Simple linear regression, chapter 4.6

Grethe Hystad

November 16, 2012

The data on the next two pages provides the average January minimum temperature in degrees Fahrenheit with the latitude and longitude of 56 U.S. cities from 1931-1960.
Source DASL at
http://lib.stat.cmu.edu/DASL/Datafiles/USTemperatures.html
with reference: Peixoto, J.L. (1990) A property of well-formulated polynomial regression models. American Statistician, 44, 26-30.
Also found in: Hand, D.J., et al. (1994) A Handbook of Small Data Sets, London: Chapman & Hall, 208-210.

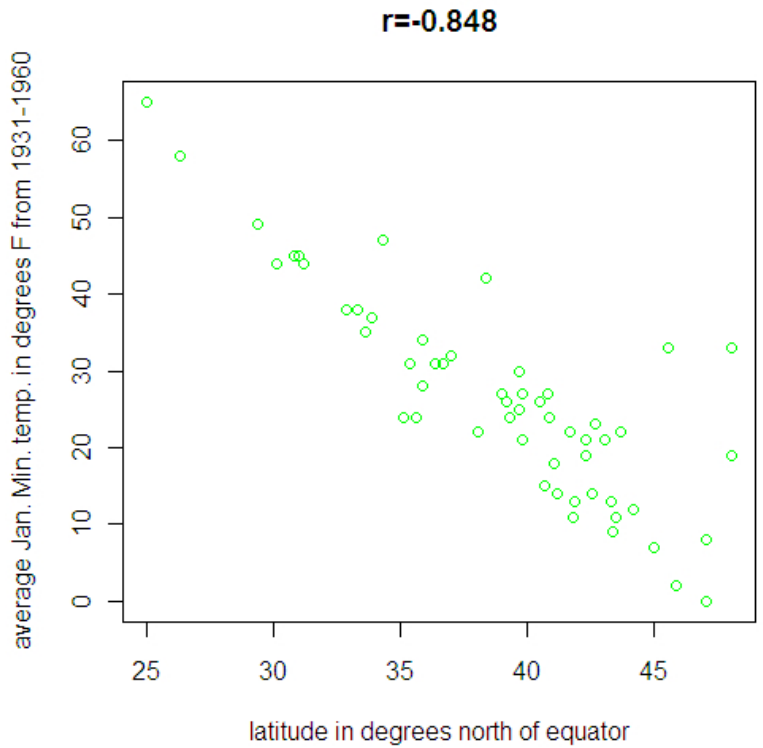
1. City: City, State postal abbreviation
2. JanTemp: Average January minimum temperature in degrees F. from 1931-1960
3. Lat: Latitude in degrees north of the equator
4. Long: Longitude in degrees west of the prime meridian

The Data:

City	JanTemp	Lat	Long	
Mobile, AL	44	31.2	88.5	
Montgomery, AL	38	32.9	86.8	
Phoenix, AZ	35	33.6	112.5	
Little Rock, AR	31	35.4	92.8	
Los Angeles, CA	47	34.3	118.7	
San Francisco, CA		42	38.4	123.0
Denver, CO	15	40.7	105.3	
New Haven, CT	22	41.7	73.4	
Wilmington, DE	26	40.5	76.3	
Washington, DC	30	39.7	77.5	
Jacksonville, FL		45	31.0	82.3
Key West, FL	65	25.0	82.0	
Miami, FL	58	26.3	80.7	
Atlanta, GA	37	33.9	85.0	
Boise, ID	22	43.7	117.1	
Chicago, IL	19	42.3	88.0	
Indianapolis, IN		21	39.8	86.9
Des Moines, IA	11	41.8	93.6	
Wichita, KS	22	38.1	97.6	
Louisville, KY	27	39.0	86.5	
New Orleans, LA	45	30.8	90.2	
Portland, ME	12	44.2	70.5	
Baltimore, MD	25	39.7	77.3	
Boston, MA	23	42.7	71.4	
Detroit, MI	21	43.1	83.9	
Minneapolis, MN	2	45.9	93.9	
St. Louis, MO	24	39.3	90.5	

City	JanTemp	Lat	Long	
Helena, MT	8	47.1	112.4	
Omaha, NE	13	41.9	96.1	
Concord, NH	11	43.5	71.9	
Atlantic City, NJ		27	39.8	75.3
Albuquerque, NM	24	35.1	106.7	
Albany, NY	14	42.6	73.7	
New York, NY	27	40.8	74.6	
Charlotte, NC	34	35.9	81.5	
Raleigh, NC	31	36.4	78.9	
Bismarck, ND	0	47.1	101.0	
Cincinnati, OH	26	39.2	85.0	
Cleveland, OH	21	42.3	82.5	
Oklahoma City, OK		28	35.9	97.5
Portland, OR	33	45.6	123.2	
Harrisburg, PA	24	40.9	77.8	
Philadelphia, PA		24	40.9	75.5
Charleston, SC	38	33.3	80.8	
Nashville, TN	31	36.7	87.6	
Amarillo, TX	24	35.6	101.9	
Galveston, TX	49	29.4	95.5	
Houston, TX	44	30.1	95.9	
Salt Lake City, UT		18	41.1	112.3
Burlington, VT	7	45.0	73.9	
Norfolk, VA	32	37.0	76.6	
Seattle, WA	33	48.1	122.5	
Spokane, WA	19	48.1	117.9	
Madison, WI	9	43.4	90.2	
Milwaukee, WI	13	43.3	88.1	
Cheyenne, WY	14	41.2	104.9	

Scatter plot of average Jan min. temp. versus latitude



There is a strong association between latitude and temperature except for the coasts where the temperature is moderate due to the ocean.

Scatter plot

The pattern in a scatterplot can be described by

- The direction
- the form
- strength of the relationship of the variables.

positive association
negative association

Definition

Given n observations $(x_1, y_1), ..., (x_n, y_n)$. The sample covariance s_{xy} is defined as

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

Notice that when $x = y$, we have $s_{xx} = var(x)$.

If we standardize s_{xy}, we obtain the sample correlation coefficient.

Definition

The sample correlation coefficient is

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

The correlation coefficient

- The correlation measures the **strength** and **direction** of the linear relationship between two variables.
- We have that $-1 \leq r \leq 1$.
- **Positive association** between the variables corresponds to $r > 0$.
- **Negative association between** the variables corresponds to $r < 0$.
- **Strength:** Values of r close to -1 or 1 indicates a close linear relationship between the variables.
- Values of r close to zero, indicates a weak linear relationship between the variables.

Example

We will calculate the correlation coefficient, r, in the previous example. Let $x =$ latitude and let $y =$average minimum Jan. temp. We have $\overline{x} = 26.51786$ and $\overline{y} = 38.96964$, $s_x = 13.37976$, and $s_y = 5.37854$. Hence,

$$
\begin{aligned}
r &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right) \\
&= \frac{1}{56-1}\sum_{i=1}^{56}\left(\frac{x_i - 26.51786}{13.37976}\right)\left(\frac{y_i - 38.96964}{5.37854}\right) \\
&= -0.8480352.
\end{aligned}
$$

In R:

```
> cor(x,y)
[1] -0.8480352
```

Simple linear regression

- We are interested in the relation between two variables.
- In the previous example, we looked at the relationship between the latitude and the average January minimum temperature. From the scatter plot and the correlation coefficient, it looks like there is approximately a linear relationship between these two variables.

Simple linear regression

The equation of a straight line relating these two variables is

$$Y = \alpha_1 + \beta x + \epsilon, \tag{1}$$

where

- α_1 is the y-intercept and β is the slope and are **unknown** constants.
- ϵ is a random error that is assumed to be $N(0, \sigma^2)$ and is uncorrelated to the other errors (i.e the value of one error is independent of the other errors.)
- The error ϵ is the difference between the observed value of y and the straight line $\alpha_1 + \beta x$ since the data points do not fall exactly on a straight line.

Simple linear regression

- The equation in (1) is called a **linear regression model**.
- x is called an independent variable, a predictor variable, or an explanatory variable.
- Y is called a dependent variable or response variable.
- Since (1) only have one independent variable x, it is called a **simple linear regression model**.
- It is convenient to consider the predictor x as controlled or measured with negligible error.
- The response variable Y is a **random variable**.

