Föreläsning 14: Försöksplanering

Matematisk statistik

David Bolin
Chalmers University of Technology
Oktober 14, 2015

Modellbeskrivning

Vi har gjort mätningar av en responsvariabel Y för fixerade värden på förklarande variabler $x_1, \ldots, x_p$, som kan väljas utan fel.
Vi ansätter en linjär modell för $(Y_i, x_{1,i}, \ldots, x_{p,i}), i = 1, \ldots, n$:

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_p x_{p,i} + \varepsilon_i \tag{1}$$

där ε_i är oberoende $\mathsf{N}(0, \sigma^2)$ slumpvariabeler som beskriver mätfelet och β_i är parametrar som beskriver beroendet.
Vi kan formulera modellen på matrisform som

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$$

där

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{p,1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \cdots & x_{p,n} \end{pmatrix} \quad \mathbf{E} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$$

Parameterskattning

Sats

Givet att $\mathbf{X}^T\mathbf{X}$ är inverterbar fås minstakvadrat-skattningen av $\boldsymbol{\beta}$ som $\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Skattningarna är väntevärdesriktiga och vi har

$$V(\beta_i^*) = \sigma^2(\text{diagonalelement nummer i+1 i } (\mathbf{X}^T\mathbf{X})^{-1}).$$

Vidare är $s^2 = Q_0/(n-p-1)$ där

$$Q_0 = \sum_{i=1}^{n}(y_i - \beta_0^* - \beta_1^* x_{1,i} - \ldots - \beta_p^* x_{p,i})^2 = \mathbf{Y}^T\mathbf{Y} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y}.$$

Sats

Med $\mu_Y^*(\mathbf{x_0}) = \beta_0^* + \beta_1^* x_{0,1} + \cdots + \beta_p x_{0,p}$ är
$\mathrm{V}(\mu_Y^*(\mathbf{x_0})) = \sigma^2 \mathbf{x_0}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0$

Egenskaper hos skattningarna

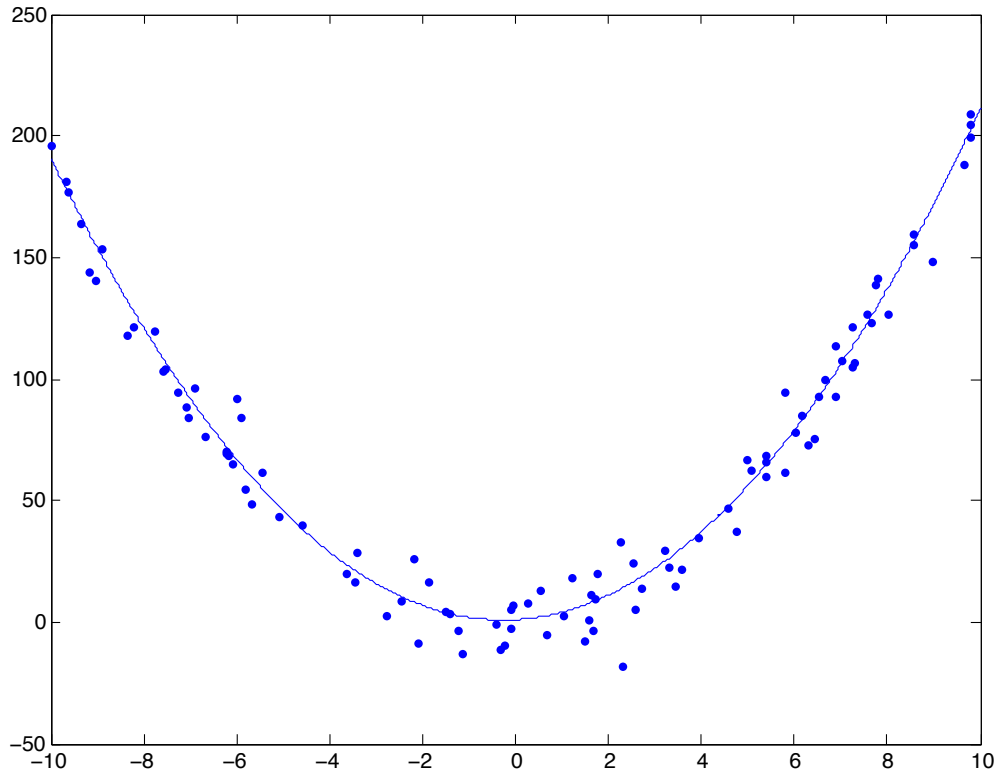
- Om ε_i är normalfördelade är också β_i^* normalfördelade.
- $(n-p-1)s^2/\sigma^2 \sim \chi^2(n-p-1)$.
- med $\theta = \beta_i$ eller $\mu_Y(\mathbf{x}_0)$ gäller $(\theta^* - \theta)/d(\theta^*) \sim t(n-p-1)$.
- Ett konfidensintervall för θ är

