STATS 305 Notes[1]

Art Owen[2]

Autumn 2013

[1]The class notes were beautifully scribed by Eric Min. He has kindly allowed his notes to be placed online for stat 305 students. Reading these at leasure, you will spot a few errors and omissions due to the hurried nature of scribing and probably my handwriting too. Reading them ahead of class will help you understand the material as the class proceeds.

[2]Department of Statistics, Stanford University.

Contents

Chapter 1

Overview

1.1 The Math of Applied Statistics

Given x, we'd like to predict y. But what exactly are x and y? Many potential combinations exist.

x, y	$\mathbb{R}$	$\{0,1\}$	K groups	Ordered groups	$\mathbb{R}^p$	$\mathbb{N}$	...
1 group							
2 groups							
K groups							
Ordered groups							
$\mathbb{R}$							
$\mathbb{R}^p$							
$\vdots$							

Table 1.1: Table of potential x's and y's.

The class's main focus is on the shaded column. However, the math behind this first column can often be transferred to the others without too much problem.

Keep in mind that our work in statistics is generally based on many untenable assumptions. However, some assumptions are more or less harmful to the analysis than others.

1.2 The Linear Model

We seek to predict $y \in \mathbb{R}$ using data $(x_i, y_i), i = 1, ..., n$. For now, assume that the data is i.i.d. Also, x_i may sometimes be fixed.

Given x, our "best" predictor of y is

$$\mathrm{E}[Y|X = x] = \mu(x)$$

This expression minimizes $\mathrm{E}[(Y - m(x))^2]$ over functions of m.

Proof.

$$\begin{aligned}
\mathrm{E}[(Y - m(x))^2 | X = x] =& \mathrm{E}[(Y - \mu(x) + \mu(x) - m(x))^2 | X = x] \\
=& \mathrm{E}[(Y - \mu(x))^2 | X = x] \\
&+ 2\mathrm{E}[(Y - \mu(x))(\mu(x) - m(x))| X = x] \\
&+ (\mu(x) - m(x))^2 \\
=& \mathrm{V}(Y|X = x) + 2(0) + (\mu(x) - m(x))^2 \\
=& \mathrm{V}(Y|X = x) + \mathrm{Bias}^2
\end{aligned}$$

□

This is the standard bias-variance trade-off. We cannot change variance. (We also assume that y has a finite variance.) However, our choice of $m(x)$ can minimize the bias.

1.2.1 Other Extensions

The proof above is slightly unsatisfactory since it already "knows" the conclusion. Instead, we could take a first-order condition:

$$\frac{d}{dm}\mathrm{E}[(Y - m)^2 | X = x] = 0 \qquad \Rightarrow \qquad m(x) = \mathrm{E}[Y|X = x]$$

This yields an extremum which must obviously be a minimum.

We were dealing with mean squared error above. What about absolute error? For that, $\mathrm{E}[(Y - m)|X = x]$ is minimized by using $m = \mathrm{med}(Y|X = x)$ (the median). In some cases, we may want $\mathrm{V}(Y|X = x)$, or a quantile like $Q^{0.999}(Y|X = x)$.

1.3 Linearity

Suppose we have data on boiling points of water at different levels of air pressure.

We can fit a line through this data and toss out all other potential information.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

By drawing a single line, we also assume that the residual errors are independent and have the same variance. That is,

$$\varepsilon \sim (0, \sigma^2)$$

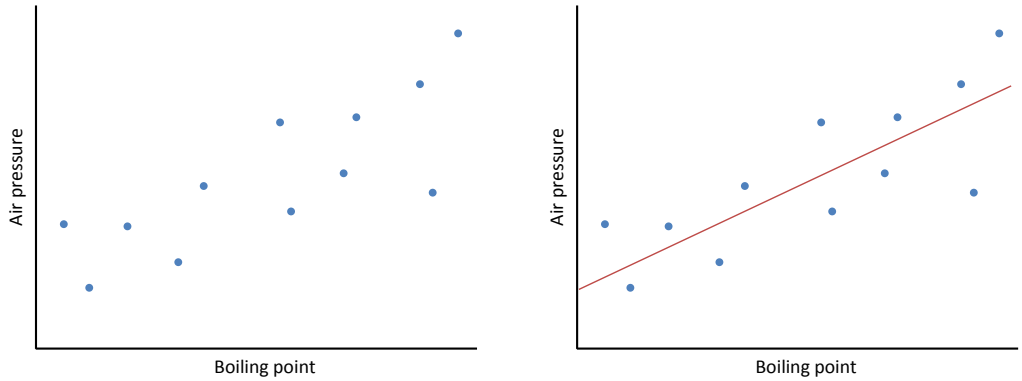


Figure 1.1: Sample data of boiling points at different air pressures, with and without linear fit.

Doing so overlooks the possibility that each point's error may actually not be random, but the effect of other factors we do not analyze directly (such as whether the experiment was done in the sun, done indoors, used a faulty thermometer, etc.). Nonetheless, the linear model is powerful in that it summarizes the data using three values: $\beta_0, \beta_1, \varepsilon_i$.

But what about the data in Figure 1.2?

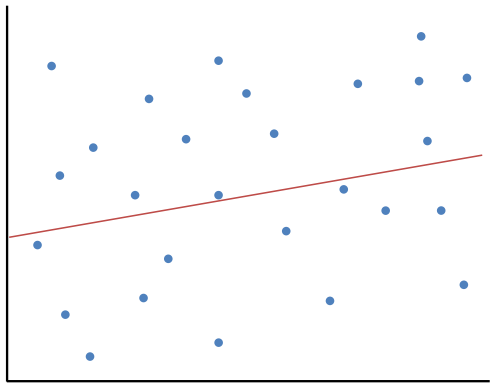


Figure 1.2: Other data with a fitted line.

This is also a simple linear model; the key difference is that the variance in the residuals here is much, much higher.

1.4 Beyond Simple Linearity

The definition of a linear model goes further than a straight line. More generally,

$$y_i = \beta_0 + \beta_i x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i \quad \text{has } p \text{ predictors and } p-1 \text{ parameters}$$

For example, we might want to find patterns in or predict the sale prices of homes. It could be that

$$\begin{aligned} y_i &= \text{sale price of home } i \\ x_{i1} &= \text{square footage} \\ x_{i2} &= \text{number of bedrooms} \\ &\vdots \\ x_{i147} &= \text{number of fireplaces} \end{aligned}$$

This sort of model is *linear* in β. It is *not* linear in x.

- Even if we had that $\beta_{i94} = (\text{Number of sinks})^2$, the model is linear in β.
- Sometimes, we may want to think of β_0 such that it is $\beta_0 x_{i0}$ where $x_{i0} = 1 \ \ \forall i$.

Linear models can go even further.

1.4.1 Polynomial Regression

$$\mathrm{E}[Y] = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k \qquad (x \in \mathbb{R})$$

If the relationship between x and y is a smooth curve, we can approximate it arbitrarily well with a large enough k.

- However, this is suspect, and we should expect overall bad performance with prediction.
- For small k, however, this can be quite useful.

1.4.2 Two Groups

What if we want to compare two groups, such as male vs. female, nickel vs. copper, or treatment vs. control?

$$\mathrm{E}[Y] = \begin{cases} \beta_0 + \beta_1 & x = 1 \\ \beta_0 & x = 0 \end{cases}$$

where one of the two groups is indicated as $x = 1$. Note that β_1 is the "extra" effect from being $x = 1$; it is the key parameter of interest here.

More commonly, we would write this function as

$$\mathrm{E}[Y] = \beta_0 + \beta_1 x$$

where x is a *dummy variable*.

1.4.3 k Groups

The logic above extends when having more than two groups. Let:

$$x_1 = \begin{cases} 1 & \text{if group 2} \\ 0 & \text{otherwise} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{if group 3} \\ 0 & \text{otherwise} \end{cases} \qquad \dots \qquad x_{k-1} = \begin{cases} 1 & \text{if group } k \\ 0 & \text{otherwise} \end{cases}$$

where one group is chosen as a point of reference. Then, we get that

$$\mathrm{E}[Y] = \beta_0 + \beta_1 x_1 + \beta_2 x_x + ... + \beta_{k-1} x_{k-1}$$

where group 1 has mean β_0 and group $j > 1$ has mean $\beta_0 + \beta_{j-1}$.

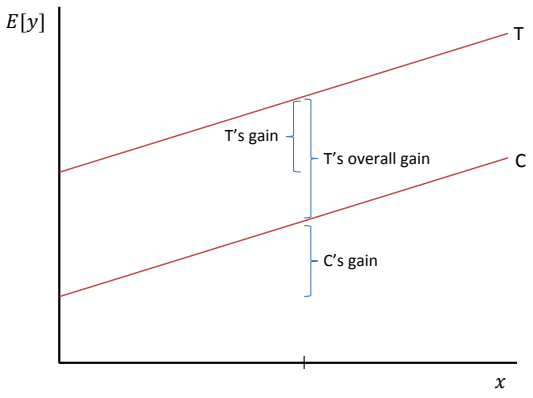
Another way to consider k groups is through the *cell mean model*, which we express as the following:

$$\mathrm{E}[Y] = \beta_1 1\{x = 1\} + \beta_2 1\{x = 2\} + ... + \beta_k 1\{x = k\}$$

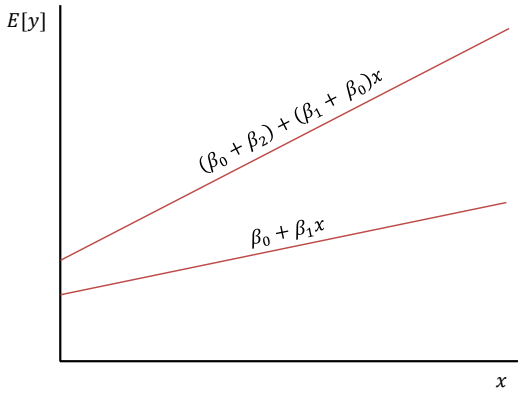
Note that the cell mean model has no intercept term.

1.4.4 Different Slopes

Suppose we are comparing a treatment and control group. It could be the case that both groups experience the same effect based on time, in which case the slopes of their two lines are the same (Figure 1.3a). But what if the two groups also have different slopes (Figure 1.3b)?



(a) Same slopes can be dealt with using a dummy variable.



(b) Different slopes can be dealt with using interactions.

Figure 1.3: Two cases of different slopes.

$$\mathrm{E}[Y] = \beta_0 + \beta_1 x + \beta_2 z + \beta_2 xz$$

where x is time and z is an indicator for treatment. This allows for the two groups to have different slopes.

1.4.5 Two-Phase Regression

There may be cases where the slope of the line changes at a certain point. For instance, the performance of an average human kidney begins to decline at age 40. See Figure 1.5b. How can we express these kinds of situations?

$$\mathrm{E}[Y] = \beta_0 + \beta_1 x + \beta_2 (x - t)_+ + \varepsilon_i \quad \text{where} \quad z_+ = \max(0, z) = \begin{cases} z & z \geq 0 \\ 0 & z < 0 \end{cases}$$

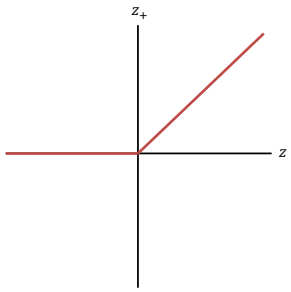
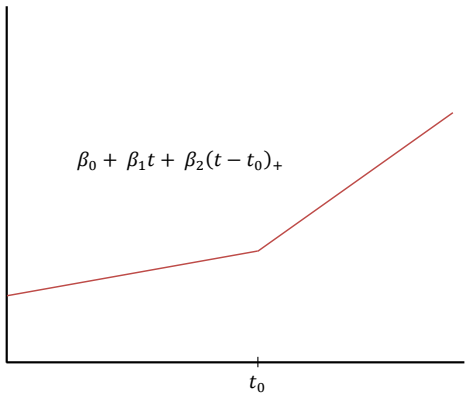
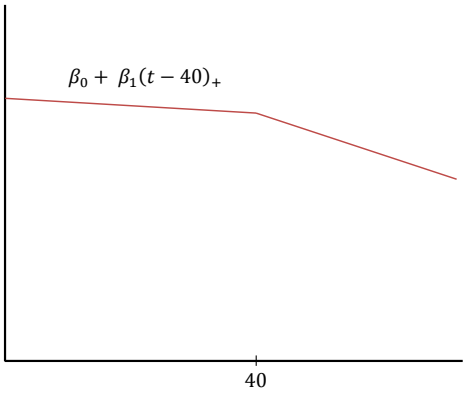


Figure 1.4: Visualization of z_+.



(a) Accounting for sudden increase at a known t_0.



(b) Accounting for decline in kidney performance at 40.

Figure 1.5: Examples of two-phase regression models.

1.4.6 Periodic Functions

What about cyclical data, such as calendar time? December 31 is not that distinct from January 1. How can we deal with an x that has a torus-like shape?

$$\mathrm{E}[Y] = \beta_0 + \beta_1 \sin(2\pi x) + \beta_2 \cos(2\pi x) + \beta_3 \sin(4\pi x) + ... \quad \text{where } 0 \leq x \leq 1$$

1.4.7 Haar Wavelets

Haar wavelets (on $[0, 1]$) are a sequence of square-shaped functions that can be added together in order to create a specific function. Think of Fourier analysis using "square" functions.

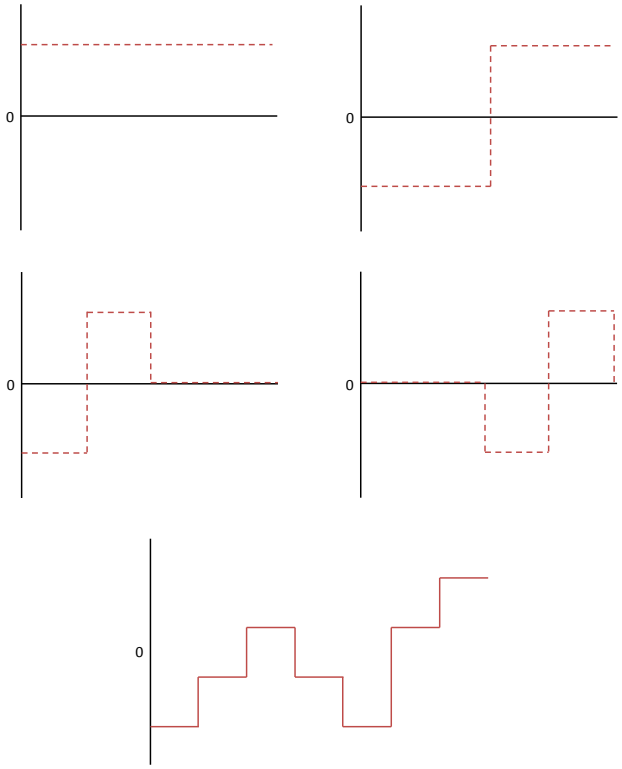


Figure 1.6: Four Haar wavelets. When added, they could form the last stepwise function.

We could add a sinusoidal part to this model, and it would still be linear. Remember:

Sums of linear models are linear models.

1.4.8 Multiphase Regression

Say that we want to model a relationship that changes over time. Suppose that time goes for k periods. We can use the following:

$$\mathrm{E}[Y] = \beta_0 + \beta_1(x - t_1)_+ + \beta_2(x - t_2)_+ + ... + \beta_k(x - t_k)_+$$

This can approximate piecewise functions and may be preferable to methods such as (excessive) polynomial regression.

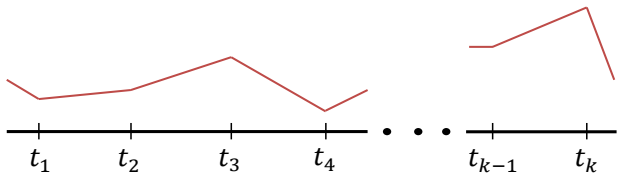


Figure 1.7: Example of multiphase regression.

1.5 Concluding Remarks

Despite these models' differences, the mathematics underlying them is all linear and practically the same.

Then what are examples of non-linear models?

- $y_i = \beta_0 \left(1 - e^{-\beta_1 t_i}\right) + \varepsilon_i$ is not linear because β_1 is in an exponent.
- $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 (x_i - \beta_3)_+ + \varepsilon_i$ is not linear because of β_3.
- $y_i + \sum_{j=1}^{k} \beta_j e^{-\frac{1}{2}||x_i - \mu_j||^2}$ is almost linear but has a small Gaussian bump in the middle.

Chapter 2

Setting Up the Linear Model

Last time, we discussed the big picture of applied statistics, with focus on the linear model.

Next, we will delve into more of the basic math, probability, computations, and geometry of the linear model. These components will be similar or the same across all forms of the model. We will then explore the actual ideas behind statistics; there, things will differ model by model.

There are six overall tasks we would like to perform:

1. Estimate β
2. Test, e.g. $\beta_7 = 0$
3. Predict y_{n+1} given x_{n+1}
4. Estimate σ^2
5. Check the model's assumptions
6. Make a choice among linear models

Before anything, let's set up some notation.

2.1 Linear Model Notation

$$X_i \in \mathbb{R}^d \qquad Y_i \in \mathbb{R}$$

$$Y_i = \sum_{j=1}^{p} Z_{ij}\beta_j + \varepsilon_i \quad \text{where} \quad Z_{ij} = j^{th} \text{ function of } X_i$$

We also call these j^{th} functions *features*. Note that the dimensions of X (d) do not have to be equal to the number of features (p). For example,

$$Z_i = \begin{bmatrix} 1 & x_{i1} & \dots & x_{ip} \end{bmatrix} \qquad Z_i = \begin{bmatrix} 1 & x_{i1} & x_{i2} & x_{i1}^2 & x_{i2}^2 \end{bmatrix} \quad \text{(quadratic regression)}$$

In the first case, $p = d + 1$. In the second, $p = 5$ and $d = 2$. Features are not the same as variables.

2.2 Two Potential Models

In a regression, we are given data $\{(x_1, y_1), ..., (x_n, y_n)\}$. Using this, we seek to estimate $\mu(x) = \mathrm{E}[Y|X = x]$. We can conceive of this in two different ways. The two are often confused or not differentiated well, and the choice may be non-trivial. However, most of the course will focus on the first approach below.

2.2.1 Regression Model

In the regression model, we assume the following:

- $x_1, ..., x_n$ are *fixed* numbers.
- Given $x_1, ..., x_n$,

$$y_i \overset{\text{i.i.d.}}{\sim} \mathcal{L}(Y|X = x)$$

That is, y_i is an independent random variable with some distribution given x_i. Since these are i.i.d., y_i only depends on x_i and not on any $x_{j\neq i}$.

2.2.2 Correlation Model

In the correlation model, we instead assume that

$$(x_i, y_i) \text{ are i.i.d. from } F_{x,y}$$

That is, they are jointly distributed.

This model also has the following characteristic:

$$\begin{aligned}\mathrm{V}(Y) &= \mathrm{E}[\mathrm{V}(Y|X)] + \mathrm{V}(\mathrm{E}[Y|X]) \\ &= \mathrm{E}[\sigma^2(x)] + \mathrm{V}(\mu(x)) \quad \text{where} \quad \sigma^2(x) = \mathrm{V}(Y|X = x)\end{aligned}$$

It happens to be that $0 \leq \dfrac{\mathrm{V}(\mu(x))}{\mathrm{V}(Y)} \leq 1$. The middle term is what we consider R^2, or a measure of how much variance in Y is explained by X.

2.3 The Linear Model

There are three ways to express the linear model.

First,

$$Y_i = \sum_{j=1}^{p} Z_{ij}\beta_j + \varepsilon_i \qquad \text{where} \qquad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

Note that the i.i.d. notation for ε_i is interchangeable with "ind." However, both of these are distinct from $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_i^2)$ (where each error term has its own variance) and $\varepsilon_i \overset{\text{ind.}}{\sim} (0, \sigma^2)$ (where the distribution is unknown).

Second,

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad Z = \begin{bmatrix} z_{11} & \dots & z_{ip} \\ \vdots & \ddots & \vdots \\ z_{n1} & \dots & z_{np} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

Third,

$$A = Z\beta + \varepsilon \qquad \text{where} \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

We can also express this third version as $A \sim \mathcal{N}(Z\beta, \sigma^2 I)$.

Let's dig deeper into the vector/matrix form of the regression model. But first, we require some probability and matrix algebra review.

2.4 Math Review

Suppose that $A \in \mathbb{R}^{m \times n}$ is a random $m \times n$ matrix. Expected values on this matrix work component-wise:

$$\mathrm{E}\left[\begin{pmatrix} x_{11} & \dots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mn} \end{pmatrix}\right] = \begin{pmatrix} \mathrm{E}[x_{11}] & \dots & \mathrm{E}[x_{1n}] \\ \vdots & \ddots & \vdots \\ \mathrm{E}[x_{m1}] & \dots & \mathrm{E}[x_{mn}] \end{pmatrix}$$

Now, suppose there exist two non-random matrices A and B. Then,

$$\mathrm{E}[AX] = A\,\mathrm{E}[X] \qquad \mathrm{E}[XB] = \mathrm{E}[X]\,B$$

Say that $A \in \mathbb{R}^n$ and $A \in \mathbb{R}^n$ are random column vectors. Then,

- $\text{Cov}(X, Y) \equiv \text{E}[(X - \text{E}[X])(Y - \text{E}[Y])^T] \in \mathbb{R}^{n \times m}$
- $\text{Var}(Y) = \text{Cov}(Y, Y) \in \mathbb{R}^{m \times m}$
  - Variances are on the diagonal, and covariances are on the off-diagonal.
- $\text{Cov}(AX, BY) = A\ \text{Cov}(X, Y)\ B^T$
- $\text{V}(AX + b) = A\text{V}(X)A^T$ (where A is random and b is non-random)

The variance matrix $\text{V}(X)$ is positive semi-definite. That is, for a vector C,

$$0 \leq \text{V}(C^T X) = C^T \text{V}(X) C$$

Indeed, $\text{V}(X)$ is positive definite unless $\text{V}(C^T X) = 0$ for some $C \neq 0$.

2.4.1 Quadratic Forms

Suppose that $A \in \mathbb{R}^{n \times n}$. Written explicitly, the quadratic form is a scalar value defined as

$$x^T A x = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} x_i x_j$$

We can assume symmetry for A; $A_{ij} = A_{ji}$. (If it is not symmetric, we can instead use $\frac{A+A^T}{2}$.[1])

This doesn't look directly relevant to us. So why do we care about quadratic forms? Consider how we estimate variance. We know that

$$\hat{\sigma}^2 \propto \sum_{i=1}^{n} (y_i - \overline{Y})^2 \quad \text{where} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

In matrix notation, this variance estimate can be written as:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}^T \begin{bmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & & \vdots \\ \vdots & & \ddots & \vdots \\ -\frac{1}{n} & \dots & \dots & 1 - \frac{1}{n} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} y_i (y_i - \bar{y})$$

[1] This is true because

$$x^T A x = (x^T A x)^T = x^T A^T x = x^T \left(\frac{1}{2} A + \frac{1}{2} A^T \right) x$$

The first equality is true because the transpose of a scalar is itself. The second equality is from the fact that we are averaging two quantities which are themselves equal. Based on this, we know that only the symmetric part of A contributes to the quadratic form.

If you subtract $\overline{Y}$ from the lone y_i, this does not change the overall result. So then,

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

This is the original expression for estimating variance.

What are some other cases where quadratic forms matter to us? R^2 is one example, since

$$R^2 = 1 - \frac{Y^T A_1 Y}{Y^T A_2 Y}$$

Finding Variance

Quadratic forms are important in finding an unbiased estimate for variance.

Suppose that $\mathrm{E}[A] = \mu \in \mathbb{R}^n$ and $\mathrm{V}(A) = \Sigma \in \mathbb{R}^{n \times n}$. Then,

$$\mathrm{E}[Y^T AY] = \mu^T A\mu + \mathrm{tr}(A\Sigma)$$

We will often engineer this so that $\mu = 0$ so that the key thing we care about is just $\mathrm{tr}(A\Sigma)$. But where does this come from?

$$\begin{aligned} Y^T AY &= [\mu + (Y - \mu)]^T A[\mu + (Y - \mu)] \\ &= \mu^T A\mu + \mu^T A(Y - \mu) + (Y - \mu)^T A\mu + (Y - \mu)^T A(Y - \mu) \end{aligned}$$

$$\mathrm{E}[Y^T AY] = \mu^T A\mu + \mathrm{E}[\underbrace{(Y - \mu)^T}_{1\times n} \underbrace{A}_{n\times n} \underbrace{(Y - \mu)}_{n\times 1}]$$

Note that the quantity in the expected value is a 1×1 matrix (basically a scalar). We can therefore pull terms out. Also, recall the "trace trick": If A and B are matrices such that AB and BA both exist, then $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. In the case of a 1×1 matrix, the trace is just the number.

With all of this in mind,

$$\begin{aligned}
E[Y^TAY] &= \mu^TA\mu + \mathrm{E}[(Y-\mu)^TA(Y-\mu)] \\
&= \mu^TA\mu + \mathrm{E}[\mathrm{tr}[(Y-\mu)^TA(Y-\mu)]] \\
&= \mu^TA\mu + \mathrm{E}[\mathrm{tr}[A(Y-\mu)(Y-\mu)^T]] \quad \text{using trace trick} \\
&= \mu^TA\mu + \mathrm{tr}\mathrm{E}[A(Y-\mu)(Y-\mu)^T] \\
&= \mu^TA\mu + \mathrm{tr}(A\mathrm{E}[(Y-\mu)(Y-\mu)^T]) \quad \text{since } A \text{ was on "outside"} \\
&= \mu^TA\mu + \mathrm{tr}(A\Sigma)
\end{aligned}$$

So what? If we assume that $\mu = 0$ and variance is $\sigma^2\Sigma$, then

$$\mathrm{E}[Y^TAY] = \sigma^2\mathrm{tr}(A\Sigma)$$

Note that Σ is a known value. Then, we search the universe to find an A such that $\mathrm{tr}(A\Sigma) = 1$, which would give us an unbiased estimate of σ^2.

But what if there are multiple matrices that satisfy this equality? In more formal terms, what if $\mathrm{tr}(A_1\Sigma) = \mathrm{tr}(A_2\Sigma) = 1$? It is actually quite likely that multiple matrices result in unbiased σ^2 estimates (and if so, then these matrices can be combined and also satisfy the equality). Which A do we pick?

- We could choose the one with the smallest $\mathrm{V}(Y^TAY)$, but this is analytically difficult/messy to figure out.
- Instead, for normal data, we prefer a small $\mathrm{tr}(A^2)$.

Coming up, we will consider *normal* random vectors. Properties of normality are very useful; they give clean and sharp answers (whether the assumption of normality is sound or not!).

Chapter 3

The Normal Distribution

Last time, we reviewed some matrix algebra. Now, let's discuss the normal distribution and its relatives.

A standard normal distribution is $Z \sim \mathcal{N}(0,1)$ and is characterized by the following (where $z \in \mathbb{R}$):

$$f(z) = \phi(z) = \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}} \qquad \Phi(z) = \int_{-\infty}^{z} \phi(x)dx \qquad z = \Phi^{-1}(u)$$

These are the pdf, cdf, and quantiles, respectively.

3.1 Friends of $\mathcal{N}(0,1)$

Many distributions we care about are based on the (standard) normal.

3.1.1 χ^2

The χ^2 distribution is based on the sum of squared normals.

Assume that $z_1, ..., z_k$ are i.i.d. $\mathcal{N}(0,1)$. Then,

$$\sum_{i=1}^{k} z_i^2 \sim \chi_k^2$$

3.1.2 t-distribution

Suppose that $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi_{(n)}^2$. Then,

$$\frac{z}{\sqrt{\frac{x}{n}}} \sim t_{(n)}$$

Note that $t_{(n)}$ is symmetric, and that $t_{(n)} \to \mathcal{N}(0,1)$ as $n \to \infty$. The t-distribution also has a heavier tail than the standard normal.

3.1.3 F-distribution

The F-distribution is the ratio of two χ^2 distributions. Using n and d to denote numerator and denominator respectively, we define:

$$\frac{\frac{1}{n}\chi^2_{(n)}}{\frac{1}{d}\chi^2_{(d)}} \sim F_{n,d}$$

We would typically expect this ratio to be close to 1.

3.2 The Multivariate Normal

Suppose we have a vector Z

$$Z = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix} \qquad \text{where} \qquad z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

The multivariate normal is a distribution of $Y = AZ + b$.

$$Y \sim \mathcal{N}(\mu, \Sigma) \qquad \mu = \mathrm{E}[Y] = b \qquad \Sigma = \mathrm{V}(Y) = AA^T$$

The characteristic function for the MVN is

$$\phi_Y(t) = \mathrm{E}[e^{it^T y}] = e^{it^T\mu - \frac{1}{2}t^T\Sigma t}$$

If Σ is invertible (that is, there exists Σ^{-1}), then y has the following density:

$$(2\pi)^{-\frac{n}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right)$$

This expression seems more convenient. However, in many cases, Σ is not invertible and we have to resort to the characteristic function.[1]

3.2.1 Linear Transformations

Suppose that we partition a vector Y in two. Then, we get the following:

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma) = \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

Y_1 and Y_2 are independent iff $\text{Cov}(Y_1, Y_2) = \Sigma_{12} = 0$. (To prove this, plug this into the original characteristic function formula.)

Furthermore, if we assume that $Y \sim \mathcal{N}(\mu, \Sigma)$, then

$$AY + b \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$$

That is, linear transformations do not affect the normality. The normal distribution remains normal. This is a unique and very important property of the normal distribution which makes it so appealing to use (even when it may not be accurate).

3.2.2 Normal Quadratic Forms

Suppose that $Y \sim \mathcal{N}(\mu, \Sigma)$ and that Σ^{-1} exists. Then,

$$(Y - \mu)^T \Sigma^{-1} (Y - \mu) \sim \chi^2_{(n)}$$

Why is this the case? First, note that Σ is symmetric and positive definite. Then, we can write that $\Sigma = P^T \Lambda P$, where P is an $n \times n$ orthogonal matrix and $\Lambda = \text{diag}(\lambda_1, ..., \lambda_n)$ where $\lambda_j > 0$.

Then, define $Z = \Lambda^{-1} P (Y - \mu)$, which is a standardized version of Y. Vector Z then is normally distributed

$$Z \sim \mathcal{N}(0, \Lambda^{-1/2} P \Sigma P^T \Lambda^{-1/2}) = \mathcal{N}(0, \Lambda^{-1/2} P P^T \Lambda P P^T \Lambda^{-1/2}) = \mathcal{N}(0, I_n)$$

It follows that the elements of Z are independent and that $Z_i \sim \mathcal{N}(0, 1)$. So,

[1] Suppose we define $e_i = y_i - \overline{Y}$. Then,

$$V\left(\sum_{i=1}^{n} e_i \right) = 0$$

because $\sum e_i = \sum (y_i - \overline{Y}) = n\overline{Y} - n\overline{Y} = 0$ and the variance of a constant is zero. Since this is a characteristic of regressions by design, we almost always are dealing with singular, non-invertible matrices.

$$(Y-\mu)^T\Sigma^{-1}(Y-\mu) = Z^TZ \sim \chi^2_{(n)}$$

3.2.3 Rotation

Suppose that $Z \sim \mathcal{N}(0, I)$ and that $Y = QZ$, where Q is an orthogonal matrix—that is, $Q^TQ = I$. Then, it also happens to be that

$$Y = QZ \sim N(0, I)$$

Any orthogonal matrix that flips/rotates Z does not affect the properties of the distribution. Again, this is unique to normal distributions and will prove useful later.

3.2.4 More on Independence

Let's go back to a partition of Y. As stated before,

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma) = \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

Y_1 and Y_2 are independent iff $\Sigma_{12} = 0$. It seems to follow that functions of independent vectors should also be independent. That is, it should be that $g_1(Y_1)$ and $g_2(Y_2)$ are independent for all g_1, g_2. We can use this "trick" to learn more about the distribution of Y.

Suppose that Y_i is i.i.d. $\mathcal{N}(\mu, \sigma^2)$. Then, we can define a vector populated by $\overline{Y}$ and $y_i - \overline{Y}$.

$$\begin{bmatrix} \overline{Y} \\ y_1 - \overline{Y} \\ \vdots \\ y_n - \overline{Y} \end{bmatrix} = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \\ & I_n - \frac{1}{n}J_n & & \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \text{where} \qquad J = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & \ddots & & 1 \\ 1 & & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

$$\begin{bmatrix} \overline{Y} \\ y_1 - \overline{Y} \\ \vdots \\ y_n - \overline{Y} \end{bmatrix} = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \\ & I_n - \frac{1}{n}J_n & & \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & I - \frac{1}{n}J \end{bmatrix} \right)$$

Ultimately, we find that $\overline{Y}$ is independent of $y_1 - \overline{Y}, ..., y_n - \overline{Y}$. This is an interesting result that allows us to say that

$$t = \sqrt{n}\frac{(\overline{Y} - \mu)/\sigma}{\sqrt{\frac{1}{n-1}\frac{\sum(y_i - \overline{Y})^2}{\sigma^2}}}$$

With the normal distribution, you can mine the data twice and get two independent entities—one for the numerator, and one for the denominator. This is a very useful finding.

3.2.5 Conditional Distributions

Suppose that we have partitioned data of Y with known Y_1. What is Y_2 given Y_1? That is, what is $\mathcal{L}(Y_2|Y_1 = y_1)$? It will also be normal, but in what way?

