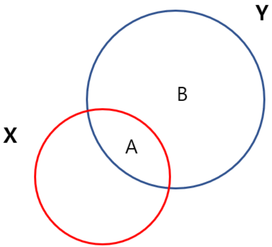
Multiple Linear Regression

CON 2012: Consumer Big Data Analysis I

Instructor: Tae-Young Pak

Spring 2022

A Venn diagram representation of the SLR



A: variation in y explained by x

B: variation in y not explained by x

$$R^2 = \frac{\text{explained variation in y}}{\text{variation in y}} = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$$
$$= \frac{A}{A+B}$$

Data preprocessing using residual

- Practically, residual indicates variation in y that is not explained by the model
- In the SLR case, it is just area B
- We can exploit this fact to isolate variation in y that is not explained by “something”

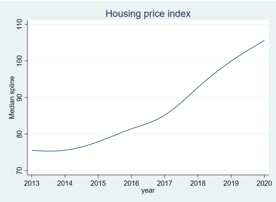
Example: isolating variation in housing price index not explained by money supply

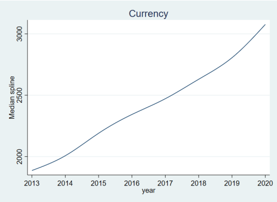
- Housing price is going up
- Some politicians argue that it is due to low interest rate and increased money supply
- This is not a matter of debate; just remove variation in housing price index due to changes in money supply and see how the remaining variation changes over time
- But how? $\Rightarrow$ regress housing price index on money supply and take residuals
- This residual indicates variation in housing price index not explained by money supply ($\Rightarrow$ i.e., what has happened to housing price index if we take out the influence of fluctuating money supply)

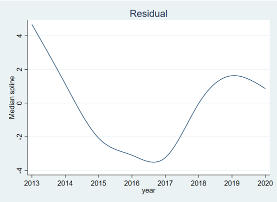
Example: isolating variation in housing price index unrelated to money supply

Changes in housing price index, currency in circulation, and residual:

- Housing price index from the KB bank monthly housing price trend ▸ click here
- M2 currency data from the Bank of Korea Economic Statistics System ▸ click here







Multiple linear regression: regression with more than one predictor

- So far, we have seen simple linear regressions where a single predictor x was used to model the outcome variable y
- In many applications, there is more than one variable that influences the outcome
- Multiple linear regression offers a more realistic setup where the outcome variable y is explained by multiple predictor

Examples:

- The market price of a house depends on location, the number of bedrooms, the number of bathrooms, the year the house was built, the square footage of the lot and a number of other factors
- Your letter grade would depend on hours of study, completion of prerequisites, and other possible distractors

Multiple linear regression (MLR) with k predictors:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

- Much like the SLR case, our goal here is to estimate PRL, $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$, using data
- The estimated regression is written as, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$
- Uses the LS estimator to estimate beta parameters

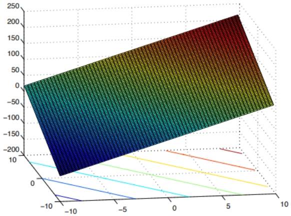
Example: The simplest multiple regression model is the regression with two predictors:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

If the estimated model takes the following form,

$$\hat{y} = 50 + 10x_1 + 7x_2$$

then, it is a plane in a three dimensional space with different slopes in x_1 and x_2 direction.



Residual sum of squares (RSS):

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_k x_{ki})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

$\Rightarrow$ Just like the SLR case, the LS estimator is obtained by minimizing the RSS with respect to betas

$\Rightarrow$ Two solutions: analytic approach, gradient descent

Goodness of fit:

- We've learned one measure of model fit, R^2
- The problem of R^2 is that it never goes down; it always increase with more predictors irrespective of whether they predict the outcome variable or not
- This means that R^2 will always favor the model with more predictors $\Rightarrow$ this is no good (curse of dimensionality; overfitting)
- A solution to this problem is to use **adjusted** R^2
- **Adjusted** R^2 is R^2 times a penalty term, which penalizes R^2 for the addition of a predictor with no explanatory power
- From this time on, model fit (or, goodness of fit) refers to **adjusted** R^2

Adjusted R^2

Adjusted $R^2 = 1 - \frac{RSS(n-1)}{TSS(n-k-1)}$

Interpretation of $\hat{\beta}$:

If the estimated model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$,

$\hat{\beta}_1 = \frac{\partial \hat{y}}{\partial x_1}$: changes in y for a unit increase in x_1 while holding other x variables constant

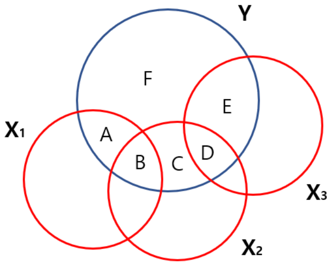
$\hat{\beta}_2 = \frac{\partial \hat{y}}{\partial x_2}$: changes in y for a unit increase in x_2 while holding other x variables constant

$\vdots$

$\hat{\beta}_k = \frac{\partial \hat{y}}{\partial x_k}$: changes in y for a unit increase in x_k while holding other x variables constant

$\Rightarrow \hat{\beta}_k$: marginal effect of x_k on y

A Venn diagram representation of the MLR



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

$$R^2 = \frac{\text{explained variation in y}}{\text{variation in y}} == 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$$
$$= \frac{A+B+C+D+E}{A+B+C+D+E+F}$$

MLR example: predicting salary of the major league baseball players using their stats.

R code:

```
library(ISLR)
mlb <- data.frame(Hitters)

mlb <- mlb[, c(-14:-16, -20)]

mlb <- na.omit(mlb)

corr.mat <- cor(mlb)
round(corr.mat, 2)

res.lm <- lm(Salary ~ HmRun, data = mlb)
summary(res.lm)
res.lm <- lm(Salary ~ HmRun + Errors, data = mlb)
summary(res.lm)
res.lm <- lm(Salary ~ HmRun + Errors + RBI, data = mlb)
summary(res.lm)
res.lm <- lm(Salary ~ HmRun + Errors + RBI + Assists, data = mlb)
summary(res.lm)
res.lm <- lm(Salary ~ ., data = mlb)
summary(res.lm)
```

R results:

	AtBat	Hits	HmRun	Runs	RBI	Walks	Years	CAtBat	CHits	CHmRun	CRuns	CRBI	CWalks	Assists	Errors	Salary
AtBat	1.00	0.96	0.56	0.90	0.80	0.62	0.01	0.21	0.23	0.21	0.24	0.22	0.13	0.34	0.33	0.39
Hits	0.96	1.00	0.53	0.91	0.79	0.59	0.02	0.21	0.24	0.19	0.24	0.22	0.12	0.30	0.28	0.44
HmRun	0.56	0.53	1.00	0.63	0.85	0.44	0.11	0.22	0.22	0.49	0.26	0.35	0.23	-0.16	-0.01	0.34
Runs	0.90	0.91	0.63	1.00	0.78	0.70	-0.01	0.17	0.19	0.23	0.24	0.20	0.16	0.18	0.19	0.42
RBI	0.80	0.79	0.85	0.78	1.00	0.57	0.13	0.28	0.29	0.44	0.31	0.39	0.23	0.06	0.15	0.45
Walks	0.62	0.59	0.44	0.70	0.57	1.00	0.13	0.27	0.27	0.35	0.33	0.31	0.43	0.10	0.08	0.44
Years	0.01	0.02	0.11	-0.01	0.13	0.13	1.00	0.92	0.90	0.72	0.88	0.86	0.84	-0.09	-0.16	0.40
CAtBat	0.21	0.21	0.22	0.17	0.28	0.27	0.92	1.00	1.00	0.80	0.98	0.95	0.91	-0.01	-0.07	0.53
CHits	0.23	0.24	0.22	0.19	0.29	0.27	0.90	1.00	1.00	0.79	0.98	0.95	0.89	-0.01	-0.07	0.55
CHmRun	0.21	0.19	0.49	0.23	0.44	0.35	0.72	0.80	0.79	1.00	0.83	0.93	0.81	-0.19	-0.17	0.52
CRuns	0.24	0.24	0.26	0.24	0.31	0.33	0.88	0.98	0.98	0.83	1.00	0.95	0.93	-0.04	-0.09	0.56
CRBI	0.22	0.22	0.35	0.20	0.39	0.31	0.86	0.95	0.95	0.93	0.95	1.00	0.89	-0.10	-0.12	0.57
CWalks	0.13	0.12	0.23	0.16	0.23	0.43	0.84	0.91	0.89	0.81	0.93	0.89	1.00	-0.07	-0.13	0.49
Assists	0.34	0.30	-0.16	0.18	0.06	0.10	-0.09	-0.01	-0.01	-0.19	-0.04	-0.10	-0.07	1.00	0.70	0.03
Errors	0.33	0.28	-0.01	0.19	0.15	0.08	-0.16	-0.07	-0.07	-0.17	-0.09	-0.12	-0.13	0.70	1.00	-0.01
Salary	0.39	0.44	0.34	0.42	0.45	0.44	0.40	0.53	0.55	0.52	0.56	0.57	0.49	0.03	-0.01	1.00

