Simple Linear Regression II

Econ UA 266: Intro to Econometrics

Recap

- Estimation of the simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Interpretation of β_0 and β_1
- Derivation of OLS estimators:

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n \quad \text{and} \quad \hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$$

- Interpretation of $\hat{\beta}_0$ and $\hat{\beta}_1$
- R^2 and standard error of the regression
- Assumptions of the linear regression model
- Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$

Consistency of OLS

- At the end of last class we showed:

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2} \rightarrow_p \frac{\sigma_{XY}}{\sigma_X^2}$$

- Using our rules for covariance:

$$\begin{aligned}\sigma_{XY} = \text{Cov}(X, Y) &= \text{Cov}(X, \beta_0 + \beta_1 X + u) \\ &= \beta_1 \text{Cov}(X, X) + \text{Cov}(X, u) \\ &= \beta_1 \sigma_X^2 + \text{Cov}(X, u)\end{aligned}$$

hence

$$\frac{\sigma_{XY}}{\sigma_X^2} = \beta_1 + \frac{\text{Cov}(X, u)}{\sigma_X^2}$$

- Therefore, $\hat{\beta}_1$ is a consistent estimator of β_1 if $\text{Cov}(X, u) = 0$

Consistency of OLS

- How do we show $\mathrm{Cov}(X, u) = \mathrm{E}[Xu] - \mathrm{E}[X]\mathrm{E}[u] = 0$?
- recall Assumption 1:
  The conditional distribution of u_i given X_i has mean zero ($\mathrm{E}[u_i|X_i] = 0$)

- Law of iterated expectations: the expected value of a random variable Z is the expected value of its conditional expectation

$$\mathrm{E}[Z_i] = \mathrm{E}[\mathrm{E}[Z_i|X_i]]$$

- therefore: $\mathrm{E}[X_i u_i] = \mathrm{E}[\mathrm{E}[X_i u_i|X_i]] = \mathrm{E}[X_i \mathrm{E}[u_i|X_i]] = 0$
- and: $\mathrm{E}[u_i] = \mathrm{E}[\mathrm{E}[u_i|X_i]] = 0$
- so: $\mathrm{Cov}(X, u) = 0$ if Assumption 1 holds

Consistency of OLS

- Therefore we have shown that:

$$\hat{\beta}_1 \to_p \beta_1$$

under our Assumptions

- Similarly for $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n \to_p \mathrm{E}[Y] - \beta_1 \mathrm{E}[X]$$

- where: $\mathrm{E}[Y] = \mathrm{E}[\beta_0 + \beta_1 X + u] = \beta_0 + \beta_1 \mathrm{E}[X]$
- hence: $\mathrm{E}[Y] - \beta_1 \mathrm{E}[X] = \beta_0$
- and so:

$$\hat{\beta}_0 \to_p \beta_0$$

Estimation Uncertainty

Our estimates depend on the sample at hand. Another sample might have produced different estimates. How do we deal with that?

- Test of hypotheses
- Confidence intervals
- We'll use the fact that the sampling distributions of

$$\frac{\hat{\beta}_0 - \beta_0}{s.e.(\hat{\beta}_0)} \quad \text{and} \quad \frac{\hat{\beta}_1 - \beta_1}{s.e.(\hat{\beta}_1)}$$

are approximately $\mathrm{N}(0,1)$ when n is large

Class size and test score

Last week we estimated the linear regression model:

$$testscr_i = \beta_0 + \beta_1 str_i + u_i$$

from data on 420 California school districts

The sample regression function is

$$\widehat{testscr} = 698.93 - 2.28 str$$

which points to a negative r'ship between class size and test score

Question: how do we know this negative relationship is statistically significant?

Class size and test score

Formally, we want to test the hypotheses:

$H_0 : \beta_1 = 0$ (i.e. no relationship between CS and avg TS)

$H_1 : \beta_1 < 0$ (i.e. negative r'ship between CS and avg TS)

We do so in three steps:

1. Compute the standard error of $\hat{\beta}_1$
2. Compute the t-statistic

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{s.e.(\hat{\beta}_1)}$$

   where $\beta_{1,0}$ is the hypothesized value of β_1 under H_0 (here $\beta_{1,0} = 0$)
3. Compute the p-value: smallest significance level at which the null hypothesis could be rejected, based on the t-statistic.
   Reject H_0 at the 5% level if p-value < 0.05

Idea of p-value

Thought experiment: suppose H_0 was true. What's the prob. we'd observe a test statistic at least as extreme as the one we observe?

This probability is the p-value

- if p-value is small: reject H_0
- if p-value is large: do not reject H_0

We know that

$$\frac{\hat{\beta}_1 - \beta_{1,0}}{s.e.(\hat{\beta}_1)}$$

is approx. distributed as $N(0,1)$ under H_0 (the approx. becomes "better" as n gets larger)

We can use this approximation to calculate the p-value.

Idea of p-value

Because it's a left-tailed test, more extreme means more negative:

$$\begin{aligned} p\text{-value} &= P_{H_0}\left(\frac{\hat{\beta}_1 - \beta_{1,0}}{s.e.(\hat{\beta}_1)} < t^{act}\right) \\ &\approx P\left(Z < t^{act}\right) \\ &= \Phi(t^{act}) \end{aligned}$$

where

- t^{act} is the actual value of the test statistic
- Φ is standard normal cdf

Regression summary

```
lm(formula = testscr ~ str, data = ca)

Residuals:
    Min      1Q  Median      3Q     Max
-47.727 -14.251   0.483  12.822  48.540

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
str          -2.2798     0.4798  -4.751 2.78e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.58 on 418 degrees of freedom
Multiple R-squared:  0.05124,Adjusted R-squared:  0.04897
F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

Class size and test score

1. Compute the standard error of $\hat{\beta}_1$:

$$s.e.(\hat{\beta}_1) = 0.4798$$

2. Compute the t-statistic

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{s.e.(\hat{\beta}_1)} = \frac{-2.2798 - 0}{0.4798} = -4.751$$

3. Compute the p-value: (this is a left-tail test: "more extreme" here means more negative)