Refer to page 46 of the "Probability Review" notes. However, the main finding is that

$$Y_2 \sim \mathcal{N}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$$

Here, let's consider a special simple case. Suppose we have that

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right)$$

Using the formula given above, we find that

$$\begin{aligned} Y|X = x \quad &\sim \quad \mathcal{N}\left(\mu_y + \rho\sigma_x\sigma_y\frac{1}{\sigma_x^2}(x - \mu_x), \sigma_y^2 - \rho\sigma_x\sigma_y\frac{1}{\sigma_x^2}\rho\sigma_x\sigma_y\right) \\ &= \mathcal{N}\left(\mu_y + \rho\frac{x - \mu_x}{\sigma_x}\sigma_y, \sigma_y^2(1 - \rho)\right) \end{aligned}$$

Note that $\frac{x-\mu_x}{\sigma_x}$ is the number of standard deviations x is from μ_x, and then $\rho\frac{x-\mu_x}{\sigma_x}$ is the number of standard deviations for Y. The size of ρ determines the relationship between x and y's standard deviations. That is, when X is $\Delta = \frac{x-\mu_x}{\sigma_x}$ standard deviations above its mean μ_x, then it should be that Y is $\rho\Delta$ standard deviations above its mean μ_y. Also note that the change in μ_y is linear.

As for variance, we see that it does not depend on X. Since $1 - \rho \leq 1$, certainty will generally improve with more data. As stated in the notes (with amendments to maintain consistent notation):

> It is noteworthy and special that, when $(X^T, Y^T)^T$ is multivariate normal, then $\mathrm{V}(Y|X = x)$ does not depend on which exact x was observed. Observing $X = x$ shifts the expected value of Y by a linear function of x but makes a variance change (usually a reduction) that is independent of x.

3.3 Non-Central Distributions

3.3.1 Non-Central χ^2

Suppose that $X_i \overset{ind}{\sim} \mathcal{N}(a_i, 1)$ where $\lambda = \sum_{i=1}^{n} a_i^2$ measures overall non-centrality across all n observations. Then, non-central χ^2 is defined by

$$\sum_{i=1}^{n} X_i^2 \sim \chi'^{\,2}_{(n)}(\lambda)$$

Note the prime symbol to indicate non-centrality. This distribution is useful for picking appropriate sample sizes for studies, among other tasks.

There is also a small miracle hidden in this expression. Everything depends only on the a_i's and the sum of their squares. All sorts of permutations of a_i could result in the same overall distribution.

Perhaps obvious, but if $\lambda = 0$, then we yield the regular χ^2 distribution.

3.3.2 Non-Central F Distribution

This is defined by a non-central χ^2 distribution divided by a central χ^2 distribution. More formally,

$$\frac{\frac{1}{n_1}\,\chi'^{\,2}_{n_1}(\lambda)}{\frac{1}{n_2}\,\chi^2_{n_2}}$$

3.3.3 Doubly Non-Central F Distribution

Simply the quotient of two non-central χ^2 distributions.

$$\frac{\frac{1}{n_1}\,\chi'^{\,2}_{n_1}(\lambda_1)}{\frac{1}{n_2}\,\chi'^{\,2}_{n_2}(\lambda_2)}$$

Chapter 4

Linear Least Squares

We use least squares so readily because the math is relatively easy while giving us the greatest marginal gain. It is parsimonious.

4.1 The Best Estimator

Conceptually, the "best" β is the one that minimizes

$$\mathrm{E}[(Y - Z\beta)^T(Y - Z\beta)] = \sum_{i=1}^{n} \mathrm{E}[(y_i - z_i\beta)^2]$$

But we don't know β. Therefore, we pick a $\hat{\beta}$ that minimizes

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - z_i\beta)^2$$

This is the sample sum of squares. Notice the multiplication by $\frac{1}{2}$; this is done to simplify the first derivative.

There are multiple ways to approach this problem: calculus, geometry, and distributions.

4.1.1 Calculus Approach

Set the first derivative equal to zero and solve.

$$\frac{d}{d\beta_j} = \sum_{i=1}^{n}(y_i - z_i\hat{\beta})(-z_{ij}) = 0 \quad \text{where} \quad j = 1, ..., p \quad \text{are columns of } Z$$

Expressed differently,

$$Z^T(Y - Z\hat{\beta}) = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

The expression above is called the *normal equation* ("normal" here meaning "perpendicular"). Rearranging, we get:

$$Z^T Y = Z^T Z\hat{\beta}$$

So, the minimizer $\hat{\beta}$ satisfies the normal equation. Notice the use of Z instead of X to distinguish between the number of variables and number of functions. In any case, this gives us the classic result:

$$\hat{\beta} = (Z^T Z)^{-1} Z^T Y$$

Proposition 1. If $(Z^T Z)^{-1}$ exists, then $\hat{\beta} = (Z^T Z)^{-1} Z^T Y$ is unique.

Then, we can predict y_{n+1} by evaluating $z_{n+1}\hat{\beta}$. If it is the case that $(Z^T Z)$ is not invertible, then there are an infinite number of solutions since various combinations of just two plausible solutions are also solutions.

4.1.2 Geometric Approach

See Figure 4.1. Essentially, $\hat{\beta}$ minimizes $||Y - Z\beta||^2$, where Y is a vector and $Z\beta$ is a hyperplane. We drop a perpendicular onto the hyperplane.

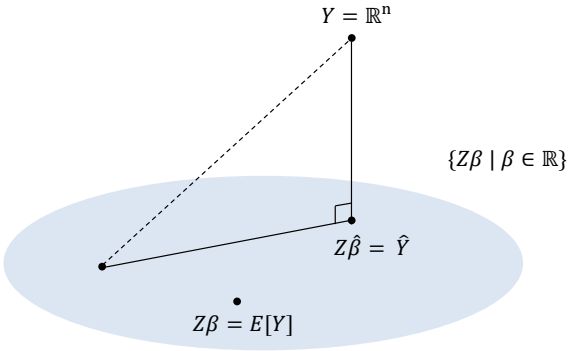


Figure 4.1: Projection onto the hyperplane.

4.1.3 Distributional Theory of Least Squares

If $\hat{\beta} = (Z^T Z)^{-1} Z^T Y$, then it follows that $\mathrm{E}[\hat{\beta}] = \mathrm{E}[(Z^T Z)^{-1} Z^T Y]$. Also, since we assume that $Y = Z\beta + \varepsilon$, we can write

$$\hat{\beta} = (Z^T Z)^{-1} Z^T (Z\beta + \varepsilon) = \beta + (Z^T Z)^{-1} Z^T \varepsilon$$

Furthermore, if we assume that X is fixed and non-random (which is a convenient assumption, whether it's sound or not), then Z must be fixed and non-random, as well. Then, we get the following:

$$\begin{aligned} \mathrm{E}[\hat{\beta}] &\overset{\text{fixed } X}{=} (Z^T Z)^{-1} Z^T \mathrm{E}[Y] \\ &= (Z^T Z)^{-1} Z^T (Z\beta) \\ &= \beta \end{aligned}$$

$\hat{\beta}$ is thus an unbiased estimator of β.[1] (We know that $\mathrm{E}[Y] = Z\beta$ since $Y = Z\beta + \varepsilon$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$.)

How about the variance? Using that $\mathrm{E}[\hat{\beta}] = \beta + (Z^T Z)^{-1} Z^T \varepsilon$,

$$\begin{aligned} \mathrm{V}(\hat{\beta}) &= \mathrm{V}((Z^T Z)^{-1} Z^T Y) \\ &= (Z^T Z)^{-1} Z^T \mathrm{V}(\varepsilon) Z (Z^T Z)^{-1} \\ &= (Z^T Z)^{-1} Z^T (\sigma^2 I) Z (Z^T Z)^{-1} \\ &= \sigma^2 (Z^T Z)^{-1} \end{aligned}$$

As we would intuitively expect, larger values of $Z^T Z$ result in smaller variances.

So, we get the following:

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (Z^T Z)^{-1})$$

If X were random, then the result is still valid: $\mathrm{E}[\hat{\beta}] = \mathrm{E}[\mathrm{E}[\hat{\beta}|X]] = \mathrm{E}[\beta] = \beta$.

[1] Another proof of unbiasedness based on the other expression for $\hat{\beta}$:

$$\mathrm{E}[\hat{\beta}] = \mathrm{E}[\beta + (Z^T Z)^{-1} Z^T \varepsilon] = \beta + (Z^T Z)^{-1} Z^T \mathrm{E}(\varepsilon) = \beta$$

4.2 The Hat Matrix and t-tests

We have established that $\hat{Y} = Z\hat{\beta} = Z(Z^TZ)^{-1}Z^TY$ where $\hat{Y}$ is the vector $\begin{bmatrix}\hat{\mathrm{E}}[y_1] & \dots & \hat{\mathrm{E}}[y_n]\end{bmatrix}^T$. We often call the expression $Z(Z^TZ)^{-1}Z^T$ the *hat matrix* (since it puts a hat on Y). The hat matrix has important properties that make it essential in the linear least squares model.

We plot the residuals below.

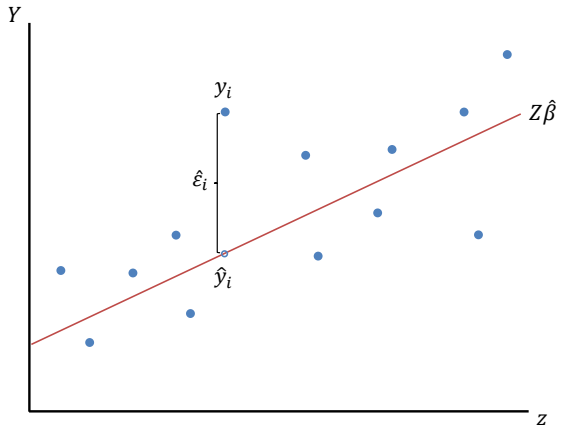


Figure 4.2: Depiction of residuals.

Note that

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \quad \Rightarrow \quad \hat{\varepsilon} = Y - \hat{Y} \quad \Rightarrow \quad \hat{\varepsilon} = (I - H)Y$$

H has two important properties. First, it is symmetric: $H = H^T$. Second, it is idempotent: $H = H^2$. The matrix algebra to prove these is simple. Idempotency also has a geometric interpretation. H drops a perpendicular to the (hyper)plane. Additional H's also drop a perpendicular to the (hyper)plane, but we are already on the hyperplane. Therefore, multiplying by another H makes no difference.

By idempotency, $H^2Y = H\hat{Y}$. In fact, $(I - H)$ is also idempotent:

$$(I - H)(I - H) = I - H - H + H^2 = I - H - H + H = I - H$$

4.3 Distributional Results

Theorem 1. For $Y \sim \mathcal{N}(Z\beta, \sigma^2 I)$, given that Z is full rank (i.e. (Z^TZ) is invertible),

- $\hat{\beta} \sim \mathcal{N}(\beta, (Z^TZ)^{-1}\sigma^2)$
- $\hat{Y} \sim \mathcal{N}(Z\beta, H\sigma^2)$

- $\hat{\varepsilon} \sim \mathcal{N}(0, (I-H)\sigma^2)$

Furthermore, $\hat{\varepsilon}$ is independent of $\hat{\beta}$ and $\hat{Y}$. (This independence helps with t-tests, where the numerator and denominator must be independent.)

The full proof is in the notes. Finding mean and variance is easy. Here, we will prove the independence result.

Proof. To prove independence in a Gaussian setting, we simply need to calculate covariance and see if it equals zero.

$$\begin{aligned}
\text{Cov}(\hat{Y}, \hat{\varepsilon}) &= \text{Cov}(HY, (I-H)Y) \\
&= H\text{Cov}(Y,Y)(I-H)^T \\
&= H(\sigma^2 I)(I-H) \\
&= \sigma^2(H-H) \\
&= 0
\end{aligned}$$

□

4.3.1 Distribution of $\hat{\varepsilon}$

Another crucial part worth noting is that $\sum_{i=1}^n \hat{\varepsilon}_i^2 \sim \sigma^2 \chi^2_{(n-p)}$. This will prove very useful soon. But why is this result true?

Based on past results, we know that $\sum_{i=1}^n \hat{\varepsilon}_i^2 = \hat{\varepsilon}^T\hat{\varepsilon} = [(I-H)Y]^T(I-H)Y$. Furthermore, we can express $(I-H)Y$ in different terms:

$$\begin{aligned}
(I-H)Y &= [I - Z(Z^TZ)^{-1}Z^T][Z\beta + \varepsilon] \\
&= [I - Z(Z^TZ)^{-1}Z^T]\varepsilon \\
&= (I-H)\varepsilon
\end{aligned}$$

So $Y = \varepsilon$. Going back to $[(I-H)Y]^T(I-H)Y$, we see that

$$[(I-H)Y]^T(I-H)Y = \varepsilon^T(I-H)^T(I-H)\varepsilon = \varepsilon^T(I-H)\varepsilon$$

Now, say that $I - H = P\Lambda P^T$, where P is an orthogonal matrix and Λ is a diagonal matrix. Since $(I-H)$ is idempotent,

$$\begin{aligned}(P\Lambda P^T)^2 &= P\Lambda P^T \\ P\Lambda P^T P\Lambda P^T &= P\Lambda P^T \\ P\Lambda^2 P^T &= P\Lambda P^T \\ \Lambda^2 &= \Lambda \\ \lambda_i &= \lambda_i^2\end{aligned}$$

It must then be the case that $\lambda_i \in \{0, 1\}$ for all i.

Then, going back once again, we see that $\sum_{i=1}^n \hat{\varepsilon}_i^2 = \varepsilon^T(I - H)\varepsilon = \varepsilon^T P\Lambda P\varepsilon$. Define

$$\eta = P^T\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

We can do this because we've previously seen that rotations on the normal distribution have no effect. Then,

$$\sum_{i=1}^n \hat{\varepsilon}_i^2 \quad = \quad \eta^T\Lambda\eta \quad = \quad \sum_{i=1}^n \eta_i^2\lambda_i \quad = \quad \sigma^2\lambda^2_{\sum\lambda_i} \quad = \quad \sigma^2\chi^2_{\text{rank}(I-H)} \quad = \quad \sigma^2\chi^2_{(n-p)}$$

Z has full rank p, H has full rank p, and $I-H$ has rank $n-p$. The eigenvalues of H are $1-\lambda_i \in \{0, 1\}$ when $I - H$ are $\lambda_i \in \{0, 1\}$.

As such, symmetry and idempotency give us the χ^2 distribution.

4.4 Applications

Say that $\beta \in \mathbb{R}^{p\times 1}$ and $c \in \mathbb{R}^{1\times p}$. We get that

$$\frac{c\hat{\beta} - c\beta}{s\sqrt{c(Z^TZ)^{-1}c^T}} \quad \sim \quad t_{(n-p)} \quad \text{where} \quad s^2 = \frac{1}{n-p}\sum_{i=1}^n \hat{\varepsilon}_i^2$$

Note that $s^2 \sim \frac{\sigma^2\chi^2_{(n-p)}}{n-p}$. The denominator is thus an estimate of standard deviation. Note that

$$\mathrm{V}(c\hat{\beta}) = c(Z^TZ)^{-1}c^T\sigma^2 \;\Rightarrow\; \widehat{\mathrm{V}}(c\hat{\beta}) = c(Z^TZ)^{-1}c^Ts^2 \quad \text{and} \quad \mathrm{E}[s^2] = \frac{1}{n-p}\mathrm{E}[\sigma^2\chi^2_{(n-p)}] = \sigma^2$$

Say that $c = \begin{bmatrix}0 & \dots & 0 & 1 & 0 & \dots & 0\end{bmatrix}$ where the jth entry is the only 1. Then, if we hypothesize that $\beta_j = 0$, then we get the t-statistic

$$\frac{\hat{\beta}_j - 0}{s\sqrt{c(Z^TZ)^{-1}c^T}}$$

4.4.1 Another Approach to the t-statistic

Let:

$$U = \frac{\hat{\beta}_j - 0}{\sigma\sqrt{c(Z^TZ)^{-1}c^T}} \sim \mathcal{N}(0,1) \qquad V = \frac{s^2}{\sigma^2} \sim \frac{\chi^2_{(n-p)}}{n-p} \qquad \text{where } U, V \text{ are independent}$$

Then,

$$\frac{U}{\sqrt{V}} \sim t_{(n-p)}$$

We get the t-distribution.

These results on the t-distribution are interesting. The t-statistic itself is actually an unknown value; since β is unknown, $c\beta - c\hat{\beta}$ is unknown, so the whole statistic is an unknown value. However, the distribution of this statistic *is* known, which allows us to quantify the unknown value.

4.5 Examples of Non-Uniqueness

4.5.1 The Dummy Variable Trap

Suppose that $y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$ (observations are ith observation in group j). If we include a dummy variable for all k groups in the data, we get a Z that is not full-rank, since the first column of Z, which is all 1's (for the intercepts) is the sum of the other columns. So, the rank of Z is k, but Z has $k+1$ columns. It is not invertible.

If this is the case, then

$$\begin{aligned} y_{ij} &= \mu + \alpha_j + \varepsilon_{ij} \\ &= (\mu + \lambda) + (\alpha_j - \lambda) + \varepsilon_{ij} \end{aligned}$$

So, any choice of λ wouldn't change the distribution. The left-hand side is never changed, which means we cannot identify γ, which means we cannot uniquely estimate β. ($\alpha_i \to \alpha_i + \lambda$ where

$\lambda \in \mathbb{R}$.) The solution is therefore non-unique. To resolve this, we can omit a column of data and get full rank.

However, note that $(\alpha_i - \lambda) - (\alpha_j - \lambda) = \alpha_i - \alpha_j$. So, we can still estimate differences between groups when the model is overparameterized. That's not to say that being overparameterized is good, though.

4.5.2 Correlated Data

If perfectly correlated data is put into an equation, the matrix Z will be rank deficient and not invertible, leading to infinite answers. For example, consider including both temperature in degrees Fahrenheit and in degrees Celsius into a model; they are perfectly correlated. Think of the line that the data sits on; a plane fitting that data would wobble, and all variations of the wobbling would be equally valid predictions.

The "kitchen sink" approach to models may also lead to rank-deficient matrices. For instance, including all monthly sales and the annual sales in the same regression (because you'd like to know what is significant) would not be a good idea.

To resolve this kind of issue, consider one of three methods:

- Get rid of variables until the matrix is full rank. (Easily the most common method.)
- Impose linear constraints like $\sum \alpha_i = 0$ or $\beta_7 = 0$. (The latter example is analogous to dropping that variable.)
- Choose the β that minimizes $\hat{\beta}^T\hat{\beta}$ such that $Z^TZ\hat{\beta} = Z^TY$. (That is, choose the shortest least squares solution.)

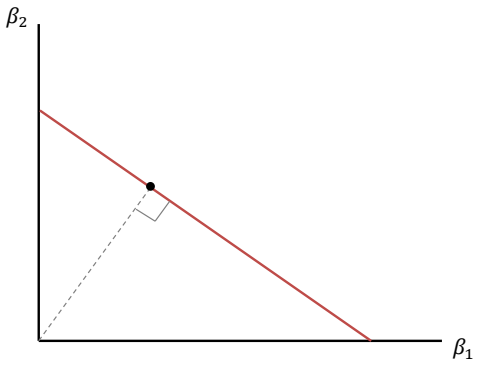


Figure 4.3: The blue line is the set of least square solutions. The dot (the perpendicular point) is the shortest least squares solution, and the most preferred.

4.6 Extra Sum of Squares

(See notes for detailed proofs.)

Our full model is specified as $Y = Z\beta + \varepsilon$ where $Z \in \mathbb{R}^{n\times p}$. Say that we take just a subset of q columns in Z and call it $\tilde{Z} \in \mathbb{R}^{n\times q}$ where $q < p$. Then we have a submodel specified as $Y = \tilde{Z}\gamma + \varepsilon$.

An analogous way to think of it is to set $\beta_j = 0$ or $\beta_j = \beta_k = 0$ (which "gets rid" of those variables by imposing the value zero on them).

Suppose we try fitting the model using both the full model and submodel in order to get $\hat{\beta}$ and $\hat{\gamma}$. To see how each one does, we find the residual sum of squares (RSS).

$$RSS_{\text{FULL}} = \sum_i (Y_i - Z_i\hat{\beta})^2$$
$$RSS_{\text{SUB}} = \sum_i (Y_i - \tilde{Z}_i\hat{\gamma})^2$$

Since the submodel has constraints imposed on it, it cannot possibly perform better than the full model. That is, we know that $RSS_{\text{SUB}} \geq RSS_{\text{FULL}}$. If the difference between the two measures is large, we probably don't trust the submodel. If the difference is small, the submodel may be a valid representation of the data. We need a standardized measure of this difference that takes into account (1) the variance of the data itself and (2) the amount of constraints imposed on the submodel. This results in the following statistic:

$$F = \frac{\frac{1}{p-q}(RSS_{\text{SUB}} - RSS_{\text{FULL}})}{\frac{1}{n-p}RSS_{\text{FULL}}}$$

The denominator, which is our estimate $s^2 = \hat{\sigma}^2$, "standardizes" the measure. $\frac{1}{p-q}$ accounts for the amount of constraints on the submodel. Unsurprisingly, the F statistic follows an F-distribution:

$$F \sim \text{F}_{p-q,n-p}$$

with the null hypothesis H_0 that the submodel is true.

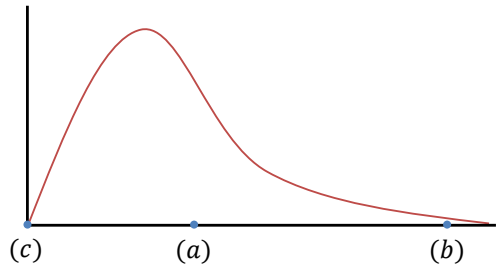


Figure 4.4: The F-distribution. At point (a), the submodel seems reasonable (we cannot reject the null). At point (b), the submodel is not reasonable (we reject the null. At point (c), things are too good. The data was probably preprocessed or manipulated.

4.7 Gauss Markov Theorem

This theorem justifies our use of the least squares estimate.

Theorem 2. Let $\text{E}[Y] = Z\beta$ and $\text{V}(Y) = \sigma^2 I$ where $Z \in \mathbb{R}^{n\times p}$ and rank $p < n$. Furthermore, let $\hat{\beta}$ be the least squares estimate.

If $a \in \mathbb{R}^n$ has $\text{E}[a^T Y] = c\beta$, then $\text{V}(a^T Y) \geq \text{V}(c\hat{\beta})$.

$c\hat{\beta} = (Z^TZ)^{-1}Z^TY$ is a linear combination of $\hat{\beta}$. The theory thus states that if you can find other unbiased estimators, their variances must be at least as large as $\hat{\beta}$. We should therefore use $\hat{\beta}$, the least squares estimator.

Also important is what is *not* mentioned in the theorem. The theorem is valid as long as Y has uncorrelated components, and the theorem does not require normality.

Proof. We only begin it here.

$$\mathrm{V}(a^TY) = V(c\hat{\beta} + a^TY - c\hat{\beta})$$

The proof follows by showing that $\mathrm{Cov}(c\hat{\beta}, a^TY - c\hat{\beta}) = 0$. Furthermore, a is subject to constraints that make this statement true (if there were no constraints, many values of a would lead to non-zero covariance). □

The key consequence of Gauss Markov Theorem: To beat the least squares estimate, you need *bias* or *non-normality*.

4.8 Computing LS Solutions

While we typically just use the computer to do estimates, we should understand why/how the computation works. We especially want to focus on doing these calculations using *Singular Value Decomposition*, or SVD.

Why do we want to know about SVD? At least three reasons.

1. SVD handles rank-deficient Z and also can help identify which covariates are causing the deficiency.
2. SVD is numerically stable. That is, it is not as subject to round-off error, as opposed to naively solving $Z^TZ\beta = Z^TY$ by Gaussian elimination. This is admittedly a very rare concern, but still a nice advantage.
3. SVD is commonly used in modern multivariate statistics for analysis.

4.8.1 The SVD

$$Z_{n\times p} = U_{n\times n}\Sigma_{n\times p}V^T_{p\times p}$$

where U, V are orthogonal, and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, ..., \sigma_k)$ where $k = \min(n, p)$ and $\sigma_i \geq 0$ for all i. (Note that we usually order the indices so that $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_{\min(n,p)} \geq 0$.)

Consider the case where $n > p$.

$$Z\beta = U\Sigma(V^T\beta)$$

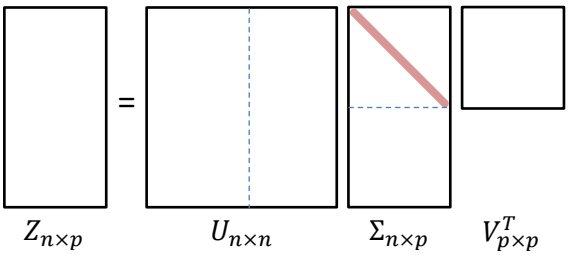


Figure 4.5: The normal ("fat") singular value decomposition. Note that the bottom portion of Σ and the right portion of U "drop out" of the analysis because Σ is populated with zeroes everywhere except on the diagonal.

Three steps are involved in the right-hand side, starting with β.

1. Multiplication by orthogonal matrix V^T "rotates" β.
2. Multiplication by Σ "stretches" components of $V^T\beta$ by the σ_i's.
3. Multiplication by U "rotates" $\Sigma(V^T\beta)$.

However, this structure is superfluous. All the zero rows in Σ (the rows not involved in the diagonal) simply multiply by zero. As a result, much of the right-hand side of U disappears during multiplication. We can therefore remove these parts and get the *skinny* SVD.

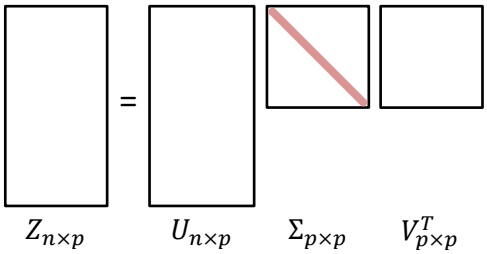


Figure 4.6: The skinny singular value decomposition.

We can go one step further in removing components in $\Sigma_{n\times n}$ where $\sigma_i = 0$ and adjusting accordingly.

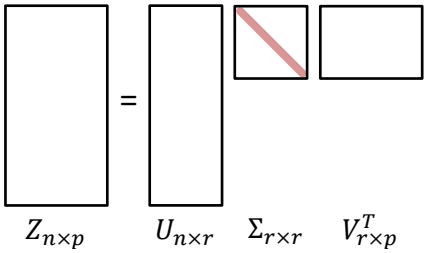


Figure 4.7: An even skinnier singular value decomposition.

The rank r of Z is the number of $\sigma_i > 0$. Note that $\sigma_p \leq \dfrac{||Z\beta||}{||\beta||} \leq \sigma_1$.

Furthermore, $Z^TZ = V\Sigma^T\Sigma V^T$, where the eigenvalues of Z^TZ are $\sigma_1^2, \sigma_2^2, ..., \sigma_p^2$.

Then, we get that

$$\begin{aligned}
||Y - Z\beta||^2 &= ||Y - U\Sigma V^T\beta||^2 \\
&= ||U^TY - \Sigma V^T\beta||^2 \\
&= ||Y^* - \Sigma\beta^*||^2 \qquad (\text{where } Y^* = U^TY, \beta^* = V^T\beta) \\
&= \sum_{i=1}^{r}(y_i^* - \sigma_i\beta_i^*)^2 + \sum_{i=r+1}^{p}(y_i^* - 0\beta_i^*)^2 + \sum_{i=p+1}^{n}(y_i^* - 0\beta_i^*)^2 \\
&= \sum_{i=1}^{r}(y_i^* - \sigma_i\beta_i^*)^2
\end{aligned}$$

This represents a diagonal matrix.

We find that β is a least squares solution if and only if

$$\beta_i^* = \begin{cases} \frac{y_i^*}{\sigma_i} & i = 1, ..., r \\ \text{anything} & i = r+1, ..., p \end{cases}$$

This describes a set of least squares solutions. When rank $Z < p$, the shortest solution is the one where $\beta_i^* = 0$ for $r + 1 \leq i \leq p$.

Then, $\hat{\beta} = V\beta^*$.

This process followed a four-step formula:

1. $Z = U\Sigma V^T$
2. $Y^* = U^TY$
3. $\beta_j^* = \begin{cases} \frac{y_j^*}{\sigma_j} & \sigma_j \neq 0 \text{ or } \sigma_j > \varepsilon\sigma_1 \\ 0 & \sigma_j = 0 \text{ or } \sigma_j \leq \varepsilon\sigma_1 \end{cases}$
4. $\hat{\beta} = V\beta^*$

What if there are thousands of Y vectors? Then, do the SVD of Z just once to get Z. After that, do steps 2-4 of the recipe for each vector. The cost of doing SVD calculations is $O(np^2)$ where n is the number of observations and p is the number of parameters.

4.9 R^2 and ANOVA Decomposition

Consider a surface $\{Z\beta|\beta \in \mathbb{R}^n\}$. Suppose that matrix Z has a first column of all 1's. On the surface is drawn a modeled line described by $\left\{ \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \mu \middle| \mu \in \mathbb{R}^n \right\}$ that attempts to capture the data.

How do we gauge how well the model accounts for the variation in the data? We use *an*alysis *o*f *va*riance, or ANOVA. See Figure 4.8.

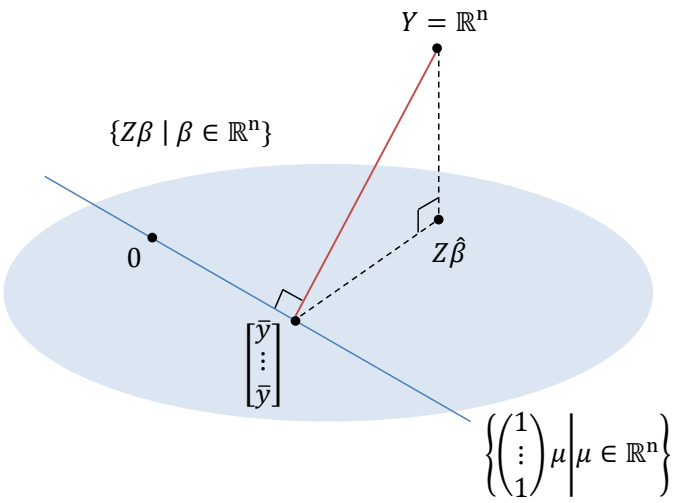


Figure 4.8: A visualization of ANOVA.

Use the Pythagorean Theorem to find the distance between the actual data Y and the closest point on the line. By Pythagorean Theorem,

$$\sum_{i=1}^{n}(y_i - \overline{Y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2$$

$$SST = SSE + SSZ$$

$$\text{Total sum of squares} \;=\; \text{Sum of squares of errors} \;+\; \text{Sum of "explained"}$$

Then,

$$R^2 = \frac{SSZ}{SST} = 1 - \frac{SSE}{SST}$$

So R^2 is a measure of the percentage of variance explained by the model. This is a very naive measure, since tossing in lots of covariates will necessarily make R^2 get closer to 1.

Next, we move away from the mathematics and focus more on actual statistics.

Chapter 5

Intro to Statistics

Let's start with the simplest case: a square Z and a population of Y_i values without X_i's. Assume that $Y_i \sim F$. That is, we are dealing with only one distribution, and we just care about $\mu = \mathrm{E}[Y]$.

The basic linear model here is $Y_i = \mu + \varepsilon_i$, where $Z = \begin{bmatrix} 1 & \dots & 1 \end{bmatrix}^T$ and $\beta = \mu \in \mathbb{R}^1$.

Say that we dove and caught 17 abalone and measured their ages. We care about ages because healthier populations have larger Y_i's while overfished populations have smaller Y_i's. Suppose we recorded the following:

$$\{15, 7, 9, 10, 7, 8, 20, 16, 9, 19, 14, 10, 11, 10, 10, 12, 7\}$$

5.1 Mean and Variance

What is the average age of the population? That is, we want $\mathrm{E}[Y]$, which is an unknown value. All we know (and our best guess) is the sample average, which is $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = 11.4$. How does this value relate to the population average? How good is our estimate? We need to determine the distribution of the data to know that.

If Y_i is i.i.d., then $\mathrm{V}(\overline{Y}) = \frac{\sigma^2}{n}$ where $\sigma^2 = \mathrm{V}(Y_i)$. However, σ^2 is an unknown value.

The natural estimate is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

There's a small issue here. The natural estimate minimizes variance by design, to the point where it does better than the actual population mean μ itself. The natural estimate is typically a bit too small. We adjust using an *unbiased* estimate:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \qquad \text{where} \qquad \mathrm{E}[s^2] = \sigma^2$$

In our abalone example, $s^2 = 16.4$. Therefore, $\mathrm{V}(\overline{Y}) = \hat{\mathrm{V}}(\overline{Y}) = \frac{1}{n}s^2 = 0.964$. So roughly speaking, our estimate of the population mean would be about

$$\overline{Y} \pm 2\sqrt{\hat{\mathrm{V}}(\overline{Y})} = 11.4 \pm 2\sqrt{0.964} \approx [9.5, 13.5]$$

5.2 A Staircase

We have an estimate of variance. But then how good is our guess of the variance? Is $\hat{\mathrm{V}}(\overline{Y})$ a good estimate of $\mathrm{V}(\overline{Y})$? That depends on $\mathrm{V}(\hat{\mathrm{V}}(\overline{Y}))$. We know that

$$\mathrm{V}(s^2) = \sigma^4 \left(\frac{2}{n-1} + \frac{\kappa}{n} \right)$$

where κ is kurtosis, which is the fourth moment (and looks at fatness of tails). We don't know κ, so we plug things in and get $\hat{V}(\hat{\mathrm{V}}(\overline{Y}))$.

