Lecture 5

Today's agenda – more on OLS regressions

1. goodness of fit measures

2. standard errors of the OLS regression coefficients

3. some examples

We have a regression model for an outcome y with explanatory variables x. Write the population model as:

$$y_i = x_i'\beta^* + u_i.$$

In a sample (size N) the FOC for the OLS estimator $\hat{\beta}$ is:

$$\begin{aligned}
\sum_{i=1}^{N} x_i(y_i - x_i'\hat{\beta}) &= 0 \qquad \Rightarrow \frac{1}{N}\sum_{i=1}^{N} x_i(y_i - x_i'\hat{\beta}) \\
\Rightarrow \frac{1}{N}\sum_{i=1}^{N} x_i y_i &= \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)\hat{\beta} \\
\Rightarrow \hat{\beta} &= \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} x_i y_i
\end{aligned}$$

How well does the model fit? A "null" model is one with just a constant. If we fit that model, we know the estimated coefficient will be the mean of y_i. The "sum of squares" of y_i is:

$$SS = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

which is the sum of the squared prediction errors from a model with only a constant. From our OLS model, the corresponding "sum of squared residuals" is

$$SSR = \sum_{i=1}^{N} (y_i - x_i\hat{\beta})^2 = \sum_{i=1}^{N} \hat{u}_i{}^2.$$

We define:

$$R^2 = 1 - (SSR/SS)$$

$\sum_i x_{i1}(y_i - x_i\hat{\beta}) = 0$

$\hat{\beta}_1$

$\frac{1}{N}\sum_i (y_i - \hat{\beta}_1) = 0$

Then,

$SSR = SS$

$R^2 = 1$.

$$
\begin{aligned}
R^2 &= 1 - (SSR/SS) \\
&= 1 - \frac{\frac{1}{N}\sum_{i=1}^{N} \hat{u}_i{}^2}{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}
\end{aligned}
$$

The lower bound is 0: in this case $SSR = SS$ and the model is no better than a constant. The upper bound is 1: in this case $SSR = 0$ and the model explains y perfectly.

One problem with R^2 is that is can only get better as you add more $x's$ – there is no penalty for adding extra variables even if they don't add much. Adjusted R^2, denoted by $\bar{R}^2$, corrects for the use of extra explanatory variables. The idea is to replace the two terms in R^2 with "degrees-of-freedom-corrected" terms:

$$R^2 = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N} \hat{u}_i{}^2}{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

$$\bar{R}^2 = 1 - \frac{\frac{1}{N-K}\sum_{i=1}^{N} \hat{u}_i{}^2}{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

$$\bar{R}^2 = 1 - \frac{\frac{1}{N-K}\sum_{i=1}^{N} \hat{u}_i{}^2}{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

The term $\frac{1}{N-K}\sum_{i=1}^{N} \hat{u}_i{}^2$ is usually called the "mean squared error" (MSE) of the model. The intuition for dividing by $N-K$ is this: if you have K explanatory variables, and you had a sample of size $N = K$, then you could perfectly predict all the observations, and $SSR = 0$. So the "degrees of freedom" in the SSR is $N-K$. It is also conventional to estimate the variance of the regression residual by the MSE:

$$\widehat{Var}[u_i] = \frac{1}{N-K}\sum_{i=1}^{N} \hat{u}_i{}^2$$

Finally, people call $\sqrt{MSE}$ the "standard error of the regression". It is an estimate of the *standard deviation* of u_i.

Inference

After we estimate the coefficients in a model, we want to know how likely it is that our estimates are "close" to the truth (where "truth" means the population regression coefficients). This will depend on the sample size, how much variability there is in y, and on how easy it is to separate out the effects of the various elements of x. We begin with the "classic" case where the $X's$ are "fixed" and we have normally distributed residuals.

indep
variance

We assume the population model is:

$$y_i = x_i'\beta^* + u_i.$$

We also assume:

1. we have an independent sample of size N
2. the elements of x_i are not linearly dependent $[x_i' x_i]$ is invertible
3. given the $x's$ each of the u_i are iid normals: $u_i \sim N(0,\sigma_u^2)$

Under the previous assumptions, the vector of OLS coefficients is distributed normally conditional on the sample $x's$:

$$\widehat{\beta} - \beta^* \sim \underline{N(0, V)}$$

Variance of matrix

$$V = \frac{\sigma_u^2}{N} S_{xx}^{-1}, \quad and \quad S_{xx} = \frac{1}{N} \sum_{i=1}^{N} x_i x_i'.$$

