Linear Regression with Categorical Predictors

Lin Wang

Previous Example: Portland Cement Formulation

- Now let's consider linear regression for the Portland cement data

■ TABLE 2.1
Tension Bond Strength Data for the Portland Cement Formulation Experiment

j	Modified Mortar y_{1j}	Unmodified Mortar y_{2j}
1	16.85	16.62
2	16.40	16.75
3	17.21	17.37
4	16.35	17.12
5	16.52	16.98
6	17.04	16.87
7	16.96	17.34
8	17.15	17.02
9	16.59	17.08
10	16.57	17.27

Linear regression with categorical predictors:
X is the cement formulation with two levels 0 (unmodified) and 1 (modified)
Y is the tension bond strength

```
> x=c(rep(1,10),rep(0,10))
> x
 [1] 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0
> y=c(16.85,16.40,17.21,16.35,16.52,17.04,16.96,
17.15,16.59,16.57,16.62,16.75,17.37,17.12,16.98,
16.87,17.34,17.02,17.08,17.27)
```

```
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4220 -0.2065  0.0080  0.2400  0.4460

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.04200    0.08989 189.590   <2e-16 ***
x           -0.27800    0.12712  -2.187   0.0422 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2843 on 18 degrees of freedom
Multiple R-squared:  0.2099,    Adjusted R-squared:  0.166
F-statistic: 4.782 on 1 and 18 DF,  p-value: 0.0422
```

Requires normal assumption for p-value

What if the categorical predictor has three levels?

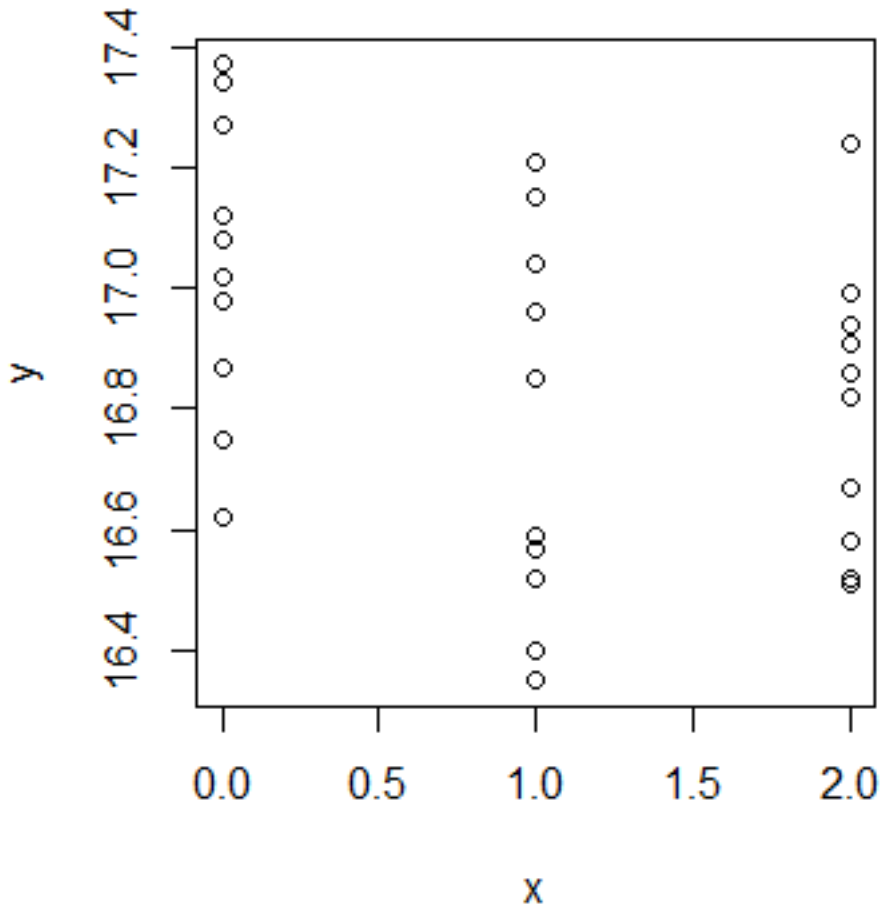
- For example, three formulations of cement

