Regression and Analysis of Variance
Some Principles and Some Cautions

Lynne Billard

University of Georgia
lynne@stat.uga.edu

$4^{th}$ International Summer School on Data Science -
Split, Croatia
September 9-13, 2019



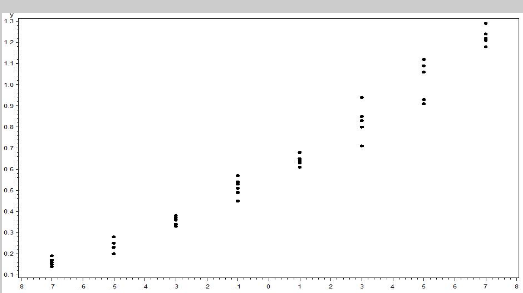
## In the beginning ...





## Regression - Models

Regression Analysis - Data from Emery, Lees, and Tootill (1951)
Observe $Y$ = Growth *Lectobacillus leichmannii*
at $X$ = Dose of Vitamin $B_{12}$ (scaled dose)

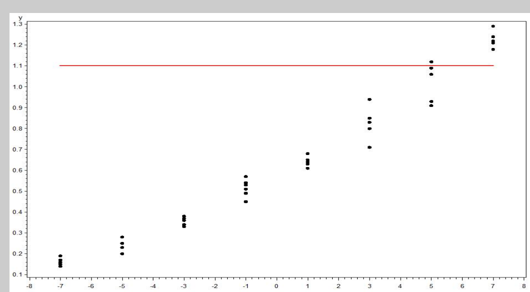



Is there a relationship between Growth and Dose, $Y$ and $X$?



## Regression - Models

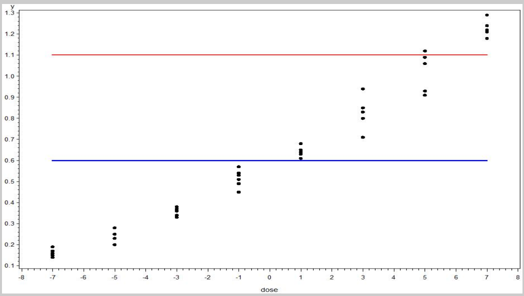
Regression Analysis
Observe $Y$ = Growth *Lectobacillus leichmannii*
at $X$ = Dose of Vitamin $B_{12}$ (scaled dose)




Is this the relationship between Growth and Dose, $Y$ and $X$?

## Regression - Models

Regression Analysis
Observe $Y$ = Growth *Lectobacillus leichmannii*
at $X$ = Dose of Vitamin $B_{12}$ (scaled dose)

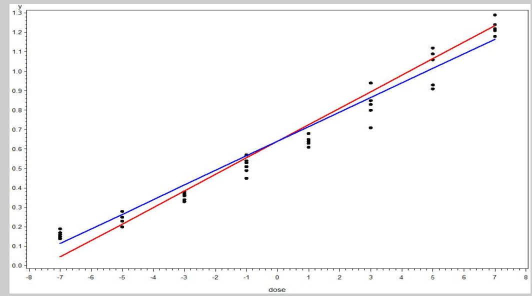


Or, this relationship between Growth and Dose, $Y$ and $X$?



## Regression - Models

Regression Analysis
Observe $Y$ = Growth *Lectobacillus leichmannii*
at $X$ = Dose of Vitamin $B_{12}$ (scaled dose)



How do we decide what is the relationship between $Y$ and $X$?



**Multiple Regression Model:**

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e$$

$Y$ = dependent variable
$X_1, \ldots, X_p$ are predictor/regression variables
$\beta_1, \ldots, \beta_p$ are regression coefficients
$\beta_0$ is intercept on $Y$- axis of regression equation
$e$ error term, – $e_i$'s independent with mean 0 and variance $\sigma_e^2$

Q? – How to estimate the parameters, $(\beta_0, \beta_1, \ldots, \beta_p)$, $\sigma_e^2$,
for observations $(Y_i, X_i)$, $i = 1, \ldots, n$

Minimize sum of squares (SS)
$SS = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y - \beta_0 - \beta_1 X_1 - \cdots - \beta_p X_p)^2$

If errors are normally distributed, these estimators are same as maximum likelihood estimators (mle)



Minimize sum of squares (SS)
$SS = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y - \beta_0 - \beta_1 X_1 - \cdots - \beta_p X_p)^2$

$p = 1$,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{SS_{xy}}{SS_{xx}}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$
$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$$

$p = p$, write $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$, Data $(Y_i, X_{ij},\ j = 1, \ldots, p)$

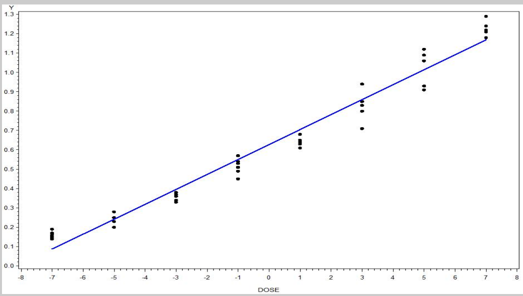
$$\hat{\boldsymbol{\beta}} = [(\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}})]^{-1}(\mathbf{Y} - \bar{Y})'(\mathbf{X} - \bar{\mathbf{X}})$$
$$\hat{\beta}_0 = \bar{\mathbf{Y}} - \hat{\boldsymbol{\beta}}\bar{\mathbf{X}}$$
$$\bar{X}_j = \frac{1}{n}\sum_{i=1}^n X_{ij}, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$$

## Regression - Models

Regression Analysis

Observe $Y$ = Growth *Lectobacillus leichmannii*

at $X$ = dose of Vitamin $B_{12}$ (scaled dose)




Linear relationship between $Y$ and $X$, i.e.,

$$Y = 0.627 + 0.077\ X$$



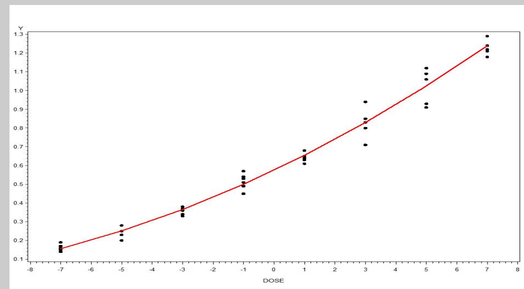
## Regression - Models

Regression Analysis

Observe $Y$ = Growth *Lectobacillus leichmannii*

at $X$ = dose of Vitamin $B_{12}$ (scaled dose)

$X_1 \equiv X,\ X_2 \equiv X^2 \quad (p = 2)$




Quadratic relationship between $Y$ and $X$, i.e.,

$$Y = 0.5750 + 0.0773\ X + 0.0025\ X^2$$

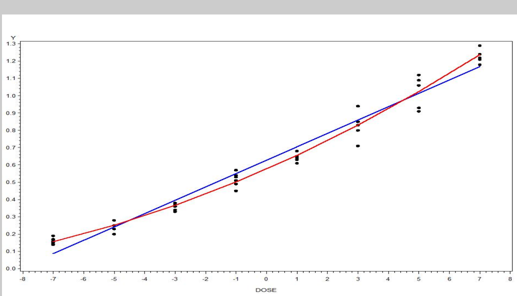


## Regression - Models

Regression Analysis

Observe $Y$ = Growth *Lectobacillus leichmannii*

at $X$ = dose of Vitamin $B_{12}$ (scaled dose)




Linear relationship: $Y = 0.6271 + 0.0773\ X$

Quadratic relationship: $Y = 0.5750 + 0.0773\ X + 0.0025\ X^2$

How do we choose between these?



## Regression - Is the Model a Good Fit?

We have: $Y = 0.5750 + 0.0773\ X + 0.0025\ X^2$

Is the model a Good Fit? – Is it an Adequate Fit?

Do we need the $X^2$ term in the model (here)?

Recall the general multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e$$

$e$ error term, – $e_i$'s independent with mean 0 and variance $\sigma_e^2$

We want to do hypothesis test: (Or, a confidence interval for $\beta_j$)

$$H_0 : \beta_j = \beta_{j0} \text{ against } H_1 : \beta_j \neq \beta_{j0}$$

In particular: $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$

Need distribution of $\hat{\beta}_j$: Can show

$$\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2((\mathbf{1},\mathbf{X})'(\mathbf{1},\mathbf{X}))^{-1})$$

(In our example, $X \equiv X_1$ and $X^2 \equiv X_2$)

## Regression - Is the Model a Good Fit?

We want to test the hypothesis:

$H_0 : \beta_j = \beta_{j0}$ against $H_1 : \beta_j \neq \beta_{j0}$

Need distribution of $\hat{\beta}_j$: Can show

$$\hat{\boldsymbol{\beta}} \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2((\mathbf{1},\mathbf{X})'(\mathbf{1},\mathbf{X}))^{-1})$$

$Var(\hat{\boldsymbol{\beta}}) = \sigma^2((\mathbf{1},\mathbf{X})'(\mathbf{1},\mathbf{X}))^{-1})$ is estimated by:

$$\widehat{Var}(\hat{\boldsymbol{\beta}}) = S^2((\mathbf{1},\mathbf{X})'(\mathbf{1},\mathbf{X}))^{-1}), \quad S^2 = \frac{1}{(n-p-1)}(\mathbf{Y}'\mathbf{Y} - \mathbf{Y}'(\mathbf{1},\mathbf{X})\hat{\boldsymbol{\beta}})$$

Test statistic (TS) is:

$$TS = \frac{\hat{\beta}_j - \beta_{j0}}{S\sqrt{\widehat{Var}(\hat{\beta}_j)}} \sim t_{n-p-1,\alpha/2}$$

(errors normally distributed) Note: $t^2_{\nu,\alpha/2} = F_{1,\nu,\alpha}$

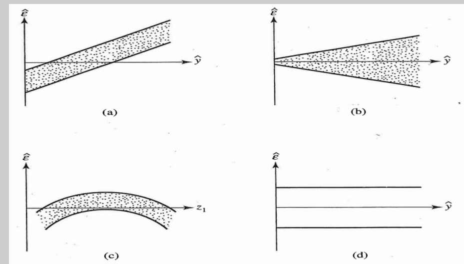
Our example, to test

$H_0 : \beta_2 = 0, \rightsquigarrow \quad TS = (0.00249)/(0.00037) = 6.68, \ p < .0001$



