

# Multiple Predictors in OLS Regression, Lecture 1

Paul E. Johnson<sup>1</sup> <sup>2</sup>

<sup>1</sup>Dept. of Political Science

<sup>2</sup>Dept. of Psychology

# Outline

- 1 Basic assumptions
- 2 Multiple Regression
- 3 New Issues in Multiple Regression
  - Need Matrix Algebra
  - Visualization
  - Multicollinearity
- 4 Hypothesis Testing
  - The Fancy t-test
  - F Test
- 5 Regression and Categorical Predictors
- 6 Appendix: Derivation without Matrices
- 7 Appendix: Matrices
- 8 Appendix: Digression on “Partial Residual” and “Partial Predicted Value”
- 9 Practice Problems

# Outline

- 1 Basic assumptions
- 2 Multiple Regression
- 3 New Issues in Multiple Regression
  - Need Matrix Algebra
  - Visualization
  - Multicollinearity
- 4 Hypothesis Testing
  - The Fancy t-test
  - F Test
- 5 Regression and Categorical Predictors
- 6 Appendix: Derivation without Matrices
- 7 Appendix: Matrices
- 8 Appendix: Digression on “Partial Residual” and “Partial Predicted Value”
- 9 Practice Problems

# The Theoretical model

- An output variable is created as the weighted sum of several “predictors” and a random error term
- Formally

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + e_i \quad (1)$$

- where
  - $i$  is a “row” of data.
  - The same model applies for each row, so  $i \in \{1, \dots, N\}$

# The Critical Assumptions

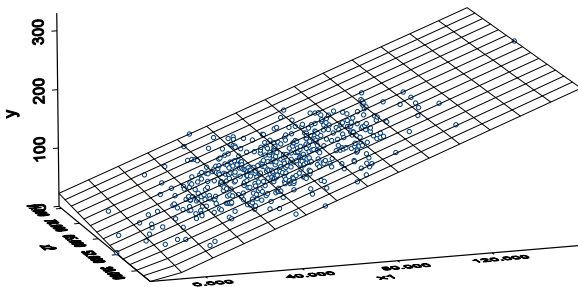
- $\beta_0, \beta_1, \beta_2$  are real-valued constants. I often write  $\beta_j$  for  $j \in 1, \dots, p$ . This case,  $p = 3$ .
- $e_i$  is drawn from a distribution for which  $E[e_i] = 0$  and has constant variance  $\text{Var}[e_i] = \sigma_e^2$ 
  - I don't assume it is  $Normal(0, \sigma_e^2)$  because that is stronger than required, but we will need that—or some other specific distributional assumption—if we do “maximum likelihood analysis”
- None of the error terms are correlated with each other,  $\text{Cov}(e_i, e_j) = 0$  and
- Errors are uncorrelated with the predictors,  $\text{Cov}(x_i, e_i) = 0$

# Picture this in 3d

Multiple Regression

$$y = 10 + 1.5x_1 + 0.5x_2$$

Note:  $\text{std.dev}(e) = 0$ , all points are exactly "in" the plane

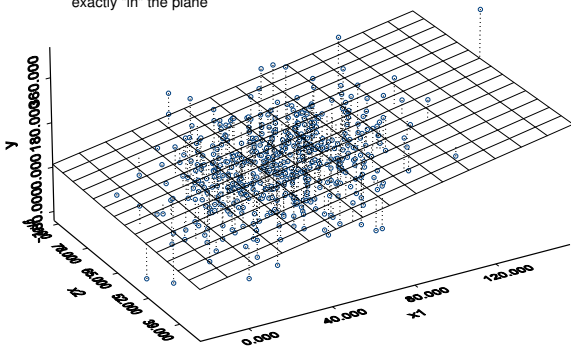


# Or this

Multiple Regression

$$y = 10 + 1.5x_1 + 0.5x_2 + e$$

Note:  $\text{std.dev}(e) = 100$ , not all points are exactly "in" the plane



# Visualization is always a challenge

- Difficult to visualize more dimensions
- In the `rockchalk` package, I have functions like `plotSlopes`, `plotCurves`, and `plotPlane` that are intended to help with that, but we never overcome the problem that we can't visualize 4 or more dimensions

# Outline

- 1 Basic assumptions
- 2 Multiple Regression**
- 3 New Issues in Multiple Regression
  - Need Matrix Algebra
  - Visualization
  - Multicollinearity
- 4 Hypothesis Testing
  - The Fancy t-test
  - F Test
- 5 Regression and Categorical Predictors
- 6 Appendix: Derivation without Matrices
- 7 Appendix: Matrices
- 8 Appendix: Digression on “Partial Residual” and “Partial Predicted Value”
- 9 Practice Problems

# OLS with Many Predictors

- Theory? Just add more predictors.

$$y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \beta_3 X3_i + \beta_4 X4_i + \beta_5 X5_i + \beta_6 X6_i + \beta_7 X7_i + e_i$$

$$E[e_i] = 0, E[e_i^2] = \sigma_e^2.$$

- Goal? Estimates  $\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \dots]$ .
- Calculate predicted values:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X1_i + \hat{\beta}_2 X2_i + \hat{\beta}_3 X3_i + \hat{\beta}_4 X4_i + \hat{\beta}_5 X5_i + \hat{\beta}_6 X6_i + \hat{\beta}_7 X7_i$$

- OLS Minimizes the unweighted Sum of Squared residuals

$$\begin{aligned} \min_{\hat{\beta}} S(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) &= \sum_{i=1}^N (y_i - \hat{y}_i)^2 & (2) \\ &= \sum_{i=1}^N (y_i - \hat{\beta}_0 - \hat{\beta}_1 X1_i - \hat{\beta}_2 X2_i - \dots)^2 \end{aligned}$$

## Compare/contrast regression with one predictor

- Least squares criterion still works!
- Minimizing estimates of  $\hat{\beta}_j$  exist (as long as no predictor is completely redundant).
- If all of the predictors are perfectly uncorrelated, then each coefficient depends only on its own predictor.
- Otherwise (and this is the usual situation)  
Every coefficient *can* depend on data for *all* of the variables.

# Gauss Markov Theorem: OLS is Best Linear Unbiased Estimator (BLUE)

- IF
  - you fit the “right” model,  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + e_i$
  - and the error term assumptions are satisfied,
- THEN The OLS Estimates  $\hat{\beta}_j$  ( $j = 0, 1, 2, \dots, p$ ) are
  - 1 unbiased,  $E[\hat{\beta}_j] = \beta_j$
  - 2 consistent: as  $N$  grows, the expected value of the gap between  $\hat{\beta}_j$  and  $\beta_j$  shrinks
  - 3 efficient: lower variance than any other linear estimator
- $\hat{\beta}_j / \text{std.err}(\hat{\beta}_j)$  is distributed as a  $t$  statistic

# JobPerformance example dataset

Look in my guides/stat/DataSets folder.

- Variables (converted to lower case)

age = employee age;

tenure = years on the job;

female = gender (0 = male, 1 = female);

wbeing = psychological well-being

satis = job satisfaction

jobperf = job performance;

turnover = turnover intentions (0 = no, 1 = yes);

iq = iq score;

# One Predictor: $wbeing \sim satis$

```
summary(mod1 <- lm(wbeing ~ satis, data=dat))
```

```
Call:
lm(formula = wbeing ~ satis, data = dat)

Residuals:
    Min       1Q   Median       3Q      Max
-3.2746 -0.9095  0.0905  0.7254  3.0905

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.08366    0.27485   14.858 < 2e-16 ***
satis         0.36516    0.04503    8.109 4.32e-15 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.158 on 478 degrees of freedom
Multiple R2: 0.1209, Adjusted R2: 0.1191
F-statistic: 65.76 on 1 and 478 DF, p-value: 4.323e-15
```

wbeing ~ satis + jobperf + tenure + age

```
Call:
lm(formula = wbeing ~ satis + jobperf + tenure + age, data = dat)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-3.2248 -0.7518 -0.0114  0.6933  2.7181
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.77194    0.44789   1.723  0.085449 .
satis        0.26371    0.04132   6.382  4.17e-10 ***
jobperf      0.40240    0.03828  10.513 < 2e-16 ***
tenure       0.02088    0.01761   1.186  0.236151
age          0.03391    0.01018   3.332  0.000929 ***
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.032 on 475 degrees of freedom
```

```
Multiple R2: 0.3064, Adjusted R2: 0.3006
```

```
F-statistic: 52.46 on 4 and 475 DF, p-value: < 2.2e-16
```

