

BIOS 600: Principles of Statistical Inference

Linear Regression

Fall 2012

Reading

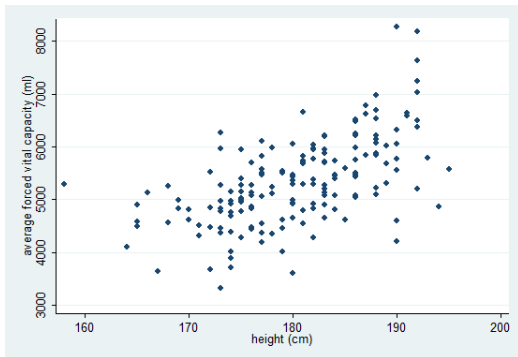
- ▶ Pagano and Gauvreau, Chapter 18
- ▶ Slope-intercept form video
- ▶ For more information on regression, take BIOS 545!

Goal

Regression is an important topic, and you are strongly encouraged to take a regression course. The goal of our coverage of regression in the next two lectures is to give you the skills you need to understand results in a subject-area journal that are obtained from fitting a linear regression model. The goal is not to give you all the skills you need to fit such models yourself on a regular basis (that requires a course!).

Example: EPA Study

The EPA's Health Effects Research Laboratory conducted a study of 172 adult males on campus. We examine the correlation between average forced vital capacity (ml) and height (cm).



How do we quantify the relationship between height and FVC?

Correlation and Regression

- ▶ *Correlation* measures the direction and strength of linear relationships

Correlation and Regression

- ▶ *Correlation* measures the direction and strength of linear relationships
- ▶ *Regression* can be used to further quantify these relationships

Correlation and Regression

- ▶ *Correlation* measures the direction and strength of linear relationships
- ▶ *Regression* can be used to further quantify these relationships
- ▶ A *regression line* summarizes the relationship between explanatory or predictor variables (e.g., height, X) and response or outcome variables (e.g., FVC or Y).

Correlation and Regression

- ▶ *Correlation* measures the direction and strength of linear relationships
- ▶ *Regression* can be used to further quantify these relationships
- ▶ A *regression line* summarizes the relationship between explanatory or predictor variables (e.g., height, X) and response or outcome variables (e.g., FVC or Y).
- ▶ The fitted regression line can be used to predict the outcome for a given set of predictor values.

Correlation and Regression

- ▶ *Correlation* measures the direction and strength of linear relationships
- ▶ *Regression* can be used to further quantify these relationships
- ▶ A *regression line* summarizes the relationship between explanatory or predictor variables (e.g., height, X) and response or outcome variables (e.g., FVC or Y).
- ▶ The fitted regression line can be used to predict the outcome for a given set of predictor values.
- ▶ Our predictions do have error, called *residuals*

Correlation and Regression

- ▶ *Correlation* measures the direction and strength of linear relationships
- ▶ *Regression* can be used to further quantify these relationships
- ▶ A *regression line* summarizes the relationship between explanatory or predictor variables (e.g., height, X) and response or outcome variables (e.g., FVC or Y).
- ▶ The fitted regression line can be used to predict the outcome for a given set of predictor values.
- ▶ Our predictions do have error, called *residuals*
- ▶ The *least-squares regression* line minimizes the squared error.

Assumptions of Regression Model

- ▶ Equal variances (homogeneity)

We will discuss later how to check some of these assumptions.

Assumptions of Regression Model

- ▶ Equal variances (homogeneity)
- ▶ Independent observations

We will discuss later how to check some of these assumptions.

