

BIOS 600: Principles of Statistical Inference

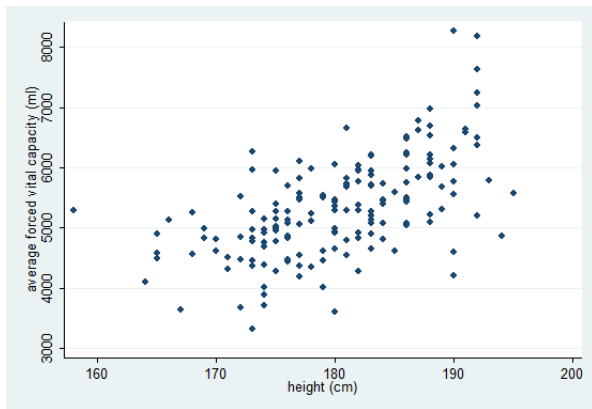
Multiple Linear Regression

Fall 2012

- ▶ Pagano and Gauvreau, Chapter 19
- ▶ For more information on regression, take BIOS 545!

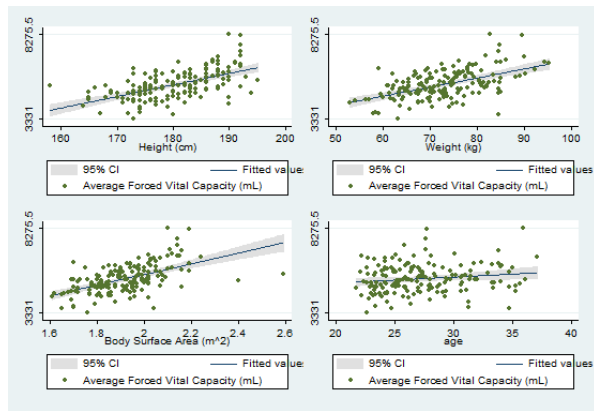
Example: EPA Study

The EPA's Health Effects Research Laboratory conducted a study of 172 adult males on campus. Last time we examined the regression of average forced vital capacity (ml) on height (cm).



Do we really think height is the important factor?

Example: EPA Study



Maybe the relationship is due to height, or weight, or surface area...how can we be sure?

Multiple Regression

The multiple regression model extends the simple linear regression model by incorporating more than one explanatory variable. The assumptions are similar to those of the simple linear regression model. This type of model is often called a *multivariable* (not multivariate, a widely-used misnomer) model.

A multiple regression model is often used to control for confounders or predictors that explain important variability in the response. Example:

- ▶ Knowing a patient's is 63" tall may not give you much information about her fasting glucose

Multiple Regression

The multiple regression model extends the simple linear regression model by incorporating more than one explanatory variable. The assumptions are similar to those of the simple linear regression model. This type of model is often called a *multivariable* (not multivariate, a widely-used misnomer) model.

A multiple regression model is often used to control for confounders or predictors that explain important variability in the response. Example:

- ▶ Knowing a patient's is 63" tall may not give you much information about her fasting glucose
- ▶ However, if you also know that her weight is 250 lbs., you can predict fasting glucose much more accurately

Multiple Regression

The multiple regression model extends the simple linear regression model by incorporating more than one explanatory variable. The assumptions are similar to those of the simple linear regression model. This type of model is often called a *multivariable* (not multivariate, a widely-used misnomer) model.

A multiple regression model is often used to control for confounders or predictors that explain important variability in the response. Example:

- ▶ Knowing a patient's is 63" tall may not give you much information about her fasting glucose
- ▶ However, if you also know that her weight is 250 lbs., you can predict fasting glucose much more accurately
- ▶ If you know she had gestational diabetes 10 years ago, you can do even better!

