Multiple linear regression

Regression Analysis

STAT5002

The University of Sydney

May 2025

THE UNIVERSITY OF SYDNEY

Regression Analysis

Topic 13: Multiple linear regression

Topic 14: Model selection

Topic 15: Logistic regression

Outline

Today:

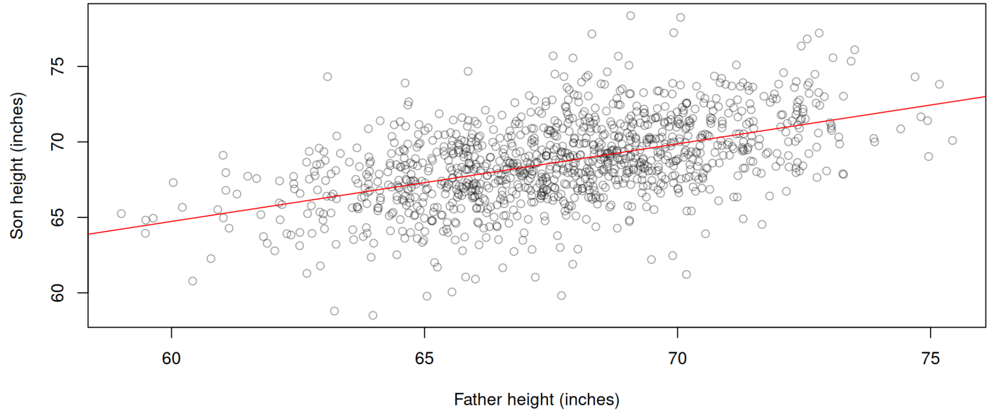
- Inference for simple linear regression models (one independent variable)
- Case study (using transformation)
- Multiple linear regression models (multiple independent variables)

Next week:

- Model selection
- Logistic regression

Pearson's data

```r
# install.packages('UsingR')
suppressMessages(library(UsingR))
library(UsingR)  # Loads another collection of datasets
data(father.son)  # This is Pearson's data.
data = father.son
x1 = data$fheight  # fathers' heights
y = data$sheight  # sons' heights
plot(x1, y, xlab = "Father height (inches)", ylab = "Son height (inches)", col = adjustcolor("black",
    alpha.f = 0.35))
abline(lm(y ~ x1), col = "red")
```



- `x1` contains the fathers' heights (independent/explanatory variable)
- `y` contains sons' heights (dependent/response variable)

Pearson's correlation coefficient (r)

- It is the **mean** of the **product** of the variables in **standard units**, indicating both the sign and strength of the **linear association**.

$$\hat{r} = \frac{\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- $x_{1,i}$ represents the i-th observed point of the 1st independent variable x_1.
- Later we will use this notation to handle multiple independent variables $x_1, x_2, \ldots, x_p$.

```r
cor(x1, y)
```

```
[1] 0.5013383
```

The correlation coefficient is **shift and scale invariant**.

```r
cor(0.2 * x1 + 3, 3 * y - 1)
```

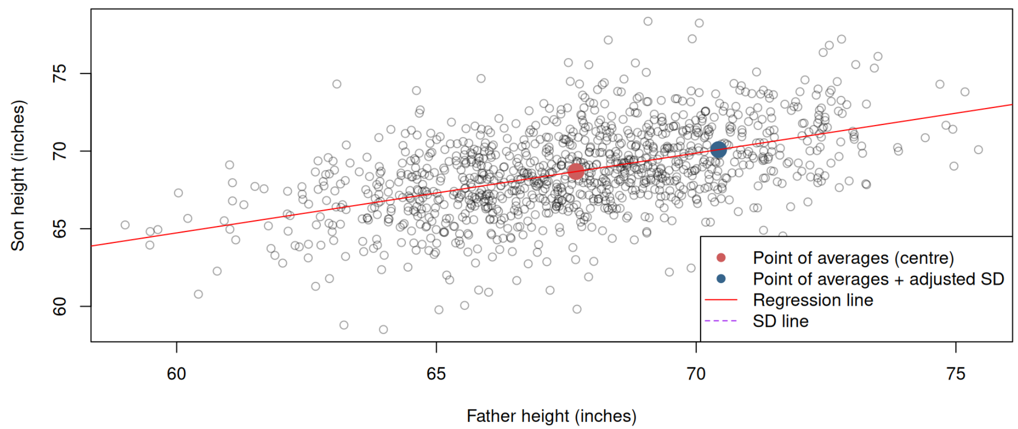
```
[1] 0.5013383
```

The correlation coefficient is not affected by interchanging the variables.

```r
cor(y, x1)
```

```
[1] 0.5013383
```

Regression line



- The regression line $\hat{y} = \hat{b}_0 + \hat{b}_1 \cdot x_1$ connects $(\bar{x}_1, \bar{y})$ and $(\bar{x}_1 + \hat{s}_{x_1}, \bar{y} + \hat{r} \cdot \hat{s}_y)$ by estimating **regression corfficients**
  - the **slope** $\hat{b}_1 = \hat{r}\frac{\hat{s}_y}{\hat{s}_{x_1}}$ and the **intercept** $\hat{b}_0 = \bar{y} - \hat{b}_1 \cdot \bar{x}_1$.
- The **resisual** of the regression model

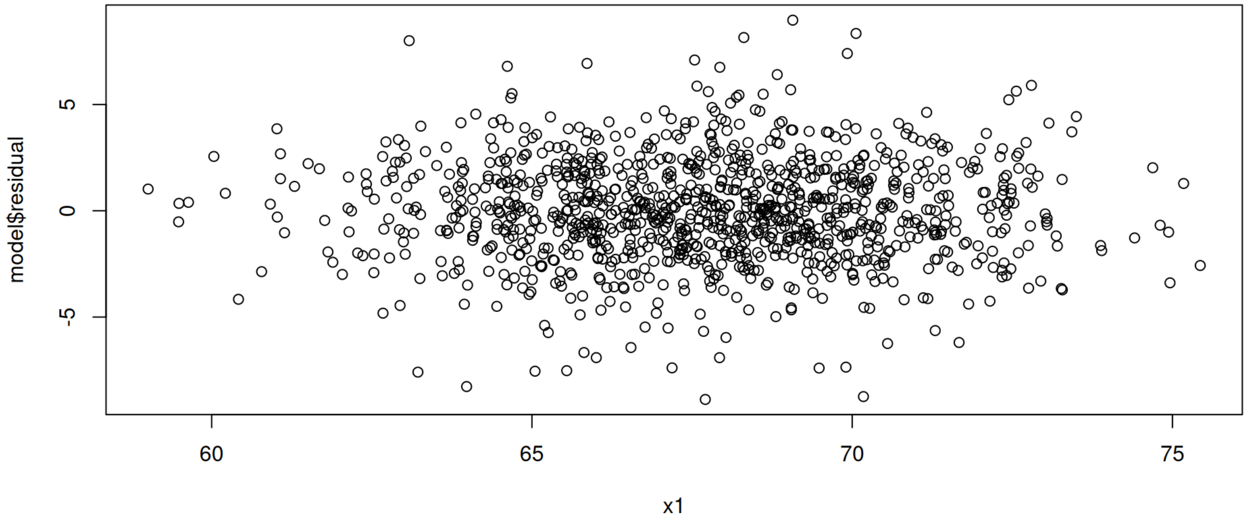
$$e_i = y_i - \hat{y}_i = y_i - (\underbrace{\hat{b}_0}_{\text{intercept}} + \underbrace{\hat{b}_1}_{\text{slope}} x_{1,i}).$$

- We use $\hat{s}_{x_1}$ and $\hat{s}_y$ to denote sample SDs here, as σ is used for the SD of residual later.

```r
model = lm(y ~ x1)
model
```

```

Call:
lm(formula = y ~ x1)

Coefficients:
(Intercept)           x1
    33.8866       0.5141
```

```r
plot(x1, model$residual)
```



- If the linear model is appropriate, the residual plot should be a random scatter of points.
- The variance of the random scatter should not change with the location of x_1 (**homoscedasticity**).