Assumptions of Regression Model

- ▶ Equal variances (homogeneity)
- ▶ Independent observations
- ▶ Linear relationship between predictor and response (can be relaxed using polynomials or splines)

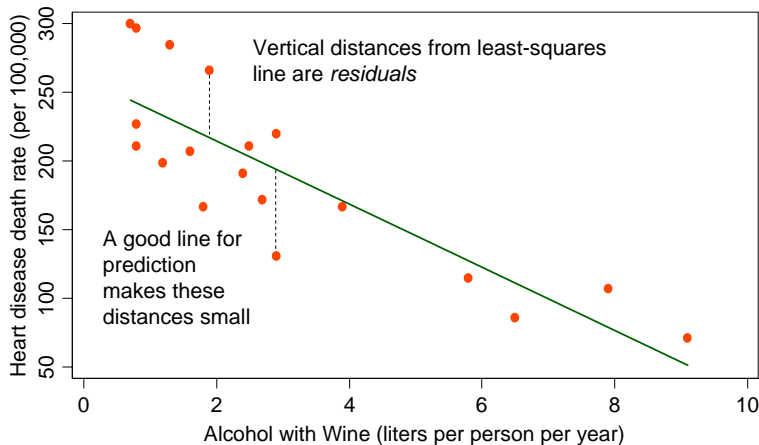
We will discuss later how to check some of these assumptions.

Assumptions of Regression Model

- ▶ Equal variances (homogeneity)
- ▶ Independent observations
- ▶ Linear relationship between predictor and response (can be relaxed using polynomials or splines)
- ▶ Conditional on covariates, the response follows a normal distribution (normal errors)

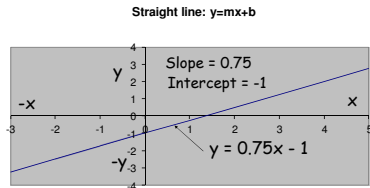
We will discuss later how to check some of these assumptions.

Example: Wine consumption and heart disease death rate



Review: Slope-intercept form of a line

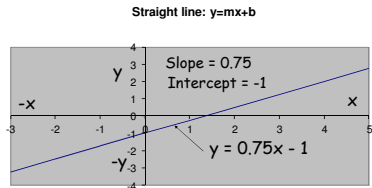
► Here, $m = 0.75$ and $b = -1$



Recall the
slope-intercept form of
a line, $y = mx + b$.

Review: Slope-intercept form of a line

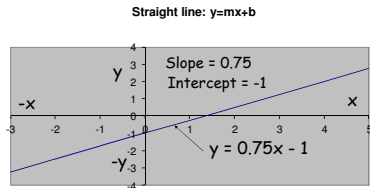
- ▶ Here, $m = 0.75$ and $b = -1$
- ▶ b is the y-intercept, or where the line crosses the y-axis. It is the predicted value of y when $x = 0$



Recall the slope-intercept form of a line, $y = mx + b$.

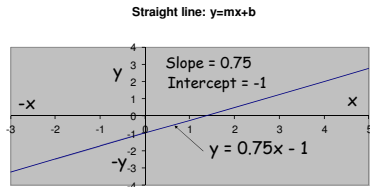
Review: Slope-intercept form of a line

- ▶ Here, $m = 0.75$ and $b = -1$
- ▶ b is the y-intercept, or where the line crosses the y-axis. It is the predicted value of y when $x = 0$
- ▶ m is the slope, which tells us the predicted increase in y when x changes by 1 unit



Recall the slope-intercept form of a line, $y = mx + b$.

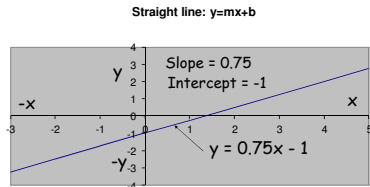
Review: Slope-intercept form of a line



- ▶ Here, $m = 0.75$ and $b = -1$
- ▶ b is the y-intercept, or where the line crosses the y-axis. It is the predicted value of y when $x = 0$
- ▶ m is the slope, which tells us the predicted increase in y when x changes by 1 unit
- ▶ What is our predicted y when $x = 2$?

Recall the slope-intercept form of a line, $y = mx + b$.

Review: Slope-intercept form of a line



Recall the slope-intercept form of a line, $y = mx + b$.