Multiple Regression

The model is given by $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$,
where

- ▶ p is the total number of predictor or explanatory variables

Multiple Regression

The model is given by $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$, where

- ▶ p is the total number of predictor or explanatory variables
- ▶ y_i is the outcome (dependent variable) of interest

Multiple Regression

The model is given by $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$, where

- ▶ p is the total number of predictor or explanatory variables
- ▶ y_i is the outcome (dependent variable) of interest
- ▶ β_0 is the intercept parameter

Multiple Regression

The model is given by $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$, where

- ▶ p is the total number of predictor or explanatory variables
- ▶ y_i is the outcome (dependent variable) of interest
- ▶ β_0 is the intercept parameter
- ▶ $\beta_1, \beta_2, \dots, \beta_p$ are the slope parameters

Multiple Regression

The model is given by $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$, where

- ▶ p is the total number of predictor or explanatory variables
- ▶ y_i is the outcome (dependent variable) of interest
- ▶ β_0 is the intercept parameter
- ▶ $\beta_1, \beta_2, \dots, \beta_p$ are the slope parameters
- ▶ $x_{1i}, x_{2i}, \dots, x_{pi}$ are predictor variables

Multiple Regression

The model is given by $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$, where

- ▶ p is the total number of predictor or explanatory variables
- ▶ y_i is the outcome (dependent variable) of interest
- ▶ β_0 is the intercept parameter
- ▶ $\beta_1, \beta_2, \dots, \beta_p$ are the slope parameters
- ▶ $x_{1i}, x_{2i}, \dots, x_{pi}$ are predictor variables
- ▶ ε_i is the error (like the β s, it is not observed)

Multiple Regression

The model is given by $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$, where

- ▶ p is the total number of predictor or explanatory variables
- ▶ y_i is the outcome (dependent variable) of interest
- ▶ β_0 is the intercept parameter
- ▶ $\beta_1, \beta_2, \dots, \beta_p$ are the slope parameters
- ▶ $x_{1i}, x_{2i}, \dots, x_{pi}$ are predictor variables
- ▶ ε_i is the error (like the β s, it is not observed)
- ▶ Assumptions are essentially the same as in simple linear regression

Multiple Regression

The model is given by $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$, where

- ▶ p is the total number of predictor or explanatory variables
- ▶ y_i is the outcome (dependent variable) of interest
- ▶ β_0 is the intercept parameter
- ▶ $\beta_1, \beta_2, \dots, \beta_p$ are the slope parameters
- ▶ $x_{1i}, x_{2i}, \dots, x_{pi}$ are predictor variables
- ▶ ε_i is the error (like the β s, it is not observed)
- ▶ Assumptions are essentially the same as in simple linear regression
- ▶ Interpretations are conditional on other covariates in model (more next)

Multiple Regression

Consider the model $FVC_i = \beta_0 + \beta_1 HEIGHT_i + \beta_2 WEIGHT_i + \varepsilon_i$.
The parameter interpretations are below.

- ▶ β_0 represents the expected FVC for someone who is 0cm tall and who weighs 0 kg (nonsense of course – don't extrapolate to infinitesimally small humans!)

Let's fit the model and look at the results.

Multiple Regression

Consider the model $FVC_i = \beta_0 + \beta_1 HEIGHT_i + \beta_2 WEIGHT_i + \varepsilon_i$.
The parameter interpretations are below.

- ▶ β_0 represents the expected FVC for someone who is 0cm tall and who weighs 0 kg (nonsense of course – don't extrapolate to infinitesimally small humans!)
- ▶ β_1 represents the expected increase in FVC for a 1 cm increase in height, holding the weight constant

Let's fit the model and look at the results.

Multiple Regression

Consider the model $FVC_i = \beta_0 + \beta_1 HEIGHT_i + \beta_2 WEIGHT_i + \varepsilon_i$.
The parameter interpretations are below.

- ▶ β_0 represents the expected FVC for someone who is 0cm tall and who weighs 0 kg (nonsense of course – don't extrapolate to infinitesimally small humans!)
- ▶ β_1 represents the expected increase in FVC for a 1 cm increase in height, holding the weight constant
- ▶ β_2 represents the expected increase in FVC for a 1 kg increase in weight, holding height constant

Let's fit the model and look at the results.

```
. regress avgfvc height weight
```

Source	SS	df	MS
Model	45172573.6	2	22586286.8
Residual	70054357.3	167	419487.17
Total	115226931	169	681816.16

Number of obs = **170**
 F(2, 167) = **53.84**
 Prob > F = **0.0000**
 R-squared = **0.3920**
 Adj R-squared = **0.3848**
 Root MSE = **647.68**

avgfvc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
height	47.14238	8.90984	5.29	0.000	29.55194	64.73282
weight	28.53403	7.458033	3.83	0.000	13.80986	43.25821
_cons	-5237.113	1345.064	-3.89	0.000	-7892.635	-2581.592

Hypotheses of interest may include the following.

