Introduction to Statistical Methods
SOC-GA 2332

Lecture 5: Regression II

Siwei Cheng

NEW YORK UNIVERSITY

The Linear Regression Model: Hypothesis Testing

- Now, let's move from estimation to hypothesis testing.
- Our research questions are often about whether there is an association between the variation in Y and the variation in X.
- In the language of the linear regression model, we test whether the coefficient on X, β_1, is significantly different from zero.
- Formally, our null and alternative hypotheses are:
  H_0: $\beta_1 = 0$ (Slope equals zero).
  H_1: $\beta_1 \neq 0$ (Slope differs from zero).
  Or: $\beta_1 > 0$ / $\beta_1 < 0$ (Slope is greater than or less than zero).

The Linear Regression Model: Hypothesis Testing

- Let's go back to $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Remember that when calculating the two OLS estimates, our calculation depends on the **sample**.
- We introduced randomness when we draw the sample from the population, because if our samples included different elements, our calculated $\hat{\beta}_0$ and $\hat{\beta}_1$ are likely to be different. Hence, we would like to know the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$.
- More generally, the population does not have to exist. Randomness is always introduced as long as there is a data generating process through which our observed sample is generated.
- We know from an earlier point about the property of OLS estimator that the means of the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ are their corresponding true parameters.

The Linear Regression Model: Hypothesis Testing

- What is the standard deviation of the sampling distribution? That is, what is the **standard error** of each parameter estimates?
- To calculate the standard error, we start with the **mean square error (MSE)**:

$$MSE = \frac{SSE}{n-2} = \frac{\Sigma(y-\hat{y})^2}{n-2} \tag{1}$$

- And we take the square root:

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}} \tag{2}$$

The Linear Regression Model: Hypothesis Testing

- The standard error of β_1 is a function of MSE. Specifically, it is:

$$se_{\beta_1} = \frac{s}{\sqrt{\Sigma(x - \bar{x})^2}} \tag{3}$$

- Note that $\Sigma(x - \bar{x})^2 = (n - 1)s_x^2$.
- So the denominator is larger when the **sample variance** of x (s_x^2) is larger, and the denominator is larger when the sample size n is larger.
- This implies that the standard error is smaller (and thus β_1 tends to be more precise) when the variance of X in our sample is larger, and when we have a larger sample.

The Linear Regression Model: Hypothesis Testing

- Now let's go back to hypothesis testing. The test statistic is:

$$t = \frac{\hat{\beta}_1 - 0}{se_{\beta_1}} \quad (4)$$

- When H_0 is true, t follows t-distribution with degree of freedom n-2 (we are using n observations to estimate 2 parameters). We compare the calculated t (based on our sample and the above formulas) with the t-distribution to obtain p-value.
- If p-value is smaller than our pre-determined α, we reject the null hypothesis, or we say that the estimated slope is significantly different from zero.

The Linear Regression Model: Hypothesis Testing

- We can also construct the confidence interval of β_1 (just as we did for the CI of mean) as: $\hat{\beta}_1 \pm t \cdot (se_{\beta_1})$.
- If this confidence interval does not contain zero, then we reject the null hypothesis. Otherwise we fail to reject the null hypothesis.

Example

- Let's look at the previous example of the relationship between occupational education level and occupational income.
- We have our estimate for β_1: $\hat{\beta}_1 = 898.862$. So how to construct a 95% confidence interval of β_1?
- First, from Equation (1) we know that we need SSE, which is $\Sigma(y - \hat{y})^2$. We let STATA calculate MSE by the following codes:

```
. gen income_predicted=-2854.114+898.862*educat

. gen residual=income-income_predicted

. gen residual_sq=residual^2

. egen SSE=total(residual_sq)

. gen MSE=SSE/(102-2)

. di MSE
12133920
```

Example

- So we have se_{β_1} as follows:

$$se_{\beta_1} = \frac{s}{\sqrt{\Sigma(x-\bar{x})^2}} = \frac{\sqrt{MSE}}{\sqrt{n-1}\cdot s_x} \tag{5}$$

$$= \frac{\sqrt{12133920}}{\sqrt{102-2}\cdot 2.728444} = 127.66902 \tag{6}$$

- The 95% confidence interval for β_1 would be (approximating t distribution by normal distribution):

$$898.862 \pm 1.96 \cdot 127.66902 = [648.6, 1149.1]$$

- The 95% confidence interval does not contain 0 (The P-value is smaller than 0.001 (significant on the 0.01 level)), so we reject the null hypothesis that the slope is zero.

Coefficient of Determination: R-squared

Coefficient of Determination: R-squared

- We know that regression describes a relationship between X and Y. We are trying to explain variations in Y by variations in X. So how much of the variations in Y can be explained by X?
- If we do not have information on X, then the total amount of variation in Y in the sample is called Total Sum of Squares (TSS):

$$TSS = \Sigma(y_i - \bar{y})^2$$

- That is, the grand mean is our best guess for everyone!

Coefficient of Determination: R-squared

- We know that regression describes a relationship between X and Y. We are trying to explain variations in Y by variations in X. So how much of the variations in Y can be explained by X?
- If we do not have information on X, then the total amount of variation in Y in the sample is called Total Sum of Squares (TSS):

$$TSS = \Sigma(y_i - \bar{y})^2$$

- That is, the grand mean is our best guess for everyone!
- We learned earlier that the sum of squared residuals from a regression model is called the Sum of Squared Errors (SSE):

$$SSE = \Sigma(y_i - \hat{y}_i)^2$$

Coefficient of Determination: R-squared

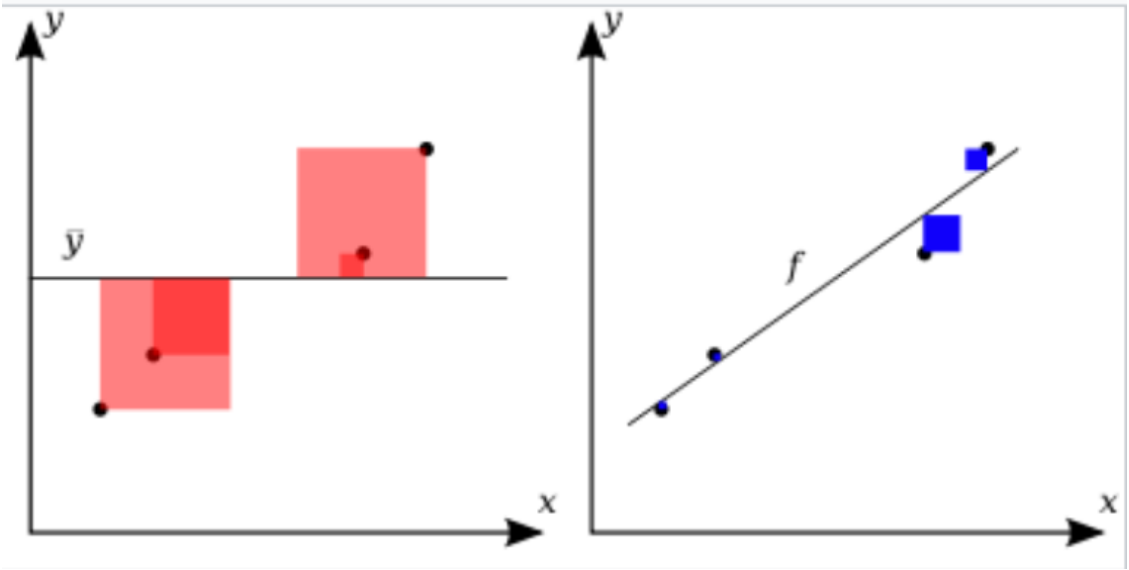
- By providing information on the X variable, we are able to reduce the amount of variation in Y from TSS to SSE. The proportion reduction in TSS is called **R-squared** (coefficient of determination):

$$R^2 = \frac{TSS - SSE}{TSS} \tag{7}$$

- It can be shown that $R^2 = \rho^2$ in the case of bivariate linear regressions.
- For example, earlier we know that the correlation between income and education in an occupation is 0.5776, so we know that the R^2 in this bivariate regression is $0.5776^2 = 0.3336$. We can also find this in the STATA output.

