Simple Regression —Chapters 1 and 2 of Wooldridge's textbook

(Jing Li, Miami University)

Big Picture

In this lecture you will learn

1. Conditional Mean and Population Regression Function
2. Ordinary Least Squares (OLS)
3. Interpret intercept and slope coefficients of a simple regression
4. Standard error and t statistic
5. R-squared
6. Fitted value and residual

Causation vs Prediction

1. Eco 311 focuses on causation other than prediction. The main tool we learn is called regression, which shows how y and x are related.
2. y is quantity demanded and x is price. We want to know one percent increase in price causes how many percent decrease in quantity (elasticity)
3. y is GDP and x is tax. We want to estimate one dollar tax cut causes how many increase in GDP (multiplier)
4. y is air quality and x is the number of nuclear reactors in Germany. We want to show how shutting down nuclear power plants affects air quality.
5. y is death rate and x is vaccination rate. We want to test whether covid vaccine is effective.
6. Ideally, we need to hold everything else constant (ceteris paribus) in order to show causal effect of x on y. For instance, we hold constant income, preference, prices of other goods, etc when drawing a demand curve. Causal study is challenging since in reality ceteris paribus is hardly satisfied.

Simple Regression

1. A simple regression model is
$$y = \beta_0 + \beta_1 x + u \tag{1}$$
   (a) y is dependent variable, the one we want to explain or predict
   (b) x is independent variable (regressor), the one we use to explain or predict y
   (c) u is error term representing unobserved other factors that affect y
   (d) β_0 is intercept (constant term)
   (e) β_1 is slope coefficient
2. Error term introduces randomness. Without u, y can be perfectly predicted using x
3. A model is as good as its assumptions. One assumption here is constant marginal effect $\frac{dy}{dx} = \beta_1$. Later we will relax that assumption
4. Sample data for y and x are given, while β_0 and β_1 are unknown. Our goal is to use sample to estimate those two unknown parameters

Population Regression Function (PRF)

1. Let's make the second assumption of zero conditional mean for the error term

$$E(u|x) = 0 \qquad (\texttt{zero conditional mean assumption}) \tag{2}$$

   Basically some of unobserved factors have positive effects and some have negative effects. Assumption (2) says that the average effect is zero

2. Under assumption (2) we can derive the population regression function (PRF)

$$E(y|x) = \beta_0 + \beta_1 x \tag{3}$$

   PRF states that conditional mean of y changes in a linear way as x changes.