Performance benchmark of linear regression model

- The regression line is the **best** (optimal) linear model, as it minimises the sum of the squared residuals $\sum_{i=1}^{n} e_i^2$ among all linear models (lines).
- We can use the **coefficient of determination** (r^2) to summarise the performance of a regression line.
  - ➠ The sum of squared residuals (or SSE for sum of squared errors) for the regression line

$$\widehat{\text{SSE}} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \underbrace{(\hat{b}_0 + \hat{b}_1 \cdot x_{1,i})}_{\hat{y}_i}\right)^2$$

    measures **variation in y left unexplained by the regression line**.
  - ➠ The sum of squared total variations (SST) of y (sum of squared deviations)

$$\widehat{\text{SST}} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

    measures of **the total amount of variation in observed y values** without relying on the independent variable x_1.
- $\widehat{\text{SST}} \geq \widehat{\text{SSE}}$ as the regression is optimal for sum of squared errors

- The proportion of variation in the observed y that **cannot** be explained by the simple linear regression model is given by

$$\frac{\widehat{\mathrm{SSE}}}{\widehat{\mathrm{SST}}}$$

which is always ≤ 1.

- Thus, the proportion of variation in the observed y that **can** be explained by the simple linear regression model (aka **coefficient of determination**) is

$$\frac{\widehat{\mathrm{SST}} - \widehat{\mathrm{SSE}}}{\widehat{\mathrm{SST}}} = 1 - \frac{\widehat{\mathrm{SSE}}}{\widehat{\mathrm{SST}}} = \hat{r}^2$$

  - ➡ It is exactly the **squared correlation coefficient** (a number between 0 and 1) for simple linear regression models

- The higher the value of the coefficient of determination, the more successful is the simple linear regression model in explaining variation of the dependent variable y.

Possible extensions

- The correlation coefficient indicates the strength of the linear association of a sample.
  - ➠ Do the data suggest a significant linear association in the population?
  - ➠ We will extend the T-test to test this.
- What happen if we have multiple independent variables, $x_1, x_2, \ldots, x_p$?
  - ➠ We need to fit a linear model

$$\hat{y} = \hat{b}_0 + \hat{b}_1 \cdot x_1 + \hat{b}_2 \cdot x_2 + \cdots + \hat{b}_p \cdot x_p$$

  - ➠ How can we interpret this?
  - ➠ How to select the most relevant independent variables from $x_1, x_2, \ldots, x_p$ to achieve a similar performance as using all independent variables?

Inference for simple linear regression models

Probablistic view of simple linear regression models

The simple linear regression model aims to predict the outcome of a dependent/response variable Y, which is a random draw, using a independent/explanatory variable x_1 and the model

$$Y_i = \underbrace{b_0 + b_1 \cdot x_{1,i}}_{\text{the "population" linear model}} + \varepsilon_i$$

for $i = 1, \ldots, n$ indexing an observation in the data set.

- The errors ε_i are **random draws** taken from an "error box" with **mean** 0 **and a fixed (population) SD** σ.
- For any given $x_{1,i}$, the regression line $b_0 + b_1 \cdot x_{1,i}$ is the expected value of Y_i.
    - The intercept is the expected value of Y_i when $x_1 = 0$.
    - The slope is the amount we expect Y to change by when x_1 increases one unit,
        - i.e. for a one unit increase in x_1 we expect Y to change by b_1 (could be an increase or decrease depending on the sign) in average.
- We estimate the population intercept and slope (b_0, b_1) using observed $(x_{1,i}, y_i), i = 1, \ldots, n$.
- We also need to estimate σ of the error box from the residuals of the fitted model. How?

Assumptions

$\boxed{\text{A}}$ We make the following assumptions:

1. The errors ε_i are independently drawn from an "error box" with mean 0 and SD σ.

- So the variability of $\boldsymbol{\varepsilon_i}$ does not depend on $\boldsymbol{x}$ (and thus homoscedasticity).
- We can check the residual plot for checking homoscedasticity
- However, the independence between the errors is usually dealt with in the experimental design phase (before data collection).
  - ➠ We didn't cover the design the experiment in this unit (let's skip the check for independence).

Assumptions

2. The “error box” should be normal-shaped.

- The estimated coefficients $(\hat{b}_0, \hat{b}_1)$ not only give the best line fitting the observed sample (in terms of minimizing the sum of squared residuals);
- but also correctly estimate of the population coefficients (b_0, b_1) in expectation under the normal error box assuption (the derivation is beyond the scope here).
- Use the QQ plot to check normality.

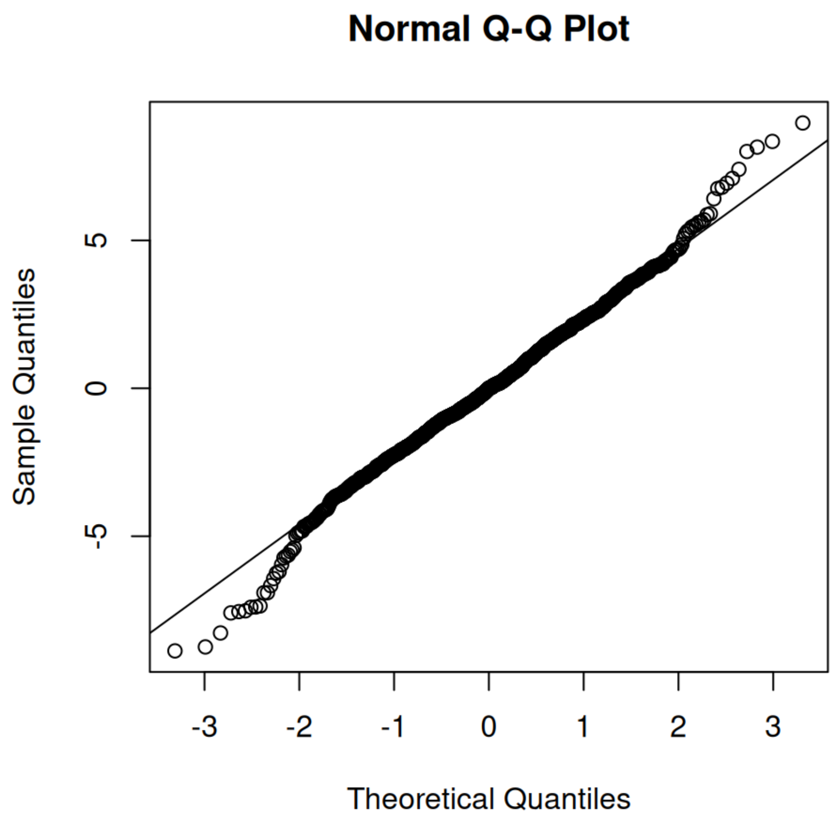
3. Linearity: should be checked using graphical summaries.

- Either the scatter plot or the residual plot can be used for this (see Topic 4).

For the first two assumptions, we simply write $\varepsilon_i \sim (\text{iid})\ N(0, \sigma^2)$.

Normality of Pearson's data

```r
qqnorm(model$residual)
qqline(model$residual)
```



- slight deviations away from the qqline towards the tails, but most of the quantile points follow the QQ line. It is reasonable to assume normality here.

Inference: T-test

Recall the population model for simple linear regression

$$Y_i = b_0 + b_1 \cdot x_{1,i} + \varepsilon_i, \quad \varepsilon_i \sim \text{(iid) } N(0, \sigma^2)$$

$\boxed{\text{H}}$ Typically, we are interested in the hypotheses

- $H_0 : b_1 = 0$ there is no linear relationshop between x and Y.
- Alternatives:
  - $H_1 : b_1 \neq 0$: there is linear relationshop between x and Y.
  - $H_1 : b_1 > 0$ (or $b_1 < 0$): there is positive (or negative) linear relationshop between x and Y.

$\boxed{\text{T}}$ To do this, we use a T-statistic

$$T = \frac{\hat{b}_1 - b_1}{\widehat{SE}(\hat{b}_1)} = \frac{\hat{b}_1}{\widehat{SE}(\hat{b}_1)} \sim t_{n-2}$$

- The estimated slope $\hat{b}_1$ can be viewed as a random draw following a normal-shaped box.
- What is the (estimated) standard error of the slope estimate $\widehat{SE}(\hat{b}_1)$?
- Why does the Student's t-distribution have $n - 2$ degrees of freedom?

