Introduction to Linear Models: One Mean I

A model for one random sample from one population

response (random)		deterministic model for the mean (not random)		error (random)
$\downarrow$		$\downarrow$		$\downarrow$
Y	$=$	μ	$+$	ϵ

Random sample $\Longrightarrow$ (nearly) independent outcomes

data		population mean		random error		sample size
$\downarrow$		$\downarrow$		$\downarrow$		$\downarrow$
Y_i	$=$	μ	$+$	ϵ_i	for $i = 1, \ldots,$	n
				$\uparrow$		
				errors independent since sample random		

Introduction to Linear Models: One Mean II

ϵ_i = "errors" = difference between data and model: $\epsilon_i = Y_i - \mu$.

This simple model (the constant μ) is a **linear model** for the mean:

$\mu_i = \beta_0 + \beta_1 x_i$ for constants β_0 and β_1 and n fixed values $x_1, x_2, \ldots, x_n$.

Here, $\beta_0 = \mu$ and $\beta_1 = 0$, and no other variable x is considered.

Introduction to Linear Models: One Mean III

Estimate the model (mean) using the Least Squares Principle

What value $\widehat{\mu}$ will **minimize sum of squared errors**?

$$\text{Sum of Squared Errors} = \text{SSE} = \sum_{i=1}^{n}(Y_i - \widehat{\mu})^2 .$$

$\widehat{\mu}$ = estimate for μ

We already know that $\widehat{\mu} = \overline{Y}$ minimizes the SSE

Introduction to Linear Models: One Mean IV

Suppose the errors have probability distribution

$$\epsilon_i \sim (\underset{\downarrow}{0} , \underset{\downarrow}{\sigma^2}) .$$

(mean: 0; variance: σ^2)

Then, the population mean (and mean for a random individual) is

$$E(Y_i) = E(\mu + \epsilon_i) = \mu + E(\epsilon_i) = \mu + 0 = \mu$$

and variance

$$Var(Y_i) = Var(\mu + \epsilon_i) = Var(\epsilon_i) = \sigma^2$$

since μ is a constant and for any r.v. ϵ_i, $Var(a + b\epsilon_i) = b^2\ Var(\epsilon_i)$

Introduction to Linear Models: One Mean V

Calculate an unbiased estimate the error variance: SSE/df

We've used the estimate $\widehat{\sigma^2} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$

and we know this estimate is unbiased: $E(S^2) = \sigma^2$

How is this related to the suggested estimator SSE/df ?

$$\frac{\text{SSE}}{\text{df}} = \frac{\text{SSE}}{n-1} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1} = s^2 \quad \leftarrow \text{unbiased for } \sigma^2 .$$

$\uparrow$

We subtract 1 to find the degrees of freedom (df), since exactly 1 parameter (μ) is estimated for the linear model for the mean

Introduction to Linear Models: One Mean VI

When the errors have a normal distribution, the standardized ("studentized") LS estimator is t_{df}-distributed

If $\epsilon_i \sim N(0, \sigma)$ then $Y_i \sim N(\mu, \sigma)$, and

$$t = \frac{\overline{Y} - E(\overline{Y})}{SE(\overline{Y})} = \frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

has a t-distribution with $n - 1$ df.

Introduction to Linear Models: One Mean VII

$100(1-\alpha)\%$ Confidence interval for μ:

$$\overline{y} \pm t^*_{n-1,1-\alpha/2} \frac{S}{\sqrt{n}}$$

Introduction to Linear Models: Two Means I

A model for two random samples from two populations

response from population j (random)		model for mean of population j (not random)		error (random)
↓		↓		↓
Y_j	$=$	μ_j	$+$	ϵ

data		population mean		random errors			
↓		↓		↓	↓		sample sizes n_1, n_2
Y_{ji}	$=$	μ_j	$+$	ϵ_{ji}	$j = 1, 2$	$i = 1, ..., n_j$	

Random samples $\Longrightarrow$ independent samples and outcomes (nearly)

$\Longrightarrow$ the random errors are independent

Introduction to Linear Models: Two Means II

This model (separate mean μ_j for each population $j = 1, 2$) is also a **linear model** for the mean response:

$\mu_j = \beta_0 + \beta_1 x_{ji}$

for constants β_0 and β_1 and $(n_1 + n_2)$ fixed values

$x_{11} = 0, x_{12} = 0, \ldots, x_{1n_1} = 0$, for data from population 1

$x_{21} = 1, x_{22} = 1, \ldots, x_{2n_2} = 1,$ for data from population 2

The x-values denote group (population) membership.

