Inference

ME104: Linear Regression Analysis
Kenneth Benoit

August 15, 2012

Stata output resvisited

```
. reg votes1st spend_total incumb minister spendXinc

      Source |       SS       df       MS              Number of obs =     462
-------------+------------------------------           F(  4,   457) =  229.05
       Model |  2.9549e+09     4   738728297           Prob > F      =  0.0000
    Residual |  1.4739e+09   457  3225201.58           R-squared     =  0.6672
-------------+------------------------------           Adj R-squared =  0.6643
       Total |  4.4288e+09   461  9607007.17           Root MSE      =  1795.9

------------------------------------------------------------------------------
    votes1st |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 spend_total |   .2033637   .0114807    17.71   0.000     .1808021    .2259252
      incumb |   5150.758   536.3686     9.60   0.000     4096.704    6204.813
    minister |   1260.001   474.9661     2.65   0.008      326.613     2193.39
   spendXinc |  -.1490399   .0274584    -5.43   0.000    -.2030003   -.0950794
       _cons |   469.3744   161.5464     2.91   0.004     151.9086    786.8402
------------------------------------------------------------------------------
```

R^2

- R^2

$$R^2 = \frac{SSM}{TSS} \tag{1}$$
$$= \frac{\sum(\hat{y}_i - \bar{\hat{y}})^2}{\sum(y_i - \bar{y})^2} \tag{2}$$
$$= \frac{2.9549e+09}{4.4288e+09} \tag{3}$$
$$= 0.667201 \tag{4}$$

- Adjusted R^2

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-k-1} \tag{5}$$
$$= 1 - (1 - 0.6672)\frac{462-1}{462-4-1} \tag{6}$$
$$= 0.6642871 \tag{7}$$

"Root MSE" (and F)

- Root MSE = estimate of σ

$$\hat{\sigma} = \sqrt{SSE/df_{resid}} \tag{8}$$
$$= \sqrt{(1.4739e+09)/457} \tag{9}$$
$$= \sqrt{3225201.58} \tag{10}$$
$$= 1795.885 \tag{11}$$

- F-test
  This is the test of the null hypothesis that the joint effect of all independent variables is zero — more on this shortly

How to compute R^2 and $\hat{\sigma}$ from this output?

Valid cases:	4274	Dependent variable:	disprls
Missing cases:	0	Deletion method:	None
Total SS:	2094312.971	Degrees of freedom:	4251
R-squared:	0.887	Rbar-squared:	0.886
Residual SS:	237366.238	Std error of est:	7.472
F(22,4251):	1511.642	Probability of F:	0.000

Variable	Estimate	Standard Error	t-value	Prob >|t|	Standardized Estimate	Cor with Dep Var
CONSTANT	50.437041	0.164518	306.573980	0.000	---	---
HSL*m	-8.501692	2.525842	-3.365885	0.001	-0.082936	-0.215230
HSL	-34.443131	2.579062	-13.354906	0.000	-0.329397	-0.216392
SL*m	-6.526475	2.525842	-2.583881	0.010	-0.063667	-0.223110
SL	-37.302552	2.579062	-14.463611	0.000	-0.356743	-0.225317
MSL*m	-7.828347	2.525842	-3.099302	0.002	-0.076367	-0.217458
MSL	-35.371193	2.579062	-13.714750	0.000	-0.338273	-0.218966
dH*m	-8.292628	2.525842	-3.283115	0.001	-0.080896	-0.207012
dH	-33.823319	2.579062	-13.114581	0.000	-0.323470	-0.208080
LRH*m	-6.953863	2.525842	-2.753087	0.006	-0.067836	-0.224528
LRH	-37.002049	2.579062	-14.347095	0.000	-0.353869	-0.226579
LRDr*m	-7.023068	2.525842	-2.780486	0.005	-0.068511	-0.222815
LRDr	-36.755473	2.579062	-14.251488	0.000	-0.351511	-0.224798
LRI*m	-7.679571	2.525842	-3.040401	0.002	-0.074916	-0.217981
LRI	-35.579349	2.579062	-13.795460	0.000	-0.340263	-0.219566
ImpHA*m	-10.721835	2.525842	-4.244856	0.000	-0.104594	-0.157791
ImpHA	-26.278325	2.579062	-10.189101	0.000	-0.251313	-0.156677
EqP*m	-21.029154	2.525842	-8.325603	0.000	-0.205144	-0.147518
EqP	-14.500895	2.579062	-5.622546	0.000	-0.138679	-0.141654
Dan*m	-7.355209	2.525842	-2.911983	0.004	-0.071752	-0.220301
Dan	-36.153527	2.579062	-14.018091	0.000	-0.345755	-0.222081
Ad*m	-20.961184	2.525842	-8.298693	0.000	-0.204481	-0.145850
Ad	-14.401702	2.579062	-5.584085	0.000	-0.137731	-0.139978

OLS computation of "best fitting line"

- This is incredibly powerful:

$$\begin{aligned} Y &= X\beta + \epsilon \\ X'Y &= X'X\beta + X'\epsilon \\ X'Y &= X'X\beta + 0 \\ (X'X)^{-1}X'Y &= \beta + 0 \\ \beta &= (X'X)^{-1}X'Y \end{aligned}$$

- But it does not tell us how uncertain is our estimate of β in any probabalistic, comparative sense

Distributional assumptions

- Distributions are used to assess uncertainty
- Many standard parametric tests are associated with interpreting uncertainty of regression results, including those from the CLRM/OLS
  - t distributions
  - z distributions
  - F distributions
  - χ^2 distributions
- In small samples, the applicability of these distributions depends on the errors being distributed normally
- In larger samples, the asymptotic properties of these distributions means the results hold even when errors are not distributed normally