- ▶ Here, $m = 0.75$ and $b = -1$
- ▶ b is the y-intercept, or where the line crosses the y-axis. It is the predicted value of y when $x = 0$
- ▶ m is the slope, which tells us the predicted increase in y when x changes by 1 unit
- ▶ What is our predicted y when $x = 2$?
- ▶ In statistics, we often represent the slope with β_1 and the intercept with β_0 , and these values are usually not known but must be estimated.

Regression

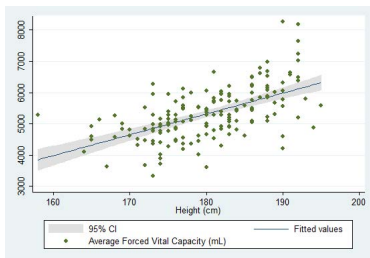
In regression, our we use one variable (or more) to try to predict values of another. In regression, one variable (Y) is the *response* or *outcome* or *dependent variable* and the other (X) is the *predictor* or *explanatory variable* or *independent variable*.

This distinction is critical. The regression of Y on X is not equal to the regression of X on Y .

The regression of Y on X can be used to predict Y based on fixed values of X .

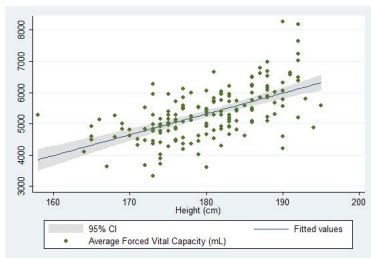
Example: EPA Study

- The regression line of FVC on height is shown



Here, Y is the FVC and X is the height.

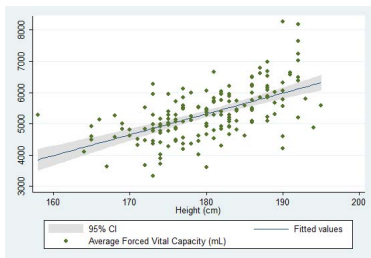
Example: EPA Study



- ▶ The regression line of FVC on height is shown
- ▶ Points represent actual data values

Here, Y is the FVC and X is the height.

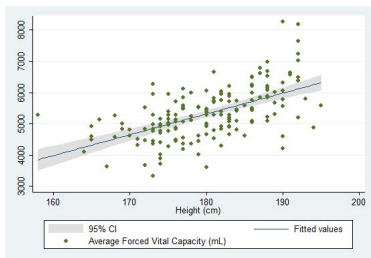
Example: EPA Study



- ▶ The regression line of FVC on height is shown
- ▶ Points represent actual data values
- ▶ Note the line is not a perfect fit

Here, Y is the FVC and X is the height.

Example: EPA Study



Here, Y is the FVC and X is the height.

- ▶ The regression line of FVC on height is shown
- ▶ Points represent actual data values
- ▶ Note the line is not a perfect fit
- ▶ r^2 tells us about how much of the variability in FVC is explained by height (33%); as r^2 gets closer to 1, the points get tighter around the line

The Model

The model is given by $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where

- ▶ y_i is the outcome (dependent variable) of interest

The Model

The model is given by $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where

- ▶ y_i is the outcome (dependent variable) of interest
- ▶ β_0 is the intercept parameter (P&G use α for this parameter, but β_0 is more typical notation)

The Model

The model is given by $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where

- ▶ y_i is the outcome (dependent variable) of interest
- ▶ β_0 is the intercept parameter (P&G use α for this parameter, but β_0 is more typical notation)
- ▶ β_1 is the slope parameter

The Model

The model is given by $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where

- ▶ y_i is the outcome (dependent variable) of interest
- ▶ β_0 is the intercept parameter (P&G use α for this parameter, but β_0 is more typical notation)
- ▶ β_1 is the slope parameter
- ▶ x_i is a predictor variable

The Model

The model is given by $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where

- ▶ y_i is the outcome (dependent variable) of interest
- ▶ β_0 is the intercept parameter (P&G use α for this parameter, but β_0 is more typical notation)
- ▶ β_1 is the slope parameter
- ▶ x_i is a predictor variable
- ▶ ε_i is the error (like β_0 and β_1 , it is not observed)

Model Assumptions

Assumptions of the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ include:

- ▶ The outcomes y_i are *independent*. This is violated if our study contains repeated outcome measures on an individual, if siblings are enrolled, etc.

