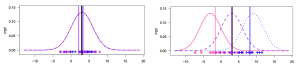
Experimental Design - controlled, can identify the effect of an action (gets causation)
observational studies cant account for confounding effects (can only get association)
i.e) confounder is smoking, observational study - does drinking lead to lung cancer
response variable = outcome being measured **(no treatment? => observational study)**
factor = explanatory var under the control of the experimenter (level = values of factor)
treatment = combination of experimental conditions applied to experimental unit (incl. control)
experimental unit = the unit on which the treatment is applied (i.e subject)
blocking variable = nuisance, variations/variable not interested in this study
randomization, placebo, blinding, control in experiments can eliminate bias & confounders
blocking and match-paired designs can reduce variation in the data

- the response variable: *presence or absence of a birth defect*
- the experimental units: *the women*
- the factor: *the dose level of folic acid is the factor*
- the levels of the factor: *there are two levels, 0mg and 0.8mg*
- number of treatments: *two*
- whether blocking was applied: *no*

Primer type	Application method: Dipping	Spraying
1	4.0, 4.5, 4.3	5.4, 4.9, 5.6
2	5.6, 4.9, 5.4	5.8, 6.1, 6.3
3	3.8, 3.7, 4.0	5.5, 5.0, 5.0

the response variable: *adhesive force*
the experimental units: *the aircraft wings*
the factors: *primer type and application method*
the levels of the factors: *two for application method, three for primer type*
number of treatments: *six*

1-way ANOVA - used to compare ≥3 groups (quant response, categorical factor/explanatory var)
If sample size is large or data is normal, better to use ANOVA than Kruskal-Wallis (also one factor only)
Assume: residuals (mean-point/within g) are normal, *obs within grp are normal*, *all grps have same variance*
obs within each grp & in diff grps are ind., groups are formed by randomly sampled from population
H0: all population means are the same
Ha: at least one population mean is different from the others



1. find overall sample mean: (μ1 + μ2 + μ3 + μ4) / 4
- taking **group mean** ($\bar{y}$) only works if u have **same # of replicates per group ($\bar{y}_i$ is mean of group i)**
- one-way ANOVAs are for single factor experiments
  - factor has **g levels/groups** (i)
  - each group has **t observations** (j)

$$\sum_{i=1}^{4}\sum_{j=1}^{5}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{4}\sum_{j=1}^{5}(y_{ij}-\bar{y}_{i\cdot})^2 + \sum_{i=1}^{4}\sum_{j=1}^{5}(\bar{y}_{i\cdot}-\bar{y})^2$$

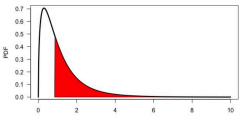
2. find sum of squares: total SS (var) around overall mean = within group SS + between group SS
- if mean rubber strength b/w machines **same -> diff (H0 -> Ha)**, total SS + between SS would increase
- **F statistic Ha > F statistic H0**

$$X \sim \chi^2_m \quad Y \sim \chi^2_n \quad \frac{X/m}{Y/n} \sim F_{m,n} \quad \text{m,n = dof}$$

$$\frac{\text{Within group SS}}{\sigma^2} \sim \chi^2_{g(t-1)}. \qquad \frac{\text{Between group SS}}{\sigma^2} \sim \chi^2_{g-1}.$$

3. test statistic = mean square ratio = SS/dof =
- test stat should follow $F_{3,16}$ dist. if H0 is true

$$\frac{\text{Between group SS}/(g-1)}{\text{Within group SS}/g(t-1)} \sim F_{g-1,g(t-1)}$$



Source	Sum of Squares	d.o.f.	MS	F
Between	6.82	3	2.273	
Within	43.02	16	2.689	$\frac{\text{Between group MS}}{\text{Within group MS}} = 0.845$
Total	49.84	19		

We reject H0 if:
- between SS > within SS
- F > $F_{α,g-1,n-g}$

p-value = area under the curve to the right of 0.845
MSE = within group SS

Multiple comparisons - posthoc method performing set of pairwise comparisons to see which grps differ
- test stat to compare two groups (where variance is MSE) => pairwise comparison

$$t = \left|\frac{\bar{y}_{l\cdot} - \bar{y}_{m\cdot}}{\sqrt{\frac{\text{MSE}}{n_l} + \frac{\text{MSE}}{n_m}}}\right| > t_{g(n-1)}$$

reject H0 if bigger than t-table val (see boneferroni for t-table val)
there are (n) = (4) = 6 possible pairwise comparisons to make within groups
(2) (2)