Distribution of $\hat{\beta}$

- $\hat{\beta} \sim N(\beta, (X'X)^{-1}\sigma^2)$ in repeated samples
- The variances of $\hat{\beta}$ will be the diagonal elements from the variance-covariance matrix of $\hat{\beta}$
- Problem: variance covariance matrix is not usually known, because σ^2 is not usually known (and $\text{Var}(\beta) = \sigma^2 / \sum(x_i - \bar{x})^2$)
- But by using s^2 as an estimate of σ^2, we can use the square root of the kth diagonal element of the variance-covariance matrix to estimate the standard error of $\hat{\beta}_k$, which will be t-distributed

t-tests for individual $\hat{\beta}$

- As just stated, the sampling distribution of $\hat{\beta}$ will be t-distributed, with $n - k - 1$ degrees of freedom
- The empirical t-value will be the coefficient estimate divided by its standard error
- This yields a t-value that is compared with the critical value for t with $n - k - 1$ degrees of freedom
- Example in Stata:

OLS Example in Stata

```
. reg votes1st spend_total incumb minister spendXinc

      Source |       SS       df       MS              Number of obs =     462
-------------+------------------------------           F(  4,   457) =  229.05
       Model |  2.9549e+09     4   738728297           Prob > F      =  0.0000
    Residual |  1.4739e+09   457  3225201.58           R-squared     =  0.6672
-------------+------------------------------           Adj R-squared =  0.6643
       Total |  4.4288e+09   461  9607007.17           Root MSE      =  1795.9

------------------------------------------------------------------------------
    votes1st |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 spend_total |   .2033637   .0114807    17.71   0.000     .1808021    .2259252
      incumb |   5150.758   536.3686     9.60   0.000     4096.704    6204.813
    minister |   1260.001   474.9661     2.65   0.008      326.613     2193.39
   spendXinc |  -.1490399   .0274584    -5.43   0.000    -.2030003   -.0950794
       _cons |   469.3744   161.5464     2.91   0.004     151.9086    786.8402
------------------------------------------------------------------------------
```

Interpretation of regression coefficients

- Each coefficient in a multiple regression model describes the association between an explanatory variable and the response, controlling for the other explanatory variables
  - These are known as partial associations
- For example, consider the coefficient β_k of X_k:

$$\mu = (\alpha + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1}) + \beta_k X_k = (\text{Others}) + \beta_k X_k$$

- Here the "(Others)" bit depends on the other explanatory variables $X_1, \ldots, X_{k-1}$ but not on X_k
- If now X_k increases by 1 unit and the other explanatory variables remain unchanged, μ changes by β_k units

Interpretation of regression coefficients

- In general,
  - The coefficient β_j of an explanatory variable X_j shows the change in the expected value of Y when X_j increases by 1 unit while all the other explanatory variables are held constant
  - ...or, in other words, the expected change in Y when X_j increases by 1 unit, "controlling for" the other explanatory variables
- For example, in the model shown below, the (estimated) coefficient of education is $\hat{\beta}_{\text{education}} = 0.990$
- Thus every one-year increase in education completed increases the expected General Health Index by 0.99 points, controlling for age and family income

A fitted model for GHI

Response variable: General Health Index					
Explanatory variable	$\hat{\beta}$	s.e.	t	P-value	95% CI
Constant	59.42				
Age	-0.128	0.032	-4.029	< 0.001	(-0.190; -0.066)
Education	0.990	0.143	6.906	< 0.001	(0.709; 1.272)
Fam. Income	0.275	0.063	4.345	< 0.001	(0.151; 0.398)
$\hat{\sigma} = 14.6$; $R^2 = 0.061$; $n = 1699$; $df = 1695$					

The uninteresting parameters

- The remaining two parameters of the model are necessary but uninteresting for substantive interpretation:
- The constant term α is the expected value of Y when all explanatory variables are 0
- The residual standard deviation σ is the standard deviation of Y given (any) single set of values for $X_1, \ldots, X_k$
  - i.e. the standard deviation "around the fitted regression surface" at any given point of it

Estimation of the parameters

- The fitted values of Y are

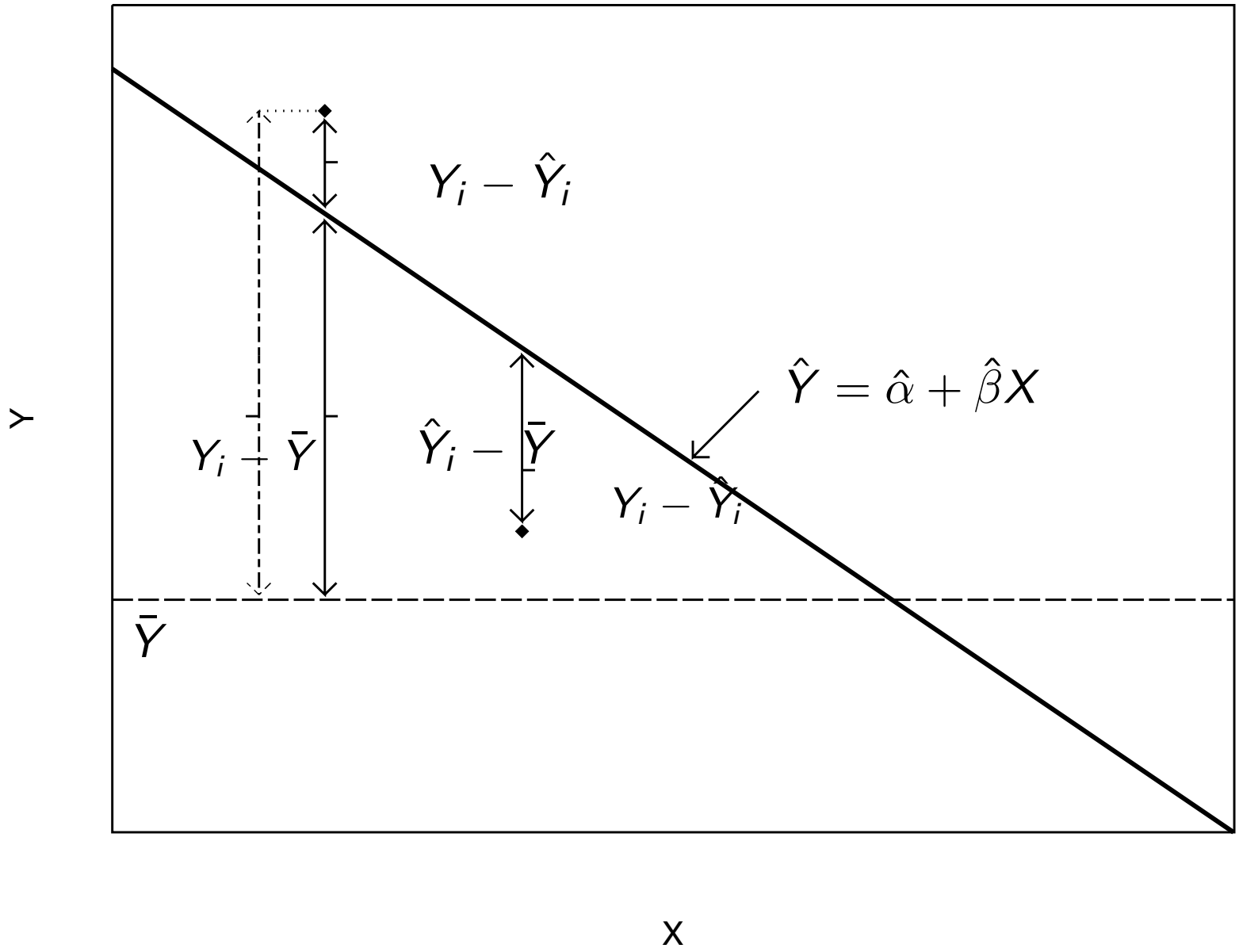
$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki}$$

