Multiple Regression

- Numeric Response variable (y)
- p Numeric predictor variables ($p < n$)
- Model:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

- Partial Regression Coefficients: $\beta_i \equiv$ effect (on the mean response) of increasing the i^{th} predictor variable by 1 unit, **holding all other predictors constant**
- Model Assumptions (Involving Error terms ε)
  - Normally distributed with mean 0
  - Constant Variance σ^2
  - Independent (Problematic when data are series in time/space)

Example - Effect of Birth weight on Body Size in Early Adolescence

- Response: Height at Early adolescence (n =250 cases)
- Predictors (p=6 explanatory variables)
  - Adolescent Age (x_1, in years -- 11-14)
  - Tanner stage (x_2, units not given)
  - Gender (x_3=1 if male, 0 if female)
  - Gestational age (x_4, in weeks at birth)
  - Birth length (x_5, units not given)
  - Birthweight Group (x_6=1,...,6 <1500*g* (1), 1500-1999*g*(2), 2000-2499*g*(3), 2500-2999*g*(4), 3000-3499*g*(5), >3500*g*(6))

Source: Falkner, et al (2004)

Least Squares Estimation

- Population Model for mean response:

$$E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- Least Squares Fitted (predicted) equation, minimizing *SSE*:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \qquad SSE = \sum \left(Y - \hat{Y} \right)^2$$

- All statistical software packages/spreadsheets can compute least squares estimates and their standard errors

Analysis of Variance

- Direct extension to ANOVA based on simple linear regression
- Only adjustments are to degrees of freedom:
  - $DF_R = p$ $\quad DF_E = n\text{-}p^*$ $\quad (p^*=p+1=\#\text{Parameters})$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F*
Model	***SSR***	***p***	***MSR* = *SSR*/*p***	***F* = *MSR*/*MSE***
Error	***SSE***	***n-p****	***MSE* = *SSE*/(*n-p**)**	
Total	***TSS***	***n*-1**		

$$R^2 = \frac{TSS - SSE}{TSS} = \frac{SSR}{TSS}$$

Testing for the Overall Model - *F*-test

- Tests whether **any** of the explanatory variables are associated with the response
- H_0: $\beta_1 = \cdots = \beta_p = 0$ (None of the x^s associated with y)
- H_A: Not all $\beta_i = 0$

$$T.S.: F_{obs} = \frac{MSR}{MSE} = \frac{R^2 / p}{(1 - R^2)/(n - p^*)}$$

$$R.R.: F_{obs} \geq F_{\alpha, p, n-p^*}$$

$$P - val: P(F \geq F_{obs})$$

Example - Effect of Birth weight on Body Size in Early Adolescence

- Authors did not print ANOVA, but did provide following:
  - n=250 p=6 R^2=0.26
- H_0: $\beta_1 = \cdots = \beta_6 = 0$ H_A: Not all $\beta_i = 0$

$$T.S.: F_{obs} = \frac{MSR}{MSE} = \frac{R^2/p}{(1-R^2)/(n-p^*)} =$$

$$= \frac{0.26/6}{(1-0.26)/(250-7)} = \frac{.0433}{.0030} = 14.2$$

$$R.R.: F_{obs} \geq F_{\alpha,6,243} = 2.13$$

$$P-val: P(F \geq 14.2)$$

Testing Individual Partial Coefficients - *t*-tests

- Wish to determine whether the response is associated with a single explanatory variable, after controlling for the others

- H_0: $\beta_i = 0$ H_A: $\beta_i \neq 0$ (2-sided alternative)

$$T.S.: t_{obs} = \frac{\hat{\beta}_i}{s_{\hat{b}_i}}$$

$$R.R.: |t_{obs}| \geq t_{\alpha/2, n-p^*}$$

$$P-val: 2P(t \geq |t_{obs}|)$$

Example - Effect of Birth weight on Body Size in Early Adolescence

Variable	b	SE_b	$t=b/SE_b$	P-val (z)
Adolescent Age	2.86	0.99	2.89	.0038
Tanner Stage	3.41	0.89	3.83	<.001
Male	0.08	1.26	0.06	.9522
Gestational Age	-0.11	0.21	-0.52	.6030
Birth Length	0.44	0.19	2.32	.0204
Birth Wt Grp	-0.78	0.64	-1.22	.2224