- ▶ $H_0 : \beta_1 = \beta_2 = 0$ versus $H_A : \text{either } \beta_1 \text{ or } \beta_2 \text{ or both are not } 0$. This tests whether height, weight, or both have any effect on FVC and is called an overall, group, or 'chunk' test.

Hypotheses of interest may include the following.

- ▶ $H_0 : \beta_1 = \beta_2 = 0$ versus $H_A : \text{either } \beta_1 \text{ or } \beta_2 \text{ or both are not } 0$. This tests whether height, weight, or both have any effect on FVC and is called an overall, group, or 'chunk' test.
- ▶ $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$. This tests whether height is associated with FVC, controlling for weight. The answer may differ from the one you got in a simple linear regression model if height and weight are related.

Hypotheses of interest may include the following.

- ▶ $H_0 : \beta_1 = \beta_2 = 0$ versus $H_A : \text{either } \beta_1 \text{ or } \beta_2 \text{ or both are not } 0$. This tests whether height, weight, or both have any effect on FVC and is called an overall, group, or 'chunk' test.
- ▶ $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$. This tests whether height is associated with FVC, controlling for weight. The answer may differ from the one you got in a simple linear regression model if height and weight are related.
- ▶ $H_0 : \beta_2 = 0$ versus $H_A : \beta_2 \neq 0$. This tests whether weight is associated with FVC, controlling for height.

EPA Study: Overall F test

The *overall F test* tests the hypothesis

$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$ (all non-intercept parameters are 0).

This is a test of whether any of our predictors are related to the response. This is an F test like those we used in ANOVA and has numerator df equal to the number of parameters being tested (p), and denominator df equal to the total sample size minus the number of mean parameters in the model ($n - p - 1$ because the intercept is also a parameter).

In the EPA study, our p-value for the overall F test is < 0.0001 and we conclude that at least one predictor is related to FVC.

EPA Study: Testing Height

We can also test parameters individually, as we did in simple linear regression, except that the df for the t-test is $n - p - 1$ instead of $n - 2$ (in simple linear regression, we always have $p=1$). First we test whether height matters after adjusting for weight, or $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$. We have a t-statistic of 5.29 on with $df=167$ and corresponding $p < 0.001$. We reject H_0 and conclude that height is related to FVC after controlling for weight. In particular, at the same weight, a subject who is 1cm taller than another would be expected to have lung function that is 47.1 (29.6, 64.7) mL greater.

EPA Study: Testing Weight

- ▶ What is H_0 (in Greek symbols and in words!)?
- ▶ What is H_A ?
- ▶ What is the value of the test statistic?
- ▶ What is the reference distribution?
- ▶ What is the p-value?
- ▶ What is the test decision?
- ▶ How do we interpret the results?

Interaction

Sometimes, the effect of one variable depends on the value of another. For example, the impact of a 10 lb. weight gain on cardiovascular disease risk may be different for a man who is 5'3" tall and a man who is 6'3" tall. To model such a relationship (often called *effect modification* because one variable modifies the effect of another), we create an *interaction term*. This is created simply by multiplying two predictors x_1 and x_2 to create a new predictor, x_1x_2 . When interaction terms are in a model, interpretations can become tricky.

We can see whether the effect of height on FVC is modified by weight as follows.

First we create an interaction term by multiplying height and weight. Stata: `generate htwt=height*weight`.

Interaction

Model:

$$FVC_i = \beta_0 + \beta_1 HEIGHT_i + \beta_2 WEIGHT_i + \beta_3 HEIGHT_i WEIGHT_i + \varepsilon_i$$

```
. regress avgfvc height weight htwt
```

Source	SS	df	MS
Model	46870577.6	3	15623525.9
Residual	68356353.3	166	411785.261
Total	115226931	169	681816.16

Number of obs = **170**
F(3, 166) = **37.94**
Prob > F = **0.0000**
R-squared = **0.4068**
Adj R-squared = **0.3960**
Root MSE = **641.7**

avgfvc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
height	-73.90973	60.26272	-1.23	0.222	-192.8899	45.07044
weight	-276.3201	150.3085	-1.84	0.068	-573.083	20.44267
htwt	1.685493	.8300283	2.03	0.044	.0467198	3.324265
_cons	16597.4	10834.77	1.53	0.127	-4794.317	37989.12