# Make a Regression Table

	One Predictor		Multiple Predictors	
	Estimate	(S.E.)	Estimate	(S.E.)
(Intercept)	4.084***	(0.275)	0.772	(0.448)
satis	0.365***	(0.045)	0.264***	(0.041)
jobperf	–		0.402***	(0.038)
tenure	–		0.021	(0.018)
age	–		0.034***	(0.010)
N	480		480	
RMSE	1.158		1.032	
$R^2$	0.121		0.306	
adj $R^2$	0.119		0.301	

\* $p \leq 0.05$  \*\*  $p \leq 0.01$  \*\*\*  $p \leq 0.001$

## Other Differences

- Var-Covar matrix larger
- $\hat{\beta}_j$  can “bounce” when other variables are added (multicollinearity)
- $R^2_{adjusted}$ , versus  $R^2$ .
- Reported  $F$  test still not very useful, but we will create a useful  $F$  that compares 2 models.

# Let's Stop And Do Some T-Tests

- Step 1. State the model:  
 $wbeing_i = \beta_0 + \beta_1satis + \beta_2jobperf + \beta_3tenure + \beta_4age + e_i$
- Step 2. Choose one variable, state  $H_0 : \beta_j = ?$ , alt. hypo  $H_A : \beta_j \neq ?$
- Step 3. Sketch t distribution for  $\nu = 480$  degrees of freedom. Are we doing two-tailed test?
- Step 4. Calculate

$$\hat{t}_j = \frac{\hat{\beta}_j - ?}{std.err.(\hat{\beta}_j)}$$

- Step 5. If  $\hat{t}_j$  is in the extreme parts of the tails (according to your sketch in step 3), reject  $H_0$

# Construct the Confidence Interval for one slope in isolation

- Recall CI for a regression coefficient is  $\hat{\beta}_j \pm t_\alpha \times \text{std.err}(\hat{\beta}_j)$
- $t_\alpha$  is the “target value” of the t statistic that you would use when conducting a hypothesis test (depends on  $\nu$ , usually 1.65, 1.98, 2.2)
- If  $\hat{t}_j$  leads us to reject  $H_0$ , we know the CI will not overlap with the null hypo value ?

# Confidence Intervals

```
confint(mod2)
```

	2.5 %	97.5 %
(Intercept)	-0.10815275	1.65203161
satis	0.18250889	0.34490299
jobperf	0.32718702	0.47760555
tenure	-0.01371126	0.05547715
age	0.01391151	0.05390162

Which means that 1) the probability that the true  $\beta_j$  is in that interval is 0.95, or 2) if we repeated this experiment, the probability that an estimate from a sample will fall in that interval is 0.95.

# Var-Covar Matrix Shows Interdependence Among All Estimators

```
round(vcov(mod2), 3)
```

	(Intercept)	satis	jobperf	tenure	age
(Intercept)	0.201	-0.007	-0.008	0.001	-0.003
satis	-0.007	0.002	0.000	0.000	0.000
jobperf	-0.008	0.000	0.001	0.000	0.000
tenure	0.001	0.000	0.000	0.000	0.000
age	-0.003	0.000	0.000	0.000	0.000

- We use these values to conduct the “fancy t-test” (see below).
- The square root of the main diagonal is the standard error column from the regression output.

```
round(sqrt(diag(vcov(mod2))), 3)
```

(Intercept)	satis	jobperf	tenure	age
0.448	0.041	0.038	0.018	0.010

# Outline

- 1 Basic assumptions
- 2 Multiple Regression
- 3 New Issues in Multiple Regression**
  - Need Matrix Algebra
  - Visualization
  - Multicollinearity
- 4 Hypothesis Testing
  - The Fancy t-test
  - F Test
- 5 Regression and Categorical Predictors
- 6 Appendix: Derivation without Matrices
- 7 Appendix: Matrices
- 8 Appendix: Digression on “Partial Residual” and “Partial Predicted Value”
- 9 Practice Problems

# The Design Matrix Has Many Columns

- In R, run a multiple regression, something large like

```
m1 <- lm(y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7, data = dat)
```

- Then run `model.matrix(m1)`. That's the "design matrix", the numeric representation of all variables.
- Result shows that "the intercept" is really a column of 1's, the same for each case, and other predictors

$$X = \begin{bmatrix} \textit{intercept} & X1 & X2 & X3 & X4 & X5 & X6 & X7 \\ 1 & 19 & 1 & 0.1 & 1 & 0 & 22 & 155 \\ 1 & 22 & 2 & 1.1 & 0 & 1 & 42 & 199 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 8 & 4 & 0.2 & 1 & 1 & 77 & 77 \end{bmatrix} \quad (3)$$

# Use Matrix Algebra

- Theoretical model

$$y = X\beta + e \quad (4)$$

- $y$  is an  $(N \times 1)$  column,  $X$  is an  $(N \times p)$  (where  $p = 7 + 1$ ) rectangular matrix,  $e$  is  $(N \times 1)$ .

- I moved the matrix stuff into Appendix 2.

- Results. OLS leads to a symbolic solution:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\widehat{\text{Var}}(\hat{\beta}) = \widehat{\sigma_e^2} (X^T X)^{-1} \quad (5)$$

- Modern computers use advanced, numerically stable formula based on matrix decompositions. At the current time, the  $X^T X$  or  $(X^T X)^{-1}$  matrices are **never** actually calculated.

# What's important in the formula?

- The matrix expression

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- $(X^T X)$  is the “sum of squares and cross-products” matrix, similar to variance of  $x$  in one-predictor model
- In the predicted value formula  $\hat{y} = X\hat{\beta}$ , replace  $\hat{\beta}$ .

$$\hat{y} = X(X^T X)^{-1} X^T y \quad (6)$$

- The matrix  $H = X(X^T X)^{-1} X^T$  is size  $N \times N$ . It serves as a weighting matrix that translates the outcome  $y$  into the predicted values.
  - $H$  is called the “hat matrix”. See why?  $\hat{y} = Hy$

# Plotting Multiple Regressions in 2 Dimensions

- Predicted value plots require we set values for *all predictors*, even the ones that are not in the illustration.
- In the R framework, the `newdata` object in a `predict` statement must include values for all of the predictors in the model

# Example: Income and Sex in Chilean Survey

	M1	
	Estimate	(S.E.)
(Intercept)	0.024	(0.032)
income	1.069*	(0.497)
sexM	-0.142***	(0.039)
N	2591	
RMSE	0.998	
$R^2$	0.007	
adj $R^2$	0.006	

\* $p \leq 0.05$ \*\*  $p \leq 0.01$ \*\*\* $p \leq 0.001$

# The R commands for that were

```
library(car)
Chile$income <- Chile$income/1000000
chmod3 <- lm(statusquo ~ income + sex, data = Chile)
outreg(chmod3, tight = FALSE)
```

If you don't use  $\text{\LaTeX}$ , consider getting the newer testing version of rockchalk and running

```
outreg(chmod3, tight = FALSE, type = "html")
```

# Look what predictOMatic does

Adjust each predictor, one at a time, keeping the others fixed at exemplar values.

```
predictOMatic(chmod3, predVals = "margins")
```

```
$income
  income sex      fit
1 0.0025  F 0.02700871
2 0.0075  F 0.03235188
3 0.0150  F 0.04036663
4 0.0350  F 0.06173930
5 0.2000  F 0.23806385

$sex
  income sex      fit
1 0.03386144  F  0.0605226
2 0.03386144  M -0.0815923
```

Read ?predictOMatic. Many customizations are built in. Adjust the divider algorithm

```
predictOMatic(chmod3, predVals = "margins", divider = "std.dev")
```

# Look what predictOMatic does ...

```
$income
  income sex      fit
1  -0.05  F -0.02909456
2  -0.01  F  0.01365079
3   0.03  F  0.05639613
4   0.07  F  0.09914148
5   0.11  F  0.14188682

$sex
   income sex      fit
1 0.03386144  F  0.0605226
2 0.03386144  M -0.0815923
```

