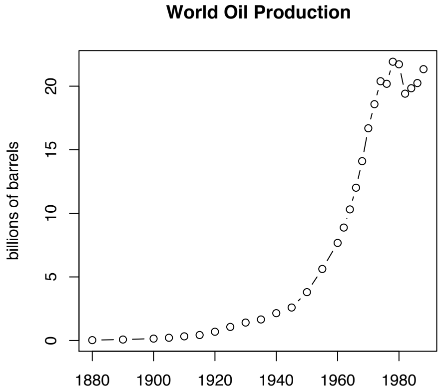
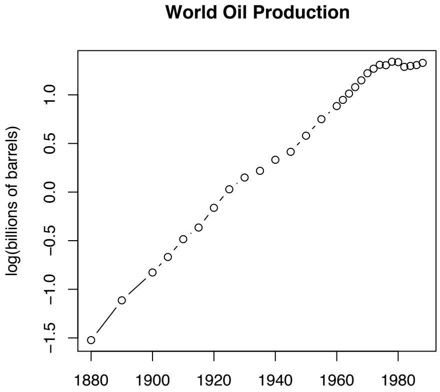
Topic 4
Correlation and Regression
Transformed Variables

Outline

Worldwide Oil Production

Example. The modern history of petroleum began in the 19th century with the refining of kerosene from crude oil. The industry grew through the 1800s, driven by the demand for kerosene and oil lamps. The internal combustion engine introduced in the early part of the 20th century provided a demand that continues to sustained the industry.





Worldwide Oil Production

The relationship between the explanatory variable and response variable can be made to be linear with a simple transformation, the common logarithm. The explanatory variable remains year. With these variables, we can use a regression line to help describe the data.

$$\log y_i = \alpha + \beta x_i + \epsilon_i.$$

```
> summary(lm(logbarrel~year))
Call:
lm(formula = logbarrel ~ year)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.159e+01  1.301e+00  -39.64   <2e-16 ***
year         2.675e-02  6.678e-04   40.05   <2e-16 ***
---
Multiple R-Squared: 0.9834, Adjusted R-squared: 0.9828
```

Worldwide Oil Production

The R output gives $r^2 = 0.9828$. Thus, the correlation, $r = 0.9914$, is very nearly one and the data lies very close to the regression line

$$\widehat{\log(barrel)} = -51.59 + 0.02675 \cdot year.$$

If we rewrite the equation in exponential form, we obtain

$$\widehat{barrel} = A10^{0.02675 \cdot year} = Ae^{\hat{k} \cdot year}.$$

Thus, $\hat{k}$ gives the instantaneous growth rate that best fits the data. This is obtained by converting from a common logarithm to a natural logarithm.

$$\hat{k} = 0.02675 \ln 10 = 0.0616$$

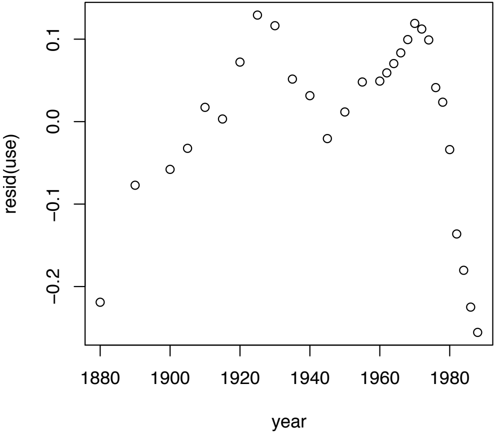
. Thus, the use of oil sustained annual growth of 6% over a span of a hundred years.

Worldwide Oil Production

For the residual plot.

```
> use<-lm(logbarrel~year)
> plot(year,resid(use))
```

Exercise. Give some aspects of world history that could explain the structure in the residual plots.



Michaelis-Menten Kinetics

Consider the chemical reaction in which an enzyme catalyzes the action on a substrate.

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \overset{k_2}{\rightarrow} E + P$$

- E_0 is the total amount of enzyme.
- E is the free enzyme.
- S is the substrate.
- ES is the substrate-bound enzyme.
- P is the product.
- $V = d[P]/dt$ is the production rate.

The symbol $[\,\cdot\,]$ indicates concentration.

Michaelis-Menten Kinetics

The enzyme, E_0, is either free or bound to the substrate. Its total concentration

$$[E_0] = [E] + [ES], \quad \text{and, thus} \quad [E] = [E_0] - [ES].$$

Our goal is to relate the production rate V to the substrate concentration $[S]$.