Standard error of slope estimate $SE(\hat{b}_1)$

- This derivation (three slides) is NOT for assessment!
- Fixing $x_{1,i}$, estimated slope $\hat{b}_1$ can be viewed as a random draw depending on Y_i

$$\hat{b}_1 = \hat{r} \times \frac{\hat{s}_Y}{\hat{s}_X} = \frac{\sum_{i=1}^n (x_{1,i} - \bar{x}_1)(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_{1,i} - \bar{x}_1)^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \times \frac{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2}}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}}$$

which gives (using the population model $Y_i = b_0 + b_1 \cdot x_i + \varepsilon_i$)

$$\hat{b}_1 = \frac{\sum_{i=1}^n (x_{1,i} - \bar{x}_1)(Y_i - \bar{Y})}{\sum_{i=1}^n (x_{1,i} - \bar{x}_1)^2} = \frac{\sum_{i=1}^n (x_{1,i} - \bar{x}_1)(b_0 + b_1 \cdot x_i + \varepsilon_i - \bar{Y})}{(n-1)\hat{s}_X^2}$$

- Recall that $\sum_{i=1}^n (x_{1,i} - \bar{x}_1) = 0$, so

$$\sum_{i=1}^n (x_{1,i} - \bar{x}_1)(b_0 + b_1 \cdot x_i + \varepsilon_i - \bar{Y}) = \sum_{i=1}^n (x_{1,i} - \bar{x}_1)(b_1 \cdot (x_i - \bar{x}) + \varepsilon_i)$$

as we can add or substact constants in the second bracket without changing the numerator

- Then, we have the slope estimate

$$\hat{b}_1 = \frac{b_1 \sum_{i=1}^n (x_{1,i} - \bar{x}_1)^2 + \sum_{i=1}^n (x_{1,i} - \bar{x}_1)\varepsilon_i}{(n-1)\hat{s}_X^2} = b_1 + \frac{\sum_{i=1}^n (x_{1,i} - \bar{x}_1)\varepsilon_i}{(n-1)\hat{s}_X^2}$$

- Rearranging, we have

$$\hat{b}_1 - b_1 = \sum_{i=1}^n \underbrace{\left(\frac{x_{1,i} - \bar{x}_1}{(n-1)\hat{s}_X^2}\right)}_{w_i} \varepsilon_i = \sum_{i=1}^n w_i \cdot \varepsilon_i$$

where the weights w_i only depend on observations of the independent variable x_1.

- The linear combination $\hat{b}_1 - b_1 = \sum_{i=1}^n w_i \cdot \varepsilon_i$ also follows a normal curve and has
  - expected value: $\sum_{i=1}^n w_i \cdot E(\varepsilon_i) = 0$
  - squared standard error:

$$\sum_{i=1}^n w_i^2 \cdot SE(\varepsilon_i)^2 = \sum_{i=1}^n w_i^2 \cdot \sigma^2 = \sigma^2 \sum_{i=1}^n w_i^2$$

- Note that the sum of squared weights is

$$\sum_{i=1}^{n} w_i^2 = \sum_{i=1}^{n} \left(\frac{x_{1,i} - \bar{x}_1}{(n-1)\hat{s}_X^2} \right)^2 = \frac{\sum_{i=1}^{n} (x_{1,i} - \bar{x}_1)^2}{\left((n-1)\hat{s}_X^2\right)^2} = \frac{(n-1)\hat{s}_X^2}{\left((n-1)\hat{s}_X^2\right)^2} = \frac{1}{(n-1)\hat{s}_X^2}$$

- This way,

$$SE(\hat{b}_1) = SE(\hat{b}_1 - b_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n} (x_{1,i} - \bar{x}_1)^2}}$$

- Since adding constant to a random draw does not change its standard error (as the SD of the box remains the same), we have

$$SE(\hat{b}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n} (x_{1,i} - \bar{x}_1)^2}}$$

Estimate $SE(\hat{b}_1)$ from an observed sample

- We need to estimate σ (the population SD of the error box for ε_i) to get $\widehat{SE}(\hat{b}_1)$. But what is a sensible estimate for σ?

$$\widehat{\sigma} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\Big(y_i - (\hat{b}_0 + \hat{b}_1 \cdot x_{1,i})\Big)^2} = \sqrt{\frac{\widehat{\mathrm{SSE}}}{n-2}}$$

We lose **two** degrees of freedom for estimating two parameters (the intercept and the slope).

- **For multiple independent variables, $x_1, x_2, \ldots, x_p$, since we need to estimate $p+1$ parameters (+1 for the intercept), the degrees of freedom of the estimated $\widehat{\sigma}$ is $n-(p+1)$.**
- In summary, we have

$$\widehat{SE}(\hat{b}_1) = \sqrt{\frac{1}{n-(p+1)} \frac{\sum_{i=1}^{n}\Big(y_i - (\hat{b}_0 + \hat{b}_1 \cdot x_{1,i})\Big)^2}{\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)^2}}$$

T-statistic

$\boxed{\mathrm{T}}$ The T-statistic for the estimated slope takes the form

$$T = \frac{\hat{b}_1 - b_1}{\widehat{SE}(\hat{b}_1)} \sim t_{n-(p+1)}$$

where

$$\widehat{SE}(\hat{b}_1) = \frac{\widehat{\sigma}}{\sqrt{\text{sum of squared deviations in } x_1}} = \sqrt{\frac{1}{n-(p+1)} \frac{\text{sum of squared residual}}{\text{sum of squared deviations in } x_1}}$$

- For simple linear regression models, $p = 1$

Pearson's data

- Estimate $\widehat{\sigma}$ and $\widehat{SE}(\hat{b}_1)$

```r
n = length(x1)  # sample size
n
```

```
[1] 1078
```

```r
sse = sum(model$residual^2)
sig.hat = sqrt(sse/(n - 2))  # estimated SD of the error model
round(sig.hat, 3)
```

```
[1] 2.437
```

```r
dev.x = x1 - mean(x1)
sqrt.sum.sq.dev.x = sqrt(sum(dev.x^2))  # sqrt of sum of squared deviations in x1
est.se = sig.hat/sqrt.sum.sq.dev.x
round(est.se, 5)
```

```
[1] 0.02705
```

- Calculate the coefficient of determination (r^2)

```r
dev.y = y - mean(y)
sum.sq.dev.y = sum(dev.y^2)  # sum of squared deviations in y
1 - sse/sum.sq.dev.y
```

```
[1] 0.2513401
```

- Calculate observed test statistic and two-sided P-value

```
b1.hat = model$coefficients[2]  # the second parameter is the slope
stat = b1.hat/est.se
round(stat, 2)
```

```
   x1
19.01
```

```
p.value = 2 * pt(abs(stat), df = n - 2, lower.tail = F)
p.value
```

```
          x1
1.121268e-69
```

- P-value for the slope is close to zero, which rejects H_0 for most commonly used false alarm rates;
  - ➟ indirectly suggests that there is linear relationship between x (father's height) and Y (son's height) in the population.

Use `summary(model)`

```r
summary(model)  # where model = lm (y ~ x1)
```

```

Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.8772 -1.5144 -0.0079  1.6285  8.9685 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 33.88660    1.83235   18.49   <2e-16 ***
x1           0.51409    0.02705   19.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.437 on 1076 degrees of freedom
Multiple R-squared:  0.2513,	Adjusted R-squared:  0.2506 
F-statistic: 361.2 on 1 and 1076 DF,  p-value: < 2.2e-16
```

- **2nd row below** `Coefficients` shows the slope $\hat{b}_1$, $\widehat{SE}(\hat{b}_1)$, observed T-statistics, and two-sided P-value
- **3rd row from the bottom** shows
  - the estimated SD of the error model, $\widehat{\sigma}$, which is the SE of the residual ε_i;
  - and the degrees of freesom $n - (p+1)$.
- **2nd row from the bottom** shows the `Multiple R-squared`, which is the coefficient of determination

Confidence intervals for regression corfficients

- Confidence intervals (e.g., 99%) for regression corfficients can be constructed in the usual way
- Find (symmetric) multipliers $-\ell = u$ such that

$$P\left(\ell \leq \frac{\hat{b}_1 - b_1}{\widehat{SE}(\hat{b}_1)} \leq u\right) = 0.99, \quad T = \frac{\hat{b}_1 - b_1}{\widehat{SE}(\hat{b}_1)} \sim t_{n-(p+1)}$$

