LECTURE 8:
PREDICTIONS

Jan Zouhar

Introductory Econometrics

Two kinds of predictions

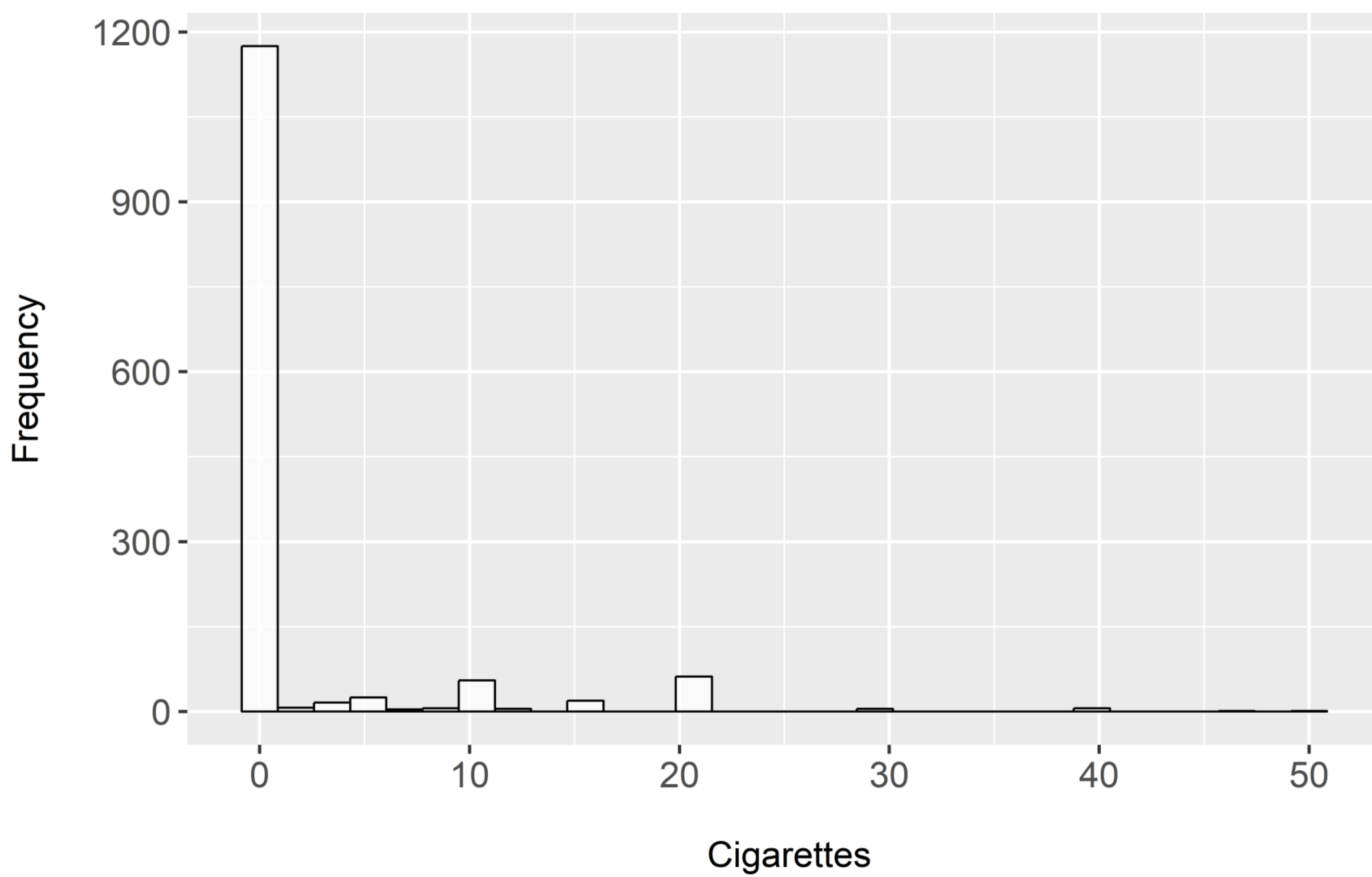