The law of mass action turns the chemical reactions into differential equations. In particular, the reactions, focusing on the substrate-bound enzyme and the product, gives the equations

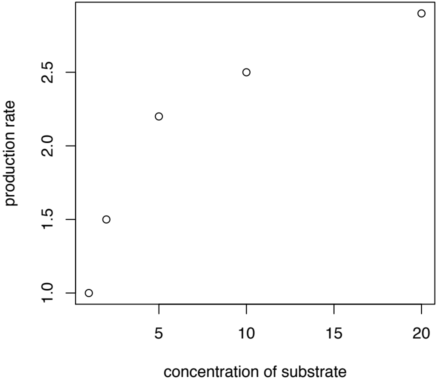
$$\frac{d[ES]}{dt} = k_1[E][S] - (k_{-1} + k_2)[ES] \quad \text{and} \quad V = \frac{d[P]}{dt} = k_2[ES].$$

We can meet our goal by finding an equation for $V = k_2[ES]$ that depends only on $[S]$.

Michaelis-Menten Kinetics

Let's look at data,

$[S]$ (mM)	1	2	5	10	20
V (nmol/sec)	1.0	1.5	2.2	2.5	2.9



Michaelis-Menten Kinetics

To use linear regression, we will have to transform the data. The Michaelis-Menten transformation applies to situations in which the concentration of the substrate-bound enzyme changes much more slowly than those of the product and substrate.

$$0 \approx \frac{d[ES]}{dt} = k_1[E][S] - (k_{-1} + k_2)[ES]$$

$$[ES] \approx \frac{k_1[E][S]}{k_{-1} + k_2} = \frac{[E][S]}{K_m} = \frac{([E_0] - [ES])[S]}{K_m},$$

The ratio $K_m = (k_{-1} + k_2)/k_1$ of the rate of loss of the substrate-bound enzyme to its production is called the Michaelis constant.

Michaelis-Menten Kinetics

We now solve for $[ES]$.

$$[ES] \approx \frac{([E_0] - [ES])[S]}{K_m} \quad \text{and} \quad [ES] \approx [E_0]\frac{[S]}{K_m + [S]}$$

Under this approximation, the production rate of the product is:

$$V = \frac{d[P]}{dt} = k_2[ES] = k_2[E_0]\frac{[S]}{K_m + [S]} = V_{\max}\frac{[S]}{K_m + [S]}$$

Here, $V_{\max} = k_2[E_0]$ is the maximum production rate. To perform linear regression, we need to have a function of V be linearly related to a function of $[S]$. This is achieved via taking the reciprocal of both sides of this equation.

$$\frac{1}{V} = \frac{[S] + K_m}{V_{\max}[S]} = \frac{1}{V_{\max}} + \frac{K_m}{V_{\max}}\frac{1}{[S]}$$

Lineweaver-Burke double reciprocal plot

Thus, we have a *linear relationship* between

$$\frac{1}{V}, \text{ the response variable, and } \frac{1}{[S]}, \text{ the explanatory variable}$$

subject to experimental error. This linear relationship, called the Lineweaver-Burke double reciprocal plot, provides a useful method for analysis of the Michaelis-Menten equation. For the data,

```
> S<-c(1,2,5,10,20);V<-c(1.0,1.5,2.2,2.5,2.9)
> Sinv<-1/S; Vinv<-1/V
> lm(Vinv~Sinv)
Call:
lm(formula = Vinv ~ Sinv)
Coefficients:
(Intercept)         Sinv
     0.3211       0.6813
```

Exercises

Exercise. Use the regression line $\widehat{\log(barrel)} = -51.59 + 0.02675 \cdot year$ to predict use of oil today. Compare it to actual use.

Exercise. Give the Lineweaver-Burke double reciprocal plot for

$$\widehat{\frac{1}{V}} = \frac{1}{V_{\max}} + \frac{K_m}{V_{\max}} \frac{1}{[S]} = 0.3211 + 0.6813 \cdot \frac{1}{[S]}.$$

Determine K_m and $V_{\max}$. Give the biochemical meaning for the intercepts with the horizontal and vertical axes.

NB The horizontal axis intercept has a negative value. The portion of the line in the third quadrant is not biochemically reasonable and is generally shown by a dotted line.