For β^* to be precise, N big, σ small, S_{xx} small [large range of x's]

To prove this we write:

$$
\begin{aligned}
\widehat{\beta} &= \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} x_i y_i \\
&= S_{xx}^{-1}\frac{1}{N}\sum_{i=1}^{N} x_i(x_i'\beta^* + u_i) \\
&= S_{xx}^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)\beta^* + S_{xx}^{-1}\frac{1}{N}\sum_{i=1}^{N} x_i u_i \\
&= \beta^* + \sum_{i=1}^{N} a_i(X) u_i
\end{aligned}
$$

where $a_i(X) = \frac{1}{N} S_{xx}^{-1} x_i$ is a $K \times 1$ vector that is a function of all the $x_i's$ (X).

$$\widehat{\beta} - \beta^* = \sum_{i=1}^{N} a_i(X) u_i$$

So conditional on X, each row of $\widehat{\beta} - \beta^*$ is just a weighted sum of the $u_i's$. Thus, conditional on X, $\widehat{\beta} - \beta^*$ is normally distributed! We have:

$$E[\widehat{\beta} - \beta^* | X] = E[\sum_{i=1}^{N} a_i(X) u_i | X] = \sum_{i=1}^{N} a_i(X) E[u_i | X] = 0$$

What about the variance of $\widehat{\beta} - \beta^*$?

To calcuate the variance of $\widehat{\beta}$, note:

$$\widehat{\beta} - \beta^* = \sum_{i=1}^{N} a_i u_i$$

There are K rows of this expression:

$$\begin{array}{ll} \widehat{\beta_1} - \beta_1^* & \overset{1}{\underset{N}{\rightarrow}} \sum_{i=1}^{N} a_{1i} u_i \\ \widehat{\beta_2} - \beta_2^* & = \sum_{i=1}^{N} a_{2i} u_i \\ \cdots & \cdots \\ \widehat{\beta_K} - \beta_K^* & \overset{1}{\underset{N}{\rightarrow}} \sum_{i=1}^{N} a_{Ki} u_i \end{array}$$

$$(\hat{\beta}_j - \beta_j^*) = \sum_i a_{ji} u_i$$

known #'s

$$Cov(\hat{\beta}_j - \beta_j^*, \hat{\beta}_k - \beta_k^*) = Cov\left(\sum_i a_{ji} u_i, \sum_j a_{kj} u_j\right)$$

since $u_i \sim N(0, \sigma^2) | X$,

So:

* relevant on heteroskedasticity

$$Var[\hat{\beta}_j - \beta_j^* | X] = \sigma_u^2 \sum_{i=1}^{N} a_{ji}^2 \geq E\left[\left(\sum_i a_{ji} u_j\right) \cdot \left(\sum_i a_{ki} u_j\right)\right]$$

consistent with

$$\geq \sigma_u^2 a_{j1} a_{k1} + \sigma_u^2 a_{j2} a_{k2}$$

outer product

$$Cov[\hat{\beta}_j - \beta_j^*, \hat{\beta}_k - \beta_k^* | X] = \sigma_u^2 \sum_{i=1}^{N} a_{ji} a_{ki}$$

since the $u_i's$ are independent of each other. In matrix form:

$$Var[\hat{\beta} - \beta^* | X] = \sigma_u^2 \sum_{i=1}^{N} a_i a_i'$$

Assume 2 coefficients

$$Var\begin{pmatrix} \hat{\beta}_1 - \beta_1^* \\ \hat{\beta}_2 - \beta_2^* \end{pmatrix} \Big| X \Big)$$

$$\begin{array}{|c|c|} \hline Var(\hat{\beta}_1 - \beta_1^*) & Cov(\hat{\beta}_2 - \beta_2^*, \hat{\beta}_1 - \beta_1^*) \\ \hline Cov(\hat{\beta}_1 - \beta_1^*, \hat{\beta}_2 - \beta_2^*) & Var(\hat{\beta}_2 - \beta_2^*) \\ \hline \end{array}$$

Thus

it's a matrix

$$\begin{aligned} Var[\hat{\beta} - \beta^* | X] &= \sigma_u^2 \sum_{i=1}^{N} a_i a_i' \\ &= \sigma_u^2 \sum_{i=1}^{N} \frac{1}{N^2} S_{xx}^{-1} x_i x_i' S_{xx}^{-1} \\ &= \frac{1}{N} \sigma_u^2 S_{xx}^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} x_i x_i' \right) S_{xx}^{-1} \\ &= \frac{1}{N} \sigma_u^2 S_{xx}^{-1} \end{aligned}$$

