Analysis of Covariance
Regression with Dummy Variables

Intro

- Consider a SLR model: $Y = \beta_0 + \beta_1 x + \varepsilon$. Suppose x=height, Y=weight.

  Question: Does the country of birth influence this equation/relationship?

  For example, take 2 countries: US and China.

  Data set will contain 3 variables: Y=weight, x1=height, x2 = country.

  X2: 0 if a person born in the US, 1 if the person born in China.

- We call X2 **"DUMMY VARIABLE".** It is really an indicator.

- Include x2 in the regression model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- To answer the question regarding possible difference in relation between height and weight based on the country of birth, we test if Y is affected by country, meaning if Y is associated with X2.

Testing for dummy variable

- Test Ho: $\beta_2 = 0$ versus Ha: $\beta_2 \neq 0$

  Test is the usual t-test (or F test) for the significance of a regression coefficient.

- If we do not reject Ho, our model is: $\hat{Y} = \hat{\beta}_o + \hat{\beta}_1 x_1$. This model is the same for both countries.

- If we reject Ho, our regression model is: $\hat{Y} = \hat{\beta}_o + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$.

  If Ho rejected we get different models for the two countries:

  Model for the USA, x2= 0: $\hat{Y} = \hat{\beta}_o + \hat{\beta}_1 x_1$

  Model for China, x2=1: $\hat{Y} = \hat{\beta}_o + \hat{\beta}_1 x_1 + \hat{\beta}_2$ = $\hat{\beta}_o + \hat{\beta}_2 + \hat{\beta}_1 x_1$

  The model equations are parallel, but differ in intercept.

Checking slopes

- **To check if the two models also have different slopes use:**

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon = (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) x_1 + \varepsilon.$$

The term containing $x_1 x_2$ is called "the interaction term".
The interaction is between x1 and x2.

- **Assuming that intercepts are different, to check if there is a significant interaction between x1 and x2:**

Test H*o: $\beta_3 = 0$ versus H*a: $\beta_3 \neq 0$

If we reject Ho*, then we conclude $\beta_3 \neq 0$. Then both slopes are intercepts are different for the two models.

If we do not reject H*o, we conclude the models have different intercepts, but the same slopes.

Checking slopes, contd.

- We can also try another path:

To start with checking if including x2 improves the model, we have to compare a simple model $Y = \beta_0 + \beta_1 x_1 + \varepsilon$
To the complex model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$

Ho: $\beta_2 = \beta_3 = 0$ versus Ha: at least one of β_3 or $\beta_2 \neq 0$

Partial (nested) F test:

$$F = \frac{(SSE_S - SSE_C)/(df_s - df_c)}{SSE_C/df_c}$$

If n observations, $df_s = n - 2, df_c = n - 4$.

If we reject Ho, then β_3 or $\beta_2 \neq 0$, so then we check if $\beta_3 \neq 0$ or $\beta_2 \neq 0$ or both.

Testing contd.

We will need to test

Test H*o: $\boldsymbol{\beta_3 = 0}$ versus H*a: $\boldsymbol{\beta_3 \neq 0}$

If we reject Ho*, then we conclude $\boldsymbol{\beta_3 \neq 0}$. ETC.

Suppose, we concluded that $\boldsymbol{\beta_2 \neq 0 \; and \;\; \beta_3 = 0}$. Then, the models will have different intercept and same slopes.

Model for the USA, x2=0: $\widehat{Y} = \widehat{\boldsymbol{\beta}}_o + \widehat{\boldsymbol{\beta}}_1 \boldsymbol{x}_1$

Model for China, x2=1: $\widehat{Y} = \widehat{\boldsymbol{\beta}}_o + \widehat{\boldsymbol{\beta}}_2 + \widehat{\boldsymbol{\beta}}_1 \boldsymbol{x}_1$

If we conclude that both $\boldsymbol{\beta_2 \neq 0 \; and \;\; \beta_3 \neq 0}$ then the models will have different slopes and intercepts.

Model for the USA, x2=0: $\widehat{Y} = \widehat{\boldsymbol{\beta}}_o + \widehat{\boldsymbol{\beta}}_1 \boldsymbol{x}_1$

Model for China, x2=1: $\widehat{Y} = \widehat{\boldsymbol{\beta}}_o + \widehat{\boldsymbol{\beta}}_2 + (\widehat{\boldsymbol{\beta}}_1 + \widehat{\boldsymbol{\beta}}_3) \boldsymbol{x}_1$

