Simple Linear Regression

- Suppose we observe bivariate data (X, Y), but we do not know the regression function $E(Y|X = x)$. In many cases it is reasonable to assume that the function is linear:

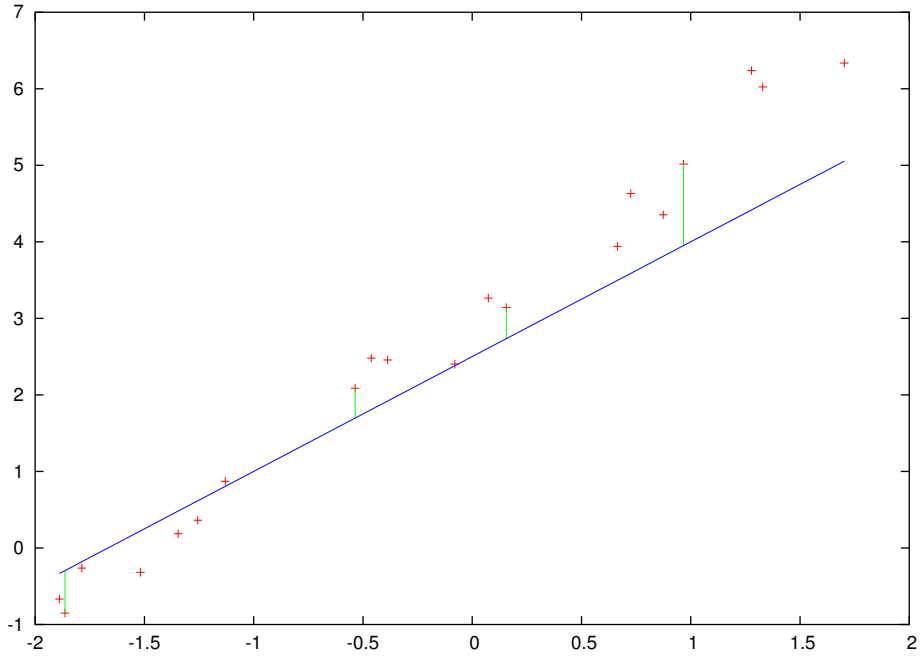
$$E(Y|X = x) = \alpha + \beta x.$$

  In addition, we assume that the distribution is homoscedastic, so that $\sigma(Y|X = x) = \sigma$.

  We have reduced the problem to three unknowns (parameters): α, β, and σ. Now we need a way to estimate these unknowns from the data.

- For fixed values of α and β (not necessarily the true values), let $r_i = Y_i - \alpha - \beta X_i$ (r_i is called the residual at X_i).

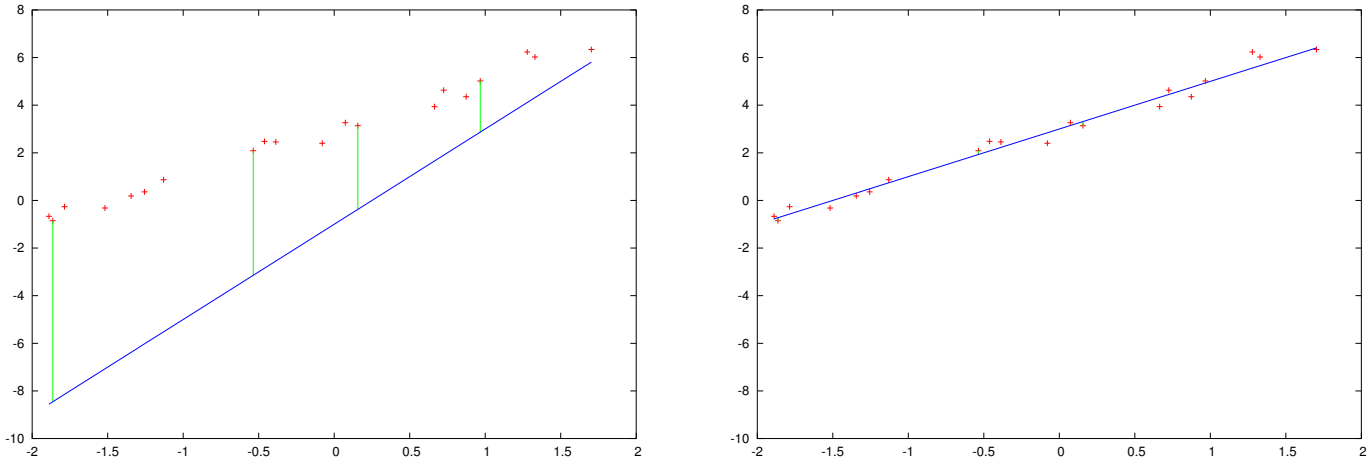
  Note that r_i is the vertical distance from Y_i to the line $\alpha + \beta x$. This is illustrated in the following figure:



A bivariate data set with $E(Y|X = x) = 3 + 2X$, where the line $Y = 2.5 + 1.5X$ is shown in blue. The residuals are the green vertical line segments.

- One approach to estimating the unknowns α and β is to consider the sum of squared residuals function, or SSR.

  The SSR is the function $\sum_i r_i^2 = \sum_i (Y_i - \alpha - \beta X_i)^2$. When α and β are chosen so the fit to the data is good, SSR will be small. If α and β are chosen so the fit to the data is poor, SSR will be large.



Left: a poor choice of α and β that give high SSR. Right: α and β that give nearly the smallest possible SSR.

- It is a fact that among all possible α and β, the following values minimize the SSR:

$$\begin{aligned} \widehat{\beta} &= \text{cov}(X,Y)/\text{var}(X) \\ \widehat{\alpha} &= \bar{Y} - \widehat{\beta}\bar{X}, \end{aligned}$$

These are called the least squares estimates of α and β.

The estimated regression function is

$$\widehat{E}(Y|X=x) = \widehat{\alpha} + \widehat{\beta}x$$

and the fitted values are

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}x_i.$$

- Some properties of the least square estimates:

    1. $\hat{\beta} = \text{cor}(X, Y)\hat{\sigma}_Y/\hat{\sigma}_X$, so $\hat{\beta}$ and $\text{cor}(X, Y)$ always have the same sign – if the data are positively correlated, the estimated slope is positive, and if the data are negatively correlated, the estimated slope is negative.

    2. The fitted line $\hat{\alpha} + \hat{\beta}x$ always passes through the overall mean $(\bar{X}, \bar{Y})$.

    3. Since $\text{cov}(cX, Y) = c \cdot \text{cov}(X, Y)$ and $\text{var}(cX) = c^2 \cdot \text{var}(X)$, if we scale the X values by c then the slope is scaled by $1/c$. If we scale the Y values by c then the slope is scaled by c.

- Once we have $\hat{\alpha}$ and $\hat{\beta}$, we can compute the residuals r_i based on these estimates, i.e.

$$r_i = Y_i - \hat{\alpha} - \hat{\beta} X_i.$$

The following is used to estimate σ:

$$\hat{\sigma} = \sqrt{\frac{\sum_i r_i^2}{n-2}}.$$

- It is also possible to formulate this problem in terms of a model, which is a complete description of the distribution that generated the data.

  The model for linear regression is written:

$$Y_i = \alpha + \beta X_i + \epsilon_i,$$

  where α and β are the population regression coefficients, and the ϵ_i are iid random variables with mean 0 and standard deviation σ. The ϵ_i are called errors.

- Model assumptions:

    1. The means all fall on the line $\alpha + \beta X$.

    2. The ϵ_i are iid (no heteroscedasticity).

    3. The ϵ_i have a normal distribution.

  Assumption 3 is not always necessary. Least squares estimates $\widehat{\alpha}$ and $\widehat{\beta}$ are still valid when the ϵ_i are not normal (as long as 1 and 2 are met).

  However hypothesis tests, CI's, and PI's (derived below) depend on normality of the ϵ_i.

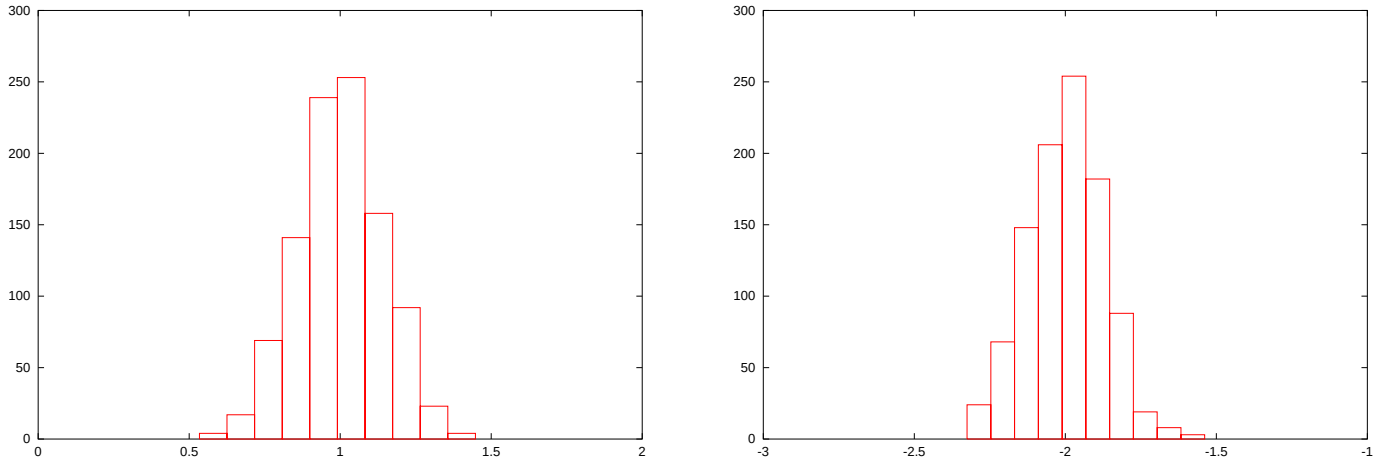
- Since $\widehat{\alpha}$ and $\widehat{\beta}$ are functions of the data, which is random, they are random variables, and hence they have a distribution.

  This distribution reflects the sampling variation that causes $\widehat{\alpha}$ and $\widehat{\beta}$ to differ somewhat from the population values α and β.

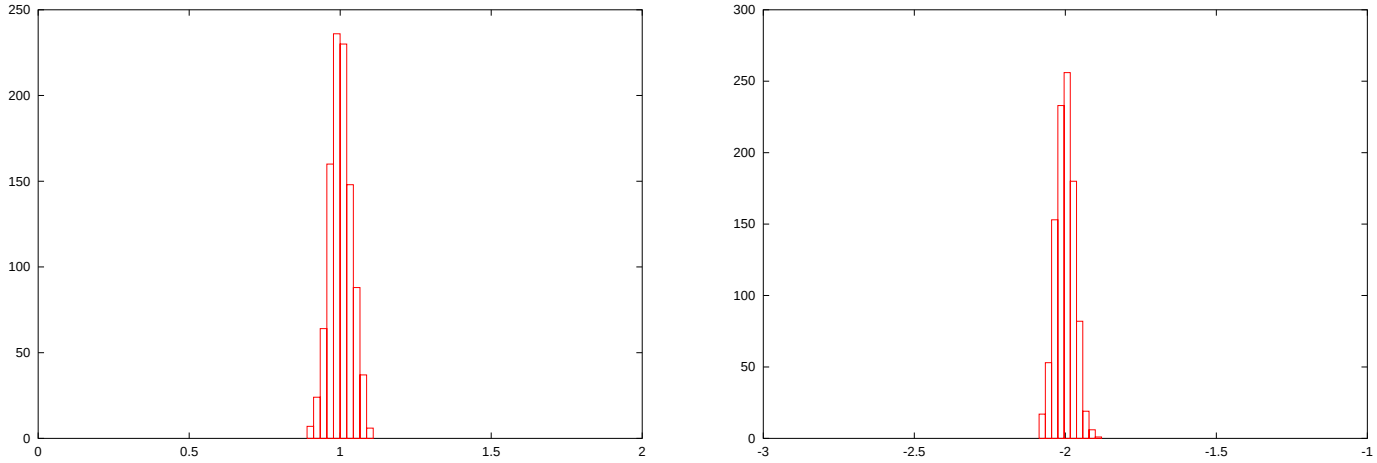
  The sampling variation is less if the sample size n is large, and if the error standard deviation σ is small.

  The sampling variation of $\widehat{\beta}$ is less if the X_i values are more variable.