- consider the model

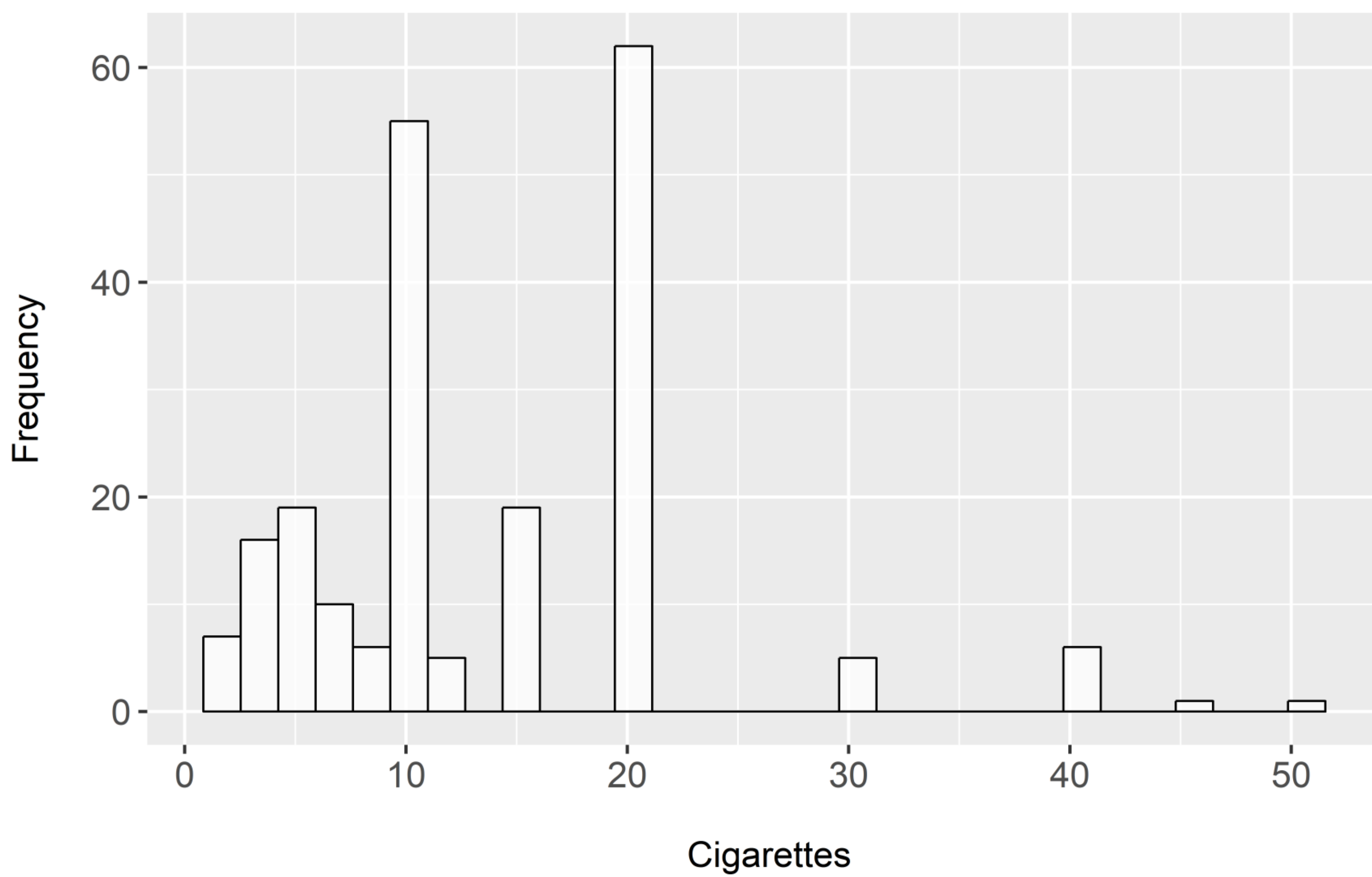
$$bweight = \beta_0 + \beta_1 cigs + u$$