Simple linear regression

- The mean $E(Y) = \alpha_1 + \beta x$ of (1) is linear in the variables α_1 and β.
- To estimate the unknown parameters α_1 and β, we observe the random variable Y for each of n different values of x, say x_1, x_2,...,x_n. Once the n independent experiments have been performed, we have n pairs of data (x_1, y_1), (x_2, y_2),.., (x_n, y_n).
- We want to fit a straight line to the data points (x_1, y_1), (x_2, y_2),.., (x_n, y_n).
- Thus our model is

$$Y_i = \alpha_1 + \beta x_i + \epsilon_i,$$

  where ϵ_i for $i = 1, 2, .., n$ are independent and $N(0, \sigma^2)$.
- We have that for each x, the value of Y differ from $E(Y) = \alpha_1 + \beta x$ by a random amount ϵ.

Simple linear regression

- The variance of Y_i at a particular value of x_i is determined by the variance of the random error which is σ^2.
- The constants α_1 and β are called the **regression coefficients**.
- The slope β is the change in $E(Y)$ caused by a unit change in x.
- The y-intercept α_1 is the mean of the distribution, $E(Y)$, when $x = 0$. If $x = 0$ is not contained in the range of x, α_1 has no practical interpretation.

Simple linear regression

A model for the mean like

$$Y = \alpha + \beta x + \gamma x^2$$

is called a **linear model** because it is linear in the parameters α, β, and γ. Thus, $\alpha e^{\beta x}$ is not a linear model because it is not linear in α and β.

Simple linear regression

We have the sample regression model, $Y_i = \alpha_1 + \beta x_i + \epsilon_i$. We will now find point estimates for α_1, β, and σ^2. We will for convenience redefine the predictor variable x_i as the deviation from its own average $x_i - \overline{x}$. We have

$$\begin{aligned} Y_i &= \alpha_1 + \beta(x_i - \overline{x}) + \beta\overline{x} + \epsilon_i \\ &= (\alpha_1 + \beta\overline{x}) + \beta(x_i - \overline{x}) + \epsilon \\ &= \alpha + \beta(x_i - \overline{x}) + \epsilon_i, \end{aligned}$$

where we defined $\alpha = \alpha_1 + \beta\overline{x}$, where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Thus, $\alpha_1 = \alpha - \beta\overline{x}$.

Y_1,..., Y_n are mutually independent normal random variables (since ϵ_1,...,ϵ_n independent normal random variables) with respective means $\alpha + \beta(x_i - \overline{x})$ for $i = 1, 2, ..., n$ and unknown variance σ^2, i.e Y_i is $N(\alpha + \beta(x_i - \overline{x}), \sigma^2)$. Their likelihood function is

$$L(\alpha, \beta, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\sum_{i=1}^{n} \frac{[y_i - \alpha - \beta(x_i - \overline{x})]^2}{2\sigma^2}\right).$$

To maximize $\ln(L(\alpha, \beta, \sigma^2)$ is equivalent to minimize

$$-\ln(L(\alpha, \beta, \sigma^2) = \frac{n}{2}\ln(2\pi\sigma^2) + \sum_{i=1}^{n} \frac{[y_i - \alpha - \beta(x_i - \overline{x})]^2}{2\sigma^2}.$$

We must select α and β to minimize

$$H(\alpha, \beta) = \sum_{i=1}^{n} [y_i - \alpha - \beta(x_i - \overline{x})]^2.$$

$$|y_i - \alpha - \beta(x_i - \overline{x})| = |y_i - E(Y)|$$

is the vertical distance from the points (x_i, y_i) to the line $E(Y) = \alpha + \beta(x - \overline{x})$. Thus, $H(\alpha, \beta)$ represents the sum of squares of those distances. Selecting α and β so that the sum of squares is minimized means that we are using the **method of least squares to fit the straight line to the data.**

Thus we must solve for α and β in the equations,

$$\frac{\partial H(\alpha, \beta)}{\partial \alpha} = 0$$

and

$$\frac{\partial H(\alpha, \beta)}{\partial \beta} = 0.$$

(Details of this is given in class.) We obtain the maximum likelihood estimators,

$$\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} y_i = \overline{y}$$

and

$$\hat{\beta} = r \frac{s_y}{s_x}.$$

To find the maximum likelihood estimator of σ^2, we solve

$$\frac{\partial \ln(L(\alpha, \beta, \sigma^2))}{\partial \sigma^2} = 0$$

and obtain

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} [y_i - \hat{\alpha} - \hat{\beta}(x_i - \overline{x})]^2.$$

Result

Let $\hat{y}_i = \hat{\alpha} + \hat{\beta}(x_i - \overline{x})$ be the predicted mean value of y_i.

The straight line fitted by the method of least square is given by

$$\hat{y} = \hat{\alpha} + \hat{\beta}(x - \overline{x}),$$

where

$$\hat{\alpha} = \overline{y}$$

and

$$\hat{\beta} = r\frac{s_y}{s_x}.$$

The variance of the error, ϵ, can be estimated by

$$\hat{\sigma}^2 = \frac{1}{n}\sum^{n}[y_i - \hat{\alpha} - \hat{\beta}(x_i - \overline{x})]^2.$$

Residuals

Definition

The difference

$$e_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta}(x_i - \overline{x})$$

for $i = 1, 2, ..., n$ between the observed value y_i and the predicted value $\hat{y}_i$ is called the i^{th} residual.

The sum of the residuals should be equal to zero. In practice, due to round off error, the observed residuals, $y_i - \hat{y}_i$, sometimes differs slightly from zero.

We have the following properties of the least-square fit

- $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} y_i - \hat{y}_i = 0.$
- $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i.$
- $\sum_{i=1}^{n} x_i e_i = 0$
- $\sum_{i=1}^{n} \hat{y}_i e_i = 0$
- The regression line passes through $(\overline{x}, \overline{y})$.

Example

Lets calculate the regression line for the first example.
Let $x =$ latitude and let $y =$average minimum Jan. temp. We have

$\overline{x}$	38.9696
$\overline{y}$	26.5179
s_x	5.3785
s_y	13.37976
r	-0.8480

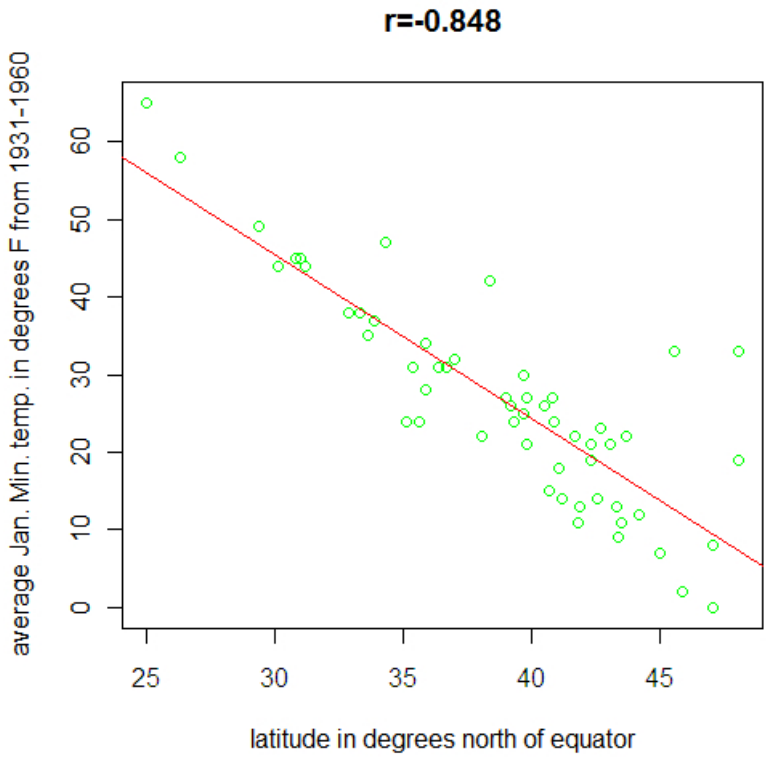
Then $\hat{\alpha} = \overline{y} = 26.5179$ and
$\hat{\beta} = r\frac{s_y}{s_x} = -0.8480\frac{13.37976}{5.3785} = -2.109588$. Thus, the regression line is

$$\hat{y} = \hat{\alpha} + \hat{\beta}(x - \overline{x}) = 26.5179 - 2.109588(x - 38.9696) = 108.728 - 2.110x.$$