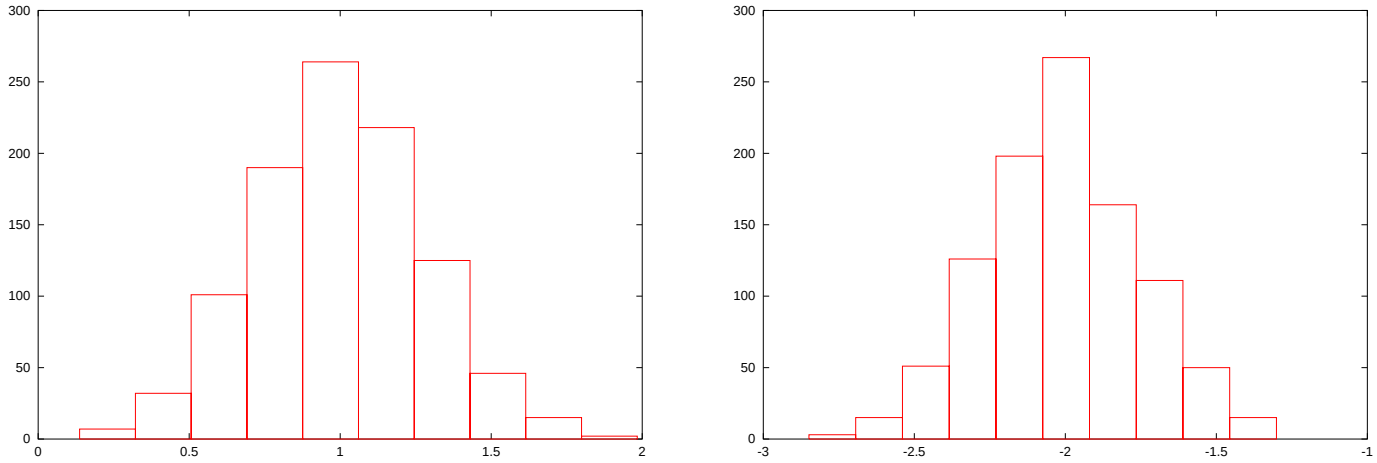
  We will derive formulas later. For now, we can look at histograms.



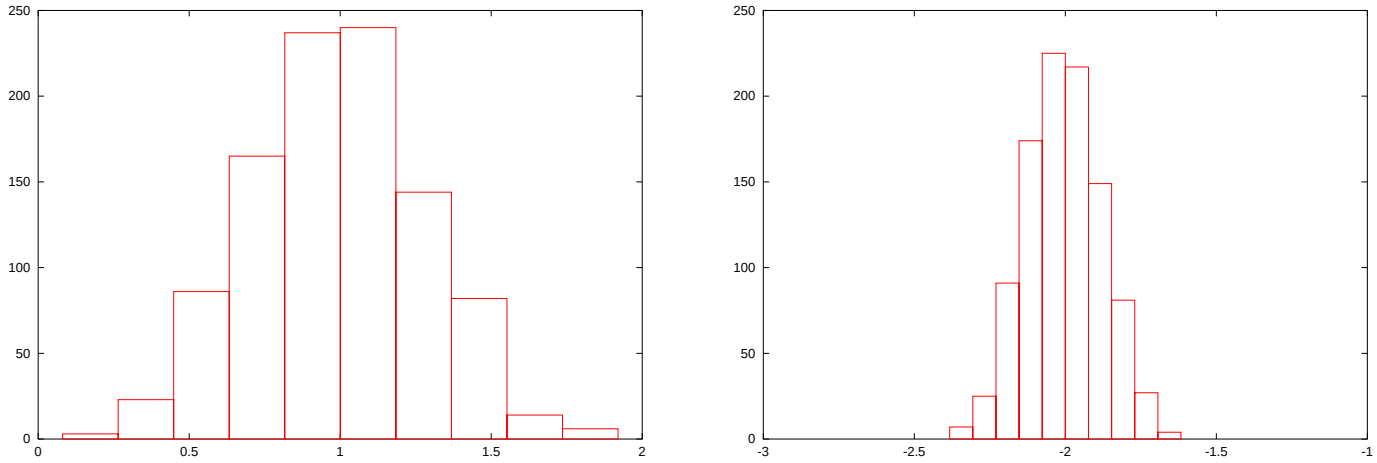
Sampling variation of $\hat{\alpha}$ (left) and $\hat{\beta}$ (right) for 1000 replicates of the simple linear model $Y = 1 - 2X + \epsilon$, where $\mathrm{SD}(\epsilon) = 2$, the sample size is $n = 200$, and $\sigma_X \approx 1.2$.



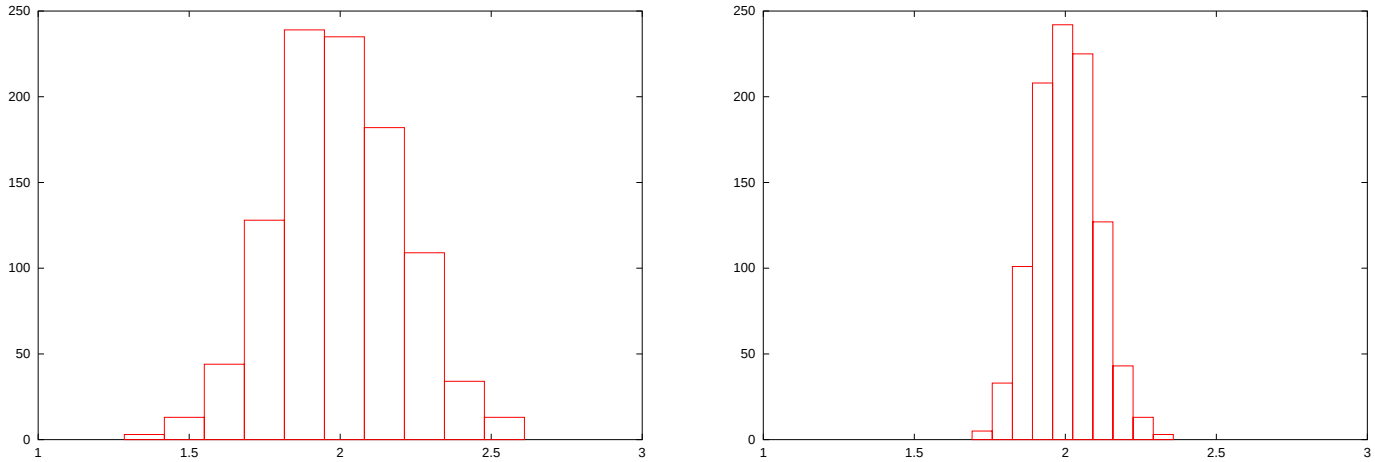
Sampling variation of $\hat{\alpha}$ (left) and $\hat{\beta}$ (right) for 1000 replicates of the simple linear model $Y = 1 - 2X + \epsilon$, where $\mathrm{SD}(\epsilon) = 1/2$, the sample size is $n = 200$, and $\sigma_X \approx 1.2$.



Sampling variation of $\hat{\alpha}$ (left) and $\hat{\beta}$ (right) for 1000 replicates of the simple linear model $Y = 1 - 2X + \epsilon$, where $\mathrm{SD}(\epsilon) = 2$, the sample size is $n = 50$, and $\sigma_X \approx 1.2$.



Sampling variation of $\hat{\alpha}$ (left) and $\hat{\beta}$ (right) for 1000 replicates of the simple linear model $Y = 1 - 2X + \epsilon$, where $\mathrm{SD}(\epsilon) = 2$, the sample size is $n = 50$, and $\sigma_X \approx 2.2$.



Sampling variation of $\widehat{\sigma}$ for 1000 replicates of the simple linear model $Y = 1 - 2X + \epsilon$, where $\mathrm{SD}(\epsilon) = 2$, the sample size is $n = 50$ (left) and $n = 200$ (right), and $\sigma_X \approx 1.2$.

Sampling properties of the least squares estimates

- The following is an identity for the sample covariance:

$$\begin{aligned} \text{cov}(X, Y) &= \frac{1}{n-1} \sum_i (Y_i - \bar{Y})(X_i - \bar{X}) \\ &= \frac{1}{n-1} \sum_i Y_i X_i - \frac{n}{n-1} \bar{Y} \bar{X}. \end{aligned}$$

The average of the products minus the product of the averages (almost).

A similar identity for the sample variance is

$$\begin{aligned} \text{var}(Y) &= \frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2 \\ &= \frac{1}{n-1}\sum_i Y_i^2 - \frac{n}{n-1}\bar{Y}^2. \end{aligned}$$

The average of the squares minus the square of the averages (almost).

- An identify for the regression model $Y_i = \alpha + \beta X_i + \epsilon_i$:

$$\begin{aligned} \frac{1}{n}\sum Y_i &= \frac{1}{n}\sum_i \alpha + \beta X_i + \epsilon_i \\ \bar{Y} &= \alpha + \beta \bar{X} + \bar{\epsilon}. \end{aligned}$$

- Let's get the mean and variance of $\widehat{\beta}$:

  An equivalent way to write the least squares slope estimate is

$$\widehat{\beta} = \frac{\sum_i Y_i X_i - n\bar{Y}\bar{X}}{\sum_i X_i^2 - n\bar{X}^2}.$$

  Now if we substitute $Y_i = \alpha + \beta X_i + \epsilon_i$ into the above we get

$$\widehat{\beta} = \frac{\sum_i (\alpha + \beta X_i + \epsilon_i) X_i - n(\alpha + \beta\bar{X} + \bar{\epsilon})\bar{X}}{\sum_i X_i^2 - n\bar{X}^2}.$$

Since

$$\begin{aligned}\sum_i(\alpha+\beta X_i+\epsilon_i)X_i &= \alpha\sum X_i+\beta\sum_i X_i^2+\sum_i\epsilon_i X_i\\ &= n\alpha\bar{X}+\beta\sum_i X_i^2+\sum_i\epsilon_i X_i\end{aligned}$$

we can simplify the expression for $\widehat{\beta}$ to get

$$\widehat{\beta}=\frac{\beta\sum_i X_i^2-n\beta\bar{X}^2+\sum_i\epsilon_i X_i-n\bar{\epsilon}\bar{X}}{\sum_i X_i^2-n\bar{X}^2},$$

and further to

$$\widehat{\beta}=\beta+\frac{\sum_i\epsilon_i X_i-n\bar{\epsilon}\bar{X}}{\sum_i X_i^2-n\bar{X}^2}$$

To apply this result, by the assumption of the linear model $E\epsilon_i = E\overline{\epsilon} = 0$, so $E\text{cov}(X, \epsilon) = 0$, and we can conclude that $E\widehat{\beta} = \beta$.

This means that $\widehat{\beta}$ is an unbiased estimate of β – it is correct on average.

If we observe an independent SRS every day for 1000 days from the same linear model, and we calculate $\widehat{\beta}_i$ each day for $i = 1, \ldots, 1000$, the daily $\widehat{\beta}_i$ may differ from the population β due to sampling variation, but the average $\sum_i \widehat{\beta}_i/1000$ will be extremely close to β.

- Now that we know $E\widehat{\beta} = \beta$, the corresponding analysis for $\widehat{\alpha}$ is straightforward. Since

$$\widehat{\alpha} = \bar{Y} - \widehat{\beta}\bar{X},$$

then

$$E\widehat{\alpha} = E\bar{Y} - \beta\bar{X},$$

and since $\bar{Y} = \alpha + \beta\bar{X} + \bar{\epsilon}$, so $E\bar{Y} = \alpha + \beta\bar{X}$, thus

$$E\widehat{\alpha} = \alpha + \beta\bar{X} - \beta\bar{X} = \alpha,$$

so α is also unbiased.

- Next we would like to calculate the standard deviation of $\widehat{\beta}$, which will allow us to produce a CI for β.