3. Remember both (2) and (3) are assumptions. They can fail in reality

Interpretation and T Test

1. Interpretation of β_0 and β_1 are based on

$$\beta_0 = E(y|x=0) \tag{4}$$

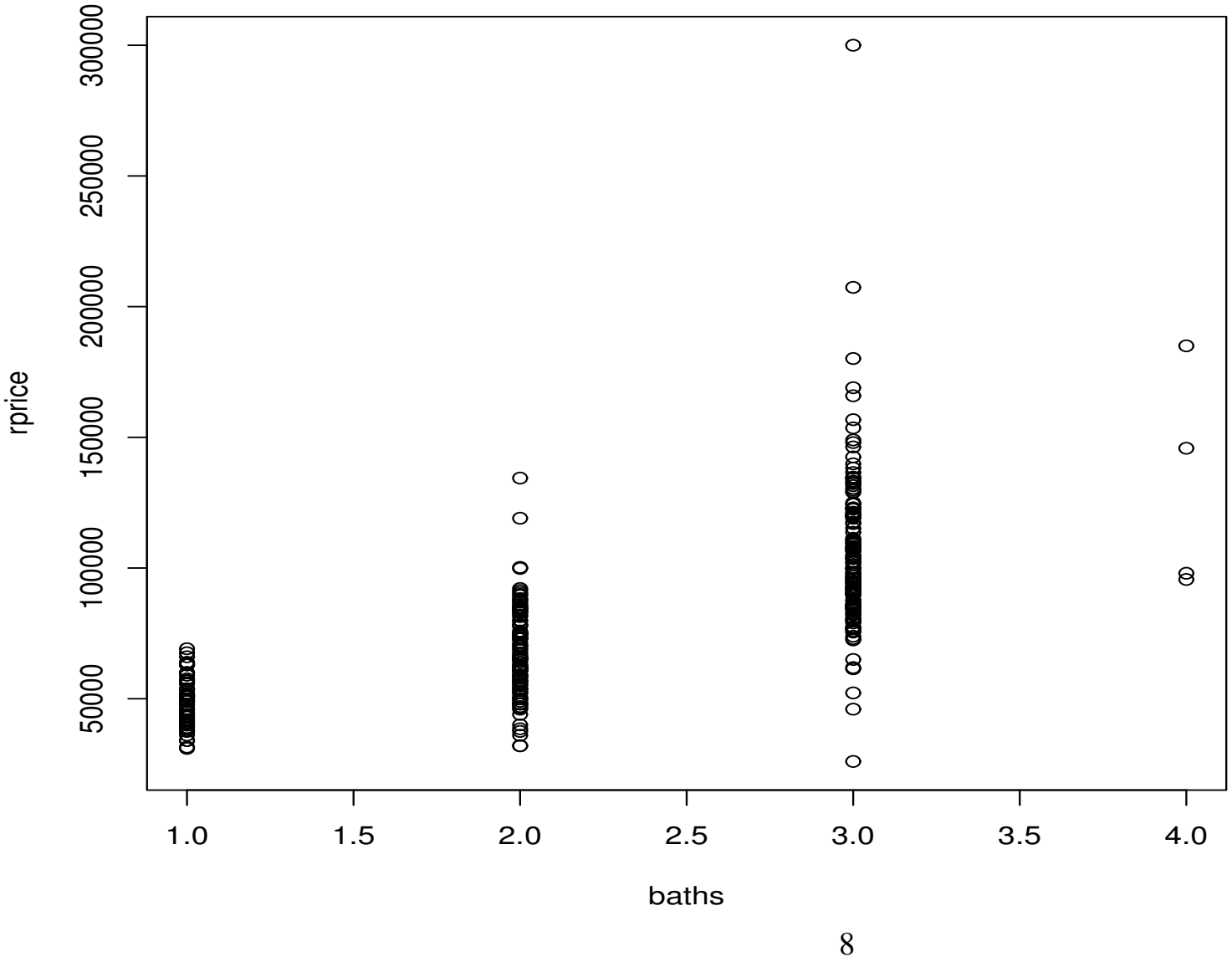
$$\beta_1 = \frac{dE(y|x)}{dx} \tag{5}$$

2. In words
   (a) β_0 is the average y when $x = 0$
   (b) Equation (5) implies that when $dx = 1$, $dE(y|x) = \beta_1$. Thus β_1 is the change in average y when x changes by one unit
   (c) By default, R function **lm** always tests $H_0 : \beta_1 = 0$. Under this null hypothesis, y is not responsive to x, so x does not matter.
   (d) x matters for y, or β_1 is statistically significant, when $H_0 : \beta_1 = 0$ is rejected.
   (e) For instance, if y is death rate and x is vaccination rate, $H_0 : \beta_1 = 0$ implies that the covid-19 vaccine is useless. We can show covid-19 vaccine is useful by rejecting H_0.
   (f) $H_0 : \beta_1 = 0$ is rejected when (i) absolute t value of β_1 is greater than 1.96; (2) p value is less than 0.05; (iii) 0 is outside confidence interval for β_1

Example 1

```
> ad = "https://www.fsb.miamioh.edu/lij14/400_house.txt"
> data = read.table(url(ad), header=T)
> head(data)
  age area baths rprice
1  48 1660     1  60000
2  83 2612     2  40000
3  58 1144     1  34000
4  11 1136     1  63900
5  48 1868     1  44000
6  78 1780     3  46000
> attach(data)
> plot(baths, rprice)
```