```
> pnorm(-4.751)
[1] 1.012066e-06
```

Therefore, we reject H_0 at 5% level (and 1% level)

Conclusion: there is strong statistical evidence for a negative relationship between class size and average test scores

Class size and test score

We could also test whether there is any significant relationship between class size and average test scores

$H_0 : \beta_1 = 0$ (i.e. no relationship between CS and avg TS)

$H_1 : \beta_1 \neq 0$ (i.e. a relationship between CS and avg TS)

1. Compute the standard error of $\hat{\beta}_1$: $s.e.(\hat{\beta}_1) = 0.4798$
2. Compute the t-statistic $t = -4.751$
3. Compute the p-value: (note this is a two-tailed test, so "more extreme" means larger in absolute value)

```
> 2*pnorm(-abs(-4.751))
[1] 2.024131e-06
```

Therefore, we reject H_0 at 5% level (and 1% level)

Conclusion: there is strong statistical evidence for a relationship between class size and average test scores

Class size and test score

We could also test whether class size is positively associated with average test scores

$H_0 : \beta_1 = 0$ (i.e. no relationship between CS and avg TS)

$H_1 : \beta_1 > 0$ (i.e. positive r'ship between CS and avg TS)

1. Compute the standard error of $\hat{\beta}_1$: $s.e.(\hat{\beta}_1) = 0.4798$
2. Compute the t-statistic $t = -4.751$
3. Compute the p-value: (note this is a right-tailed test, so "more extreme" here means more positive)

```
> 1-pnorm(-4.751)
[1] 0.999999
```

Therefore, we do not reject H_0 at any conventional level of signif.

Conclusion: there is no statistical evidence for a positive r'ship between class size and average test scores

Testing hypotheses: summary

1. Compute the standard error: $s.e.(\hat{\beta}_i)$
2. Compute the t-statistic

$$t = \frac{\hat{\beta}_i - \beta_{i,0}}{s.e.(\hat{\beta}_i)}$$

3. Compute the p-value:

$\Phi(t)$	for left-tailed test: $H_1 : \beta_i < \beta_{i,0}$
$2\Phi(-\lvert t \rvert)$	for two-tailed test: $H_1 : \beta_i \neq \beta_{i,0}$
$1 - \Phi(t)$	for right-tailed test: $H_1 : \beta_i > \beta_{i,0}$

and reject H_0 if p-value $<$ level of significance; then conclude.

R and p-values

The p-values in R output are:

- always for $H_0 : \beta_i = 0$ against $H_0 : \beta_i \neq 0$
- computed using

$$\frac{\hat{\beta}_i - \beta_{i,0}}{s.e.(\hat{\beta}_i)} \sim t_{n-2}$$

  which is only valid if all the u_i are Normally distributed
- real-world data are not Normally distributed

We instead use the approximation

$$\frac{\hat{\beta}_i - \beta_{i,0}}{s.e.(\hat{\beta}_i)} \approx N(0,1)$$

Difference goes away for large n, as $t_{n-2} \to N(0,1)$ as $n \to \infty$

Confidence intervals

A 95% confidence interval for β_i is the set of values $\beta_{i,0}$ that cannot be rejected using a two-sided test of $H_0 : \beta_i = \beta_{i,0}$

$$95\% \text{ CI for } \beta_i = [\hat{\beta}_i - 1.96 \times s.e.(\hat{\beta}_i), \hat{\beta}_i + 1.96 \times s.e.(\hat{\beta}_i)]$$

If 95% CI doesn't contain zero, then we'd reject $H_0 : \beta_i = 0$ against $H_1 : \beta_i \neq 0$ at the 5% level

Class size and test score

We can easily calculate the CIs directly:

```
> 698.93295-1.96*9.46749
[1] 680.3767
> 698.93295+1.96*9.46749
[1] 717.4892

> -2.27981-1.96*0.47983
[1] -3.220277
> -2.27981+1.96*0.47983
[1] -1.339343
```

Confidence intervals

We can also construct CIs for predicted effects: the change in the expected value of Y as X varies by Δx

A 95% confidence interval for the predicted effect is:

$$95\%\ \text{CI} = [(\hat{\beta}_1 - 1.96 \times s.e.(\hat{\beta}_1))\Delta x, (\hat{\beta}_1 + 1.96 \times s.e.(\hat{\beta}_1))\Delta x]$$

(rearrange the limits if necessary)

95% CI for β_1 is $[-3.22, -1.34]$

95% CI for predicted effect of decrease of class size by 3 is

$$[-1.34 \times -3, -3.22 \times -3] = [4.01, 9.67]$$

i.e. we'd expect, with confidence, that average test scores would be between 4.01 and 9.67 points higher

Heteroskedasticity and homoskedasticity

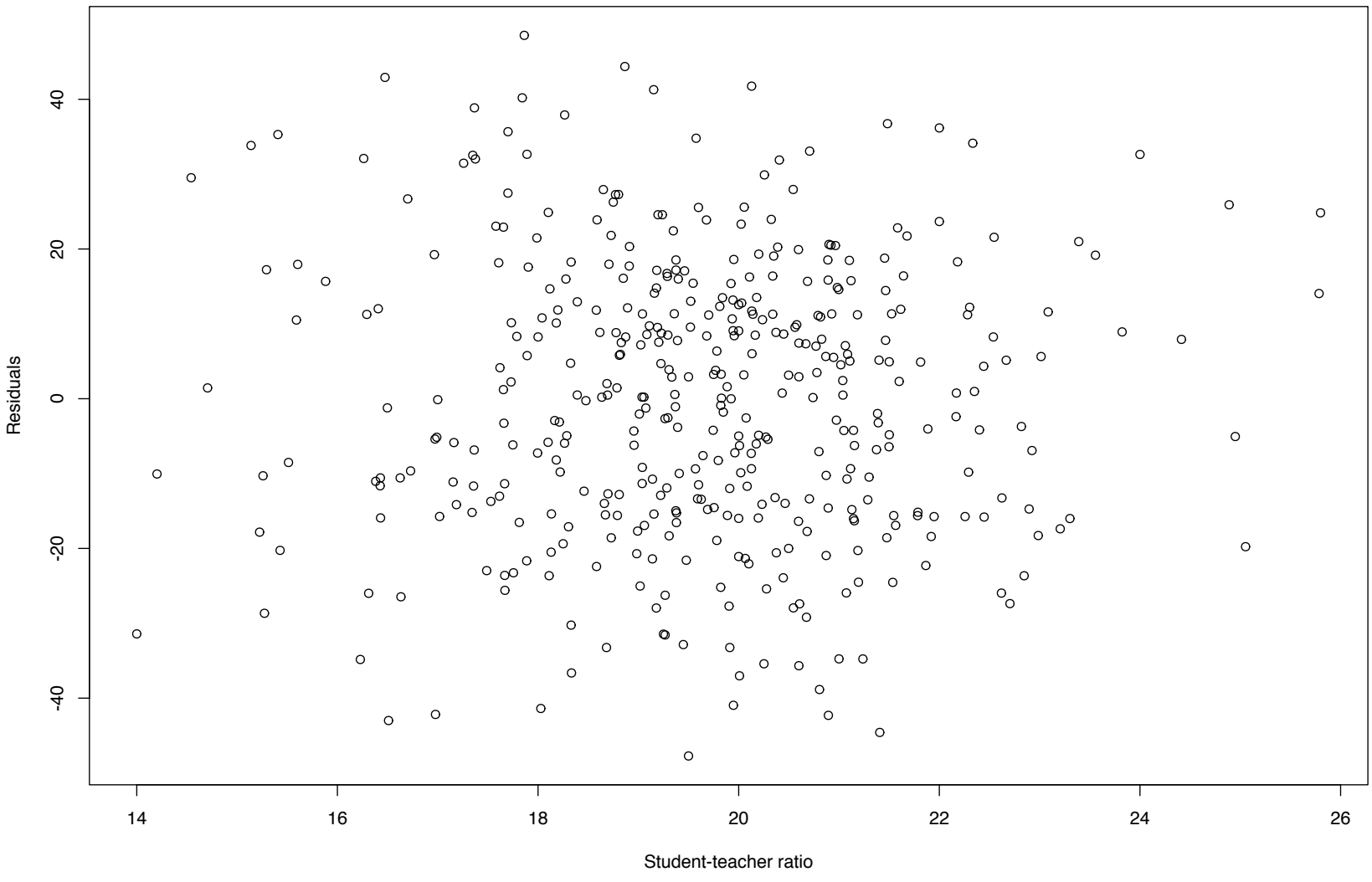
We say the error term u_i is:

- homoskedastic if $\text{Var}(u_i|X_i)$ is constant (i.e. doesn't depend on X_i)
- otherwise, we say it is heteroskedastic

Consequences of heteroskedasticity:

- OLS estimators are unbiased, consistent, asymptotically normally distributed (just like before)
- by default, software such as R reports conventional homoskedasticity-only standard errors
- these are the wrong standard errors under heteroskasticity
- therefore, our t-tests and confidence intervals are wrong unless we use the correct standard errors

Heteroskedasticity and homoskedasticity



Heteroskedasticity-robust standard errors

If we suspect the errors are heteroskedastic, then we must use heteroskedasticity-robust standard errors (aka robust std errors)

- heteroskedasticity-robust std errors remain valid even if the errors are actually homoskedastic

Rule of thumb: always use heteroskedasticity-robust std errors

In R you can use the command `coeftest` with `vcovHC` specified to obtain the correct output

```
coeftest(z,df=Inf,vcov=vcovHC(z,type="HC1"))
```

you'll need to install and load the packages `lmtest` and `sandwich` to be able to use these commands

Class size and test score with correct std errors

```
> install.packages("lmtest")
> install.packages("sandwich")
> require(lmtest)
> require(sandwich)
> coeftest(z,df=Inf,vcov=vcovHC(z,type="HC1"))

z test of coefficients:

             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 698.93295   10.36436 67.4362 < 2.2e-16 ***
str          -2.27981    0.51949 -4.3886 1.141e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Class size and test score with correct std errors

	Heterosk.-robust	Homosk.-only
$s.e.(\hat{\beta}_0)$	10.36	9.47
$s.e.(\hat{\beta}_1)$	0.52	0.48
95% CI for β_1	$[-3.30, -1.26]$	$[-3.22, -1.34]$
p-value $H_1 : \beta_1 < 0$	5.70×10^{-6}	1.01×10^{-6}
p-value $H_1 : \beta_1 \neq 0$	1.14×10^{-5}	2.02×10^{-6}

Summary so far

You have data on Y and X and want to estimate the model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

1. Estimate β_0 and β_1 using OLS. Report the estimates:

$$\widehat{Y}_i = \underset{s.e.(\hat{\beta}_0)}{\hat{\beta}_0} + \underset{s.e.(\hat{\beta}_1)}{\hat{\beta}_1} \times X_i \qquad R^2 = \qquad s_{\hat{u}} =$$

   (use heteroskedasticity-robust standard errors)

2. Hypothesis tests for β_0 and β_1:
   - Calculate t statistic $t = \frac{\hat{\beta}_i - \beta_{i,0}}{s.e.(\hat{\beta}_i)}$
   - Calculate p value from N(0,1) distribution

