Experimental Design and Data Analysis, Lecture 7

Eduard Belitser

VU Amsterdam

Lecture Overview

1. contingency tables
    1. chisquare test
    2. Fisher test
2. simple linear regression
3. multiple linear regression

contingency tables

Setting

An experiment with:

- a count of individuals or units in different categories of two factors.

Interest is in a possible dependence of the two factors.

EXAMPLE Study possible dependency between blood group and disease by counting the number of patients having a certain blood group (A, B or O) and a certain disease (stomach cancer, kidney cancer, no disease).

EXAMPLE Study possible dependency between web layout and size of a company by counting the number of companies of a certain size (small, moderate, large) using a certain web design (relative, fixed, elastic, liquid) .

EXAMPLE Consider the following (fictive) counts amongst 60 VU-students:

	exact	arts	total
men	23	17	40
women	7	13	20
total	30	30	60

Question: `study` and `gender` independent?

Design

Design A:

- Take a random sample of experimental units from the relevant population.
- Count for each cross-category the number of units falling into that cross-category.

Design B:

- Take for each category of the first (row) factor a random sample of experimental units.
- Count for each category of the second factor the number of units falling into that cross-category.

Design C:

- Take for each category of the second (column) factor a random sample of experimental units.
- Count for each category of the first factor the number of units falling into that cross-category.

Analysis (1)

The general form of a contingency table is

n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\cdot}$
n_{21}	n_{22}	$\cdots$	n_{2J}	$n_{2\cdot}$
$\vdots$		$\ddots$	$\vdots$	$\vdots$
n_{I1}	n_{I2}	$\cdots$	n_{IJ}	$n_{I\cdot}$
$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot J}$	$n_{\cdot\cdot}$

We want to test whether the two factors are independent (under design A):

$$H_0 : \textit{row variable and column variable are independent}.$$

Or, we want to test whether the distributions are homogeneous over rows (design B) or columns (design C):

$$H_0 : \textit{the distributions over row (column) factors are equal}.$$

Analysis (2)

Let $n = n_{\cdot\cdot}$ be the total number of observaions. Under the null hypothesis of no dependence (or homogeneity), the counts are expected to be in proportion:

$$E_{ij} = np_{ij} = np_{i\cdot}p_{\cdot j} = n\frac{n_{i\cdot}}{n}\frac{n_{\cdot j}}{n} = \frac{n_{i\cdot}n_{\cdot j}}{n}.$$

Expected counts in the example data set:

	exact	arts	total
men	?	?	40
women	?	?	20
total	30	30	60

$\Longrightarrow$

	exact	arts	total
men	$60 \cdot \frac{40}{60} \cdot \frac{30}{60}$	$60 \cdot \frac{40}{60} \cdot \frac{30}{60}$	40
women	$60 \cdot \frac{20}{60} \cdot \frac{30}{60}$	$60 \cdot \frac{20}{60} \cdot \frac{30}{60}$	20
total	30	30	60

The test statistic is based on the (appropriately normalized) differences between the expected counts E_{ij} under H_0 and the observed counts n_{ij}:

$$T = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(I-1)(J-1)}, \quad \text{(approx. a chisquare distribution).}$$

The p-value is always right-sided: $p_{right} = P(T > t)$. Why?
Condition: For the test to be reliable, at least 80% of the E_{ij}'s should be at least 5.

In R: `chisq.test(data)`

Analysis in R — data input

First, we need to create a table of the counts in the form of a matrix.

The following data consists of grade counts in an elementary statistics class, classified by the students' majors.

```
> grades=matrix(c(8,15,13,14,19,15,15,4,7,3,1,4),byrow=TRUE,ncol=3,nrow=4,
+ dimnames=list(c("A","B","C","D-F"),c("Psychology","Biology","Other")))
> grades
    Psychology Biology Other
A            8      15    13
B           14      19    15
C           15       4     7
D-F          3       1     4
```

For the calculations on the next slide, *R* needs the data in a `matrix` object, rather than in a `table` or `dataframe` format.

