Statistical Foundations (2)

Jinliang Yang
Sept. 14, 2018

Announcements

First Exam

- Next Monday, Sept. 17th, 2018
- **7:30 am**

Announcements

First Exam

- Next Monday, Sept. 17th, 2018
- **7:30 am**

Stat Notes

- Updated Notes
- Source Code on Overleaf for the notes

Announcements

First Exam

- Next Monday, Sept. 17th, 2018
- **7:30 am**

Stat Notes

- Updated Notes
- Source Code on Overleaf for the notes

About HTML slides

- Using `R` packge Xaringan through R Markdown.

Expectation and Variance

Expectation $E(X)$:

$$E(f(X)) = \sum_{i=1}^{k} f(x_i) Pr(X = x_i)$$

$$E[X] = 0 \times (1-p)^2 + 1 \times [2p(1-p)] + 2 \times p^2 = 2p$$

Expectation and Variance

Expectation $E(X)$:

$$E(f(X)) = \sum_{i=1}^{k} f(x_i) Pr(X = x_i)$$

$$E[X] = 0 \times (1-p)^2 + 1 \times [2p(1-p)] + 2 \times p^2 = 2p$$

Variance $Var(X)$:

$$\begin{aligned} Var(X) &= E[X^2] - E[X]^2 \\ &= 2p(1-p) \end{aligned}$$

Probabilities

Joint Probability

Two random variables to **occur together**. In the `Milk Yield` exmaple, the joint probability of $Pr(G = aa, MY > 300)$?

Probabilities

Joint Probability

Two random variables to **occur together**. In the `Milk Yield` exmaple, the joint probability of $Pr(G = aa, MY > 300)$?

Marginal Probability

A sum of **mutually exclusive** and **exhaustive** set of events. The marginal probability of $Pr(G = Aa)$ for all possible MY?

Probabilities

Joint Probability

Two random variables to **occur together**. In the `Milk Yield` exmaple, the joint probability of $Pr(G = aa, MY > 300)$?

Marginal Probability

A sum of **mutually exclusive** and **exhaustive** set of events. The marginal probability of $Pr(G = Aa)$ for all possible MY?

Conditional Probability

$$Pr(X = x|Y = y) = \frac{Pr(X = x, Y = y)}{Pr(Y = y)}$$

What is the conditional probability of $Pr(MY \leq 100|G = Aa)$?

Genotype (X) and Milk Yield (MY)

Genotype (G)	$MY \leq 100$	$100 < MY \leq 300$	$MY > 300$	**Marginal** $Pr(G)$
aa	0.10	0.04	0.02	0.16
Aa	0.14	0.18	0.16	0.48
AA	0.06	0.10	0.20	0.36
Marg. Prob.	0.30	0.32	0.38	1.00

Genotype (X) and Milk Yield (MY)

Genotype (G)	$MY \leq 100$	$100 < MY \leq 300$	$MY > 300$	**Marginal** $Pr(G)$
aa	0.10	0.04	0.02	0.16
Aa	0.14	0.18	0.16	0.48
AA	0.06	0.10	0.20	0.36
Marg. Prob.	0.30	0.32	0.38	1.00

Statistical Independence

$$
\begin{aligned}
&Pr(X = x_i | Y = y_j) \\
&= Pr(X = x_i | Y = y_k) \\
&= Pr(X = x_i)
\end{aligned}
$$

Genotype (X) and Milk Yield (MY)

Genotype (G)	$MY \leq 100$	$100 < MY \leq 300$	$MY > 300$	**Marginal** $Pr(G)$
aa	0.10	0.04	0.02	0.16
Aa	0.14	0.18	0.16	0.48
AA	0.06	0.10	0.20	0.36
Marg. Prob.	0.30	0.32	0.38	1.00

Statistical Independence

$$\begin{aligned} & Pr(X = x_i | Y = y_j) \\ & = Pr(X = x_i | Y = y_k) \\ & = Pr(X = x_i) \end{aligned}$$

$$Pr(X = x_i, Y = y_j) = Pr(X = x_i) \times Pr(Y = y_j)$$

Genotype (X) and Milk Yield (MY)

Genotype (G)	$MY \leq 100$	$100 < MY \leq 300$	$MY > 300$	**Marginal** $Pr(G)$
aa	0.10	0.04	0.02	0.16
Aa	0.14	0.18	0.16	0.48
AA	0.06	0.10	0.20	0.36
Marg. Prob.	0.30	0.32	0.38	1.00

