Normal Linear Model

STA211/442 Fall 2013

Suggested Reading

- Davison's *Statistical Models*, Chapter 8
- The general mixed linear model is defined in Section 9.4, where it is first applied.

General Mixed Linear Model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$$

- $\mathbf{X}$ is an $n \times p$ matrix of known constants
- $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants.
- $\mathbf{Z}$ is an $n \times q$ matrix of known constants
- $\mathbf{b} \sim N_q(\mathbf{0}, \boldsymbol{\Sigma}_b)$ with $\boldsymbol{\Sigma}_b$ unknown
- $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, where $\sigma^2 > 0$ is an unknown constant.

Fixed Effects Linear Regression

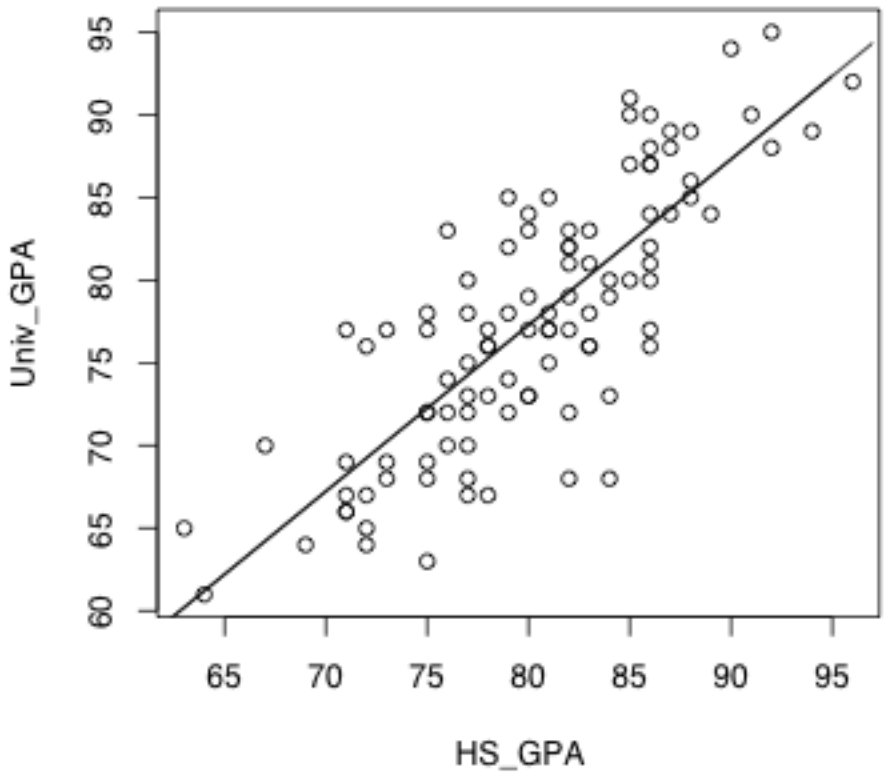
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\mathbf{X}$ is an $n \times p$ matrix of known constants
- $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants
- $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, where $\sigma^2 > 0$ is an unknown constant.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad \widehat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

$$\mathbf{e} = (\mathbf{Y} - \widehat{\mathbf{Y}})$$

“Regression” Line



Regression Means Going Back

- Francis Galton (1822-1911) studied "Hereditary Genius" (1869) and other traits
- Heights of fathers and sons
  - Sons of the tallest fathers tended to be taller than average, but shorter than their fathers
  - Sons of the shortest fathers tended to be shorter than average, but taller than their fathers
- This kind of thing was observed for lots of traits.
- Galton was deeply concerned about "regression to mediocrity."

Measure the same thing twice, with error

$$\begin{aligned} Y_1 &= X + e_1 \\ Y_2 &= X + e_2 \end{aligned}$$

$X \sim N(\mu, \sigma_x^2)$

e_1 and e_2 independent $N(0, \sigma_e^2)$

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_x^2 + \sigma_e^2 & \sigma_x^2 \\ \sigma_x^2 & \sigma_x^2 + \sigma_e^2 \end{bmatrix} \right)$$