■ TABLE 2.1
Tension Bond Strength Data for the Portland Cement Formulation Experiment

j	Modified Mortar y_{1j}	Unmodified Mortar y_{2j}
1	16.85	16.62
2	16.40	16.75
3	17.21	17.37
4	16.35	17.12
5	16.52	16.98
6	17.04	16.87
7	16.96	17.34
8	17.15	17.02
9	16.59	17.08
10	16.57	17.27

A third mortar

16.58
16.82
16.51
17.24
16.86
16.52
16.91
16.99
16.94
16.67



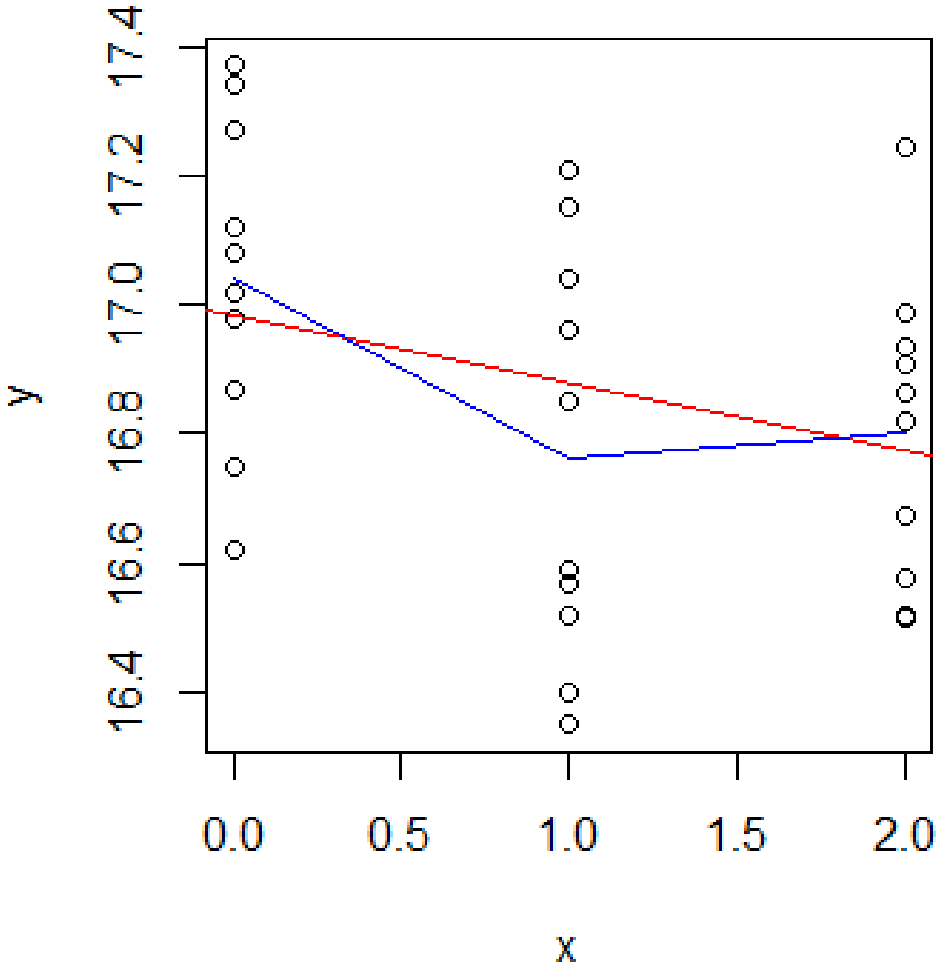
Is the solution below feasible?

```
> x
 [1] 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 2 2 2 2 2 2 2 2 2 2

> y
 [1] 16.85 16.40 17.21 16.35 16.52 17.04 16.96 17.15 16.59 16.57 16.62
[12] 16.75 17.37 17.12 16.98 16.87 17.34 17.02 17.08 17.27 16.58 16.82
[23] 16.51 17.24 16.86 16.52 16.91 16.99 16.94 16.67

 > lm(y~x)

 Call:
 lm(formula = y ~ x)

 Coefficients:
 (Intercept)            x
      16.989       -0.119
```

```
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5200 -0.2370  0.0500  0.1842  0.4890

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.98900    0.07936 214.083   <2e-16 ***
x           -0.11900    0.06147  -1.936    0.063 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2749 on 28 degrees of freedom
Multiple R-squared:  0.118,     Adjusted R-squared:  0.08655
F-statistic: 3.748 on 1 and 28 DF,  p-value: 0.06303
```

Not significant! Contradicts with our previous result that the first two types of mortar are different!