Scatter Plot



Remarks

1. We use **read.table** function to read data in txt format
2. We use House data to illustrate the simple regression, and we are interested in how *baths* (number of bathrooms) affects *rprice* (the real price of a house).
3. We use function **plot** to draw a scatter plot in which each point represents a house. The x-coordinate is baths; the y-coordinate is rprice
4. Thanks to other factors, houses with the same number of bathrooms can have different prices. For instance, there are a variety of prices for houses with 2 bathrooms
5. Using statistics jargon, there is a conditional distribution of rprice given baths=2. The center of that conditional distribution is conditional mean $E(rprice|baths = 2)$
6. Regressor x is also called conditioning variable

Conditional Means

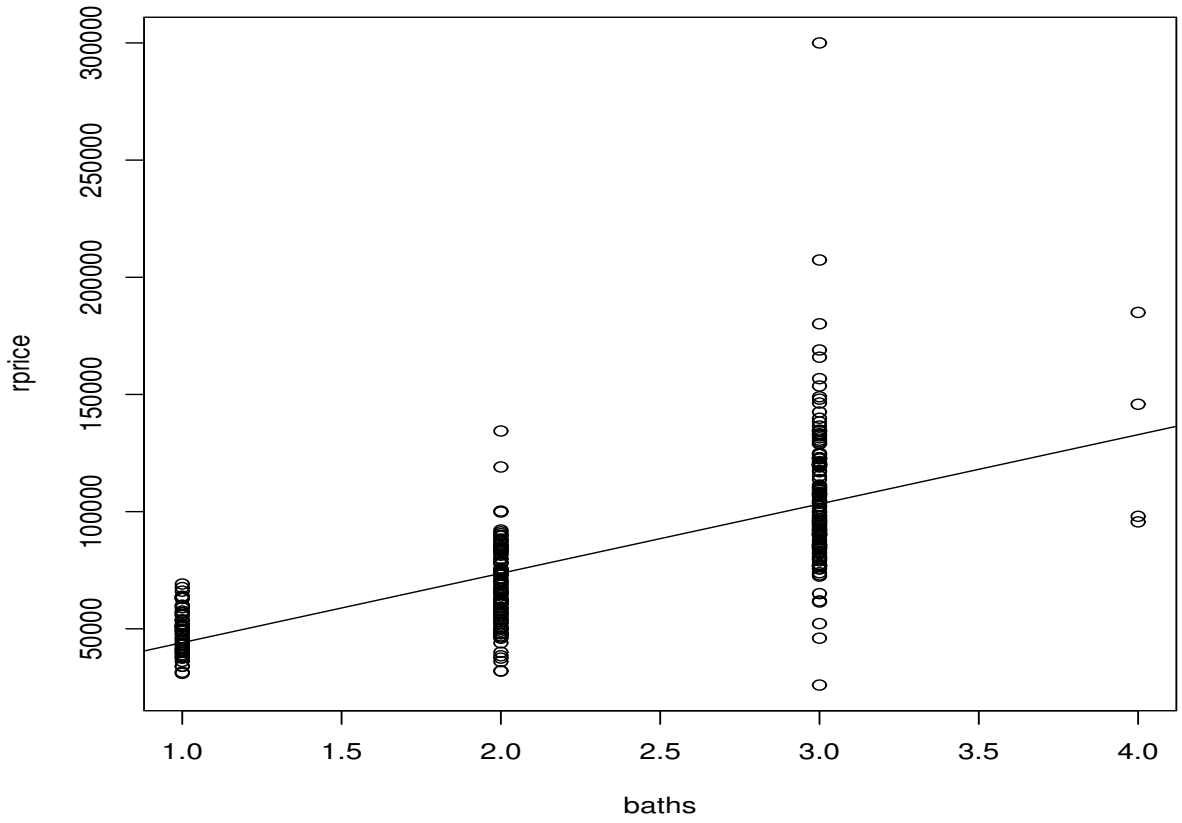
```
> mean(rprice[baths==1])
[1] 49817.64
> mean(rprice[baths==2])
[1] 67670.65
> mean(rprice[baths==3])
[1] 105365.7
```

Remarks

1. We use **mean** function to obtain three conditional means
2. There is only one unconditional mean, but multiple conditional means depending on how many values x can take
3. The first conditional mean is $E(rprice|baths = 1) = 49817.64$. It can be interpreted as the average price for houses with one bathroom
4. The second and third conditional means are $E(rprice|baths = 2) = 67670.65$ and $E(rprice|baths = 3) = 105365.7$
5. It seems that as baths rises, conditional mean rises as well, which supports (3)
6. Exercise: how to use R to find unconditional mean $E(rprice)$?