- **multiple comparisons problem**: each pairwise comparisons are not ind., chance of T1 error increases
- **Bonferroni's equality:** divide α by all possible pairwise comparisons => now the chance of making a T1 error is less than 5% (T1 error = rejecting H0 when true = alpha)
- corrected sig level: 0.05 / 6 = 0.00833 => two-sided alternative: 1 - 0.00833/2 = 0.9958 => *plug to t-table*

Colony (l,m)	$\bar{y}_{l\cdot} - \bar{y}_{m\cdot}$	t	Significant?
1,2	-29.4	6.22	Y
1,3	-32.2	6.81	Y
1,4	-30.6	6.47	Y
2,3	-2.8	0.59	N
2,4	-1.2	0.25	N
3,4	1.6	0.34	N

- t-table val (dof 16, 0.9958) = 3
- colony 1 differs from other colonies
- on average, colony 1 depart about 30 days earlier than the others
- note:

$$\text{Var}(\bar{y}_{l\cdot} - \bar{y}_{m\cdot}) = \frac{\sigma^2}{n_l} + \frac{\sigma^2}{n_m} \quad ESD(\bar{y}_{l\cdot} - \bar{y}_{m\cdot}) = \sqrt{\frac{\text{MSE}}{n_l} + \frac{\text{MSE}}{n_m}}$$

Contrasts - extending two sample t-test for general multi–level single factor
- i.e) want to compare group bio with level A, B and group non-bio with level C
  1. $\bar{y}_A + \bar{y}_B - 2\bar{y}_C = 0.46$ or $\frac{\bar{y}_A + \bar{y}_B}{2} - \bar{y}_C$
  2. $\begin{aligned}\text{Var}(\bar{y}_A + \bar{y}_B - 2\bar{y}_C) &= \text{Var}(\bar{y}_A) + \text{Var}(\bar{y}_B) + \text{Var}(2\bar{y}_C) \\ &= \text{Var}(\bar{y}_A) + \text{Var}(\bar{y}_B) + 2^2\text{Var}(\bar{y}_C) \\ &= \sigma^2/n + \sigma^2/n + 2^2(\sigma^2/n) \\ &= 6(\sigma^2/5)\end{aligned}$
  3. $t^* = (\bar{y}_A + \bar{y}_B - 2\bar{y}_C)/\sqrt{6(MSE/5)}$

- **H₀: A,B has no difference from C**
- use MSE for σ² => equiv to pooled variance
- pooled variance = sd1² + sd2² + sd3² / 3
- b/c MSE uses data from all group

- why use contrast? if we pooled data from 2 grps into 1, we can't account for small diff these grps have
- $c_a + c_b - 2c_c = 0$, this constraint on coeff/weights with grp mean (y) of resulting stat => **contrast**
- under H0, mean of ALL groups should be same, which means coeff/weights should sum to 0

Under H_0: $\sum_{i=1}^{g} c_i\mu_i = 0$ (where μ_i is the expected response in group i), standardized contrast follow a t distribution with n-g d.o.f. where n= total number of obs

- if n is not equal amongst groups, use this formula: this is step 2, so use c properly => **MSE = variance σ²**

$$\sqrt{MSE\sum_{i=1}^{g}\frac{c_i^2}{n_i}}$$

2-way ANOVA - used to look at the effect of 2 factors on a continuous response variable

$\hat{\sigma}^2$ = Within group mean square
- Also known as: Mean square error (MSE)
- $\text{MSE} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{t}(y_{ij} - \bar{y}_{i\cdot})^2}{g(t-1)}$

$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $\epsilon_{ij} \sim N(0, \sigma^2)$
- μ: overall mean $\hat{\mu} = \bar{y}$
- α_i: group (or treatment) effect for group i $\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}$
- ϵ_{ij}: independent error variables

- **main effect**: constant effect of one factor across all values the other factors

	A -	A +
B -	20	40
B +	30	50

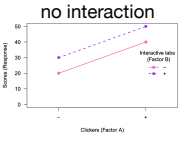
Main effect of A: $\bar{y}_{A+} - \bar{y}_{A-} = \frac{50+40}{2} - \frac{30+20}{2} = 20$

