# Sabermetric Manifesto

## Intro

* **Sabermetrics** = “the search for objective knowledge about baseball”
  + attempts to answer objective questions about baseball
    - "Which player on Red Sox contributed most to team’s offense?”, “How many HR's will Ken Griffey hit next year?”
  + cannot deal w/ subjective judgments (also important to the game):
    - “Who is your favorite player?” or “That was a great game.”
* **Statistics =**the best objective record of the game available
  + large part of Sabermetrics involves understanding \*\*\***how to use statistics\*\*\*** + which statistics are useful for what purposes, etc.
  + do NOT need to know a lot about math to understand sabermetrics, only some \*\*\***idea of how statistics can be used and misused\*\*\***
* Since Sabermetrics = an ***objective*** study of the game, necessary to \*\*\***use logical reasoning in arguments\*\*\***
* **Hypothesis** can be developed from info you have, either from statistics or observations;
  + a claim which cannot be *directly* tested can be evaluated by studying conclusions which would follow.
  + \*Ex:\* “Pitching is X% of baseball,” which has been said with X between 15-80%.
    - want to test the claim “Pitching is 75% of baseball.”
    - If true, we'd conclude teams w/ best pitching = much more likely to win the pennant than teams w/ best hitting.
    - **However, this isn’t the case**.
    - League leaders in fewest RA (which is both pitching *and* fielding) win pennant about 1/2 the time + league leaders in RS (includes all of hitting) win just as often.
      * NOTE the definition of **offense** here = if measuring hitting by an **incomplete measure such as BA,**you'd conclude that pitching is much more important
    - Other unreasonable conclusions: team w/ 75% of its value in pitching would never trade a regular pitcher for a regular hitter, thus the claim must be rejected.
    - But if 75% is replaced by a number close to 40%, conclusions become reasonable.

## General Principles

* goal of a baseball team = win more games than any other team.
* 1 team has **very little control over # of games other teams win**, so **goal = win as many as possible**
* Therefore, it is of interest to measure player contribution to the team wins
* Clear relationship between team's RS and RA allowed + its W/L  (not perfect, but very strong.)
* Good formula, determined empirically from data by Bill James, is a team’s W/L ratio = square of the ratio between RS and RA
  + Thus team scoring + allowing same # of runs = win + lose same # of games, finishing @ .500
  + team which scores 800 runs + allows 700 wins 64 games for every 49 it loses, which projects to a 92-70 record over a season (comes very close to actual records of most teams)
* **Basic goal of sabermetrics = evaluate a measure for a given purpose.**
  + **\*\*\*most common uses of statistics\*\*\* = evaluate past performance** (determine who should win MVP ) + **to predict future performance** (evaluate a trade that was just made).
  + In both cases, interested in **measuring contribution to games won and lost**.
* Baseball statistics can measure individual performance, ***independent of what other players do***
  + While importance of an individual event depends on situation, the effect of the situations on the importance of the statistic over a large sample such as a season is not great.
    - When a batter hits a single, this describes what *he* did; when a quarterback throws a ten-yard pass, the guard who took out a linebacker gets no statistical credit.
    - batter who received a single is properly credited for a success; the 10-yd. pass may have been a failure if it was 3rd down w/ 13 yards to go
    - Thus it is reasonable for the **goal of a baseball statistic to be to measure a player’s individual contribution to runs or wins.**
* Given goal, it's possible to **evaluate a statistic**.
* \*\*\*Baseball statistics can be evaluated, have same types of flaws, + be misused/misinterpreted in the same ways as *non*-baseball statistics\*\*\*
* 1st natural question to ask about a statistic = "*Does the statistic measure an important contribution to that goal*?”
  + \*Ex:\* ERA = # of runs a pitcher allows == *almost all a pitcher contributes to winning games*
  + BA does fairly well b/c it counts hits, but it ignores power + walks, which are also important parts of offense.
  + Few statistics fail badly here:
    - those which measure things happen rarely (like HBP)
    - those having little to do w/ winning games (ex: % of a batter’s outs that are SO's)
    - or both
    - Non-baseball \*Ex\*: # of crimes in a *city* last year = important if you want to know something about safety of city; # of crimes on a *single street* says very little about safety of whole city.
* 2nd + usually most important, question to ask is, “**How well does the statistic measure the player’s own contribution**?”
  + Many ways a statistic, baseball-related or not, can fail here.
  + \*\*\***Virtually every statistic fails in some way to some extent**\*\*\*, so the \*\*\***best statistics are those w/ only minor failings + relatively few of them\*\*\***
  + \*Ex:\* Player should be evaluated for what *HE*does, not for what teammates or manager do = a major problem w/ such statistics as RS
    - Unless a batter hits a HR or steals home, he needs teammate contributions to actually score, + cannot do much to cause them to get hits once on base.
    - Thus, batting in front of best HR hitters = score a lot, whether or not you have a good ability to score runs.
    - If batting you 8th on an NL team, won’t score many runs when you do get on base.
* **Good statistic should not measure *outside* effects over which player has no control (e.g. the park)**
  + Good non-baseball \*Ex:\* = high death rate in Miami.
    - population of Miami = older than population of most other cities;
    - thus, regardless of quality of medical care in Miami, we expect a high death rate.
  + Likewise: easier to score in Fenway than in Oakland
    - Therefore, pitcher w/ 3.60 ERA in Oakland could pitch just as well in Fenway, helping his team win games just as much, but have a 4.00 ERA
  + Will sometimes see "park-adjusted numbers" = designed to eliminate this effect;
    - Ex: pitcher above might have a 3.80 park-adjusted ERA in either park.
      * Note: this is adjusting for value of the pitcher’s performance, not the actual performance;
      * 4.00 ERA for a Red Sox pitcher is just as valuable to his team regardless of how it is split between home + road games.
* **\*\*\*If a player’s statistics change considerably when changing teams, parks, or lineup positions, this suggests the outside effect has a major effect on the statistics.\*\*\***
* **\*\*\*If the statistic remains consistent when outside conditions change, this means it is measuring the player’s own contribution\*\*\***
  + Pitchers w/ good ERA’s tend to keep them when changing teams, so park effect is not a serious problem.
  + Hitters who score a lot in the leadoff spot score many fewer runs if dropped to 6th in the lineup, which means runs scored were mostly created by lineup position rather than the batter.
* **In addition to problems w/ outside effects, there can be problems with measurement.**
  + **\*\*\*No statistic can be useful w/out proper context/a measure of opportunities\*\*\*.**
  + There were more crimes committed in NY than in Boston last year, but this doesn’t say much about relative safety of the cities;
    - * to make such a comparison, you would need to compare crime ***rates***.
  + If a batter has 150 hits, ***what does that mean***? => If 500 AB's, he is good at getting hits; if 650 AB's, he is poor.
* **\*\*\*This is a problem w/ most counting statistics\*\*\***
  + BA places hits in a reasonable context, + this is recognized b/c batting title goes to player w/ the highest BA, NOT player w/ most hits.
* Similarly**, \*\*\*a statistic may not be useful if it tries to measure something with a very small sample size or number of occurrences\*\*\***
  + best pitchers @ throwing shutouts often don’t lead league, b/c league leader normally has ~5, + it’s quite common for a pitcher who usually throws 3 shutouts a year to get 7 in 1 year.
  + In contrast, best SO pitchers DO lead league in SO's (or SO's per 9 innings), b/c their totals are in the hundreds, + a pitcher capable of getting 250 strikeouts in 240 innings might get 230, but not 150.
* Same problem comes up w/ non-baseball statistics.
  + If 2/3 of people polled in a city plan to vote Democratic, it means nothing if it was 4/6 + not much if 40/60, but quite a lot if 400/600.
  + **\*\*\*This is the major flaw w/ many statistics often used on TV\*\*\***
    - statistic such as, “Wade Boggs is hitting .154 against Baltimore pitchers w/ runners in scoring position” means nothing b/c sample is probably 2 hits in 13 AB's
  + Sabermetricians agree w/ most fans that such stats are ridiculous + are there only to hold interest of (mostly statistically illiterate) TV audience
* Once you have some idea of how well a statistic measures player’s own contribution to the goal, the \*\*\*final question to ask is, “**Is there a better way to measure the same thing**?”\*\*\*
  + *A statistic which has problems w/ the other questions but has no reasonable alternative measurement may still be useful.*
  + In contrast, a statistic such as RS, which can be *replaced by other statistics = very little value.*
  + A player’s own contribution to his total RS can be measured by his ability to get on base (already measured very well by OBP), +, to a lesser extent, to advance himself once he gets on base (measured by extra-base hits, + by stolen bases + caught stealing).
* Now, given these criteria, you can **evaluate a statistical conclusion**.
  + If you dispute the conclusion, your argument may be valid if based on these criteria
    - i.e. need to find something NOT measured by the statistic, or IS measured but *shouldn’t be*
    - \*Ex:\* can argue Mike Schmidt is a good hitter, even though his career average = .267, b/c he hit 548 HR + drew 1507 walks.
    - These are valid arguments, b/c **BA gives the same value to homers + singles**, + *does not count walks at all.*
    - Likewise, Ozzie Smith is *not a great offensive* player, but is *still an excellent* player, b/c of his *defense* (**no *offensive* statistic measures his *overall* value)**
* But you *CANNOT* dispute a statistical conclusion w/ a claim based on something that is *already included in the statistic*, or something which is *improperly measured* by your claim.
  + *NOT* reasonable to say Brooks Robinson was great at getting hits b/c of his 2848 hits;
  + *Correct* measure of how well he got hits = his .267 BA, which led to such a high hit total b/c his other skills allowed him to have a very long career
* **Turning 1 of the above examples around, you CAN'T claim Schmidt could not possibly be a great hitter, despite his .527 SLG, by looking at his BA, as BA is already counted in the slugging average.**

## Sabermetric stats

* A good, *complete* measure of individual offense would satisfy criteria above for a valuable statistic better than any of the traditional offensive measures.
* Therefore, sabermetricians often use/develop such statistics.
  + For measuring pitching: less need for such a statistic, b/c ERA + RA already count # of RA *by a pitcher*
* At team level, a **good measure of offense should have a strong correlation with RS**
  + i.e. should be possible to predict RS reasonably well from such a measure;
    - best teams by this measure should score a lot, while worst teams should score little
  + Measures such as BA do NOT do this
    - common for teams w/ best BA to be below average in RS.
  + RS *itself* obviously measures *team* offense very well, but creates a problem when trying to measure *individual* contributions;
  + it isn’t easy to measure directly how much a batter helped/hurt his team score.
* Several ways to develop a statistic which measures team offense.
  + Probably most natural way = say " score by getting runners on base *+ then advancing them"*
  + **Thus, a team’s RS should be *proportional* to # of runners it gets on base AND to the frequency w/ which it advances the runners.**
* OBP measures # of runners on base, while SLG (slugging average) =1 way to measure advancement
  + Note: an out reduces SLG, b/c it makes it less likely that any runners on base will be advanced
* Thus team **RS should be correlated w/ OBP\*SLG.**
* The \*\*\***test** of a statistic of this type **= how well it** **agrees with reality**\*\*\*
  + If you compare teams OBP\*SLG to their RS, you find a very good correlation w/ a standard error of just 24 runs.
  + For comparison, SD of RS in 1 season = ~70 runs (error returned if you predicted that all teams would be average in RS)
  + Meanwhile, BA alone has a standard error of 54 runs.
  + **24-run Std. Error covers everything OBP\*SLG does NOT measure/measures improperly**
    - includes such factors as baserunning + imperfections in the formula, but **much of the difference is chance.**
* Now, need to make an individual statistic by measuring a player’s contribution;
* OBP\*SLG is NOT the correct measure for a player b/c *he usually doesn’t drive himself in.*
* **Instead, want to multiply *his* OBP by the *team’s* SLG, + *his* SLG by *team’s* OBP.**
  + Since league (+ individual teams’) SLG are usually ~1.2 times OBP, **each point of a player’s OBP has 1.2 times the effect on OBP\*SLG that a point of his SLG has**.
  + Thus our measure **= (1.2\*OBP)+SLG.**
  + For simplicity, we often ignore factor of 1.2 + refer to **OPS, On-base Plus Slugging**
* When using **OPS**, remember that **OBP is slightly *undervalued*** + that SB's have NOT been counted.
* Using same process for other models of offense gives other measures, which give slightly different values for different elements of offense.
* The **choice of which measure** to use **depends** on which ones you **have handy**, the **purpose** for which you want **to use** it, + some **personal preferences**.
* But if you use any well-designed measure of offense, you won’t be wrong.
  + May find that a player w/ 2 more **Runs Created** than another is 003 worse in OPS, but such differences aren’t important
  + either way, you will reach the reasonable conclusion that they are very close.
* The *complete* measures of offense give a good estimate of the *value of the individual categories,* such as walks, HR’s, + outs, which make them up.
* The value of a player’s HR’s = the effect they have on OPS, or any similar statistic, + the importance of HR’s thus depends on this value + their frequency.

## Evaluating official statistics

* We can now apply the criteria to the official statistics
* Not reasonable to go through the arguments for *every* statistic, but useful to look @ statistics which cause the most frequent arguments.
* **RBI =** commonly used as a measure of a player’s offense, b/c they’re the only statistics which are **easily available** which look **like a complete measure**
  + *As a result, MVP is more likely to be league leader in RBI than in any other category*
* Of course, they are NOT a complete measure 🡺 ability to drive in runs is an *important* part of offense, but **not the whole thing**.
  + **This does NOT make RBI’s meaningless, only incomplete**.
* Real problem w/ RBI’s = 2nd question (**How well does statistic measure player’s own contribution)**
  + RBI’s measure a lot of things which are NOT the player’s *own* contribution.
  + CANNOT drive in runners who are *not on base* (except w/ HR’s), but *your own batting doesn’t put them there*;
    - bat behind good players = get a lot of chances.
  + In fact, league leaders in RBI = much more likely to be players **who batted w/ the most teammates on base or in scoring position** (*not the batter’s contribution*) + NOT those who hit best w/ runners on base or in scoring position.
  + **Thus RBI = a better measure of who had the most chances to drive in runners than of who was the best at driving in runners.**
* 3rd test = **Is there a better way to measure the same thing**
  + There IS a better measure of ability to drive in runners
  + **Hits** drive runners in from scoring position + therefore, a player w/ many hits = good at this part of driving runners in.
  + Likewise, **extra-base hits** drive runners in from 1st base, + HR’s drive in from home plate.
  + ***Slugging average (SLG) =*** *player’s ability to get hits, extra-base hits, + HR’s, so it measures his ability to drive in runs, w/ park effects = the only significant bias.*
  + **Thus, RBI’s = *NOT* as useful a measure of offense or a measure of ability to drive in runs**
* Other statistic subject to many of the same problems = pitcher’s W-L record (compare it to ERA)
  + Both measure something which is *clearly important*, since a pitcher’s goal = win games
  + the way he does this = preventing opponents from scoring.
  + *But both have some problems measuring the pitcher’s own contribution*
  + a comparison of their value depends on these problems.
  + 1st problem = runs are allowed by the WHOLE defense, not JUST by the pitcher
    - slightly more of a problem with W-L, as ERA eliminates runs due to **errors**, but NOT due to fielders that’re out of position, run slowly, or make weak throws.
  + At the MLB-level it isn’t a serious problem
    - good pitchers can still have good ERA’s (+ RA) even w/ teams of poor fielders.
  + **W-L record is one of the few categories which is immune to park effects**
    - there is 1 win in every game in every park.
    - ERA has slight problem w/ park effects 🡺 makes it more useful w/ park adjustment
  + **Most important factor = effect of the team offense.**
    - **\*\*\*Offense has almost no effect on ERA, but has a considerable effect on W-L\*\*\***.
    - A game is not won just by the pitcher (despite name of thestatistic), but by the team which scores more than allows.
    - In a single season, pitcher w/ best W-L record in the league = just as likely to be pitcher w/ best run support as the pitcher w/ fewest RA.
    - The run support is NOT pitcher’s contribution (except for batting in NL).
    - If there were pitchers who could cause teammates to score more for them, it would make sense to give the pitchers some of the credit.
    - But there is no tendency for pitchers who had support better than their team’s average in 1 season to have it again in the following season.
    - Nor does a pitcher have any control over whether he gets to pitch on a good offensive team
  + B/c of effect of run support, single-season WL = NOT a good measure of pitcher’s own value.
  + **ERA is available + is a better measure of what you actually want to know**.
  + **However, a career W-L reduces luck in run support by using a much larger sample size**.
  + In addition, pitchers rarely spend full careers w/ poor or good teammates.
  + Thus, a career W-L for a long career (several hundred decisions) = decent measure of pitcher’s own performance
    - it’s about as useful as a career ERA without park adjustments.
* Have now dealt w/ most common measures of batting + pitching + makes sense to now deal w/ most common measure of fielding.
* **Fielding average** = problem w/ 1st test (*Does statistic measure important contribution to that goal)*
  + While *defense is important*, an *incomplete measure of defense is NOT*.
* League leader in errors @ third usually makes about 30; leader in fielding average makes about 10.
  + These aren’t enough plays to make a difference of very many runs.
* More important part of fielding = **ability to prevent hits;**
* If 3rd baseman can’t reach a ball in the hole, or knocks it down but has no play, he won’t be charged w/ an error, *but the batter will get a hit which has the SAME effect.*
* **Errors are about as useful as a measure of defense as SO’s are as a measure of batting average.** 
  + They measure ONE way to fail to make a play;
  + while it is the most obvious failure, **all failures count the same on the scoreboard**.
* Fielder w/ poor range = a poor fielder whether he makes few or many errors, just as a hitter who hits too many grounders or popups can be a poor hitter even though he puts ball in play.
* While fielding average also has problems w/ park effects + scorer’s biases, the \*\*\***incompleteness is the most serious problem**\*\*\*
* Still, since it DOES measure SOMETHING useful, + fielders who are good at other things tend not to make errors (fielding % has a good correlation w/ wins), it would be a useful measure **in the absence of anything else**.
* So, it still has *some* value, particularly in concluding that players w/ very low fielding averages can’t handle their positions, but *it should be used in conjunction with putouts, assists, and an attempt to understand any biases in the numbers.*
* For recent players, better measure of overall defense = **Defensive Average (DA) =** makes fielding average unnecessary.
  + Basis for DA = division of playing field into zones of responsibility for fielders.
  + When a ball is hit into a fielder’s zone, it is **charged as an opportunity for that fielder**
    - if fielder turns it into an out, he receives credit for a play made.
      * Thus, all ground balls near 3rd are charged as chances for 3rd baseman;
        + a good 3rd baseman will make plays on most of them.
    - If he *fails* to make a play, effect is the same whether his throw is wild (error) or late (scored a single)
  + **fielding average does not tell you anything more.**
  + **Defensive average should be put to the same tests as any other statistic**
    - does reasonably well in the 1st test 🡺 measures a player’s ability to turn balls in play into outs, which covers *most* of his defensive play but *not all*
      * such skills as turning a double play + throwing out runners trying to stretch hits are NOT counted
    - does well in 2nd test (although still has some problems, mostly w/ park effects)
      * Pitchers cannot introduce bias simply by being left-handed (thus allowing a lot of ground balls to 3rd base + fly balls to left)
      * Good pitchers *may* help fielders’ DA *slightly* by allowing fewer hard-hit balls
      * Fielders do NOT have a great effect *on each other’s DA*, although there will be a *small* effect for plays such as low throws a good 1st baseman can handle
      * **All these effects will cause problems w/ almost any measure of fielding**
    - For 3rd test, DA = best measure of ability to make a play in the field *that we have*
      * isn’t perfect, but is *complete enough + accurate enough* to be useful.
* Thus the established statistics, used for reasons of tradition, may be *good* measures (ERA) or poor measures (RBI’s)
* \*\*\*Their value does NOT depend on *tradition* or *names*, but **depends on how well they meet the basic tests of any statistic**\*\*\*