R results:

```
Call:
lm(formula = Salary ~ HmRun, data = mlb)

Residuals:
    Min      1Q  Median      3Q     Max
-748.73 -275.49  -79.27  184.72 1829.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  330.594     43.551   7.591 5.64e-13 ***
HmRun         17.671      2.995   5.900 1.13e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 424.6 on 261 degrees of freedom
Multiple R-squared:  0.1177,Adjusted R-squared:  0.1143
F-statistic: 34.81 on 1 and 261 DF,  p-value: 1.125e-08
```

```
Call:
lm(formula = Salary ~ HmRun + Errors, data = mlb)

Residuals:
    Min      1Q  Median      3Q     Max
-747.95 -276.62  -80.35  184.84 1829.63

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 331.8135    55.6398   5.964 8.02e-09 ***
HmRun        17.6699     3.0011   5.888 1.20e-08 ***
Errors       -0.1406     3.9780  -0.035    0.972
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 425.4 on 260 degrees of freedom
Multiple R-squared:  0.1177,Adjusted R-squared:  0.1109
F-statistic: 17.34 on 2 and 260 DF,  p-value: 8.548e-08
```

R results:

```
Call:
lm(formula = Salary ~ HmRun + Errors + RBI, data = mlb)

Residuals:
    Min      1Q  Median      3Q     Max 
-891.81 -244.35  -74.08  172.32 2021.19 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  147.511     61.623   2.394   0.0174 *  
HmRun         -9.687      5.559  -1.743   0.0826 .  
Errors        -6.895      3.937  -1.751   0.0811 .  
RBI           10.881      1.902   5.720 2.93e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 401.6 on 259 degrees of freedom
Multiple R-squared:  0.2166,Adjusted R-squared:  0.2076 
F-statistic: 23.87 on 3 and 259 DF,  p-value: 1.128e-13
```

```
Call:
lm(formula = Salary ~ HmRun + Errors + RBI + Assists, data = mlb)

Residuals:
    Min      1Q  Median      3Q     Max 
-876.98 -245.03  -71.01  174.30 1993.03 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 150.6908    61.8343   2.437   0.0155 *  
HmRun        -8.2976     5.8837  -1.410   0.1597    
Errors       -9.5253     5.3525  -1.780   0.0763 .  
RBI          10.5173     1.9687   5.342 2.02e-07 ***
Assists       0.1853     0.2552   0.726   0.4684    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 401.9 on 258 degrees of freedom
Multiple R-squared:  0.2182,Adjusted R-squared:  0.2061 
F-statistic: 18.01 on 4 and 258 DF,  p-value: 4.698e-13
```

R results:

```
Call:
lm(formula = Salary ~ ., data = mlb)

Residuals:
     Min       1Q   Median       3Q      Max 
-1045.42  -198.12   -46.87   130.82  1978.47 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 139.2160    85.7474   1.624 0.105745    
AtBat        -1.9485     0.6491  -3.002 0.002958 ** 
Hits          7.8422     2.4651   3.181 0.001654 ** 
HmRun         3.4940     6.3897   0.547 0.584999    
Runs         -2.9357     3.0791  -0.953 0.341299    
RBI          -0.2564     2.6782  -0.096 0.923816    
Walks         6.8261     1.8796   3.632 0.000342 ***
Years        -4.6778    12.7686  -0.366 0.714415    
CAtBat       -0.2115     0.1399  -1.512 0.131909    
CHits         0.2425     0.6950   0.349 0.727434    
CHmRun       -0.4994     1.6770  -0.298 0.766128    
CRuns         1.3952     0.7691   1.814 0.070902 .  
CRBI          0.9359     0.7175   1.304 0.193302    
CWalks       -0.7088     0.3404  -2.082 0.038332 *  
Assists       0.2441     0.2265   1.077 0.282339    
Errors       -1.3608     4.5180  -0.301 0.763515    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 328.3 on 247 degrees of freedom
Multiple R-squared:  0.5008,Adjusted R-squared:  0.4705 
F-statistic: 16.52 on 15 and 247 DF,  p-value: < 2.2e-16
```

