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

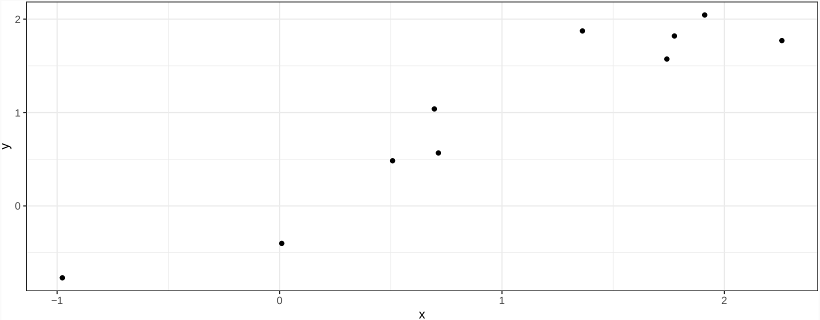
```
## # A tibble: 10 x 2
##          x      y
##      <dbl>  <dbl>
##  1  1.36     1.87
##  2  0.714    0.567
##  3  1.91     2.04
##  4  1.78     1.82
##  5  1.74     1.57
##  6  0.696    1.04
##  7  0.508    0.484
##  8  2.26     1.77
##  9  0.00973 -0.401
## 10 -0.977   -0.770
```

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

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

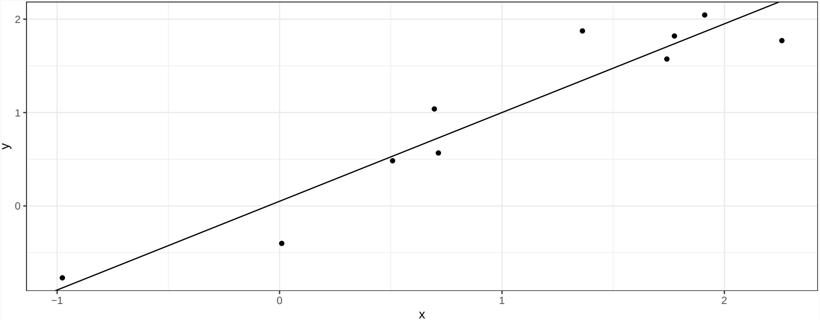
Understanding the regression line

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



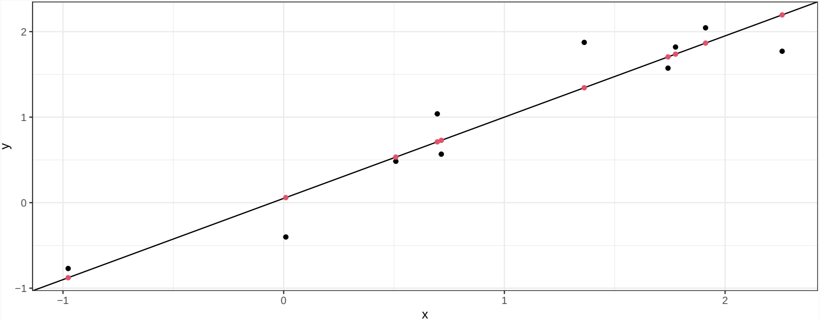
Understanding the regression line: adding the fit

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



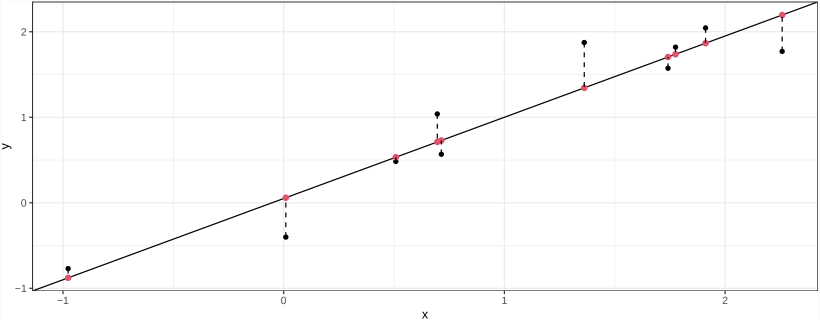
Understanding the regression line: adding $\hat{y}$

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



Understanding the regression line: adding ε

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



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): $V(\varepsilon|X) = V(\varepsilon)$

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:

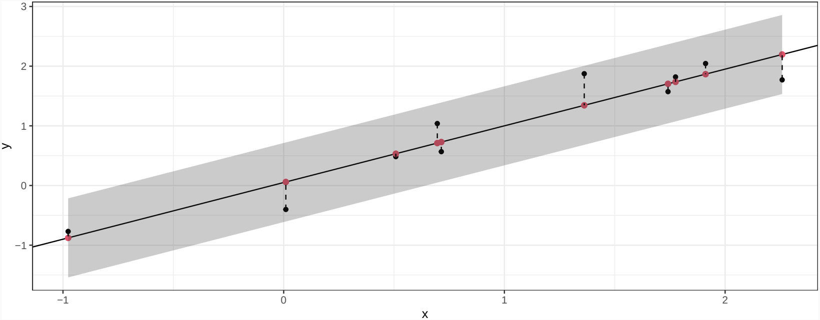
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \varepsilon$$