Conditional distribution of Y_2 given $Y_1=y_1$ for a general bivariate normal

$$N\left(\mu_2 + \frac{\sigma_2}{\sigma_1}\rho(y_1-\mu_1), (1-\rho^2)\sigma_2^2\right)$$

$$= \quad N\left(\mu + \rho(y_1-\mu), (1-\rho^2)(\sigma_x^2+\sigma_e^2)\right)$$

So $E(Y_2|Y_1=y_1) = \mu + \rho(y_1-\mu)$,
where $\rho = \frac{\sigma_x^2}{\sigma_x^2+\sigma_e^2}$.

$$E(Y_2|Y_1 = y_1) = \mu + \rho(y_1 - \mu)$$

- If y_1 is above the mean, average y_2 will also be above the mean
- But only a fraction (rho) as far above as y_1.
- If y_1 is below the mean, average y_2 will also be below the mean
- But only a fraction (rho) as far below as y_1.

- This exactly the “regression toward the mean” that Galton observed.

Regression toward the mean

- Does not imply systematic change over time
- Is a characteristic of the bivariate normal and other joint distributions

- Can produce very misleading results, especially in the evaluation of social programs

Regression Artifact

- Measure something important, like performance in school or blood pressure.
- Select an extreme group, usually those who do worst on the baseline measure.
- Do something to help them, and measure again.

$$E(Y_2|Y_1 = y_1) = \mu + \rho(y_1 - \mu)$$

- If the treatment does **nothing**, they are expected to do worse than average, but better than they did the first time – completely artificial!

A simulation study

- Measure something twice with error: 500 observations
- Select the best 50 and the worst 50
- Do two-sided matched t-tests at alpha = 0.05