Interaction

How do we interpret the estimates? For a given height and weight, our predicted FVC is

$\widehat{FVC}_i = \hat{\beta}_0 + \hat{\beta}_1 HEIGHT_i + \hat{\beta}_2 WEIGHT_i + \hat{\beta}_3 HEIGHT_i WEIGHT_i$
and for someone at the same weight but 1 cm taller, the predicted FVC is

$$\begin{aligned}\widehat{FVC}_{i'} &= \hat{\beta}_0 + \hat{\beta}_1 (HEIGHT_i + 1) + \hat{\beta}_2 WEIGHT_i + \\ &\quad \hat{\beta}_3 (HEIGHT_i + 1) WEIGHT_i \\ &= \hat{\beta}_0 + \hat{\beta}_1 HEIGHT_i + \hat{\beta}_1 + \hat{\beta}_2 WEIGHT_i \\ &\quad + \hat{\beta}_3 HEIGHT_i WEIGHT_i + \hat{\beta}_3 WEIGHT_i\end{aligned}$$

Subtracting, we have $\widehat{FVC}_{i'} - \widehat{FVC}_i = \hat{\beta}_1 + \hat{\beta}_3 WEIGHT_i$, which is the expected change in FVC for a 1cm change in height.

What is the expected change in FVC for a 1kg change in weight?

One common problem in multiple regression is *collinearity*, which occurs when multiple highly correlated variables are used as predictors. In this case, the model can become unstable (often seen as standard errors that get huge and lead to huge confidence interval estimates), and it can be hard to assess the impact of the predictors.

An extreme example would be including height in cm and height in inches in the same model. Because the variables provide the same information, the model wouldn't be able to determine which one was the appropriate predictor.

Diagnosing Collinearity

Consider the model

$$FVC_i = \beta_0 + \beta_1 HEIGHT_i + \beta_2 WEIGHT_i + \beta_3 AREA_i + \varepsilon_i$$

```
. regress avgfvc height weight area
```

Source	SS	df	MS
Model	45237157.9	3	15079052.6
Residual	69989773.1	166	421625.139
Total	115226931	169	681816.16

Number of obs = 170
F(3, 166) = 35.76
Prob > F = 0.0000
R-squared = 0.3926
Adj R-squared = 0.3816
Root MSE = 649.33

avgfvc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
height	86.15486	100.0784	0.86	0.391	-111.4357	283.7454
weight	83.74776	141.272	0.59	0.554	-195.1738	362.6693
area	-4974.349	12709.73	-0.39	0.696	-30067.9	20119.2
_cons	-6750.966	4096.296	-1.65	0.101	-14838.52	1336.589

Yikes! Now nothing looks predictive! However, we have three important clues.

1. Height was an important predictor alone in the model (as are weight and area if you fit those)
2. Standard errors and interval estimates are huge
3. The overall F test is significant

Collinearity

A significant overall F test with no significant individual variable test is a typical sign of collinearity. Check out the correlations among the three predictors.

```
. correlate height weight area  
(obs=170)
```

	height	weight	area
height	1.0000		
weight	0.6117	1.0000	
area	0.8352	0.9453	1.0000

Look at weight and area! They contain much of the same information, and only one should be used (which one depends on your knowledge of how body size relates to FVC).

There is no fixed criterion for correlation to exclude a variable for collinearity. It is possible to construct examples where the correlation is very high, but collinearity is not a problem because the information about the outcome in the two variables is different.

Predictors in regression models do not have to be continuous. We can easily incorporate predictors that are categorical (e.g., gender, marital status, type of health insurance) or counts as well.

Categorical variables are often coded 0/1 and called 'dummy' or 'indicator' variables. We'll look at an example of this in a recent scientific paper.

Study of Child BMI, TV-watching Time, and Parental Obesity

A Turkish study recently examined the relationship between a child's BMI, tv-watching time, and parental obesity. Variables were defined as follows.