Coefficient of Determination: R-squared

- If we have no other variables, then the best we can do is to predict Y using the mean of Y. In this case, the variance of Y will be the "residual," and the R^2 is zero.
- The better the linear regression (on the right) fits the data in comparison to the simple average (on the left graph), the closer the value of R^2 is to 1.

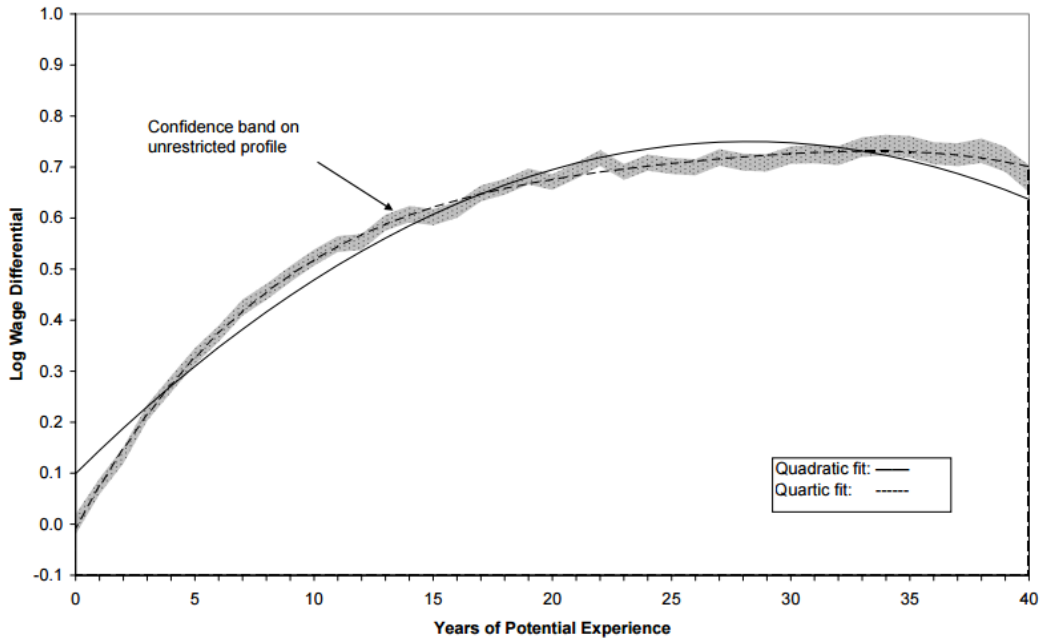


Some more thoughts on linear regressio models...

Non-linear relationships?

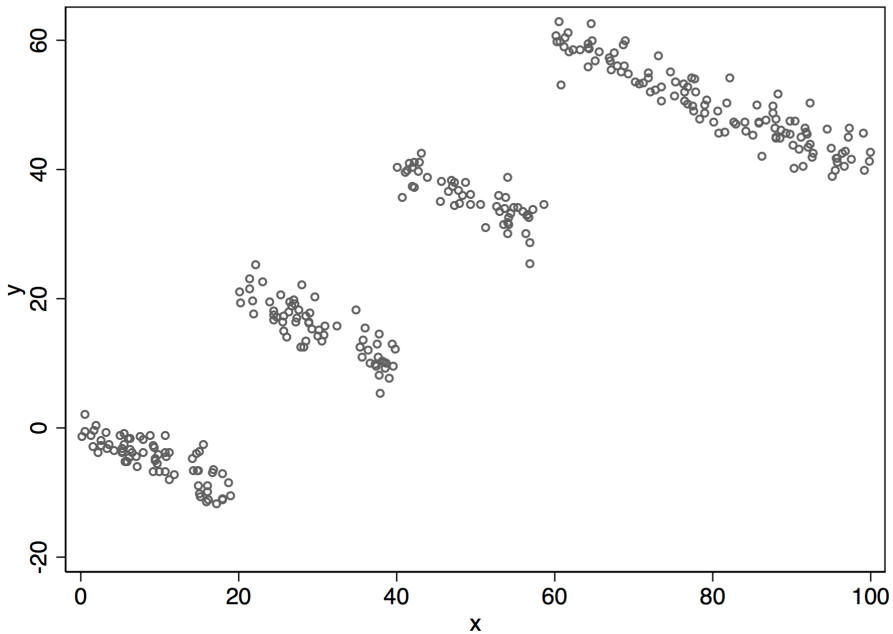
- What would you do when the relationship is non-linear in X? From Lemieux 2003:

Figure 11.8: Experience Profiles for Men, 1999-2001 CPS



Ecological Fallacy

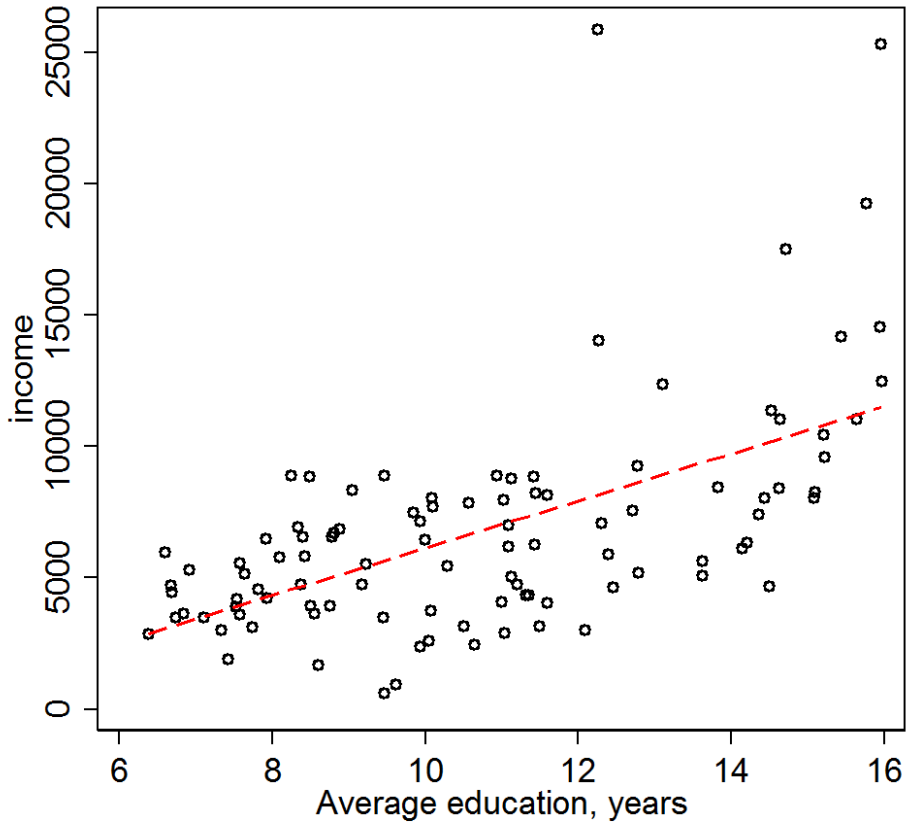
- Does the association between groups always consistent with the association within groups?
- Look at the following problem of an ecological fallacy:



- Ecological fallacy occurs when the relationship between two variables differs within subgroups compared to that observed for the aggregated data. (Simpson's paradox)

Sensitivity to Outliers

- Is the OLS estimator sensitive to outliers in the data?
- It's always a good idea to see the scatterplot of your data first:



Sensitivity to Outliers

- Now let's run the regression excluding those with income greater than 20,000 (potentially outliers):

```
. regress income educat if income<=20000
```