But then how good of an estimate is that? We'd want to find $\mathrm{V}(\hat{\mathrm{V}}(\hat{\mathrm{V}}(\overline{Y})))$, $\mathrm{V}(...\hat{\mathrm{V}}(\hat{\mathrm{V}}(\overline{Y}))...)$ and so on. There is an infinite regress, where we estimate

$$\mathrm{V}(\hat{\mathrm{V}}^{(k)}(\overline{Y})) \qquad \text{using} \qquad \hat{\mathrm{V}}^{k+1}(\overline{Y})$$

Tukey called this the "staircase of inference." This problem is not good, nor is it fully resolvable. Most people just stop at the mean or variance and accept the estimates of those. If the data were perfectly normal, then μ and σ^2 can be found since higher moments are 0. But then if you want to test for normality of the data, you start going up the staircase once again. There is ultimately no way to avoid this problem; we cannot eliminate a twinge of doubt in any of our findings. We simply make some assumptions that sweep much of this under the rug.

5.3 Testing

Suppose we want to know whether the average age is actually less than 10. (It could be the case that above 10 is good, while below 10 is unhealthy.) Recall that we got a mean estimate of 11.4, but that the range was $[9.5, 13.5]$.

Let $\mu = \mathrm{E}[Y_i]$ and denote $\mu_0 = 10$. Then our *null hypothesis* is $H_0 : \mu = 10$. Meanwhile, our *alternative hypothesis* is $H_A : \mu \neq 10$. Or, we could do a one-tailed test where $H_A : \mu < 10$ or $H_A : \mu > 10$; or a test where $H_0 : \mu \leq 10$ and $H_A : \mu > 10$. It all depends on what you're looking for. If observed data is sufficiently unlikely under the null, we reject. If not, we fail to reject. (Note that not rejecting is not the same as accepting the null. An infinite range of values for μ would not be rejected, yet that does not mean they are all true.)

5.3.1 One Sample t-test

Assume $Y_1, ..., Y_n \sim \mathcal{N}(\mu, \sigma^2)$. We test the proposal $\mu = \mu_0$ using the t-statistic:

$$t = \sqrt{n}\frac{\overline{Y} - \mu_0}{s}$$

If $\mu = \mu_0$, then $t \sim t_{(n-1)}$. The t-distribution looks like the normal, though with slightly heavier tails. If the null is true, then our t-statistic is a sample from a common part of the distribution. If we get an extreme t-statistic, it's highly unlikely to have been observed if $\mu = \mu_0$. We reject $H_0 : \mu = \mu_0$ for unlikely t-values.

The t-statistic has a couple nice qualities. First, it is dimensionless. Second, we allows us to get a known distribution while dealing with an unknown value. It is a "pivotal quantity."

In our example, the $t = 1.438$. We fail to reject the null. However, $P(t = 1.438) = 0$, just like $P(t = 19.4) = 0$; yet we fail to reject the first and reject the second. Why is this reasonable? We care about *areas* of a distribution rather than points. Our tests are based on areas under the tails.

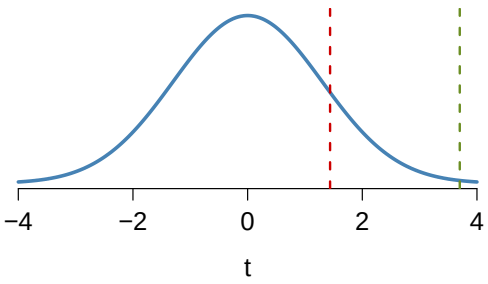


Figure 5.1: A t-distribution with the t-value from the abalone data in red. We fail to reject. If we saw a value such as the one in green on the right, we would reject the null hypothesis.

If $H_A : \mu \leq \mu_0$ (two-tailed test), reject if

$$p = P(|t_{n-1}| \geq t_{obs}) \text{ is small}$$

If $H_A : \mu > \mu_0$ (one-tailed test), reject if

$$p = P(t_{n-1} \geq t_{obs}) \text{ is small}$$

These probabilistic quantities are p-values. If the chance of getting what we saw under the null is tiny, we reject the null. Our interpretation is either that H_0 is false or a very rare event occurred.

For our abalone data, $p = P(|t_{16}| \geq 1.438) = 2 \times 0.085 = 0.17$. If we use the classic test at level $\alpha = 0.05$, we fail to reject the null. More generally, we reject the null if $p < \alpha$, where α is commonly 0.05,[1] but also 0.01 or 0.001.

[1]This is a bad norm. By the time you make choices about specifying your model, incentives exist to use models that have a low p. While this is true for any value of α, 0.05 allows for false findings 5% of the time. And if data

Reporting p-values is more informative than simply saying "$p < 0.05$."

One-Tailed Warning

A one-tailed test uses the following p-value:

$$p = P\left(t_{n-1} \geq \sqrt{n}\frac{\overline{Y} - \mu_0}{s}\right) = \frac{1}{2}P\left(|t_{n-1}| \geq \sqrt{n}\frac{\overline{Y} - \mu_0}{s}\right)$$

That is, the p value is half of that we'd see with a two-tailed test. Watch out when people report one-tailed tests to salvage statistical significance. Generally, we should use two-tailed tests. If one-tailed tests are not justified/specified ex ante, they are suspicious to use later.

A couple potential reasons may exist for using one-tailed tests.

1. Maybe $\mu < \mu_0$ is impossible.
2. Maybe $\mu > \mu_0$ is important to discover while $\mu < \mu_0$ has no normative value.

However, these are shaky reasons. (1) is tenuous because rare events do happen, and more often than one might think. (2) is suspect because the result $\mu < \mu_0$ may end up being useful. (Think about the accidental discovery of weak adhesives that were later used for sticky notes.) We may not want to apply normative judgements beforehand.

5.4 Practical vs. Statistical Significance

If $p \leq 0.001$, we say that "it is statistically significant at the 0.1% level."

Suppose that we want to analyze the strength/weight limits of cardboard boxes, where

$$\begin{aligned} Y &= \text{ strength of cardboard box in pounds} \\ \mu_0 &= 80 \text{ pounds} \\ \overline{Y} &= 80.2 \text{ pounds} \\ p &= 10^{-8} \end{aligned}$$

This is highly statistically significant, but the extra 0.2 probably is not a big deal. It is not practically significant. Practical significance is much harder to nail down than simply computing statistical significance. It depends greatly on context: What will the data be used for? What matters? These are *not* statistical issues.

Mechanically, statistical significance is more likely when n is large and/or s is small. Indeed, consider the following quotation:

> "p measures the sample size." - R. Olshen

is not normal (and it usually is not), chances are that false discoveries actually occur at the rate of 0.07 or 0.08. The 0.05 threshold was set by Ronald Fisher when working with limited agricultural data decades ago. With our computational power and access to data, we should set higher thresholds.

Theory states that $p \doteq e^{-k \times n}$, where n is the sample size and k depends on μ, μ_0, σ.

We can summarize significance using this table.

	Practically Significant	**Practically Insignificant**
Statistically Significant	Learning something useful	n is too large
Statistically Insignificant	n is too small	Maybe we should keep H_0...

5.5 Confidence Intervals

Briefly,

$$\begin{aligned}
0.95 &= P\left(|t| \leq t_{n-1}^{0.975}\right) \\
&= P\left(-t_{n-1}^{0.975} \leq t \leq t_{n-1}^{0.975}\right) \\
&= P\left(-t_{n-1}^{0.975} \leq \sqrt{n}\,\frac{\overline{Y}-\mu}{s} \leq t_{n-1}^{0.975}\right) \\
&= P\left(\underbrace{\overline{Y} - \frac{s}{\sqrt{n}}\,t_{n-1}^{0.975}}_{\text{Random } L} \leq \underbrace{\mu}_{\text{Fixed and unknown}} \leq \underbrace{\overline{Y} + \frac{s}{\sqrt{n}}\,t_{n-1}^{0.975}}_{\text{Random } U}\right)
\end{aligned}$$

A 95% confidence interval is the range such that $P(L \leq \mu \leq U) = 0.95$. It is a random interval that contains the true value 95% of the time. At least, that is how it works in theory; it is almost never that neat in reality.

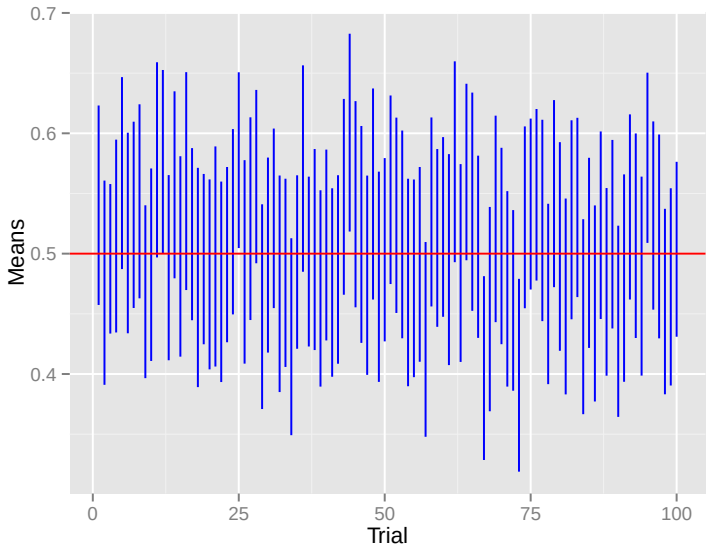


Figure 5.2: 100 confidence intervals. About 5% capture the true mean, which is 50% in this case.

Chapter 6

Power and Correlations

Recall the concept of hypothesis testing: We reject a null hypothesis H_0 (or fail to) at some level α.

- We report p-values: the smallest α at which rejection is possible. Usually, $p \sim U(0,1)$ under H_0.
- Confidence interval: $100(1-\alpha)\%$
  - Lower and upper bounds L and U are random, while μ is fixed.
  - $P(L \leq \mu \leq U) = 1 - \alpha$
  - Conveys practical significance.

Figure 6.1: Statistical significance and lack of statistical significance.

6.1 Significance

In theory, confidence intervals are nice and neat. With 95% confidence intervals, we expect 95% of the confidence intervals to include the true mean. But in reality, this is rarely the case. Suppose we measured the distance between the earth and the sun on occasion between 1900 and 1960. Of 14 estimates over these 60 years, only two overlap with μ. (See Figure 6.2.) Why is this the case? Because there are other non-statistical biases that statistical uncertainty does not capture. Maybe the equipment is faulty. Or maybe there is some human error involved. A couple examples by Ioannis illustrate these sorts of patterns where many intervals do not overlap with μ or even show systematic patterns over time. (See Figure 6.3.)

We can visualize statistically significant findings in journals and other publications with a *funnel plot*.

As the sample size of a study goes up, the ability to find smaller statistically significant effects increases. This is a function of the data (sample size), not because of some inherent property of

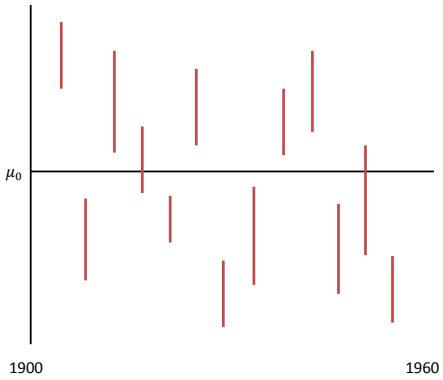


Figure 6.2: Examples of sun data.

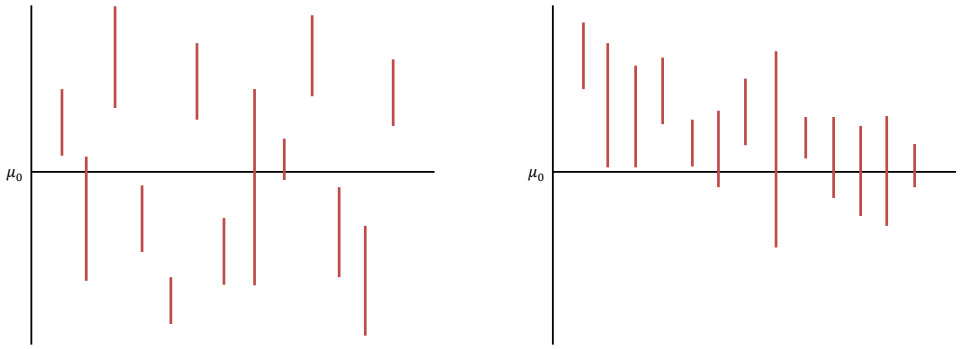


Figure 6.3: Examples of imperfect but more realistic confidence intervals.

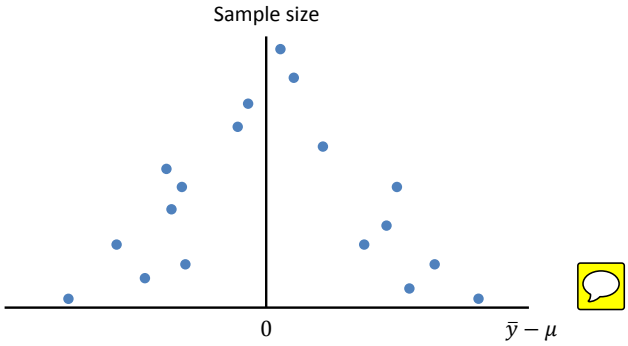


Figure 6.4: Funnel plot.

the things being studied. It also reflects an academic bias. As such, the math and empirics do not always line up. We may be capturing human/non-statistical aspects of studies.

It is important to note:

$$\text{Non-significance} \not\Rightarrow \text{Non-effect}$$

It may simply be that our sample size was too small to find the effect.

6.1.1 Finding Non-Significance

Suppose a drug maker wants to measure the potency of a generic medication. It is clearly in their interest for the generic medication to not find a statistically significant (negative) difference between the generic medicine and the name-brand medicine.

Say that the generic drug is deemed acceptable if its potency is within a 10% window of the name-brand drug. Also suppose that $\overline{Y}$, the found potency of the generic medicine, is within this window. This is insufficient ground to say the drugs are similar. The whole confidence interval for $\overline{Y}$ has to be inside that $\pm 10\%$ range.

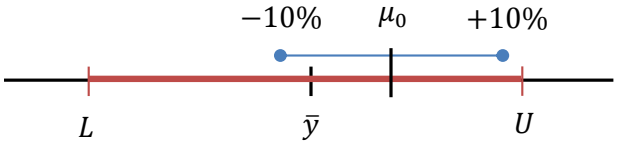


Figure 6.5: Indeterminate drug data.

6.1.2 Brief History of the t-test

William Gosset (who wrote under the pen name "Student") worked at Guinness. He knew that $\sqrt{n}\frac{\overline{Y}-\mu}{s}$ was only *approximately* normally distributed, but that this was especially tenuous for small n. He wanted a test statistic that would perform better (in order to evaluate the quality of Guinness stout). To do this, he simulated draws $X_1, ..., X_n$ using finger lengths of prisoners and guessed the statistic. Ultimately, he developed the following test statistic for small n:

$$\sqrt{\pi\, df}\frac{\Gamma\left(\frac{df+1}{2}\right)}{\Gamma\left(\frac{df}{2}\right)}\left(1+\frac{t^2}{df}\right)^{\frac{-df+1}{2}}$$

Later, Fisher proved this guess to be correct.

6.2 Power

Recall the standard t-test:

- $Y_1, ..., Y_n \sim \mathcal{N}(\mu, \sigma^2)$
- $H_0 : \mu = \mu_0$
- Reject if $|t| \geq t_{n-1}^{1-\frac{\alpha}{2}}$

Power (typically denoted $1-\beta$) is the chance of rejecting H_0. Formally,

$$\begin{aligned}
\text{Power} &= P\left(|t| > t_{n-1}^{1-\frac{\alpha}{2}}\right) = 1-\beta \\
&= P\left(t^2 > \left(t_{n-1}^{1-\frac{\alpha}{2}}\right)^2\right) \\
&= P\left(n\frac{(\overline{Y}-\mu_0)^2}{s^2} > F_{1,n-1}^{1-\alpha}\right) \\
&= P\left(\frac{\left(\frac{\sqrt{n}(\overline{Y}-\mu+\mu-\mu_0)}{\sigma}\right)^2}{\frac{s^2}{\sigma^2}} > F_{1,n-1}^{1-\alpha}\right)
\end{aligned}$$

The numerator is distributed $\mathcal{N}\left(\sqrt{n}\frac{\mu-\mu_0}{\sigma}, 1\right)^2$, which is a non-central χ^2 distribution. The denominator $\frac{s^2}{\sigma^2}$ is distributed $\frac{1}{n-1}\chi^2_{n-1}$. The numerator and denominator are independent. Putting these two results together, we get a non-central F-distribution.

$$1-\beta = P\left(F'_{1,n-1}\left(n\left(\frac{\mu-\mu_0}{\sigma}\right)^2\right) > F_{1,n-1}^{1-\alpha}\right)$$

We can visualize this below. At μ_0, we want a 5% chance of getting a false positive and a 95% of getting a significant result.

Note that β depends on five parameters: $\alpha, n, \mu, \mu_0, \sigma$. However, power is really dependent on the last three, which constitute the effect size (how many standard deviations apart the two means are). Specifically,

$$\Delta = \frac{\mu-\mu_0}{\sigma}$$

Clearly, we can detect bigger effects more easily. If we make the effect size (the non-centrality parameter $n\left(\frac{\mu-\mu_0}{\sigma}\right)^2$) twice as big, we only require a sample that is a quarter of the size.

6.2.1 Power Example

Suppose that we believe $\mu_0 = 10$ and have some reason to believe/guess that $\sigma = 5$. We are interested in whether $|\mu - \mu_0| \geq 1$. We want to find significance at the $\alpha = 0.05$ level, and also want to accept $\beta = 0.05$ false positives. Then, the non-centrality parameter is

$$\lambda = \left(\frac{\mu-\mu_0}{\sigma}\sqrt{n}\right)^2 = \frac{n}{25}$$

And power $= P(F'_{1,n-1}\left(\frac{n}{25}\right) > F^{0.95}_{1,n-1})$. The smallest n necessary to meet the conditions above ($\alpha = 0.05, \beta = 0.05$) is 327. In some applications, this may be a prohibitively high/costly number. So what if we were okay with 80% power ($\beta = 0.2$) instead? Then, $n = 199$. (Much as $\alpha = 0.05$ is an accepted standard for significance, $\beta = 0.2$ is a widely accepted standard for power.)

If your study can only afford $n = 30$, then power is 18.5% (there is only an 18.5% chance of rejecting the null). In that case, it may not be worth trying the study at all; just save your money.

6.3 Variance Estimation

Sometimes, we care more about σ^2 than about μ. A good example of this is quality control; we care more about consistency (informally speaking).

So, if $Y_i \sim \mathcal{N}(\mu, \sigma^2)$ and $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$, then $\mathrm{E}[s^2] = \sigma^2$ and $\mathrm{V}(s^2) = ... = \frac{2\sigma^4}{n-1}$.

Under the normality assumption, we can get a confidence interval for the variance estimate. Based on what we have above, $\sqrt{\mathrm{V}(s^2)} = \sigma^2\sqrt{\frac{2}{n-1}}$. Then, we can get the confidence interval from a pivot quantity.

$$P\left(\frac{s^2}{(\chi^2_{n-1})^{1-\frac{\alpha}{2}}/(n-1)} \leq \sigma^2 \leq \frac{s^2}{(\chi^2_{n-1})^{\frac{\alpha}{2}}/(n-1)}\right)$$

Note that the upper tail of χ^2 gives the lower bound of the confidence interval since the χ^2 value is in the denominator of the fraction. Also note that Central Limit Theorem does not apply here. While CLT helps to rescue estimates of μ when the data is non-normal (think of the kurtosis question on Problem Set 1), variances are another story.

If we denote the degrees of freedom as ν, then we can create the following table:

$\frac{s^2\nu}{(\chi^2_\nu)^{0.975}}$	ν	$\frac{s^2\nu}{(\chi^2_\nu)^{0.025}}$
$0.2s^2$	1	$1018.25s^2$
$0.27s^2$	2	$39.5s^2$
$0.32s^2$	3	$13.9s^2$
$0.36s^2$	4	$8.26s^2$
$0.39s^2$	5	$6.02s^2$

Table 6.1: Changes in the confidence interval for variance estimates with respect to degrees of freedom.

Obviously, fewer degrees of freedom make for worse estimates of s^2. At very low degrees of freedom, the confidence intervals are atrocious. (This is a serious issue for expensive studies where a single observation may cost large amounts of money.) Therefore, sample sizes (and thus ν) must be quite large to get estimates with a 10% or 5% range. Estimating variance is simply more difficult then estimating means.

Moreover, if $\kappa \neq 0$ (kurtosis exists), you are in deeper trouble. We need (1) additional assumptions,

(2) more data, or (3) bootstrapping methods to attempt to salvage variance estimates in the face of kurtosis. More formally,

$$\mathrm{V}(s^2) = \sigma^4 \left[\frac{2}{n-1} + \frac{\kappa}{n} \right]$$

6.4 Correlations in Y_i

Let $Y_i \sim (\mu, \sigma^2)$ (no assumption about distribution), but $\rho_{ij} = \mathrm{Corr}(Y_i, Y_j) \neq 0$ (and $|p| \leq 1$). There are at least a couple ways to account for this correlation.

6.4.1 Autoregressive Model

$$\rho_{ij} = \rho^{i-j} \qquad \text{and} \qquad Y_i = U_i + \gamma Y_{i-1}$$

The U_i's are independent. The model assumes that correlations between observations decrease exponentially with time, space, or whatever degree of "distance" separates the observations.

6.4.2 Moving Average Model

We may instead think that an observation at i may depend to some extent on only the previous observation's white noise; it is basically a "holdover" effect.

$$\rho_{ij} = \begin{cases} 1 & i = i \\ \rho & |i-j| = 1 \\ 0 & |i-j| > 1 \end{cases} \qquad \text{and} \qquad Y_i = U_i + \gamma U_{i-1}$$

In the moving average model,

$$\mathrm{E}[\overline{Y}] = \mu \qquad \mathrm{V}(\overline{Y}) = \ldots = \frac{\sigma^2}{n} \left[1 + 2\rho \frac{n-1}{n} \right] \qquad \mathrm{E}[s^2] = \ldots = \sigma^2 \left[1 - \frac{2\rho}{n} \right]$$

In the second expression (variance of $\overline{Y}$), if $\rho > 0$, then variance increases. In the third expression, $\rho > 0$ causes $\mathrm{E}[s^2]$ to go down. (As correlation between covariates increases, the two become more alike and the data loses some of its "variety.") Hence, correlations have opposite effects.

What about the test statistic?

$$t = \sqrt{n}\frac{\overline{Y} - \mu}{s} \to \mathcal{N}(0, 1 + 2\rho)$$

We get this result from Slutsky's Theorem, since $\sqrt{n}(\overline{Y} - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2(1 + 2\rho))$ and $s \xrightarrow{p} \sigma$.

So, for $\rho > 0$,

- The real distribution is wider than the typical t_{n-1} distribution
- Confidence intervals are too short by a factor of $\sqrt{1 + 2\rho}$ (fail to cover enough)
- p-values are too small, and we reject too many H_0's

And opposite effects if $\rho < 0$.

To make matters worse, correlations are hard to detect, especially in small samples. It could be that correlations undermine your analysis but may be too small to detect. Our assumption of independent observations may be far more dangerous than the normality assumption.

6.4.3 Non-Normality

Say that $Y_i \sim F$. What if the data is non-normal? What happens to the t-statistic? (With normal data, μ, σ^2 may be $\neq 0$ but $\gamma, \kappa = 0$. Not when non-normal.) Recall that

$$t = \sqrt{n}\frac{\overline{Y} - \mu}{s}$$

In the 1940s, they wrestled with this using brute-force math.

$$\text{Corr}(\overline{Y}, s^2) \to \frac{\gamma}{\sqrt{\kappa + 2}}$$

Why is this the case? If data shifts up, both $\overline{Y}$ and s^2 go up. If data shifts down, $\overline{Y}$ goes down but s^2 still goes up.

So how can we figure out what t is? Formulas exist for the mean and variance of this statistic.

$$\text{E}[t] \doteq -\frac{\gamma}{2\sqrt{n}}$$

The t-statistic skews in the opposite direction because the increase in s^2 outpaces any changes in $\overline{Y}$.

$$\text{V}(t) \doteq 1 + \frac{1}{n}\left(2 + \frac{7}{4}\gamma^2\right)$$

Variance does not move around quite as much as the mean, since variance involves division by n instead of $\sqrt{n}$. Note that κ (kurtosis) does not enter into the calculation of variance or mean of t; it all depends only on skewness of the data.

$$\text{Skew}(t) = -\frac{2\gamma}{\sqrt{n}}$$

Again, this expression goes in the opposite direction and disappears with n. All three expressions involve division by some form of n, meaning that they all disappear with increasing n as a consequence of CLT.

Then, in the 1980s, Peter Hall proved some results about coverage.

- For central tests:

$$P(|t| \leq t^{1-\frac{\alpha}{2}}) = 1 - \alpha + O(n^{-1})$$

- For one-sided tests:

$$P(t \leq t^{1-\alpha}) = 1 - \alpha + O(n^{-1/2})$$
$$P(t \geq t^{\alpha}) = 1 - \alpha + O(n^{-1/2})$$

With large enough n, the "unbelievable" assumption of normality becomes less relevant because of CLT. It is the other assumptions that prove to be more dangerous. (The discussion of correlation from before is a key example.)

6.4.4 Outliers

We can sometimes encounter rare huge values, either as a consequence of errors or a heavy-tailed distribution. A really high value would drag the sample mean up, or a really low value could pull it down. Do we keep the value or not?

In the second case, outliers are an extreme form of non-normality. We will look at this more later.

Chapter 7

Two-Sample Tests

7.1 Setup

Now, we move on the $k = 2$ populations, and 2 samples.

Say that $X \in \{0, 1\}$, or $\{1, 2\}$, or whatever binary form we wish to use. Then, for $X \in \{0, 1\}$,

$$Z = \left[\begin{array}{cc} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \hline 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{array}\right] \qquad \beta = \begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix} = \begin{pmatrix} \mathrm{E}[Y|X = 0] \\ \mathrm{E}[Y|X = 1] \end{pmatrix}$$

That is, we assign either a 0 or 1 to each of the two groups. Each group has a total of n_0 or n_1 observations, constituting a total of $n = n_0 + n_1$.

However, there is another way of capturing two different groups. Say that we have the following Z:

$$Z = \left[\begin{array}{cc} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \hline 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{array}\right]$$

where Group 0 gets 1's in both columns and Group 1 gets 1's in just one column. Then, $\beta_1 = \mathrm{E}[Y|X = 1] - \mathrm{E}[Y|X = 0]$, with the usual null hypothesis that $H_0 : \beta_1 = 0$.

In this case, we get a covariance matrix where $Z^T Z = \begin{pmatrix} n & n_1 \\ n_1 & n_1 \end{pmatrix}$. The (2,2) element in that matrix

matters most to us here, because it is part of our t-statistic in the two-sample case:

$$t = \frac{\hat{\beta}_1 - 0}{s\sqrt{(Z^T Z)^{-1}_{22}}} = ... = \frac{\overline{Y}_1 - \overline{Y}_0}{s\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}}$$

We can only estimate variance, s:

$$\begin{aligned} s^2 &= \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \\ &= \frac{1}{n-2}\left[\sum_{i:X_i=0}(Y_i - \hat{\beta}_0)^2 + \sum_{i:X_i=1}(Y_i - \hat{\beta}_0 - \hat{\beta}_1)^2\right] \\ &= \frac{1}{n-2}\left[\sum_{i:X_i=0}(Y_i - \overline{Y}_0)^2 + \sum_{i:X_i=1}(Y_i - \overline{Y}_1)^2\right] \end{aligned}$$

But there could be an issue here. We assumed that $Y = Z\beta + \varepsilon$ and that $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. However, we could have that $\mathrm{V}(Y|X=0) \neq \mathrm{V}(Y|X=1)$. Or it could be the case that $Y_{ij} \sim \mathcal{N}(\mu_j, \sigma_j^2)$ for $j = 0, 1$ and $i = 1, ..., n_j$. The two sources of data may be completely different, so it is easy to believe that their variances could be completely different too. What can we do about that?

In an ideal world,

$$\mathrm{V}(\overline{Y}_1 - \overline{Y}_0) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}$$

We know our estimate, however, is

$$s^2 = \frac{1}{n-2}\left[\sum_{i:X_i=0}(Y_i - \overline{Y}_0)^2 + \sum_{i:X_i=1}(Y_i - \overline{Y}_1)^2\right]$$

Then,

$$\mathrm{E}[s^2] = \frac{1}{n-2}\left[(n_0 - 1)\sigma_0^2 + (n_1 - 1)\sigma_1^2\right]$$

And

$$t = \frac{\overline{Y}_1 - \overline{Y}_0 - \Delta}{s\sqrt{\frac{1}{n_0} + \frac{1}{n_1}}}$$

where Δ is from the null hypothesis. So, actual variances are proportional to inverse sample sizes. The estimate of s^2, however, is proportional to sample size. This seems concerning, unless the two sample sizes are identical (in which case proportionality based on sample size or inverse sample size does not matter). But what about the worse cases?

If our estimate of variance is good, then it should be that

$$\frac{\mathrm{V}(\overline{Y}_1 - \overline{Y}_0 - \Delta)}{E\left[s^2\left(\frac{1}{n_0} + \frac{1}{n_1}\right)\right]} \approx 1$$

Suppose that $\sigma_1^2 > \sigma_0^2$ and $n_1 < n_0$. (Say that $n_1 = Rn_0$ and $\sigma_1^2 = \theta\sigma_0^2$ for $\theta > 1, R < 1$.) How is our estimate of variance affected?

$$\frac{\mathrm{V}(\overline{Y}_1 - \overline{Y}_0 - \Delta)}{E\left[s^2\left(\frac{1}{n_0} + \frac{1}{n_1}\right)\right]} = \frac{\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}}{\left(\frac{1}{n_0} + \frac{1}{n_1}\right)\left(\frac{(n_1-1)\sigma_1^2 + (n_0-1)\sigma_0^2}{n_1+n_0-2}\right)} \approx \frac{\theta + R}{R\theta + 1}$$

How do we apply this? Suppose that we knew $\theta = 2, R = \frac{1}{2}$. Then, we get that

$$\frac{2 + \frac{1}{2}}{1 + 1} = \frac{\frac{5}{2}}{2} = \frac{5}{4}$$

This result has several implications.

- s^2 is too small by $\frac{5}{4}$.
- s is too small by $\sqrt{\frac{5}{4}}$.
- Our confidence intervals are too short by $\sqrt{\frac{5}{4}}$.
- Our p-values will be too small.
- We will reject the null hypothesis too often, even as $n \to \infty$.

7.2 Welch's t

Efforts have been made to overcome this issue. One well-known case is Welch's t, which is defined as:

$$t' = \frac{\overline{Y}_1 - \overline{Y}_0 - \Delta}{\sqrt{\frac{s_1^2}{n_0} + \frac{s_1^2}{n_1}}} \qquad \text{where} \qquad s_j^2 = \frac{1}{n_j - 1}\sum_{j=1}^{n_j}(Y_{ij} - \overline{Y}_j)^2$$

Welch's t has an asymptotic result: $t' \to \mathcal{N}(0,1)$ if $\mu_1 - \mu_0 = \Delta$ and $\min n_j \to \infty$. That is nice, but Welch's t also has no t distribution. What do we do with finite, smaller sample sizes? How do we know the degrees of freedom to choose in evaluating the test statistic?

We match the second and fourth moments of $\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}$ to $\frac{\sigma^2}{\nu}\chi^2_{(\nu)}$. This sounds awful, but it actually works out nicely. Satterthwaite suggested the following:

$$\nu = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{\sigma_1^2}{n_1}\right)^2 + \frac{1}{n_0 - 1}\left(\frac{\sigma_0^2}{n_0}\right)^2}$$

If we plug in s_1^2, s_0^2, we yield a new $\hat{\nu}$. This is the appropriate degrees of freedom that should be used when we then look for the appropriate $t(\hat{\nu})$.

Note that $\hat{\nu}$ falls within a certain range:

$$\min n_j - 1 \leq \hat{\nu} \leq n_0 + n_1 - 2$$

The left-hand term is the correct degrees of freedom if $\frac{\sigma_j^2}{\sigma_{1-j}^2} = \infty$ (that is, one has tons of variance while the other has basically none). The right-hand term is correct if $\sigma_0^2 = \sigma_1^2$.

This is all well and good, but it has a very small substantive effect on our estimates. The t-statistic will most likely be between 1.96 (the ideal result) and 2. So in the end, this is really not a big deal.

7.3 Permutation Tests

The two-sample case is the first instance where we can consider doing permutation tests. Say that $Y_i \overset{\text{ind.}}{\sim} F_j$ for $j = 0, 1$ and $i = 1, ..., n_j$. Suppose the null hypothesis is that the two samples are from the same distribution: $H_0 : F_0 = F_1$. Then, the data could be generated via

1. Sample n observations from same distribution F for $n = n_0 + n_1$
2. Randomly pick n_0 from Group 0, and pick n_1 from Group 1

The second option could be done in $\binom{n_1 + n_0}{n_0}$ ways. So, try computing $\overline{Y}_1 - \overline{Y}_0$ under all permutations.

$$p = \frac{\text{number of permutations with } (\overline{Y}_1 - \overline{Y}_0) \geq |\text{observed } \overline{Y}_1 - \overline{Y}_0|}{\binom{n_1 + n_0}{n_0}}$$

If the permutation value is too large, then sample N permutations at random, and then:

$$p = \frac{1 + \text{number of sampled permutations with } (\overline{Y}_1 - \overline{Y}_0) \geq |\text{observed } \overline{Y}_1 - \overline{Y}_0|}{N + 1}$$

Asymptotically, if you run the permutation test enough times, it approaches the two-sample t-test.