Statistical Independence

$$
\begin{aligned}
&Pr(X = x_i | Y = y_j) \\
&= Pr(X = x_i | Y = y_k) \\
&= Pr(X = x_i)
\end{aligned}
$$

$$Pr(X = x_i, Y = y_j) = Pr(X = x_i) \times Pr(Y = y_j)$$

$$
\begin{aligned}
&Pr(MY > 300, X_{AA}) \\
&Pr(MY > 300) \times Pr(X_{AA})
\end{aligned}
$$

Genotype (X) and Milk Yield (MY)

Genotype	$MY = 100$	$MY = 150$	$MY = 300$	**Marginal** $Pr(G)$
aa	0.10	0.04	0.02	0.16
Aa	0.14	0.18	0.16	0.48
AA	0.06	0.10	0.20	0.36
Marg. Prob.	0.30	0.32	0.38	1.00

What are the genotype effects, or $E(MY|X_{AA)}$, $E(MY|X_{aa)}$, $E(MY|X_{Aa)}$?

Genotype (X) and Milk Yield (MY)

Genotype	$MY = 100$	$MY = 150$	$MY = 300$	**Marginal** $Pr(G)$
aa	0.10	0.04	0.02	0.16
Aa	0.14	0.18	0.16	0.48
AA	0.06	0.10	0.20	0.36
Marg. Prob.	0.30	0.32	0.38	1.00

What are the genotype effects, or $E(MY|X_{AA)}$, $E(MY|X_{aa)}$, $E(MY|X_{Aa)}$?

Conditional Expectation

The expectation (=mean) for variable X conditional on variable $Y = y$ is:

$$\begin{aligned} E(X|Y = y) &= \sum_{i=1}^{k} x_i Pr(X = x_i|Y = y) \\ &= \sum_{i=1}^{k} x_i \frac{Pr(X = x_i, Y = y)}{Pr(Y = y)} \end{aligned}$$

Genotype (X) and Milk Yield (MY)

Genotype	$MY = 100$	$MY = 150$	$MY = 300$	**Marginal** $Pr(G)$
aa	0.10	0.04	0.02	0.16
Aa	0.14	0.18	0.16	0.48
AA	0.06	0.10	0.20	0.36
Marg. Prob.	0.30	0.32	0.38	1.00

What are the genotype effects, or $E(MY|X_{AA)}$, $E(MY|X_{aa)}$, $E(MY|X_{Aa)}$?

Genotype (X) and Milk Yield (MY)

Genotype	$MY = 100$	$MY = 150$	$MY = 300$	**Marginal** $Pr(G)$
aa	0.10	0.04	0.02	0.16
Aa	0.14	0.18	0.16	0.48
AA	0.06	0.10	0.20	0.36
Marg. Prob.	0.30	0.32	0.38	1.00

What are the genotype effects, or $E(MY|X_{AA})$, $E(MY|X_{aa})$, $E(MY|X_{Aa})$?

$$E(MY|X_{AA}) = \sum_{i=1}^{3} MY_i Pr(MY = MY_i|X = X_{AA})$$
$$= 100 \times 0.06/0.36 + 150 \times 0.10/0.36 + 300 \times 0.20/0.36 = 81/0.36 = 225$$

Genotype (X) and Milk Yield (MY)

Genotype	$MY = 100$	$MY = 150$	$MY = 300$	**Marginal** $Pr(G)$
aa	0.10	0.04	0.02	0.16
Aa	0.14	0.18	0.16	0.48
AA	0.06	0.10	0.20	0.36
Marg. Prob.	0.30	0.32	0.38	1.00

What are the genotype effects, or $E(MY|X_{AA})$, $E(MY|X_{aa})$, $E(MY|X_{Aa})$?

$$E(MY|X_{AA}) = \sum_{i=1}^{3} MY_i Pr(MY = MY_i|X = X_{AA})$$
$$= 100 \times 0.06/0.36 + 150 \times 0.10/0.36 + 300 \times 0.20/0.36 = 81/0.36 = 225$$

$$E(MY|X_{aa}) = \sum_{i=1}^{3} MY_i Pr(MY = MY_i|X = X_{aa})$$
$$= 100 \times 0.10/0.16 + 150 \times 0.04/0.16 + 300 \times 0.02/0.16 = 22/0.16 = 137.5$$

$$E(MY|X_{Aa}) = \sum_{i=1}^{3} MY_i Pr(MY = MY_i|X = X_{Aa})$$
$$= 100 \times 0.14/0.48 + 150 \times 0.18/0.48 + 300 \times 0.16/0.48 = 202/0.48 = 420.8$$

Covariance

To quantify to what extent the two variables **co-vary**.