Reasons?



The linear assumption is the key!

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

which is assuming each unit increase of X has the same increase on Y.

The linear assumption does not make sense for categorical predictors!

Coding System for Categorical Predictors

Dummy coding

Level	New variable 1 (x1)	New variable 2 (x2)
1	1	0
2	0	1
3	0	0

- Our model is $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$, here i is for the ith observation
- The model can be written as $y_{ij} = \mu_i + \epsilon_{ij}$, (means model) for the jth replicate of the ith level
- We may also write the model as $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, (effects model) with the restriction that $\tau_1 + \tau_2 + \tau_3 = 0$

Equivalences (3 levels shown)

Dummy (treatment): $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

Mapping: $\beta_0 = \mu_3$, $\beta_1 = \mu_1 - \mu_3$, $\beta_2 = \mu_2 - \mu_3$

Cell means: $y_{ij} = \mu_i + \varepsilon_{ij}$

Effects (sum) coding: $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, with $\Sigma\, n_i\, \tau_i = 0$

Grand mean: $\mu = (n_1\, \mu_1 + n_2\, \mu_2 + n_3\, \mu_3) / N$; $\tau_i = \mu_i - \mu$; balanced $\Rightarrow \Sigma\, \tau_i = 0$

Parameterization	What coefficients mean	Advantages	Disadvantages
Dummy coding (ref = level 3)	β_0 reference mean; β_k difference vs reference	Natural with control; direct vs control tests; works well in GLMs	Asymmetric; depends on reference; pairwise diffs need contrasts; no grand mean directly
Cell-means model (~ 0 + group)	Each μ_i is the group mean	Transparent for reporting means; easy plotting	Need to drop intercept; global tests need contrasts; in GLMs μ_i on link scale
Effects (sum) coding	μ grand mean; τ_i deviation from grand mean	Symmetric across levels; good with no baseline	Less convenient for vs control

Practical guidance

Use dummy when you have a clear control and want direct "vs control" estimates.

Use cell-means when you mainly report group or factorial cell means.

Use effects coding when no level is special and you want symmetric deviations or ANOVA-style summaries.

Coding for a variables with 4 levels:

Level of race	New variable 1 (x1)	New variable 2 (x2)	New variable 3 (x3)
1 (Hispanic)	1	0	0
2 (Asian)	0	1	0
3 (African American)	0	0	1
4 (white)	0	0	0

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

$$y_{ij} = \mu_i + \epsilon_{ij}$$

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \text{ with } \tau_1 + \tau_2 + \tau_3 + \tau_4 = 0$$

```
> x
 [1] 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 2 2 2 2 2 2 2 2 2 2
> y
 [1] 16.85 16.40 17.21 16.35 16.52 17.04 16.96 17.15 16.59 16.57 16.62
[12] 16.75 17.37 17.12 16.98 16.87 17.34 17.02 17.08 17.27 16.58 16.82
[23] 16.51 17.24 16.86 16.52 16.91 16.99 16.94 16.67
```

Now let's consider in R how to build the model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$

```
> x1=c(rep(1,10),rep(0,10),rep(0,10))
> x1
 [1] 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
> x2=c(rep(0,10),rep(0,10),rep(1,10))
> x2
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1
> lm(y~x1+x2)

Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
     17.042       -0.278       -0.238
```

```
> summary(lm(y~x1+x2))

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4220 -0.2165  0.0270  0.1935  0.4460

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.04200    0.08493 200.664   <2e-16 ***
x1          -0.27800    0.12011  -2.315   0.0285 *
x2          -0.23800    0.12011  -1.982   0.0578 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2686 on 27 degrees of freedom
Multiple R-squared:  0.1883,    Adjusted R-squared:  0.1282
F-statistic: 3.132 on 2 and 27 DF,  p-value: 0.05982
```

What are the null hypothesis?

