19. Multiple Linear Regression

Readings: Kleinbaum, Kupper, Nizam, and Rosenberg (KKNR): Chs. 6, 9-10

SAS: PROC REG

Homework: Homework 7 due by 11:59 pm on October 31
Homework 8 due by 11:59 pm on November 7
Bonus Matrix Algebra Homework due by 11:59 pm on November 14

Overview

A) Re/Preview of Topics
B) Correlation
C) Multiple Linear Regression
D) MLR: Additional Diagnostics
E) MLR: Inference for the Independent Variables
F) Quick Review

A. Review (Lecture 18) / Current (Lecture 19) / Preview (Lecture 20)

Lecture 18:

- Simple Linear Regression Example (one covariate)
- Motivation for Multiple Linear Regression (more than one covariate)
- Complete Summary includes the following:
  - Point estimate (beta)
  - Measure of uncertainty (p-value)
  - Interval estimate (95% CI)
  - Decision (reject/ fail to reject)

Lecture 19:

- Correlation in terms of linear regression
- Multiple Linear Regression example
- Comparing full and reduced models (partial F-test)

Lecture 20:

- Linear Algebra Review
- Least Squares Estimation (LSE) using matrices

B. (Pearson) Correlation

Both correlation and regression are used to describe the linear relationship between two continuous variables (e.g., height and weight).

In general, correlation tends to be used when there is no identified response variable. It measures the strength and direction of the linear relationship between 2 variables.

Regression is used when there are identified response and explanatory variables. It is used to describe the relationship (quantitatively) between 2 or more variables.

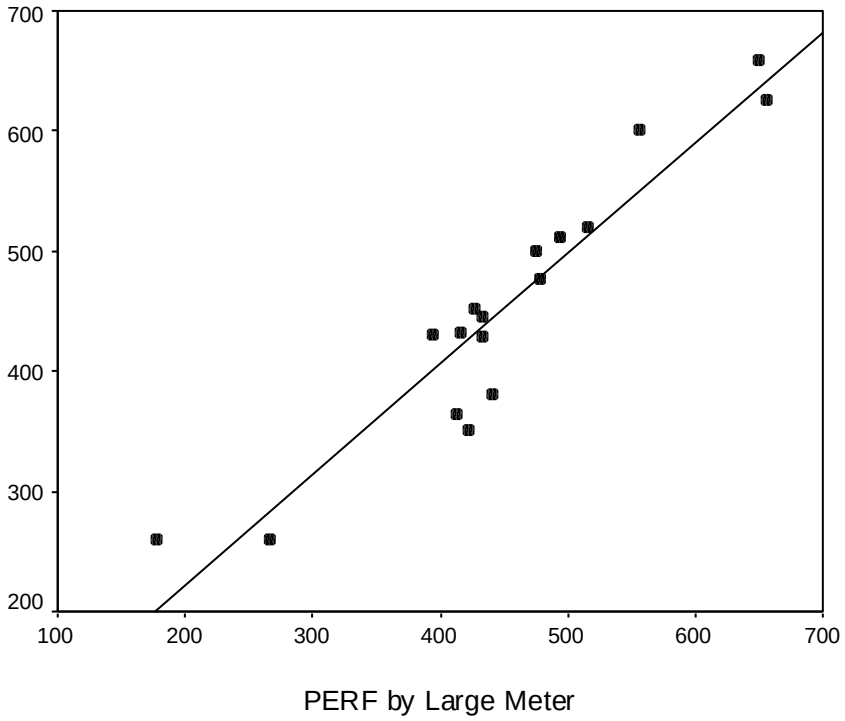
The **Pearson correlation coefficient** measures the strength of the ***linear*** association between two variables. It can be used to estimate the population correlation, ρ, and is defined by:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Recall: $S_X^2 = \frac{S_{XX}}{n-1}$, $S_Y^2 = \frac{S_{YY}}{n-1}$, $Cov(X,Y) = \frac{S_{XY}}{n-1}$

Example: Correlating Two Lung Function Meters (Wright vs. Mini-Wright)

Peak expiratory flow rate (PERF) can be measured using either the Wright Peak Flow Meter or the Mini-Wright Peak Flow Meter. Do their measurements "correlate"?



For this set of data, r=0.943, a strong, positive linear correlation.

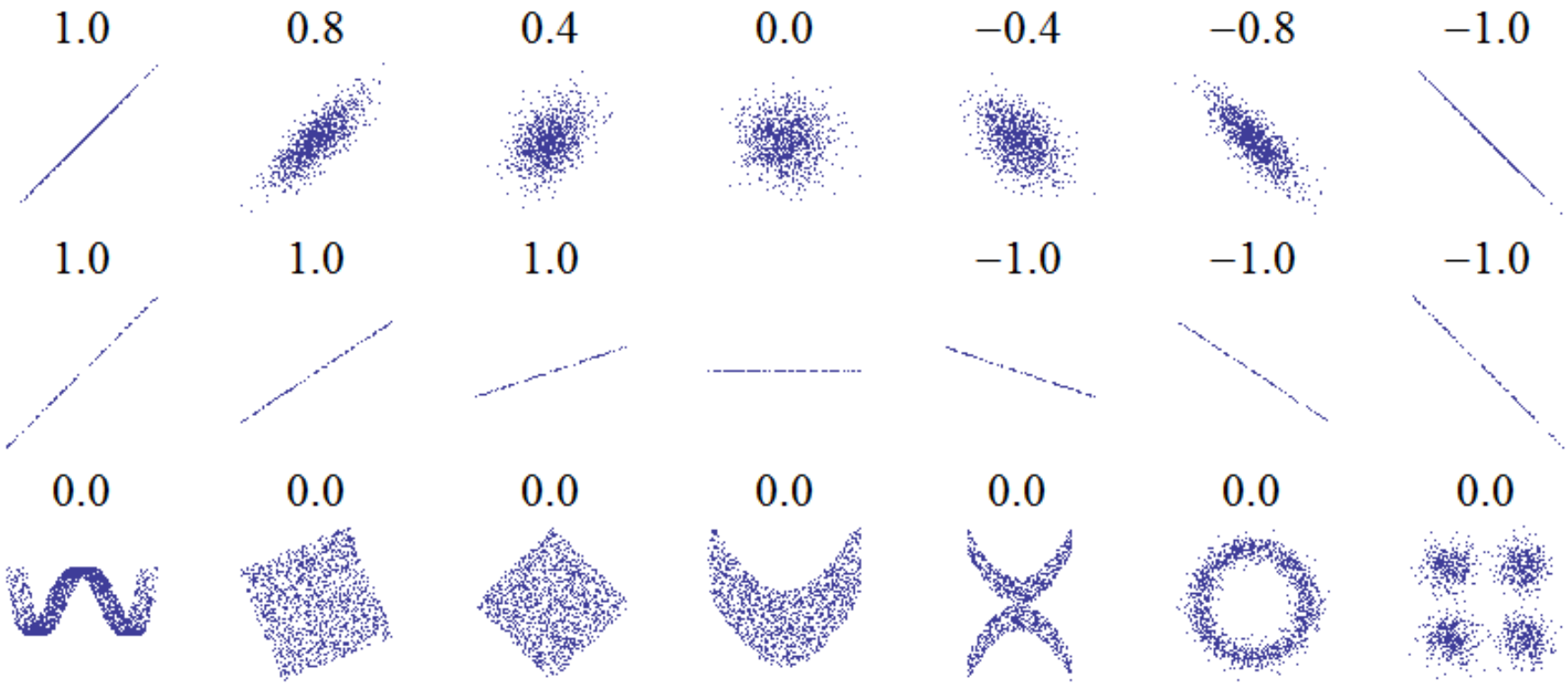
Note that this is a very different question than do their measurements "agree".