Permutation tests apply to very general statistics—not just $\overline{Y}_1 - \overline{Y}_0$. We could look at medians, interquartile ranges, etc. That is one strength of these tests. However, in testing that two distributions are exactly equal, it is *not* that $H_0 : \mathrm{E}[Y_1] = \mathrm{E}[Y_0]$. The permutation test can only test that the distributions are *identical*, and this is totally different from testing that two samples have identical means.

Chapter 8

k Groups

Once we enter territory where $k \geq 3$, it all pretty much becomes the same.

8.1 ANOVA Revisited

Suppose there are k groups. Say that $\mathrm{E}[Y_{ij}] = \mu_i \in \mathbb{R}$ for $j = 1, ..., n_i$.

The null hypothesis is now $H_0 : \mu_1 = \mu_2 = ... = \mu_k$.

- For $k = 2$, we use $\mu_2 = \mu_1$.
- For $k = 3$, we use $\mu_2 - \mu_1 = \mu_3 - \mu_2 = 0$.
- We have to make $\binom{k}{2}$ comparisons.

8.1.1 Cell Means Model

We use the means model, which we have seen before:

$$Z = \begin{bmatrix} 1 & & \\ \vdots & & \\ 1 & & \\ & 1 & \\ & \vdots & \\ & 1 & \\ & & \ddots \\ & & 1 \\ & & \vdots \\ & & 1 \end{bmatrix} \qquad \beta = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_k \end{bmatrix}$$

The null hypothesis here is that

$$H_0 : C\beta = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{k-1}$$

where

$$C = \begin{bmatrix} 1 & -1 & 0 & ... & 0 \\ 1 & 0 & -1 & ... & 0 \\ \vdots & & \ddots & \ddots & \\ 1 & & & 0 & -1 \end{bmatrix} \in \mathbb{R}^{(k-1)\times k} \qquad \text{or} \qquad C = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & 1 & -1 & \\ & & & & \ddots \end{bmatrix}$$

In the cell means model, $\hat{\mu}_i = ... = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \equiv \overline{Y}_{i\cdot}$.

8.1.2 Effects Model

We can also use the effects model.

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \sum \alpha_i = 0$$

8.2 Hypothesis Testing

Let's test that $C\beta = 0$. We can try it the hard way using brute force algebra and distributional theory. That process gives us:

$$\frac{\frac{1}{r}(\hat{\beta}-\beta)^T C^T (C(Z^TZ)^{-1}C^T)^{-1} C(\hat{\beta}-\beta)}{s^2} \sim F_{r,N-k}$$

where $r = \text{rank}(C) = k-1$. Note that scaling C by some constant does not affect this statistic, since the constant would cancel out.

The easy way is to fit the model assuming H_0, then try fitting *not* assuming H_0. Then, compare the sum of squared errors.

- H_0 : all $\hat{\mu}_i = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij} \equiv \overline{Y}_{\cdot\cdot}$ (where $N = \sum_i n_i$)
- This results in an F distribution:

$$F = \frac{\frac{1}{r}(SS_{\text{SUB}} - SS_{\text{FULL}})}{\frac{1}{n-k}SS_{\text{FULL}}} \sim F_{k-1,N-k}$$

The total sum of squares in this $k \geq 3$ ANOVA can be found using the ANOVA identity (which is a Pythagorean relation):

$$\sum_i \sum_j (Y_{ij} - \overline{Y}_{\cdot\cdot})^2 = \sum_i \sum_j (Y_{ij} - \overline{Y}_{i\cdot})^2 + \sum_i \sum_j (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2$$

This is $SS_{\text{TOTAL}} = SS_{\text{WITHIN}} + SS_{\text{BETWEEN}}$. The total sum of squares equals sum of squares within group i plus and sum of squares between all groups.

Then, the F test is 

$$F = \frac{\frac{1}{k-1} SS_{\text{TOTAL}} - SS_{\text{WITHIN}}}{SS_{\text{WITHIN}}} = \frac{\frac{1}{k-1} SS_{\text{BETWEEN}}}{\frac{1}{N-k} SS_{\text{WITHIN}}} \sim F'_{k-1,N-k}(\lambda) \qquad \lambda = \frac{1}{\sigma^2} \sum_{i=1}^{k} n_i \alpha_i^2$$

Hence, everything only depends on the right-hand terms: sum of squares within and sum of squares between. (Remember why we include the fractions $\frac{1}{k-1}$ and $\frac{1}{N-k}$ in these expressions: they help capture accurate amount of noise by counteracting the additional number of observations.)

The smaller the variance is, the more power we get. We also get more power from larger sample sizes and larger α. More specifically, note:

$$\frac{1}{\sigma^2} \sum_{i=1}^{k} n_i \alpha_i^2 = N \times \frac{1}{\sigma^2} \times \sum_{i=1}^{k} \frac{n_i}{N} \alpha_i^2 = N \times \text{ ratio of variances}$$

$\frac{1}{\sigma^2}$ speaks to overall variance. Meanwhile, $\sum \frac{n_i}{N}$ captures the fraction of the data that is in group i (the probability of being in group i), and α_i^2 captures the "deviation" of group i. Put together, $\sum \frac{n_i}{N} \alpha_i^2$ gives the variance of group i. So, if $\sum \frac{n_i}{N} \alpha_i^2 > \sigma^2$, then the ratio is greater than 1 and we can detect the variance of group i.

8.3 Lab Reliability

Suppose we manufacture allergy medicine and want to know how much antihistamine is in each pill. We send ten pills to seven labs and have them measure how much antihistamine is in each pill.

It could be the case that our pills have slightly varying levels of active ingredient in them. (That's likely.) But it is also possible that the seven labs are measuring the pills differently. We went to test the null hypothesis that there is no significant difference in measurements across all the labs—that variation between them is most likely due to chance.

Say that we get the following data from the seven labs.

To evaluate our hypothesis, we make an ANOVA table.

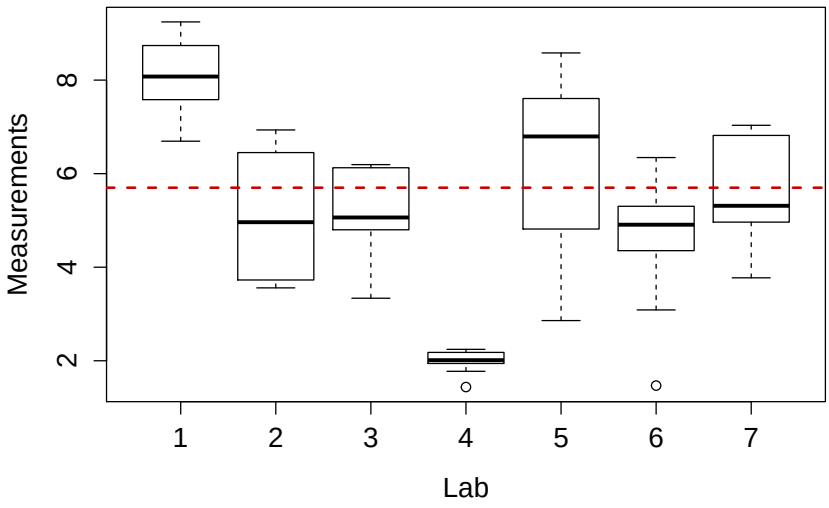


Figure 8.1: Boxplots of measurements, with actual value represented by the dotted line.

Source	df	SS	MS	F	p
Lab (error between labs)	6	0.125	0.021	5.66	< 0.005
Error (error within labs)	63	0.231	0.0037		
Total	69	0.356			

MS stands for mean square, which is $\frac{SS}{df}$. We determine F by the equation $F = \frac{MS_{\text{GROUP}}}{MS_{\text{ERROR}}}$.

Note the importance of the degrees of freedom here. With between-lab error, there are six "chances" to catch additional noise; there are 63 with within-lab error. Therefore, if error was purely random, we would expect the sum of squared errors (SS) to be about 10.5 times higher for within-lab error than between-lab error. However, SS_{ERROR} is less than double SS_{LAB}. This suggests that the error is not purely random.

Furthermore, with purely random errors, we would expect the two MS values to be equal. But MS_{LAB} is more than five times greater; we get more than five times the variation between labs than we would expect if there was only noise. The high F-value and correspondingly low p-value help us reject the null hypothesis. The labs are measuring differently.

MS speaks to statistical significance, while SS speaks to practical significance. F and p speak to whether or not the results are statistically significant.

8.4 Contrasts

So, which groups/treatments differ? We could use a t-test. For example, we can look for differences between groups 1 and 2:

$$\frac{\overline{Y}_{1\cdot} - \overline{Y}_{2\cdot}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{(N-k)}$$

where s comes from the *pooled* $s^2 = MSE = \frac{SSE}{N-k}$. But, there is something odd about this statistic. Simply by adding more groups that affect the pooled variance, we can affect the findings we get about groups 1 and 2. (This can go either way; either something that wasn't statistically significant before becomes significant, or vice versa, depending on what kind of additional variance information additional groups give.) This effect can be either good or bad.

On the other hand, we can use *contrasts*. Suppose we want to test the effectiveness of three detergents with phosphates against four detergents without phosphates. Our null is that phosphate detergents are no different from non-phosphate detergents. Then, our test statistic is

$$\frac{\frac{\overline{Y}_{1\cdot}+\overline{Y}_{2\cdot}+\overline{Y}_{3\cdot}}{3} - \frac{\overline{Y}_{4\cdot}+\overline{Y}_{5\cdot}+\overline{Y}_{6\cdot}+\overline{Y}_{7\cdot}}{4}}{s\sqrt{\frac{1}{9}\left(\frac{1}{n_1}+\frac{1}{n_2}+\frac{1}{n_3}\right) + \frac{1}{16}\left(\frac{1}{n_4}+\frac{1}{n_5}+\frac{1}{n_6}+\frac{1}{n_7}\right)}} \sim t$$

Or, say we are comparing a single product with some "treatment" against four products without that treatment. Then,

$$\frac{\overline{Y}_{1\cdot} - \frac{1}{4}(\overline{Y}_{2\cdot} + \overline{Y}_{3\cdot} + \overline{Y}_{4\cdot} + \overline{Y}_{5\cdot})}{s\sqrt{n_1 + \frac{1}{16}\left(\frac{1}{n_2}+\frac{1}{n_3}+\frac{1}{n_4}+\frac{1}{n_5}\right)}} \sim t$$

So, we have that

$$\frac{\sum_{i=1}^k \lambda_i \overline{Y}_{i\cdot}}{s\sqrt{\sum_{i=1}^k \frac{\lambda_i^2}{n_i^2}}} \sim t \qquad \text{where} \qquad \sum_{i=1}^k \lambda_i = 0 \quad \text{and} \quad \sum_{i=1}^k \lambda_i^2 \neq 0$$

With contrasts, the numerator is $\sum_i \lambda_i \overline{Y}_{i\cdot}$ for a sample. For the population, it would be $\sum_i \lambda_i \mu_i = \sum_i \lambda_i(\mu + \alpha_i) = \sum_i \lambda_i \alpha_i$.

8.4.1 Another Example

When $k = 2$, there is only one contrast to be done. However, when $k \geq 3$, the number grows substantially.

Say we perform an experiment with potatoes. We may or may not treat a potato field with potassium, and we may or may not treat a potato field with sulfur. There are therefore four different treatment groups in this experiment, and three contrasts.

	Potassium	No Potassium
Sulfur	$\overline{Y}_1$	$\overline{Y}_2$
No Sulfur	$\overline{Y}_3$	$\overline{Y}_4$

We can examine several effects.

- For the effect of sulfur, look at $\overline{Y}_1 + \overline{Y}_2 - \overline{Y}_3 - \overline{Y}_4$.
- For the effect of potassium, look at $\overline{Y}_1 + \overline{Y}_3 - \overline{Y}_2 - \overline{Y}_4$.
- For the interaction (doing nothing or both; synergy vs. dis-synergy), look at $\overline{Y}_1 - [\overline{Y}_4 + (\overline{Y}_2 - \overline{Y}_4) + (\overline{Y}_3 - \overline{Y}_4)] = \overline{Y}_1 - \overline{Y}_2 + \overline{Y}_3 + \overline{Y}_4$.

Effect	1	2	3	4
Sulfur	+	+	−	−
Potassium	+	−	+	−
Interaction	+	−	+	−

8.4.2 The Most Sensitive Contrast

What contrast is most likely to help reject the null? (What contrast provides the most power?) If we square the t-statistic and do some algebra, we find:

$$t^2 = \frac{(\sum_i \lambda_i \overline{Y}_{i\cdot})^2}{s^2 \sum_i \frac{\lambda_i^2}{n_i}} \sim \ldots = F'_{1,N-k}\left(\frac{(\sum_i \lambda_i \mu_i)^2}{\sigma^2 \sum_i \frac{\lambda_i^2}{n_i}}\right) = F'_{1,N-k}\left(\frac{(\sum_i \lambda_i \alpha_i)^2}{\sigma^2 \sum_i \frac{\lambda_i^2}{n_i}}\right)$$

Therefore, the bigger the ratio, the more power we have. So how do we get the most power given these conditions, and the constraint that $\sum_i \lambda_i = 0$?

When n_i's are all equal, then the most powerful contrast λ is the one that maximizes $\frac{\sum_i \lambda_i \alpha_i}{\sigma^2 \sum_i \frac{\lambda_i^2}{n}}$ by taking $\lambda \propto \alpha$. So, we would use $\lambda_i = \alpha_i$, or $\lambda_i = c \cdot \alpha_i$ where $c \neq 0$. This is just the treatment pattern. The most powerful contrast hence looks like a cell mean or effect pattern.

So, suppose we have $k = 4$. Then, to make the best contrast, we use

$$\lambda = \begin{bmatrix} -3 & -1 & 1 & 3 \end{bmatrix} \quad \Rightarrow \quad (-3)\overline{Y}_{1\cdot} + (-1)\overline{Y}_{2\cdot} + (1)\overline{Y}_{3\cdot} + (3)\overline{Y}_{4\cdot}$$

What if $k = 5$? Then,

$$\lambda = \begin{bmatrix} -2 & -1 & 0 & 1 & 2 \end{bmatrix}$$

This ensures that the sums of the λ_i's add up to zero. However, it also means that the middle group gets no weight in the analysis, which seems odd.

What if we thought there should be a quadratic relationship? Then,

$$\lambda = \begin{bmatrix} 2 & -1 & -2 & -1 & 2 \end{bmatrix}$$

because

$$\lambda \propto (i-3)^2 - \frac{1}{5}\sum_{j=1}^{n}(j-3)^2$$

(The logic behind these choices of λ is that the strongest contrast is proportional/linear in nature. So, if we describe a line as $m(x-\bar{x})$, then $\lambda \propto m(x-\bar{x}) \propto (x-\bar{x})$.)

8.5 Some Recap

When thinking about k different groups, we care about contrasts. Sample contrasts are

$$C_1 = \sum \lambda_i \overline{Y}_{1\cdot}$$
$$C_2 = \sum \eta_i \overline{Y}_{i\cdot}$$

where $\sum \lambda_i = \sum \eta_i = 0$. The population contrasts are then $\sum \lambda_i \alpha_i$ and $\sum \eta_i \alpha_i$.

Under H_0, we can do t-tests using the assumptions that

$$C_1 \sim \mathcal{N}\left(0, \sigma^2 \sum \frac{\lambda_i^2}{n_i}\right)$$
$$C_2 \sim \mathcal{N}\left(0, \sigma^2 \sum \frac{\eta_i^2}{n_i}\right)$$

We know that

$$\text{Cov}(C_1, C_2) = \sigma \sum_{i=1}^{n} \frac{\lambda_i \eta_i}{n_i}$$

If covariance equals 0, then the contrasts C_1 and C_2 are independent. That is, orthogonal contrasts are such that $\sum_{i=1}^{k} \frac{\lambda_i \eta_i}{n_i} = 0$. If so, then we can test the two things separately. For example, suppose C_1 is linear and C_2 is quadratic. If the two are orthogonal, whether a linear relationship exists is independent of whether a quadratic relationship exists. This is very different from a case where $\text{Corr}(C_1, C_2) = 0.8$, and correlation between the two is too high to make any meaningful separate tests.

We can design contrasts to be orthogonal. Our choice of contrasts also depends on what we want to test.

8.5.1 The "Cheat" Contrast

There is one special contrast that we could consider a "cheat" contrast. If any contrast would reject the null hypothesis, this would be it. (It is considered a "cheat" because we use the data to determine the values of the contrast instead of choosing the contrasts ex ante.)

In this contrast,

$$\lambda_i = \overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot} \qquad \text{when} \qquad n_i = n,\ i = 1, ..., k$$

This contrast maximizes

$$t^2 = \frac{\left(\sum_i \lambda_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})\right)^2}{s^2 \sum_i \frac{\lambda_i^2}{n}}$$

because we pick the λ_i's to be proportional to the $\overline{Y}_{i\cdot}$'s. We can simplify this expression further.

$$\begin{aligned} t^2 = \frac{\left(\sum_i \lambda_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})\right)^2}{s^2 \sum_i \frac{\lambda_i^2}{n}} &= \frac{(\sum_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})\overline{Y}_{i\cdot})^2}{\frac{s^2}{n}\sum_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2} \\ &= \frac{n(\sum_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2)^2}{s^2 \sum_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2} \\ &= \frac{n\sum_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2}{s^2} \\ &= \frac{SS_{\text{BETWEEN}}}{\text{MSE}} \end{aligned}$$

We know the distribution of this. $t^2 \sim (k-1)F_{k-1,N-k}$. (We get the $(k-1)$ because $\frac{SSE_{\text{BET}}}{k-1}$ gives us a MSE, and the quotient of two MSE's gives us the F-distribution.) Note that this is *not* $F_{1,N-k}$, which is what we use with pre-stated contrasts.

This test statistic rejects more often than it should. It accounts for what used to be called maximal "data snooping," leveraging the data as much as possible.

8.6 Multiple Comparisons

This only applies when $k \geq 3$.

When we reject H_0, we don't directly know which groups differ. How can we find out which groups are motivating rejection of the null? There are multiple potential approaches, each with their shortcomings.

8.6.1 *t*-tests

We could use a standard t-test, where we find $|\overline{Y}_{i_1\cdot} - \overline{Y}_{i_2\cdot}|$ and refer to $t_{N-k}^{1-\frac{\alpha}{2}} s\sqrt{\frac{1}{n_{i_1}} + \frac{1}{n_{i_2}}}$.

However, if we did this for every pair within the k groups, we require $\binom{k}{2}$ tests. These additional tests inflate the possibility of finding false positives. For example, if $k = 10$, there are 45 chances to reject the null, so $\mathbb{P}(\text{reject}) > \alpha$. (We would expect at least two false positives on average here.)

8.6.2 Fisher's Least Significant Difference (LSD)

R.A. Fisher had his own suggestion, which is not a very good idea. Still, it's instructive.

Run an F-test on H_0.

- If you can reject, then do all $\binom{k}{2}$ tests.
- If you cannot reject, stop testing.

This is potentially better than just doing t-tests since the F-test ensures that there's only a 5% chance of false discovery. However, it is also possible that you can reject via the F-test but that no single t-test turns out to be significant. For example, suppose we have the following data where each group's true μ_i is plotted.

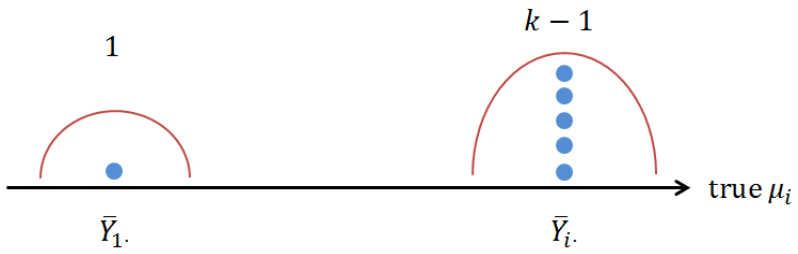


Figure 8.2: The one stray μ_i triggers a significant F-test, and then we will have $C(k,2) \cdot \alpha$ expected false discoveries.

8.6.3 Bonferroni's Union Bound

The method is simple: Do $\binom{k}{2}$ tests at level $\frac{\alpha}{\binom{k}{2}}$. That is, we use $t_{N-k}^{1-\frac{\alpha}{2\cdot C(k,2)}}$. This ensures that

$$\mathbb{P}(\text{reject anything}|H_0 \text{ true}) \leq \frac{\alpha}{\binom{k}{2}} \times \binom{k}{2} = \alpha$$

But, as we might see later, this process may be very/too conservative.

8.6.4 Tukey's Standardized Range Test

For simplicity, take all $n_i = n$.

Here, we find the distribution of the following:

$$\max_{1\le i\le i'\le k} \frac{|\overline{Y}_{i\cdot} - \overline{Y}_{i'\cdot}|}{s\sqrt{\frac{2}{n}}} = \frac{\overline{Y}_{(k)\cdot} - \overline{Y}_{(1)\cdot}}{s\sqrt{\frac{2}{n}}} \qquad \text{where} \qquad \overline{Y}_{(1)\cdot} \le \overline{Y}_{(2)\cdot} \le \ldots \le \overline{Y}_{(k)\cdot}$$

(The test statistic is assumed to be normal.) Then, we tabulate the critical points. Basically, we are finding the most extreme between-group difference, then making all comparisons using that largest difference as a "measuring stick" for the rest.

8.6.5 Scheffé

This is based on the "cheat contrast" method.

We reject average contrasts when 0 is not in the following confidence interval:

$$\sum_i \lambda_i \overline{Y}_{i\cdot} \pm \sqrt{(k-1)F^{1-\alpha}_{k-1,N-k}}\, s \left(\sum_i \frac{\lambda_i^2}{n_i}\right)^{1/2}$$

for all pairs $\lambda_i = 1, \lambda_{i'} = -1$. This is a fairly conservative method.

Note that Tukey and Scheffé can give different results. Tukey's method is based on quadrilaterals, while Scheffé's is based on ellipsoids. See Figure 8.3.

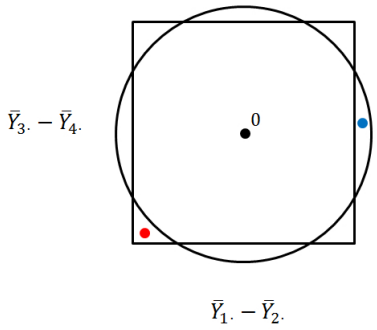


Figure 8.3: The square is Tukey's method, while the circle is Scheffé's. Results inside the shapes cannot reject, while those outside can. Thus, the red point on the left would be only rejected by Scheffé, while the blue dot on the right would only be rejected by Tukey.

8.6.6 Benjamini and Hochberg

Suppose we want to do m hypothesis tests where m is quite large. With normal t-tests, we would expect $m\alpha$ false discoveries, which is far too loose and has too easy of a standard. On the other

hand, Bonferroni gives false discoveries at a rate of $\frac{\alpha}{m}$, which could be way too conservative and preclude meaningful discoveries by having a really high bar applied to all tests.

Instead, we might want to choose the proportion of false discoveries we are willing to have. We could say that we're willing to accept a 10% rate of false discoveries because we still feel that those findings could be useful. How can we do this? Benjamini and Hochberg provide one (of many potential) solutions.

First, consider the following table.

	H_0 not rejected	H_0 rejected	
H_0 true	U	V	m_0
H_1 false	T	S	m_1
	$m - R$	R	m

In terms of the table, the false discovery proportion is $FDP = \begin{cases} \frac{V}{R} & R > 0 \\ 0 & R = 0 \end{cases}$.

The false discovery rate is then $FDR = \mathrm{E}[FDP]$.

B&H propose the following process.

- Perform each test, and then sort the p-values from lowest to highest:

$$p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$$

- Define

$$\ell_i = \frac{i\alpha}{m}$$

  which is just a line scaled by m.

- Define

$$R = \max\{i | p_{(i)} \leq \ell_i\}$$

  That is, R is the largest i at which the corresponding $p_{(i)}$ value is below the ℓ line.

- Reject all H_0's that have $p_{(i)} \leq T = p_{(R)}$.

Assuming that all p-values are independent, $FDR \leq \frac{m_0}{m}\alpha \leq \alpha$.

What if the p-values are correlated? (Maybe whether one gene is significant depends on another gene that's being separately tested.) Then, we use $\ell_i = \frac{i\alpha}{mC_m}$. This pulls down the line to the worst-case scenario, where

$$C_m = \sum_{i=1}^{m} \frac{1}{i} \approx \log(m)$$

But in practice, people tend to just use the independent version even if correlations/dependencies may exist.

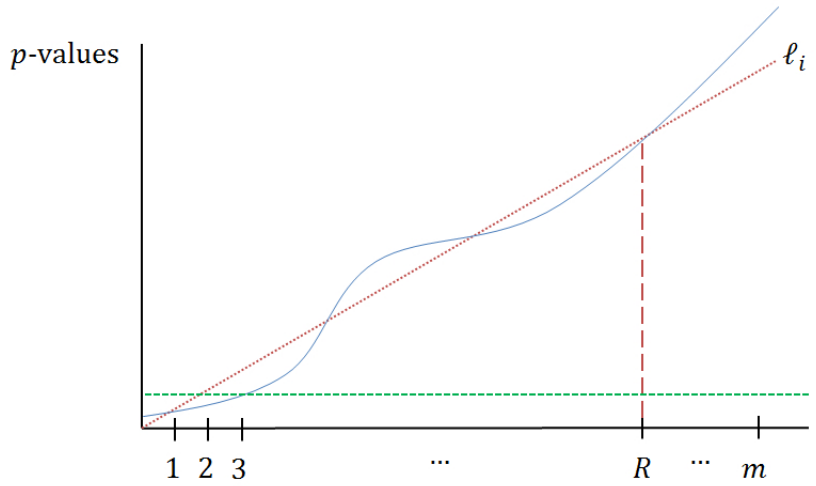


Figure 8.4: Visualization of B&H's process. The curvy blue line is the ordered p-values. The green line at the bottom is what we might get with Bonferroni's method; it clearly is a very stringent test.

8.7 Creating Groups

We've been talking about comparing groups, but what *are* these groups? How are they formed? We can think of at least three methods, in increasing order of rigor/preferability.

1. Historical control: Compare those being treated now with comparable units from the past.
    - This method is problematic in that we cannot account for any time trends and can never really nail down why the observed effects exist, or how much of it is attributable to the treatment.
    - Historical controls are not trustworthy, regardless of data size or the use of algorithms to try to account for time.
2. Contemporaneous but non-random experiment (starting a new education initiative where people can apply, drug trials where people choose to participate)
3. Randomized experiment

The choice of method has *no* effect on t-statistics, t-tests, p-values, F-distributions, confidence intervals, etc. All of these statistics are *agnostic* to the method of data collection and will produce the same results. This is where careful thought goes beyond simple application of statistical methods.

Chapter 9

Simple Regression

We are dealing with the case where $X_i \in \mathbb{R}$ and $Y_i \in \mathbb{R}$.

The term "regression" comes from a classic investigation of parents' and children's heights. Compared to parents' heights, children seemed to "regress" toward the mean.

If it was the case that

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_Y\sigma_X & \sigma_Y^2 \end{pmatrix} \right)$$

then we get an elliptical contour, and a regression line on such data is

$$Y = \mu_Y + \rho\frac{\sigma_X}{\sigma_Y}(X - \mu_X)$$

Meanwhile, note that the "45-degree" line is represented by $Y = \mu_Y + \frac{\sigma_X}{\sigma_Y}(X - \mu_X)$. Therefore, regression lines are flatter than the 45-degree line (the regression line is only 45 degrees of $\sigma_Y = \sigma_X$). If no correlation existed, then the data would be distributed as a circle and the regression line would be flat.

9.1 Regression Fallacy

Suppose we have data like the one below, comparing before and after. It would appear that the effect of treatment is a loss overall. Is that all we can say? It also appears that those who were worst before treatment made some gains. Doesn't that seem right?

But there's also a post-treatment loss for those that were the highest pre-treatment. In fact, if we drew points at each pre-treatment level, the points would form a regression line. If there exists any correlation between x and y (before and after treatment), then we will get a flatter regression line. So of course, we expect to see lower-than-average x align with higher-than-average y. This is a simple statistical artifact.

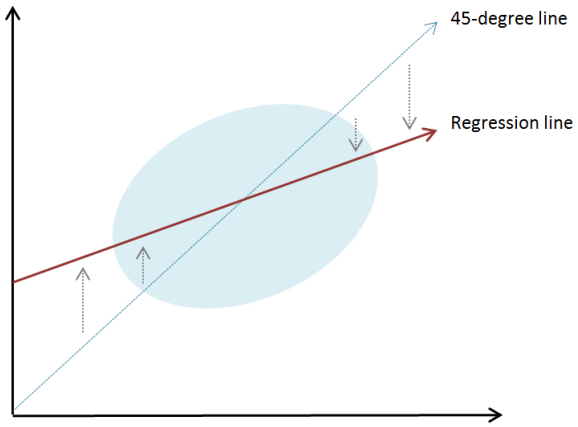
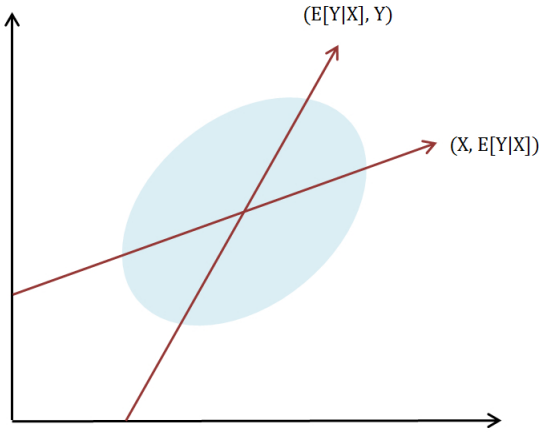


Figure 9.1: The linear regression line is flatter than the 45-degree line.



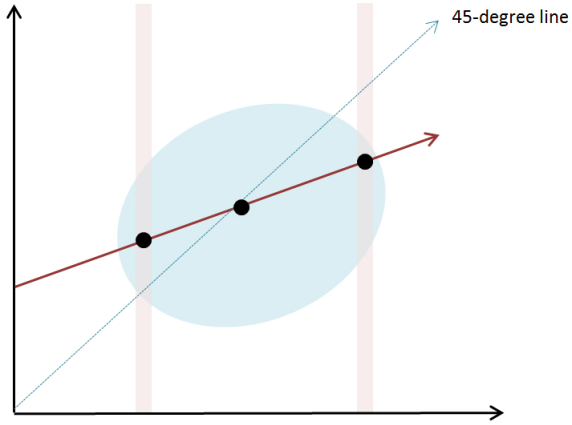


Figure 9.2: If we looked only at the two narrow bands, it looks like those who were first low improved, while those who were first high got worse. But these are simply artifacts of the regression line.

Be careful of people who make these kinds of interpretations of data. It's a fallacy and is meaningless.

9.2 The Linear Model

The following are building blocks of the standard simple linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Z = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \hat{\beta} = (Z^T Z)^{-1} Z^T Y$$

$$Z^T Z = \begin{bmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{bmatrix} \times n \qquad \overline{x} = \frac{1}{n}\sum x_i \qquad \overline{x^2} = \frac{1}{n}\sum x_i^2$$

$$Z^T Y = n \begin{bmatrix} \overline{y} \\ \overline{xy} \end{bmatrix} \qquad \overline{xy} = \frac{1}{n}\sum x_i y_i$$

So, we get that

$$\hat{\beta} = \begin{bmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{bmatrix} \begin{bmatrix} \overline{y} \\ \overline{xy} \end{bmatrix} \frac{1}{\overline{x^2} - \overline{x}^2} \qquad \text{where the fraction is a determinant}$$

Breaking this down,

$$\hat{\beta}_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{S_{XY}}{S_{XX}} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} = \frac{\sum_i (x_i - \overline{x}) y_i}{\sum_i (x_i - \overline{x})^2}$$

$$\hat{\beta}_0 = \frac{\overline{y}\,\overline{x^2} - \overline{x}\,\overline{xy}}{\overline{x^2} - \overline{x}^2} = \frac{\overline{y}\,\overline{x^2} - \overline{x}\,\overline{xy} + \overline{y}\,\overline{x}^2 - \overline{y}\,\overline{x}^2}{\overline{x^2}\,\overline{x}^2} = \overline{y} + \frac{-\overline{x}\,\overline{xy} - \overline{x}^2\overline{y}}{\overline{x^2} - \overline{x}^2} = \overline{y} - \overline{x}\frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \overline{y} - \hat{\beta}_1 \overline{x}$$

What does this tell us? $\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}$. So, that means that $(\overline{x}, \overline{y})$ is sitting on the regression line.

This was the hard way to derive parts of a simple regression line. There is an easier way.