$$I_\theta = [\theta^* \pm t_{\alpha/2}(n-p-1)d(\theta^*)]$$

- För att testa $H_0 : \theta = \theta_0$ mot $H_1 : \theta \neq \theta_0$ används teststorheten $T = (\theta^* - \theta_0)/d(\theta^*)$ som har kritiskt område $C_\alpha = \{T : |T| > t_{\alpha/2}(n-p-1)\}$.
- Ett prediktionsintervall för en framtida observation $Y(\mathbf{x}_0)$ ges av

$$I_{Y(\mathbf{x}_0)} = \left[\mu_Y^*(\mathbf{x}_0) \pm t_{\alpha/2}(n-p-1) \cdot s\sqrt{1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}\right]$$

Polynomregression



- Vi har gjort mätningar av en responsvariabel Y för fixerade värden på en förklarande variabel x. Ansätt modellen

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \varepsilon_i \qquad (2)$$

- Specialfall av multipel regerssion med $x_k = x^k$.

Förklaringsgrad och modellval

- För en given modell har vi

$$Q_0 = \mathbf{Y}^T\mathbf{Y} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} = S_{yy} - \text{SSR}$$

- Alltså kan den totala variationen i datan, S_{yy} delas upp i en del som förklaras av linjen, SSR, och en del som är residualvariationen: $S_{yy} = Q_0 + \text{SSR}$
- Andelen av variationen som förklaras av modellen brukar betecknas $R^2 = \frac{\text{SSR}}{S_{yy}}$.
- En viktig fråga är hur vi väljer vilka förklarande variabler som ska ingå i en multipel regressionsmodell.
- För att undersöka om x_p bör tas med kan vi utföra hypotestest

$$H_0 : \beta_p = 0$$
$$H_1 : \beta_p \neq 0$$

genom att bilda teststorheten $\beta_p^*/d(\beta_p^*)$ med kritiskt område $C_\alpha = \{T : |T| > t_{\alpha/2}(n-p-1)\}$.

Stegvis regression (I)

Låt $(\beta_0, \beta_i, \beta_j)$ beteckna modellen $Y = \beta_0 + \beta_i x_i + \beta_j x_j + \varepsilon$ och låt $Q_0(\beta_0, \beta_i, \beta_j)$ vara residualvariationen för modellen.

Steg 1
Välj (β_0, β_{i_1}) där $i_1 = \arg\min_i Q_0(\beta_0, \beta_i)$. Om β_{i_1} är signifikant, gå till Steg 2a, annars välj (β_0) som modell och avsluta modellval.

Steg 2a
Välj $(\beta_0, \beta_{i_1}, \beta_{i_2})$ där $i_2 = \arg\min_i Q_0(\beta_0, \beta_{i_1}, \beta_i)$. Om β_{i_2} är signifikant, gå till Steg 2b, annars välj (β_0, β_{i_1}) som modell och avsluta modellval.