Analysis in R — testing (1)

```
> rowsums=apply(grades,1,sum); colsums=apply(grades,2,sum)
> total=sum(grades); expected=(rowsums%*%t(colsums))/total
> round(expected,0)
     Psychology Biology Other
[1,]         12      12    12
[2,]         16      16    16
[3,]          9       9     9
[4,]          3       3     3
> sum((grades-expected)^2/expected) #realization of statistics T
[1] 12.18346
> 1-pchisq(12.18346,6)   #p-value for the observed T=12.18346
[1] 0.05799897
```

Less than 80% of the expected counts are above 5. Hence, the approximation by a chi-square test is not reliable.

Analysis in R — testing (2)

Of course, no need to perform all these computations, just use build-in R command: chisq.test, which executes the χ^2-test.

```
> z=chisq.test(grades); z
                      Pearson's Chi-squared test
data:  grades
X-squared = 12.1835, df = 6, p-value = 0.058

Warning message:
In chisq.test(grades) : Chi-squared approximation may be incorrect
```

R gives a warning because the chi-squared approximation in this case is not reliable. In such a case one can use the setting simulate.p.value=TRUE, which computes a p-value in a bootstrap fashion. This may yield a very different p-value.

```
> chisq.test(grades,simulate.p.value=TRUE)
Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  grades
X-squared = 12.1835, df = NA, p-value = 0.05647
```

Analysis in R — testing (3)

You can extract information from z=chisq.test(grades): z$expected gives the table of expected values, z$observed recovers the observed values.
We can look at the (square root) contributions of each cell to the chi-squared statistics, by using residuals(z) (or z$residuals), to determine which observed values deviate most from the expected under H_0.

```
> residuals(z)  # = (z$observed-z$expected)/sqrt(z$expected)
    Psychology    Biology      Other
A   -1.2032599  0.8992005  0.3193881
B   -0.5630451  0.7872412 -0.2170232
C    2.0838439 -1.5668929 -0.5434979
D-F  0.1749697 -1.0110751  0.8338764
```

- From this table we see that psychology students have relatively more C's,
- biology students have relatively less C's,
- psychology students have relatively less A's,

than expected under H_0 (the differences are not significant though ($p \approx 0.06$)).

Alternatively, we can look at the standardized residuals using z$stdres (=(z$observed - z$expected)/sqrt(V), where V is the residual cell variance, see Agresti, 2007, section 2.4.5) and compare to $z_{\alpha/2}$=qnorm(0.975)=1.96.

Fisher's exact test for 2x2-tables

For 2x2-tables it is possible to compute an exact p-value, that does not use approximation or simulation. This is called Fisher's exact test.
Data on right- and left-handed people, classified according to gender.

```
> handed=matrix(c(2780,3281,311,300),nrow=2,ncol=2,byrow=TRUE,
+ dimnames=list(c("right-handed","other"),c("men","women")))
> handed
             men women
right-handed 2780  3281
left-handed   311   300
```

We can compare this to picking without replacement 3091 balls from a vase which contains 6672 balls, 6061 white and 611 red. The number of white balls amongst the picked 3091 balls is $n_{11} = 2780$.

n_{11}	...	6061
...	...	611
3091	3581	6672

$\Longrightarrow$

n_{11}	$6061 - n_{11}$
$3091 - n_{11}$	$3581 - (6061 - n_{11})$

The number n_{11} determines all other numbers. Fisher's exact test is based on this number. Under the null hypothesis of no dependence between the two factors it has a hypergeometric distribution.

Analysis in R — testing

```
> fisher.test(handed)
        Fisher's Exact Test for Count Data

data:  handed
p-value = 0.01918
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.6894895 0.9688105
sample estimates:
odds ratio
 0.8173619
> chisq.test(handed)
        Pearson's Chi-squared test with Yates' continuity correction

data:  handed
X-squared = 5.4542, df = 1, p-value = 0.01952
```

The chisquare approximation is also fine for these data. The odds ratio is computed as $\frac{2780/311}{3281/300} = 0.8173619$ and can be interpreted as "for one right-handed women there is ≈ 0.82 right-handed men", there are relatively more left handed men than women.

simple linear regression

Setting

An experiment with:

- a numerical outcome Y ("dependent variable"),
- a numerical explanatory variable X ("independent variable").