$$\begin{aligned} Cov(X,Y) &= E([X - E(X)][Y - E(Y)]) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

where,

$$E(XY) = \sum_i \sum_j x_i y_j Pr(X = x_i, Y = y_j)$$

A plant example

The number of florets per spikelet:

$$Y_i = \sum_{j=1}^{j=m} X_{ij}\alpha_j + \epsilon_i = G_i + \epsilon_i$$

Genotype	P	G	E	Prob
t/t	3	2.8	0.2	0.20
t/t	2	2.8	-0.8	0.05
T/t	3	2.6	0.4	0.30
T/t	2	2.6	-0.6	0.20
T/T	3	2.2	0.8	0.05
T/T	2	2.2	-0.2	0.20

A plant example

The number of florets per spikelet:

$$Y_i = \sum_{j=1}^{j=m} X_{ij}\alpha_j + \epsilon_i = G_i + \epsilon_i$$

Genotype	P	G	E	Prob
t/t	3	2.8	0.2	0.20
t/t	2	2.8	-0.8	0.05
T/t	3	2.6	0.4	0.30
T/t	2	2.6	-0.6	0.20
T/T	3	2.2	0.8	0.05
T/T	2	2.2	-0.2	0.20

What is the covariance between G and P?

$$Cov(G, P) = E(GP) - E(G)E(P)$$

What is the covariance between G and P?

$$Cov(G, P) = E(GP) - E(G)E(P)$$

```
dt <- data.frame(Genotype=c("t/t", "t/t", "T/t", "T/t", "T/T", "T/T"),
                 P=c(3, 2, 3, 2, 3, 2),
                 G=c(2.8, 2.8, 2.6, 2.6, 2.2, 2.2),
                 E=c(0.2, -0.8, 0.4, -0.6, 0.8, -0.2),
                 Prob=c(0.20, 0.05, 0.30, 0.20, 0.05, 0.20))
kable(dt) # print the table
```

Genotype	P	G	E	Prob
t/t	3	2.8	0.2	0.20
t/t	2	2.8	-0.8	0.05
T/t	3	2.6	0.4	0.30
T/t	2	2.6	-0.6	0.20
T/T	3	2.2	0.8	0.05
T/T	2	2.2	-0.2	0.20

What is the covariance between G and P?

$$Cov(G, P) = E(GP) - E(G)E(P)$$

```
dt <- data.frame(Genotype=c("t/t", "t/t", "T/t", "T/t", "T/T", "T/T"),
                 P=c(3, 2, 3, 2, 3, 2),
                 G=c(2.8, 2.8, 2.6, 2.6, 2.2, 2.2),
                 E=c(0.2, -0.8, 0.4, -0.6, 0.8, -0.2),
                 Prob=c(0.20, 0.05, 0.30, 0.20, 0.05, 0.20))
kable(dt[1,]) # only print the first line to save space
```

Genotype	P	G	E	Prob
t/t	3	2.8	0.2	0.2

What is the covariance between G and P?

$$Cov(G, P) = E(GP) - E(G)E(P)$$

```
dt <- data.frame(Genotype=c("t/t", "t/t", "T/t", "T/t", "T/T", "T/T"),
                 P=c(3, 2, 3, 2, 3, 2),
                 G=c(2.8, 2.8, 2.6, 2.6, 2.2, 2.2),
                 E=c(0.2, -0.8, 0.4, -0.6, 0.8, -0.2),
                 Prob=c(0.20, 0.05, 0.30, 0.20, 0.05, 0.20))
kable(dt[1,]) # only print the first line to save space
```

Genotype	P	G	E	Prob
t/t	3	2.8	0.2	0.2

```
sum(with(dt, G*P*Prob)) # E(GP)
```

```
## [1] 6.55
```

```
sum(with(dt, G*Prob)) # E(G)
```

```
## [1] 2.55
```

```
sum(with(dt,  P*Prob)) # E(P)
```

```
## [1] 2.55
```

What is the covariance between G and P?

$$\begin{aligned} Cov(G, P) &= E(GP) - E(G)E(P) \\ &= 6.55 - (2.55)^2 = 0.0475 \end{aligned}$$

What is the covariance between G and P?

$$\begin{aligned} Cov(G, P) &= E(GP) - E(G)E(P) \\ &= 6.55 - (2.55)^2 = 0.0475 \end{aligned}$$