Source	SS	df	MS	Number of obs	=	100
				F(1, 98)	=	59.46
Model	415369013	1	415369013	Prob > F	=	0.0000
Residual	684601386	98	6985728.43	R-squared	=	0.3776
				Adj R-squared	=	0.3713
Total	1.1000e+09	99	11110812.1	Root MSE	=	2643.1

income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educat	758.5689	98.37479	7.71	0.000	563.3473	953.7904
_cons	-1672.395	1082.481	-1.54	0.126	-3820.544	475.7534

Sensitivity to Outliers

- Some practical strategies to deal with outliers:
- The simple strategy is to drop them!
- But when data size is large and you have a number of variables, it is not always straightforward to detect outliers.
- Taking the log may help:

```
. gen logincome=log(income)

. regress logincome educat
```

Source	SS	df	MS	Number of obs	=	102
				F(1, 100)	=	42.94
Model	10.6141108	1	10.6141108	Prob > F	=	0.0000
Residual	24.7168977	100	.247168977	R-squared	=	0.3004
				Adj R-squared	=	0.2934
Total	35.3310085	101	.349811966	Root MSE	=	.49716

logincome	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educat	.1188136	.018131	6.55	0.000	.0828422	.1547849
_cons	7.384684	.200818	36.77	0.000	6.986266	7.783101

Sensitivity to Outliers

- Another strategy is the run the **quantile regression** with quantile set at the 50th percentile, which predicts conditional median instead of *conditional mean*:

```
. qreg income educat, quantile(0.5)
Iteration  1:  WLS sum of weighted deviations =  120436.36

Iteration  1: sum of abs. weighted deviations =  120611.73
Iteration  2: sum of abs. weighted deviations =  118824.04
Iteration  3: sum of abs. weighted deviations =  118301.94

Median regression                               Number of obs =        102
  Raw sum of deviations   143188 (about 5902)
  Min sum of deviations 118301.9                Pseudo R2     =     0.1738
```

income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educat	672.5275	138.8062	4.85	0.000	397.14	947.9149
_cons	-876.0548	1537.412	-0.57	0.570	-3926.236	2174.126

Sensitivity to Outliers

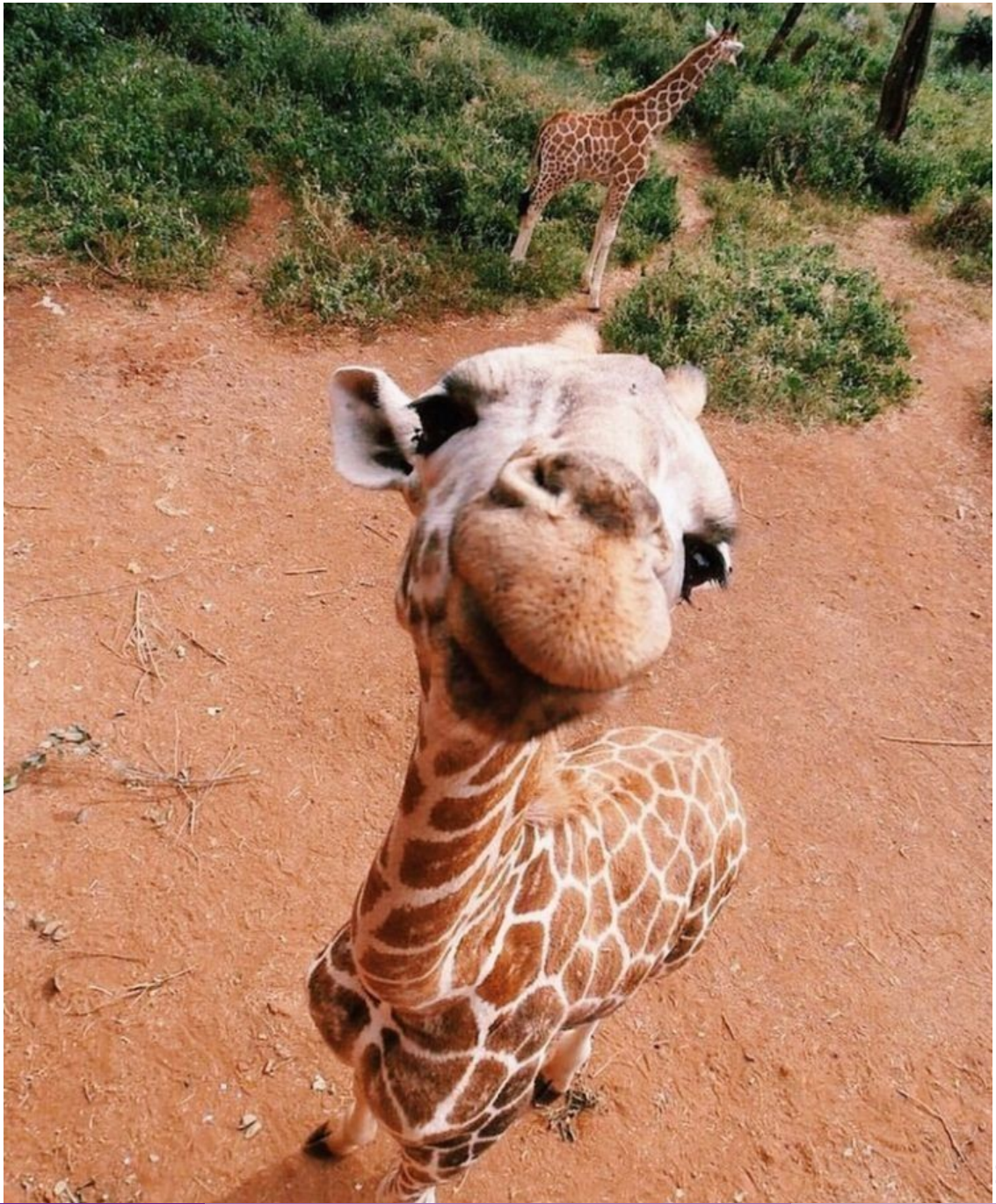
```
. qreg income educat if income<=20000, quantile(0.5)
Iteration  1:  WLS sum of weighted deviations =  101796.81

Iteration  1: sum of abs. weighted deviations =  101643.05
Iteration  2: sum of abs. weighted deviations =  101303.44
Iteration  3: sum of abs. weighted deviations =  101239.31

Median regression                               Number of obs =        100
  Raw sum of deviations 123496.5 (about 5811)
  Min sum of deviations 101239.3                Pseudo R2     =     0.1802
```

income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educat	650.4835	137.392	4.73	0.000	377.8336	923.1335
_cons	-705.6459	1511.812	-0.47	0.642	-3705.788	2294.496

```
. di 650.4835/672.5275
.96722216
```



Multiple Regression

The Multivariate Case

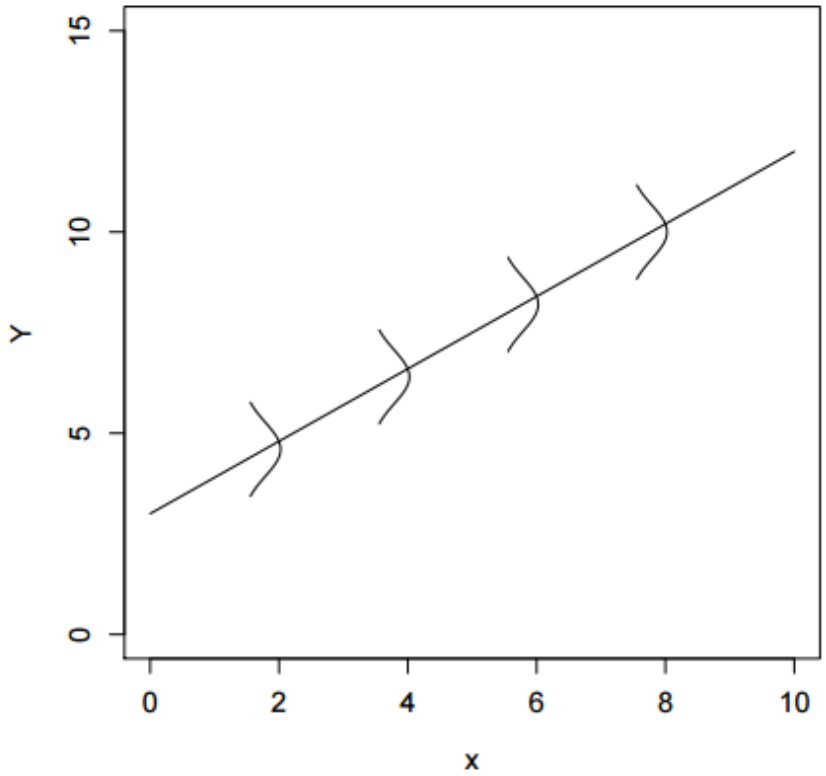
- Now, we are extending the equation to the multivariate case with k independent variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \epsilon. \quad (8)$$