Variance of the OLS coefficients with the others

S_{xx}

estimated

A very important example: $x_i' = (1, x_{2i})$ – we have a constant and 1 other regressor.

$$
\begin{aligned}
S_{xx} &= \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} 1 \\ x_{2i} \end{pmatrix}(1 \quad x_{2i}) \\
&= \frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix} 1 & x_{2i} \\ x_{2i} & x_{2i}^2 \end{pmatrix} \\
&= \begin{pmatrix} 1 & \bar{x}_2 \\ \bar{x}_2 & \frac{1}{N}\sum_{i=1}^{N} x_{2i}^2 \end{pmatrix}
\end{aligned}
$$

$$
S_{xx}^{-1} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N} x_{2i}^2 - \bar{x}_2^2}\begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N} x_{2i}^2 & -\bar{x}_2 \\ -\bar{x}_2 & 1 \end{pmatrix}
$$

$$S_{xx}^{-1} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N} x_{2i}^2 - \bar{x}_2^2} \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N} x_{2i}^2 & -\bar{x}_2 \\ -\bar{x}_2 & 1 \end{pmatrix}$$

The (2,2) element is:

$$\begin{aligned} \left(S_{xx}^{-1}\right)_{2,2} &= \frac{1}{\frac{1}{N}\sum_{i=1}^{N} x_{2i}^2 - \bar{x}_2^2} \\ &= \frac{1}{\frac{1}{N}\sum_{i=1}^{N} (x_{2i} - \bar{x}_2)^2} \\ &= \frac{1}{\widehat{\sigma}^2(x_2)} \end{aligned}$$

So, conditional on X,

$$\widehat{\beta}_2 - \beta_2^* \sim N(0, \frac{1}{N}\frac{\sigma_u^2}{\widehat{\sigma}^2(x_2)})$$

which is a formula you should recall from 140/141. Notice that we don't know σ_u^2: we estimate it using the MSE:

$$\widehat{\sigma}_u^2 = \frac{1}{N-2}\sum_{i=1}^{N} \widehat{u}_i{}^2$$

which makes sense because we are assuming each u_i has the same variance.

In a standard regression package, the default standard errors (or default "estimated sampling errors") of the OLS coefficients are calculated "as if" the classic normal model were true. In particular, the reported variance/covariance matrix of the estimated OLS coefficients is:

$$\frac{1}{N}\widehat{\sigma}_u^2 S_{xx}^{-1}$$

where $S_{xx} = \frac{1}{N}\sum_i x_i x_i'$. Notice the three terms:

- the $1/N$ term reflects the sample size

- the $\widehat{\sigma}_u^2$ term reflects our estimate of the variability in y_i conditional on x_i

- the S_{xx}^{-1} term reflects the difficulty of pulling apart the contributions of the different rows of x_i

So we have $\widehat{\beta} - \beta^* \sim N(0, V)$, and an estimate of V, $\widehat{V} = \frac{1}{N}\widehat{\sigma}_u^2 S_{xx}^{-1}$. How do we conduct inference?

1. To test the null $\beta_k^* = 0$: we form the ratio: $\widehat{\beta}_k / \sqrt{\widehat{V}_{kk}}$ and compare this to a $N(0,1)$ T-test

If $|\widehat{\beta}_k / \sqrt{\widehat{V}_{kk}}| > 1.96$ then we "reject the null" under a 2-sided test at 95% confidence

2. How do we test $b\beta_k^* - c\beta_j^* = 0$? $Var(aX+bY) = a^2 var(X) + b^2 var(Y) + 2ab\,cov(X,Y)$

Note $Var[b\beta_k^* - c\beta_j^*] = b^2 Var[\beta_k^*] + c^2 Var[\beta_j^*] - 2bc Cov[\beta_k^*, \beta_j^*]$

So we pull out the corresponding elements of $\widehat{V}$, and form the ratio:

$$\frac{b\widehat{\beta}_k - c\widehat{\beta}_j}{b^2\widehat{V}_{kk} + c^2\widehat{V}_{jj} - 2bc\widehat{V}_{kj}}$$

$\beta_1^* \neq 0$ $\beta_2^* \neq 0$ β_{17} (mean) 0

There is a potentially serious limitation of the "classic normal" model. In lots of cases, we think that $E[y_i|x_i]$ is not really linear. The classic case says that

$$E[y_i|x_i] = x_i'\beta^*$$

When that is not true, we know that the "true error" for the ith observation is

$$\begin{aligned} u_i &= y_i - E[y_i|x_i] + E[y_i|x_i] - x_i'\beta^* \\ &= \varepsilon_i + v_i \end{aligned}$$

In this case the error at each value of x_i can have a non-zero mean, and possibly a variance that depends on x_i. So we'll have to "tech up" our sampling errors!

