# 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**

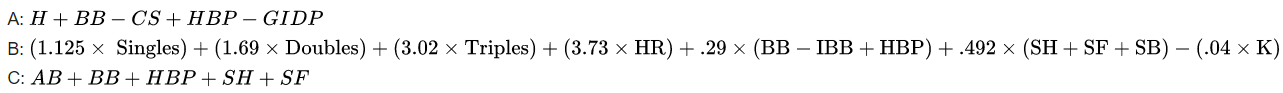
## VI. Conclusion

* 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:



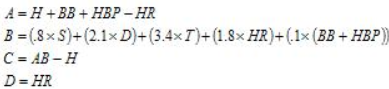
* + - RISP = runners in scoring position, ROB = runners on base.
    - subscripts indicate required condition for the formula.
      * Ex: HRISP = hits while runners are in scoring position
  + This is then figured for every member of a team + an estimate of total team RS is added up.
  + **Actual total 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

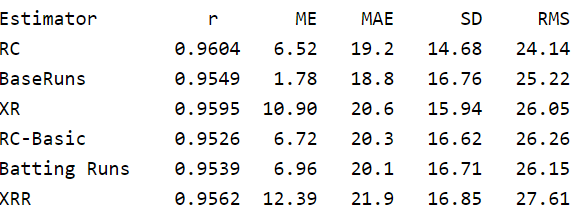
* For typical fan, sabermetrics doesn’t represent anything as theoretical as scientific inquiry.
* Rather, **sabermetrics = associated w/ new + unfamiliar statistics**.
* **OPS** = most famous of those new stats
  + gone from nearly unknown statistic in early 80s, to barely used a decade ago, to mainstream 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/)

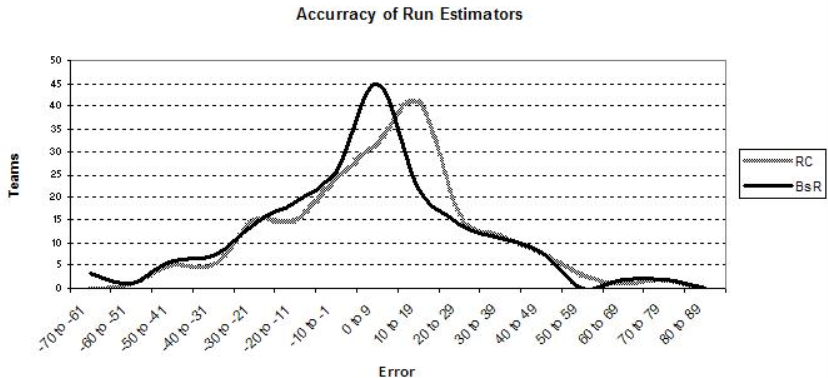
* Closer look at increasingly fashionable **On-Base Plus Slugging (OPS)** statistic in order to see how it compared as a proxy for offensive production vs. run estimators + other measures + just why it stacked up so well against them.
* **Correlation coefficients** showed that OPS tracks extremely well 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:



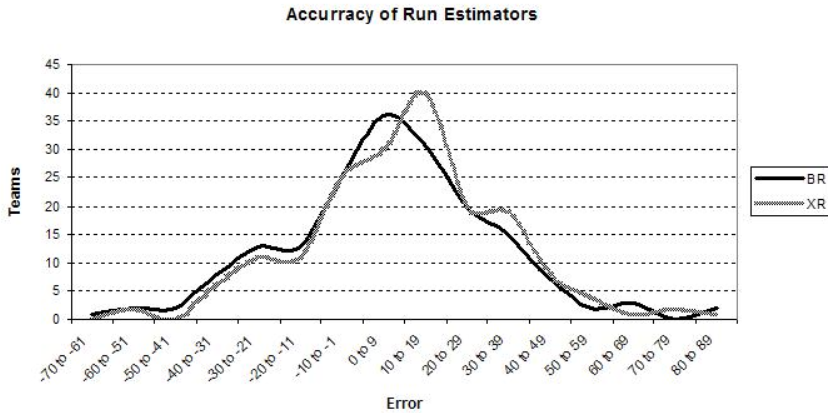
* + Interesting aspect of BaseRuns = B + C components are *combined* to calculate **score rate**
    - attempts to *estimate how often base runners score*.
  + When this ratio is multiplied by # of base runners (A) + added to # of HR’s, you get an estimate of runs would be scored as shown by **BsR**
  + 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.**



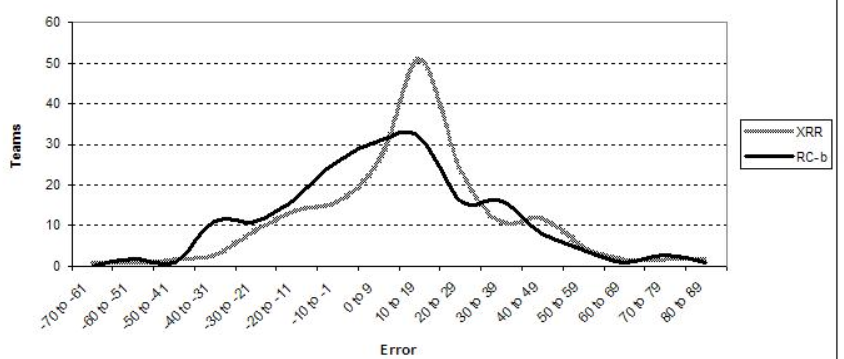
* See **BsR** has lowest MAE, while RC has lowest SD.
* Taking both into consideration, **RC outperforms BsR** by a fraction *when compared using RMS*.
* Another way to think about this = While BsR, *on average*, comes closer than RC, the *distribution of its errors is larger* (for the variant used here, 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



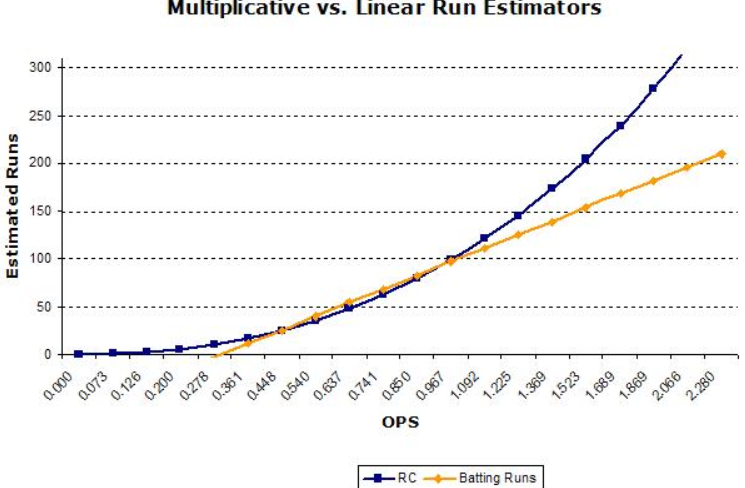
* Apex of the curve for BsR = higher + centered around 0-9 grouping, indicating it more often gets closer to the mark, hence its lower MAE.
* However, left-hand tail of RC remains under BsR’s, indicating that **when it underestimates actual runs scored, it does so w/ a smaller error**.
* RC also sneaks 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.



