Regression and uncertainty part 2: stochastic error

Frank Edwards

The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0: The value of y when x is equal to zero

The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0: The value of y when x is equal to zero

β_1: The average increase in y when x increases by one unit

The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0: The value of y when x is equal to zero

β_1: The average increase in y when x increases by one unit

ε: The distance between the line $y = \beta_0 + \beta_1 X$ and the actual observed values of y. Allows us to estimate the line, even when x and y do not fall exactly on a line.

The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0: The value of y when x is equal to zero

β_1: The average increase in y when x increases by one unit

ε: The distance between the line $y = \beta_0 + \beta_1 X$ and the actual observed values of y. Allows us to estimate the line, even when x and y do not fall exactly on a line.

Understanding the regression line

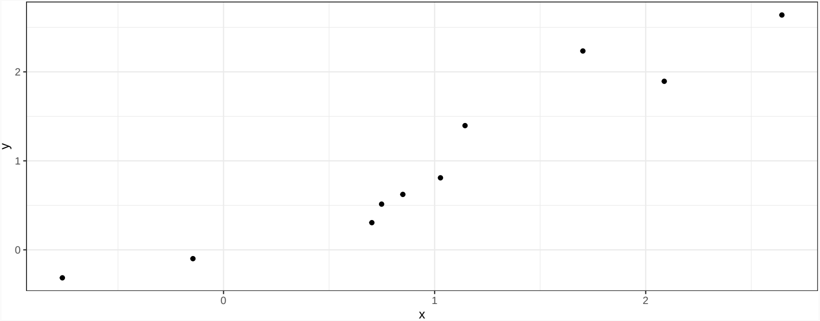
```
## # A tibble: 10 x 2
##         x       y
##     <dbl>   <dbl>
##  1  0.849  0.623
##  2  1.03   0.809
##  3  2.09   1.89
##  4 -0.763 -0.315
##  5  1.70   2.23
##  6  0.749  0.514
##  7  0.703  0.305
##  8 -0.145 -0.0992
##  9  1.14   1.40
## 10  2.65   2.64
```

$\hat{\beta}_0 = 0.05$, $\hat{\beta}_1 = 0.95$

- Estimate $\hat{Y}$. Recall that $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
- Estimate ε. Recall that $\varepsilon = Y - \hat{Y}$