9.2.1 The Easier Way

Consider this instead, where we shift all the x_i's so that they have a mean of zero:

$$y_i = \alpha + \beta_1 (x_i - \overline{x}) + \varepsilon \qquad \alpha = \beta_0 + \beta_1 \overline{x} \qquad \beta_0 = \alpha - \beta_1 \overline{x}$$

$$Z = \begin{bmatrix} 1 & x_1 - \overline{x} \\ \vdots & \vdots \\ 1 & x_n - \overline{x} \end{bmatrix} \qquad Z^T Z = \begin{bmatrix} n & 0 \\ 0 & S_{XX} \end{bmatrix} \qquad Z^T Y = n \begin{bmatrix} n\overline{y} \\ S_{XY} \end{bmatrix}$$

Then,

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \quad \text{(as before)} \qquad \hat{\alpha} = \overline{y}$$

9.3 Variance of Estimates

9.3.1 $\hat{\beta}_1$

Most of the time, when we do a simple regression, we care most about $\hat{\beta}_1$. What is the variance of this estimate?

$$\mathrm{V}(\hat{\beta}_1) = V\left(\frac{\sum_i (x_i - \overline{x}) y_i}{S_{XX}}\right) = \frac{\sum_i (x_i - \overline{x})^2}{S_{XX}^2}\sigma^2 = \frac{\sigma^2}{S_{XX}} = \frac{\sigma^2}{n\left[\frac{1}{n}\sum (x_i - \overline{x})^2\right]}$$

In essence, this is $\frac{\text{Variance of noise}}{n \cdot \text{Variance of groups}}$. That is similar to what we saw in ANOVA.

So, what affects $\mathrm{V}(\hat{\beta}_1)$?

- σ^2: Larger variance of error term.
- n: Variance of $\hat{\beta}_1$ decreases with n.
- $\frac{1}{n}\sum (x_i - \overline{x})^2$: This is the variance of X itself. Larger variance of X decreases variance of $\hat{\beta}_1$.
  - We want this variance to be large. Large variance in X leads to small Y variance.
  - Note that variance in X is not the same as variance in the error of X.
  - Larger X variance gives more leverage to draw an accurate regression line. If all the data is tightly clustered around some x, our estimate is uncertain.

Minimizing $\hat{\beta}_1$

These results indicate what the optimal X is that we should use to minimize $\mathrm{V}(\hat{\beta}_1)$.

For $A \leq X \leq B$, an optimal design has $\frac{n}{2}$ observations at $x_i = A$ and $x_i = B$. Then, send $B \to \infty$ and $A \to -\infty$.

However, this optimal design may be problematic. A *two-level* design like this cannot test for linearity; that is an assumption we make. Therefore, we may be missing non-linearities in the data

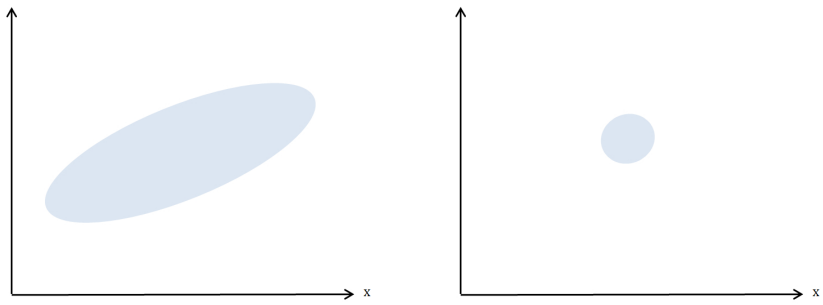


Figure 9.3: The data cloud on the left has a large S_{XX}, and therefore smaller variance on $\hat{\beta}_1$. The data cloud on the right has a small S_{XX} and therefore causes much larger variance in $\hat{\beta}_1$.

by only caring about two points. It could be that the relationship is non-linear in between A and B; or that while it is linear between A and B, it is not linear outside of this region. We should only consider this design if we are certain about linearity. (To hedge bets, you might collect some data points in between.) Also problematic, though maybe less so, is the feasibility of heading toward $\pm\infty$.

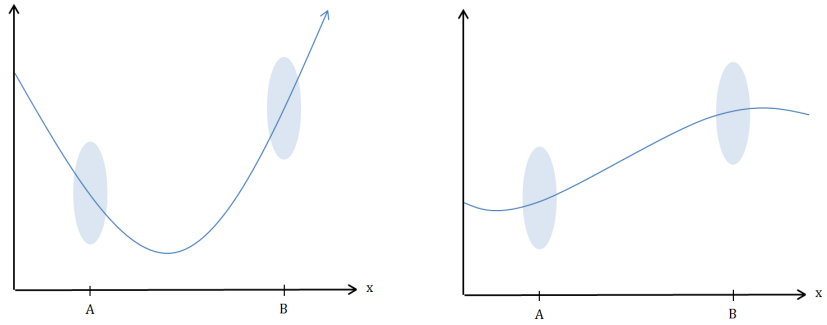


Figure 9.4: On the left is a case where non-linearity clearly exists between A and B but isn't captured with the two-level model. On the right is a case where things are linear between A and B but not elsewhere.

So, a large S_{XX} is good to have, but we should be careful about taking this to the extreme. On a related note, a small S_{XX} is bad.

An Example

Consider a study on the relationship between high school GPA and GPA in the first year of college, which may have an overall positive relationship as shown in the figure below.

However, say that the Stanford admissions office does a study on this relationship. The admissions office only gets applications with high GPAs (a selection effect), which sharply narrows variance in X, or S_{XX}. Using this limited data, the regression line may turn out to be flat or simply be hard to nail down; it could look like high school GPA is unrelated to freshman year performance at Stanford. As such, having artificial thresholds on the inclusion of data (which often happens in studies, especially when justified as removing "outliers") can undermine good analysis.

Even worse might be a case where Stanford admits some select students who had poor GPAs but compensated by being remarkable in other ways. Say that these students perform incredibly well

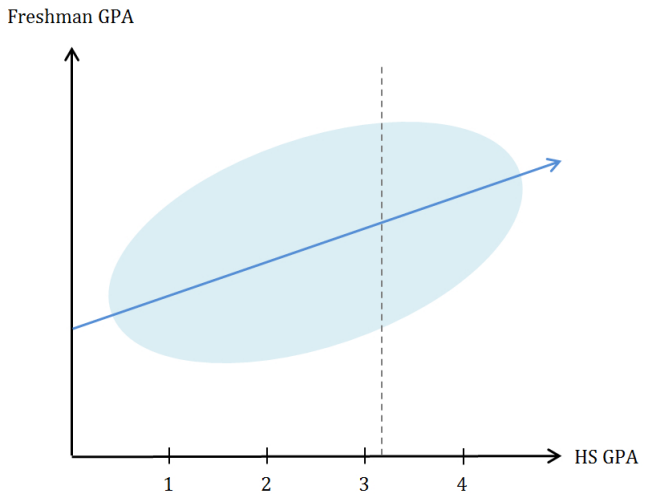


Figure 9.5: If the study was done using only data right of the dotted line, we may not find the positive relationship.

in the first year of college. Then, our observed data looks the way it does below and makes analysis even worse.

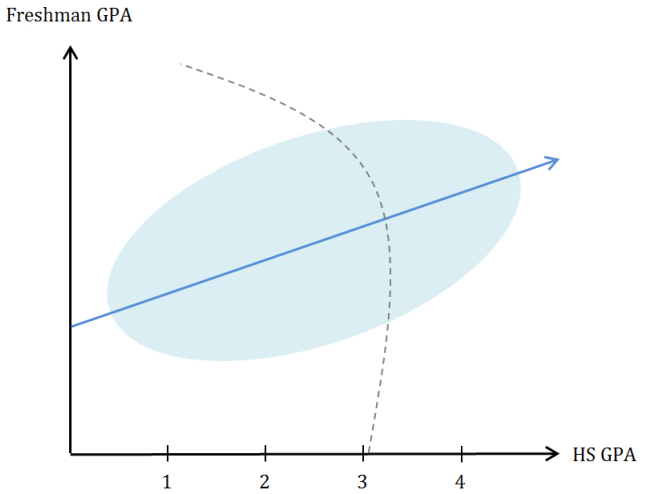


Figure 9.6: If the study was done and included some exceptional students, the results may be even worse.

9.3.2 $\hat{\beta}_0$

How about the intercept term? What is its variance?

$$\mathrm{V}(\hat{\beta}_0) = \ldots = \frac{\sigma^2}{n}\frac{\overline{x^2}}{\overline{x^2} - \overline{x}^2} = \frac{\sigma^2}{n}\frac{1}{1 - \frac{\overline{x}^2}{\overline{x^2}}} \qquad \left(= \frac{\sigma^2}{n}\frac{\overline{x^2}}{S_{XX}}\right)$$

$\mathrm{V}(\hat{\beta}_0)$ tends to be large when we are dealing with large values of X that have small variance. That

is, data tightly clustered around some x far away from 0 lead to imprecise $\hat{\beta}_0$ estimates. See the figure below.

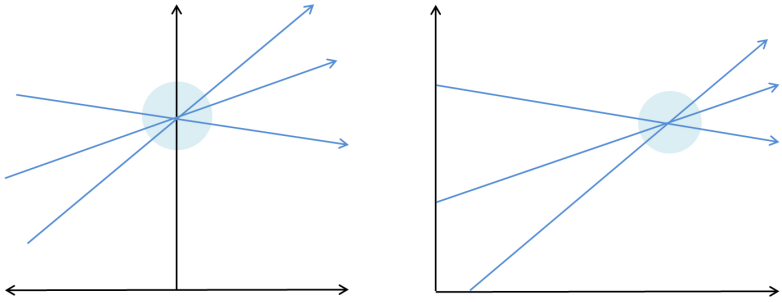


Figure 9.7: On the left is a case with a large $\frac{\text{mean}}{\text{SD}}$. With data clustered far away from 0, it is hard to nail down β_0 (or for that matter, β_1). On the right is a case with a small $\frac{\text{mean}}{\text{SD}}$. With data clustered tightly around 0, we can nail down β_0 if not β_1.

9.3.3 $\hat{\beta}_0 + \hat{\beta}_1 x$

$\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x$ is simply estimated using $\hat{\beta}_0 + \hat{\beta}_1 x$. What about the variance of the entire regression line?

$$\mathrm{V}(\hat{\beta}_0 + \hat{\beta}_1 x) = \ldots = \sigma^2 \left[\frac{1}{n} + \frac{(x - \overline{x})^2}{S_{XX}}\right] = \frac{\sigma^2}{n}\left[1 + \left(\frac{x - \overline{x}}{\sqrt{\mathrm{V}(X)}}\right)^2\right]$$

The variance of the regression line therefore depends on x. The term $\frac{x_i - \overline{x}}{\sqrt{\mathrm{V}(X)}}$ measures how many standard deviations the point x_i is away from the mean. The best accuracy/lowest variance occurs at $x = \overline{x}$; there, variance is simply $\frac{\sigma^2}{n}$ (as good as n observations at $x_i = \overline{x}$). The further away we get from $\overline{x}$, the larger the variance/uncertainty becomes. This suggests the danger of extrapolating data.

Regression (Confidence) Bands

This is the typical test statistic that we've seen before, applied to the regression line:

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x - (\beta_0 + \beta_1 x)}{s\sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{S_{XX}}}} \sim t_{(n-2)}$$

Using this, we can create a confidence interval for $\hat{\beta}_0 + \hat{\beta}_1 x$.

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{n-2}^{1-\frac{\alpha}{2}} \cdot s\sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{S_{XX}}}$$

If we calculate this across all of x, we get a confidence band.

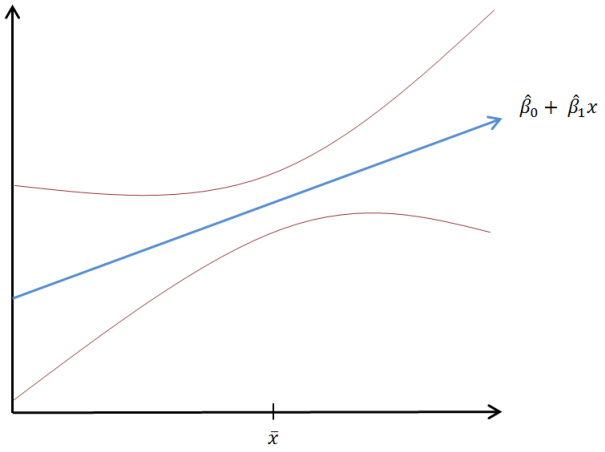


Figure 9.8: A confidence band. Note that it is tightest $\left(\frac{\sigma^2}{n}\right)$ at $\bar{x}$.

Note: $\sqrt{... + \frac{(x-\bar{x})^2}{...}}$. Therefore, the confidence band is a hyperbola with linear asymptotes, again showing a danger of extrapolating. (Potential non-linearity is another.)

More generally, for some $Y = Z\beta + \varepsilon$ where $Z_0 \in \mathbb{R}^{1\times p}$, the confidence band is

$$Z_0\beta \in Z_0\hat{\beta} \pm t_{n-p}^{1-\frac{\alpha}{2}} \cdot s\sqrt{Z_0(Z^TZ)^{-1}Z_0^T}$$

This allows us to get confidence bands for polynomial regressions and the like.

Prediction Bands

Now suppose that we get a new data point x_{n+1} and try to predict y_{n+1}. We want some form of interval where this prediction will land. Note that

$$\begin{aligned} y_{n+1} &= \beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1} \\ \hat{y}_{n+1} &= \hat{\beta}_0 + \hat{\beta}_1 x_{n+1} + 0 \end{aligned}$$

Our estimate of ε_{n+1} is zero; that's our best guess, anyway. Then,

$$y_{n+1} - \hat{y}_{n+1} = \beta_0 + \beta_1 x_{n+1} - (\hat{\beta}_0 + \hat{\beta}_1 x_{n+1}) + \varepsilon_{n+1}$$

where $\hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$ and ε_{n+1} are independent and normally distributed. The two parts are separable. Then, we get that

$$\mathrm{V}(\hat{\beta}_0 + \hat{\beta}_1 x_{n+1} + \varepsilon_{n+1}) = \mathrm{V}(\hat{\beta}_0 + \hat{\beta}_1 x_{n+1}) + \sigma^2$$

Our associated t-statistic for prediction intervals is therefore

$$\frac{y_{n+1} - \hat{\beta}_0 - \hat{\beta}_1 x_{n+1}}{s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{XX}}}} \sim t_{(n-2)}$$

The key difference is the additional 1 in the denominator. The prediction intervals give wider bands due to additional uncertainty from one additional observation.

Prediction bands are risky to use because they make a strong assumption of normality. $\hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$ can be relatively safely assumed to be normal due to Central Limit Theorem; $\hat{\beta}_0, \hat{\beta}_1$ are the result of averaging over all the observations. However, ε_{n+1} is just "averaging" over one (new) thing. If ε_{n+1} is not normally distributed (and we obviously don't know whether it is or not), then this statistic is in trouble.

More generally, if we take the average of m new y's at x_{n+1}:

$$\hat{\beta}_0 + \hat{\beta}_1 x_{n+1} \pm t_{n-2}^{1-\frac{\alpha}{2}} \cdot s\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{XX}}}$$

Note that if $m \to \infty$, we get the confidence interval. If $m = 1$, we get the prediction interval. The related test statistic is:

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 x_{n+1} - (\beta_0 + \beta_1 x_{n+1}) - \bar{\varepsilon}}{s\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{S_{XX}}}} \sim t_{n-2}$$

where $\bar{\varepsilon}$ is the average of the m future errors. Note that uncertainty increases as s increases, x_{n+1} gets further away from $\overline{X}$, etc.

9.4 Simultaneous Bands

So far, our confidence intervals have been pointwise rather than universal. That is, we are finding the interval around a specific x at which there is a 95% chance that the average of m y's lies in this band. (See Figure 9.9.) But what if we wanted to find a "confidence interval" for the entire regression line? A region in which we are 95% certain that the entire regression line sits?

A *simultaneous band* contains $\beta_0 + \beta_1 x$ at *all* $x \in \mathbb{R}$ with probability $1 - \alpha$ (typically 95%). That is, with probability $1 - \alpha$,

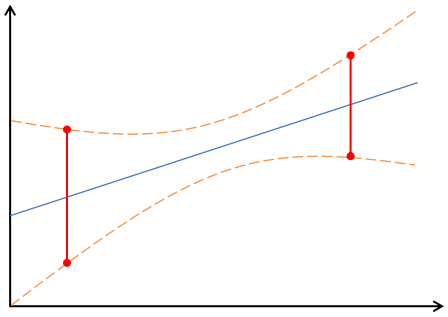


Figure 9.9: In doing confidence intervals so far, we looked at specific points like the two indicated. However, these did not speak to our confidence about the entire regression line.

$$(\hat{\beta} - \beta)^T (Z^T Z)^{-1} (\hat{\beta} - \beta) \leq s^2 \cdot p \cdot F_{p,n-p}^{1-\alpha}$$

This expression produces an ellipsoid for β centered around $\hat{\beta}$. The true β_0, β_1 lies in this ellipsoid with probability 95%.

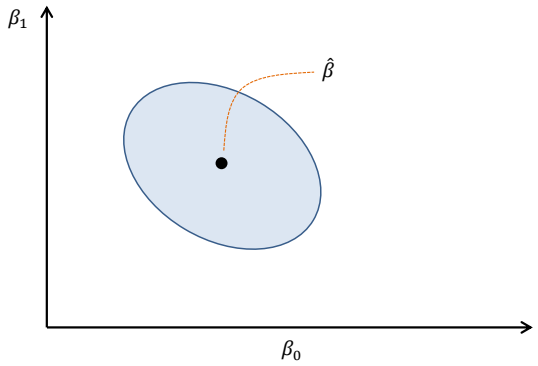


Figure 9.10: An ellipsoid. We have 95% confidence that the true β is inside this ellipsoid.

Like in Figure 9.10, this ellipsoid may be rotated. If we instead use the form $\alpha + \beta(x - \overline{X})$ for the regression line (where α stands for the intercept and not any probability), then $(Z^T Z)^{-1}$ becomes a diagonal matrix and the ellipsoid lies flat. This ellipsoid contains the true α, β with probability 95%.

Let's look at this ellipsoid in more detail. We can consider five specific points along this ellipsoid: the actual estimate of α, β and then four "extreme" points. Two capture the lowest and highest slopes, while two capture the lowest and highest intercepts.

If we skate around the perimeter of this ellipsoid and draw the respective lines, we get the familiar looking band. However, the band has become a bit wider than in the pointwise case. And specifically, if we look for

$$\max \beta_0 + \beta_1 x_{n+1} \quad \text{s.t.} \quad \beta_0, \beta_1 \quad \text{are in the ellipsoid}$$

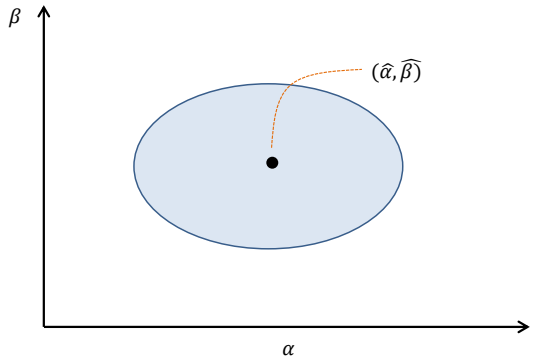


Figure 9.11: An ellipsoid based on $\alpha + \beta(x - \overline{X})$. We have 95% confidence that the true (α, β) is inside this ellipsoid.

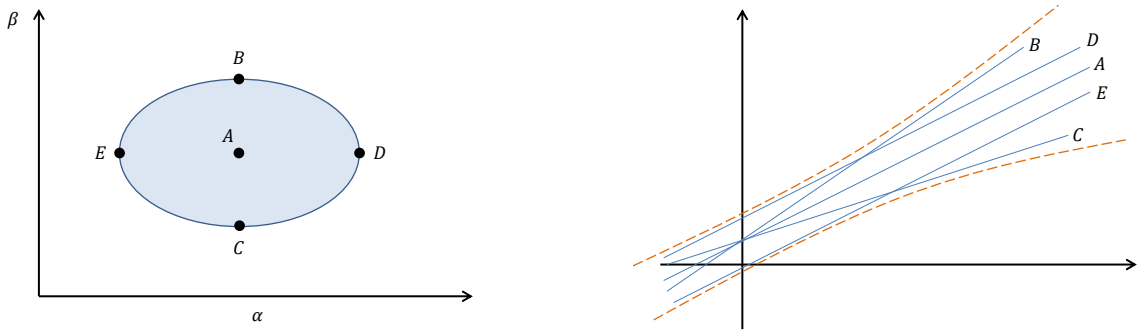


Figure 9.12: A is the original estimate. Going around the ellipsoid, B is where the slope is highest; C is where the slope is lowest; D is where the intercept is highest; E is where the intercept is lowest. Mapping the coefficients to a plot, they begin to form a hyperbolic band.

then we get the three following bands that correspond to the confidence, prediction, and average-of-m bands:

$$\hat{\beta}_0 + \hat{\beta}_1 x_{n+1} \pm \sqrt{2F_{2,n-2}^{1-\alpha}} \cdot s\sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{XX}}}$$

$$\hat{\beta}_0 + \hat{\beta}_1 x_{n+1} \pm \sqrt{2F_{2,n-2}^{1-\alpha}} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{XX}}}$$

$$\hat{\beta}_0 + \hat{\beta}_1 x_{n+1} \pm \sqrt{2F_{2,n-2}^{1-\alpha}} \cdot s\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{XX}}}$$

Instead of being t-distributed, these are now F-distributed. These bands are called *Working Hotelling Bands*. (Working is a last name and not an adjective.) We can generalize this to p dimensions:

$$Z_{n+1}\hat{\beta} \pm \sqrt{pF_{p,n-p}^{1-\alpha}} \cdot s\sqrt{\frac{1}{n} + Z_{n+1}(Z^TZ)^{-1}Z_{n+1}^T}$$

So, to cover the next $(x_{n+1}, \overline{Y}_{n+1})$ where $\overline{Y}_{n+1}$ stands for the average of the m new observations, we get

$$Z_{n+1}\hat{\beta} \pm \sqrt{pF_{p,n-p}^{1-\alpha}} \cdot s\sqrt{\frac{1}{m} + \frac{1}{n} + Z_{n+1}(Z^TZ)^{-1}Z_{n+1}^T}$$

9.5 Calibration

In some sense, this is the opposite problem of confidence intervals. Consider the simple linear model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Suppose we observe some y_{n+1} and want to find x_{n+1}. How can we do this?

Typically, people will simply try regressing x on y. This would work perfectly well if $\begin{pmatrix} x \\ y \end{pmatrix}$ were distributed bivariate normal. This would even be acceptable if $X|Y$ looked like a linear regression model. However, this method can fail if X is fixed by design. For instance, say that we tested something at three different temperatures. Then, flipping the axes does not really seem like a good idea.

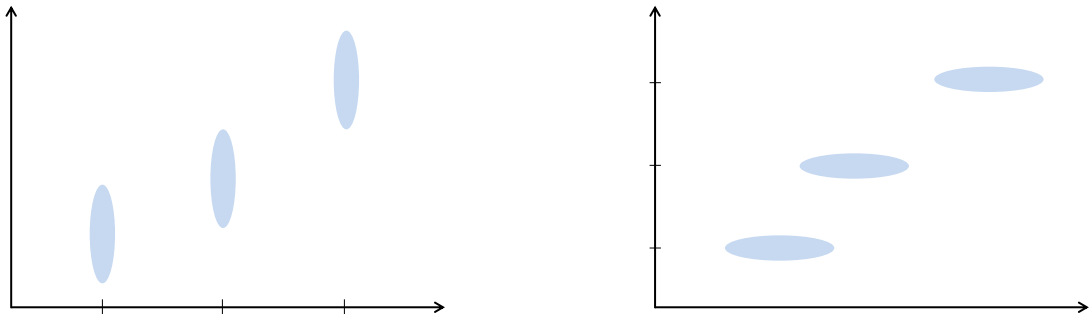


Figure 9.13: The original data is plotted on the left, with X potentially taking one of three values. On the right, we flip the axes to attempt calibration; this is clearly problematic.

So instead, we go back to the pivot. If $\overline{Y}_{n+1}$ represents the average of m points of new data,

$$\frac{\overline{Y}_{n+1} - \hat{\alpha} - \hat{\beta}(x_{n+1} - \bar{x})}{s\sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{S_{XX}}}} \sim t_{n-2}$$

That is, the t-distribution does not care about the labels of x and y. Using this pivot, the question then becomes: Which x_{n+1} gives $[...] \leq t_{n-2}^{1-\frac{\alpha}{2}}$? We could get an answer through simulation, algebra, or geometry. Here, we consider the geometric approach.

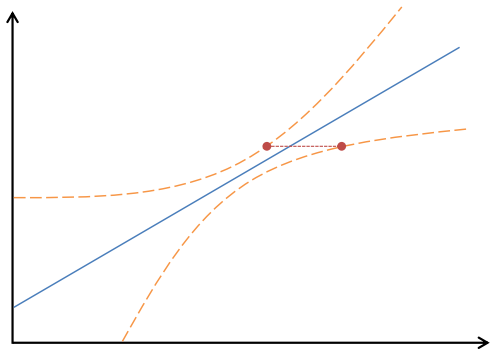


Figure 9.14: The horizontal line is our 95% calibration.

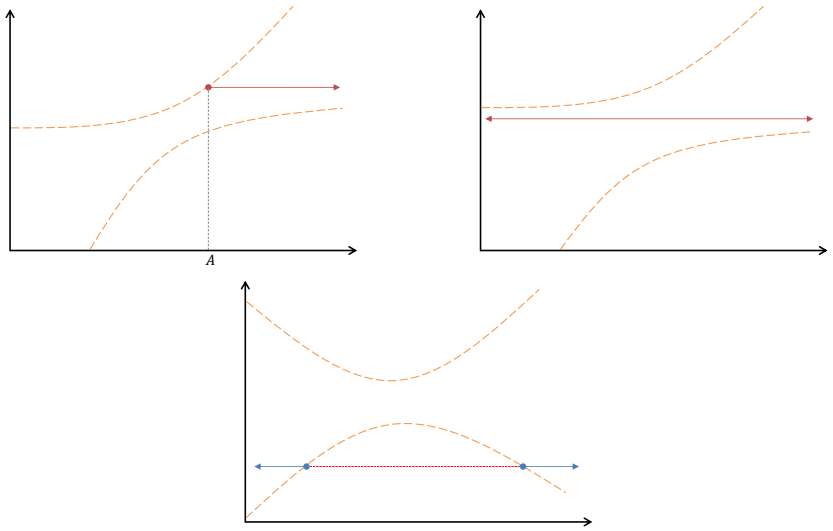


Figure 9.15: In the first case, we get that $x_{n+1} \in [A, \infty)$. In the second, we get $(-\infty, \infty)$ (and somehow only at the 95% confidence level). In the third case, we know that x_{n+1} is not in the red dotted region, but somewhere on the blue line segments extending to $\pm\infty$. None of these are useful results.

This is an example of *calibration*. However, in certain cases, calibration can fail spectacularly (due to x_{n+1} being in both the numerator and denominator of the statistic). Three examples follow in Figure 9.15.

When faced with these sorts of results, some people attempt to justify workarounds. However, any of these anomalies suggests that $\hat{\beta}$ is not well-established. Some potential reasons include:

- You need more data.
- You are looking for too precise of a confidence interval (like 99.99%).
- You misspecified the model (assuming linearity when the data is really non-linear).

Moreover, we have only been discussing the case of one x. Just imagine how much more complicated calibration becomes when dealing with two or more...

9.6 R^2 for $x \in \mathbb{R}$

Recall that in any regression,

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \overline{Y})^2}$$

If we mess around with some algebra, we also get that

$$R^2 = \frac{S_{XY}^2}{S_{XX}S_{YY}} = \left(\frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}\right)^2 = \hat{\rho}^2$$

So, R^2 is the squared correlation between X and Y.

9.7 Regression through the Origin

Suppose we have county-level data on population and the number of people afflicted with cancer. The standard regression line is such that $\mathrm{E}[Y] = \beta_0 + \beta_1 X$. It seems reasonable (if not intuitive/obvious) that if there are no people in a county, then $\mathrm{E}[Y|X = 0] = 0$. So in a case like this, we might consider forcing the regression line to pass through the origin.

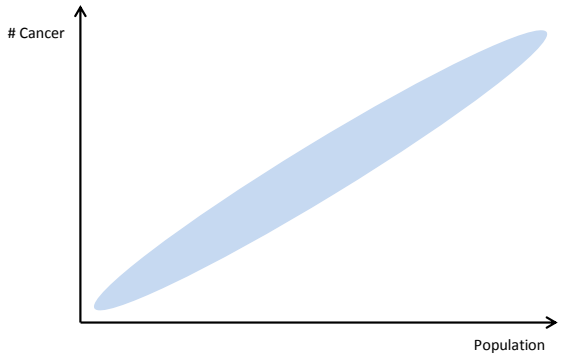


Figure 9.16: Data on population and number of cancer patients in counties.

But consider another example. Suppose that we measure the relationship between amount of fuel and distance traveled by a car. However, we only gather data further removed from the origin.

Obviously, a car without gas will not go anywhere, so perhaps $\mathrm{E}[Y|X = 0] = 0$ here, as well. But if we drew a least-squares line through this data, it would not pass through the origin. So which line—least squares or through the origin—should we use? If there is no data near the origin and linearity is in doubt, then we probably should not regress through the origin. But then, it is important to be careful about making extrapolations using the data (like predicting distances traveled on no gas).

As a technical note,

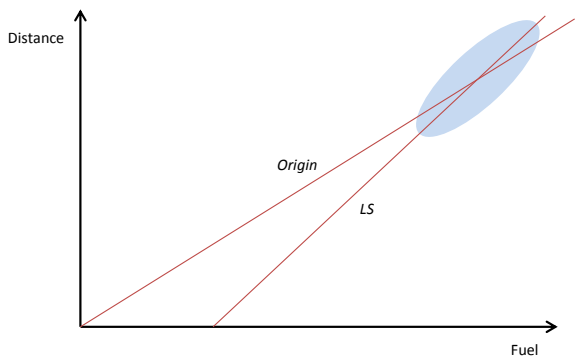


Figure 9.17: Data on fuel and distance traveled by a car. Note that the best-fitting regression line and the regression with no intercept term give different conclusions.

```
lm(Y ~ X - 1)
```

kills the intercept and forces the regression line through the origin in `R`.

9.7.1 The Futility of R^2

There is no meaningful R^2 for regressions forced through the origin. Recall that $\min \sum(Y_i - X_i\beta)^2$ yields $\hat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2} = \frac{S_{XY}}{S_{XX}}$.

We previously wrote R^2 in two different ways:

$$R^2 = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \overline{Y})^2}$$

While these were equivalent in the normal regression model, these two values may not be the same with regressions forced through the origin. There is no good fix for this, so there is effectively no R^2 here.

Chapter 10

Errors in Variables

(For more details, see Weisberg 4.6.3.)

We have always assumed that some degree of error exists in Y. However, it could also be the case that measurement error exists in X. How can we capture and account for this? How does the true Y correspond to the true X, both unadulterated by noise? Consider the following:

$$X_i = U_i + \delta_i \qquad Y_i = V_i + \varepsilon_i \qquad V_i = \beta_0 + \beta_1 U_i$$
$$\text{where} \qquad \delta_i \sim (0, \sigma_0^2) \qquad \varepsilon_i \sim (0, \sigma_1^2) \qquad \text{are independent of each other}$$

We are effectively performing a regression on data we do not actually see.

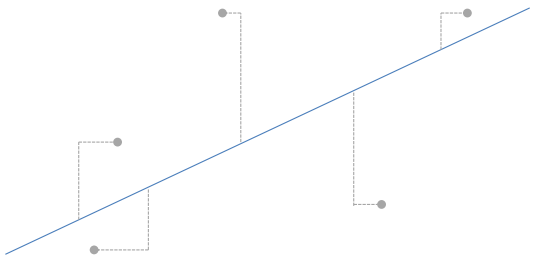


Figure 10.1: An example of the kind of data we are dealing with, with errors in both dimensions.

Note that

$$\mathrm{E}[Y|X] = \mathrm{E}[\beta_0 + \beta_1 U + \varepsilon|X] = \beta_0 + \beta_1 \mathrm{E}[U|X] = \beta_0 + \beta_1[X - \underbrace{\mathrm{E}[\delta|X]}_{\text{generally } \neq 0}]$$

If we just did a normal regression, we would get $\beta_1 \mathrm{E}[U|X]$, which is biased and not what we are looking for.

This extra noise biases β towards zero. The additional error in X causes the regression line to flatten; δ dampens the magnitude of β.

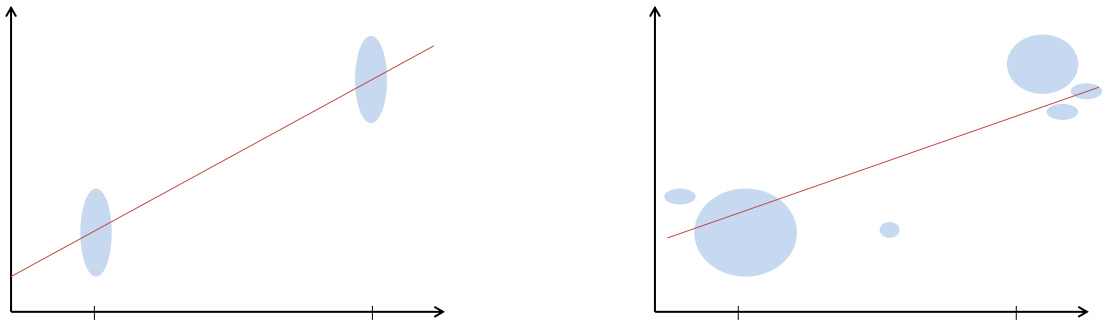


Figure 10.2: On the left is the regression line fitted to data with no noise in the explanatory variables. On the right, we add noise and note that the regression line becomes less steep because of it.

To illustrate this, suppose the following:

$$\begin{pmatrix} U_i \\ \delta_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_u \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_0^2 \end{pmatrix} \right)$$

(If $\sigma_0^2 >> \sigma_u^2$, then we are mostly seeing noise.) The algebra of the normal distribution then implies that

$$\mathrm{E}[U|X = x] = ... = \frac{\sigma_u^2 X + \sigma_0^2 \mu_u}{\sigma_u^2 + \sigma_0^2}$$

This is a weighted combination of relative variances.