$\mu_1 = \beta_0 + \beta_1(0) = \beta_0$ and $\mu_2 = \beta_0 + \beta_1(1) = \beta_0 + \beta_1 = \mu_1 + \beta_1$.

That is, $\beta_0 = \mu_1$ and $\beta_1 = \mu_2 - \mu_1$.

So, $H_o : \mu_1 = \mu_2$ is equivalent to $H_o : \beta_1 = 0$.

Introduction to Linear Models: Two Means III

Estimate the model (means) by the Least Squares Principle

What values of $\widehat{\mu_1}$ and $\widehat{\mu_2}$ will **minimize sum of squared errors**?

$$\text{SSE} = \sum_{j=1}^{2}\sum_{i=1}^{n_j}(\ Y_{ji} - \ \widehat{\mu_j} \ )^2$$

$$\uparrow$$

$\widehat{\mu_j}$ = estimate of mean for population j, $j = 1, 2$

To minimize the SSE with respect to both $\widehat{\mu_1}$ and $\widehat{\mu_2}$:

1. take the derivative of the SSE with respect to each,
2. set both derivatives $= 0$ and
3. solve for $\widehat{\mu_1}$ and $\widehat{\mu_2}$ as functions of the data (Y_{ji}).
4. 2nd derivative test for minimum (vs. maximum)

Introduction to Linear Models: Two Means IV

$$\frac{\mathrm{d}}{\mathrm{d}\widehat{\mu_1}}\,\mathrm{SSE} \;=\; \sum_{i=1}^{n_1} 2\,(Y_{i1}-\widehat{\mu_1})(-1) \;=\; 0$$

$$\Rightarrow \quad \sum_{i=1}^{n_1} Y_{i1} \;=\; \sum_{i=1}^{n_1} \widehat{\mu_1} \;=\; n_1\widehat{\mu_1}$$

$$\Rightarrow \quad \widehat{\mu_1} \;=\; \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{i1} \;=\; \overline{Y}_1$$

Similarly, $\widehat{\mu_2} = \overline{Y}_2$ is a solution.

Can show that $\overline{Y}_1$ and $\overline{Y}_2$ indeed minimize the *SSE*.

Introduction to Linear Models: Two Means V

So, we have estimates for the mean component of the linear model.

Now, estimate variance...

Again, we consider random errors have distribution $\epsilon_{ji} \sim (0,\ \sigma^2)$.

Thus, model specifies that the two populations have same variance.

That is, $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

After all, why compare means if the variances differ widely?

$$E(Y_{1i}) = E(\mu_1 + (\mu_2 - \mu_1)(0) + \epsilon_{1i}) = \mu_1 + 0 = \mu_1$$

$$Var(Y_{1i}) = Var(\mu_1 + \epsilon_{1i}) = Var(\epsilon_{1i}) = \sigma^2$$

Similarly $E(Y_{2i}) = \mu_2$, and $Var(Y_{2i}) = \sigma^2$ for all $i = 1, \ldots, n_2$.

Introduction to Linear Models: Two Means VI

Calculate an unbiased estimate the error variance: SSE/df

An unbiased estimate for σ^2 has the form

$$\widehat{\sigma^2} = \frac{\text{SSE}}{\text{df}} = \frac{\text{SSE}}{n_1 + n_2 - 2}$$

The degrees of freedom are $\text{df} = n_1 + n_2 - 2$ since $n_1 + n_2$ is the total sample size, and 2 is the number of parameters estimated in the model component for the mean (μ_1, μ_2).

$$\widehat{\sigma^2} = \frac{\text{SSE}}{\text{df}} = \frac{\sum_{j=1}^{2}\sum_{i=1}^{n_j}(\ Y_{ji} - \widehat{\mu_j}\ )^2}{n_1 + n_2 - 2}$$

$$= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = S_p^2 = \text{“pooled variance”}$$

We can show that S_p^2 is an unbiased estimate for σ^2 : $E(S_p^2) = \sigma^2$.