Multicollinearity

- When two or more predictors are highly correlated, it is difficult to get the reliable estimates of their impact on y
- To put it simply, you cannot tease apart the net effect of each predictor
- Everything else being constant, the standard error associated with $\hat{\beta}$ will be inflated by the extent of the correlation with other predictors in the model
- This potential problem is known as **multicollinearity**

Multicollinearity

- It could be the reason why predictors that are believed to be key in predicting y do not result statistically significant when conducting hypothesis test
- Not a mistake in model specification, but rather an undesirable characteristic of data

Consequences:

- ⋄ High $se(\hat{\beta})$; low t-values; null hypothesis not rejected
- ⋄ Little influence on out-of-sample prediction accuracy
- ⋄ Not an issue in ML unless one is interested in interpreting a predictor's effect on y

Solution: remove the predictors causing high collinearity or merge the similar predictors into one

- ⋄ Feature selection (more broadly, regularization)
- ⋄ Principal Component Analysis (PCA)

Detecting multicollinearity

- Variance inflation factor (VIF): an index of how much $var(\hat{\beta})$ is inflated due to the correlation with other predictors
- VIF of the j-th predictor:

$$VIF_j = \frac{1}{1 - R_j^2}$$

  R_j^2 is the R^2 from a regression of x_j on other predictors

- $VIF_j > 10$ is evidence that the estimation of β_j is being substantially affected by multicollinearity

R code:

```
res.lm <- lm(Salary ~ ., data = mlb)
summary(res.lm)

install.packages("car")
library(car)
vif(res.lm)
```

R results:

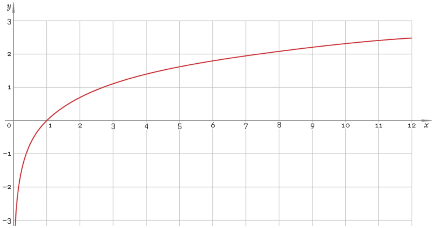
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 139.2160    85.7474   1.624 0.105745
AtBat        -1.9485     0.6491  -3.002 0.002958 **
Hits          7.8422     2.4651   3.181 0.001654 **
HmRun         3.4940     6.3897   0.547 0.584999
Runs         -2.9357     3.0791  -0.953 0.341299
RBI          -0.2564     2.6782  -0.096 0.923816
Walks         6.8261     1.8796   3.632 0.000342 ***
Years        -4.6778    12.7686  -0.366 0.714415
CAtBat       -0.2115     0.1399  -1.512 0.131909
CHits         0.2425     0.6950   0.349 0.727434
CHmRun       -0.4994     1.6770  -0.298 0.766128
CRuns         1.3952     0.7691   1.814 0.070902 .
CRBI          0.9359     0.7175   1.304 0.193302
CWalks       -0.7088     0.3404  -2.082 0.038332 *
Assists       0.2441     0.2265   1.077 0.282339
Errors       -1.3608     4.5180  -0.301 0.763515
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 328.3 on 247 degrees of freedom
Multiple R-squared:  0.5008,Adjusted R-squared:  0.4705
F-statistic: 16.52 on 15 and 247 DF,  p-value: < 2.2e-16