Controlling for all other predictors, adolescent age, Tanner stage, and Birth length are associated with adolescent height measurement

Comparing Regression Models

- Conflicting Goals: Explaining variation in *Y* while keeping model as simple as possible (parsimony)
- We can test whether a subset of *p-g* predictors (including possibly cross-product terms) can be dropped from a model that contains the remaining *g* predictors. H_0: $\beta_{g+1}=\ldots=\beta_p=0$
  - Complete Model: Contains all *p* predictors
  - Reduced Model: Eliminates the predictors from H_0
  - Fit both models, obtaining sums of squares for each (or R^2 from each):
    - Complete: SSR_c , SSE_c (R_c^2)
    - Reduced: SSR_r , SSE_r (R_r^2)

Comparing Regression Models

- H_0: $\beta_{g+1} = \ldots = \beta_p = 0$ (After removing the effects of $X_1, \ldots, X_g$, none of other predictors are associated with Y)
- H_a: H_0 is false

$$\mathrm{TS}: F_{obs} = \frac{(SSR_c - SSR_r)/(p-g)}{SSE_c/[n-p^*]} = \frac{\left(R_c^2 - R_r^2\right)/(p-g)}{\left(1-R_c^2\right)/[n-p^*]}$$

$$RR: F_{obs} \geq F_{\alpha, p-g, (n-p^*)}$$

$$P = P(F \geq F_{obs})$$

P-value based on *F*-distribution with *p*-*g* and *n*-p* d.f.

Models with Dummy Variables

- Some models have both numeric and categorical explanatory variables (Recall **gender** in example)
- If a categorical variable has m levels, need to create m-1 dummy variables that take on the values 1 if the level of interest is present, 0 otherwise.
- The baseline level of the categorical variable is the one for which all m-1 dummy variables are set to 0
- The regression coefficient corresponding to a dummy variable is the difference between the mean for that level and the mean for baseline group, controlling for all numeric predictors

Example - Deep Cervical Infections

- Subjects - Patients with deep neck infections
- Response (Y) - Length of Stay in hospital
- Predictors: (One numeric, 11 Dichotomous)
  - Age (x_1)
  - Gender (x_2=1 if female, 0 if male)
  - Fever (x_3=1 if Body Temp > 38C, 0 if not)
  - Neck swelling (x_4=1 if Present, 0 if absent)
  - Neck Pain (x_5=1 if Present, 0 if absent)
  - Trismus (x_6=1 if Present, 0 if absent)
  - Underlying Disease (x_7=1 if Present, 0 if absent)
  - Respiration Difficulty (x_8=1 if Present, 0 if absent)
  - Complication (x_9=1 if Present, 0 if absent)
  - WBC > 15000/mm^3 (x_{10}=1 if Present, 0 if absent)
  - CRP > 100μg/ml (x_{11}=1 if Present, 0 if absent)

Source: Wang, et al (2003)

Example - Weather and Spinal Patients

- Subjects - Visitors to National Spinal Network in 23 cities Completing SF-36 Form
- Response - Physical Function subscale (1 of 10 reported)
- Predictors:
  - Patient's age (x_1)
  - Gender (x_2=1 if female, 0 if male)
  - High temperature on day of visit (x_3)
  - Low temperature on day of visit (x_4)
  - Dew point (x_5)
  - Wet bulb (x_6)
  - Total precipitation (x_7)
  - Barometric Pressure (x_7)
  - Length of sunlight (x_8)
  - Moon Phase (new, wax crescent, 1st Qtr, wax gibbous, full moon, wan gibbous, last Qtr, wan crescent, presumably had 8-1=7 dummy variables)

Source: Glaser, et al (2004)

Modeling Interactions

- Statistical Interaction: When the effect of one predictor (on the response) depends on the level of other predictors.
- Can be modeled (and thus tested) with cross-product terms (case of 2 predictors):
  - $E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
  - $X_2 = 0 \Rightarrow E(Y) = \alpha + \beta_1 X_1$
  - $X_2 = 10 \Rightarrow E(Y) = \alpha + \beta_1 X_1 + 10\beta_2 + 10\beta_3 X_1$