## Other sabermetric arguments

* **Similar analysis must also be used in evaluating a hypothesis which depends on a statistical argument**
* **If a hypothesis leads to conclusions which don’t correspond w/ real game of baseball, it needs to be revised.**
  + Ex: natural question in predicting a player’s future performance in the major league = how useful minor-league #’s will be in a prediction.
  + problems w/ using minor-league #’s b/c there are *extreme* park effects + differences between leagues.
  + However, once you \*\*\*adjust a player’s minor-league #’s for these effects, + then make a specific adjustment for the difference between AA/AAA ball vs. majors, you *may* have something meaningful\*\*\*
  + There IS a method for making these corrections = **Minor-League Equivalency (MLE)**
    - will be useful if it works when tested against real world.
    - In fact, works almost as well as past major-league performance in predicting future major-league performance.
    - Most players w/ MLE’s which say they will hit .300 will hit close to .300 as rookies, just as most players who hit .300 last year will.
      * **Of course, neither prediction is perfect**
* Another issue sabermetricians have studied + discussed = **existence of clutch hitters**.
  + Clutch hits themselves certainly exit, but many players have reputations as players who will hit best w/ the game on the line
  + this is a hypothesis which CAN be tested; *are there any players with such an ability?*
  + Again, it’s necessary to **look at what actually happens, + what would happen if there were no clutch ability at all or if clutch hitting was a significant ability.**
  + Even if a .250 hitter were just a pair of coinswhich got a hit when both were heads, some .250 hitters would hit.400 during 1 season in the late innings of close games (3% chancein 80 AB)
  + So, the existence of such #’s doesn’t prove anything.
  + Butif there IS an ability, players who hit well in clutch in the past will continue to do so.
  + This CAN be tested, + HAS been == only*very weak evidence*, + but is clear that whatever abilitythere IS *does not mean much in baseball terms*.
  + There may be .267hitters who are actually as valuable as .268 hitters b/c of good clutch #’s
  + **But if replacing .268 w/ .275 = a conclusion inconsistent w/ what actually happens**

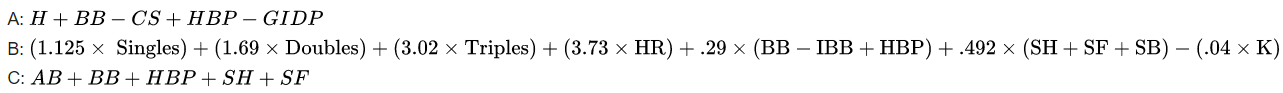
## VI. Conclusion

* Baseball statistics = useful **only if they enhance understanding of the game**.
* Therefore, they **should be judged by how well they measure what actually happens in the game**.
* **Meaningless statistics should be ignored or replaced + deficient statistics improved**.
* **Well-designed statistics should be used as an important part of discussion about the game + players**

# Runs Created Primer

* **Runs created (RC)** = invented by Bill James to **estimate the # of runs a hitter contributes to his team**
* Why James believes RC = an essential thing to measure:
  + W/ regard to an offensive player, 1st key question = *how many runs have resulted from what he’s done w/ the bat + on the basepaths.*
  + Willie McCovey hit .270 in his career, w/ 353 doubles, 46 triples, 521 HR’s, + 1,345 walks -- but his job was not to hit doubles, nor singles, nor triples, nor draw walks or even hit HR’s, but to put runs on the scoreboard.
  + *How many runs resulted from all of these things?*
* RC attempts to answer this question.
* Conceptual framework of RC 🡺 RC = (A\*B)/C
  + Where **A = On-base factor, B = Advancement factor, C = Opportunity factor**
* Most basic RC formula: **RC =** H = hits, BB = base on balls, TB = total bases, AB = at-bats.
  + can also be expressed as **OBP × SLG × AB** or **OBP × TB**
    - where OBP = on-base %, SLG = slugging average, AB = at-bats, TB = total bases.
* "Stolen base" version of RC = 
  + expands on basic formula by accounting for a player's base-stealing ability.
* "Technical" version of RC = 
  + accounts for all basic, easily available offensive statistics.
  + HBP = hit by pitch, GIDP = grounded into double play, IBB = intentional base on balls,

SH = sacrifice hit, SF = sacrifice fly

* *Earlier versions* of RC *overestimated* # of runs created by players w/ extremely high on-base and slugging factors (Babe Ruth, Ted Williams, Barry Bonds)
  + This is b/c these formulas *placed a player in an offensive context of players equal to himself;* 
    - As if player is assumed to be on base for *himself* when he hits HR’s (impossible)
    - In reality, a great player is interacting w/ offensive players whose contributions are *inferior to his*.
    - 2002 version corrects this by placing the *player in the context of his real-life team +* also takes into account performance in clutch situations.
* 2002 VERSION: 
* The makes the initial individual RC estimate = 
  + If **situational hitting info** is available, the following should be added to the above total:



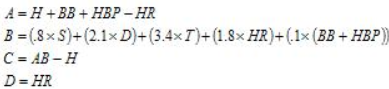
* + - RISP = runners in scoring position, ROB = runners on base.
    - subscripts indicate required condition for the formula.
      * Ex: HRISP = hits while runners are in scoring position
  + This is then figured for every member of a team + an estimate of total team RS is added up.
  + **Actual total of team RS is then divided by estimated total team RS, yielding a ratio of real to estimated team RS**
  + The above individual RC estimate is then multiplied by this ratio, to yield a RC estimate for the individual
* Same info provided by RC can be expressed as a **rate stat**, rather than a raw # of runs contributed
  + usually expressed as RC per some # of outs, e.g. RC/27 (standard 9-inning game)
* **RC = believed to be an accurate measure of an individual's offensive contribution b/c when used on whole teams, the formula normally closely approximates how many runs team actually scores**
* Even basic version of RC usually predicts a team's run total w/in a 5% margin of error
  + Other, more advanced versions are even more accurate.

# Sabermetric Primer on RC

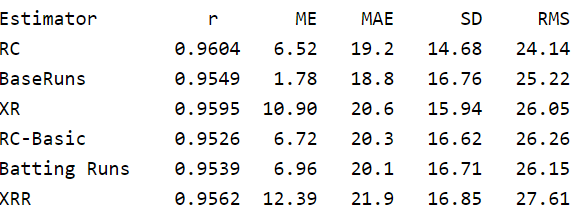
* For typical fan, sabermetrics doesn’t represent anything as theoretical as scientific inquiry.
* Rather, **sabermetrics = associated w/ new + unfamiliar statistics**.
* **OPS** = most famous of those new stats
  + gone from nearly unknown statistic in early 80s, to barely used a decade ago, to mainstream now (appears on Topps baseball cards)
* Have also been stats like **Linear Weights, Runs Created, Extrapolated Runs, WAR**, + so on.
* Can still argue that sabermetrics isn’t really about those statistics
* Rather, the statistics have been proven to be useful based on evidence that Sabermetricians have uncovered
* RC, for instance, = a statistic created by Bill James in late ‘70svia thinking that a team’s job on offense = score runs (more runs = better)
  + Suppose you didn’t know how many runs a team scored + wanted to make an estimate based on its **batting line**:
  + 
  + How many runs would you guess that team scored that year?
  + If made to guess, you’d probably look over few years of team statistics, try to find some team that was reasonably close, + use that as baseline.
  + Might find a team that hit .267 w/ less power, + scored 788 runs + figure “well, this team hit only .263, but had a few more HR’s, so maybe they’d cancel out, so we’d guess the same 788 runs
  + But, wait, this team had ~20 more walks than the other team, so maybe I should bump up my estimate to 800 or something.”
* What James probably did was work through logic like that + after some trial + error, come up w/ the RC formula, intended to provide a formal way of estimating how a batting line translates into runs.
* In most basic form, RC = 
  + Plugging in #’s from above batting line == (2371) (2055) / (6121) == 796 runs
  + This was actually the batting line for ‘85 Baltimore + they actually scored 818 runs
    - estimate is off by 22 runs, which is fairly typical.
* **Why is RC important?** Why do we need RC if we already know the Orioles scored 818 runs?
* Knowing there is a predictable relationship between a batting line + runs is useful when *we don’t know how many runs we actually have*.
  + Ex: can use RC on an *individual* player’s batting line like Pujols in ‘09:
  + 
  + Using the basic **RC** formula, we can estimate if a given major league team had a batting line like Pujols, it would score ~149 runs.
  + That batting line would comprise about 15 games == gives ~10 runs per game.
  + Can then conclude that if you put together a lineup of 9 Pujols clones, on average they’d score ~10 runs/game (average MLB team scores 4.5-5.0)
* Can compare Pujols to Joe Mauer, or Adam Lind, or Alex Rodriguez, to help inform conclusions on how much each contributed to his team, or even to arguments about which player deserves MVP
* RC = 1 of the most famous statistics used to evaluate offense
  + Others include Pete Palmer’s "**Linear Weights**," Jim Furtado’s “**Extrapolated Runs**,” David Smyth’s “**Base Runs**.”
  + All are very good estimators.
  + Which is best? 🡺 depends, as no estimator is perfect + all have strengths + weaknesses
* **1 way to compare the various estimators = test them for accuracy**.
  + Apply them to the last (say) 50 years == should give ~700 team-seasons.
  + Have them each estimate runs for all 700 teams + see which ones do the best.

# [A Closer Look at Run Estimation](https://www.fangraphs.com/tht/a-closer-look-at-run-estimation/)

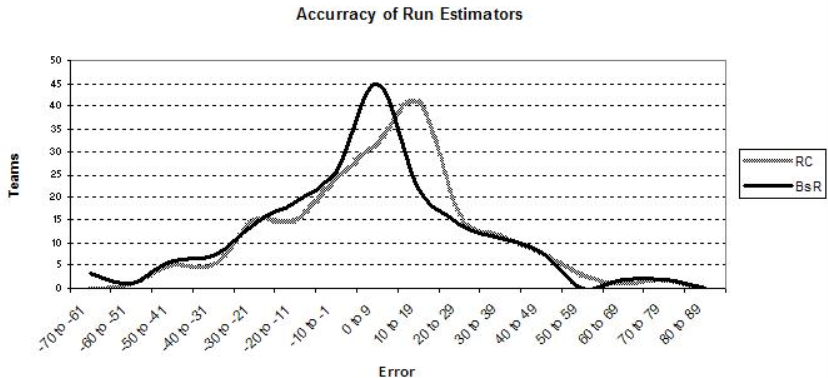
* Closer look at increasingly fashionable **On-Base Plus Slugging (OPS)** statistic in order to see how it compared as a proxy for offensive production vs. run estimators + other measures + just why it stacked up so well against them.
* **Correlation coefficients** showed that OPS tracks extremely well w/ actual RS *at the team level*, + as a result it actually belongs in the same group as the more complex run estimators, such James’ **RC**, Palmer’s **Batting Runs**, Smyth’s **Base Runs**, + Furtado’s **Extrapolated Runs**.
* **Simpler statistics (HR’s, walks, SLG %, OBP, BA) aren’t in the same class as these others when it comes to correlating w/ runs scored**
* **Close correlation w/ team-level RS justifies OPS as shorthand way of characterizing a player’s offensive contribution**
* In short, it has the **right mix of simplicity + accuracy to be a useful approximation for “back of the envelope” calculations of player value**.
* Concluded using algebra to show that, when broken into constituent parts, OPS = fundamentally a kind of **linear approximation** of these more complex formulas.
* While using correlation coefficients = acceptable way of comparing runs estimators w/ other measures like OPS, can illustrate different ways of comparing pure run estimation formulas + then tackle the questions that inevitably bubble to the surface when this topic is discussed.
* Before looking at other ways to compare run estimators, 1st be clear on **what we mean by a run estimator** + take a look @ **how they’re constructed**.
* Literal sense: a **run estimator = formula that attempts to calculate # of runs that “should” be scored, given some set of offensive elements.**
* i.e. **run estimator =** formula that models run scoring when applied to a set of *counting statistics* (bats, hits, HR, etc.)
* Then, by definition, statistics like BA, SLG, OPS = NOT run estimators, while RC, **Batting Runs**, **BaseRuns** ARE.
* Key strength of run estimators = they take into account the offensive **context** (outs + opportunities in the form of **plate appearances**) to return # of runs.
* *Even though OPS correlates well w/ run production, it lacks this crucial element*,
* If you compare team A w/ OPS = .750 vs. team B w/ OPS = .760, it doesn’t give a sense for how many runs a team should’ve scored nor the difference between the 2 teams
* Since run estimators = sometimes a bit of a religious issue among initiated, let’s 1st provide a few caveats.
* 1) Note that these formulas were *developed against the backdrop of specific data sets*.
  + Ex: versions of Extrapolated Runs here were created to best fit the period from ‘55-‘97, Batting Runs = originally developed using data from ‘01-‘77, **RC** = more recently tweaked
  + All could (+ in a perfect world should) be tailored to better fit a particular set of *teams*,
    - Ex: a specific league + year
  + The reason these formulas could be modified when working w/ different data sets is b/c the **offensive context changes from season to season + between leagues**.
    - Ex: A triple = worth more in a low run scoring environment like ‘68 (“year of the pitcher”) than in higher run scoring world (2001)
    - Conversely, cost of an out = smaller when runs are scarcer than when runs are plentiful (since in the latter, *each out = a greater opportunity lost*).
  + Notice all of formulas below apply various **weights** to offensive elements, either in terms of runs or base runner advancement.
  + As a result, those weights given to triples, outs, + other elements should rightfully vary in accordance w/ context.
  + The fact that analysts don’t typically take time to adjust for variations in context is a simplification used for convenience
  + That said, creators of both Batting Runs + BaseRuns include explicit ways to customize formulas for team + league context
    - Ex: **“score rate”** 🡺 Batting Runs formula can be tweaked by changing **negative run value** of an out
      * This adjustment (the **league batting factor**/**ABF**) is used to ensure total Batting Runs = zero for the given league + year.
  + Along these lines, adjustments could be made to the out value for both Extrapolated Runs formulas below as well.
  + Incidentally, in order to be able to compare Batting Runs w/ other estimators, we can also adjust the value of an out to transform the formula from one that returns runs above average into one that returns total runs
* 2) There’re multiple versions of each formula in the public domain
  + don’t all include the same offensive elements, so results shouldn’t be taken as proof a particular construction is necessarily superior to another.
  + Results may vary if using more complex variants or using a different data set.
* 3) Should *also* be noted Clay Davenport’s **Equivalent Average (EqA)** (used by Baseball Prospectus) is not included b/c technically it is NOT a *run estimator* 
  + requires a separate formula to transform EqA to produce Equivalent Runs (EqR))
  + also did not include Paul Johnson’s **Estimated Runs Produced (ERP)**, since it is essentially a *derivative of Batting Runs + Extrapolated Runs formulas*
* The Formulas
* **Runs Created (RC) =** 
  + This = more complex version of the formula introduced in The Bill James Handbook 2005 + includes SB + minor categories such as grounded into double play, intentional walks, sacrifice hits, + sacrifice flies.
  + formula contains A, B, + C components representing base runners, advancement, + opportunities, respectively, that are calculated as 
  + These components’ combination = **Runs Created Basic (RC-Basic) =** , one of the most basic versions of the formula created by James in the ‘70s.
  + This is essentially equivalent to OBP\*SLG + uses the same basic A, B, and C components
* **Batting Runs (BR) =** 
  + *offensive component* of Palmer’s **Linear Weights** system, originally derived using CPU simulation.
  + formula above = based on version from the 2004 edition of The Baseball Encyclopedia, where value of an **out** = -.10 rather than adjusted for the league as noted above, where typical ABF values for particular leagues + years range from -.23 to -.28.
    - Value here = lower b/c **ABF** is calculated so that **BR** returns runs *above average*.
    - In order to compare this formula to others, it’s necessary to **split the ABF** (value of an out) into 2 parts:
      * inning-ending value + advancement value.
    - Basis for this = straightforward + has been written about by Tangotiger.
  + **Value of an out** (or any offensive event for that matter) can be thought of as **sum of the value of reaching base, value the event has in moving runners over, + the value it has related to ending the inning**.
  + Using the run environment of 4.3 runs/g (average from ‘01-‘77), each out is “worth” -.16 runs, in terms of its *inning-ending value* (4.3/27).
  + Subtracting -.16 from -.26 yields an *advancement value of -.10* (used in this BR formula)
  + Notice the formula also includes SB + CS
* **Extrapolated Runs (XR) =** 
  + This = full version of the formula published by Furtado in ‘99 Big Bad Baseball Annual
  + Includes intentional walks, sacrifice hits + flies, as well as grounded into double play.
  + *Furtado created his formulas using linear regression.*
* **Extrapolated Runs Reduced =** 
  + This = Furtado’s simpler formula + also includes stolen bases.
* **BaseRuns (BsR) =** 
  + This = formula Smyth developed in early ’90s, for which you can find many variations.
  + variation above was published on Tangotiger’s site several years ago + is fairly basic, since it does not contain minor elements other than hit by pitch.
  + Since then, Smyth published updated version in June ‘05
  + This formula contains *A, B, C, + D components representing base runners*, the *advancement of those runners*, # of outs, + HR’s, calculated as follows:



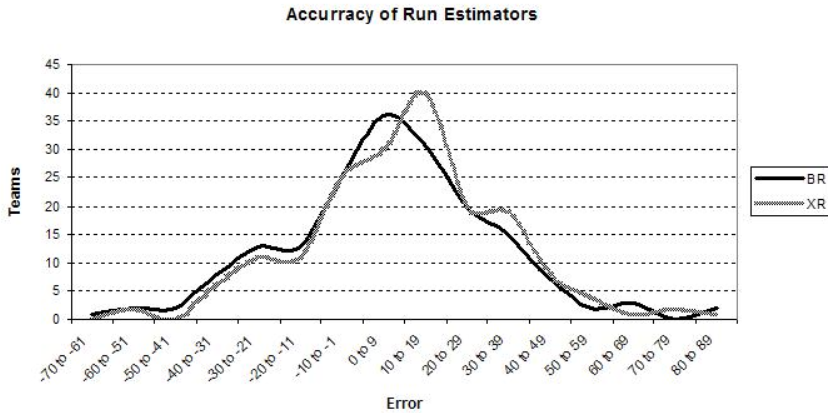
* + Interesting aspect of BaseRuns = B + C components are *combined* to calculate **score rate**
    - attempts to *estimate how often base runners score*.
  + When this ratio is multiplied by # of base runners (A) + added to # of HR’s, you get an estimate of runs would be scored as shown by **BsR**
  + Proponents of BaseRuns argue this model = only one that truly models how runs are scored
  + Some simple calculations using *extreme run environments* (ex: average beer league softball game) prove that out
  + MLB is not played in extreme environments, so all formulas tend to give similar results when OBP ranges from .300-.400 + SLG from .300-.500.
* The Results
* Previously used correlation coefficients, **r,** as a measure for how well a statistic like OPS or RC stands in for offensive performance.
* While r = a good start, it *only measures the strength of the linear relationship between 2 sets of values*, w/ -1 = perfectly negative/inverse relationship + 1 = perfect positive correlation.
  + i.e. high r for OPS (.948) indicates the general ordering of both RS + OPS is very close
* Although a run estimator may have an **r** very close to 1, that *doesn’t necessarily mean it does a great job of predicting actual # of runs a team scores*.
  + Ex: could create a statistic based on OPS, like OPS+63, that has a high **r** by virtue of its’ incorporating OPS, but doesn’t come close to returning correct # of runs.
  + As a result, also consider measures like avg. error/team, + their spread across all teams.
* The following table includes those measures applied to 180 teams from ‘00-’05 w/ columns = r, **mean error** (average of the sum of the estimate minus the # of runs a team scored), **mean absolute error** (average of the sum of the absolute values of the estimate minus the # of runs team scored), **standard deviation** (measures spread of the differences between # of runs scored + the estimate) + **root mean square** (known in statistics as a **power mean** = a combination of MAE + SD calculated as 
* RMS represents the *magnitude of the varying quantity*.
  + E.g. **RMS = handy way to take into consideration both average error + distribution of those errors + distill them in a single # for comparison.**



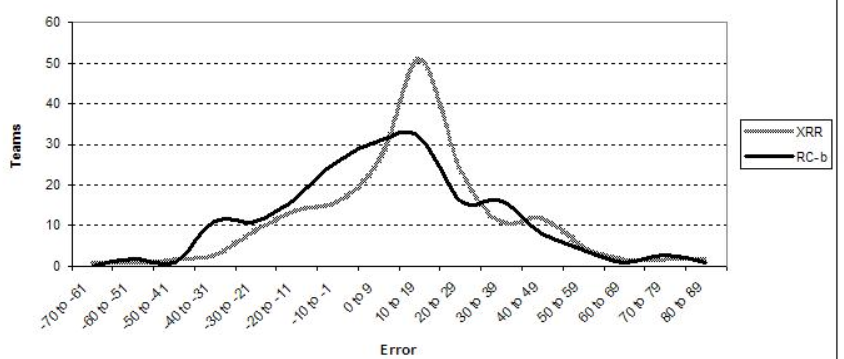
* See **BsR** has lowest MAE, while RC has lowest SD.
* Taking both into consideration, **RC outperforms BsR** by a fraction *when compared using RMS*.
* Another way to think about this = While BsR, *on average*, comes closer than RC, the *distribution of its errors is larger* (for the variant used here, at least) than any other estimator except XRR.
* It is therefore more likely that a specific BsR value is either closer or wider of the mark.
  + Ex: using BsR: 71/180 teams had estimates w/in 10 runs, while the next highest was 61 for Batting Runs.
  + At the same time, however, BsR had estimates for 7 teams that were off 60+ runs, while all other formulas except XRR had fewer than that.
* Also see that although XRR has a higher correlation coefficient than BsR, + it also had both the highest MAE + SD, + was therefore ranked last in RMS.
* This drives the point home that **correlation coefficient isn’t necessarily the best way to compare run estimators**
* Can view distribution of errors graphically by looking at # of teams that fall w/in certain error ranges



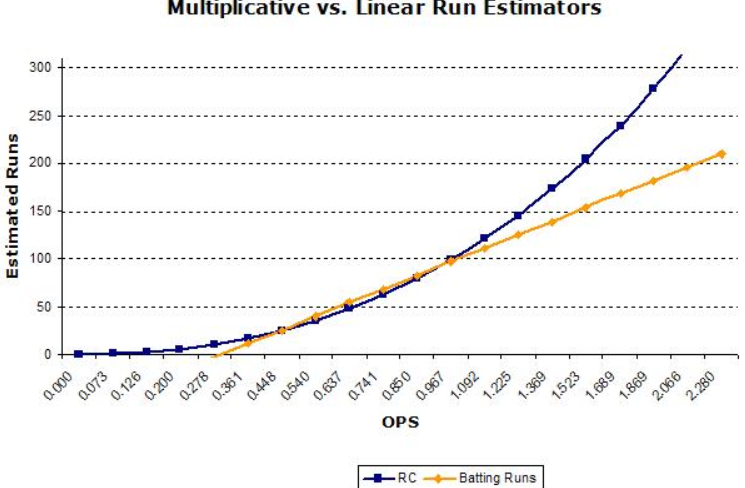
* Apex of the curve for BsR = higher + centered around 0-9 grouping, indicating it more often gets closer to the mark, hence its lower MAE.
* However, left-hand tail of RC remains under BsR’s, indicating that **when it underestimates actual runs scored, it does so w/ a smaller error**.
* RC also sneaks under BsR on the far right-hand side as well.
* Shifted nature of the graphs also illustrates that RC tends to overestimate actual runs moreso than BsR, as evidenced by its ME = 6.52, compared to 1.78 for BsR.
* 2nd graph showing XR along with its cousin BR.



* Obvious that XR tends to have a more compact distribution, although BR = more centered in -10-10 run error range.
* Also clear that XR = less prone to underestimate RS + more prone to overestimate it, as evidenced by higher ME of 10.90 for XR vs. 6.96 for BR.
* For completeness here are RC-Basic and XRR.



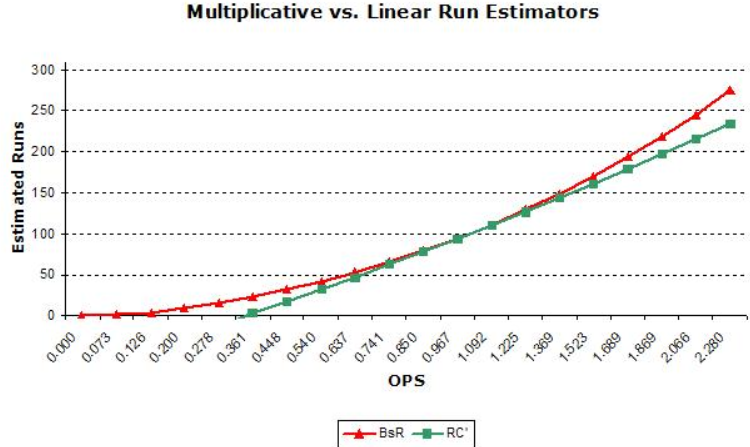
* Interestingly, although XRR has highest MAE = 21.9 + SD = 16.85, its distribution appears more compressed RC-Basic’s.
* Reason = **extreme values carry greater weight in the calculations** + XRR was off by 40+ runs on 27 of teams, while RC-Basic did so just 22 times.
* In addition, XRR curve = clearly shifted right as evidenced by ME = 12.39, so that it has, by far, the fewest # of teams in the -10-10 range.
* Although different run estimators have different distributions + tend to make the same kinds of errors overall, note that **they also make errors on the same teams**.
* Ex: all estimators agree ‘02 Phillies “should” have scored more than 710 runs by a wide margin
  + RC = 782 BR = 796 BsR = 786 XRR = 795 XR = 789 RC-Basic = 778
* As a result, differences in estimations for any 1 team is fairly small, as in this case here where it varies by 18 runs.
* Variables that tend to throw estimates off include especially good or especially poor hitting w/ runners on base (‘02 Phils hit .250/.338/.398 w/ runners on+ .266/.340/.441 w/ bases empty), or offenses where 1-2individuals carry the lion’s share of the load.
* But that brings us to the larger question regarding these run estimator = **What About the Players?**
* Keep in mind each formula = a run *estimator*, + all do a pretty good job of estimating runs *at the team level*.
* From best to worst, they vary by just 2.1 runs/team in ME, or just over a 1/10th of a run/game
* All told, the MAE ranges from 2.4% to 2.8% per team + as a result, any of these formulas can reasonably be used to estimate how many runs a team should have scored given a set of events.
* As mentioned previously, *several can be made more accurate by adjusting for the league context.*
* Where most of the value of these formulas lies, however, is in **applying them to individual players + from there calculating more advanced measures, like Batting Runs Above Replacement (BRAR) or Win Shares, in order to pin a value on a player’s contribution.**
* Note that this idea = nothing new.
* Today, when run estimators are applied to individuals, there are 2 approaches analysts use, depending on what they’re trying to measure.
* For some, run estimation for individuals = exercise in trying to determine how many of runs a team actually scored should be credited to an individual player
* i.e. aim to put a player in his team context.
* This = the approach taken by James in annual Bill James Handbook + in assigning responsibility for actual RS in his allocation of offensive **Win Shares**.
* For others, goal = exactly the opposite 🡺 attempt to take a player *out of his team context* to assign a value to a player’s offensive contribution + *compare him to players on other teams and eras*.
* In both cases, however, 1st must answer the question of **whether run estimation formulas designed + validated @ the team level can actually be applied to individuals**.
* At first glance, answer should obviously be yes.
* After all, *if a team can be projected to score X runs given a specific # of AB, hits, doubles, etc., then a player can be said to have created/contributed/produced Y runs given his AB, hit, doubles, etc.*
* *However*, statisticians are quick to point out **inferences about individuals based on aggregate data don’t always hold.** = **ecological fallacy**
* Ex: ‘00 presidential election 🡺 a study reveals a strong correlation between states w/ higher %’s of African American voters + states voting predominately for Bush.
* Problem = from such aggregate data, you can’t infer African American voters are more likely to vote Republican (In fact, 90% of African Americans voted for Gore)
* The fact that African American voters made up a smaller % of the total # of voters + that southern states contain a much greater % of white voters who voted for Bush conspired to bring about this result.
* Many analysts don’t agree that run estimators suffer from this problem
* Key difference between ecological fallacy examples + run scoring in baseball = in other examples, there’s an **interaction** of **multiple groups w/ different attributes that act independently** (black + white voters w/ different voting patterns + population sizes.
* In baseball, individual players create runs for a team, working in concert towards the *same goal* achieved in the *same way*
* Logically, coefficients derived for run creation @ a team level must apply to the individual.
* So, assuming formulas can be applied to individuals, there is a 2nd issue that often comes up
* In looking @ the formulas above, notice there is a fundamental difference in their construction that allows us to place each into 1 of 2 camps.
* RC + BsR = **multiplicative** formulas, while BR + XR = **linear** formulas.
* RC + BsR model run scoring in a **non-linear + interdependent fashion** w/ respect to offensive events, while **linear estimators** like BR’s **apply run weights** to offensive events + **sum totals**.
* Result = in multiplicative formulas, offensive elements interact w/ each other to produce run estimates
* Weights used in multiplicative formulas should therefore be thought of as applying advancement values to offensive elements.
* *As a result, multiplicative formulas give higher estimates for runs as frequency of offensive events increase, while linear estimators, as the name implies, increase in a straight line.*
* This difference is illustrated graphically below, where as OPS increases (in context of 650 plate appearances), **multiplicative estimator** increases faster than the linear estimator.