- What proportion of the time do the worst 50 show significant average improvement?
- What proportion of the time do the best 50 show significant average deterioration?

```
> sig2x = 10; sig2e = 10; n = 500; set.seed(9999)
> X = rnorm(n,100,sqrt(sig2x))
> e1 = rnorm(n,0,sqrt(sig2e)); e2 = rnorm(n,0,sqrt(sig2e))
> Y1 = X+e1; Y2 = X+e2; D = Y2-Y1 # D measures "improvement"
> low50 = D[rank(Y1)<=50]; hi50 = D[rank(Y1)>450]
> t.test(low50)
```

```
One Sample t-test

data:  low50
t = 7.025, df = 49, p-value = 6.068e-09
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.874234 5.177526
sample estimates:
mean of x
  4.02588
```

```
> t.test(hi50)


One Sample t-test

data:  hi50
t = -5.3417, df = 49, p-value = 2.373e-06
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -4.208760 -1.907709
sample estimates:
mean of x
-3.058234
```

```
> t.test(low50)$p.value
[1] 6.068008e-09
> t.test(low50)$estimate
mean of x
  4.02588
> t.test(low50)$p.value<0.05 && t.test(low50)$estimate>0
[1] TRUE
```

```
> # Estimate probability of artificial difference
> M = 100 # Monte Carlo sample size
> dsig = numeric(M)
> for(i in 1:M)
+     {
+     X = rnorm(n,100,sqrt(sig2x))
+     e1 = rnorm(n,0,sqrt(sig2e)); e2 = rnorm(n,0,sqrt(sig2e))
+     Y1 = X+e1; Y2 = X+e2; D = Y2-Y1 # D measures "improvement"
+     low50 = D[rank(Y1)<=50]
+     ttest = t.test(low50)
+     dsig[i] = ttest$p.value<0.05 && ttest$estimate>0
+     }
> dsig
```

```
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[40] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[79] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Summary

- Source of the term “Regression”
- Regression artifact
  - Very serious
  - People keep re-inventing the same mistake
  - Can’t really blame the policy makers
  - At least the statistician should be able to warn them
  - The solution is random assignment
  - Taking difference from a baseline measurement may still be useful

Multiple Linear Regression

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

Statistical **MODEL**

- There are *p-1* explanatory variables
- For each *combination* of explanatory variables, the conditional distribution of the response variable *Y* is normal, with constant variance
- The conditional population mean of *Y* depends on the *x* values, as follows:

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$

Control means hold constant

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$\frac{\partial}{\partial x_3} E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_3$$

So β_3 is the rate at which $E[Y|\boldsymbol{x}]$ changes as a function of x_3 with all other variables held constant at fixed levels.

Increase x_3 by one unit holding other variables constant

$$
\begin{aligned}
& \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad + \beta_3 (x_3 + 1) \quad + \beta_4 x_4 \\
- \quad & (\beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad + \beta_3 \, x_3 \qquad\quad + \beta_4 x_4) \\
= \quad & \beta_3 (x_3 + 1) - \beta_3 x_3 \\
= \quad & \beta_3
\end{aligned}
$$

So β_3 is the amount that $E[Y|\mathbf{x}]$ changes when x_3 is increased by one unit and all other variables are held constant at fixed levels.

It's model-based control

- To "hold x_1 constant" at some particular value, like x_1=14, you don't even need data at that value.
- Ordinarily, to estimate $E(Y|X_1=14,X_2=x)$, you would need a lot of data at X_1=14.
- But look:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot 14 + \widehat{\beta}_2 x$$

Statistics b estimate parameters beta

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$\widehat{Y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$$

Categorical Explanatory Variables

- X=1 means Drug, X=0 means Placebo