- Recall that in the bivariate case, we want to estimate two parameters: β_0 and β_1. In the multivariate case, we want to estimate $k + 1$ parameters.
- How to interpret the parameters in Eq.(8)?
  1. β_0 is the intercept, that is, the expected value of Y when all $X's$ are set to zero.
  2. β_1 is the slope on X_1, which measures the change in Y with respect to X_1, *holding other factors fixed*.
  3. β_k is the slope on X_k, which measures the change in Y with respect to X_k, *holding other factors fixed*.

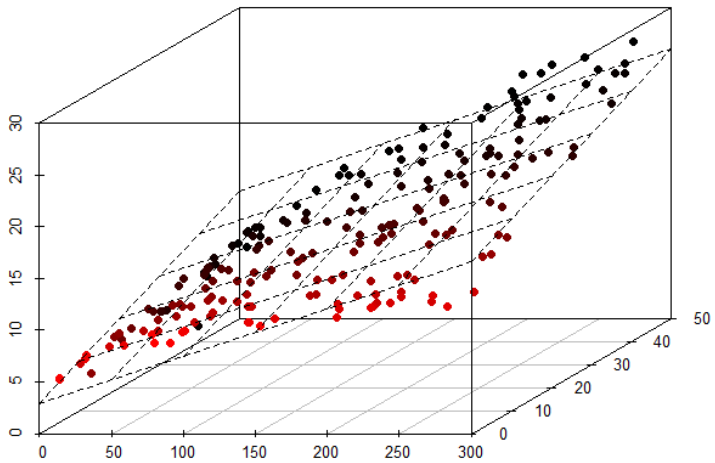
The Multivariate Case

- Remember that in the bivariate regression model, we predict the expectation of Y using X by fitting a straight line in a two-dimensional plain, so that the conditional mean of Y falls on this line:



The Multivariate Case

- When we work with two instead of one predictors, we are fitting a flat plane in a three-dimentional space, so that the conditional mean of Y falls on this plane:



- No matter how many explanatory variables we include in our model, there will always be factors we cannot include. Those factors may be unobservable (including "luck"), or they are not measured in the dataset we work with. They are in the error term.

An Example: Predicting College GPA

- Let's consider a simple example of predicting college GPA using high school GPA and ACT score.

$$GPA_{col} = \beta_0 + \beta_1 GPA_{hs} + \beta_2 ACT + \epsilon. \tag{9}$$

- In Eq.(9), once we have our estimation for the $\beta's$, we can predict change in GPA by the change in GPA_{hs} and the change in ACT:

$$\Delta GPA_{col} = \hat{\beta}_1 \Delta GPA_{hs} + \hat{\beta}_2 \Delta ACT \tag{10}$$

- How to interpret the $\beta's$ in this equation?

Partial Regression Coefficients

- The $\beta_1, \beta_2, ...\beta_k$ in Eq.(8) are also called **partial regression coefficients**. That is, holding other independent variables constant in the model, then a unit change in X_k will cause a change by β units in the predicted value of Y, or the expectation of Y ($E(Y|X)$).
- For example, consider the following prediction equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \tag{11}$$

Partial Regression Coefficients

- When $x_1 = a$, $x_2 = b$, we have:

$$y^* = \hat{\beta}_0 + \hat{\beta}_1 a + \hat{\beta}_2 b \tag{12}$$

- And when $x_1 = a + 1$, $x_2 = b$ (that is, one unit increase in x_1 but holding x_2 constant), we have:

$$y^{**} = \hat{\beta}_0 + \hat{\beta}_1 (a + 1) + \hat{\beta}_2 b \tag{13}$$

- Taking the difference between Eq.(13) and Eq.(12) we get:

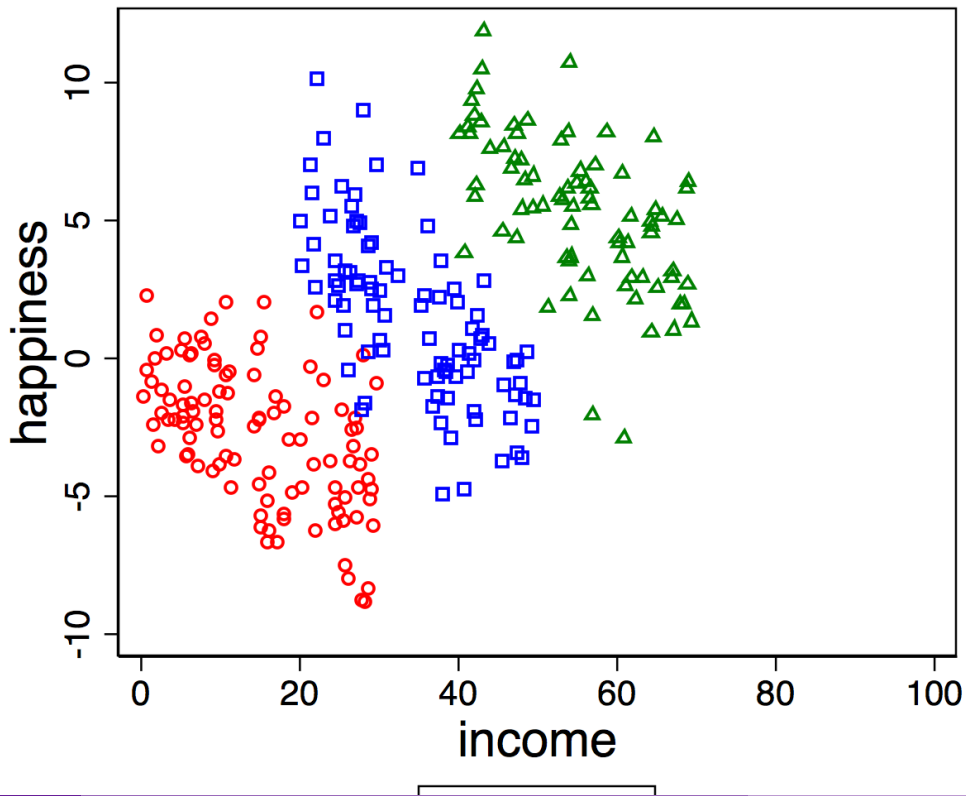
$$\begin{aligned} \Delta y &= y^{**} - y^* && (14) \\ &= (\hat{\beta}_0 + \hat{\beta}_1 (a + 1) + \hat{\beta}_2 b) - (\hat{\beta}_0 + \hat{\beta}_1 a + \hat{\beta}_2 b) && (15) \\ &= \beta_1 && (16) \end{aligned}$$

"Holding Others Constant"

- What does "holding others constant" (i.e. *Ceteris Paribus*) mean?
  1. It means that we are looking at the effect of one predictor (x_1), assuming other predictors are held at the same value.
  2. In other words, we are comparing y among individuals with different values on x_1 but the same value on all other variables.
  3. Importantly, the multiple regression model provides a "holding others constant" interpretation, even though the variables in reality are not independent from each other.
  4. For example, we know that high school GPA and ACT score is correlated in the sample of college students, but the coefficients from the multiple regression model will tell us the effect of high school GPA under the assumption that ACT is held constant.

"Holding Others Constant"

- Let's look at a visual illustration of "conditional on" or "holding other factors constant":
- In the figure below, the different colors/shapes of the dots represents individuals' income (X-level) and level of happiness (Y-axis) in three different countries (red, blue, and green).