3. Confidence intervals for β_0 and β_1: $\hat{\beta}_i \pm z_{crit} \times s.e.(\hat{\beta}_i)$
   where $z_{crit} = 1.645$ for 90% CI and 1.96 for 95% CI

Dummy variables

Linear regression coefficients have a special interpretation when X is a dummy variables (0 or 1)

`Star.csv` contains data from the Tennessee schooling experiment

In week 1, we regressed average test scores ($avgscore$) on a dummy variable for being assigned to a small class ($small$)

The model is:

$$avgscore_i = \beta_0 + \beta_1 small_i + u_i$$

which implies (by Assumption 1)

$$\mathrm{E}[avgscore_i | small_i = 0] = \beta_0$$
$$\mathrm{E}[avgscore_i | small_i = 1] = \beta_0 + \beta_1$$

so β_1 is the difference in means between the two subpopulations

Tennessee experiment

```
# import the data
tn <- read.csv("Star.csv")
# make average test score variable
tn$avgscore <- (tn$tmathssk+tn$treadssk)/2
# make small dummy variable
tn$small <- (tn$classk=='small.class')*1
# make aide dummy variable
tn$aide <- (tn$classk=='regular.with.aide')*1
# regress test score on class size dummy
z <- lm(avgscore~small,data=subset(tn,aide==0))
# print output
summary(z)
# print output with robust standard errors
coeftest(z,df=Inf,vcov=vcovHC(z,type="HC1"))
```

Tennessee experiment

```
Call:
lm(formula = avgscore ~ small, data = subset(tn, aide == 0))

Residuals:
     Min       1Q   Median       3Q      Max
-141.471  -26.525   -3.471   22.529  160.475

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 458.9710     0.8351 549.615  < 2e-16 ***
small         7.0544     1.2256   5.756 9.32e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 37.35 on 3731 degrees of freedom
Multiple R-squared:  0.008801,Adjusted R-squared:  0.008536
F-statistic: 33.13 on 1 and 3731 DF,  p-value: 9.32e-09
```

Tennessee experiment

```
z test of coefficients:

             Estimate Std. Error  z value  Pr(>|z|)
(Intercept) 458.97100    0.81789 561.1617 < 2.2e-16 ***
small         7.05439    1.22946   5.7378 9.592e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tennessee experiment

So, we'd report the results as:

$$\widehat{avgscore} = \underset{(0.818)}{458.97} + \underset{(1.229)}{7.054} \times small_i \quad R^2 = 0.009 \quad s_{\hat{u}} = 37.35$$

A test of $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 > 0$ has p-value

```
> 1-pnorm(5.7378)
[1] 4.795714e-09
```

so we'd reject H_0 and conclude smaller class sizes have a positive effect on test scores. Equivalently, average test scores in small classes are significantly higher than those in large classes

Note our interpretation here is causal because students were randomly assigned to small and large classes

Dummy variables

In fact, with dummy variables one can show that the OLS estimates are simply:

$$\hat{\beta}_0 = (\text{sample mean of } Y_i \text{ for those with } D_i = 0)$$
$$\hat{\beta}_1 = (\text{sample mean of } Y_i \text{ for those with } D_i = 1) - \hat{\beta}_0$$