Example:

- mom-baby data from problem set #1

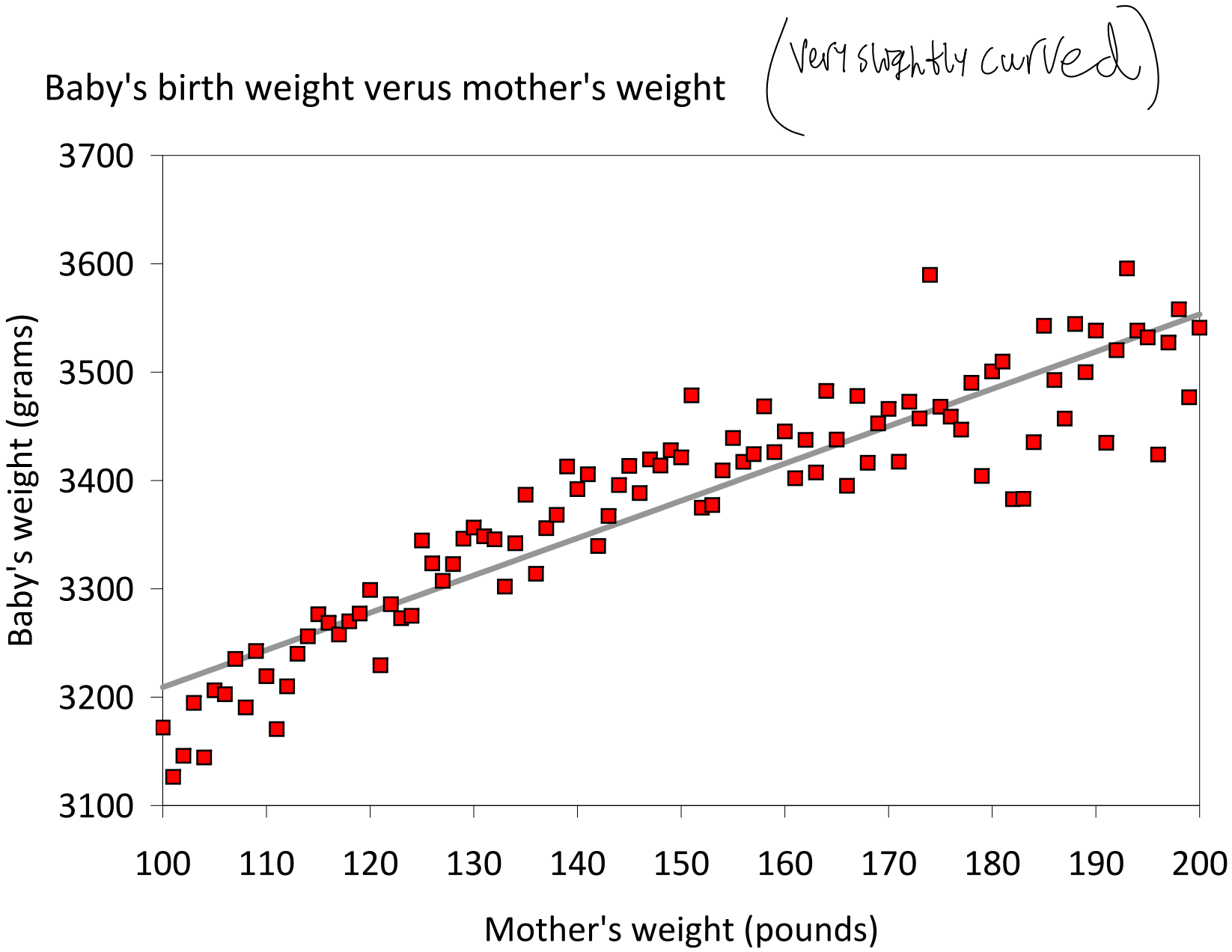
- can fit a linear model BUT the R-sq is pretty low - lots of dispersion