> vif(res.lm)
    AtBat       Hits      HmRun       Runs        RBI
22.224918  30.083441   7.612143  15.034975  11.681897
    Walks      Years     CAtBat      CHits     CHmRun
 4.051448   9.108170 248.889543 493.383447  46.196936
    CRuns       CRBI     CWalks    Assists     Errors
157.764854 130.860319  19.638823   2.625633   2.166027
```

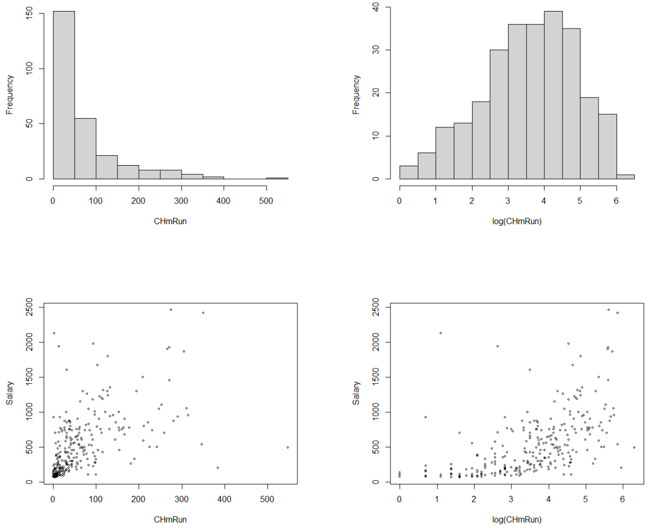
Log transformation

- Converts a skewed distribution to a normal distribution/less-skewed distribution by reducing scale
- Produces approximately equal spreads
- Often converts a curved relationship between predictor and outcome into a linear relationship

$\Rightarrow$ Improves model fit!!!

Log function




Frequency
CHmRun
Frequency
log(CHmRun)
Salary
CHmRun
Salary
log(CHmRun)

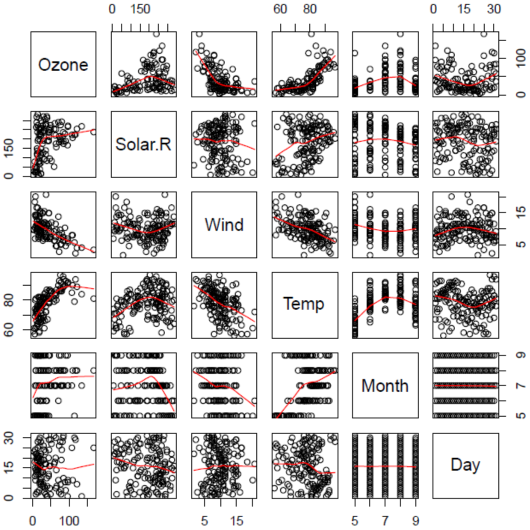
R code:

```
hist(mlb$CHmRun, main=NULL, xlab="CHmRun")
plot(mlb$CHmRun, mlb$Salary, cex=.6, ylab="Salary", xlab="CHmRun")
res.lm <- lm(Salary ~ CHmRun, data = mlb)
summary(res.lm)

mlb$log_CHmRun <- log(mlb$CHmRun+1)
hist(mlb$log_CHmRun, main=NULL, xlab="log(CHmRun)")
plot(mlb$log_CHmRun, mlb$Salary, cex=.6, ylab="Salary",
    xlab="log(CHmRun)")
res.lm <- lm(Salary ~ log_CHmRun, data = mlb)
summary(res.lm)
```

R results:

```
Call:
lm(formula = Salary ~ CHmRun, data = mlb)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 336.4512    31.0408  10.839   <2e-16 ***
CHmRun        2.8809     0.2891   9.964   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 384.7 on 261 degrees of freedom
Multiple R-squared:  0.2756,Adjusted R-squared:  0.2728
F-statistic: 99.27 on 1 and 261 DF,  p-value: < 2.2e-16