Main effect of B: $\bar{y}_{B+} - \bar{y}_{B-} = \frac{50+30}{2} - \frac{40+20}{2} = 10$

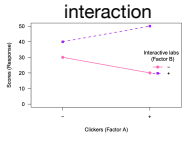
- **no interaction = slope is equal** (lines will never touch)
- when there is interaction, it makes no sense to talk about individual factor(s) or main effects (effect of A may not be 20 whatever level of B is)

Plotting the treatment means can help us identify potential interactions

no interaction



interaction



2-way ANOVA interaction — how do we know if the interaction is significant
- if there is no interaction => $\bar{y}_{22\cdot} - \bar{y}_{21\cdot} \approx \bar{y}_{12\cdot} - \bar{y}_{11\cdot}$.
- otherwise, interaction effect = $(\bar{y}_{22\cdot} - \bar{y}_{12\cdot}) - (\bar{y}_{21\cdot} - \bar{y}_{11\cdot})$

Prof	Clickers: No	Yes
Jones	68.52	79.83
Smith	69.42	65.58

Interaction SS: $\sum_{i,j,k}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y})^2$
- Degrees of freedom: $(c-1)(r-1)$

Error SS: $\sum_{i,j,k}(y_{ijk} - \bar{y}_{ij\cdot})^2$
- Degrees of freedom: $rc(m-1)$

Between Factor A SS: $\sum_{i,j,k}(\bar{y}_{i\cdot\cdot} - \bar{y})^2$
- Degrees of freedom: $c-1$

Between Factor B SS: $\sum_{i,j,k}(\bar{y}_{\cdot j\cdot} - \bar{y})^2$
- Degrees of freedom: $r-1$

Factor B \ Factor A	1	2	
1	$\bar{y}_{11\cdot}$	$\bar{y}_{21\cdot}$	$\bar{y}_{1\cdot\cdot}$
2	$\bar{y}_{12\cdot}$	$\bar{y}_{22\cdot}$	$\bar{y}_{2\cdot\cdot}$
	$\bar{y}_{1\cdot\cdot}$	$\bar{y}_{2\cdot\cdot}$	$\bar{y}$

Source	SS	dof	MS	F	
Clickers	139.5	1	139.5	19.69	<- MSC/MSE
Prof	445.6	1	445.6	62.88	<- MSP/MSE
Interaction	573.8	1	573.8	80.97	<- MSI/MSE
Error	255.1	36	7.1		
Total		39			

- for ex) MSI = $\frac{\sum_{i,j,k}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y})^2}{(r-1)(c-1)}$
- r = 2 in A;; c = 2 in B;; m =10 students
- test stat should be compared to $F_{1,36}$ which is $F_{\text{dof interaction, dof error}}$

f-val > f-table to reject null hypothesis = significant
- if A,B,AB are significant (Y = A + B + AB + ε)
- remove interaction term if not significant (Y = A + B + ε) => ε~N(0, MSE²)
- you cannot remove the main effect in the presence of interactions
- Y = A + B + AB + e is the model expression for a two-way ANOVA, which is used to test for differences in group means between A and B and to determine if the interaction effect is significant. Specifically, it tests whether the group mean differences in A depend on the levels of grp B

Two-way ANOVA divides variability into 4 components:
- Interactions between row and column
  - Differences between rows that are not the same at each column vice versa)
- Variability among columns
- Variability among rows
- Residual or error
  - Variation among replicates not related to systematic differences between rows and columns

If the interaction term is significant:
- We keep the interaction term in the model
- Interpretation of the main effects becomes difficult
  - Since the effect of one factor depends on the level of the other factor

We are often interested in the main effects
- The interpretations of the main effects are simpler in the absence of interaction term
  - Thus, if the interaction term is not significant, we should remove it from the ANOVA model

It is not meaningful to discuss main effects in the presence of interactions

The observed test statistic for the interaction term, and both factors are larger than the crictial value, as such all terms are significant at the 5% level. Hence we would adopt an interaction model of the form

$$Y = A + B + AB + \varepsilon,$$

where A is instructor effect, B is the clicker effect. AB is the interaction term, and ε is the error term.

