Sections 3.9 and 6.8: Transformations

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

Linear Regression Assumptions

$$Y = X\beta + \epsilon, \qquad \epsilon \sim N_n(0, \sigma^2 I)$$

Assumptions

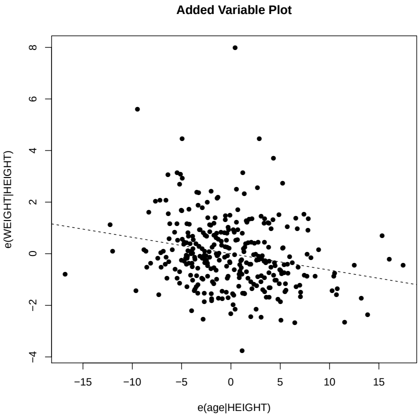
- **L**inear relationship
- **I**ndependent observations
- **N**ormally distributed residuals
- **E**qual variance across X's
- Plus need to check for influential points and outliers: one or a few observations should not dominate the model fit

10.1 Added variable plots (partial regression plot)

- Consider a pool of predictors $x_1, \ldots, x_k$.
- Regress Y_i vs. all predictors *except* x_j, call the residuals $e_i(Y|\mathbf{x}_{-j})$.
- Regress x_j vs. all predictors *except* x_j, call the residuals $e_i(x_j|\mathbf{x}_{-j})$.
- The *added variable plot* for x_j is $e_i(Y|\mathbf{x}_{-j})$ vs. $e_i(x_j|\mathbf{x}_{-j})$.
- The least squares estimate b_j obtained from fitting a line (through the origin) to the plot *is the same* as one would get from fitting the full model $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots \beta_k x_{ik} + \epsilon_i$ (Proof this in homework).
- Gives an *idea* of the relationship between Y and X_j adjusting for all other variables in the model.

Added variable plots: IGROWUP Data

$$e(WEIGHT|HEIGHT) = \beta_0 + \beta_1 \times e(age|HEIGHT)$$



Transformations of variables (Section 3.9 & p. 236)

- Some violations of our model assumptions may be fixed by transforming one or more predictors $x_1, \ldots, x_k$ or Y.
- If the *only* problem is a nonlinear relationship between Y and the predictors, i.e. constant variance seems okay, a transformation of one or more of the $x_1, \ldots, x_k$ is preferred.
- If non-constant variance appears in one or more plots of Y versus the predictors, a transformation in Y can help...or make it worse!
- *Data analysis is an art.* The best way to learn how to analyze data is to analyze data.
- A nonlinear relationship *could* manifest itself the scatterplot matrix of Y_i versus x_{ij} for $j = 1, \ldots, k$, or the residuals e_i versus x_{ij} from an initial fit.
- The chosen transformation should roughly mimic the relationship seen in the plot.

Transformations for $x_{i1}, \ldots, x_{ik}$

	Prototype Regression Pattern	Transformations of X
(a)		$X' = \log_{10} X$ $\quad X' = \sqrt{X}$
(b)		$X' = X^2$ $\quad X' = \exp(X)$
(c)		$X' = 1/X$ $\quad X' = \exp(-X)$

Transforming the response

If there is evidence of nonconstant error variance, a transformation of Y can often fix things. Examples include:

- $Y^* = \log(Y)$
- $Y^* = \sqrt{Y}$
- $Y^* = 1/Y$

All of these are included in the Box-Cox family of transformations.

For some data, a transformation in Y may be followed by one or more transformations in the $x_{i1}, \ldots, x_{ik}$.