- If σ_u^2 is really large, the noise is washed out and X takes over.
- If σ_0^2 is really large, then μ_u takes over.

Furthermore,

$$\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 \mathrm{E}[U|X = x] = \beta_0 + \beta_1 \frac{\sigma_0^2 \mu_u}{\sigma_0^2 + \sigma_u^2} + \beta_1 \frac{\sigma_u^2}{\sigma_u^2 + \sigma_0^2} X$$

The last term , $\beta_1 \frac{\sigma_u^2}{\sigma_u^2 + \sigma_0^2} X$, weakens the slope because $\frac{\sigma_u^2}{\sigma_u^2 + \sigma_0^2} \in [0, 1]$.

10.1 The Normal Case

What can we do about this? We can try the most straightforward case where we say *everything* is normally distributed.

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_{xy}^2 \\ \sigma_{xy}^2 & \sigma_y^2 \end{pmatrix} \right)$$

Notice that our error model has six parameters: $\beta_0, \beta_1, \mu_u, \sigma_u^2, \sigma_0^2, \sigma_1^2$. However, this normally distributed problem has five parameters: $\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_{xy}$. That is, a model with six parameters is somehow described by five parameters. Since we basically have six unknowns and five equations, the problem is not identifiable; in other words, it is underdetermined. This is true even with infinite data. We need an additional constraint from outside the data to solve this.

Often, we use the (untestable) constraint/assumption that $\frac{\sigma_1^2}{\sigma_0^2} = r > 0$. Or, we can directly specify the value of σ_1^2 or σ_0^2. (Notice that in all previous regressions, we had simply been assuming that $\sigma_0^2 = 0$. That is, we've implicitly been making an additional constraint all along.)

In the case where $\sigma_1^2 = \sigma_0^2$, we get an *orthogonal regression*, defined by

$$\min_{U_i, V_i, \beta_0, \beta_1} \sum_i (X_i - U_i)^2 + (Y_i - V_i)^2 \quad \text{s.t.} \quad V_i = \beta_0 + \beta_1 U_i$$

Instead of minimizing vertical sum of squares, this model minimizes the *orthogonal* sum of squares. See Figure 10.3.

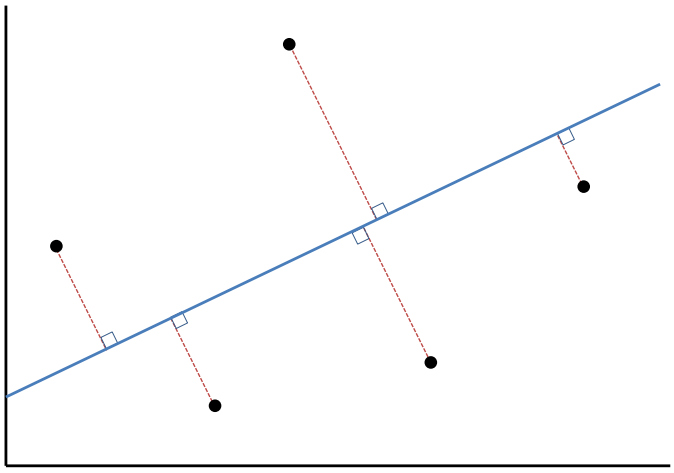


Figure 10.3: Orthogonal regression.

Note that changing the units can affect the results (especially the variances), so be careful. This method is also called orthogonal or total least squares.

Overall, this approach is hard to do, even with normal data. Generally, errors in some X_{ij} can bias $\hat{\beta}_k, k \neq j$. (That is, error in some observation can mess up *all* other estimates.)

Chapter 11

Random Effects

More detailed theoretical notes are online.

Random effects are not often covered in statistics courses, but can be very important to avoiding large mistakes.

Suppose we have Y_{ij} where $i = 1, ..., k$ and $j = 1, ..., n_i$ groups. This looks similar to past set-ups. However, now k represents k groups that are sampled from a larger population. Note the difference here. In *fixed effects* models, there are only k levels in the whole universe that we care about. For example, we may simply care about the effects of three different painkillers. However, in a *random effects* model, there are more then k levels in the population. For example, we may have thirty subjects in a drug trial, but we do not specifically care about the thirty subjects themselves. We are more concerned with the population from which they came; we want to generalize beyond the thirty subjects.

Another example: Consider four lathe operators. If we want to look at the result of woodwork, we may get variance in results from each of the four operators, and/or from other considerations (the quality of the wood, the maintenance of the lathes, weather, and the like). In this case, we may care about either fixed or random effects. If we wanted to study the effects of the four operators and they were the only operators in the factory, we could use fixed effects. However, if these four operators were out of 100 in the factory that happened to be there on the test day, we might want to use random effects. So, the choice of fixed/random is really about your goals and circumstances; it is not a strictly statistical choice.

11.1 The Model

$$Y_{ij} = \mu + a_i + \varepsilon_{ij} \qquad a_i \sim \mathcal{N}(0, \sigma_A^2) \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_E^2)$$

a_i, ε_i are independent of each other.

In looking at random effects, we may want to know

- σ_A^2 (just the variance of the random effects)

- $\frac{\sigma_A^2}{\sigma_E^2}$ (a ratio)
- $\frac{\sigma_A^2}{\sigma_A^2+\sigma_E^2}$ (a proportion)
- $H_0 : \sigma_A = 0$ (a test for relevance)

Here,

$$\text{Corr}(Y_{ij}, Y_{i'j'}) = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2} 1\{i = i'\}$$

And, as before,

$$\sum_i \sum_j (\overline{Y}_{ij} - \overline{Y}_{\cdot\cdot})^2 = \underbrace{\sum_i \sum_j (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2}_{SSA \text{ or } SS_{\text{between}}} + \underbrace{\sum_i \sum_j (\overline{Y}_{ij} - \overline{Y}_{i\cdot})^2}_{SSE \text{ or } SS_{\text{within}}}$$

In this case,

$$SSE \sim \sigma_E^2 \chi^2_{N-k}$$

If we assume $n = n_i$ (all groups have same size), then

$$\frac{1}{n} SSA = \sum_{i=1}^{k} (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 \sim \left(\sigma_A^2 + \frac{\sigma_E^2}{n}\right) \chi^2_{k-1}$$

If we put SSE and SSA together, we have two independent χ^2 tests and thus an F distribution:

$$F = \frac{\frac{1}{k-1} SSA}{\frac{1}{N-k} SSE} = \frac{\frac{n}{k-1}\left(\sigma_A^2 + \frac{\sigma_E^2}{n}\right) \chi^2_{k-1}}{\frac{1}{N-k}\sigma_E^2 \chi^2_{N-k}} \sim \left(1 + \frac{n\sigma_A^2}{\sigma_E^2}\right) F_{k-1,N-k}$$

If σ_A^2 is high, then F is also high. If $\sigma_A^2 = 0$, then we just get a normal F-distribution. As such, we can test a null like $H_0 : \sigma_A^2 = 0$. If we get that $F > F^{1-\alpha}_{k-1,N-k}$, we can reject the null.

This looks a lot like the test we used for fixed effects. However, it isn't. With FE, we used a non-central F' with the sum of α_i's. That is,

$$F'_{k-1,N-k}\left(\frac{n}{\sigma^2} \sum_{i=1}^{k} \alpha_i^2\right)$$

We used this to estimate the α_i's given the data.

11.2 Estimation

In the random effects model, in order to estimate the effect of group a_i, we use

$$\tilde{a}_i = \mathrm{E}[a_i | Y_{ij} \quad i = 1, ..., k \quad j = 1, ..., n_i] = ... = \frac{n\sigma_A^2}{\sigma_E^2 + n\sigma_A^2}(\overline{Y}_{i\cdot} - \mu)$$

Then, we get that

$$\mu + \tilde{a}_i = \frac{n\sigma_A^2}{\sigma_E^2 + n\sigma_A^2}\overline{Y}_{i\cdot} + \left(1 - \frac{n\sigma_A^2}{\sigma_E^2 + n\sigma_A^2}\right)\mu$$

So, we are dealing with an estimate that is between μ (the overall average) and the group average. The first term is the weighted sum of averages within the group, while the second term is the "background" mean. If σ_E^2 is huge and/or n is small, then we have no good measure of group i, so we place more emphasis on μ as an estimate. When σ_A^2 is high and/or n is large, we have a much better estimate of the group mean and place less weight on the overall mean μ. That is, in the second case, we are able to avoid "shrinkage" (movement toward the population mean).

11.3 Two-Factor ANOVA

Covariates of interest may be either considered as fixed or random. What happens when two of them are interacted? What kind of "massive mistakes" could we make in the last case? Suppose we

Interaction	What to do
Fixed × fixed	Normal regression
Random × random	Not too bad; just change the test and use the F test earlier in these notes
Fixed × random	We could make massive mistakes here; exercise caution

Table 11.1: Overview of two-factor combinations and what to do.

tested three painkillers on five subjects, but we have 10,000 observations over these five subjects. If we use fixed effects on the painkillers but don't use random effects on the five subjects (implicitly assuming that the five subjects are the entire universe), then our results could be completely wrong.

11.3.1 Fixed × Fixed

Suppose:

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$
$$\text{where} \qquad k = 1, ..., n_{ij} \qquad i = 1, ..., I \qquad j = 1, ..., J \qquad n_{ij} = n$$

That is, in each i, j cell, we have k observations. Then, we can look for μ_{ij} in the following way:

$$\mu_{ij} = \underbrace{\mu + \underbrace{\alpha_i}_{\text{row effect}} + \underbrace{\beta_j}_{\text{column effect}}}_{\text{additive}} + \underbrace{(\alpha\beta)_{ij}}_{\text{interaction}}$$

Note that the interaction term $(\alpha\beta)_{ij}$ is optional. Also, we treat $\alpha\beta$ as a single symbol in and of itself. We do this instead of using a new letter because we'd run out of Greek letters too quickly.

Furthermore, we have the following constraints:

$$0 = \sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij}$$

These constraints make sense. If one of these sums did not add up to zero, we could simply mean-shift everything so that it did add up to zero. We do this because we would like to make μ the "grand" average and have α_i, β_j as adjustments/deviations from this overall mean depending on being in group i and group j.

Decomposition

The decomposition, including the interaction, has four parts (sums of squares):

$$SSA + SSB + SSAB + SSE$$

These are captured by:

$$\begin{aligned}\sum_i \sum_j \sum_k (Y_{ijk} - \overline{Y}_{...})^2 = &\sum_i \sum_j \sum_k (\overline{Y}_{i..} - \overline{Y}_{...})^2 \\ &+ \sum_i \sum_j \sum_k (\overline{Y}_{.j.} - \overline{Y}_{...})^2 \\ &+ \sum_i \sum_j \sum_k (\overline{Y}_{ij.} - \overline{Y}_{i..} - \overline{Y}_{.j.} + \overline{Y}_{...})^2 \\ &+ \sum_i \sum_j \sum_k (\overline{Y}_{ijk} - \overline{Y}_{ij.})^2\end{aligned}$$

This is the sum of row variance, column variance, SS_{BETWEEN}, SS_{WITHIN}.

Distributionally,

$$SSA \sim \sigma^2 \chi'^2_{I-1}\left(\frac{nJ\sum_i \alpha_i^2}{\sigma^2}\right)$$

$$SSB \sim \sigma^2 \chi'^2_{J-1}\left(\frac{nI\sum_j \beta_j^2}{\sigma^2}\right)$$

$$SSAB \sim \sigma^2 \chi'^2_{(I-1)(J-1)}\left(\frac{n\sum_i\sum_j (\alpha\beta)_{ij}^2}{\sigma^2}\right)$$

$$SSE \sim \sigma^2 \chi^2_{IJ(n-1)}$$

(Note that SSE is the only one that uses the regular, central χ^2 distribution.) Then, we can perform following null hypothesis tests:

$$\alpha_i = 0 \qquad F_A = \frac{MS_A}{MS_E} \sim F_{(I-1),IJ(n-1)}$$

$$\beta_j = 0 \qquad F_B = \frac{MS_B}{MS_E} \sim F_{(J-1),IJ(n-1)}$$

$$(\alpha\beta)_{ij} = 0 \qquad F_{AB} = \frac{MS_{AB}}{MS_E} \sim F_{(I-1)(J-1),IJ(n-1)}$$

If $n = 1$

Gathering data can be cost-prohibitive. (To move from $n = 1$ to $n = 2$ involves doubling the data.) So, it may be tempting to only get one observation in every group and just get more groups.

However, if we do this, there is no SSE because there are zero degrees of freedom. If there is no SSE, then all the F-tests above are undefined.

Is there a workaround? We could try calculating $\frac{MS_A}{MS_{AB}}$ instead. But, we will lose some statistical power if $(\alpha\beta)_{ij}$ gets large (because F will get small). Therefore, if we can reject the null using this measure, then we can safely reject the null because we were able to reject despite lower power.

11.3.2 Random × Random

We have the following set-up:

$$Y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk}$$

$$\begin{aligned} a_i &\sim \mathcal{N}(0, \sigma_A^2) \\ b_j &\sim \mathcal{N}(0, \sigma_E^2) \\ (ab)_{ij} &\sim \mathcal{N}(0, \sigma_{AB}^2) \\ \varepsilon_{ijk} &\sim \mathcal{N}(0, \sigma_E^2) \end{aligned}$$

where all distributions are independent, $i = 1, ..., I$; $j = 1, ..., J$; and $k = 1, ..., n$. Then, we have the following sum of squares:

$$\begin{aligned} SS_A &\sim (\underbrace{\sigma_E^2}_{\text{noise}} + n\sigma_{AB}^2 + \underbrace{nJ\sigma_A^2}_{\text{variance of } A \text{ out in ``the wild''}}) \, \chi_{I-1}^2 \\ SS_B &\sim (\sigma_E^2 + n\sigma_{AB}^2 + nI\sigma_B^2) \, \chi_{J-1}^2 \\ SS_{AB} &\sim (\sigma_E^2 + n\sigma_{AB}^2) \, \chi_{(I-1)(J-1)}^2 \\ SS_E &\sim \sigma_E^2 \, \chi_{IJ(n-1)}^2 \end{aligned}$$

Note that these are are multiples of centered χ^2 distributions. Since they are independent, we can combine them to do F-tests. Say that we wanted to test A. We could try the following:

$$F_A = \frac{MS_A}{MS_E} \sim \frac{\sigma_E^2 + n\sigma_{AB}^2 + nJ\sigma_A^2}{\sigma_E^2} \, F_{I-1, IJ(n-1)}$$

But there is a problem here. Even if $\sigma_A^2 = 0$, we get $\frac{\sigma_E^2 + n\sigma_{AB}^2}{\sigma_E^2} > 1$. So we do not get the normal F back, and we do not get a pure test of A. Instead, this also ends up testing a combination of A and B, too. A better approach would be:

$$F_A = \frac{MS_A}{MS_{AB}} \sim \left(1 + \frac{nJ\sigma_A^2}{\sigma_E^2 + n\sigma_{AB}^2}\right) \, F_{I-1, (I-1)(J-1)}$$

This does give us a proper test of $H_0 : \sigma_A = 0$.

11.3.3 Random × Fixed

This is also called *mixed effects*. It is also where things get scary.

Imagine that we want to test $I = 3$ medicines on $J = 10$ subjects. We will perform $n = 5$ replicates of the experiment. We want to learn about just these three drugs' effects, but on the population at large (not just the ten test subjects). We can therefore think of using fixed effects on the medicines and random effects on the subjects.

We have the following set-up:

$$Y_{ijk} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + \varepsilon_{ijk}$$

$$\begin{aligned} \sum_i \alpha_i &= 0 \\ b_j &\sim \mathcal{N}(0, \sigma_E^2) \\ (ab)_{ij} &\sim \mathcal{N}(0, \sigma_{AB}^2) \\ \varepsilon_{ijk} &\sim \mathcal{N}(0, \sigma_E^2) \end{aligned}$$

where all distributions are independent, $i = 1, ..., I$; $j = 1, ..., J$; and $k = 1, ..., n$. Moreover, we have the additional constraint that

$$\sum_{i=1}^{I} (\alpha b)_{ij} = 0 \quad \forall j \text{ with probability 1}$$

(So for each subject, the interactive effects add up to zero.) Then, we have the following sum of squares:

$$\begin{aligned} SS_A &\sim (\sigma_E^2 + n\sigma_{AB}^2)\ \chi'^2_{I-1}\ \left(\frac{nJ\sum_i \alpha_i^2}{\sigma_E^2 + n\sigma_{AB}^2}\right) \\ SS_B &\sim (\sigma_E^2 + nI\sigma_B^2)\ \chi^2_{J-1} \\ SS_{AB} &\sim (\sigma_E^2 + n\sigma_{AB}^2)\ \chi^2_{(I-1)(J-1)} \\ SS_E &\sim \sigma_E^2\ \chi^2_{IJ(n-1)} \end{aligned}$$

Again, we want to find ways to test the main effects and have statistics where we recover the regular distribution when $\sigma^2 = 0$. We end up finding the following:

- To test A, use $\dfrac{MS_A}{MS_{AB}}$.
- To test B, use $\dfrac{MS_B}{MS_E}$.
- To test AB, use $\dfrac{MS_{AB}}{MS_E}$.

All of these result in (potentially non-central) F-distributions due to independence.

Note that in order to test fixed effects, we used MS_{AB} (mean squared for the interaction) in the denominator. To test random effects, we used MS_E in the denominator. But then in the case of two fixed effects, we use MS_E, and in the case of two random effects, we used MS_{AB}. The pattern appears to flip.

11.3.4 Other Remarks

Even with $n \to \infty$, we really only have ten data points. Getting more data on these same subjects will not affect the results much; there are diminishing returns to more replicates. Watch out for studies that perform many replications and abuse this data without accounting for random effects.

Once a lot of random and fixed effects are in a single model, there may not be a mathematically sound way to construct the F-test. The best methods are still being debated in statistics. This is a good reason to limit your analysis.

Chapter 12

Multiple Regression

Now, we finally move on to the more general case where $x \in \mathbb{R}^p$.

Suppose we regress Y_i on $Z_{i1}, Z_{i2}, ..., Z_{ip}$. Usually, $Z_{i1} = 1$ to account for the intercept.

One big question we might have is which j's to include in a model. What features should we use? Thinking about this, note that

$$\sum_{i=1}^{n} \mathrm{V}(\hat{Y}_i) = \sum_{i=1}^{n} H_{ii}\sigma^2 = \sigma^2 \mathrm{tr}(H) = p\sigma^2$$

where H is the hat matrix. Therefore, with more variables comes more variance. With fewer variables, we get more bias. This speaks to the classic bias-variance trade-off.

12.1 Example: Body Fat

Consider a body fat analysis that involves 252 men. We obtain proportion of body fat in each subject by obtaining their density (by submerging them in water). Each person's weight in air and weight underwater are obtained, as well as measures of things like skin folds, wrist size, abdomen size, etc. The goal is to find an easy formula for estimating a person's percentage of body fat without having to go through the process of getting their density in such a convoluted and unpleasant way.

The study has two responses: density and proportion of body fat.

We define

$$\begin{aligned} D &= \text{density} = \frac{\text{g}}{\text{cm}^3} \\ A &= \text{proportion lean mass} \\ B &= \text{proportion fat mass} \\ a &= \text{lean density} = 1.1 \\ b &= \text{fat density} = 0.9 \\ A + B &= 1 \end{aligned}$$

Then, Siri's formula says

$$D = \frac{1}{\frac{A}{a} + \frac{B}{b}} = \frac{1}{\frac{1-B}{a} + \frac{B}{b}}$$

This rewrites density all in terms of fat proportions. We get that

$$B = \frac{4.95}{D} - 4.5 \qquad D = \frac{\text{weight in air}}{\frac{\text{weight in air} - \text{weight in water}}{0.997}} - \text{lung volume}$$

With this in mind, refer to the handout from class. Some points from there worth repeating here:

- The "fullest" regression seems to perform incredibly well ($R^2 = 0.9996$). However, density is one of the predictors here (and also has a ridiculous t-statistic of -373). Since density is something that we are trying to predict, we should not include it as a predictor. In fact, the whole purpose of the study was to avoid using density calculations to estimate body fat.
- Should we drop coefficients that do not appear significant? It may not be a good idea. Just because we could not prove statistical significance does not prove that it actually is insignificant.
- Abdomen size is a pretty good predictor of body fat. A regression with just that has $R^2 = 0.65$. Once we add age, age is also highly significant but $R^2 = 0.6636$. (In a regression with only age, $R^2 = 0.09$.) This marginal change in R^2 is because age and abdomen size are highly correlated. It is why the R^2's are not additive.
- Always check the data for outliers and anomalies. Always be scared of them.

12.2 Some Considerations

12.2.1 A "True" β_j

In $Y = Z\beta + \varepsilon$, the value of β_j depends on which other predictors are in the model. Therefore, there is no "real" coefficient; such a term is not well-defined, because it depends on what is included. A "true" β_j exists for each and every subset of covariates used in a model.

The values also depend on the population being sampled. So if a study uses only on-campus college students, we are hard-pressed to speak to anything about the general population.

12.2.2 Naive Face-Value Interpretation

Again, the regression model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$$

We could think of β_j as $\frac{\partial}{\partial X_j} Y$ or $\frac{\partial}{\partial X_j} \mathrm{E}[Y]$. While this seems intuitive, it is also unreliable. Consider the case where $X_2 = X_1^2$. Then, changing one predictor alone is not even possible. Furthermore, changing X_1 alone may cause changes to a variety of $X_{k \neq i}$ (like a ripple effect).

12.2.3 Wrong Signs

Suppose we wanted to find patterns in housing prices.

$$\begin{aligned} Y &= \text{value of house} \\ X_1 &= \text{age} \\ X_2 &= \text{square footage} \\ X_3 &= \#\text{ bedrooms} \\ X_4 &= \#\text{ bathrooms} \\ &\vdots \\ X_{77} &= \text{patio size} \\ &\vdots \\ X_{242} &= \#\text{ chimneys} \end{aligned}$$

If we regressed housing prices on everything, we might get that $\beta_3 < 0$. This does not seem right; additional bedrooms should increase the price of a house. However, this negative coefficient could occur because there are lots of correlations in the data. For instance, adding a bedroom to a house usually adds square footage. It may be more proper to add a bedroom and to add the square footage of a typical bedroom in order to get a more accurate effect. These are things that must be considered by the individual; the algebra of regression does not know/tell you to do this.

12.2.4 Correlation, Association, Causality

Remember that regressions with observational data only capture correlations/associations and say nothing about causation. It could be that $X \to Y$, $Y \to X$, or some $Z \to X, Y$.

Suppose that $X =$ price, and $Y =$ sales volume. We might find that higher prices lead to higher sales. This seems odd and seems to suggest that a company should produce an infinite quantity

of a product for infinite sales. But, the data actually shows the importance of Z = old vs. new customer. Specifically, the company charges higher prices to old (and thus loyal) customers, while offering lower prices to lure new customers. If the company did not lure many new customers but had good business with the loyal clientele, that would explain the odd result and show that Z is responsible for both.

Chapter 13

Interplay between Variables

Recall that the coefficient of X is not fixed; it depends on what other variables are included. So, there is no such thing as a "true" coefficient.

For example, suppose we perform a regression using data in Figure 13.1. We could use age as the only predictor. Then, $Y = \beta_0 + \beta_1 AGE$, and we get that $\beta_1 > 0$. However, if we also accounted for group, then $Y = \beta_0 + \beta_1 AGE + \beta_2 GROUP$, and we get that $\beta_1 < 0$.

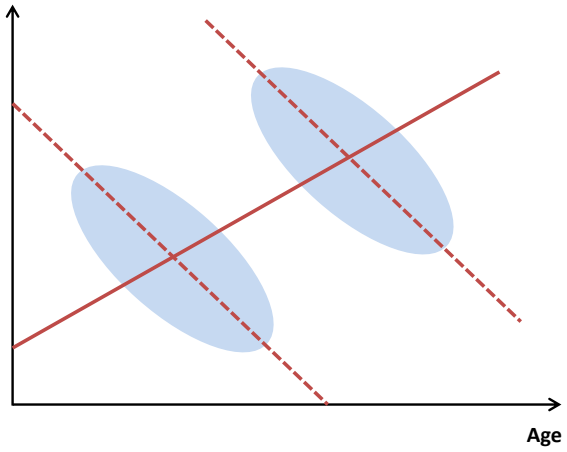


Figure 13.1: With only age, we get a positive coefficient. Controlling for groups, we get negative coefficients on age.

Or suppose we have data that looks like Figure 13.2, which is categorical data. Overall, a positive relationship seems to exist. However, if we conditioned on (controlled for) groups Z, we get a negative relationship. But then if we accounted for W (groups within each Z) and Z, we get something positive again.

As such, two people could reasonably claim completely different coefficients using the same data.

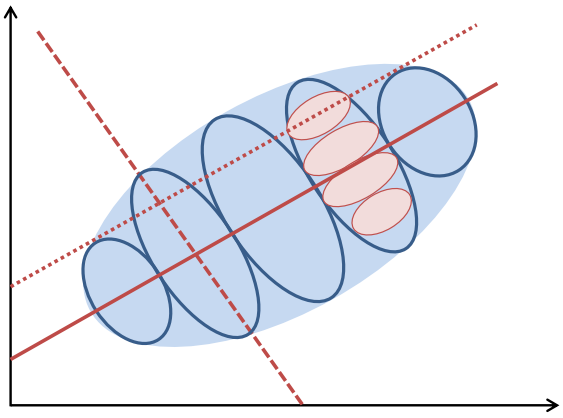


Figure 13.2: The sign of our coefficient with respect to X changes depending on what we control.

13.1 Simpson's Paradox

Association is not causality.

13.1.1 Hospitals

Say that we examine two hospitals, A and B. In each hospital, we record how many admitted patients survive (S) and die (D). There are two subsets of patients: those that are mildly sick and those that are very sick. The data is below.

	S	D
A	50	0
B	80	10

	S	D
A	25	25
B	0	10

$\Rightarrow$

	S	D
A	75	25
B	80	20

Table 13.1: Records for mildly sick, very sick, and totals, respectively.

Hospital A is more successful in dealing with mildly sick patients. It also does a better job overall with very sick patients. However, when we add the tables together, Hospital B comes out looking better. This is because B got more easy cases. Be careful adding up tables across multiple dimensions or features. In some cases, this subsetted data may not even be available.

13.1.2 Smoking and Cancer

A common claim is that smokers have higher propensity for lung cancer. Fisher questioned this claim, suggesting that perhaps a genetic factor exists that causes a predisposition to both more smoking and higher likelihood of cancer. (That is, rather than $X \to Y$, there exists a Z such that $Z \to X, Y$.) In order to test this, Fisher found twins and looked at their smoking patterns. He found that in identical twins, 76% of pairs both smoked or did not smoke. However, in fraternal twins, only 49% matched in smoking patterns. He concluded that there probably is a genetic component.

However, Cornfield refuted this, saying that the difference in percentages is not big enough to be conclusive. Smokers have a nine-fold increased risk factor for cancer. (When smoking two or more packs a day, that rises to 60.) With such large effects, we need to find big correlations to justify the existence of an outside variable like genetics.

We will see a regression-based example of this later.

13.2 Competition and Collaboration

Variables can both complete and collaborate for significance in a regression.

In the case of competition,

- $\hat{\beta}_1$ is significant if X_2 *is not* in the model.
- $\hat{\beta}_2$ is significant if X_1 *is not* in the model.

In the case of collaboration,

- $\hat{\beta}_1$ is significant if X_2 *is* in the model.
- $\hat{\beta}_2$ is significant if X_1 *is* in the model.

13.2.1 Competition

Competition will occur if X_1, X_2 are highly correlated, since if we already have one variable in a regression, adding the second does not contribute much more useful information. See Figure 13.3 below, where X_1, X_2 have a high positive correlation. Then, when we estimate β, we could get a confidence ellipse like the one below.[1] With this ellipse, $\beta_1 = 0$ is inside, as is $\beta_2 = 0$. However, both cannot be zero simultaneously.

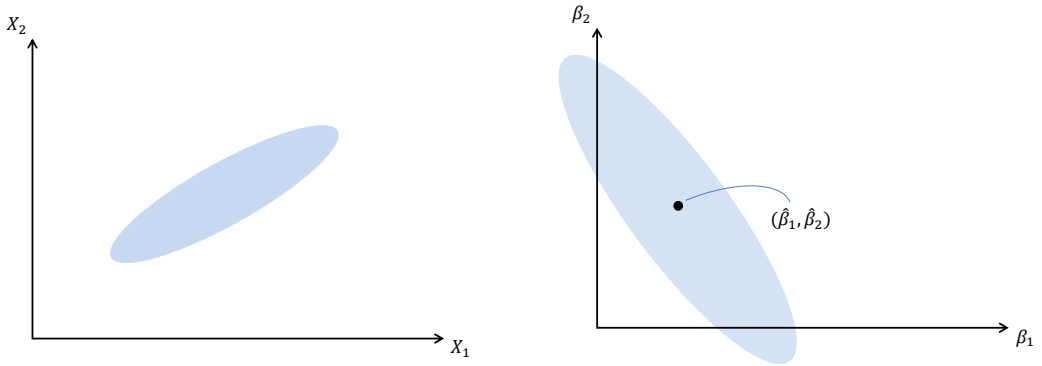


Figure 13.3: When a strong positive correlation exists between X_1 and X_2 (left), then the confidence ellipse for $(\hat{\beta}_1, \hat{\beta}_2)$ is negative and suggests that each β could actually be zero.

[1] Algebraically, if two things are positively correlated, then $\hat{\beta}_1$ and $\hat{\beta}_2$ may be negatively correlated. That is, if the signs are $Z^T Z = \begin{bmatrix} + & + \\ + & + \end{bmatrix}$, then $(Z^T Z)^{-1} = \begin{bmatrix} + & - \\ - & + \end{bmatrix}$.

In cases like this, we might just pick one variable or the other, or average the two. In practical terms, we may pick the one that is less labor-intensive or costly.

The story is flipped with X_1, X_2 are negatively correlated; the confidence ellipse runs upward.

13.2.2 Collaboration

Collaboration is not as common as the case of competition. Still, it can happen.

This cannot be seen in a visual, correlation-based way. Instead, it's math. Adding $x_j\beta_j$ reduces s^2, which makes $\hat{\beta}_{3-j}$ significant because variances are all reduced in a t-test for other predictors.

Partial Correlations

For example, say we are looking at heights and spelling aptitudes of schoolkids from ages 7 to 12. We get a positive correlation. We know that height isn't determining spelling; kids are both growing in age and spelling ability during these years. So how can we get at something more intrinsic?

We can first regression height on age, and then get residuals. The residuals are "age-adjusted height," or how much above/below an observation is from the regression line. We can also do the same by regressing spelling on age, getting "age-adjusted spelling." Then, we plot the two sets of residuals and look for some correlation. That is, we look for the *partial correlation* of X_i, X_j adjusting for X_k. Weisberg refers to this as an *added variable plot*.

So,

$$\text{Partial correlation of } X_i, X_j \text{ adjusting for } X_k = \text{Corr}(\text{Resid. for } X_i \text{ on } X_k,\ \text{Resid. for } X_j \text{ on } X_k)$$

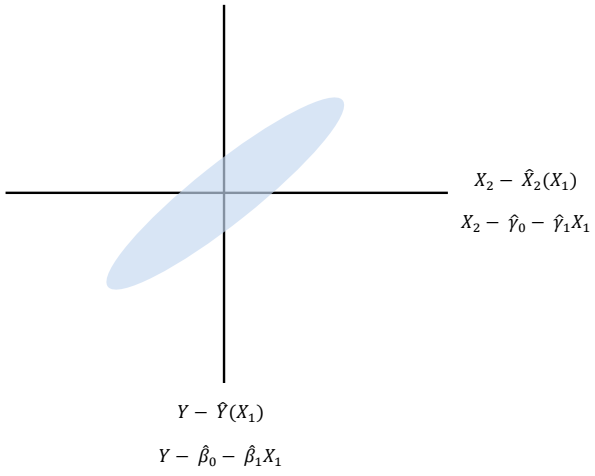


Figure 13.4: An added variable plot. Given the tight positive correlation in this plot, it is a good idea to add X_2 to the regression. Something in X_2 is adding to the predictive power of the model. If there was no correlation (residuals seemed randomly scattered around the origin), then we can probably leave X_2 out.

For a Gaussian population, we get

$$\rho_{ij \cdot k} = \rho_{ij|k} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sqrt{(1-\rho_{ik}^2)(1-\rho_{jk}^2)}}$$

(Here, we write this without hats. If we use true correlations, then we get a true partial correlation. If we use sample correlations, we get a sample partial correlation.)

This links back to Cornfield's argument. If there is an outside effect, it must be strong to overturn the high correlation that we already see. Note that if $\rho_{ij} > 0$, then in order to change the overall sign of $\rho_{ij \cdot k}$, we need $\rho_{ik}\rho_{jk} > \rho_{ij}$. If $\rho_{ij} = 0.8$, then it must be that ρ_{ik} and ρ_{jk} are greater than 0.8 in order to change the sign and overturn the observed effect. (In the smoking case, if the correlation between smoking and cancer is 0.8, then we must find genetic-smoking and genetic-cancer correlations that are each greater than 0.8. It is highly unlikely that such a strong correlation would exist without us noticing it beforehand, so the claim that genetics explain both smoking and cancer is probably not true.) The formula also shows that it is easier to overturn smaller correlations/effects.