## Regression - Model Checking

Check the residuals:




(a) Dependence in residuals – ? $\beta_0$

(b) Variance not constant

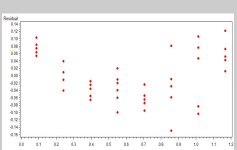
(c) Residuals versus $X_j$ – ? $X_j^2$ or $X_j X_{j'}$ needed

(d) Ideal – variance constant, no dependencies missing

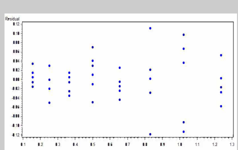


## Regression - Model Checking

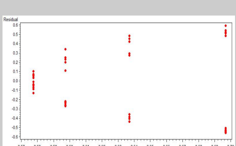
Our example – plots of residuals versus $\hat{Y}$:



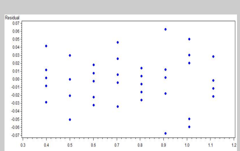
$Y = 0.627 + 0.077\ X$



$Y = 0.572 + 0.077\ X + 0.0026\ X^2$



$Y = 0.575 + 0.0025 X^2$



$Y' = log(Y) = 0.755 + 0.51\ X$

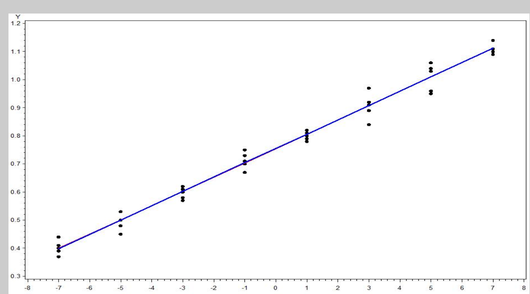


## Regression

Regression Analysis

Observe $Y$ = Growth *Lectobacillus leichmannii*

at $X$ = Dose of Vitamin $B_{12}$ (scaled dose)




Take log transform, $Y' = log(Y) = 0.755 + 0.51\ X$

## Regression - Model Checking

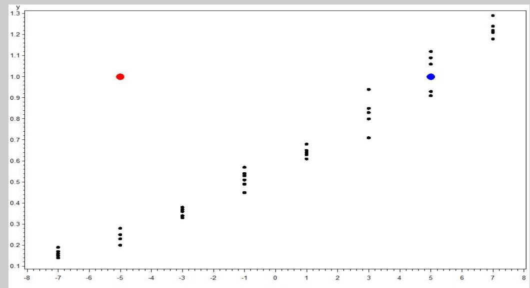
Other important checks, and what to do about them:

- Are the errors normally distributed? – QQ-plots
- Are variances constant?
- Transformations: $\log(Y)$, $\sqrt{Y}$, $1/Y$, $\log(Y+1)$, $\sqrt{(Y+1)}$, ...
- Outliers, Data cleaning
- ...



## Regression - Outlier or Typo

Regression Analysis
Observe $Y$ = Growth *Lectobacillus leichmannii*
at $X$ = Dose of Vitamin $B_{12}$ (scaled dose)

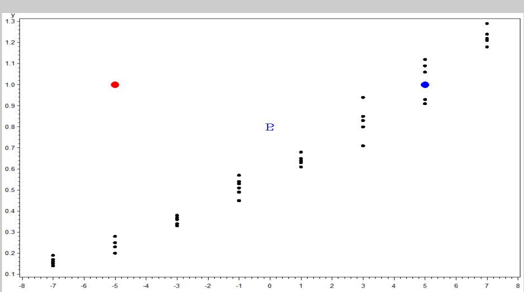



What about the observation: $Y = 1,\ X = -5$ ?
Check to see if it is a Typo: $Y = 1,\ X = 5$



## Regression - Outlier?

Regression Analysis - Data from Emery, Lees, and Tootill (1951)
Observe $Y$ = Growth *Lectobacillus leichmannii*
at $X$ = Dose of Vitamin $B_{12}$ (scaled dose)

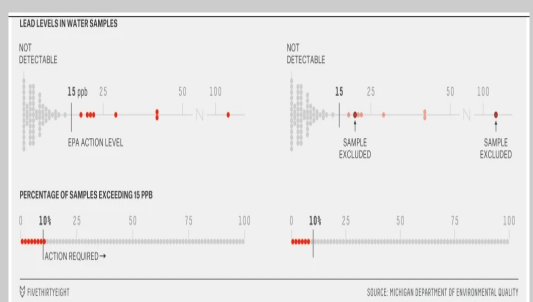



What about the observation: $Y = 0.8,\ X = 0$ ?
Challenge today is to develop (computer) methods - outliers, influence values, models, very large data sets, ...



## Regression - Outliers?

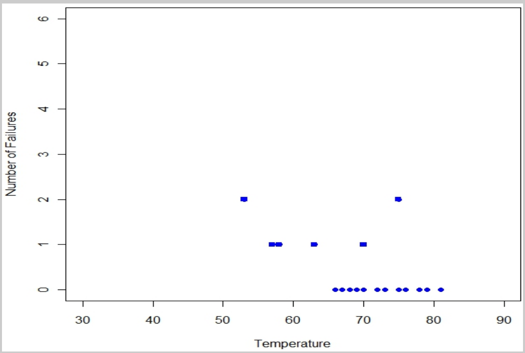
Flint River water — levels of lead




Outliers should have been included in analysis

## Regression - Outliers?

Consider this dataset




What about those zeros?



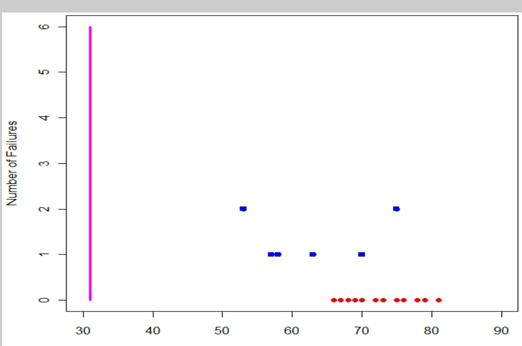
## Regression - Outliers?


Ignoring Observations can be Costly!



## Regression - Outliers?

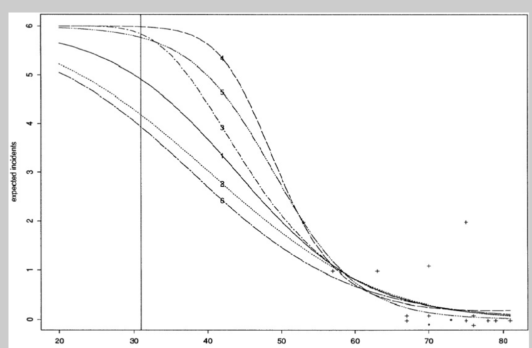
Space Shuttle Challenger - O-Ring failures and temperature.






## Regression - Outliers?

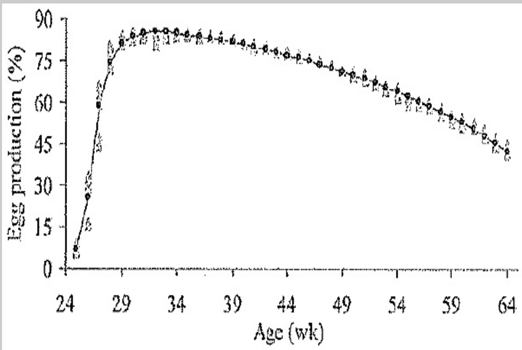
Ignoring Observations can be Costly!




Lavine (1991), Dalal *et al.* (1989)

## Other Regression Fits

Egg production - Faridi *et al.* (2011)




Growth models - many applications

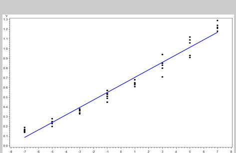




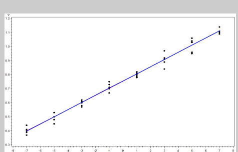
Prediction:

We have the estimated model (take $p = 1, \quad X_1 = X$ )

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$



$Y = 0.627 + 0.077\ X$



$Y' = log(Y) = 0.755 + 0.51\ X$



Prediction and Prediction Intervals: (take $p = 1, \quad X_1 = X$ )

We have the estimated model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

Prediction of E(Y) at $X_k \rightsquigarrow \hat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 X_k$

$$Var(\hat{Y}_k) = \sigma^2 \big[ \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \big]$$

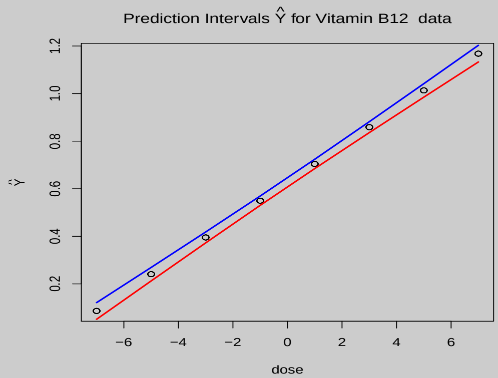
$$\hat{\sigma}^2 = S^2 = \frac{1}{(n-p-1)} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Hence, $(1-\alpha)100\%$ prediction interval for $Y_k$ at $X = X_k$ is

$$\hat{Y}_k \pm t_{n-p-1,\alpha/2} S \big[ 1 + \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \big]^{1/2}$$

(data normally distributed)

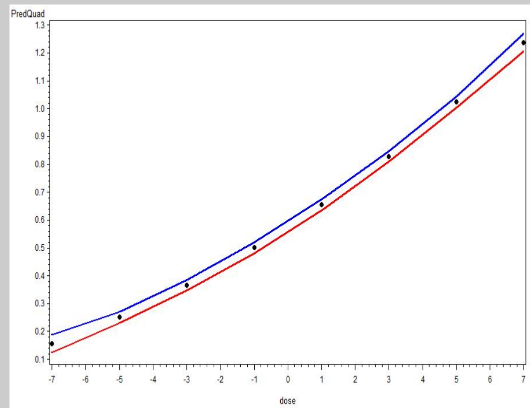
95% Prediction Interval:




$$\hat{Y}_k \pm t_{n-p-1,\alpha/2} S\left[1+\frac{1}{n}+\frac{(X_k-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right]^{1/2}$$



95% Prediction Interval: Quadratic model




$Y = 0.572 + 0.077\ X + 0.0026\ X^2$





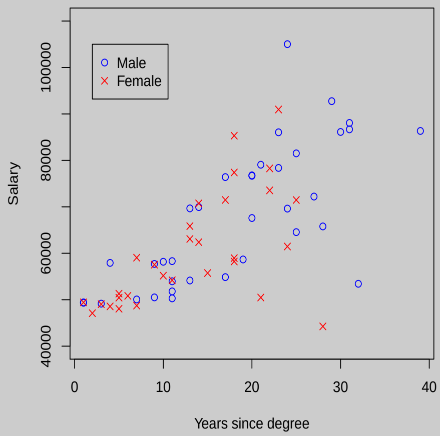
## Indicator variables

The issue of Indicator variables is too often mis-used.