Example continues, R-code

```
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    108.728       -2.110
```

Scatter plot with regression line



R-code:

$>$ plot(x,y)

$>$ *abline*(*lm*($y \sim x$)) (add the regression line)

Prediction

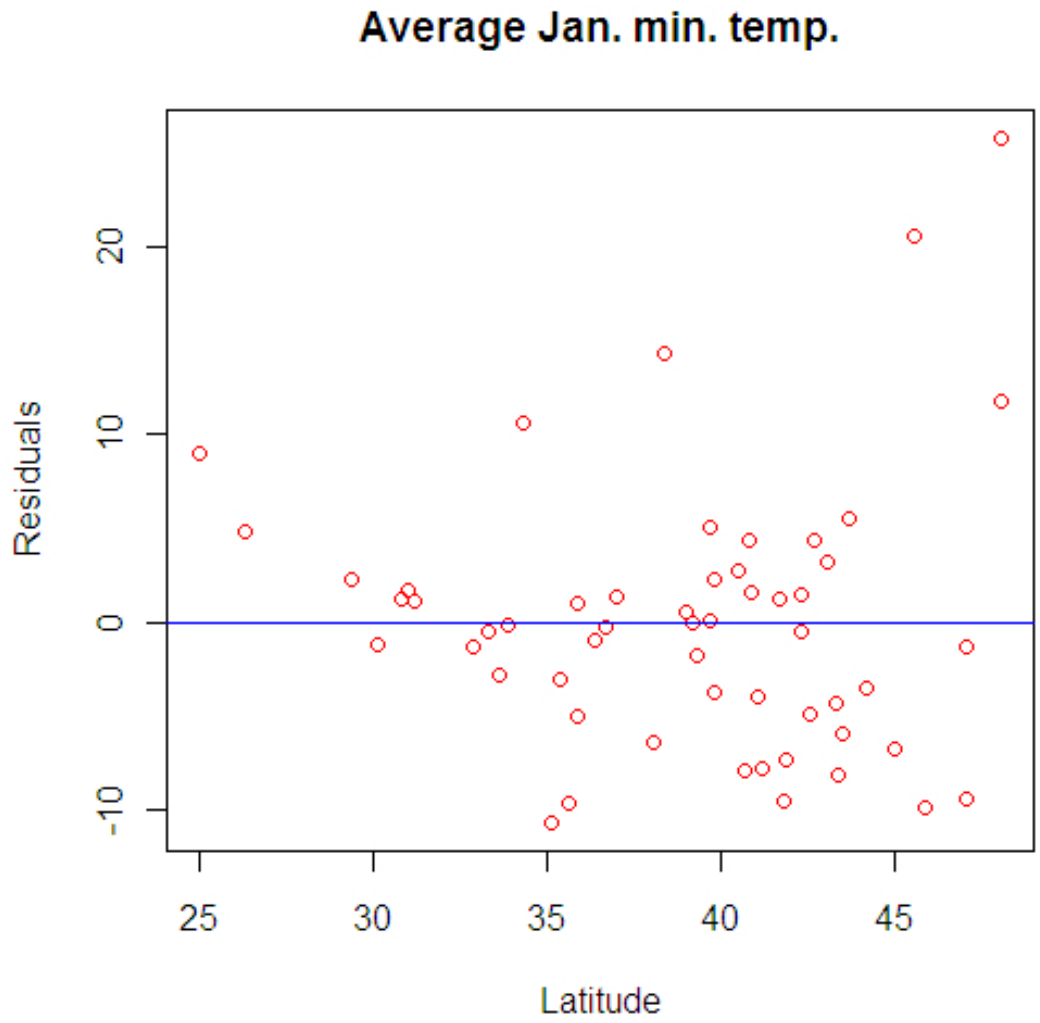
$$\hat{Y} = \hat{\alpha} + \hat{\beta}(x - \overline{x})$$

is the prediction of the value of Y for some given x.

Example

What is the prediction of $\hat{y}$ when $x = 37$ in the previous example?
We have $\hat{y} = 108.728 - 2.110x = 108.728 - 2.110 * 37 = 30.658$.
Thus when the latitude is 37 degrees north of equator, the average Jan. min. temp is predicted to be 30.658 degrees F.

Residual plot

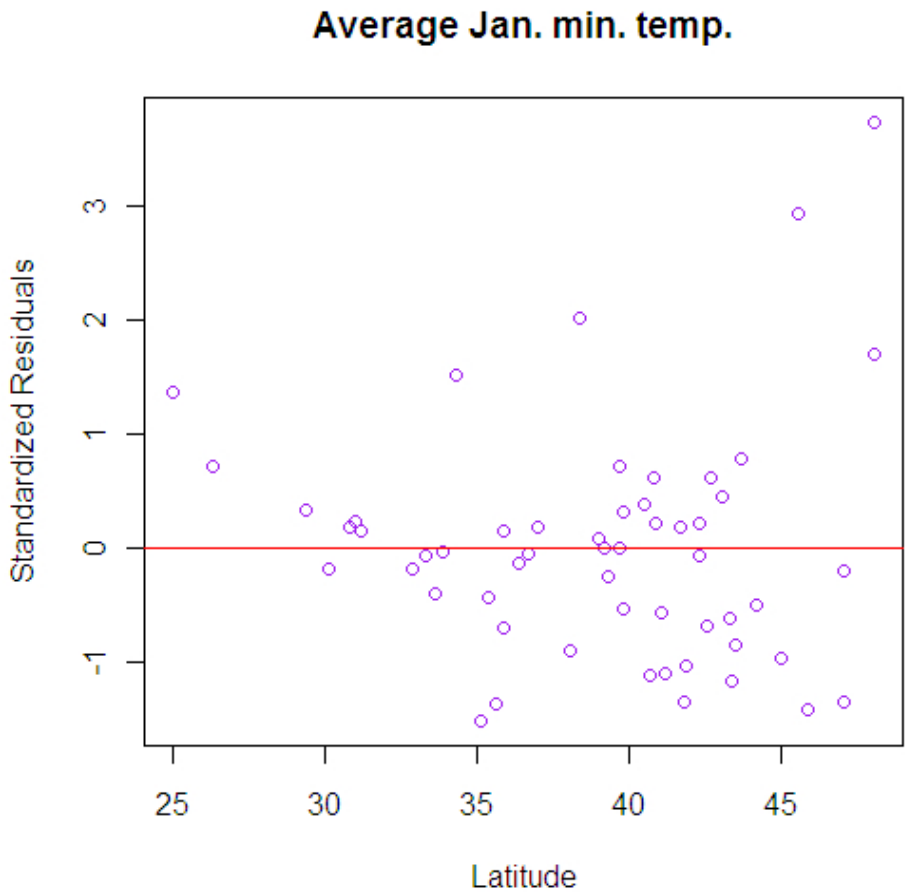


R-code for Residuals and residuals plot

R-code for residuals: $resid(lm(y \sim x))$
R-code for standardize residuals: $rstandard(lm(y \sim x))$

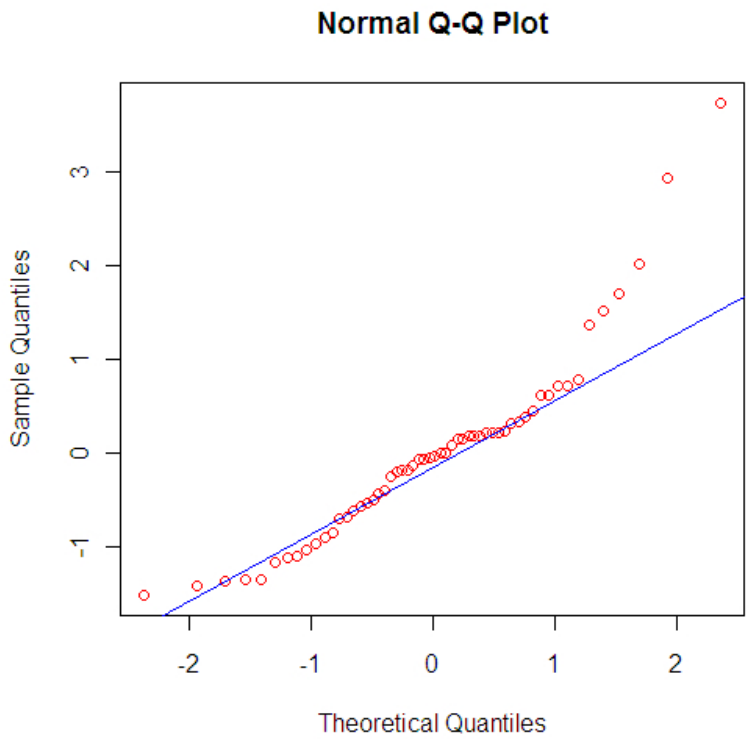
```
> plot(x,y.res,xlab="Residuals",ylab="Latitude",main="Average Jan. min. temp.", col="red")
> abline(0,0,col="blue")
```