- After rearragement, we have

$$P\left(\hat{b}_1 - u \times \widehat{SE}(\hat{b}_1) \leq b_1 \leq (\hat{b}_1 + u \times \widehat{SE}(\hat{b}_1)\right) = 0.99$$

- u is given as the 99.5% quantile (the 0.5% percentage point in the upper tail)

```
u = qt(0.995, df = n - 2)
u
```

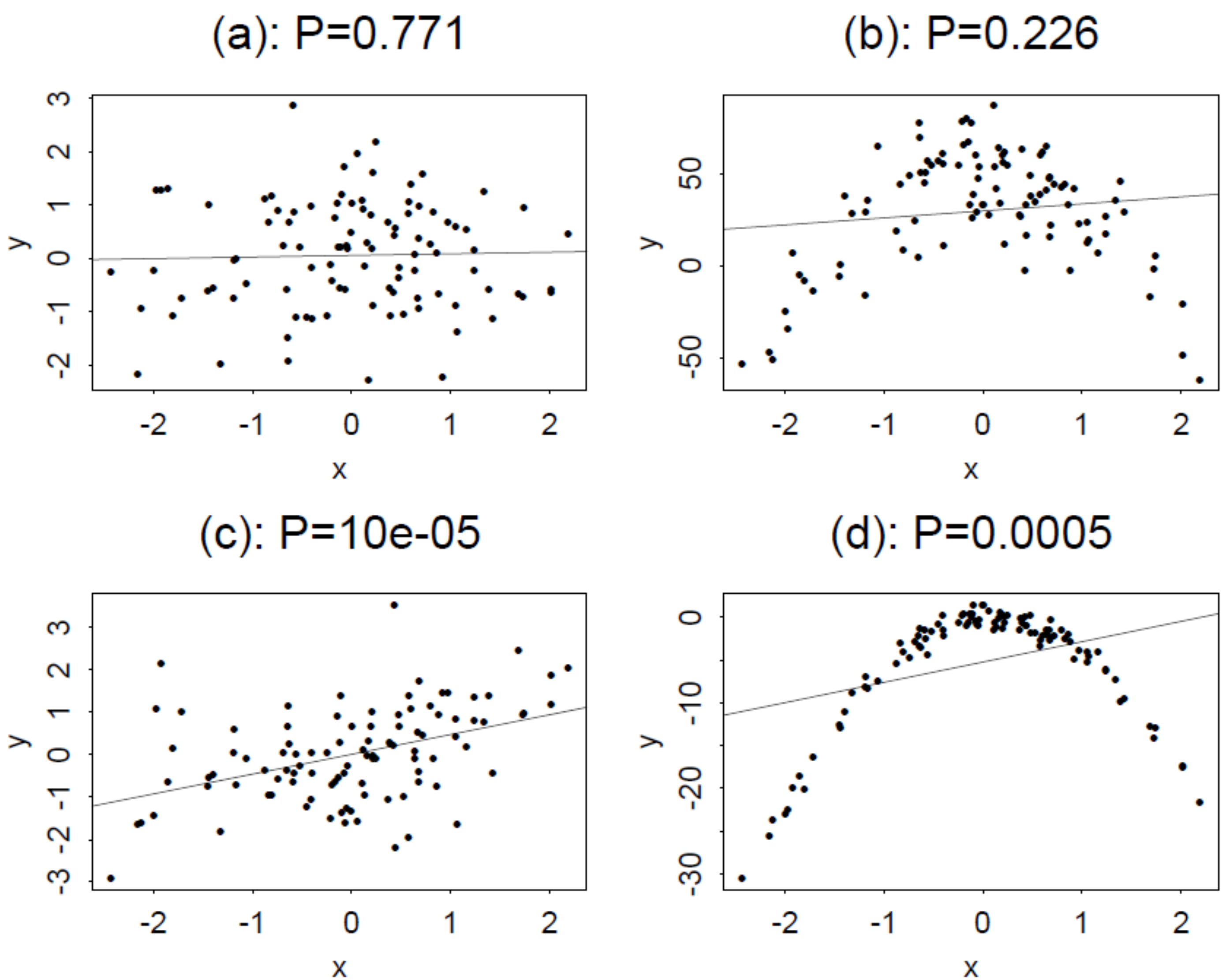
```
[1] 2.580406
```

```
round(b1.hat + c(-1, 1) * u * est.se, 3)
```

```
[1] 0.444 0.584
```

The C.I. does not contain zero (again, reject H_0 at the 1% level of significance).

P-values mean nothing if you haven't looked your data



Case study

Air pollution

The data frame `environmental` has four environmental variables taken in New York City from May to September of 1973:

- `ozone` concentration (part per billion), solar `radiation` (langley), maximum daily `temperature` (Fahrenheit) and `wind` speed (mile per hour)

```r
data("environmental", package = "lattice")
dim(environmental)
```

```
[1] 111   4
```

```r
str(environmental)
```

```
'data.frame':	111 obs. of  4 variables:
 $ ozone      : num  41 36 12 18 23 19 8 16 11 14 ...
 $ radiation  : num  190 118 149 313 299 99 19 256 290 274 ...
 $ temperature: num  67 72 74 62 65 59 61 69 66 68 ...
 $ wind       : num  7.4 8 12.6 11.5 8.6 13.8 20.1 9.7 9.2 10.9 ...
```

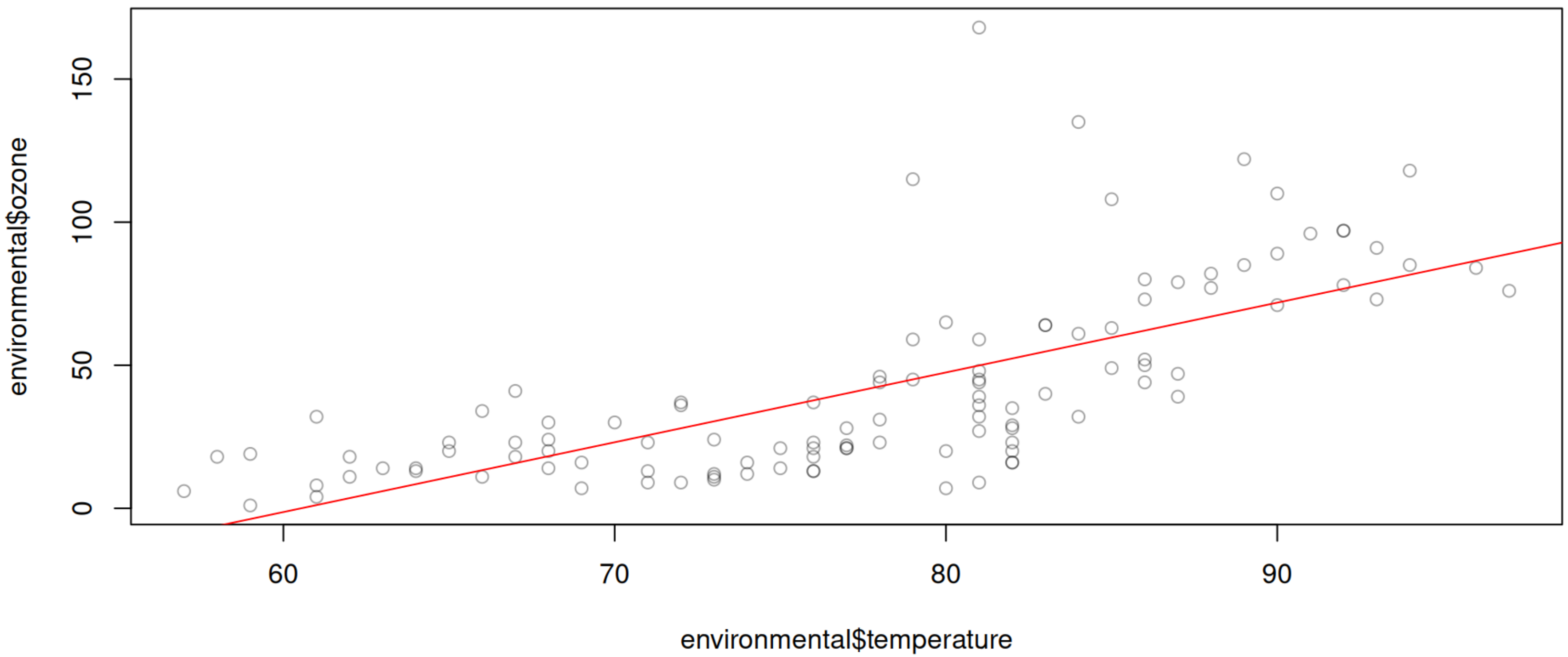
We'd like to assess whether the maximum daily temperature (x) has an influence on average ozone concentration (Y). - Let's use the 1% level of significance.

