PLSC 503 – Spring 2025
Dichotomous Covariates and Transformations

February 10, 2025

"Dummy" variables may be:

- ..."naturally" dichotomous, including
  - Structural breaks
  - Proper nouns
- "Factors":

$$\texttt{partyid} = \begin{cases} 0 = \text{Labor} \\ 1 = \text{Liberal} \\ 2 = \text{Conservative} \end{cases}$$

- Ordinal variables...
- Continuous variables...

"Dummy coding":

$$\texttt{female} = \begin{cases} 0 \text{ if male} \\ 1 \text{ if female} \end{cases}$$

vs. "Effect coding":

$$\texttt{female} = \begin{cases} -1 \text{ (or } -0.5\text{) if male} \\ \;\;1 \text{ (or 0.5) if female} \end{cases}$$

TL;DR: (generally) use the former.

For

$$Y_i = \beta_0 + \beta_1 D_i + u_i$$

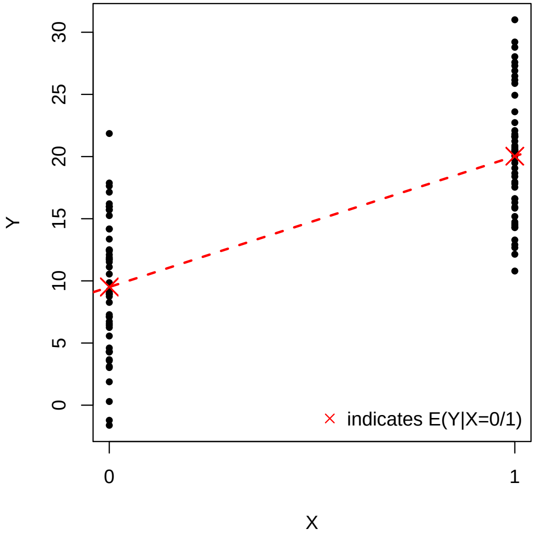
we have

$$\mathrm{E}(Y|D=0) = \beta_0$$

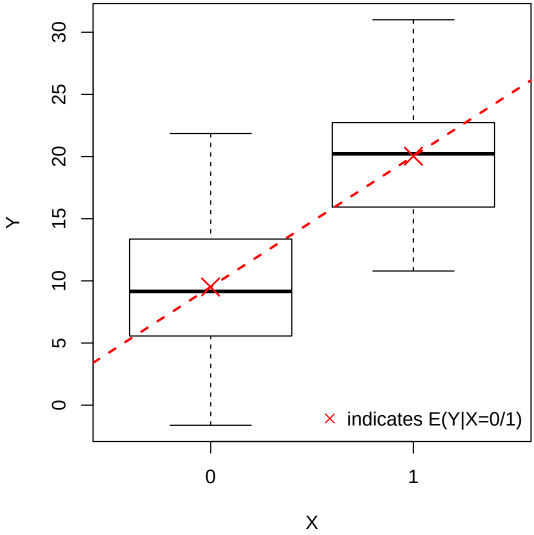
and

$$\mathrm{E}(Y|D=1) = \beta_0 + \beta_1.$$

Dichotomous X, Graphically (No!)



Dichotomous X, Graphically (Yes!)



For:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + ... + \beta_\ell D_{\ell i} + u_i$$

- $\mathrm{E}(Y|D_k = 0) \,\forall\, k \in \ell = \beta_0$,
- Otherwise, $\mathrm{E}(Y) = \beta_0 + \sum_{k=1}^{\ell} \beta_k \,\forall\, k \; s.t. \; D_k = 1$.

Note that where the D_ℓ are mutually exclusive and exhaustive:

- The expected values are the same as the within-group means.
- Identification requires that we either
  - omit a "reference category," or
  - omit β_0.

Many Dummies: Toy Example

```
> labs<-c(rep("A",3),rep("B",3),rep("C",3)) # Three groups
> D<-as.factor(labs)                          # "Factor" variable
> Y<-c(12,16,8,25,27,23,38,42,40)             # Y
> df<-data.frame(D=D,Y=Y)

> df
  D  Y
1 A 12
2 A 16
3 A  8
4 B 25
5 B 27
6 B 23
7 C 38
8 C 42
9 C 40

> # Means of Y by group:
>
> aggregate(df$Y,list(df$D),FUN=mean)
  Group.1  x
1       A 12
2       B 25
3       C 40
```

Many Dummies Example (continued)

```
> # Create binary indicators "by hand":
>
> df$DA<-ifelse(df$D=="A",1,0)
> df$DB<-ifelse(df$D=="B",1,0)
> df$DC<-ifelse(df$D=="C",1,0)
> df
  D  Y DA DB DC
1 A 12  1  0  0
2 A 16  1  0  0
3 A  8  1  0  0
4 B 25  0  1  0
5 B 27  0  1  0
6 B 23  0  1  0
7 C 38  0  0  1
8 C 42  0  0  1
9 C 40  0  0  1

> # Same thing, using fastDummies:
>
> df2<-dummy_cols(df[,1:2],select_columns=c("D"))
> df2
  D  Y D_A D_B D_C
1 A 12   1   0   0
2 A 16   1   0   0
3 A  8   1   0   0
4 B 25   0   1   0
5 B 27   0   1   0
6 B 23   0   1   0
7 C 38   0   0   1
8 C 42   0   0   1
9 C 40   0   0   1
```

```
> # Regressions:
>
> summary(lm(Y~D,data=df))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    12.00       1.63    7.35  0.00032 ***
DB             13.00       2.31    5.63  0.00134 **
DC             28.00       2.31   12.12 0.000019 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.83 on 6 degrees of freedom
Multiple R-squared:  0.961, Adjusted R-squared:  0.948
F-statistic: 73.6 on 2 and 6 DF,  p-value: 0.00006


> summary(lm(Y~DA+DB+DC,data=df))

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)    40.00       1.63    24.5 0.0000003 ***
DA            -28.00       2.31   -12.1 0.0000191 ***
DB            -15.00       2.31    -6.5   0.00063 ***
DC                NA         NA      NA        NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.83 on 6 degrees of freedom
Multiple R-squared:  0.961, Adjusted R-squared:  0.948
F-statistic: 73.6 on 2 and 6 DF,  p-value: 0.00006
```

Many Dummies: Regression (continued)

```
> summary(lm(Y~DA+DB+DC-1,data=df)) # "-1" removes the constant

Call:
lm(formula = Y ~ DA + DB + DC - 1, data = df)

Coefficients:
   Estimate Std. Error t value  Pr(>|t|)
DA    12.00       1.63    7.35   0.00032 ***
DB    25.00       1.63   15.31 0.0000049 ***
DC    40.00       1.63   24.49 0.0000003 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.83 on 6 degrees of freedom
Multiple R-squared:  0.993, Adjusted R-squared:  0.99
F-statistic:  296 on 3 and 6 DF,  p-value: 0.000000659
```

Suppose we have:

$$\texttt{PID} = \begin{cases} 1 = \text{Strong Democrat} \\ 2 = \text{Weak Democrat} \\ 3 = \text{Independent} \\ 4 = \text{Weak Republican} \\ 5 = \text{Strong Republican} \end{cases}$$