- One indicator variable
- Two or more indicator variables

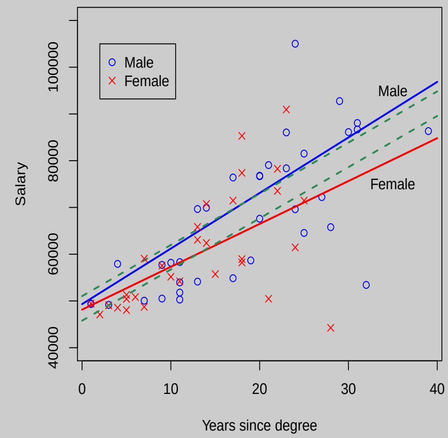
E.g., Suppose we have a response variable $Y$ = salary and $p$ regression /predictor variables, for two groups, males and females (say).

Some data:






Regression fits:






Correct Model: Take one X = years since degree ≡ years

$$Y = \beta_0 + \beta_1 \text{ years} + \beta_2 \text{ gender} + \beta_3 \text{ gender} \times \text{years} + e$$

For these data, the regression equation becomes

$\hat{Y} = 48134.0 + 917.2 \text{ years} + 1189.4 \text{ gender} + 271.3 \text{ gender} \times \text{years}$

gender = 1, males $\Rightarrow$ $\hat{Y} = 49323.4 + 1188.5 \text{ years}$

gender = 0, females $\Rightarrow$ $\hat{Y} = 48134.0 + 917.2 \text{ years}$



Incorrect Model: Take one X = years since degree ≡ years

$$Y = \beta_0 + \beta_1 \text{ years} + \beta_2 \text{ gender} + e$$

For these data, the regression equation becomes

$\hat{Y} = 46176.8 + 1064.7 \text{ years} + 2205.5 \text{ gender}$

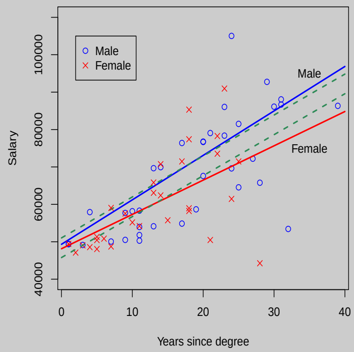
gender = 1, males $\Rightarrow$ $\hat{Y} = 48382.7 + 1064.7 \text{ years}$

gender = 0, females $\Rightarrow$ $\hat{Y} = 46176.8 + 1064.7 \text{ years}$

Correct Model:
gender = 1, males $\Rightarrow$
$Y = 49323.4 + 1188.5$ years

gender = 0, females $\Rightarrow$
$Y = 48134.0 + 917.2$ years

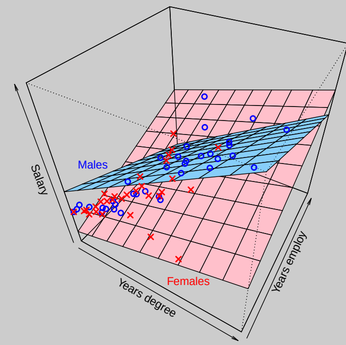
Incorrect Model:
gender = 1, males $\Rightarrow$
$\mathbf{Y = 48382.7 + 1064.7}$ years
gender = 0, females $\Rightarrow$
$\mathbf{Y = 46176.8 + 1064.7}$ years



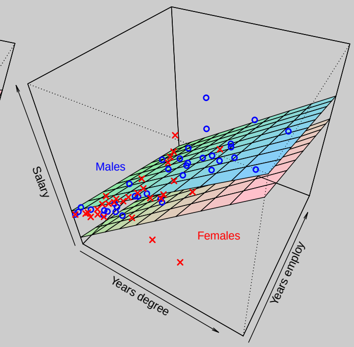



Two $X$'s: $X_1$ = years (since) degree, $X_2$ = years employed

Correct Model:




Incorrect Model:






Fits: (Regression fits)

$R^2 = \sum_{i=1}^{n} \text{residual}_i^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$

Correct model
- Males – $R_m^2 = 2802.7 \times 10^5$
- Females – $R_f^2 = 2871.2 \times 10^5$

Incorrect model
- Males – $R_m^2 = 2810.2 \times 10^5$
- Females – $R_f^2 = 2925.7 \times 10^5$

(for one $X$ model)



Fits – Tests of coincidences:
Q: Are the two regressions statistically significant?
Here, are the male and females regressions the same? Take $(X_1, X_2)$
Need model (and all male/female data)

$$\begin{aligned} Y = &\beta_0 + \beta_1 \text{ degree} + \beta_2 \text{ employ} + \beta_3 \text{ gender} \\ &+ \beta_4 \text{ gender} \times \text{degree} + \beta_5 \text{ gender} \times \text{employ} \end{aligned}$$

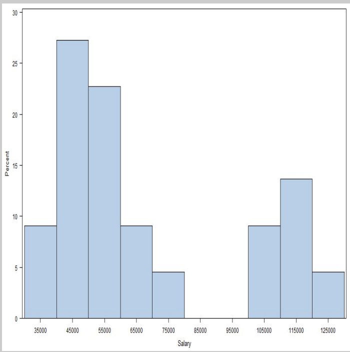
For these data the regression equation is

$$\begin{aligned} \hat{Y} = &46773.0 + 581.3 \text{ degree} + 755.4 \text{ employ} - 3085.0 \text{ gender} \\ &+ 2197.8 \text{ gender} \times \text{degree} - 2583.4 \text{ gender} \times \text{employ}. \end{aligned}$$

Test: $H_0$ : Full model (with all $\beta_0, \ldots, \beta_5$) against
$H_1$: Reduced model (with $\beta_0, \ldots, \beta_3$ only
This is an $F$-test (not the $t$-test for testing particular $\beta_j$ values)

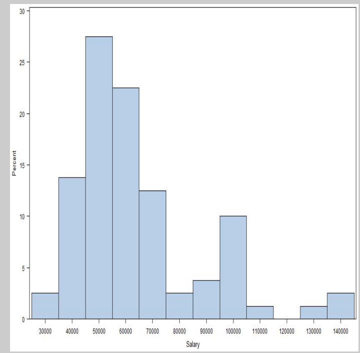
Here, $p = .0002$ – regressions are statistically significantly different
Gender only ( $H_0 : \beta_3 = 0$), $p = .0995 \rightsquigarrow$ incorrect conclusion about gender

Grouping disciplines: Normality tests





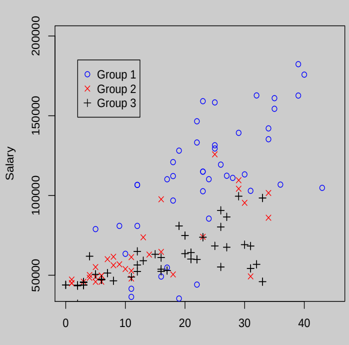

Outliers:






Disciplines / Groups: E.g., one $X$, $g$ groups -

$$Y = \beta_0 + \beta_1 X + \beta_2 \text{ group}_1 + \cdots + \beta_g \text{ group}_{g-1}$$
$$+ \beta_{g+1} \text{ group}_1 \times X + \cdots + \beta_{2g} \text{ group}_{g-1} \times X + e$$






Disciplines / Groups: E.g., one $X$, $g$ groups -

$$Y = \beta_0 + \beta_1 X + \beta_2 \text{ group}_1 + \cdots + \beta_g \text{ group}_{g-1}$$
$$+ \beta_{g+1} \text{ group}_1 \times X + \cdots + \beta_{2g} \text{ group}_{g-1} \times X + e$$

For these data, the regression equation is:

$$\hat{Y} = 43131.6 + 1018.7 \text{ years} + 428.4 \text{ group}_1 - 2634.8 \text{ group}_2$$
$$+ 1990.4 \text{ group}_1 \times \text{ years} + 722.5 \text{ group}_2 \times \text{ years.}$$

$\text{group}_1 = 1$ and $\text{group}_2 = 0$, gives, for Group 1,

$$\hat{Y} = 43560.0 + 3009.1 \text{ years;}$$

$\text{group}_1 = 0$ and $\text{group}_2 = 1$, gives, for Group 2,

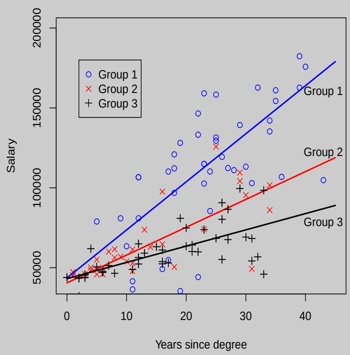
$$\hat{Y} = 40469.8 + 1741.2 \text{ years;}$$

$\text{group}_1 = 0$ and $\text{group}_2 = 0$, gives, for Group 3,

$$\hat{Y} = 43131.6 + 1018.7 \text{ years.}$$

Disciplines / Groups: E.g., one $X$, $g$ groups -

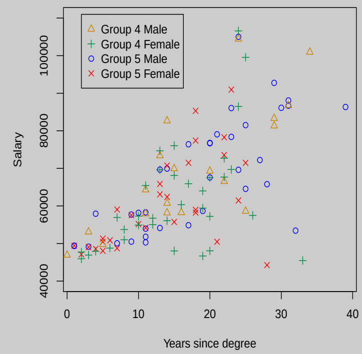
$$Y = \beta_0 + \beta_1 X + \beta_2\ \text{group}_1 + \cdots + \beta_g\ \text{group}_{g-1}$$
$$+ \beta_{g+1}\ \text{group}_1 \times X + \cdots + \beta_{2g}\ \text{group}_{g-1} \times X + e$$