Further Design - randomized complete block design (RCBD): designing experiments when there's a nuisance factor, analyze data with blocking factor
RBCD: *each block* (i.e body parts) have the *same number of experimental units* and a *complete set of treatments* (i.e scanner type) => 2 lung on scanA,B each, 2 brain on scanA,B each… *assign treatment to each unit of a block randomly*
blocking: method for removing extraneous sources of variation
block by matching: matching similar experimental units into groups (i.e subjects, twins, factory where the units are made) => blocks are the group of similar units
block by reusing: use the same unit multiple times (i.e give a subject different drugs at different times) => unit/subject becomes the block; time slot for that unit becomes the experimental unit
block by subdividing: dividing a large block of land into smaller subplot

Use scanner A for brain patients, scanner B for lung patients, etc
- confounding the effect of scanner type with the body part being scanned

ANOVA on **scanners** as factor => ANOVA on **scanners, body** parts as factors:
- scanners SS and overall variability in the response variable **doesn't change**
- **residuals (SSE) decreases** bc we explain some of total variability with the blocking factor (i.e. body part) so there's less var attributed to the error
- since MSE affects f-val, f-val will change as well

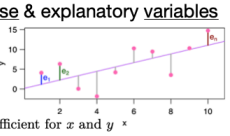
One-way ANOVA:
- Overall variability = Between group SS + Within group SS
- Overall variability = Scanner SS + Error SS

Two-way ANOVA
- Overall variability = Column SS + Row SS + Error SS
- Overall variability = Scanner SS + Body part SS + Error SS

Including blocking variable in analysis accounted for some variability, and thus helped properly measure the variability attributed to the factor of interest (speed)

Confounding: situation where effect of a factor can't be separated from another
Regression - describe relationship b/w contin. response & explanatory variables

Linear model $y = \beta_0 + \beta_1 x + \varepsilon$
Regression line $y = b_0 + b_1 x$
Predicted/fitted val $\hat{y} = b_0 + b_1 x$
Error term $\varepsilon \sim N(0, \sigma^2) = y - \hat{y}$



y: response variable
x: explanatory variable
$\bar{y}$: sample mean of y
$\bar{x}$: sample mean of x
r: sample correlation coefficient for x and y
s_Y: sample standard deviation of y
s_X: sample standard deviation of x
e_i: **vertical** distance from the point to the line fitted
- point above the line → e_i is positive

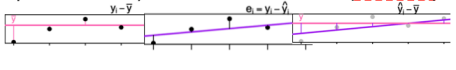
b_0 is an estimate of the population intercept, β_0
b_1 is an estimate of the population slope, β_1

Slope: $b_1 = r\frac{s_Y}{s_X}$ Intercept: $b_0 = \bar{y} - b_1\bar{x}$

$$b_1 = \frac{rs_Y}{s_X} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

variance of residuals $s^2 = \frac{\sum_i e_i^2}{n-2}$
- regression line of X on Y != Y on X
- point $(\bar{x}, \bar{y})$ will always lie on the regression line
- gradient/slope/reg coeff (b): increasing X by 1 unit inc. Y by b_1 (rate of change)
- **sum of residuals**/err $\sum_{i=1}^{n} e_i$ should be **close to 0** for **real values x, y**
- **sum of residuals**/err is **always 0** in a regression line/**model x, y_hat**
- **regression line *minimizes the sum of the squares of errors*** (to find params)
- we use the err² bc its always + and sum of residuals in a model is always 0
- var. of res. has n-2 division, bc we have 2 params/unknown: slope, intercept we lose a degree of freedom for each parameter estimated
- R² =% of the variation (in response variable) explained by the model (usefulnes)

$$R^2 = \frac{\sum_{i=1}^{7}(y_i - \bar{y})^2 - \sum_{i=1}^{7} e_i^2}{\sum_{i=1}^{7}(y_i - \bar{y})^2}$$



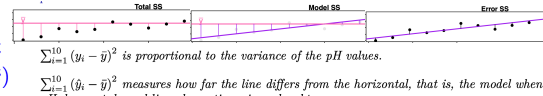
Analysis of Variance (F-test) to test: does relationship exist or no?
H_0: $\beta_1 = 0$ no relation
H_a: $\beta_1 \neq 0$ relationship

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

SS Total = SS Model + SS Error
SS about mean = SS due to model + SS about model

$$\begin{aligned}\hat{y}_i &= b_0 + b_1 x_i \\ &= \bar{y} - b_1\bar{x} + b_1 x_i \\ &= \bar{y} + b_1(x_i - \bar{x})\end{aligned}$$