# Try R's Termplot for Inspecting Regressions

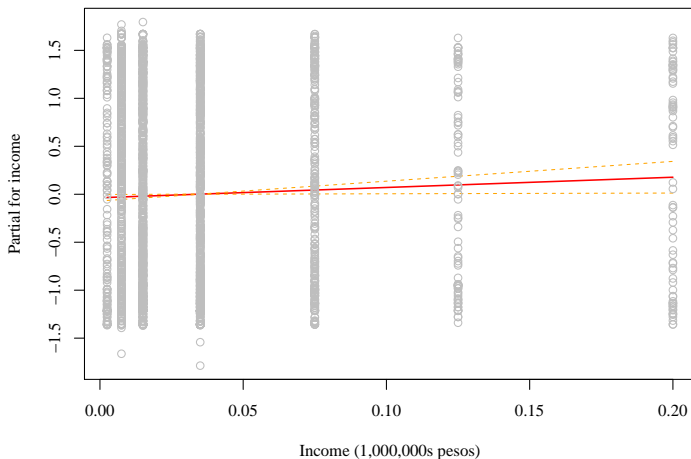
- `termplot()` will cycle through the variables.
- You run

```
termplot(mod1, se = TRUE, partial.resid = TRUE)
```

to see all of them.

- `termplot` will interactively show one predictor at a time

# Termplo: Depict the Income Effect



```
termplot(chmod3, se = TRUE, partial.resid = TRUE, terms = "income",  
         xlab = "Income (1,000,000s pesos)")
```

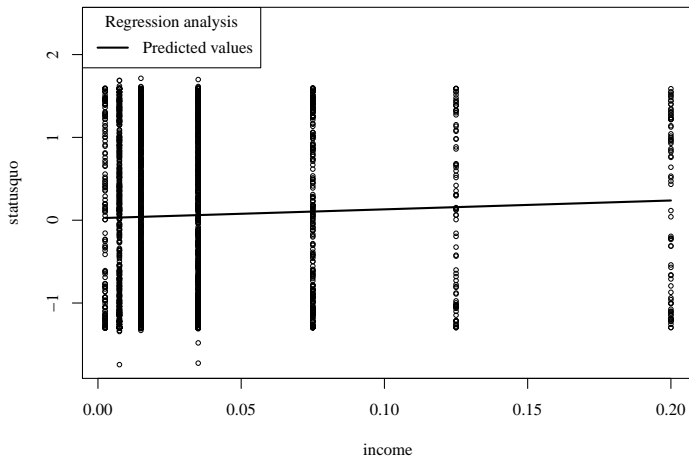
## Why is the Vertical Axis called "partial"?

- There is a Fully worked out derivation in Appendix C
- "Partial for income": What's left to predict" after removing the effect of other variables?
- Termpplot sets all variables at their numerical average.
- `partial.resid = TRUE` plots the "partial residuals" (removed part explained by other variables)
- `se = TRUE` prints the "pointwise standard errors" (uncertainty after using other variables)

# rockchalk::plotSlopes

- Choose one numeric predictor for the  $X$ -axis.
- The `modx` argument allows different lines for example values of another predictor (a “moderator”).
- See 3D in `plotPlane` and `addLines`.

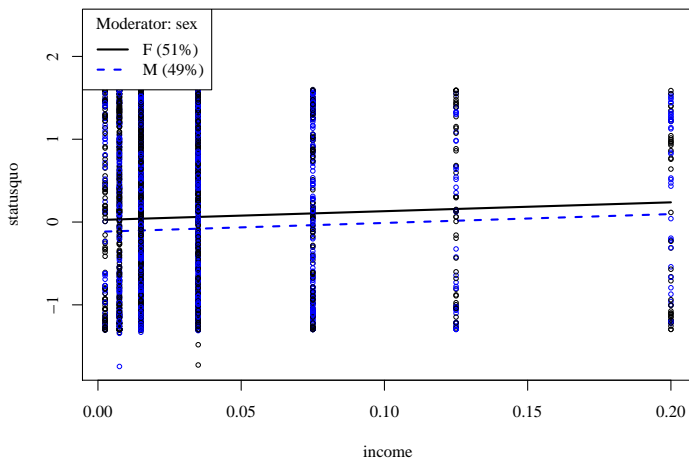
# plotSlopes for the Chilean regression



```
plotSlopes(chmod3, plotx = "income")
```

The default sets the sex variable at the mode, not the mean

# plotSlopes with separate lines for the Sexes



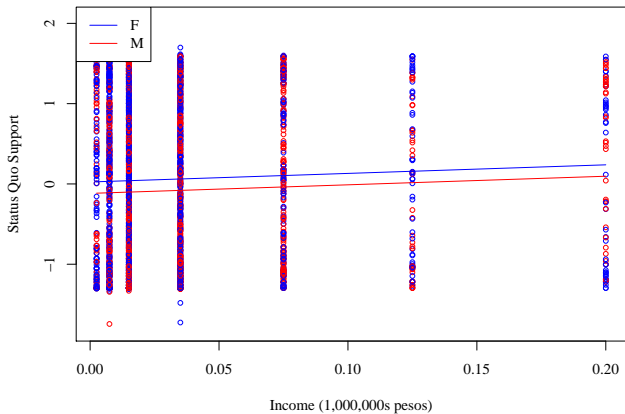
```
plotSlopes(chmod3, plotx = "income", modx = "sex")
```

## If you want to draw those line-by-line, some work is required

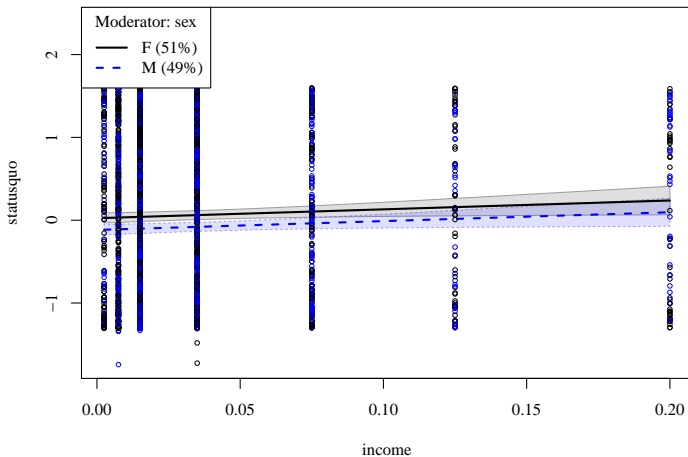
```
mycols <- c("blue", "red")
plot(statusquo ~ income, data = Chile, xlab = "Income (1,000,000s
      pesos)", ylab = "Status Quo Support", col = mycols[Chile$sex],
      cex = 0.7 )
incomeSeq <- seq(min(Chile$income, na.rm = TRUE),
                 max(Chile$income, na.rm = TRUE), length.out = 5)
newdat <- expand.grid(income = incomeSeq, sex = levels(Chile$sex))
newdat$pred <- predict(chmod3, newdata = newdat)
by(newdat, newdat$sex, function(datsub) lines(pred ~ income, dat =
      datsub, col = mycols[datsub$sex]))
legend("topleft", levels(newdat$sex), col = mycols, lty = 1, bg = "
      white")
```

Most R users will have to work through this process at one time or another.