The purpose is to explain Y by a numerical function of X. Extrapolation to nonmeasured values of X is desirable.

EXAMPLE Chemical production process with outcome total yield and explanatory variable temperature.

EXAMPLE Educational study with outcome score on final exam and explanatory variable number of pupils per teacher.

EXAMPLE Quality of a genetic algorithm to determine the minimal value of a criterion function with outcome CPU time needed to find true minimum and explanatory variable mutation probability.

Design

- Fix a set of values X of the explanatory variable.
- Perform the corresponding experiments and measure the outcome Y.

It is natural to let the explanatory variable X vary over a grid of values in its range of interest.
Regression analysis is also often used in nonexperimental situations, with the explanatory variable not under control.

Analysis

Data

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N).$$

The simple linear regression model assumes that

$$Y_n = \beta_0 + \beta_1 X_n + e_n, \quad n = 1, 2, \ldots, N,$$

where errors $e_1, \ldots, e_N$ are viewed as a random sample from $N(0, \sigma^2)$.

We test the null hypothesis $H_0 : \beta_1 = 0$ that the explanatory variable does *not* influence the outcome. We also want to estimate the parameters β_0, β_1.

The function $x \mapsto \beta_0 + \beta_1 x$ is a line with intercept (value at $x = 0$) β_0 and slope (change per unit) β_1. This is a simple function and may give a bad fit!

Analysis in R — data input

The column `total` of the dataset `sat.txt` is the average score on the *scolastic aptitude test* of pupils in a US-state in 1994/95; the column `expend` is the amount of dollars spent per pupil in the state.

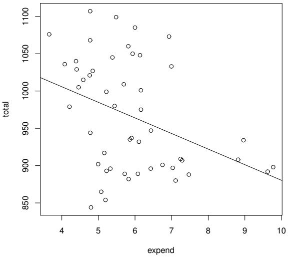
```
> sat=read.table("sat.txt",header=TRUE); sat1=sat[,c(1,7)]
> sat1[1:4,]
          expend total
Alabama    4.405  1029
Alaska     8.963   934
Arizona    4.778   944
Arkansas   4.459  1005
```

Analysis in R — graphics, estimation and testing

```
> sat1lm=lm(total~expend,data=sat1); summary(sat1lm)
[ some output deleted ]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1089.294     44.390  24.539  < 2e-16 ***
expend       -20.892      7.328  -2.851  0.00641 **
```

The parameters β_0 and β_1 are estimated to be 1089.294 and -20.892. The p-value for testing $H_0 : \beta_1 = 0$ is 0.00641. The slope is significantly negative!

```
> plot(total~expend,data=sat1)
> abline(sat1lm)
```



Compare to Pearson's correlation test

Compare simple linear regression to Pearson's correlation test (treated earlier) which tests whether the response and explanatory variable (in our case columns `total` and `expand`) are uncorreleted.

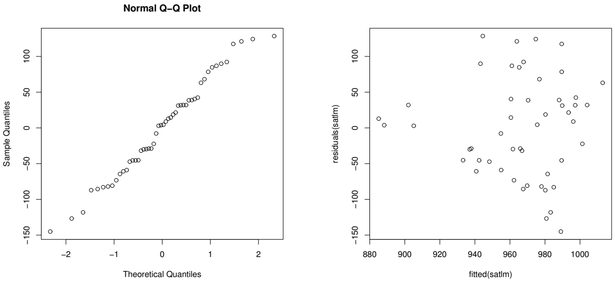
```
> cor.test(sat1$total,sat1$expend)

Pearson's product-moment correlation

data:  sat1$total and sat1$expend
t = -2.8509, df = 48, p-value = 0.006408
```

Notice that the p-value of the correlation test between response and covariate is equal to the p-value for testing the zero slope in simple linear regression. In fact this is the same test, $H_0 : \rho = 0$ is the same as $H_0 : \beta_1 = 0$.