- ▶ BMI_i : body mass index (kg/m^2) of child i

Suppose they fit the model

$$BMI_i = \beta_0 + \beta_1 TV_i + \beta_2 obese1_i + \beta_3 obese2_i + \varepsilon_i$$

Study of Child BMI, TV-watching Time, and Parental Obesity

A Turkish study recently examined the relationship between a child's BMI, tv-watching time, and parental obesity. Variables were defined as follows.

- ▶ BMI_i : body mass index (kg/m^2) of child i
- ▶ TV_i : average tv-watching time of child i (hours/day)

Suppose they fit the model

$$BMI_i = \beta_0 + \beta_1 TV_i + \beta_2 obese1_i + \beta_3 obese2_i + \varepsilon_i$$

Study of Child BMI, TV-watching Time, and Parental Obesity

A Turkish study recently examined the relationship between a child's BMI, tv-watching time, and parental obesity. Variables were defined as follows.

- ▶ BMI_i : body mass index (kg/m^2) of child i
- ▶ TV_i : average tv-watching time of child i (hours/day)
- ▶ $obese1_i$: =1 if exactly 1 biological parent is obese and 0 otherwise

Suppose they fit the model

$$BMI_i = \beta_0 + \beta_1 TV_i + \beta_2 obese1_i + \beta_3 obese2_i + \varepsilon_i$$

Study of Child BMI, TV-watching Time, and Parental Obesity

A Turkish study recently examined the relationship between a child's BMI, tv-watching time, and parental obesity. Variables were defined as follows.

- ▶ BMI_i : body mass index (kg/m^2) of child i
- ▶ TV_i : average tv-watching time of child i (hours/day)
- ▶ $obese1_i$: =1 if exactly 1 biological parent is obese and 0 otherwise
- ▶ $obese2_i$: =1 if both biological parents are obese and 0 otherwise

Suppose they fit the model

$$BMI_i = \beta_0 + \beta_1 TV_i + \beta_2 obese1_i + \beta_3 obese2_i + \varepsilon_i$$

Study of Child BMI, TV-watching Time, and Parental Obesity

A Turkish study recently examined the relationship between a child's BMI, tv-watching time, and parental obesity. Variables were defined as follows.

- ▶ BMI_i : body mass index (kg/m^2) of child i
- ▶ TV_i : average tv-watching time of child i (hours/day)
- ▶ $obese1_i$: =1 if exactly 1 biological parent is obese and 0 otherwise
- ▶ $obese2_i$: =1 if both biological parents are obese and 0 otherwise
- ▶ Children with non-obese parents form the *reference or comparison group*.

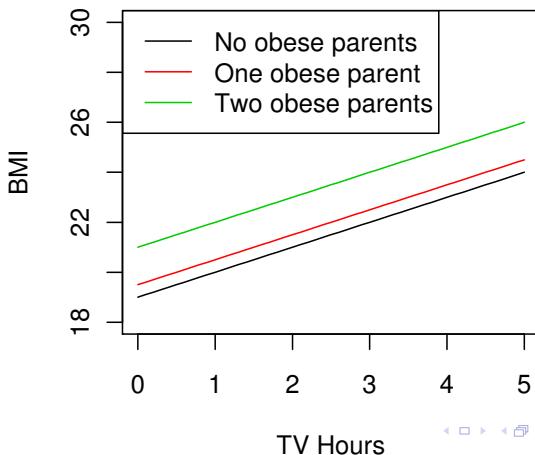
Suppose they fit the model

$$BMI_i = \beta_0 + \beta_1 TV_i + \beta_2 obese1_i + \beta_3 obese2_i + \varepsilon_i$$

Estimates

Suppose we get the estimates $\hat{\beta}_0 = 19$, $\hat{\beta}_1 = 1$, $\hat{\beta}_2 = 0.5$, and $\hat{\beta}_3 = 2$. It's easiest to interpret using a plot.

BMI by TV Time and Parental Obesity



Cholesterol Study

Prairie et al. (2012, *Journal of Women's Health*) examined cholesterol levels among women with major depression. How do you interpret the quantities in the table?