* However, in practice, this **difference has little effect when the formulas are applied to team statistics**, since baseball is not played anywhere near the extreme ends of this graph (team OPS values typically hover in the .730-.780 range)
* However, **when applied to individuals**, these **differences are immediately noticeable** as players like Pujols + Bonds *benefit from their own offensive elements interacting w/ each other*
* Of course, a player does not interact w/ his own statistics to create runs, but *rather w/ his teammates*
* This led James in 2002 to modify his formula for individuals + place the player *in the context of 8 other players w/* a .300 OBP + .400 SLG, *by changing how the A, B, and C components are combined.*



* A factor = modified to include 8 other players w/ a .300 OBP (8 x .300 = 2.4), B factor augmented w/ 8 players w/ a .400 SLG (8 x .400 = 3), + C factor includes plate appearances for all 9 players.
* After performing **(AxB)/C**, RC by the other 8 players *are removed* by multiplying plate appearances for 1 player by .9.
* This works since the RC by 8 typical players = 10% of plate appearances (quirk of using the .300 OBP and .400 SLG)
* Noted he also modified the formula to adjust for player’s performance w/ runners in scoring position (following graph shows the adjusted version along with BsR)



* BsR doesn’t have the problem to the same degree, since HR’s are largely separated + as such don’t interact as heavily w/ other elements.
* Also notice it rather nicely intersects at 0, while adjustments to RC serve to make the formula essentially linear + force it to intersect X @ an OPS = ~.350.
* *This illustrates why proponents of BsR find the formula so elegant + claim it is the only formula that accurately models run scoring.*
* Finally, applying formulas to individuals has the same issue as when applying to teams == **context**.
* In actuality, **offensive context** changes not only w/ leagues + seasons but also w/ teams, batting order, parks, + so on.
* Although various coefficients applied to offensive elements won’t change that much for a particular lineup position on a particular team, there IS a subtle difference between them.
* Therefore, an analyst attempting to allocate a team’s RS to individual players should technically adjust the formula for as many of these context issues as possible.
* The Score: Where does that leave us?
* As mentioned previously, there are few topics which divide the performance analysis community more than run estimation.
* However, when you boil it down, there are 2 important points for the thinking fan.
* 1) RS can be predicted from the *combination of offensive elements quite closely using any of the popular formulas, w/ the caveat that performance falls w/in reasonable ranges.*
* 2) The *formulas can indeed be applied to individuals*.
* Together that means thinking fans can + should embrace these tools + thank their creators Lane, Palmer, James, Furtado, and Smyth, which help all of us understand a little better, + hopefully appreciate the game a little more

# Pitching and Defense: How Much Control Do Hurlers Have

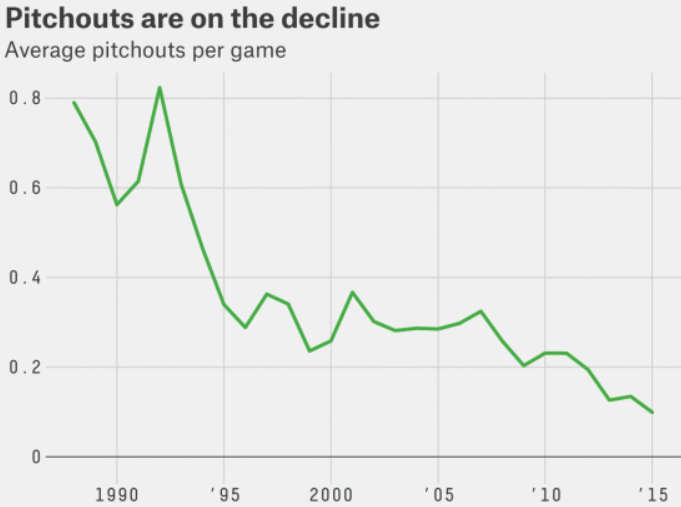
* A belief 🡺 hits allowed are not a particularly meaningful statistic in evaluation of pitchers.
* 1 basic issues in evaluating pitchers = to what extent defense is responsible for results.
* Baseball Prospectus 2000, 1 of Woolner‘s “Hilbert Problems” for baseball = **issue of separating defense + pitching** 🡺 “Pitching + fielding are so intertwined they seem impossible to separate.”
* Divide pitcher’s stat line into what defense can + can’t affect
  + Defense Independent: Walks, SO, HR (essentially), Hit Batsmen, Intentional Walks
  + Defense Dependent: W, L, Innings, Runs, Earned Runs, HA, Sacrifice Hits, Sacrifice Flies.
* Any stats *derived* from defense-dependent ones (OPS against, ERA = also defense dependent.
* Idea = to express things defense can’t affect in 1 area, check results, then check those areas where defense can have an effect + analyze how much of performance = pitching + how much = defense.
* **Defense Independent Pitching Stats** **(DIPS) =** representation of a pitcher’s stat line w/out any possible influence from defense behind a pitcher.
* Ex: Various rates for walks, SO, HR, hit batsmen, etc. were calculated *as a function of batters faced* + inserted into pitcher’s line.
* Then how many batters faced were remaining was calculated + league-average rates for all other component stats: innings, hits, doubles, triples, etc. are assigned
* So, for all stats that defense *could* affect, every pitcher was now on equal footing.
* Results using Burba‘s line in ‘00, looked something like this:
  + Actual:  Defense Independent: 
* HR, walks, + SO changed little (only due to **park effects** + a few other minor factors).
* H + innings pitched changed by a decent amount (at least in this case)
* Next step = look at rest of pitcher’s stat line + somehow **divine how much was the result of pitcher’s work.**
* To do this 🡺look @ range of values for **Defense Independent ERA** + compared how close they were to the range of values of **actual ERA**.
* Ex: if range of Defense Independent ERA = 4.00-5.00, it would be a good indication that there’s a lot about pitching not covered in the stat, b/c ERAs have a much larger range than that.
  + *That didn’t happen 🡺* range was virtually the same as *actual* ERA, w/ best pitchers having **DIPS ERAs** near 2.40 + worst having DIPS ERAs near 7.00.
  + Surprising, as range was expected to be quite a bit more narrow than that.
* Then, look @ behavior of **Hits Per Balls in Play = [(H-HR)/(BFP-HR-BB-SO-HB)].**
* No matter the calculations, there is little if any difference among ML pitchers in ability to prevent hits on balls hit in field of play.
* It is a controversial statement that counters a significant portion of 110 years of pitcher evaluation.
* Facts that led to this conclusion:
  + range of ERAs for pitchers is almost as large *without* defense-dependent stats as it is w/ them 🡺 speaks to the fact that *there can be massive differences in ability of pitchers before even considering impact of defense.*
  + Pitchers who’re best @ preventing hits on balls in play 1 year are often worst the next.
    - ‘98, Maddux = 1 of best rates, ’99 one of the worst, ’00 1 of the better ones again
    - ‘99, Pedro Martinez = 1 of the worst; ’00 = the best.
    - This happens a lot
  + Little correlation between what a pitcher does 1 year in the stat + what he will do the next.
    - In other words, what hits per balls in play in ‘00 tells us next to nothing about ‘01.
    - This is NOT true in other significant stats (walks, SO, HR).
    - Walks + SO correlate very well + HR correlate *somewhat* well == a crucial fact.
    - \*\*\*1 of the more critical aspects of statistical analysis = **determining how well a statistic reflects an ability\*\*\***
    - This = the test given to clutch hitting, catcher game-calling, pitcher W/L records, etc.
    - \*\*\*1 of the 1st things asked when addressing this == **“Does the stat correlate well with itself from year to year?”\*\*\***
    - 1 reason “clutch hitting” is questioned == “clutch hitters” change from year to year, which **indicates it probably isn’t the hitter as much as it’s other factors**.
    - The answer to whether hits per balls in play correlates well from year to year is a fairly solid “No.”
  + Can better predict pitcher’s hits/balls in play from the rate of the rest of pitcher’s team than from pitcher’s own rate
    - pretty self-explanatory 🡺 effects of having same team defense + home park appear to be significant determinants in creating what little correlation there is in the stat
  + Take pitchers w/ similar stats in every other **component category** (+ other peripheral factors like age, throwing hand, team hits per balls in play rates, etc.) but **large differences in hits allowed** (+ therefore in innings pitched).
    - When you group pitchers into 2 categories (high vs. low hits), the following year high-hit pitchers do NOT give up significantly more hits per balls in play (.292 vs. .291) than low-hit pitchers + the groups have identical ERAs.
    - This is a difficult point to overcome if you want to show that preventing hits per balls in play is a significant ability of *pitchers*.
    - If, *all other things equal*, there is no difference, the conclusion becomes clearer.
  + Similarly, if you take pitchers w/ comparable stats in every other component category, but have as large as possible a difference in SO, then separate pitchers into high + low-SO categories
    - high-SO pitchers continue to strike out more hitters, while also giving up far fewer hits + having significantly lower ERAs.
    - This is the natural opposite of the above point.
    - If above point = true, then logically, this point ought to be true as well + it is.
  + Range of career rates of hits per balls in play for pitchers w/ a significant # of innings is about the same as the range expected from random chance.
    - This is true even though we know some pitchers may’ve had consistent advantages over others, as these rates are unadjusted for park or league.
    - Vast majority of pitchers who’ve pitched significant innings have career rates between .280-.290.
  + When adjusting for environmental advantages (DH, park effects, etc,), the range becomes even smaller.
    - Leaders in this stat have had significant environmental advantages while most trailers have had disadvantages.
    - After adjustments, the range is well w/in the realm expected from chance alone
  + A stat like **Component ERA** (or any similarly stat calculating ERA from the rest of a pitcher’s performance), while *correlating better w/ next-year ERA than ERA itself*, does NOT correlate nearly as well w/ next-year ERA as *it does if performing the same calculation while using average hits-allowed rate of the team for which he pitched*.
    - Thisadvantage of **“team average” rate** grows to rather largeproportions as # of innings pitched in the season shrinks more + more.
    - 2 key points here:
      * doesn’t appear to be any “hidden quality” aspect to the stat as #’s come out as they should if the above are all true: (can better predict ERA w/out HA than w/ them)
      * Using a reliever’s **hit rate** seems to be an extremely suspect way of evaluating relievers (see Ayala in ‘98 + ’99)
* There are a few lesser + somewhat anecdotal points to be made that, while not critical, are nonetheless good concepts to understand:
  + Hard to diagnose pitchers that’re very good at preventing hits on balls in play (often hear Randy Johnson, Jamie Moyer, Andy Pettitte in protest of the concept, but by any definition you want to use, they are NOT particularly good in the stat)
  + Pedro Martinez has been quoted as saying that *the batter* determines what happens once he hits the ball + Greg Maddux described his scoreless-inning streak as “mostly luck” as hard hit balls that had been falling in were being caught.
  + We only have 38 innings’ worth of non-pitchers’ pitching (ex: Brent Mayne) = too small a sample on which to draw conclusions
    - but it’s something to think about that non-pitchers were not any worse than regular pitchers in the stat + in fact were a good bit better.
  + Pitchers are often dubbed as “unpredictable”, + HA = by far the most unpredictable of **component stats**.
    - In other words, is 1 of the main culprits of pitcher unpredictability.
  + ***There is no significant cross-correlation*** 🡺 a high # of HR allowed doesn’t really mean anything in determining how many hits per balls in play pitcher will allow.
    - closest = an inverse relationship w/ SO (lots of SO means fewer hits per balls in play), but that relationship is very weak + could be the result of unrelated factors
    - There was no significant hits-per-balls-in-play advantage found in SO study above.
* Many people, after reading these points, think I’m saying all pitchers give up the same amount of hits 🡺 not true + not what I’m saying.
* Randy Johnson gives up fewer hits than Scott Karl not b/c batters hit balls harder off Karl than Johnson, but b/c they hit balls more *often* off Karl than Johnson.
* Aside from walks, there are 2 basic outcomes for a pitcher: batter hits ball or batter strikes out.
* W/ latter, result = almost always out + w/ former, many things can happen, including a base hit
* *So why is this all true*? 🡺Theories can be checked out + some are more difficult to verify.
  + Scouting: MLB scouting network is set up to sift through an enormous pool of potential players to get to a group that might be MLB pitchers.
    - To do so, often employ tactics many might call unfair in effort to reduce pool to manageable # (don’t take guys < 5’10” + every one must throw a certain speed fastball, etc.)
    - 1 of these factors may be weeding out a subset of pitchers for which above theory is not true.
  + High talent level. = theory that there’s a certain limit as to *how good* you can get at preventing hits on balls in play + that in order to even come close to the ML’s you have to have reached this.
    - often comes up in clutch-hitting discussions.
  + Too many variables. suggests the ability may/may not exist, but that the # of variables involved in the outcome of balls in play are so numerous + difficult to control for that any ability gets lost
    - i.e. **the noise completely masks any signal.**
  + A misunderstanding of how the batter/pitcher dynamic works.
    - Some argue that despite all the #’s, the above can’t be true b/c it means a screaming line drive hit into right-center-field gap is as likely to be an out as a pop-up to shortstop
    - This deserves further discussion:
      * 1 critical point of misunderstanding = issue of “blame.”
      * When a ball gets crushed into gap in right-center, some think we’re saying the defense deserves the blame, not pitcher.
      * Counter w/ “Neither is to blame, the batter who is to blame”
      * Ex: HR Derby 🡺 Watch for batted balls that would clearly be outs, as pitcher is *trying* to give up HR, so does he deserve credit for a pop-up?
      * In MLB, a pitch could = a pop-up or a line drive + it all depends on what the batter does w/ it.
      * Conventional wisdom on dynamic between pitcher + batter may be slightly inaccurate
* The critical thing to understand= **ML pitchers don’t appear to have the ability to prevent hits on balls in play.**
* Many possible reasons why this is the case w/ no concrete idea as to why it is