- Population mean is $E[Y|X=x] = \beta_0 + \beta_1 x$
- For patients getting the drug, mean response is $E[Y|X=1] = \beta_0 + \beta_1$
- For patients getting the placebo, mean response is $E[Y|X=0] = \beta_0$

Sample regression coefficients for a binary explanatory variable

- X=1 means Drug, X=0 means Placebo
- Predicted response is $\widehat{Y} = b_0 + b_1 x$
- For patients getting the drug, predicted response is

$$\widehat{Y} = b_0 + b_1 = \overline{Y}_1$$

- For patients getting the placebo, predicted response is

$$\widehat{Y} = b_0 = \overline{Y}_0$$

Regression test of b_1

- Same as an independent t-test
- Same as a oneway ANOVA with 2 categories
- Same t, same F, same p-value.

Drug A, Drug B, Placebo

- x_1 = 1 if Drug A, Zero otherwise
- x_2 = 1 if Drug B, Zero otherwise
- $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- Fill in the table

Group	x_1	x_2	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$
A			$\mu_1 =$
B			$\mu_2 =$
Placebo			$\mu_3 =$

Drug A, Drug B, Placebo

- x_1 = 1 if Drug A, Zero otherwise
- x_2 = 1 if Drug B, Zero otherwise
- $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Group	x_1	x_2	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	0	0	$\mu_3 = \beta_0$

Regression coefficients are *contrasts* with the category that has no indicator – the *reference* category

Indicator dummy variable coding with intercept

- Need p-1 indicators to represent a categorical explanatory variable with p categories
- If you use p dummy variables, trouble
- Regression coefficients are **contrasts** with the category that has no indicator
- Call this the **reference category**

Now add a quantitative variable (covariate)

- x_1 = Age
- x_2 = 1 if Drug A, Zero otherwise
- x_3 = 1 if Drug B, Zero otherwise
- $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Drug	x_2	x_3	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$(\beta_0 + \beta_2) + \beta_1 x_1$
B	0	1	$(\beta_0 + \beta_3) + \beta_1 x_1$
Placebo	0	0	$\beta_0 \quad + \beta_1 x_1$

Parallel regression lines

Effect coding

- *p-1* dummy variables for *p* categories
- Include an intercept
- Last category gets -1 instead of zero
- What do the regression coefficients mean?

Group	x_1	x_2	$E[Y\|\boldsymbol{X}=\boldsymbol{x}]=\beta_0+\beta_1 x_1+\beta_2 x_2$
A	1	0	$\mu_1=\beta_0+\beta_1$
B	0	1	$\mu_2=\beta_0+\beta_2$
Placebo	-1	-1	$\mu_3=\beta_0-\beta_1-\beta_2$

Meaning of the regression coefficients

Group	x_1	x_2	$E[Y \mid \boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	-1	-1	$\mu_3 = \beta_0 - \beta_1 - \beta_2$

$$\mu = \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) = \beta_0$$

The grand mean

With effect coding

- Intercept is the *Grand Mean*
- Regression coefficients are deviations of group means from the grand mean.
- They are the non-redundant *effects*.
- Equal population means is equivalent to zero coefficients for all the dummy variables
- Last category is not a reference category

Group	x_1	x_2	$E[Y \mid \boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	-1	-1	$\mu_3 = \beta_0 - \beta_1 - \beta_2$

Add a covariate: Age = x_1

Group	x_2	x_3	$E[Y \mid \boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$\mu_1 = \beta_0 + \beta_2 \quad + \beta_1 x_1$
B	0	1	$\mu_2 = \beta_0 + \beta_3 \quad + \beta_1 x_1$
Placebo	-1	-1	$\mu_3 = \beta_0 - \beta_2 - \beta_3 + \beta_1 x_1$

Regression coefficients are deviations from the average conditional population mean (conditional on x_1).

So if the regression coefficients for all the dummy variables equal zero, the categorical explanatory variable is unrelated to the response variable, controlling for the covariate(s).