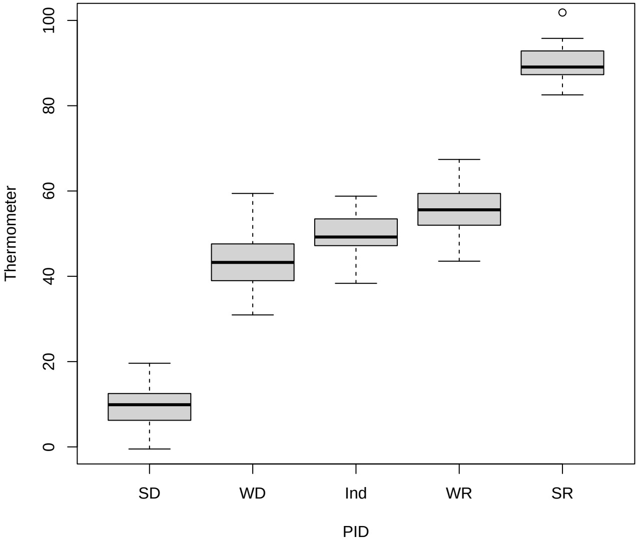
We might estimate:

$$\texttt{Thermometer}_i = \beta_0 + \beta_1(\texttt{PID}_i) + u_i$$

Alternatively, we could "dummy out" PID:

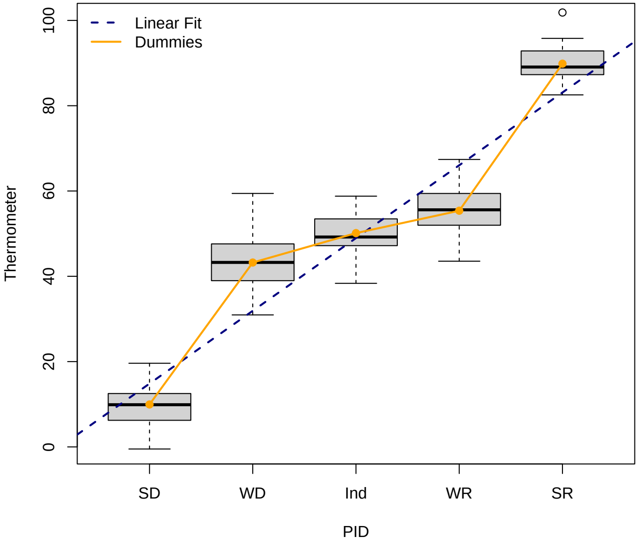
$$\texttt{Thermometer}_i = \beta_1(\texttt{SD}_i) + \beta_2(\texttt{WD}_i) + \beta_3(\texttt{Ind}_i) + \beta_4(\texttt{WR}_i) + \beta_5(\texttt{SR}_i) + u_i$$

Ordinal, Illustrated



```
> # Regressions:
>
> fit1<-lm(Therm~as.numeric(PID))
> summary(fit1)

Coefficients:
                Estimate Std. Error t value            Pr(>|t|)
(Intercept)       -2.233      1.575   -1.42                0.16
as.numeric(PID)   17.067      0.476   35.88 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.56 on 198 degrees of freedom
Multiple R-squared:  0.867, Adjusted R-squared:  0.866
F-statistic: 1.29e+03 on 1 and 198 DF,  p-value: <0.0000000000000002


> fit2<-lm(Therm~PID-1)
> summary(fit2)

Coefficients:
       Estimate Std. Error t value            Pr(>|t|)
PIDSD     9.949      0.792    12.6 <0.0000000000000002 ***
PIDWD    43.227      0.854    50.6 <0.0000000000000002 ***
PIDInd   50.132      0.866    57.9 <0.0000000000000002 ***
PIDWR    55.380      0.758    73.1 <0.0000000000000002 ***
PIDSR    89.855      0.854   105.2 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.19 on 195 degrees of freedom
Multiple R-squared:  0.991, Adjusted R-squared:  0.991
F-statistic: 4.5e+03 on 5 and 195 DF,  p-value: <0.0000000000000002
```

Ordinal X (continued)



E.g.,

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + u_i$$

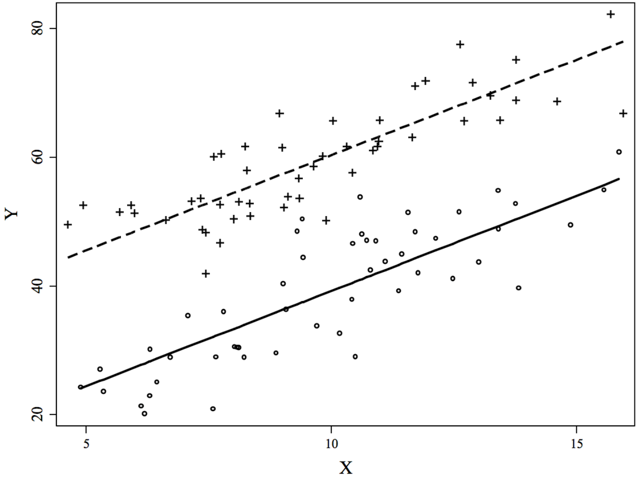
we have

$$\mathrm{E}(Y|X, D = 0) = \beta_0 + \beta_2 X_i$$

and

$$\mathrm{E}(Y|X, D = 1) = (\beta_0 + \beta_1) + \beta_2 X_i.$$

Dichotomous + Continuous X



Examples: SCOTUS (OT1953-1985)

From the "Phase II" SCOTUS database...

```
> summary(SCOTUS)
       id            term           Namici          lctdiss          multlaw
 Min.   :   1   Min.   :53.00   Min.   : 0.000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:1791   1st Qu.:64.00   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :3581   Median :72.00   Median : 0.000   Median :0.0000   Median :0.0000
 Mean   :3581   Mean   :71.12   Mean   : 0.842   Mean   :0.1509   Mean   :0.1490
 3rd Qu.:5371   3rd Qu.:79.00   3rd Qu.: 1.000   3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   :7161   Max.   :85.00   Max.   :39.000   Max.   :1.0000   Max.   :1.0000
                NA's   : 4.00                    NA's   :4.0000   NA's   :5.0000
    civlibs           econs            constit           lctlib
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :  0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:  0.0000
 Median :1.0000   Median :0.0000   Median :0.0000   Median :  0.0000
 Mean   :0.5009   Mean   :0.1709   Mean   :0.2536   Mean   :  0.3742
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:  1.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :  1.0000
                                                    NA's   :120.0000
```

All civil rights & economics cases:

```
> SCOTUS$civil.econ<-SCOTUS$civlibs + SCOTUS$econs
```

Factors:

```
> SCOTUS$termdummies<-factor(SCOTUS$term)
> is.factor(SCOTUS$termdummies)
[1] TRUE
> summary(SCOTUS$termdummies)
  53   54   55   56   57   58   59   60   61   62   63   64   65   66   67
 126  109  128  162  196  165  157  160  148  189  223  156  187  201  285

  68   69   70   71   72   73   74   75   76   77   78   79   80   81
 207  185  227  262  269  267  223  253  254  244  244  221  255  269

  82   83   84   85 NA's
 277  298  301  309    4
```