    $= (\alpha + 10\beta_2) + (\beta_1 + 10\beta_3) X_1$
- The effect of increasing X_1 by 1 on $E(Y)$ depends on level of X_2, unless $\beta_3 = 0$ (t-test)

Regression Model Building

- Setting: Possibly a large set of predictor variables (including interactions).
- Goal: Fit a parsimonious model that explains variation in *Y* with a small set of predictors
- Automated Procedures and all possible regressions:
  - Backward Elimination (Top down approach)
  - Forward Selection (Bottom up approach)
  - Stepwise Regression (Combines Forward/Backward)
  - C_p, *AIC*, *BIC*- Summarizes each possible model, where "best" model can be selected based on each statistic

Backward Elimination

- Select a significance level to stay in the model (e.g. SLS=0.20, generally .05 is too low, causing too many variables to be removed)
- Fit the full model with all possible predictors
- Consider the predictor with lowest *t*-statistic (highest *P*-value).
  - If $P >$ SLS, remove the predictor and fit model without this variable (must re-fit model here because partial regression coefficients change)
  - If $P \leq$ SLS, stop and keep current model
- Continue until all predictors have *P*-values below SLS

Forward Selection

- Choose a significance level to enter the model (e.g. SLE=0.20, generally .05 is too low, causing too few variables to be entered)
- Fit all simple regression models.
- Consider the predictor with the highest *t*-statistic (lowest *P*-value)
  - If $P \leq$ SLE, keep this variable and fit all two variable models that include this predictor
  - If $P >$ SLE, stop and keep previous model
- Continue until no new predictors have $P \leq$ SLE

Stepwise Regression

- Select SLS and SLE (SLE<SLS)
- Starts like Forward Selection (Bottom up process)
- New variables must have $P \leq$ SLE to enter
- Re-tests all "old variables" that have already been entered, must have $P \leq$ SLS to stay in model
- Continues until no new variables can be entered and no old variables need to be removed

All Possible Regressions – C_p and *PRESS*

- Fit every possible model. If K potential predictor variables, there are 2^K-1 models.
- Label the Mean Square Error for the model containing all K predictors as MSE_K
  - C_p: For each model, compute SSE and C_p where p^* is the number of parameters (including intercept) in model
  - *PRESS*: Fitted values for each observation when that observation is not used in model fit.

$$C_p = \frac{SSE}{MSE_K} - (n - 2p^*) \qquad PRESS = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{i(i)} \right)^2$$

- C_p: Select the model with the fewest predictors that has $C_p \approx p^*$
- *PRESS*: Choose model with minimum value for *PRESS*

All Possible Regressions – *AIC*, *BIC*

- Fits every possible model. If K potential predictor variables, there are 2^K-1 models.
- For each model, compute *SSE* and *AIC* and *BIC* where p^* is the number of parameters (including intercept) in model

$$AIC = n\ln\left(\frac{SSE}{n}\right) + 2p^* \qquad BIC = n\ln\left(\frac{SSE}{n}\right) + \left[\ln\left(n\right)\right]p^*$$

• Select the model that minimizes the criterion. BIC puts a higher penalty (for most sample sizes) and tends to choose "smaller" models. Note that various computing packages use different variations, but goal is to choose model that minimizes measure.

Regression Diagnostics

- Model Assumptions:
  - Regression function correctly specified (e.g. linear)
  - Conditional distribution of *Y* is normal distribution
  - Conditional distribution of *Y* has constant standard deviation
  - Observations on *Y* are statistically independent
- Residual plots can be used to check the assumptions
  - Histogram (stem-and-leaf plot) should be mound-shaped (normal)
  - Plot of Residuals versus each predictor should be random cloud
    - U-shaped (or inverted U) $\Rightarrow$ Nonlinear relation
    - Funnel shaped $\Rightarrow$ Non-constant Variance
  - Plot of Residuals versus Time order (Time series data) should be random cloud. If pattern appears, not independent.