Effect coding is very useful when there is more than one categorical explanatory variable and we are interested in *interactions* --- ways in which the relationship of an explanatory variable with the response variable *depends* on the value of another explanatory variable.

Interaction terms correspond to products of dummy variables.

Analysis of Variance

And testing

Analysis of Variance

- Variation to explain: **Total Sum of Squares**

$$\text{SSTO} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

- Variation that is still unexplained: **Error Sum of Squares**

$$\text{SSE} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

- Variation that is explained: **Regression** (or Model) **Sum of Squares**

$$\text{SSR=SSTO-SSE} = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$$

ANOVA Summary Table

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	$p-1$	SSR	$MSR = SSR/(p-1)$	$F = \frac{MSR}{MSE}$	p-value
Error	$n-p$	SSE	$MSE = SSE/(n-p)$		
Total	$n-1$	$SSTO$			

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

Proportion of variation in the response variable that is explained by the explanatory variables

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SSTO}}$$

Hypothesis Testing

- Overall F test for all the explanatory variables at once,
- t-tests for each regression coefficient: Controlling for all the others, does that explanatory variable matter?
- Test a collection of explanatory variables controlling for another collection,
- Most general: Testing whether sets of linear combinations of regression coefficients differ from specified constants.

Controlling for mother's education and father's education, are (any of) total family income, assessed value of home and total market value of all vehicles owned by the family related to High School GPA?

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_5 x_5$$

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$$

(A false promise because of measurement error in education)

Full vs. Reduced Model

- You have 2 sets of variables, A and B
- Want to test B controlling for A
- Fit a model with both A and B: Call it the **Full Model**
- Fit a model with just A: Call it the **Reduced Model**
- It's a likelihood ratio test (exact)

$$R_F^2 \geq R_R^2$$

When you add *r* more explanatory variables, R^2 can only go up

- By how much? Basis of F test.

$$F = \frac{(SSR_F - SSR_R)/r}{MSE_F}$$

- Denominator MSE = SSE/df for full model.
- Anything that reduces MSE of full model increases F
- Same as testing H_0: All betas in set B (there are *r* of them) equal zero

General H_0: **Lβ = h** (**L** is rxp, row rank r)

$$
\begin{aligned}
F &= \frac{(SSR_F - SSR_R)/r}{MSE_F} \\
&= \frac{(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})}{r\,MSE_F}
\end{aligned}
$$

Distribution theory is well within reach.

Distribution theory for tests, confidence intervals and prediction intervals

Remember

- If $\mathbf{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(k)$
- Zero covariance implies independence for the multivariate normal.
- For the regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$,

  $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \sim N_p\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$

  $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ is multivariate normal.

  $\text{SSE} = \mathbf{e}'\mathbf{e}$.

It's like the independence of $\overline{X}$ and S^2

$$C(\mathbf{e}, \widehat{\boldsymbol{\beta}}) = E(\mathbf{e}\widehat{\boldsymbol{\beta}}') - E(\mathbf{e})E(\widehat{\boldsymbol{\beta}}') = E(\mathbf{e}\widehat{\boldsymbol{\beta}}') = \mathbf{0}$$

So $\widehat{\boldsymbol{\beta}}$ AND $\mathbf{e}$ are independent.

So $\widehat{\boldsymbol{\beta}}$ AND $SSE = \mathbf{e}'\mathbf{e}$ are independent.

Test statistic is