# Difficulty of teaching that sequence → rockchalk package



# Confidence intervals for predicted values

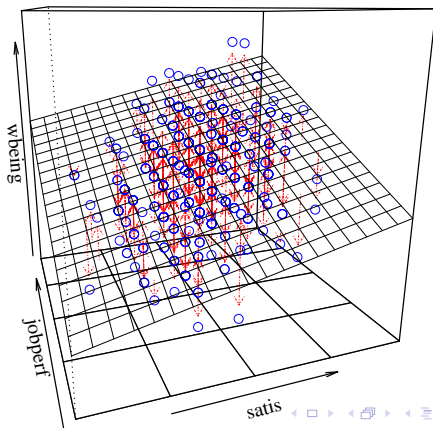


```
plotSlopes(chmod3, plotx = "income", modx = "sex", interval = "confidence")
```

## plotPlane with mod2 (well being)

Recall: `mod2 <- lm(wbeing ~ satis + jobperf + tenure + age, data=dat)`

```
plotPlane(mod2, plotx1 = "satis", plotx2 = "jobperf", drawArrows =  
  TRUE, alwd = 0.45, alength = 0.08)
```



# Multicollinearity: Its a Problem

- Regression works great, as long as the predictors are truly “separate” from (“orthogonal to”) each other.
- If predictors co-vary, difficult to separate their impacts.
- $Var(\hat{\beta}_j)$  is “inflated”
- The Usual problem: people want to put in too many similar variables.
  - do you hate eating salad because it has “lettuce”, or “spinach”, or “leafy greens”, or “vegetables”, or “no meat”, or...
- Several lectures on this coming up

# Outline

- 1 Basic assumptions
- 2 Multiple Regression
- 3 New Issues in Multiple Regression
  - Need Matrix Algebra
  - Visualization
  - Multicollinearity
- 4 Hypothesis Testing**
  - The Fancy t-test
  - F Test
- 5 Regression and Categorical Predictors
- 6 Appendix: Derivation without Matrices
- 7 Appendix: Matrices
- 8 Appendix: Digression on “Partial Residual” and “Partial Predicted Value”
- 9 Practice Problems

# Overview

- T test still works one-variable-at-a-time.
- New tests
  - The Wald test: compare 2 coefficients (fancy t-test)
  - The F test: compare nested models

# Wald (*Fancy t*) test

- t test of the hypothesis that 2 coefficients are statistically significantly different (Nicknamed fancy t-test).
- In  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}$ , are the effects of  $x_2$  and  $x_3$  are exactly the same?
  - Ex: does a dollar spent in community college equal a dollar spent at Harvard University
- The null hypothesis:  $H_0 : \beta_2 = \beta_3$ , same as  $H_0 : \beta_2 - \beta_3 = 0$ .
- Numerator in  $\hat{t}$  obvious, denominator will require some effort

$$\hat{t} = \frac{\hat{\beta}_2 - \hat{\beta}_3}{\text{std.err.}(\hat{\beta}_2 - \hat{\beta}_3)}$$

# Denominator

- *std.err.* = square root of estimated variance

$$\text{std.err.}(\hat{\beta}_2 - \hat{\beta}_3) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_2 - \hat{\beta}_3)} \quad (7)$$

- Fill in the blanks

$$\text{Var}(\hat{\beta}_2 - \hat{\beta}_3) = \text{Var}(\hat{\beta}_2) + \text{Var}(\hat{\beta}_3) - 2\text{Cov}(\hat{\beta}_2, \hat{\beta}_3) \quad (8)$$

- Recall the rule

$$\text{Var}(k_1 W + k_2 Z) = k_1^2 \text{Var}(W) + k_2^2 \text{Var}(Z) + 2k_1 k_2 \cdot \text{Cov}(W, Z) \quad (9)$$

# Great Paper: Will make you feel smarter

*Gelman, Andrew and Hal Stern. November 1, 2006. The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. The American Statistician. v, 60(4): 328-331. doi:10.1198/000313006X152649.*

- Compare 2 coefficients, say
  - $\hat{\beta}_2 = 1.1$  with  $std.err.(\hat{\beta}_2) = 0.6$ .
  - $\hat{\beta}_5 = 1.2$  with  $std.err(\hat{\beta}_5) = 0.5$

By the usual reasoning,  $\hat{\beta}_2$  is not statistically significant, but  $\hat{\beta}_5$  is. Gelman & Stern point out this fallacy:

*Therefore, the effect of variable  $x_5$ ; is larger than the effect of  $x_2$ .*

- It is a fallacy because we we did not test that hypothesis! (Implicitly assumed  $\beta_2$  different from  $\beta_5$ )
- Can conduct "fancy" t test to see!

# Test $H_0: \beta_j = \beta_k$

- Idea: do two variables have “the same effect” on the output?
- Null hypothesis

$$H_0 : \beta_j = \beta_k$$

same as

$$H_0 : \beta_j - \beta_k = 0$$

$$\hat{t} = \frac{\widehat{\beta_j - \beta_k}}{\text{std.err.}(\widehat{\beta_j - \beta_k})}$$

- It is easy to see  $\widehat{\beta_j - \beta_k}$ , is the actually the difference of the two estimates,  $\widehat{\beta_j} - \widehat{\beta_k}$ .

And the answer will be:

- Use this test statistic

$$\hat{t} = \frac{\hat{\beta}_j - \hat{\beta}_k}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_j) + \widehat{\text{Var}}(\hat{\beta}_k) - 2\widehat{\text{Cov}}(\hat{\beta}_j, \hat{\beta}_k)}}$$

- Estimates for denominator obtained from the var/covar matrix.

# Gelman/Stern Effect In our Example

- Easy to “eyeball” estimates and think “this effect is bigger than that”
- However, difference may not be “statistically significantly greater than 0”.
- Example: Is “age” effect bigger than “tenure” effect?

	Satisfaction	
	Estimate	(S.E.)
(Intercept)	0.772	(0.448)
satis	0.264***	(0.041)
jobperf	0.402***	(0.038)
tenure	0.021	(0.018)
age	0.034***	(0.010)
N	480	
RMSE	1.032	
$R^2$	0.306	
adj $R^2$	0.301	

\* $p \leq 0.05$ \*\*  $p \leq 0.01$ \*\*\* $p \leq 0.001$

# The Satisfaction Regression

## ■ Variance-Covariance matrix

	(Intercept)	satis	jobperf	tenure	age
(Intercept)	0.20061	-0.00693	-0.00804	0.00085	-0.00308
satis	-0.00693	0.00171	-0.00028	-0.00008	-0.00002
jobperf	-0.00804	-0.00028	0.00146	-0.00002	0.00003
tenure	0.00085	-0.00008	-0.00002	0.00031	-0.00009
age	-0.00308	-0.00002	0.00003	-0.00009	0.00010

$$\begin{aligned}
 t &= \frac{0.02088 - 0.03391}{\sqrt{(3.099521 \times 10^{-04} + 1.035459 \times 10^{-04} - 2(-8.893854 \times 10^{-05}))}} \\
 &= \frac{-0.01303}{\sqrt{0.0005913751}} = -0.5358126 \quad (10)
 \end{aligned}$$

- tenure effect may be statistically significantly greater than 0
- but its effect is not statistically significantly greater than that of age.

# F test

## ■ Nested Models:

- A “larger model” entirely contains a “smaller model”.
- Smaller model results by setting some coefficients to 0

## ■ Same: You want to “drop a bunch of variables”: Is the smaller model “as good”

## ■ Why is this better than conducting several t tests?

- Possible to find examples where we can't reject  $H_0 : \beta_j = 0$  for several variables, but  $F$  test may reject idea that they all are equal to 0.
- Recall problem of “weak power” in a t test.

# F test for groups of parameters

- Test a block of coefficients: are all equal to 0?
- Suppose 3 “efficacy” predictors ( $x$ 's), 3 “narcissism” predictors ( $z$ 's) in the Full Model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 z_{1i} + \beta_5 z_{2i} + \beta_6 z_{3i} + e_i$$

- Wonder, is it possible that all 3 narcissism predictors have 0 effect?
- Would a reduced model be as good?

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$$

- Principle of “parsimony” prefers models with fewer unknowns

# F Test for Subsets of Regressors

- 1 “Full” or “unrestricted” model: all coefficients included.
  - 1 Error sum of squares  $ESS_U$
- 2 “Restricted” model, some coefficients excluded. Meaning we assumed some coefficients equal 0.
  - 1 Error sum of squares for the restricted model  $ESS_R$
- 3 Compare the two models to find out if the restriction makes a difference.

**Caution:** 1) The models must be estimated on the SAME data and 2) they must be nested (one results from deleting terms from other)

# F test

- The F statistic is

$$F(q, N - k - 1) = \frac{(ESS_R - ESS_{UR})/q}{ESS_{UR}/N - k - 1}$$

- $N$  be the sample size.
- $k$  be the number of independent variables used in the full model.
- $q$  be the number of variables that are excluded from the restricted model.