  Beginning with

$$\widehat{\beta} = \beta + \frac{\sum_i \epsilon_i X_i - n\bar{\epsilon}\bar{X}}{\sum_i X_i^2 - n\bar{X}^2}$$

  and applying the identity $\mathrm{var}(U-V) = \mathrm{var}(U)+\mathrm{var}(V)-2\mathrm{cov}(U,V)$:

$$\mathrm{var}(\widehat{\beta}) = \frac{\mathrm{var}(\sum_i \epsilon_i X_i) + \mathrm{var}(n\bar{\epsilon}\bar{X}) - 2\mathrm{cov}(\sum_i \epsilon_i X_i, n\bar{\epsilon}\bar{X})}{(\sum_i X_i^2 - n\bar{X}^2)^2}.$$

  Simplifying

$$\mathrm{var}(\widehat{\beta}) = \frac{\sum_i X_i^2 \mathrm{var}(\epsilon_i) + n^2\bar{X}^2\mathrm{var}(\bar{\epsilon}) - 2n\bar{X}\sum_i X_i \mathrm{cov}(\epsilon_i, \bar{\epsilon})}{(\sum_i X_i^2 - n\bar{X}^2)^2}.$$

Next, using $\mathrm{var}(\epsilon_i) = \sigma^2$, $\mathrm{var}(\bar{\epsilon}) = \sigma^2/n$:

$$\mathrm{var}(\widehat{\beta}) = \frac{\sigma^2 \sum_i X_i^2 + n\sigma^2 \bar{X}^2 - 2n\bar{X} \sum_i X_i \mathrm{cov}(\epsilon_i, \bar{\epsilon})}{(\sum_i X_i^2 - n\bar{X}^2)^2}.$$

$$\begin{aligned} \mathrm{cov}(\epsilon_i, \bar{\epsilon}) &= \sum_j \mathrm{cov}(\epsilon_i, \epsilon_j)/n \\ &= \sigma^2/n. \end{aligned}$$

So we get

$$\begin{aligned} \mathrm{var}(\widehat{\beta}) &= \frac{\sigma^2 \sum_i X_i^2 + n\sigma^2 \bar{X}^2 - 2n\bar{X} \sum_i X_i \sigma^2/n}{(\sum_i X_i^2 - n\bar{X}^2)^2} \\ &= \frac{\sigma^2 \sum_i X_i^2 + n\sigma^2 \bar{X}^2 - 2n\bar{X}^2 \sigma^2}{(\sum_i X_i^2 - n\bar{X}^2)^2}. \end{aligned}$$

Alomst done:

$$\begin{aligned}
\mathrm{var}(\widehat{\beta}) &= \frac{\sigma^2 \sum_i X_i^2 - n\bar{X}^2\sigma^2}{(\sum_i X_i^2 - n\bar{X}^2)^2} \\
&= \frac{\sigma^2}{\sum_i X_i^2 - n\bar{X}^2} \\
&= \frac{\sigma^2}{(n-1)\mathrm{var}(X)},
\end{aligned}$$

and

$$\mathrm{sd}(\widehat{\beta}) = \frac{\sigma}{\sqrt{n-1}\widehat{\sigma}_X}.$$

- The slope SD formula is consistent with the three factors that influenced the precision of $\hat{\beta}$ in the histograms:

    1. greater sample size reduces the SD

    2. greater σ^2 increases the SD

    3. greater X variability ($\hat{\sigma}_X$) reduces the SD.

- A similar analysis for $\widehat{\alpha}$ yields

$$\text{var}(\widehat{\alpha}) = \sigma^2 \frac{\sum X_i^2/n}{(n-1)\text{var}(X)}.$$

  Thus $\text{var}(\widehat{\alpha}) = \text{var}(\widehat{\beta}) \sum X_i^2/n$.

  Due to the $\sum X_i^2/n$ term the estimate will be more precise when the X_i values are close to zero.

  Since $\widehat{\alpha}$ is the intercept, it's easier to estimate when the data is close to the origin.

- Summary of sampling properties of $\widehat{\alpha}$, $\widehat{\beta}$:

  Both are unbiased: $E\widehat{\alpha} = \alpha$, $E\widehat{\beta} = \beta$.

$$\text{var}(\widehat{\alpha}) = \sigma^2 \frac{\sum X_i^2/n}{(n-1)\text{var}(X)}.$$

$$\text{var}(\widehat{\beta}) = \frac{\sigma^2}{(n-1)\text{var}(X)}$$

Confidence Intervals for $\widehat{\beta}$

- Start with the basic inequality for standardized $\widehat{\beta}$:

$$P(-1.96 \leq \sqrt{n-1}\hat{\sigma}_X \frac{\widehat{\beta} - \beta}{\sigma} \leq 1.96) = 0.95$$

then get β alone in the middle:

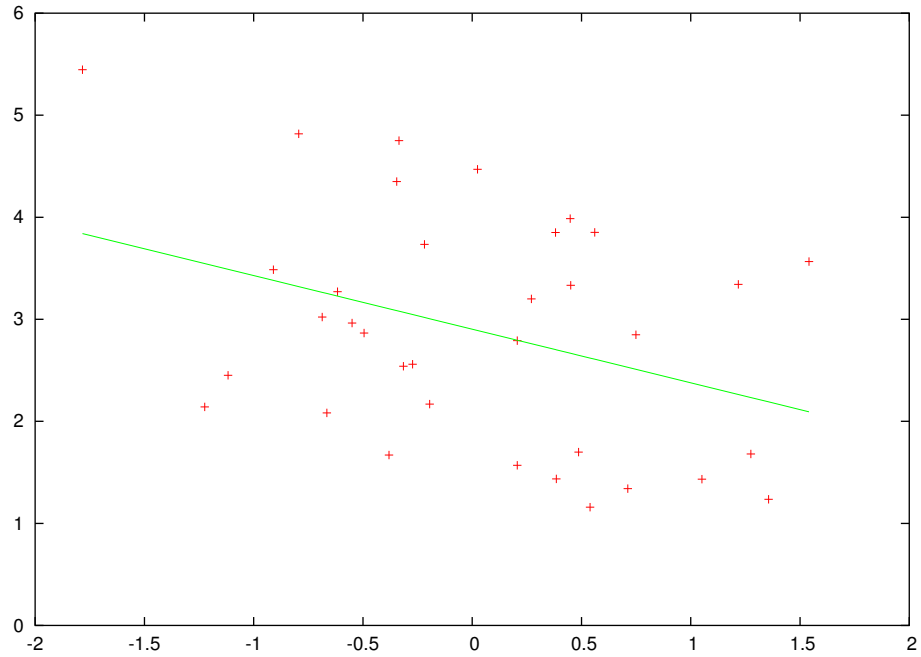
$$P(\widehat{\beta} - 1.96\frac{\sigma}{\sqrt{n-1}\hat{\sigma}_X} \leq \beta \leq \widehat{\beta} + 1.96\frac{\sigma}{\sqrt{n-1}\hat{\sigma}_X}) = .95,$$

Replace 1.96 with 1.64, etc. to get CI's with different coverage probabilities.

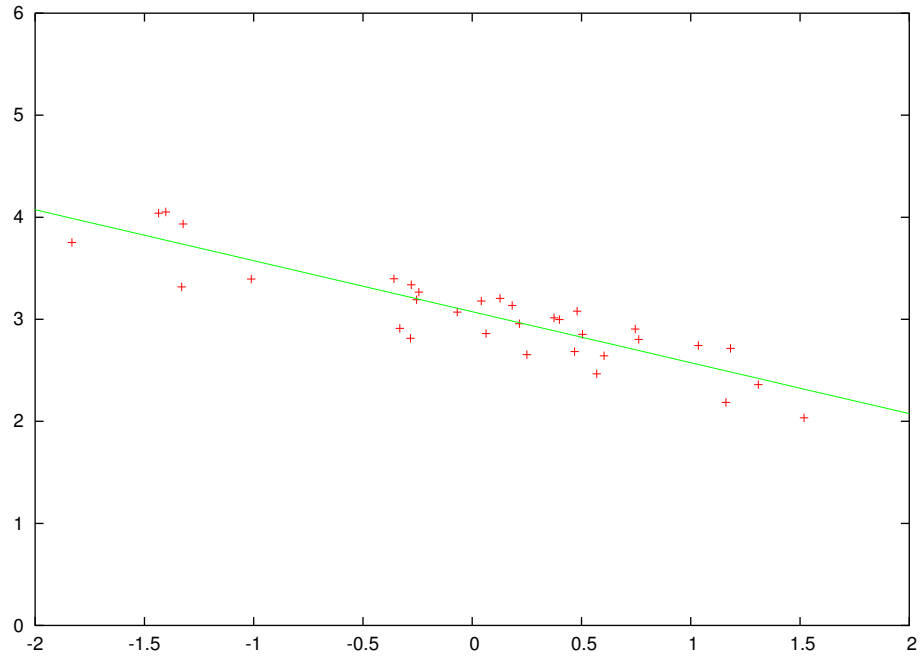
- Note that in general we will not know σ, so we will need to plug-in $\widehat{\sigma}$ (defined above) for σ.

  This plug-in changes the sampling distribution to t_{n-2}, so to be exact, we would replace the 1.96 in the above formula with $Q_T(.975)$, where Q_T is the quantile function of the t_{n-2} distribution.

  If n is reasonably large, the normal quantile will be an excellent approximation.



35 points generated accoring to the model $Y = 3 - X/2 + \epsilon$, where the population standard deviation of ϵ is $\sigma = .8$. The least squares slope estimate is $\widehat{\beta} = -.53$ and the estimate of the error standard deviation is $\widehat{\sigma} = 1.08$. The X standard deviation is $\widehat{\sigma}_X = .79$. A 95% (approximate) CI for β is $-.53 \pm .45$.



35 points generated accoring to the model $Y = 3 - X/2 + \epsilon$, where the population standard deviation of ϵ is $\sigma = .2$. The least squares slope estimate is $\widehat{\beta} = -.50$ and the estimate of the error standard deviation is $\widehat{\sigma} = .23$. The X standard deviation is $\widehat{\sigma}_X = 1.04$. A 95% (approximate) CI for β is $-.50 \pm .07$.

Hypothesis tests for $\widehat{\beta}$

- We can test the hypothesis $\beta = 0$ against alternatives such as $\beta \neq 0$, $\beta > 0$, and $\beta < 0$.

  For example, suppose we are testing the 2-sided alternative $\beta \neq 0$. A suitable test statistic would be

$$T = \frac{\widehat{\beta}\sqrt{n-1}\widehat{\sigma}_X}{\widehat{\sigma}},$$

  which has a t_{n-2} distribution (which may be approximate with a normal distribution if n is not too small).

- *Example:* Suppose we the 35 data points shown in the first plot above, and we calculate $T = -2.29$. Using the t_{33} distribution gives a p-value of .029 (a standard normal distribution gives .022 as the p-value).

Confidence intervals for the regression line

- The fitted value at X, denoted $\widehat{Y}$, is the Y coordinate of the estimated regression line at X:

$$\widehat{Y} = \widehat{\alpha} + \widehat{\beta}X$$

The fitted value is an estimate of the regression function $E(Y|X)$ evaluated at the point X, so we may also write $\widehat{E}(Y|X)$.

Fitted values may be calculated at any X value. If X is one of the observed X values, say $X = X_i$, write $\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}X_i$.

- Since $\widehat{Y}_i$ is a random variable, we can calculate its mean and variance.

  To get the mean, recall that $E\hat{\alpha} = \alpha$ and $E\hat{\beta} = \beta$. Therefore

$$\begin{aligned} E\widehat{Y}_i &= E(\hat{\alpha} + \hat{\beta}X_i) \\ &= E\hat{\alpha} + E\hat{\beta} \cdot X_i \\ &= \alpha + \beta X_i \\ &= EY_i \end{aligned}$$

  Thus $\widehat{Y}_i$ is an unbiased estimate of $E(Y|X)$ evaluated at $X = X_i$.

- To calculate the variance, begin with the following:

$$\begin{aligned}
\text{var}\widehat{Y}_i &= \text{var}(\widehat{\alpha} + \widehat{\beta}X_i) \\
&= \text{var}\widehat{\alpha} + \text{var}(\widehat{\beta}X_i) + 2\text{cov}(\widehat{\alpha}, \widehat{\beta}X_i) \\
&= \text{var}\widehat{\alpha} + X_i^2\text{var}\widehat{\beta} + 2X_i\text{cov}(\widehat{\alpha}, \widehat{\beta}) \\
&= \sigma^2(\sigma_X^2 + \bar{X}^2)/n\sigma_X^2 + X_i^2\sigma^2/n\sigma_X^2 + 2X_i\text{cov}(\widehat{\alpha}, \widehat{\beta})
\end{aligned}$$

To derive $\text{cov}(\widehat{\alpha}, \widehat{\beta})$, similar techniques as were used to calculate $\text{var}\widehat{\alpha}$ and $\text{var}\widehat{\beta}$ can be applied. The result is

$$\text{cov}(\widehat{\alpha}, \widehat{\beta}) = -\frac{\sigma^2\bar{X}}{n\sigma_X^2}.$$

Simplifying yields

$$\text{var}\widehat{Y}_i = \frac{\sigma^2}{n\sigma_X^2}(\sigma_X^2 + \bar{X}^2 + X_i^2 - 2X_i\bar{X}),$$

which reduces further to

$$\text{var}\widehat{Y}_i = \frac{\sigma^2}{n\sigma_X^2}(\sigma_X^2 + (X_i - \bar{X})^2).$$

An equivalent expression is

$$\text{var}\widehat{Y}_i = \frac{\sigma^2}{n}\left(1 + \left(\frac{X_i - \bar{X}}{\sigma_X}\right)^2\right).$$

To simplify notation define

$$\sigma_i^2 = \frac{1}{n}\left(1 + \left(\frac{X_i - \bar{X}}{\sigma_X}\right)^2\right)$$

so that var$\widehat{Y}_i = \sigma^2\sigma_i^2$.

Key point: Difficulty in estimating the mean response varies with X, and the variance is smallest when $X_i = \bar{X}$.