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

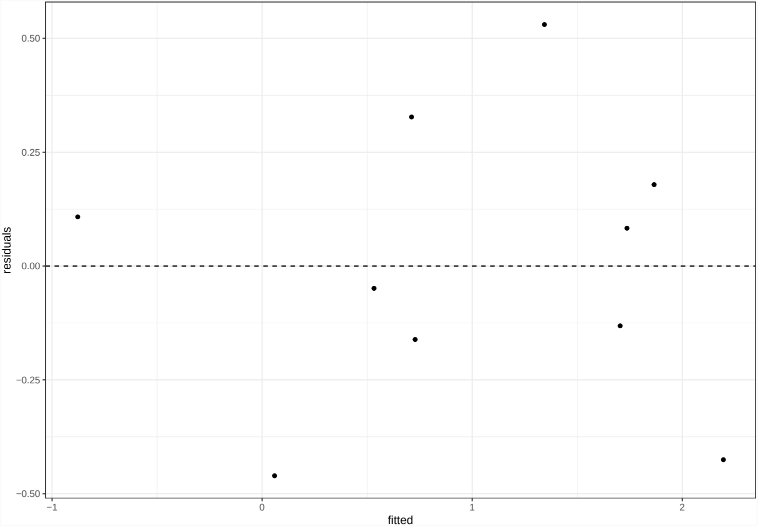
As Normal, with linear vector of means:

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

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: 90 x 2
##    hours_tv_per_school_day school_night_hours_sleep
##                      <dbl>                    <dbl>
##  1                       4                        8
##  2                       1                        7
##  3                       2                        7
##  4                       2                        5
##  5                       2                        6
##  6                       1                        7
##  7                       3                        7
##  8                       5                        7
##  9                       2                        8
## 10                       5                       10
## # i 80 more rows
```

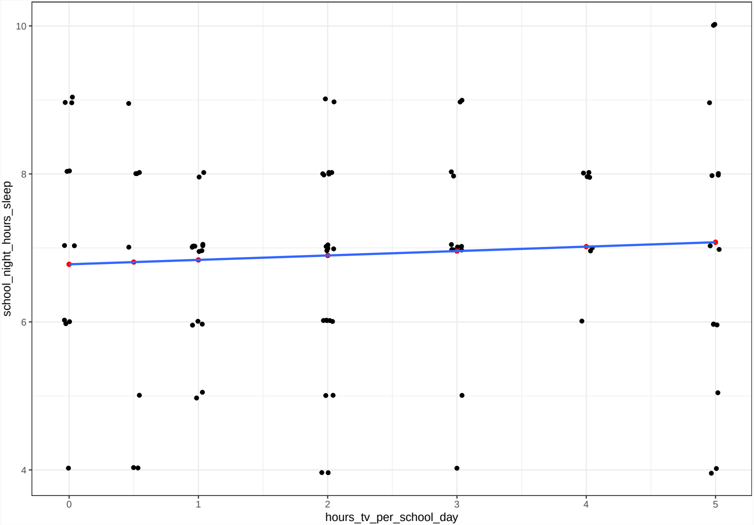
Fit a model for sleep duration predicted by tv watching

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

tidy(m1)
```

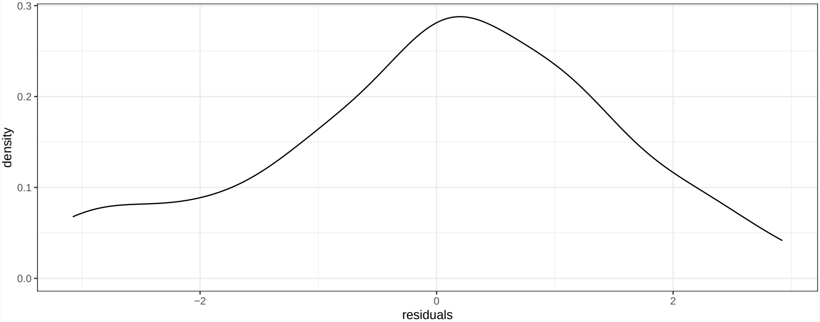
```
## # A tibble: 2 x 5
##   term                    estimate std.error statistic  p.value
##   <chr>                      <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)               6.78      0.261     26.0   4.42e-43
## 2 hours_tv_per_school_day   0.0596    0.0945     0.630 5.30e- 1
```

Evaluate the regression line



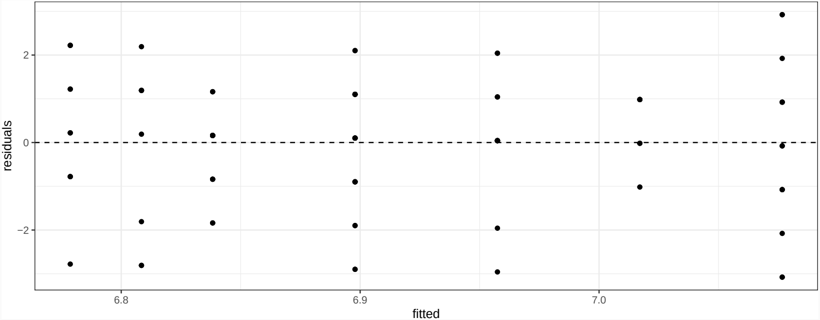
What is the distribution of the residuals? Are they Normal?

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



What about residuals vs fitted?

```
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 distinct pattens in their residuals