Gender and Discipline: E.g. $X$ = years; group = 0,1; gender = 0,1

$$Y = \beta_0 + \beta_1\ \text{years} + \beta_2\ \text{gender} + \beta_3\ \text{gender} \times \text{years}$$
$$+ \beta_4\ \text{group} + \beta_5\ \text{group} \times \text{years}$$
$$+ \beta_6\ \text{gender} \times \text{group} + \beta_7\ \text{gender} \times \text{group} \times \text{years} + e$$






Gender and Discipline: E.g. $X$ = years; group = 0,1; gender = 0,1

$$Y = \beta_0 + \beta_1\ \text{years} + \beta_2\ \text{gender} + \beta_3\ \text{gender} \times \text{years}$$
$$+ \beta_4\ \text{group} + \beta_5\ \text{group} \times \text{years}$$
$$+ \beta_6\ \text{gender} \times \text{group} + \beta_7\ \text{gender} \times \text{group} \times \text{years} + e$$

For these data, the regression equation is:

$$\hat{Y} = 48134.0 + 917.2\ \text{years} + 1189.4\ \text{gender} + 271.3\ \text{gender} \times \text{years}$$
$$- 2550.7\ \text{group} + 241.0\ \text{group} \times \text{years}$$
$$- 3446.5\ \text{gender} \times \text{group} + 234.7\ \text{gender} \times \text{group} \times \text{years}.$$

Group 4, Males: group = 1, gender = 1 $\Rightarrow$
$\hat{Y} = 43326.2 + 1664.2$ years

**Group4, Females:** group = 1, gender =0 $\Rightarrow$
$\hat{Y} = 45583.3 + 1158.2$ **years**

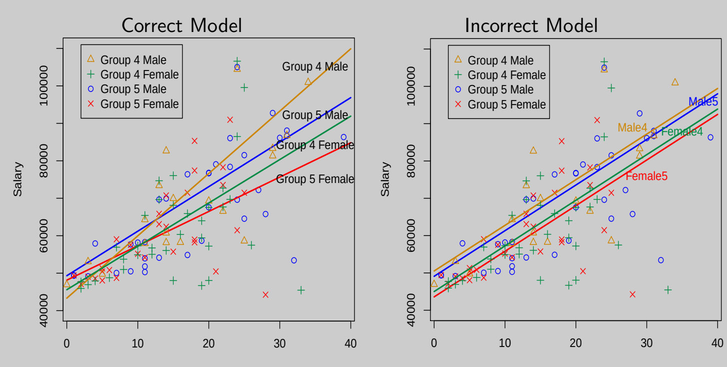
Group 5, Males: group = 0, gender = 1 $\Rightarrow$
$\hat{Y} = 49323.4 + 1188.5$ years

Group 5, Females: group = 0, gender = 0 $\Rightarrow$
$\hat{Y} = 48134.0 + 917.2$ years



Gender and Discipline: E.g. $X$ = years; group = 0,1; gender = 0,1

$$Y = \beta_0 + \beta_1\ \text{years} + \beta_2\ \text{gender} + \beta_3\ \text{gender} \times \text{years}$$
$$+ \beta_4\ \text{group} + \beta_5\ \text{group} \times \text{years}$$
$$+ \beta_6\ \text{gender} \times \text{group} + \beta_7\ \text{gender} \times \text{group} \times \text{years} + e$$



One more aspect:
Let us re-visit some of our calculations – Suppose $p = 1$, ie one X

Estimated variance $= \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$   or, write
Error Sum of Squares = SSE $= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \equiv$ Residual variation
Total SS = SSY $= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \equiv$ Total unexplained variation
Regression SS $= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \equiv$ Explained variation

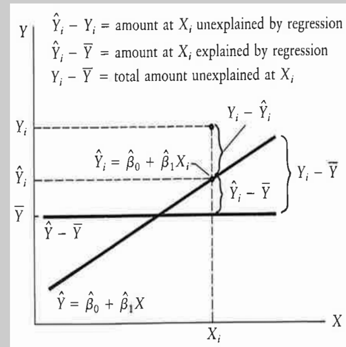
Can show:
Total SS = Regression SS + Residual SS,   or
Total SS = Model SS + Residual SS

Can set out as an analysis of variance table: (MS = SS/df)

| Source | df | SS | MS | F |
|---|---|---|---|---|
| Regression | $p - 1$ | SSY - SSE | | |
| Residual | $n - p - 1$ | SSE | | |
| Total | $n - 1$ | SSY | | |



Total SS = Regression SS + Residual SS

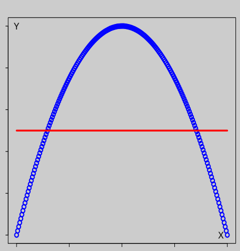



(Kleinbaum Kupper and Muller, 1988)



Finally, Correlation:
Estimated Correlation function =
$$\hat{\rho} = r = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{[\sum_{i=1}^{n} (Y_i - \bar{Y})^2 \sum_{i=1}^{n} (X_i - \bar{X})^2]^{1/2}} = \frac{SS_X}{SS_Y} \hat{\beta}_1$$

$\rho$ is a measure of the **linear** relationship between X and Y




$\hat{\rho} = 0$



## In the beginning ... Half-way!

1 Regression Analyses ✓
   - Models – predictor variables?
   - Prediction intervals
   - Model Checking, Fits, Residuals, Normality, ...
   - Outliers
   - Indicator variables?
2 Analysis of Variance (ANOVA)
   - One-way design, two-way design, ...
   - ANOVA as multiple regression
   - Covariance
   - Repeated measures
   - Covariance in repeated measures design

## Analysis of Variance (ANOVA)

ANOVA: – Some data:
Time $Y$ to complete task for $a = 4$ workers (factor A),
with $r = 6$ replications per worker

| Worker | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 20.1 | 20.5 | 23.1 | 22.8 | 25.0 | 25.2 |
| 2 | 17.2 | 16.9 | 20.0 | 19.8 | 22.1 | 22.1 |
| 3 | 16.0 | 15.9 | 17.2 | 17.1 | 24.3 | 23.8 |
| 4 | 18.8 | 19.1 | 23.7 | 23.2 | 24.3 | 21.8 |

Q: $H_0$ : Are times same for all workers?

This is a one-way analysis of variance design,
or, completely randomized design



## Analysis of Variance (ANOVA)

ANOVA: – Model:

$$Y_{ij} = \mu_i + e_{ij},\ i = 1, \ldots, a, \quad j = 1, \ldots, r_i,$$
$$= \mu + \tau_i + e$$

where

$Y_{ij}$ = observation for $j^{th}$ replication of treatment $i$

$\mu_i = i^{th}$ treatment mean

$\tau_i$ = effect of $i^{th}$ treatment ($A_i$)

$\mu$ = overall (grand) mean

$e_{ij} = ij^{th}$ observational error

Assume $e_{ij} \sim IN(0, \sigma^2)$; $\quad \sum \tau_i = 0$

Test $H_0$ : Are times same for all workers?
i.e., $H_0 : \mu_i = \mu,\ i = 1, \ldots, a$
i.e., $H_0 : \tau_i = 0, \quad i = 1, \ldots, a$



ANOVA: – One-way Model:

Can show Sum of Squares (SS)

Total SS = (A)SS + Residual SS

where

$$(A)SS = \sum_{i=1}^{a} r_i(\bar{Y}_{i\cdot} - \bar{Y})^2 = \sum_{i=1}^{a} Y_{i\cdot}^2/r_i - CF$$
$$\text{Total SS} = \sum_{i=1}^{a}\sum_{j=1}^{r_i}(Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{a}\sum_{j=1}^{r_i} Y_{ij}^2 - CF$$

hence,
Residual SS = Total SS - (A)SS

$$Y_{i\cdot} = \sum_{j=1}^{r_i} Y_{ij}, \quad CF = (\sum_{ij} Y_{ij})^2/N, \quad N = \sum_i r_i$$

| Source | df | SS | MS | F | E(MS) |
|---|---|---|---|---|---|
| A | $a-1$ | $\sum_i Y_{i\cdot}^2/r_i - CF$ | ① | | $\sigma^2 + \frac{1}{(a-1)}\sum_i r_i\tau_i^2$ |
| Residual | $N-a$ | $\sum_{ij}(Y_{ij} - \bar{Y}_{i\cdot})^2$ | ② | | $\sigma^2$ |
| Total | $N-1$ | $\sum_{ij}(Y_{ij} - \bar{Y})^2$ | | | |

df = degrees of freedom, MS = mean squares = SS/df



ANOVA: – One-way Model:
We have

| Source | df | SS | MS | F | E(MS) |
|---|---|---|---|---|---|
| A | $a-1$ | $\sum_i Y_{i\cdot}^2/r_i - CF$ | ① | ①/② | $\sigma^2 + \frac{1}{(a-1)}\sum_i r_i\tau_i^2$ |
| Residual | $N-a$ | $\sum_{ij}(Y_{ij} - \bar{Y}_{i\cdot})^2$ | ② | | $\sigma^2$ |
| Total | $N-1$ | $\sum_{ij}(Y_{ij} - \bar{Y})^2$ | | | |

Test $H_0$ : Are times same for all workers?
i.e., $H_0 : \mu_i = \mu,\ i = 1, \ldots, a$
i.e., $H_0 : \tau_i = 0,\ i = 1, \ldots, a$
i.e., $H_0 : E((A)MS) = \sigma^2$

Test statistic is: $TS = (A)MS/ResidualMS =$ ①/② $\sim F_{a-1,N-a}$

# Analysis of Variance (ANOVA)

ANOVA: – Our data:

$Y$ = Time, for $a = 4$ workers (factor A), $r = 6$ replications

| Worker | $Y_{ij}$ | | | | | |
|---|---|---|---|---|---|---|
| 1 | 20.1 | 20.5 | 23.1 | 22.8 | 25.0 | 25.2 |
| 2 | 17.2 | 16.9 | 20.0 | 19.8 | 22.1 | 22.1 |
| 3 | 16.0 | 15.9 | 17.2 | 17.1 | 24.3 | 23.8 |
| 4 | 18.8 | 19.1 | 23.7 | 23.2 | 24.3 | 21.8 |

| Source | df | SS | MS | F | $p$ |
|---|---|---|---|---|---|
| Worker | 3 | 55.633 | 18.544 | 2.41 | .0967 |
| Residual | 20 | 153.660 | 7.683 | - | - |
| Total | 23 | 209.293 | | | |

$p = .0967 > .05$ suggests the workers all the same
– at least for this model and this analysis

However,



Two-way model:

| Worker | $Y_{ijk}$ | | | | | |
|---|---|---|---|---|---|---|
| | Computer1 | | Computer2 | | Computer3 | |
| 1 | 20.1 | 20.5 | 23.1 | 22.8 | 25.0 | 25.2 |
| 2 | 17.2 | 16.9 | 20.0 | 19.8 | 22.1 | 22.1 |
| 3 | 16.0 | 15.9 | 17.2 | 17.1 | 24.3 | 23.8 |
| 4 | 18.8 | 19.1 | 23.7 | 23.2 | 24.3 | 21.8 |

$Y$ = Time, for $a = 4$ workers (factor A), $b = 3$ computers (factor B), and $r = 2$ replications for each worker-computer combination

Write model as:

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + e_{ijk}, \;\; i = 1, \ldots, a, j = 1, \ldots, b, k = 1, \ldots, r$$

where

$$A_i = \text{effect of } i^{th} \text{ level of } A,$$
$$B_j = \text{effect of } j^{th} \text{ level of } B,$$
$$(AB)_{ij} = \text{effect of interaction of } A_i \text{ with } B_j,$$

$\sum_i A_i = 0, \;\; \sum_j B_j = 0, \;\; \sum_i (AB)_{ij} = 0, \;\; \sum_j (AB)_{ij} = 0$
$e_{ijk} \sim IN(0, \sigma^2)$



Model: $Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + e_{ijk}$

Can show:

$$\text{Total SS} = \text{(A)SS} + \text{(B)SS} + \text{(AB)SS} + \text{ErrorSS}$$

ANOVA Table is

| Source | df | SS | MS | F | E(MS) |
|---|---|---|---|---|---|
| A | $a-1$ | $\sum_i rb(\bar{Y}_{i\cdot\cdot} - \bar{Y})^2$ | ① | | $\sigma^2 + \frac{br}{(a-1)} \sum_i A_i^2$ |
| B | $b-1$ | $\sum_j ra(\bar{Y}_{\cdot j\cdot} - \bar{Y})^2$ | ② | | $\sigma^2 + \frac{ar}{(b-1)} \sum_j B_j^2$ |
| AB | $(a-1)(b-1)$ | $\sum_{ij} r(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y})^2$ | ③ | | $\sigma^2 + \frac{r}{(a-1)(b-1)} \sum_{ij} (AB)_{ij}^2$ |
| Error | $(r-1)ab$ | Difference | ④ | | $\sigma^2$ |
| Total | $rab-1$ | $\sum_{ijk}(Y_{ijk} - \bar{Y})^2$ | | | |



Model: $Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + e_{ijk}$

| Source | df | SS | MS | F | E(MS) |
|---|---|---|---|---|---|
| A | $a-1$ | $\sum_i rb(\bar{Y}_{i\cdot\cdot} - \bar{Y})^2$ | ① | ①/④ | $\sigma^2 + \frac{br}{(a-1)} \sum_i A_i^2$ |
| B | $b-1$ | $\sum_j ra(\bar{Y}_{\cdot j\cdot} - \bar{Y})^2$ | ② | ②/④ | $\sigma^2 + \frac{ar}{(b-1)} \sum_j B_j^2$ |
| AB | $(a-1)(b-1)$ | $\sum_{ij} r(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y})^2$ | ③ | ③/④ | $\sigma^2 + \frac{r}{(a-1)(b-1)} \sum_{ij} (AB)_{ij}^2$ |
| Error | $(r-1)ab$ | Difference | ④ | | $\sigma^2$ |
| Total | $rab-1$ | $\sum_{ijk}(Y_{ijk} - \bar{Y})^2$ | | | |

$H_0$ : All workers same
$H_0 : A_i = 0, \; \forall \; i$
$H_0 : E((A)MS) = \sigma^2$
$TS = \frac{(A)MS}{ErrorMS} \sim F_{a-1,(r-1)ab}$

$H_0$ : All computers same
$H_0 : B_j = 0, \; \forall \; j$
$H_0 : E((B)MS) = \sigma^2$
$TS = \frac{(B)MS}{ErrorMS} \sim F_{b-1,(r-1)ab}$

$H_0$ : No interactions between worker and computer
$H_0 : (AB)_{ij} = 0, \; \forall \; i, j$
$H_0 : E((AB)MS) = \sigma^2$
$TS = (AB)MS/ErrorMS \sim F_{(a-1)(b-1),(r-1)ab}$

Our data:

| Worker | Computer1 | | Computer2 | | Computer3 | |
|---|---|---|---|---|---|---|
| | $Y_{ijk}$ | | | | | |
| 1 | 20.1 | 20.5 | 23.1 | 22.8 | 25.0 | 25.2 |
| 2 | 17.2 | 16.9 | 20.0 | 19.8 | 22.1 | 22.1 |
| 3 | 16.0 | 15.9 | 17.2 | 17.1 | 24.3 | 23.8 |
| 4 | 18.8 | 19.1 | 23.7 | 23.2 | 24.3 | 21.8 |

| Source | df | SS | MS | F | $p$ |
|---|---|---|---|---|---|
| Worker | 3 | 55.633 | 18.544 | 18.544 | $< .0001$ |
| Computer | 2 | 121.561 | 60.780 | 200.38 | $< .0001$ |
| Worker$\times$Computer | 6 | 28.459 | 4.743 | 15.64 | $< .0001$ |
| Error | 12 | 3.640 | 0.303 | - | - |
| Total | 23 | 209.293 | | | |



One-way model:

| Source | df | SS | MS | F | $p$ |
|---|---|---|---|---|---|
| Worker | 3 | 55.633 | 18.544 | 2.41 | .0967 |
| Residual | 20 | 153.660 | 7.683 | - | - |
| Total | 23 | 209.293 | | | |

Two-way model:

| Source | df | SS | MS | F | $p$ |
|---|---|---|---|---|---|
| Worker | 3 | 55.633 | 18.544 | 18.544 | $< .0001$ |
| Computer | 2 | 121.561 | 60.780 | 200.38 | $< .0001$ |
| Worker$\times$Computer | 6 | 28.459 | 4.743 | 15.64 | $< .0001$ |
| Error | 12 | 3.640 | 0.303 | - | - |
| Total | 23 | 209.293 | | | |



One-way model: D1

| Source | df | SS | MS | F | $p$ |
|---|---|---|---|---|---|
| Worker | 3 | 55.633 | 18.544 | 2.41 | .0967 |
| Residual | 20 | 153.660 | 7.683 X | - | - |
| Total | 23 | 209.293 | | | |

Model: $Y_{ij} = \mu + A_i + e_{ij}$

Two-way model: D2

| Source | df | SS | MS | F | $p$ |
|---|---|---|---|---|---|
| Worker | 3 | 55.633 | 18.544 | 18.544 | $< .0001$ |
| Computer | 2 | 121.561 | 60.780 | 200.38 | $< .0001$ |
| Worker$\times$Computer | 6 | 28.459 | 4.743 | 15.64 | $< .0001$ |
| Error | 12 | 3.640 | 0.303 ✓ | - | - |
| Total | 23 | 209.293 | | | |

Model: $Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + e_{ijk}$

Residual SS (of D1) $=$ B SS $+$ (A$\times$ B)SS $+$ Error SS (of D2)

(D1) $e_{ij} \equiv B_j + (AB)_{ij} + e_{ijk}$ (D2)



So far, we have assumed the $(a = 4)$ workers and $(b = 3)$ computers were 4 specific workers and 3 specific computers – conclusions apply to these only. Fixed effects model (also called parametric model.

Other options: df, SS, MS same; E(MS) and F-values change

- Random effects model – workers randomly selected from (population of) workers, and computers randomly selected from (population of) computers
  Conclusions apply to all workers and all computers
- Mixed effects model – specific workers considered, and computers randomly selected from (population of) computers
  Conclusions apply to these specific workers for all computers
- Mixed effects model – workers randomly selected from (population of) workers, and specific computers used
  Conclusions apply to all workers for these specific computers

General model: $Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + e_{ijk}$

Conditions include

- Fixed model (Parametric)-
  $\sum_i A_i = 0$, $i^{th}$ A effect
  $\sum_j B_j = 0$, $j^{th}$ B effect
  $\sum_{ij} (AB)_{ij} = 0$, $(ij)^{th}$, AB interaction effect
- Random model -
  $A_i \sim IN(0, \sigma_a^2)$, $E(A_i) = 0 \ \forall\ i$, $i^{th}$ A effect
  $B_j \sim IN(0, \sigma_b^2)$, $E(B_j) = 0 \ \forall\ j$, $j^{th}$ B effect
  $(AB)_{ij} \sim IN(0, \sigma_{ab}^2)$, $E(AB_{ij}) = 0 \ \forall\ (i, j)$, $(ij)^{th}$ interaction effect, $A$ and $B$ independent
- Mixed model - (take A fixed, B random)
  $\sum_i A_i = 0$, $i^{th}$ A effect
  $B_j \sim IN(0, \sigma_b^2)$, $E(B_j) = 0 \ \forall\ j$, $j^{th}$ B effect
  $(AB)_{ij} \sim N(0, \sigma_{ab}^2)$, $\sum_i (AB)_{ij} = 0 \ \forall\ j, E(A_{ij}) = 0 \ \forall\ i$
- All models: $e_{ijk} \sim IN(0, \sigma^2)$, $\forall\ i, j, k$