Properties of Correlations

A correlation can range in value between –1 and 1:

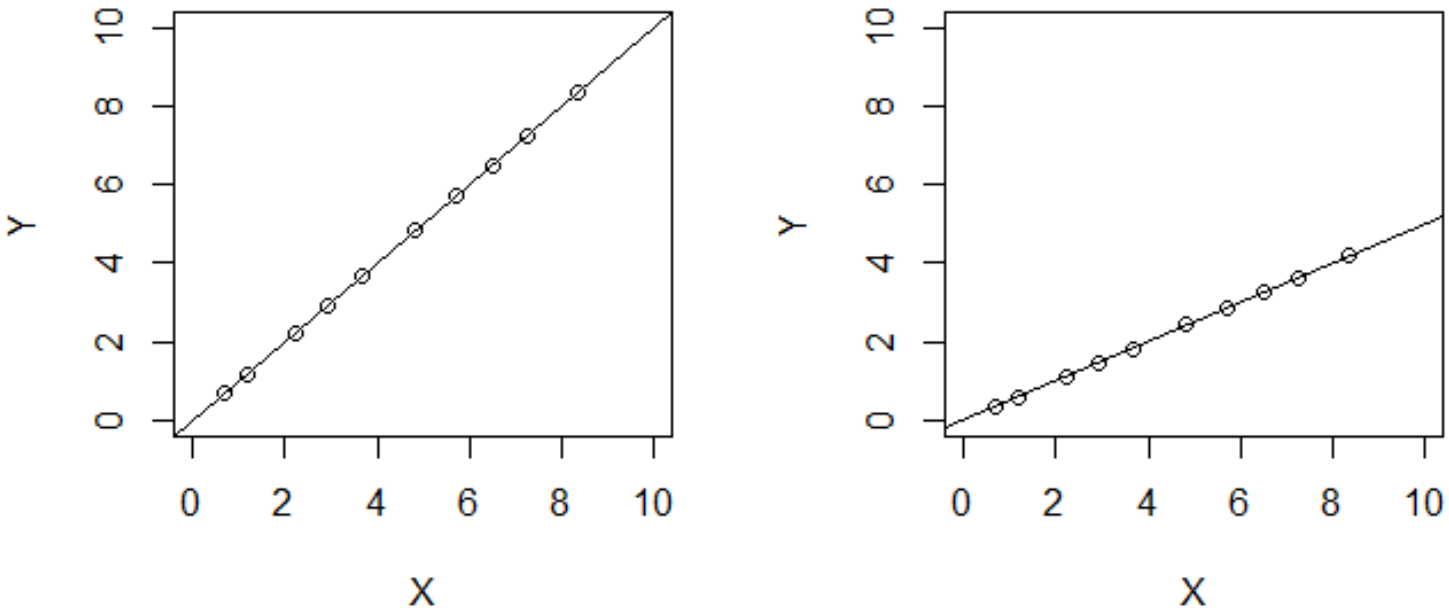
- *r* is a dimensionless quantity; i.e., *r* is independent of the units of measurement of *X* and *Y*.
- If the correlation is greater than 0, then as *X* increases *Y* increases and the two variables are said to be ***positively correlated***. *r* = 1 is perfect positive correlation.
- If the correlation is less than 0, then as *X* increases *Y* decreases and the two variables are said to be ***negatively correlated***. *r* = -1 is perfect negative correlation.
- If the correlation is 0 then there is no *linear* relationship between *X* and *Y*. The two variables are said to be ***uncorrelated***. The correlation is 0 when the covariance of X and Y is 0.
- The correlation coefficient is a measure of the strength of the linear trend relative to the variability of the data around that trend. Thus, it is dependent both on the magnitude of the trend and the magnitude of the variability in the data.

Examples of various Pearson correlation coefficients:



Source: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Correlation is **not** a measure of the magnitude of the slope of the regression line.

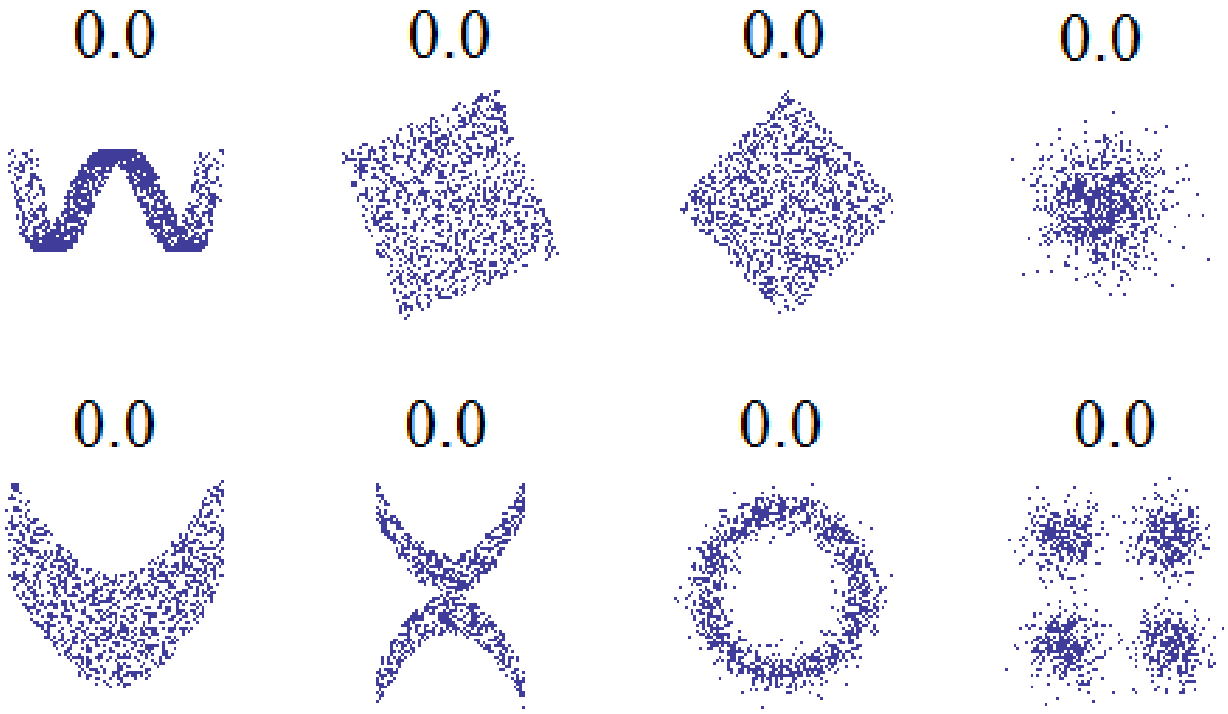


Both plots have $r = 1$, but very different slopes!

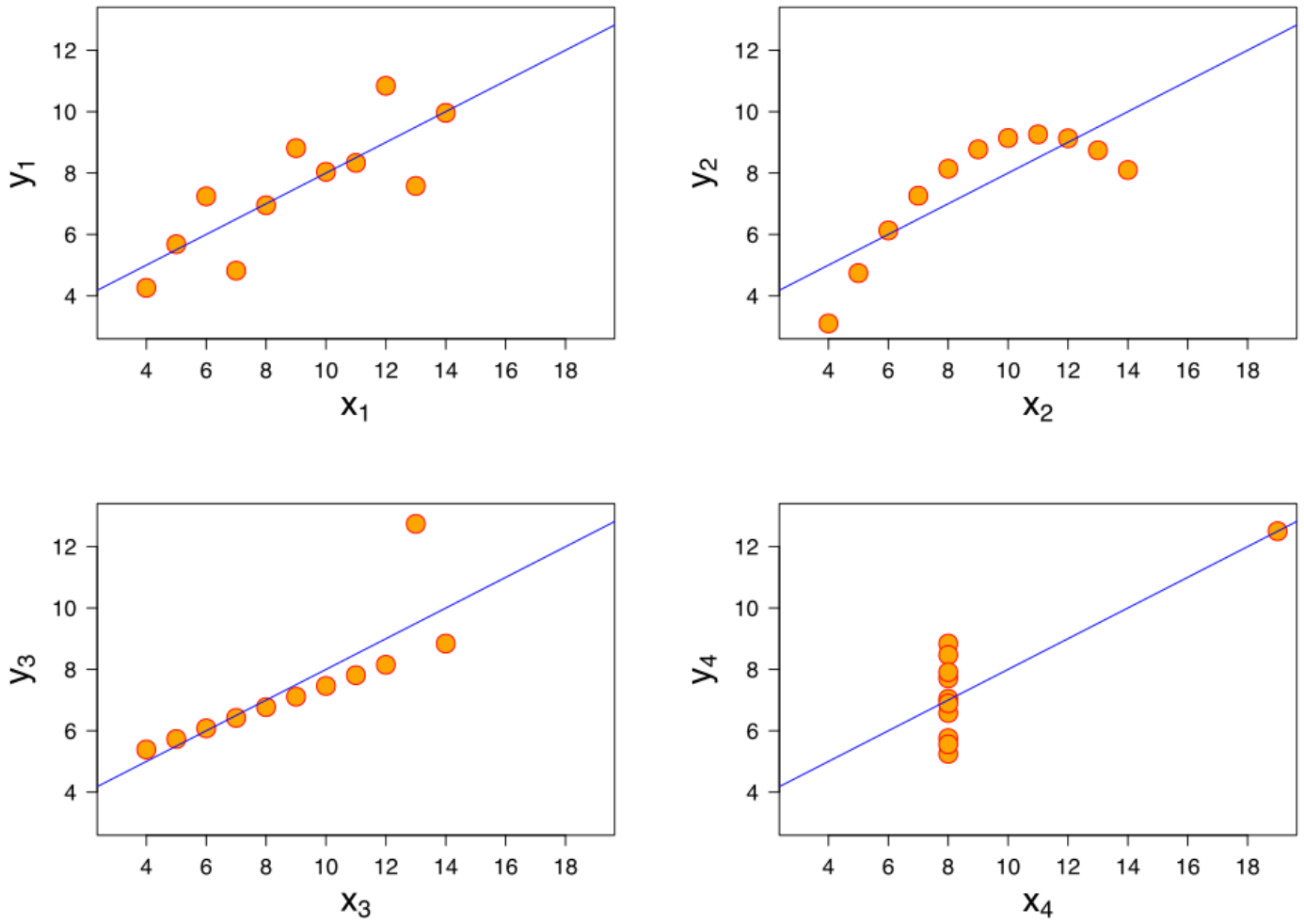
In linear regression we can (and do) estimate the amount *Y* increases/decreases with a one-unit change in *X*. However, this slope does depend on the units of measurement of *X* and *Y*, whereas the correlation is dimensionless.