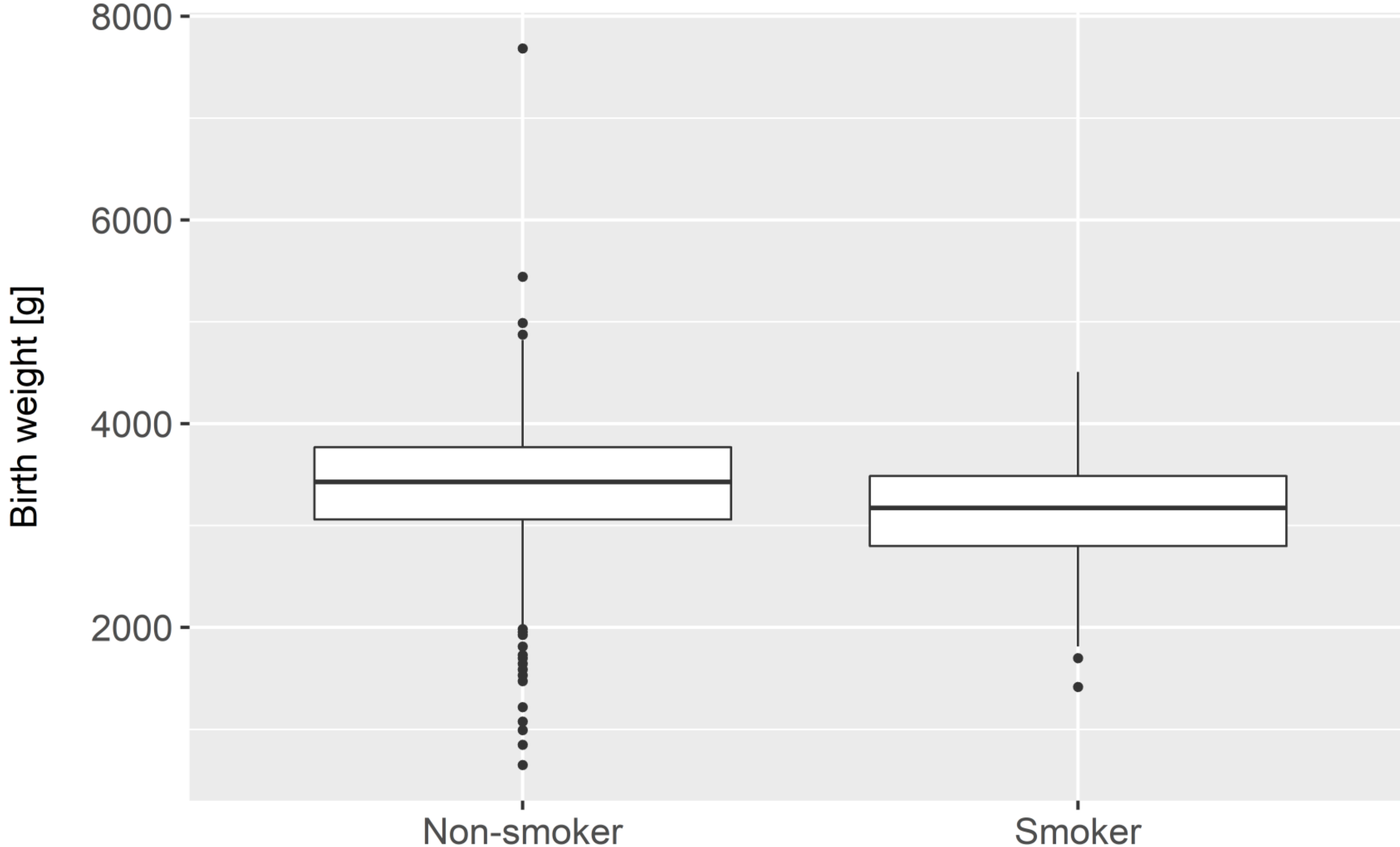
estimated as

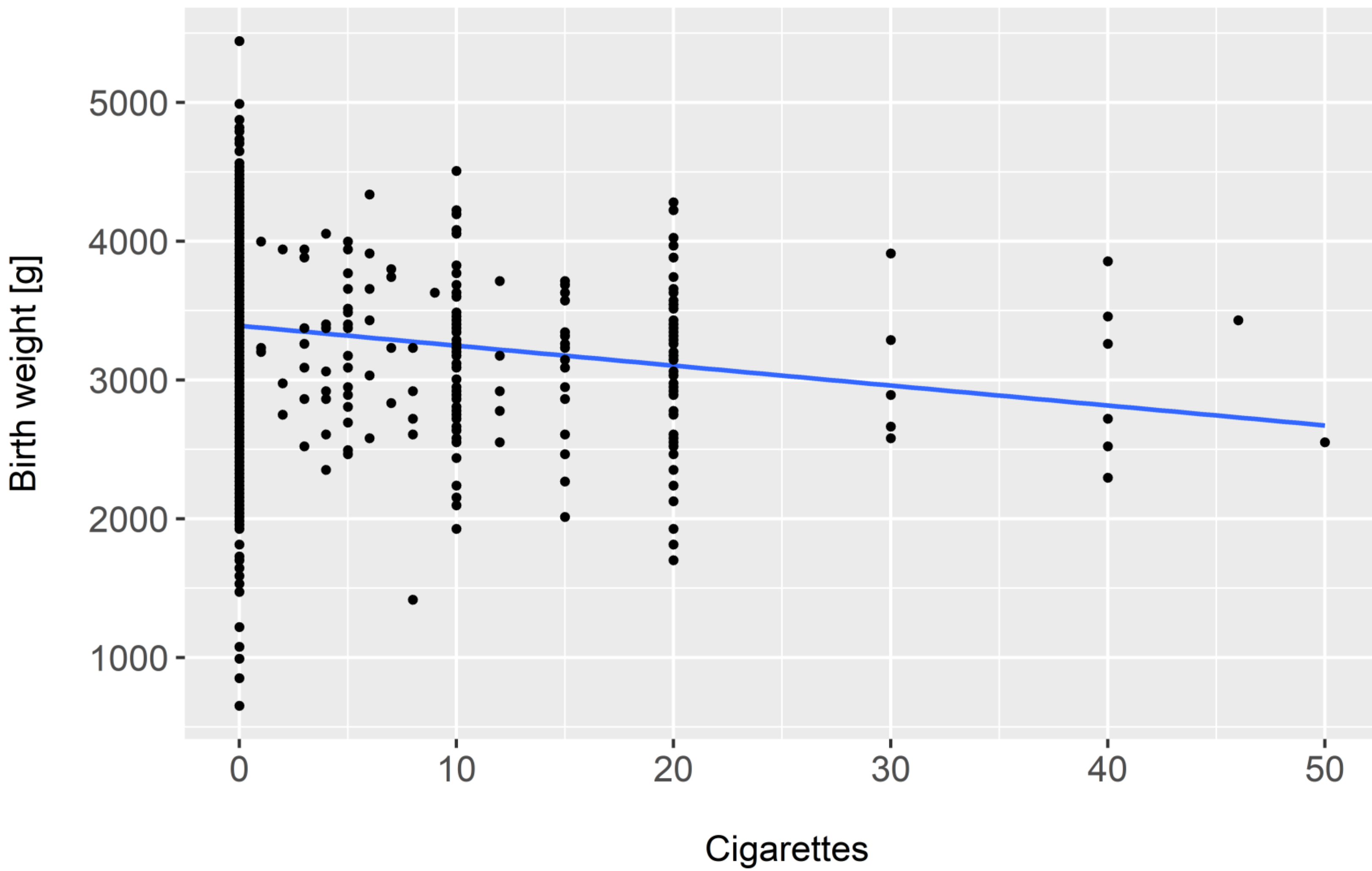
$$bweight = 3{,}395 - 14.57\, cigs$$

- assume we know a pregnant woman who smokes 10 cigarettes a day
- there are two kinds of predictions we might be interested in:
  1. predict the actual weight of the woman's baby, denoted $bweight_P$
  2. predict E[$bweight$ | $cigs$ = 10], the average birth weight for a mother smoking 10 cigarettes a day, denoted θ
- point prediction is the same in both cases: 3,395 − 14.57 × 10, denoted $\hat{\theta}$
- what differs is the 95% CI (or, the standard errors)
  1. 95% CI for the birth weight of a baby of a particular mother (smoking 10 cigarettes a day), typically called the **prediction interval**
  2. 95% CI for a mean in the category of mothers (smoking 10 cigarettes a day), i.e. 95% CI for θ


