Lecture 12

Minimum Mean Square Error Linear Prediction

Predicting a r.v. $X$ by a constant $a$: best $a$?

$$\text{mmse value of } a : \min_a E(X - a)^2$$

$$= \min_a \{E(X^2) - 2aEX + a^2\}$$

occurs at $a = EX$, minimum value $E(X^2) - (EX)^2 = \text{Var}(X)$

Predicting the deviation $Y - EY$ of a r.v. $Y$ from its mean as a multiple $b$ times $X - EX$, where $X$ is a secondly jointly distributed r.v: best $b$?

$$\text{mmse value of } b : \min_b E((Y - EY) - b(X - EX))^2$$

$$= \min_b \{E(Y - EY)^2 - 2bE(Y - EY)(X - EX) + b^2E(X - EX)\}$$

$$= \min_b \{\text{Var}(Y) - 2b\text{Cov}(X, Y) + b^2\text{Var}(X)\}$$

occurs at $b = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$, minimum value $\text{Var}(Y)[1 - \rho^2]$.

where $\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{SD}(X)\text{SD}(Y)}$

Regression and the Inheritance of Quantitative Traits

- Mendel gave us the story we now accept about the inheritance of qualitative traits. We’ll revisit it. 1865.

- Galton discovered the phenomenon of regression. We’ll revisit it. 1877.

- Pearson attempted to reconcile the two and failed, concluding that quantitative traits did not get inherited according to Mendel’s laws. 1900.

- Yule pointed out his error. 1906.

- Fisher followed up on Yule and developed a full genetical theory for quantitative traits consistent with Mendel. 1912, 1918.
We begin with an important idea from **population genetics** due independently to Hardy and Weinberg, 1906. Suppose that \(A\) and \(a\) are two allelic forms of a gene inherited according to Mendel’s laws, and suppose that individuals mate “at random” with respect to any phenotypic effects of this gene.

Suppose that the frequency of \(AA\), \(Aa\) and \(aa\) individuals in the population is \(P\), \(2Q\) and \(R\) \((P + 2Q + R = 1)\) in the parental generation. We seek the **joint distribution** of the genotypes between parent and offspring. That is, I want the full table of probabilities of the form:

\[
\text{pr(father is } Aa, \text{ son is } aa)\\
\]

Here there is no sex effect but I’ll take parent = father, offspring = son for definiteness, and to relate to lab 3.

\[
\text{pr(father is } Aa, \text{ son is } aa) = \text{pr(father is } Aa)\text{pr(son is } aa|\text{father is } Aa)
\]

Now, \(\text{pr(father is } Aa) = 2Q\), but for \(\text{pr(son is } aa|\text{father is } Aa)\) we need to consider the mother.

\[
\text{pr(son is } aa|\text{father is } Aa) = \text{pr(son is } aa, \text{ mother is } AA|\text{father is } Aa) \\
+ \text{pr(son is } aa, \text{ mother is } Aa|\text{father is } Aa)
\]

\[
+ \text{pr(son is } aa, \text{ mother is } aa|\text{father is } Aa)
\]

\[
= \text{pr(m is } AA|\text{f is } Aa)\text{pr(s is } aa| m is } AA, f \text{ is } Aa) \\
+ \text{pr(m is } Aa|f \text{ is } Aa)\text{pr(s is } aa| m is } Aa, f \text{ is } Aa)
\]

\[
+ \text{pr(m is } aa|f \text{ is } Aa)\text{pr(s is } aa| m is } aa, f \text{ is } Aa)
\]

\[
= P \times 0 + 2Q \times \frac{1}{2} + R \times \frac{1}{2} \text{ by random mating and Mendel}
\]

<table>
<thead>
<tr>
<th>Parental genotype</th>
<th>Offspring genotype</th>
<th>Marginal</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>(P(P + Q))</td>
<td>(P)</td>
</tr>
<tr>
<td>Aa</td>
<td>(P(Q + R))</td>
<td>(2Q)</td>
</tr>
<tr>
<td>aa</td>
<td>0</td>
<td>(R)</td>
</tr>
<tr>
<td>Marginal</td>
<td>((P + Q)^2)</td>
<td>((Q + R)^2)</td>
</tr>
</tbody>
</table>