Use F table,  $q$  degrees of freedom for the numerator and  $N-k-1$  for the denominator. If the number you get is bigger than that, reject the null.

## F Test in Regression Output is Different

- In default output, the F reported tests the Null hypothesis: *all slope coefficients are 0*.
- With  $k$  predictors, null hypothesis

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

- Some interpret this as a significance test on  $R^2$ . Is estimated  $R^2$  significantly greater than 0?

# Review

- Full, Unrestricted Model (all Variables Included):

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_{k-q} X_{(k-q)i} + \beta_{k-q+1} X_{(k-q+1)i} + \dots + \beta_k X_{ki} + e_i$$

- Restricted Model removes  $q$  coefficients (sets them at 0).

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_{k-q} X_{(k-q)i} + e_i$$

- If the full model “fits better”, it will have smaller Error Sum of Squares.
- F examines Null that  $\beta_{k-q+1} = \beta_{k-q+2} \dots = \beta_k = 0$ .

# Can Get Same from $R^2$

- Using R-squares

$$F(q, N - k - 1) = \frac{(R_{UR}^2 - R_R^2)/q}{(1 - R_{UR}^2)/(N - k - 1)} = \frac{(ESS_R - ESS_{UR})/q}{ESS_{UR}/N - k - 1} \quad (11)$$

- Sometimes thought of as a way of answering: Is the improvement from  $R_R^2$  to  $R_{UR}^2$  great enough to justify the number of coefficients being estimated?

# I need to work out more examples here

- I wrote more examples in lecture notes from the R course.  
http:  
[//pj.freefaculty.org/guides/Rcourse/regression-glm-2](http://pj.freefaculty.org/guides/Rcourse/regression-glm-2)
- Example with coefficients from a categorical predictor below

# Be aware of the danger of fluctuating N

- “Nested models” assumption requires same data used in 2 regressions.
- “Listwise Deletion” may cause N to accidentally differ, R will throw an error.
- I proceed as follows
  - 1 Fit the “big” model

```
m1 <- lm(y ~ x1 + x2 + x3 + z1 + z2, data = dat)
```

- 2 Fit the small model on the data subset from the big one

```
m2 <- lm(y ~ x1 + x2 + x3, data = model.frame(m1))
```

- 3 Then use the anova function to conduct the test

```
anova(m2, m1, test = F)
```

# Outline

- 1 Basic assumptions
- 2 Multiple Regression
- 3 New Issues in Multiple Regression
  - Need Matrix Algebra
  - Visualization
  - Multicollinearity
- 4 Hypothesis Testing
  - The Fancy t-test
  - F Test
- 5 Regression and Categorical Predictors**
- 6 Appendix: Derivation without Matrices
- 7 Appendix: Matrices
- 8 Appendix: Digression on “Partial Residual” and “Partial Predicted Value”
- 9 Practice Problems

# R: Model Matrix vs Design Matrix

- The “data frame” has numeric and factor (categorical) variables.
- Software converts factor variables to numeric columns for use in regression.

factor	numeric contrast
sex	sexMale
“Male”	1
“Female”	0
“Female”	0
“Male”	1

Call “sexMale”

- a “dummy variable”, or
- an “indicator variable”, or
- a “contrast variable.”

# Multi-valued Categorical Predictor

- 4 category religion variable

relig	religChristian	religJewish	religMuslim	Intercept
"Christian"	1	0	0	1
"Jewish"	0	1	0	1
"Muslim"	0	0	1	1
"Buddhist"	0	0	0	1

3 "dummy variables" for a 4 category variable

Along with the intercept

Coefficient estimates for religChristian, religJewish, religMuslim indicate difference of each religion from Buddhist.

- Various ways exist to re-do the columns of 0's and 1's, the default in R is "Treatment Contrasts"

## Ex: Model Frame and Design Matrix

- Lets create an example with these variables

`xnum1` a numeric predictor

`xcat2` a factor variable {"Female", "Male"}

`z` a factor variable {"A", "B", "C", "D"}

	min	med	max	mean	sd	skewness
<code>xnum1</code>	-3.128437	0.08685732	2.168475	0.009257184	0.9585762	-0.4158011
	0.6408297	100	0			
<code>y</code>	-2.470202	0.06019190	1.951928	0.011305065	0.9418138	-0.1899865
	-0.3395058	100	0			

# Ex: Model Frame and Design Matrix ...

```
$xcat2
$xcat2$table
Female   Male
      53    47

$xcat2$stats
      nobs      nmiss      entropy normedEntropy
100.000000  0.000000  0.9974016  0.9974016

$z
$z$table
  A  B  C  D
27 20 18 35

$z$stats
      nobs      nmiss      entropy normedEntropy
100.000000  0.000000  1.9498154  0.9749077

attr(,"class")
[1] "summarizedFactors"
attr(,"maxLevels")
[1] 5
```

# Ex: Model Frame and Design Matrix ...

```
attr(,"stats")
[1] "entropy"      "normedEntropy" "nobs"          "nmiss"
attr(,"digits")
[1] 2
```

## ■ The top of the data frame

```
head(dat, 15)
```

	xnum1	xcat2	z	y
1	0.221608622	Female	B	0.26596283
2	-0.231276910	Female	D	0.21272973
3	-0.281808460	Female	C	-0.23402274
4	0.012702720	Male	D	0.08226325
5	0.640549486	Male	D	0.32527083
6	0.000329053	Male	D	0.05663577
7	1.876086067	Female	B	1.08698168
8	1.846589937	Male	D	-0.61146485
9	-0.717189068	Female	C	-0.47003519
10	-0.112679417	Female	A	0.24509022
11	-0.200304178	Female	A	-0.31550858
12	1.354768108	Female	A	-0.29179589
13	0.894995240	Female	D	0.82910002

## Ex: Model Frame and Design Matrix ...

```
14  0.145769927  Male A  0.64123188
15 -0.216373991  Male B  0.69446722
```

- Behind the scenes, software creates the numeric “design matrix”.

```
mod1mm <- model.matrix(mod1)
head(mod1mm, 15)
```

```
(Intercept)      xnum1  xcat2Male  zB  zC  zD
1             1  0.221608622          0  1  0  0
2             1 -0.231276910          0  0  0  1
3             1 -0.281808460          0  0  1  0
4             1  0.012702720          1  0  0  1
5             1  0.640549486          1  0  0  1
6             1  0.000329053          1  0  0  1
7             1  1.876086067          0  1  0  0
8             1  1.846589937          1  0  0  1
9             1 -0.717189068          0  0  1  0
10            1 -0.112679417          0  0  0  0
11            1 -0.200304178          0  0  0  0
12            1  1.354768108          0  0  0  0
13            1  0.894995240          0  0  0  1
14            1  0.145769927          1  0  0  0
15            1 -0.216373991          1  1  0  0
```

## Ex: Model Frame and Design Matrix ...

- The default method creates “treatment contrasts” (dummy “indicator” variables)

```
contrasts(dat$z)
```

	B	C	D
A	0	0	0
B	1	0	0
C	0	1	0
D	0	0	1

- And the estimated coefficients are

## Ex: Model Frame and Design Matrix ...

```
Call:
lm(formula = y ~ xnum1 + xcat2 + z, data = dat)

Residuals:
    Min       1Q   Median       3Q      Max
-1.77519 -0.54207  0.04199  0.53220  1.83903

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.052043   0.174434  -0.298   0.766
xnum1        0.499466   0.087440   5.712 1.3e-07 ***
xcat2Male    0.006291   0.166862   0.038   0.970
zB           0.289755   0.240265   1.206   0.231
zC           0.432044   0.249772   1.730   0.087 .
zD          -0.228432   0.207446  -1.101   0.274

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8086 on 94 degrees of freedom
Multiple R2: 0.3002, Adjusted R2: 0.263
F-statistic: 8.065 on 5 and 94 DF, p-value: 2.306e-06
```

## Ex: Model Frame and Design Matrix ...

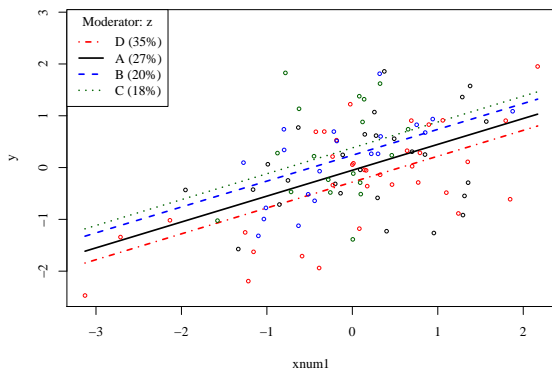
Understand those as estimates of this model:

$$y_i = \beta_0 + \beta_1 x_{num1_i} + \beta_2 x_{cat2Male_i} + \beta_3 zB_i + \beta_4 zC_i + \beta_5 zD_i + e_i \quad (12)$$

- A is the “baseline” group, it is “in” the intercept.
- The coefficient  $\hat{\beta}_1$  is the slope as  $\hat{y}_i$  depends on the numeric predictor
- The coefficient  $\hat{\beta}_2$  is the difference predicted for Males (compared against females)
- The coefficients  $\hat{\beta}_3, \hat{\beta}_4$ , and  $\hat{\beta}_5$  are “intercept shifters”, how B, C, and D are different from A.

