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

Probability model for linear regression:

\[ Y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \quad \text{independent deviations} \]

\[ \alpha + \beta x_i \quad \text{mean response at } x = x_i \]

**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 \(\alpha, \beta\) 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:

\[ Y_i = 65 + \frac{1}{3} x_i + \epsilon_i, \quad x_i = 1, 2, \ldots, 30 \]

Model is exactly true, by construction

An equivalent statement of the LM model: Assume \(x_i\) fixed, \(Y_i\) independent, and

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

**Remark**: Suppose that \((X_i, Y_i)\) are a random sample from a bivariate normal distribution with means \((\mu_X, \mu_Y)\), variances \(\sigma_X^2, \sigma_Y^2\) and correlation \(\rho\). 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

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

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

**Note**: \(\sigma\) is SD of \(y\) given \(x\). It is less than the unconditional SD of \(Y, \sigma_Y\) (without \(x\) fixed). e.g.

consider extreme case when \(y = x; \sigma = 0\), but \(\sigma_Y = \sigma_X\).

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

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

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

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,

\(\epsilon_i\) = deviations in model

**Estimate of** \(\sigma^2\): Again, use the squared scale:

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

**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 $\mu$ – we estimate one parameter)

Here, two parameter estimates: $a$ estimates $\alpha$, and $b$ estimates $\beta$.

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

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

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

   \[ a \sim N(\alpha, \sigma_a^2), \quad b \sim N(\beta, \sigma_b^2), \quad (n-2)\frac{s^2}{\sigma^2} \sim \chi^2_{n-2} \]

   (and a, b, and $s^2$ are independent).

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

$SE_{\text{Est}}$ means $\sigma_{\text{Est}}$ with $\sigma$ replaced by its estimate $s$. Thus

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

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

- slope $b \pm t_{1-\frac{\alpha}{2},(n-2)}SE_b$
- intercept $a \pm t_{1-\frac{\alpha}{2},(n-2)}SE_a$

Tests Described here for slope $\beta$, (but same for intercept $\alpha$).

For $H_0: \beta = \beta_0$, use $t = \frac{b - \beta_0}{\frac{SE_b}{\sigma_b}}$ which under $H_0$ has the $t_{n-2}$ distribution.

P values: One sided: (e.g.) $H_1: \beta < \beta_0$, $P = P(t_{n-2} < t_{\text{obs}})$

Two sided: $H_1: \beta \neq \beta_0$, $P = 2P(t_{n-2} > |t_{\text{obs}}|)$

(Aside: Why $t_{n-2}^2$? 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

\[ t = \frac{(b - \beta_0)/\sigma_b}{\sqrt{\frac{SE^2_b}{\sigma^2_b}}} \]

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

\[ (b - \beta_0)/\sigma_b \sim N(0,1) \quad \text{and} \quad \frac{SE^2_b}{\sigma^2_b} = \frac{s^2}{\sigma^2} \sim \chi^2_{n-2} \]

and it can be shown that $b$ and $s^2$ are independent. This shows that the test statistic $t \sim t_{n-2}$.)
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 \( \mu_{Y|x^*} = \alpha + \beta x^* \)

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

\[ \hat{\mu} - t_{1-\alpha/2,n-2} SE \leq \hat{\mu} \leq \hat{\mu} + t_{1-\alpha/2,n-2} SE \]

> predict(lm(y ~ x), new, interval="confidence", se.fit=T)
$fit

<table>
<thead>
<tr>
<th>fit</th>
<th>lwr</th>
<th>upr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.82756</td>
<td>-3.165220</td>
</tr>
<tr>
<td>2</td>
<td>-1.55835</td>
<td>-2.690602</td>
</tr>
<tr>
<td>3</td>
<td>-1.28914</td>
<td>-2.222694</td>
</tr>
<tr>
<td>4</td>
<td>-1.01992</td>
<td>-1.766869</td>
</tr>
</tbody>
</table>

Vs 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.

\[ SE_{\hat{y}(x^*)} = s^2_{Y|x} + SE_{\mu_{Y|x^*}}^2 \]

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 \( y^{new}(x^*) \)

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

> predict(lm(y ~ x), data, interval="prediction", se.fit=T)
$fit

<table>
<thead>
<tr>
<th>fit</th>
<th>lwr</th>
<th>upr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.827565</td>
<td>-3.99873</td>
</tr>
<tr>
<td>2</td>
<td>-1.558351</td>
<td>-3.60936</td>
</tr>
<tr>
<td>3</td>
<td>-1.289137</td>
<td>-3.23751</td>
</tr>
<tr>
<td>4</td>
<td>-1.019922</td>
<td>-2.88608</td>
</tr>
</tbody>
</table>