Analysis in R — diagnostics

We can use the data to check whether the assumptions on the errors $e_n = Y_n - \beta_0 - \beta_1 X_n$ are not totally untrue.
The residuals are $\hat{e}_n = Y_n - \hat{\beta}_0 - \hat{\beta}_1 X_n$; the fitted values $\hat{Y}_n = \hat{\beta}_0 + \hat{\beta}_1 X_n$.
The residuals should look normal, and their spread should not vary with the fitted values.

```
> qqnorm(residuals(sat1lm))
> plot(fitted(sat1lm),residuals(sat1lm))
```



(The two plots look OK.)

multiple linear regression

Setting and design

Setting: an experiment with

- a numerical outcome Y ("dependent variable");
- p numerical explanatory variables $X_1, \ldots, X_p$ ("independent variables", "predictors").

The purpose is to explain Y by a numerical function of $X_1, \ldots, X_p$.

EXAMPLE Chemical production process with outcome total yield and explanatory variables temperature and pressure.

EXAMPLE Educational study with outcome score on final exam and explanatory variables teacher salaries and number of pupils per teacher.

Design:

- Fix a set of combinations $(X_1, \ldots, X_p)$ of explanatory variables.
- Perform the corresponding experiments and measure the outcome Y.

It is natural to let each explanatory variable vary over a grid and use all their possible combinations, but this may necessitate many experiments. (Regression analysis is also often used in non-experimental situations, with the explanatory variables not under control.)

Analysis

Data $Y_n, X_{n1}, X_{n2}, \dots, X_{np}$, $n = 1, \dots, N$. The linear regression model:

$$Y_n = \beta_0 + \beta_1 X_{n1} + \dots + \beta_p X_{np} + e_n, \quad n = 1, \dots, N, \quad (\text{matrix notation } Y = X\beta + e)$$

where errors $e_1, e_2, \dots, e_N$ are viewed as a random sample from $N(0, \sigma^2)$, $\beta_0, \dots, \beta_p$ are unknown population parameters,.

We test the null hypotheses $H_0 : \beta_j = 0$ that the jth explanatory variable does *not* influence the outcome for $j = 1, \dots, p$.
We also want to estimate the parameters β_j's.

Possible explanatory variables (prediction variables):

- all x_j different $Y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + e$,
- powers of x_j's $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + e$,
- interactions between x_j's $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e$.

Essential: all models are linear in the β_j's, but not necessarily in the x_j's.

Estimating parameters, SSE

To find the best parameters we minimize the sum of squared differences:

$$\min_{\beta_0,\ldots\beta_k} \sum_{n=1}^{N}(Y_n-\beta_0-\beta_1X_{n1}-\ldots-\beta_pX_{np})^2 = \sum_{n=1}^{N}(Y_n-\hat\beta_0-\hat\beta_1X_{n1}-\ldots-\hat\beta_pX_{np})^2 = SSE,$$

where $\hat\beta_0,\ldots,\hat\beta_p$ are the least squares estimates for the β's, the Sum of Squared Errors (SSE) and the estimated variance of the errors e_n are

$$SSE = \sum_{n=1}^{N}(Y_n-\hat\beta_0-\hat\beta_1x_{n1}-\ldots-\hat\beta_px_{nk})^2 = \sum_{n=1}^{N}\hat e_n^2, \quad \hat\sigma^2 = s^2 = \frac{SSE}{n-p-1}.$$

$\hat\sigma^2$ is the estimated variance of the errors e_n, $\hat e_n = Y_n - \hat\beta_0 - \hat\beta_1X_{n1} - \ldots - \hat\beta_pX_{np}$ is the n-th residual (the estimated error e_n of the n^{th} observation).

In R: `lm(y~x1+...+xp,data=...)`

Coefficient of determination R^2

- The coefficient of determination R^2 compares the models

$$Y = \beta_0 + e \qquad \text{and} \qquad Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + e.$$

- For the model on the left, $\hat{\beta}_0 = \bar{Y}$ with $SS_y = \sum_{i=n}^{N} (Y_n - \bar{Y})^2$.
- For the model on the right, we have already computed SSE.
- The coefficient of determination R^2 is

$$R^2 = \frac{SS_y - SSE}{SS_y} = \frac{\sum_{n=1}^{N} (Y_n - \bar{Y})^2 - \sum_{n=1}^{N} \hat{e}_n^2}{\sum_{n=1}^{N} (Y_n - \bar{Y})^2}.$$

- $0 \leq R^2 \leq 1$ because always $SS_y \geq SSE \geq 0$.
- R^2 is also called the proportion of explained variance.
- R^2 yields a global check on the multiple linear regression model. The higher R^2 the more variation the model explains.
- If $p = 1$, then $R^2 = r^2$ (the squared correlation between X_1 and Y).