  for all observations $i = 1, \ldots, n$ in the sample
- We would like to select the values of the estimated coefficients $\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_k$ so that the fitted values are a good match to observed values $Y_1, \ldots, Y_n$
- This is done by finding estimates which minimize the *error sum of squares*

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

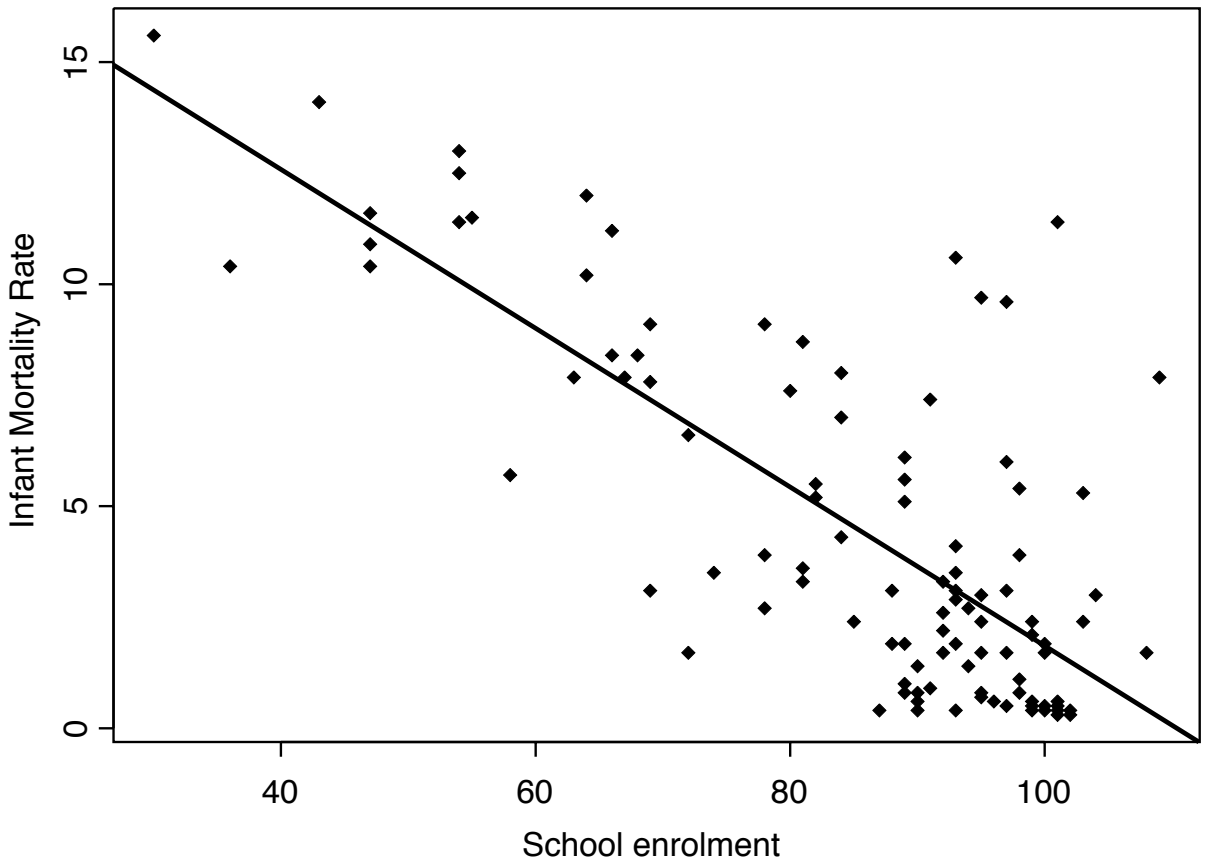
- These are the (ordinary) least squares (OLS) estimates of the coefficients

Estimation of the parameters



Estimation of the parameters

Reminder: For the simple (one-X) linear model, the fitted values define a straight line:



Estimation of the parameters

- Least squares estimates $\hat{\beta}_j$ of the coefficients are easily calculated by a computer
- Also produced are estimated standard errors $\hat{\text{se}}(\hat{\beta}_j)$ of the estimated coefficients
- Also produced is an estimate $\hat{\sigma}$ of the residual standard deviation