The smallest value of var$\widehat{Y}_i$ occurs when $X_i = \bar{X}$, which is var$\widehat{Y}_i = \sigma^2/n$.

This is the same as the variance of the sample mean in a univariate analysis.

Thus for a given sample size n, an estimate of the conditional mean $E(Y|X = x)$ is more variable than an estimate of the marginal mean EY, except for estimating $E(Y|X = \bar{X})$, which is equally variable as the estimate of EY.

This makes sense, since the fitted value at $\bar{X}$ is

$$\begin{aligned} \widehat{\alpha} + \widehat{\beta}\bar{X} &= (\bar{Y} - \text{cov}(X,Y)\bar{X}/\text{var}(X)) + \text{cov}(X,Y)\bar{X}/\text{var}(X) \\ &= \bar{Y}, \end{aligned}$$

which has variance σ^2/n.

- We now know the mean and variance of $\widehat{Y}_i$. Standardizing yields

$$P(-1.96 \leq \frac{\widehat{Y}_i - (\alpha + \beta X_i)}{\sigma\sigma_i} \leq 1.96) = .95,$$

equivalently

$$P(\widehat{Y}_i - 1.96\sigma\sigma_i \leq \alpha + \beta X_i \leq \widehat{Y}_i + 1.96\sigma\sigma_i) = .95.$$

This gives a 95% CI for EY_i.

- Since σ is unknown we must plug-in $\hat{\sigma}$ for σ in the CI. Thus we get the approximate CI

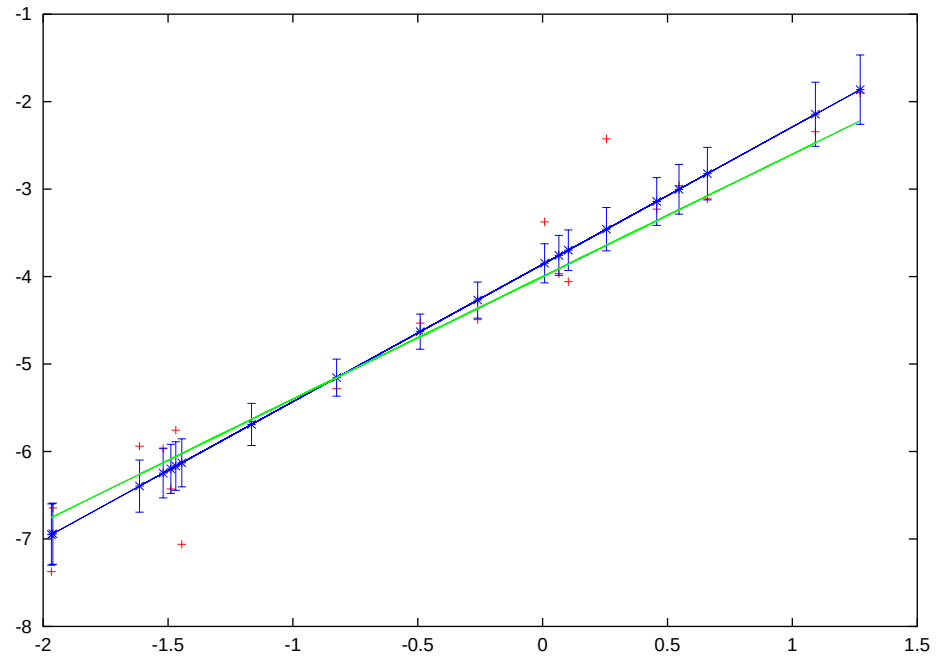
$$P(\hat{Y}_i - 1.96\hat{\sigma}\sigma_i \leq \alpha + \beta X_i \leq \hat{Y}_i + 1.96\hat{\sigma}\sigma_i) \approx 0.95.$$

We can make the coverage probability exactly 0.95 by using the t_{n-2} distribution to calculate quantiles:

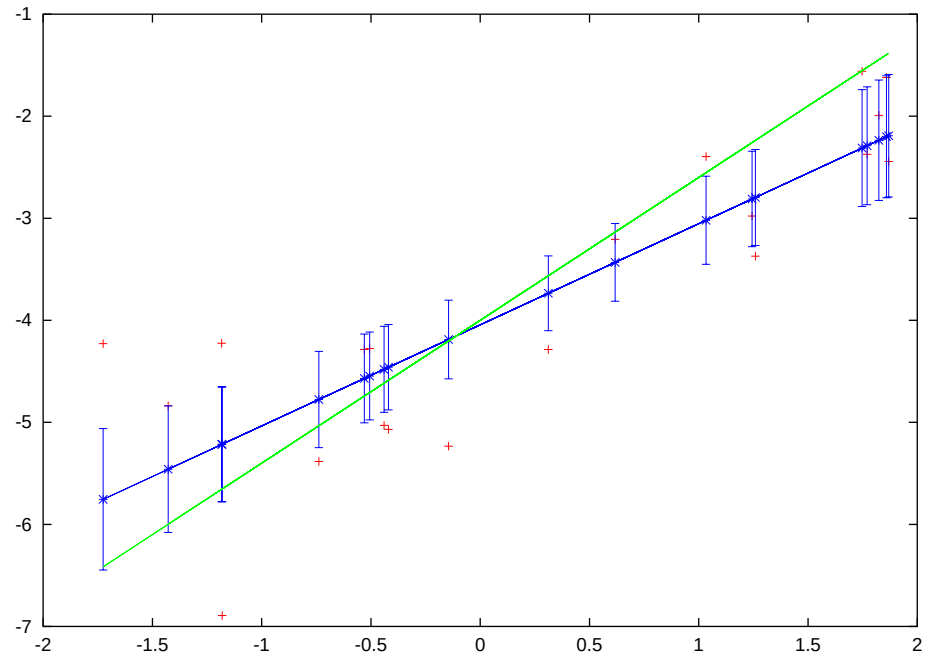
$$P(\hat{Y}_i - Q(0.975)\hat{\sigma}\sigma_i \leq \alpha + \beta X_i \leq \hat{Y}_i + Q(0.975)\sigma\sigma_i) = 0.95.$$

- The following show CI's for the population regression function $E(Y|X)$. In each data figure, a CI is formed for each X_i value. Note that the goal of each CI is to cover the green line, and this should happen 95% of the time.

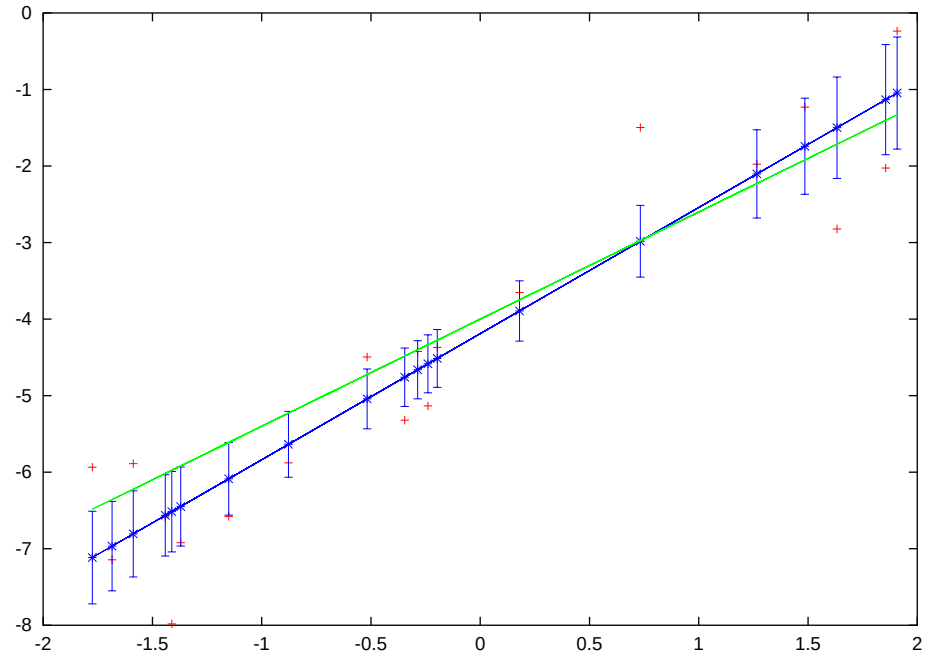
  Note also that the CI's are narrower for X_i close to $\bar{X}$ compared to X_i that are far from $\bar{X}$. Also note that the CI's are longer when σ is greater.



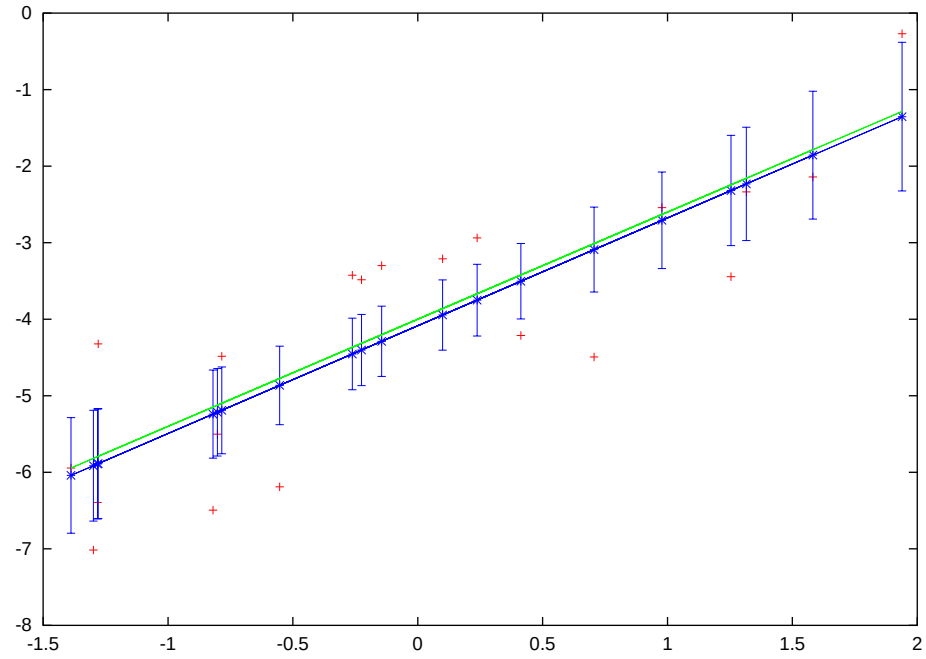
The red points are a bivariate data set generated according to the model $Y = -4+1.4X+\epsilon$, where $\mathrm{SD}(\epsilon) = .4$. The green line is the population regression function, the blue line is the fitted regression function, and the vertical blue bars show 95% CI's for $E(Y|X = X_i)$ at each X_i value.



The red points are a bivariate data set generated according to the model $Y = -4 + 1.4X + \epsilon$, where $\mathrm{SD}(\epsilon) = 1$. The green line is the population regression function, the blue line is the fitted regression function, and the vertical blue bars show 95% CI's for $E(Y|X = X_i)$ at each X_i value.



This is an independent realization from the model shown in the previous figure.



Another independent realization.

Prediction intervals

- Suppose we observe a new X point X^* after having calculated $\widehat{\alpha}$ and $\widehat{\beta}$ based on an independent data set. How can we predict the Y value Y^* corresponding to X^*?

  It makes sense to use $\widehat{\alpha} + \widehat{\beta} X^*$ as the prediction. We would also like to quantify the uncertainty in this prediction.

- First note that $E(\hat{\alpha} + \hat{\beta}X^*) = \alpha + \beta X^* = EY^*$, so the prediction is unbiased.

  Calculate the variance of the prediction error:

$$\begin{aligned}
\text{var}(Y^* - \hat{\alpha} - \hat{\beta}X^*) &= \text{var}Y^* + \text{var}(\hat{\alpha} + \hat{\beta}X^*) - 2\text{cov}(Y^*, \hat{\alpha} + \hat{\beta}X^*) \\
&= \sigma^2 + \sigma^2(1 + ((X^* - \bar{X})/\sigma_X)^2)/n \\
&= \sigma^2(1 + (1 + ((X^* - \bar{X})/\sigma_X)^2)/n) \\
&= \sigma^2(1 + \sigma_*^2).
\end{aligned}$$

  Note that the covariance term is 0 since Y^* is independent from the data used to fit the model.

When n is large, α and β are very precisely estimated, so σ_* is very small, and the variance of the prediction error is $\approx \sigma^2$ – nearly all of the uncertainty comes from the error term ϵ.