Regressions (vs. *t*-tests...)

```
> fit1<-with(SCOTUS, lm(Namici~civlibs))
> summary(fit1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.91774    0.03661  25.069  < 2e-16 ***
civlibs     -0.15136    0.05173  -2.926  0.00344 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.189 on 7159 degrees of freedom
Multiple R-squared:  0.001195, Adjusted R-squared:  0.001055
F-statistic: 8.563 on 1 and 7159 DF,  p-value: 0.003442
```

```
> with(SCOTUS, t.test(Namici~civlibs))

 Welch Two Sample t-test

data:  Namici by civlibs
t = 2.9258, df = 7114.116, p-value = 0.003446
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.04995001 0.25277126
sample estimates:
mean in group 0 mean in group 1
      0.9177392       0.7663786
```

Dummy vs. effect coding:

```
> SCOTUS$civlibeffect<-SCOTUS$civlibs
> SCOTUS$civlibeffect[SCOTUS$civlibs==0]<-(-1)
> fit2<-with(SCOTUS, lm(Namici~SCOTUS$civlibeffect))
> summary(fit2)

Call:
lm(formula = Namici ~ SCOTUS$civlibeffect)

Residuals:
   Min     1Q Median     3Q    Max
-0.918 -0.918 -0.766  0.082 38.234

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.84206    0.02586  32.559  < 2e-16 ***
SCOTUS$civlibeffect -0.07568    0.02586  -2.926  0.00344 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.189 on 7159 degrees of freedom
Multiple R-squared:  0.001195, Adjusted R-squared:  0.001055
F-statistic: 8.563 on 1 and 7159 DF,  p-value: 0.003442
```

```
> fit3<-with(SCOTUS, lm(Namici~lctdiss+multlaw+civlibs+
+                          econs+constit+lctlib))
> summary(fit3)

Call:
lm(formula = Namici ~ lctdiss + multlaw + civlibs + econs + constit +
    lctlib)

Residuals:
   Min     1Q Median     3Q    Max
-2.582 -0.976 -0.472 -0.260 37.086

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.47245    0.05273   8.960  < 2e-16 ***
lctdiss      0.36760    0.07173   5.125 3.06e-07 ***
multlaw      0.61306    0.07445   8.235  < 2e-16 ***
civlibs     -0.21255    0.06022  -3.530 0.000419 ***
econs        0.08772    0.07652   1.146 0.251691
constit      0.53793    0.06372   8.442  < 2e-16 ***
lctlib       0.50309    0.05396   9.323  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.15 on 7033 degrees of freedom
  (121 observations deleted due to missingness)
Multiple R-squared:  0.05013, Adjusted R-squared:  0.04932
F-statistic: 61.86 on 6 and 7033 DF,  p-value: < 2.2e-16
```

Change Over Time: Linear Trend

```
> fit4<-with(SCOTUS, lm(Namici~lctdiss+multlaw+civlibs+
+                          econs+constit+lctlib+term))
> summary(fit4)

Call:
lm(formula = Namici ~ lctdiss + multlaw + civlibs + econs + constit +
    lctlib + term)

Residuals:
   Min     1Q Median     3Q    Max
-2.968 -0.906 -0.428  0.143 36.958

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.726962   0.202367 -13.475  < 2e-16 ***
lctdiss      0.359494   0.070415   5.105 3.39e-07 ***
multlaw      0.649932   0.073109   8.890  < 2e-16 ***
civlibs     -0.289314   0.059295  -4.879 1.09e-06 ***
econs        0.199464   0.075419   2.645  0.00819 **
constit      0.515435   0.062559   8.239  < 2e-16 ***
lctlib       0.339891   0.053901   6.306 3.04e-10 ***
term         0.046142   0.002821  16.354  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.11 on 7032 degrees of freedom
  (121 observations deleted due to missingness)
Multiple R-squared:  0.08493, Adjusted R-squared:  0.08402
F-statistic: 93.24 on 7 and 7032 DF,  p-value: < 2.2e-16
```

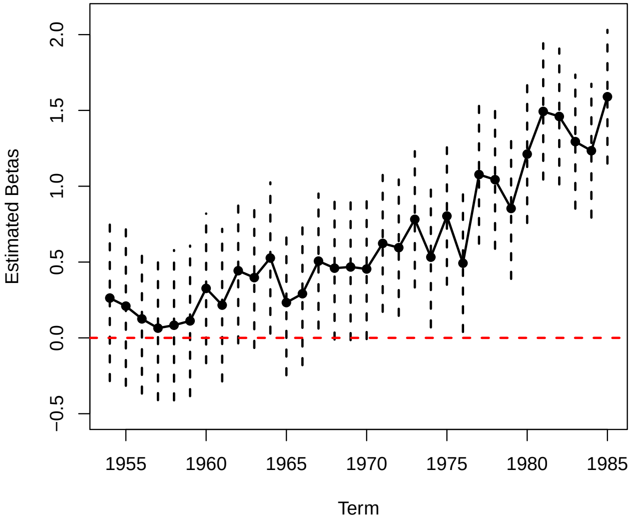
Change Over Time: Using `factor`

```
> fit5<-with(SCOTUS, lm(Namici~lctdiss+multlaw+civlibs+
+                       econs+constit+lctlib+as.factor(term)))
> summary(fit5)

Call:
lm(formula = Namici ~ lctdiss + multlaw + civlibs + econs + constit +
    lctlib + as.factor(term))

Residuals:
   Min     1Q Median     3Q    Max
-3.064 -0.920 -0.384  0.106 36.831

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.16153    0.19530  -0.827 0.408200
lctdiss            0.34558    0.07067   4.890 1.03e-06 ***
multlaw            0.64348    0.07334   8.774  < 2e-16 ***
civlibs           -0.27137    0.05967  -4.548 5.51e-06 ***
econs              0.20039    0.07581   2.643 0.008232 **
constit            0.54280    0.06297   8.620  < 2e-16 ***
lctlib             0.33863    0.05458   6.205 5.80e-10 ***
.
.
.
```