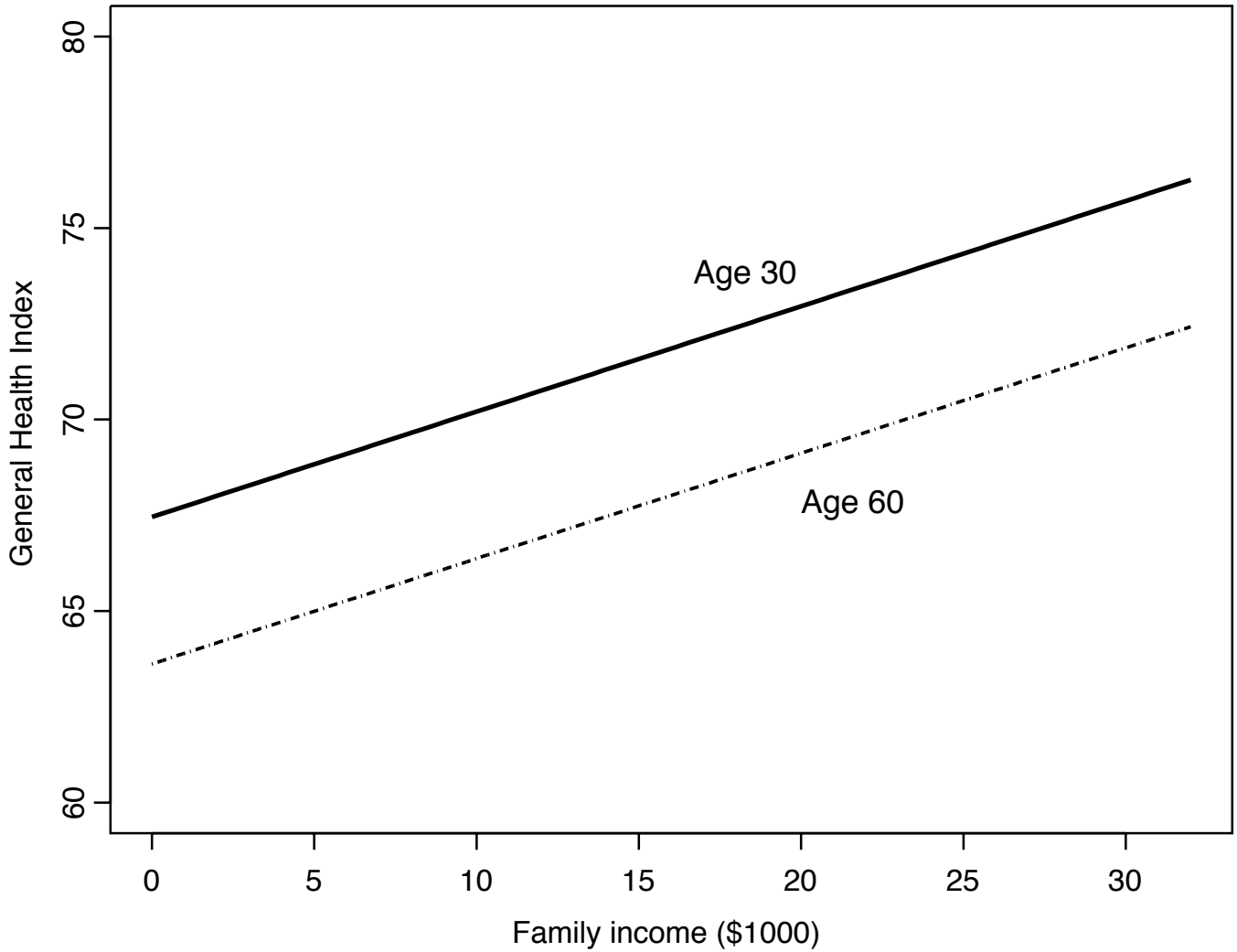
Fitted values for interpretation

- A fitted model can be interpreted using the regression coefficients $\hat{\beta}_j$ as well as fitted (predicted) values

$$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k$$

  for Y, calculated at some representative values of the explanatory variables
- Methods of presentation:
  - Plots of fitted values given a continuous explanatory variable, fixing others at some values
  - Tables of fitted values, given an array of values for the explanatory variables
- See examples for GHI below, given age, education and income

Fitted values of GHI given income



(Education fixed at 12 years)

Fitted values of GHI

	Education								
	0			12			18		
Income	0	10	32	0	10	32	0	10	32
Age 16	57.4	60.1	66.2	69.2	72.0	78.0	—	—	—
35	54.9	57.7	63.7	66.8	69.6	75.6	72.8	75.5	81.6
62	51.5	54.2	60.3	63.4	66.1	72.2	69.3	72.1	78.1

Inference for a single regression coefficient

- Consider the following null hypothesis for the coefficient of an explanatory variable X_j:

$$H_0 : \beta_j = 0$$

against the alternative hypothesis

$$H_a : \beta_j \neq 0,$$

both with no claims about the coefficients of the other explanatory variables

- In other words,

  H_0 : There is no partial association between X_j and Y, controlling for the other explanatory variables

Inference for a single regression coefficient

- This is tested using the t-test statistic

$$t = \frac{\hat{\beta}_j}{\hat{\text{se}}(\hat{\beta}_j)}$$

- When H_0 is true, the sampling distribution of t is a t distribution with $n - (k + 1)$ degrees of freedom (k is the number of explanatory variables)
- The P-value is calculated just as before
- If the null hypothesis is not rejected, the implication is that X_j may be dropped from the model (while keeping the other explanatory variables in the model)

Inference for a single regression coefficient

- For example, in the model for GHI given age, education and income, the coefficient of education is $\hat{\beta}_{\text{education}} = 0.990$ and $\hat{\text{se}}(\hat{\beta}_{\text{education}}) = 0.143$, so

$$t = \frac{0.990}{0.143} = 6.91$$

  for which $P < 0.001$
- Thus there is strong evidence of a partial association between education and GHI, even controlling for age and income

Inference for a single regression coefficient

- Similarly, $P < 0.001$ for tests of the effects of both age and income, so both of these have a partial effect as well
- If, however, we add work experience to the model, the test of its coefficient has $P = 0.563$
  - Length of work experience has no partial effect on GHI, once we control for age, education and income, so it does not need to be included in the model

Inference for a single regression coefficient

Variable	Model (1)	(2)	(3)	(4)	(5)
Age	−0.138	−0.089	−0.184	−0.128	−0.142
	(< 0.001)	(0.004)	(< 0.001)	(< 0.001)	(< 0.001)
Education	—	1.157	—	0.990	0.981
		(< 0.001)		(< 0.001)	(< 0.001)
Income	—	—	0.391	0.275	0.277
			(< 0.001)	(< 0.001)	(< 0.001)
Experience	—	—	—	—	0.002
					(0.563)
(Constant)	74.777	58.801	72.383	59.417	59.723
R^2	0.012	0.051	0.035	0.061	0.061