Measure	Total cholesterol	
	coefficient ^a	p
Age	2.0491	0.0005
Race (Reference: white)		0.2568
Black	-12.1145	
Other	-5.7358	
Education level (Reference: < HS)		0.0466
High school	13.6378	0.2796
Some college	29.0213	0.0175
College	19.8883	0.1597
Graduate school	54.5697	0.0098
Marital status (Reference: single)		0.5358
Married/cohabiting	8.5027	
Divorced/separated	-1.1925	
Prepregnancy BMI	0.2806	0.5400
During most recent pregnancy		
Diabetes	14.4835	0.4470
High BP	-14.9235	0.1680
Preeclampsia	-19.4453	0.1488
Vaginal delivery	8.3138	0.2910
Gestational age at delivery (weeks)	1.1676	0.5749
Infant feeding method (Reference: breast)		0.9931
Formula	-0.8219	
Both	0.0239	
Parity	3.4616	0.2968
Parity (Reference: 1)		0.2811
2	-2.4821	
3+	11.4416	
Gravidity	2.1312	0.3035

Body Dissatisfaction Study

Calado et al.
(2011)
examined
body dissatisfaction as a
function of
mass media
exposure in
Spanish
adolescents.
How do you
interpret the
quantities in
the table?

Table 3

Multiple Regression Analyses Predicting Body Dissatisfaction Taking the EDI-BD Total Scores as Dependent Variable

DV: EDI-BD Score	F	Beta	t	R ²
Total sample	79.2 [‡]			0.50
Disordered Eating (EAT-26)		2.42	7.41 [‡]	
Self-esteem (SES)		-0.25	-8.63 [‡]	
BMI		0.27	10.1 [‡]	
Internalization (I)		0.12	3.46 [‡]	
Gender		0.14	4.16 [‡]	
Awareness (C)		0.11	3.36 [‡]	
Dieting topic		0.22	5.31 [‡]	
Fitness topic		0.11	3.47 [‡]	
Beauty topic		-0.12	-2.87 [†]	
Males				
DV: EDI-BD score	37.8 [‡]			0.35
Self-esteem (SES)		-0.33	-7.20 [‡]	
BMI		0.30	6.76 [‡]	
Disordered Eating (EAT-26)		0.22	4.71 [‡]	
		0.15	3.44 [‡]	
Awareness (C) fitness topic		-0.11	-2.40*	
Females				
DV: EDI-BD score	111.6 [‡]			0.54
Disordered Eating (EAT-26)		0.29	6.77 [‡]	
Internalization (I)		0.28	6.44 [‡]	
BMI		0.28	8.01 [‡]	
Self-esteem (SES)		-0.23	-5.94 [‡]	

Abbreviations: BD, body dissatisfaction from EDI-BD; Beauty, content about TV and magazine beauty topics; BMI, body mass index; C, awareness from SATAQ-R; Dieting, content about TV and magazine dieting topics; DV, dependent variable; Fitness, content about TV and magazine fitness topics; I, internalization from SATAQ-R.

* $p < .05$.

† $p < .01$.

‡ $p < .001$.

In Utero PM_{2.5} Exposure and Lung Function

Jedrychowski
et al. (2010)
examined
associations
between *in
utero*
exposure to
PM_{2.5} and
lung function
(FEV₁) in
preschoolers.
How do you
interpret the
quantities in
the table?

Table 7. Prenatal PM_{2.5} exposure (in quartiles) on lung function (FEV₁) of 5-year-olds adjusted to potential confounders

Predictors	Coefficient	[95% CI]	P
Age (in months)	29.0	[2.96, 55.0]	0.029
Height (cm)	18.4	[12.9, 23.9]	<0.001
Gender of child (girls)	-35.4	[-81.9, 11.0]	0.134
Prenatal ETS	-17.2	[-93.4, 59.1]	0.657
Postnatal ETS	47.5	[-58.1, 153.1]	0.375
Wheezing	-54.6	[-107, -1.36]	0.045
Birthweight g (quartiles)			
≤3165 g	0.00	Reference	0.238
3166–3425 g	13.0	[-53.9, 79.8]	
3426–3719 g	-3.17	[-68.8, 62.4]	
≥3720 g	54.3	[-14.2, 122.7]	
Prenatal PM _{2.5} level (quartiles)			
<20.95 µg/m ³	0.00	Reference	0.008
20.95–32.42 µg/m ³	-32.8	[-98.6, 32.9]	
32.43–52.6 µg/m ³	-39.8	[-105, 26.1]	
>52.6 µg/m ³	-87.7	[-151, -23.6]	