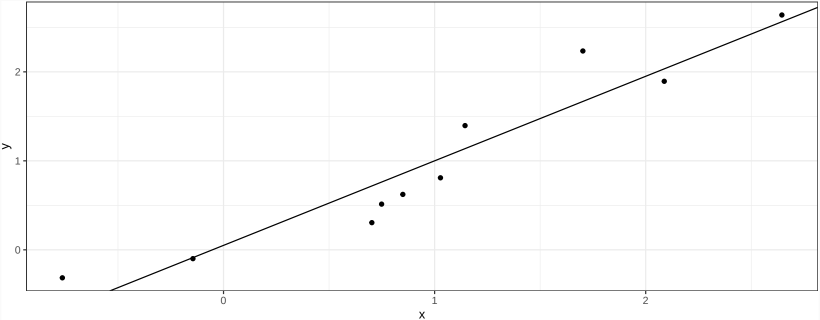
Understanding the regression line

$\hat{\beta}_0 = 0.05$, $\hat{\beta}_1 = 0.95$

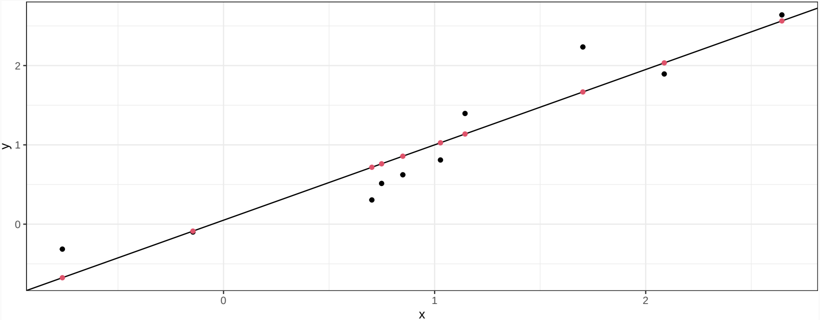


Understanding the regression line: adding the fit

$\hat{\beta}_0 = 0.05$, $\hat{\beta}_1 = 0.95$

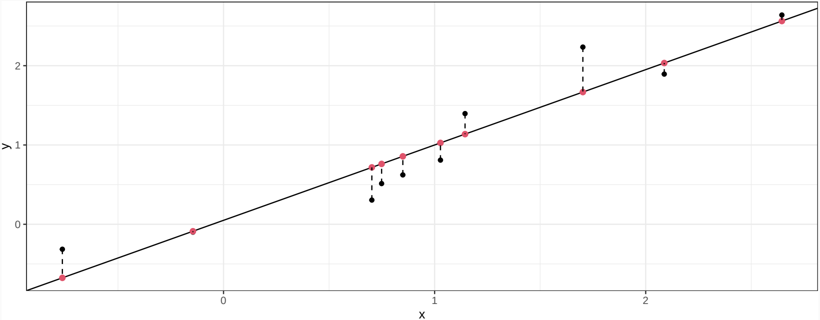


$\hat{\beta}_0 = 0.05$, $\hat{\beta}_1 = 0.95$



Understanding the regression line: adding ε

$\hat{\beta}_0 = 0.05$, $\hat{\beta}_1 = 0.95$



For estimates of β to be unbiased and consistent, the following assumptions must be met:

1. The linear model approximates the data generating process

Assumptions of a linear regression model

For estimates of β to be unbiased and consistent, the following assumptions must be met:

1. The linear model approximates the data generating process
2. Exogeneity of errors: $E(\varepsilon) = 0$, errors uncorrelated with predictors

Assumptions of a linear regression model

For estimates of β to be unbiased and consistent, the following assumptions must be met:

1. The linear model approximates the data generating process
2. Exogeneity of errors: $E(\varepsilon) = 0$, errors uncorrelated with predictors
3. Linear independence of predictors

Assumptions of a linear regression model

For estimates of β to be unbiased and consistent, the following assumptions must be met:

1. The linear model approximates the data generating process
2. Exogeneity of errors: $E(\varepsilon) = 0$, errors uncorrelated with predictors
3. Linear independence of predictors
4. Constant error variance (Homoskedasticity): $V(\varepsilon|X) = V(\varepsilon)$

Assumptions of a linear regression model

For estimates of β to be unbiased and consistent, the following assumptions must be met:

1. The linear model approximates the data generating process

Assumptions of a linear regression model

For estimates of β to be unbiased and consistent, the following assumptions must be met:

1. The linear model approximates the data generating process
2. Exogeneity of errors: $E(\varepsilon) = 0$, errors uncorrelated with predictors

Assumptions of a linear regression model

For estimates of β to be unbiased and consistent, the following assumptions must be met:

1. The linear model approximates the data generating process
2. Exogeneity of errors: $E(\varepsilon) = 0$, errors uncorrelated with predictors
3. Linear independence of predictors

Assumptions of a linear regression model

For estimates of β to be unbiased and consistent, the following assumptions must be met:

1. The linear model approximates the data generating process
2. Exogeneity of errors: $E(\varepsilon) = 0$, errors uncorrelated with predictors
3. Linear independence of predictors
4. **Constant error variance (Homoskedasticity):** $V(\varepsilon|X) = V(\varepsilon)$

Ways to express an OLS model

As linear with Normal errors:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

Ways to express an OLS model

As linear with Normal errors:

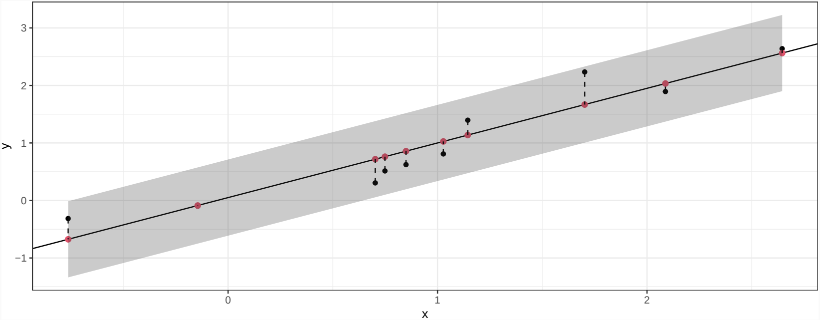
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