# The Art of Pitch Framing

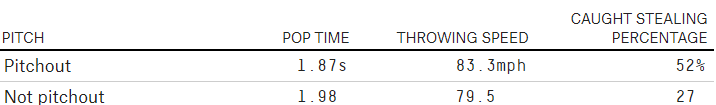
* Focus on how catchers “**frame**” pitches to make them look more like strikes + distinction between players starts to stand out.
* Depending on how caught, 2 pitches almost identical on way to plate can look a lot different once getting to the glove
* Set up farther outside 🡺 even though a pitch is farther from plate, can catch it in the center of his body, while reaching for the ball == drawing attention to its distance from strike zone.
* 1 goes down to 1 knee to present a lower, more stable target while the other’s head jerks sharply downward the instant after catching the pitch.
* 1 remains still, while the other’s glove, descending to meet the pitch, dips even more after catching, which sends ball farther outside the zone + forces him to jerk glove back up in exaggerated fashion
* 1’s glove never gets any lower when receiving the pitch, makes a much more subtle upward movement, + takes about half as much time for glove to come to rest.
* Differences in technique don’t only pop up on *some* pitches but are present for almost every pitch, as constant as a pitcher’s motion or a batter’s stance.
* Body leans toward the outside of the plate vs. staying almost perfectly still, raises + clenches hand, creating a potential distraction for umpire vs. remaining tucked behind, head + glove dip before pronounced pull vs. timing reception to catch the ball w/ glove already on the rise.
* Baseball = often described as chess match between batter + pitcher but is more between batter + pitcher in which, once in a while, the catcher grabs the board + moves someone’s piece.
* Sep. 2011 article, “Spinning Yarn: Removing the Mask,” attempted to determine what **catcher receiving** was worth.
* By studying where strikes are typically called + establishing which pitchers were getting more/fewer strikes than “should” have, given where pitches crossed the plate, it was able to isolate **effect of the catcher**.
* Concluded **pitch framing** can make a major impact + is also more consistent year to year than even reliable offensive metrics like OBP or SLG.
* i.e. it’s not *insignificant*, *+ it’s not just noise* but a valuable skill that persists season to season.
* Found Molina (best receiver) = worth 35 runs above avg. per 120 games, + Doumit (worst) = worth 26 runs *below avg.*.
* A later sophisticated model for framing accounted for most potentially confounding factors: umpire, ballpark, batter, ball-strike count, + pitch location + type.
* Molina has saved teams 111 runs (or, using standard 10-runs-to-a-win conversion, ~11 wins) b/c of framing from 2008-2013
  + only other catcher w/ a higher run total over same time period = Brian McCann @ 122, who has caught more than 2x as many pitches.)
* Doumit *cost* his teams 155 runs 🡺 0.50 runs added by Molina + 0.55 runs subtracted by Doumit per 100 pitches, an enormous difference
  + For comparative purposes, Bonds’s bat during 2001-2004 seasons, when he basically broke baseball, was worth ~0.78 runs above average per game.
* Granted, *those run totals aren’t typical* as Molina + Doumit = outliers (no one else comes close to costing as much per pitch as Doumit)
  + only Lucroy has approached Molina levels of framing effectiveness via some advantages over Molina = 11 years younger w/ a better bat, which allows him to stay in lineup
  + Since his debut in 2010, he’s saved almost 2x as many runs due to framing as any other catcher, aside from Molina.
* Still, consider implications: Giancarlo Stanton, 1 of the most coveted young players, was worth as much (in terms of **Wins Above Replacement**) over past 2 seasons as Molina’s framing alone was worth in a *part-time* role over the past 5+, yet Molina’s value, unlike Stanton’s, is largely overlooked.
* That’s w/out factoring in any value Molina added by calling games, handling pitching staff, + controlling running game.
* Best of all, Molina’s done it all for an average of $1.5M/season, in an era when a single win on the FA market usually runs teams ~$5 million.
* No surprise Molina plays for the Rays, low-payroll competitors who’ve found ways to make a $ go further than any other org.
* Nor is it a coincidence that Molina, a career backup catcher whom standard Sabermetric stats peg as a replacement-level player, became a 1st-time starter + appeared in a career-high 102 games in his age-37 season, barely 6 months after Fast’s study appeared on Baseball Prospectus.
* 11 wins sounds like too much for a single part-time pitch receiver to add in just over 5 seasons, but think about how many opportunities to gain + lose strikes there are over 162 games.
* Durable catcher can catch ~10k called pitches/single season + many pitches are clear-cut calls.
* Leaves hundreds to thousands of pitches in shadowy border region between ball + strike, where a good receiver can perform catching equivalent of turning water into wine.
* Even if an extra strike doesn’t send batter back to dugout, it puts him in a less-favorable count + makes him less likely to do damage later in the AB.
* Some put avg. value of turning a single ball into a strike = 0.13 runs
  + Do that a few times per game, as Molina does, run total climbs quickly
* “If a catcher can perfect a great way of receiving a ball + gets the ball maybe a half a ball outside — or even a ball outside — off the corners consistently, he’s worth his weight in gold,”
* Receiving skills weren’t *not* valued before, but it’s easier to discount a certain skill w/ no #’s attached to it. (i.e. clubhouse chemistry.)
* Before framing became a sabermetric buzzword, Molina’s receiving skills were known but nebulous.
* Offensive struggles were easier to see + almost as easy to quantify.
* Now that we can count his contributions on *both* sides of the ball, what he does on defense is impossible to dismiss + Rays are reaping rewards
* Catchers like this = not a renewable resource.
* Teams could learn to teach framing efficiently at lower levels + have hands on a major market inefficiency
* Thanks to advances in technology + analysts, we have a much better sense of what a catcher w/ good receiving skills is worth + also know which big leaguers qualify as best.
* What we don’t know, necessarily, is *where good framing catchers come from*. Born or made? And if made, *how* do you make them?
* Pitch-tracking tech has penetrated minors, only so far w/ only ~40 minor league ballparks have a PITCHf/x system installed, + ~25 are equipped w/ Trackman (competing ball-tracking tech that relies on Doppler radar).
* Around 20/30 parent clubs have at least 1 affiliate w/ 1 of the 2 systems in place, w/ highest concentration at Triple-A.
* Teams pool much of the info collected (the more of your own PITCHf/x or Trackman data you share, the more you receive in return (although clubs generally keep info from non-league events held in their stadiums (workouts, high school + college games) to themselves)
* Still, 65 ballparks w/ ball-tracking tech just scratches the surface of professional baseball below the big leagues.
* Even in 2013, 7 years after 1st PITCHf/x systems installed in big league ballparks, most minors (to say nothing of amateur + international baseball) are a ball-tracking blind spot.
* Means receiving skills still have to be evaluated “old-fashioned” way (w/ actual eyes) + slightly less old-fashioned way (via video)
* Not really a problem, b/c search for good framers (like almost every other pursuit in modern front office) works best w/ **input from both stats + scouts**.
* Over small samples, experienced scout/instructors can tell more about someone’s receiving skills than a CPU connected to a camera, especially @ levels where it’s more difficult for catchers to frame pitches b/c pitchers can’t hit targets consistently.
* Opinions vary on exactly how long it takes to evaluate a catcher’s receiving skills, but consensus is the assessment comes quickly, some say 5-10 pitches.
* Others = more conservative estimates = couple games w/ different pitchers, b/c every pitcher’s different”
* Snap judgment = just the start.
* Once acquiring catcher w/ some raw talent for framing, real work begins.
* No one-size-fits-all plan for grooming great receivers (+ not every catching instructor agrees on proper way to receive pitches), but there are, generally, 2 types of improvements a player can make.
  + mechanical change 🡺 can be adopted immediately
  + incremental increase in comfort + confidence that comes w/ experience, repetition, + familiarity w/ a pitching staff
* Some catchers = framing “fixer-uppers” = just a few alterations away from becoming much more valuable properties.
* Lower the target, turn thumb, adjust crouch = easy stuff
* May feel unnatural at first w/ some time to get used to, but can produce improvement overnight
* More grueling method of improvement over the long run produces greater gains + is as simple as it is painstakingly slow: *catch* 🡺 in games, in bullpen, in side sessions, + again in dreams
  + “Baseball = a habit. It’s a repetition. You’ve got to repeat things every day, + they come”
* Bullpen drills = big part of repetition as pitching machines can simulate any type of pitch @ any speed in any location, + apprentice pitch framers exhaust all possibilities.
* Drills aren’t a time to fill quota of practice pitches + move on to something more interesting but a time to concentrate + consolidate lessons.
  + breaking balls, left-handed, right-handed, + move catcher around behind the plate to work on backhand pitches + forehand pitches, some soft wiffle-balls to simulate movement, working on catching w/out a glove, check to make sure they’re breathing back there.
* Motivated catchers don’t just do drills but watch video of themselves *and* opponents @ the position
* Eventually, all this pays off in game action.
* If it sounds a little monotonous, it is = another reason why, even if decades down the line every place where baseball is played comes pre-installed w/ a PITCHf/x (or FIELDf/x) system, there will still be a role for scouts.
* No matter how all-seeing eyes in the sky above ballparks become, teams will still need to know what’s going on *inside* players.
* Character + work ethic are at least as important in determining whether a catcher reaches his ceiling as a receiver as it is in determining how good a hitter he’ll be.
* Naturally, younger you start to work + the more naturally gifted, the quicker it will come, but those gifts aren’t a prerequisite.
* Now teams know what a great receiver is worth, + players know they know, so both sides have more incentive than ever to focus on framing
* Losing teams led by creative executives = able to experiment + innovate.
* If going to be bad no matter what, might as well try to be bad in a way that’ll make you better eventually
* Astros = rebuilding club 🡺 Rather than risk jeopardizing long-term potential by trying to strike a balance between rebuilding + respectability, decided to raze their roster, trading every veteran who wasn’t tied down + spending hardly anything on FA.
* path led to a lot of losing, w/ much more to come but it’s also allowed them to completely start over, in just about every capacity
* Restocked formerly barren minor league system, climbing from 26th to 9th in Baseball Prospectus organizational rankings in 1 season, blazed new trails in pitcher usage patterns + defensive shifts, + assembled a new-school, cross-disciplinary front office w/ intellectual talent from atypical baseball backgrounds.
* Astros = exploring ways in which they can make receiving skills a strength (can’t see effort reflected yet at ML level + it wouldn’t matter much, anyway)
  + few extra strikes right now wouldn’t make Astros’ immediate outlook any less hopeless
* It’s *below* the surface where the framing future is taking shape + where Astros are seeking out potential benefits of developing strong receivers before they reach big league level.
* By stocking minors w/ good framers, Astros could accelerate pitching prospects development/make them more attractive to other teams by bolstering stats w/ catchers who earn them extra strikes
* W/ catchers themselves, can simply wait until catchers enter the arbitration process + start to make more than worth, at which point they can promote those prospects.
* By the time Astros decide to make winning at the big league level a priority, could have a few catchers capable of expanding the strike zone.
* 1st thing to look for = moving in the direction of framing
* Next stage = implementing development strategy designed to make the most of each catcher’s receiving skills (appears to be buy-in @ all levels of the organization)
* If new Houston regime has its way, practice will translate into wins a few years from now + if the approach succeeds, it’s bound to inspire copycats,
* All it takes = for other teams to start asking the same simple question: “Why not try to get better?”
* Any team that has already started targeting + trying to develop good framers are only the vanguard.
* Market for catchers w/ superior receiving skills will grow more crowded as long as it looks like an area where clubs can get an edge
  + a lot easier to teach somebody how to frame a pitch than to hit homers + drive in runs.
* If you take the Astros’ plan + comments about copycats to a logical conclusion, @ some point in the not-too-distant future, almost every team (save, perhaps, a few w/ gifted offensive catchers for whom framing aptitude is less paramount) could have someone squatting behind the plate + stealing extra strikes.
* But sweeping changes to a sport rarely come w/out unforeseen consequences.
* That kind of mass movement toward catchers w/ strong receiving skills would upset delicate balance between batter + pitcher;
* If you think baseball’s SO rates are high now, wait until the 1st wave of Stepford framers arrives.
* If umpires start to see nothing but good receivers, might adjust zones, much like a hitter adjusts to a pitch he’s shown too often.
* Then the framing bubble would burst, as previous advantages have when every organization discovered them
* Also possible that greater awareness of framing could hasten the end of umpiring as we know it.
* Discovery of framing has opened up a new field of research, but what’s good for baseball writers isn’t always good for baseball.
* Even if intent isn’t to criticize umpires, it’s impossible to write about framing w/out drawing noting the rulebook strike zone = more of an abstract concept than something that exists.
* The more attention catcher’s ability to influence a strike zone receives, the more likely MLB will act to automate it.
* If the human element goes, framing will go w/ it == calls might be more accurate, but we’d lose the art that is a perfect frame.
* Market-induced demise of framing might arrive 1 day, but it’s not here yet.
* Career-wise, this is the best time in baseball history to be a solid receiver.
* Runs are runs, whether scored or saved + now that we know how many runs they’re saving, former fringe players have become commodities that every team wants.

# Sabermetrics Is Killing Bad Dugout Decisions

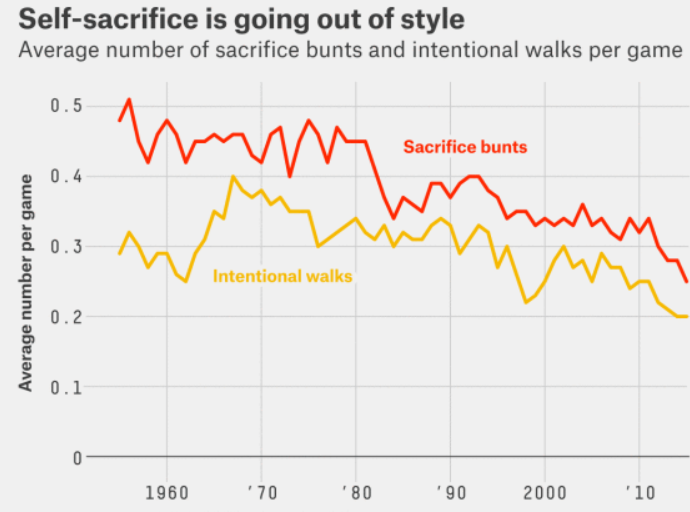
* Teams evaluate data differently in 2016 than 30 years earlier
* Front-office pages include job titles like “quantitative analyst” or “decision scientist”
* In-game, the evolution would be obvious if you knew where to look, broadcast by tactical details like how few hitters shorten swings to make contact or how willingly fielders play out of typical positions
  + 7 infield shifts/game last season, up from less than 1/ game as recently as 2011
* What’s missing from most games might be just as instructive as what’s more common.
* Among most conspicuous = pitchouts, intentional walks, + sacrifice bunts = 3 tactics in use since the 19th century that’ve seen reputations suffer in age of sabermetrics.
* None are extinct, but all are endangered, w/ each reaching its lowest recorded level last season.
* All 3 force teams to accept something bad in order to increase odds of getting something good.
  + **Pitch out** = sacrifice possibility of a strike for better chance of nabbing a base stealer
  + **sacrifice bunt** = give up an out to raise probability of scoring a run on a single (or an additional sacrifice out)
  + **intentional walk** = concede a baserunner to set up a matchup w/ a weaker batter they believe is more likely to let them escape from the inning unscathed.
* In recent years, those costs + benefits have been better quantified + research has suggested the moves have long been abused.
* **pitchout’s decline** = the most precipitous:



* downward trend = attributable almost entirely to decisions made in dugout.
* Former catcher John Baker: “100% of pitchouts were called from dugout” + Baseball Prospectus writer Sam Miller showed in 2013 that pitchouts are probably counterproductive.
  + pitchouts = attempt to control the running game w/out wasting a pitch on a pitchout.
  + If difference in time for catcher’s release is negligible, what is the point of wasting a pitch?
* When managers guess right w/ a pitchout, there is a tangible benefit.
* According to a combination of Statcast data from ‘15 + play-by-play data from 2011-15 provided by MLB Advanced Media, pitchouts on attempted steals of 2nd enable catchers to release the ball more quickly, put more power behind throws + catch runners at a markedly higher rate than when they have to worry about receiving a regular pitch:



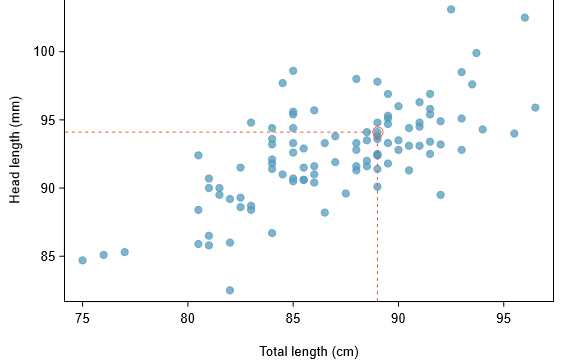
* **Problem = managers guess right relatively rarely**.
* According to MLBAM’s, only 19% of pitchouts from 2011-2015 coincided w/ steal attempts
* For every pitchout that did so, more than 4 resulted in a ball while runner stayed glued to the bag.
* ~36% of those pitches would’ve been balls anyway (league-wide strike rate over that span = 64%)
* But, that still means pitchouts turned strikes into balls more often than giving catchers a better chance of stopping a steal.
* Now that we know how much worse hitters perform after falling behind in the count, we also know that losing even a single strike comes at a significant cost.
* Likely the decline in pitchouts is tied at least loosely to a corresponding decline in SB attempts, which *in turn* is tied to a growing awareness of costs of being caught stealing, another way in which managers have curbed excesses.
* Even in a high-steal environment, a pitchout-heavy stratagem would be tough to justify.
* For pitchout to make sense, managers would have to guess right almost 50% of the time, especially given that **the tactic doesn’t act as a deterrent to future steals**.
* More + more managers are conceding they can’t pull off the impossible.
* MLBAM’s records claim the ‘15 Red Sox = 1st team to play full season w/out a single pitchout, whose coaching staff’s philosophy wasn’t dictated by a front-office study
* “Tried to put it in the hands of the pitcher + catcher = varying hold times, making sure that we school guys enough to have unloading times where they’re controlling the running game + minimizing that w/out artificially doing it through a pitchout.”
* In ‘15, Reds + Mariners managers pitched out 30 + 28 times, respectively = most anachronistic skippers in respective leagues by far
* While falling # of pitchouts might seem predictable in a low-steal era, other tactics have fallen out of favor in an environment where, at first glance, they’d seem positioned to thrive.
* Sac bunts + free passes have fallen despite league-wide decrease in scoring (tends to make sacrifices + intentional walks less costly) + pitchers’ deepening offensive ineptitude, which gives them greater incentive to bunt + makes batters hitting ahead of them more likely to be IBB’d



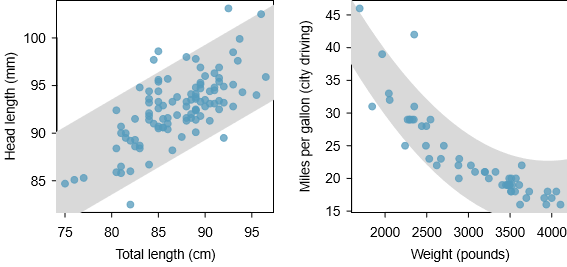
* Run-expectancy tables based on records from hundreds of thousands of rallies + routine innings revealed that, in many cases, sac bunts impair a team’s offensive outlook, while intentional walks improve an opponent’s.
* World Series-winning manager Ned Yost (old-school rep persisted despite sabermetric bookshelf + extremely laissez-faire style) = @ forefront of anti-IBB trend.
* Royals have ranked last in IBBs allowed for past 2 seasons, thanks to Yost’s appreciation for the %’s
* On the other side of the ball, “sac bunting is bad” had been a sabermetric maxim for decades before “Moneyball” made it mainstream.
* Even now, Oakland’s bunt total brings up the rear, but more clubs are keeping the A’s company.
* Neuropsychologically speaking, it’s easy to see why a manager would be tempted to try to prevent a steal, push across a run or bypass a scary batter.
* But baseball’s ever-tighter embrace of advanced stats is making managers smarter + more likely to take the long view and less likely to go by their gut.