1200
900
600
300
0
Frequency
0
10
20
30
40
50
Cigarettes


Frequency
60
40
20
0
0
10
20
30
40
50
Cigarettes


8000
6000
4000
2000
Birth weight [g]
Non-smoker
Smoker


Birth weight [g]
5000
4000
3000
2000
1000
0
10
20
30
40
50
Cigarettes

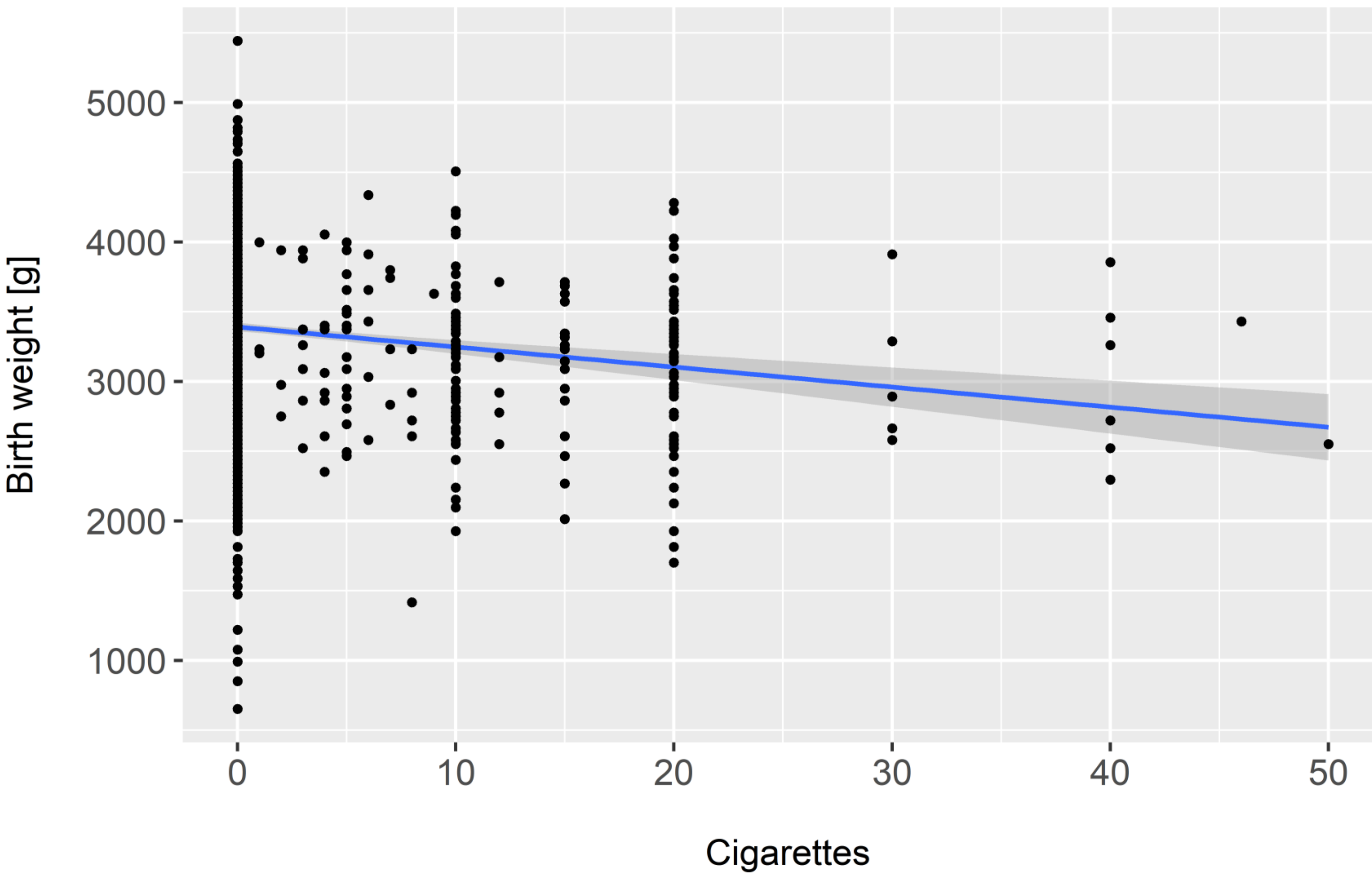
95% CI for ϑ

- we are interested in the 95% CI for $\theta = \beta_0 + 10\beta_1$
- it is useful to rewrite the estimated equation as

$$\begin{aligned} bweight &= \beta_0 + \beta_1 cigs + u \\ &= \underbrace{\beta_0 + 10\beta_1}_{\theta} + \beta_1 (\underbrace{cigs - 10}_{A}) + u \\ &= \theta + \beta_1 A + u \end{aligned}$$

- this gives us a simple procedure to find the 95% for E[$bweight \mid cigs = 10$]