(P-values in parentheses)

Confidence intervals

- A confidence interval for a single regression coefficient β_j is

$$\hat{\beta}_j \pm t_{\alpha/2}^{(n-(k+1))} \widehat{\text{se}}(\hat{\beta}_j)$$

  where $t_{\alpha/2}^{(n-(k+1))}$ is the multiplier from the $t_{n-(k+1)}$ distribution for the required confidence level
  - or approximately from the standard normal distribution, e.g. 1.96 for 95% intervals
- For example, 95% confidence interval for the coefficient of education in the model discussed above is

$$0.990 \pm 1.96 \times 0.143 = (0.709; 1.272)$$

An example

- Data from the Rand Health Insurance Experiment (HIE): see S. 4.1 of the coursepack
- $n = 1699$ respondents to a survey at the start of the study
- **Response variable** Y: Respondent's diastolic blood pressure at the end of the study
- Various explanatory variables considered today for illustration

Dummy variables

- Categorical explanatory variables are included in regression models as dummy variables (indicator variables)
  - Variables with only two values, 0 and 1
  - 1 if a subject's value of a categorical variable is in a particular category, 0 if not
- For example, a person's sex may be entered as the dummy variable for men:

$$X = \begin{cases} 1 & \text{if the person is male} \\ 0 & \text{otherwise} \end{cases}$$

or as the dummy for women (but not both)

Coefficients of dummy variables

- Consider a model with only dummy for men as explanatory variable:

$$\begin{aligned} \text{For men:} \quad \mathrm{E}(Y) &= \alpha + \beta X = \alpha + \beta \times 1 = \alpha + \beta \\ \text{For women:} \quad \mathrm{E}(Y) &= \alpha + \beta X = \alpha + \beta \times 0 = \alpha \\ \text{Difference:} & \qquad\qquad\qquad\qquad\qquad\qquad \beta \end{aligned}$$

- In short, the coefficient of the dummy variable for men is the expected difference in Y between men and women

Coefficients of dummy variables

- This is the two-group model considered in lecture 2, and β is the group (sex) difference in expected Y
- Least squares estimates are here

$$\begin{aligned} \hat{\alpha} &= \bar{Y}_{\text{women}} \\ \hat{\beta} &= \bar{Y}_{\text{men}} - \bar{Y}_{\text{women}} \end{aligned}$$

- The t-test for the null hypothesis that $\beta = 0$ and confidence interval for β are the same as the inference for group difference of means in lecture 2
- This is the simplest example of an Analysis of Variance (ANOVA) model: linear regression models with only categorical explanatory variables

Coefficients of dummy variables

- More generally, dummy variables may be included in multiple linear models together with other (continuous or dummy) explanatory variables
- In general, the coefficients of dummy variables are interpreted as expected differences in Y between units at different levels of categorical variables, controlling for other variables in the model
- A t-test for the hypothesis that such a coefficient is 0 is a test of no such difference

Example from HIE data

- Models for diastolic blood pressure at exit (Y), given
  - Control variables: initial blood pressure, age and sex (as dummy for men)
  - Dummy variable for free health care (0 for all other insurance plans)
- The coefficient of the free-care dummy is -1.544, with $P = 0.013$
  - Thus the expected blood pressure at exit is 1.544 points lower for participants on free care than for those on some other plan, controlling for initial blood pressure, age and sex
  - This difference is statistically significant, at the 5% level
  - The 95% confidence interval for this difference is (-2.76; -0.33)

Example from HIE data

Response variable: Diastolic blood pressure at exit				
Explanatory variable	$\hat{\beta}$	s.e.	t	P-value
Constant	31.98			
Initial BB	0.249	0.027	9.10	< 0.001
Age	-0.186	0.026	-7.25	< 0.001
Sex: male	3.938	0.977	4.03	< 0.001
Free health care	-1.544	0.621	-2.49	0.013

Variables with more categories

- Dummy variables are also used for explanatory variables with more than two categories
  - e.g. smoking status: never smoked/ex-smoker/current smoker
- Dummy variables for all but one of the categories are included in the model
- The category without a dummy is the reference (baseline) category
- The coefficient of the dummy of a category is the expected difference in Y between that category and the baseline
  - Differences between non-baseline categories are given by differences of their coefficients
- The choice of the baseline is arbitrary: the model is the same, whatever the choice

Variables with more categories

Response variable: blood pressure at exit			
	Model		
Variable	(1)	(2)	(3)
Past blood pressure	0.573	0.573	0.573
Sex			
Female	0	-2.033	0
Male	2.033	0	2.033
Smoking status			
Never smoked	0	1.239	1.382
Ex-smoker	-1.239	0	0.143
Current smoker	-1.382	-0.143	0
(Constant)	35.383	36.177	34.001

Example from HIE data

- Again, models for diastolic blood pressure at exit (Y), with initial blood pressure, age and sex as control variables
- Insurance plan now entered as a five-category variable, with 95% coinsurance plan as the reference level
- Each t-test of the coefficient of the dummy for a particular insurance plan tests the hypothesis that there is no difference in expected blood pressure between that plan and the reference plan, controlling for the other variables
  - The only significant difference is for the free care plan, with coefficient -2.009: This is the difference in expected blood pressure between free care and 95% plans, controlling for initial BB, age and sex

Example from HIE data

Response variable: Diastolic blood pressure at exit				
Explanatory variable	$\hat{\beta}$	s.e.	t	P-value
(Other coefficients not shown)				
Insurance plan:				
95% coinsurance	0	—	—	—
50% coinsurance	0.091	1.382	0.07	0.947
25% coinsurance	-0.772	0.979	-0.79	0.430
Individual deductible	-0.727	0.947	-0.77	0.442
Free care	-2.009	0.866	-2.32	0.020

The General F-test

- This is used to test multiple-coefficient hypotheses of the form

$$H_0 : \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0,$$

against the alternative

$$H_a : \text{ at least one of } \beta_{g+1}, \beta_{g+2}, \ldots, \beta_k \text{ is not } 0$$