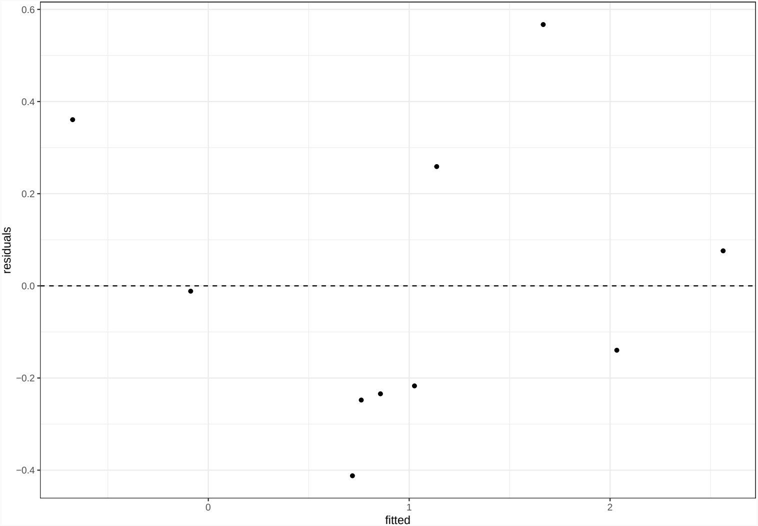
As Normal, with linear vector of means:

$$y \sim N(\beta X, \sigma^2)$$

What this means: 95% of observations should fall in this zone



One way to visualize: residuals vs fitted



Let’s try this with real data

Let’s try this with real data

```
## [1] "4"            "1"            "2"            NA             "3"
## [6] "5+"           "<1"           "do not watch"
```

```
## [1] "8"   "7"   "5"   "6"   "10+" "<5"  NA    "9"
```

```
## # A tibble: 64 x 13
##      age gender grade hispanic race                     height weight helmet_12m
##    <dbl> <chr>  <dbl> <chr>    <chr>                     <dbl>  <dbl> <chr>
##  1    16 female    11 not      Black or African Americ~   1.5    52.6 never
##  2    17 male      11 not      White                      1.78   74.8 rarely
##  3    17 male      11 not      White                      1.75  107.  never
##  4    18 male      12 not      Black or African Americ~   1.7    80.3 never
##  5    16 male      10 not      White                      1.78   81.6 always
##  6    16 male      10 not      Black or African Americ~   1.63   56.7 never
##  7    14 male       9 hispanic White                      1.63   54.4 never
##  8    17 male      11 not      Black or African Americ~   1.83   92.5 never
##  9    15 male       9 not      Black or African Americ~   1.69   67.8 did not r~
## 10    17 male      10 not      White                      1.78   66.7 never
## # i 54 more rows
## # i 5 more variables: text_while_driving_30d <chr>, physically_active_7d <dbl>,
## #   hours_tv_per_school_day <dbl>, strength_training_7d <dbl>,
## #   school_night_hours_sleep <dbl>
```

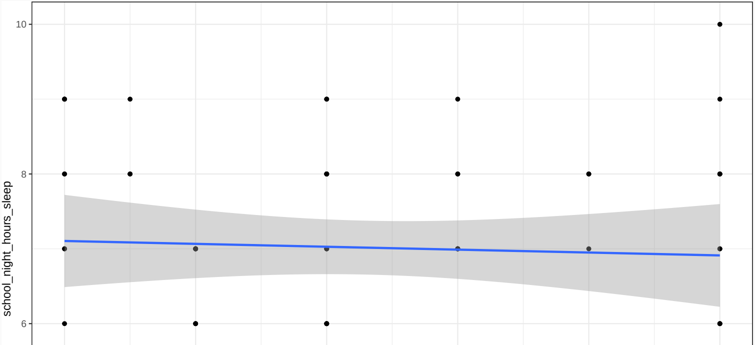
Fit a model for sleep duration predicted by tv watching

```r
m1 <- lm(school_night_hours_sleep ~ hours_tv_per_school_day, data = dat)

tidy(m1)
```