Steg 2b
Testa om β_{i_1} är signifikant i $(\beta_0, \beta_{i_1}, \beta_{i_2})$. Behåll modellen om så är fallet, annars välj (β_0, β_{i_2}). Gå till steg 3a.

Stegvis regression (II)

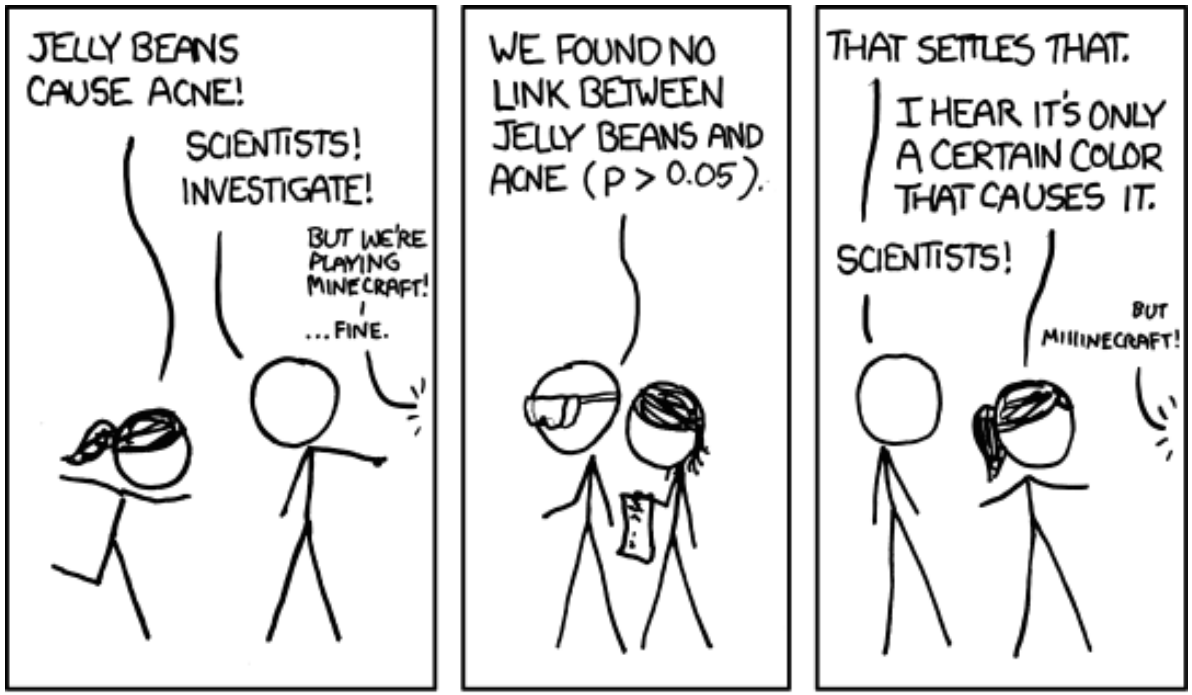
Fortsätt iterera stegen i modellvalet:

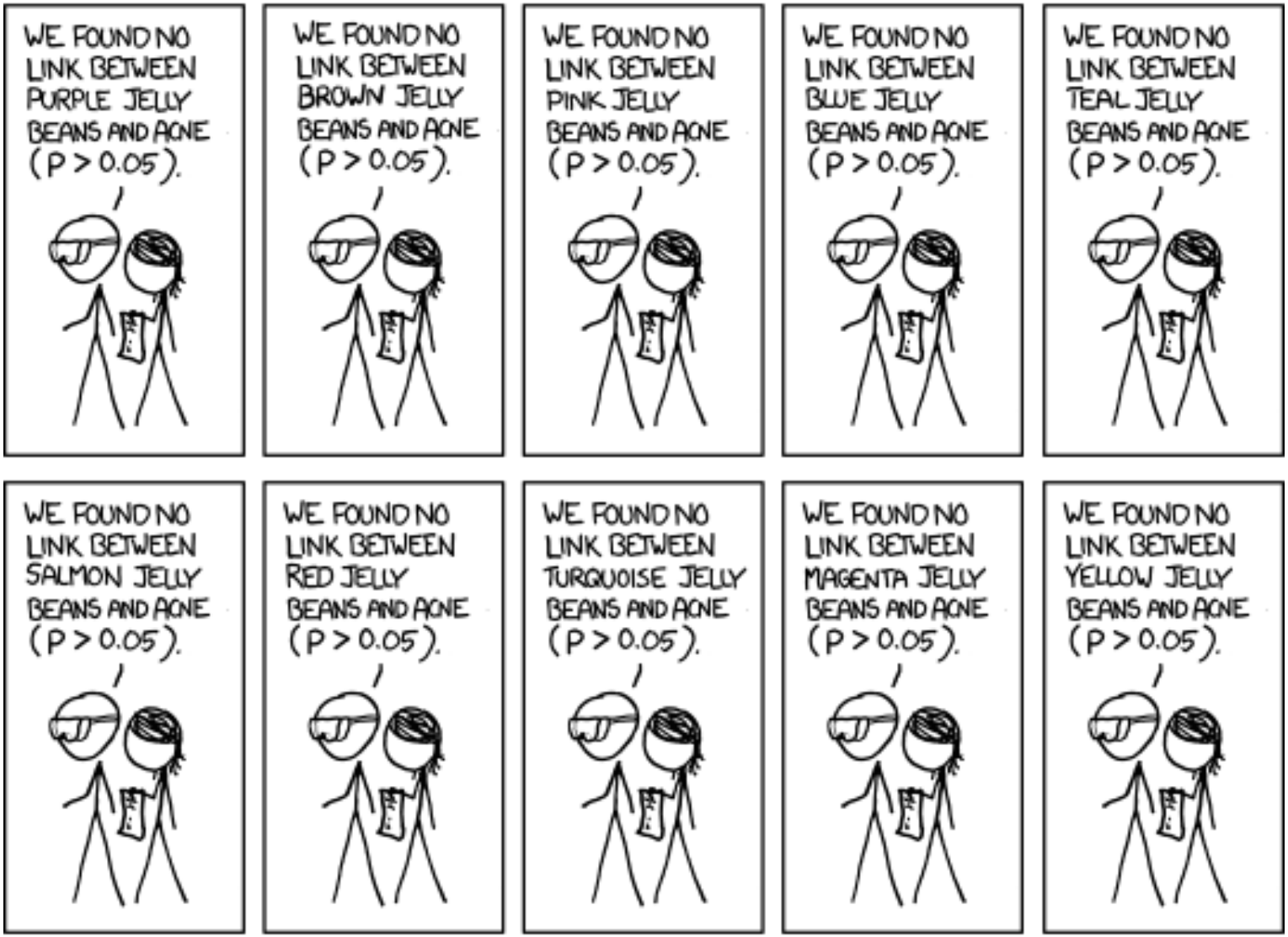
Steg m:a
Aktuell modell är $(\beta_0, \beta_{i_1}, \ldots, \beta_{i_k})$. Välj $(\beta_0, \beta_{i_1}, \ldots, \beta_{i_k}, \beta_{i_{k+1}})$ där $i_{k+1} = \arg\min_i Q_0(\beta_0, \beta_{i_1}, \ldots, \beta_{i_k}, \beta_i)$. Om $\beta_{i_{k+1}}$ är signifikant, gå till Steg m:a, annars välj $(\beta_0, \beta_{i_1}, \ldots, \beta_{i_k})$ som modell och avsluta modellval.

