INTRODUCING LINEAR REGRESSION MODELS

Straight line regression

- Response or Dependent variable y
- Predictor or Independent variable x
- Measurement error model: repeat values $i = 1, \ldots, n$,

$$y_i = \alpha + \beta x_i + \epsilon_i$$

- ϵ_i : independent errors (sampling, measurement, lack of fit)
- Typically/initially: $\epsilon \sim N(0, \sigma^2)$
- Analysis and inference:
  - Estimate parameters $(\alpha, \beta, \sigma^2)$
  - Assess model fit — adequate? good? if inadequate, how?
  - Explore implications: $\beta, \beta x$
  - Predict new ("future") responses at new $x_{n+1}, \ldots$

BIG PICTURE:

- Understanding variability in y as a function of x
- Exploring $p(y|x)$ for different x values
- One aspect: *Regression function* $E(y|x)$ as x varies
- Special case: normal, linear in mean
  - Other cases: binomial y, success prob depends on x
  - e.g., Dose-response models

- How much variability does x explain?
- Normal models: Variance measures "variability"

- Observational studies versus Designed studies
  - "Random" x versus "Controlled" x

- Bivariate data (y_i, x_i) BUT focus on x_i fixed
- "Special" status of response variable
- Several or many predictor variables

e.g., POLLUTION LEVELS, MERCEDES USED CAR PRICES, OLD FAITHFUL GEYSER TIMES, SEX BIAS IN SALARIES, UNIVERSITY TUITION LEVELS, EEG DATA, ABALONE SHELL FISH AGES, ... etc

SAMPLE SUMMARY STATISTICS

- Sample means $\bar{x}, \bar{y}$
- Sample variances s_x^2, s_y^2

$$s_y^2 = S_{yy}/(n-1), \qquad s_x^2 = S_{xx}/(n-1)$$

- ... and sample COVARIANCE

$$s_{xy} = S_{xy}/(n-1)$$

where

- $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ – "Total Variation in response"
- $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$
- $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

"Sums of squares" S_{xx}, S_{yy}, S_{xy}

– measures of total variation and *covariation*

Standardised scale for covariance:

SAMPLE CORRELATION:

$$r = \frac{s_{xy}}{s_x s_y}$$

$-1 < r < 1$, measure of dependence

S-Plus: `var(y), var(x), cor(y,x)`

SQUARED ERRORS AND “FIT” OF CHOSEN LINES

Measurement error version of model: $y_i = \alpha + \beta x_i + \epsilon_i$

For any chosen α, β,

$$Q(\alpha, \beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

measures “fit” of chosen line $\alpha + \beta x$ to response data

LEAST SQUARES LINE:

- Choose $\hat{\alpha}, \hat{\beta}$ to *minimise* $Q(\alpha, \beta)$
- *Least squares estimates* (LSE) $\hat{\alpha}, \hat{\beta}$
- (Venerable/ad-hoc) “principal” of *least squares estimation*

LEAST SQUARES ESTIMATES

FACTS:

$$\hat{\beta} = \frac{s_{xy}}{s_x^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Or

$$\hat{\beta} = r\left(\frac{s_y}{s_x}\right)$$

$\hat{\beta}$ *is correlation coefficient* r, *corrected for relative scales of* $y : x$ *so that the units of the "fitted values"* $\hat{\beta}x$ *are on scale of* y

R^2 measure of model fit:

Simplest model: $\beta = \hat{\beta} = 0$ so y_i are a normal random sample

$$\hat{\alpha} = \bar{y}, \qquad Q(\bar{y}, 0) = S_{yy} = \text{total sum of squares}$$

Any other model fit: Residual Sum of Squares $Q(\hat{\alpha}, \hat{\beta})$

DEFINE: $R^2 = 1 - Q(\hat{\alpha}, \hat{\beta})/S_{yy}$

– proportion of variation “explained” by model –

FACT: $R^2 = r^2$ *(algebra ...)*

- “Multiple regression” generalisation later
- Higher %variation explained is better: Higher correlation