Model Assumptions

Assumptions of the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ include:

- ▶ The outcomes y_i are *independent*. This is violated if our study contains repeated outcome measures on an individual, if siblings are enrolled, etc.
- ▶ For a specified value of x , which is measured without error, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. We often call the mean $\beta_0 + \beta_1 x_i = \mu_i$ and note that this implies $\varepsilon_i \sim N(0, \sigma^2)$.

Model Assumptions

Assumptions of the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ include:

- ▶ The outcomes y_i are *independent*. This is violated if our study contains repeated outcome measures on an individual, if siblings are enrolled, etc.
- ▶ For a specified value of x , which is measured without error, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. We often call the mean $\beta_0 + \beta_1 x_i = \mu_i$ and note that this implies $\varepsilon_i \sim N(0, \sigma^2)$.
- ▶ Straight line relationship holds (can relax this to some extent using polynomial regression or other methods)

Model Assumptions

Assumptions of the model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ include:

- ▶ The outcomes y_i are *independent*. This is violated if our study contains repeated outcome measures on an individual, if siblings are enrolled, etc.
- ▶ For a specified value of x , which is measured without error, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. We often call the mean $\beta_0 + \beta_1 x_i = \mu_i$ and note that this implies $\varepsilon_i \sim N(0, \sigma^2)$.
- ▶ Straight line relationship holds (can relax this to some extent using polynomial regression or other methods)
- ▶ The variance σ^2 is constant across all values of x (analogous to the equal variances assumption in the t-test); this is called homogeneity of variance

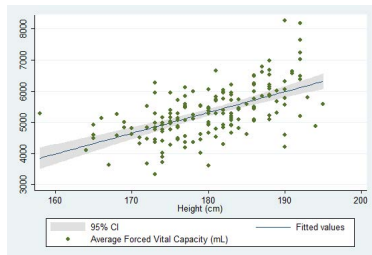
Least Squares

- Fitted line equation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Dots are ordered pairs (x_i, y_i)

Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$



Least Squares

- Fitted line equation:

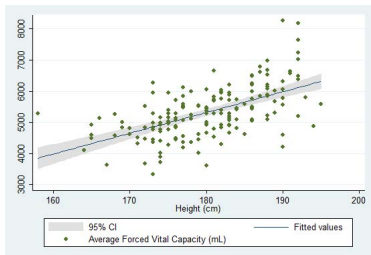
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Residual

$$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i = y_i - \hat{y}_i$$

Dots are ordered pairs (x_i, y_i)

Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$



Least Squares

- Fitted line equation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Residual

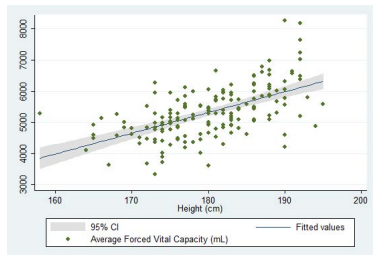
$$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i = y_i - \hat{y}_i$$

- Least squares selects $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the error *sum of squares*

$\sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ is minimized

Dots are ordered pairs (x_i, y_i)

Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$



Least Squares

- ▶ Fitted line equation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- ▶ Residual

$$\hat{\varepsilon}_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i = y_i - \hat{y}_i$$

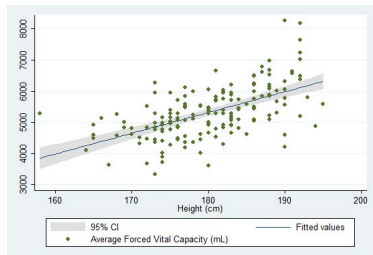
- ▶ Least squares selects $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the error *sum of squares*

$$\sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \text{ is minimized}$$

- ▶ Watch how it works with the regression applet

Dots are ordered pairs (x_i, y_i)

Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$



Least Squares Regression: Slope

The estimates that minimize the error sum of squares (*least squares estimates*) are $\hat{\beta}_1 = r_{xy} \frac{s_y}{s_x}$, where r_{xy} is the estimated correlation between x and y , s_y is the sd of y , and s_x is the sd of x . So the estimated slope is directly related to the correlation coefficient, scaled by the ratio of sd's of y and x . If y and x both have estimated variance 1 (and thus $\text{sd}=1$), the $\beta_1 = r_{xy}$.

Least Squares Regression: Intercept

The least squares estimate of the intercept is $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, where \bar{y} and \bar{x} are the sample means of y and x . Note that when $\bar{x} = 0$, then the estimated intercept is just the sample mean of y . A covariate x is sometimes centered at its mean (or some other meaningful value) in order to ensure β_0 has a meaningful interpretation, though this is not required.

Regression of FVC on Height in EPA Data

```
. regress avgfvc height
```

Source	SS	df	MS
Model	38448970.8	1	38448970.8
Residual	76856871.5	170	452099.244
Total	115305842	171	674303.172

```
Number of obs = 172
F( 1, 170) = 85.05
Prob > F = 0.0000
R-squared = 0.3335
Adj R-squared = 0.3295
Root MSE = 672.38
```

avgfvc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
height	66.60359	7.222244	9.22	0.000	52.34676 80.86042
_cons	-6673.894	1303.322	-5.12	0.000	-9246.674 -4101.114

The t-tests are for the hypothesis $H_0 : \beta_1 = 0$ and $H_0 : \beta_0 = 0$. Note also that we get 95% CI's for each parameter and indeed for each fitted value on the line (look at [another regression applet](#) for an illustration). To get plot:

```
Stata graph twoway lfitci avgfvc height || scatter
avgfvc height
```

Predicted Means

The regression equation can be used to get predicted means at any value of x . $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. For the FVC data, we have

$$\hat{y}_i = -6673.9 + 66.6x_i.$$

So the predicted mean FVC for students who are 6 feet tall (182.88cm) is

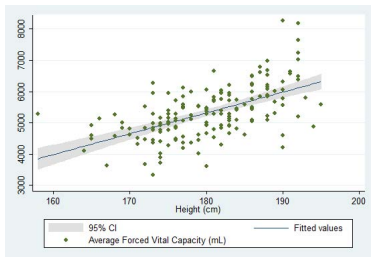
$$\hat{y} = -6673.9 + 66.6(182.88) = 5506ml.$$

Interpreting the Intercept

Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Least Squares Estimates:

$\hat{\beta}_0 = -6673.9$, $\hat{\beta}_1 = 66.6$



- Intercept β_0 is the expected value of y when $x = 0$

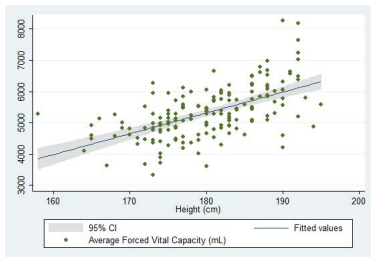
Interpreting the Intercept

Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Least Squares Estimates:

$\hat{\beta}_0 = -6673.9$, $\hat{\beta}_1 = 66.6$

- ▶ The intercept should not be directly interpreted when the range of x does not include 0! In EPA data, there are no men whose height is 0cm



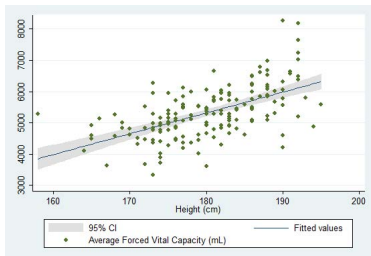
- ▶ Intercept β_0 is the expected value of y when $x = 0$