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$$



$\sum_{i=1}^{10}(y_i - \bar{y})^2$ *is proportional to the variance of the pH values.*

$\sum_{i=1}^{10}(\hat{y}_i - \bar{y})^2$ *measures how far the line differs from the horizontal, that is, the model when pH does not depend linearly on time since slaughter.*

$\sum_{i=1}^{10}(y_i - \hat{y}_i)^2$ *measures variation in the pH levels about the fitted line.*

If Y didn't depend on X, plot scatters about mean y_bar & only var is due to noise (b_0 ~= noise. b_1 ~= 0)
- if **relationship exists**: model/regression SS > error SS
- if **relationship DNE**: model/regression SS < error SS
- assumption ε ~ N(0, σ²) works well if n is big

Under H_0, $F \sim F_{1,n-2}$

Under H_a, $F > F_{1,n-2}$ reject H_0, test stat falls in far upper tail of F-dist
- i.e) **β_1 != 0, relationship exists, MSM/MSE to be > 1**

Source	DoF	SS	MS	F
Model	1	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	MSM	MSM/MSE
Error	$n-2$	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	MSE	
Total	$n-1$	$\sum_{i=1}^n (y_i - \bar{y})^2$		

Regression Estimators - quantify uncertainty around a parameter estimate

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

- params are β_0, β_1, σ
- estimators are all unbiased
- estimators b_0, b_1 and σ^2 are all random variables (fxns of the data which are subject to random error according to the model)

$$\beta_0\text{: } b_0 = \bar{y} - b_1\bar{x}$$
$$\beta_1\text{: } b_1 = \frac{rS_Y}{S_X}$$
$$\sigma^2\text{: } s^2 = \frac{\sum_i e_i^2}{n-2}$$

★ s^2 is the MSE

If $\mathrm{Var}(Y_i|X = x_i) = \sigma^2$ for all i, it can be shown that

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

- Var(b_1) is inversely proportional to σ²
- more spread out x (bigger x), lower the variance of the estimate of the slope
- **predictor, explanatory**, or independent variables = **x**
- **response, observable**, or dependent variables = **y**

$$\mathrm{Var}(b_0) = \frac{\sigma^2 \sum x_i^2}{n\sum_{i=1}^n (x_i - \bar{x})^2}$$

- to reduce Var(b_1), we may want a large range of x values (more spread out)
- replace known variance σ² with estimated variance s² => formula for sd(b_1) => ***b_1 = estimate***

CI: *estimate* $\pm t^*_{n-2} \times se$(estimate)

$b_1 = 0.0731$

$t^*_{n-2} = 2.01$

$$se(b_1) = \sqrt{\frac{s^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$s^2 = \frac{\sum_i e_i^2}{n-2} = \mathrm{MSE} = 61.47$

$$S_X = 46.45 = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{49}}$$

$\sum_{i=1}^n (x_i - \bar{x})^2 = 49 \times 46.45^2$

- we know sd/Sx, we know Sx formula => simple algebra to find (xi - x-)

- **Confidence interval** for β_1= $0.0731 \pm 2.01 \times \sqrt{\frac{61.47}{49 \times 46.45^2}}$

$b_1 \pm t_{n-2,\alpha/2} \times \mathrm{se}(b_1)$

- **if 0 is in CI, then we accept H_0**

ANOVA test for regression
- H_0: $\beta_1 = 0$
- H_1: $\beta_1 \neq 0$
- Test statistic: $F = MSM/MSE \sim F_{1,n-2}$

Altenative test (t-test)
- H_0: $\beta_1 = 0$
- Test statistic: $t = \frac{b_1}{\mathrm{se}(b_1)} \sim t_{n-2}$

H_a: $\beta_1 \neq 0$ $\rightarrow t^*_{48,0.975} = 2.01$
H_a: $\beta_1 > 0$ $\rightarrow t^*_{48,0.95} = 1.68$
H_a: $\beta_1 < 0$ $\rightarrow t^*_{48,0.05} = -1.68$

- one–sided: H_a : $\beta_1 < 0$ or H_a : $\beta_1 > 0$
- two–sided: H_a : β_1 != 0

ANOVA test (f-val) OR t-test
- t^2 = F bc a random variable t_m^2 is a random variable $F_{1,m}$
- **For H_a: β != 0 t-test and F-test are equivalent**
  (f-test is always two-tailed bc it accounts for sq vals only)
  (f-test is always positive, so it can't be equivalent to 1-tail)

Parameter estimates $\hat{\beta}_0 = b_0$ and $\hat{\beta}_1 = b_1$ and their standard errors $\mathrm{se}(b_0)$ and $\mathrm{se}(b_1)$ can be used for inference

Testing $H_0 : \beta_1 = 0$ *vs* $H_a : \beta_1 \neq 0$
- $t = b_1/se(b_1)$

Why should we not test for H_0: intercept/β_0 = 0?