The prediction interval

$$P(-1.96 \leq \frac{Y^* - \hat{\alpha} - \hat{\beta}X^*}{\sigma\sqrt{1+\sigma_*^2}} \leq 1.96) = .95,$$

can be rewritten

$$P(\hat{\alpha}+\hat{\beta}X^*-1.96\sigma\sqrt{1+\sigma_*^2} \leq Y^* \leq \hat{\alpha}+\hat{\beta}X^*+1.96\sigma\sqrt{1+\sigma_*^2}) = .95.$$

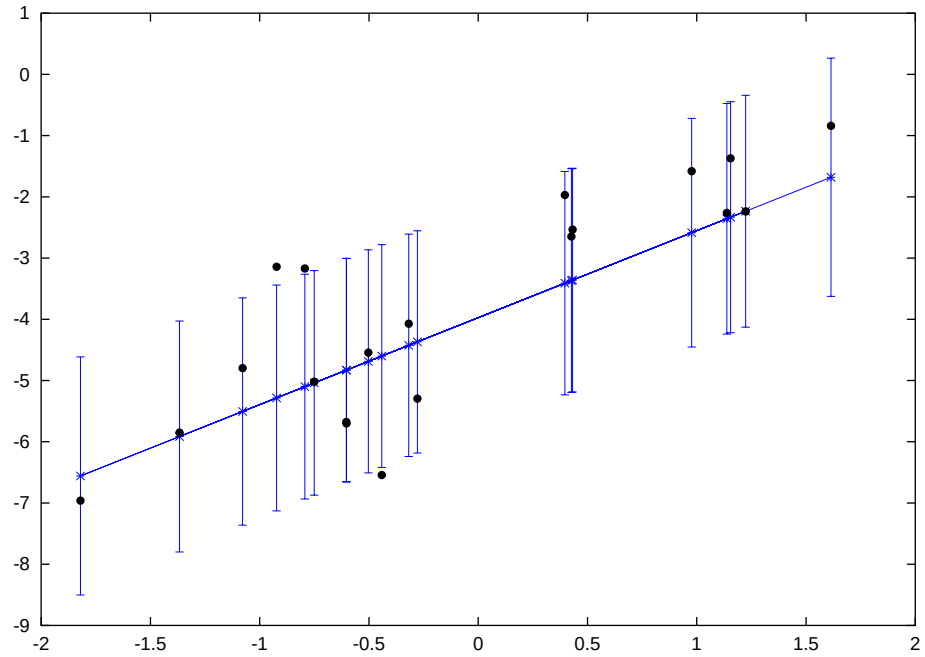
- As with the CI, we will plug-in $\hat{\sigma}$ for σ, making the coverage approximate:

$$P(\hat{\alpha}+\hat{\beta}X^*-1.96\hat{\sigma}\sqrt{1+\sigma_*^2} \leq Y^* \leq \hat{\alpha}+\hat{\beta}X^*+1.96\hat{\sigma}\sqrt{1+\sigma_*^2}) \approx .95.$$

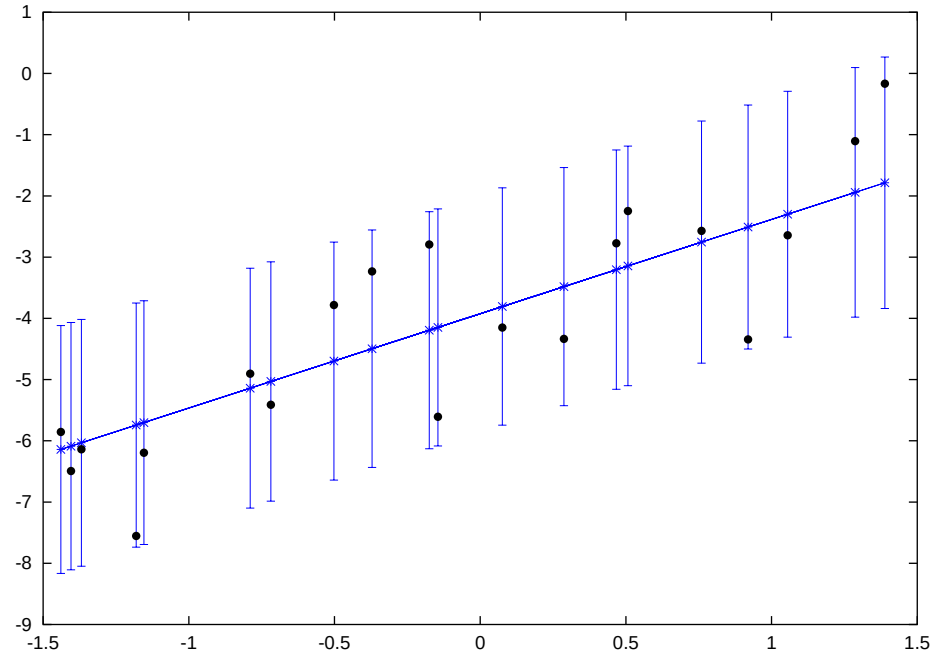
For the coverage probability to be exactly 95%, 1.96 should be replaced with $Q(0.975)$, where Q is the t_{n-2} quantile function.

- The following two figures show fitted regression lines for a data set of size $n = 20$ (the fitted regression line is shown but the data are not shown). Then 95% PI's are calculated at each X_i, and an independent data set of size $n = 20$ is generated at the same set of X_i values. The PI's should cover the new data values 95% of the time.

  The PI's are slightly narrower in the center, but this is hard to see unless n is quite small.



A set of $n = 20$ bivariate observations were generated according to the model $Y = -4 + 1.4X + \epsilon$, where $\mathrm{SD}(\epsilon) = 1$. Based on these points (which are not shown), the fitted regression line (shown in blue) was determined. Next an independent set was generated (black points), with one point having each X_i value from the original data. The vertical blue bars show 95% PI's at each X_i value.



An independent replication of the previous figure.

Residuals

- The residual r_i is the difference between the fitted and observed values at X_i: $r_i = Y_i - \hat{Y}_i$.

  The residual is a random variable since it depends on the data.

  Be sure you understand the difference between the residual (r_i) and the error (ϵ_i):

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \epsilon_i \\ Y_i &= \hat{\alpha} + \hat{\beta} X_i + r_i \end{aligned}$$

- Since $E\epsilon_i = 0$, $EY_i = \alpha + \beta X_i$. Thus $Er_i = EY_i - E\widehat{Y}_i = 0$.

  Calculate the sum of the residuals:

$$\begin{aligned} \sum r_i &= \sum Y_i - \sum \widehat{Y}_i \\ &= \sum Y_i - n\widehat{\alpha} - \widehat{\beta} \sum X_i. \end{aligned}$$

  So the average residual is $\bar{r} = \bar{Y} - \widehat{\alpha} - \widehat{\beta}\bar{X}$.

  Since $\widehat{\alpha} = \bar{Y} - \widehat{\beta}\bar{X}$, it follows that $\bar{r} = 0$.

- Each residual r_i estimates the corresponding error ϵ_i. The ϵ_i are iid, however the r_i are not iid.

  We already saw that $Er_i = 0$. To calculate $\mathrm{var} r_i$, begin with:

$$\begin{aligned} \mathrm{var} r_i &= \mathrm{var} Y_i + \mathrm{var} \widehat{Y}_i - 2\mathrm{cov}(Y_i, \widehat{Y}_i) \\ &= \sigma^2 + \sigma^2 \sigma_i^2 - 2\mathrm{cov}(Y_i, \widehat{Y}_i). \end{aligned}$$

It is a fact that $\text{cov}(Y_i, \widehat{Y}_i) = \sigma^2\sigma_i^2$, thus

$$\text{var}r_i = \sigma^2 + \sigma^2\sigma_i^2 - 2\sigma^2\sigma_i^2 = \sigma^2(1 - \sigma_i^2).$$

Since a variance must be positive, it must be true that $\sigma_i^2 \leq 1$. This is easier to see by rewriting σ_i^2 as follows:

$$\sigma_i^2 = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_j (X_j - \bar{X})^2}.$$

It is true that

$$\frac{(X_i - \bar{X})^2}{\sum_j (X_j - \bar{X})^2} \leq \frac{n-1}{n},$$

but we will not prove this.

If the sample size is $n = 2$, then $(X_1 - \bar{X})^2 = (X_2 - \bar{X})^2$, so

$$\frac{(X_i - \bar{X})^2}{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2} = \frac{n-1}{n} = \frac{1}{2},$$

so the variance of r_i is zero in that case.

This makes sense since the regression line fits the data with no residual when there are only two data points.

The residuals r_i are less variable than the errors ϵ_i since $\sigma_i^2 \sigma^2 \leq \sigma^2$. Thus the fitted regression line is closer to the data than the population regression line. This is called overfitting.

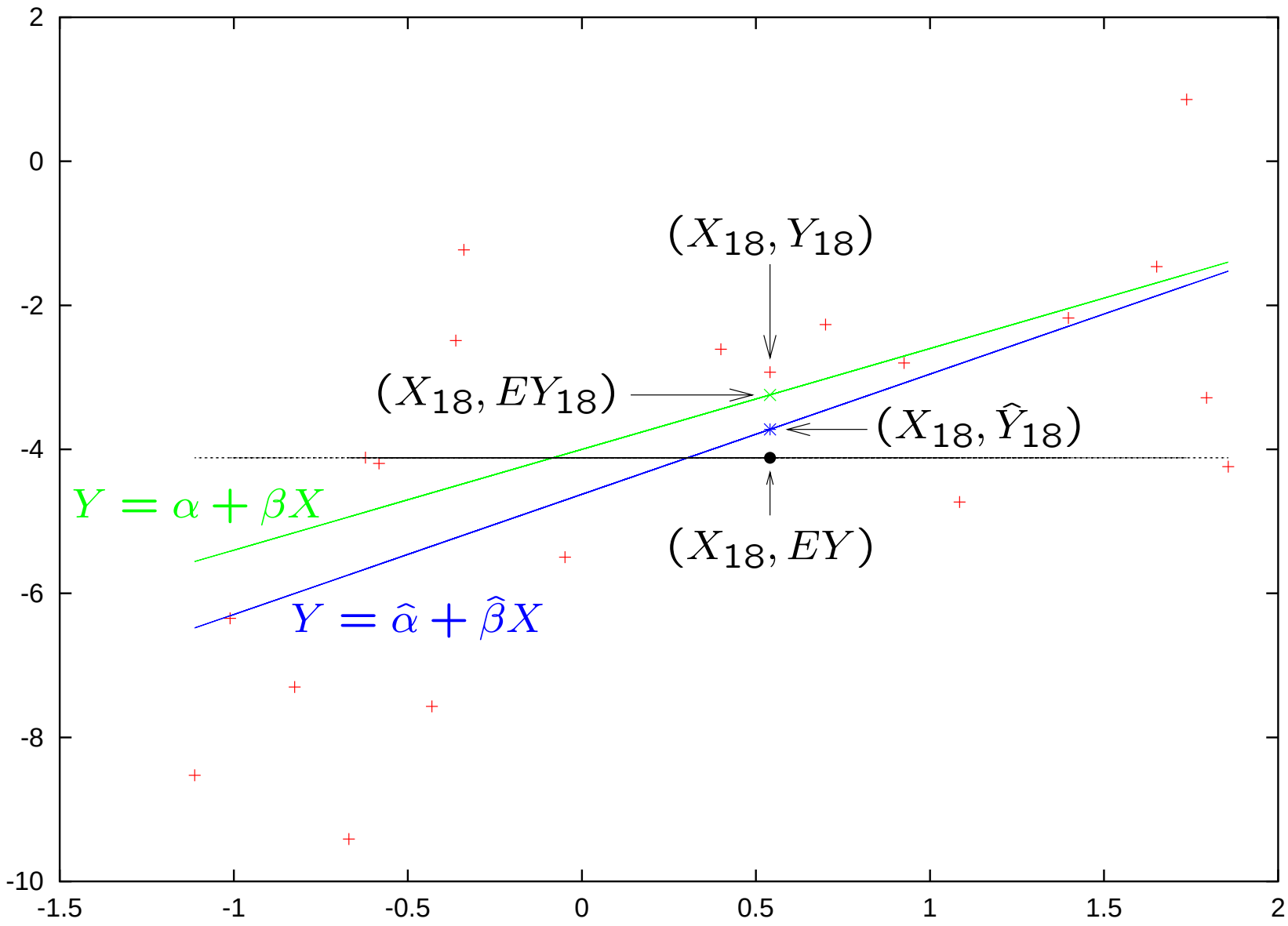
Sums of squares

- We would like to understand how the following quantitites are related:
  - $Y_i - \bar{Y}$ (observed minus marginal mean)
  - $Y_i - \widehat{Y}_i = r_i$ (residual: observed minus linear fit)
  - $\widehat{Y}_i - \bar{Y}$ (linear fit minus marginal mean).