Now: df, SS, MS same;
E(MS) and F-values change – $E(MS)$ tables are now:

| Source | Fixed | Random |
|---|---|---|
| A | $\sigma^2 + \frac{br}{(a-1)} \sum_i A_i^2$ | $\sigma^2 + r\sigma_{ab}^2 + rb\sigma_a^2$ |
| B | $\sigma^2 + \frac{ar}{(b-1)} \sum_j B_j^2$ | $\sigma^2 + r\sigma_{ab}^2 + ra\sigma_b^2$ |
| AB | $\sigma^2 + \frac{r}{(a-1)(b-1)} \sum_{ij} (AB)_{ij}^2$ | $\sigma^2 + r\sigma_{ab}^2$ |
| Error | $\sigma^2$ | $\sigma^2$ |

| Source | Mixed (A fixed) | Mixed (B fixed) |
|---|---|---|
| A | $\sigma^2 + r\sigma_{ab}^2 + \frac{br}{(a-1)} \sum_i A_i^2$ | $\sigma^2 + rb\sigma_a^2$ |
| B | $\sigma^2 + ra\sigma_b^2$ | $\sigma^2 + \sigma_{ab}^2 + \frac{ar}{(b-1)} \sum_j B_j^2$ |
| AB | $\sigma^2 + r\sigma_{ab}^2$ | $\sigma^2 + r\sigma_{ab}^2$ |
| Error | $\sigma^2$ | $\sigma^2$ |



Our data:

| Source | df | SS | MS | F | $p$ |
|---|---|---|---|---|---|
| Worker | 3 | 55.633 | 18.544 | 18.544 | $< .0001$ |
| Computer | 2 | 121.561 | 60.780 | 200.38 | $< .0001$ |
| Worker×Computer | 6 | 28.459 | 4.743 | 15.64 | $< .0001$ |
| Error | 12 | 3.640 | 0.303 | - | - |
| Total | 23 | 209.293 | | | |

| | | Fixed | | Random | | Mixed | |
|---|---|---|---|---|---|---|---|
| Source | df | F | $p$ | F | $p$ | F | $p$ |
| A Worker | 3 | 18.544 | $< .0001$ | 3.91 | .0732 | F: 3.91 | .0732 |
| B Computer | 2 | 200.38 | $< .0001$ | 12.81 | .0068 | R: 200.38 | $< .0001$ |
| A×B | 6 | 15.64 | $< .0001$ | 15.64 | $< .0001$ | 15.64 | $< .0001$ |
| Error | 12 | - | - | - | - | - | - |
| Total | 23 | | | | | | |



ANOVA as Multiple Regression:
A study on a new citrus-flavored soft drink was undertaken to see what were the color preferences for customers. The observations are $Y =$ number of cases sold per 1000 in the study. There were 5 replications for each of 4 colors.

| Color $i$ | Observations $Y_{ij}$ | | | | | $Y_{i.}$ | $\bar{Y}_i$ |
|---|---|---|---|---|---|---|---|
| colorless | 26.5 | 28.7 | 25.1 | 29.1 | 27.2 | 136.6 | 27.32 |
| pink | 31.2 | 28.3 | 30.8 | 27.9 | 29.6 | 147.8 | 29.56 |
| orange | 27.9 | 25.1 | 28.5 | 24.2 | 26.5 | 132.2 | 26.44 |
| lime | 30.8 | 29.6 | 32.4 | 31.7 | 32.8 | 157.3 | 31.46 |

The one-way ANOVA model is:

$$Y_{ij} = \mu + \tau_i + e_{ij}, \ i = 1, \ldots, 4, \ j = 1, \ldots, 5$$

The multiple regression model ($p = 3$) is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_i, \ i = 1, \ldots, 20$$

| Color | $X_{i1}$ | $X_{i2}$ | $X_{i3}$ | $Y$ |
|---|---|---|---|---|
| colorless | 0 | 0 | 0 | 26.5 |
| | 0 | 0 | 0 | 28.7 |
| | 0 | 0 | 0 | 25.1 |
| | 0 | 0 | 0 | 29.1 |
| | 0 | 0 | 0 | 27.2 |
| pink | 1 | 0 | 0 | 31.2 |
| | 1 | 0 | 0 | 28.3 |
| | 1 | 0 | 0 | 30.8 |
| | 1 | 0 | 0 | 27.9 |
| | 1 | 0 | 0 | 29.6 |
| orange | 0 | 1 | 0 | 27.9 |
| | 0 | 1 | 0 | 25.1 |
| | 0 | 1 | 0 | 28.5 |
| | 0 | 1 | 0 | 24.2 |
| | 0 | 1 | 0 | 26.5 |
| lime | 0 | 0 | 1 | 30.8 |
| | 0 | 0 | 1 | 29.6 |
| | 0 | 0 | 1 | 32.4 |
| | 0 | 0 | 1 | 31.7 |
| | 0 | 0 | 1 | 32.8 |

$$\begin{aligned} X_1 &= 1, \text{ if pink,} \\ &= 0, \text{ otherwise;} \end{aligned}$$

$$\begin{aligned} X_2 &= 1, \text{ if orange,} \\ &= 0, \text{ otherwise;} \end{aligned}$$

$$\begin{aligned} X_3 &= 1, \text{ if lime,} \\ &= 0, \text{ otherwise.} \end{aligned}$$



Analyses Outputs:

ANOVA:

| Source | df | SS | MS | F | $p$ |
|---|---|---|---|---|---|
| Color | 3 | 76.8455 | 25.6152 | 10.49 | .0005 |
| Error | 16 | 39.0840 | 2.44275 | - | - |
| Total | 19 | 115.9295 | | | |

Model

$\bar{Y}_1 = 27.32,\ \bar{Y}_2 = 29.56,\ \bar{Y}_3 = 26.44,\ \bar{Y}_4 = 31.46$

Regression:

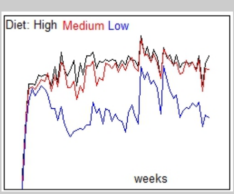
| Source | df | SS | MS | F | $p$ |
|---|---|---|---|---|---|
| Model | 3 | 76.8455 | 25.6152 | 10.49 | .0005 |
| Error | 16 | 39.0840 | 2.44275 | - | - |
| Total | 19 | 115.9295 | | | |

Regression

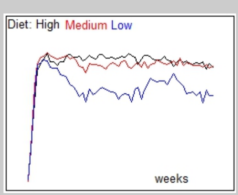
$\hat{\beta}_0 = 27.32,\ \hat{\beta}_1 = 2.24,\ \hat{\beta}_2 = -0.88,\ \hat{\beta}_3 = 4.14$



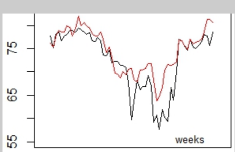
## Twist - Covariate present




Average Daily Feed Intake



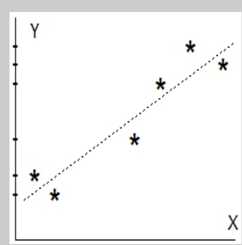

Egg Production




Room Temperature is a **covariate**



## Standard Simple Regression



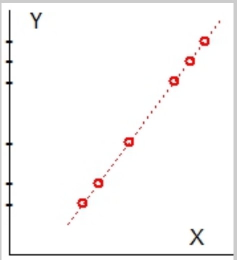

$$Y = \alpha + \beta X + e$$

where

- $Y$ = response variable
- $X$ = predictor variable, covariate
- $\beta$ = regression parameter, slope
- $\alpha$ = intercept on $Y$-axis
- $e$ = observation error; $E(e) = 0,\ Var(e) = \sigma^2$

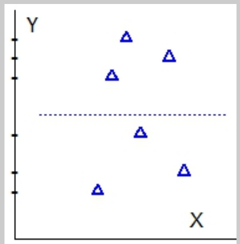
## Standard Simple Regression

Consider - Two Extremes - with same $Y$ observation values




$Y = \alpha + \beta X + e$

No observation error $\sigma^2 = 0$



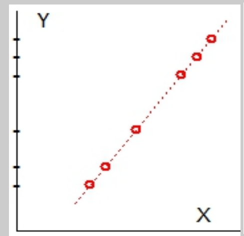

$Y = \alpha + \beta X + e$

No regression present $\beta = 0$



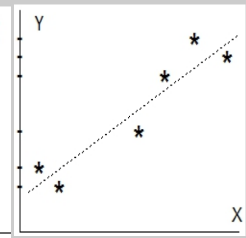
## Standard Simple Regression

Consider - Two Extremes - with same $Y$ observation values



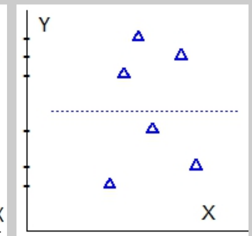

$Y = \alpha + \beta X + e$

No error $\sigma^2 = 0$




$Y = \alpha + \beta X + e$

Error and regression



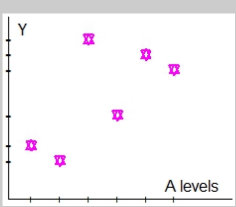

$Y = \alpha + \beta X + e$

No regression $\beta = 0$



## One Factor ANOVA

Analysis of variance - one factor A – same $Y$ values




$Y = \mu + A_i + e, \quad \sigma^2 = 0$

where

$Y$ = response variable

$\mu$ = overall mean

$A_i$ (or, $\tau$) = effect of $A_i$

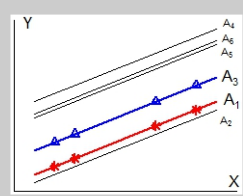
$e$ = observation error; $Var(e_i) = \sigma^2$

(various model conditions)



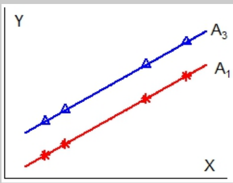
## Regression and ANOVA

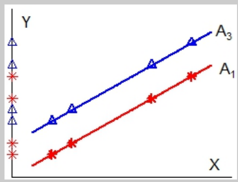
Analysis of covariance - one factor A



## Regression and ANOVA

Take levels $A_1$ and $A_3$:



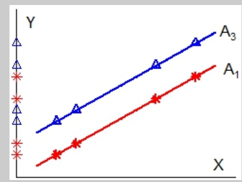




There are NO observation errors present, $\sigma^2 = 0$
Variations in $Y$ due solely to regression and level of $A$