Similarly, to calculate $Cov(G, E) = E(GE) - E(G)E(E)$:

```
sum(dt$G * dt$E * dt$Prob) # E(GE)
```

```
## [1] -5.551115e-17
```

```
sum(dt$E * dt$Prob) # E(E)
```

```
## [1] 0
```

$$\begin{aligned} Cov(G, E) &= E(GE) - E(G)E(E) \\ &= 0 - (2.55) \times 0 = 0 \end{aligned}$$

Correlation

A mutual relationship between two variables.

$$\begin{aligned} r_{XY} &= Corr(X, Y) \\ &= \frac{Cov(X, Y)}{\sqrt{\overline{Var(X)Var(Y)}}} \end{aligned}$$

Correlation

A mutual relationship between two variables.

$$\begin{aligned} r_{XY} &= Corr(X, Y) \\ &= \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} \end{aligned}$$

The correlation coefficent between G and P

$$r_{GP} = \frac{Cov(G, P)}{\sqrt{Var(G)Var(P)}}$$

$$r_{GP} = \frac{Cov(G, P)}{\sqrt{Var(G)Var(P)}}$$

$Cov(G, P) = 0.0475$

$$r_{GP} = \frac{Cov(G, P)}{\sqrt{Var(G)Var(P)}}$$

$Cov(G, P) = 0.0475$

$Var(G) = E(G^2) - E(G)^2$

```
sum(dt$G^2 * dt$Prob) - sum(dt$G * dt$Prob)^2
```

```
## [1] 0.0475
```

$$r_{GP} = \frac{Cov(G,P)}{\sqrt{Var(G)Var(P)}}$$

$Cov(G,P) = 0.0475$

$Var(G) = E(G^2) - E(G)^2$

```
sum(dt$G^2 * dt$Prob) - sum(dt$G * dt$Prob)^2
```

```
## [1] 0.0475
```

$Var(P) = E(P^2) - E(P)^2$

```
sum(dt$P^2 * dt$Prob) - sum(dt$P * dt$Prob)^2
```

```
## [1] 0.2475
```

$$r_{GP} = \frac{Cov(G, P)}{\sqrt{Var(G)Var(P)}}$$

$Cov(G, P) = 0.0475$

$Var(G) = E(G^2) - E(G)^2$

```
sum(dt$G^2 * dt$Prob) - sum(dt$G * dt$Prob)^2
```

```
## [1] 0.0475
```

$Var(P) = E(P^2) - E(P)^2$

```
sum(dt$P^2 * dt$Prob) - sum(dt$P * dt$Prob)^2
```

```
## [1] 0.2475
```

$$\begin{aligned} r_{GP} &= \frac{Cov(G, P)}{\sqrt{Var(G)Var(P)}} \\ &= \frac{0.0475}{\sqrt{0.0475 \times 0.2475}} \\ &= 0.438 \end{aligned}$$

Regression

The regression of Y on X:

$$\hat{y} = E(Y|X)$$

This is also called the **best predictor** of Y given X.

Regression

The regression of Y on X:

$$\hat{y} = E(Y|X)$$

This is also called the **best predictor** of Y given X.

Regression can be used to define **a linear model**:

$$y = \hat{y} + e$$

where e is called the residual.

Regression

The regression of Y on X:

$$\hat{y} = E(Y|X)$$

This is also called the **best predictor** of Y given X.

Regression can be used to define **a linear model**:

$$y = \hat{y} + e$$

where e is called the residual.

Another definition of the simple linear regression model:

$$\begin{aligned} y &= \bar{y} + \beta_{YX}(x - \bar{x}) + e \\ \text{with } \bar{y} &= E(Y) \\ \beta_{YX} &= \frac{Cov(Y, X)}{Var(X)} \end{aligned}$$

Predict G based on P

$$G = \bar{G} + \beta_{GP}(P - \bar{P}) + e$$
$$\beta_{GP} = \frac{Cov(G, P)}{Var(P)}$$

Predict G based on P

$$G = \bar{G} + \beta_{GP}(P - \bar{P}) + e$$
$$\beta_{GP} = \frac{Cov(G, P)}{Var(P)}$$

$$\begin{aligned}\beta_{GP} &= \frac{Cov(G, P)}{Var(P)} \\ &= 0.0475/0.2475 = 0.192 \\ \bar{P} &= E(P) = 2.55 \\ \bar{G} &= E(G) = 2.55\end{aligned}$$

Predict G based on P

$$G = \bar{G} + \beta_{GP}(P - \bar{P}) + e$$
$$\beta_{GP} = \frac{Cov(G, P)}{Var(P)}$$