Linear Regression Assumptions: Transformation of Variables

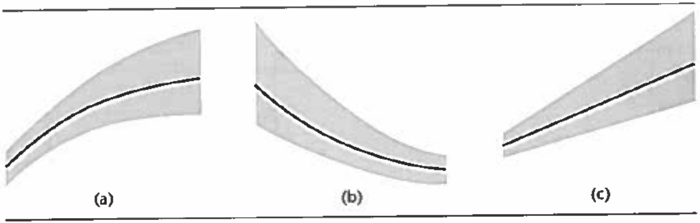
$$Y = X\beta + \epsilon, \qquad \epsilon \sim N_n(0, \sigma^2 I)$$

Assumptions

- **L**inear relationship
- **I**ndependent observations
- **N**ormally distributed residuals
- **E**qual variance across X's
- Plus need to check for influential points and outliers: one or a few observations should not dominate the model fit

Transformations for Y

Prototype Regression Pattern



Transformations on Y

$Y' = \sqrt{Y}$

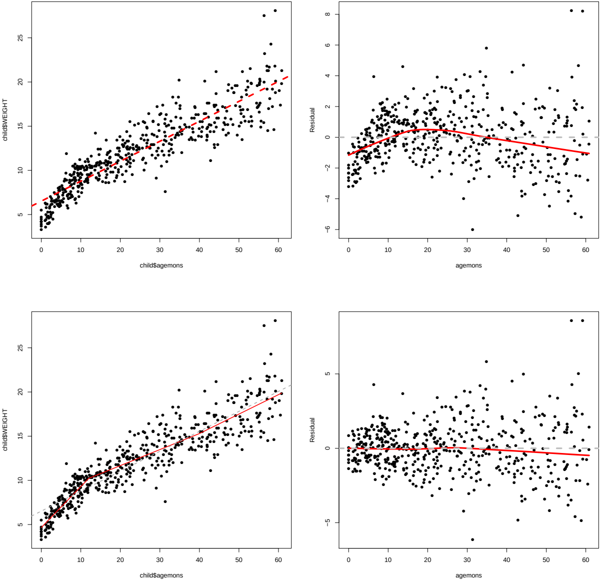
$Y' = \log_{10} Y$

$Y' = 1/Y$

Non-constant variance

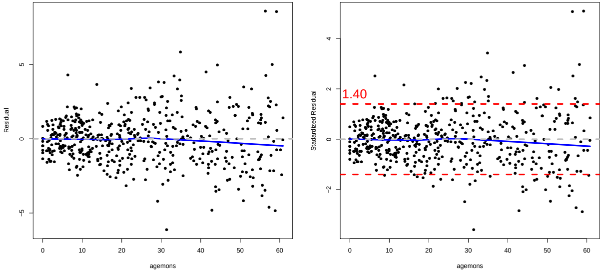
- **Breusch-Pagan test** (pp. 118–119): tests whether the log error variance increases or decreases linearly with the predictor(s). Where $Y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma_i^2)$, set $\log \sigma_i^2 = \alpha_0 + \alpha_1 x_{i1} + \cdots \alpha_k x_{ik}$ and test $H_0 : \alpha_1 = \cdots = \alpha_k = 0$, i.e. $\log \sigma_i^2 = \alpha_0$. Requires large samples & assumes normal errors.
- **Brown-Forsythe test** (pp. 116–117): Robust to non-normal errors. Requires user to break data into groups and test for constancy error variance across groups (not natural for continuous data).
- Graphical methods have advantage of checking for *general violations*, not just violation of a specific type.

Residual plots: IGROWUP data



Standardized Residuals

$$\text{standardized } Residual_i = \frac{Residual_i}{\text{standard deviation of } Residual_i}$$



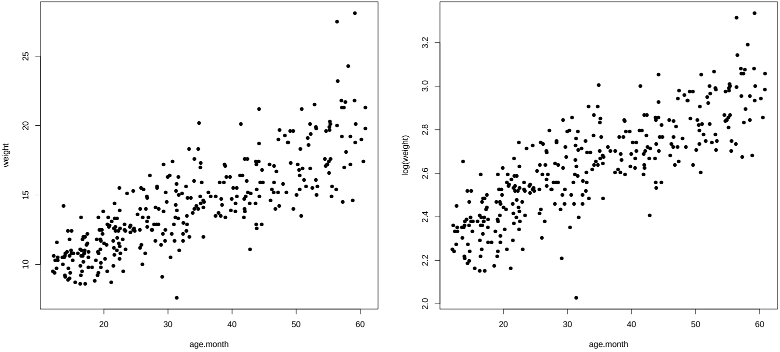

```
>stand.resid<-rstandard(fit1sp)
```

```
>library(lmtest)
#### Breusch-Pagan test
>bptest(WEIGHT ~ agemons + age12 + age30, data=child)

studentized Breusch-Pagan test

data:  WEIGHT ~ agemons + age12 + age30
BP = 49.639, df = 3, p-value = 9.537e-11
```