  All three average out to zero over the data:

$$\frac{1}{n}\sum Y_i - \bar{Y} = \frac{1}{n}\sum r_i = \frac{1}{n}\sum \widehat{Y}_i - \bar{Y} = 0.$$

- The following figure shows $n = 20$ points generated from the model $Y = -4 + 1.4X + \epsilon$, where SD(ϵ) = 2. The green line is the population regression line, the blue line is the fitted regression line, and the black line is the constant line $Y = EY$. Note that another way to write EY_{18} is $E(Y|X = X_{18})$.


2
0
-2
-4
-6
-8
-10
-1.5
-1
-0.5
0
0.5
1
1.5
2
(X_{18}, Y_{18})
(X_{18}, EY_{18})
$(X_{18}, \widehat{Y}_{18})$
(X_{18}, EY)
$Y = \alpha + \beta X$
$Y = \widehat{\alpha} + \widehat{\beta} X$

- We will begin with two identities. First,

$$\begin{aligned}\widehat{Y}_i &= \widehat{\alpha} + \widehat{\beta} X_i \\ &= \bar{Y} - \widehat{\beta}\bar{X} + \widehat{\beta} X_i \\ &= \bar{Y} + \widehat{\beta}(X_i - \bar{X}).\end{aligned}$$

As a consequence, $\widehat{Y}_i - \bar{Y} = \widehat{\beta}(X_i - \bar{X})$.

Second,

$$\begin{aligned}Y_i - \widehat{Y}_i &= Y_i - (\widehat{\alpha} + \widehat{\beta} X_i) \\ &= Y_i - (\bar{Y} - \widehat{\beta}\bar{X} + \widehat{\beta} X_i) \\ &= Y_i - \bar{Y} - \widehat{\beta}(X_i - \bar{X})\end{aligned}$$

- Now consider the following "sum of squares":

$$\begin{aligned}\sum(Y_i-\bar{Y})^2 &= \sum(Y_i-\hat{Y}_i+\hat{Y}_i-\bar{Y})^2 \\ &= \sum(Y_i-\hat{Y}_i)^2+\sum(\hat{Y}_i-\bar{Y})^2+2\sum(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y}).\end{aligned}$$

Applying the above identities to the final term:

$$\begin{aligned}\sum(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y}) &= \hat{\beta}\sum(Y_i-\bar{Y}-\hat{\beta}(X_i-\bar{X}))(X_i-\bar{X}) \\ &= \hat{\beta}\sum(Y_i-\bar{Y})(X_i-\bar{X})-\hat{\beta}(X_i-\bar{X})^2 \\ &= \hat{\beta}(n-1)\mathrm{cov}(Y,X)-(n-1)\hat{\beta}^2\mathrm{var}(X) \\ &= \hat{\beta}(n-1)\mathrm{cov}(Y,X)-(n-1)\hat{\beta}\mathrm{cov}(Y,X) \\ &= 0\end{aligned}$$

Since the mean of $Y_i - \widehat{Y}_i$ and the mean of $\widehat{Y}_i - \bar{Y}$ are both zero,

$$\sum (Y_i - \widehat{Y}_i)(\widehat{Y}_i - \bar{Y}) = (n-1)\text{cov}(Y_i - \widehat{Y}_i, \widehat{Y}_i - \bar{Y}).$$

Therefore we have shown that the residual $r_i = Y_i - \widehat{Y}_i$ and the fitted values $\widehat{Y}_i$ are uncorrelated.

We now have the following "sum of squares law":

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \widehat{Y}_i)^2 + \sum (\widehat{Y}_i - \bar{Y})^2.$$

- The following terminology is used:

Formula	Name	Abbrev.
$\sum(Y_i - \bar{Y})^2$	Total sum of squares	SSTO
$\sum(Y_i - \hat{Y}_i)^2$	Residual sum of squares	SSE
$\sum(\hat{Y}_i - \bar{Y})^2$	Regression sum of squares	SSR.

The sum of squares law is expressed: "SSTO = SSE + SSR".

- Corresponding to each "sum of squares" is a "degrees of freedom" (DF). Dividing the sum of squares by the DF gives the "mean square".

Abbrev.	DF	Formula
MSTO	n-1	$\sum(Y_i - \bar{Y})^2/(n-1)$
MSE	n-2	$\sum(Y_i - \hat{Y}_i)^2/(n-2)$
MSR	1	$\sum(\hat{Y}_i - \bar{Y})^2$

Note that the MSTO is the sample variance, and the MSE is the estimate of $\hat{\sigma}^2$ in the regression model.

The "SS" values add: SSTO = SSE + SSR and the degrees of freedom add: $n - 1 = (n-2) + 1$.

The "MS" values do not add: MSTO $\neq$ MSE + MSR.

- If the model fits that data well, MSE will be small and MSR will be large. Conversely, if the model fits the data poorly then MSE will be large and MSR will be small. Thus the statistic

$$F = \frac{\text{MSR}}{\text{MSE}}$$

can be used to evaluate the fit of the linear model (bigger F = better fit).

The distribution of F is an "F distribution with $1, n-2$ DF", or $F_{1,n-2}$.

We can test the null hypothesis that the data follow a model $Y_i = \mu + \epsilon_i$ against the alternative that the data follow a model $Y_i = \alpha + \beta X_i + \epsilon_i$ using the F statistic (an "F test"). A computer package or a table of the F distribution can be used to determine a p-value.

- In the case of simple linear regression, the F test is equivalent to the hypothesis test $\beta = 0$ versus $\beta \neq 0$. Later when we come to multiple linear regression, this will not be the case.

  A useful way to think about what the F-test is evaluating is that the null hypothesis is "all Y values have the same expected value" and the alternative is that "the expected value of Y_i depends on the value of X_i".

Diagnostics

- In practice, we may not be certain that the assumptions underlying the linear model are satisfied by a particular data set. To review, the key assumptions are:

    1. The conditional mean function $E(Y|X)$ is linear.

    2. The conditional variance function $\text{var}(Y|X)$ is constant.

    3. The errors are normal and independent.

  Note that (3) is not essential for the estimates to be valid, but should be approximately satisfied for confidence intervals and hypothesis tests to be valid. If the sample size is large, then it is less crucial that (3) be met.

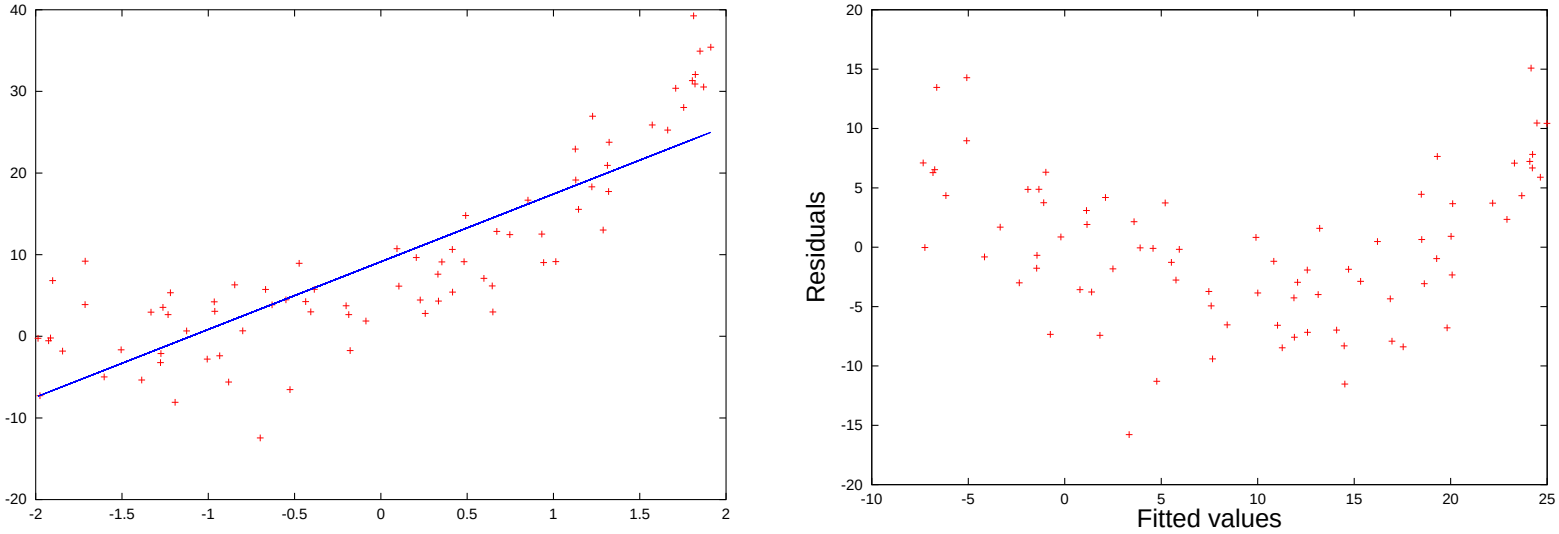
- To assess whether (1) and (2) are satsified, make a scatterplot of the residuals r_i against the fitted values $\widehat{Y}_i$.

  This is called a "residuals on fitted values plot".

  Recall that we showed above that r_i and $\widehat{Y}_i$ are uncorrelated.

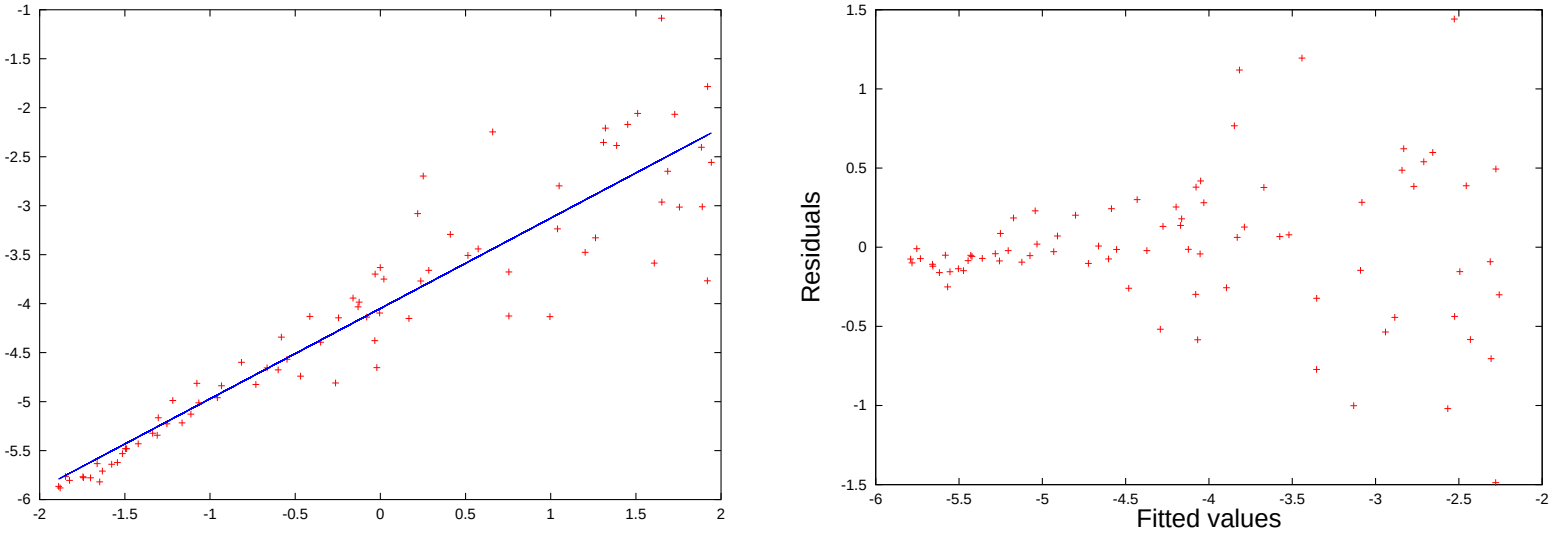
  Thus if the model assumptions are met this plot should look like iid noise – there should be no visually apparent trends or patterns.

For example, the following shows how a residual on fitted values plot can be used to detect nonlinearity in the regression function.



Left: A bivariate data set (red points) with fitted regression line (blue). Right: A diagnostic plot of residuals on fitted values.

The following shows how a residual on fitted values plot can be used to detect heteroscedasticity.



Left: A bivariate data set (red points) with fitted regression line (blue). Right: A diagnostic plot of residuals on fitted values.