Correlation is **not** a measure of the appropriateness of the straight-line model.

Examples where $r = 0$:



Correlation is **not** a measure of the appropriateness of the straight-line model.
Examples when *r* is high (*r*=0.816):

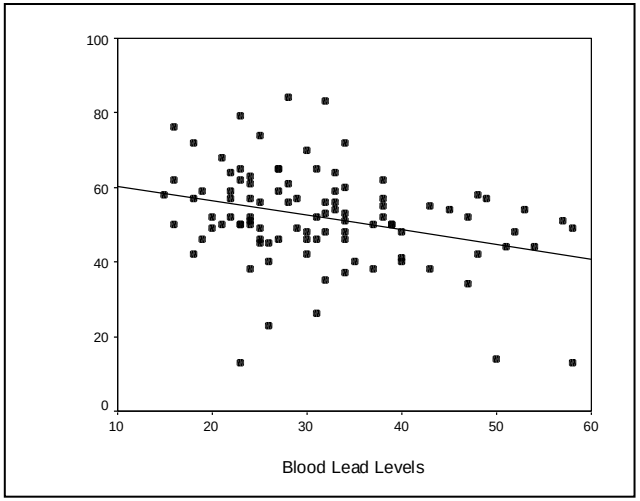


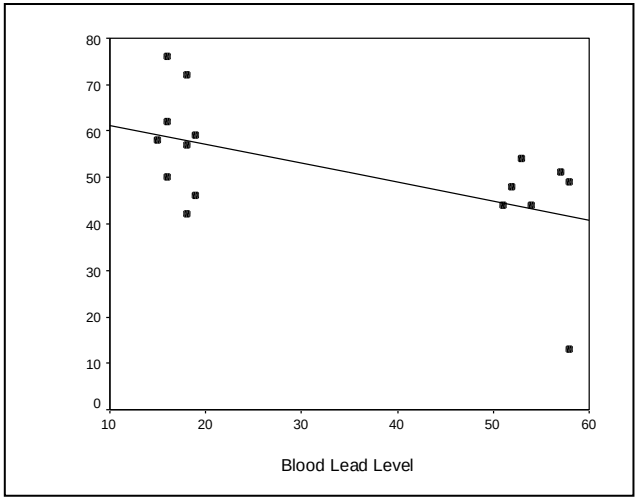
Source: https://en.wikipedia.org/wiki/Correlation_and_dependence

Assumptions

- For the correlation coefficient to be generalizable, you must have a random sample of the underlying population.
- A non-random sample will not reflect the true variability and will lead to bias in the correlation coefficient.
- The sample must consist of independent observations (i.e., one observation per subject).
- For valid inferences to be made about ρ, at least one of the two variables must be normally distributed in the population. Preferably both variables are approximately normally distributed.

Example: Neurological Function and Lead Exposure





r = -0.317

$\hat{Y}$ = 64.33 - 0.39×pb73

p = 0.001

r = -0.563

$\hat{Y}$ = 65.42 - 0.41×pb73

p = 0.023

Hypothesis Testing for Correlations

A basic test of whether or not there is a significant *linear* association between two continuous variables can be obtained by testing if the correlation coefficient is equal to 0.

$$H_0: \rho = 0 \text{ vs. } H_A: \rho \neq 0$$

A *T* statistic can be used to test the null hypothesis. The *T* statistic is defined as:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

Under the null hypothesis, the *T* statistic follows a *t*-distribution with *n*-2 degrees of freedom.

Rejecting H_0 is taken to mean that there is a significant *linear* association between the two variables.

Example: Comparing Two Lung Function Meters

$r = 0.943$

$n = 17$

$p < 0.001$

Relationship Between Correlation and Regression

In simple linear regression, one measure of how well a line fits the data was the R^2:

$$R^2 = \frac{SS_{Model}}{SS_{Total}} = \frac{\text{Variability explained by the line}}{\text{Total variability}}$$

This measure of fit is equal to the square of the correlation coefficient ($R^2 = r^2$).

The slope coefficient is equivalent to the correlation times the ratio of the standard deviation of *Y* to the standard deviation of *X*:

$$\hat{\beta}_1 = r\left(\frac{\hat{\sigma}_Y}{\hat{\sigma}_X}\right)$$

Testing if the slope coefficient is equal to 0 is equivalent to testing if the correlation coefficient is equal to 0.

r is positive, negative, or zero as $\hat{\beta}_1$ is positive, negative or zero; and vice versa.

Uses and Misuses of Correlation (Altman, *Practical Statistics for Medical Research*, 1991)

Spurious correlations involving time

- http://www.tylervigen.com/spurious-correlations
- When both variables have time trends, this will induce a (potentially spurious) correlation between them.

Restricted sampling of individuals

- Between-subject variability makes a direct impact on the calculation of the correlation coefficient.

Mixed samples

- The presence of subgroups with different associations may result in an overall correlation that is not representative of the correlation within the subgroups.

Assessing agreement

- Correlation measures association, but correlation does NOT imply causation.

Change related to initial value

- Regression to the mean

Relating a part to a whole

- X and X + Y will always appear to be correlated.

C. Multiple Linear Regression

Multiple linear regression can be used to summarize the relationship between a continuous response variable, Y, and multiple explanatory variables, X_1, X_2,…, X_k, using linear relationships.

There are a variety of reasons we may want to include additional predictors in the model, including:

- To address the scientific question
- To adjust for confounding
- To gain precision

Address the scientific question. The scientific question may dictate inclusion of predictors:

- Predictor(s) of interest: The scientific factor under investigation may need to be modeled by multiple predictors (e.g., dummy variables, polynomials). Or there may be more than one predictor of interest.
- Effect modifiers: The scientific question may relate to the detection of effect modification.
- **Confounders**: The scientific question may have been stated in terms of identifying known or suspected confounders (or mediators).

Precision. Adjusting for an additional covariate changes the standard error of the slope estimate:

- The standard error is decreased by having smaller within group variance.
- The standard error is increased by having correlations between the predictor of interest and other covariates in the model.

As in simple linear regression, we assume that $Y_i|X_i \sim N(\mu_{Y|X}, \sigma^2_{Y|X})$, but now we assume that the underlying center changes linearly with several other factors, X_1,…, X_k; that is,

$$\mu_{Y|X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Equivalently, we can assume that

$$Y_i|X_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i$$

where ϵ_i represents the random error and $\epsilon_i \sim N\left(0, \sigma^2_{Y|X}\right)$

Interpretation of Coefficients

Intercept: β_0 is the expected value of *Y* **when all X_1,…, X_k are equal to zero**.

Slope: β_j is the expected change in *Y* associated with a one-unit change in X_j **when values for all other independent variables X_i ($i \neq j$) are held constant**.

Least Squares Estimation for Multiple Linear Regression

As in simple linear regression, we seek to use the data (Y_i, X_{1i}, X_{2i},…, X_{ki}; i = 1,…, n) to estimate the parameters (β_0, β_1, β_2,…, β_k).

The model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k$$

represents the fitted values.