Call:
lm(formula = Salary ~ log_CHmRun, data = mlb)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -134.89      68.68  -1.964   0.0506 .
log_CHmRun    187.77      18.07  10.391   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 380.1 on 261 degrees of freedom
Multiple R-squared:  0.2926,Adjusted R-squared:  0.2899
F-statistic:   108 on 1 and 261 DF,  p-value: < 2.2e-16
```

What if there is a non-linear relationship?



Degree-n polynomial regression

- Non-linear relationship between x and y is modeled as an nth-degree polynomial in x.
- General form:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_n x_i^n + \varepsilon_i$$

e.g., 2nd-degree polynomial regression (quadratic regression):

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

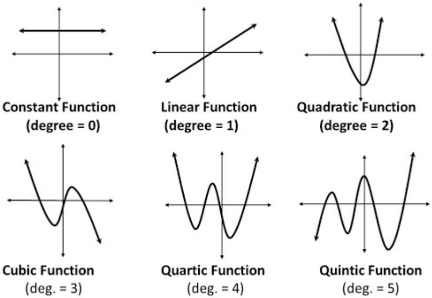
e.g., 3rd-degree polynomial regression (cubic regression):

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$$

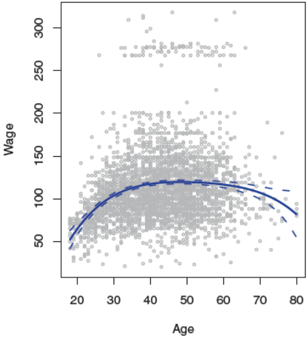
e.g., 4th-degree polynomial regression (quartic regression):

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_i$$

Why polynomials?



Wage against age using the 4th-degree polynomial regression



$$\text{Wage}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{age}_i^2 + \beta_3 \text{age}_i^3 + \beta_4 \text{age}_i^4 + \varepsilon_i$$

Things to consider for polynomial regression

- Be cautious about how many polynomial terms to include... (not too many)
- Test significance of the coefficient estimate on the highest-degree term; if the null is not rejected, go for a polynomial regression at lower degree (why?)
- Multicollinearity not a problem
- Does not yield intuitive interpretation (if $n \geq 3$)
- May not be the best model for a small sample...

Feature selection (or, variable selection) is intended to select the "best" subset of predictors...
But why bother?

- Increased interpretability of model
- Reduced training time
- Principle of parsimony
    - Unnecessary explanatory variables only add noise to the estimation of other quantities
    - Degrees of freedom are wasted
- Avoid overfitting

Prior to feature selection:

- Identify outliers and address them using a suitable transformation (e.g., log transformation)
- Scale the variables (if needed)

Backward elimination

a Estimate a fully specified model (a model that includes all predictors)

b Remove one explanatory variable with the highest p-value, greater than α_{crit}

c Re-estimate the model and goto b

d Stop when all p-values are less than α_{crit}

$\Rightarrow$ If prediction is the goal, then about 15-30% cut-off may work the best

Backward elimination: example ($\alpha_{crit} = 0.25$)

```
library(ISLR)
mlb <- data.frame(Hitters)

mlb <- mlb[, c(-14:-16, -20)]
mlb <- na.omit(mlb)

res.lm <- lm(Salary ~ ., data = mlb)
summary(res.lm)
res.lm <- lm(Salary ~ . -RBI, data = mlb)
summary(res.lm)
res.lm <- lm(Salary ~ . -RBI -CHmRun, data = mlb)
summary(res.lm)
res.lm <- lm(Salary ~ . -RBI -CHmRun -Errors, data = mlb)
summary(res.lm)
res.lm <- lm(Salary ~ . -RBI -CHmRun -Errors -Years, data = mlb)
summary(res.lm)
res.lm <- lm(Salary ~ . -RBI -CHmRun -Errors -Years -HmRun, data = mlb)
summary(res.lm)
res.lm <- lm(Salary ~ . -RBI -CHmRun -Errors -Years -HmRun -Runs, data = mlb)
summary(res.lm)
res.lm <- lm(Salary ~ . -RBI -CHmRun -Errors -Years -HmRun -Runs -CHits, data = mlb)
summary(res.lm)
```

R results:

```
Call:
lm(formula = Salary ~ ., data = mlb)

Residuals:
     Min       1Q   Median       3Q      Max
-1045.42  -198.12   -46.87   130.82  1978.47

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 139.2160    85.7474   1.624 0.105745
AtBat        -1.9485     0.6491  -3.002 0.002958 **
Hits          7.8422     2.4651   3.181 0.001654 **
HmRun         3.4940     6.3897   0.547 0.584999
Runs         -2.9357     3.0791  -0.953 0.341299
RBI          -0.2564     2.6782  -0.096 0.923816
Walks         6.8261     1.8796   3.632 0.000342 ***
Years        -4.6778    12.7686  -0.366 0.714415
CAtBat       -0.2115     0.1399  -1.512 0.131909
CHits         0.2425     0.6950   0.349 0.727434
CHmRun       -0.4994     1.6770  -0.298 0.766128
CRuns         1.3952     0.7691   1.814 0.070902 .
CRBI          0.9359     0.7175   1.304 0.193302
CWalks       -0.7088     0.3404  -2.082 0.038332 *
Assists       0.2441     0.2265   1.077 0.282339
Errors       -1.3608     4.5180  -0.301 0.763515
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 328.3 on 247 degrees of freedom
Multiple R-squared:  0.5008,Adjusted R-squared:  0.4705
F-statistic: 16.52 on 15 and 247 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Salary ~ . - RBI, data = mlb)