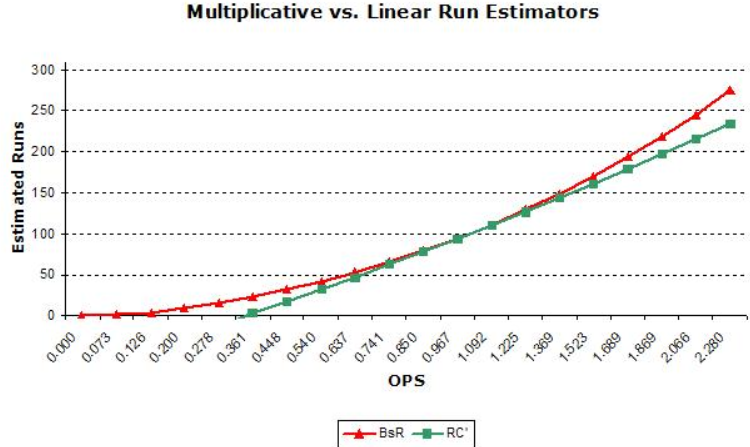
* Interestingly, although XRR has highest MAE = 21.9 + SD = 16.85, its distribution appears more compressed RC-Basic’s.
* Reason = **extreme values carry greater weight in the calculations** + XRR was off by 40+ runs on 27 of teams, while RC-Basic did so just 22 times.
* In addition, XRR curve = clearly 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