$\boxed{\text{H}}$ for the simple linear regression model: $Y_i = b_0 + b_1 \cdot x_{1,i} + \varepsilon_i$

- $H_0 : b_1 = 0$ – temperature has no linear association with ozone concentration
- $H_1 : b_1 \neq 0$ – temperature has a linear association with ozone concentration

A checking assumptions

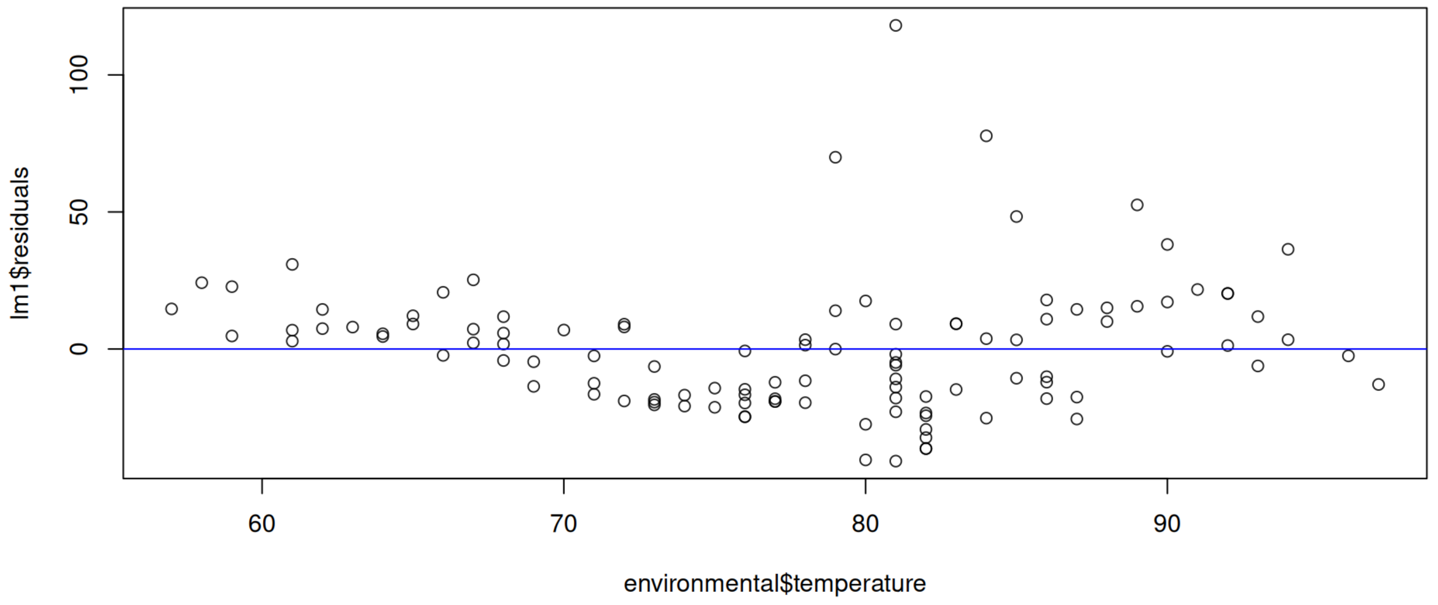
```r
plot(environmental$temperature, environmental$ozone, col = adjustcolor("black", alpha.f = 0.35))
lm1 = lm(ozone ~ temperature, environmental)
abline(lm1, col = "red")
```



- `lm(ozone ~ temperature, data=environmental)` fits a linear regression model with response variable `ozone`, explanatory variable `temperature`, and both are taken from the data frame `environmental`

A linearity

```r
plot(environmental$temperature, lm1$residuals, col = adjustcolor("black", alpha.f = 0.85))
abline(h = 0, col = "blue")
```



The residuals are above zero for low temperatures, then they go below zero for moderate temperatures, and end up again above zero for high temperatures.

- Our predictions are **systematically wrong** for certain ranges of temperature: **underestimate** the ozone level for low and high temperatures and **overestimate** the ozone level at moderate temperatures.

Transformation

- If the linearity assumption fails, there's not much point checking the other assumptions because it's not an appropriate prediction model.
- If we see a non-linear relationship between y and x we might be able to transform the data so that we have a linear relationship between the transformed variable(s).
    - ➟ What if we considered the log of ozone concentration?

```r
env.new = environmental  # create a new data frame
env.new[, "log.ozone"] = log(environmental$ozone)  # add a new variable log.ozone
env.new[, "ozone"] = NULL  # delete the old variable ozone
```

```r
lm2 = lm(log.ozone ~ temperature, env.new)
lm2
```

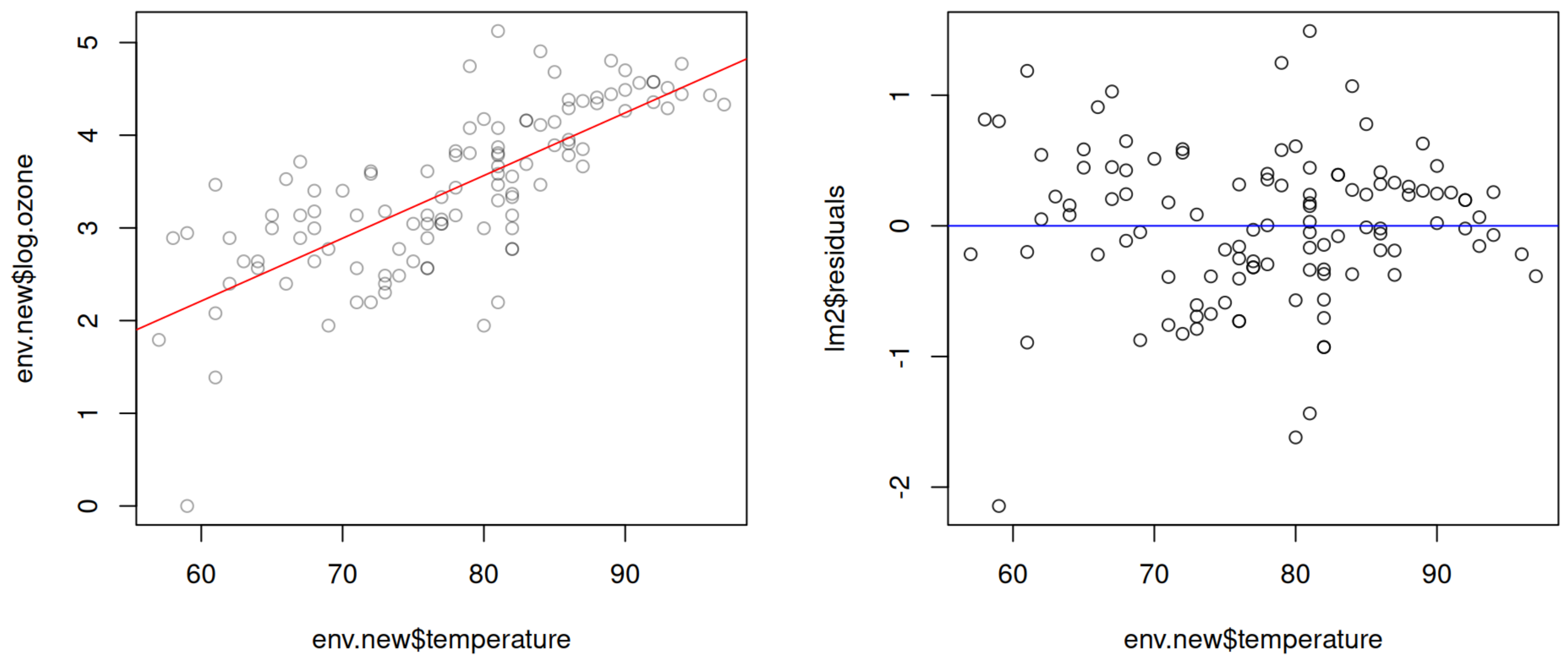
```

Call:
lm(formula = log.ozone ~ temperature, data = env.new)

Coefficients:
(Intercept)  temperature
   -1.84852      0.06767
```