Linearity of Regression (SLR)

F-Test for Lack-of-Fit (n_j observations at c distinct levels of "X")

$H_0: E(Y_i) = \beta_0 + \beta_1 X_i \quad H_A: E(Y_i) = \mu_i \neq \beta_0 + \beta_1 X_i$

Compute fitted value Y_j and sample mean $\overline{Y}_j$ for each distinct X level

Lack-of-Fit: $SS(LF) = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left(\overline{Y}_j - Y_j \right)^2 \quad df_{LF} = c - 2$

Pure Error: $SS(PE) = \sum_{j=1}^{c} \sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}_j \right)^2 \quad df_{PE} = n - c$

Test Statistic: $F_{LOF} = \dfrac{\left(SS(LF) / (c-2) \right)}{\left(SS(PE) / (n-c) \right)} = \dfrac{MS(LF)}{MS(PE)} \quad \overset{H_0}{\sim} \quad F_{c-2,n-c}$

Reject H$_0$ if $F_{LOF} \geq F(1-\alpha; c-2, n-c)$

Non-Normal Errors

- Box-Plot of Residuals – Can confirm symmetry and lack of outliers
- Check Proportion that lie within 1 standard deviation from 0, 2 SD, etc, where SD=sqrt(MSE)
- Normal probability plot of residual versus expected values under normality – should fall approximately on a straight line (Only works well with moderate to large samples)   **qqnorm(e); qqline(e)**   in R

Expected value of Residuals under Normality:

1) Rank residuals from smallest (large/negative) to highest (large/positive) Rank = k

2) Compute the percentile using $p = \frac{k - 0.375}{n + 0.25}$ and obtain corresponding z-value: $z(p)$

3) Multiply by $s = \sqrt{MSE}$ expected residual = $\sqrt{MSE}\left[z(p)\right]$

Test for Normality of Residuals

- Correlation Test
  1) Obtain correlation between observed residuals and expected values under normality (see slide 7)
  2) Compare correlation with critical value based on α=0.05 level with: 1.02-1/sqrt(10n)
  3) Reject the null hypothesis of normal errors if the correlation falls below the critical value

- Shapiro-Wilk Test – Performed by most software packages. Related to correlation test, but more complex calculations

Equal (Homogeneous) Variance

Breusch-Pagan (aka Cook-Weisberg) Test:

H_0 : Equal Variance Among Errors $\sigma^2\left\{\varepsilon_i\right\} = \sigma^2 \ \forall \ i$

H_A : Unequal Variance Among Errors $\sigma_i^2 = \sigma^2 h\left(\gamma_1 X_{i1} + \ldots + \gamma_p X_{ip}\right)$

1) Let $SSE = \sum_{i=1}^{n} e_i^2$ from original regression

2) Fit Regression of e_i^2 on $X_{i1},\ldots X_{ip}$ and obtain $SS\left(\text{Reg}^*\right)$

Test Statistic: $X_{BP}^2 = \dfrac{SS\left(\text{Reg}^*\right)/2}{\left(\sum_{i=1}^{n} e_i^2 \Big/ n\right)^2} \overset{H_0}{\sim} \chi_p^2$

Reject H_0 if $X_{BP}^2 \geq \chi^2\left(1-\alpha; p\right)$ $\quad p$ = # of predictors

Test For Independence - Durbin-Watson Test

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \quad \varepsilon_t = \rho\varepsilon_{t-1} + u_t \quad u_t \sim NID\left(0, \sigma^2\right) \quad |\rho| < 1$$

$H_0 : \rho = 0 \quad \Rightarrow \quad$ Errors are uncorrelated over time

$H_A : \rho > 0 \quad \Rightarrow \quad$ Positively correlated

1) Obtain Residuals from Regression

2) Compute Durbin-Watson Statistic (given below)

3) Obtain Critical Values from Durbin-Watson Table (on class website)

If $DW < d_L(1, n)$ Reject H_0

If $DW > d_U(1, n)$ Conclude H_0

Otherwise Inconclusive

Test Statistic: $DW = \dfrac{\sum_{t=2}^{n}\left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{n} e_t^2}$

Note 1: This generalizes to any number of Predictors (p)

Note 2: R will produce a bootstrapped based P-value

Detecting Influential Observations

- **Studentized Residuals** – Residuals divided by their estimated standard errors (like *t*-statistics). Observations with values larger than 3 in absolute value are considered outliers.
- **Leverage Values (Hat Diag)** – Measure of how far an observation is from the others in terms of the levels of the independent variables (not the dependent variable). Observations with values larger than 2p*/n are considered to be potentially highly influential, where p is the number of predictors and n is the sample size.
- **DFFITS** – Measure of how much an observation has effected its fitted value from the regression model. Values larger than 2sqrt(p*/n) in absolute value are considered highly influential. Use standardized DFFITS in SPSS.