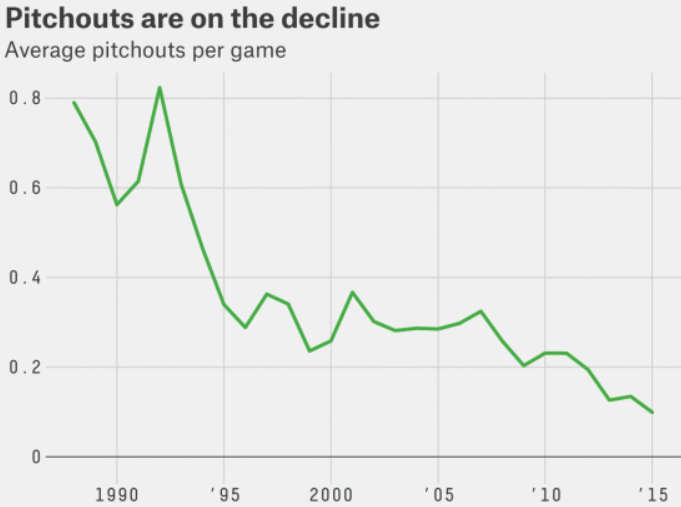
* You’re insane.” That’s generally the response I get when I
* present the information you’re about to read. I’ve been accused of being
* the “epitome of ‘pseudo-stat fan’ gibberish.” I’ve even been
* accused of being Aaron Sele writing under a pseudonym. I’m not
* entirely sure why my little way of doing things stirs the emotions of
* people to such a large extent, but apparently it does.
* My belief? Well, simply that hits allowed are not a particularly meaningful
* statistic in the evaluation of pitchers.
* Now before you accuse me of being Aaron Sele, please bear with me for a few
* paragraphs as I explain how I reached this point, and where it led from there.
* One of the basic issues in evaluating pitchers is to what extent the
* defense behind them is responsible for the results. In fact, in Baseball
* Prospectus 2000, one of Keith Woolner‘s “Hilbert Problems”
* for baseball was the issue of separating defense and pitching. As he put
* it, “Pitching and fielding are so intertwined that they seem
* impossible to separate.”
* Around the end of the 1999 season, I started to think about that problem. My
* plan was to go about dividing a pitcher’s stat line into what the defense
* can’t affect and what it’s possible that it can:
* Defense Independent:
* Walks, Strikeouts, Home Runs (essentially), Hit Batsmen, Intentional Walks
* Defense Dependent:
* Wins, Losses, Innings, Runs, Earned Runs, Hits Allowed, Sacrifice
* Hits, Sacrifice Flies.
* Any stats derived from the defense-dependent ones like OPS against or ERA
* or would also be defense dependent.
* The idea was to express the things the defense can’t affect in one area and
* check the results, then check those areas where it’s possible the defense
* can have an effect and analyze how much of the performance is pitching and
* how much is defense.
* The first thing I did was create something called “Defense Independent
* Pitching Stats.” DIPS are the representation of a pitcher’s stat line
* without any possible influence from the defense behind the pitcher. I
* calculated the various rates for walks, strikeouts, home runs, hit batsmen,
* etc. as a function of batters faced, and inserted them into the pitcher’s
* line. Then I calculated how many batters faced were remaining, and assigned
* league-average rates for all of the other component stats: innings, hits,
* doubles, triples, etc. So for all the stats that it was possible that the
* defense could affect, every pitcher was now on equal footing. The results,
* using Dave Burba‘s line in 2000, looked something like this:
* Actual
* BFP IP H HR ER BB SO ERA
* 848 191 199 19 95 91 180 4.48
* Defense Independent
* BFP IP H HR ER BB SO ERA
* 848 195 185 18 89 93 179 4.13
* As you can see, the home runs, walks, and strikeouts changed little (they
* changed at all only due to park effects and a few other minor factors). But
* hits and innings pitched changed by a decent amount, at least in this case.
* The next step was to look at the rest of a pitcher’s stat line and somehow
* divine how much of it was the result of the pitcher’s work. To do this, I
* looked at the range of values for Defense Independent ERA and compared how
* close they were to the range of values of actual ERA. For example, if the
* range of Defense Independent ERA was between 4.00 and 5.00, it would be a
* good indication that there’s a lot about pitching not covered in the stat,
* because ERAs have a much larger range than that.
* That didn’t happen. The range was virtually the same as actual ERA, with
* the best pitchers having DIPS ERAs near 2.40 and the worst having DIPS ERAs
* up near 7.00. I found this surprising, as I expected the range to narrow
* quite a bit more than that.
* Then, I looked at the behavior of Hits Per Balls in Play
* [(H-HR)/(BFP-HR-BB-SO-HB)]. That’s where the trouble really started. I
* swear to you that I did everything within my power to come to a different
* conclusion than the one I did. I ran every test, checked every stat,
* divided this by that and multiplied one thing by another. Whatever I did,
* it kept leading back to the same conclusion:
* There is little if any difference among major-league pitchers in their
* ability to prevent hits on balls hit in the field of play.
* It is a controversial statement, one that counters a significant portion of
* 110 years of pitcher evaluation. Let’s go over the facts that led me to
* this conclusion:
* As we discussed, the range of ERAs for pitchers is almost as large
* without defense-dependent statistics as it is with them. This speaks to the
* fact that there can be massive differences in the ability of pitchers
* before even considering the impact of defense.
* The pitchers who are the best at preventing hits on balls in play one
* year are often the worst at it the next. In 1998, Greg Maddux had
* one of the best rates in baseball, then in 1999 he had one of the worst. In
* 2000, he had one of the better ones again. In 1999, Pedro Martinez
* had one of the worst; in 2000, he had the best. This happens a lot.
* There is little correlation between what a pitcher does one year in the
* stat and what he will do the next. In other words, what Eric
* Milton‘s hits per balls in play was in 2000 tells us next to nothing
* about what it will be in 2001. This is not true in the other significant
* stats (walks, strikeouts, home runs). Walks and strikeouts correlate very
* well and homers correlate somewhat well.
* This is a crucial fact. One of the more critical aspects of statistical
* analysis is determining how well a statistic reflects an ability. It’s the
* test given to clutch hitting, catcher game-calling, pitcher won/loss
* records, and so on. One of the first things asked when addressing this is
* “Does the stat correlate well with itself from year to year?” One
* reason clutch hitting is questioned is that the “clutch hitters”
* change from year to year, which indicates that 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.”
* You can better predict a pitcher’s hits per balls in play from the rate
* of the rest of the pitcher’s team than from the pitcher’s own rate. This is
* pretty self-explanatory. The effects of having the same team defense and
* home park appear to be significant determinants in creating what little
* correlation there is in the stat.
* Take pitchers with similar stats in every other component category (and
* other peripheral factors like age, throwing hand, team hits per balls in
* play rates, etc.) but large differences in hits allowed (and therefore in
* innings pitched). When you group the pitchers into two
* categories–high-hits and low-hits–the following year the high-hits
* pitchers do not give up significantly more hits per balls in play (.292 to
* .291) than the low-hits pitchers, and 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, when all other things are equal, there is no difference, the
* conclusion becomes clearer.
* Similarly, if you take pitchers with comparable stats in every other
* component category, but have as large as possible a difference in
* strikeouts, then separate the pitchers into high-strikeout and
* low-strikeout categories, the high-strikeout pitchers continue to strike
* out more hitters, while also giving up far fewer hits and having
* significantly lower ERAs.
* This is the natural opposite of the fifth point. If number five is true,
* then logically number six ought to be true as well. It is.
