Inference for coefficients

Mean response at x vs. New observation at x

Linear Model (or Simple Linear Regression) for the population.
("Simple" means single explanatory variable, in fact we can easily add more variables)
– explanatory variable (independent var / predictor) – response (dependent var)

Probability model for linear regression:

$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad \begin{array}{ll} \epsilon_i \sim N(0,\sigma^2) & \text{independent deviations} \\ \alpha + \beta x_i & \text{mean response at } x = x_i \end{array}$$

Goals: unbiased estimates of the three parameters $(\alpha, \beta, \sigma^2)$ tests for null hypotheses: $\alpha = \alpha_0$ or $\beta = \beta_0$ C.I.'s for α, β or to predict$E(Y|X = x_0)$.

(A model is our 'stereotype' – a simplification for summarizing the variation in data)

For example if we simulate data from a temperature model of the form:

$$Y_i = 65 + \frac{1}{3}x_i + \epsilon_i, \qquad x_i = 1, 2, \ldots, 30$$

Model is exactly true, by construction

An *equivalent* statement of the LM model: Assume x_i fixed, Y_i independent, and

$$Y_i|x_i \sim N(\mu_{y|x_i}, \sigma^2), \qquad \mu_{y|x_i} = \alpha + \beta x_i, \text{population regression line}$$

Remark: Suppose that (X_i, Y_i) are a random sample from a bivariate normal distribution with means (μ_X, μ_Y), variances σ_X^2, σ_Y^2 and correlation ρ. Suppose that we condition on the observed values $X = x_i$. Then the data (x_i, y_i) satisfy the LM model. Indeed, we saw last time that $Y|x_i \sim N(\mu_{y|x_i}, \sigma^2_{y|x_i})$, with

$$\mu_{y|x_i} = \alpha + \beta x_i, \qquad \sigma^2_{Y|X} = (1-\rho^2)\sigma_Y^2$$

Example: Galton's fathers and sons: $\mu_{y|x} = 35 + 0.5x$; σ = 2.34 (in inches). For comparison: $\sigma_Y = 2.7$.

Note: σ is SD of y **given** x. It is less than the unconditional SD of Y, σ_Y (without *x* fixed). e.g. consider extreme case when $y = x$: σ=0, but $\sigma_Y = \sigma_X$.

Estimating parameters $(\alpha, \beta, \sigma^2)$ **from a sample**

Recall the least squares estimates: *(a,b)* chosen to minimize $\sum(y_i - a - bx_i)^2$:

$$a = \bar{y} - b\bar{x} \qquad b = r\frac{s_y}{s_x} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Fitted values: $\hat{y} = a + bx_i$

Residuals:

$$e_i = y_i - \hat{y}_i = y_i - a - bx_i = \alpha + \beta x_i + \epsilon_i - a - bx_i = (\alpha - a) + (\beta - b)x_i + \epsilon_i$$

where e_i = deviations in *sample*,
ϵ_i = deviations in *model*

Estimate of σ^2: Again, use the squared scale:

$$s^2_{Y|x} = \frac{1}{n-2}\sum_i e_i^2 = \frac{1}{n-2}\sum_i (y_i - a - bx_i)_i^2, s = \sqrt{s^2_{Y|x}} \qquad (n-2) = \text{“degrees of freedom”}$$

Why n-2??

1. Recall with SRS $Y_i \sim N(\mu, \sigma^2)$, $s^2 = \frac{1}{n-1}\sum(y_i - \bar{y})^2$

   (and $\bar{y}$ estimates μ – we estimate **one** parameter)

   Here, **two** parameter estimates: *a* estimates α , and b estimates β.

2. There are two linear constraints on e_i:

   - $\bar{e} = \bar{y} - a - b\bar{x} = \bar{y} - (\bar{y} - b\bar{x}) - b\bar{x} = 0$
   - $\sum(x_i - \bar{x})e_i = 0$

Properties of Regression Estimates. Assume model (LM): then $(a, b, s^2 = s^2_{Y|x})$ are **random variables.**

1. **Means** $(a, b, s^2 = s^2_{Y|x})$ are unbiased estimates of $(\alpha, \beta, \sigma^2)$

2. **Variances** For (a, b) variances are

$$\sigma_a^2 = Var(a) = \sigma^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right\}, \qquad \sigma_b^2 = Var(b) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