# Ex: Model Frame and Design Matrix ...

The plotSlopes function in rockchalk can make a drawing.



Parallel Lines. Slopes for A, B, C, and D are same.

## Now Come back to the F test

- The religion factor leads us to estimate 3 coefficients.
- We can formally reject the null hypothesis that all 3 religion dummy variables have no effect
- $H_o : \beta_3 = \beta_4 = \beta_5 = 0$
- You can retrieve the error sum of squares manually from the regression, i.e.,  $\text{residuals(mod1)}^2$ , but you don't need to
- R offers 2 more convenient ways.
  - 1 fit the R and UR models and then use the `anova()` function to compare the two
  - 2 use the `drop1()` function, which automatically conducts the F test for each variable "group"

# The `anova()` function

```
mod2 <- lm(y ~ xnum1 + xcat2, data = dat)
anova(mod2, mod1, test = "F")
```

## Analysis of Variance Table

Model 1:  $y \sim xnum1 + xcat2$

Model 2:  $y \sim xnum1 + xcat2 + z$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	97	67.841				
2	94	61.453	3	6.3877	3.2569	0.02505 *

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# The drop1 function

```
drop1(mod1, test = "F")
```

```
Single term deletions
```

```
Model:
```

```
y ~ xnum1 + xcat2 + z
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)	
<none>			61.453	-36.690			
xnum1	1	21.3306	82.784	-8.894	32.6278	1.302e-07	***
xcat2	1	0.0009	61.454	-38.688	0.0014	0.97001	
z	3	6.3877	67.841	-32.801	3.2569	0.02505	*

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Puzzle: compare this to conclusion of a t-test

# Outline

- 1 Basic assumptions
- 2 Multiple Regression
- 3 New Issues in Multiple Regression
  - Need Matrix Algebra
  - Visualization
  - Multicollinearity
- 4 Hypothesis Testing
  - The Fancy t-test
  - F Test
- 5 Regression and Categorical Predictors
- 6 Appendix: Derivation without Matrices**
- 7 Appendix: Matrices
- 8 Appendix: Digression on “Partial Residual” and “Partial Predicted Value”
- 9 Practice Problems

## OLS with Just Three Predictors

- First Order Conditions: Calculate partial derivative for each parameter, set them equal to 0 (finding the “bottom of the bowl”).

$$\begin{aligned}\frac{\partial S}{\partial \hat{\beta}_0} &= -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) = 0 \\ \frac{\partial S}{\partial \hat{\beta}_1} &= -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) x_{1i} = 0 \\ \frac{\partial S}{\partial \hat{\beta}_2} &= -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) x_{2i} = 0\end{aligned}\quad (13)$$

These imply the three “normal equations” (one for each estimated parameter):

$$\begin{aligned}\sum y_i &= N \hat{\beta}_0 + (\sum x_{1i}) \hat{\beta}_1 + (\sum x_{2i}) \hat{\beta}_2 \\ \sum y_i x_{1i} &= (\sum x_{1i}) \hat{\beta}_0 + (\sum x_{1i}^2) \hat{\beta}_1 + (\sum x_{1i} x_{2i}) \hat{\beta}_2 \\ \sum y_i x_{2i} &= (\sum x_{2i}) \hat{\beta}_0 + (\sum x_{1i} x_{2i}) \hat{\beta}_1 + (\sum x_{2i}^2) \hat{\beta}_2\end{aligned}\quad (14)$$

# “Scalar Math” Gets Tedious

- def. Scalar : a “number”
- As long as you do ordinary ‘scalar’ math, you get one equation per parameter
- Summation signs flying around everywhere
- Estimate a regression with 10 variables, write out 10 equations? (ugh...)

# Outline

- 1 Basic assumptions
- 2 Multiple Regression
- 3 New Issues in Multiple Regression
  - Need Matrix Algebra
  - Visualization
  - Multicollinearity
- 4 Hypothesis Testing
  - The Fancy t-test
  - F Test
- 5 Regression and Categorical Predictors
- 6 Appendix: Derivation without Matrices
- 7 Appendix: Matrices**
- 8 Appendix: Digression on “Partial Residual” and “Partial Predicted Value”
- 9 Practice Problems

# Why Bother with Matrix Algebra?

- All treatments at “the next level of statistics” are presented in matrix algebra.
- You can't really read the literature unless you invest some effort to learn this
- Stats in math or economics would require matrices

# Terms

- a vector: always a column thing, referred to as  *$N$ .of rows  $\times$  1 column*
- a matrix: vectors glued together, side by side,  
 *$N$ .of rows  $\times$   $N$  of columns*
- transpose: turns a column into a row

# Matrix View

dep. var	indep var			slopes	predicted values
$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$	$X = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{1N} & x_{2N} \end{bmatrix}$	$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}$		$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix}$	

residuals		predicted values	
$\hat{e} = y - \hat{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix}$		$\hat{y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix} = X\hat{\beta}$	

# Matrix View of Multiple Regression

- Assume:

$$Y = X\beta + e$$

- With 2 predictors, that's short for:

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} \\ 1 & X_{12} & X_{22} \\ 1 & \dots & \dots \\ 1 & X_{1N} & X_{2N} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{bmatrix}$$

# $X\beta$ is Matrix Multiplication

- $X$  is rectangular,  $\beta$  is a column. We multiply to create  $X\beta$  like so:

$$\begin{bmatrix} 1 & 19 & 1 & 0.1 & 1 & 0 & 22 & 3 \\ 1 & 22 & 2 & 1.1 & 0 & 1 & 42 & 1 \\ 1 & 17 & & & & & & \\ \vdots & \vdots & & & & & & \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \end{bmatrix} \quad (15)$$

- Matrix by Vector Multiplication: multiply and add across each row separately.

$$\text{Row1} : 1 \cdot \beta_0 + \beta_1 \cdot 19 + \beta_2 \cdot 1 + \beta_3 \cdot 0.1 + \beta_4 \cdot 1 + \beta_5 \cdot 0 + \beta_6 \cdot 22 + \beta_7 \cdot 3 \quad (16)$$

## $X\beta$ is Matrix Multiplication ...

- Do that for each row, and so the result from

$$X\beta \tag{17}$$

will be a column of  $N$  numbers, one for each row.

$$X\beta = \begin{bmatrix} \textit{some number}_1 \\ \textit{some number}_2 \\ \textit{some number}_3 \\ \textit{some number}_4 \\ \dots \\ \textit{some number}_N \end{bmatrix}$$

## Brief Matrix Multiplication

- In my course web guide collection, there are longer introductions to matrix algebra
- A row vector is a transposed column
- Multiply vectors (a row vector times a column vector). Sometimes called an “inner product”.

$$\begin{bmatrix} a & b & c & d & e \end{bmatrix} \cdot \begin{bmatrix} f \\ g \\ h \\ i \\ j \end{bmatrix} = af + bg + ch + di + ej$$

- $[1 \times 5] \cdot [5 \times 1]$  yields a  $[1 \times 1]$  result, just a single number (a “scalar”)
- One of the most common uses will calculate “sum of squares”

$$\begin{bmatrix} a & b & c & d & e \end{bmatrix} \cdot \begin{bmatrix} a \\ b \\ c \\ d \\ e \end{bmatrix} = a^2 + b^2 + c^2 + d^2 + e^2$$

# Brief Matrix Multiplication

- Multiply a matrix times a vector

$$\begin{bmatrix} a & b & c & d & e \\ r & s & t & u & v \end{bmatrix} \cdot \begin{bmatrix} f \\ g \\ h \\ i \\ j \end{bmatrix} = \begin{bmatrix} af + bg + ch + di + ej \\ rf + sg + th + ui + vj \end{bmatrix}$$

- Treat matrix as two rows, then conduct multiplication separately for each one.
- $[2 \times 5] \cdot [5 \times 1]$  yields a  $[2 \times 1]$  result
- Example:  $\hat{y} = Xb$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \dots \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ 1 & \dots & \dots \\ 1 & x_{1N} & x_{2N} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 + \hat{\beta}_1 x_{11} + \hat{\beta}_2 x_{21} \\ \hat{\beta}_0 + \hat{\beta}_1 x_{12} + \hat{\beta}_2 x_{22} \\ \dots \\ \hat{\beta}_0 + \hat{\beta}_1 x_{1N} + \hat{\beta}_2 x_{2N} \end{bmatrix}$$