Predicting the response for a particular value of the explanatory variable
- Here, predicting the number of goals for a given number of shots

For a given x^* value, do we want to predict
- The mean response?
  - ★ Here, $\hat{\mu}_{y}$, the mean number of goals for players taking x^* shots
- The response for an individual?
  - ★ Here, $\hat{y}$, the number of goals for a single player that had x^* shots

For both, the point estimate is: $b_0 + b_1 x^*$
But CIs are different because se() are different
- $se(\hat{\mu}_y) < se(\hat{y})$

In testing $H_0 : \beta_1 = 0$ we are investigating whether it is plausible that the intercept of the model is zero. In this case the null hypothesis would imply that a player who does not take any shots scores no goals. Of course that hypothesis must hold, and it would be plausible to assess whether our model concurs with realism. A problem here though is that by looking at just the top scorers in the NHL we have no data close to the origin. More generally, how the model behaves near the origin may have no physical meaning and the associated hypothesis test is of no interest.

When presenting results
- Present the standard error or confidence interval along with the point estimate of a parameter

Rejecting H_0 does not imply the model fits well
- e.g., could be non-linear relationship

ANOVA		Regression
M1	μ_1	$\beta_0 + \beta_1$
M2	μ_2	$\beta_0 + \beta_2$
M3	μ_3	$\beta_0 + \beta_3$
M4	μ_4	β_0

- point estimate for mean and point y_hat will be same but confidence interval will have different width bc sd(y_hat) > sd(mean) => extra variability in a single response compared to mean (mean doesn't have ε)

Multiple Linear Regressions - more than one predictor variable (x)
- Visualizing relationship b/w Y and (X1, X2) is difficult, especially if X1 and X2 has correlations/interaction
- **minimize least squares** pink = y_i_hat

simple linear regression $\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$

multiple linear regression $\sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}))^2$

- what is an interpretation of estimate β_2? We estimate that for day with a common POL, the average number of EMV increases by β_2 units for every one increase in TMP $ERV = \beta_0 + \beta_1 POL + \beta_2 TMP + \epsilon$.
- β_1 CI (POL) has lower endpoint is positive => likely high POL causes more ERV
  β_2 CI (TMP) has lower endpoint is positive but narrow => less likely high TMP causes more ERV
- **β CI includes 0 => no evidence that there exists a relationship between Y and X (accept H_0)**
  => good chance that the true population value is close to 0

Residual SS $\sum_{i=1}^n e_i^2$

Residual MS / MSE (estimate of σ²) $\frac{\sum_{i=1}^n e_i^2}{n-(p+1)}$

Multiple R^2 $R^2 = 1 - \frac{\text{Residual SS}}{\text{Total SS}}$ % of variation in response variable accounted for by the explanatory variables

- when >1 predictor variable, we cannot use scatterplots because there are >2 variables to plot together
- relationship b/w y and x1 may change if we add x2 in the model (i.e scatterplot of y and x1 may not show true relationship in a multi regression model because **x1 and x2 could be correlated**)
- interpretation of params: β_1 is the change in Y when X1 is changed by 1 unit (***while other predictors fixed***)

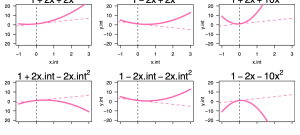
Curve Fitting in Regression - using the multiple linear regression machinery to fit curves to data

$$b_1 = \frac{dy}{dx} = \frac{db_0 + b_1 x}{dx}$$

- b1 is slope = rate of change
- **adding an explanatory variables (X) will decrease RSS**
  - **exceptions where RSS will be same** (these cannot be fitted in practice):
    1. If $y = \beta_0 + \beta_1 x_1$ with no error (i.e. perfect fit, RSS will be 0, so cannot be lower)
    2. If x_1 and x_2 are perfectly correlated
- **increasing power of regression line => will never increase residual (will always decrease or stay the same)**

$y = b_0 + b_1 x + b_2 x^2$
- **we can't say**, b1 is change in Y when x is changed by 1 unit while we hold x^2
- **instead,** change in Y is $b_1 + 2b_2 x$ when x is changed by 1 unit

For any value c, $\hat{\beta}_1 + 2\hat{\beta}_2 c$ is the estimated rate of change in the conditional mean of Y given X as X increases from $X = c$.