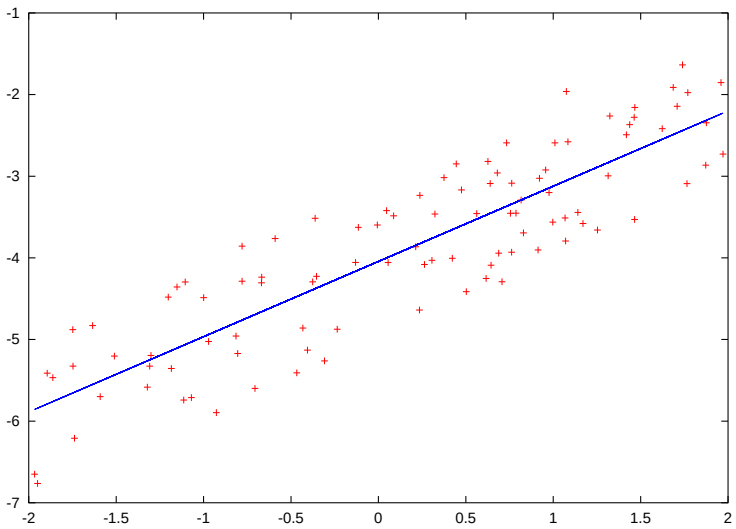
- Suppose that the observations were collected in sequence, say two per day for a period of one month, yielding $n = 60$ points. There may be some concern that the distribution has shifted over time.

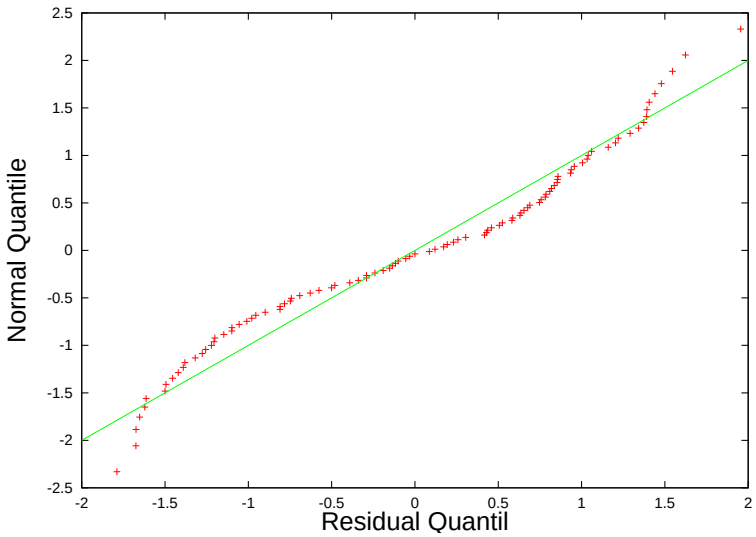
  These are called "sequence effects" or "time of measurement effects".

  To detect these effects, plot the residual r_i against time. There should be no pattern in the plot.

- To assess the normality of the errors use a normal probability plot of the residuals.

  For example, the following shows a bivariate data set in which the errors are uniform on $[-1, 1]$ (i.e. any value in that interval is equally likely to occur as the error). This is evident in the quantile plot of the r_i.





Left: A bivariate data set (red points) with fitted regression line (blue). Right: A normal probability plot of the residuals.

Outliers and leverage points

- If the assumptions of the linear model are met, the variance of the residual r_i is a bit less than σ^2.

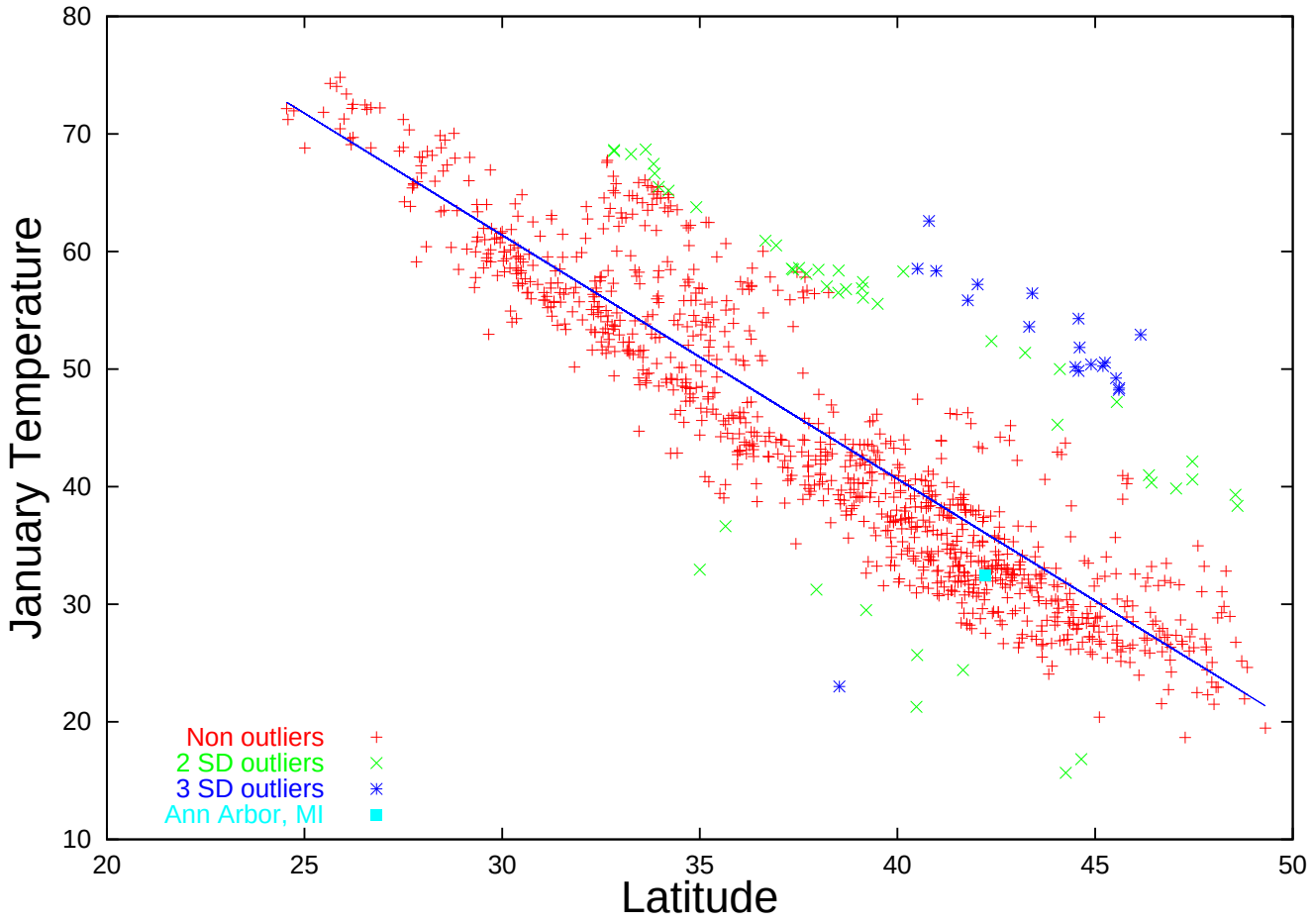
  If the residuals are approximately normal, it is very unlikely that a given residual r_i will differ from its mean (which is 0) by more than $3\hat{\sigma}$. Such an observation is called an outlier.

  An alternative is to calculate the IQR of the residuals, and consider an outlier to be any point with residual greater than 2 or 2.5 times the IQR.

- In some cases, outliers may be discarded, and the regression model refit to the remaining data. This can give a better description of the trend for the vast majority of observations.

  On the other hand, the outliers may be the most important observations in terms of revealing something new about the system being studied, so they can not simply be ignored.

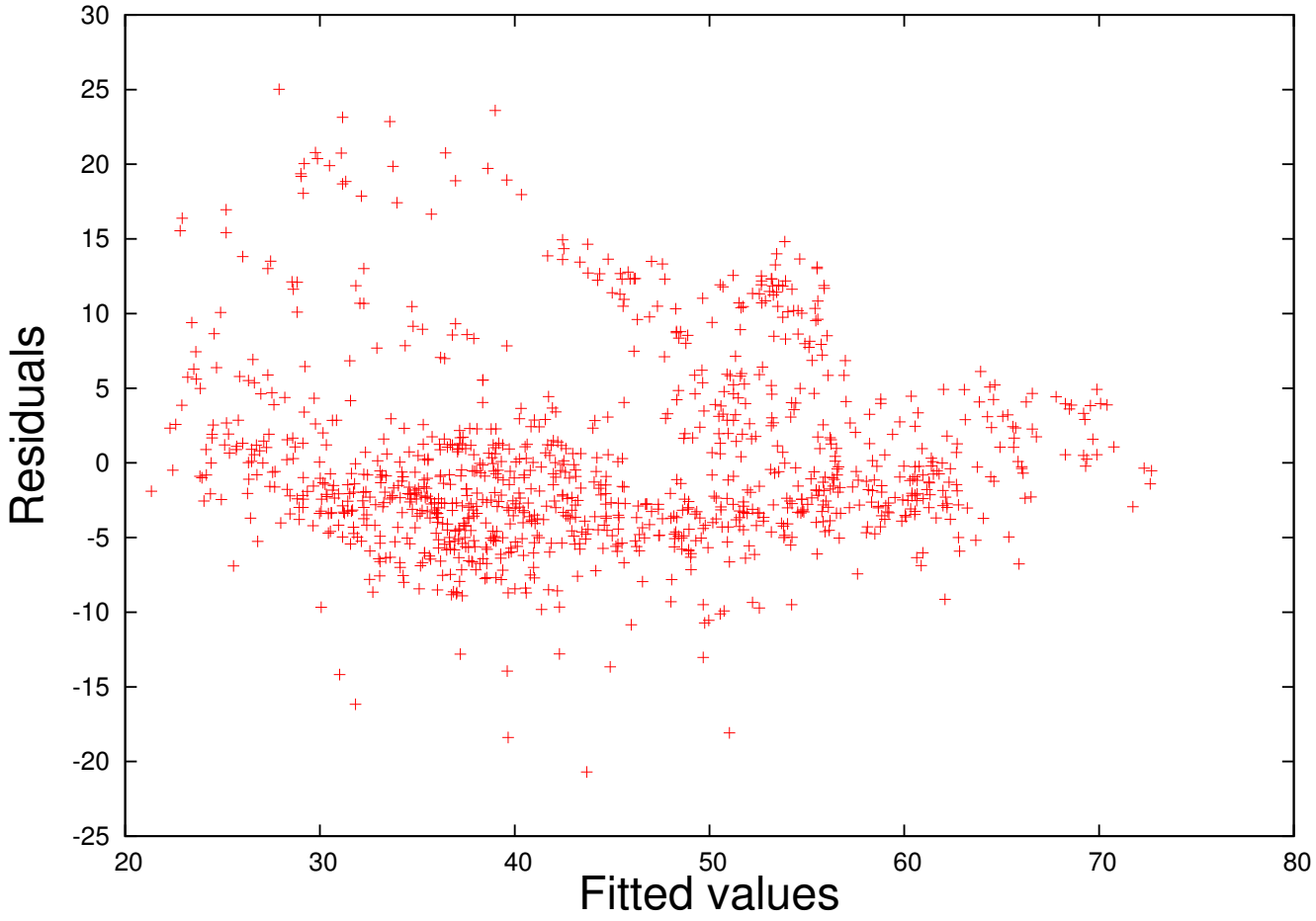
- *Example:* The following figure shows the fitted least squares regression line (blue) for the regression of January maximum average temperature on latitude. Points greater than 2 and greater than 3 times $\widehat{\sigma}$ are shown. The green points do not meet our definition of "outlier", but they are somewhat atypical.


January Temperature
Latitude
80
70
60
50
40
30
20
10
20
25
30
35
40
45
50
Non outliers
2 SD outliers
3 SD outliers
Ann Arbor, MI

It turns out that of the 19 outliers, 18 are warmer than expected, and these stations are all in northern California and Oregon.

The one outlier station that is substantially colder than expected is in Gunnison County, Colorado, which is very high in elevation (at 2,339 ft, it is the fourth highest of 1072 stations in the data set).

In January 2001, Ann Arbor, Michigan was slightly colder than the fitted value (i.e. it was a bit colder here than in other places of similar latitude).



A plot of residuals on fitted values for the regression of January maximum temperature on latitude.