Now the fitted model is:

$$\log\widehat{(\text{ozone})} = \underbrace{-1.84852}_{\hat{b}_0} + \underbrace{0.06767}_{\hat{b}_1} \times \text{temperature}$$

linearity

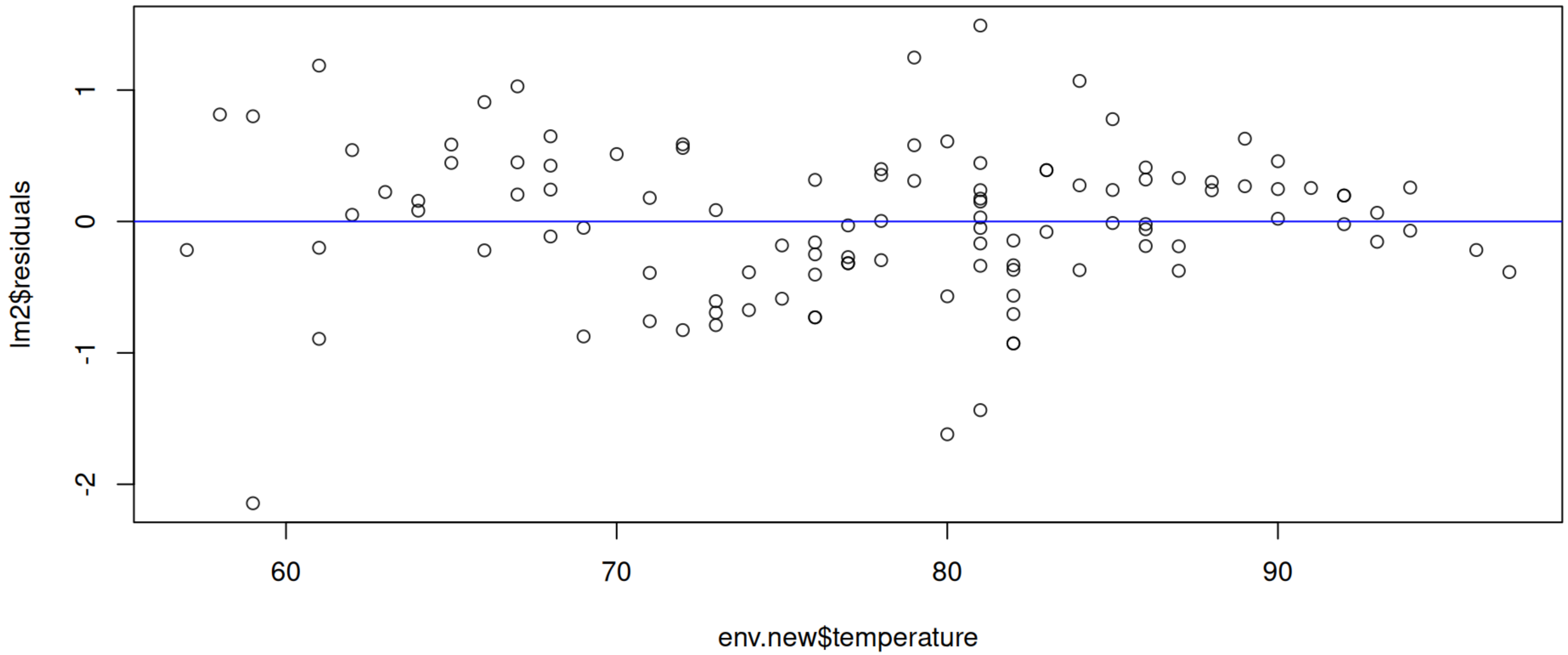
```r
par(mfrow = c(1, 2))
plot(env.new$temperature, env.new$log.ozone, col = adjustcolor("black", alpha.f = 0.35))
abline(lm2, col = "red")
plot(env.new$temperature, lm2$residuals, col = adjustcolor("black", alpha.f = 0.85))
abline(h = 0, col = "blue")
```



- No more over- and under-estimates. It seems that linearity holds between $\log(\text{ozone})$ and temperature.

A homoscedasticity

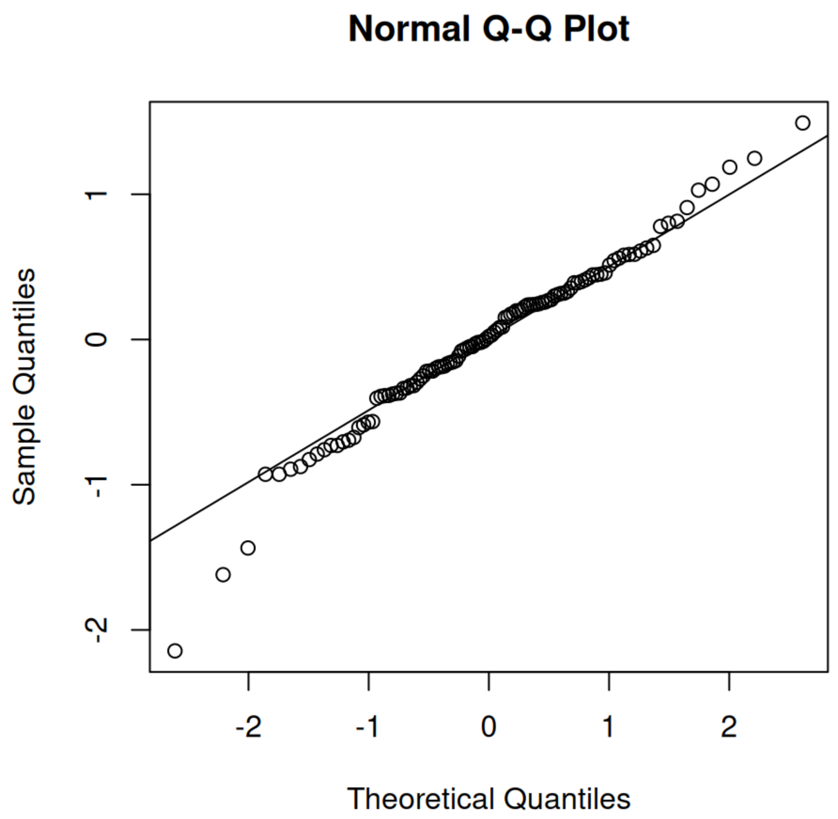
```r
plot(env.new$temperature, lm2$residuals, col = adjustcolor("black", alpha.f = 0.85))
abline(h = 0, col = "blue")
```



- Is the data homoscedastic? The spread looks reasonably constant over the range of temperature values.
  - However, in the region above 85°F, the spread might be somewhat smaller than the spread in the region below 85°F.

A normality

```r
qqnorm(lm2$residual)
qqline(lm2$residual)
```



- Apart from three points in the lower tail, the majority of the points lie quite close to QQ line. Hence, the normality assumption for the residuals is reasonably well satisfied.

How can we interpret the estimated coefficients?

$$Y_i = b_0 + b_1 x_{1,i} + \varepsilon_i$$

- The intercept is the expected value of Y_i when $x_1 = 0$.
- For a one unit increase in x_1 we expect Y to change by the slope b_1 (could be an increase or decrease depending on the sign).
- However, recall our fitted model

$$\widehat{\log(\text{ozone})} = \underbrace{-1.84852}_{\hat{b}_0} + \underbrace{0.06767}_{\hat{b}_1} \times \text{temperature}$$

- How do we interpret this model?