```
> mean(tn$avgscore[tn$aide==0&tn$small==0])
[1] 458.971
> mean(tn$avgscore[tn$aide==0&tn$small==1])-
    mean(tn$avgscore[tn$aide==0&tn$small==0])
[1] 7.054389
```

Earnings data

The data file `incomedata.Rda` contains data on 1192 individuals in the labor force in the 1990s (survey by U. of Michigan):

- `EARN` is annual earnings
- `ED` is number of years of education
- `GEN` is a dummy variable (1 for male, 0 for female)
- `HEIGHT` is height in inches
- and other demographic variables

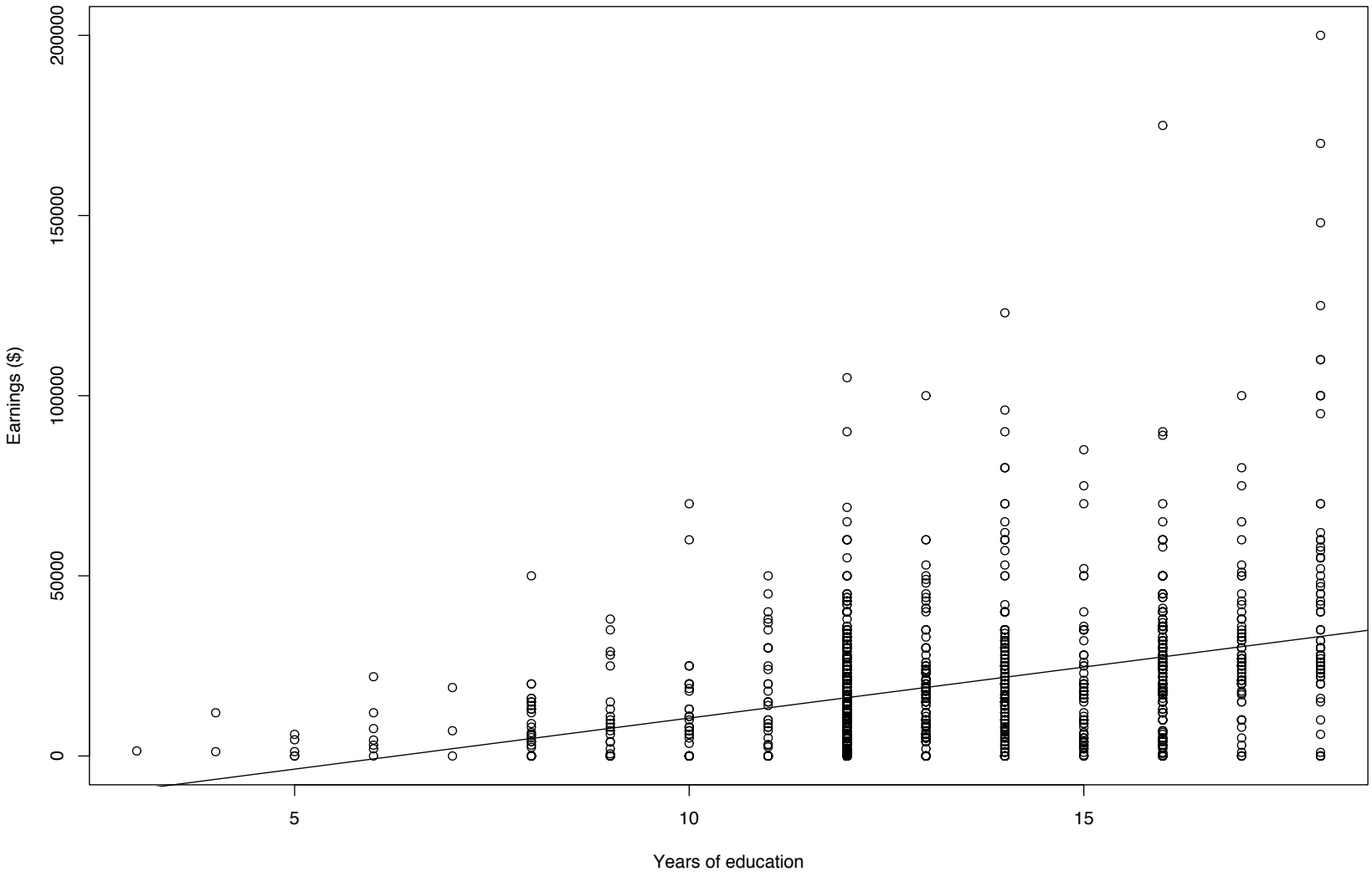
Let's start by estimating

$$\text{EARN}_i = \beta_0 + \beta_1 \text{ED}_i + u_i$$

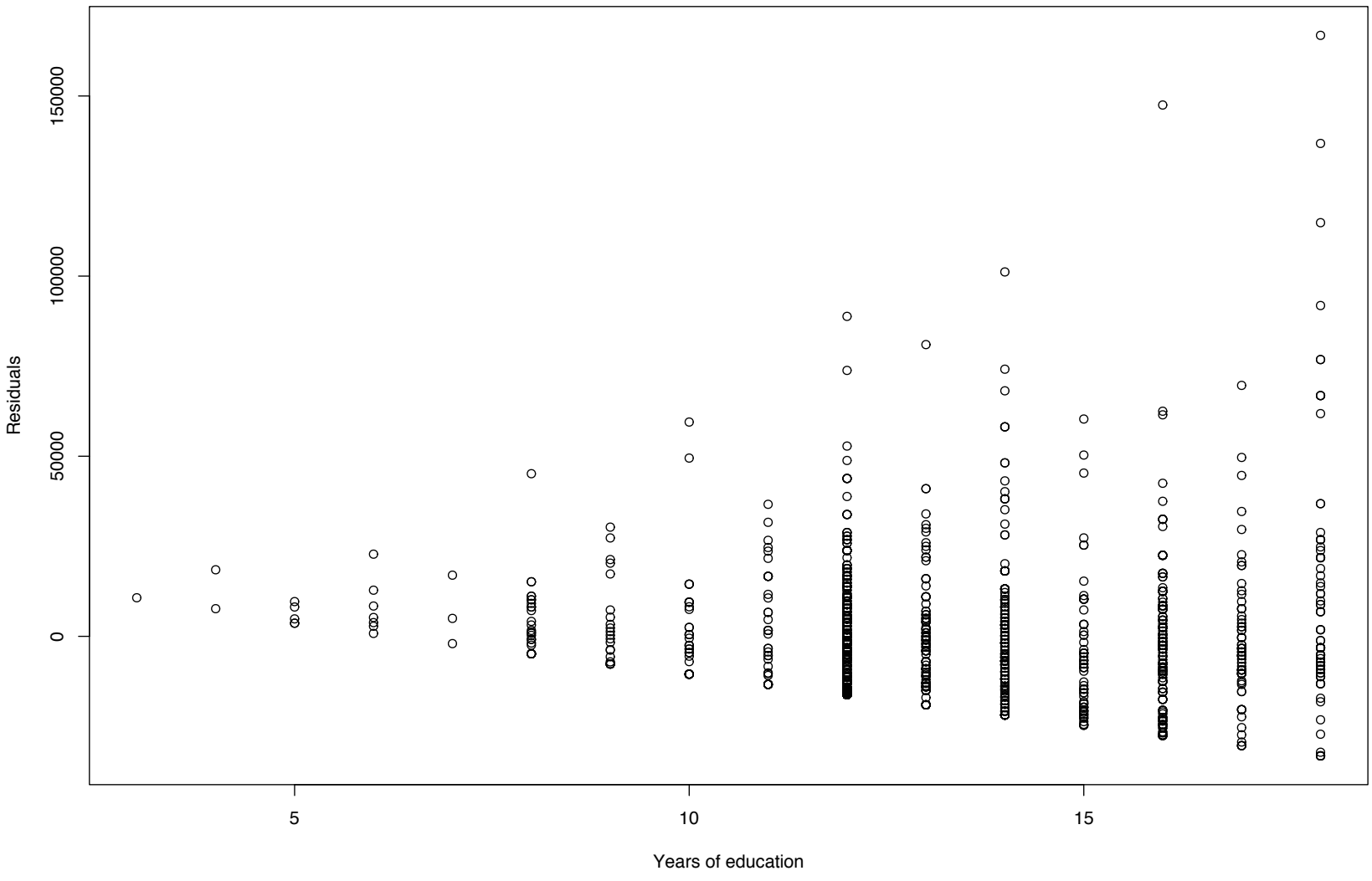
Regress earnings on education