Another example is of height, weight, and girth. It could be that $\rho_{\text{HT, WT}} > 0$, but $\rho_{\text{HT, WT} \cdot \text{GIRTH}} < 0$. The choice of what ρ we care about is a substantive and not a statistical matter.

Moreover, note that we can find partial correlations of X_1, Y adjusting for $X_2, ..., X_k$. We just regress those sets of predictors out of both X_1 and Y and find the partial correlation.

Chapter 14

Automatic Variable Selection

So far, we have talked about "manual" methods to use to choose models. We could consider having a computer pick the best model for us.

Suppose we have Y and $X_1, X_2, ..., X_q$ where q is large. Then, there exist 2^q possible regression models. A computer can pick a "best" model, but it has to be taught a bit more than to simply reduce the residuals. (If not, it will always just pick the full model, since adding variables never increases the residuals.) This links to the bias-variance trade-off: Large subsets of predictors result in higher variance, while smaller subsets of predictors result in higher bias.

14.1 Stepwise Methods

Say that we have three predictors and are wondering how many/which of them to include in a model. We plot the sum of squared residuals over all possible combinations of the predictors in Figure 14.1. For cases of small q such as this, it can calculate all possible RSS values on branches.

We can summarize this in a table, as well:

# Predictors	Best	RSS
0	$\oslash$	10
1	$\{3\}$	5
2	$\{1, 2\}$	2
3	$\{1, 2, 3\}$	1

There are two *stepwise* methods we can use to pick the ideal model. In a *forward* stepwise approach, we start with $\oslash$ and add the best predictor if it is statistically significant; we stop otherwise. In a *backward* stepwise approach, we start will all predictors and drop the least significant predictor if it is not statistically significant; we stop otherwise.

Both have potential shortcomings. For one thing, the two methods often will not agree on the same model. In the figure above, forward search stops at $\{3\}$; backward search stops at $\{1, 2\}$. But also, neither approach looks into the future. While the next predictor to be added/removed may not be important, a predictor later down the line could be incredibly important but not be reached. We

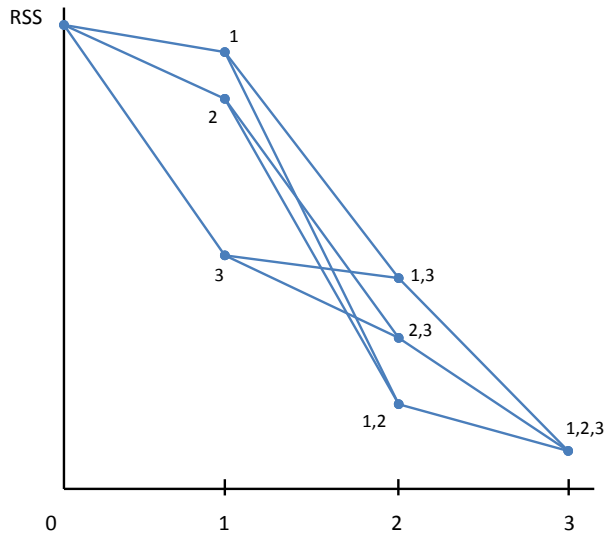


Figure 14.1: RSS as we add three predictors to the model.

might want our algorithm to look ahead a few steps (how many steps is another parameter that we have to choose). So, we might use a *hybrid* method that mixes both forward and backward steps.

Moreover, if we use these methods, it is not legitimate to use t-statistics, p-values, or confidence intervals from our final models, acting as if they were the only model we had. We were wandering through the tree. There is a serious multiplicity issue here. The distribution of the result $\hat{\beta}_j$'s is not random (after all, the selection process literally chose the predictors it found to be significant), so t-tests cannot apply here, nor will bootstrap methods resolve the issue. See Figure 14.2, and refer to Breiman's "quiet scandal" article on the course website.

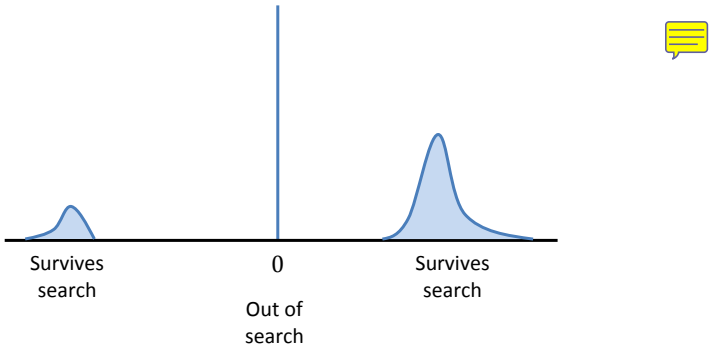


Figure 14.2: The distribution of $\hat{\beta}_j$ is not suited for hypothesis testing. From Breiman: "Over many simulated runs of the model, each time generating new random noise, and selecting, say, a subset of size four, the coefficient estimates of a given variable have a point mass at zero, reflecting the probability that the variable has been deleted. In addition, there is a continuous mass distribution over those times when the variable showed up in the final four-variable equation. The relation of this distribution to the original coefficients is obscure" (751).

On the plus side, these methods give us predictions that we can cross-validate to measure accuracy.

This also helps us leave out some variables, which could help us save money, time, and more later on.

In order to pick a model that does best without using too many predictors, we might teach a computer to do the following:

$$\min[\text{ sum of squared errors } + \text{ penalty for \# of parameters}]$$

There are many varieties of this.

14.2 Mallow's C_p

Mallow's C_p is based on expected squared error (ESE):

$$\begin{aligned} ESE &= \mathrm{E}\left[\sum_{i=1}^{n}(\hat{Y}_i - \mathrm{E}[Y_i])^2\right] \\ &= \sum_{i=1}^{n}\mathrm{V}(\hat{Y}_i) + \sum_{i=1}^{n}\mathrm{bias}_i^2 \\ &= p\sigma^2 + \sum_{i=1}^{n}\mathrm{bias}_i^2 \end{aligned}$$

We can't directly evaluate this. But, we do know the following:

$$\begin{aligned} RSS &= \sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \\ \mathrm{E}[RSS] &= \ldots = (n-p)\sigma^2 + \sum_{i=1}^{n}\mathrm{bias}_i^2 \end{aligned}$$

Therefore,

$$\begin{aligned} ESE &= \mathrm{E}[RSS] - (n-2p)\sigma^2 \\ &= \mathrm{E}[RSS] - n\sigma^2 + 2p\sigma^2 \end{aligned}$$

Roughly speaking, the $2p\sigma^2$ is a penalty for additional parameters.

So then,

$$\widehat{ESE} = RSS - (n-2p)\hat{\sigma}^2 \qquad \text{where usually} \qquad \hat{\sigma}^2 = \frac{1}{n-q}\sum_i (Y_i - \hat{Y}_i)^2$$

This estimate of $\hat{\sigma}^2$ is okay if we assume that the full model has no bias.

Then, we get that

$$C_p = \frac{\widehat{ESE}}{\hat{\sigma}^2} = \frac{RSS}{\hat{\sigma}^2} - n + 2p$$

We can try plotting C_p with respect to the number of parameters in the model, p (where q is all of them). If the model has no bias, models should be around the line with slope 2 (because of the $2p$ in the expression). See Figure 14.3.

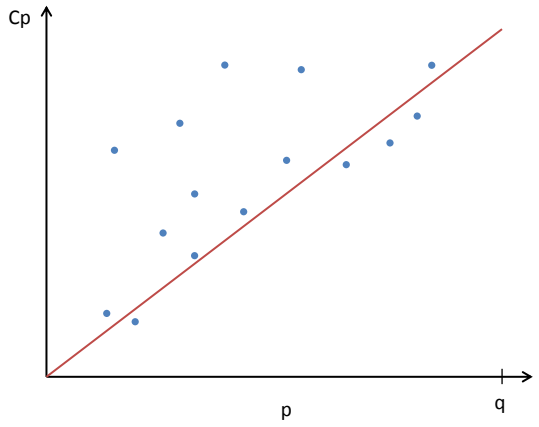


Figure 14.3: A plot of C_p. Models that are represented by dots around the line have minimal bias. Points far from the line have lots of bias.

If points are above the line, then there is some bias left unaccounted for by the model. We ultimately pick the model that minimizes C_p.

14.3 Least Squares Cross-Validation

Let $\hat{Y}_{-i}$ be the predictor of Y_i on the model, fit to $(X_{i'}, Y_{i'})$ where $i' \neq i$. That is, we run the regression on data without observation i. We are dealing with

$$\hat{Y}_{-i} = Z_i^T \hat{\beta}_{-i}$$

$\hat{\beta}_{-i}$ thus depends on the set of predictors in the model. We then measure cross-validation as:

$$\text{CV(model)} = \sum_{i=1}^{n} (Y_i - \hat{Y}_{-i})^2$$

Cross-validation has some nice properties. Importantly, it will knock out models that overfit (models with high variance and low bias). High-order polynomials are penalized in CV processes, because leaving out a single point that dictates the polynomial would cause really large residuals. CV also parallels our goal of predicting new values; it is as if we are predicting the "future" on the one $-i$ observation.

To find overall diagnostic on cross-validation, we use

$$\text{E}[CV] = n\sigma^2 + \sum_i \text{E}[(\hat{Y}_{-i} - \text{E}[Y_i])^2]$$

The long way of cross-validating for $i = 1, ..., n$ leaves out each (X_i, Y_i) and gets $\hat{\beta}_{-i}$. Then, $\hat{Y}_{-i} = Z_i^T \hat{\beta}_{-1}$. However, calculating this involves $O(n^2)$ work, which may be excessive and prohibitive. In the case of linear models, there is a shorter way of calculating cross-validation.

14.3.1 The Short Way

This method, limited to linear models, uses the following important property:

$$\hat{Y}_i = H_{ii} Y_i + (1 - H_{ii}) \hat{Y}_{-i}$$

where $H = Z(Z^T Z)^{-1} Z^T$ is the hat matrix, and H_{ii} is the i'th diagonal entry. Using this, we no longer have to perform $O(n^2)$ computations. Instead,

$$\hat{Y}_{-i} = \frac{\hat{Y}_i - H_{ii} Y_i}{1 - H_{ii}}$$

So, we can perform *just one regression* to get $\hat{Y}$ and $\text{diag}(H)$ and cross-validate everything using this. Then, the residual is

$$Y_i - \hat{Y}_i = ... = \frac{Y_i - \hat{Y}_i}{1 - H_{ii}}$$

We take the naive residual and apply an inflation factor (note that $1 - H_{ii} \geq 0$).[1] Finally, we get the cross-validation measure:

[1] If we ever get that $H_{ii} = 1$, then the left-out Y_i depends solely on its own X_i. It ignores all remaining $n - 1$ points. So if $H_{ii} = 1$, the formula breaks down because of division by zero. The breakdown of this formula here is therefore justified.

$$CV = \sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{(1 - H_{ii})^2} = \sum_{i=1}^{n} \frac{\hat{\varepsilon}_i^2}{(1 - H_{ii})^2}$$

This is also called *Allen's PRESS*. Again remember that this shortcut only works for linear models. A short method may not exist for non-linear cases.

14.3.2 Algebraic Aside

Where does $\hat{Y}_{-i} = H_{ii}Y_i + (1 - H_{ii})\hat{Y}_i$ come from? Given how important this equality is, it is worth showing.

First, recall that $H = Z(Z^TZ)^{-1}Z^T$. An element in the hat matrix is $H_{ij} = Z_i^T(Z^TZ)^{-1}Z_j$. To go further, we need an additional formula.

Theorem 3. Sherman-Morrison Formula If $A \in \mathbb{R}^{n \times n}$ is invertible, $u, v \in \mathbb{R}^n$, and $1 + v^TAu \neq 0$, then

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}$$

A small change to A changes the inverse by some predictable amount.

Then,

$$(Z^TZ)_{(-i)} = Z^TZ - Z_iZ_i^T$$

We pull down the rank of the matrix by one. By Sherman-Morrison Formula, the inverse is then

$$\begin{aligned}(Z^TZ)_{(-i)}^{-1} &= (Z^TZ)^{-1} + \frac{(Z^TZ)^{-1}Z_iZ_i^T(Z^TZ)^{-1}}{1 - Z_i^T(Z^TZ)^{-1}Z_i} \\ &= (Z^TZ)^{-1} + \frac{(Z^TZ)^{-1}Z_iZ_i^T(Z^TZ)^{-1}}{1 - H_{ii}}\end{aligned}$$

The inverse is then close to what it was before, with a small shift.

We know that $(Z^TY)_{(-i)} = Z^TY - Z_iY_i$. Then,

$$\begin{aligned}
\hat{\beta}_{-i} &= \left[(Z^T Z)^{-1} + \frac{(Z^T Z)^{-1} Z_i Z_i^T (Z^T Z)^{-1}}{1 - H_{ii}}\right] [Z^T Y - Z_i Y_i] \\
&= \hat{\beta} - (Z^T Z)^{-1} Z_i Y_i + \frac{(Z^T Z)^{-1} Z_i Z_i^T \hat{\beta}}{1 - H_{ii}} - \frac{(Z^T Z)^{-1} Z_i H_{ii} Y_i}{1 - H_{ii}} \\
&= \hat{\beta} + \frac{-(1 - H_{ii})(Z^T Z)^{-1} Z_i Y_i + (Z^T Z)^{-1} Z_i Z_i^T \hat{\beta} - (Z^T Z)^{-1} Z_i H_{ii} Y_i}{1 - H_{ii}} \\
&= \hat{\beta} - \frac{(Z^T Z)^{-1} Z_i (Y_i - \hat{Y}_i)}{1 - H_{ii}}
\end{aligned}$$

Use this to find $\hat{Y}_{-i}$:

$$\begin{aligned}
\hat{Y}_{-i} = Z_i^T \hat{\beta}_{-i} &= \hat{Y}_i - \frac{H_{ii}(Y_i - \hat{Y}_i)}{1 - H_{ii}} \\
(1 - H_{ii})\hat{Y}_{-i} &= (1 - H_{ii})\hat{Y}_i - H_{ii}(Y_i - \hat{Y}_i) \\
\hat{Y}_i &= H_{ii} Y_i + (1 - H_{ii})\hat{Y}_{-i}
\end{aligned}$$

And we get the desired equality. Again, this lets us do one linear regression and pull out the hat matrix and $\hat{Y}$ and thus do leave-one-out cross-validation much more quickly.

We pick out the model with the lowest CV. Leaving out an important variable from the model causes large increases in CV. Leaving out unimportant variables results in small, insubstantial increases in CV.

14.4 Generalized CV

The average value in H_{ii} is $\frac{p}{n}$. So what would happen if we just used $\frac{p}{n}$ everywhere in calculating CV? That is the notion behind generalized cross-validation. If $\frac{p}{n}$ is small, then

$$\begin{aligned} GCV &= \sum_{i=1}^{n} \frac{\hat{\varepsilon}_i^2}{\left(1-\frac{p}{n}\right)^2} \\ &= \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \times \frac{1}{\left(1-\frac{p}{n}\right)^2} \\ &= \left(\sum_{i=1}^{n} \hat{\varepsilon}_i^2\right)\left(1+\frac{2p}{n}\right) \qquad \text{(Taylor approximation)} \\ &= \left(\sum_{i=1}^{n} \hat{\varepsilon}_i^2\right)\frac{1+\frac{p}{n}}{1-\frac{p}{n}} \\ &= \left(\sum_{i=1}^{n} \hat{\varepsilon}_i^2\right)\frac{1}{1-\frac{2p}{n}} \end{aligned}$$

But *generalized* cross-validation is a misnomer. The case where $H_{ii} \approx \frac{p}{n}$ for all i is actually an exception to the rule; it is not a "general" case. While it is an admittedly convenient shortcut, we might prefer to stick with regular cross-validation.

14.5 AIC and BIC

Akaike's Information Criterion (AIC) is defined as

$$AIC = n \log \frac{RSS}{n} + 2p$$

Bayes's Information Criterion (BIC) is defined as

$$BIC = n \log \frac{RSS}{n} + p \log n$$

We pick the model that minimizes AIC or BIC.

14.5.1 AIC vs. BIC

Which criterion do we use? How do they differ?

Penalties

With AIC, the penalty factor for each additional predictor is 2. For BIC, it is $\log n$. In most cases, $\log n > 2$. BIC will penalize predictors more, so BIC tends to prefer smaller models than AIC.

Intent to Use

AIC is better for prediction. BIC is better at getting the "right" model. This seems odd, but the math backs up the distinction.

With BIC, we let $n \to \infty$ while keeping q (the total number of predictors) fixed. Assume that the true model is $\mathrm{E}[Y] = Z\beta$—an actually linear model where some β_i's are actually valued zero. Asymptotically, BIC will correctly differentiate between variables that are zero and non-zero. That is, it is right in knowing what should and should not be in the model.

Meanwhile, AIC will also pick out the non-zero variables, but may also pick up a few more.

On the way to ∞, AIC is more accurate in making predictions, but you never know if it has truly chosen the right model. Once n "equals" ∞, BIC is definitely correct. Asymptotically, $AIC \approx CV$. But it is also worth noting that the difference between AIC and BIC may not be large at all.

AIC and BIC are more "principled" approaches than the random walk model. Still, with any of these methods, it is dubious to use confidence intervals and tests, acting as if the model we chose is the final model. We should not say things like "a one-unit change in x leads to a β change in y."

14.6 Ridge Regression

We may be drawing a false dichotomy in saying that a predictor should be left in or out of a model in such a binary manner. The knife-edge nature of this, especially in cases with lots of close calls, may not be a good idea. Ridge regressions were created to deal with this and were especially devised to deal with nearly singular Z^TZ matrices.

The estimate from ridge regressions is $\tilde{\beta}$.

$$\tilde{\beta} = [Z^TZ + \lambda I]^{-1}Z^TY \qquad \lambda > 0$$

This method clearly shrinks $\tilde{\beta}$ to zero. As $\lambda \to \infty$, $\tilde{\beta}$ goes to 0. As $\lambda \to 0$, $\tilde{\beta}$ goes to $\hat{\beta}$.

However, we may not want to shrink the intercept. In that case, we might to the following:

$$\tilde{\beta}_\lambda = \left[Z^TZ + \lambda \begin{pmatrix} 0 & \\ & I \end{pmatrix}\right]^{-1} Z^TY \qquad \lambda > 0$$

The first diagonal entry is 0 instead of 1 in order to avoid rescaling the intercept. We call this method "ridge regression" because we are adding a "ridge" to the diagonal of Z^TZ.

Or, we can replace Y_i with $Y_i - \overline{Y}$ and Z_{ij} with $Z_{ij} - \overline{Z}_{ij}$. Then, we do a regression through the origin to get $\tilde{\beta}$ and use $\overline{Y}$ as the intercept. That is, the regression line is

$$\overline{Y} + Z\tilde{\beta}$$

Furthermore, we can use other weight matrices beyond the identity matrix. We can do some

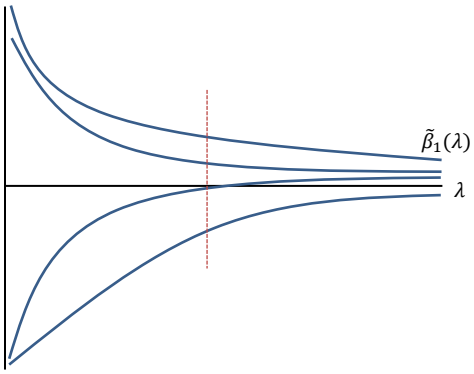


Figure 14.4: An example ridge trace plot. We might set λ equal to the vertical line, where the lines seem to stabilize. But this is very imprecise.

$$\tilde{\beta} = [Z^T Z + \lambda\mu]^{-1} Z^T Y \qquad \lambda > 0$$

where μ is some other matrix that reflects different intuitions (or weightings).

14.6.1 Optimization

In ridge regressions, we then attempt to minimize the following:

$$\sum_{i=1}^{n} (Y_i - Z_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\sum_{j=1}^{p} \beta_j^2$ is a penalty for additional parameters.

This expression also reflects the bias-variance trade-off. For large λ, we get high bias. For small λ, we get high variance.

How do we know what λ to pick? We could use (generalized) cross-validation. In the past, people would also plot out ridge traces and choose a λ where all the lines on the plot seemed to "stabilize." However, this is clearly not a rigorous method. Cross-validation seems like a better option.

Ridge regressions may be more accurate, give better predictions, are less chaotic, and are more principled than cross-validation and random walk techniques. On the flipside, ridge regressions give non-zero values for all covariates. Some other methods point out which covariates actually matter and drop the rest, which may be useful in practical terms.

14.6.2 A Bayesian Connection

We can approach ridge regression from a roughly "Bayesian" perspective. Suppose the following:

$$Y \sim \mathcal{N}(Z\beta, \sigma^2 I_n) \qquad \beta \sim \mathcal{N}(0, \tau^2 I_p)$$

The posterior distribution of β is proportional to the following:

$$\exp\left(-\frac{1}{2\sigma^2}(Y - Z\beta)^T(Y - Z\beta)\right) \times \exp\left(-\frac{1}{2\tau^2}\beta^T\beta\right)$$

The expression is all quadratics in β. This is helpful.

Then, maximizing the posterior density of β is the same as minimizing $\frac{1}{2\sigma^2}||Y - Z\beta||^2 + \frac{1}{2\tau^2}||\beta||^2$. Or, in other terms, the goal is

$$\min ||Y - Z\beta||^2 + \frac{\sigma^2}{\tau^2}||\beta||^2$$

This has the same structure as ridge regression if we say

$$\lambda = \frac{\sigma^2}{\tau^2}$$

So, if τ is huge or σ is tiny, then λ is small and only adds a small bump to the ridge and β is largely unchanged. If τ is tiny or σ is huge, then we add a big ridge along the diagonal and β shrinks dramatically toward zero.

Note that this is also the posterior mean, which is what Bayesians usually care about. Lastly, this method gives no zero estimates for β.

Calculation

How do we actually minimize this? We instead deal with this:

$$\min ||Y^* - Z^*\beta||^2$$

where

$$Y^* = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad Z^* = \begin{bmatrix} Z_{11} & \dots & Z_{1p} \\ \vdots & & \vdots \\ Z_{n1} & \dots & Z_{np} \\ & \frac{\sigma}{2}I & \end{bmatrix}$$

We append p zeroes to Y and a scaled identity matrix to Z. The top portion of Z are the squared

errors, while the identity matrix represents the penalties. In the expression $\min ||Y - Z\beta||^2 + \frac{\sigma^2}{\tau^2}||\beta||^2$, the $||Y - Z\beta||^2$ matches with the n Y values, and the $||\beta||^2$ matches with the p zeroes.

Solve this using an SVD. Remember the set-up:

$$Z = UDV^T \qquad D = \text{diag}(d_1, d_2, ..., d_p) \quad \text{where} \quad d_1 \geq ... \geq d_p \geq 0$$

The possibility of having d_i's that equal zero allows for rank deficient matrices. Then, we can rewrite $\tilde{\beta}_\lambda$ (remembering that $VV^T = I$ because V is orthogonal):

$$\begin{aligned}
\tilde{\beta}_\lambda &= (Z^T Z + \lambda I)^{-1} Z^T Y \\
&= (VD^T DV + \lambda I)^{-1} VD^T U^T Y \\
&= [V(D^T D + \lambda I)V^T]^{-1} VD^T U^T Y \\
&= V(D^T D + \lambda I)^{-1} D^T U^T Y \\
\tilde{Y}_\lambda &= Z\tilde{\beta}_\lambda \\
&= UD(D^T D + \lambda I)^{-1} D^T U^T Y \\
&= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda} U_j U_j^T Y
\end{aligned}$$

This is in contrast to the normal least squares estimate, which is the same as a ridge regression where $\lambda = 0$:

$$\hat{Y} = \tilde{Y}_0 = UU^T Y = \sum_{j=1}^{p} U_j U_j^T Y$$

So if d_j is small, there is lots of weighting on β down to zero.

We can compute ridge regressions this way. Furthermore, note a similarity to the standard hat matrix:

$$H_\lambda = UD(D^T D + \lambda I)^{-1} D^T U^T$$

Then, the "degrees of freedom" here is $\text{tr}(H_\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$. This is the expression used in GCV:

$$\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_{i\lambda})^2}{\left[1 - \frac{df(\lambda)}{n}\right]^2}$$

14.6.3 An Alternative

One other approach would be to keep the first k columns of V and dropping the last $p - k$. Then,

$$\sum_{j=1}^{k} U_j U_j^T Y$$

This is almost like principal components regression. For actual principal components regression, drop the intercept term and make the columns of Z have mean zero. (Optionally, also make the columns have equal variance.)

14.7 Principal Components

Suppose that $Z_i \in \mathbb{R}^d$, where d is large. PC regression attempts to choose dimensions with the highest variance (and thus explain the most with the fewest ingredients). In more mathematical terms,

$$\max \mathrm{V}(Z^T U) \qquad \text{s.t.} \qquad U^T U = 1$$

And usually, $\mathrm{E}[Z] = 0$. So, principal components picks out k most important dimensions and projects the data from d dimensions down to $k < d$ dimensions.

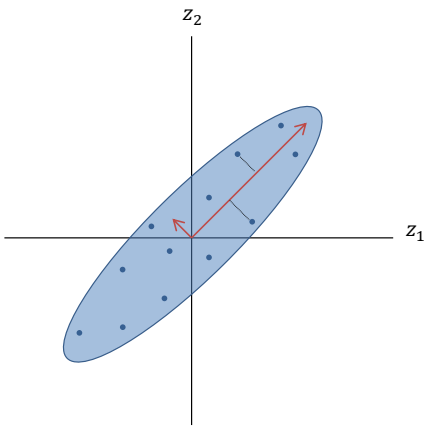


Figure 14.5: Principal component regression. The longer ray represents the first principal component; it has the highest variance.

The method helps reduce the number of covariates, which could allow for regressions when they weren't tractable before. For example, say you want to know about 30,000 genes but have samples from only 200 patients. PCR might narrow things down to just 30 key components that can be analyzed in a regression. However, this all comes at the risk of knocking out predictors that may actually best account for Y (since PCR is more about capturing variance in Z).

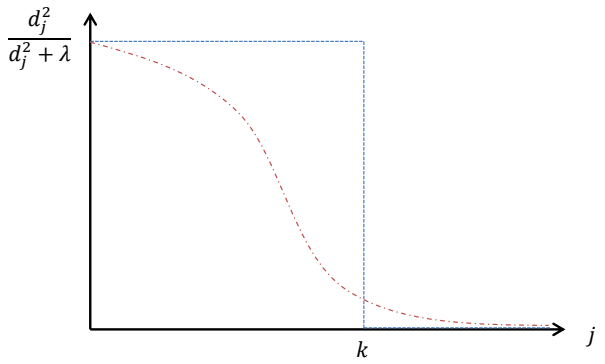


Figure 14.6: The difference of weights with ridge (red, dash-dotted line) and principal components (blue, dotted line) regressions.

14.8 L_1 Penalties

This is also known as lasso or basis pursuit. Here, we deal with the following:

$$\min \sum_{i=1}^{n} (Y_i - Z_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

In linear algebra terms, $||Y - Z\beta||_2^2 + \lambda||\beta||_2$.

Then, there are three different ways to approach this.

- Vary λ.
- Minimize $||Y - Z\beta||^2$ such that $||\beta||_2 \leq t$ and vary t. If $t = \infty$, then we just get the least squares estimate. If $t = 0$, then $\beta = 0$. So like in all previous cases, the choice between least squares and $\beta = 0$ is really a continuum.
- Minimize $||\beta||_2$ such that $||Y - Z\beta||^2 \leq t$.

The difference between lasso and ridge regressions might be useful to see visually.

L_1 allows for more exact zero estimates because the bounds are "pointy." This is good for practical terms; getting zero estimates allows for researchers to save money by not measuring things that are found to have no effect.) It also allows for more shrinkage of β.

Both lasso and ridge regressions sacrifice unbiasedness for reduced variance.

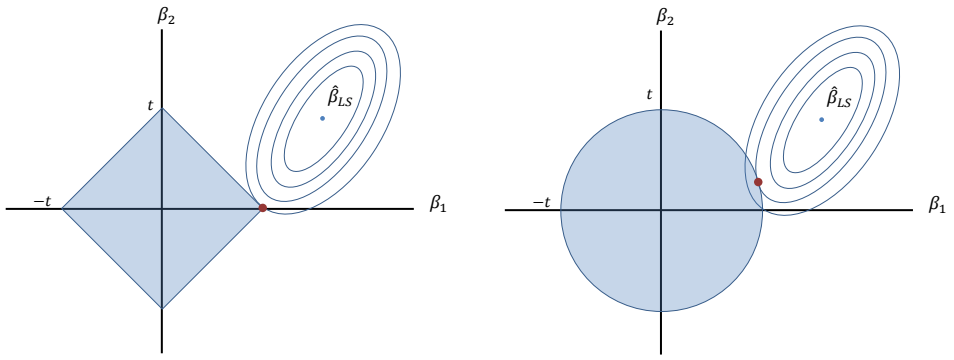


Figure 14.7: Contour plots of $||Y - Z\beta||^2$ with lasso and ridge regressions. Note that in this example, lasso makes $\beta_2 = 0$ while ridge regression gives non-zero coefficients. Lasso is generally more likely to give zero coefficients.

Chapter 15

Causality

We will take an incredibly short detour into causality. While regressions easily deal with association, identification for causality is much harder. One useful method worth considering is regression discontinuity.

15.1 Regression Discontinuity Designs

See Angrist and Pischke for more on RDDs.

Suppose we want to study the relationship between National Merit Scholarships and future earnings.

$$\begin{aligned} X_1 &= \text{Whether student receives National Merit Scholarship} \\ Y &= \text{Future earnings (or whether student went to grad school)} \end{aligned}$$

A regression does not ascertain whether $X_1 \to Y$. There could be a selection issue: Even if a positive correlation existed between the two, it could be that students more innately motivated students did well with National Merit and future earnings. Suppose we also have data on PSAT scores:

$$X_2 = \text{PSAT score}$$

National Merit basically works so that students with a PSAT score over a certain threshold τ become National Merit Finalists, while those below do not. Then, we can think about comparing the two groups of students on each side of this threshold:

$$\begin{aligned} \tau \leq X_2 \leq \tau + \varepsilon \qquad & \text{barely above threshold} \\ \tau - \varepsilon \leq X_2 < \tau \qquad & \text{barely missed it} \end{aligned}$$

The two groups are similar enough to compare. There is little reason to think that a selection issue exists between these two groups that are dancing on either side of the line.

The size of the discontinuity is then the effect size.

$$Y = \beta_0 + \beta_1 \underbrace{X_1}_{\text{binary}} + \beta_2 X_2 \qquad \text{or} \qquad \text{poly}(X_2) + \beta_1 X_1$$

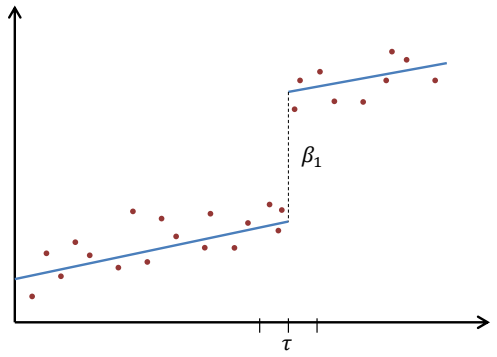


Figure 15.1: Regression discontinuity design.

Another example is the effect of class sizes on elementary school performance. It could be that class sizes are based on past performance (gifted students are placed into small classes, or delinquent students are placed into particularly small classes, etc.). This would undermine inference.

However, in Israel, a rule exists that limits classes to 40 or less. If a class gets 41 or more students, then it must be split in two. Then, we can perform a student that compares classes of size 38-40 with those of 41-42 that were split into smaller classes.

One potential failure of RDD is if a steep but continuous increase happens to occur at a real-world τ that we use in a study. Then we would believe a discontinuity exists when it is really just a non-linear relationship. But this is probably unlikely and should not stop us from using this method.

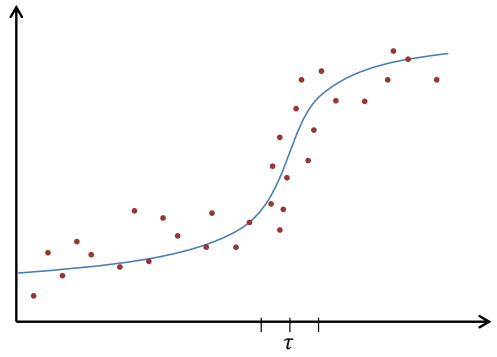


Figure 15.2: A potential but unlikely failure.

Chapter 16

Violations of Assumptions

Recall out usual linear model, $Y_i = Z_i^T\beta + \varepsilon_i$, where ε_i are independent and identically distributed $\mathcal{N}(0, \sigma^2)$. What can go wrong? There could be bias; that is, maybe $\mathrm{E}[Y_i] \neq Z_i^T\beta$. Or it could be that variances are not common across observations; that is, $\mathrm{V}(Y) \neq \sigma^2 I$. Or it could be that the residuals are not normally distributed. So what do we do when the linear model's assumptions do not hold up? How do we detect these issues?

16.1 Bias

Suppose that $\mathrm{E}[Y|X] \neq Z(X)\beta$. It might instead be the case that the expected value is some general parametric function, $g(X, \theta)$. Or, there may be some other data we have that is important. For instance, say that in reality, $\mathrm{E}[Y|X] = Z\beta + \omega\gamma$. Then,

$$\mathrm{E}[\hat{\beta}] = (Z^TZ)^{-1}Z^T(Z\beta + \omega\gamma) = \beta + (Z^TZ)^{-1}Z^T\omega\gamma$$

If $Z^T\omega = 0$, then there is no problem. But since this is generally not zero, we want to know how we might detect ω. Some options come to mind.