Slope interpretation for log-transform

$$\widehat{\log(\text{ozone})} = -1.84852 + 0.06767 \times \text{temperature}$$

- Consider two temperatures: $\text{temperature}_2 - \text{temperature}_1 = 1$, their corresponding predicted log ozone values have the difference

$$\widehat{\log(\text{ozone})}_2 - \widehat{\log(\text{ozone})}_1 = 0.06767 \times (\text{temperature}_2 - \text{temperature}_1) \approx 0.07,$$

  ➟ Interpreting the slope: a one degree increase in temperature results in a 0.07 unit **increase** in log ozone, on average.

- The ratio between two ozone readings can be approximated by

$$\frac{\widehat{\text{ozone}}_2}{\widehat{\text{ozone}}_1} = \exp\left(\widehat{\log(\text{ozone})}_2 - \widehat{\log(\text{ozone})}_1\right) \approx \exp(0.07) \approx 1 + 0.07$$

- A nicer way to interpret this is: a one degree increase in temperature results in an approximate 7% **increase** in ozone, on average.
- In general, for log-linear models $\log(Y) = b_0 + b_1 \cdot x_1$,

  ➟ On average, a one unit increase in x_1 will result in a $b_1\%$ change in Y (only works for small b_1).

Inference on the slope coefficient

```
summary(lm2)
```

```

Call:
lm(formula = log.ozone ~ temperature, data = env.new)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.14417 -0.32555  0.02066  0.34234  1.49100 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.848518   0.455080  -4.062  9.2e-05 ***
temperature  0.067673   0.005807  11.654  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5804 on 109 degrees of freedom
Multiple R-squared:  0.5548,	Adjusted R-squared:  0.5507 
F-statistic: 135.8 on 1 and 109 DF,  p-value: < 2.2e-16
```

$\boxed{\text{T}}$ observed T-statistics $t = 11.654$, d.f. $= 109$

$\boxed{\text{P}}$ the P-value is $< 2e - 16$ for the two-sided alternative

$\boxed{\text{C}}$ We reject H_0 at the 1% level of significance, suggesting there is a linear association between log ozone and temperature

Multiple linear regression models

Air pollution

The coefficient of determination (r^2) in the ozone example is 0.5548. - We can say that temperature explains 55% of the observed variation in the logarithm of ozone concentration.

> Can we do better if we use more variables to help explain the logarithm of ozone concentration?

```r
dim(env.new)
```

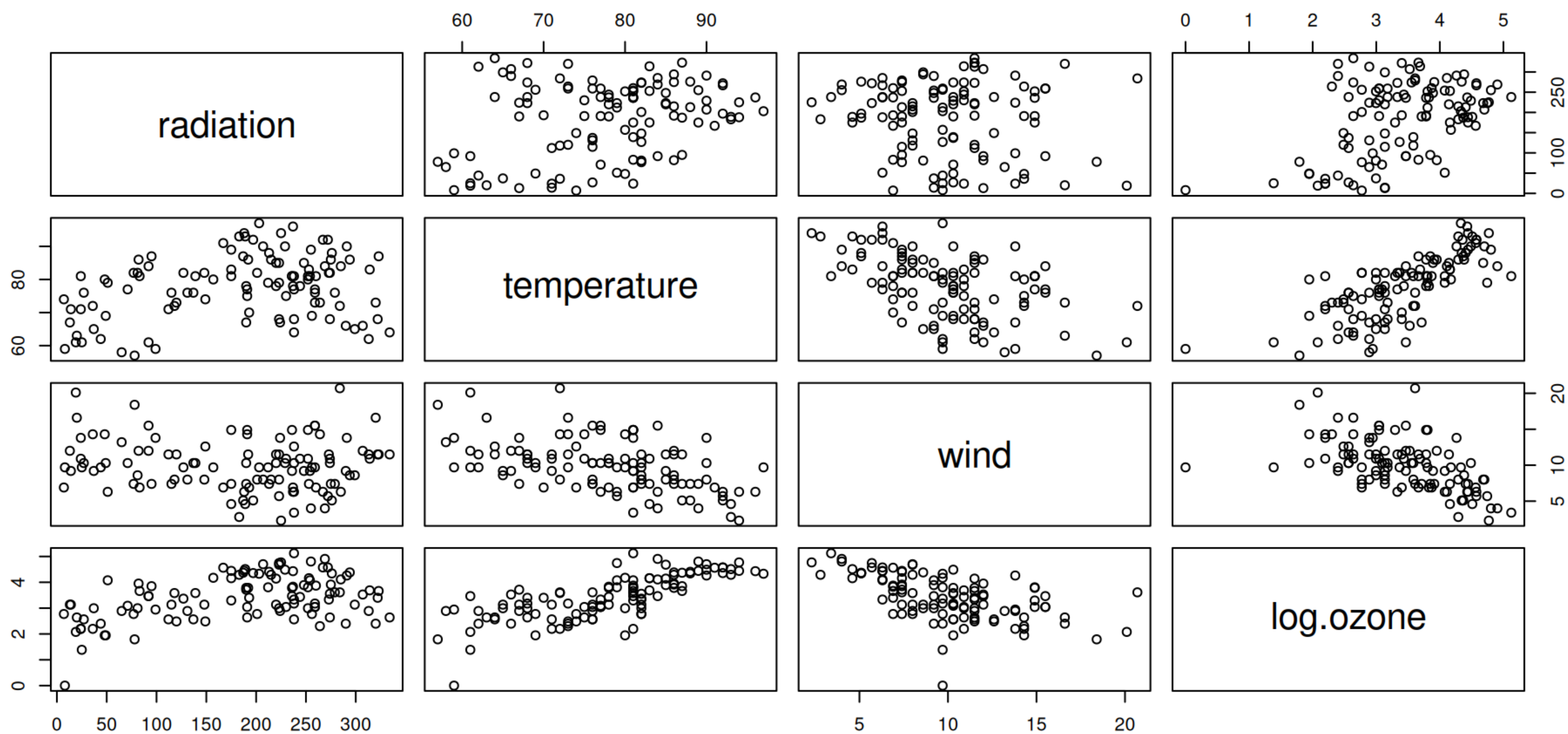
```
[1] 111   4
```

```r
str(env.new)
```

```
'data.frame':	111 obs. of  4 variables:
 $ radiation  : num  190 118 149 313 299 99 19 256 290 274 ...
 $ temperature: num  67 72 74 62 65 59 61 69 66 68 ...
 $ wind       : num  7.4 8 12.6 11.5 8.6 13.8 20.1 9.7 9.2 10.9 ...
 $ log.ozone  : num  3.71 3.58 2.48 2.89 3.14 ...
```

Pairwise scatter plot

```r
pairs(env.new)
```



Pairwise correlation

```r
round(cor(env.new), 2)
```

```
            radiation temperature  wind log.ozone
radiation        1.00        0.29 -0.13      0.46
temperature      0.29        1.00 -0.50      0.74
wind            -0.13       -0.50  1.00     -0.56
log.ozone        0.46        0.74 -0.56      1.00
```

- The variable `log.ozone` appears to be positively associated with `temperature`, negatively associated with `wind`, and and (moderately) positively associated with `radiation`.

Model

Can `radiation`, `temperature` and `wind` be used to predict `log.ozone`?

$$\log(\text{ozone})_i = b_0 + b_1 \cdot \text{radiation}_i + b_2 \cdot \text{temperature}_i + b_3 \cdot \text{wind}_i + \varepsilon_i$$

```r
lm3 = lm(log.ozone ~ radiation + temperature + wind, env.new)
round(summary(lm3)$coefficients, 3)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.261      0.553  -0.472    0.638
radiation      0.003      0.001   4.518    0.000
temperature    0.049      0.006   8.078    0.000
wind          -0.062      0.016  -3.922    0.000
```

Fitted model:

$$\widehat{\log(\text{ozone})} = -0.261 + 0.003 \cdot \text{radiation} + 0.049 \cdot \text{temperature} - 0.062 \cdot \text{wind}$$

Multiple linear regression model

Multiple linear regression is a natural extension of simple linear regression that incorporates multiple independent (or explanatory) variables. It has the general form,

$$Y_i = b_0 + b_1 \cdot x_{1,i} + b_2 \cdot x_{2,i} + \ldots + b_p \cdot x_{p,i} + \varepsilon_i, \text{ where } \varepsilon_i \sim \text{(iid) } N(0, \sigma^2).$$

- The same assumption on ε_i as in the simple linear regression case.