```
# import the data
load("incomedata.Rda")
# plot earnings against education
plot(EARN~ED,data=id,xlab="Years of education",
    ylab="Earnings ($)")
# estimate regression of earnings on education
z1 <- lm(EARN~ED,data=id)
# plot line of best fit
abline(z1)
# print summary
summary(z1)
# print output with robust standard errors
coeftest(z1,df=Inf,vcov=vcovHC(z1,type="HC1"))
```

Plot of income against education



Plot of residuals against education



Regress earnings on education

```
lm(formula = EARN ~ ED, data = id)

Residuals:
   Min     1Q Median     3Q    Max
-34453 -10511  -4011   6018 164547

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -13784.5     3009.2  -4.581 5.12e-06 ***
ED            2735.4      219.3  12.471  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18320 on 1190 degrees of freedom
Multiple R-squared:  0.1156,Adjusted R-squared:  0.1148
F-statistic: 155.5 on 1 and 1190 DF,  p-value: < 2.2e-16
```

Regress earnings on education

```
z test of coefficients:

             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -13784.48    3340.11  -4.127 3.676e-05 ***
ED            2735.39     264.98  10.323 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Estimated regression model

$$\widehat{\text{EARN}}_i = \underset{(3340.11)}{-13784.5} + \underset{(264.98)}{2735.4}\text{ED}_i \qquad R^2 = 0.1156, \quad s_{\hat{u}} = 18320$$

Interpretation:

$\hat{\beta}_0$: Estimated expected earnings of someone with zero years of education is -\$13784.50

$\hat{\beta}_1$: Estimated changed in expected earnings per extra year of education is \$2735.40

R^2: 11.56% of variation in earnings is explained by variation in education

95% confidence interval for β_1:

$\hat{\beta}_1 \pm 1.96 \times s.e.(\hat{\beta}_1) = 2735.4 \pm 1.96 \times 264.98 = [2216, 3255]$

Estimated regression model

$$\widehat{\text{EARN}}_i = \underset{(3340.11)}{-13784.5} + \underset{(264.98)}{2735.4}\text{ED}_i \qquad R^2 = 0.1156, \quad s_{\hat{u}} = 18320$$

test for positive relationship between education and earnings:

$H_0 : \beta_1 = 0$ (no relationship between ED and EARN)
$H_1 : \beta_1 > 0$ (positive r'ship between ED and EARN)

t-statistic:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{s.e.(\hat{\beta}_1)} = \frac{2735.4 - 0}{264.98} = 10.32$$

p-value: this is a right-tailed test, so

$$p = 1 - \Phi(t) = 1 - \Phi(10.32) = 0.0000$$

so we reject H_0

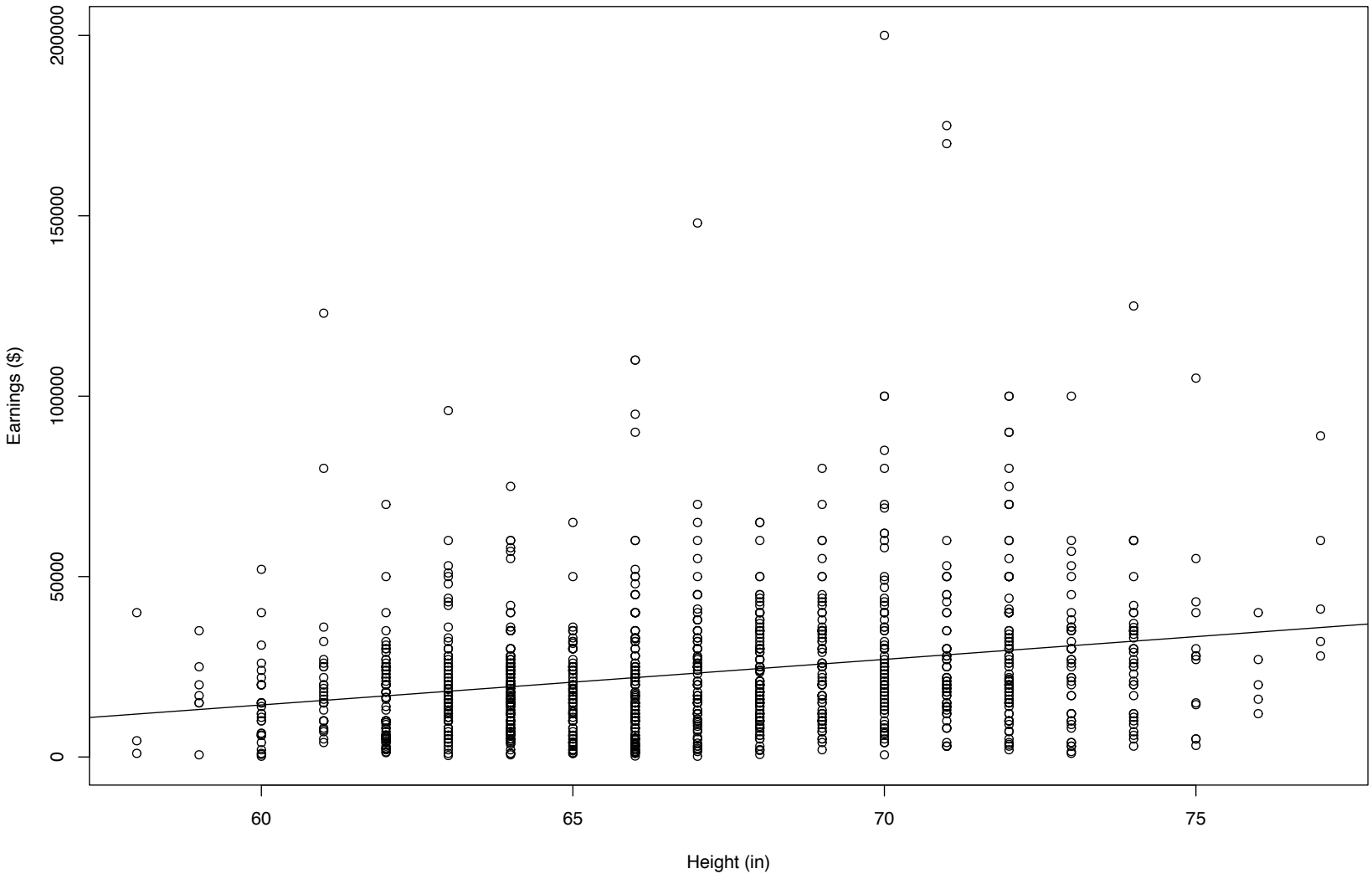
Conclusion: there is strong statistical evidence for a positive relationship between education and earnings