Transforming the response: IGROWUP data



Box-Cox transformations

Box-Cox transformations are of the type

$$Y^* = \begin{cases} Y^\lambda & \lambda \neq 0, \\ log(Y) & \lambda = 0, \end{cases}$$

where λ is estimated from the data, typically $-3 \leq \lambda \leq 3$. These include

$$\begin{array}{lll} \lambda = 2 & Y^* = Y^2 & \\ \lambda = 1 & Y^* = Y & \text{No transformation!} \\ \lambda = 0 & Y^* = \log(Y) & \text{By definition} \\ \lambda = -1 & Y^* = 1/Y & \text{Reciprocal} \\ \lambda = -2 & Y^* = 1/Y^2 & \end{array}$$

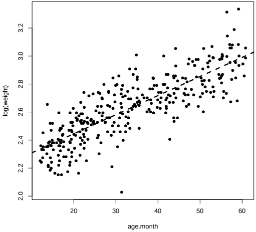
R uses `boxcox()` in the **MASS** package.

Interpretation changes with transformed data

Note: When working with transformed data, predictions and interpretations of regression coefficients are all in terms of the *transformed variables*.

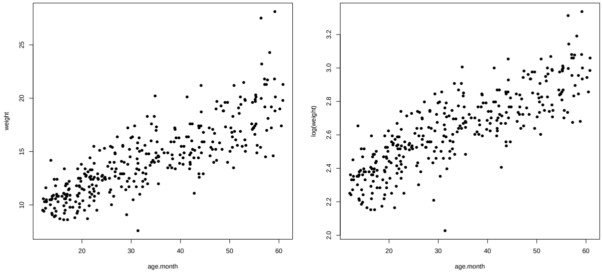
To state the conclusions in terms of the original variables, we need to do a reverse transformation...carefully.