Adding OLS Fitted Line

```
> plot(baths, rprice)
> abline(lm(rprice ~ baths))
```



Remarks

1. We hope to find an optimal straight line given by $\hat{\beta}_0 + \hat{\beta}_1 x$ that fits data best
2. We use a method called ordinary least squares to find $\hat{\beta}_0$ and $\hat{\beta}_1$, and then draw that optimal fitted line
3. The straight line shown in the scatter plot represents the estimated PRF
4. The PRF line seems to pass centers of conditional distributions
5. The PRF line is upward-sloping, implying that on average a house becomes more expensive as the number of bathrooms rises

Ordinary Least Squares (OLS)

1. OLS is a method we use to estimate β_0 and β_1 (or a method for line-fitting). There are other methods such as LAD (optional reading)
2. It is incorrect to say OLS is a model. Equation (1) is a model. OLS is the method we use to estimate that model
3. OLS solves an optimization problem of minimizing residual sum squares (RSS)

$$\min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} \hat{u}_i^2 = \min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{6}$$

$$\hat{u}_i = y_i - \hat{y}_i \tag{7}$$

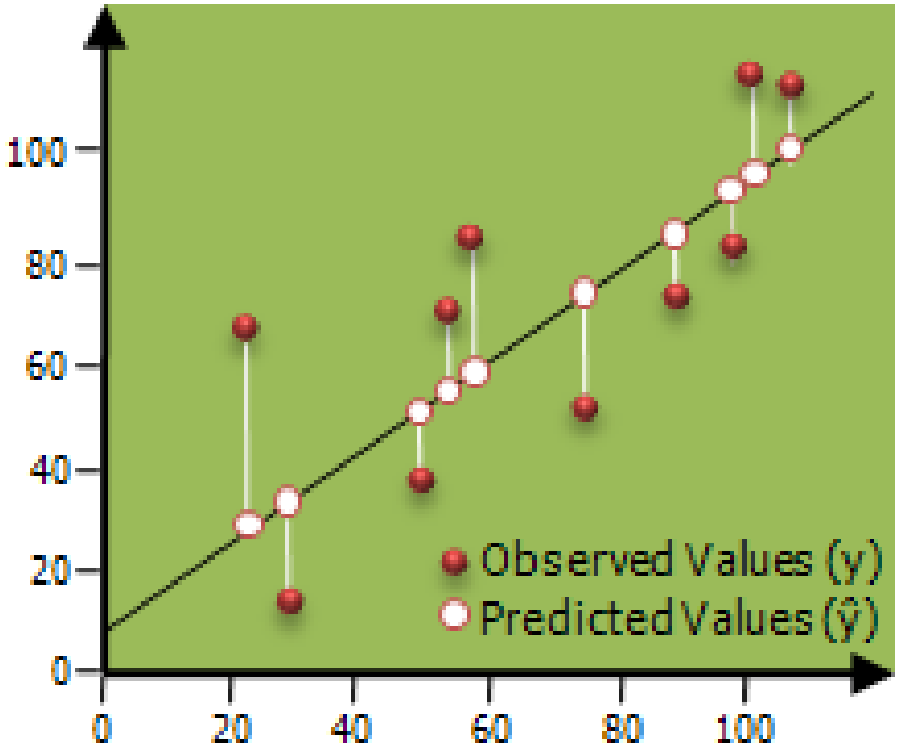
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{8}$$