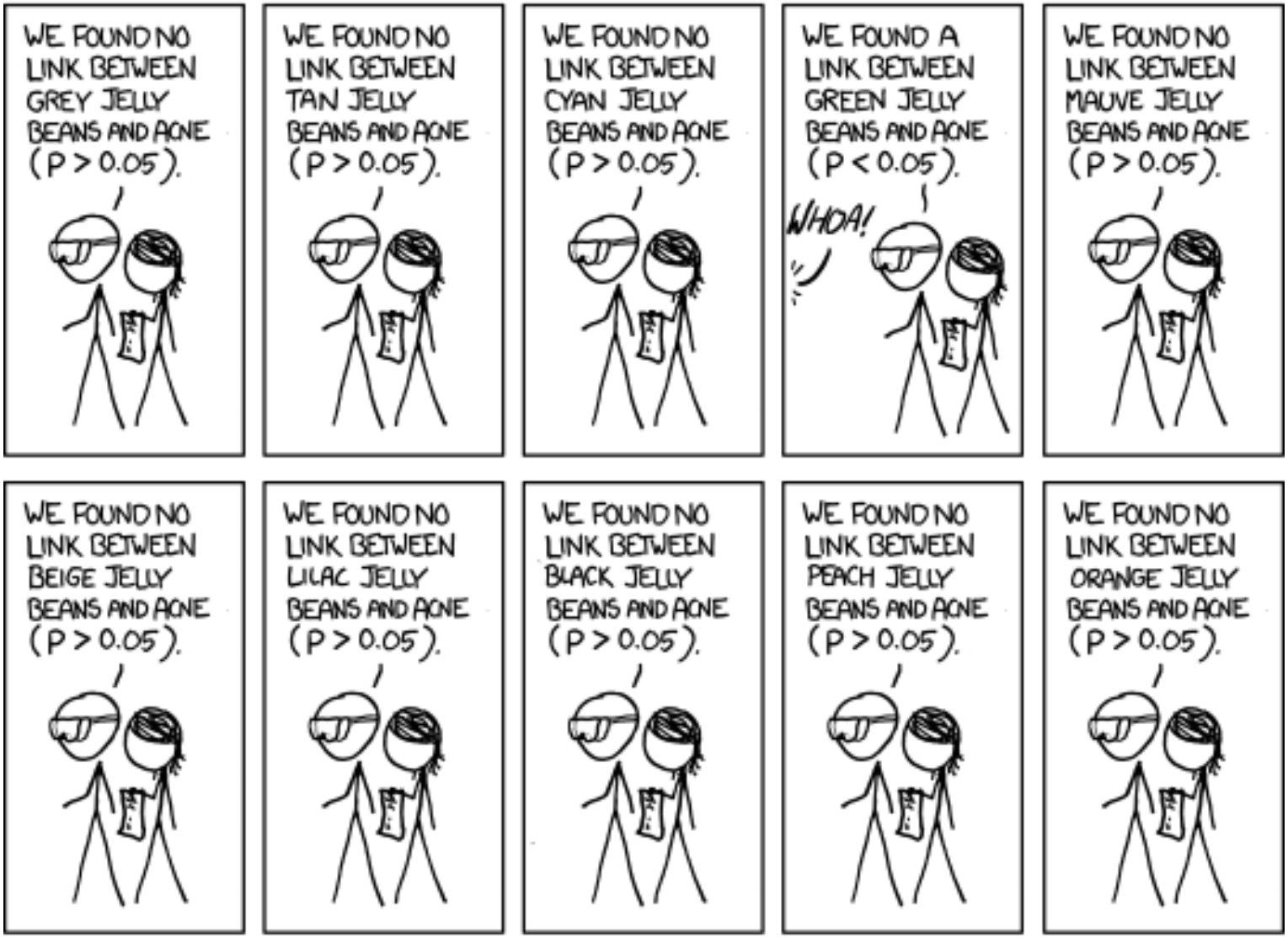
Steg m:b
Testa om någon parameter β_j är icke-signifikant i $(\beta_0, \beta_{i_1}, \ldots, \beta_{i_k}, \beta_{i_{k+1}})$. Om så är fallet, välj bort den med lägst värde på teststorheten, annars behåll alla parametrar. Gå till Steg $(m+1) : a$.