Using `factor` (continued)

```
as.factor(term)54  0.26276    0.27934   0.941 0.346918
as.factor(term)55  0.20958    0.26804   0.782 0.434309
as.factor(term)56  0.12536    0.25126   0.499 0.617859
as.factor(term)57  0.06432    0.24227   0.265 0.790654
as.factor(term)58  0.08353    0.25274   0.331 0.741025
.
.
.
as.factor(term)71  0.62313    0.23019   2.707 0.006806 **
as.factor(term)72  0.59503    0.22929   2.595 0.009476 **
as.factor(term)73  0.78179    0.22918   3.411 0.000650 ***
as.factor(term)74  0.53254    0.23636   2.253 0.024287 *
as.factor(term)75  0.80353    0.23118   3.476 0.000513 ***
as.factor(term)76  0.49269    0.23138   2.129 0.033262 *
as.factor(term)77  1.07725    0.23265   4.630 3.72e-06 ***
as.factor(term)78  1.04335    0.23243   4.489 7.27e-06 ***
as.factor(term)79  0.85363    0.23696   3.602 0.000318 ***
as.factor(term)80  1.21205    0.23183   5.228 1.76e-07 ***
as.factor(term)81  1.49347    0.22925   6.515 7.80e-11 ***
as.factor(term)82  1.46004    0.22858   6.388 1.79e-10 ***
as.factor(term)83  1.29417    0.22549   5.739 9.90e-09 ***
as.factor(term)84  1.23434    0.22517   5.482 4.36e-08 ***
as.factor(term)85  1.59037    0.22491   7.071 1.68e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.108 on 7001 degrees of freedom
  (121 observations deleted due to missingness)
Multiple R-squared:  0.0914, Adjusted R-squared:  0.08647
F-statistic: 18.53 on 38 and 7001 DF,  p-value: < 2.2e-16
```

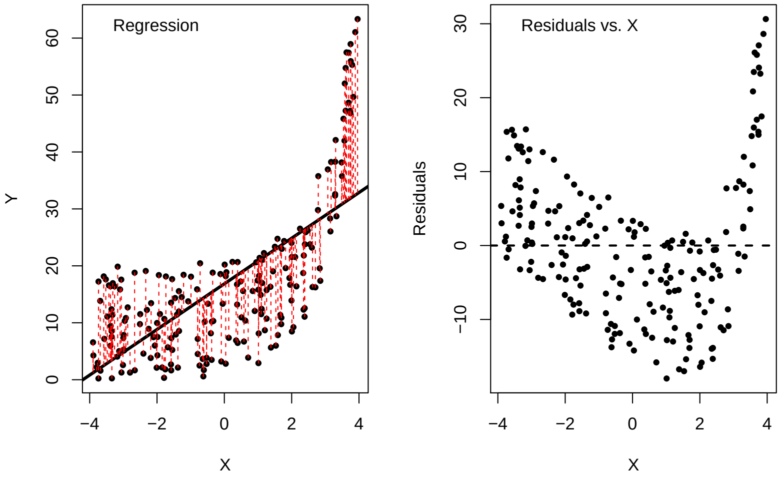
`factor` results, plotted ($1953 = 0$)



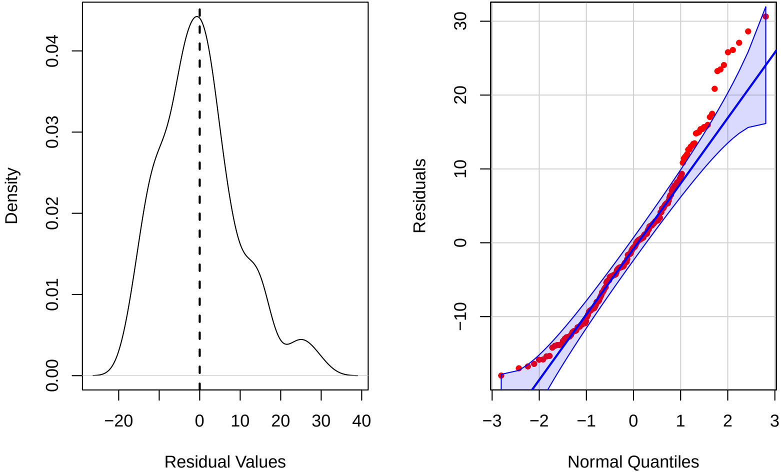
Transformations

- Normality (of u_is)
- Linearity
- Additivity
- Interpretation / Model Specification

What Difference Does It Make? (Part I)



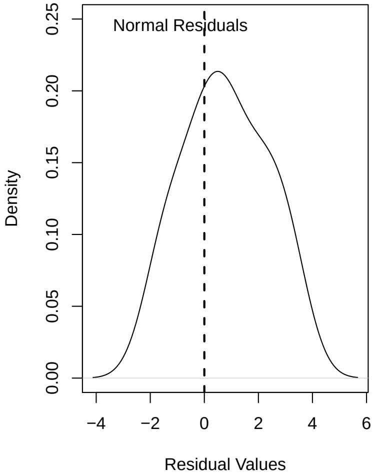
Residuals Are Still (Pretty) Normal...

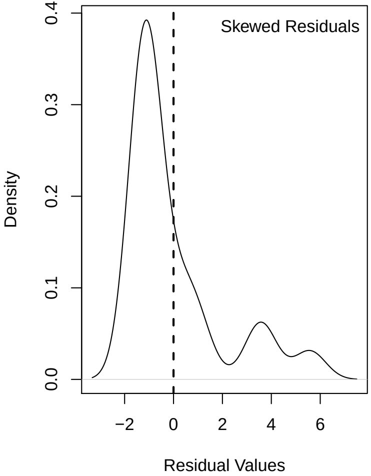


What Difference Does It Make? (Part II)

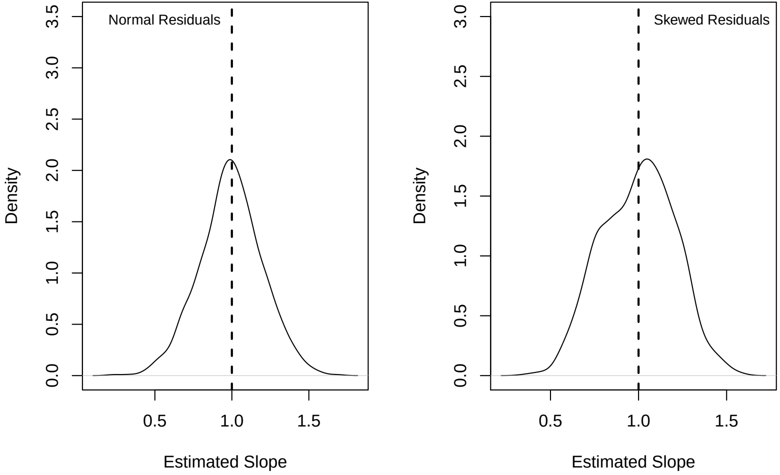
```
N <- 20 # pretty small sample size
u <- rnorm(N,0,2) # mean zero, s.d = 2

# Exponentiate:
eu <- exp(u)
eu <- eu-mean(eu) # new residuals are mean-zero
eu <- (eu/sd(eu))*2 # and also sd = 2

X <- runif(N,-4,4)
Y1 <- 0 + 1*X + 1*u
Y2 <- 0 + 1*X + 1*eu # same Xs in both
```

What Difference Does It Make? (Part II)