# Brief Matrix Multiplication

- Multiply a matrix times a matrix

$$\begin{bmatrix} a & b & c & d & e \\ r & s & t & u & v \end{bmatrix} \cdot \begin{bmatrix} f & k \\ g & l \\ h & m \\ i & n \\ j & o \end{bmatrix}$$
$$= \begin{bmatrix} af + bg + ch + di + ej & ak + bl + cm + dn + eo \\ rf + sg + th + ui + vj & rk + sl + tm + un + vo \end{bmatrix}$$

- Break into sequences of vector multiplications, row 1 · column 1, row2 · column 1, row 1 · · column 2, row 2 · column 2.
- $[2 \times 5] \cdot [5 \times 2]$  yields a  $[2 \times 2]$  result

# Transpose

- I've been avoiding the question of where row vectors come from.
- A row vector is a column vector “on its side. It has been “transposed”.

$$\beta^T = (\beta_0, \beta_1, \beta_2, \dots)$$

- A matrix transpose is “rotated” so the left column becomes the first row

$$X = \begin{bmatrix} 1 & 3 & 33 \\ 1 & 2 & 62 \\ 1 & 5 & 65 \\ 1 & 1 & 45 \\ 1 & 5 & 66 \end{bmatrix} \quad X \text{ is } 5 \times 3$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 3 & 2 & 5 & 1 & 5 \\ 33 & 62 & 65 & 45 & 66 \end{bmatrix} \quad X^T \text{ is } 3 \times 5$$

# The Sum of Squares is $(y - \hat{y})^T (y - \hat{y})$

- $\hat{y} = X\hat{\beta}$  is the predicted value column
- Estimation Procedure: Choose  $\hat{\beta}$  to minimize  $(y - \hat{y})^T (y - \hat{y})$

$$(y - \hat{y})^T (y - \hat{y}) = (y_1 - \hat{y}_1, y_2 - \hat{y}_2, y_3 - \hat{y}_3 \dots, y_N - \hat{y}_N) \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_N - \hat{y}_N \end{bmatrix} \quad (18)$$

- Which is just the matrix way of saying “sum of squared errors”

$$(y - \hat{y})^T (y - \hat{y}) = \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (19)$$

# Matrix solution

- Calculus is used to find “first order” conditions for the sum-of-squares minimizing estimates of  $\beta$ .
- The theoretical solution is often written down like this

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (20)$$

Poetry for your T-Shirts: “ $\beta$  hat equals X-transpose X inverse X-transpose y”.

- $(X^T X)$  is the covariance matrix of the input “design matrix”. It is  $p \times p$ .

# Matrix Inverse

- $(X^T X)^{-1}$  is the “inverse” of  $(X^T X)$ , meaning.
- Recall ordinary math:  $x^{-1} \cdot x = 1$ . The inverse is  $\frac{1}{x}$ , the “reciprocal”
- To translate that to matrices, we need to define the idea of 1 in a matrix.
- The Identity matrix  $I$  in matrix algebra is like 1 in scalar math

$$(X^T X)^{-1}(X^T X) = I, \text{ where } I = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (21)$$

# True and Estimated Variance of $\hat{\beta}$

- The “true variance/covariance matrix” of the estimator  $\hat{\beta}$  is given by

$$\text{Var}(\hat{\beta}) = \sigma_e^2 (X^T X)^{-1} \quad (22)$$

- Recall, that's the theoretical quantity. If your error term did have variance  $\sigma_e^2$  then the estimates of  $\hat{\beta}$  would fluctuate
- But we don't know  $\sigma_e^2$ , we must estimate it by the Root Mean Squared Error from the regression.
- So the estimated variance of  $\hat{\beta}$  replaces the true variance of the error with the estimate.

$$\widehat{\text{Var}}(\hat{\beta}) = \widehat{\sigma}_e^2 (X^T X)^{-1} \quad (23)$$

## Sum of Squares on a Matrix Level

- $X^T$  means “ $X$  transpose”, or “ $X$  turned on its side”

$$X^T = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x1_1 & x1_2 & x1_3 & & x1_N \\ x2_1 & x2_2 & x2_3 & & x2_N \end{bmatrix} \quad (24)$$

- $X^T X$  is a sum of squares matrix. See?

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x1_1 & x1_2 & x1_3 & & x1_N \\ x2_1 & x2_2 & x2_3 & & x2_N \end{bmatrix} \begin{bmatrix} 1 & x1_1 & x2_1 \\ 1 & x1_2 & x2_2 \\ \vdots & \vdots & \vdots \\ 1 & x1_N & x2_N \end{bmatrix} \quad (25)$$

$$= \begin{bmatrix} N & \sum x1_i & \sum x2_i \\ \sum x1_i & \sum x1_i^2 & \sum x1_i x2_i \\ \sum x2_i & \sum x1_i x2_i & \sum x2_i^2 \end{bmatrix} \quad (26)$$

# First Order Conditions are called the Normal Equations

- The Normal Equations

$$\begin{aligned}\sum y_i &= N\hat{\beta}_0 + (\sum x_{1i})\hat{\beta}_1 + (\sum x_{2i})\hat{\beta}_2 \\ \sum y_i x_{1i} &= (\sum x_{1i})\hat{\beta}_0 + (\sum x_{1i}^2)\hat{\beta}_1 + (\sum x_{1i}x_{2i})\hat{\beta}_2 \\ \sum y_i x_{2i} &= (\sum x_{2i})\hat{\beta}_0 + (\sum x_{1i}x_{2i})\hat{\beta}_1 + (\sum x_{2i}^2)\hat{\beta}_2\end{aligned}$$

- can be written with matrices as:

$$\begin{bmatrix} \sum y_i \\ \sum y_i x_{1i} \\ \sum y_i x_{2i} \end{bmatrix} = \begin{bmatrix} N & \sum x_{1i} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \quad (27)$$

- Look what pops out

$$X^T y = (X^T X) \hat{\beta} \quad (28)$$

# Almost done

- Once you get to this point, the work is almost done.

$$X^T y = (X^T X) \hat{\beta} \quad (29)$$

- Need to get  $\hat{\beta}$  “all by itself”. So suppose you had a matrix  $(X^T X)^{-1}$  that could work just like “dividing” in ordinary math. Then

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (30)$$

- $(X^T X)^{-1}$  is called an “inverse” of  $X^T X$ , and if we multiply the inverse of  $X^T X$  times  $(X^T X)^{-1}$  then we get the matrix equivalent of 1, which is called  $I$  (for “identity matrix”).

$$I = \left( \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right) \quad (31)$$

Multiply “anything” times  $I$  and you get “anything” back, unchanged. Like multiplying by 1 in ordinary math.

# QR Decomposition

- Recently, I've learned that stats books teach us some unrealistic calculations.
- We talk about  $(X^T X)$ , but actually forming that, or inverting it, would cause a lot of “numerical instability” due to roundoff error.
- Software does not invert  $X^T X$ , rather it decomposes  $X$  and then uses clever algorithms (QR decomposition) to preserve numerical accuracy.
- Usually, they'll say re-write the Normal Equations as

$$X^T y - (X^T X) \hat{\beta} = 0 \quad (32)$$

- or this

$$X^T (y - X \hat{\beta}) = 0 \quad (33)$$

- See <http://pj.freefaculty.org/guides/stat/Math/Matrix-Decompositions>

## Var/Cov Matrix of $\hat{\beta}$

- The variance/covariance matrix of  $\hat{\beta}$  (if we knew the variance of the error) would be

$$\text{Var}(\hat{\beta}) = \sigma_e^2 (X^T X)^{-1} \quad (34)$$

- Estimate the variance of the error term as the mean square error (same as 1 input regression)

$$\widehat{\sigma_e^2} = \text{MSE} = \frac{1}{N - k} (y - \hat{y})^T (y - \hat{y}) \quad (35)$$

- So the estimated Variance of  $\hat{\beta}$  is

$$\widehat{\text{Var}}(\hat{\beta}) = \widehat{\sigma_e^2} (X^T X)^{-1} \quad (36)$$

- And the square root of the diagonal elements of  $\widehat{\text{Var}}(\hat{\beta})$  are the “standard errors” of the  $\hat{\beta}$ 's.

# Outline

- 1 Basic assumptions
- 2 Multiple Regression
- 3 New Issues in Multiple Regression
  - Need Matrix Algebra
  - Visualization
  - Multicollinearity
- 4 Hypothesis Testing
  - The Fancy t-test
  - F Test
- 5 Regression and Categorical Predictors
- 6 Appendix: Derivation without Matrices
- 7 Appendix: Matrices
- 8 Appendix: Digression on "Partial Residual" and "Partial Predicted Value"**
- 9 Practice Problems

# Digression: Calculate Term "Partial" Residual, "Partial" Predicted Value?

- Consider model