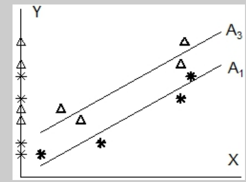


## Regression and ANOVA

Two cases – Take levels $A_1$ and $A_3$ – Same $Y$ observations:




No errors $\sigma^2 = 0$

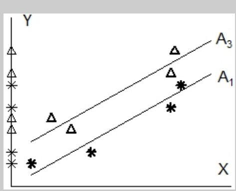



General case: errors present



## One Factor ANCOVA

Analysis of COvariance - one factor A, covariate $X$




$$Y = \mu + A_i + \gamma X + e$$

$Y$ = response variable
$\mu$ = overall mean
$A_i$ = effect of $A_i$
$X$ = predictor variable, covariate
$\gamma$ = regression parameter
$e$ = observation error; $Var(e_i) = \sigma^2$



An example:
A researcher wanted to study the effect of $a = 4$ drugs in delaying atrophy of denervated muscles in rats. Atrophy is measured by the loss in weight; but the initial weight of the muscle could not be measured (without killing the rat). Instead the initial weight $X$ of the rat was measured. After 12 days, the rats were killed and the weight $Y$ of the denervated muscle was measured.

| Drug A | | Drug B | | Drug C | | Drug D | |
|---|---|---|---|---|---|---|---|
| X | Y | X | Y | X | Y | X | Y |
| 198 | 0.34 | 233 | 0.41 | 204 | 0.57 | 186 | 0.81 |
| 175 | 0.43 | 250 | 0.87 | 234 | 0.80 | 286 | 1.01 |
| 199 | 0.41 | 289 | 0.91 | 211 | 0.69 | 245 | 0.97 |
| 224 | 0.48 | 255 | 0.87 | 214 | 0.84 | 215 | 0.87 |
| 796 | 1.66 | 1027 | 3.06 | 863 | 2.90 | 932 | 3.66 |

Model:
$Y_{ij} = \mu + \tau_i + \gamma(X_{ij} - \bar{X}) + e_{ij},\ i = 1, \ldots, a,\ j = 1, \ldots, r_i$

SSs for Y
$(A)SS_y = \sum_i r_i(\bar{Y}_i - \bar{Y})^2$
$ErrorSS_y = \sum_{ij}(Y_{ij} - \bar{Y}_{i\cdot})^2$
$TotalSS_y = \sum_{ij}(Y_{ij} - \bar{Y})^2$

SSs for X
$(A)SS_x = \sum_i r_i(\bar{X}_i - \bar{X})^2$
$ErrorSS_x = \sum_{ij}(X_{ij} - \bar{X}_{i\cdot})^2$
$TotalSS_x = \sum_{ij}(X_{ij} - \bar{X})^2$

Sum of Product - SP for XY
$(A)SP = \sum_i r_i(\bar{X}_i - \bar{X})(\bar{Y}_i - \bar{Y})$
$ErrorSP = \sum_{ij}(X_{ij} - \bar{X}_{i\cdot})(Y_{ij} - \bar{Y}_{i\cdot})$
$TotalSP = \sum_{ij}(X_{ij} - \bar{X})(Y_{ij} - \bar{Y})$

Adjustment for Regression:

$$\begin{aligned} Adjusted\ TotalSS_y &= TotalSS_y - (TotalSP)^2/TotalSS_x \\ Adjusted\ ErrorSS_y &= ErrorSS_y - (ErrorSP)^2/ErrorSS_x \\ Adjusted\ (A)SS_y &= Adjusted\ TotalSS_y - Adjusted\ ErrorSS_y \end{aligned}$$



No adjustment for regression

| Source | df | SS | MS | F | $p$ |
|---|---|---|---|---|---|
| Drugs | 3 | 0.5288 | 0.1763 | 8.52 | .0027 |
| Residual | 12 | 0.2484 | 0.0207 | - | - |
| Total | 15 | 0.7772 | | | |

$\hat{\sigma}^2 = .2484/12 = .0207$ X

With adjustment for regression

| Source | df | SS | MS | F | $p$ |
|---|---|---|---|---|---|
| Drugs | 3 | 0.2982 | 0.0994 | 6.890 | .0071 |
| Residual | 11 | 0.1587 | 0.0144 | - | - |
| Total | 15 | 0.7772 | | | |

$\hat{\sigma}^2 = .1587/11 = .0144$ ✓



As for SSs for $Y$, so should means be adjusted

| | Drug A | | Drug B | | Drug C | | Drug D | | |
|---|---|---|---|---|---|---|---|---|---|
| | X | Y | X | Y | X | Y | X | Y | |
| | 198 | 0.34 | 233 | 0.41 | 204 | 0.57 | 186 | 0.81 | |
| | 175 | 0.43 | 250 | 0.87 | 234 | 0.80 | 286 | 1.01 | |
| | 199 | 0.41 | 289 | 0.91 | 211 | 0.69 | 245 | 0.97 | |
| | 224 | 0.48 | 255 | 0.87 | 214 | 0.84 | 215 | 0.87 | |
| Totals | 796 | 1.66 | 1027 | 3.06 | 863 | 2.90 | 932 | 3.66 | |
| $\bar{Y}_i$ | | 0.4150 | | 0.7650 | | 0.7250 | | 0.9150 | Unadjusted Means |
| $\bar{Y}_i'$ | | 0.5014 | | 0.6675 | | 0.7580 | | 0.8931 | Adjusted Means |



# Example - Two-Factor Covariate Design

Example: Diet - A (High, Medium, Low protein content) feed to hens. Two different time periods week - B ($B_1, B_2$). Interest in food intake = Y; three replications for each diet×week combination. Temperature = X was also measured

| | | B - Week | | | |
|---|---|---|---|---|---|
| $A$ | | $B_1$ | | $B_2$ | |
| Diet | Hen | $X$ | $Y$ | $X$ | $Y$ |
| $A_1$ | 1 | 2 | 6 | 3 | 12 |
| High | 2 | 4 | 9 | 8 | 16 |
| | 3 | 10 | 14 | 13 | 20 |
| $A_2$ | 4 | 1 | 4 | 0 | 6 |
| Medium | 5 | 7 | 10 | 8 | 12 |
| | 6 | 9 | 7 | 8 | 8 |
| $A_3$ | 7 | 6 | 8 | 3 | 8 |
| Low | 8 | 7 | 12 | 9 | 16 |
| | 9 | 8 | 13 | 11 | 20 |

Two-factor Covariance Model:

$$Y_{ijk} = \mu + A_i + B_j + (AB)_{ij} + +\gamma(X_{ij} - \bar{X}) + e_{ijk},$$
$$i = 1, \ldots, a, \; j = 1, \ldots, b, \; k = 1, \ldots, r$$

Need – $SS_y$, $SS_x$, $SP$ (Sum of Products)

$SS_y$: $(A)SS_y = rb\sum_i(\bar{Y}_{i\cdot\cdot} - \bar{Y})^2$, $(B)SS_y = ra\sum_j(\bar{Y}_{\cdot j\cdot} - \bar{Y})^2$

$(AB)SS_y = r\sum_{ij}(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{ij\cdot})^2$

$ErrorSS_y = \sum_{ijk}(Y_{ijk} - \bar{Y}_{ij\cdot})^2 \equiv E_{yy}$, $TotalSS_y = \sum_{ijk}(Y_{ijk} - \bar{Y})^2$

$SS_x$: $(A)SS_x = rb\sum_i(\bar{X}_{i\cdot\cdot} - \bar{X})^2$, $(B)SS_x = ra\sum_j(\bar{X}_{\cdot j\cdot} - \bar{X})^2$

$(AB)SS_x = r\sum_{ij}(\bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X}_{ij\cdot})^2$

$ErrorSS_x = \sum_{ijk}(X_{ijk} - \bar{X}_{ij\cdot})^2 \equiv E_{xx}$, $TotalSS_x = \sum_{ijk}(X_{ijk} - \bar{X})^2$

$SP$: $(A)SP = rb\sum_i(\bar{X}_{i\cdot\cdot} - \bar{X})(\bar{Y}_{i\cdot\cdot} - \bar{Y})$

$(B)SP = ra\sum_j(\bar{X}_{\cdot j\cdot} - \bar{X})(\bar{Y}_{\cdot j\cdot} - \bar{Y})$

$(AB)SP = r\sum_{ij}(\bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X}_{ij\cdot})(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{ij\cdot})$

$ErrorSP = \sum_{ijk}(X_{ijk} - \bar{X}_{ij\cdot})(Y_{ijk} - \bar{Y}_{ij\cdot}) \equiv E_{xy}$

$TotalSP = \sum_{ijk}(X_{ijk} - \bar{X})(Y_{ijk} - \bar{Y})$



We need to adjust $SS_y$'s for presence of regression:

Adjusted $SS_y \equiv SS'_y \equiv$ Adj $SS_y$

$$Adj\ ErrorSS = ErrorSS_y - \frac{(ErrorSP)^2}{ErrorSS_x}$$
$$Adj\ (A)SS = (A)SS_y + \frac{(ErrorSP)^2}{ErrorSS_x} - \frac{((A)SP + ErrorSP)^2}{(A)SS_x + ErrorSS_x}$$
$$Adj\ (B)SS = (B)SS_y + \frac{(ErrorSP)^2}{ErrorSS_x} - \frac{((B)SP + ErrorSP)^2}{(B)SS_x + ErrorSS_x}$$
$$Adj\ (AB)SS = (AB)SS_y + \frac{(ErrorSP)^2}{ErrorSS_x} - \frac{((AB)SP + ErrorSP)^2}{(AB)SS_x + ErrorSS_x}$$



# Example - Two-Factor Design

Our data:

Factorial Design ignoring regression

| Source | df | SS | MS | F | $p$ |
|---|---|---|---|---|---|
| Diet A | 2 | 100.000 | 50.000 | 3.16 | .0791 |
| Week B | 1 | 68.056 | 68.056 | 4.30 | .0603 |
| A× B | 2 | 16.444 | 8.222 | 0.52 | .6077 |
| Error | 12 | 190.000 | 15.833 X | | |
| Total | 17 | 374.500 | | | |