$R^2 \approx 1$ means that the linear regression model can explain the measured response values Y very well using a linear function of the explanatory variables $(X_1, \ldots, X_p)$.
$R^2 \approx 0$ means that the linear model does not explain much.

Global model fit, relevance of individual coefficients

Test if the linear regression is adequate ($X_1, \dots, X_p$ together have significant explanatory power in the model) $H_0 : \beta_1 = \dots = \beta_p = 0$. The test statistic

$$T = \frac{R^2/k}{(1-R^2)/(n-(k+1))} \sim F_{k,n-(k+1)}, \quad \text{under } H_0.$$

The larger R^2, the larger T, the more evidence against H_0, hence we reject H_0 if T is large (R^2 is large). The test is always right-sided: if $p = P(T > t) < \alpha$, reject H_0. In R, this p-value is in the last line of `summary(lm(y~x))`.

Not all available explanatory variables have explanatory power. From all explanatory variables, we need to find relevant ones. Test $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$ for individual β_i's (usually two-sided). Test statistic: under H_0,

$$T_i = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}} \sim t_{n-(k+1)}, \quad \text{where } s^2_{\hat{\beta}_i} = \hat{\sigma}^2 \nu_{ii}, \ [\nu_{ij}] = (X^T X)^{-1}, \ Y = X\beta + e.$$

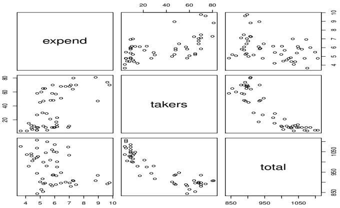
In R, the estimates $\hat{\beta}_i$, standard errors $s_{\hat{\beta}_i}$, the statistics values T_i and the p-values are all given in the output of `summary(lm(y~x))`.

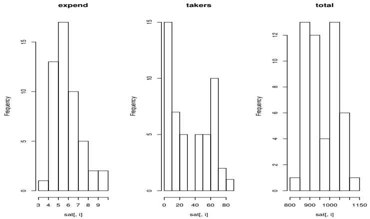
Analysis in R — data input and graphics

The dataset `sat.txt` concerns data on the Scholastic Aptitude Test (SAT) for pupils in the US in 1994/1995. The column `expend` contains the mean expenses per pupil (in $ per pupil), `ratio` is the pupil/teacher ratio, `salary` is the mean salary of teachers, `takers` is the percentage of pupils that takes the SAT. Variables `verbal` and `math` are partial scores of the total SAT score in `total` and not used in the analysis.

Create a data frame with one column of Y-values and p columns $X_1, \ldots, X_p$.

```
> sat[1:3,]
          expend ratio salary takers verbal math total
Alabama    4.405  17.2 31.144      8    491  538  1029
Alaska     8.963  17.6 47.951     47    445  489   934
> plot(sat[,c(1,4,7)]); par(mfrow=c(1,3))
> for (i in c(1,4,7)) hist(sat[,i],main=names(sat)[i])
```





Analysis in R — estimation and testing

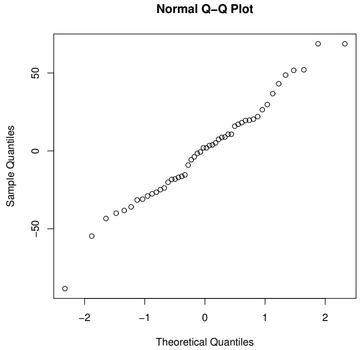
```
> satlm=lm(total~expend+takers,data=sat); summary(satlm)
[ some output deleted ]
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 993.8317    21.8332  45.519  < 2e-16 ***
expend       12.2865     4.2243   2.909  0.00553 **
takers       -2.8509     0.2151 -13.253  < 2e-16 ***
[ some output deleted ]
Residual standard error: 32.46 on 47 degrees of freedom
Multiple R-squared:  0.8195,Adjusted R-squared:  0.8118
F-statistic: 106.7 on 2 and 47 DF,  p-value: < 2.2e-16
```