* The range of career rates of hits per balls in play for pitchers with a
* significant number of innings is about the same as the range you would
* expect from random chance. This is true even though we know that some
* pitchers may have had consistent advantages over others, as these rates are
* unadjusted for park or league. The vast majority of pitchers who have
* pitched significant innings have career rates between .280 and .290.
* When you adjust for environmental advantages (the DH, park effects, and
* so on) the range becomes even smaller. The leaders in this stat (Pete
* Harnisch) have had significant environmental advantages while most of
* the trailers (yup, Aaron Sele) have had disadvantages. After these
* adjustments, the range is well within the realm you could expect from
* chance alone.
* A stat like Component ERA (or any similarly stat that calculates ERA
* from the rest of a pitcher’s performance), while correlating better with
* next-year ERA than ERA itself, does not correlate nearly as well with
* next-year ERA as it does if you perform the same calculation while using
* the average hits-allowed rate of the team for which he pitched. This
* advantage of “team average” rate grows to rather large
* proportions as the number of innings pitched in the season shrinks more and
* more.
* Two key points here: one, there doesn’t appear to be any “hidden
* quality” aspect to the stat. The numbers come out as they should if
* the above are all true: you can better predict ERA without hits allowed
* than you can with them. The other key point is that using a reliever’s hit
* rate seems to be an extremely suspect way of evaluating relievers. One of
* my favorite examples of this is Bobby Ayala in 1998 and 1999.
* There are a few lesser and somewhat anecdotal points to be made that, while
* not critical, are nonetheless good concepts to understand:
* People have a hard time diagnosing who the pitchers are that are very
* good at preventing hits on balls in play. You’ll often hear people use
* names like Randy Johnson, Jamie Moyer and Andy
* Pettitte in protest of the concept, but by any definition you want to
* use, these guys are not particularly good in the stat.
* Pitchers like Pedro Martinez and Greg Maddux have, at times, expressed
* thoughts on the matter. Martinez has been quoted as saying that the batter
* determines what happens once he hits the ball. Maddux described his
* scoreless-inning streak last year 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 (like Brent
* Mayne). That’s too small a sample on which to draw conclusions, but it
* is something to think about that these non-pitchers were not any worse than
* regular pitchers in the stat. In fact, they were a good bit better.
* Pitchers are often dubbed as “unpredictable”, and hits allowed
* is by far the most unpredictable of the component stats. In other words, it
* is one of the main culprits of pitcher unpredictability.
* There is no significant cross-correlation. That is, a high number of
* home runs allowed doesn’t really mean anything in determining how many hits
* per balls in play the pitcher will allow. The closest is an inverse
* relationship with strikeouts (lots of strikeouts means fewer hits per balls
* in play) but that relationship is very weak and could be the result of
* unrelated factors. There was no significant hits-per-balls-in-play
* advantage found in the strikeout study above.
* Many people, after reading these points, think I’m saying that all pitchers
* give up the same amount of hits. That’s not true, and of course it’s not
* what I’m saying. Randy Johnson gives up fewer hits than Scott Karl.
* That’s not because batters hit the ball harder off Karl than Johnson, but
* because they hit the ball more often off Karl than Johnson.
* Aside from walks, there are two basic outcomes for a pitcher: batter hits
* the ball or batter strikes out. With the latter, the result is almost
* always an out. With the former, all sorts of things can happen, including a
* base hit.
* So why is this all true? All I can advance are theories, some that can be
* checked out and some that are more difficult to verify. I’ll end this
* article with a list of some of the more popular ones:
* Scouting. The MLB scouting network is set up to sift through an
* enormous pool of potential players to get to the group that might be MLB
* pitchers. To do this, they often employ tactics that many might call unfair
* in an effort to reduce the pool to a manageable number. So they don’t take
* guys under 5’10” and every pitcher has to throw a certain speed
* fastball and so on. One of these factors may be weeding out a subset of
* pitchers for which the theory is not true.
* High talent level. This theory is that there’s a certain limit as
* to how good you can get at preventing hits on balls in play, and that in
* order to even come close to the major leagues you have to have reached
* this. This theory often comes up in clutch-hitting discussions.
* Too many variables. This suggests that the ability may or may not
* exist, but that the number of variables involved in the outcome of balls in
* play are so numerous and so difficult to control for that any ability gets
* lost. In other words, the noise completely masks any signal.
* A misunderstanding of how the batter/pitcher dynamic works. Some
* people will argue that despite all the numbers, the above can’t be true
* because it means that a screaming line drive hit into the
* right-center-field gap is as likely to be an out as a pop-up to the shortstop.
* This point deserves further discussion. One of the critical points of
* misunderstanding is the issue of “blame.” When a ball gets
* crushed into the gap in right-center, some think I’m saying that the
* defense deserves the blame, not the pitcher. When I counter with
* “Neither the pitcher nor defense is to blame, it’s the batter who is
* to blame,” I lose some people. Consider this example:
* When I was a kid, we used to go to the cemetery (this was our playground)
* and play a game called Lob-League. The makeup of this game was mostly
* offense and some fielding, with little to no pitching effects. The
* pitcher’s job was to lob the ball over the heart of the plate and let the
* batter hit it as hard as he wants.
* Now, let’s suppose we’re playing Lob-League and the pitcher lobs one right
* in the batter’s wheelhouse, but the batter pops it up to the shortstop. Who
* deserves credit for the pop-up? The blame argument would indicate that the
* pitcher deserves credit for inducing a pop-up despite the fact that all he
* did was lob the ball over the plate. No credit or blame would belong to the
* batter who popped up the pitch.
* A more relevant MLB example might be the Home Run Derby at the All-Star
* festivities. I encourage you to watch next year’s contest, or, if you have
* it, a videotape of past contests. Watch for batted balls that would clearly
* be outs. The pitcher is trying to give up home runs, so does he deserve
* credit for a pop-up?
* In MLB, a pitch could result in a pop-up or a line drive. It all depends on
* what the batter does with it. I think the conventional wisdom on the
* dynamic between pitcher and batter may be slightly inaccurate.
* The critical thing to understand is that major-league pitchers don’t appear
* to have the ability to prevent hits on balls in play. There are many
* possible reasons why this is the case, and I don’t really have a concrete
* idea as to why it is