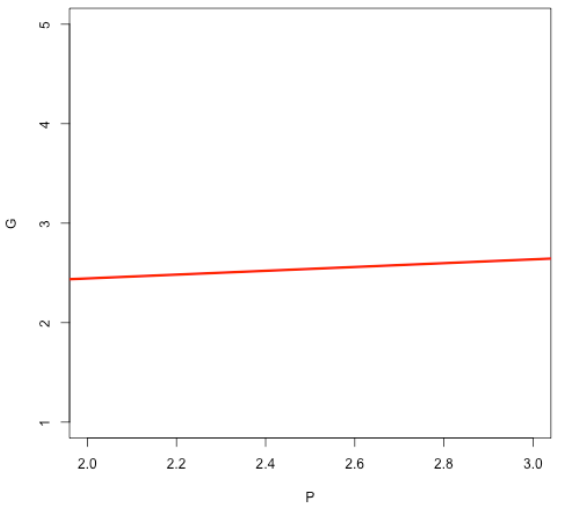
$$\begin{aligned}\beta_{GP} &= \frac{Cov(G, P)}{Var(P)} \\ &= 0.0475/0.2475 = 0.192 \\ \bar{P} &= E(P) = 2.55 \\ \bar{G} &= E(G) = 2.55\end{aligned}$$

$$\begin{aligned}G &= \bar{G} + \beta_{GP}(P - \bar{P}) + e \\ \hat{G} &= \bar{G} + \beta_{GP}(P - \bar{P}) \\ &= 2.55 + 0.192(P - 2.55)\end{aligned}$$

Prediction Model

$$
\begin{aligned}
G &= 2.55 + 0.192(P - 2.55) \\
&= 2.0604 + 0.192 \times P
\end{aligned}
$$

```
plot(x=1, y=1, ylim=c(1, 5), xlim=c(2, 3), type="n", xlab="P", ylab="G")
# a, b : single values specifying the intercept and the slope of the line
abline(a=2.0604, b=0.192, lwd=3, col="red")
```



Get predicted G ($\hat{G}$)

Using the prediction model:

$$G = 2.0604 + 0.192 \times P$$

```
dt$ghat <- 2.0604 + 0.192*dt$P
kable(dt)
```

Genotype	P	G	E	Prob	ghat
t/t	3	2.8	0.2	0.20	2.6364
t/t	2	2.8	-0.8	0.05	2.4444
T/t	3	2.6	0.4	0.30	2.6364
T/t	2	2.6	-0.6	0.20	2.4444
T/T	3	2.2	0.8	0.05	2.6364
T/T	2	2.2	-0.2	0.20	2.4444

Accuracy of prediction

The accuracy of the prediction is equal to the **correlation of $\hat{y}$ with its true value y**.

We can derive accuracy as:

$$\begin{aligned} r_{\hat{y}y} &= \frac{Cov(\hat{y}, y)}{\sqrt{Var(\hat{y})Var(y)}} \\ &= \frac{Cov(x, y)}{\sqrt{Var(x)Var(y)}} \\ &= r_{XY} \end{aligned}$$

Accuracy of prediction

Method One:

$$r_{\hat{G}G} = \frac{Cov(\hat{G}, G)}{\sqrt{Var(\hat{G})Var(G)}}$$

```
vg <- sum(dt$G^2 * dt$Prob) - sum(dt$G * dt$Prob)^2
vghat <- sum(dt$ghat^2 * dt$Prob) - sum(dt$ghat * dt$Prob)^2
cov_g_ghat <- sum(dt$ghat * dt$G * dt$Prob) - sum(dt$G * dt$Prob) * sum(dt$ghat
r_ghat_g <- cov_g_ghat / sqrt(vg*vghat)
r_ghat_g
```

```
## [1] 0.4380858
```

Accuracy of prediction

Method Two:

$$r_{GP} = \frac{Cov(G, P)}{\sqrt{Var(G)Var(P)}}$$

```
vp <- sum(dt$P^2 * dt$Prob) - sum(dt$P * dt$Prob)^2
cov_g_p <- sum(dt$P * dt$G * dt$Prob) - sum(dt$G * dt$Prob) * sum(dt$P * dt$Prob
r_g_p <- cov_g_p / sqrt(vg*vp)
r_g_p
```

```
## [1] 0.4380858
```

Decomposition of variance

$$Y = G + E$$

Variance of Y can be decomposed as **explained** and **unexplained** variance components:

$$Var(Y) = r^2_{XY} Var(Y) + (1 - r^2_{XY}) Var(Y)$$

```
vg
```

```
## [1] 0.0475
```

```
r_g_p^2 * vg
```

```
## [1] 0.009116162
```

```
(1- r_g_p^2) *vg
```

```
## [1] 0.03838384
```