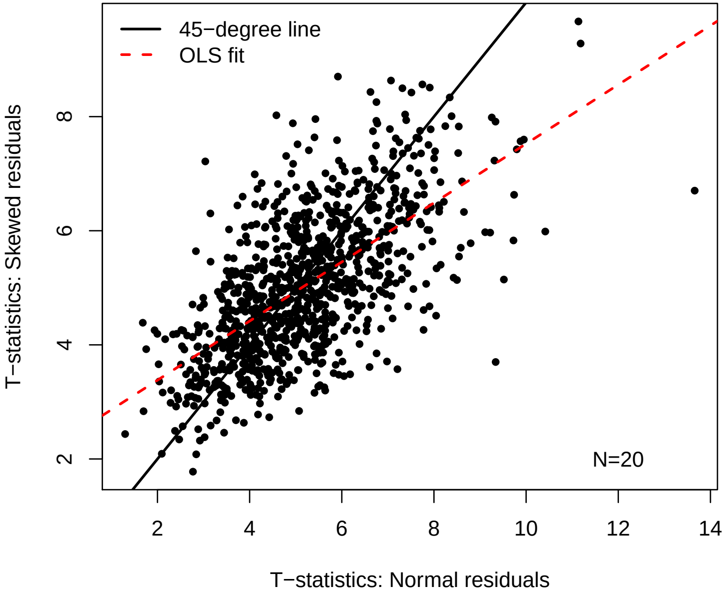




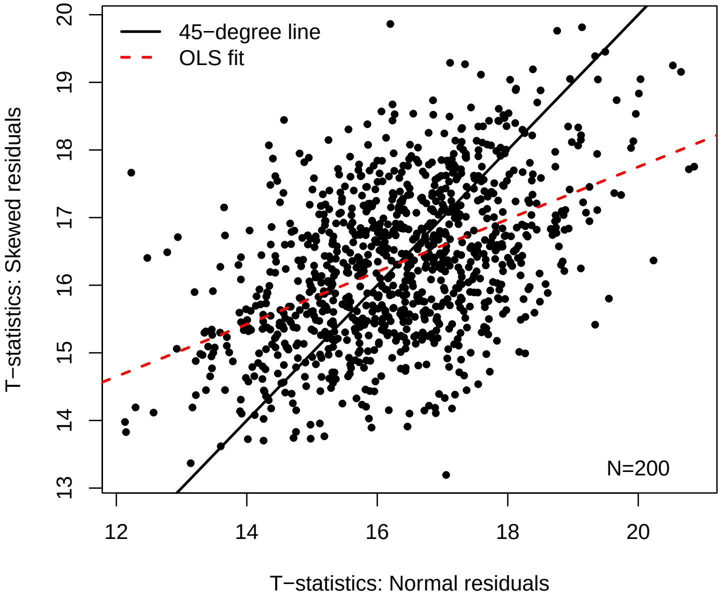
Little Effect On $\hat{\beta}$



Important Differences in Inference



With $N = 200$? Not So Much...



This:

$$Y_i = \beta_0 X_i^{\beta_1} u_i$$

becomes this:

$$\ln(Y_i) = \ln(\beta_0) + \beta_1 X_i + \ln(u_i)$$

And this:

$$\exp(Y_i) = \beta_0 + \beta_1 X_i + u_i$$

becomes this:

$$Y_i = \ln(\beta_0) + \beta_1 \ln(X_i) + \ln(u_i)$$

Monotonic Transformations

The "Ladder of Powers":

Transformation	p	$f(X)$	Fox's $f(X)$
Cube	3	X^3	$\frac{X^3-1}{3}$
Square	2	X^2	$\frac{X^2-1}{2}$
(None/Identity)	(1)	(X)	(X)
Square Root	$\frac{1}{2}$	$\sqrt{X}$	$2(\sqrt{X}-1)$
Cube Root	$\frac{1}{3}$	$\sqrt[3]{X}$	$3(\sqrt[3]{X}-1)$
Log	0 (sort of)	$\ln(X)$	$\ln(X)$
Inverse Cube Root	$-\frac{1}{3}$	$\frac{1}{\sqrt[3]{X}}$	$\frac{\left(\frac{1}{\sqrt[3]{X}}-1\right)}{-\frac{1}{3}}$
Inverse Square Root	$-\frac{1}{2}$	$\frac{1}{\sqrt{X}}$	$\frac{\left(\frac{1}{\sqrt{X}}-1\right)}{-\frac{1}{2}}$
Inverse	-1	$\frac{1}{X}$	$\frac{\left(\frac{1}{X}-1\right)}{-1}$
Inverse Square	-2	$\frac{1}{X^2}$	$\frac{\left(\frac{1}{X^2}-1\right)}{-2}$
Inverse Cube	-3	$\frac{1}{X^3}$	$\frac{\left(\frac{1}{X^3}-1\right)}{-3}$

Using higher-order power transformations (e.g. squares, cubes, etc.) "inflates" large values and "compresses" small ones; conversely, using lower-order power transformations (logs, etc.) "compresses" large values and "inflates" (or "expands") smaller ones.

Power Transformations: Two Issues

1. X must be *positive*; so:

$$X^* = X + (|X_l| + \epsilon)$$

with (CZ's Rule of Thumb):

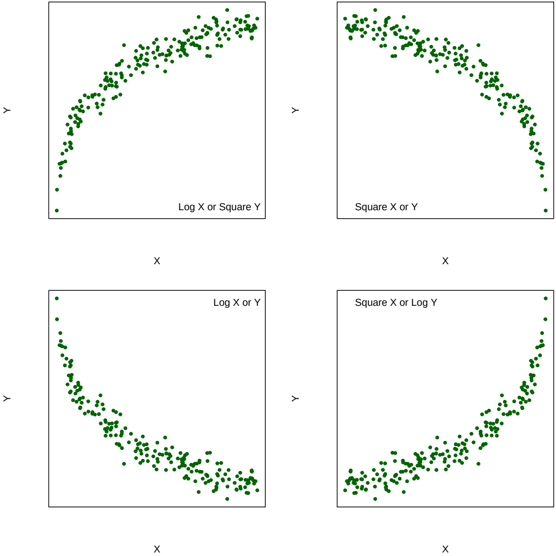
$$\epsilon = \frac{X_{l+1} - X_l}{2}$$

2. Power transformations generally require that:

$$\frac{X_h}{X_l} > 5 \text{ (or so)}$$

Which Transformation?

Mosteller and Tukey's "Bulging Rule":



For:

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i,$$

then:

$$\mathrm{E}(Y) = \exp(\beta_0 + \beta_1 X_i)$$

and so:

$$\frac{\partial \mathrm{E}(Y)}{\partial X} = \exp(\beta_1).$$

Similarly, for:

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$$

we have:

$$\frac{\partial \mathrm{E}(Y)}{\partial \ln(X)} = \beta_1.$$

So doubling X (say, from X_ℓ to $2X_\ell$):

$$\begin{aligned}
\Delta \mathrm{E}(Y) &= \mathrm{E}(Y|X = 2X_\ell) - \mathrm{E}(Y|X = X_\ell) \\
&= [\beta_0 + \beta_1 \ln(2X_\ell)] - [\beta_0 + \beta_1 \ln(X_\ell)] \\
&= \beta_1[\ln(2X_\ell) - \ln(X_\ell)] \\
&= \beta_1 \ln(2)
\end{aligned}$$

Specifying:

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + ... + u_i$$

means:

$$\text{Elasticity}_{YX} \equiv \frac{\%\Delta Y}{\%\Delta X} = \beta_1.$$

IOW, a one-percent change in X leads to a $\hat{\beta}_1$-percent change in Y.