Interactions and Quadratic Functions

Interactions

- Note that one important implication of the model specification in Eq.(11) is that the *partial effects* of x_1 and x_2 on y do not depend on either other.
- In other words, the partial effects of x_1 and x_2 only depends on their own corresponding coefficient (β_1 and β_2 respectively).
- But what if the effect of x_1 and x_2 depends on each other?

Interactions

- For example, when we study the effect of parental education level (edu) and parental income (inc) on the child's cognitive development (cog), we may specify the following prediction equation:

$$E(cog|edu, inc) = \beta_0 + \beta_1 edu + \beta_2 inc \tag{17}$$

- However, we may expect that not only do parental education and parental income have their independent effect on child's cognitive development, but also the effect of parental education is greater among higher-income parents.

Interactions

- This means that there may be an *interaction effect* between parental education and parental income. So we re-specify Eq.(17) as follows:

$$E(cog|edu, inc) = \beta_0 + \beta_1 edu + \beta_2 inc + \underbrace{\beta_3 \cdot edu \times inc}_{\text{interaction term}} \quad (18)$$

- When the interaction term is non-zero (i.e. $\beta_3 \neq 0$), it means that the effect of *edu* depends on the value of *inc*, or alternatively speaking, the effect of *inc* depends on the value of *edu*.

Interactions

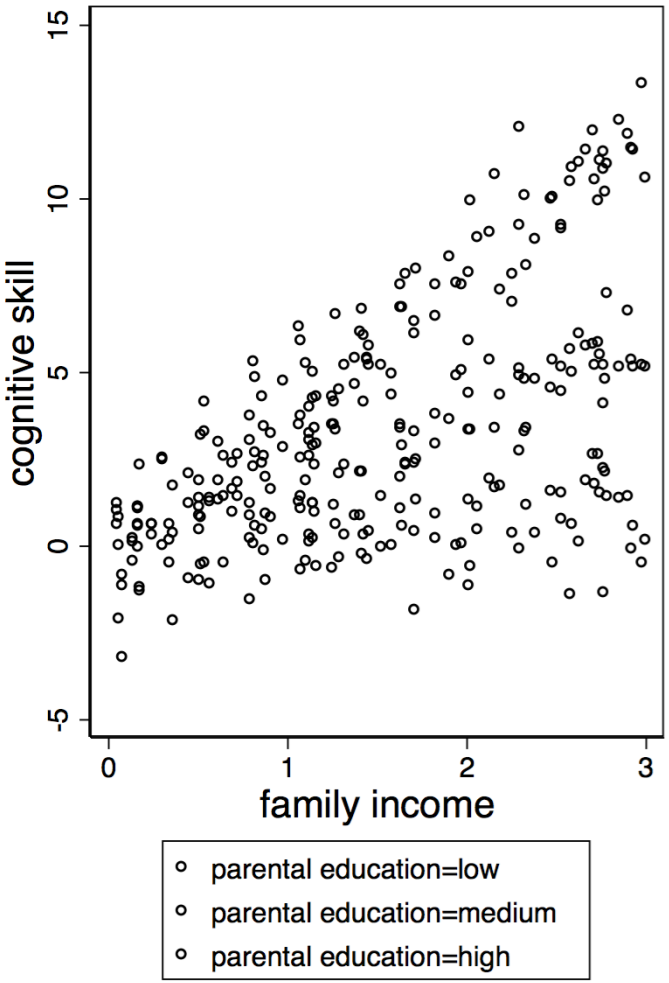
$$E(cog|edu, inc) = \beta_0 + \beta_1 edu + \beta_2 inc + \underbrace{\beta_3 \cdot edu \times inc}_{\text{interaction term}} \quad (19)$$

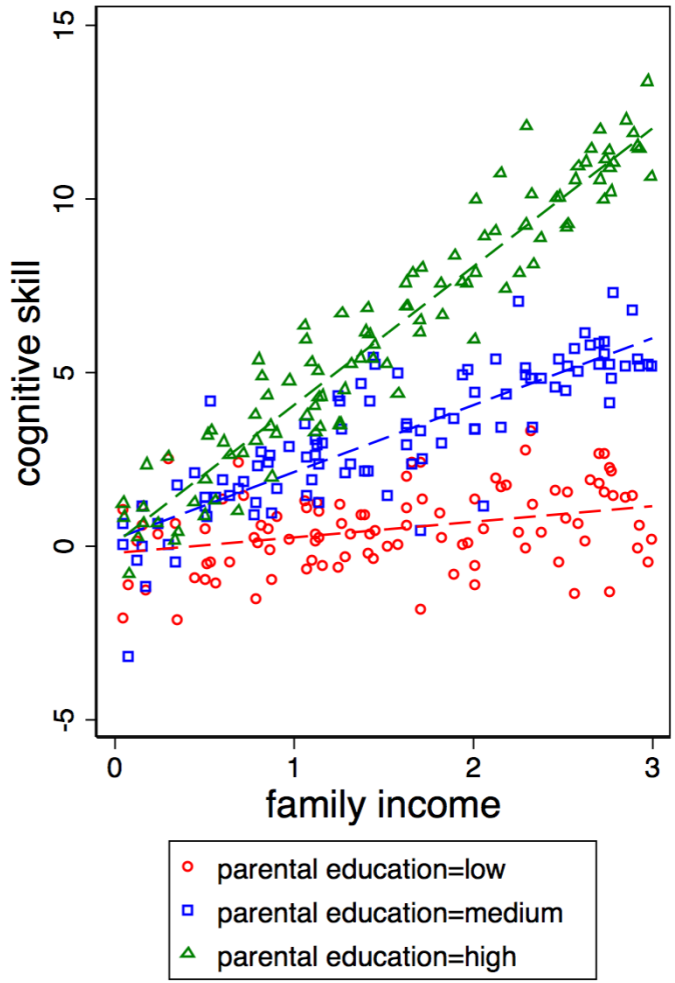
- Specifically, if $\beta_3 > 0$, then the effect of parental education is greater among higher-earning parents, or we can say the effect of parental income is greater among highly-educated parents.
- If $\beta_3 < 0$, then the effect of parental income is greater among lower-earning parents, or we can say the effect of parental income is greater among poorly-educated parents.
- Question to think about: what if $\beta_3 = 0$? What does this mean?

↳ no interaction effect

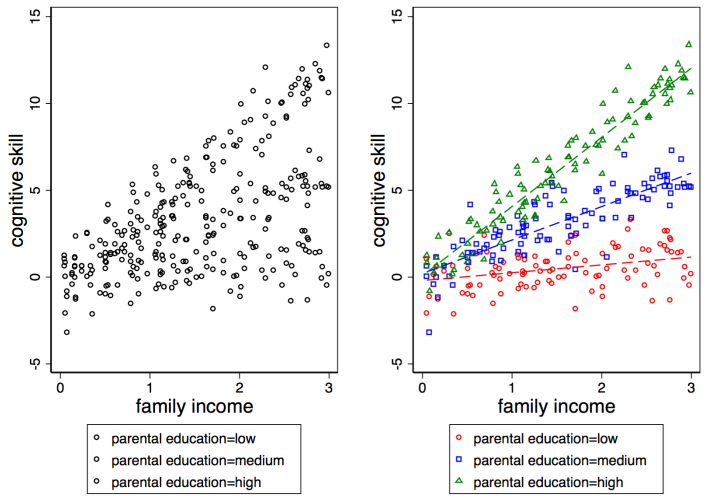
Interactions: A Visual Illustration

- For a visual illustration, let's look at the two figures below:





Interactions: A Visual Illustration



- The figure on the left does not differentiate between families with different parental education levels, while the figure on the right looks at the effect of family income at low, medium, and high levels of parental education.
- Can you tell from the figure whether there is likely to be a positive or negative interaction coefficient (β_3)?
- In practice, looking at the relationship between two variables *at different levels of a third variable* is helpful for understanding the **heterogeneity** in the effects.