- The most common application of this is testing the coefficients of dummy variables for different categories of a categorical explanatory variable simultaneously
  - e.g. in the example above, the coefficients of the dummies for four insurance plans
  - If this is not rejected, none of the plans differ from the reference plan (and thus they also do not differ from each other), i.e. insurance plan has no effect on blood pressure, given the control variables

The General F-test

- To carry out the F-test, first fit two models:
  - The **restricted model** (M_0), where the variables of the null hypothesis are omitted
  - The **full model** (M_a), where the variables of the null hypothesis are included
- In this example,
  - M_0 includes initial BB, age and sex
  - M_a includes initial BB, age and sex, **and** the four insurance plan dummies
- Then compare the R^2 values (or error sums of squares SSE) between the two models

The General F-test

- The F-test statistic is

$$\begin{aligned} F &= \frac{(SSE_0 - SSE_a)/(k_a - k_0)}{SSE_a/[n-(k_a+1)]} \\ &= \frac{(R_a^2 - R_0^2)/(k_a - k_0)}{(1-R_a^2)/[n-(k_a+1)]} \end{aligned}$$

- **Large** values of this are evidence **against** the null hypothesis that the restricted model is correct
  - In that case $R_a^2 - R_0^2$ is large, i.e. the full model has a "much" higher R^2
- The sampling distribution of F is an F distribution with $k_a - k_0$ and $n-(k_a+1)$ degrees of freedom
  - In practice, P-values obtained with a computer

The General F-test

- In the example, $n = 1045$ and
  - $R_a^2 = 0.3536$ and $k_a = 7$ for the full model
  - $R_0^2 = 0.3491$ and $k_a = 3$ for the restricted model, so

$$F = \frac{(0.3536 - 0.3491)/4}{(1 - 0.3536)/1037} = 1.80$$

  for which $P = 0.127$
- Thus the null hypothesis is not rejected: no evidence of differences between insurance plans in their effect on blood pressure, controlling for the other three variables

Interactions

- There is an interaction between two explanatory variables, if the effect of (either) one of them on the response variable depends on *at which value* the other one is controlled
- Included in the model by using products of the two explanatory variables as additional explanatory variables in the model
- Example: data for the 50 United States, average SAT score of students (Y) given school expenditure per student (X) and % of students taking the SAT in three groups (low, middle and high)
  - The %-variable included as two dummy variables, say D_M for middle and D_L for low

Interactions

- A model without interactions:

$$\mathrm{E}(Y) = \alpha + \beta_1 D_L + \beta_2 D_M + \beta_3 X$$

- Here the partial effect of expenditure is β_3, same for all values of the %-variable
- Add now the products $(D_L X)$ and $(D_M X)$, to get the model

$$\mathrm{E}(Y) = \alpha + \beta_1 D_L + \beta_2 D_M + \beta_3 X + \beta_4 (D_L X) + \beta_5 (D_M X)$$

- This model states that there is an interaction between school expenditure and the %-variable
  - Why?

Interactions

- Consider the effect of X at different values of the dummy variables:

$$
\begin{aligned}
&\mathrm{E}(Y) \\
&= \alpha + \beta_1 D_L + \beta_2 D_M + \beta_3 X + \beta_4 (D_L X) + \beta_5 (D_M X) \\
&= \alpha + \beta_3 X && \text{For high-\% states} \\
&= (\alpha + \beta_2) + (\beta_3 + \beta_5) X && \text{For mid-\% states} \\
&= (\alpha + \beta_1) + (\beta_3 + \beta_4) X && \text{For low-\% states}
\end{aligned}
$$

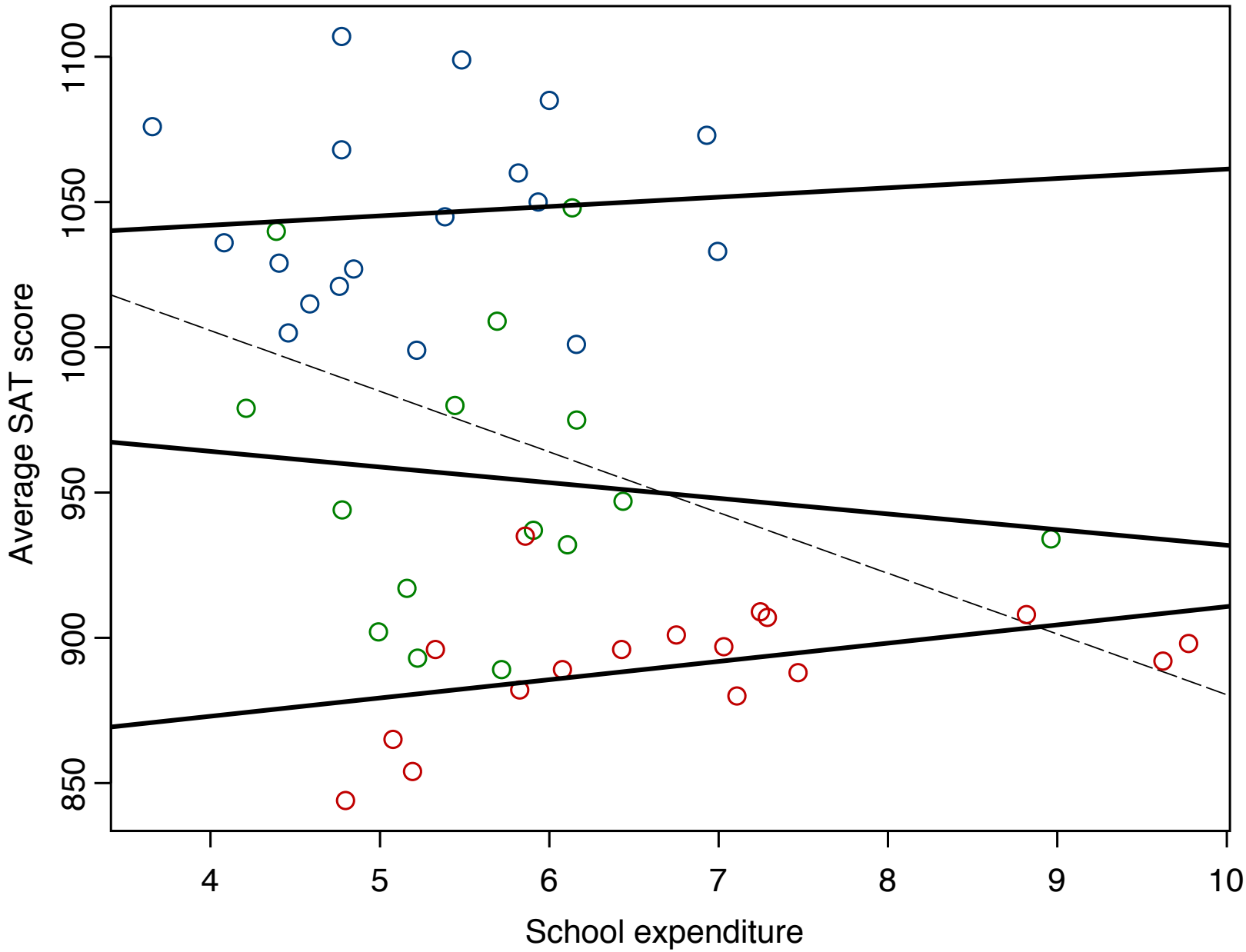
- In other words, the coefficient of X depends on the value at which D_L and D_M are fixed

