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 $\epsilon \sim N(0, \sigma^2)$
- 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 $p(y|x)$ as a function of x
- 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., logistic regression, 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 take x_i fixed
- “Special” status of response variable
- Several or many predictor variables

SAMPLE SUMMARY STATISTICS

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

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

- sample covariance

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

where the "Sums of Squares" are:

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

Standardised scale for covariance:

SAMPLE CORRELATION:

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

$$-1 < r < 1, \text{ measure of dependence}$$

for a single predictor, $r^2 = R^2$

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* $Q(\alpha, \beta)$
- Geometric interpretation
- *Least squares estimates* (LSE) $\hat{\alpha}, \hat{\beta}$
- (Venerable/ad-hoc) "principal" of least squares estimation
- Least squares fit is also the MLE

LEAST SQUARES ESTIMATES

FACTS:

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

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

Note also

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

of use in theoretical derivations

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: $R^2 = r^2$

- "Multiple regression" generalisation later
- Higher %variation explained is better: Higher correlation
- Measures linear correlation
  - not general dependence
  - not causation

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:

$$s^2 = \frac{RSS}{n-2} = \sum_{i=1}^{n} \frac{\hat{\epsilon}_i^2}{n-2}$$

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

  note: $n-2$ degrees of freedom, not $n-1$

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

Conjugate Priors for Regression

- Normal is conjugate for α and β
- Inverse-Gamma is conjugate for σ^2

Common re-parameterization: Precision $\phi = \frac{1}{\sigma^2}$

- Gamma is conjugate for ϕ

A Reference Prior for Regression

Take limits of conjugate prior as the prior variance goes to infinity (information goes to zero)

- $\lim_{t \to \infty} N(0, t)$ is proportional to a constant
- $\lim_{\substack{a \to 0, \\ b \to 0}} \Gamma(a, b)$ or $\Gamma^{-1}(a, b)$ is proportional to the inverse

$$p(\alpha, \beta, \sigma^2) \propto \frac{1}{\sigma^2} \qquad\qquad p(\alpha, \beta, \phi) \propto \frac{1}{\phi}$$

THEORY FOR INFERENCE: REFERENCE POSTERIOR

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 (HPD or equal-tailed):

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

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
- 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:

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

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 ≡ large F ≡ 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
- Can use S-Plus function `predict.lm()`

Model fit assessment/implications: Explore predictive distributions

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

- Standardise: $r_i = \hat{\epsilon}_i / s$
- Check normality assumption
- Treat ϵ_i as "new data" – look at structure, other predictors

Other predictors?