1. We could plot $\hat{\varepsilon}_i$ versus ω. If there is a strong linear correlation between the residuals, or any other functional dependence (such as $\hat{\varepsilon} \sim \omega^2$ or $\hat{\varepsilon} \sim |\omega|$), then we should be suspicious.

2. A better method is to plot $\hat{\varepsilon}$ against the residual of ω on Z (which gives as the added variable plot). Note that since ε is orthogonal to Z, there is no need to also compute the further residual of ε on Z.

3. Perhaps the easiest thing to do is to add ω to the regression model and test goodness of fit via the extra sum of squares.

In general, consider all possible administrative things the data could depend on. For instance, plot $\hat{\varepsilon}_i$ versus i. Or perhaps two data sets put together are not i.i.d. Look for trends and blocks, considering file order, calendar time, and the like. In principle, we should do this for every variable. In practice, this may not be possible.

16.1.1 Detection

How can we detect a more general lack of fit?

Suppose the data was collected so that we have repeated observations for the same value of X, as the case in Figure 16.1.

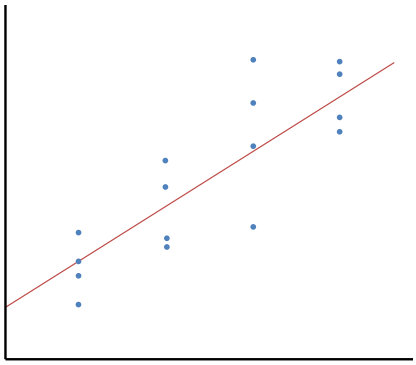


Figure 16.1: Example data with repeated observations and best-fitting line.

In particular, suppose that there are n_i observations with $X = X_i$. (In the figure, $n_i = 4$.) Then, the model can we written as $Y_{ij} = Z_i^T\beta + \varepsilon_{ij}$. Then, one way to test for goodness of fit is to compare the model $Z^T\beta$ to the model $Y_{ij} = \mu(X_i) + \varepsilon_{ij}$, where

$$\hat{\mu}(X_i) = \frac{1}{n_i}\sum_{j=1}^{n_1} Y_{ij}$$

Then, calculate the *extra sum of squares*. This requires finding the *pure* and *lack of fit sum of squares*:

$$\text{pure error sum of squares} = \sum_{i=1}^{N}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\cdot})^2$$

$$\text{lack of fit sum of squares} = \sum_{i=1}^{N}\sum_{j=1}^{n_i}(\overline{Y}_{i\cdot} - Z_i^T\hat{\beta})^2$$

Note that the second summand does not depend on j (so we could take out factors of n_i). The total error of sum squares is then:

$$\begin{aligned}\text{total error sum of squares} &= \sum_{i=1}^{N}\sum_{j=1}^{n_i}(\overline{Y}_{ij} - Z_i^T\hat{\beta})^2 \\ &= \text{lack of sum of squares} + \text{pure error sum of squares}\end{aligned}$$

The corresponding F statistic to be used in a hypothesis test is then

$$F = \frac{\frac{1}{N-p}(\text{lack of fit sum of squares})}{\frac{1}{\sum(n_i-1)}(\text{pure error sum of squares})}$$

Even if one does not have repeated observations at identical X values, it is still possible to play this trick if there are groups of observations at very similar X values. Grouping this way is a possibility. However, since there will then be some spread amongst the X values, the pure error sum of squares will be inflated and the F test is less likely to be significant (since the denominator of the F statistic is growing).

Also note that even if there are repeated observations for the same X value, there still may be a reason to worry that there are not "genuine" replicas. If there is some correlation within the data on a given day, then the pure error sum of squares will be too small (positive correlation). This could happen if, for instance, the X values are days of the week and the operating taking the data changed from day to day (and different operators collect the data differently).

16.1.2 Transformations

Another way out of a situation where the linear model poorly fits the data is to transform the data. For example, a popular model in economics is the *Cobb-Douglas* model:

$$Y = \beta_0 X_i^{\beta_1} ... X_p^{\beta_p}(1+\varepsilon)$$

where ε is the typical normally distributed residual term. We can transform this into a linear model by taking a logarithm:

$$\log(Y) = \log(\beta_0) + \beta_1 \log(X_1) + ... + \beta_p \log(x_p) \log(1+\varepsilon)$$

Note that this would not work for $Y = \beta_0 X_i^{\beta_1} ... X_p^{\beta_p} + \varepsilon$.

A word of caution about taking logarithms: A higher R^2 on a log scale may not mean a more accurate prediction of Y. For example, consider $e^{Z\hat{\beta}}$ in which errors would be exponentiated. (Moreover, note that this model is biased due to Jensen's inequality.) Logarithms can pull in large outlying values and can make small values become outliers. This latter problem (and the problem of evaluating $\log(0)$) can be avoided by considering $\log(1+Y_i)$ or $\log\left(\frac{Y^*}{2} + Y_i\right)$, where Y^* is the smallest positive Y_i.

Box Cox Transformations

A general class of transformations (parameterized by λ) called Box Cox transformations is given by the following:

$$\hat{Y} = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(Y) & \lambda = 0 \end{cases}$$

The -1 in the numerator above comes from the desire to to have $\hat{Y} = \log(Y)$ as $\lambda \to 0$. For example, $\lambda = 1$ corresponds to $\hat{Y} = Y$; $\lambda = 2$ corresponds to $\hat{Y} = Y^2$; $\lambda = 0$ corresponds to $\log(Y)$; and so on.

One way to use this transformation is to fit $\tilde{Y}(\lambda) = Z\beta + \varepsilon$, and then estimate λ by maximum likelihood, perhaps numerically using a grid-based search. Often, the data are compatible with a wide range of λ values. For $\lambda \in \Lambda$, $\tilde{Y} = \frac{Y^\lambda - 1}{\lambda}$ or $\log(Y)$. Correspondingly,

$$\hat{\beta}(\lambda) = (Z^T Z)^{-1} Z^T \tilde{Y}$$

$$\hat{\sigma}(\lambda) = ... = \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - Z_i \beta)^2$$

16.2 Heteroskedasticity

Suppose $Y \sim \mathcal{N}(Z\beta, \sigma^2 V)$ where V is full rank and not necessarily I (and I is the ideal case). One prominent example could be when $V = \text{diag}(1/m_1, ..., 1/m_p)$. In that case, the Y_i are independent and $\text{V}(Y_i) = \frac{\sigma^2}{m_i}$. This could be the case if a different amount of data goes into each measurement. There could also be correlations in the data, such as

$$V = \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix}$$

in which each measurement is correlated with the ones before and after. (For example, if the measurements occur over time, nearby times might be correlated.)

We can deal with this issue using *generalized least squares*. Say that

$$\text{V} = P^T \Lambda P$$

where P is an orthogonal matrix and $\Lambda = \text{diag}(\lambda_1, ..., \lambda_n)$. Furthermore, let

$$D = \Lambda^{-1/2} P \qquad D^T D = ... = V^{-1}$$

So basically, D functions "like $\frac{1}{\sqrt{V}}$"—a multidimensional version of $\frac{1}{\sigma}$. Then,

$$\begin{aligned} DY &= DZ\beta + D\varepsilon \\ \tilde{Y} &= \tilde{Z}\beta + \tilde{\varepsilon} \end{aligned}$$

$$\begin{aligned} \mathrm{V}(\tilde{\varepsilon}) &= DVD^T\sigma^2 \\ &= \Lambda^{-1/2}PP^T\Lambda PP^T\Lambda^{-1/2}\sigma^2 \\ &= \sigma^2 I \end{aligned}$$

This result for $\tilde{\varepsilon}$, along with Gauss-Markov Theorem, says that the normal OLS-like method of finding β still works.

$$\begin{aligned} \hat{\beta}_{GLS} &= (\tilde{Z}^T\tilde{Z})^{-1}\tilde{Z}^T Y \\ &= (Z^T V^{-1} Z)^{-1} Z^T V^{-1} Y \end{aligned}$$

$$\hat{\beta}_{OLS} = (Z^T Z)^{-1} Z^T Y$$

Furthermore, it is important to note that Gauss-Markov Theorem suggests

$$\mathrm{V}(\hat{\beta}_{GLS}) \leq \mathrm{V}(\hat{\beta}_{OLS})$$

When we say one matrix is less than the other, we mean that $\mathrm{V}(\hat{\beta}_{GLS}) - \mathrm{V}(\hat{\beta}_{OLS})$ is negative semi-definite, or that $\mathrm{V}(c^T\hat{\beta}_{GLS}) \leq \mathrm{V}(c^T\hat{\beta}_{OLS})$ for all $c \in \mathbb{R}^p$.

16.2.1 A Special Case

Consider the special case where V is a diagonal matrix populated with $\sigma_1^2, ..., \sigma_n^2$. Then, we get

$$\left[\sum_{i=1}^{n} \sigma_i^{-2} Z_i Z_i^T\right]^{-1} \sum_{i=1}^{n} \sigma_i^{-2} Z_i Y_i$$

This is a weighted least squares. Be sure that division is by σ^2 and not σ. Intuitively, if Y_i is the average of $m_i \geq 1$ observations, then $\sigma_i^2 = \frac{\sigma^2}{m_i}$. Then, we are taking weights proportional to m_i.

16.2.2 Consequences

What happens if we should have used generalized least squares but find β_{OLS} instead (and thus don't account for substantial heteroskedasticity)? We know that both β_{OLS} and β_{GLS} are unbiased,

and that $\mathrm{V}(\bar{\beta}_{GLS}) \leq \mathrm{V}(\bar{\beta}_{OLS})$. More specifically, if we took the OLS approach, we get the following estimate for variance:

$$\hat{\mathrm{V}}_{OLS}(\beta_{OLS}) = s^2(Z^T Z)^{-1}$$

However, if we used the GLS approach,

$$\mathrm{V}(\beta_{OLS}) = \mathrm{V}((Z^T Z)^{-1} Z^T Y) = (Z^T Z)^{-1} Z^T V Z (Z^T Z)^{-1} \sigma^2 \neq \sigma^2 (Z^T Z)^{-1}$$

We do not get the correct variance estimate (unless $V = I$). So while our mean estimate is right, we get the wrong $\hat{\mathrm{V}}$, which gives us the wrong confidence intervals and test results. If $\hat{\mathrm{V}}$ is too small, then our confidence intervals are too short and p is too small. If $\hat{\mathrm{V}}$ is too large, then our confidence intervals are too wide and p is too large. So, doing OLS instead of GLS in this case will not give us dependable or useful estimates.

More generally, say that in reality, $\mathrm{V}(Y) = \sigma^2 V_1$ but that we think it is $\sigma^2 V_2$. Then, we get that

$$\begin{aligned} \hat{\beta}_{GLS} &= (Z^T V_2^{-1} Z)^{-1} Z^T V_2^{-1} Y \\ \hat{\mathrm{V}}(\hat{\beta}_{GLS}) &= (Z^T V_2^{-1} Z) s^2 \end{aligned}$$

and again, this is the incorrect variance estimate.

16.2.3 Detection

One standard method is to plot $\hat{\varepsilon}_i$ versus X_i. The two figures below are not things we want to see.

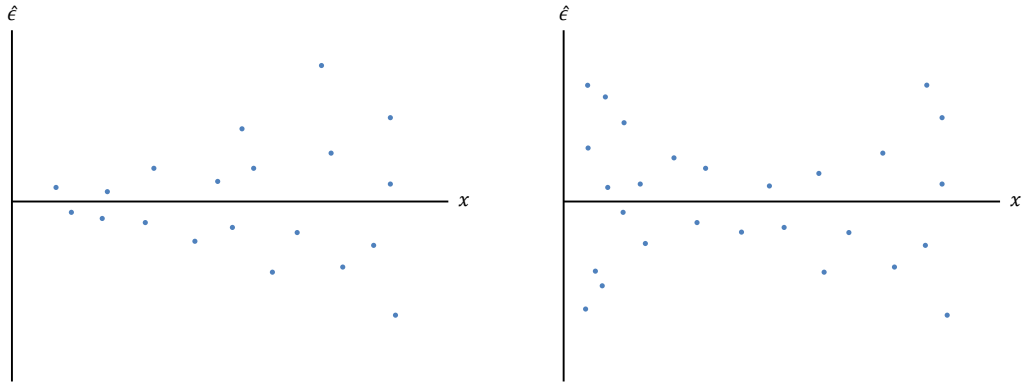


Figure 16.2: Two residual plots. In the first, variance of Y increases with X. This is a common pattern in data. Less common but possible is the second example, where variance increases at the extremes.

But these figures only work when there is one predictor. When there are many X's, then we can't draw this. Instead, we can plot $\hat{\varepsilon}_i$ to $\hat{Y}_i$ and look for similar patterns.

Another approach is to plot $\hat{\varepsilon}_i$ on $\hat{\varepsilon}_{i-1}$. If we get a plot like the one in Figure 16.3, then there is a dependence in errors. Specifically, there exists correlation between predictors that needs to be addressed using time-series methods. While $\hat{\beta}$ is still unbiased, we get erroneous variances so we cannot do proper statistical tests.

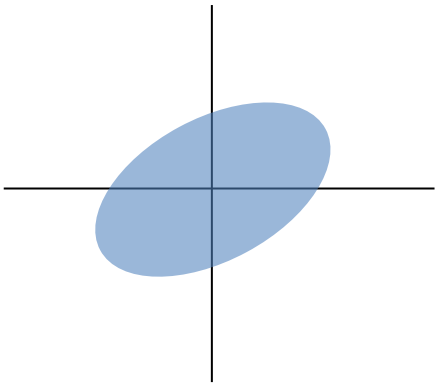


Figure 16.3: A plot of $\hat{\varepsilon}_i$ on $\hat{\varepsilon}_{i-1}$; a case in which correlation exists.

Speaking of correlations, we can compute them:

$$\hat{\rho}_k = \frac{\frac{1}{n}\sum_{i=k+1}^{n} \hat{\varepsilon}_i \hat{\varepsilon}_{i-1}}{\frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2}$$

This is the autocorrelation at lag k. It looks a lot like sample correlation. It can also be plotted using a *autocorrelation function* (acf) plot:

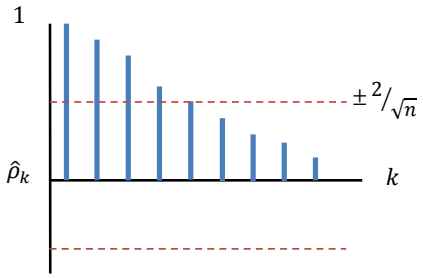


Figure 16.4: An acf plot. $\mathrm{V}(\hat{\rho}_k) = \frac{1}{n}$. Under the null hypothesis, $\rho_k = 0$.

The acf plot tells us when and where autocorrelation exists and how to fix it. This goes beyond the scope of the class, but we might build correlations into our matrices before doing the regression. See the `nmle` package in `R` for some of these iterative algorithms which estimate V, then estimate the parameters, then re-estimate V, then re-estimate the parameters, etc.

16.3 Non-Normality

Ideally, we hope that $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Even $\varepsilon \dot{\sim} \mathcal{N}(0, \sigma^2)$ is all right. Most of the time, CLT fixes mild cases of non-normality. More specifically, three conditions need to be met for CLT to alleviate

non-normality. These three properties are fairly mild requirements, which is why non-normality typically is not a huge deal.

1. Need that eigenvalues of $Z^TZ \to \infty$. Z^TZ is a matrix version of n. Even if all but one eigenvalues go to infinity, CLT does not work. We cannot make any useful estimations using a single observation on one dimension.

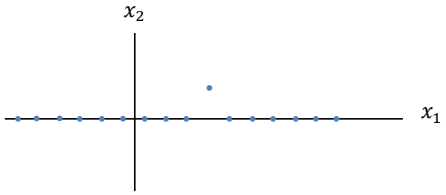


Figure 16.5: Hypothetical data of two predictors where one eigenvalue does not go to infinity. We can't estimate anything based on this single observation on x_2. CLT does not help here.

2. No Z_{ij} can be too large, or else it may have undue influence on the estimation.

3. ε_i cannot be too heavy-tailed.

16.3.1 Detection

How can we detect non-normality? One approach is to form

$$\hat{\varepsilon}_i = Y_i = \hat{Y}_i = Y_i - Z_i^T\hat{\beta}$$

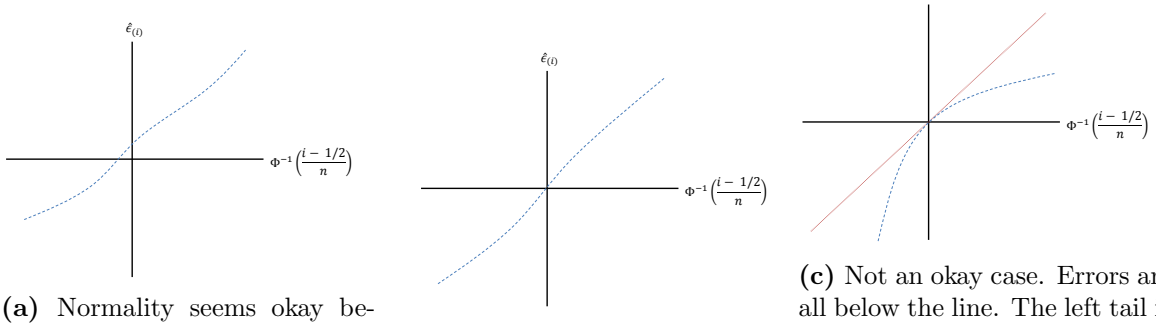
Note that $\hat{\varepsilon}_i \neq \varepsilon_i$. Then, you could consider the distribution of the ratios $\frac{\hat{\varepsilon}_i}{s}$, which is intuitively an estimate of the uncertainty from one data point over an estimate of σ. A more careful computation shows that

$$\mathrm{V}(\hat{\varepsilon}_i) = \sigma^2(1 - H_{ii})$$

where H is the hat matrix. Then, a possibly better ratio to consider is $\frac{\hat{\varepsilon}_i}{s\sqrt{1-H_{ii}}}$, which is corrected for the bias in the above equation. Another possibility is to form $\frac{Y_i - \hat{Y}_{i-1}}{S_{-i}}$, where X_{-i} is the quantity X formed with all data except the ith observation. Intuitively, this is the estimate of the data value Y_i from all the other data, protected from inflation of s.

One very easy approach is QQ-plots: Sort $\hat{\varepsilon}_i$ so that $\hat{\varepsilon}_{(1)} \leq \hat{\varepsilon}_{(2)} \leq ... \leq \hat{\varepsilon}_{(n)}$. Then, plot these residuals versus $\Phi^{-1}\left(\frac{i-1/2}{n}\right)$ (use `qnorm` in `R`). Ideally, the plot should follow a straight line where the slope is σ and the intercept is μ (which should be 0 with residuals).

Non-normality is usually a non-issue, except when it comes to outliers.



(a) Normality seems okay because the line is straight, but we want the line to pass through 0.

(b) An okay case.

(c) Not an okay case. Errors are all below the line. The left tail is heavier than the Gaussian, while the right tail is lighter than the Gaussian.

Figure 16.6: Three QQ-plots.

16.4 Outliers

Say that there is some $\hat{\varepsilon}_i$ that is extremely large, like 12 standard deviations away from the mean. This is almost certainly not from a normal distribution and could really mess up our model's fit. What do we do?

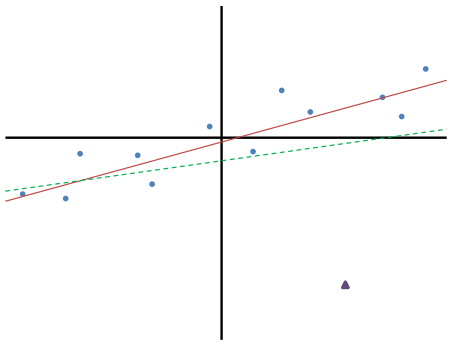


Figure 16.7: Potential effects of an outlier. Without the outlier, we might get the red solid line as the regression line. However, the outlier could pull the whole line down.

If we see an outlier, we could just remove it from the data and explain why it is an outlier. (Perhaps there is known measurement error or definite evidence that a mistake was made). But if we do not know why it is an outlier, we can still take it out of the data and say so, but it starts looking like manipulation.

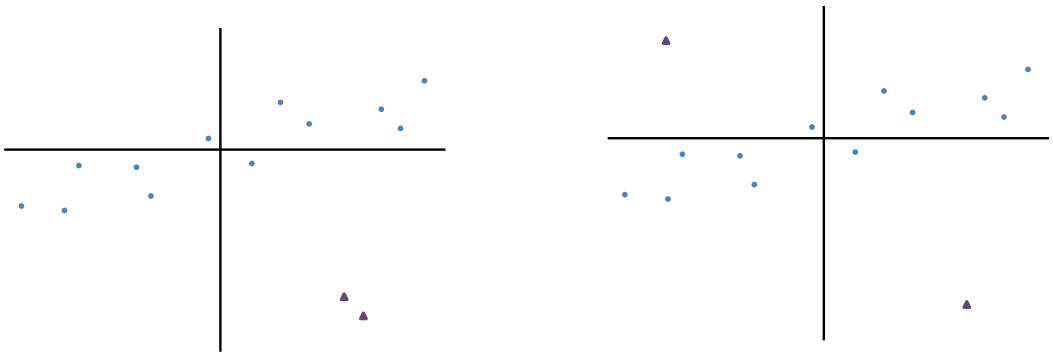
16.4.1 Detection

Large $|\hat{\varepsilon}_i|$ is a clue of an outlier. However, outliers can actually inflate overall standard errors, which means we are less likely to detect outliers since all $|\hat{\varepsilon}_i|$'s are deflated. A better indication may be finding large $\frac{|\hat{\varepsilon}_{-i}|}{s_i}$. This is a *leave-one-out* residual that does not allow an outlier to inflate standard errors (since it is left out of the data used in the model). However, this obviously does

not solve the problem if there is more than one outlier, and it computationally not feasible to do leave-m-out analysis for all m-tuples.

16.4.2 Potential Solutions

Auto-removal algorithms are not sure-fire solutions, either. They could fail in at least two ways. There could be *masking*: pairs of outliers make each other look like non-outliers.



(a) Pairs of outliers that make each other look like non-outliers.

(b) Both outliers pull the regression line down, but conspire so that you would never identify either.

Figure 16.8: Two cases of masking.

There could also be *swamping*: outliers may make good data points look bad.

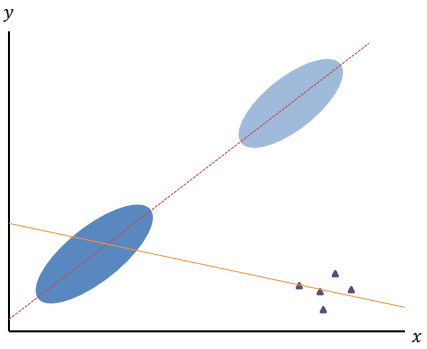


Figure 16.9: A case of swamping. The outliers, if influential enough, could force the regression line down to the orange dotted line when the actual data would suggest the red solid line.

Robust regression methods try to be less sensitive to outliers. One approach is $L1$ regression, where we minimize absolute sums of errors:

$$\min_{\beta} \sum_{i=1}^{n} |Y_i - Z_i^T \beta|$$

Another is looking only at the intercept:

$$\min \sum_{i=1}^{n} |Y_i - \beta_0| \qquad \text{and we get that} \qquad \beta_0 = \text{median}(Y_1, ..., Y_n)$$

But these are not really robust. $L1$ can still fail if outliers are "bad" enough (have really large leverage). There is at least one algorithm that could solve this.

Least Trimmed Means

In least trimmed means regressions, we minimize sum of squared errors for only a portion of the data.

Suppose $|\hat{\varepsilon}|_{(i)}$ are $|Y_i - Z_i^T \beta|$ are sorted from smallest to largest. Then, least trimmed means does the following estimation:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{\lfloor 0.8n \rfloor} |\hat{\varepsilon}|_{(i)}^2$$

That is, we take the smallest 80% of squared residuals and find the β that minimizes the sum of only those. The good news is that the model is robust if less than 20% of the data is outliers (and we'd hope so, or else there are serious issues with the data or they're not outliers at all). The bad news is that it is incredibly hard to compute this, especially more than 15 predictors; this is a non-convex optimization problem.

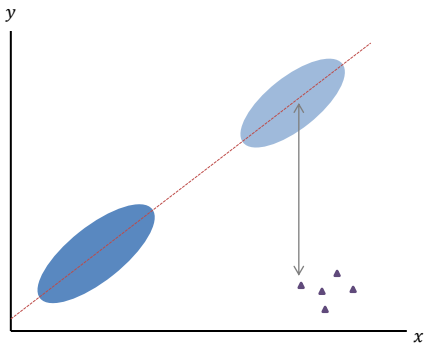


Figure 16.10: Using least mean squares, we get the correct regression line and can correctly identify the gigantic residuals on the outliers.

Chapter 17

Bootstrapped Regressions

Recall that the standard model makes the following assumptions:

$$Y = Z\beta + \varepsilon \qquad \mathrm{E}[\varepsilon] = 0 \qquad \mathrm{E}[Z\varepsilon] = 0$$

As such, residuals have mean zero and are uncorrelated with the predictors. Data is in (X_i, Y_i) pairs, and $Z_i = Z(X_i)$.

But sometimes, the standard model using the data may fall short. The data may be too small or assumptions about the general population may be wrong. In these cases, we may want to make empirical estimates but without imposing distribution assumptions. We use bootstrapping to do this, and now with regressions. There are at least four methods.

17.1 Bootstrapped Pairs

Resample pairs of data, drawing (X^{*b}, Y^{*b}), where pairs are $i = 1, ..., n$ and number of bootstrappings are $b = 1, ..., B$. Then, we estimate the following repeatedly:

$$\hat{\beta}^{*b} = (Z^{*bT} Z^{*b})^{-1} Z^{*bT} Y^{*b}$$

We can then estimate $\mathrm{Cov}(\hat{\beta})$ by using the sample covariance of $\hat{\beta}^*$. It is especially good for t-statistics:

$$\frac{c^T(\hat{\beta}^* - \hat{\beta})}{s^* c^T (Z^{*T} Z^*)^{-1} c} \approx \frac{c^T(\hat{\beta} - \beta)}{s c^T (Z^T Z)^{-1} c}$$

This method has its upsides and downsides.

- Good: Corrects for unequal variance across observations.

- Bad: The method can break if $Z^{*T}Z^*$ is singular. But as long as this is an uncommon occurrence, this ultimately is not a huge deal.

17.2 Bootstrapped Residuals

Instead of looking at the data, we can look at the errors. The linear regression fits

$$Y_i = Z_i^T\hat{\beta} + \hat{\varepsilon}_i$$

In this method, we resample ε_i^{*b} from $\hat{\varepsilon}_i$'s (or from $\frac{\hat{\varepsilon}_i}{\sqrt{1-H_{ii}}}$, where $i = 1, ..., n$ and $b = 1, ..., B$). Once we do that, we take

$$Y_i^{*b} = Z_i^T\hat{\beta} + \varepsilon_i^{*b} \qquad \Rightarrow \qquad \hat{\beta}^{*b} = (Z^TZ)^{-1}Z^TY^{*b}$$

Again, there are upsides and downsides.

- Good: Always uses the same Z^TZ, so we don't get singular Z^TZ's unless the original data was singular.
- Good: X_i's are fixed.
- Bad: Wires in the assumption that ε_i are i.i.d., and especially that they have common/equal variance. This is not good if heteroskedasticity exists.
- Bad: So, does not correct for unequal variance.

17.2.1 Unequal Variances

What do we do if $\varepsilon_i \overset{ind.}{\sim} (0, \sigma_i^2)$? Then,

$$\begin{aligned} \mathrm{V}(\hat{\beta}) = ... &= (Z^TZ)^{-1}Z^T \operatorname{diag}(\sigma_1^2 ... \sigma_n^2)\ Z(Z^TZ)^{-1} \\ &= (Z^TZ)^{-1}\sum_{i=1}^{n}\sigma_i^2 Z_i Z_i^T (Z^TZ)^{-1} \end{aligned}$$

where Z_i is a column vector of observation i (a transpose). Of course, we do not know the real σ_i's. Each bootstrapping technique estimates this differently.

For resampled residuals,

$$(Z^TZ)^{-1}(Z^TZ\hat{\sigma}^2)(Z^TZ)^{-1} = (Z^TZ)^{-1}\hat{\sigma}^2 \qquad \text{where} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

For resampled pairs, as $B \to \infty$,

$$(Z^T Z)^{-1} \sum_{i=1}^{n} \hat{\sigma}_i^2 Z_i Z_i^T (Z^T Z)^{-1} \qquad \text{where} \qquad \hat{\sigma}_i^2 = \hat{\varepsilon}_i^2 = (Y_i - Z_i \hat{\beta})^2$$

Note that $\hat{\sigma}_i^2 Z_i Z_i^T$ is a fourth-order term, which may make it unstable in dealing with heavy-tailed distributions. Some statisticians do not like this estimate for this reason. (In fact, statisticians have not all agreed on what techniques are best in bootstrapping.)

Using $\hat{\varepsilon}_i$ to estimate σ_i is a bold and fairly unsound move at the individual observation level. However, summing over n observations makes for a decent estimate. This is problematic if some σ_i^2 are tiny are gigantic, but that's always a problem. Any problem that breaks CLT will break this too. But if CLT is broken, then we have bigger issues.

These resampling approaches that deal with different variances were/are called *Huber-White estimators* of $\mathrm{V}(\hat{\beta})$, which we still see in many computational statistical packages.

17.3 Wild Bootstrap

Consider the following model:

$$Y_i^{*b} = Z_i^T \hat{\beta} + \varepsilon_i^{*b}$$
$$\varepsilon_i^{*b} \text{ independent} \qquad \mathrm{E}[\varepsilon_i^{*b}] = 0 \qquad \mathrm{V}(\varepsilon_i^{*b}) = \hat{\varepsilon}_i^2$$

So we have a model where residuals are $\pm(Y_i - Z_i^T \hat{\beta})$ with probability 0.5 each. This gives us a model that has fixed Z_i's and allows for unequal variances—two characteristics we like.

If we wanted, we could also say something about skewness: $\mathrm{E}[\hat{\varepsilon}_i^{*b3}] = \hat{\varepsilon}_i^3$. Using a two-point distribution can allow us to match a desired mean, variance, and skewness.

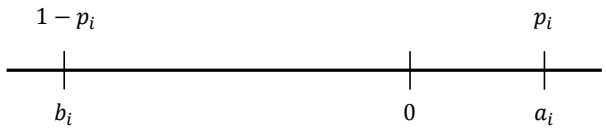


Figure 17.1: A two-point distribution with probability p_i of the value a_i and probability $1 - p_i$ of the value b_i for each observation i.

However, the method is not good at dealing with lack of fit (that is, $\mathrm{E}[Y] \neq Z\beta$).

17.4 Weighted Likelihood Bootstrap

The typical MLE $\hat{\beta}$ puts equal weights $\frac{1}{n}$ on observations $i = 1, ..., n$. Bootstrapped pairs puts random multinomial weights on observations $i = \frac{\#\text{boot}=i}{n}$. However, we do not have to limit ourselves to multinomially distributed weights. Instead, say

$$N_i^* \overset{\text{ind.}}{\sim} \exp(1) \qquad W_i^* = \frac{N_i}{\sum_{i=1}^n N_k}$$

Then,

$$\hat{\beta}^* = \left(\sum_{i=1}^n W_i Z_i Z_i^T \right)^{-1} \left(\sum_{i=1}^n W_i Z_i Y_i \right)$$

We thus reweight using exponentially distributed random variables. This helps to avoid singularities and deals with unequal variances.

Chapter 18

Course Summary

The class can be thought of as a sequence of steps.

18.1 The Steps

1. We start with data X and Y where $Y \in \mathbb{R}$. We want to predict Y from X, and looked at how to do so for (among others):
   - 1 group
   - 2 groups
   - $k > 2$ groups
   - $k \times r$ groups
   - $\mathbb{R}$
   - $\mathbb{R}^p$
2. We identify features Z, such as $\begin{bmatrix} 1 & X_i & ... & X_p \end{bmatrix}$.
   - We can add in non-linear features like X_3^2 or $1\{X_2 \leq 7\}$, etc.
   - The inclusion/exclusion of features can be based on intuition, experience, science, and even machine learning techniques.
3. We apply a model. The basic one is $Y \sim \mathcal{N}(Z\beta, \sigma^2 I)$. This is an incredibly powerful basic model.
4. We estimate $\hat{\beta}$ and $\hat{\sigma}$.
5. With those estimates, we can derive confidence intervals, p-values, hypothesis tests, and power calculations. Finding $\hat{\beta}$ and $\hat{\sigma}$ involves linear algebra and some computation. The diagnostics primarily depend on distribution theory.
6. We interpret the regressions. Involved in this is concerns about causality. Also, recall that a true β_j depends on whether Z_k is in the model. As such, "true" parameter values are actually moving targets.

7. We choose a model using measures like AIC, BIC, C_p, GCV, lasso, ridge regressions, etc. These are optimized for prediction, but not necessarily substantive understanding of causal mechanisms. (Which "truth" are you getting a confidence interval or p-value for?)
8. We consider problems and fixes to the model.
    - Non-normality: CLT fixes most cases.
    - Non-constant variance: Bootstrapping or Huber-White variances.
    - Bias: No good solutions. Context and expertise on topic are useful. Lack of fit sum of squares and pure error measures may help identify bias but can do little to solve it.
    - Outliers: Some methods exist, but these are tough to overcome in high dimensions.
    - Correlations within data: Time series methods.