An Example: Cell Phones and Wealth

Data are from the World Development Indicators (2018 *only*)...

- `Region` - The geographical region of the country
- `country` - The name of the country (useful for labeling, etc.)

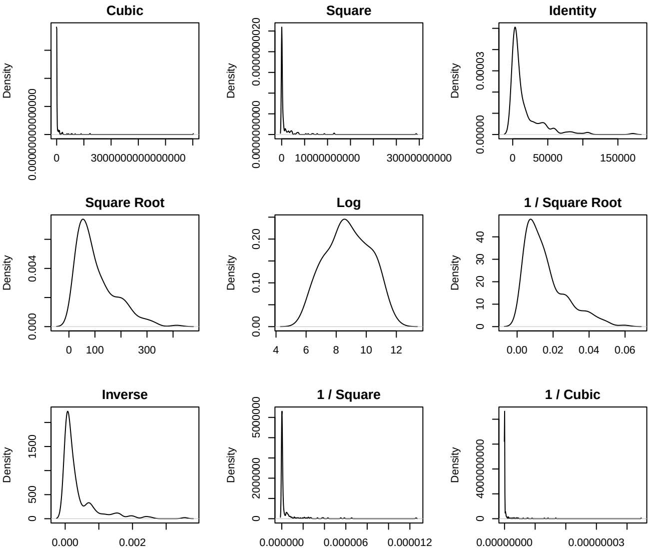
  .

  .

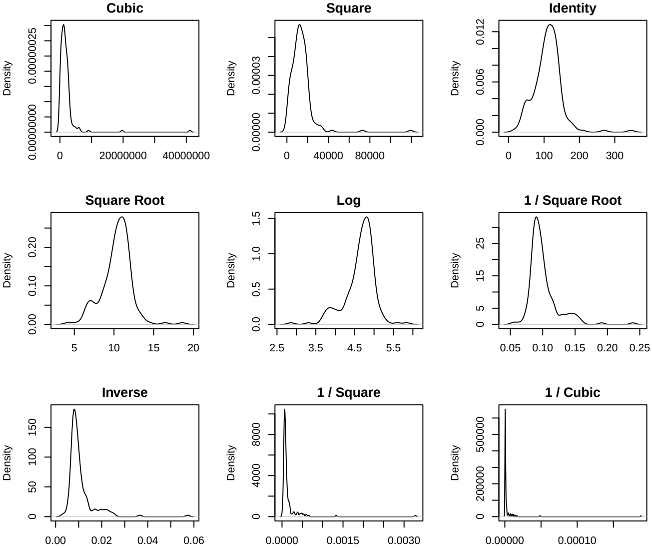
  .
- `GDPPerCapita` - GDP per capita (constant 2010 $US)
- `MobileCellSubscriptions` - Mobile / cellular subscriptions per 100 people

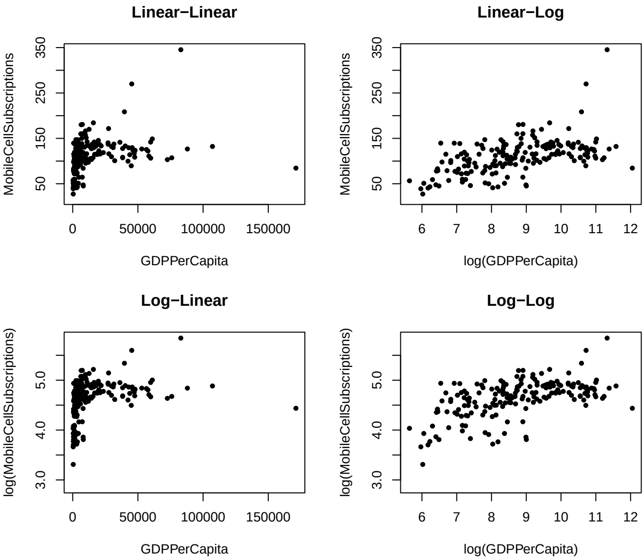
```
> with(WDI, describe(MobileCellSubscriptions))
   vars   n mean   sd median trimmed  mad  min max range skew kurtosis   se
X1    1 190  110 38.7    112     110 30.4 17.5 345   328  1.3     7.37 2.81

> with(WDI, describe(GDPPerCapita))
   vars   n  mean    sd median trimmed  mad min    max  range skew kurtosis   se
X1    1 203 16507 23853   6262   11418 7508 282 171224 170942 2.75     10.2 1674
```

“Ladder of Powers”: Wealth / GDP



"Ladder of Powers": Mobile Subscriptions




Linear−Linear
MobileCellSubscriptions
GDPPerCapita
Linear−Log
log(GDPPerCapita)
Log−Linear
log(MobileCellSubscriptions)
Log−Log
50
150
250
350
0
50000
100000
150000
6
7
8
9
10
11
12
3.0
4.0
5.0

Untransformed (linear-linear):

```
> linlin <- with(WDI, lm(MobileCellSubscriptions~I(GDPPerCapita/1000)))
> summary(linlin)

Residuals:
    Min      1Q  Median      3Q     Max
-114.17  -20.99   -0.76   19.38  196.73

Coefficients:
                     Estimate Std. Error t value  Pr(>|t|)
(Intercept)           101.541      3.282   30.94   < 2e-16 ***
I(GDPPerCapita/1000)    0.567      0.120    4.74 0.0000043 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36.6 on 181 degrees of freedom
  (32 observations deleted due to missingness)
Multiple R-squared:  0.111, Adjusted R-squared:  0.106
F-statistic: 22.5 on 1 and 181 DF,  p-value: 0.00000426
```

Logging X:

```
> linlog <- with(WDI, lm(MobileCellSubscriptions~log(GDPPerCapita/1000)))
> summary(linlog)

Residuals:
   Min     1Q Median     3Q    Max
 -72.6  -17.7   -3.9   15.5  198.5

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)               84.33       4.00   21.08  < 2e-16 ***
log(GDPPerCapita/1000)    14.15       1.72    8.23  3.6e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.1 on 181 degrees of freedom
  (32 observations deleted due to missingness)
Multiple R-squared:  0.272, Adjusted R-squared:  0.268
F-statistic: 67.7 on 1 and 181 DF,  p-value: 3.63e-14
```