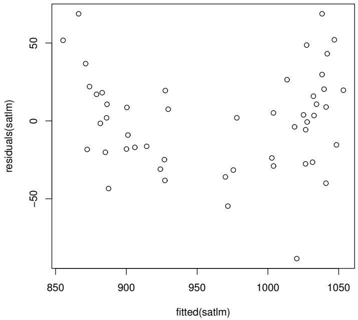
The estimates $\hat{\beta}_n$ are in the column `Estimate`. The $(1-\alpha)$ CI's for the β_i's are $\beta_i = \hat{\beta}_i \pm t_{\alpha/2,n-(k+1)} s_{\hat{\beta}_i}$, obtained in R by `confint(satlm)`. Since more explanatory variables always explain more, R^2 always increases with more variables. The R^2 adjusted for p predictors: $R^2_{adj} = 1 - \frac{n-1}{n-(p+1)}(1-R^2)$. The more variables, the more conservative R^2_{adj} becomes (as compared to R^2), it can be used to choose between models with different amounts of variables. But the interpretation of R^2_{adj} is not fraction of explained variance anymore.

Analysis in R — diagnostics

The residuals $\hat{e}_n = Y_n - \hat{\beta}_0 - \hat{\beta}_1 X_{n,1} - \cdots - \hat{\beta}_p X_{n,p}$ (in R: `residuals(model)`); the fitted values $\hat{Y}_n = \hat{\beta}_0 + \hat{\beta}_1 X_{n,1} + \cdots + \hat{\beta}_p X_{n,p}$ (in R: `fitted(model)`).

```
> qqnorm(residuals(satlm))
> plot(fitted(satlm),residuals(satlm))
```





The fitted-residuals plot has a Y-shape, whereas no specific shape should be seen.

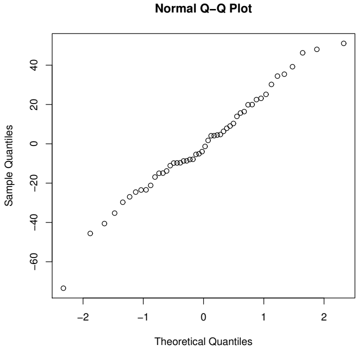
Analysis in R — estimation and testing

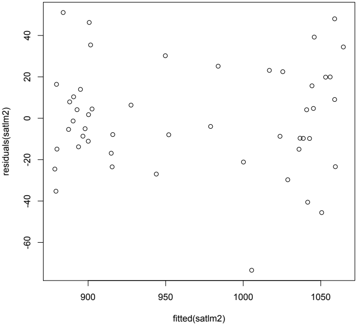
```
> sat$takers2=sat$takers^2
> satlm2=lm(total~expend+takers+takers2,data=sat)
> summary(satlm2)
[ some output deleted ]
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.052e+03  2.082e+01  50.511  < 2e-16 ***
expend       7.914e+00  3.498e+00   2.262   0.0285 *
takers      -6.381e+00  7.036e-01  -9.068 8.30e-12 ***
takers2      4.741e-02  9.161e-03   5.175 4.87e-06 ***
```

This fits a model that is quadratic in `takers`. The function $x \mapsto \alpha + \beta x + \gamma x^2$ is a parabola, one step up in complexity from a linear function.

Analysis in R — diagnostics

```
> qqnorm(residuals(satlm2))
> plot(fitted(satlm2), residuals(satlm2))
```





Both plots look OK.

If the assumptions fail?

One can consider:

- transforming the outcomes (e.g., use $\log Y$, Y^3).
- transforming the explanatory variables (e.g. use $\log X$, X^2).
- adding powers $X_i^2, X_i^3, \ldots$ of the regression variables.
- adding "interactions" like $X_i X_j$.
- performing nonparametric or additive regression.
- something else (there is no fix that always works).

To finish

Today we discussed:

- contingency tables
  - chisquare test
  - Fisher test
- simple linear regression
- multiple linear regression

Next time: more on linear regression.