Recall, the difference between the fitted values and the observed values are called the ***residuals*** (our observed estimate) [as compared to ***errors*** (unobserved difference from *true* value)]:

$$\hat{e}_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k)$$

The coefficients are chosen to minimize the sum of the squared errors (i.e., the residual sum of squares or error sum of squares).

The error sum of squares measures the “left-over” variability in the response after accounting for the variability explained by X_1, X_2,…, X_k.

Multiple Regression Assumptions

The assumptions for multiple linear regression are the same as for simple linear regression:

- **Existence**: For each specific combination of values of the independent variables X_1, X_2,…, X_k, Y is a random variable with a certain probability distribution having finite mean and variance.

- **Linearity**: The mean value of Y for each specific combination of X_1, X_2,…, X_k is a linear function of X_1, X_2,…, X_k.

- **Independence**: The Y observations are statistically independent of one another.

- **Homoscedasticity**: The variance of Y ($\sigma^2_{Y|X_1,X_2,\ldots,X_k}$) is the same for any fixed combination of X_1, X_2,…, X_k.

- **Normality**: For any fixed combination of X_1, X_2,…, X_k, the residuals are normally distributed. (This assumption is primarily used for hypothesis testing and CIs.)

Example: FEV

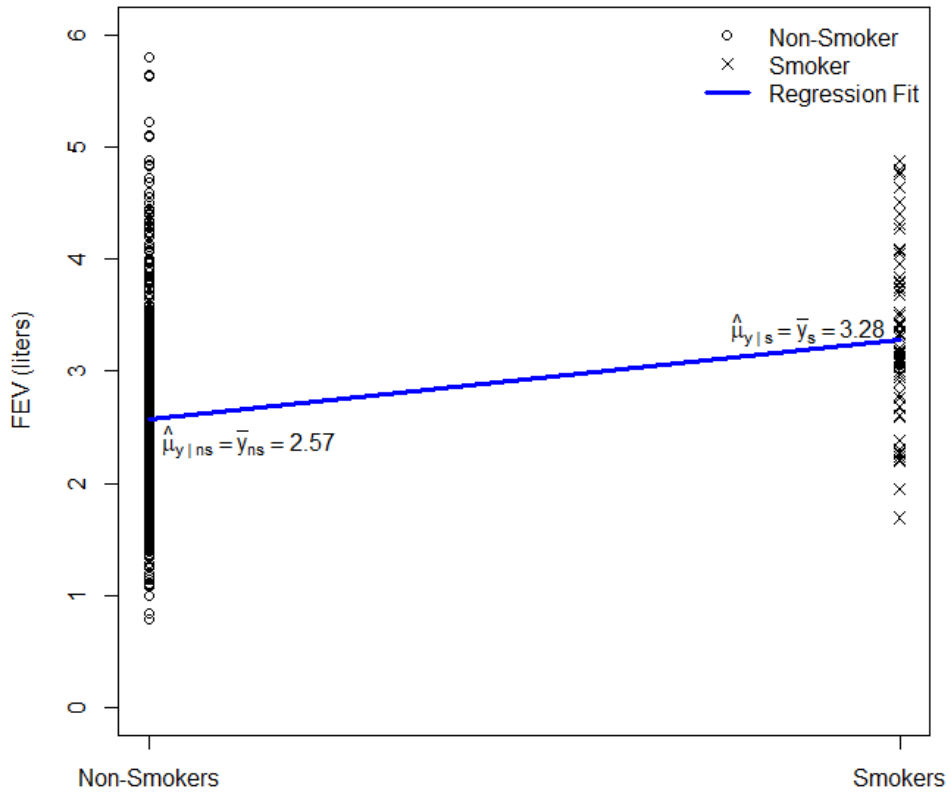
```
PROC REG DATA=fev;
    MODEL fev = csmoke;   /*0=Non-smoker 1=Smoker */
RUN;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	29.56968	29.56968	41.79	<.0001
Error	652	461.35015	0.70759		
Corrected Total	653	490.91984			

Root MSE	0.84119	**R-Square**	0.0602
Dependent Mean	2.63678	**Adj R-Sq**	0.0588
Coeff Var	31.90198		

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	
Intercept	**1**	2.56614	0.03466	74.04	<.0001	$\hat{\mu}_{Y\|ns} = 2.56614$
csmoke	**1**	0.71072	0.10994	6.46	<.0001	$\hat{\mu}_{Y\|s} = 2.56614 + 0.71072 = 3.27686$

Example: FEV



Least-squares regression line: $\hat{Y} = 2.57 + 0.71 \times smoke,$

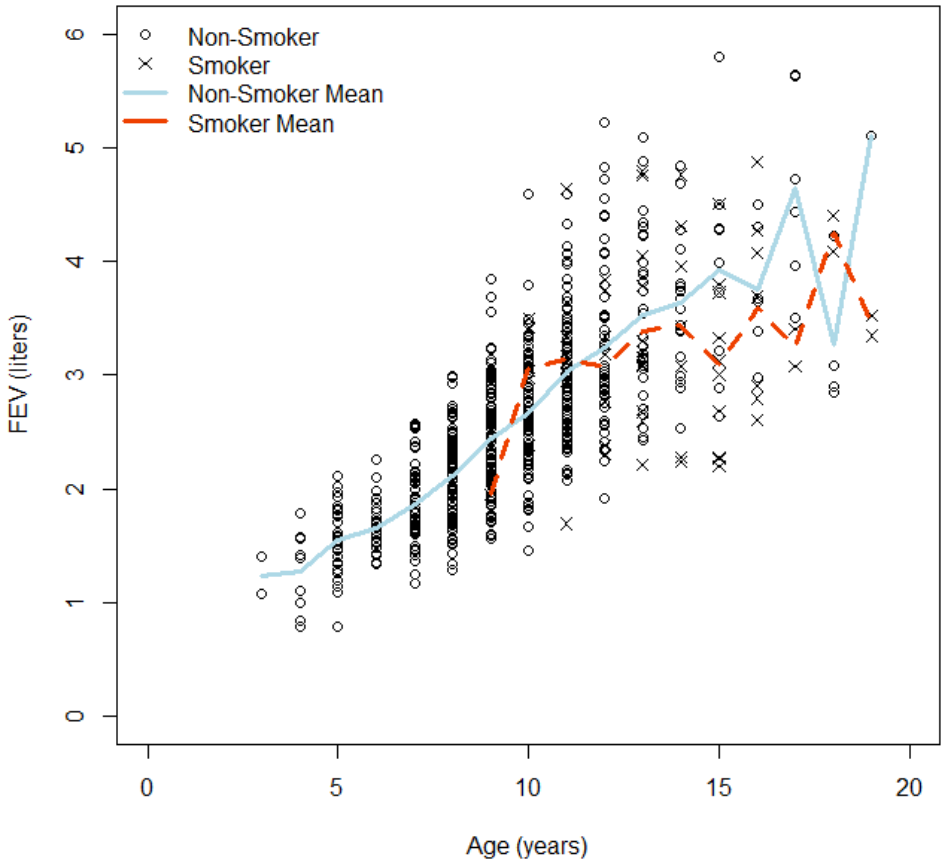
where smoke = 0 for non-smokers and smoke = 1 for smokers.

Interpretation: Smokers have better lung function than non-smokers ($p < 0.0001$).
(Does this make clinical sense? Smokers tend to be older than non-smokers and older children have higher FEV than young children.)

***New question*:** For a group of children at a specific age, do smokers have lower FEV compared to non-smokers?

- Option 1: We can perform a stratified analysis and compare smokers to non-smokers within age strata (see below).

- Option 2: With multiple linear regression, we can get a single estimate of the average effect of smoking on FEV, adjusting for differences due to age.



Stratified Analysis:

Age Group	Smokers	Non-smokers	FEV Smokers	FEV Non-smokers	Smoke-NonSmoke difference	T statistic	p-value
3-8	0	215	-	1.86 (0.42)	-	-	-
9-10	6	169	2.88 (0.60)	2.54 (0.51)	0.34	-1.57	.118
11-12	16	131	3.11 (0.67)	3.11 (0.64)	0.00	0.01	.993
13-14	20	48	3.40 (0.83)	3.57 (0.68)	-0.17	0.87	.389
15-16	17	15	3.30 (0.82)	3.85 (0.81)	-0.55	1.92	.065
17-19	6	11	3.64 (0.50)	4.19 (1.03)	-0.55	1.21	.244

Multiple Linear Regression Analysis

```
PROC REG data=fev;
    MODEL fev = csmoke age; /* Fit model with both age and csmoke */
RUN;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	283.05825	141.52913	443.25	<.0001
Error	651	207.86159	0.31930		
Corrected Total	653	490.91984			

Root MSE	0.56506	**R-Square**	0.5766
Dependent Mean	2.63678	**Adj R-Sq**	0.5753
Coeff Var	21.43003		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	0.36737	0.08144	4.51	<.0001
csmoke	**1**	-0.20899	0.08075	-2.59	0.0099
age	**1**	0.23060	0.00818	28.18	<.0001

Interpretation (not complete):

Smokers have worse lung function compared to non-smokers of the same age (p = 0.0099).