$$F = \frac{(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})}{r\,MSE}$$

Numerator and denominator are independent.

Independent chi-squares

$$F = \frac{(\mathbf{L}\widehat{\boldsymbol{\beta}}-\mathbf{h})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}}-\mathbf{h})/r}{MSE}$$

If $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$ is true, then $\mathbf{L}\widehat{\boldsymbol{\beta}} \sim N_r\left(\mathbf{h}, \mathbf{L}\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\right)$, and

$$\begin{aligned} & (\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})'(\mathbf{L}\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h}) \\ = \quad & \frac{1}{\sigma^2}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h}) \sim \chi^2(r). \end{aligned}$$

So if $SSE/\sigma^2 \sim \chi^2(n-p)$, we have a ratio of independent chi-squares, each divided by its degrees of freedom.

SSE/σ^2 is chi-squared

$$\mathbf{Y} \sim N_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right), \text{ so}$$

$$\begin{aligned} & (\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\left(\sigma^2\mathbf{I}_n\right)^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \\ = \quad & \frac{1}{\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \\ \sim \quad & \chi^2(n) \end{aligned}$$

This is the sum of two independent random variables, one of which is chi-squared.

Add and subtract $\widehat{\mathbf{Y}}$

$$
\begin{aligned}
& \frac{1}{\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \\
= \quad & \frac{1}{\sigma^2}(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}} \;+\; \mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}} \;+\; \mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}) \\
= \quad & \ldots \\
= \quad & \frac{1}{\sigma^2}(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) + \;0\; + \frac{1}{\sigma^2}(\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}) \\
= \quad & \frac{1}{\sigma^2}\mathbf{e}'\mathbf{e} + (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\left(\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}\right)(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})
\end{aligned}
$$

Because $\widehat{\boldsymbol{\beta}} \sim N_p\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$, the second term is $\chi^2(p)$.

Have $\frac{1}{\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) = \frac{1}{\sigma^2}\mathbf{e}'\mathbf{e} + (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\left(\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X}\right)(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$

- $C = A + B$
- A and B are independent.
- $C \sim \chi^2(n)$
- $B \sim \chi^2(p)$
- So by a homework problem, $A = \frac{SSE}{\sigma^2} \sim \chi^2(n-p)$

$$
\begin{aligned}
F &= \frac{(\mathbf{L}\widehat{\boldsymbol{\beta}}-\mathbf{h})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}}-\mathbf{h})/r}{MSE} \\
&= \frac{\frac{1}{\sigma^2}(\mathbf{L}\widehat{\boldsymbol{\beta}}-\mathbf{h})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}}-\mathbf{h})/r}{\frac{1}{\sigma^2}\mathbf{e}'\mathbf{e}/(n-p)}
\end{aligned}
$$

Notice the similarity: $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$

$$W_n = (\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})' \left(\mathbf{L}\widehat{\mathbf{V}}_n\mathbf{L}'\right)^{-1} (\mathbf{L}\widehat{\boldsymbol{\theta}}_n - \mathbf{h})$$

$$F = \frac{(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})/r}{MSE}$$

$$= (\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})'(\mathbf{L}\widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})/r$$

Prediction interval

- Given a sample of size n, have $\widehat{\boldsymbol{\beta}}$ and MSE
- Have a vector of explanatory variables $\mathbf{x}_{n+1}$