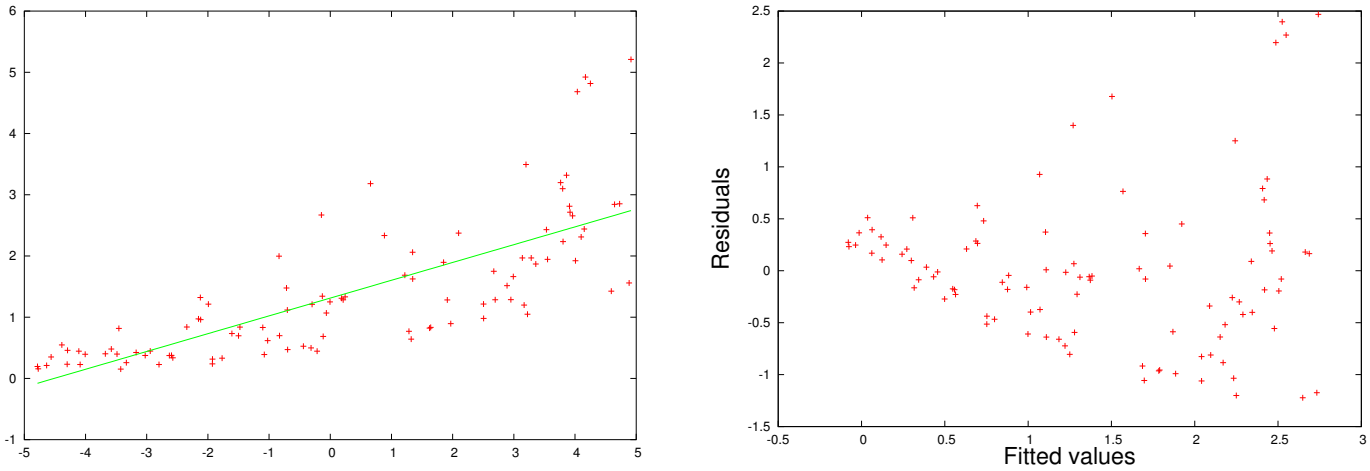
Transformations

- If the assumptions of the linear model are not met, it may be possible to transform the data so that a linear fit to the transformed data meets the assumptions more closely.

  Your options are to transform Y only, transform X only, or transform both Y and X.

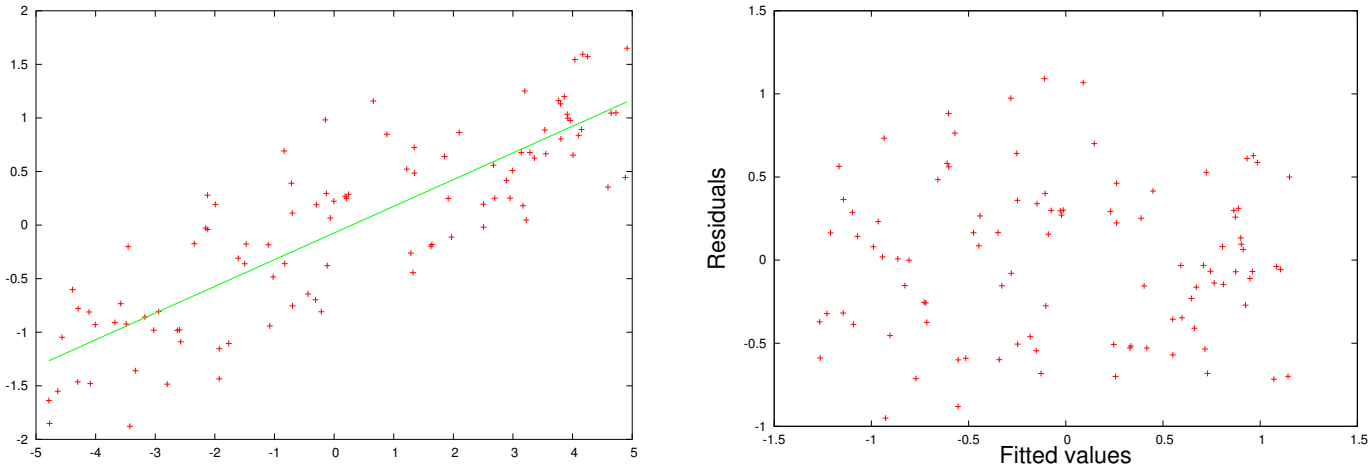
  The most useful transforms are the log transform $X \rightarrow \log(X+c)$ and the power transform $X \rightarrow (X+c)^q$.

  The following example shows a situation where the errors do not seem to be homoscedastic.



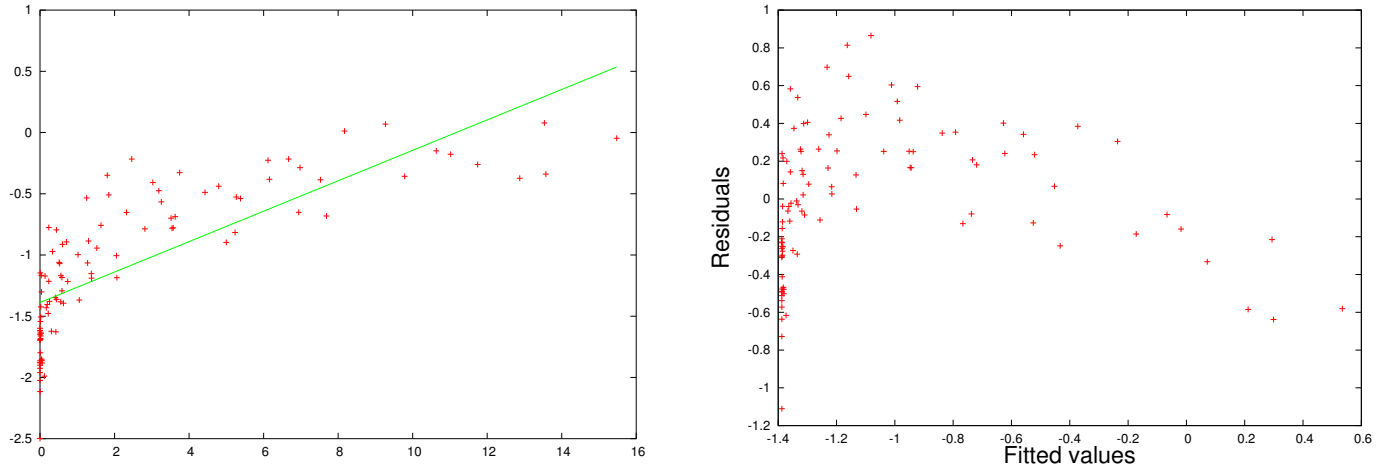
Left: Scatterplot of the raw data, with the regression line drawn in green. Right: Scatterplot of residuals on fitted values.

Here is the same example where the Y variable was transformed to $\log(Y)$:



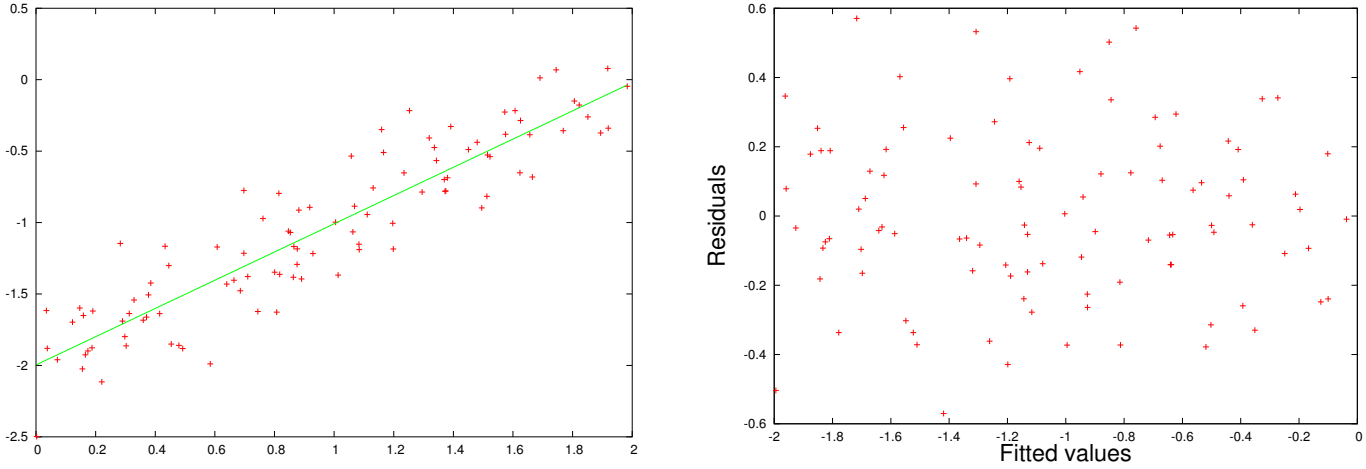
Left: Scatterplot of the transformed data, with the regression line drawn in green. Right: Scatterplot of residuals on fitted values.

- Another common situation occurs when the X values are skewed:



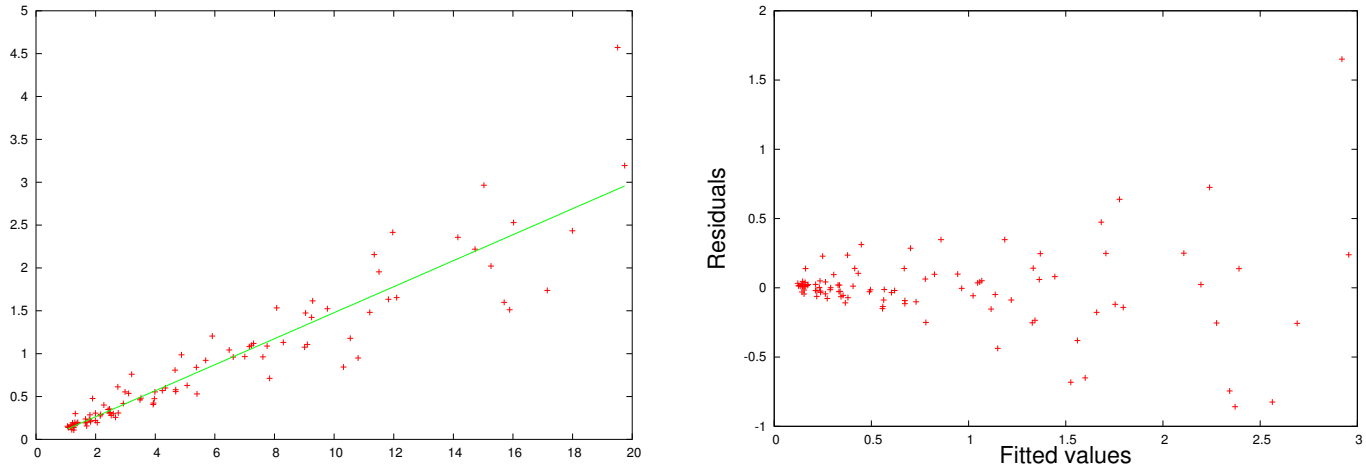
Left: Scatterplot of the raw data, with the regression line drawn in green. Right: Scatterplot of residuals on fitted values.

In this case transforming X to $X^{1/4}$ removed the skew:



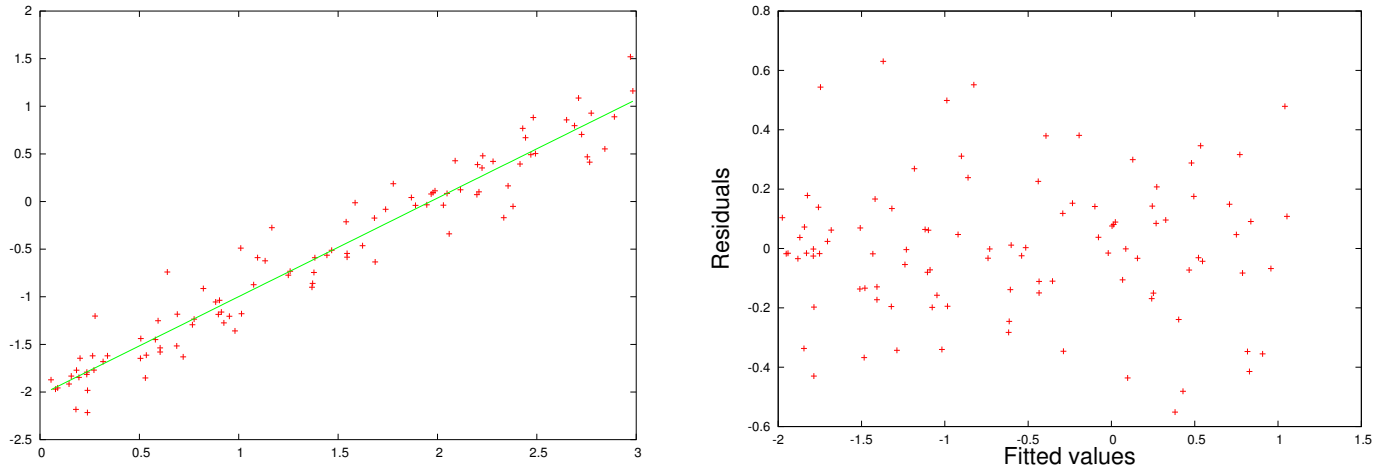
Left: Scatterplot of the transformed data, with the regression line drawn in green. Right: Scatterplot of residuals on fitted values.

- Logarithmically transforming both variables (a "log/log" plot) can reduce both heteroscedasticity and skew:



Left: Scatterplot of the raw data, with the regression line drawn in green. Right: Scatterplot of residuals on fitted values.

after the transform...



Left: Scatterplot of the transformed data, with the regression line drawn in green. Right: Scatterplot of residuals on fitted values.