Interactions

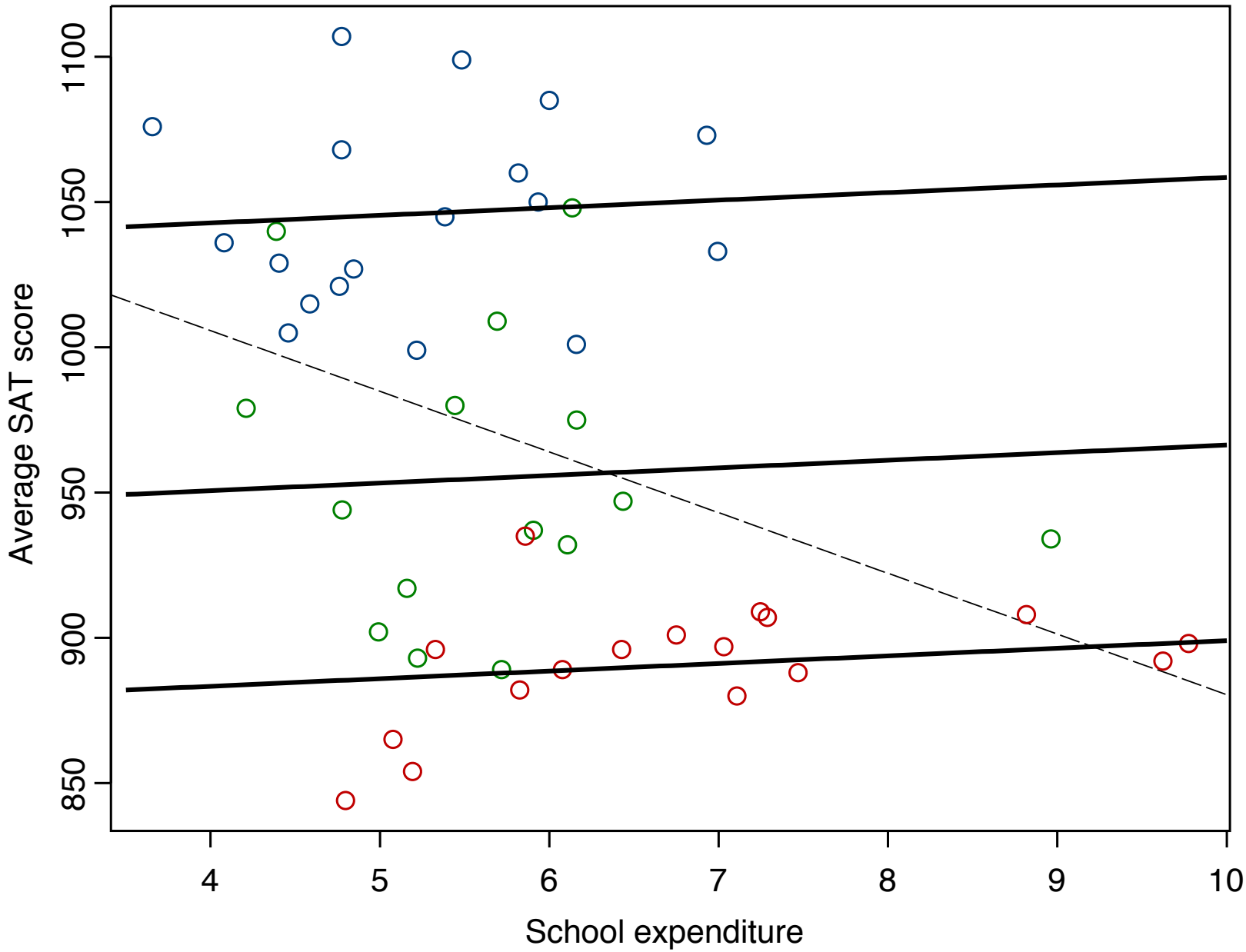
- The estimated coefficients in this example are

$$\begin{aligned}
\mathrm{E}(Y) &= 847.9 + 181.3D_L + 137.8D_M + 6.3X \\
&\qquad -3.2(D_L X) - 11.7(D_M X) \\
&= 847.9 + 6.3X && \text{for high-\% states} \\
&= 1029.2 + 3.1X && \text{for low-\% states} \\
&= 985.7 - 5.4X && \text{for mid-\% states}
\end{aligned}$$

Model with interaction



...and without


Average SAT score
1100
1050
1000
950
900
850
4
5
6
7
8
9
10
School expenditure

Testing for interactions

- A standard test of whether the coefficient of the product variable (or variables) is zero is a test of whether the interaction is needed in the model
  - t-test or (if more than one product variable) F-test
- In the example, we use an F-test, comparing

$$\begin{aligned} \text{Full model} \quad & \mathrm{E}(Y) = \alpha + \beta_1 D_L + \beta_2 D_M + \beta_3 X \\ & \qquad + \beta_4 (D_L X) + \beta_5 (D_M X) \\ \text{vs. Restricted m.} \quad & \mathrm{E}(Y) = \alpha + \beta_1 D_L + \beta_2 D_M + \beta_3 X \end{aligned}$$

  i.e. a test of $H_0 : \ \beta_4 = \beta_5 = 0$
- Here $F = 0.61$ and $P = 0.55$, so the interaction is not in fact significant