Taking the partial derivatives

- To further illustrate the partial effect of *edu* and *inc* on *cog*, let's take the partial derivatives of Eq.(19) with respect to *edu* and *inc* separately:

$$\frac{\partial E(cog|edu, inc)}{\partial edu} = \beta_1 + \beta_3 \cdot inc \tag{20}$$

$$\frac{\partial E(cog|edu, inc)}{\partial inc} = \beta_2 + \beta_3 \cdot edu \tag{21}$$

- Or, we can use the "Δ" language, assuming that income is fixed, then the change in the conditional mean of the outcome variable is equal to:

$$\Delta E(cog|edu, inc) = \beta_1 \Delta edu + \beta_3 \cdot inc \cdot \Delta edu \tag{22}$$

Heterogeneous Effect

- In empirical research, researchers often use the interaction terms to examine the heterogeneity in the effect of key independent variables.
- In this example, testing for the statistical significance and the direction of the interaction term (we'll learn how to test it in the next section) may have important implications for policy analysis.
- For example, if the interaction term has a non-zero coefficient, then, when offering tax subsidies to low-income families with children, we will need to take into account the education attainment of the parents in order to calculate the expected policy effect.

Heterogeneous Effect

- Another example: suppose we are interested in the effect of gender and marital status on earnings. But we also know that marriage may have different effect on earnings for men and women.
- What should we include as the interaction term in the model? And how would you interpret this?

marital status x gender

Non-linear Model Specification

- Let's now consider another type of model specification in multiple regression analysis.
- Let Y=attitudes towards immigration policies (attitude). A higher value on Y indicate greater support for path-to-citizen for illegal immigrants. We may hypothesize that Y depends on X=frequency of contact with immigrants (freq).
- However, we may think that the effect of frequency of contact with immigrants decreases as a person is in contact with more and more immigrant. In this case, we can specify a *quadratic* function:

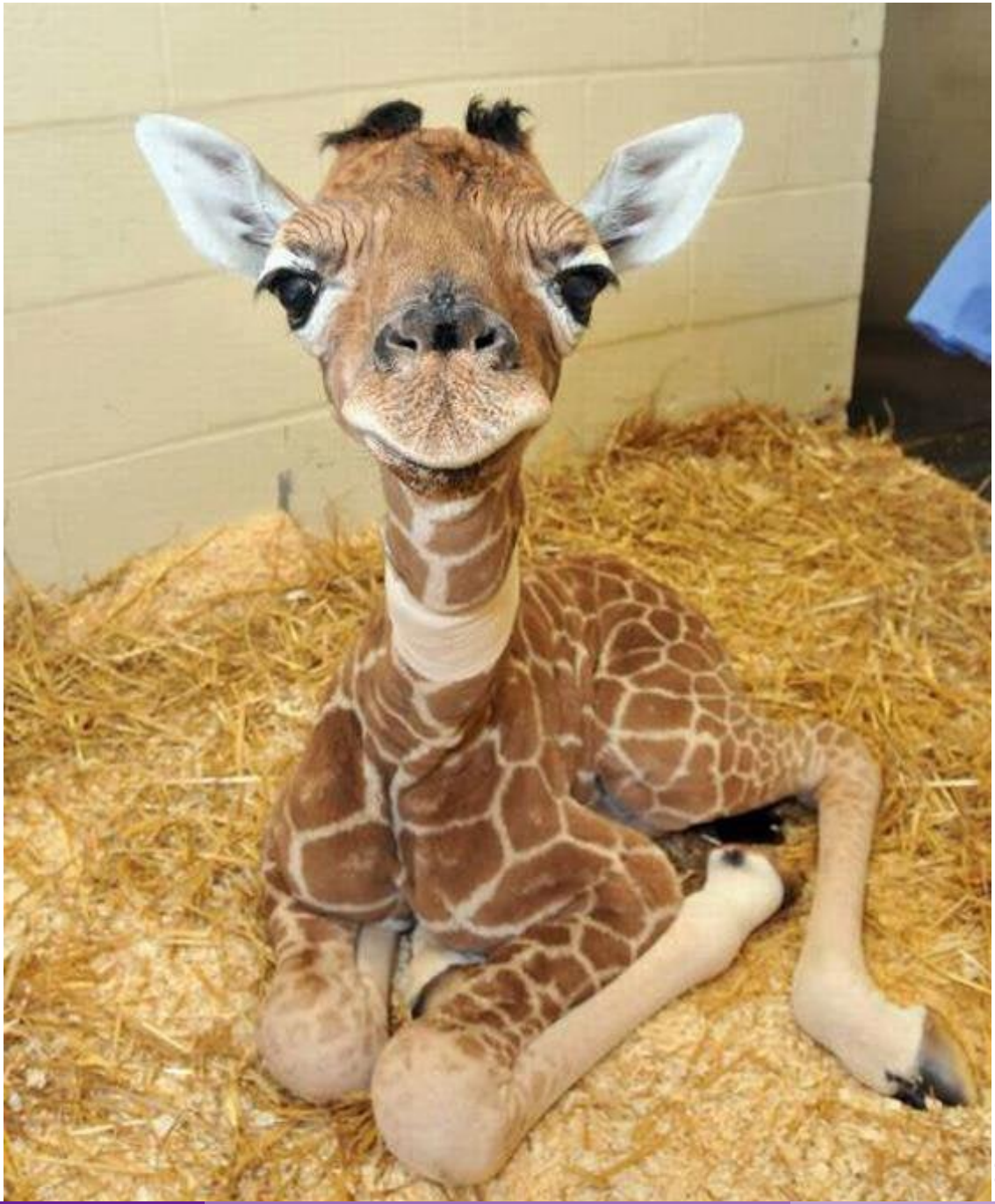
$$\Delta E(attitude|freq) = \beta_0 + \beta_1 \cdot freq + \beta_2 \cdot freq^2 \qquad (23)$$

Non-linear Model Specification

- The **quadratic function** allows for the effect of *freq* on *attitude* to depend on the level of *freq*. Why? Let's take the derivative of E(*attitude* — *freq*) with respect to *freq*:

$$\frac{dE(attitude|freq)}{d(freq)} = \beta_1 + \beta_2 \cdot freq \tag{24}$$

- That is, the rate by which the outcome variable changes in response to changes in *freq* depends on the level of *freq*.



Estimating Model Parameters

- Next, let's estimate the $\beta's$ using information from the sample.
- Recall that in the previous lecture, we discussed the OLS estimator, which is the estimators that minimizes the residual sum of squares. Similarly, in the multiple regression, the OLS estimator is a set of $\hat{\beta}'s$ that minimizes $\Sigma(y - \hat{y})^2$.
- Going back to our earlier graphic illustration, an OLS estimator in the two-predictor case will find the plane that minimizes the sum of squared vertical differences between the observations and the plane.
- So, what are the $\beta's$ that minimizes the sum of squared residuals? Consider the following prediction equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon. \tag{25}$$

Estimating Model Parameters

- It can be shown that the OLS estimators for β_0, β_1, and β_2 are:

$$\hat{\beta}_1 = \frac{Cov(x_1, y) \cdot Var(x_2) - Cov(x_2, y) \cdot Cov(x_1, x_2)}{Var(x_1) \cdot Var(x_2) - Cov(x_1, x_2)^2} \tag{26}$$

$$\hat{\beta}_2 = \frac{Cov(x_2, y) \cdot Var(x_1) - Cov(x_1, y) \cdot Cov(x_1, x_2)}{Var(x_2) \cdot Var(x_2) - Cov(x_1, x_2)^2} \tag{27}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x_1} - \hat{\beta}_2 \bar{x_2} \tag{28}$$

Estimating Model Parameters

Let's take a closer look at $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{Cov(x_1, y) \cdot Var(x_2) - Cov(x_2, y) \cdot Cov(x_1, x_2)}{Var(x_1) \cdot Var(x_2) - Cov(x_1, x_2)^2} \tag{29}$$

- It will be helpful to know that the slope estimates depends on these factors:
  1. Variances of independent variables;
  2. Covariance between x_1 and x_2
  3. Covariance between independent variables x_1, x_2 and the dependent variable y.