# Chapter 5: Introduction to linear regression

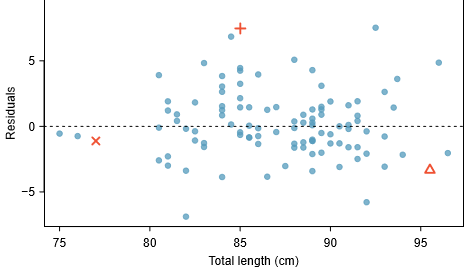
* Linear regression = very powerful statistical technique where we try to model relationships w/ a straight line
* perfect linear relationship = know exact value of y just by knowing value of x (unrealistic)
* Family income X, would provide useful info about how much ﬁnancial support Y a college may oﬀer a prospective student but there’d still be variability in ﬁnancial support, even when comparing students w/ families w/ similar ﬁnancial backgrounds.
* **Linear regression *assumes* relationship between variables can be modeled by straight line: y = β0 + β1x** where β0 + β1 = model parameters estimated using data (written as **point estimates** = b0 + b1)



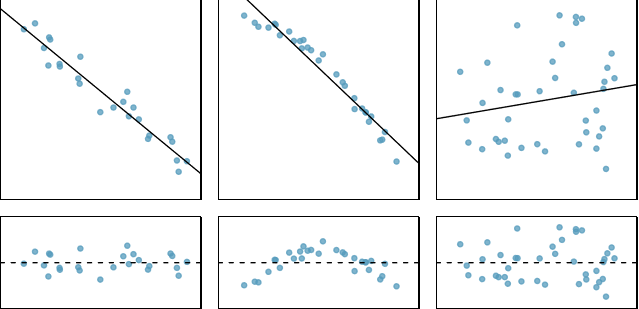
* head length + total length of 104 brushtail possums from Australia are associated.
* above average total length = tend to have above average head lengths.
* relationship is not perfectly linear, but could be helpful to partially explain connection between these variables w/ a straight line.
* **Straight lines should only be used when data appear to have a linear relationship** vs. a curved line



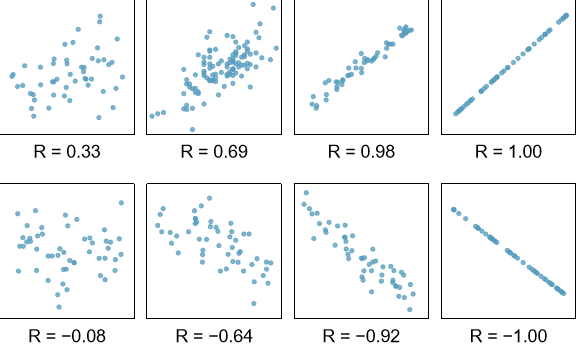
* If data show a nonlinear trend, use more advanced/complex, non-linear techniques
* We want to describe relationship between head length + total length using a line w/ total length as predictor, x, to predict head length, y 🡺 **y^ = 41 + 0.59x**
* equation predicts possum w/ total length = 80 cm will have a head length = **y^ = 41 + 0.59\*80 = 88.2**
* **y^ estimate** may be viewed as an **average** 🡺 predict possums w/ total length = 80 cm will have **an average head length = 88.2 mm**.
* **\*\*\*Absent further info about an 80 cm possum, prediction for head length that uses the average is a reasonable estimate\*\*\***
* **Residuals** = leftover variation in data after accounting for model ﬁt: **Data = Fit + Residual**
* **=** vertical distance from observation to the line, is positive (Each observation has a residual)
* If observation above regression line, residual = positive, if below the line, residual = negative
* **\*\*\*Goal in picking right linear model = have residuals to be as small as possible, in terms of its absolute value\*\*\***
* Residual = diﬀerence between observed + expected
* residual of the ith observation (xi, yi) = diﬀerence of observed response (yi) + response predicted based on model ﬁt (yi^ = identify by plugging xi into model) 🡺 
* Guided Practice 5.4: If a model underestimates an observation, will the residual be positive or negative? What about if it overestimates the observation?
* If underestimated true value, residual = positive, if overestimated, residual = negative
* Guided Practice 5.5 Compute residuals for observations (85.0, 98.6) + (95.5,94.0) using a linear relationship y^ = 41 + 0.59x
* 98.6 - 91.15 = 7.45 94 - 97.345 = -3.345
* Residuals = helpful in evaluating how well linear model ﬁts data set, often displayed in **residual plot**



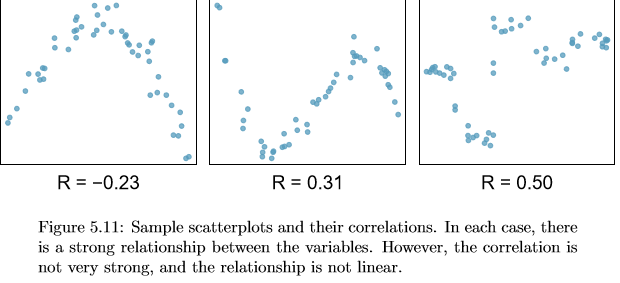
* Residuals are plotted @ original horizontal locations but w/ vertical coordinate = residual.
* residual plot = sort of like tipping scatterplot over so regression line is horizontal
* **\*\*\*1 purpose of residual plots** = **ID** characteristics/**patterns** **still apparent in data after ﬁtting a model**\*\*\*



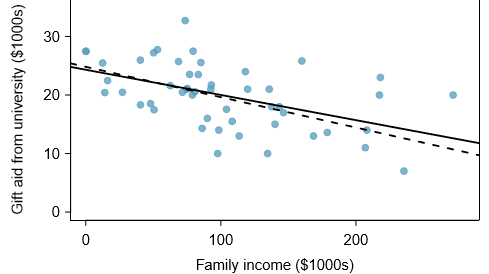
* + **Middle model’s residual plot has upside-down U pattern = non-linear structure of true *f***
  + 1st + 3rd datasets’ residuals = no obvious patterns + appear to be scattered randomly around dashed line that = 0
  + But, 3rd plot = very little upwards trend, so reasonable to try to ﬁt a linear model to the data.
  + However, unclear if there’s statistically signiﬁcant evidence the slope parameter is diﬀerent from 0.
  + **\*\*\*Point estimate of the slope parameter, b1,** is NOT = 0, but can wonder if this could just be due to chance\*\*\*
* **Correlation (r)** = strength of linear relationship (always between -1 and 1) between 2 variables.
* Can compute correlation via formula, but it’s complex () so generally perform calculations on a CPU

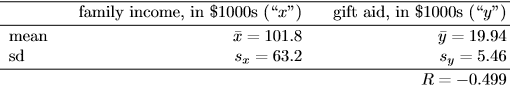
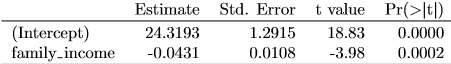


* **correlation** = intended to **quantify the strength of a linear trend**.
* **\*\*\*Nonlinear trends, even when strong, sometimes produce correlations that do not reﬂect the strength of the relationship\*\*\***



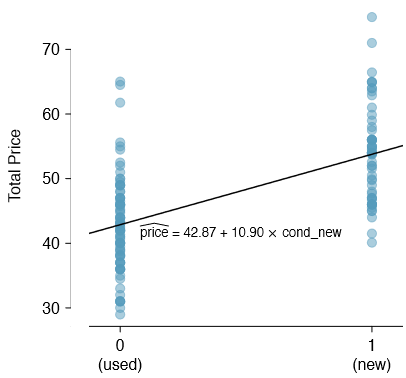
* Guided Practice: It appears no straight line would ﬁt any of the datasets represented above. Instead, try drawing nonlinear curves on each plot. Once you create a curve for each, describe what is important in the ﬁt.
  + Curve should be close to the DP’s such that residuals are minimized
* **Least squares regression** = more rigorous approach.



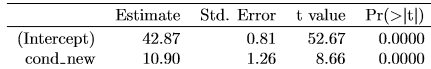
* family income + gift aid data from a random sample of 50 students in ‘11 freshman class of Elmhurst College in Illinois
* Negative trend/correlation🡺 students w/ higher family incomes tend to have lower gift aid
* Begin by thinking about “**best**” 🡺 Mathematically want a line w/ small residuals, so criterion could *minimize the sum of the residual magnitudes:*  (dashed line in above plot)
* More common = **choose line that minimizes SUM OF SQUARED RESIDUALS**: (solid line)
* most commonly used method.
* Computing this line = much easier by hand + in most statistical software.
* **\*\*\*In many applications, residual 2x as large as another residual == *more* than 2x as bad + squaring the residuals accounts for this discrepancy\*\*\***
* \*\*\*There ARE applications where *sum of the residual magnitudes* may be more useful, + there are plenty of other criteria we might consider\*\*\*
* Can write equation of least squares regression line as 
* Other than estimating parameters using observed data, can apply \*\*\*2 properties of the least squares line:\*\*\*
* Slope of least squares line can be estimated by  where sx + sy = sample SD’s
* **If x­- = mean of the predictor + y¯ = mean of outcome, point (x¯, y¯) is ON the least squares line**
* Use b­­0 + b1 as **point estimates** of parameters β0 + β1.. However, this book only applies the least squares criterion.
* Guided Practice:  🡺 How to plot point (101.8,19.94) on to verify it falls on the least squares line
  + **draw straight line up from x-value = ~101 + draw a horizontal line @ y = ~20 + these lines should intersect on the least squares line**
* Guided Practice: Using summary statistics, compute slope for regression line of gift aid against family income
  + **-.499\*(5.46/63.2) == -0.04310981012**
* Given slope + point on a line, (x0,y0), equation for line can be written as
* Common exercise to become more familiar w/ foundations of least squares regression = use basic summary stats + point-slope form to produce least squares line.
* TIP: IDing least squares line from summary statistics:
  + Estimate slope parameter, β1, by calculating b1 using 
  + Noting point (x¯, y¯) is ON least squares line, use x0 = x¯ and y0 = y¯ along w/ slope b1 in the point-slope equation: 
  + Simplify the equation.
* \*\*Ex\*\*: Using point (101.8,19.94) from the **sample means** + slope estimate b1 = −0.0431, ﬁnd least-squares line for predicting aid based on family income.
* Apply **point-slope equation** using the DP + slope b1:  = 
* Expanding RHS + adding 19.94 to each side to simplify: 
* CPU output: 
  + 1st row = β0, 2nd row = β1, \*\*\***1st col = point estimate for β1, 2nd col, 2nd row = standard error for a point estimate** (0.0108), **3rd col = t-test statistic for the null H0: β1 = 0** (−3.98), **last col = p-value for the t-test statistic for the null H0: β1 = 0 + the 2-sided H1 (0.0002)**
* Suppose a high school senior is considering Elmhurst College. Can she simply use the linear equation that we have estimated to calculate her ﬁnancial aid from the university?
  + \*\*\*May use it as an **ESTIMATE**, though some qualiﬁers are important 🡺 data all come from 1 freshman class, the way aid is determined by the university may change year to year\*\*\*
  + Also, \*\*\***equation will provide an imperfect estimate**\*\*\* = linear equation is good at capturing the trend in the data, but no individual aid will be perfectly predicted
* **\*\*\*Interpreting parameters** in a regression model = often one of the most important steps in the analysis\*\*\*
* Example 5.18: Slope + intercept estimates for Elmhurst data = -0.0431 and 24.3. What do these numbers really mean?
* Family income = $0 predicts **AVERAGE** financial aid = $24.3K, + for each $1K increase in family income, financial aid would **ON AVERAGE** decreases in -.0431 financial aid units ($1K) = −$43.10
* higher family income corresponds to less aid b/c coeﬃcient of family income < 0 in model.
* \*\*\*Must be cautious in this interpretation: while there is a *real* association, **we cannot interpret a causal connection between the variables because these data are observational\*\*\***
  + **Increasing a student’s family income may not cause the student’s aid to drop.** (reasonable to contact + ask if relationship is causal/if aid decisions are partially based on family income)
* **\*\*\*Meaning of the intercept is relevant to this application since family income for some students at Elmhurst is $0\*\*\***
* **\*\*\*In other applications, intercept may have little/no practical value if there are no observations where x is near 0\*\*\***
* Interpreting parameters estimated by least squares: **Slope** describes estimated diﬀerence in outcome if predictor for a case happened to be 1 unit larger. **Intercept** describes average outcome if predictor = 0 + **that the linear model is valid all the way to x = 0** (in many applications = not case)
* **\*\*\*Extrapolation is treacherous\*\*\***
* Linear models have real limitations 🡺 it’s simply a **modeling framework** + the **truth = almost always much more complex** than simple line (don’t know how data outside of training will behave)
* Use aid model income to estimate aid of student w/ family income = $1 million (Recall units of family income = $1K’, so calculate aid as 
* Model predicts student will have -$18,800 in aid, i.e. pay extra on top of tuition to attend
* **Extrapolation** = applying a model estimate to values outside the realm of the original data
* Generally, a **linear model is only an *approximation*** of **real relationship** **between** **variables**.
* \*\*\*Extrapolating = making an \*\*\***unreliable**\*\*\* bet that the *approximate* linear relationship will be valid in places where it has not been explored\*\*\*
* Compared to R, more common to explain strength of a linear ﬁt using, R-squared, R2
* If provided a linear model, might like to describe how closely data cluster around linear ﬁt + **R2 describes amount of variation in the response explained by the least squares line**.
* Ex: Aid data 🡺 variance of response = 
* Applying our least squares line, model reduces uncertainty in predicting aid using a student’s family income.
* **\*\*\*Variability in *residuals* describes how much variation remains *after* using model\*\*\***  == In short, there was a reduction of:



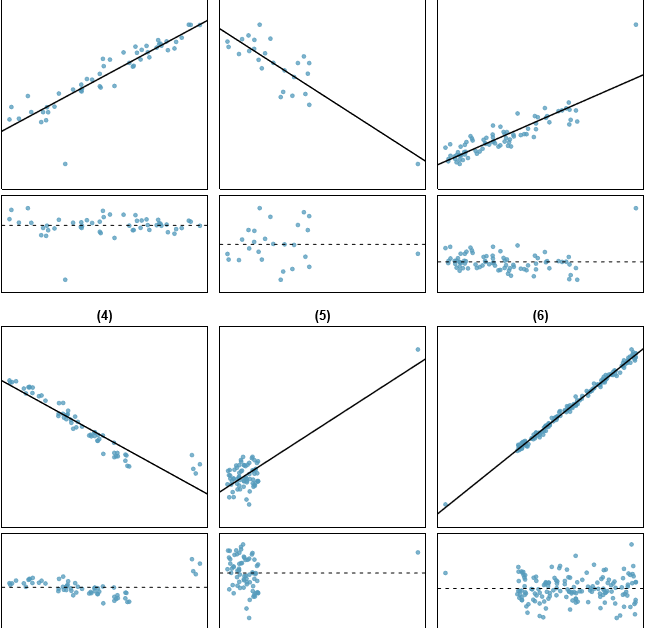
* i.e. we get ~25% reduction in data’s variation by using info about family income for predicting aid *using a linear model.*
* This corresponds exactly to R2 = 0.25, compared to R = −0.499
* Guided Practice: If a linear model has a very strong negative relationship w/ correlation r = -0.97, how much variation in the response is explained by the explanatory variable?
* Compute R2 🡪 -.972 = .9409 🡺 ~94% of variation in response is explained by model
* **Categorical variables** are also useful in predicting outcomes.
* Ex: categorical predictor w/ 2 levels 🡺 Ebay auctions for a video game, Mario Kart for Nintendo Wii, where both total price of auction + condition of the game were recorded
* Want to predict total price based on **game condition = used and new**.



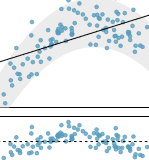
* To incorporate game condition into a regression, **convert the categories into a numerical form** using an **indicator (dummy) variable** called condnew, which takes value = 1 when game is new + 0 when game is used.
* Using this indicator variable, linear model is 



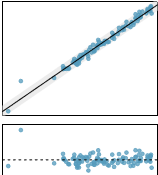
* Interpret the 2 parameters estimated in the model for the price of Mario Kart in eBay auctions.
  + Intercept value of 42.87 estimates that, on average, used games selling price = $42.87, and if new, on average, new games selling price = $42.87 + $10.9 = $53.77
  + Coefficient for indicator says on average, new games sell for ~$10.90 more than used games
* **\*\*\*TIP: Interpreting model estimates for categorical predictors. Estimated intercept = value of response for 1st category (i.e. category corresponding to indicator value = 0) + estimated slope = avg. change in response between the 2 categories\*\*\***
* **Outliers** in regression = observations that fall **far** from the “cloud” of points
* **\*\*\*Outliers** = especially important b/c they **can have a strong inﬂuence on least squares line\*\*\***
* 6 plots, each w/ fit + residuals (each dataset has *at least 1* outlier)



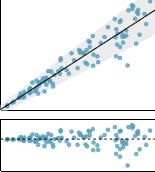
* (1) 1 outlier, only appears to *slightly* inﬂuence the line
* (2) 1 outlier, though quite close to least squares line = suggests not very inﬂuential
* (3) 1 outlier, appears to pull least squares line up on the right 🡺 see how line around the primary cloud doesn’t appear to ﬁt very well
* (4) 4 outliers, appears to be inﬂuencing line somewhat strongly (pulling upwards), making least square line ﬁt poorly almost everywhere (**might be interesting explanation 🡪 investigate**)
* (5) No obvious trend in main cloud + **outlier appears to largely control the slope** 
  + *data w/ no clear trend were assigned a line w/ a large trend simply due to one outlier*
* (6) 1 outlier far from the cloud, but falls quite close to line + doesn’t appear to be very inﬂuential
* **\*\*\*Leverage Points** = fall *horizontally* away from the center of the cloud + **tend to pull harder on the line 🡺 call them *high leverage points\*\*\****
* **High leverage points can strongly inﬂuence slope of a least squares line.**
* **Inﬂuential Point 🡺** ahigh leverage DP *DOES* appear to inﬂuence slope of line (cases (3), (4), (5))
* \*\*\*Usually can say a point is **inﬂuential** if, *had we ﬁtted the line without it*, the inﬂuential point would have been unusually far from the least squares line\*\*\*.
* \*\*\*Tempting to remove outliers 🡺 **Don’t do this without a very good reason**\*\*\*
* Models that ignore exceptional (+ interesting) cases often perform poorly.
* Ex: If financial ﬁrm ignored largest market swings (“outliers”), they’d soon go bankrupt by making poorly thought-out investments.
* Caution: **Don’t ignore outliers when ﬁtting a ﬁnal model**
* Outliers should NOT be removed or ignored **without a good reason**.
* Whatever ﬁnal model is ﬁt to the data would NOT be very helpful if it ignores the most exceptional cases
* **Caution: Outliers for a categorical predictor w/ 2 levels Be cautious about using categorical predictor when 1 level = very few observations as the few observations become inﬂuential points**
* Inference for linear regression
* **There is uncertainty** in estimates of slope + y-intercept for a regression line.
* **\*\*\*When performing inference on a least squares line, we generally require the following:\*\*\***
* **Linearity**: Data should show linear trend + if nonlinear, apply more advanced regressions

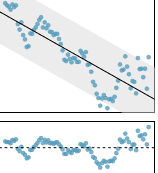


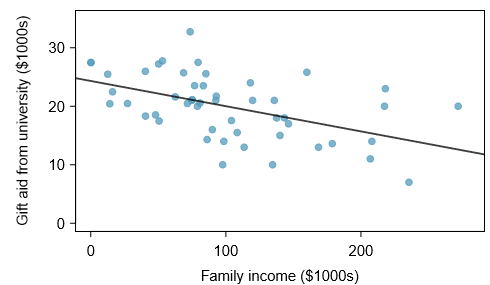
* **Nearly normal residuals**: When this = unreasonable, it is ***usually b/c of outliers or concerns about inﬂuential points***.