Standardized residual plot



R-code: $rstandard(lm(y \sim x))$ (standardized residuals)

The normal probability plot



R-code: $qqnorm(rstandard(lm(y \sim x)), col = "red")$
$qqline(rstandard(lm(y \sim x)), col = "blue")$

Result (proof given in class)

The sample mean of $\hat{y}_1, ..., \hat{y}_n$ is

$$\overline{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \overline{y}.$$

The sample variance of $\hat{y}_1, ..., \hat{y}_n$ is

$$s^2_{\hat{y}} = \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = r^2 s^2_y.$$

Hence,

$$r^2 = \frac{s^2_{\hat{y}}}{s^2_y}.$$

r^2

- r^2 is the fraction of the variation in the values of $y_1, ..., y_n$ that is explained by the regression of y on x.
- In the previous example we had $r = -0.848$ and hence $r^2 = 0.719104$. Thus, 71.9% of the variation in the values of $y_1, ..., y_n$ is due to the regression of y on x.

$\hat{\alpha}$

We will now find distributions of $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\sigma}^2$. Since the estimator $\hat{\alpha}$ is a linear function of independent and normally distributed random variables, $\hat{\alpha}$, has a normal distribution with mean

$$E(\hat{\alpha}) = \alpha$$

and variance

$$Var(\hat{\alpha}) = \frac{\sigma^2}{n}.$$

Thus, $\hat{\alpha}$ is $N(\alpha, \frac{\sigma^2}{n})$.
(Proof is given in class.)

$\hat{\beta}$ is a linear function of $Y_1, \ldots Y_n$ and hence has a normal distribution with mean

$$E(\hat{\beta}) = \beta$$

and variance

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Thus $\hat{\beta}$ is $N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)$.

(Proof is given in class.)

Since Y_i, $\hat{\alpha}$ and $\hat{\beta}$ have normal distributions, we know that

$$\frac{[Y_i - \alpha - \beta(x_i - \overline{x})]^2}{\sigma^2},$$

$$\frac{(\hat{\alpha} - \alpha)^2}{\frac{\sigma^2}{n}} \quad \text{and} \quad \frac{(\hat{\beta} - \beta)^2}{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

each has a chi-square distribution with one degree of freedom. Since Y_1, Y_2,..., Y_n are mutually independent,

$$\frac{\sum_{i=1}^{n}[Y_i - \alpha - \beta(x_i - \overline{x})]^2}{\sigma^2}$$

is $\chi^2(n)$.

Since

$$\frac{\sum_{i=1}^{n}[Y_i - \alpha - \beta(x_i - \overline{x})]^2}{\sigma^2}$$

$$= n\frac{(\hat{\alpha} - \alpha)^2}{\sigma^2} + \frac{(\hat{\beta} - \beta)^2}{\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 + \frac{\sum_{i=1}^{n}[Y_i - \hat{\alpha} - \hat{\beta}(x_i - \overline{x})]^2}{\sigma^2},$$

where the term on the left hand side is $\chi^2(n)$ and the first two terms on the right hand side are $\chi^2(1)$.
it can be shown that

$$\frac{\sum_{i=1}^{n}[Y_i - \hat{\alpha} - \hat{\beta}(x_i - \overline{x})]^2}{\sigma^2} = \frac{n\hat{\sigma^2}}{\sigma^2},$$

is $\chi^2(n-2)$. We have lost two degrees of freedom of replacing α with $\hat{\alpha}$ and β with $\hat{\beta}$. It can be proved that $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\sigma^2}$ are mutually independent.

Confidence interval for β.

$$T_1 = \frac{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\,\frac{\hat{\beta}-\beta}{\sigma}}{\sqrt{\frac{n\hat{\sigma}^2}{\sigma^2(n-2)}}} = \frac{\hat{\beta}-\beta}{\sqrt{\frac{n\hat{\sigma}^2}{(n-2)\sum_{i=1}^{n}(x_i-\overline{x})^2}}}$$

has a t-distribution with $n-2$ degrees of freedom. Thus,

$$P\left[-t_{\frac{\gamma}{2}}(n-2) \leq \frac{\hat{\beta}-\beta}{\sqrt{\frac{n\hat{\sigma}^2}{(n-2)\sum_{i=1}^{n}(x_i-\overline{x})^2}}} \leq t_{\frac{\gamma}{2}}(n-2)\right] = 1-\gamma$$

Hence

$$\hat{\beta} \pm t_{\frac{\gamma}{2}}(n-2)\sqrt{\frac{n\hat{\sigma}^2}{(n-2)\sum_{i=1}^{n}(x_i-\overline{x})^2}}$$

is a $100(1-\gamma)\%$ confidence interval for β.

Confidence interval for α.

$$T_2 = \frac{\frac{\sqrt{n}(\hat{\alpha}-\alpha)}{\sigma}}{\sqrt{\frac{n\hat{\sigma}^2}{\sigma^2(n-2)}}} = \frac{\hat{\alpha}-\alpha}{\sqrt{\frac{\hat{\sigma}^2}{n-2}}}$$

has a t-distribution with $(n-2)$ degrees of freedom. The endpoint for a $100(1-\gamma)\%$ confidence interval for α is

$$\hat{\alpha} \pm t_{\frac{\gamma}{2}}(n-2)\sqrt{\frac{\hat{\sigma}^2}{(n-2)}}$$

Confidence interval for σ^2.

$$\frac{n\hat{\sigma}^2}{\sigma^2}$$

has a chi-square distribution with $n-2$ degrees of freedom. A $100(1-\gamma)\%$ confidence interval for σ^2 is

$$\left[\frac{n\hat{\sigma}^2}{\chi^2_{\frac{\gamma}{2}(n-2)}}, \frac{n\hat{\sigma}^2}{\chi^2_{1-\frac{\gamma}{2}(n-2)}}\right]$$

Test about the slope of the regression line

If we are testing

$$H_0 : \beta = \beta_0 \quad \text{against} \quad H_1 : \beta > \beta_0,$$

we reject H_0 if the test statistics

$$T_1 = \frac{\hat{\beta} - \beta_0}{\sqrt{\frac{n\hat{\sigma}^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \geq t_\gamma(n-2).$$

Alternatively, we can compute $T_1 = t_0$, say and determine the p-value$=P(T(n-2) \geq t_0)$ and reject H_0 if p-value $\leq \gamma$.

Test of hypothesis

One sided hypothesis test about the slope of the regression line:

$$H_0 : \beta = \beta_0 \quad \textit{against} \quad H_1 : \beta > \beta_0.$$

We reject H_0 at the $(\gamma \times 100)\%$ significance level if

$$t_1 \geq t_\gamma(n-2). \ (\textbf{Critical region})$$

Test of hypothesis

One sided hypothesis test about the slope of the regression line:

$$H_0 : \beta = \beta_0 \quad \textit{against} \quad H_1 : \beta < \beta_0.$$

We reject H_0 at the $(\gamma \times 100)\%$ significance level if

$$t_1 \leq -t_\gamma(n-2). \ (\textbf{Critical region})$$

Test of hypothesis

One sided hypothesis test about the slope of the regression line:

$$H_0 : \beta = \beta_0 \quad \textit{against} \quad H_1 : \beta \neq \beta_0.$$

We reject H_0 at the $(\gamma \times 100)\%$ significance level if

$$|t_1| \geq t_{\gamma/2}(n-2). \ (\textbf{Critical region})$$

Example continues

Example

Using the previous data set, test the hypothesis,

$$H_0 : \beta = 0 \quad \text{against} \quad H_1 : \beta \neq 0.$$

Construct a 95% confidence interval. We have $\sum_{i=1}^{n}(x_i - \bar{x})^2 = (n-1)var(x) = 1591.078$ and $\hat{\sigma}^2 = 49.377$ so

$$T_1 = \frac{\hat{\beta} - \beta}{\sqrt{\frac{n\hat{\sigma}^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}} = \frac{-2.11 - 0}{\sqrt{\frac{56(49.377)}{(54)(1591.078)}}} = -11.76$$

$t_{0.025}(54) = 2.0049$ (In R: $qt(0.975, 54)$) Thus, we reject H_0, so there is a linear relationship between the variables.
The p-value$= 2P(T \geq 11.76) = 2.22044610^{-16}$.

