observed_data[,-1])) # notice we've added the all-ones vector
  cov_ols = sigma^2 * solve(t(X) %*% X)             # covariance of \hat{beta}|X_1,...,X_n
  standard_errors[ii,] = sqrt(diag(cov_ols))
}

# Plot estimates of the intercept, with error bars
plot(x = coefs[,1], y = 1:n_iters,
     xlim = c(-2,2),
     pch = 16,
     main = "OLS estimates of intercept")
for(ii in 1:n_iters)
{
  segments(x0 = coefs[ii,1] - standard_errors[ii,1],y0 = ii,
           x1 = coefs[ii,1] + standard_errors[ii,1],y1 = ii,
           col = ifelse(abs(coefs[ii,1]) < standard_errors[ii,1],"green","red")
  )
}
abline(v = 0)
```




```
# Plot estimates of beta[1], with error bars
plot(x = coefs[,2], y = 1:n_iters,
     xlim = c(-2,2),
     pch = 16,
     main = "OLS estimates of beta[1]")
for(ii in 1:n_iters)
{
  segments(x0 = coefs[ii,2] - standard_errors[ii,2],y0 = ii,
           x1 = coefs[ii,2] + standard_errors[ii,2],y1 = ii,
           col = ifelse(abs(coefs[ii,2] - beta[1]) < standard_errors[ii,2],"green","red")
  )
}
abline(v = beta[1])
```




```
# Plot estimates of beta[2], with error bars
plot(x = coefs[,3], y = 1:n_iters,
     xlim = c(-2,2),
     pch = 16,
     main = "OLS estimates of beta[2]")
for(ii in 1:n_iters)
{
  segments(x0 = coefs[ii,3] - standard_errors[ii,3],y0 = ii,
           x1 = coefs[ii,3] + standard_errors[ii,3],y1 = ii,
           col = ifelse(abs(coefs[ii,3] - beta[2]) < standard_errors[ii,3],"green","red")
  )
}
abline(v = beta[2])
```



There are three plots, one for each parameter. In each plot there are 100 points and error bars, one for each run of the simulation. The points correspond to the OLS estimate, and the error bars represent plus and minus 1 standard error. The vertical line marks the true value of the parameter which does not change from trial to trail. The line segments are green if the true parameter is "trapped' ' by the error bars, and red if not.

For each parameter, we see that across the different runs of the simulation, the true parameter is usually, but not always, within 1 standard error of the estimate. This means the rule of thumb is usually but not always correct. In fact, if we ran many simulations (say $1,000,000$ instead of 100) we would see that the rule of thumb is correct about 68.5% of the time.

The purpose of this simulation – besides giving a little more practice with the mechanics of the linear model in R – is to give a very intuitive interpretation of standard error and the rule of thumb. In fact, most statisticians prefer this **frequency-based** interpretation of probability, by appealing to what would happen on average over these many hypothetical datasets. Statisticians who think in this way are called **frequentists**.