1. Create a new variable $A = cigs - 10$.
   Gretl: (Add → Define new variable → A = cigs − 10)
2. Regress *bweight* on A. The intercept (constant) in this equation is θ, and the 95% CI for the intercept is constructed as usual, i.e. $\hat{\theta} \pm c \cdot \mathrm{se}(\hat{\theta})$, where c is either the number 2, or, if more precision is required, the 97.5$^{\text{th}}$ percentile of t with $n - k - 1$ degrees of freedom.
   Gretl: (Analysis → Confidence intervals for coefficients)



Birth weight [g]
5000
4000
3000
2000
1000
0
10
20
30
40
50
Cigarettes

Prediction intervals

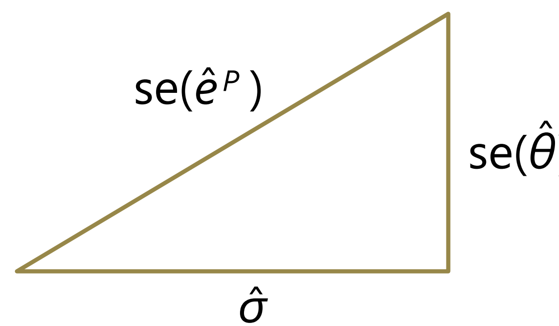
- predicted value: $bweight^P = \beta_0 + \beta_1(10) + u \;=\; \theta + u$
- point prediction: $\hat{\theta}$
  - this makes sense as $\mathrm{E}\,u = 0$
- **prediction error**: $\hat{e}^P = bweight^P - bweight_P \quad (= \theta + u - \hat{\theta})$
- two sources of the predictions error:
  1. the population regression function is not estimated precisely; simply put, $\hat{\theta} \neq \theta$
  2. random variation around the mean: u
- we have: $\mathrm{var}(\hat{e}^P) \;=\; \mathrm{var}(\theta + u - \hat{\theta}) \;=\; \mathrm{var}\,\hat{\theta} + \mathrm{var}\,u \;=\; \mathrm{var}\,\hat{\theta} + \sigma^2$
- therefore, a natural estimator of the standard deviation of prediction error is

$$\mathrm{se}(\hat{e}^P) = \sqrt{\left[\mathrm{se}(\hat{\theta})\right]^2 + \hat{\sigma}^2}$$