   (a) $\hat{\beta}_0$ is estimated intercept; $\hat{\beta}_1$ is estimated slope

   (b) $\hat{y}$ is fitted value or predicted value

   (c) $\hat{u}$ is residual (prediction error)—the difference between true value y and predicted value $\hat{y}$, see the next graph

Ordinary Least Squares (OLS)—Line-Fitting



Ordinary Least Squares (OLS)—continued

1. Basically OLS finds $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing total squared prediction errors
2. Prediction errors are squared to avoid cancellation
3. The red dots are actual data (observed values). White dots are on the fitted line, so they are predicted values.
4. The vertical gaps between red and white dots are residuals (prediction errors). OLS minimizes total squared prediction errors

Ordinary Least Squares (OLS)—continued

1. Solving problem (6) entails taking (partial) derivative of RSS with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$, and setting them to zero (called first order condition FOC)

$$\frac{\partial RSS}{\partial \hat{\beta}_0} = 0 \Rightarrow \sum \hat{u}_i = 0 \quad (\texttt{FOC I}) \tag{9}$$

$$\frac{\partial RSS}{\partial \hat{\beta}_1} = 0 \Rightarrow \sum \hat{u}_i x_i = 0 \quad (\texttt{FOC II}) \tag{10}$$

2. The solutions of solving this system of two equations are the estimated coefficients:

$$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{S_{x,y}}{S_x^2} \tag{11}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{12}$$

3. From (11) it is clear that the sign of $\hat{\beta}_1$ depends only on the sample covariance. $\hat{\beta}_1$ is positive (negative) if x and y are positively (negatively) related.

4. Exercise: prove that $\hat{\beta}_1 = \rho \frac{S_y}{S_x}$

Intuition

1. FOC (9) implies that average prediction error is zero when OLS is used

$$\frac{\sum \hat{u}_i}{n} = \bar{\hat{u}} = 0 \tag{13}$$

   In other words, the OLS fitted line passes the center of data

2. FOC (10) implies that prediction error is uncorrelated with x when OLS is used

$$\frac{\sum \hat{u}_i (x_i - \bar{x})}{n-1} = cov(\hat{u}, x) = 0 \tag{14}$$

   In other words, the OLS uses x efficiently so that no information about x is left in the prediction error

Standard Error

1. After obtaining $\hat{\beta}_0$ and $\hat{\beta}_1$, we compute residual using (7) and (8) for each observation
2. The standard deviation of the error term is called residual standard error or standard error of regression (denoted by $\hat{\sigma}$), and is estimated as

$$\hat{\sigma} = \sqrt{\frac{\sum \hat{u}_i^2}{n-k-1}} \tag{15}$$

   where k is the number of regressors. For a simple regression $k = 1$.

3. Then we can show the standard error of $\hat{\beta}_1$, which measures sampling uncertainty, is

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}} = \frac{\hat{\sigma}}{\sqrt{(n-1)S_x^2}} \tag{16}$$

   The last equality holds because $S_x^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$. Alternatively, $se(\hat{\beta}_1)$ can be obtained via bootstrap (optional reading)

4. Exercise: How to reduce $se(\hat{\beta}_1)$?

T Test

1. The central limit theorem implies that the t statistic follows standard normal distribution

$$t - statistic_{\hat{\beta}_1} = panda = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim N(0,1) \tag{17}$$

   By default **lm** function tests $H_0 : \beta_1 = 0$ (x does not matter). We reject the null hypothesis if t statistic exceeds 1.96 in absolute value

2. In order to show x matters, we hope to increase the t statistic or decrease the standard error of $\hat{\beta}_1$. To do so, according to (16), we can (i) increase sample size n; (2) increase the variation in x

3. R also reports the p-value. H_0 is rejected if the p-value is less than 0.05

4. A third approach is, H_0 is rejected when 0 is outside confidence intervals

R-squared

1. A regression effectively decomposes y into an explained part (fitted value) and an unexplained part (residual)

$$y_i = \hat{y}_i + \hat{u}_i \tag{18}$$

2. We want to know to what degree the model can explain y. Toward that end we define total sum squares (TSS) and residual sum squares (RSS) as

$$TSS \equiv \sum (y_i - \bar{y})^2, \quad RSS \equiv \sum \hat{u}_i^2 \tag{19}$$