Example: FEV

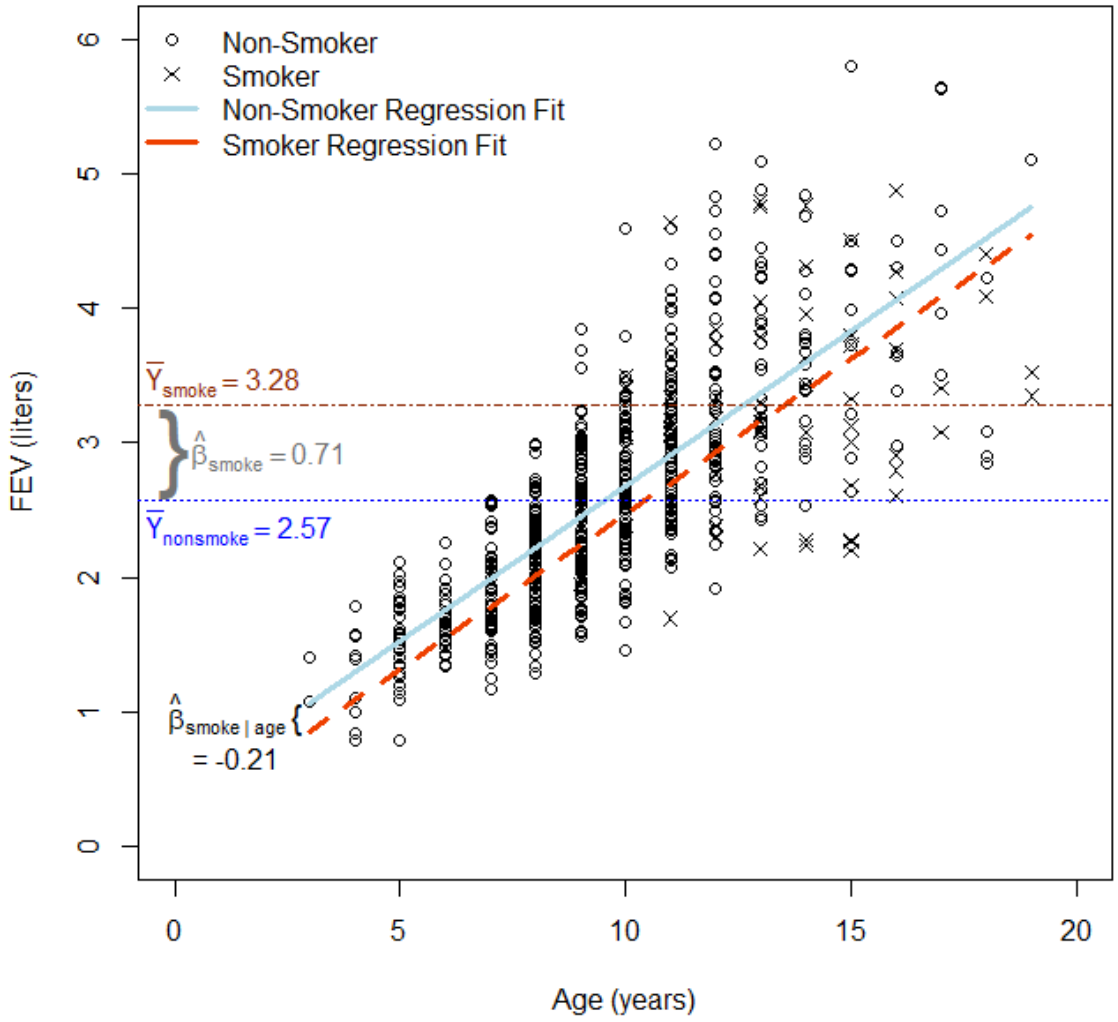
Least-squares regression line:

$\hat{Y} = 0.37 + 0.23 \times age - 0.21 \times smoke$

SLR without adjusting for age:

$\hat{Y} = 2.57 + 0.71 \times smoke$

MLR Interpretation: After adjusting for differences in age, smokers have poorer lung function than non-smokers. On average, FEV is 0.21 liters *lower* in smokers compared to non-smokers *of the same age*.



- Note that the effect of age on FEV is assumed to be the same for smokers and non-smokers.
- Note that the effect of smoking on FEV is assumed to be the same for every age.
- We can relax these restrictions by including interaction terms in the model (*more about this in Lecture 22*).

D. Multiple Linear Regression: Additional Diagnostics Beyond SLR

Partial regression plots (or ***partial plots, added variable plots, adjusted variable plots***) characterize the relationship between the dependent variable *Y* and an independent variable *X*, adjusting for the other covariates in the multiple regression model C_1, C_2, …, C_k.

The following steps can be followed to produce a partial plot:

(1) Perform a multiple regression of *Y* on the covariates C_1, C_2, …, C_k and save the residuals.

(2) Perform a multiple regression of *X* on the covariates C_1, C_2, …, C_k and save the residuals.

(3) The partial plot is the scatterplot of the residuals from step (1) against the residuals from step (2).

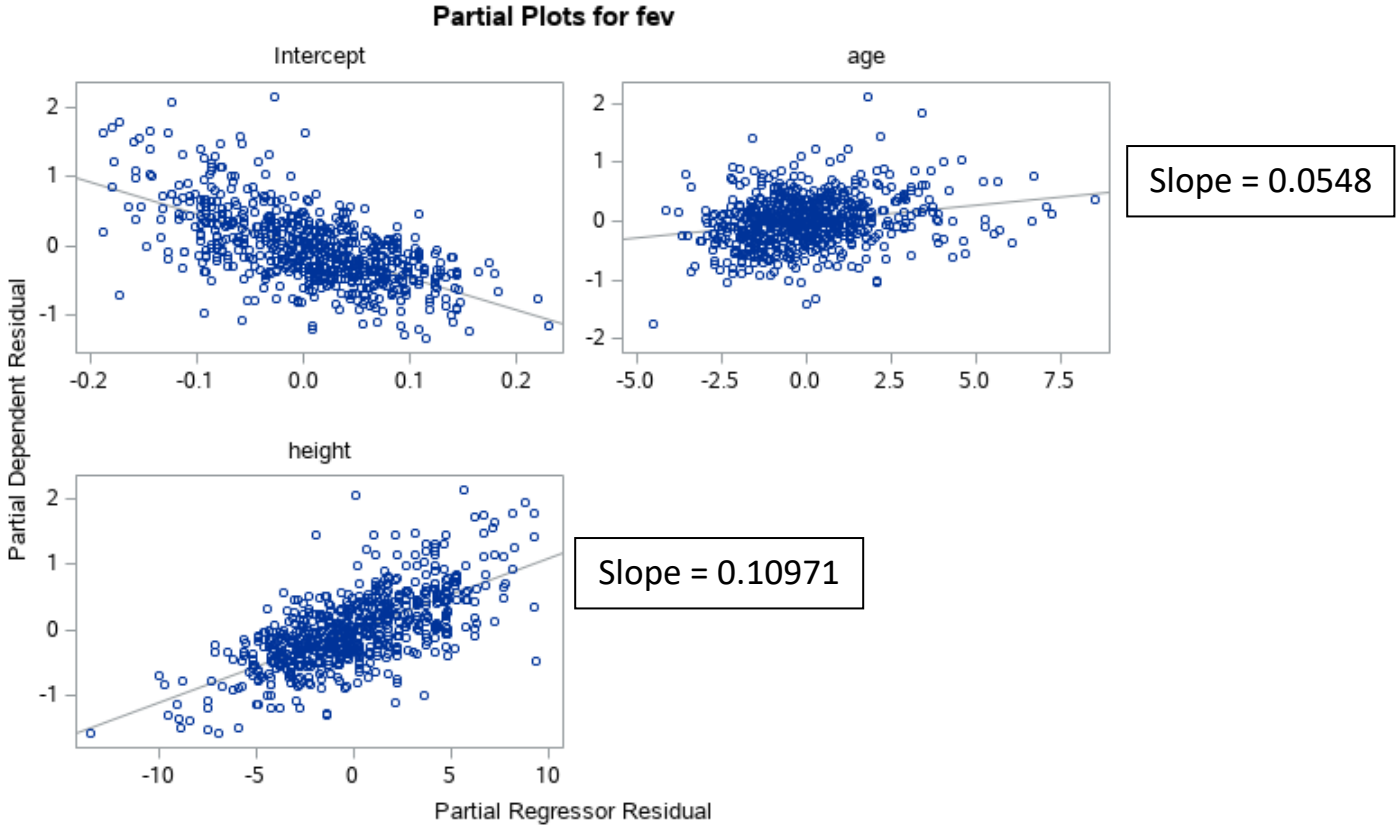
The slope in the partial plot will be the same as the slope for *X* in the multiple regression model of *Y* on *X*, C_1, C_2, …, C_k.