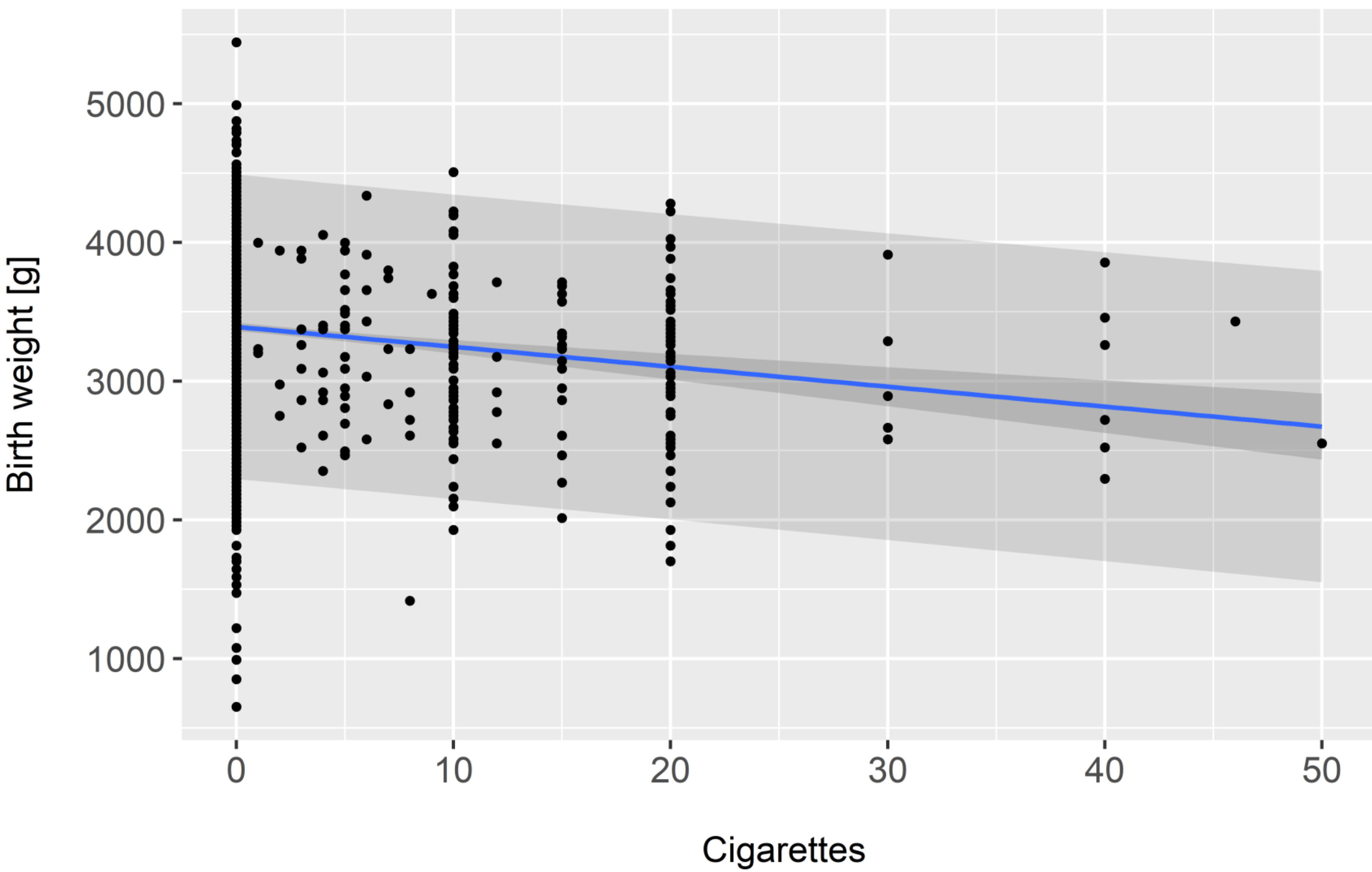
- and the 95% CI is $\hat{\theta} \pm c \cdot \mathrm{se}(\hat{e}^P)$

Prediction intervals

- obtaining the prediction interval for a pregnant woman who smokes 10 cigarettes a day:
  1. Create a new variable $A = cigs - 10$.
     Gretl: (Add → Define new variable → A = cigs – 10)
  2. Regress *bweight* on A. The intercept (constant) in this equation is θ, its std. error is $\mathrm{se}(\hat{\theta})$. The regression output will probably contain either $\hat{\sigma}^2$ or $\hat{\sigma}$. (In Gretl, $\hat{\sigma}$ is called S.E. of regression).
  3. Calculate $\mathrm{se}(\hat{e}^P) = \sqrt{\left[\mathrm{se}(\hat{\theta})\right]^2 + \hat{\sigma}^2}$.
  4. Calculate the 95% prediction interval as $\hat{\theta} \pm c \cdot \mathrm{se}(\hat{e}^P)$.



- with multiple regression models the procedure is analogous:
  - all explanatory variables need to be specified, say $x_1 = c_1$, …, $x_k = c_k$
  - we regress y on $(x_1 - c_1),\ldots,(x_k - c_k)$ in step 2, the rest is the same