Interpreting the Intercept

Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Least Squares Estimates:

$$\hat{\beta}_0 = -6673.9, \hat{\beta}_1 = 66.6$$



- ▶ Intercept β_0 is the expected value of y when $x = 0$

- ▶ The intercept should not be directly interpreted when the range of x does not include 0! In EPA data, there are no men whose height is 0cm
- ▶ The predicted FVC for someone who is 0cm tall is nonsensical: -6674 (-9247, -4101). *Extrapolating the line beyond the range of the data almost always eventually leads to nonsense predictions. Confine inferences to points within the range of x*

Interpreting the Slope

The slope is the expected change in y corresponding to a one-unit change in x

Consider the following.

$$\blacktriangleright \hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x$$

Interpreting the Slope

The slope is the expected change in y corresponding to a one-unit change in x

Consider the following.

▶ $\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x$

▶ $\hat{y}_{x+1} = \hat{\beta}_0 + \hat{\beta}_1(x+1) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_1$

Interpreting the Slope

The slope is the expected change in y corresponding to a one-unit change in x

Consider the following.

- ▶ $\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x$
- ▶ $\hat{y}_{x+1} = \hat{\beta}_0 + \hat{\beta}_1(x+1) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_1$
- ▶ $\hat{y}_{x+1} - \hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_1 - \hat{\beta}_0 - \hat{\beta}_1 x = \hat{\beta}_1$

Interpreting the Slope

The slope is the expected change in y corresponding to a one-unit change in x

Consider the following.

- ▶ $\hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x$
- ▶ $\hat{y}_{x+1} = \hat{\beta}_0 + \hat{\beta}_1(x + 1) = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_1$
- ▶ $\hat{y}_{x+1} - \hat{y}_x = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_1 - \hat{\beta}_0 - \hat{\beta}_1 x = \hat{\beta}_1$
- ▶ FVC slope and corresponding 95% interval estimate: 66.6 (52.3, 80.9). For each 1 cm increase in height, we expect a 66.6 ml increase in FVC.

Hypothesis Test about the Slope

The primary hypothesis test of interest is usually $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$. The t statistic and p-value for this test are provided on the output, and we conclude that $\beta_1 \neq 0$. FVC increases with height: for each 1 cm increase in height, we expect a 66.6 ml increase in FVC.

The hypothesis test is constructed as

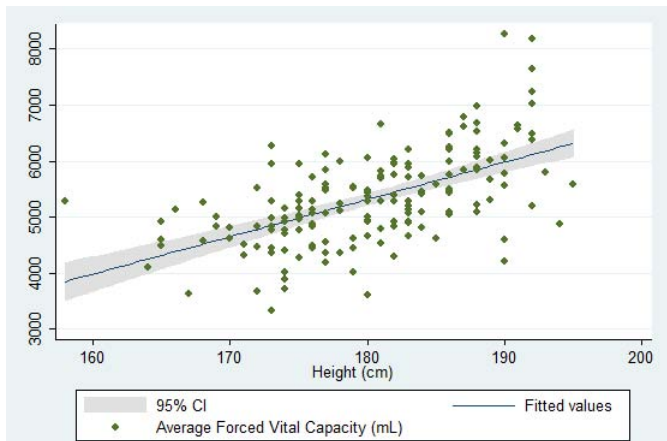
$$t = \frac{\hat{\beta} - \beta_{NULL}}{\hat{se}(\hat{\beta})},$$

where β_{NULL} is the hypothesized value of β (usually 0 but not always). Note P&G use the notation β_0 instead of β_{NULL} though we use β_0 for the intercept instead of a hypothesized value. The degrees of freedom are $n - 2$ (sample size minus total number of mean parameters in model).

Note that the hypothesis $H_0 : \beta_1 = 0$ is equivalent to the hypothesis $H_0 : \rho = 0$ where ρ is the correlation between x and y .