Example continues

Example

The standard error is $\sqrt{\frac{n\hat{\sigma}^2}{(n-2)\sum_{i=1}^{n}(x_i-\overline{x})^2}} = 0.1794$, so a 95% confidence interval for β is

$$\hat{\beta} \pm t_{\frac{\alpha}{2}}(n-2)\sqrt{\frac{n\hat{\sigma}^2}{(n-2)\sum_{i=1}^{n}(x_i-\overline{x})^2}} = -2.11 \pm (2.0049)(0.1794)$$

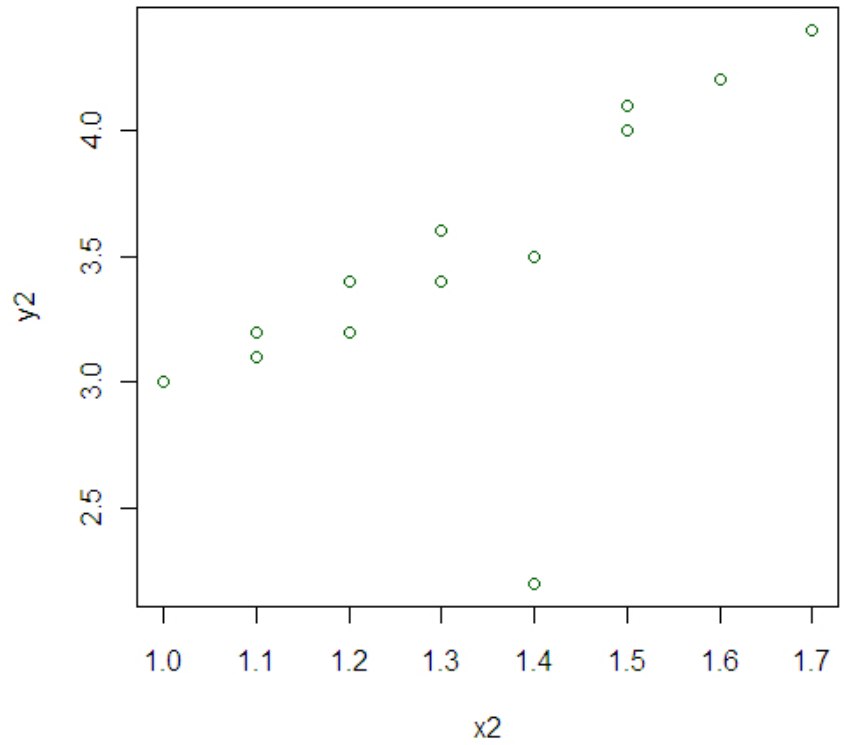
or

$$(-2.470, -1.750).$$

Summary statistics

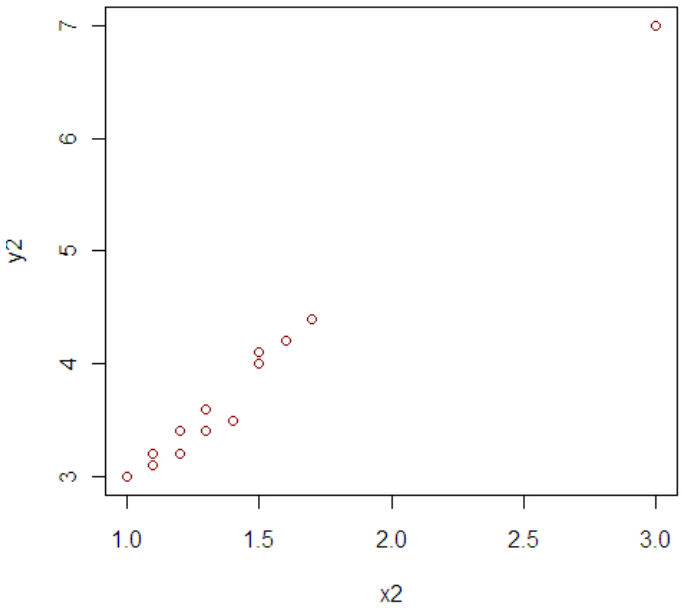
```
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6812  -4.5018  -0.2593   2.2489  25.7434

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 108.7277     7.0561   15.41   <2e-16 ***
x            -2.1096     0.1794  -11.76   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.156 on 54 degrees of freedom
Multiple R-squared: 0.7192,     Adjusted R-squared: 0.714
F-statistic: 138.3 on 1 and 54 DF,  p-value: < 2.2e-16
```

Influential observation



An influential point has an unusual y-value and "pulls" the regression model in its direction.

Leverage point



A **leverage point** is a point that has an unusual x-coordinate but it lies almost on the regression line fitting the rest of the sample points. This point will not affect the estimate of the regression coefficient but it will affect the summary statistics such as r^2 and the standard errors of the regression coefficients.

Model assumptions

In our regression analysis, we have assumed the following:

- There is a linear relationship between the response y and the regressors x.
- The random error, ϵ, has mean zero.
- The random error, ϵ, has constant variance σ^2.
- The errors are uncorrelated.
- The errors are normally distributed.

One way to diagnose violations of the regression assumptions is to study the **residuals**.

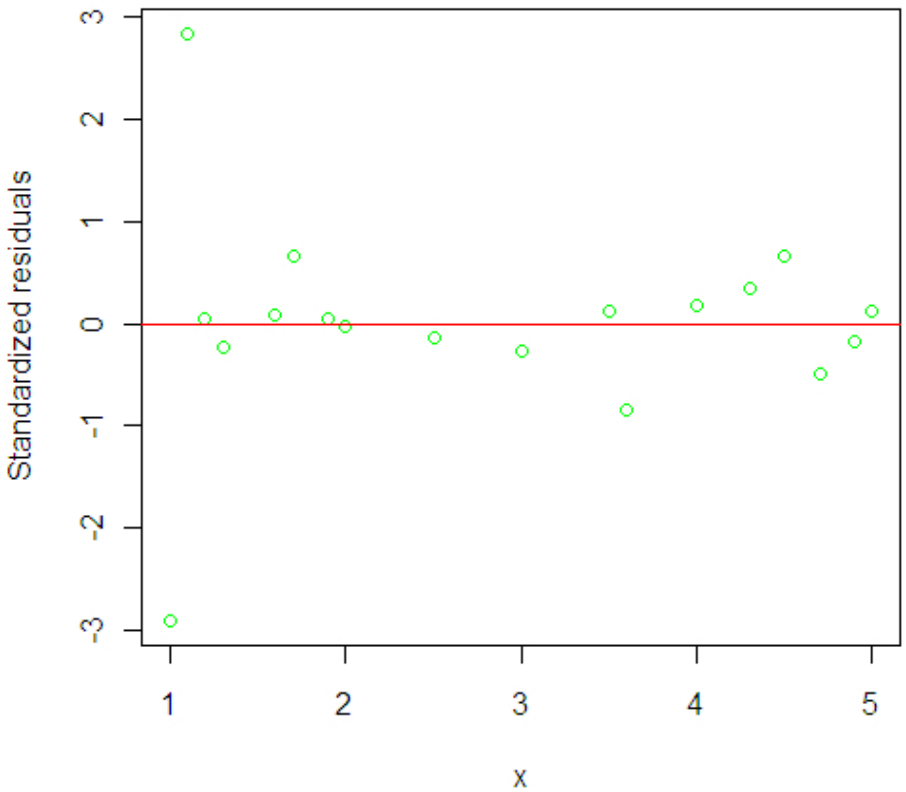
Residual plots

- The residuals is a measure of the variability in the response variable not explained by the regression model.
- We can also consider the residuals as the observed values of the model errors. Thus, if the model errors do not follow the model assumptions as described above, it will be shown in the residuals.
- By analyzing the residuals, one can discover violations of the initial assumptions and types of model inadequacies.
- In this course, we will analyze residual by plotting them against the regressor x or the predictor $\hat{y}$.