Often it's convenient to write the model in matrix format,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)'$, $\boldsymbol{\beta} = (b_0, b_1, b_2, \ldots, b_p)'$, $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I})$ and

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \ldots & x_{p,1} \\ 1 & x_{1,2} & x_{2,2} & \ldots & x_{p,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1,n} & x_{2,n} & \ldots & x_{p,n} \end{bmatrix},$$

is the **design matrix** depending on observed independent variables, where $\boldsymbol{x}_i' = (1, x_{1,i}, x_{2,i}, \ldots, x_{p,i})$ is the vector of independent variables for the $\boldsymbol{i}$th observation.

Fitting a multiple linear regression model

The optimal fit (least squares solution) is:

$$\begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_p \end{bmatrix} = \widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

which gives the coefficients that minimise the sum of squared residuals $\sum_{i=1}^{n} = e_i^2$ where the residual is defined as

$$e_i = y_i - \underbrace{(\hat{b}_0 + \hat{b}_1 \cdot x_{1,i} + \hat{b}_2 \cdot x_{2,i} + \ldots + \hat{b}_p \cdot x_{p,i})}_{\text{fitted regression model } \hat{y}_i}$$

- We will only consider using R to solve this for obtaining the estimated regression coefficients.

Interpretation

The estimated coefficients ($\hat{b}$'s) are now interpreted as **conditional on** the other variables

- each $\hat{b}_j$ reflects the predicted change in y associated with a one unit increase in the independent variable x_j, holding the other variables constant.

$$\widehat{\log(\text{ozone})} = -0.261 + 0.003 \cdot \text{radiation} + 0.049 \cdot \text{temperature} - 0.062 \cdot \text{wind}$$

- A one degree (Fahrenheit) increase in temperature results in a 4.9% **increase** in ozone on average, holding radiation and wind speed constant.
- A one langley increase solar radiation results in a 0.3% **increase** in ozone on average, holding radiation and wind constant.
- A one mile per hour increase in average wind speed results in a 6.2% **decrease** in ozone on average, holding radiation and temperature constant.

Coefficient of determination

The coefficient of determination (r^2) value has the same interpretation: proportion of total variability in Y explained by the regression model.

- Simple linear regression model

```
summary(lm2)$r.squared
```

```
[1] 0.5547615
```

- "Full" model

```
summary(lm3)$r.squared
```

```
[1] 0.664515
```

- Including more parameters can better explain the dependent variable.
- Note that for multiple linear regression, we can use $1 - \frac{\widehat{\mathrm{SSE}}}{\widehat{\mathrm{SST}}}$ to calculate the coefficient of determination.

```
SSE = sum(lm3$residuals^2)
SST = sum((env.new$log.ozone - mean(env.new$log.ozone))^2)
1 - SSE/SST
```

```
[1] 0.664515
```

- However, we cannot simply sum over the squared correlation coefficients; in fact, it is the squared correlation between the fitted model $\hat{y}_i$ and the observed data y_i (let's skip this).

Inference on regression coefficients

- We can also apply the T test to regression coefficients of multiple regression models.

$$Y_i = b_0 + b_1 \cdot x_{1,i} + b_2 \cdot x_{2,i} + \ldots + b_p \cdot x_{p,i} + \varepsilon_i, \text{ where } \varepsilon_i \sim \text{(iid) } N(0, \sigma^2).$$

- The T-test aims at testing if independent variable x_j has a significant linear relationship with the dependent variable Y, **after adjusting for all other independent variables in the model.**
  - ➡ In other words, after considering all other independent variables $1, x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p$ (as well as the intercept), we want to test if there is a linear relationship between x_j and Y.
- Equivalently, taking the effect of all other independent variables out of the dependent variable

$$U_i = Y_i - (b_0 + b_1 \cdot x_{1,i} + \cdots + b_{j-1} \cdot x_{j-1,i} + b_{j+1} \cdot x_{j+1,i} + \cdots + b_p \cdot x_{p,i})$$

we have the new model

$$U_i = b_j \cdot x_{j,i} + \varepsilon_i, \text{ where } \varepsilon_i \sim \text{(iid) } N(0, \sigma^2)$$

and want to test if there is a linear relationship between x_j and U.

Let's consider `wind` (x_3) and a two-sided alternative as an example, using the 1% level of significance.

- $\boxed{\text{H}}$ hypotheses
  - $H_0 : b_3 = 0$ – after adjusting for all other independent variables, there is no linear relationship between `wind` and `log.ozone`
  - $H_1 : b_3 \neq 0$ – after adjusting for all other independent variables, there is a linear relationship between `wind` and `log.ozone`

- $\boxed{\mathrm{T}}$ The test statistic is

$$T = \frac{\hat{b}_3 - b_3}{\widehat{SE}(\hat{b}_3)} \sim t_{n-(p+1)}$$

where $b_3 = 0$ under H_0 and $p = 3$; however, the estimated standard error takes a different form

$$\widehat{SE}(\hat{b}_3) = \hat{\sigma} \times \sqrt{[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{33}}$$

- ➠ as before, we have the estimated SD of the residual error

$$\hat{\sigma} = \sqrt{\frac{1}{n-(p+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n-(p+1)} SSE}$$

- ➠ the term $[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{33}$ is the last element of the matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$
  - ➠ is analogous to $1/\sqrt{\sum_{i=1}^{n}(x_{1,i} - \bar{x}_1)^2}$ in the simple linear regression case (see Page 19);
  - ➠ but counting the linear dependency among all independent variables (the derivation is beyond the scope here)
- ➠ so, we rely on the R function `summary()` to work this out.

```r
dim(env.new)
```

```
[1] 111   4
```

```r
summary(lm3)
```

```

Call:
lm(formula = log.ozone ~ radiation + temperature + wind, data = env.new)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.06212 -0.29968 -0.00223  0.30767  1.23572 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.2611739  0.5534102  -0.472 0.637934    
radiation    0.0025147  0.0005567   4.518 1.62e-05 ***
temperature  0.0491630  0.0060863   8.078 1.07e-12 ***
wind        -0.0615925  0.0157037  -3.922 0.000155 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5085 on 107 degrees of freedom
Multiple R-squared:  0.6645,	Adjusted R-squared:  0.6551 
F-statistic: 70.65 on 3 and 107 DF,  p-value: < 2.2e-16
```

- The estimated SE for $\hat{b}_3$ is $\mathbf{0.0157037}$.
- The observed T-statistic is $\mathbf{-3.922}$ and the degrees of freedom is $n - (p+1) = \mathbf{107}$.
- $\boxed{\text{P}}$ The corresponding two-sided P-value is $\mathbf{0.000155}$.
- $\boxed{\text{C}}$ We reject H_0 at the 1% level of significance, indirectly suggesting there is a linear relationship between `wind` and `log.ozone`, after adjusting for all other independent variables.

Warning: dangers of multicollinearity

When some of the independent variables are highly correlated (multicollinearity) then we can find that the fitted multiple regression models can have

- $\hat{b}_j$ coefficients with counter-intuitive signs;
- terms with large estimated standard errors $\widehat{SE}(\hat{b}_j)$; and
- rather large (counter-intuitive) P-values.

Often removing some of the independent variables

- changes all of the above with very little impact on the coefficient of determination (r^2)
- This comes from the fact that $\hat{b}_j$ reflects the additional information provided by variable x_j given that all the other variables have been fitted.

See the lab today and more examples next week.

Assessment expectations

Assessment expectations

- Simple linear regression
  - ➠ Know how to work out the slope and intercept
  - ➠ Know how to work out the coefficient of determination r^2 given the correlation coefficient
- Multiple linear regression
  - ➠ We will rely on R outputs for both the regression coefficients and r^2
- T-test for simple and multiple linear regression
  - ➠ We don't expect you to work out standard errors of the estimated regression coefficients by hand;
  - ➠ but given the estimated regression coefficients and standard errors, you should be able to work out the test statistic and P-value.
  - ➠ Know how to get the confidence interval for given standard errors.
  - ➠ Know how to get the degrees of freedom, $n-(p+1)$, for estimating the SD of the residual error, and hence the degrees of freedom to be used in Student's t-distribution.