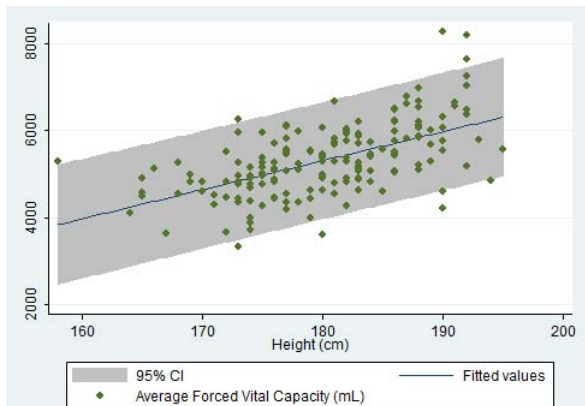
Predicted Means

The **predict** command can be used to get predicted means for values of x . The previous plot shows predicted means and a 95% confidence interval for the predicted regression line.



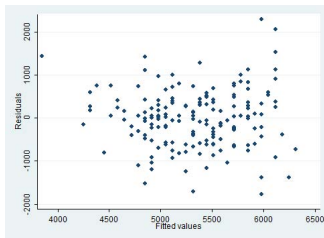
Predictions for New Individuals

We can also make predictions for new individuals not in our sample, but there is more variability associated with these individual predictions than for means. Code for this is `graph twoway lfitci avgfvc height, stdf || scatter avgfvc height`.



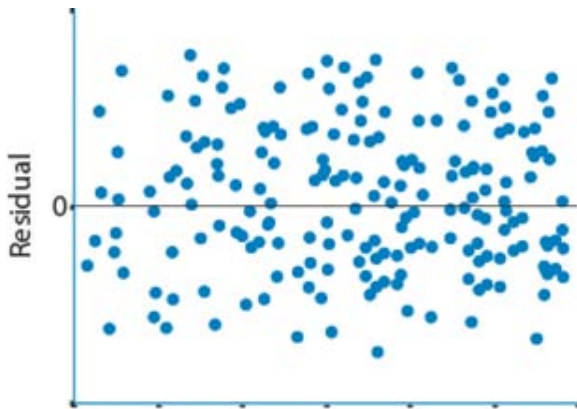
Regression Diagnostics: Equal Variances

We should always check assumptions of regression to see whether the method is valid in our setting. To check equal variances, we can use a plot of the residuals ($\hat{\epsilon}_i$) by the predicted (or fitted) values (\hat{y}_i), often called an “R by P” plot. To get this plot in Stata after fitting a regression model, just type **rvfplot**.

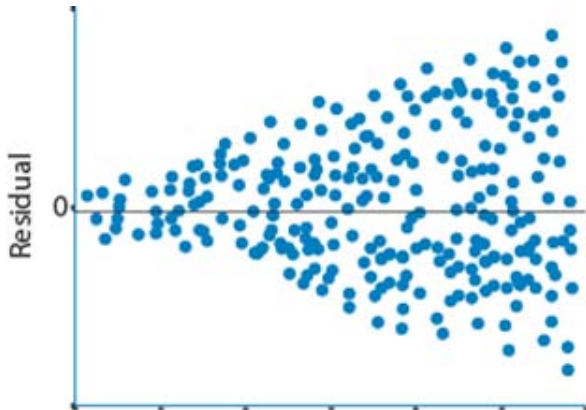


We expect to see evenly-spaced dots along the y axis. Patterns or trends are evidence something is wrong.