Birth weight [g]
5000
4000
3000
2000
1000
0
10
20
30
40
50
Cigarettes

Predicting y when log(y) is the dependent variable

- consider the model

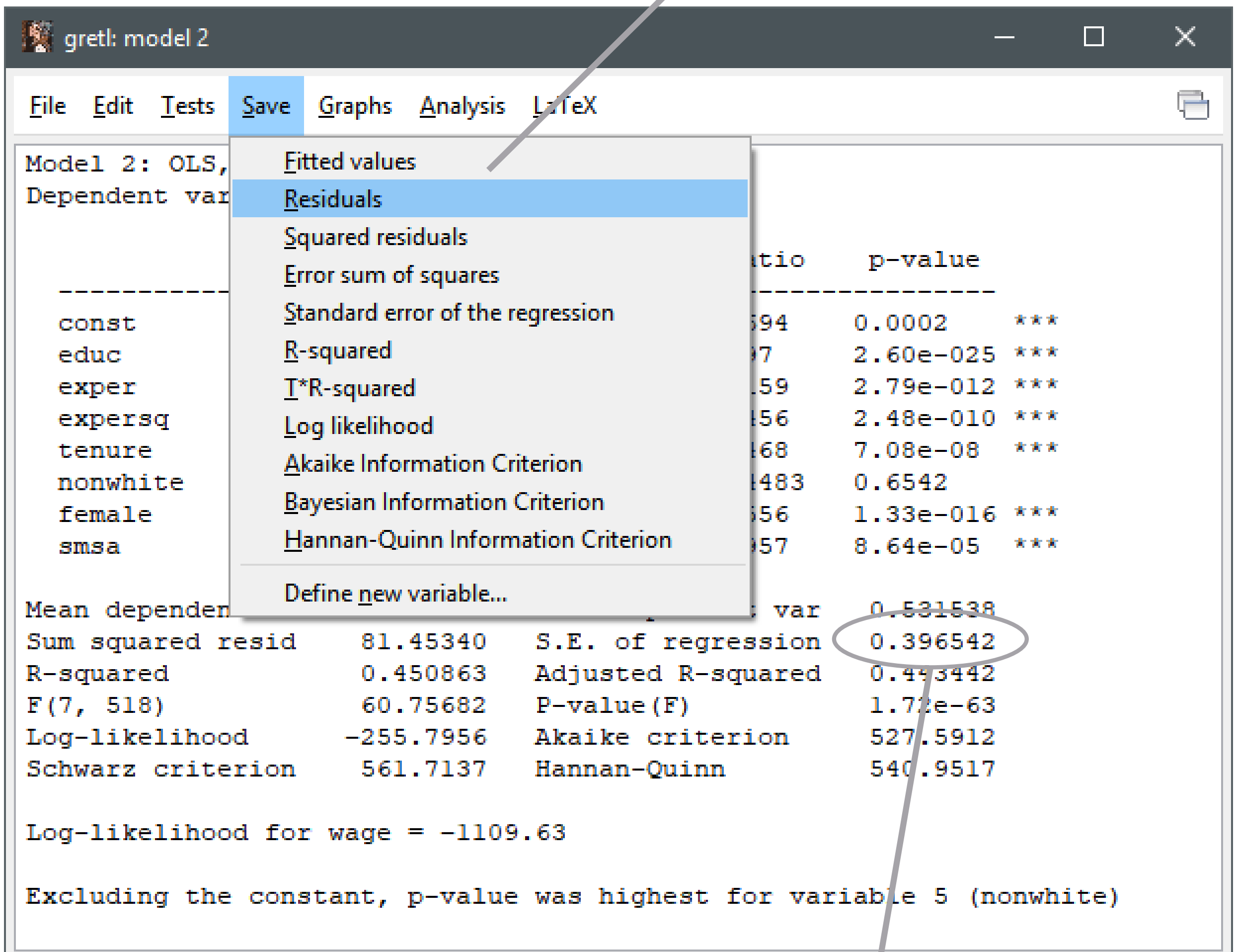
$$\log y = \beta_0 + \beta_1 x + u \tag{1}$$

- how do we predict, say, $\mathrm{E}[y \mid x = 10]$?
- the model implies that $y = e^{\beta_0 + \beta_1 x + u}$
- therefore:

$$\begin{aligned} \mathrm{E}[y \mid x] &= \mathrm{E}[e^{\beta_0 + \beta_1 x + u} \mid x] \\ &= \underbrace{e^{\beta_0 + \beta_1 x}}_{A} \cdot \underbrace{\mathrm{E}[e^{u} \mid x]}_{B} \end{aligned}$$

- A is consistently estimated as $e^{\hat{\beta}_0 + \hat{\beta}_1 x}$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are parameter estimates from (1) and x is replaced with the specified value
- B is more tricky; here we'll discuss two options:
  1. If we assume that u is normally distributed, then $B = \exp(\sigma^2 / 2)$
  2. Duan's (1983) estimator: estimate B as the sample mean of exponentiated residuals from (1), i.e. $B = \frac{1}{n} \sum_{i=1}^{n} \exp(\hat{u}_i)$

$$\hat{B} = \frac{1}{n}\sum_{i=1}^{n} \exp(\hat{u}_i) = 1.0807$$



$$\hat{B} = \exp(0.396542^2 / 2) = 1.0818$$

Lecture 8:
Predictions

Jan Zouhar

Introductory Econometrics