**Joint distribution table**

The other entries are built up similarly. Check the row marginal probabilities: as assumed. Check the column marginal probabilities: o.k. for \(AA, aa\); for \(Aa\): either use summing to 1 or:

\[
2(P + Q)(Q + R) = 2PQ + 2Q^2 + 2PR + 2QR
\]

\[
= PQ + Q(P + 2Q + R) + 2PR + QR
\]

\[
= P(Q + R) + Q + R(P + Q)
\]

2
If we let \( P + Q = p \) — the \( A \)-allele gene frequency we have derived the Hardy-Weinberg equilibrium genotype frequencies in the offspring generation:

\[
\begin{align*}
AA & \quad p^2 \\
Aa & \quad 2p(1-p) = 2pq \text{ where } q = 1 - p \\
aa & \quad q^2
\end{align*}
\]

If we now suppose that the population was already in Hardy Weinberg equilibrium via random mating, the joint distribution of parental-offspring genotypes is:

<table>
<thead>
<tr>
<th>Parental genotype</th>
<th>Offspring genotype</th>
<th>( AA )</th>
<th>( Aa )</th>
<th>( aa )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( AA )</td>
<td></td>
<td>( p^2 )</td>
<td>( p^2q )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( Aa )</td>
<td></td>
<td>( q^2 )</td>
<td>( pq )</td>
<td>( pq^2 )</td>
</tr>
<tr>
<td>( aa )</td>
<td></td>
<td>( 0 )</td>
<td>( pq^2 )</td>
<td>( q^3 )</td>
</tr>
</tbody>
</table>

Following Yule we’ll put \( p = q = 1/2 \).

<table>
<thead>
<tr>
<th>Parental genotype</th>
<th>Offspring genotype</th>
<th>( AA )</th>
<th>( Aa )</th>
<th>( aa )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( AA )</td>
<td></td>
<td>( 1/8 )</td>
<td>( 1/8 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( Aa )</td>
<td></td>
<td>( 1/8 )</td>
<td>( 1/4 )</td>
<td>( 1/8 )</td>
</tr>
<tr>
<td>( aa )</td>
<td></td>
<td>( 0 )</td>
<td>( 1/8 )</td>
<td>( 1/8 )</td>
</tr>
</tbody>
</table>

Now we consider a quantitative trait governed solely by this gene (a bit unrealistic, but we’ll proceed, with Yule). The population mean value of this trait will be something, but we are interested in fluctuations about this mean generated by Mendelian inheritance.

Suppose that individuals with genotypes \( AA \), \( Aa \) and \( aa \) are \( \alpha \), \( \beta \) and \( \gamma \) units from the mean. Then \( \alpha + 2\beta + \gamma = 0 \) (since genotype frequencies are 1:2:1). The population variance of the trait for both parent and offspring is:

\[
\text{Variance} = \frac{1}{4} \alpha^2 + \frac{1}{2} \beta^2 + \frac{1}{4} \gamma^2
\]

and the parent-offspring covariance is:

\[
\text{Cov(Par,Off)} = \frac{1}{8} \alpha^2 + \frac{1}{8} \alpha \beta + \frac{1}{8} \beta \alpha + \frac{1}{4} \beta^2 + \frac{1}{8} \beta \gamma + \frac{1}{8} \gamma \beta + \frac{1}{8} \gamma^2
\]

\[
= \frac{1}{8} (\alpha^2 + 2\alpha \beta + 2\beta^2 + 2\beta \gamma + \gamma^2)
\]

Thus the population linear regression coefficient \( b \) of offspring on parent trait is

\[
b = \frac{\text{Cov(Par,Off)}}{\text{SD(Par) SD(Off)}} = \frac{\frac{1}{8} (\alpha^2 + 2\alpha \beta + 2\beta^2 + 2\beta \gamma + \gamma^2)}{\frac{1}{4} (\alpha^2 + 2\beta^2 + \gamma^2)}
\]
\[
i.e. \quad b = \frac{1}{2} + \frac{\beta(\alpha + \gamma)}{\alpha^2 + 2\beta^2 + \gamma^2}
= \frac{1}{2} \quad \text{if } \beta = 0
\]

If there was dominance, say \(\alpha = \beta\), we’d find \(\gamma = -3\alpha\) and
\[
\frac{\text{Cov(Par, Off)}}{\text{SD(Par)}\text{SD(Off)}} = \frac{1}{3}
\]

Interestingly, Pearson thought (following Mendel) there should be dominance, got this 1/3,
whereas 1/2 is observed, and thought Mendel must fail for QTs. Yule pointed out that
additivity restores Mendel.