Residual and normal probability plots

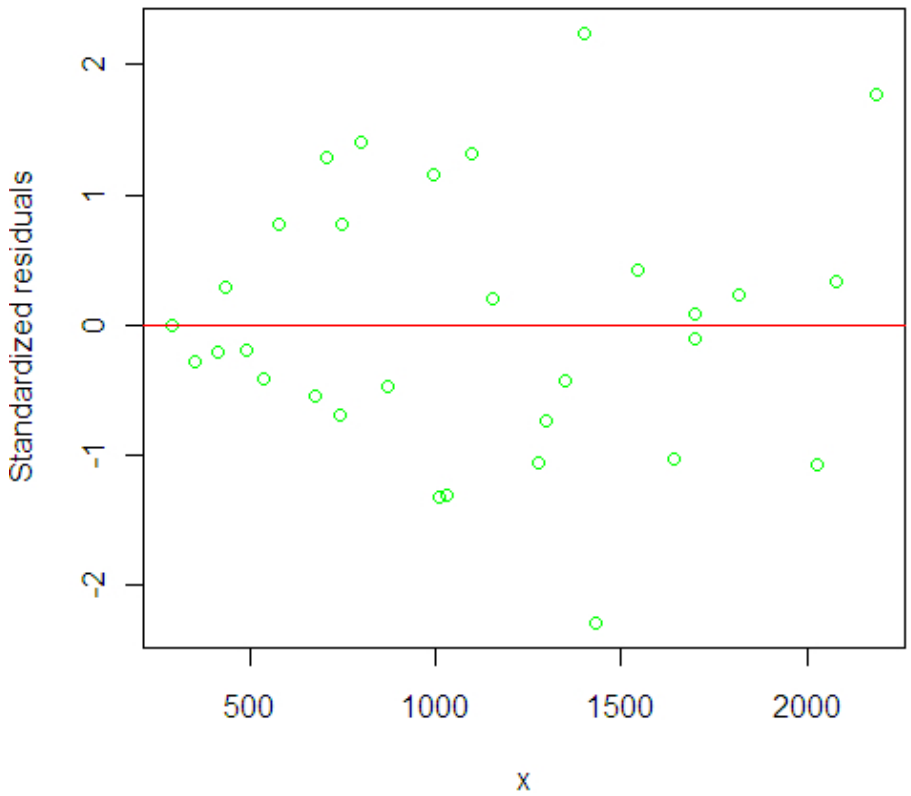
- It is useful to scale the residual to find observations that are outliers. One way to do this is to standardize the residuals so that the resulting residuals have mean zero and approximate unit variance. A large standardize residual, say greater than 3, indicates a potential outlier.
- In R:, $rstandard(lm(y \sim x))$.
- One way to check the normality assumption is to draw a normal probability plot of the residuals. The cumulative normal distribution will be plotted as a straight line and the residuals should lie approximately on the straight line. If the residuals do not lie approximately on the straight line, it indicates that the underlying distributions is not normal. (See previous problem).

Patterns for linear residual plot



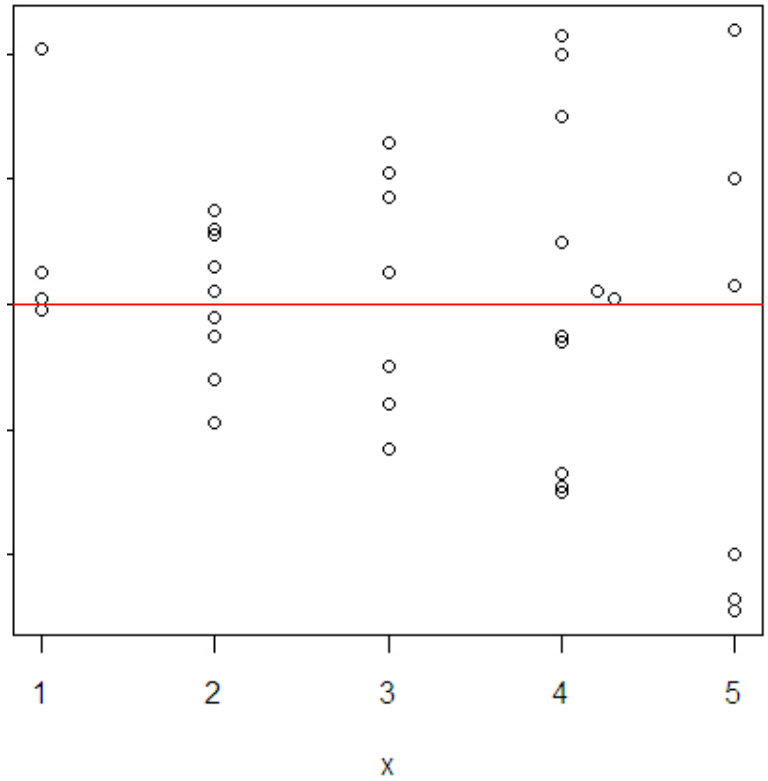
The residuals can be contained in a horizontal band which indicates no model defects.

Patterns for residual plot, double bow



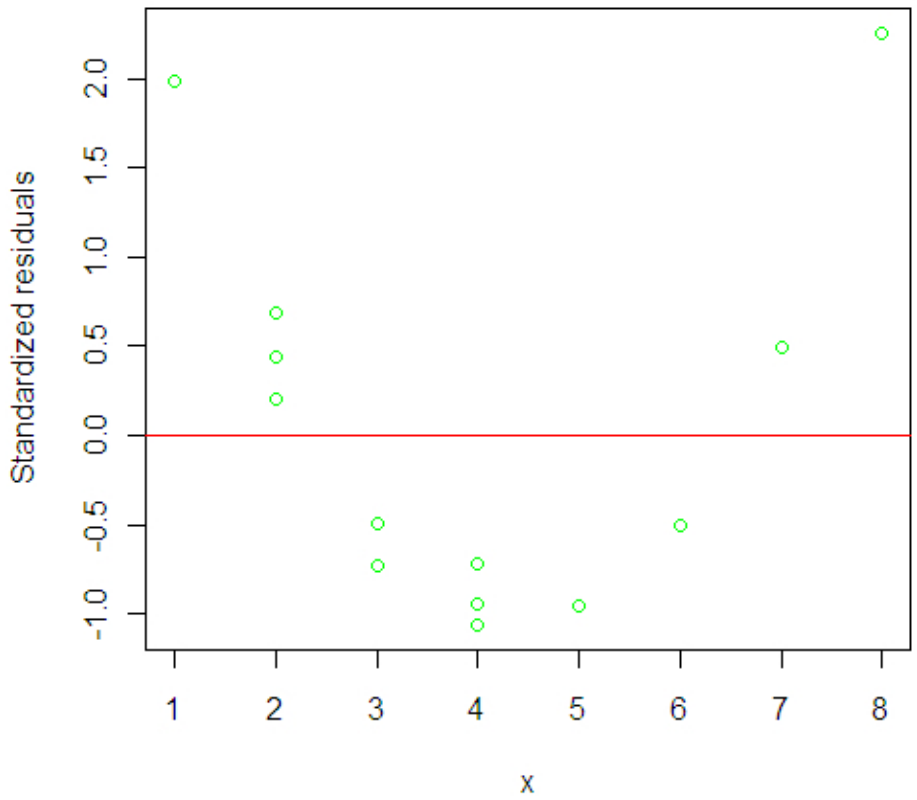
Indicates that the variance is not constant.

Patterns for residual plot, funnel



The outward-opening funnel or cone indicates that the variance increases as y increases.

Patterns for nonlinear residual plot



A curved plot, indicates non-linearity.

Transformations

- To take care of the inequality of variance one usually applies a transformation either to the regressor x or to the response variable y.
- If the standard deviation of y seems to increase with the value of y, logarithmic transformation $\log(y)$, the square transformation $\sqrt{y}$ or the transformation y^p for some value of p might be an appropriate one.
- If the x-values seemed skewed to the right, $\log(x)$, $\sqrt{x}$ or x^p $(p < 1)$ might be an appropriate one.
- If the scatter plot of y against x shows some curvature, we may be able to transform the variables and use the resulting linear forms to represent the data.