Detecting Influential Observations

- **DFBETAS** – Measure of how much an observation has effected the estimate of a regression coefficient (there is one DFBETA for each regression coefficient, including the intercept). Values larger than 2/sqrt(n) in absolute value are considered highly influential.
- **Cook's D** – Measure of aggregate impact of each observation on the group of regression coefficients, as well as the group of fitted values. Values larger than 4/n are considered highly influential.
- **COVRATIO** – Measure of the impact of each observation on the variances (and standard errors) of the regression coefficients and their covariances. Values outside the interval 1 +/- 3p*/n are considered highly influential.

Variance Inflation Factors

- **Variance Inflation Factor (VIF)** – Measure of how highly correlated each **independent variable** is with the other predictors in the model. Used to identify **Multicollinearity**.
- Values larger than 10 for a predictor imply large inflation of standard errors of regression coefficients due to this variable being in model.
- Inflated standard errors lead to small *t*-statistics for partial regression coefficients and wider confidence intervals

Remedial Measures

- Nonlinear Relation – Add polynomials, fit exponential regression function, or transform *Y* and/or *X*
- Non-Constant Variance – Weighted Least Squares, transform *Y* and/or *X*, or fit Generalized Linear Model
- Non-Independence of Errors – Transform *Y* or use Generalized Least Squares
- Non-Normality of Errors – Box-Cox tranformation, or fit Generalized Linear Model
- Omitted Predictors – Include important predictors in a multiple regression model
- Outlying Observations – Robust Estimation

Nonlinearity: Polynomial Regression

- When relation between Y and X is not linear, polynomial models can be fit that approximate the relationship within a particular range of X
- General form of model:

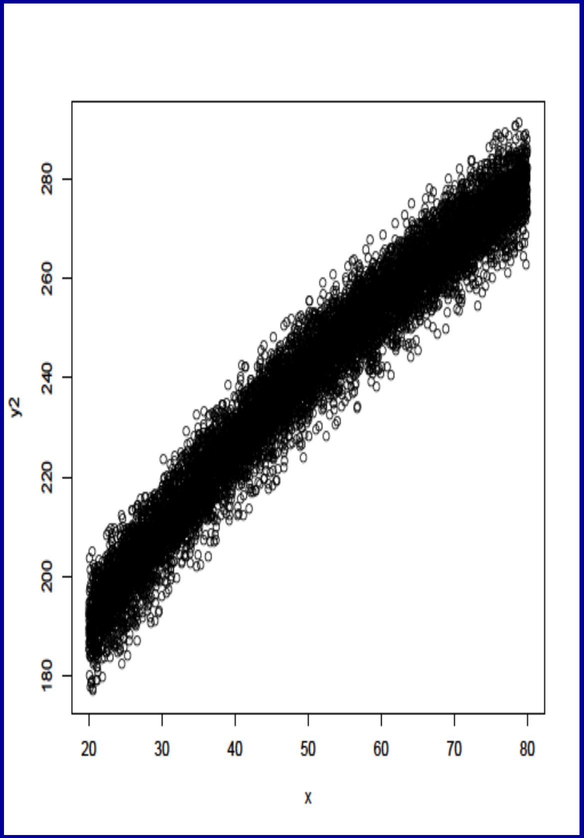
$$\boxed{E(Y) = \alpha + \beta_1 X + \cdots + \beta_p X^p}$$