- looking more closely we see that there is a "specification error"

$$u_i = y_i - E[y_i|x_i] + E[y_i|x_i] - x_i'\beta^* = \varepsilon_i + v_i$$

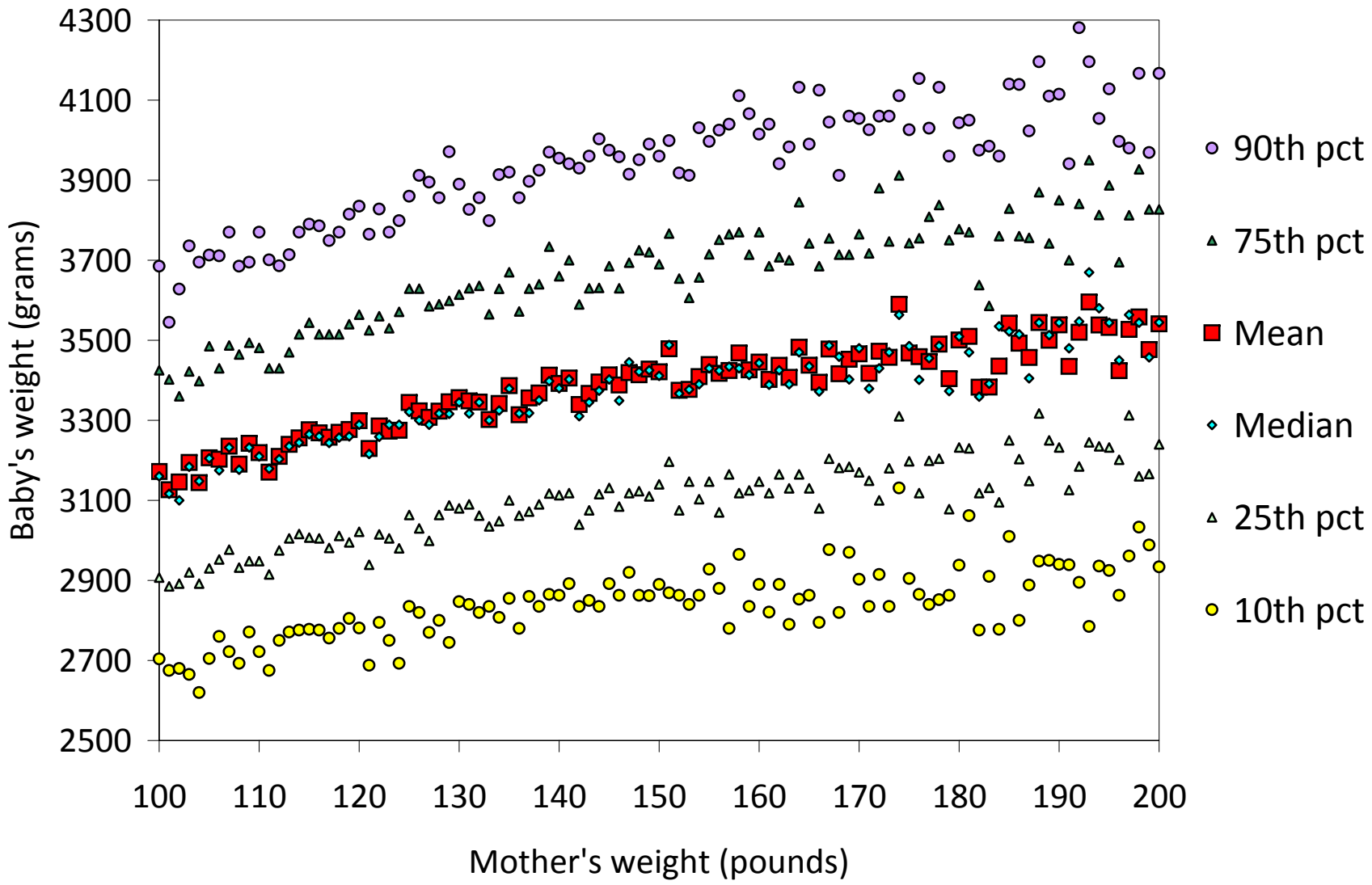
- some curvature in simple linear fit

- and heteroskedasticity


Baby's birth weight verus mother's weight
(very slightly curved)
Baby's weight (grams)
3700
3600
3500
3400
3300
3200
3100
100
110
120
130
140
150
160
170
180
190
200
Mother's weight (pounds)

$E[u_i|x_i] \neq 0$

$Var(u_i|x_i) \neq \sigma_u^2$



Variation in baby's birth weight verus mother's weight