Transformations

Example

Consider the exponential function

$$y = \alpha x^{\beta} \epsilon.$$

This function can be transformed to a straight line by logarithmic transformations,

$$\log(y) = \log(\alpha) + \beta \log(x) + \log(\epsilon)$$

or

$$y' = \alpha' + \beta x' + \epsilon'.$$

It is required that the transformed errors, $log(\epsilon)$, are independent and normally distributed with mean zero and variance σ^2.

Example

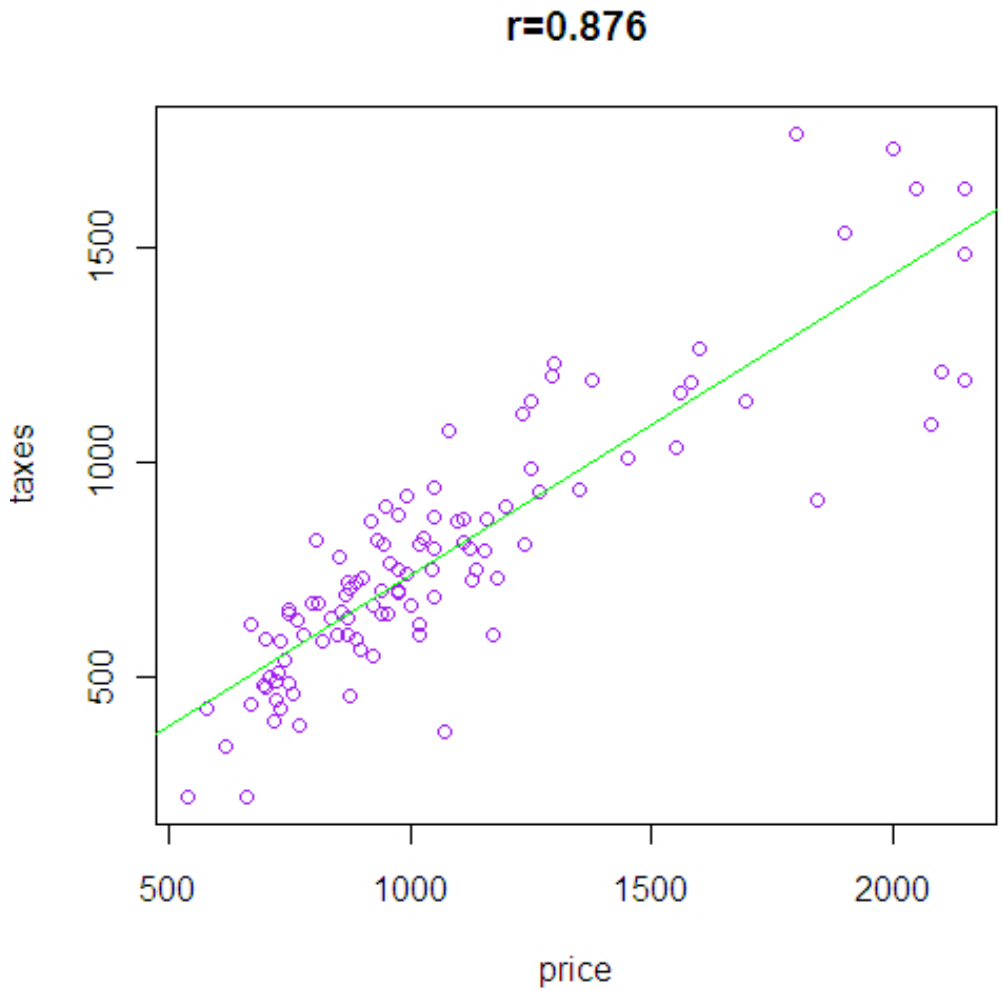
Example

The data provided on the next page shows a random sample of records of resales of homes from Feb 15 to Apr 30, 1993 in Albuquerque. We want to investigate how taxes change in response to changing market value of homes.
Let X be the selling price in hundreds of dollars.
Let Y be the annual taxes in dollars.
Reference: Albuquerque Board of Realtors at http://lib.stat.cmu.edu/DASL/Datafiles/homedat.html

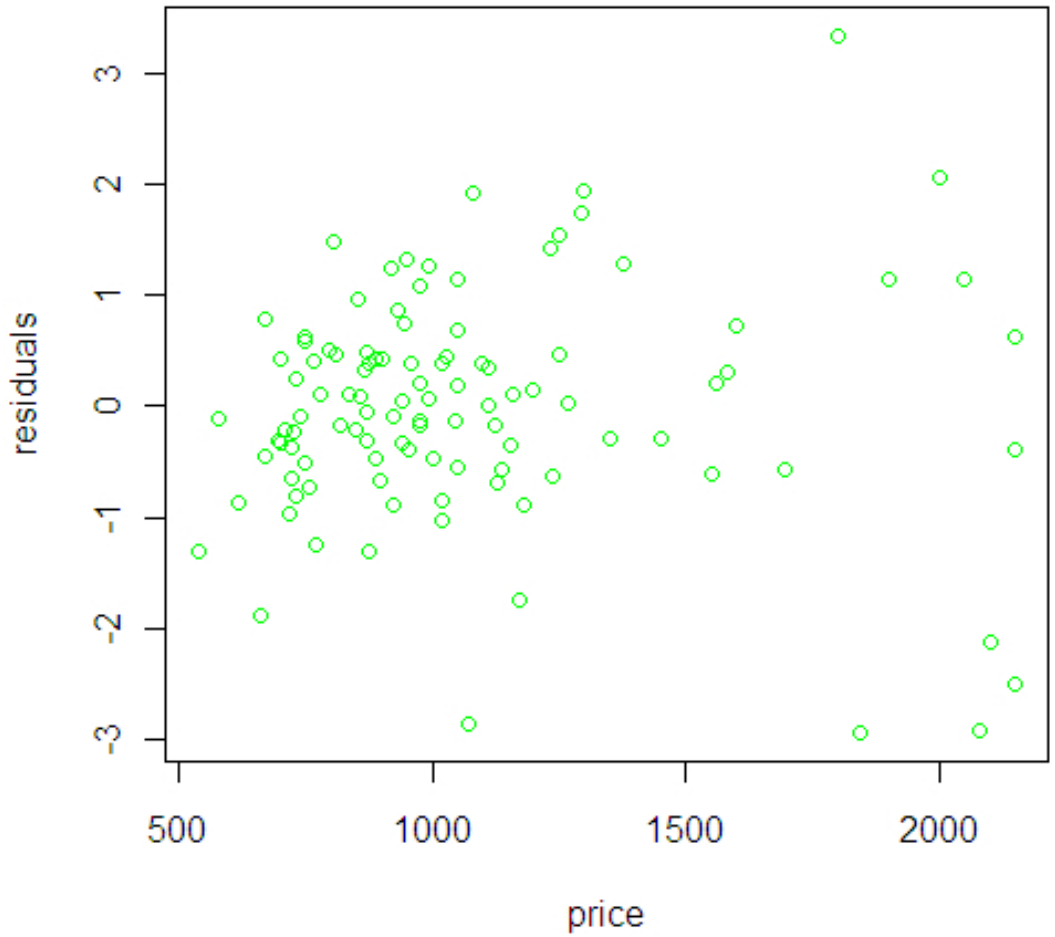
Example

```
> z=read.delim("C:/Documents and Settings/ghystad/Desktop/Math 361/homedata.txt")
> x=z[,1]
> y=z[,8]
> x
  [1] 2050 2080 2150 2150 1999 1900 1800 1560 1449 1375 1270 1250 1235 1170 1180
 [16] 1155 1110 1139  995  995  975  975  900  960  860 1695 1553 1020 1020  922
 [31]  925  899  850  890  870  700  720  749  731  725  670 2150 1599 1350 1299
 [46] 1250 1239 1200 1125 1100 1080 1050 1049  955  934  875  889  855  835  810
 [61]  805  799  750  759  750  730  729  710  670  619 1295  975  939  820  780
 [76]  770  700  540 1070 2100  725  660  580 1844 1580  699 1160 1109 1129 1050
 [91] 1045 1050 1020 1000 1030  975  950  940  920  945  874  872  870  869  766
[106]  739
> y
  [1] 1639 1088 1193 1635 1732 1534 1765 1161 1010 1191  930  984 1112  600  733
 [16]  794  867  750  923  743  752  696  731  768  653 1142 1035  626  600  668
 [31]  553  566  600  591  599  477  398  656  585  490  440 1487 1265  939 1232
 [46] 1141  810  899  800  865 1076  875  690  648  820  456  723  780  638  673
 [61]  821  671  649  461  486  427  513  504  622  342 1200  700  701  585  600
 [76]  391  591  223  376 1209  447  225  426  915 1189  481  867  816  725  800
 [91]  750  944  811  668  826  880  900  647  866  810  707  721  638  694  634
[106]  541
```