Examples of R by P Plots: Equal Variances



Examples of R by P Plots: Unequal Variances

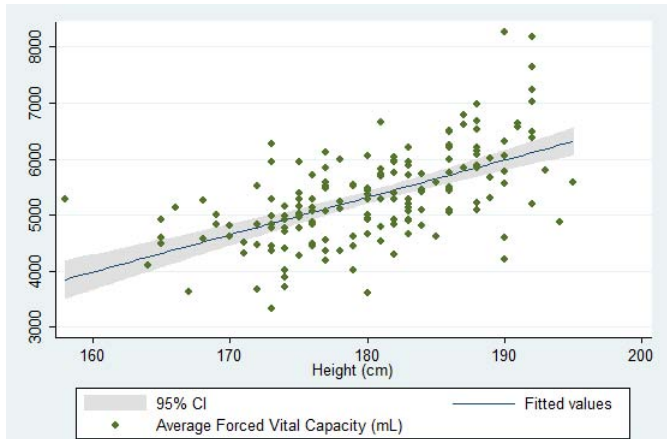


Regression Diagnostics: Independence

We just have to think through this one. For the FVC data, independence could be compromised if some of the men lived in the same (smoking) dormitory and had shared passive smoking exposure, or if some siblings made it into the sample.

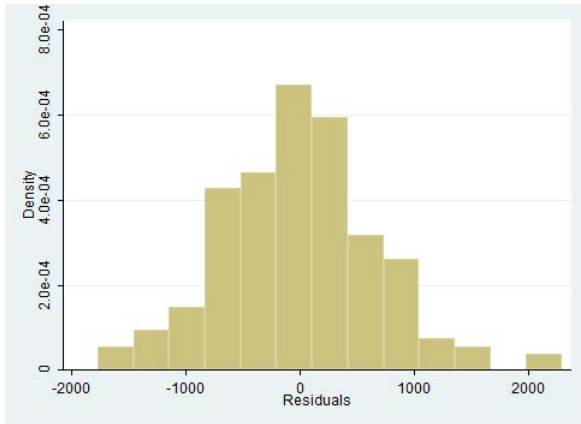
Regression Diagnostics: Linearity

To judge linearity, look at our plot of the regression line superimposed on the data points. Is the line generally consistent with the point locations, or is it missing a nonlinear pattern?



Regression Diagnostics: Normality of Errors

Let's check out a histogram of the residuals to see how normal they look. First, we generate the residuals in Stata by typing `predict r, resid`. Then we plot them by typing `histogram r`.

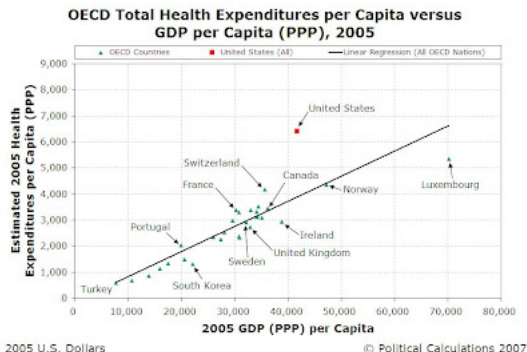


LDL Cholesterol and Age

Investigators are interested in the relationship between LDL cholesterol and age and have collected data from a group of 200 UNC professors. They wish to test the null hypothesis that age is unrelated to LDL cholesterol against the alternative that age and LDL cholesterol are related.

1. Write an appropriate linear regression model
2. Write the null hypothesis test in terms of the model's parameters

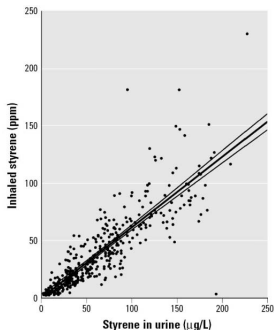
Health Expenditures



1. Which are the dependent (outcome) and independent (predictor) variables?
2. What are your best guesses for $\hat{\beta}_0$ and $\hat{\beta}_1$?
3. How do you interpret the US data relative to the fitted line?

Styrene Exposures

About 90,000 workers, including those who make boats, tubs, and showers, are potentially exposed to styrene, a primarily synthetic chemical linked to adverse effects on the central nervous system.



A researcher is conducting a meta-analysis of health effects of styrene exposure, and some studies measure styrene in urine, while others measure inhaled styrene. Based on the figure, what are your best estimates of the slope and intercept for the regression of inhaled styrene on styrene in urine? What issues may arise in a meta-analysis if studies using different styrene measures are combined?