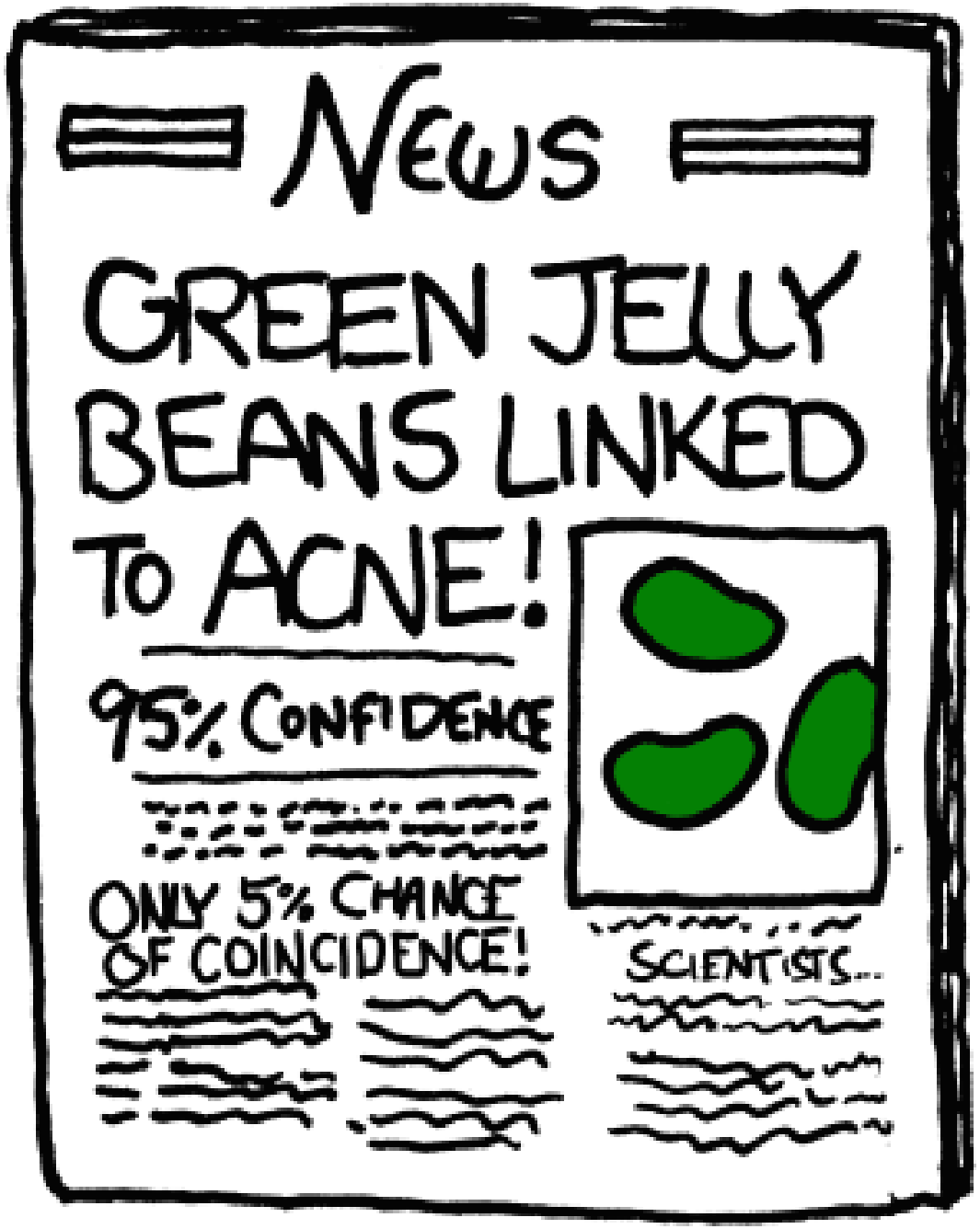
Massignifikans

- En invändning mot stegvis regression är att det inte alltid går att komma fram till en bra modell genom att endast lägga till en förklarande variabel i taget.
- En annan invändning är att valet av α i testerna i viss mån godtyckligt eftersom vi gör upprepade tester.
- Detta brukar kallas *massignifikansproblemet*:




WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Statistiska undersökningar

Vi kan dela in statistiska undersökningar i två huvudkategorier:

- Deskriptiva: Syftar till att beskriva egenskaper hos någon population.
- Analytiska: Fokuserar på att jämföra hur olika förklarande variabler (eller behandlingar) påverkar en population.

Analytiska studier kan i sin tur delas upp i olika typer:

- Experimentella: Man kan på förhand bestämma vilka enheter som utsätts för viken behandling. Detta är viktigt för att kunna dra slutsatser om hur olika förklarande variabler påverkar processen.
- Observationsstudier: Det är på förhand givet vilka enheter som får vilken behandling. Dessa är oftast enklare att få data för men är mycket svårare att analysera.

Experimentella undersökningar

- Vi har tidigare undersökt hur vi kan skatta polynomiella (tex linjära eller kvadratiska) samband mellan en responsvariabel och en eller flera förklarande variabler.
- Vi ska nu undersöka mer generella situationer:
  - Sambandet behöver inte beskrivas av någon enkel funktion
  - Förklarande variabler kan vara både numeriska och kategoriska
- Ibland delar man upp variabler som kan påverka processen i olika kategorier:
  - Faktorer: Variabler som vi kan styra själva under experimentet.
  - Kovariater: Variabler som vi kan mäta men ej styra.
  - Övriga variabler som har en systematisk påverkan på processen.
- Ett experiment där man undersöker flera faktorer kallas ett *faktorförsök*.
- De olika nivåerna som används för de olika faktorerna kallas för *faktornivåer.*

Experimentella undersökningar

- Målsättningar för en bra experimentell undersökning bör vara:
  - Den ska eliminera systematiska fel.
  - Den ska vara effektiv, dvs ge goda parameterskattningar med rimligt antal observationer.
  - Den ska tillåta tillräckligt generella slutsatser.
- Det finns flera sätt att uppnå dessa mål:
  - Randomisering: Låt till exempel slumpen bestämma vilka enheter som får vilken behandling. Detta är en viktig teknik, men kan vara svårt att genomföra i praktiken, och kan ge låg precision vid små populationer.
  - Undersök en homogen delpopulation för att minska variationen, tex genom att hålla vissa faktorer konstanta. En nackdel är att detta kan minska förmågan att dra generella slutsatser.
  - Blockindelning/stratifiering: Dela in i homogena delpopulationer och gör jämförelser separat inom varje delpopulation, väg sedan samman till ett slutgiltigt resultat.
  - Studera flera faktorer samtidigt i ett flerfaktorförsök.

Undersökning av en faktor i taget

- Ofta undersöker man i ett experiment endast en faktor.
- Vill man undersöka en ny faktor görs sedan ett nytt försök.
- Detta är ofta en dålig ieé!
- Ett problemet är att vi har svårt att upptäcka *samspelseffekter*.
- Samspel innebär att effekten av en viss faktor varierar beroende på värdet av andra faktorer.
- Även om inga samspelseffekter finns så kräver flerfaktorförsök färre mätningar för samma precision.
- Det är dessutom svårt att hitta optimala värden på faktorerna om de undersöks separat.