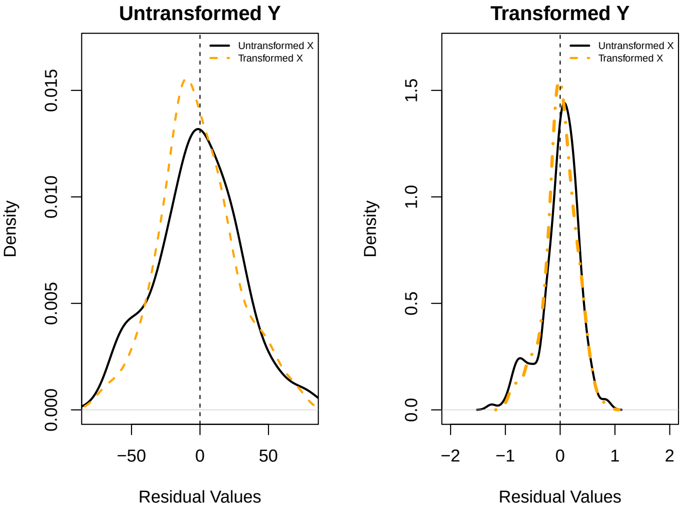
Logging Y:

```
> loglin <- with(WDI, lm(log(MobileCellSubscriptions)~I(GDPPerCapita/1000)))
> summary(loglin)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2519 -0.1540  0.0554  0.2192  0.8560

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.56082    0.03157  144.47  < 2e-16 ***
I(GDPPerCapita/1000)  0.00516    0.00115    4.48 0.000013 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.352 on 181 degrees of freedom
  (32 observations deleted due to missingness)
Multiple R-squared:  0.0998, Adjusted R-squared:  0.0949
F-statistic: 20.1 on 1 and 181 DF,  p-value: 0.0000132
```

Logging X and Y:

```
> loglog <- with(WDI, lm(log(MobileCellSubscriptions)~log(GDPPerCapita/1000)))
> summary(loglog)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9364 -0.1502  0.0114  0.1873  0.8311

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)              4.3752     0.0372  117.52  < 2e-16 ***
log(GDPPerCapita/1000)   0.1444     0.0160    9.02  2.6e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.308 on 181 degrees of freedom
  (32 observations deleted due to missingness)
Multiple R-squared:  0.31, Adjusted R-squared:  0.306
F-statistic: 81.4 on 1 and 181 DF,  p-value: 2.64e-16
```

Density Plots of $\hat{u}_i$s



(One) simple solution: Polynomials...

- First-order / linear ($P = 1$):

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Second-order / quadratic ($P = 2$):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

- Third-order / cubic ($P = 3$):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i$$

- ...pth-order ($P = p$):

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + ... + \beta_p X_i^p + u_i$$

Understanding Polynomials

Read coefficients "left to right." So, for the quadratic:

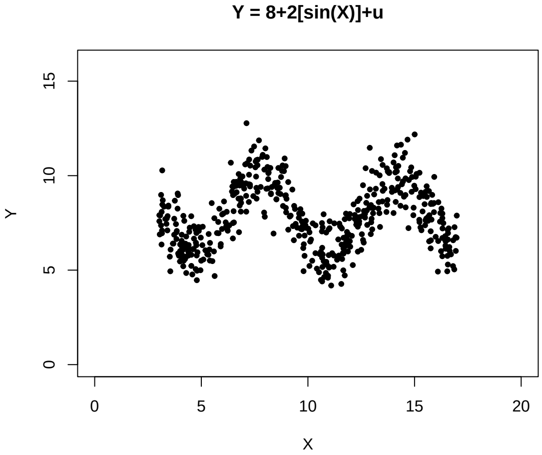
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

then:

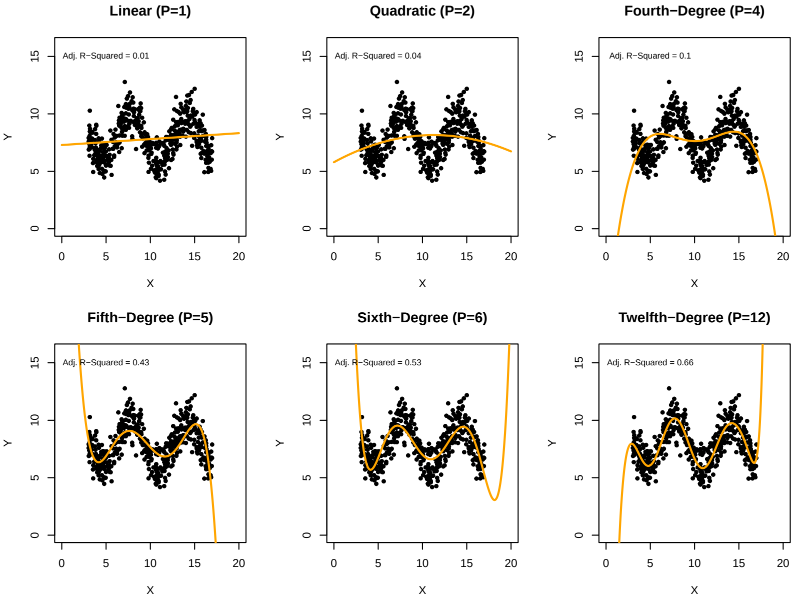
	$\hat{\beta}_2$		
$\hat{\beta}_1$	< 0	$= 0$	> 0
< 0	E(Y) decreases in X at an increasing rate	E(Y) decreases linearly in X	E(Y) decreases in X at low values of X, but increases in X at high values of X
$= 0$	E(Y) decreases in X^2	E(Y) is (quadratically) unrelated to X	E(Y) increases in X^2
> 0	E(Y) increases in X at low values of X, but decreases in X at high values of X	E(Y) increases linearly in X	E(Y) increases in X at an increasing rate

Polynomials: Simulated Example

```
> N<-500
> set.seed(7222009)
> X<-runif(N,3,17)
> Y<-8+2*sin(X)+rnorm(N)
```



Some Polynomial Regressions



Check out the $P = 12$ regression:

```
> summary(R.12)

Call:
lm(formula = Y ~ X + I(X^2) + I(X^3) + I(X^4) + I(X^5) + I(X^6) +
    I(X^7) + I(X^8) + I(X^9) + I(X^10) + I(X^11) + I(X^12))

Coefficients: (1 not defined because of singularities)
                  Estimate      Std. Error t value Pr(>|t|)
(Intercept)  -92.9216900503  816.2489413571   -0.11     0.91
X            149.4055086103 1212.1158531764    0.12     0.90
I(X^2)      -104.6855472388  788.3527710495   -0.13     0.89
I(X^3)        46.9192208616  296.7267961681    0.16     0.87
I(X^4)       -14.5570311719   71.9111574249   -0.20     0.84
I(X^5)         3.1175787785   11.8013895994    0.26     0.79
I(X^6)        -0.4537511003    1.3406553442   -0.34     0.74
I(X^7)         0.0442854569    0.1056218156    0.42     0.68
I(X^8)        -0.0028398562    0.0056659448   -0.50     0.62
I(X^9)         0.0001145568    0.0001974535    0.58     0.56
I(X^10)       -0.0000026340    0.0000040301   -0.65     0.51
I(X^11)        0.0000000263    0.0000000366    0.72     0.47
I(X^12)                  NA              NA      NA       NA

Residual standard error: 0.986 on 488 degrees of freedom
Multiple R-squared:  0.669, Adjusted R-squared:  0.662
F-statistic: 89.7 on 11 and 488 DF,  p-value: <2e-16
```

“Raw” vs. Orthogonal Polynomials (continued)

What’s going on?