The model
$$Y = \beta_0 + \beta_1 X + \epsilon$$
postulates a **linear relationship** between X and Y, whereas
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$
describes a **quadratic relationship**.

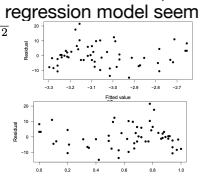
But: statisticians consider **both** these models as **linear models**
- Because in both cases the expected value of Y given X is modeled as a linear function of the **parameters**, the regression coefficients
- Can use linear model software to fit both models

polynomial models with order q $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_q X^q + \epsilon$
Which polynomial order model is the most consistent with the data?
1. Set $q = 1$.
2. Fit model q.
3. If the 95% CI for β_q includes zero then stop, and output model $q-1$ as the answer. Otherwise, increase q by one, and go to Step 2.

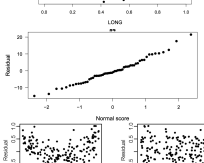
- note that this algorithm can give us q=0 for the answer ("intercept-only")
- methods for multiple regression can be used in polynomial regression
- **limitation** for a polynomial model: it may be unstable if the order is higher than 2 (i.e relation may change if diff samples are taken from the same population)

All linear models are empirical models - may be useful to describe the observed data, but may not reflect true relationship b/w predictors & response => predictions outside observed data range can be dangerous (i.e extrapolation)

Residuals in Regression - check if assump. of model is reasonable for data in a linear regression, we assume (residual plots + normal QQ plots)
- mean of the response is a **linear** function of the predictors
- errors are **independent** and **normal** RV with mean 0 and **constant variance**
- look at residuals (estimated errors) to assess if conditions of a multiple linear regression model seem to hold for our data



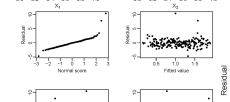
- **residuals (obs - fit) vs fitted values (y_hat)**
- we want patternless "band" centred around 0
- pattern (i.e funnel) suggest inadequate model
- inadeq? may want to change response variable



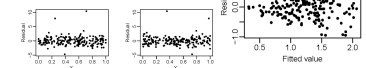
- **residuals vs predictor variables (x)**
- we want patternless "cloud" centred around 0
- only accounts for 1 var, not generalizable

- normal score plots of residuals
- we want an approximately straight line thru origin
- check departures from normality (skewnes, heavy tails)

- data w/ X_1,X_2
- X_1 residuals is nonlinear, add X_1^2



- same, but there's 3 outliers => remove outliers



- same, but resid vs fit has funnel => try transformation (i.e sq of y) as a remedy when the variation in residuals is inc/dec w/ fitted vals

- residual = whats left after model is removed => should resemble noise
- H_0: residual is normal w/ σ^2 that does not change w/ the explanatory variable

$\log Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

Implies
- $Y = e^{\beta_0} e^{\beta_1 X_1} e^{\beta_2 X_2} e^{\epsilon}$

1 unit increase in X_1
- ★ (for a constant X_2 value)

Results in
- ★ $e^{\beta_1} = e^{-0.0273} = 0.973$
- ★ So a $(0.973 - 1) \times 100 = -2.7\%$ rate of change

Residual plots allow us to check:
- **goodness-of-fit**: does the model fit the data well?
- **constant variance**: is the variance of y constant?
- **outliers**: are there any outliers?
- **transformations**: any transformations needed for y and x_j's?