# The Art of Pitch Framing

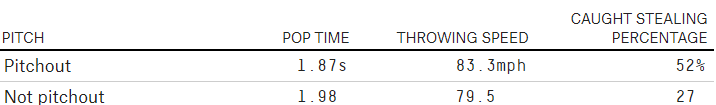
* ’s the day before a foul tip off the bat of Rajai Davis will fracture Francisco Cervelli’s hand, and the Yankees’ still-intact starting catcher is in excellent spirits. After spending almost all of 2012 in the minors, he’s happy to be back in the Yankees clubhouse. He’s also happy to be off to a good start with the bat, a start that’s about to get better; in a few hours, he’ll take Mark Buehrle deep for his third home run of the season. But how Cervelli hits is secondary, even to Cervelli.
* “I’ve been focused on my defense, and that’s it,” Cervelli says. “And I’m going to keep doing that no matter what happens with my bat.”
* A lot of eyebrows, and maybe a few middle fingers, were raised over the winter, when the Yankees — the team with the catching legacy of Dickey, Berra, Howard, Munson, and Posada, not to mention the $200 million–plus payroll — entrusted their catching duties to Cervelli and backup catcher Chris Stewart, a duo that entered the season with a combined .249/.315/.332 line in the big leagues. In the past, the Yankees would have dipped into the free-agent market and signed someone with a bigger bat and a bigger name — A.J. Pierzynski, perhaps, who was coming off a 27-homer season, or another offense-first option like Mike Napoli, who signed with the rival Red Sox. Both players agreed to one-year contracts, so they wouldn’t have hampered the Yankees’ goal of getting under the $189 million luxury tax threshold for 2014.1
* Instead, they stuck with two players who are earning barely more than the major league minimum. And they’ll probably be better off. Cervelli and Stewart can do more to help the Yankees win with a subtle shift of the glove than Mariano Rivera can with his cutter, than Brett Gardner can in the outfield, than Ichiro can with his arm and his base-running ability combined. They have an ability that not only doesn’t show up in the box score but doesn’t show up in advanced stats like UZR and WAR. Baseball teams have always known it existed, but they haven’t known what it was worth until now. And one need only look at the lineup card to see how valuable the Yankees believe it is.
* “They’re both exceptional defenders,” Yankees GM Brian Cashman said of Cervelli and Stewart in a recent interview with Mike Ferrin on SiriusXM’s MLB Network Radio. “Tremendous pitch framers. We’re big into that.”
* Take a look at these two pitches from 2012:
* They’re both four-seam fastballs thrown by right-handed pitchers to left-handed hitters. They both pass through the strike zone 21 inches off the ground, between 11.7 and 12.9 inches from the center of home plate. They both hit their targets, so the catchers know where they’re headed and have time to prepare. And they’re both called by the same umpire, Sam Holbrook.
* In fact, the two pitches are similar in just about every respect but their outcomes. The top one, thrown by James Shields last July, is a strike, but the bottom one, thrown by Liam Hendriks last June, is a ball.
* We can’t say for sure why only one was a strike; maybe Holbrook was just feeling generous when the first pitch crossed the plate. But we do know one important variable that differs between the two pitches — the catcher. The pitch on the top was caught by the Rays’ Jose Molina, one of baseball’s best receivers. The pitch on the bottom was caught by the Twins’ Ryan Doumit, one of the worst. And that may have made all the difference.
* Focus on how catchers “frame” pitches to make them look more like strikes, or talk to guys who are good at it, and the distinction between players like Molina and Doumit starts to stand out. Depending on how they’re caught, two pitches that are almost identical on their way to the plate can look a lot different once they get to the glove.
* Concentrate on the catchers in those clips. Molina sets up farther outside, so even though the pitch to him is farther from the plate, he catches it in the center of his body. Doumit has to reach for the ball, drawing attention to its distance from the strike zone. The bases are empty in both clips, giving the catchers the freedom to set up any way they want without worrying about base runners. But only Molina goes down to one knee to present a lower, more stable target. Doumit’s head jerks sharply downward the instant after he catches the pitch. Molina’s remains still. And Doumit’s glove, descending to meet the pitch, dips even more after he catches it. This sends the ball farther outside the zone and forces him to jerk the glove back up in an exaggerated fashion. Molina’s glove never gets any lower than it is when he receives the pitch. He makes a much more subtle upward movement, and it takes about half as much time for his glove to come to rest.
* These differences in technique don’t only pop up on some pitches. They’re present for almost every pitch, as constant as a pitcher’s motion or a batter’s stance. Go back to 2011, when Molina (top) was with Toronto and Doumit (bottom) played for Pittsburgh. You’ll see the same differences:
* Although the changes in camera angle and batter position make it difficult to tell, these pitches are even closer together, according to PITCHf/x: 1.26 feet and 1.29 feet from the center of the strike zone, respectively. Both are four-seamers on 0-1 counts called by umpire Mike Winters. Again, we see the call go one way for Molina and the other for Doumit. And again, it’s not tough to tell why. Doumit’s whole body leans toward the outside of the plate, while Molina’s stays almost perfectly still. Doumit raises and clenches his right hand, creating a potential distraction for the umpire; Molina’s right hand remains tucked behind him. Doumit’s head and glove again dip before his pronounced pull. Molina times his reception to catch the ball with his glove already on the rise.
* Once you train yourself to see it, it’s almost impossible to stop seeing it. Baseball is often described as a chess match between batter and pitcher. But it’s more like a chess match between batter and pitcher in which, once in a while, the catcher grabs the board and moves someone’s piece.
* In a September 2011 article titled “Spinning Yarn: Removing the Mask,” Mike Fast, then an analyst for Baseball Prospectus (and now an analyst for the Houston Astros), attempted to determine what catcher receiving was worth. By studying where strikes are typically called and establishing which pitchers were getting more or fewer strikes than they “should” have, given where their pitches crossed the plate, Fast was able to isolate the effect of the catcher. He concluded that pitch framing can make a major impact, and it also is more consistent from year to year than even reliable offensive metrics like on-base percentage or slugging percentage. In other words, it’s not insignificant, and it’s not just noise. It’s a valuable skill that persists from season to season.
* Fast found that Molina, the best receiver, was worth 35 runs above average per 120 games, and Doumit, the worst, was worth 26 runs below average. After Houston hired Fast, another analyst named Max Marchi succeeded him at Baseball Prospectus2 and brought with him a sophisticated model for framing that accounted for most of the potentially confounding factors: the umpire, the ballpark, the batter, the ball-strike count, and the pitch location and type. According to Marchi, who has consulted for a major league club and whose work has been mentioned by Rays manager Joe Maddon, Molina has saved his teams 111 runs — or, using the standard 10-runs-to-a-win conversion, about 11 wins — because of framing from 2008 to 2013. (The only other catcher with a higher run total over that same time period, Brian McCann at 122, has caught more than twice as many pitches.) Doumit, on the other end of the receiving spectrum, cost his teams 155 runs. That comes out to 0.50 runs added by Molina and 0.55 runs subtracted by Doumit per 100 pitches, an enormous difference.3 For comparative purposes, Barry Bonds’s bat during the 2001-2004 seasons, when he basically broke baseball, was worth about 0.78 runs above average per game.
* Granted, those run totals aren’t typical. Molina and Doumit are outliers; no one else comes close to costing his team as much per pitch as Doumit, and only Brewers catcher Jonathan Lucroy has approached Molina levels of framing effectiveness. (Lucroy has some advantages over Molina. He’s 11 years younger and has a better bat, which allows him to stay in the lineup. Since his debut in 2010, he’s saved almost twice as many runs due to framing as any other catcher, aside from Molina.)
* Still, consider the implications. Giancarlo Stanton, one of the most coveted young players in the game, was worth as much4 (in terms of Wins Above Replacement) over the past two seasons as Molina’s framing alone was worth in a part-time role over the past five-plus, yet Molina’s value, unlike Stanton’s, is largely overlooked. And that’s without factoring in any value Molina added by calling games, handling his pitching staff, and controlling the running game. Best of all, Molina’s done it all for an average of $1.5 million per season, in an era when a single win on the free-agent market usually runs teams around $5 million. It’s no surprise that Molina plays for the Rays, low-payroll competitors who’ve found ways to make a dollar go further than any other organization. Nor is it a coincidence that Molina, a career backup catcher whom standard sabermetric stats peg as a replacement-level player, became a first-time starter and appeared in a career-high 102 games in his age-37 season, barely six months after Fast’s study appeared on Baseball Prospectus.
* Eleven wins sounds like too much for a single part-time pitch receiver to add in just over five seasons, but think about how many opportunities to gain and lose strikes there are over a 162-game schedule. A durable catcher can catch around 10,000 called pitches in a single season. Many of those pitches are clear-cut calls. But that still leaves hundreds, maybe thousands, of pitches in the shadowy border region between ball and strike, where a good receiver can perform the catching equivalent of turning water into wine. Even if an extra strike doesn’t send the batter back to the dugout, it puts him in a less-favorable count and makes him less likely to do damage later in the at-bat. Dan Turkenkopf, another former Baseball Prospectus staffer who was recently hired by the Rays, put the average value of turning a single ball into a strike at 0.13 runs. If you do that a few times per game, as Molina does, the run total climbs quickly.5