```
mod <- lm(y ~ X1 + X2 + X3)
```

- An ordinary "residual" is the "observed value" minus a "fitted value"

$$y_i - \hat{y}_i = y_i - \{\hat{\beta}_0 + \hat{\beta}_1 X1_i + \hat{\beta}_2 X2_i + \hat{\beta}_3 X3_i\} \quad (37)$$

- The predicted value for the mean case always equal to the mean of  $y_i$

$$\bar{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \bar{X1} + \hat{\beta}_2 \bar{X2} + \hat{\beta}_3 \bar{X3} \quad (38)$$

## Digression: "term predictions"

- How much of the prediction's quality is due to each variable?
- Begin with  $\hat{y}_i$ ,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} \quad (39)$$

- Then add zero to that, but do it in the sneaky math way: We both add AND subtract  $\hat{\beta}_0 + \hat{\beta}_1 \overline{X1} + \hat{\beta}_2 \overline{X2} + \hat{\beta}_3 \overline{X3}$ .
- Rearrange so that  $\hat{y}_i$  is seen as a sum of the "mean prediction" and "term" predictions

$$\begin{aligned} \hat{y}_i = & \quad \{ \hat{\beta}_0 + \hat{\beta}_1 \overline{X1} + \hat{\beta}_2 \overline{X2} + \hat{\beta}_3 \overline{X3} \} \\ & + \{ \hat{\beta}_1 (X_{1i} - \overline{X1}) + \hat{\beta}_2 (X_{2i} - \overline{X2}) + \hat{\beta}_3 (X_{3i} - \overline{X3}) \} \quad (40) \end{aligned}$$

- Which is:

## Digression: "term predictions" ...

$$\hat{y}_i = \bar{y} + \{\hat{\beta}_1(X1_i - \bar{X1}) + \hat{\beta}_2(X2_i - \bar{X2}) + \hat{\beta}_3(X3_i - \bar{X3})\} \quad (41)$$

- We've found a way to see each predicted value as a variation about the observed average of  $y$ .
- In R, `predict(mod, type="terms")` returns the last "things" as separate columns:

$$\overbrace{\hat{\beta}_1(X1_i - \bar{X1})}^{X1}, \quad \overbrace{\hat{\beta}_2(X2_i - \bar{X2})}^{X2}, \quad \overbrace{\hat{\beta}_3(X3_i - \bar{X3})}^{X3} \quad (42)$$

- These are the "partial predicted values" used by `termplot`

## Digression: And a partial residual is...

- The "full residual" is the difference between the observed value and each partial prediction:

$$residual_i = y_i - \hat{y}_i = y_i - \{\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}\}$$

- An R "partial residual" for  $X_{1j}$ : remove  $X_{1j}$ , replace it with  $\overline{X_1}$ .
- A partial residual:

$$\begin{aligned} \text{partial residual}(X_{1i}) &= y_i - \{\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}\} \\ &\quad + \{\hat{\beta}_1 (X_{1i} - \overline{X_1})\} \\ &= y_i - \{\hat{\beta}_0 + \hat{\beta}_1 \overline{X_1} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}\} \end{aligned}$$

- Like an ordinary residual, except the variable of interest is replaced by its mean in the calculation of the predicted value.
- Interpretation
  - We have removed the effect of variation in  $X_{1i}$ , but accounted for the other variables.

# Outline

- 1 Basic assumptions
- 2 Multiple Regression
- 3 New Issues in Multiple Regression
  - Need Matrix Algebra
  - Visualization
  - Multicollinearity
- 4 Hypothesis Testing
  - The Fancy t-test
  - F Test
- 5 Regression and Categorical Predictors
- 6 Appendix: Derivation without Matrices
- 7 Appendix: Matrices
- 8 Appendix: Digression on “Partial Residual” and “Partial Predicted Value”
- 9 Practice Problems

# Problem 1

Download the PublicSpending data from: <http://pj.freefaculty.org/guides/stat/DataSets/PublicSpending>. Fit a model that predicts EX as a linear function of all of the other variables in the data set except STATE. That's the "full model"

- A. *Make a professionally acceptable regression table to display the results.*
- B. *Write a paragraph that states the null hypothesis and conducts a significance test for any of the parameter estimates.*
- C. *When I run that, the results indicate that the slope coefficients for YOUNG and GROW are not statistically different from 0. But I can't say for sure if YOUNG's effect is different from GROW. Conduct a fancy t-test to find out, write up your answer.*

## Problem 1 (cont)

*D. Run a regression model that predicts EX as a function of YOUNG. Make a table, discuss the difference of the estimate of the slope estimate for YOUNG compared with the estimate from the full model.*

*E. Run R's termplot function to display the relationship for each independent variable separately. (If you don't use R, look for a replacement function in your software. If you can't find one, we have to settle for scatterplots of the individual variables against EX, I guess.)*

## Problem 2

Take the “full” model from the previous question. Conduct the F test to see if there is no significant effect among the variables GROW, OLD, and YOUNG. The F formula depends on the difference in the Error Sum of Squares in the full and reduced models.

R has an easy routine to do this test

```
mod1 <- lm( ~fill infull model ~)
mod2 <- lm(~omit the 3 variables ~)
anova(mod2, mod1, test="F")
```

## Problem 3

Get any dataset you like that has 1 more-or-less continuous DV and 3 or more IV. You could take one of our example datasets from <http://pj.freefaculty.org/guides/stat/DataSets> (except for PublicSpending). Suppose the variables are called  $y$ ,  $x_1$ ,  $x_2$ ,  $x_3$ ,  $x_4$  and so forth.

First, estimate a sea of regression models. Suppose you select any one variable, say  $x_3$ , and run the regression to estimate  $y_i = \beta_0 + \beta_3 x_{3i} + e_i$ . Then add the other variables one at a time, so estimate  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_3 x_{3i} + e_i$ ,  $y_i = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} + e_i$ ,  $y_i = \beta_0 + \beta_3 x_{3i} + \beta_4 x_{4i} + e_i$ . Then estimate one model with all of the variables included.

- 1 I wonder how much the slope estimates “bounce around” from one fit to the next. Concentrate on  $\hat{\beta}_3$  in the various models. I guess we better look at the standard error of  $\beta_3$  as well.
- 2 For the model with all of the variable, run R’s termplot function. Don’t forget the se and partial options, as demonstrated in my example above.