The simple linear regression of the residuals from step (1) on the residuals from step (2), will give you the same beta coefficient as the beta coefficient for *X* ($\hat{\beta}_X$) in the multiple linear regression model regression *Y* on *X*, C_1, C_2, …, C_k.

Example: FEV and age, adjusted for height

```
PROC REG DATA=fev;
    MODEL fev = age height / PARTIAL;
RUN;
```

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-4.61047	0.22427	-20.56	<.0001
age	1	**0.05428**	0.00911	5.96	<.0001
height	1	**0.10971**	0.00472	23.26	<.0001



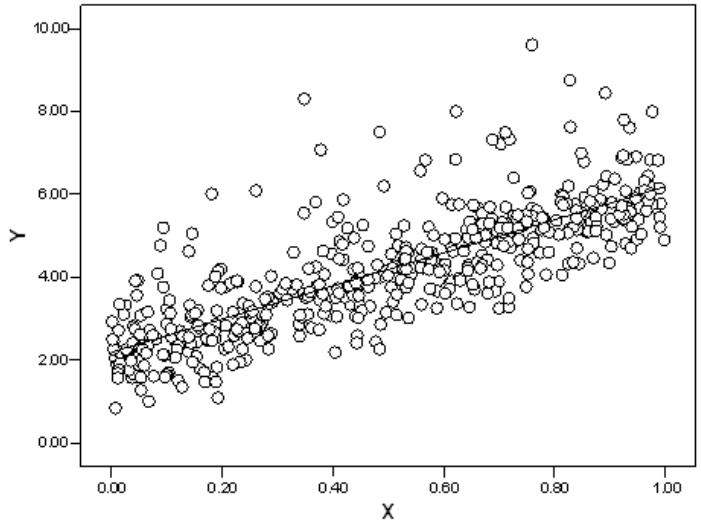
Omitted Covariates

Transformation of the response or predictor variable(s) is not always the best solution when violations of the assumptions are detected.

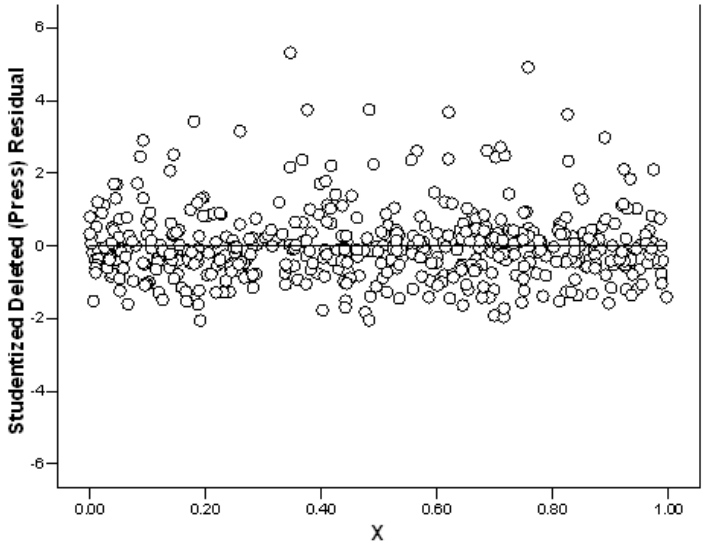
Sometimes the violation can be corrected by the addition of an additional covariate to the model.

Example: Non-normality of the errors can sometimes occur when an important predictor variable is excluded from the regression model. Assume we only include X and Y, we have the following plots:

Scatterplot



Residual Plot

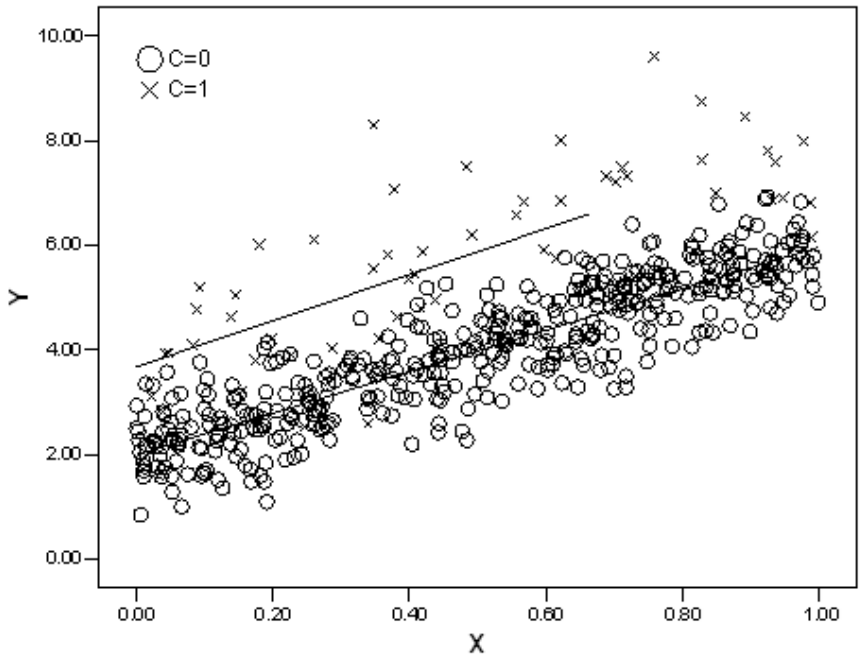


Non-Normal Errors Example: Simple Regression Model (NOT Adjusting for C)

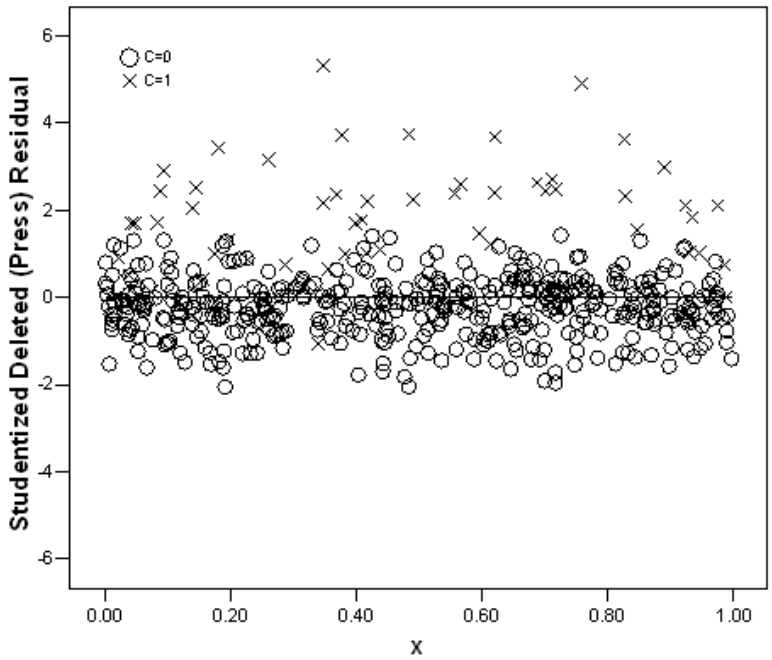
Assume we know the omitted covariate (C) and fit the same figures, but with different points for the two levels of the covariate:

$$E[Y] = \beta_o + \beta_x X$$

Scatterplot



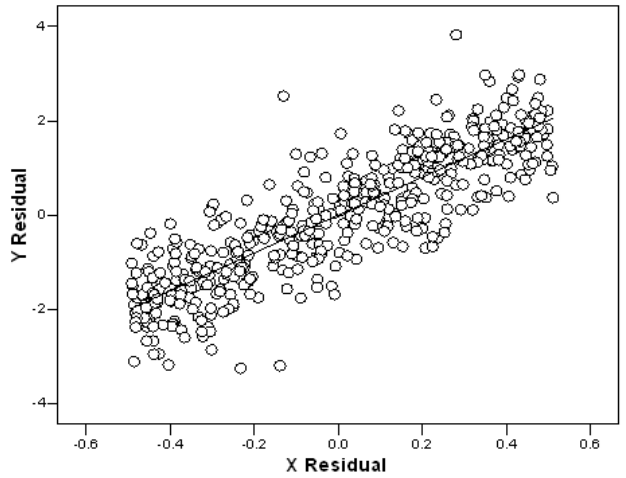
Residual Plot



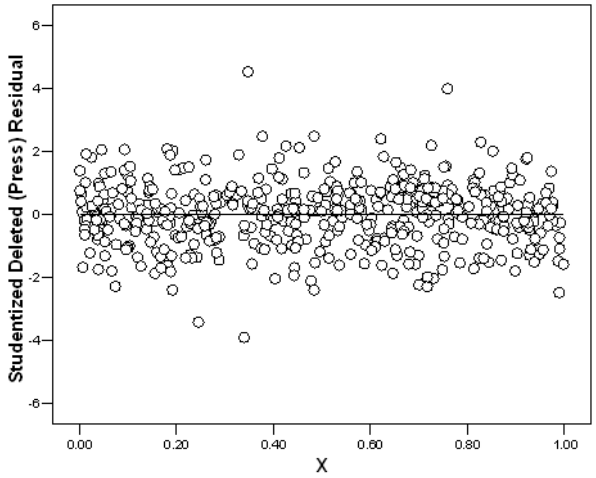
Non-Normal Errors Example: Multiple Regression Model (Adjusting for C)

Now assume we include the omitted covariate: $E[Y] = \beta_o + \beta_x X + \beta_c C$

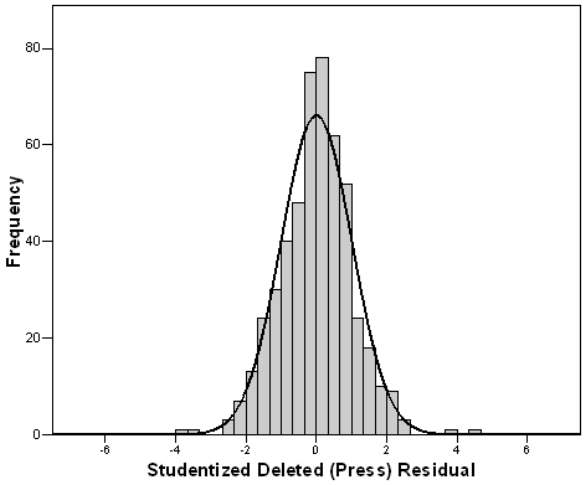
Partial Plot



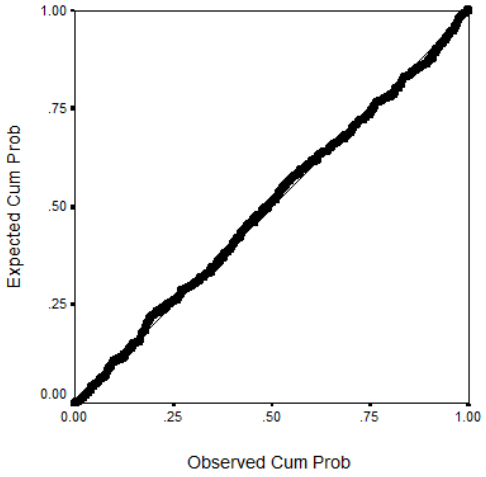
Residual Plot



Histogram



Normal Probability Plot



NOTE: Since there are two variables in this model (X and C) you should plot the residuals versus both X and C or plot the residuals versus y-hat. We omit here for space.

E. Multiple Linear Regression: Inference about the independent variable(s)

Once we have fit a multiple regression model, there are various types of tests we may wish to perform to make inferences about the underlying parameters:

- ***Overall test***: Taken collectively, does the *entire set* of independent variable contribute significantly to the prediction of *Y*? (***F* test**; Partial *F* test)

- ***Test for addition of a single variable***: Does the addition of *one* particular independent variable of interest add significantly to the prediction of *Y* over and above that achieved by other independent variables already present in the model? (Partial *F* test; ***t* test**)

- ***Test for addition of a group of variables***: Does the addition of some *group* of independent variables of interest add significantly to the prediction of *Y* over and above that achieved by other independent variables already present in the model? (**Partial *F* test**)

Comparing Full and Reduced Models

Hypothesis tests about the coefficients can be interpreted as a comparison between two models:

- the *full* (or *complete*) model
- the *reduced* model (some subset of the full model)

The "complete model" simplifies to the "reduced model" if the null hypothesis is true.

Simple Linear Regression (SLR): Overall test

H_0: $\beta_1 = 0$
H_A: $\beta_1 \neq 0$

Full model: $E[Y] = \beta_0 + \beta_1 X_1$
Reduced model: $E[Y] = \beta_0$

Multiple Linear Regression (MLR): Overall test

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
H_A: At least one of the $\beta_k \neq 0$

Full model: $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$
Reduced model: $E[Y] = \beta_0$

MLR: Test for adding a single variable

H_0: $\beta_1^* = 0$

H_A: $\beta_1^* \neq 0$

Full model: $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \beta_1^* X_1^*$

Reduced model: $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$

MLR: Test for adding a group of variables

H_0: $\beta_1^* = \beta_2^* = \ldots = \beta_k^* = 0$

H_A: At least one of the $\beta_k^* \neq 0$

Full model: $E[Y] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \beta_1^* X_1^* + \ldots + \beta_k^* X_k^*$

Reduced model: $E[Y] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$

Overall *F* Test in Multiple Linear Regression

We previously used the *F* test for the one-way ANOVA and simple linear regression.

The *F* statistic is the ratio of the regression mean square (the variability explained by our regression model) to the residual mean square (the variability remaining after fitting our regression line):

$$F = \frac{MS_{Model}}{MS_{Error}}$$

The *F* statistic can also be thought of as a ratio of two independent estimates of variance:

$$F = \frac{\hat{\sigma}_0^2}{\hat{\sigma}_{Y|X}^2}$$

- The term $\hat{\sigma}_0^2$ estimates $\sigma_{Y|X}^2$ if the null hypothesis is true and estimates some quantity larger than $\sigma_{Y|X}^2$ when the null hypothesis is not true.

- The term $\hat{\sigma}_{Y|X}^2$ estimates $\sigma_{Y|X}^2$ whether the null hypothesis is true or not.

Partial *F* Test

The Partial *F* test assesses whether the addition of any specific independent variable(s), given others already in the model, significantly contributes to the prediction of *Y*.

- Does the addition of any *k* independent variables X_1^*, X_2^*,…, X_k^* to the model (the full model) cause a significant reduction in the error sum of squares when compared to a model excluding these variables (the reduced model).

The *F* statistic for our Partial *F* test can be calculated as:

$$F = \frac{[SS_{\text{model}}(full) - SS_{\text{model}}(reduced)]/k}{MS_{error}(full)} \sim F_{k,n-p-k-1}$$

where

- *k* = number of regression coefficients set to 0 for H_0
- *p* = number of regression coefficients in the reduced model
- *k+p* = number of regression coefficients in the full model

Under the null hypothesis, this *F* statistic has an *F* distribution with *k* and $n - p - k - 1$ degrees of freedom.

Note: Some texts, such as KKNM, refer to the Partial *F* test for the addition of a group of variables as the Multiple-Partial *F* test.

Example: Overall F Test for FEV Data

Overall test: Taken collectively, do age and smoking status contribute significantly to the prediction of FEV?