Normal scores plots allow us to check:
- Whether ϵ follow Normal distributions or not
- A near straight line shows that the Normality assumption may be OK

Dummy variables -regressions on qualitative variables (categories/groups as X)

Machine	M1	M2	M3	M4
	18.3	15.9	19.9	15.1
	16.9	19.2	15.7	16.2
	15.8	17.7	17.8	17.5
	18.6	15.4	18.9	20.1
	17.5	16.4	20.3	18.3
Mean:	17.42	16.92	18.52	17.44
s.d.:	1.126	1.535	1.850	1.926

	x1	x2	x3	y	
[1,]	1	0	0	18.3	← M1
⋮					
[6,]	0	1	0	15.9	← M2
⋮					
[16,]	0	0	0	15.1	← M4

ANOVA or regression both works

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$.

ANOVA's H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

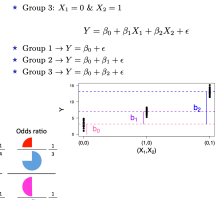
Multiple regression H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

i.e) H_0: Y = β₀ + ε (no linear relation)

choose M4 as base/reference group (0,0,0)
X1,X2,X3 are dummy vars

General model can be written: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$
- X1 (elevation, contin.), X2 (direction, 1/0) are two factors (X2 has 2 levels)
- there are 2 completely different lines (dir 1/0) with diff slopes and intercepts
- β_3 measures interaction

Assumptions for ANOVA and regression are similar

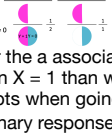


Odds ratio - variables X: 0/1, Y:0/1 => (a+b+c+d=n)

	Y = 0	Y = 1
X = 0	a	b
X = 1	c	d

risk = prob of event
odds = p/(1-p)

where p is prob of success=> 1 fail = odds# success
Risk ratio (RR), Odds ratio (OR) measures associatio
b/w X and Y => **OR=1, no association**



$RR = \frac{P(Y=1|X=1)}{P(Y=1|X=0)}$

$OR = \frac{P(Y=1|X=1)/P(Y=0|X=1)}{P(Y=1|X=0)/P(Y=0|X=0)} = \frac{P(Y=1|X=1)/(1-P(Y=1|X=1))}{P(Y=1|X=0)/(1-P(Y=1|X=0))}$

0 < OR < inf, further away OR is from 1, stronger the a association b/w X and Y
OR = 16, Odds of Y = 1 are 16 times higher when X = 1 than when X = 0
RR = increase in proportion of successful attempts when going X=0 to X=1

Logistic Regression - linear regression for binary response variable Y~bin

probability of Y=1 given X:
logit(PR(Y=1|X)) = β_0 + β_1X
PR(Y=1|X) = expit(β_0 + β_1X)
logit(p) = log(p/1-p) = log(Odds)

if X is also binary:
log(Odds(Pr(Y=1|X)) = β_0 + β_1X
Odds(Y=1|X) = $e^{\beta 0 + \beta 1 X}$

$\frac{\mathrm{Odds}(Y=1|X=1)}{\mathrm{Odds}(Y=1|X=0)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$

- odds of getting cancer (Y=1) when you smoke (X = 1) is $e^{\beta 1}$ the odds of getting cancer if you don't smoke (X = 0)

$\frac{\mathrm{Odds}(Y=1|X=c+\Delta)}{\mathrm{Odds}(Y=1|X=c)} = \frac{e^{\beta_0+\beta_1(c+\Delta)}}{e^{\beta_0+\beta_1 c}} = e^{\beta_1 \Delta}$

- $e^{\beta 1}$ is the odds ratio when Δ=1
- continuous: $e^{\beta 1 \Delta}$ is the odds ratio corresponding to an unit change in X
- multi explanatory variables?

$\mathrm{logit}\,(Pr(Y = 1|X_1, X_2)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad Pr(Y = 1|X_1, X_2) = \mathrm{expit}(\beta_0 + \beta_1 X_1 + \beta_2 X_2)$

- H_0: $\beta_2 = 0$ (linear relationship)
- H_a: β_2 != 0 (quadratic relationship)

which model does CI include 0?

- $expit(\beta_0)$ *is the chance that* $Y = 1$ *if* $X_1 = 0$ *and* $X_2 = 0$.
- e^{β_1} *is the multiplicative change in the odds of* $Y = 1$ *associated with a unit increase in* X_1*, when* X_2 *is held fixed.*
- e^{β_2} *is the multiplicative change in the odds of* $Y = 1$ *associated with a unit increase in* X_2 *when* X_1 *is held fixed.*
- *Echoing a point we stressed for multiple linear regression, the "held fixed" is important when interpreting the coefficient associated with explanatory variables.*