Estimating Model Parameters

- If you still remember, the estimator for β_1 in the bivariate case (previous lecture) is given by:

$$\hat{\beta}_1 = \frac{Cov(x, y)}{Var(x)} \tag{30}$$

- You can think of the bivariate regression estimator in Eq.(30) as a special case for the multiple regression case where $Cov(x_1, x_2) = 0$.

Extension: Writing Estimators in Matrix Forms

- In the multivariate case with k predictors, the formula for the OLS estimators will be very complicated.
- However, it will be more succinct (and actually very helpful) to express the model and estimators using matrix forms.
- First, we can express the sample data with number of observations n and number of independent variables k using a n by k matrix:

$$\begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \dots & x_{nk} \end{bmatrix} \tag{31}$$

Extension: Writing Estimators in Matrix Forms

- The outcome variable Y in the data can be expressed as a $n \times 1$ vector

$$\begin{bmatrix} y_1 \\ y_2 \\ .. \\ y_n \end{bmatrix} \tag{32}$$

- We can express the unknown parameters as a $(k+1) \times 1$ vector. (Why $(k+1) \times 1$? Don't forget the constant term!)

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ .. \\ \beta_k \end{bmatrix} \tag{33}$$

Extension: Writing Estimators in Matrix Forms

- Putting them together we have:

$$\begin{bmatrix} y_1 \\ y_2 \\ .. \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ .. \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ .. \\ \epsilon_n \end{bmatrix} \tag{34}$$

- We can rewrite Eq.(34) as:

$$Y = X\beta + \epsilon, \tag{35}$$

where Y, X, β, and ϵ are matrices or vectors, and the operations are matrix operations.

Extension: Writing Estimators in Matrix Forms

- It can be shown that the vector of OLS estimators for vector β can be written as:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{36}$$

Hypothesis Testing in Multiple Regression

Hypothesis Testing in Multiple Regression

1. Next, let's look at the sample standard deviation of the OLS estimators:
2. The formula for the standard error of $\hat{\beta}_1$ is:

$$se_{\beta_1} = \sqrt{\frac{s_\epsilon^2}{(n-k-1)Var(x_1)(1-r_{x_1,x_2}^2)}} \tag{37}$$

3. As the formula suggests, the standard error of $\hat{\beta}_1$ is smaller when:
   (1) The variance of the corresponding variable x_1 is larger.
   (2) The correlation between the two independent variables are small.

rho/ρ is small

Hypothesis Testing in Multiple Regression

- Two additional notes on the standard error of β before we move on:
- First, In the special case where we have two uncorrelated variables (i.e. $r^2_{x_1,x_2} = 0$), Eq.(37) becomes:

$$se_{\beta_1} = \sqrt{\frac{s^2_\epsilon}{(n-k-1)Var(x_1)}} \tag{38}$$

(39)

This becomes the same standard error for the binary regression case (k=1) we showed in the previous lecture.

Hypothesis Testing in Multiple Regression

- Second, In a more general case of a multiple regression model with k independent variables, the standard error of the coefficient on x_j, β_j, can be written as:

$$se_{\beta_j} = \sqrt{\frac{s_\epsilon^2}{(n-k-1)Var(x_j)} \cdot \frac{1}{1-R_j^2}}, \tag{40}$$

  where R_j^2 is the proportion of the total variation in x_j that can be explained by the other independent variables.

- The term $\frac{1}{1-R_j^2}$ is termed the variance inflation factor (VIF). Because R_j^2 lies between zero and one, the VIF is greater than one. The term captures the "inflation" in the sampling variance (and thus the standard error) of β_j due to the correlation between x_j and other independent variables.

Hypothesis Testing in Multiple Regression

- The t-statistic for the significance test of whether $\hat{\beta}_1$ is significantly different from zero is:

$$t = \frac{\beta_1 - 0}{se_{\beta_1}} = \frac{\beta_1}{se_{\beta_1}}, \tag{41}$$

- When $\beta_1 = 0$ (i.e. H_0 is true), t follows t distribution with the degree of freedom equal to $n - k - 1$. Sample - k vars. - 1
- If β_1 is statistically significant, it means that conditional on other independent variables, the variable x_1 has a statistically significant effect on the outcome variable y.

Testing whether *any* of the explanatory variables are statistically different from zero

- Next, we test whether the explanatory variables collectively have effect on y.
- We test the null hypothesis H_0: $\beta_1 = \beta_2 ... = \beta_k = 0$ (i.e. y independent of all the explanatory variables) against the alternative hypothesis H_a: At least one $\beta_j \neq 0$.
- The test judges whether using the linear regression function with $x_1, x_2, ...x_k$ together to predict y is better than using $\bar{y}$.

The F Statistic

- The test statistic is written as:

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)}, \tag{42}$$

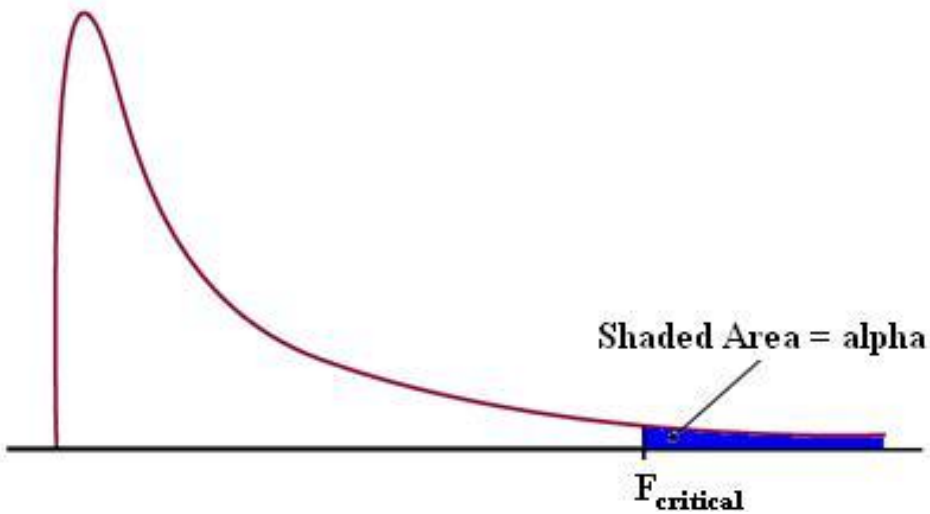
where R^2 is the coefficient of multiple determination (varying from zero to one), similar to what we have defined in the binary regression case:

$$R^2 = \frac{TSS - SSE}{TSS} = \frac{\Sigma(y-\bar{y})^2 - \Sigma(y-\hat{y})^2}{\Sigma(y-\bar{y})^2} \tag{43}$$

- We can think of the F-statistic as the **explained** variance divided by the **unexplained** variance.

The F Statistic

- When H_0 is true, the F-statistic follows F distribution with $df = (k, n - k - 1)$.



- When the calculated F-statistic is greater than the critical value, we reject the null hypothesis, and the evidence is leaning towards our alternative hypothesis (at least one of the $\beta's$ is non-zero).

The F Statistic for Nested Models

- Two models are **nested** if both contain the same terms and one has at least one additional term.

$$\text{Restricted model: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon. \quad (44)$$

$$\text{Full model: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon. \quad (45)$$

- The first model is **nested** within second model.
- How do we decide whether the more complex (full) model contributes additional information about the association between y and the predictors?
- This is equivalent to testing the null hypothesis H_0: $\beta_3 = \beta_4 = 0$ against the alternative hypothesis H_a: At least one $\beta_j \neq 0, j = 3$ or 4..