* “If a catcher can perfect a great way of receiving the ball, and he gets the ball maybe a half a ball outside — or even a ball outside — off the corners consistently, I think he’s worth his weight in gold,” says Steve Yeager, a 15-year big league veteran and the Dodgers’ catching coach. Molina is listed at 250 pounds, so depending on the karat, his weight in gold would be worth much more than he’s making. But even a 24-karat Molina wouldn’t go for as much as Marchi’s model suggests he’s worth to a team.
* It’s not as if receiving skills weren’t valued when Yeager was playing; as more than one person (OK, almost every person) I spoke to pointed out, the word “catch” is right there in the name of the position. But it’s easier to discount a certain skill when there are no numbers attached to it. (Just try to talk to the more dogmatic species of stathead about clubhouse chemistry.) Before framing became a sabermetric buzzword, Molina’s receiving skills were known but nebulous. His offensive struggles, on the other hand, were easier to see, and almost as easy to quantify. But now that we can count his contributions on both sides of the ball, what he does on defense is impossible to dismiss. The Rays, a team that’s smart enough to see that, are reaping the rewards (and so is Molina, though it’s a little late for him to cash in).
* Jose’s younger brother Yadier and retired older brother Bengie share(d) his receiving skill to a lesser degree, and all three inspire awe from their peers at the position. “I think the three Molinas, they come from another planet,” says Yankees catcher Francisco Cervelli. “I’ve never seen anything like that in my life.” But as many Molinas as there are, they’re not a renewable resource. At some point, the last Molina will retire, and if receiving skills are still undervalued among major league catchers, a team that could learn to teach framing efficiently at the lower levels, minting Molina-like receivers while other teams are still running out Ryan Doumits, might have its hands on a major market inefficiency.
* Thanks to advances in technology and analysts like Fast and Marchi, we have a much better sense than before of what a catcher with good receiving skills is worth. We also know which big leaguers qualify as the best in the business. What we don’t know, necessarily, is where good framing catchers come from. Are good receivers born or made? And if they’re made, how do you make them?
* Chris Stewart Video is widely available through the minors, but the pitch-tracking technology that has made the Molina name more famous has penetrated only so far. According to one front-office source, about 40 minor league ballparks have a PITCHf/x system installed, and about 25 are equipped with Trackman (a competing ball-tracking technology that relies on Doppler radar). Around 20 of the 30 parent clubs have at least one affiliate with one of the two systems in place, with the highest concentration at Triple-A. The teams pool much of the information that’s collected; the more of your own PITCHf/x or Trackman data you elect to share with others, the more you receive in return (although clubs generally keep info from non-league events held in their stadiums — workouts, high school and college games — to themselves).
* Still, 65 ballparks with ball-tracking technology just scratches the surface of professional baseball below the big leagues. Even in 2013, seven years after the first PITCHf/x systems were installed in big league ballparks, most of the minors — to say nothing of amateur and international baseball — are a ball-tracking blind spot. That means that receiving skills still have to be evaluated the old-fashioned way (with actual eyes) and the slightly less old-fashioned way (via video).6
* That’s not really a problem, because the search for good framers, like almost every other pursuit in the modern front office, works best with input from both stats and scouts. Over small samples, an experienced scout or instructor can tell you more about someone’s receiving skills than a computer connected to a camera can, especially at levels where it’s more difficult for catchers to frame pitches because the pitchers can’t hit the target consistently. Opinions vary on exactly how long it takes to evaluate a catcher’s receiving skills, but the consensus is that the assessment comes quickly.
* Yankees catcher Chris Stewart confidently says he can spot a strong receiver in five to 10 pitches. Others offer more conservative estimates. “I think you need a couple of games to give a truthful evaluation,” Yeager says. “You have to see a guy catch a couple of days, and you have to see him catch three or four different pitchers, because every pitcher’s different.”
* That snap judgment is just the start. Once you’ve acquired a catcher with some raw talent for framing, the real work begins. There’s no one-size-fits-all plan for grooming great receivers — and not every catching instructor agrees on the proper way to receive pitches — but there are, generally, two types of improvements a player can make. There’s the mechanical change, which can be adopted immediately, and the incremental increase in comfort and confidence that comes with experience, repetition, and familiarity with a pitching staff.
* Some catchers are framing fixer-uppers, just a few exterior alterations away from becoming much more valuable properties. For Diamondbacks bullpen coach Glenn Sherlock, a former minor league catcher and an experienced catching instructor whom Arizona GM Kevin Towers credits with turning Miguel Montero into a good framer, the first fix is glove height. “I think catchers’ targets at the bottom of the zone is so important,” he says. “You see a lot of catchers who have high targets and take the balls out of the strike zone.”7
* Stewart recalls a mechanical tweak that helped him in 2008: “I had my thumb pointing down. That’s kind of the natural way you learn to catch. But [Yankees bench coach Tony Pena] had me turn my glove so my thumb was more pointing toward the second baseman area instead of pointing down at the plate. It helped me work on balls coming on my left side easier. I didn’t have to turn my whole arm.”
* Yeager focuses first on a catcher’s stance. “I think you can correct some guys that are doing something just by the positioning of his arm,” he says. “Sometimes they have a tendency to get the arm inside of the knee, or they get it on the outside of the knee, and … their knee sometimes gets in the way … Everything in the positioning is a key.”
* Lower the target, turn your thumb, adjust your crouch. That’s the easy stuff. It may feel unnatural at first, and it takes some time to get used to, but it can produce improvement overnight. The more grueling method of improvement, the one that over the long run produces the greater gains, is as simple as it is painstakingly slow: catch. Catch in games, catch in the bullpen, catch in side sessions, and then do it again in your dreams after you finally fall asleep. “Baseball is a habit,” Cervelli says. “It’s a repetition. You’ve got to repeat things every day, and they come.”
* Bullpen drills are a big part of that repetition. Pitching machines can simulate any type of pitch, at any speed, in any location, and apprentice pitch framers exhaust all the possibilities. The drills aren’t a time to fill your quota of practice pitches and move on to something more interesting. They’re a time to concentrate and consolidate lessons.
* “We do some things with the pitching machines, breaking balls, left-handed, right-handed,” Sherlock says. “We move the catcher around behind the plate to work on backhand pitches and forehand pitches. Work on some soft wiffle-balls to simulate movement, and just working on catching it without a glove … Also, something that we do is check to make sure that they’re breathing back there. You start watching the catchers go through these drills and see that they’re holding their breath.”
* Motivated catchers don’t just do drills. They watch video, both of themselves and their opponents at the position. “Watch video of guys that are good,” Stewart says. “Watch them and see how they do it and try to intertwine it into your game.”
* Eventually, all of this pays off in game action. If it sounds a little monotonous, well, it is. That’s another reason why, even if decades down the line every place where baseball is played comes pre-installed with a PITCHf/x (or FIELDf/x) system, there will still be a role for scouts. No matter how all-seeing the eyes in the sky above ballparks become, teams will still need to know what’s going on inside their players. Character and 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.
* “We can make you better if you have an open mind and you’re willing to work and willing to try certain things,” says Yeager, who mentions Russell Martin, A.J. Ellis, and Tim Federowicz as some of his best students.8 “[You’ve] got to want to get back there and take the time of squatting and blocking balls, transferring balls, throwing balls, receiving balls, getting your hands beat up, getting foul tips into you. You’ve got to realize that you’re going to get beat up physically back there.”
* The good news is that if you do have the masochistic impulse to get drilled by baseballs, you can have a future in framing. Naturally, the younger you start to work, and the more naturally gifted you are, the quicker it will come, but those gifts aren’t a prerequisite. “If you see a guy that has some very soft hands, you know that at one point he’s going to get it,” says Tony Pena, a four-time Gold Glover who now serves as the Yankees’ bench coach and catching instructor.9 “There’s no limit. There’s going to be a guy that plays for a long time and then becomes a really good catcher over years.”
* “I don’t think everybody can be the same,” Cervelli says. “[But] if you work, if you let somebody coach you, it can happen. You can be close.” Stewart concurs. “Obviously there’s something inside that person that allows him to do what he’s able to do, but it’s more of a repetition skill, I think, than a God-given talent.”
* Being born without the hands of a Molina brother is a handicap, but it’s a handicap that can be overcome. Now that teams know what a great receiver is worth, and players know that they know, both sides have more incentive than ever to focus on framing, both before and after each backstop’s big league debut.