Regress earnings on height

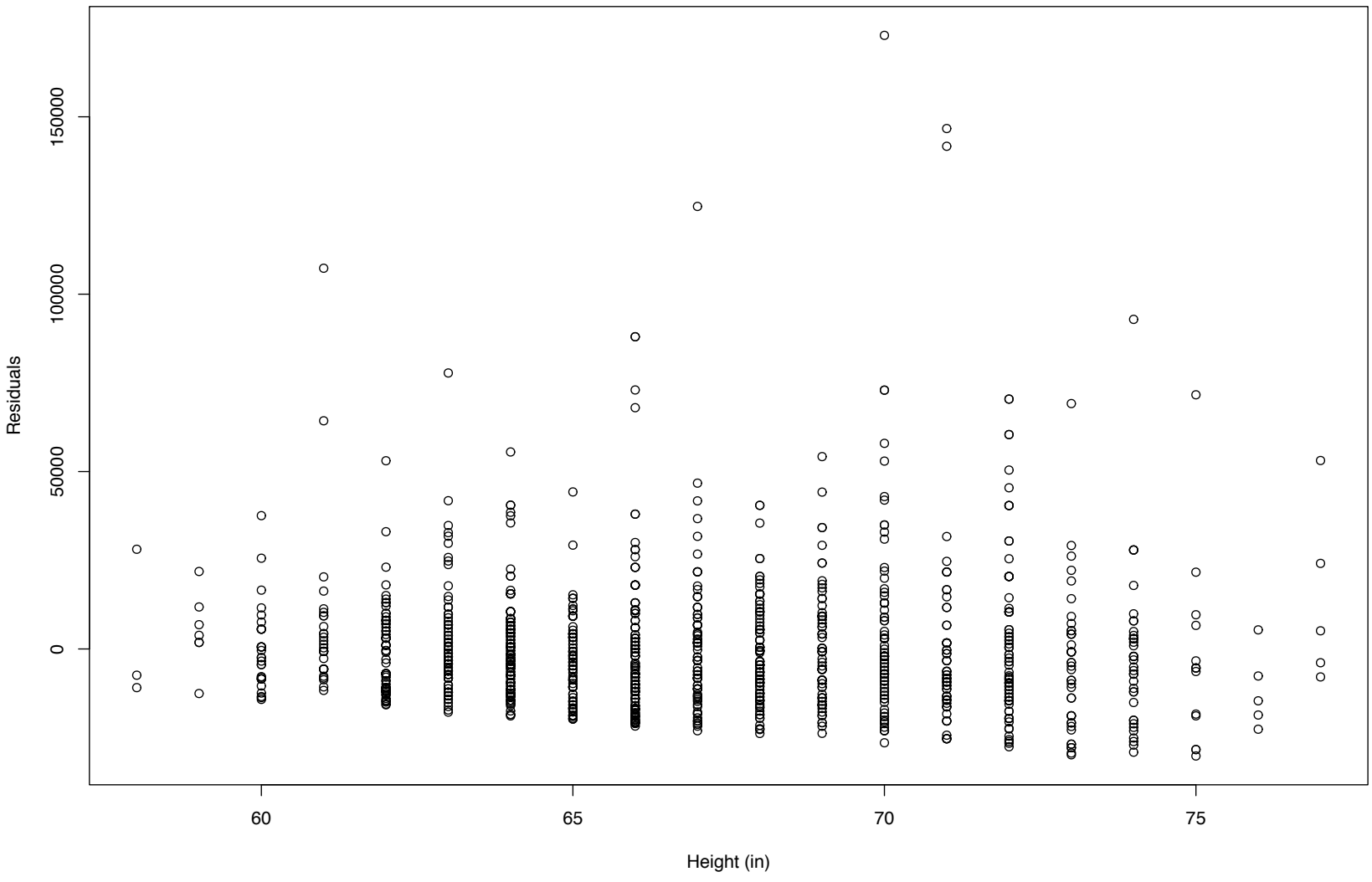
```
# plot earnings against height
plot(EARN~HEIGHT,data=id,xlab="Height (in)",
    ylab="Earnings ($)")
# estimate regression of earnings on education
z2 <- lm(EARN~HEIGHT,data=id)
# plot line of best fit
abline(z2)
# print summary
summary(z2)
# print output with robust standard errors
coeftest(z2,df=Inf,vcov=vcovHC(z2,type="HC1"))
```

Code for Stata R Statistics Tutor New York City

Plot of earnings against height



Plot of residuals against height



Regress earnings on height

```
lm(formula = EARN ~ HEIGHT, data = id)

Residuals:
   Min     1Q Median     3Q    Max
-30166 -11309  -3428   6527 172953

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -61316.3     9525.2  -6.437 1.76e-10 ***
HEIGHT        1262.3      142.1   8.883  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18870 on 1190 degrees of freedom
Multiple R-squared:  0.06218,Adjusted R-squared:  0.06139
F-statistic:  78.9 on 1 and 1190 DF,  p-value: < 2.2e-16
```