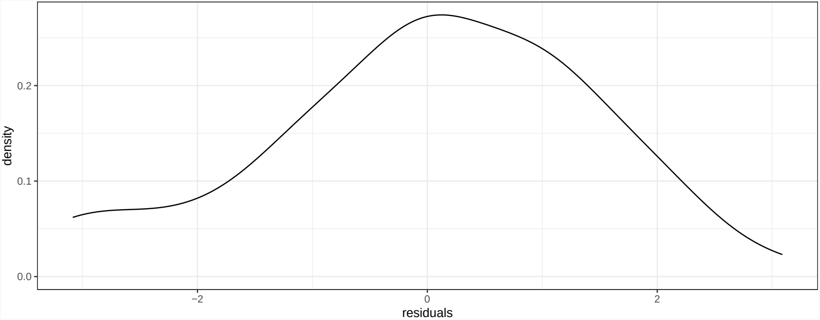
```
## # A tibble: 2 x 5
##   term                    estimate std.error statistic  p.value
##   <chr>                      <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)               7.10       0.308    23.1   4.15e-32
## 2 hours_tv_per_school_day  -0.0386     0.109    -0.356 7.23e- 1
```

```r
ggplot(dat, aes(x = hours_tv_per_school_day, y = school_night_hours_sleep)) + geom_point() +
    geom_smooth(method = "lm")
```



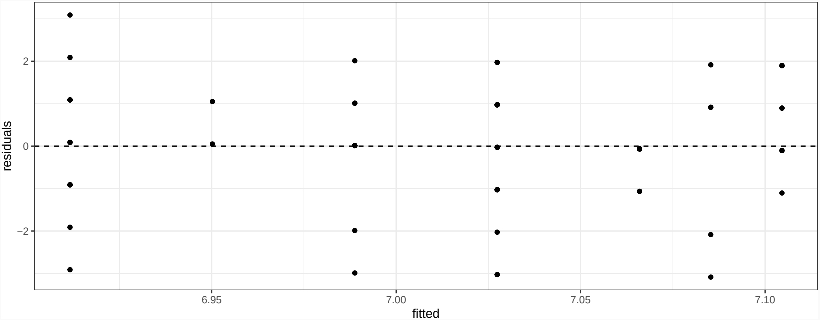
What is the distribution of the residuals? Are they Normal?

```
ggplot(data = data.frame(residuals = resid(m1)), aes(x = residuals)) + geom_density()
```



What about residuals vs fitted?

```r
ggplot(data = data.frame(fitted = fitted(m1), residuals = resid(m1)), aes(x = fitted,
    y = residuals)) + geom_point() + geom_hline(yintercept = 0, lty = 2)
```



Looks ok! Now what?

Because our model is not *heteroskedastic* (non-constant error variance), we can make valid predictions from it!

Looks ok! Now what?

Because our model is not *heteroskedastic* (non-constant error variance), we can make valid predictions from it!

Before, we estimated the sampling distribution of $E(y)$ using information on the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

Looks ok! Now what?

Because our model is not *heteroskedastic* (non-constant error variance), we can make valid predictions from it!

Before, we estimated the sampling distribution of $E(y)$ using information on the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$

But that's not the only source of uncertainty in our model!

Sources of random variation in our model

$$\hat{\beta}_0 \sim N(\beta_0, s^2_{\beta_0})\ \hat{\beta}_1 \sim N(\beta_1, s^2_{\beta_1})\ y = \beta_1 + \beta_2 + \varepsilon\ \varepsilon \sim N(0, \sigma^2)$$

What does epsilon represent?

Exercise: Make predictions from m1 for y.

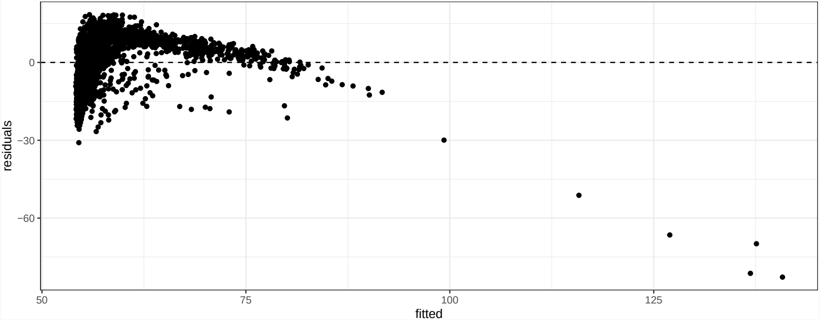
What does non-constant error variance look like?