Introduction to Linear Models: Two Means VII

When errors have normal distribution, standardized ("studentized") mean is t_{df}-distributed

If $\epsilon_{ji} \sim N(0, \sigma)$, then $Y_{1i} \sim N(\mu_1, \sigma)$ and $Y_{2i} \sim N(\mu_2, \sigma)$,

$$t = \frac{(\overline{Y}_1 - \overline{Y}_2) - E(\overline{Y}_1 - \overline{Y}_2)}{SE(\overline{Y}_1 - \overline{Y}_2)} = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}}$$

$$= \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Introduction to Linear Models: Two Means VIII

100$(1-\alpha)$% Confidence interval for $\mu_1 - \mu_2$:

$$(\overline{y}_1 - \overline{y}_2) \pm t^*_{n_1+n_2-2,1-\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Introduction to Linear Models: Two Means IX

An alternate model: allow for non-constant error variance

For the "constant variance" linear model,
we must believe the two populations have same variance:

$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2.$

An informal test (we skipped Section 7.3):
Is the ratio of the sample variances no more than than 2?

$$\text{if} \quad \frac{\text{largest}(S_1^2, S_2^2)}{\text{smallest}(S_1^2, S_2^2)} \leq 2, \quad \text{then} \quad \sigma_1^2 = \sigma_2^2 = \sigma^2 \quad \text{is plausible}$$

Introduction to Linear Models: Two Means X

What if this ratio is larger than 2?

Cannot assume population variances equal, so model becomes:

		population		random		
data		mean		errors		sample sizes n_1, n_2
$\downarrow$		$\downarrow$		$\downarrow$		$\downarrow$
Y_{ji}	$=$	μ_j	$+$	ϵ_{ji}		$j = 1, 2 \quad i = 1, ..., n_j$.

This looks the same as before.

However, errors have distribution $\epsilon_{ji} \sim (0, \sigma_j^2),\ j = 1, 2.$

This situation, suggests the statistic :

$$\text{“}t\text{”} = \frac{(\overline{Y}_1 - \overline{Y}_2) - E(\overline{Y}_1 - \overline{Y}_2)}{SE(\overline{Y}_1 - \overline{Y}_2)} = \frac{(\overline{Y}_1 - \overline{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Introduction to Linear Models: Two Means XI

This statistic **does not have a t-distribution.**
(The probability distribution is unknown!)

However, all statistical software uses a t-distribution to approximate p-values and confidence intervals for this situation.

. . . but what degrees of freedom to use?

Suggestion for conservative p-values (too large) and confidence intervals (too wide): $\text{df} = \min(n_1, n_2) - 1$.

Conservative because "less data" = more uncertainty

Just use the conservative df when working "by hand"
If analyzing data in R, you'll get a fancy df approximation.

Introduction to Linear Models: Two Means XII

Caution:
If you believe that the population variances are not equal, then the two populations differ more than just in their means.

Are you sure then that you want to base your decision about a difference between the populations based only on a test for a difference in means?

A comparison based only on the two means may not be appropriate.

Introduction to Linear Models: Two Variables I

response (random)		mean depends on another variable (x) (not random)		error (random)
$\downarrow$		$\downarrow$		$\downarrow$
Y	$=$	$\mu(x)$	$+$	ϵ

(Each x can be viewed as a populaton)
We consider the x-values as "fixed" and model the probability distribution of Y "conditional" on the observed x-values.

A simple model considers the mean to be a linear function of x :
$\mu(x) = \mu_{Y|x} = E(Y|x) = \beta_0 + \beta_1 x.$

This model is called the **simple linear regression model**.