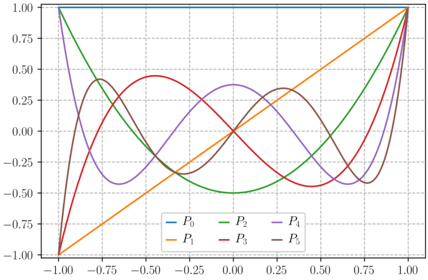
- The “raw” polynomial terms are (often strongly) correlated with each other...

```
> cor(X,I(X^2))
[1] 0.984
```

- $\rightarrow$ large standard errors / imprecision in the estimates
- Can also lead to numerical instability in estimation...

An alternative is to use *orthogonal polynomials*...

- Think of these as orthogonal (uncorrelated) versions of the polynomials above
- There are many of them; probably the most commonly-used are the Legendre polynomials:



- The math is a bit complex; the R command is `poly()`

"Raw" polynomials using poly():

```
> P.12R<-lm(Y~poly(X,degree=12,raw=TRUE))
> summary(P.12R)

Call:
lm(formula = Y ~ poly(X, degree = 12, raw = TRUE))

Coefficients: (1 not defined because of singularities)
                                         Estimate      Std. Error t value Pr(>|t|)
(Intercept)                         -92.9216900503  816.2489413571   -0.11     0.91
poly(X, degree = 12, raw = TRUE)1   149.4055086103 1212.1158531764    0.12     0.90
poly(X, degree = 12, raw = TRUE)2  -104.6855472388  788.3527710495   -0.13     0.89
poly(X, degree = 12, raw = TRUE)3    46.9192208616  296.7267961681    0.16     0.87
poly(X, degree = 12, raw = TRUE)4   -14.5570311719   71.9111574249   -0.20     0.84
poly(X, degree = 12, raw = TRUE)5     3.1175787785   11.8013895994    0.26     0.79
poly(X, degree = 12, raw = TRUE)6    -0.4537511003    1.3406553442   -0.34     0.74
poly(X, degree = 12, raw = TRUE)7     0.0442854569    0.1056218156    0.42     0.68
poly(X, degree = 12, raw = TRUE)8    -0.0028398562    0.0056659448   -0.50     0.62
poly(X, degree = 12, raw = TRUE)9     0.0001145568    0.0001974535    0.58     0.56
poly(X, degree = 12, raw = TRUE)10   -0.0000026340    0.0000040301   -0.65     0.51
poly(X, degree = 12, raw = TRUE)11    0.0000000263    0.0000000366    0.72     0.47
poly(X, degree = 12, raw = TRUE)12              NA              NA      NA       NA

Residual standard error: 0.986 on 488 degrees of freedom
Multiple R-squared:  0.669, Adjusted R-squared:  0.662
F-statistic: 89.7 on 11 and 488 DF,  p-value: <2e-16
```

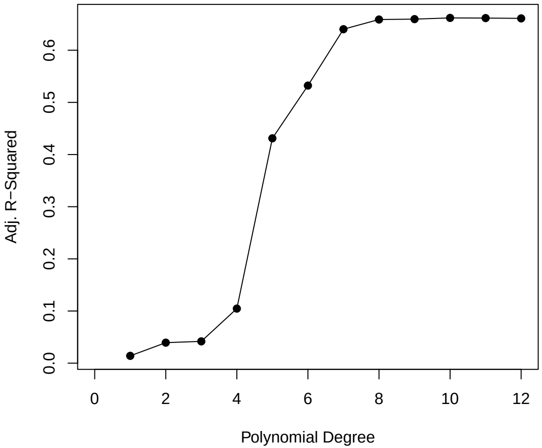
Orthogonal polynomials:

```
> P.12<-lm(Y~poly(X,degree=12))
> summary(P.12)

Call:
lm(formula = Y ~ poly(X, degree = 12))

Coefficients:
                      Estimate Std. Error t value       Pr(>|t|)
(Intercept)             7.7989     0.0441  176.80        < 2e-16 ***
poly(X, degree = 12)1   4.7901     0.9863    4.86 0.00000161352 ***
poly(X, degree = 12)2  -6.2379     0.9863   -6.32 0.00000000058 ***
poly(X, degree = 12)3   2.5039     0.9863    2.54          0.011 *
poly(X, degree = 12)4  -9.5937     0.9863   -9.73        < 2e-16 ***
poly(X, degree = 12)5 -21.5763     0.9863  -21.87        < 2e-16 ***
poly(X, degree = 12)6  12.0295     0.9863   12.20        < 2e-16 ***
poly(X, degree = 12)7  12.4067     0.9863   12.58        < 2e-16 ***
poly(X, degree = 12)8  -5.2176     0.9863   -5.29 0.00000018541 ***
poly(X, degree = 12)9  -1.4389     0.9863   -1.46          0.145
poly(X, degree = 12)10  2.0529     0.9863    2.08          0.038 *
poly(X, degree = 12)11  0.7097     0.9863    0.72          0.472
poly(X, degree = 12)12 -0.3987     0.9863   -0.40          0.686
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.986 on 487 degrees of freedom
Multiple R-squared:  0.669, Adjusted R-squared:  0.661
F-statistic: 82.1 on 12 and 487 DF,  p-value: <2e-16
```

What Degree Polynomial?

```
> for(degree in 1:12) {
+   fit <- lm(Y~poly(X,degree))
+   assign(paste("P", degree, sep = "."), fit)
+ }
> anova(P.1,P.2,P.3,P.4,P.5,P.6,P.7,P.8,P.9,P.10,P.11,P.12)
Analysis of Variance Table

Model  1: Y ~ poly(X, degree)
Model  2: Y ~ poly(X, degree)
Model  3: Y ~ poly(X, degree)
Model  4: Y ~ poly(X, degree)
Model  5: Y ~ poly(X, degree)
Model  6: Y ~ poly(X, degree)
Model  7: Y ~ poly(X, degree)
Model  8: Y ~ poly(X, degree)
Model  9: Y ~ poly(X, degree)
Model 10: Y ~ poly(X, degree)
Model 11: Y ~ poly(X, degree)
Model 12: Y ~ poly(X, degree)
   Res.Df  RSS Df Sum of Sq      F        Pr(>F)
1     498 1409
2     497 1370  1        39  40.00 0.00000000058 ***
3     496 1364  1         6   6.44         0.011 *
4     495 1272  1        92  94.60       < 2e-16 ***
5     494  807  1       466 478.51       < 2e-16 ***
6     493  662  1       145 148.74       < 2e-16 ***
7     492  508  1       154 158.22       < 2e-16 ***
8     491  481  1        27  27.98 0.00000018541 ***
9     490  479  1         2   2.13         0.145
10    489  474  1         4   4.33         0.038 *
11    488  474  1         1   0.52         0.472
12    487  474  1         0   0.16         0.686
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Plotting R^2_{adj} for different polynomial degrees...



Polynomial Tips

Good things...

- Polynomials are flexible functional forms for nonlinear marginal associations
- They are also easy to fit, and easily interpretable

Cautions...

- Polynomials can be prone to overfitting, which...
- ...can lead to poor out-of-sample generalizability / predictive power
- This is especially true outside the observed values of the data (extrapolation)

- **Theory is valuable**.
- **Try different things**.
- **Look at plots**.
- **It takes practice**.