Data with non-constant error variance will show distinctive patterns in their residuals

```
library(gapminder)
m2 <- lm(lifeExp ~ gdpPercap, data = gapminder)
```

Non-contant error variance

```r
ggplot(data.frame(fitted = fitted(m2), residuals = resid(m2)), aes(x = fitted, y = residuals)) +
    geom_point() + geom_hline(yintercept = 0, lty = 2)
```



Problems

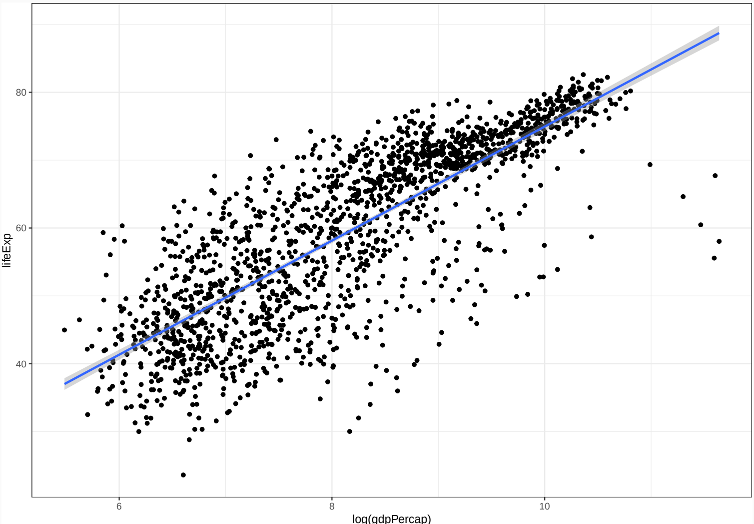
Here there are actually two problems:

1. The linear model does not reflect the data generating process (the relationship between x and y)
2. Error variance is not constant

OK, so log it

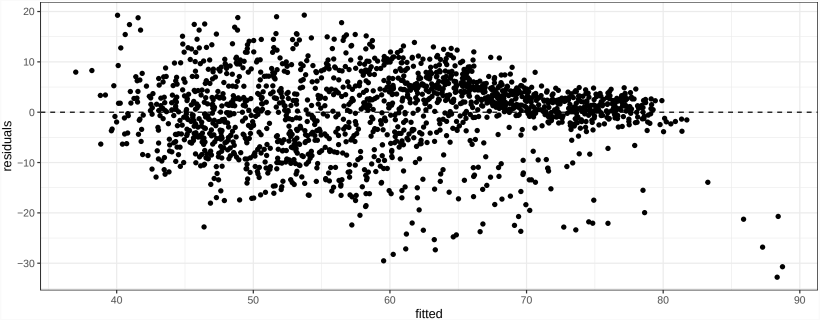
```
library(gapminder)
m3 <- lm(lifeExp ~ log(gdpPercap), data = gapminder)

ggplot(gapminder, aes(x = log(gdpPercap), y = lifeExp)) + geom_point() + geom_smooth(method = "lm")
```



Non-contant error variance

```r
ggplot(data.frame(fitted = fitted(m3), residuals = resid(m3)), aes(x = fitted, y = residuals)) +
    geom_point() + geom_hline(yintercept = 0, lty = 2)
```



In summary

In linear regression, we assume that the residuals follow a normal distribution.

In linear regression, we assume that the residuals follow a normal distribution.

We can write this two ways (they are equivalent):

1. $y = \beta_0 + \beta_1 X + \varepsilon; \varepsilon \sim N(0, \sigma^2)$
2. $\mu = \beta_0 + \beta_1 X; y \sim N(\mu, \sigma^2)$

In summary

We have multiple sources of random error in our model.

1. We have uncertainty in our estimates of β, that we approximate using the Central Limit Theorem. This is the uncertainty we have about the location of $E(y)$ or μ
2. We have uncertainty in y, which is ordinary sampling error. For linear regression, we are going to assume that conditional on X, this error is Normally distributed