H_0: $\beta_{csmoke} = \beta_{age} = 0$
H_A: At least one of the β's $\neq 0$

Full model: $E[Y] = \beta_0 + \beta_{csmoke}X_{csmoke} + \beta_{age}X_{age}$
Reduced model: $E[Y] = \beta_0$

Full Model (Output from Slide 21)

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	**283.05825**	141.52913	443.25	<.0001
Error	651	207.86159	**0.31930**		
Corrected Total	653	490.91984			

Root MSE	0.56506	**R-Square**	0.5766
Dependent Mean	2.63678	**Adj R-Sq**	0.5753
Coeff Var	21.43003		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	0.36737	0.08144	4.51	<.0001
csmoke	**1**	-0.20899	0.08075	-2.59	0.0099
age	**1**	0.23060	0.00818	28.18	<.0001

Reduced Model:

```
PROC REG data=fev;
    MODEL fev = ;
RUN;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	0	**0**	.	.	.
Error	653	490.91984	0.75179		
Corrected Total	653	490.91984			

Root MSE	0.86706	**R-Square**	0.0000
Dependent Mean	2.63678	**Adj R-Sq**	0.0000
Coeff Var	32.88326		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	2.63678	0.03390	77.77	<.0001

Overall *F* Test (using Partial *F* Test):

$$F = \frac{[SS_{\text{model}}(full) - SS_{\text{model}}(reduced)]/k}{MS_{error}(full)} = \frac{(283.05825-0)/2}{0.31930} = \frac{MS_{Model}(Full)}{MS_{Error}(Full)} = 443.25 \sim F_{2,651}$$

Note: This does not necessarily mean *both* smoking and age are significant predictors. Perhaps a more parsimonious model is adequate (i.e. only one variable is necessary).

Example: Partial F Test for FEV Data

Test for addition of a single variable: Does the addition of *smoking status* add significantly to the prediction of *FEV* over and above that achieved by age?

H_0: $\beta_{csmoke} = 0 \mid \beta_{age}$ in model
H_A: $\beta_{csmoke} \neq 0 \mid \beta_{age}$ in model

Full model: $E[Y] = \beta_0 + \beta_{csmoke}X_{csmoke} + \beta_{age}X_{age}$
Reduced model: $E[Y] = \beta_0 + \beta_{age}X_{age}$

Partial F test (using output on next slide):

$$F = \frac{[SS_{\text{model}}(full) - SS_{\text{model}}(reduced)]/k}{MS_{error}(full)} = \frac{(283.05825\text{-}280.91916)/1}{0.31930} = 6.6993 \sim F_{1,654-1-1-1} = F_{1,651}, p = 0.0099$$

How else could we have tested H_0: $\beta_{csmoke} = 0$?

$$t = \frac{\beta_{csmoke}}{SE(\beta_{csmoke})} = \frac{-0.20899}{0.08075} = \text{-}2.59 \sim t_{651}, p = 0.0099$$

Full Model (Output from Slide 21):

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	**283.05825**	141.52913	443.25	<.0001
Error	651	207.86159	**0.31930**		
Corrected Total	653	490.91984			

Reduced Model:

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	**280.91916**	280.91916	872.18	<.0001
Error	652	210.00068	0.32209		
Corrected Total	653	490.91984			

Root MSE	0.56753	**R-Square**	0.5722
Dependent Mean	2.63678	**Adj R-Sq**	0.5716
Coeff Var	21.52349		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	0.43165	0.07790	5.54	<.0001
age	**1**	0.22204	0.00752	29.53	<.0001

```
PROC REG;
 MODEL fev = age;
RUN;
```

Test for addition of a group of variables: Does the addition of height and sex add significantly to the prediction of *FEV* over and above that achieved by age and smoking status?

H_0: $\beta_{height} = \beta_{sex} = 0$ Full model: $E[Y] = \beta_0 + \beta_{csmoke}X_{csmoke} + \beta_{age}X_{age} + \beta_{height}X_{height} + \beta_{sex}X_{sex}$

H_A: β_{height} and/or $\beta_{sex} \neq 0$ Reduced model: $E[Y] = \beta_0 + \beta_{csmoke}X_{csmoke} + \beta_{age}X_{age}$

Full Model:

```
PROC REG;
    MODEL fev = csmoke age height male;
    Sex_Height: TEST male, height; /* Specify partial F test */
RUN;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	**380.64028**	95.16007	560.02	<.0001
Error	649	110.27955	**0.16992**		
Corrected Total	653	490.91984			

Root MSE	0.41222	**R-Square**	0.7754
Dependent Mean	2.63678	**Adj R-Sq**	0.7740
Coeff Var	15.63332		

Full Model continued:

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	-4.45697	0.22284	-20.00	<.0001
csmoke	**1**	-0.08725	0.05925	-1.47	0.1414
age	**1**	0.06551	0.00949	6.90	<.0001
height	**1**	0.10420	0.00476	21.90	<.0001
male	**1**	0.15710	0.03321	4.73	<.0001

Reduced Model (Output from Slide 21):

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	**283.05825**	141.52913	443.25	<.0001
Error	651	207.86159	0.31930		
Corrected Total	653	490.91984			

Partial *F* test:

$$F = \frac{[SS_{\text{model}}(full) - SS_{\text{model}}(reduced)]/k}{MS_{error}(full)} = \frac{(380.64028\text{-}283.05825)/2}{0.16992} = 287.14 \sim F_{2,654-2-2-1} = F_{2,649}, p < 0.0001$$

Partial *F* test in SAS:

Test Sex_Height Results for Dependent Variable fev				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	2	48.79102	**287.14**	<.0001
Denominator	649	0.16992		

t test for Individual Coefficients

The *t* test can be used to assess whether the addition of *one* particular independent variable of interest adds significantly to the prediction of *Y* over and above that achieved by other independent variables already present in the model.

The *t* statistic can be calculated as: $t = \frac{\widehat{\beta}^*}{SE(\widehat{\beta}^*)}$

Under the null hypothesis, this *t* statistic has a *t*-distribution with $n - p - 1$ degrees of freedom, where *p* is the number of independent variables in the model (note that *p* does not include the intercept).

The *t* test will give the same result as the partial *F* test for the addition of a single variable.

Example: t-test for Individual Coefficients for FEV Data (Output from Slide 38)

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	-4.45697	0.22284	-20.00	<.0001
csmoke	**1**	-0.08725	0.05925	-1.47	0.1414
age	**1**	0.06551	0.00949	6.90	<.0001
height	**1**	0.10420	0.00476	21.90	<.0001
male	**1**	0.15710	0.03321	4.73	<.0001

Tests for addition of a single variable:

Does the addition of *smoking status* add significantly to the prediction of *FEV* over and above that achieved by age, sex, and height?

Does the addition of *age* add significantly to the prediction of *FEV* over and above that achieved by smoking status, sex, and height?

Does the addition of *sex* add significantly to the prediction of *FEV* over and above that achieved by smoking status, age, and height?

Does the addition of *height* add significantly to the prediction of *FEV* over and above that achieved by smoking status, age, and sex?

E. Quick Review

What is the most straightforward way to test the following?

1. H_0: $\beta_1^* = \beta_2^* = \ldots = \beta_k^* = 0$ **Partial F-test**
   H_A: At least one of the $\beta_k^* \neq 0$

   Full model: $E[Y] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \beta_1^* X_1^* + \ldots + \beta_k^* X_k^*$
   Reduced model: $E[Y] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$

2. H_0: $\beta_1^* = 0$ **t-test**
   H_A: $\beta_1^* \neq 0$

   Full model: $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \beta_1^* X_1^*$
   Reduced model: $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$

3. H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ **F-test**
   H_A: At least one of the $\beta_k \neq 0$

   Full model: $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$
   Reduced model: $E[Y] = \beta_0$

General SAS Output Guide

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	p	SS_{Model}	SS_{Model} / p	MS_{Model} / MS_{Error}	$P(F_{p,n-p-1} > F)$ $H_0: \beta_1 = \beta_2 = 0$
Error	n-p-1	SS_{Error}	SS_{Error} / (n-p-1)		
Corrected Total	n-1				

Root MSE	XXXX	**R-Square**	SS_{Model}/SS_{Total}
Dependent Mean	XXXX	**Adj R-Sq**	$1-(1-R^2)\left(\frac{n-1}{n-p-1}\right)$
Coeff Var	XXXX		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	**1**	$\hat{\beta}_0$	$SE(\hat{\beta}_0)$	$\hat{\beta}_0 / SE(\hat{\beta}_0)$	$P(t_{n-p-1} > t)$ $H_0: \beta_0 = 0$
Covariate1	Cov1	**1**	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$\hat{\beta}_1 / SE(\hat{\beta}_1)$	$P(t_{n-p-1} > t)$ $H_0: \beta_1 = 0$
Covariate2	Cov2	**1**	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$\hat{\beta}_2 / SE(\hat{\beta}_2)$	$P(t_{n-p-1} > t)$ $H_0: \beta_2 = 0$

The 1 df does not relate to the t-statistic which has n-p-1 degrees of freedom.