Introduction to Linear Models: Two Variables II

The linear model is $Y = \beta_0 + \beta_1 x + \epsilon$,

and we observe pairs (x_i, Y_i). For each observation:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \ , \\ & \epsilon_i \sim \text{ independent } (0, \sigma^2) \qquad i = 1, 2, \ldots, n \\ \Rightarrow \quad & Y_i - (\beta_0 + \beta_1 x_i) = \epsilon_i = \text{error or "residual"} \end{aligned}$$

$$\begin{aligned} E(Y_i) &= E(\beta_0 + \beta_1 x_i + \epsilon_i) &&= \beta_0 + \beta_1 x_i + E(\epsilon_i) \\ &= \beta_0 + \beta_1 x_i \end{aligned}$$

$$\begin{aligned} Var(Y_i) &= Var(\beta_0 + \beta_1 x_i + \epsilon_i) &&= Var(\epsilon_i) \\ &= \sigma^2 \end{aligned}$$

This is a considerable assumption: constant variance for the response variable (Y) for every value of predictor variable (x).

Introduction to Linear Models: Two Variables III

Estimate for $\mu(x)$ by the "least-squares" method: minimize the **sum of squared errors** with respect to the parameters included in the model specified for the mean:

$$\text{SSE} \;=\; \sum_{i=1}^{n}(\; Y_i \;-\widehat{Y}_i\;)^2 \qquad \text{where} \qquad \widehat{Y}_i = \widehat{\mu}(x_i) = b_0 + b_1 x_i$$

LS estimates b_0 for β_0 (the intercept) and b_1 for β_1 (the slope):

$$b_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad\qquad b_0 = \bar{Y} - b_1\overline{x}\;.$$

Introduction to Linear Models: Two Variables IV

Simple algebra will yield another expression for $b_1 = r\frac{S_y}{S_x}$

where r is the sample correlation coefficient

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{Y_i-\bar{Y}}{S_y}\right)\left(\frac{x_i-\bar{x}}{S_x}\right) \ .$$

Introduction to Linear Models: Two Variables V

Why do we keep using the Least Squares Principle (LS)
to determine the "best" estimate?

Sure, it minimizes the SSE, but what's so great about that?

Assuming the variances are equal, consider any estimator of a linear function of the x's that is

1. unbiased and
2. a linear combination of the y-values (the data)

Then, the estimator found by the least squares method has less variance than all of the other estimators.

LS estimators are BLUE: Best Linear Unbiased Estimators

This is called the *Gauss-Markov Theorem*

Introduction to Linear Models: Two Variables VI

Claim:

Like all least squares estimates for parameters in a linear model, the estimates are unbiased:

$E(b_0) = \beta_0$, and $E(b_1) = \beta_1$.

And,

$$Var(b_1) = \sigma^2 \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$Var(b_0) = Var(\bar{Y}) + \bar{x}^2 Var(b_1) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)$$

In the regression model, errors have distribution $\epsilon_i \sim (0, \sigma^2)$. So

$$b_1 \sim N\left(\beta_1, \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}} \right), \quad b_0 \sim N\left(\beta_0, \sigma \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right)} \right)$$

Introduction to Linear Models: Two Variables VII

σ is unkown: calculate an unbiased estimate of the error variance: SSE/df

An unbiased estimate for σ^2 has the form

$$\widehat{\sigma}^2 \;=\; S^2 \;=\; \frac{\text{SSE}}{\text{df}} \;=\; \frac{\text{SSE}}{n-2} = \frac{\sum_{i=1}^{n} r_i^2}{n-2},$$

where each **residual** $r_i = Y_i - \hat{\mu}(x_i) = Y_i - (b_0 + b_1 x_i)$.

The degrees of freedom are $\text{df} = n - 2$ since n is the total sample size, and 2 is the number of parameters estimated in the model component for the mean (β_0, β_1).

Introduction to Linear Models: Two Variables VIII

Compare models:

$$Y_i = \mu + \epsilon_i$$

$$Y_i = \mu(x_i) + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Do we need the more complex model?