 = non-normal residuals due to 2 outliers

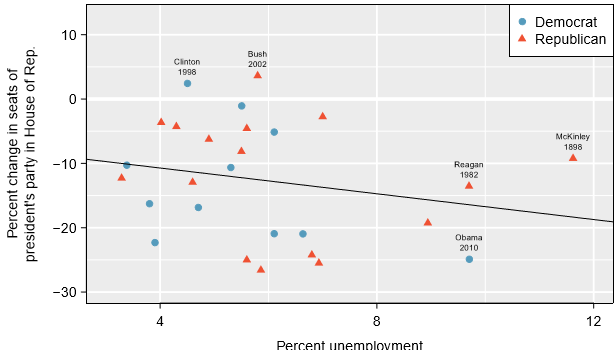
* **Constant variability** = *variability of points around least squares line remains roughly constant.*



* **Independent observations**. Be cautious about applying regression to data collected *sequentially* in a **time series**
  + Such data may have an underlying structure that should be considered in a model/analysis
  + Ex: time series where independence is violated 🡺 
* Should we have concerns about applying inference to the Elmhurst data?

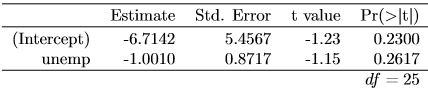


* + trend appears to linear, data fall around the line (no obvious outliers), variance = roughly constant
  + also NOT time series observations 🡺 would be reasonable to analyze model w/ **inference**
* House of Reps Elections occur every 2 years, coinciding every 4 years w/ Presidential election + House elections in middle of a Presidential term = **midterm elections**.
* In America’s 2-party system, 1 political theory suggests higher unemployment rate = worse President’s party will do in the midterm elections.
* To assess validity of this claim, compile historical data + look for a connection, considering every midterm election from 1898-2010, w/ exception of elections during Great Depression.

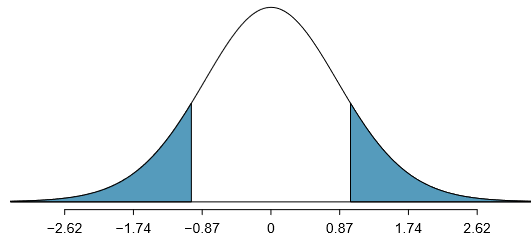


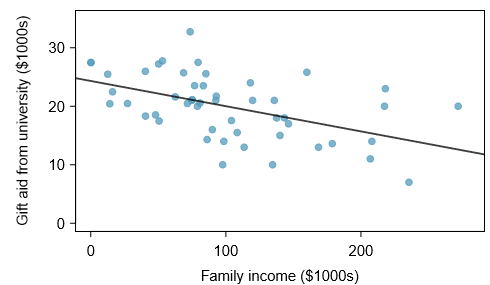
* Consider % change in # of seats of President’s party (e.g. % change # of seats for Democrats in 2010) against unemployment rate.
* Examining data 🡺 no clear deviations from linearity, constant variance condition, or normality of residuals (don’t examine a normal probability plot here) + while data are collected sequentially, a separate analysis made sure of no any apparent correlation between successive observations
* Guided Practice: Data for Great Depression (1934 + 1938) were removed b/c unemployment rate was 21% + 18%, respectively. Do you agree they should be removed for this investigation?
  + Yes, b/c very unusual numbers that could have large influence on model
  + Each of these DP’s would have **very high leverage** on ANY least-squares regression line, + **years w/ such high unemployment may not help us understand what would happen in other years where unemployment is only modestly high**
  + On the other hand, these are exceptional cases, + we’d be discarding important info if excluded from a ﬁnal analysis
* Negative slope but \*\*\*it (+ the y-intercept) are only **estimates** of the parameter values\*\*\*.
* Is this convincing evidence that the “true” linear model ***f*** has a negative slope? 🡺 Do the data provide strong evidence the political theory is accurate?
* **\*\*\*Frame investigation into a 2-sided statistical hypothesis test**\*\*\*
  + *2-sided test since a statistically signiﬁcant result in either direction would be interesting*
* **H0: β1 = 0 🡺 true linear model has slope = 0.**
* **HA: β1 <> 0 🡺 true linear model has a slope diﬀerent than 0 + higher unemployment = greater loss for President’s party in the House, or vice-versa**
* **\*\*\*Reject H0 in favor of HA if data provide strong evidence the true slope parameter < 0\*\*\***
* To assess the hypotheses, we ID a **standard error for the estimate**, **compute** an appropriate **test statistic,** + **ID the p-value**
* Generally label test statistic w/ ***t***, since it **follows the t-distribution**.



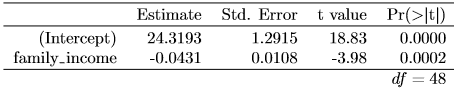
* Ex: What do the 1st + 2nd columns represent?
* **1st col = least squares *estimates*, b0 + b1, of the parameters β0 + β1**
* **2nd col = standard errors of *each estimate***
* In the hypotheses, null value for slope = 0, so compute the test statistic using the T (or Z) score formula: 
* Look for the 2-tailed p-value using the probability table for the t-distribution



* + **See sampling distribution for b1, \*\*\**if the null hypothesis was true*\*\*\*.**
  + Shaded tail = p-value for the hypothesis test evaluating for convincing evidence higher unemployment corresponds to greater loss of House seats for President’s party in midterms
* dF = 25 gives absolute value of the test statistic (1.15) is smaller than any value listed 🡺 means the tail area + therefore also the p-value *is larger* than 0.200 (2 tails).
* B/c p-value is so large, we **fail to reject the null**.
* That is, the data do NOT provide convincing evidence that unemployment = good predictor of how well a president’s party will do in midterms for the House
* Can also ID the t-test statistic from the software output 🡺 2nd row (unemp) + 3rd column (t-value).
* 2nd row + last column = p-value for the 2-sided hypothesis test where the null value = 0
* **Inference for regression**: Usually rely on statistical software to ID **point estimates**+ **standard errors** for parameters of a regression line.
* After verifying conditions hold for ﬁtting a line, can use the methods for the t-distribution to create CI’s for regression parameters or to evaluate hypothesis tests.
* \*\*\***Caution: Don’t carelessly use the p-value from regression output**\*\*\*
* last column in regression output often lists p-values for *one particular hypothesis* = 2-sided test where null value = 0.
* If hypothesis test should be ONE-sided or a **comparison** is being made to *a value other than zero*, be cautious about using the software output to obtain the p-value
* Ex: How sure are we the slope is statistically signiﬁcantly diﬀerent from 0 for Elmhurst data? That is, do we think a formal hypothesis test would reject the claim that true slope of the line should be = 0?

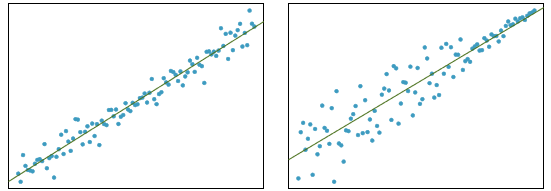
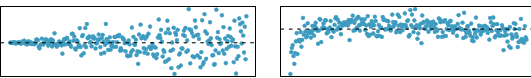


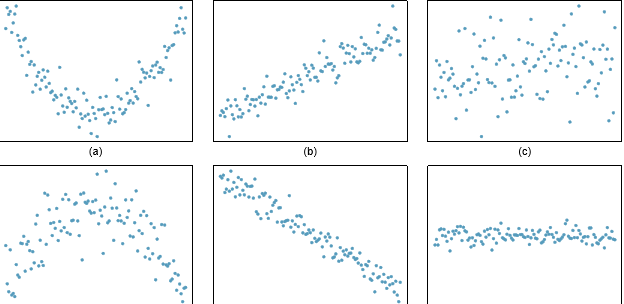
* While relationship is not *perfect*, there IS an evident decreasing trend in the data 🡺 suggests the hypothesis test will reject the null claim that slope = 0
* Guided Practice: Use below to formally evaluate the following hypotheses.
* H0: True coeﬃcient for family income = 0 HA: True coeﬃcient for family income is NOT zero



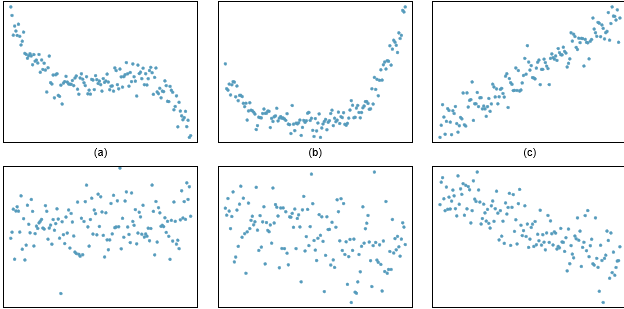
* Point estimate of slope of line = -0.0431, standard error of this estimate = 0.0108, + t-test statistic = -3.98
* **p-value corresponds exactly to the 2-sided test we’re interested in 🡺 0.0002 = so small that we reject the null + conclude family income + ﬁnancial aid at Elmhurst College for freshman entering in the year 2011 are negatively correlated + true slope parameter is indeed < 0,**
* **TIP: Always check assumptions** 🡺 If conditions for a regression line do NOT hold, methods presented above should NOT be applied.
* **Standard error/distribution assumption of the point estimate (assumed to be normal when applying the t-test statistic) may not be valid.**
* We considered t-statistic as a way to evaluate strength of evidence for a hypothesis test
* Could focus on R2 (proportion of variability in response explained by predictor)
* If this R2 proportion is large 🡺 suggests linear relationship exists between the variables.
* If small 🡺 evidence provided by the data may not be convincing.
* Concept of **considering amount of variability in response explained by predictor** is a key component in some statistical techniques.
  + **Analysis of variance (ANOVA)** uses this general principle.
* The method states that **if enough variability is explained away by the categories, we’d conclude the mean varied between the categories.**
* **On the other hand, we might NOT be convinced if only a little variability is explained.**
* ANOVA can be further employed in advanced regression modeling to evaluate the of explanatory variables

# 5.5 Exercises

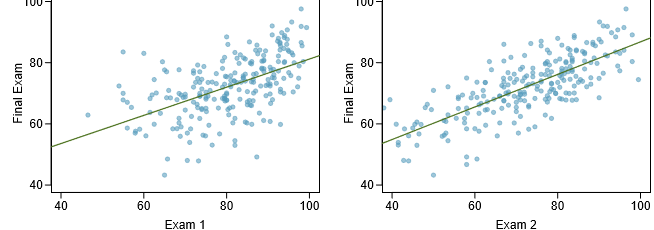
* Visualize the residuals: Scatterplots below each have a superimposed regression line. If we were to construct a residual plot (residuals vs. x) for each, describe what those plots would look like.
  + - 
  + 1) Residuals = cloud of points huddled around horizontal line = 0 w/ constant variance, as regression line is quite close to majority of DP’s
  + 2) Residuals = more scattered around horizontal line = 0 for lower values of X, huddling close together as X increases (tail-end of plot) == fan-shape pattern == bad model fit
* Trends in residuals. Shown below = 2 plots of residuals remaining after ﬁtting a linear model to 2 diﬀerent sets of data. Describe important features + determine if a linear model would be appropriate for these data. Explain reasoning.
  + 
  + 1) Fan-shape = a pattern, linear model not a good choice == **Strong relationship, but it’s non-linear**
  + 2) DP’s have mostly constant variance after initial underestimation in the lower values of X
* Identify relationships:. For each plot, ID strength of the relationship + whether ﬁtting a linear model would be reasonable.



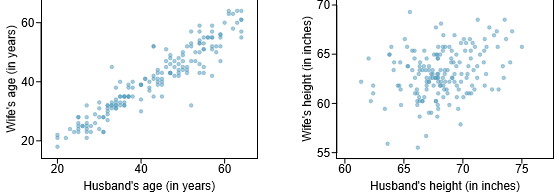
* + 1 = strong, non-linear 2 = moderately strong, linear 3 = weak*, could use* linear
  + 4 = weak-moderate, non-linear 5 = strong, linear 6 = no/weak relationship, linear



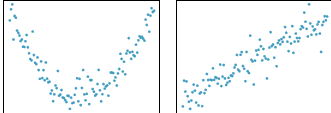
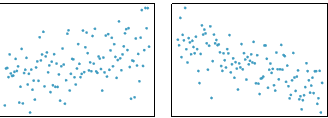
* + 1 = strong, non-linear, 2 = strong, non-linear 3 = strong, linear
  + 4 = moderate, linear could work 5 = weak, linear could work 6 = moderate, linear
* 2 scatterplots show relationship between ﬁnal + mid-semester exam grades recorded during several years for a course at a university.



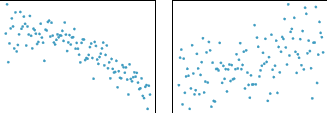
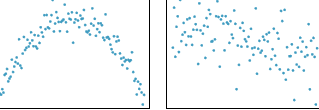
* + (a) Based on these graphs, which of the 2 exams has the strongest correlation w/ the ﬁnal?
    - Exam 2, as the DP’s are more closely aligned to the regression line
  + Can you think of a reason why correlation between exam chosen above + the ﬁnal is higher?
    - Further on in course = more knowledge of what final exam entailed
* The Great Britain Oﬃce of Population Census + Surveys once collected data on a random sample of 170 married couples in Britain, recording age (in years) + heights (converted here to inches) of the husbands + wives. Left scatterplot = wife’s age plotted against husband’s age, Right = wife’s height plotted against husband’s height.



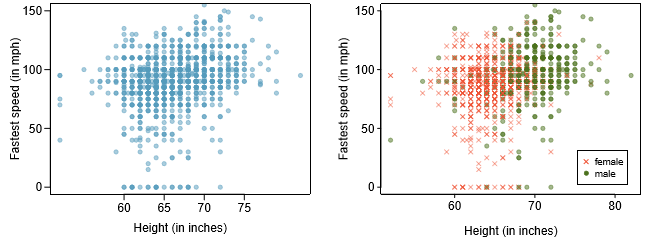
* + (a) Describe the relationship between husbands’ and wives’ ages.
    - Strong, positive relationship (as husband ages, wife ages linearly)
  + (b) Describe the relationship between husbands’ and wives’ heights.
    - No real relationship, possibly weak linear
    - As husband’s grow taller, it’s not necessarily true that wives do as well
  + (c) Which plot shows a stronger correlation? Explain your reasoning.
    - Plot 1 = more closely aligns with a line, plot 2 = more of a random scatter
  + (d) Data on heights were originally collected in cm, + then converted to inches. Does this conversion aﬀect the correlation between husbands’ and wives’ heights?
    - **\*\*\*No, Change of units will NOT impact mathematical relationship\*\*\***
* Match calculated correlations to corresponding scatterplot.

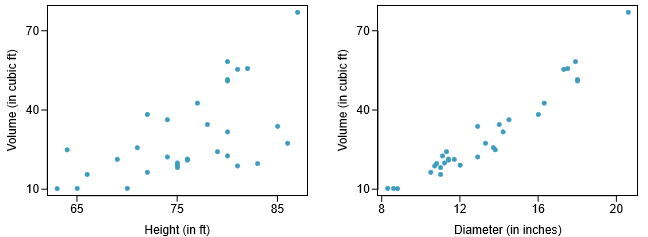
 

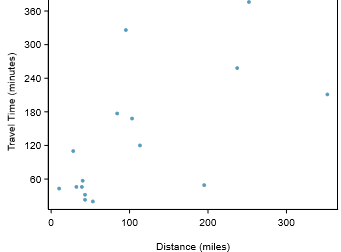
* + (a) R = −0.7 (b) R = 0.45 (c) R = 0.06 (d) R = 0.92
  + **1 = c 2 = 4 3 = b 4 = a**

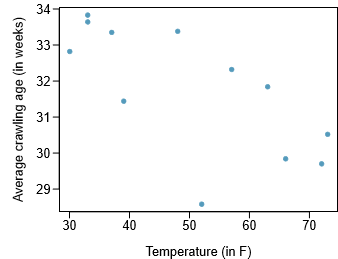
* + . (a) R = 0.49 (b) R = −0.48 (c) R = −0.03 (d) R = −0.85
  + **1 = d 2 = a 3 = c 4 = b**
* 1302 students ﬁlled out a survey, asked about height, fastest speed ever driven, + gender. Left scatterplot = relationship between height + fastest speed, Right scatterplot = breakdown by gender in this relationship.