S-Plus: *linear model* fitting function: `lm(x)`, See examples

EXAMINING MODEL FIT

- Fitted values $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$
- Residuals $\hat{\epsilon}_i = y_i - \hat{y}_i$ ... estimates of ϵ_i
- Residual sum of squares $Q(\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^{n} \hat{\epsilon}_i^2$

  – measures remaining/residual variation in response data –
- Residual sample variance:

$$s^2 = \sum_{i=1}^{n} \hat{\epsilon}_i^2 / (n-2)$$

- s^2 is a point estimate of σ^2 from fitted model

  n.b., $n-2$ degrees of freedom, not $n-1$

  – "lose" one degree of freedom for each model parameter α, β –

THEORY FOR INFERENCE: REFERENCE POSTERIOR

Anticipating later theory, some key aspects of the REFERENCE posterior for $(\alpha, \beta, \sigma^2)$

- (marginal) posterior for β is T distribution with $n-2$ d.o.f.

$$T_{n-2}(\hat{\beta}, s^2 v_\beta^2)$$

  where $v_\beta^2 = 1/S_{xx}$

- s^2 is the posterior estimate of σ^2 – residual variance

Key to assessing *significance* of regression fit and measuring the "explanatory power" of chosen predictor x

Intervals:

$$\hat{\beta} \pm (s v_\beta) t_{p/2}$$

where $t_{p/2}$ is 100(p/2)% quantile of standard T_{n-2}

“TESTING” SIGNIFICANCE OF THE REGRESSION FIT

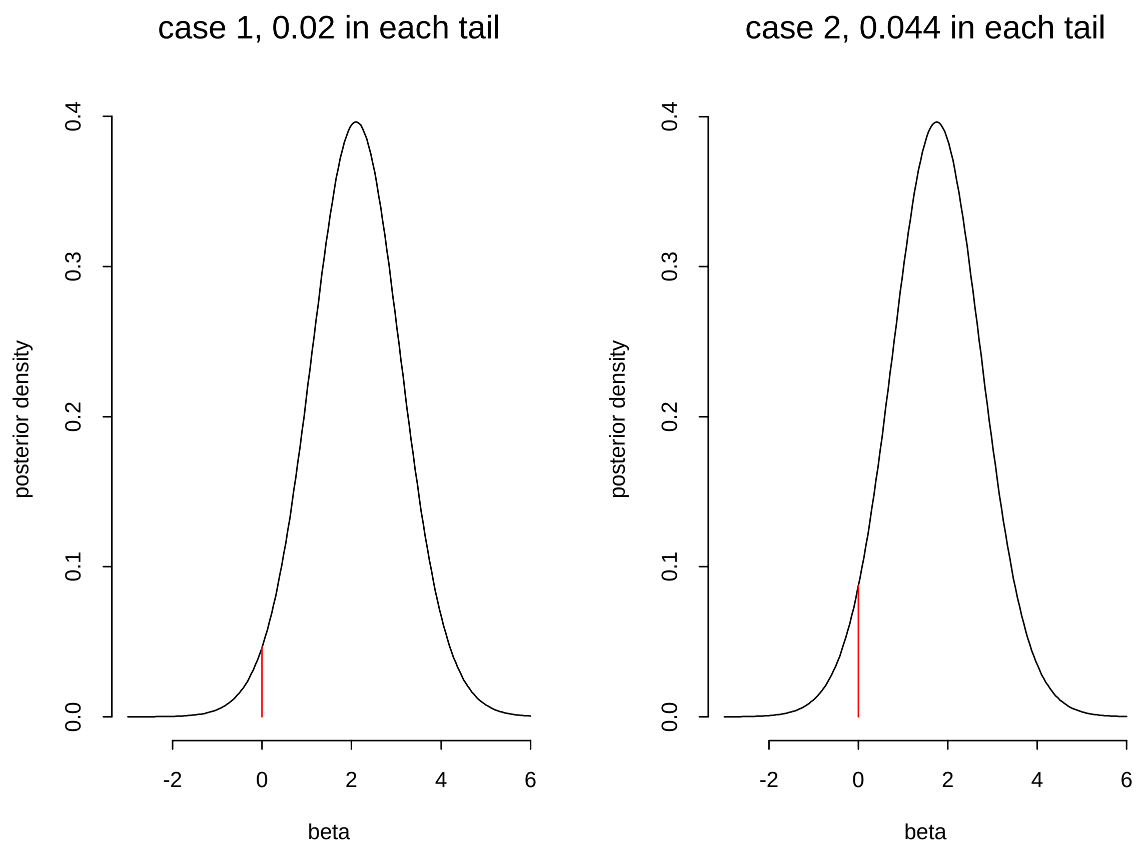
Question: How probable is $\beta = 0$ under the posterior?