What is the null hypothesis for this test?

Now let's consider how to build the model $y_{ij} = \mu_i + \epsilon_{ij}$

```
> x1=c(rep(1,10),rep(0,10),rep(0,10))
> x1
 [1] 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
> x2=c(rep(0,10),rep(1,10),rep(0,10))
> x2
 [1] 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0
> x3=c(rep(0,10),rep(0,10),rep(1,10))
> x3
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1

> lm(y~x1+x2+x3-1)

Call:
lm(formula = y ~ x1 + x2 + x3 - 1)

Coefficients:
   x1     x2     x3
16.76  17.04  16.80
```

Consistent with the previous model!

How to Handle The Effects Model

$y_{ij} = \mu + \tau_i + \epsilon_{ij}$, with $\tau_1 + \tau_2 + \tau_3 = 0$

We can write it as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

0

0 1 ... 1

```
> x0=c(rep(1,30,),0)
> x1=c(rep(1,10),rep(0,10),rep(0,10),1)
> x1
 [1] 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
> x2=c(rep(0,10),rep(1,10),rep(0,10),1)
> x2
 [1] 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1
> x3=c(rep(0,10),rep(0,10),rep(1,10),1)
> x3
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1
> y=c(y,0)
> y
 [1] 16.85 16.40 17.21 16.35 16.52 17.04 16.96 17.15 16.59 16.57 16.62
[12] 16.75 17.37 17.12 16.98 16.87 17.34 17.02 17.08 17.27 16.58 16.82
[23] 16.51 17.24 16.86 16.52 16.91 16.99 16.94 16.67  0.00
> lm(y~x0+x1+x2+x3-1)

Call:
lm.default(formula = y ~ x0 + x1 + x2 + x3 - 1)

Coefficients:
    x0       x1       x2       x3
16.870   -0.106    0.172   -0.066
```

Consistent with the previous model!

Why Coding Matters

- Categorical predictors must be coded
- Choice of coding changes the meaning of the intercept and coefficients
- Predictions do not change with different codings, but interpretations do
- Coding affects hypothesis tests

Coding Methods

- Dummy coding
- Effect (Sum-to-Zero) Coding
- Cell Means Coding (No Intercept)
- Orthogonal Polynomial Contrasts
- Helmert Coding

Orthogonal Polynomial Contrasts

- Components are orthogonal (uncorrelated)
- The construction is a little complicated, but we will see some examples

Example with 3 levels

Level (x)	0	1	2
Linear (unscaled)	-1	0	1
Quadratic (unscaled)	1	-2	1

Orthogonal Polynomial Contrasts

Example with 5 levels

Level (x)	0	1	2	3	4
Linear (unscaled)	-2	-1	0	1	2
Quadratic (unscaled)	2	-1	-2	-1	2
Cubic (unscaled)	-1	2	0	-2	1
Quartic (unscaled)	1	-4	6	-4	1

- Model: $y = \mu + \beta_L \cdot \text{Linear} + \beta_Q \cdot \text{Quadratic} + \beta_C \cdot \text{Cubic} + \ldots + \varepsilon$
- t-tests on β_L, β_Q, … answer “is there a linear trend?”, “is there curvature?”, etc.

Helmert Contrasts

- Reparameterize a K-level factor into K−1 sequential comparisons.
- Each contrast compares level j+1 to the mean of levels 1..j
- Orthogonal when group sizes are equal. Intercept remains the overall mean.

	[,1]	[,2]	[,3]
L1	-1	-1	-1
L2	1	-1	-1
L3	0	2	-1
L4	0	0	3

Interpretation and when to use

Test if the next level differs from the average of previous levels.

Prefer when no single baseline is special and you want orthogonal decomposition

Advantages: Orthogonal. Symmetric, no baseline.

Disadvantages: Depends on level order. Less intuitive than vs control.