Regress earnings on height

```
z test of coefficients:

             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -61316.28    9909.43 -6.1877 6.106e-10 ***
HEIGHT        1262.33     150.84  8.3687 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Earnings by height

Estimated regression model:

$$\widehat{\text{EARN}}_i = \underset{(9909.43)}{-61316.28} + \underset{(150.84)}{1262.33}\text{HEIGHT}_i \quad R^2 = 0.06218, \quad s_{\hat{u}} = 18870$$

Interpretation:

$\hat{\beta}_0$: Estimated expected earnings of someone zero inches tall is -\$61316.28

$\hat{\beta}_1$: Estimated change in expected earnings per extra inch of height is \$1262.33

R^2: 6.2% of variation in earnings is explained by variation in height

Earnings by height

$$\widehat{\text{EARN}}_i = \underset{(9909.43)}{-61316.28} + \underset{(150.84)}{1262.33}\text{HEIGHT}_i \quad R^2 = 0.06218, \quad s_{\hat{u}} = 18870$$

test for positive relationship between height and earnings:

$H_0 : \beta_1 = 0$ (no relationship between height and earnings)
$H_1 : \beta_1 > 0$ (positive r'ship between height and earnings)

t-statistic:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{s.e.(\hat{\beta}_1)} = \frac{1262.33 - 0}{150.84} = 8.39$$

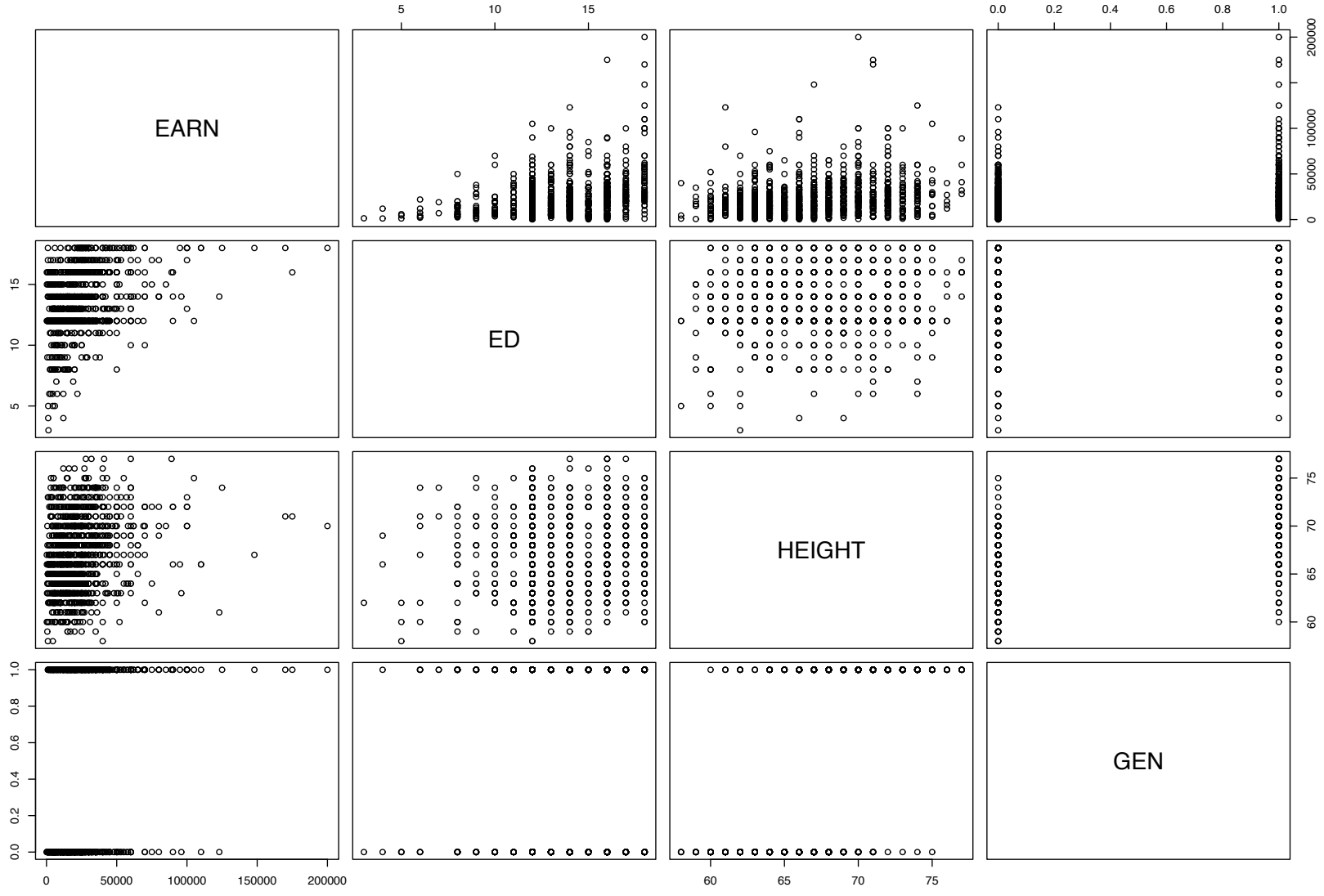
p-value: this is a right-tailed test, so

$$p = 1 - \Phi(t) = 1 - \Phi(8.39) = 0.0000$$

so we reject H_0

Conclusion: there is strong statistical evidence for a positive relationship between height and earnings

Earnings, education, height and gender



Earnings by height, gender and education

```
z test of coefficients:

             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -29030.01   11413.70 -2.5434   0.01098 *
HEIGHT         183.53     177.19  1.0358   0.30030
ED            2638.54     253.80 10.3963 < 2.2e-16 ***
GEN          10083.50    1397.31  7.2163  5.34e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```