Example: IGROWUP Data



$$E[ln(Weight)] = \beta_0 + \beta_1 \times agemons$$

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.1722076  0.0187270  115.99   <2e-16 ***
todd$agemons 0.0136300  0.0005147   26.48   <2e-16 ***
```

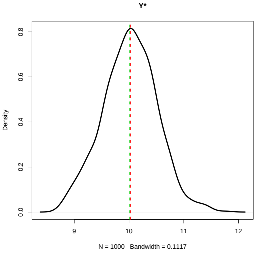


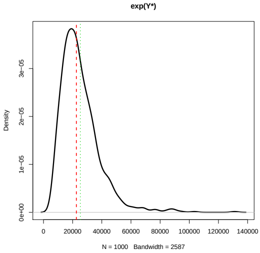
$$E[ln(Weight)] = \beta_0 + \beta_1 \times agemons$$

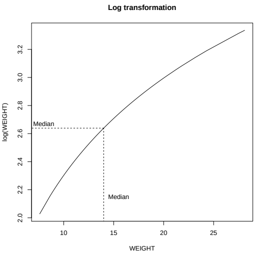
```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.1722076  0.0187270  115.99   <2e-16 ***
todd$agemons 0.0136300  0.0005147   26.48   <2e-16 ***
> pred<-predict(fit3, new=data.frame(agemons=25), interval="confidence")
> pred
       fit      lwr      upr
1 2.512957 2.496285 2.529629
```

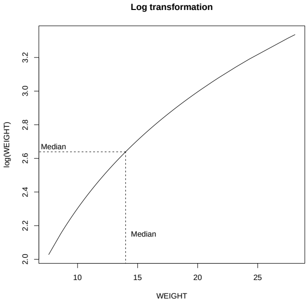
- What is the "expected" weight of a 25 month-old child? $e^{(2.172+25*0.0136)} \approx 12.34$ (95% CI: $e^{2.496}$, $e^{2.530}$)=(12.14, 12.55). The median weight of a 25 month-old child is 12.34 (95% CI: 12.14, 12.55).

$$\widehat{Y^*} \sim N(X\beta, \sigma^2)$$
$$\widehat{\log(Y)} \sim N(X\beta, \sigma^2)$$









$$E[log(Weight)] \sim N(X\beta, \sigma^2)$$

- What is the expected change in weight for each 1-month increase in age?
  The median weight changes by a factor of e^{β_1} for each 1-month increase in age.

$$\begin{aligned} \text{Median}_{Y|X=x} &= e^{\beta_0+\beta_1 x} \\ \text{Median}_{Y|X=(x+1)} &= e^{\beta_0+\beta_1 (x+1)} \\ \frac{\text{Median}_{Y|X=(x+1)}}{\text{Median}_{Y|X=x)}} &= e^{\beta_1} \end{aligned}$$

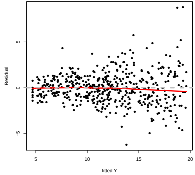
The Delta Method

- What is the standard error of $e^{\hat{\beta}_1}$ (the median)?

$$
\begin{aligned}
g(x) &= g(\theta) + g'(\theta)(x-\theta) + ...\text{(Taylor expansion)} \\
E[g(x)] &\approx g(\theta) + g'(\theta)E[(x-\theta)] = g(\theta) \qquad [E(x) = \theta] \\
var[g(x)] &\approx E[g(x) - g(\theta)]^2 = (g'(\theta))^2 E[(x-\theta)^2] \\
var(g(x)) &= (\frac{\partial g(x)}{\partial x})^2 var(x). \qquad \text{Let } x = \beta_1 \\
var(e^{\hat{\beta}_1}) &= (e^{\hat{\beta}_1})^2 \times var(\hat{\beta}_1) \\
\text{std error}(e^{\hat{\beta}_1}) &= \sqrt{(e^{\hat{\beta}_1})^2 \times var(\hat{\beta}_1)}
\end{aligned}
$$

Non-Constant Variance: Iteratively Re-weighted Least Square (IRLS)

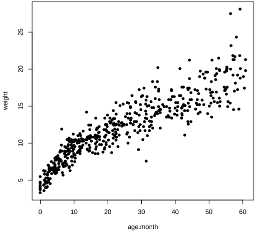
If variances are of scientific interest, the following model can be considered:



$$
\begin{aligned}
\mathbf{Y} &= \mathbf{X}\beta + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N_n(0, \Sigma), \\
\Sigma &= \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ \dots & & & \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}. \\
\hat{\beta} &= (X'WX)^{-1}X'WY.
\end{aligned}
$$

More details in Lecture 12.

Non-Constant Variance: Robust Estimator



$$Y_i = X_i\beta + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma_i^2)$$

- $\hat{\beta}_i$ unbiased: $\mathrm{E}(\hat{\beta}) = \beta$
- But se($\hat{\beta}$) would be wrong $\rightarrow$ inefficient
- 95% CI, t-test, p-value would be wrong also
- Use bootstrap, robust or empirical approaches for estimating se($\hat{\beta}$).

Robust Estimator: IGROWUP Data

```
>fit1sp<-lm(WEIGHT ~ agemons + age12 + age30, data=child)
> summary(fit1sp)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.699571   0.262318  17.916  < 2e-16 ***
agemons      0.454219   0.030906  14.697  < 2e-16 ***
age12       -0.266450   0.042567  -6.260 8.42e-10 ***
age30        0.009773   0.024950   0.392    0.695
> summary(fit1sp, robust=T)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.699571   0.144109  32.611  < 2e-16 ***
agemons      0.454219   0.018968  23.947  < 2e-16 ***
age12       -0.266450   0.031364  -8.495 2.37e-16 ***
age30        0.009773   0.030328   0.322    0.747
> bootvar<-boots(child$agemons, child$WEIGHT)
> bootvar
         Estimate  Std.Error
[1,]  4.713311477 0.15161826
[2,]  0.452656368 0.02008021
[3,] -0.264007819 0.03295003
[4,]  0.007629656 0.03082049
```