- Define $\widehat{Y}_{n+1} = \mathbf{x}'_{n+1}\widehat{\boldsymbol{\beta}}$
- Want an interval that will contain Y_{n+1} with high probability, say 0.95.
- Call it a 95% *prediction interval*.

$$T = \frac{Y_{n+1} - \mathbf{x}'_{n+1}\widehat{\boldsymbol{\beta}}}{\sqrt{MSE(1 + \mathbf{x}'_{n+1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{n+1})}} \sim t(n-p)$$

$$\begin{aligned} 1-\alpha &= Pr\{-t_{\alpha/2} < T < t_{\alpha/2}\} \\ &= Pr\left\{\widehat{Y}_{n+1} - t_{\alpha/2}\sqrt{MSE(1 + \mathbf{x}'_{n+1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{n+1})}\right. \\ &\qquad \left. < Y_{n+1} < \widehat{Y}_{n+1} + t_{\alpha/2}\sqrt{MSE(1 + \mathbf{x}'_{n+1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{n+1})}\right\} \end{aligned}$$

or

$$\mathbf{x}'_{n+1}\widehat{\boldsymbol{\beta}} \pm t_{\alpha/2}\sqrt{MSE(1 + \mathbf{x}'_{n+1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{n+1})}$$

Back to full versus reduced model

$$R_F^2 \geq R_R^2$$

$$F = \frac{(SSR_F - SSR_R)/r}{MSE_F}$$

F test is based not just on change in R^2, but upon

$$a = \frac{R_F^2 - R_R^2}{1 - R_R^2}$$

Increase in explained variation expressed as a fraction of the variation that the reduced model does *not* explain.

$$F = \left(\frac{n-p}{r}\right)\left(\frac{a}{1-a}\right)$$

- For any given sample size, the bigger *a* is, the bigger *F* becomes.
- For any a ≠0, *F* increases as a function of *n.*
- So you can get a large *F* from strong results and a small sample, or from weak results and a large sample.

Can express *a* in terms of *F*

$$a = \frac{rF}{n - p + rF}$$

- Often, scientific journals just report *F*, numerator df = *r*, denominator df = *(n-p),* and a *p*-value.
- You can tell if it's significant, but how strong are the results? Now you can calculate it.
- This formula is less prone to rounding error than the one in terms of R-squared values

When you add explanatory variables to a model (with observational data)

- Statistical significance can appear when it was not present originally
- Statistical significance that was originally present can disappear
- Even the signs of the b coefficients can change, reversing the interpretation of how their variables are related to the response variable.
- Technically, omitted variables cause regression coefficients to be inconsistent.

A few More Points

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

- Are the x values really constants?
- Experimental versus observational data
- Omitted variables
- Measurement error in the explanatory variables

Recall Double Expectation

$$E\{Y\} = E\{E\{Y|X\}\}$$

E{Y} is a constant. E{Y|X} is a random variable, a function of X.

$$E\{E\{Y|X\}\} = \int E\{Y|X = x\}\, f(x)\, dx$$

Beta-hat is (conditionally) unbiased

$$E\{\widehat{\boldsymbol{\beta}}|\mathbf{X}=\mathbf{x}\}=\boldsymbol{\beta}$$

Unbiased unconditionally, too

$$E\{\widehat{\boldsymbol{\beta}}\}=E\{E\{\widehat{\boldsymbol{\beta}}|\mathbf{X}\}\}=E\{\boldsymbol{\beta}\}=\boldsymbol{\beta}$$

Perhaps Clearer

$$
\begin{aligned}
E\{\widehat{\boldsymbol{\beta}}\} &= E\{E\{\widehat{\boldsymbol{\beta}}|\mathbf{X}\}\} \\
&= \int \cdots \int E\{\widehat{\boldsymbol{\beta}}|\mathbf{X} = \mathbf{x}\}\, f(\mathbf{x})\, d\mathbf{x} \\
&= \int \cdots \int \boldsymbol{\beta}\, f(\mathbf{x})\, d\mathbf{x} \\
&= \boldsymbol{\beta} \int \cdots \int f(\mathbf{x})\, d\mathbf{x} \\
&= \boldsymbol{\beta} \cdot 1 = \boldsymbol{\beta}.
\end{aligned}
$$

Conditional size α test, Critical region *A*

$$Pr\{F \in A | \mathbf{X} = \mathbf{x}\} = \alpha$$

$$\begin{aligned} Pr\{F \in A\} &= \int \cdots \int Pr\{F \in A | \mathbf{X} = \mathbf{x}\} f(\mathbf{x})\, d\mathbf{x} \\ &= \int \cdots \int \alpha f(\mathbf{x})\, d\mathbf{x} \\ &= \alpha \int \cdots \int f(\mathbf{x})\, d\mathbf{x} \\ &= \alpha \end{aligned}$$