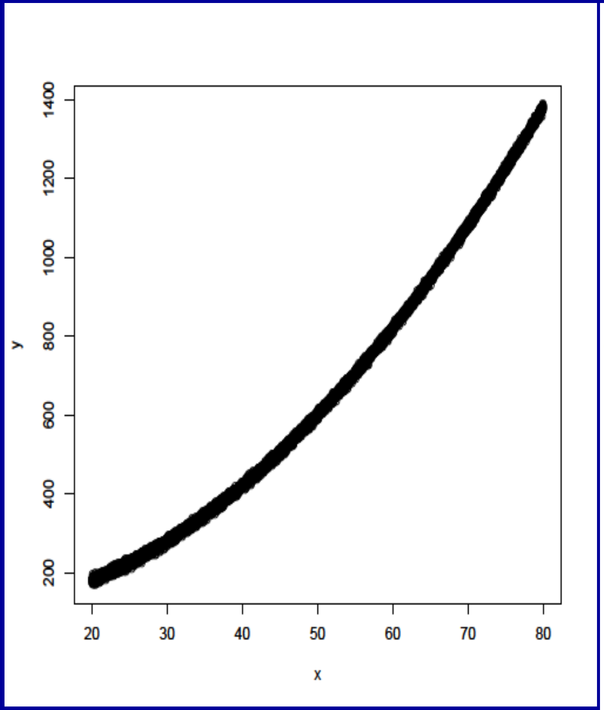
- Second order model (most widely used case, allows one "bend"):

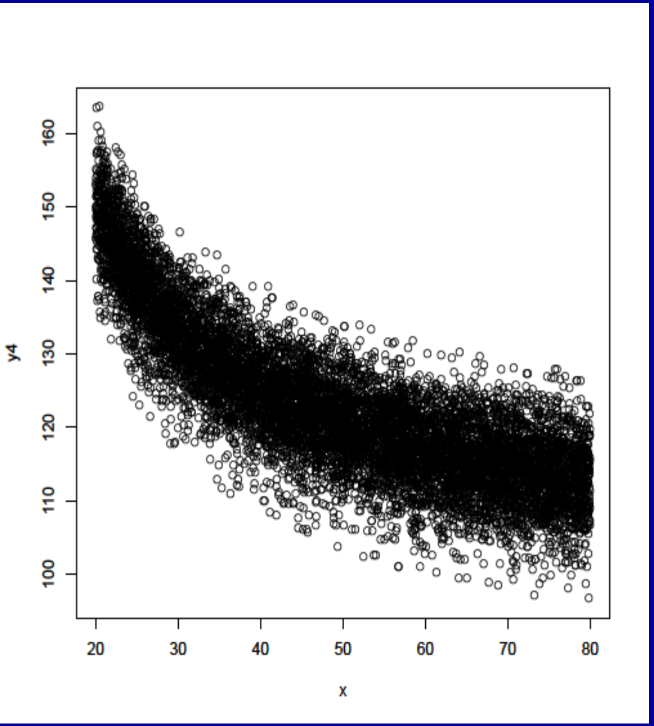
$$\boxed{E(Y) = \alpha + \beta_1 X + \beta_2 X^2}$$