Answer:

- Compute posterior probability on β values with lower posterior density than $\beta = 0$
- “Measures” probability of β “less likely” than $\beta = 0$
- Informal “test” of significance –
  Probability in tails = significance level = (Bayesian) p–value
- Symmetric posterior density: double one tail area
- S-Plus: `2*(1-pt(abs(T), n-2))` where
  - T=$\hat{\beta}/sv_\beta$ – *standardised T Statistic*

Classical testing terminology:

"The regression on x is significant at the 5% level (or 1%, etc) if the $p-$value is smaller than 0.05 (or 0.01, etc)"



F TESTS, ANOVA AND DEVIANCES

F test of regression fit:

Theory: If $t \sim T_k(0,1)$ then $F = t^2 \sim F_{1,n-2}$

So

- $p-$value $= Pr(F \geq f_{obs})$
- $f_{obs} = \hat{\beta}^2 / s^2 v_\beta^2$
- T and F tests are equivalent: same $p-$value
- S-Plus output: quotes T values, $p-$values in coefficient table and F test result

F TESTS, ANOVA AND DEVIANCES

Deviances = Sums of squares: Deviance decomposition ...

$$S_{yy} = Q(\hat{\alpha}, \hat{\beta}) + \hat{\beta}^2 / v_\beta^2$$

- Total deviance $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$
- Residual deviance $Q(\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- *Fitted or explained deviance:* $\hat{\beta}^2 / v_\beta^2$
  – here equal to $s^2 f_{obs}$ –
- Large deviance explained $\equiv$ large F $\equiv$ significant regression
- ANOVA: analysis of variance (deviance)

HONEST PREDICTION FROM FITTED MODEL

Question: What is the posterior predictive distribution for a new case,

$$y_{n+1} = \alpha + \beta x_{n+1} + \epsilon_{n+1}$$

Answer: Also a Student t distribution with $n-2$ d.o.f.

$$y_{n+1} \sim T_{n-2}(\hat{y}, s^2 v_y^2)$$

- Mean is $\hat{y} = \hat{\alpha} + \hat{\beta} x_{n+1}$
- Spread: $s^2 v_y^2 = s^2 + s^2 w^2$ ...
  - $s^2 w^2$ – posterior uncertainty about $\alpha + \beta x_{n+1}$ depends on x_{n+1}, spread is higher for x_{n+1} far from $\bar{x}$
  - additional variability $+s^2$ due to ϵ_{n+1}, estimating σ^2 by s^2

Form of predictive distribution anticipates theory

– later under Multiple linear regression –

S-Plus function: `predict()` handles all the details

Examples in S-Plus code: pollution data, mercedes used prices, etc

Model fit assessment/implications: Explore predictive distributions

Residual analysis: Graphical exploration of fitted residuals $\hat{\epsilon}_i$

- Standardise: $r_i = \hat{\epsilon}_i / s$
- Approximately standard normal? qqplot, etc
- RESIDUALS = RESPONSE MINUS FIT:

  Treat ϵ_i as "new data" – look at structure, other predictors

Other predictors?