Hypothesis Testing

- The summary of a linear model and the ANOVA table of a linear model summarize different hypothesis tests.
- Regression coefficients must be understood in the context of which other variables are included in the model.
- The R command `summary` shows results of \( t \)-tests that test if a parameter is zero in a model that includes all other coefficients.
- The R command `anova` shows the results of \( F \)-tests that test if a parameter is zero in a model including only parameters listed earlier in the table.
- However, tests in the ANOVA table use the full model to estimate error for all tests.
R Commands to Fit Models

```r
> toxic = read.table("toxic.txt", header = T)
> str(toxic)
> attach(toxic)
> toxic0.lm = lm(effect ~ 1)
> toxic1.lm = lm(effect ~ dose)
> toxic2.lm = lm(effect ~ weight)
> toxic12.lm = lm(effect ~ dose + weight)
> toxic21.lm = lm(effect ~ weight + dose)
```

Differences in Tests

- We need to distinguish between

\[ H_0: [\beta_1 = 0 | \beta_0] \]

(i.e., \( \beta_1 = 0 \) given that \( \beta_0 \) is in the model), and

\[ H_0: [\beta_1 = 0 | \beta_0, \beta_2] \]

(i.e., \( \beta_1 = 0 \) given that \( \beta_0, \beta_2 \) are in the model).
Hypothesis Testing

Example

Pesticide Example (cont.)

```r
> summary(toxic12.lm)
Call: lm(formula = effect ~ dose + weight)
Residuals:
     Min      1Q  Median      3Q     Max
-0.081512 -0.023945 -0.003421  0.022278  0.094310
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.22281   0.08364   2.664  0.01698 * 
dose        0.65139   0.17305   3.764  0.00170 **
weight      -1.13321   0.18044  -6.280 1.10e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.0466 on 16 degrees of freedom
Multiple R-Squared: 0.7796, Adjusted R-squared: 0.752
F-statistic: 28.3 on 2 and 16 DF,  p-value: 5.57e-06
```

```r
> anova(toxic12.lm)
Analysis of Variance Table
Response: effect
Df  Sum Sq Mean Sq F value Pr(> F)
 dose 1  0.03724 0.03724 17.152 0.0007669 ***
weight 1  0.08563 0.08563 39.440 1.097e-05 ***
Residuals 16  0.03474 0.00217
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null p-value

- $H_0: \beta_1 = 0 \mid \beta_0, \beta_2$ 0.00170
- $H_0: \beta_2 = 0 \mid \beta_1, \beta_2$ 1.10e-05
- $H_0: [\beta_0 = 0 \mid \beta_1, \beta_2]$ 0.01698
- $H_0: [\beta_1 = \beta_2 = 0 \mid \beta_0]$ 5.57 x 10^-6
- $H_0: [\beta_1 = 0 \mid \beta_0]$ 0.0007669
- $H_0: [\beta_2 = 0 \mid \beta_0, \beta_1]$ 1.097e-05

Coefficient Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>Intercept</th>
<th>Dose</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>toxic0.lm</td>
<td>0.3714</td>
<td></td>
<td></td>
</tr>
<tr>
<td>toxic1.lm</td>
<td>0.6049</td>
<td>-0.3206</td>
<td></td>
</tr>
<tr>
<td>toxic2.lm</td>
<td>0.5226</td>
<td>-0.5258</td>
<td></td>
</tr>
<tr>
<td>toxic12.lm</td>
<td>0.2226</td>
<td>0.6514</td>
<td>-1.133</td>
</tr>
<tr>
<td>toxic21.lm</td>
<td>0.2228</td>
<td>0.6514</td>
<td>-1.133</td>
</tr>
</tbody>
</table>

- Coefficient interpretation depends on other variables present in the model!
The usual model is:
\[ y_i = \beta_0 + \beta_1 x_i + e_i, \text{ where } e_i \sim \text{iid } N(0, \sigma^2) \]

Suppose we accept \( H_0: \beta_0 = 0 \). Then the model reduces to:
\[ y_i = \beta_1 x_i + e_i, \text{ where } e_i \sim \text{iid } N(0, \sigma^2). \]

Model parameters are \( \beta_1, \sigma^2. \)

The equations for least squares regression change.

The least squares criterion becomes to minimize \( \sum_{i=1}^{n} (y_i - (b_1 x_i))^2 \).

The solution to this problem is:
\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}. \]

The estimated variance is \( \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i)^2}{n-1} \)

Inference using regression through the origin has one more degree of freedom for error as one fewer parameter is needed for the mean.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>( \frac{(\sum_{i=1}^{n} x_i y_i)^2}{\sum_{i=1}^{n} x_i^2} )</td>
<td>( \frac{(\sum_{i=1}^{n} x_i y_i)^2}{\sum_{i=1}^{n} x_i^2} )</td>
</tr>
<tr>
<td>Error</td>
<td>( n - 1 )</td>
<td>By subtraction</td>
<td>SSErove ( (n - 1) )</td>
</tr>
<tr>
<td>Total</td>
<td>( n )</td>
<td>( \sum_{i=1}^{n} y_i^2 )</td>
<td>–</td>
</tr>
</tbody>
</table>
Regression Through the Origin

Example

```r
> x = c(40, 43, 49, 52, 54, 55, 60)
> y = c(44, 34, 46, 53, 55, 56, 51, 55, 60)
> origin = data.frame(x = x, y = y)
> origin.lm = lm(y ~ x, data = origin)
> summary(origin.lm)$coefficients

                Estimate Std. Error  t value  Pr(>|t|)
(Intercept) -5.585586   13.265566  -0.42106  0.6863388
x            1.105856    0.260159   4.25069  0.0037899

> anova(origin.lm)

Analysis of Variance Table

Response: y

             Df Sum Sq Mean Sq  F value    Pr(>F)
(Intercept) 1 361.98 361.980 18.06800 0.0037899 **
 Residuals   7 140.24  20.034

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fit Without an Intercept

```r
> origin2.lm = lm(y ~ x - 1, data = origin)
> summary(origin2.lm)$coefficients

                Estimate Std. Error  t value Pr(>|t|)
 x            0.9970085    0.0277148 35.97382 3.905618e-10

> anova(origin2.lm)

Analysis of Variance Table

Response: y

             Df  Sum Sq Mean Sq F value    Pr(>F)
 x            1 23260.2 23260.2 1294.10 3.906e-10 ***
 Residuals   8  143.8  18.036

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

> par(mar = c(4, 4, 0, 0), las = 1, pch = 16)
> plot(x, y, xlim = c(0, 60))
> abline(origin.lm)
> abline(origin2.lm, col = "red", lty = 2)
```