Teckentabell för 2^3-försök

Försök	Medel	μ	A	B	C	AB	AC	BC	ABC
(1)	$\bar{y}_{111}$	+	-	-	-	+	+	+	-
a	$\bar{y}_{211}$	+	+	-	-	-	+	+	+
b	$\bar{y}_{121}$	+	-	+	-	-	-	-	+
ab	$\bar{y}_{221}$	+	+	+	-	+	-	-	-
c	$\bar{y}_{112}$	+	-	-	+	+	-	-	+
ac	$\bar{y}_{212}$	+	+	-	+	-	-	-	-
bc	$\bar{y}_{122}$	+	-	+	+	-	+	+	-
abc	$\bar{y}_{222}$	+	+	+	+	+	+	+	+

Yates schema

Vid räkning för hand på ett 2^k-försök kan beräkningarna underlättas avsevärt genom att använda Yates schema:

1. Skriv effekterna i standardordning i en kolumn (tex (1), a, b, ab, c, ac, bc, abc)
2. Skriv ner motvarande medelvärden i en ny kolumn och para värdena.
3. Bilda en ny kolumn (1) där första hälften av värdena är alla parsummor av värden i föregående kolumn, och där andra hälften av värdena är alla pardifferenser. Para värdena.
4. Bilda en ny kolumn (2) genom att upprepa Steg 3, fortsätt med detta tills kolumn (k) bildats.
5. Dela värdena i kolumn (k) för att få skattningarnas värde i standardordningen.

Yates schema (exempel)

Försök	Medel	(1)	(2)	(3)	Effekt	Skattning
(1)	522	1068	-	-	-	-
a	546	-	-	-	-	-
b	557	-	-	-	-	-
ab	581	-	-	-	-	-
c	567	-	-	-	-	-
ac	579	-	-	-	-	-
bc	597	-	-	-	-	-
abc	609	-	-	-	-	-

Yates schema (exempel)

Försök	Medel	(1)	(2)	(3)	Effekt	Skattning
(1)	522	1068	-	-	-	-
a	546	1138	-	-	-	-
b	557	-	-	-	-	-
ab	581	-	-	-	-	-
c	567	-	-	-	-	-
ac	579	-	-	-	-	-
bc	597	-	-	-	-	-
abc	609	-	-	-	-	-

Yates schema (exempel)

Försök	Medel	(1)	(2)	(3)	Effekt	Skattning
(1)	522	1068	-	-	-	-
a	546	1138	-	-	-	-
b	557	1146	-	-	-	-
ab	581	-	-	-	-	-
c	567	-	-	-	-	-
ac	579	-	-	-	-	-
bc	597	-	-	-	-	-
abc	609	-	-	-	-	-

Yates schema (exempel)

Försök	Medel	(1)	(2)	(3)	Effekt	Skattning
(1)	522	1068	-	-	-	-
a	546	1138	-	-	-	-
b	557	1146	-	-	-	-
ab	581	1206	-	-	-	-
c	567	-	-	-	-	-
ac	579	-	-	-	-	-
bc	597	-	-	-	-	-
abc	609	-	-	-	-	-

Yates schema (exempel)

Försök	Medel	(1)	(2)	(3)	Effekt	Skattning
(1)	522	1068	-	-	-	-
a	546	1138	-	-	-	-
b	557	1146	-	-	-	-
ab	581	1206	-	-	-	-
c	567	24	-	-	-	-
ac	579	-	-	-	-	-
bc	597	-	-	-	-	-
abc	609	-	-	-	-	-

Yates schema (exempel)

Försök	Medel	(1)	(2)	(3)	Effekt	Skattning
(1)	522	1068	2206	4558	$\hat{\mu}$	569.75
a	546	1138	2352	72	$\hat{A}$	9
b	557	1146	48	130	$\hat{B}$	16.25
ab	581	1206	24	0	$\hat{AB}$	0
c	567	24	70	146	$\hat{C}$	18.25
ac	579	24	60	-24	$\hat{AC}$	-3
bc	597	12	0	-10	$\hat{BC}$	-1.25
abc	609	12	0	0	$\hat{ABC}$	0