Factorial Design adjusting for regression

| Source | df | SS | MS | F | $p$ |
|---|---|---|---|---|---|
| Diet A | 2 | 54.420 | 27.210 | 6.32 | .0149 |
| Week B | 1 | 40.708 | 40.708 | 9.45 | .0106 |
| A× B | 2 | 3.285 | 1.642 | 0.38 | .6916 |
| Error | 11 | 47.369 | 4.306 ✓ | - | - |
| Total | 17 | 145.782 | | | |



As for SSs, so means should be adjusted for regression/covariate

Unadjusted Means

| Diet | $B_1$ | $B_2$ | $\bar{Y}_i$ |
|---|---|---|---|
| $A_1$ | 9.667 | 16.00 | 12.833 |
| $A_2$ | 7.000 | 8.667 | 7.833 |
| $A_3$ | 11.000 | 14.667 | 12.833 |
| $\bar{Y}_j$ | 9.222 | 13.111 | |

Adjusted Means

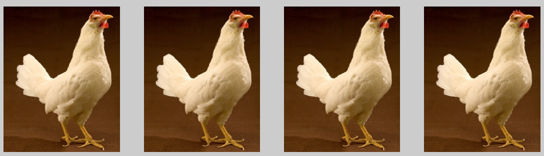
| Diet | $B_1$ | $B_2$ | $\bar{Y}_i$ |
|---|---|---|---|
| $A_1$ | 10.655 | 14.729 | 12.692 |
| $A_2$ | 7.706 | 9.655 | 8.681 |
| $A_3$ | 10.576 | 13.678 | 12.127 |
| $\bar{Y}_j$ | 9.646 | 12.687 | |

## Repeated Measures

Consider an experimental design with one factor $A$ with $a$ levels

Standard factorial design: One level of $A$ is assigned to any one subject/hen/... This is, the $a$ levels are assigned to $a$ different subjects/hens/...

Repeated measures design: All levels of $A$ are assigned to each subject/hen/... That is, the $a$ levels are assigned to the same subject/hen/...





## Repeated Measures

Consider an experimental design with one factor $A$ with $a$ levels

Standard factorial design: One level of $A$ is assigned to any one subject/hen/... This is, the $a$ levels are assigned to $a$ different subjects/hens/...

Repeated measures design: All levels of $A$ are assigned to each subject/hen/... That is, the $a$ levels are assigned to the same subject/hen/...

**The models and analyses differ** depending on design

Standard factorial design:

Total SS = (A)SS + Residual/Error SS

Repeated measures design:

Total SS = Between Subjects SS + Within Subjects SS

Total SS = Between Subjects SS + (A)SS + Error SS

Total SS = Between Subjects SS + (A)SS + A×Subjects SS + Error SS



## Repeated Measures - Two factors

Suppose there are two factors, $A$ with $a$ levels, and $B$ with $b$ levels

There are $a \times b$ combinations of $A$ and $B$ (for each replication)

Standard factorial design: One combination of $A$ and $B$ is assigned to any one subject/hen/... That is, the $a \times b$ combinations are assigned to $a \times b$ different subjects/hens/...

Repeated measures design: There are three possible designs –

- All combinations of $A$ and $B$ are assigned to each subject/hen/... That is, all $a \times b$ combinations are assigned to the same hen...
- One level of $A$ and all $b$ levels of $B$ are assigned to each subject/hen/... That is, all levels of $B$ are assigned to the same subject/hen/...but that hen only has one level of $A$; need $a$ hens

$A_1$

$A_2$



## Repeated Measures - Two factors

Suppose there are two factors, $A$ with $a$ levels, and $B$ with $b$ levels

There are $a \times b$ combinations of $A$ and $B$ (for each replication)

Standard factorial design: One combination of $A$ and $B$ is assigned to any one subject/hen/... That is, the $a \times b$ combinations are assigned to $a \times b$ different subjects/hens/...

Repeated measures design: There are three possible designs –

- All combinations of $A$ and $B$ are assigned to each subject/hen/... That is, all $a \times b$ combinations are assigned to the same subject/hen/...
- One level of $A$ and all $b$ levels of $B$ are assigned to each subject/hen/... That is, all levels of $B$ are assigned to the same subject/hen/...but that hen only has one level of $A$; need $a$ hens
- One level of $B$ and all $a$ levels of $A$ are assigned to each subject/hen/... That is, all levels of $A$ are assigned to the same subject/hen/...but that hen only has one level of $A$; need $b$ hens

**The models and analyses differ** depending on (design, effects,...)

## Example - Two-Factor Repeated Measures Design

Example: Diet - A (High, Medium, Low protein content) feed to hens. Two different time periods week - B ($B_1, B_2$). Interest in food intake = Y; three replications for each diet×week combination. Temperature = X was also measured. Now, same hen was used each week.

| | | B - Week | | | |
|---|---|---|---|---|---|
| *A* | | $B_1$ | | $B_2$ | |
| Diet | Hen | *X* | *Y* | *X* | *Y* |
| $A_1$ | 1 | 2 | 6 | 3 | 12 |
| High | 2 | 4 | 9 | 8 | 16 |
| | 3 | 10 | 14 | 13 | 20 |
| $A_2$ | 4 | 1 | 4 | 0 | 6 |
| Medium | 5 | 7 | 10 | 8 | 12 |
| | 6 | 9 | 7 | 8 | 8 |
| $A_3$ | 7 | 6 | 8 | 3 | 8 |
| Low | 8 | 7 | 12 | 9 | 16 |
| | 9 | 8 | 13 | 11 | 20 |



## Example - Two-Factor Repeated Measures

Factorial Design:

| Source | df | SS | MS | F | *p* |
|---|---|---|---|---|---|
| Diet A | 2 | 100.000 | 50.000 | 3.16 | .0791 |
| Week B | 1 | 68.056 | 68.056 | 4.30 | .0603 |
| A× B | 2 | 16.444 | 8.222 | 0.52 | .6077 |
| Error | 12 | 190.000 | 15.833 | | |
| Total | 17 | 374.500 | | | |

Repeated Measures Design:

| Source | df | SS | MS | F | *p* |
|---|---|---|---|---|---|
| Between Hens | 8 | 277.000 | | | |
| Diet A | 2 | 100.000 | 50.000 | 1.69 | .2609 |
| Hens(A) | 6 | 177.000 | 29.500 | | |
| Within Hens | 9 | 97.500 | | | |
| Week B | 1 | 68.056 | 68.056 | 31.41 | .0014 |
| A× B | 2 | 16.444 | 8.222 | 3.79 | .0861 |
| Error | 6 | 13.000 | 2.167 | | |
| Total | 17 | 374.500 | | | |



## Example - Two-Factor Repeated Measures - Covariate

Repeated Measures Design:

| Source | df | SS | MS | F | *p* |
|---|---|---|---|---|---|
| Between Hens | 8 | 277.000 | | | |
| Diet A | 2 | 100.000 | 50.000 | 1.69 | .2609 |
| Hens(A) | 6 | 177.000 | 29.500 | | |
| Within Hens | 9 | 97.500 | | | |
| Week B | 1 | 68.056 | 68.056 | 31.41 | .0014 |
| A× B | 2 | 16.444 | 8.222 | 3.79 | .0861 |
| Error | 6 | 13.000 | 2.167 | | |
| Total | 17 | 374.500 | | | |

Repeated Measures Design - Covariate:

| Source | df | SS | MS | F | *p* |
|---|---|---|---|---|---|
| Between Hens | | | | | |
| Diet A (adj)$^\dagger$ | 2 | 54.259 | 27.130 | 3.06 | **.1357** |
| Hens(A) (adj)$^\dagger$ | 5 | 44.370 | 8.974 | | |
| Within Hens | 9 | 97.500 | | | |
| Week B (adj)$^\dagger$ | 1 | 31.547 | 31.547 | 52.61 | **.0008** |
| A× B (adj)$^\dagger$ | 2 | 2.339 | 1.170 | 1.95 | **.2365** |
| Error | 5 | 2.998 | **0.600** | | |
| Total | 17 | 374.500 | | | |

$^\dagger$ Adjusted for presence of covariate



## Example - Two-Factor Design – Our data

Factorial Design:

No adjustment for Covariate

| Source | df | F | *p* |
|---|---|---|---|
| Diet A | 2 | 3.16 | .0791 |
| Week B | 1 | 4.30 | .0603 |
| A× B | 2 | 0.52 | .6077 |
| Error | 12 | | |
| Total | 17 | | |

$\hat{\sigma}^2 = 15.833$

Adjusted for Covariate

| Source | df | F | *p* |
|---|---|---|---|
| Diet A$^\dagger$ | 2 | 6.32 | .0149 |
| Week B$^\dagger$ | 1 | 9.45 | .0106 |
| A× B$^\dagger$ | 2 | 0.38 | .6916 |
| Error$^\dagger$ | 11 | | |
| Total | 17 | | |

$^\dagger$ Adjusted for presence of covariate

$\hat{\sigma}^2 = 4.306$

Repeated Measures Design:

No adjustment for Covariate

| Source | df | F | *p* |
|---|---|---|---|
| Between Hens | 8 | | |
| Diet A | 2 | 1.69 | .2609 |
| Hens(A) | 6 | | |
| Within Hens | 9 | | |
| Week B | 1 | 31.41 | .0014 |
| A× B | 2 | 3.79 | .0861 |
| Error | 6 | | |
| Total | 17 | | |

$\hat{\sigma}^2 = 2.167$

Adjusted for Covariate

| Source | df | F | *p* |
|---|---|---|---|
| Between Hens | 8 | | |
| Diet A$^\dagger$ | 2 | 3.06 | .1357 |
| Hens(A) (adj)$^\dagger$ | 5 | | |
| Within Hens | 9 | | |
| Week B (adj)$^\dagger$ | 1 | 52.61 | **.0008** |
| A× B (adj)$^\dagger$ | 2 | 1.95 | **.2365** |
| Error | 5 | | |
| Total | 17 | | |

$^\dagger$ Adjusted for presence of covariate

$\hat{\sigma}^2 = 0.600$

Conclusion:

Moral is: ..............

Regression:
Scientific errors: when omit indicator interaction terms
Philosophical errors: when using “tainted” variables such as rank

Analysis of variance:
Are all relevant factors included?
Is a factor fixed, random?
Are observations repeated measures?
Are there covariates present?



Hvala $\sim$ $\sim$ Thankyou