Plot of data with regression line



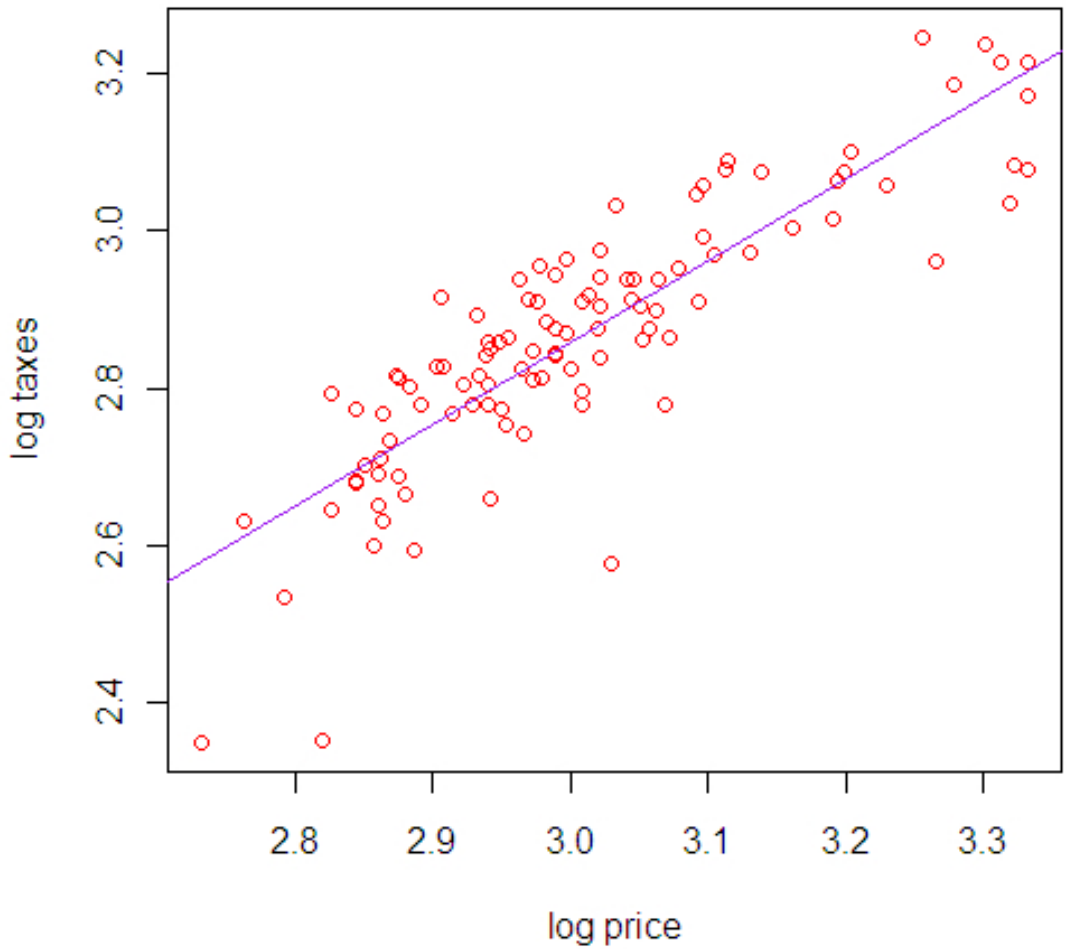
Residual plot



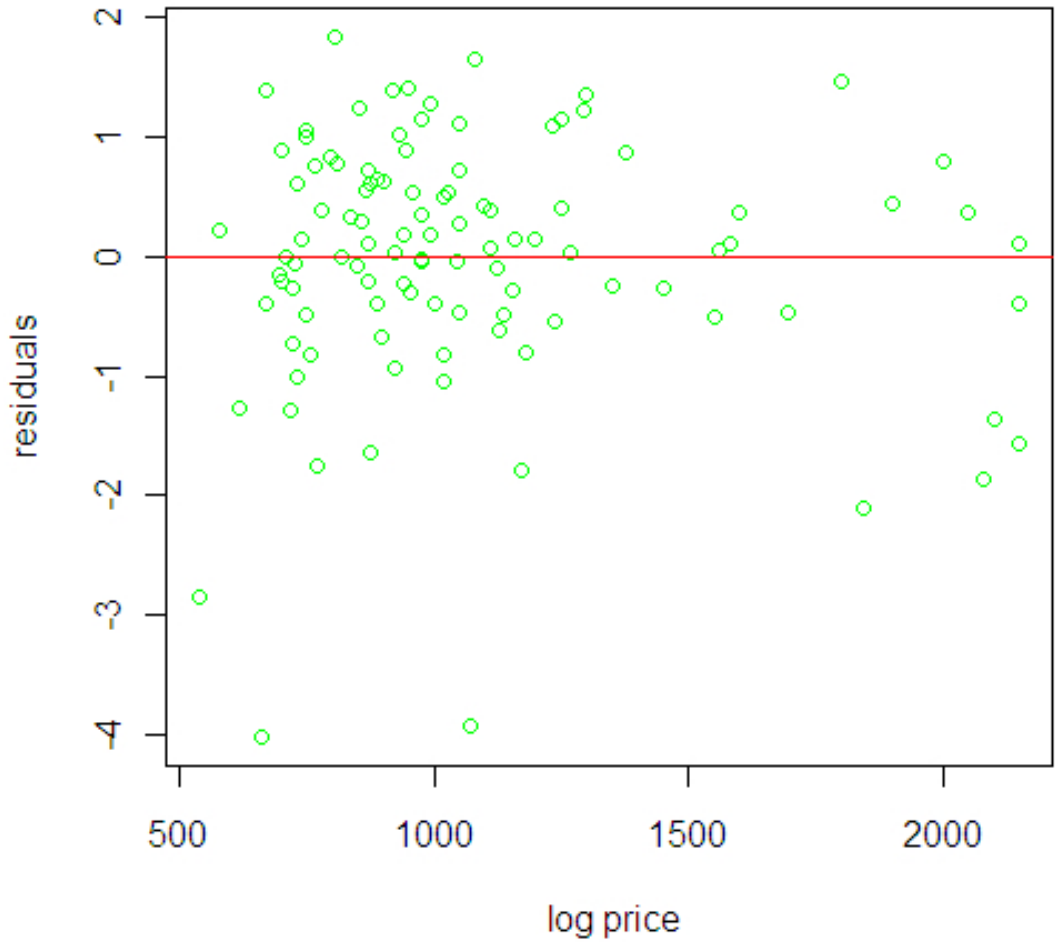
```
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-414.69  -75.15    2.09   72.16  466.15

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.11038   43.30514   0.857    0.393
x            0.70097    0.03794  18.478   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 149.7 on 104 degrees of freedom
Multiple R-squared: 0.7665,     Adjusted R-squared: 0.7643
F-statistic: 341.4 on 1 and 104 DF,  p-value: < 2.2e-16
```

We see that few undertaxed expensive homes influence the linear relationship between taxes and resale value.
We also see that as the price increases the residual error increases.
We will therefore transform the analysis to log taxes versus log price. That is, take the log of X and log of Y and investigate the linear relationship of the transformed variables.

Plot of the transformed variables with the regression line



Residual plot of the transformed variables



```
> x1=log10(x)
> y1=log10(y)
> summary(lm(y1~x1))

Call:
lm(formula = y1 ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31758 -0.03831  0.00887  0.05410  0.15451

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.27439    0.18295   -1.50    0.137
x1           1.04420    0.06075   17.19   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08589 on 104 degrees of freedom
Multiple R-squared: 0.7396,     Adjusted R-squared: 0.7371
F-statistic: 295.4 on 1 and 104 DF,  p-value: < 2.2e-16
```

We see that for the transformed variables the residual standard error decreases (more about this later) and the p-value for the intercept decreases.