Why predict a response variable from an explanatory variable?

- There may be a practical reason for prediction (buy, make a claim, price of wheat).
- It may be "science."

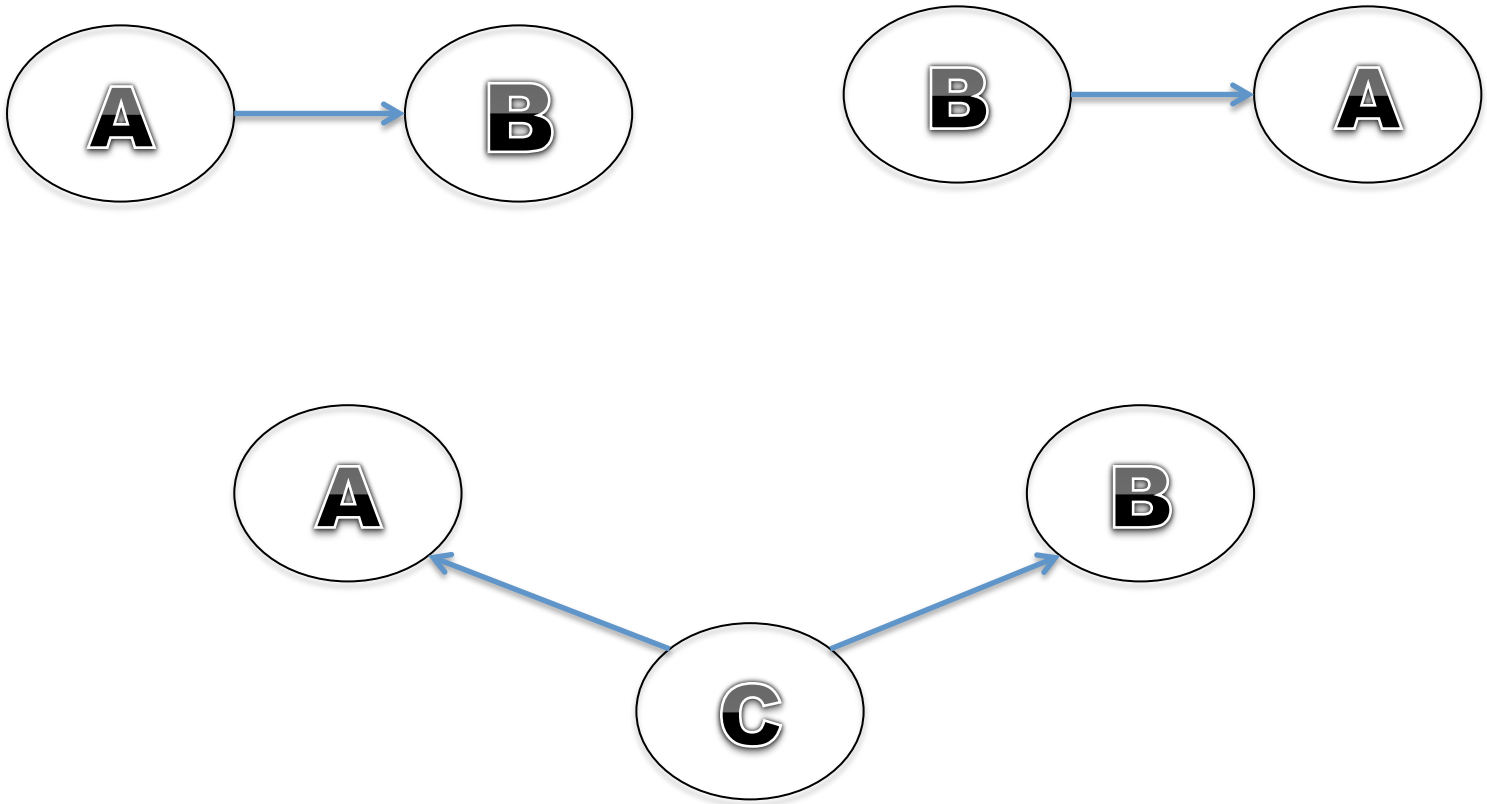
Young smokers who buy contraband cigarettes tend to smoke more.

- What is explanatory variable, response variable?

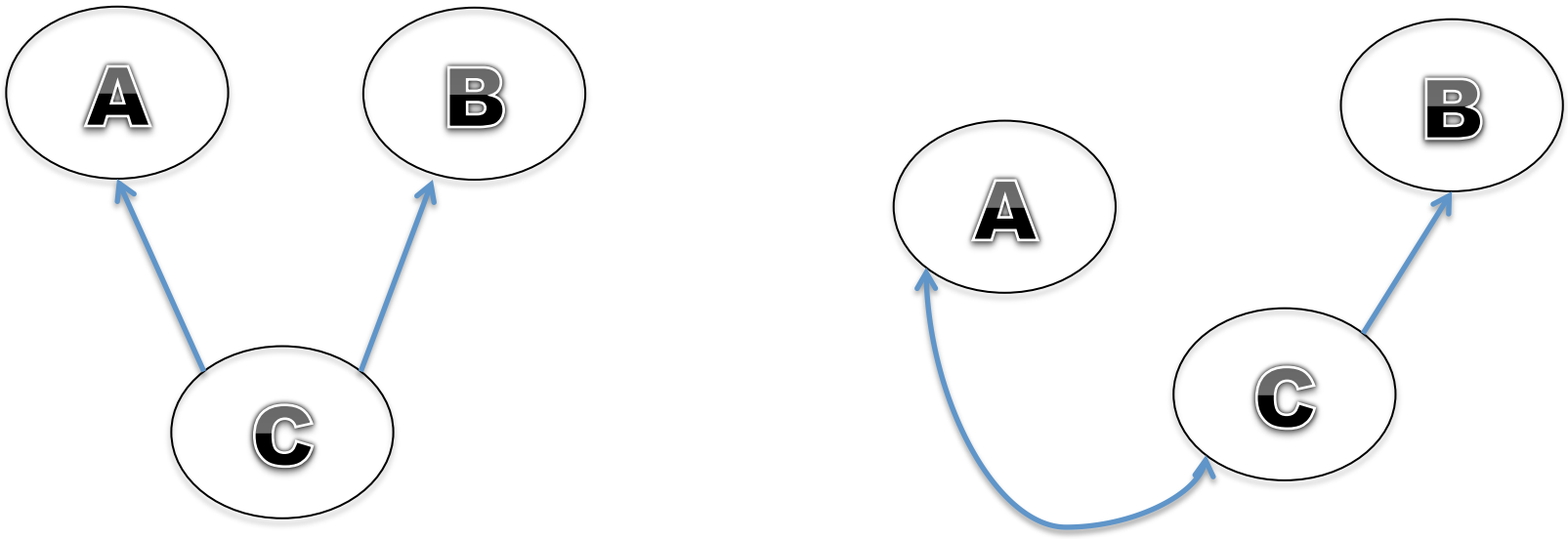
Correlation versus causation

- Model is $Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i$
- It looks like Y is being *produced* by a mathematical function of the explanatory variables plus a piece of random noise.
- And that's the way people often interpret their results.
- People who exercise more tend to have better health.
- Middle aged men who wear hats are more likely to be bald.

Correlation is not the same as causation



Confounding variable: A variable that is associated with both the explanatory variable and the response variable, causing a misleading relationship between them.



Mozart Effect

- Babies who listen to classical music tend to do better in school later on.
- Does this mean parents should play classical music for their babies?
- Please comment. (What is one possible confounding variable?)

Parents' education

- The question is DOES THIS MEAN. Answer the question. Expressing an opinion, yes or no gets a zero unless at least one potential confounding variable is mentioned.

- It may be that it's helpful to play classical music for babies. The point is that this study does not provide good evidence.

Hypothetical study

- Subjects are babies in an orphanage (maybe in Haiti) awaiting adoption in Canada. All are assigned to adoptive parents, but are waiting for the paperwork to clear.
- They all wear headphones 5 hours a day. Randomly assigned to classical, rock, hip-hop or nature sounds. Same volume.
- Carefully keep experimental condition secret from everyone
- Assess academic progress in JK, SJ, Grade 4.
- Suppose the classical music babies do better in school later on. What are some potential confounding variables?

Experimental vs. Observational studies

- **Observational**: Explanatory, response variable just observed and recorded
- **Experimental**: Cases randomly assigned to values of the explanatory variable
- Only a true experimental study can establish a causal connection between explanatory variable and response variable.

- Maybe we should talk about observational vs experimental variables.
- Watch it: Confounding variables can creep back in.

If you ignore measurement error in the explanatory variables

- Disaster if the (true) variable for which you are trying to control is correlated with the variable you're trying to test.
  - Inconsistent estimation
  - Inflation of Type I error rate
- Worse when there's a lot of error in the variable(s) for which you are trying to control.
- Type I error rate can approach one as *n* increases.

Example

- Even controlling for parents' education and income, children from a particular racial group tend to do worse than average in school.

- *Oh really? How did you control for education and income?*
- I did a regression.
- *How did you deal with measurement error?*
- Huh?

Sometimes it's not a problem

- Not as serious for experimental studies, because random assignment erases correlation between explanatory variables.
- For pure prediction (not for understanding) standard tools are fine with observational data.

More about measurement error

- R. J. Carroll et al. (2006) *Measurement Error*
- *in Nonlinear Models*

- W. Fuller (1987) *Measurement error models*.

- P. Gustafson (2004) *Measurement error and misclassification in statistics and epidemiology*

Copyright Information