* Jeff Luhnow Brian Cashman has never known the feeling, but there’s a certain freedom that comes with being in charge of a bad baseball team, especially one that isn’t expected to be good. Unburdened by the constant pressure to minimize risk and construct a competitive roster that accompanies a club like the Yankees, losing teams led by creative executives are able to experiment and innovate. If you’re going to be bad no matter what you do, you might as well try to be bad in a way that will make you better eventually.
* That’s what we’re seeing with the Astros, a rebuilding club that’s currently making the Marlins look like big spenders. Rather than risk jeopardizing their long-term potential by trying to strike a balance between rebuilding and respectability, the Astros decided to raze their roster, trading every veteran who wasn’t tied down and spending hardly anything on free agents. It’s a path that has led to a lot of losing, with much more to come. But it’s also allowed the Astros to completely start over, in just about every capacity. They’ve restocked a formerly barren minor league system, climbing from 26th to ninth in the Baseball Prospectus organizational rankings in a single season. They’ve blazed new trails in pitcher usage patterns and defensive shifts. And led by GM Jeff Luhnow, they’ve assembled a new-school, cross-disciplinary front office with intellectual talent drawn from atypical baseball backgrounds.
* One member of that front office is Mike Fast. And as you’d expect, given Fast’s focus on framing before he was hired, the Astros are exploring ways in which they can make receiving skills a strength. You can’t see that effort reflected yet at the major league level; backup catcher Carlos Corporan is about average, according to Marchi’s model, and starter Jason Castro is somewhat worse. It wouldn’t matter much, anyway. A few extra strikes in the majors right now wouldn’t make the Astros’ immediate outlook any less hopeless. It’s below the surface where Houston’s framing future is taking shape. That’s where the Astros are seeking out the potential benefits of developing strong receivers before they reach the big league level.
* By stocking the minors with good framers, the Astros could accelerate their pitching prospects’ development or make them more attractive to other teams by bolstering their stats with catchers who earn them extra strikes. With the catchers themselves, Houston can simply wait until Castro and Corporan enter the arbitration process and start to make more than they’re worth, at which point they can promote those prospects. By the time the Astros decide to make winning at the big league level a priority, they could have a few catchers capable of expanding the strike zone. They may have put the first phase of that plan into action when they acquired Double-A catching prospect Max Stassi, a 22-year-old with a reputation for strong receiving skills, in the February trade that sent Jed Lowrie to Oakland.
* The man tasked with overseeing the receiver assembly line is Mark Bailey, the former Astros catcher and bullpen coach who accepted a new position this season as the club’s roving catching instructor, and he couldn’t sound more excited. The 51-year-old is a recent convert to the church of pitch framing, and he’s eager to spread the word. Bailey has seen the research and the stats, and he’s also studied what the catchers who produce the best stats look like. Fast and Astros director of decision sciences Sig Mejdal have sent him video comparing good receivers like Jose Molina and Jonathan Lucroy with subpar receivers, and the visuals made a major impression.
* “A lot of it’s stuff that we already kind of knew and learned,” Bailey says, “but to really see it and look at the numbers … some of the video that I saw was really amazing. The good and the bad. You compare, and it’s like, ‘Whoa.'”
* That eureka moment made Bailey reevaluate what he looks for first in a catcher.
* “In the past, it’s always been the guy with the strong arm,” Bailey says. “But my opinion is moving in [the direction of framing], just by watching this and looking at some of that research … It’s always been the premium asset back there, but this is a little bit different. This takes it a step further.”
* The next stage is implementing a development strategy designed to make the most of each catcher’s receiving skills, and there appears to be buy-in at all levels of the organization. Jeff Murphy, a former minor league backstop who spent 12 seasons as the Cardinals’ bullpen catcher and catching instructor, followed Luhnow to Houston, where he’s serving in the same role this season. Although Murphy stresses still-intangible aspects of catching — like handling a pitching staff and calling games — to a greater degree than Bailey, it’s clear that he considers receiving skills important. Maybe a little too important, if you ask his exhausted pupils.
* “The first day of spring training, I told them, every morning, we’re going to be in the cage at 7 a.m.,” Murphy says. “And we’re doing receiving drills, and we’re just working on receiving … and these guys had never been through that type of work. And they said, ‘You know what, we’re not used to this.’ And I said, ‘This is our time to practice.'”
* If Bailey, Murphy, and the rest of the new Houston regime have their way, that practice will translate into wins a few years from now. And if the Astros’ approach succeeds, it’s bound to inspire copycats, even if no one writes a book or makes a movie about it. All it will take is for other teams to start asking the same simple question Bailey says he’s asked himself: “Why not try to get better?”
* “That’s the way the game is,” says Kirt Manwaring, the Giants’ former Gold Glove catcher and current catching coordinator. “Once somebody does something, once somebody’s successful at something, then they want to try to find the method behind the madness. ‘Well, what are they doing?'”
* Not every baseball lifer is as open-minded as Bailey, but the Astros, the Yankees, the Rays, and any other teams that have already started targeting and trying to develop good framers are only the vanguard. The market for catchers with superior receiving skills will grow more crowded as long as it looks like an area where clubs can get an edge. As Russell Martin says, it’s “a lot easier to teach somebody how to frame a pitch than it would be to teach them how to hit homers and drive in runs.”
* If you take the Astros’ plan, and Manwaring’s comment about copycats, to their logical conclusion, then at some point in the not-too-distant future, almost every team — save, perhaps, for a few with gifted offensive catchers for whom framing aptitude is less paramount — could have someone squatting behind the plate and stealing extra strikes. But sweeping changes to the sport rarely come without unforeseen consequences. That kind of mass movement toward catchers with strong receiving skills would upset the delicate balance between batter and pitcher; if you think baseball’s strikeout rates are high now, wait until the first wave of Stepford framers arrives. If umpires start to see nothing but good receivers, they might adjust their 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.
* It’s also possible that a greater awareness of framing could hasten the end of umpiring as we know it. The discovery of framing has opened up a new field of research for authors and aspiring sabermetricians, but what’s good for baseball writers isn’t always good for baseball. Even if the intent isn’t to criticize umpires, it’s impossible to write about framing without drawing some attention to the fact that the rulebook strike zone is more of an abstract concept than something that exists in the wild. The more attention that the catcher’s ability to influence the strike zone receives, the more likely it is that Major League Baseball will act to automate it. And if the human element goes, replaced by robo umps, then framing will go with it.
* The calls might be more accurate. But we’d lose the art that is a perfect frame.
* What Francisco Cervelli wants to talk about, more than how much better he’s gotten at framing, is how much better he’s still going to get. “I’m hungry,” he says. “I want it more than everybody. I don’t work to maintain, I work to be better and better and better.”
* Cervelli’s stats suggest that the work has been worth it. From 2009 to 2010, his receiving saved the Yankees 6.3 runs in 979 innings. But from 2011 to 2013, looking much more quiet behind the plate, he’s saved the Yankees 16.6 runs in 459⅓ innings — more than twice as many runs in less than half the playing time. Cervelli was an infielder and a pitcher before the Yankees signed him out of Venezuela and converted him to catcher, so he faced a steeper learning curve than most professionals. But he’s living proof that receiving skills can improve over time if accompanied by proper coaching and a desire to improve.
* The market-induced demise of framing might arrive one day, but it’s not here yet. Careerwise, this is the best time in baseball history to be a solid receiver. Framing is Chris Stewart’s meal ticket, too — he’s saved 16.5 runs in a little more than 8,000 pitches — and he’s well aware of it.
* “Within the last two or three years it’s taken over as one of the highly sought-after skill sets for a defensive catcher,” he says. “The sabermetrics stuff coming out, they put a value on it. We actually have a number for it. It’s not a ‘This guy’s good’ or ‘This guy’s bad,’ it’s like, ‘This guy’s this good’ and ‘This guy’s this bad.'”
* It’s early in the season, but it’s not too soon to assess the Yankees’ Cervelli/Stewart experiment. Through the game in which Cervelli was injured, the pair had combined to catch 1,792 called pitches and saved the Yankees a little more than five runs in the process. If we extend that to the number of called pitches Yankees catchers caught last season, their total contribution comes to 36 runs. A.J. Pierzynski’s extended total over the entire season has him costing the Rangers roughly seven runs. He might outhit Cervelli and Stewart, but not by nearly enough to make up for a framing disparity that size. Even if he came close, there would still be the not-so-small matter of his salary, which is almost six times higher than Cervelli’s and Stewart’s put together.
* Runs are runs, whether they’re scored or saved. Now that we know how many runs Cervelli, Stewart, and others like them are saving, those former fringe players have become commodities that every team wants.