- Must be very careful not to extrapolate beyond observed X levels

Transformations for Non-Linearity – Constant Variance





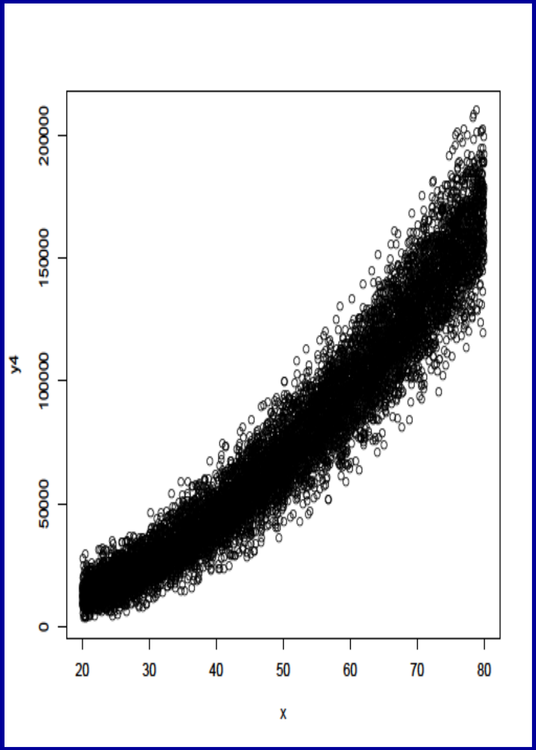


$X' = \sqrt{X}$ $X' = \ln(X)$

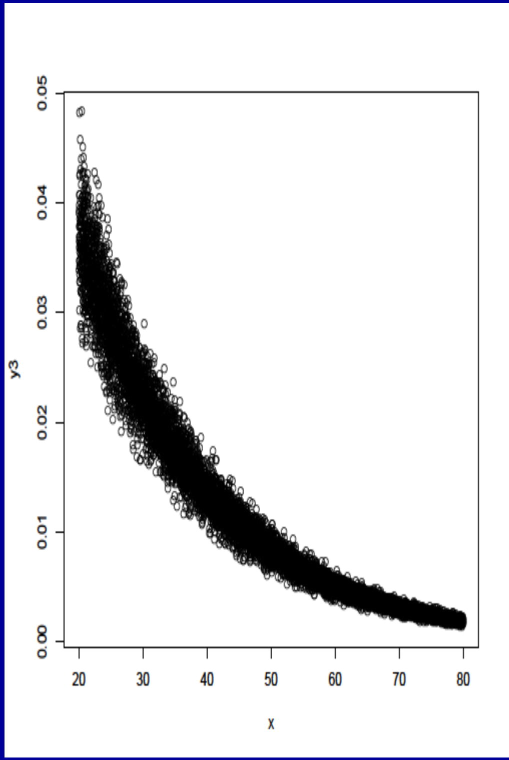
$X' = X^2$ $X' = e^X$

$X' = 1/X$ $X' = e^{-X}$

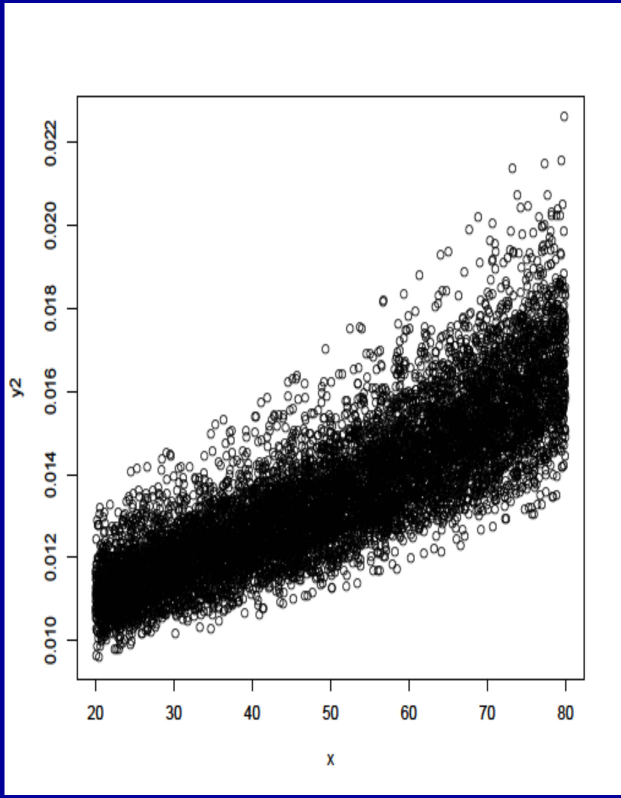
Transformations for Non-Linearity – Non-Constant Variance



Y' = √Y



Y' = ln(Y)



Y' = 1/Y

Box-Cox Transformations

- Automatically selects a transformation from power family with goal of obtaining: normality, linearity, and constant variance (not always successful, but widely used)
- Goal: Fit model: $Y' = \beta_0 + \beta_1 X + \varepsilon$ for various power transformations on Y, and selecting transformation producing minimum SSE (maximum likelihood)
- Procedure: over a range of λ from, say -2 to +2, obtain W_i and regress W_i on X (assuming all $Y_i > 0$, although adding constant won't affect shape or spread of Y distribution)

$$W_i = \begin{cases} K_1\left(Y_i^{\lambda} - 1\right) & \lambda \neq 0 \\ K_2 \ln\left(Y_i\right) & \lambda = 0 \end{cases} \qquad K_2 = \left(\prod_{i=1}^{n} Y_i\right)^{1/n} \qquad K_1 = \frac{1}{\lambda K_2^{\lambda-1}}$$