Residuals:
     Min       1Q   Median       3Q      Max
-1044.17  -197.69   -47.43   131.12  1978.49

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 139.6581    85.4517   1.634 0.103454
AtBat        -1.9548     0.6443  -3.034 0.002671 **
Hits          7.7725     2.3503   3.307 0.001083 **
HmRun         3.0243     4.0845   0.740 0.459743
Runs         -2.8669     2.9880  -0.959 0.338255
Walks         6.7908     1.8395   3.692 0.000274 ***
Years        -4.7295    12.7317  -0.371 0.710603
CAtBat       -0.2112     0.1396  -1.513 0.131599
CHits         0.2525     0.6857   0.368 0.712960
CHmRun       -0.4513     1.5970  -0.283 0.777707
CRuns         1.3851     0.7604   1.822 0.069729 .
CRBI          0.9094     0.6609   1.376 0.170018
CWalks       -0.7044     0.3365  -2.093 0.037364 *
Assists       0.2443     0.2261   1.080 0.280970
Errors       -1.3889     4.4995  -0.309 0.757826
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 327.6 on 248 degrees of freedom
Multiple R-squared:  0.5008,Adjusted R-squared:  0.4726
F-statistic: 17.77 on 14 and 248 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Salary ~ . - RBI - CHmRun, data = mlb)

Residuals:
     Min       1Q   Median       3Q      Max
-1065.51  -193.07   -47.03   131.64  1977.66

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 139.3269    85.2856   1.634 0.103596
AtBat        -1.9480     0.6427  -3.031 0.002695 **
Hits          7.6796     2.3230   3.306 0.001086 **
HmRun         2.6079     3.8026   0.686 0.493454
Runs         -2.6315     2.8643  -0.919 0.359120
Walks         6.7507     1.8306   3.688 0.000278 ***
Years        -4.5117    12.6848  -0.356 0.722383
CAtBat       -0.2209     0.1351  -1.635 0.103270
CHits         0.3855     0.4978   0.774 0.439426
CRuns         1.2429     0.5689   2.185 0.029848 *
CRBI          0.7362     0.2468   2.983 0.003142 **
CWalks       -0.6798     0.3245  -2.095 0.037204 *
Assists       0.2497     0.2248   1.111 0.267715
Errors       -1.4319     4.4886  -0.319 0.749989
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 327 on 249 degrees of freedom
Multiple R-squared:  0.5006,Adjusted R-squared:  0.4745
F-statistic:  19.2 on 13 and 249 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Salary ~ . - RBI - CHmRun - Errors, data = mlb)

Residuals:
     Min       1Q   Median       3Q      Max
-1065.03  -192.51   -47.14   133.61  1984.78

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 133.2771    83.0009   1.606 0.109595
AtBat        -1.9753     0.6358  -3.107 0.002111 **
Hits          7.7739     2.2999   3.380 0.000841 ***
HmRun         2.4845     3.7760   0.658 0.511168
Runs         -2.6618     2.8576  -0.932 0.352492
Walks         6.7798     1.8250   3.715 0.000251 ***
Years        -4.2282    12.6309  -0.335 0.738096
CAtBat       -0.2179     0.1345  -1.620 0.106505
CHits         0.3654     0.4929   0.741 0.459243
CRuns         1.2634     0.5642   2.239 0.026024 *
CRBI          0.7363     0.2464   2.988 0.003087 **
CWalks       -0.6840     0.3237  -2.113 0.035597 *
Assists       0.2050     0.1755   1.168 0.243848
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 326.4 on 250 degrees of freedom
Multiple R-squared:  0.5004,Adjusted R-squared:  0.4764
F-statistic: 20.87 on 12 and 250 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Salary ~ . - RBI - CHmRun - Errors - Years, data = mlb)