The F Statistic for Nested Models

- The test statistic can be written as:

$$F = \frac{(SSE_R - SSE_F)/p}{SSE_F/(n-(k+p+1))} \tag{46}$$

- where k is the number of predictors in the restricted model and p is the number of **additional** predictors added in the full model.
- The F statistic follows an F distribution.
- Parsimonious *models are preferable to big models as long as both have similar predictive power.*
- Note that If models are not nested, cannot use the F−test above to choose between one and another.

Coefficient of Multiple Determination

- Let's now go back a little bit to look at the coefficient of multiple determination R^2:

$$R^2 = \frac{TSS - SSE}{TSS} = \frac{\Sigma(y - \bar{y})^2 - \Sigma(y - \hat{y})^2)}{\Sigma(y - \bar{y})^2} \tag{47}$$

- R^2 equals the squared term of the correlation between outcome variable and our linear prediction: $R^2 = r_{y,\hat{y}}$.
- R^2 falls between 0 and 1. $R^2 = 1$ when all residuals are 0. $R^2 = 0$ when $\beta_1 = \beta_2 ... = \beta_k = 0$. The prediction line become a flat line with $\hat{y} = \bar{y}$.
- The larger the value of R^2, the better the set of explanatory variables collectively predict y.

Coefficient of Multiple Determination

- Note that if we add an additional explanatory variable to the model, R^2 will not decrease. That is, it is impossible to explain less variation in y by incorporating one additional variable.
- This raises an important problem: in multiple regression models, we can increase R^2 "artificially" by adding more and more explanatory variables but without meaningful interpretations.
- To account for the potentially spurious increase in R^2 by number of predictors, we often report *adjusted* R^2 as an alternative:

$$R^2_{adj} = R^2 - (1 - R^2)\frac{k}{n - k - 1}. \tag{48}$$

- The adjustment is done by placing a "**penalty**" on increase in k. To see this, note that the larger the k, the greater the term that we subtract from the original R^2.

Example: Multiple Regression

- Suppose a researcher collected a sample of 600 students with measures of students' reading test score and three psychological variables: *locus of control*, *self-concept* and *motivation*.
- The researcher is interested in knowing: (1) whether at least one of the three psychological factors has an effect on reading score; (2) whether each of the psychological variables has an affect on reading score; (3) How much variations in the student's reading score can be explained by the set of three psychological factors.

Example: Multiple Regression

- Let's summarize these variables:

```
> describe(score_minidata)
                vars   n mean    sd median trimmed   mad   min   max range  skew kurtosis   se
read               1 600 51.9 10.10  51.86   51.76 10.99 24.62 80.59 55.97  0.09    -0.25 0.41
motivation         2 600  0.0  0.82  -0.01   -0.01  0.79 -2.75  2.58  5.33  0.09    -0.07 0.03
self_concept       3 600  0.0  0.71   0.03    0.01  0.73 -2.53  2.09  4.63 -0.12    -0.09 0.03
locus_of_control   4 600  0.1  0.67   0.08    0.09  0.67 -2.00  2.21  4.20  0.08    -0.07 0.03
>
```

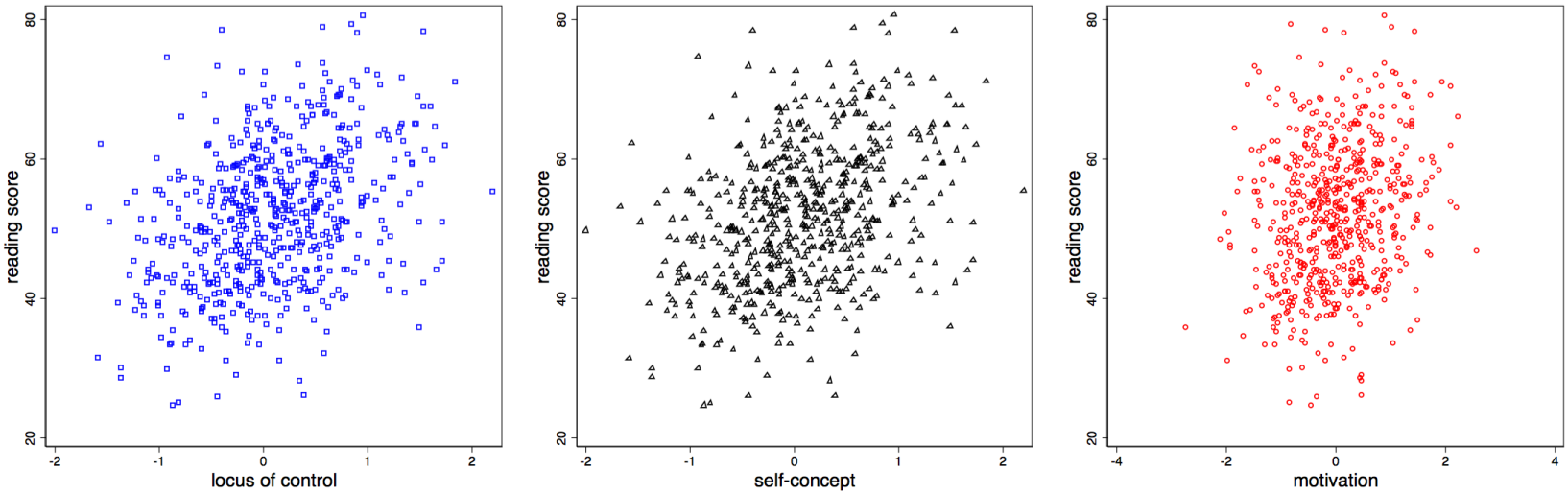
- Then we can look at the correlation between pairs of these variables.

```
> rcorr(as.matrix(score_minidata), type="pearson")
                  read motivation self_concept locus_of_control
read              1.00       0.21         0.06              0.37
motivation        0.21       1.00         0.29              0.25
self_concept      0.06       0.29         1.00              0.17
locus_of_control  0.37       0.25         0.17              1.00

n= 600
```

Example: Multiple Regression

- It's also helpful to show the scatter plots of these predictors and outcome variable:



Example: Multiple Regression

- Now, we estimate the multiple regression model in R:

```
> # regression
> linear_regression <- lm(read~motivation + self_concept + locus_of_control,data=score_minidata)
> summary(linear_regression)

Call:
lm(formula = read ~ motivation + self_concept + locus_of_control,
    data = score_minidata)

Residuals:
     Min       1Q   Median       3Q      Max
-28.3235  -6.5275  -0.3192   6.1475  29.2204

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      51.3937     0.3840 133.832  < 2e-16 ***
motivation        1.6787     0.4936   3.401 0.000717 ***
self_concept     -0.5457     0.5662  -0.964 0.335498
locus_of_control  5.2241     0.5885   8.876  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.305 on 596 degrees of freedom
Multiple R-squared:  0.1559,    Adjusted R-squared:  0.1517
F-statistic:  36.7 on 3 and 596 DF,  p-value: < 2.2e-16
```

Example: Multiple Regression

- Can you calculate the R^2 and F-statistic by hand, using the total, model and residual sum of squares?

$$R^2 = \frac{TSS - SSE}{TSS} \tag{49}$$

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)}, \tag{50}$$

- Can you interpret these coefficients, their confidence intervals, and tell me whether each of the psychological variables has a statistically significant effect on reading score?

Example: Multiple Regression

- What proportion of the total variation in reading score is explained by the three psychological variables?
- From the results, can you tell me what would be the *correlation* between the **observed** reading score and the **predicted** reading score?
- We can check our calculation by doing the prediction and calculating the correlation in R:

```
> # prediction and correlation between read_hat and read
> score_minidata$read_hat <- fitted(linear_regression)
> cor(score_minidata$read_hat,score_minidata$read)
[1] 0.3948953
>
```

- Why is the adjusted R-squared slightly smaller than the original R-squared?

Example: Multiple Regression

- We can calculate the adjusted R-squared by hand:

$$R^2_{adj} = R^2 - (1 - R^2)\frac{k}{n - k - 1} \tag{51}$$

$$= 0.1559 - (1 - 0.1559) \cdot \frac{3}{600 - 3 - 1} \tag{52}$$

$$= 0.1517 \tag{53}$$