3. Now we can compute a measurement of overall fit called R-squared

$$R^2 \equiv 1 - \frac{RSS}{TSS} \tag{20}$$

which satisfies the inequality

$$0 \leq R^2 \leq 1 \tag{21}$$

4. R-squared measures how much variation of y can be explained by the model. A model with R^2 close to 1 does a better job than a model with R^2 close to 0

Example 1—continued

```
> summary(lm(rprice~baths))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    14511       4300   3.375  0.00083 ***
baths          29583       1746  16.944  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24060 on 319 degrees of freedom
Multiple R-squared:  0.4737,    Adjusted R-squared:  0.472
F-statistic: 287.1 on 1 and 319 DF,  p-value: < 2.2e-16

> beta1hat = cov(rprice, baths)/var(baths)
> beta1hat
[1] 29582.67
```

```
> beta0hat = mean(rprice)-beta1hat*mean(baths)
> beta0hat
[1] 14510.8

> uhat = rprice-beta0hat-beta1hat*baths
> RSE = sqrt(sum(uhat^2)/(length(rprice)-2))
> RSE
[1] 24064.22
>
> R_squared = 1 - sum(uhat^2)/sum((rprice-mean(rprice))^2)
> R_squared
[1] 0.4736976

> confint(lm(rprice~baths))
                2.5 %   97.5 %
(Intercept)  6051.42 22970.18
baths       26147.82 33017.53
```

Remarks

1. Function **lm** regresses rprice onto baths (use baths to predict rprice)
2. R-squared = 0.4737, so the model explains 47 percent of variation of rprice
3. $\hat{\beta}_1 = 29583$—the interpretation is that when the number of bathrooms increases by 1, that is associated with average rprice rising by 29583
4. T statistic for $\hat{\beta}_1$ is $16.944 > 1.96$, so we reject the default null hypothesis that baths does not matter. In this case, baths is significantly and positively correlated with rprice
5. P value being less than 0.05, and 0 being outside the confidence interval (26147.82, 33017.53) lead to the same conclusion of rejecting $H_0 : \beta_1 = 0$
6. $\hat{\beta}_0 = 14511$—the interpretation is that the average rprice of a house with 0 bathroom is 14511
7. The straight line that represents $14511 + 29583baths$ is shown in the scatter plot on page 12, see formula (8)

Example 1—continued

```
> my_mod=lm(rprice~baths)
> yhat = fitted(my_mod)
> head(yhat)
1         2         3         4         5         6
44093.47  73676.15  44093.47  44093.47  44093.47 103258.82

> uhat = resid(my_mod)
> head(uhat)
1            2             3            4             5             6
15906.52557 -33676.14813 -10093.47443  19806.52557    -93.47443 -57258.821
> mean(uhat)
[1] 4.487376e-12
```

Remarks

1. Function **fitted** computes the fitted value for each house using formula (8). For instance, for a house with one bathroom, the predicted price is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 14510.8 + 29582.67(1) = 44093.47$$

2. Function **resid** computes the residual (prediction error) for each house using formula (7). For instance, for a house with one bathroom and actual price of 60000, the residual is

$$\hat{u} = y - \hat{y} = 60000 - 44093.47 = 15906.53$$

3. Note that there is a residual and a fitted value for each observation

4. We can verify FOC (9), which implies average residual is 0 (or very close to zero due to rounding error)

5. Exercise: how to use R to verify FOC (10)?

Mini Project

Please use GaltonFamilies data in the HistData package.

1. Run a simple regression that predicts childHeight using father's height
2. Interpret the slope coefficient
3. What does the t value imply?
4. Predict childHeight when father's height is 70
5. Is it a better idea to predict childHeight using mother's height? Why?
6. What is the serious drawback of the simple regression?