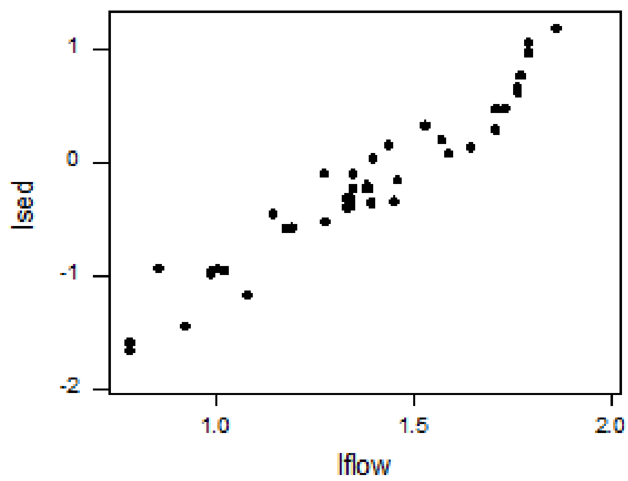
Example

We studied relationship between (log) sediment and log (flow) for two creeks in the Lake Tahoe basin: Gray Creek and Bronco Creek. Bronco Creek is in the area with little human interference, while Gray Creek is in an area more influenced by people. We would like to know if the relationship between flow and sediment is the same for the two creeks. We will use a dummy variable Z= 1 for Gray, 0 for Bronco.

DATA SET IS: covgraybronco.MPJ

The data for flow is lflow=logarithm(base 10) of the flow measured in ft^3/s and for sediment is lsed=logarithm(base 10) of sec

Scatter plot of lflow versus lsed shows a linear relationship



The model equation is: lsed=β_0+ $\beta_1$lflow+ β_2Z+ β_3Z*lflow+ε (Complex model)

First, I find out if: β_2=0 and β_3=0. In this case simple model: lsed=β_0+ $\beta_1$lflow+ε
USE significance level, say 0.1.

Test H_0: β_2=0 and β_3=0 versus H_a: at least one of β_2 and β_3 is not zero

COMPLEX MODEL
The regression equation is lsed = - 2.89 + 1.91 lflow - 0.690 z + 0.533 lflowz
Analysis of Variance

Source	DF	SS	MS	F	P
Residual Error	37	1.2171	0.0329		

SIMPLE MODEL WITH NO DUMMY TERMS
The regression equation is lsed = - 3.30 + 2.25 lflow
Analysis of Variance

Source	DF	SS	MS	F	P
Residual Error	39	1.420	0.036		

$$df_c = n - 4 = 41 - 4 = 37$$
$$df_s = 41 - 2 = 39$$

$$F = \frac{(SSE_S - SSE_C)/(df_s - df_c)}{SSE_C/df_c} = \frac{(1.42 - 1.2171)/2}{1.2171/37} = 3.084.$$

P-value =P($F_{2,\,37}$ > 3.084) ≈ 0.06 < 0.1, reject Ho. Conclusion: At least one of β_2 and β_3 is not zero

Test H_0: $\beta_3=0$ versus H_a: $\beta_3 \neq 0$

COMPLEX MODEL: $lsed=\beta_0+\beta_1 lflow+\beta_2 Z+\beta_3 Z*lflow+\varepsilon$

The regression equation is
lsed = - 2.89 + 1.91 lflow - 0.690 z + 0.533 lflow*z

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	18.3251	6.1084	185.69	0.000
Residual error	37	1.2171			

SIMPLE MODEL: $lsed=\beta_0+\beta_1 lflow+\beta_2 Z+\varepsilon$

The regression equation is
lsed = - 3.30 + 2.24 lflow + 0.0203 z

Analysis of Variance

Source	DF	SS	MS	F	P
Residual Error	38	1.4161	0.0373		

$$F=\frac{(SSE_S-SSE_C)/(df_s-df_c)}{SSE_C/df_c}=\frac{(1.4161\ -1.2171)/1}{1.2171/37}=6.05$$

P-value < 0.1, reject Ho. Conclude $\beta_3 \neq 0$

Test H_0: $\beta_2=0$ versus H_a: $\beta_2 \neq 0$

COMPLEX MODEL: $lsed=\beta_0+ \beta_1 lflow+ \beta_2 Z+ \beta_3 Z*lflow+\varepsilon$

```
 The regression equation is
lsed = - 2.89 + 1.91 lflow - 0.690 z + 0.533 lflowz
 Analysis of Variance
 Source            DF      SS           MS          F         P
Residual Error      37    1.2171       0.0329
```

SIMPLE MODEL: $lsed=\beta_0+ \beta_1 lflow+ \beta_3 Z*lflow+\varepsilon$

The regression equation is
lsed = - 3.25 + 2.19 lflow + 0.0384 lflow*z

```
 Analysis of Variance
 Source                DF      SS       MS          F         P
Residual Error         38    1.3965     0.0367
```

$$F=\frac{(SSE_S-SSE_C)/(df_s-df_c)}{SSE_C/df_c}=\frac{(1.3965\ -1.2171)/1}{1.2171/37}=5.45$$

P-value < 0.1, reject Ho. Conclude $\beta_3 \neq 0$

Conclusion of example

Since both β_2 and β_3 are not zero, then models for the two creeks will differ in slope and intercept.

General model: lsed = -2.89 + 1.91 lflow – 0.69 z + 0.533 lflow*z

Model for Gray creek, z=1: $\widehat{lsed}$= - 3.58 + 2.443 lflow.

Model for Bronco creek: $\widehat{lsed}$= - 2.89 + 1.91 lflow.