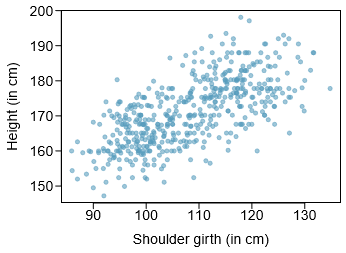
* + Describe the relationship between height and fastest speed
    - Seems relationship is quite weak, but positive, but w/ some *mayyybe* some influence: taller = more likely to drive faster
    - Many students have not driven a car (DP’s at 0 for a value of Y (speed))
    - Leverage DP’s = low height, all the same (error?) + some very tall
  + Why do you think these variables are positively associated?
    - No real clear reason in 1st plot
  + What role does gender play in the relationship between height and fastest driving speed?
    - **\*\*\*2nd plot = reveals confounder of gender 🡪 taller = more male, males = more likely to drive fast**
    - **Seems a binary classification line can bisect the cloud into males vs. females w/ males = higher speeds\*\*\***
    - **Males > females on avg. for height, also for speed (anecdotal)** 
      * **\*\*\*sociological studies conﬁrm this anecdote\*\*\***
* Scatterplots show relationship between height, diameter, + volume of timber in 31 felled black cherry trees. The diameter of the tree is measured 4.5 feet above the ground
  + - * 
  + (a) Describe the relationship between volume + height or diameter of these trees.
    - Volume + height = weak/moderate, linear, positive
    - Volume + diameter = linear, strong, positive
  + (b) Suppose you have height + diameter measurements for another black cherry tree. Which variable would be preferable to use to predict volume of timber in this tree using a simple linear regression model? Explain your reasoning.
    - Diameter b/c the relationship would have a larger correlation (more closely resembles a line)
* The Coast Starlight Amtrak train runs from Seattle to LA. Scatterplot displays distance between each stop (in miles) + amount of time it takes to travel from 1 stop to another (in minutes).



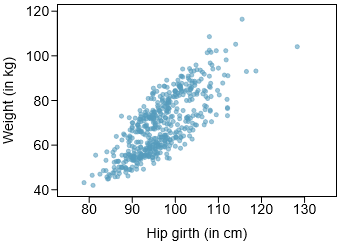
* + (a) Describe the relationship between distance and travel time.
    - Very weak, could use linear 🡺 **\*\*\*Note: small cluster in lower-left\*\*\***
    - **Did not expect such a weak relationship THB**
  + (b) How would the relationship change if travel time was instead measured in hours, and distance was instead measured in KM?
    - Would decrease range of y-axis, increase x-axis
    - **\*\*\*Would NOT change form, direction, or strength of relationship\*\*\***
      * If longer distances in miles = associated w/ longer travel time in minutes, longer distances in KM will be associated w/ longer travel time in hours.
  + (c) Correlation between travel time (in miles) + distance (in minutes) is R = 0.636. What is the correlation between travel time (in kilometers) and distance (in hours)?
    - **\*\*\*Same (changing units does NOT affect correlation)\*\*\***
* Study conducted @ University of Denver investigated whether babies take longer to learn to crawl in cold months, when often bundled in clothes that restrict movement, than in warmer months. Infants born during study year were split into 12 groups, 1 per month. Consider average crawling age of babies in each group against average temperature when babies = 6 months old (when babies often begin trying to crawl). Temperature is measured in Fahrenheit (◦F) + age in weeks.



* + (a) Describe the relationship between temperature and crawling age.
    - Appears to be a moderate, negative correlation/relationship
    - Suggests as temps increase, **on average,** the average crawling age in weeks would decrease, supporting the hypothesis above
    - 1 outlier of very low average crawling age in weeks @ around low 50’s F
  + (b) How would the relationship change if temperature was measured in degrees Celsius (◦C) and age was measured in months?
    - **changing units does NOT affect correlation**
  + (c) The correlation between temperature in ◦F and age in weeks was R = −0.70. If we converted the temperature to ◦C and age to months, what would the correlation be?
    - **Same (changing units does NOT affect correlation)**
* Researchers studying anthropometry collected body girth + skeletal diameter measurements, as well as age, weight, height, + gender for 507 physically active individuals. Scatterplot shows relationship between height + shoulder girth (over deltoid muscles), both measured in cm.



* + (a) Describe the relationship between shoulder girth and height.
    - Quite strong positive relationship = as girth increases, can assume higher height, (both in cm) *on average*
  + (b) How would the relationship change if shoulder girth was measured in inches while the units of height remained in centimeters?
    - **\*\*\*Same (changing units, \*\*\*even if just for 1 of variable\*\*\* does NOT affect correlation)\*\*\***
* Scatterplot shows relationship between weight measured in kg + hip girth measured in cm from the data described



* + (a) Describe the relationship between hip girth and weight.
    - Quite a strong, positive relationship but some outliers appear to minorly influence relationship
  + (b) How would the relationship change if weight was measured in pounds while the units for hip girth remained in centimeters?
    - **\*\*\*Same (changing units, \*\*\*even if just for 1 of variable\*\*\* does NOT affect correlation)\*\*\***
* What would be the correlation between ages of husbands + wives if men *always married woman who were*
  + (a) 3 years younger? (b) 2 years older? (c) half as old as themselves?
    - Model as: ageH = ageW + 3; (b) ageH = ageW −2; and (c) ageH = 2×ageW
    - **correlation will be exactly 1 in all three parts**
    - as husband ages 1 year, so will wife, no matter starting age
* What would be the correlation between annual salaries of males + females at a company if for a certain type of position men always made
  + (a) $5k more than women? (b) 25% more than women? (c) 15% less than women?
    - Models: salaryM = salaryW + 5000; salaryM = 1.25\*salaryW; salaryM = .75\*salaryW
    - **Same absolute value for all (1), for a) and b) = positive, for c) = negative**
* The Association of Turkish Travel Agencies reports the number of foreign tourists visiting Turkey and tourist spending by year.19 The scatterplot below shows the relationship between these two variables along with the least squares ﬁt. (a) Describe the relationship between number of tourists and spending. (b) What are the explanatory and response variables? (c) Why might we want to ﬁt a regression line to these data? (d) Do the data meet the conditions required for ﬁtting a least squares line? In addition to the scatterplot, use the residual plot and histogram to answer this question.
* 19Association of Turkish Travel Agencies, Foreign Visitors Figure & Tourist Spendings By Years.
* 252 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION
* 5.18 Nutrition at Starbucks, Part I. The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain.20 Since Starbucks only lists the number of calories on the display items, we are interested in predicting the amount of carbs a menu item has based on its calorie content. (a) Describe the relationship between number of calories and amount of carbohydrates (in grams) that Starbucks food menu items contain. (b) In this scenario, what are the explanatory and response variables? (c) Why might we want to ﬁt a regression line to these data? (d) Do these data meet the conditions required for ﬁtting a least squares line?
* 5.19 The Coast Starlight, Part II. Exercise 5.11 introduces data on the Coast Starlight Amtrak train that runs from Seattle to Los Angeles. The mean travel time from one stop to the next on the Coast Starlight is 129 mins, with a standard deviation of 113 minutes. The mean distance traveled from one stop to the next is 107 miles with a standard deviation of 99 miles. The correlation between travel time and distance is 0.636. (a) Write the equation of the regression line for predicting travel time. (b) Interpret the slope and the intercept in this context. (c) Calculate R2 of the regression line for predicting travel time from distance traveled for the Coast Starlight, and interpret R2 in the context of the application. (d) The distance between Santa Barbara and Los Angeles is 103 miles. Use the model to estimate the time it takes for the Starlight to travel between these two cities. (e) It actually takes the the Coast Starlight about 168 mins to travel from Santa Barbara to Los Angeles. Calculate the residual and explain the meaning of this residual value. (f) Suppose Amtrak is considering adding a stop to the Coast Starlight 500 miles away from Los Angeles. Would it be appropriate to use this linear model to predict the travel time from Los Angeles to this point?
* 20Source: Starbucks.com, collected on March 10, 2011, http://www.starbucks.com/menu/nutrition.
* 5.5. EXERCISES 253
* 5.20 Body measurements, Part III. Exercise 5.13 introduces data on shoulder girth and height of a group of individuals. The mean shoulder girth is 108.20 cm with a standard deviation of 10.37 cm. The mean height is 171.14 cm with a standard deviation of 9.41 cm. The correlation between height and shoulder girth is 0.67.
* (a) Write the equation of the regression line for predicting height. (b) Interpret the slope and the intercept in this context. (c) Calculate R2 of the regression line for predicting height from shoulder girth, and interpret it in the context of the application. (d) A randomly selected student from your class has a shoulder girth of 100 cm. Predict the height of this student using the model. (e) The student from part (d) is 160 cm tall. Calculate the residual, and explain what this residual means. (f) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?
* 5.21 Helmets and lunches. The scatterplot shows the relationship between socioeconomic status measured as the percentage of children in a neighborhood receiving reduced-fee lunches at school (lunch) and the percentage of bike riders in the neighborhood wearing helmets (helmet). The average percentage of children receiving reduced-fee lunches is 30.8% with a standard deviation of 26.7% and the average percentage of bike riders wearing helmets is 38.8% with a standard deviation of 16.9%. (a) If the R2 for the least-squares regression line for these data is 72%, what is the correlation between lunch and helmet? (b) Calculate the slope and intercept for the leastsquares regression line for these data. (c) Interpret the intercept of the least-squares regression line in the context of the application. (d) Interpret the slope of the least-squares regression line in the context of the application. (e) What would the value of the residual be for a neighborhood where 40% of the children receive reduced-fee lunches and 40% of the bike riders wear helmets? Interpret the meaning of this residual in the context of the application.
* 5.5.3 Types of outliers in linear regression
* 5.22 Outliers, Part I. Identify the outliers in the scatterplots shown below, and determine what type of outliers they are. Explain your reasoning.
* 5.23 Outliers, Part II. Identify the outliers in the scatterplots shown below and determine what type of outliers they are. Explain your reasoning.
* 5.24 Crawling babies, Part II. Exercise 5.12 introduces data on the average monthly temperature during the month babies ﬁrst try to crawl (about 6 months after birth) and the average ﬁrst crawling age for babies born in a given month. A scatterplot of these two variables reveals a potential outlying month when the average temperature is about 53◦F and average crawling age is about 28.5 weeks. Does this point have high leverage? Is it an inﬂuential point?
* 5.25 Urban homeowners, Part I. The scatterplot below shows the percent of families who own their home vs. the percent of the population living in urban areas in 2010.21 There are 52 observations, each corresponding to a state in the US. Puerto Rico and District of Columbia are also included.
* (a) Describe the relationship between the percent of families who own their home and the percent of the population living in urban areas in 2010. (b) The outlier at the bottom right corner is District of Columbia, where 100% of the population is considered urban. What type of outlier is this observation?
* 5.5.4 Inference for linear regression
* Visually check the conditions for ﬁtting a least squares regression line, but you do not need to report these conditions in your solutions unless it is requested.
* 21United States Census Bureau, 2010 Census Urban and Rural Classiﬁcation and Urban Area Criteria and Housing Characteristics: 2010.
* 5.5. EXERCISES 255
* 5.26 Nutrition at Starbucks, Part II. Exercise 5.18 introduced a data set on nutrition information on Starbucks food menu items. Based on the scatterplot and the residual plot provided, describe the relationship between the protein content and calories of these menu items, and determine if a simple linear model is appropriate to predict amount of protein from the number of calories.
* 5.27 Grades and TV. Data were collected on the number of hours per week students watch TV and the grade they earned in a biology class on a 100 point scale. Based on the scatterplot and the residual plot provided, describe the relationship between the two variables, and determine if a simple linear model is appropriate to predict a student’s grade from the number of hours per week the student watches TV.
* 256 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION
* 5.28 Beer and blood alcohol content. Many people believe that gender, weight, drinking habits, and many other factors are much more important in predicting blood alcohol content (BAC) than simply considering the number of drinks a person consumed. Here we examine data from sixteen student volunteers at Ohio State University who each drank a randomly assigned number of cans of beer. These students were evenly divided between men and women, and they diﬀered in weight and drinking habits. Thirty minutes later, a police oﬃcer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood.22 The scatterplot and regression table summarize the ﬁndings.
* (a) Describe the relationship between the number of cans of beer and BAC. (b) Write the equation of the regression line. Interpret the slope and intercept in context. (c) Do the data provide strong evidence that drinking more cans of beer is associated with an increase in blood alcohol? State the null and alternative hypotheses, report the p-value, and state your conclusion. (d) The correlation coeﬃcient for number of cans of beer and BAC is 0.89. Calculate R2 and interpret it in context. (e) Suppose we visit a bar, ask people how many drinks they have had, and also take their BAC. Do you think the relationship between number of drinks and BAC would be as strong as the relationship found in the Ohio State study?
* 5.29 Body measurements, Part IV. The scatterplot and least squares summary below show the relationship between weight measured in kilograms and height measured in centimeters of 507 physically active individuals.
* (a) Describe the relationship between height and weight. (b) Write the equation of the regression line. Interpret the slope and intercept in context. (c) Do the data provide strong evidence that an increase in height is associated with an increase in weight? State the null and alternative hypotheses, report the p-value, and state your conclusion. (d) The correlation coeﬃcient for height and weight is 0.72. Calculate R2 and interpret it in context. 22J. Malkevitch and L.M. Lesser. For All Practical Purposes: Mathematical Literacy in Today’s World. WH Freeman & Co, 2008.
* 5.5. EXERCISES 257
* 5.30 Husbands and wives, Part II. Exercise 5.6 presents a scatterplot displaying the relationship between husbands’ and wives’ ages in a random sample of 170 married couples in Britain, where both partners’ ages are below 65 years. Given below is summary output of the least squares ﬁt for predicting wife’s age from husband’s age.
* \
* (a) We might wonder, is the age diﬀerence between husbands and wives consistent across ages? If this were the case, then the slope parameter would be β1 = 1. Use the information above to evaluate if there is strong evidence that the diﬀerence in husband and wife ages diﬀers for diﬀerent ages. (b) Write the equation of the regression line for predicting wife’s age from husband’s age. (c) Interpret the slope and intercept in context. (d) Given that R2 = 0.88, what is the correlation of ages in this data set? (e) You meet a married man from Britain who is 55 years old. What would you predict his wife’s age to be? How reliable is this prediction? (f) You meet another married man from Britain who is 85 years old. Would it be wise to use the same linear model to predict his wife’s age? Explain.
* 5.31 Husbands and wives, Part III. The scatterplot below summarizes husbands’ and wives’ heights in a random sample of 170 married couples in Britain, where both partners’ ages are below 65 years. Summary output of the least squares ﬁt for predicting wife’s height from husband’s height is also provided in the table.
* (a) Is there strong evidence that taller men marry taller women? State the hypotheses and include any information used to conduct the test. (b) Write the equation of the regression line for predicting wife’s height from husband’s height. (c) Interpret the slope and intercept in the context of the application. (d) Given that R2 = 0.09, what is the correlation of heights in this data set? (e) You meet a married man from Britain who is 5’9” (69 inches). What would you predict his wife’s height to be? How reliable is this prediction? (f) You meet another married man from Britain who is 6’7” (79 inches). Would it be wise to use the same linear model to predict his wife’s height? Why or why not?
* 258 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION
* 5.32 Urban homeowners, Part II. Exercise 5.25 gives a scatterplot displaying the relationship between the percent of families that own their home and the percent of the population living in urban areas. Below is a similar scatterplot, excluding District of Columbia, as well as the residuals plot. There were 51 cases.
* (a) For these data, R2 = 0.28. What is the correlation? How can you tell if it is positive or negative? (b) Examine the residual plot. What do you observe? Is a simple least squares ﬁt appropriate for these data?
* 5.33 Babies. Is the gestational age (time between conception and birth) of a low birth-weight baby useful in predicting head circumference at birth? Twenty-ﬁve low birth-weight babies were studied at a Harvard teaching hospital; the investigators calculated the regression of head circumference (measured in centimeters) against gestational age (measured in weeks). The estimated regression line is dhead circumference = 3.91 + 0.78×gestational age (a) What is the predicted head circumference for a baby whose gestational age is 28 weeks? (b) The standard error for the coeﬃcient of gestational age is 0.35, which is associated with df = 23. Does the model provide strong evidence that gestational age is signiﬁcantly associated with head circumference?
* 5.5. EXERCISES 259
* 5.34 Rate my professor. Some college students critique professors’ teaching at RateMyProfessors.com, a web page where students anonymously rate their professors on quality, easiness, and attractiveness. Using the self-selected data from this public forum, researchers examine the relations between quality, easiness, and attractiveness for professors at various universities. In this exercise we will work with a portion of these data that the researchers made publicly available.23
* The scatterplot on the right shows the relationship between teaching evaluation score (higher score means better) and standardized beauty score (a score of 0 means average, negative score means below average, and a positive score means above average) for a sample of 463 professors. Given below are associated diagnostic plots. Also given is a regression output for predicting teaching evaluation score from beauty score.
* (a) Given that the average standardized beauty score is -0.0883 and average teaching evaluation score is 3.9983, calculate the slope. Alternatively, the slope may be computed using just the information provided in the model summary table. (b) Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning. (c) List the conditions required for linear regression and check if each one is satisﬁed for this model.

# Lecture