Test or confidence interval for β_1:

$$H_o : \beta_1 = 0 \quad t\text{-statistic} = \frac{b_1 - 0}{SE(b_1)}$$

$$b_1 \pm t^* SE(b_1) \quad \text{where} \quad SE(b_1) = \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

$t^* : P(T_{n-2} > t^*) = \alpha/2.$

Introduction to Linear Models: Two Variables IX

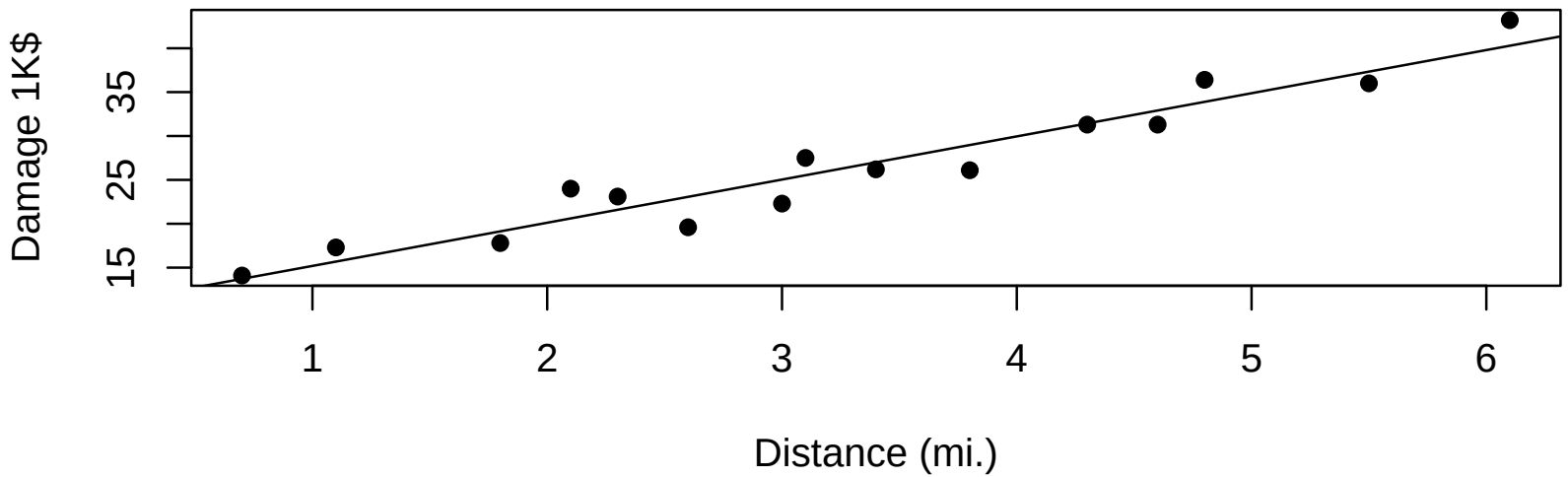
```
glimpse(fire.dat)


## Rows: 15
## Columns: 3
## $ obs    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
## $ dist   <dbl> 0.7, 1.1, 1.8, 2.1, 2.3, 2.6, 3.0, 3.1, 3.4, 3.8, 4.3, ...
## $ damage <dbl> 14.1, 17.3, 17.8, 24.0, 23.1, 19.6, 22.3, 27.5, 26.2, 2...


fire.lm <- lm(damage~dist) # Estimate the model
```

Introduction to Linear Models: Two Variables X

```
# Plot the data and the regression line
plot(dist,damage,pch=16,xlab="Distance (mi.)",
     ylab="Damage 1K$")
abline(fire.lm)
```



Introduction to Linear Models: Two Variables XI

```
summary(fire.lm)


##
## Call:
## lm(formula = damage ~ dist)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4682 -1.4705 -0.1311  1.7915  3.3915
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.2779     1.4203   7.237 6.59e-06 ***
## dist          4.9193     0.3927  12.525 1.25e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.316 on 13 degrees of freedom
## Multiple R-squared:  0.9235,Adjusted R-squared:  0.9176
## F-statistic: 156.9 on 1 and 13 DF,  p-value: 1.248e-08
```

Introduction to Linear Models: Two Variables XII

```r
# Plot residuals against x - should not show dependence on x.
# Variance should be the same for all x. Ideally residuals have
# normal distribution.
par(mfrow=c(1,2))
plot(dist,fire.lm$res,xlab="Distance (mi.)",
     ylab="Residual",pch=16)
abline(h=0)
res=(fire.lm$res-mean(fire.lm$res))/sd(fire.lm$res) # standardize residuals
qqnorm(res,main="",pch=20,xlab="",ylab="") # Check qqplot
qqline(res)
```