Residuals:
     Min       1Q   Median       3Q      Max
-1061.24  -191.19   -48.26   131.72  1990.64

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 117.0817    67.3226   1.739 0.083241 .
AtBat        -1.9362     0.6239  -3.103 0.002133 **
Hits          7.6854     2.2807   3.370 0.000871 ***
HmRun         2.4811     3.7693   0.658 0.510981
Runs         -2.6267     2.8506  -0.921 0.357694
Walks         6.7740     1.8217   3.718 0.000247 ***
CAtBat       -0.2390     0.1187  -2.013 0.045132 *
CHits         0.3982     0.4822   0.826 0.409621
CRuns         1.2868     0.5589   2.303 0.022124 *
CRBI          0.7383     0.2459   3.003 0.002946 **
CWalks       -0.6854     0.3231  -2.121 0.034882 *
Assists       0.2123     0.1738   1.221 0.223065
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 325.8 on 251 degrees of freedom
Multiple R-squared:  0.5002,Adjusted R-squared:  0.4783
F-statistic: 22.83 on 11 and 251 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Salary ~ . - RBI - CHmRun - Errors - Years - HmRun,
    data = mlb)

Residuals:
    Min      1Q  Median      3Q     Max
-1090.2  -192.8   -53.3   132.5  1998.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 114.6923    67.1491   1.708 0.088863 .
AtBat        -1.8600     0.6124  -3.037 0.002639 **
Hits          7.4479     2.2494   3.311 0.001066 **
Runs         -1.9798     2.6728  -0.741 0.459550
Walks         6.5845     1.7968   3.665 0.000302 ***
CAtBat       -0.2380     0.1185  -2.008 0.045748 *
CHits         0.3731     0.4801   0.777 0.437855
CRuns         1.2628     0.5571   2.267 0.024247 *
CRBI          0.8122     0.2186   3.716 0.000249 ***
CWalks       -0.6771     0.3225  -2.100 0.036760 *
Assists       0.1800     0.1666   1.081 0.280892
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 325.5 on 252 degrees of freedom
Multiple R-squared:  0.4993,Adjusted R-squared:  0.4794
F-statistic: 25.13 on 10 and 252 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Salary ~ . - RBI - CHmRun - Errors - Years - HmRun -
    Runs, data = mlb)

Residuals:
     Min       1Q   Median       3Q      Max
-1064.25  -190.10   -50.53   125.74  1989.48

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 119.2438    66.8077   1.785 0.075479 .
AtBat        -1.8602     0.6119  -3.040 0.002612 **
Hits          6.5516     1.8945   3.458 0.000638 ***
Walks         6.0713     1.6564   3.665 0.000301 ***
CAtBat       -0.2525     0.1168  -2.161 0.031608 *
CHits         0.5074     0.4442   1.142 0.254409
CRuns         1.0575     0.4828   2.190 0.029419 *
CRBI          0.8192     0.2182   3.755 0.000215 ***
CWalks       -0.6193     0.3126  -1.981 0.048681 *
Assists       0.2066     0.1625   1.271 0.204887
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 325.2 on 253 degrees of freedom
Multiple R-squared:  0.4982,Adjusted R-squared:  0.4804
F-statistic: 27.91 on 9 and 253 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = Salary ~ . - RBI - CHmRun - Errors - Years - HmRun -
    Runs - CHits, data = mlb)

Residuals:
     Min       1Q   Median       3Q      Max
-1129.26  -183.77   -42.84   120.66  1994.69

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 117.03909   66.81986   1.752 0.081056 .
AtBat        -2.15294    0.55594  -3.873 0.000137 ***
Hits          7.48931    1.70846   4.384 1.71e-05 ***
Walks         6.27913    1.64735   3.812 0.000173 ***
CAtBat       -0.13623    0.05738  -2.374 0.018339 *
CRuns         1.36286    0.40236   3.387 0.000818 ***
CRBI          0.82517    0.21823   3.781 0.000195 ***
CWalks       -0.78973    0.27490  -2.873 0.004412 **
Assists       0.20893    0.16261   1.285 0.200017
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 325.4 on 254 degrees of freedom
Multiple R-squared:  0.4956,Adjusted R-squared:  0.4797
F-statistic:  31.2 on 8 and 254 DF,  p-value: < 2.2e-16
```