3. (a, b) are jointly normal – in particular

$$a \sim N(\alpha, \sigma_a^2), \qquad b \sim N(\beta, \sigma_b^2), \qquad (n-2)\frac{s^2}{\sigma^2} \sim \chi^2_{n-2} \qquad \text{(and a,b, and } s^2 \text{ are independent).}$$

Confidence Intervals: at level α CI's always have the form: Est $\pm\, t_{1-\frac{\alpha}{2},(n-2)}\text{SE}_{Est}$

SE_{Est}means σ_{Est} with σ replaced by its estimate s. Thus

$$SE_b = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}, \qquad SE_a = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$$

Thus, $(1-\alpha)\%$ CI for:

slope	$b \pm t_{1-\frac{\alpha}{2},(n-2)}\text{SE}_b$
intercept	$a \pm t_{1-\frac{\alpha}{2},(n-2)}\text{SE}_a$

Tests Described here for slope β , (but same for intercept α).
For $H_0 : \beta = \beta_0$, use $t = \frac{b-\beta_0}{SE_b}$ which under H_0 has the t_{n-2} distribution.

P values:	**One sided:** (e.g.)	$H_1 : \beta < \beta_0$,	$P = P(t_{n-2} < t_{obs})$
	Two sided:	$H_1 : \beta \neq \beta_0$,	$P = 2P(t_{n-2} > \lvert t_{obs}\rvert)$

(Aside: Why t_{n-2}?
By definition $t_\nu \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2_\nu}{\nu}}}$, with numerator and denominator variable independent. But we can write the slope test statistic in the form

$$t = \frac{(b-\beta_0)/\sigma_b}{\sqrt{\frac{SE_b^2}{\sigma_b^2}}}$$

and now we can note from properties (1) – (3) that

$$(b-\beta_0)/\sigma_b \sim N(0,1) \qquad \text{and} \qquad SE_b^2/\sigma_b^2 = s^2/\sigma^2 \sim \chi^2_{n-2}$$

and it can be shown that *b* and s^2 are independent. This shows that the test statistic $t \sim t_{n-2}$.)

Vs

Mean response at x

Want to estimate $\mu_{Y|x^*} = \alpha + \beta x^*$
mean for subpopulation given $x = x^*$
Point estimate:$\hat{\mu}_{Y|x^*} = a + bx^*$
Sources of uncertainty :a, b estimated.

$$SE_{\hat{\mu}_{Y|x^*}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

CI :level $(1-\alpha)\%$ n-2 df

$$\hat{\mu}_{Y|x^*} \pm t_{1-\alpha/2,n-2} SE_{\hat{\mu}_{Y|x^*}}$$

Claim: 95% of the time , CI covers $\mu_{Y|x^*} = \alpha + \beta x^*$
ie

$$\hat{\mu} - t_{1-\alpha/2,n-2} SE_{\hat{\mu}} \leq \mu \leq \hat{\mu} + t_{1-\alpha/2,n-2} SE_{\hat{\mu}}$$

```
> predict(lm(y ~ x), new,
interval="confidence",se.fit=T)
$fit
        fit       lwr       upr
1 -1.82756 -3.165220 -0.48991
2 -1.55835 -2.690602 -0.42610
3 -1.28914 -2.222694 -0.35558
4 -1.01992 -1.766869 -0.27298
```

New Observation at x

Want to **predict** $y^{new}(x^*) = \alpha + \beta x^* + \epsilon^{new}$
new draw of y from subpopulation given $x = x^*$
Point prediction:$\hat{y}(x^*) = a + bx^*$
Sources of uncertainty :a, b estimated.
random error $\epsilon$of new observ.

$$SE^2_{\hat{y}(x^*)} = s^2_{Y|x} + SE^2_{\hat{\mu}_{Y|x^*}}$$

Prediction Interval : level $(1-\alpha)\%$ n-2 df

$$\hat{y}(x^*) \pm t_{1-\alpha/2,n-2} SE_{\hat{y}(x^*)}$$

Claim: 95% of the time , Pred. Interv. covers $y^{new}(x^*)$
ie

$$\hat{y} - t_{1-\alpha/2,n-2} SE \leq y^{new} \leq \hat{y} + t_{1-\alpha/2,n-2} SE$$

```
predict(lm(y ~ x), data,
interval="prediction",se.fit=T)
$fit
        fit       lwr     upr
1 -1.827565 -3.99873 0.34360
2 -1.558351 -3.60936 0.49266
3 -1.289137 -3.23751 0.65924
4 -1.019922 -2.88608 0.84624
```