Interactions between categorical variables

- In the previous example, the interaction was between a continuous variable and a categorical variable
- In other cases too, interactions are included as products of variables
- An example of interaction between two categorical (here binary) explanatory variables, from HIE data:
  - Response variable: blood pressure at exit
  - Two binary explanatory variables:
    - Being on free health care vs. some other plan
    - Income in the lowest 20% in the data vs. not
  - Other control variables: initial blood pressure, age and sex

Interactions between categorical variables

Variable	Coefficient
Initial blood pressure	0.483
Age	0.260
Sex: Male	3.981
Low income (lowest 20%)	2.662
Free health care	-1.299
Income×Insurance plan	-1.262
(Constant)	31.83

Interactions between categorical variables

- Which coefficients involving income and insurance plan apply to different combinations of these variables:

Free care	Low income	
	No	Yes
No	0	2.662
Yes	-1.299	0.101

(not showing the other coefficients)

where 0.101=2.662-1.299-1.262

- In other words,
  - effect of low income on blood pressure is smaller for respondents on free care than on other plans
  - effect of free care on blood pressure is bigger for low-income respondents than for high-income ones
- (Again, the interaction is not actually significant ($P = 0.42$) here, so this just illustrates the general idea)

F-tests for all predictors

- The "default" test with most regressions is the test that all $\beta = 0$ — in other words, nothing going on here
- Null hypothesis: $H_0 : \beta_0 = \ldots = \beta_k = 0$
- This is equivalent to testing a model with a set of linear constraints where all β are set to zero
- Formula:

$$F = \frac{(SST - SSE)/k}{SSE/(n - k - 1)}$$

  where:
  - SST and SSE are the total and error sums of squares
  - n is the number of observations
  - k is the number of *variables* (excluding constant!)
  - $(n - k - 1)$ is also the "residual degrees of freedom"

Example of F-test for all predictors

```
> m1 <- lm(votes1st ~ spend_total*incumb, data=dail)
> summary(m1)

Call:
lm(formula = votes1st ~ spend_total * incumb, data = dail)

Residuals:
    Min      1Q  Median      3Q     Max
-5574.9  -947.5  -214.0   893.8  7154.6

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         464.59553  162.59753   2.857  0.00447 **
spend_total           0.20414    0.01155  17.671  < 2e-16 ***
incumb             4493.32513  478.80828   9.384  < 2e-16 ***
spend_total:incumb   -0.10689    0.02254  -4.742 2.83e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1808 on 458 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.6621, Adjusted R-squared: 0.6599
F-statistic: 299.1 on 3 and 458 DF,  p-value: < 2.2e-16

> SST <- sum((m1$model[,1] - mean(m1$model[,1]))^2)
> SSE <- sum(m1$residuals^2)
> k <- 3
> (df <- m1$df.residual)
[1] 458
> (F <- ((SST-SSE)/k) / (SSE/df))
[1] 299.1095
> 1 - pf(F, k, df)
[1] 0
```

Testing just one predictor

- Null hypothesis: $H_0 : \beta_j = 0$
- Error statistic will be

$$t_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

  where t_j is t-distributed with $n - k - 1$ degrees of freedom (same as `df.residual`)

- The F statistic will be t_j^2

Example of testing just one predictor

```
> m1c <- lm(votes1st ~ spend_total + incumb, data=dail)
> SSEc <- deviance(m1c)  # the SSE
> (F2 <- (SSEc - SSE) / (SSE / df))
[1] 22.48497
> 1 - pf(F2, 1, df)
[1] 2.832798e-06
> sqrt(F2)      # will be the same as the t-test for this coefficient
[1] 4.741832
> summary(m1)$coeff
                       Estimate   Std. Error   t value     Pr(>|t|)
(Intercept)         464.5955332 162.59752848  2.857335 4.466694e-03
spend_total           0.2041449   0.01155236 17.671273 1.154515e-53
incumb             4493.3251289 478.80828470  9.384393 2.962201e-19
spend_total:incumb   -0.1068943   0.02254283 -4.741832 2.832798e-06
> anova(m1c,m1) # a much easier way to compare 2 models
Analysis of Variance Table

Model 1: votes1st ~ spend_total + incumb
Model 2: votes1st ~ spend_total * incumb
  Res.Df        RSS  Df  Sum of Sq      F    Pr(>F)
1    459 1570088906
2    458 1496614393   1   73474513 22.485 2.833e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing multiple predictors and generalized linear constraints

- Just because two variables are individually not significant, does not mean that *jointly* the variables are not significant
- The F-test can be generalized to any set of J linear constraints, as follows:

$$F = \frac{[SSE_{constrained} - SSE_{unconstrained}]/(df_{const} - df_{unconst})}{SSE_{unconstrained}/df_{unconstr}}$$

- Steps:
  1. Run the unconstrained regression, save SSE
  2. Run the constrained regression, save SSE
  3. Compute F and reject if $F > F_{df_{const}, df_{unconst}}$
- For a single β_k, this is equivalent to the t-test

Parametric confidence intervals for β

- CIs or *confidence intervals* provide an alternative way to express uncertainty for our estimates
- For a $100(1-\alpha)\%$ confidence region, any point that lies within the region represents a null hypothesis that would not be rejected at the $100\alpha\%$ level, while every point outside it represents a null hypothesis that would have been rejected
- More valuable than simple hypothesis tests because it tells us about a parameter's plausible values
- Formula: Estimate $\pm$ Critical Value $\times$ S.E.
- For β specifically:

$$\hat{\beta}_i \pm t_{n-k-1}^{\alpha/2} \hat{\sigma} \sqrt{(X'X)_{ii}^{-1}}$$

- In practice we should consider joint confidence regions, especially when $\hat{\beta}$ are (highly) correlated