# Sabermetrics Is Killing Bad Dugout Decisions

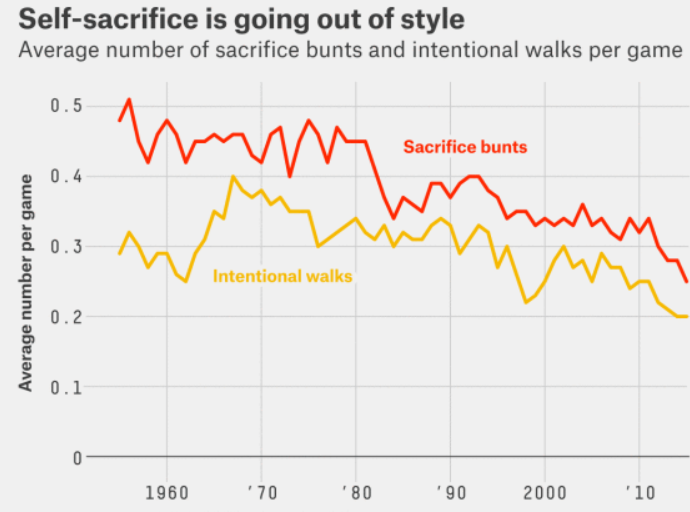
* Teams evaluate data differently in 2016 than 30 years earlier
* Front-office pages include job titles like “quantitative analyst” or “decision scientist”
* In-game, the evolution would be obvious if you knew where to look, broadcast by tactical details like how few hitters shorten swings to make contact 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
* The equation for the line is

y = 5 + 57.49x

Imagine what a perfect linear relationship would mean: you would know the exact value of y just by knowing the value of x. This is unrealistic in almost any natural process. For example, if we took family income x, this value would provide some useful information about how much ﬁnancial support y a college may oﬀer a prospective student. However, there would still be variability in ﬁnancial support, even when comparing students whose families have similar ﬁnancial backgrounds. Linear regression assumes that the relationship between two variables, x and y, can be modeled by a straight line:

y = β0 + β1x (5.1)

where β0 and β1 represent two model parameters (β is the Greek letter beta). These

β0,β1 Linear model parameters

parameters are estimated using data, and we write their point estimates as b0 and b1. When we use x to predict y, we usually call x the explanatory or predictor variable, and we call y the response. It is rare for all of the data to fall on a straight line, as seen in the three scatterplots in Figure 5.2. In each case, the data fall around a straight line, even if none of the observations fall exactly on the line. The ﬁrst plot shows a relatively strong downward linear trend, where the remaining variability in the data around the line is minor relative to the strength of the relationship between x and y. The second plot shows an upward trend that, while evident, is not as strong as the ﬁrst. The last plot shows a very weak downward trend in the data, so slight we can hardly notice it. In each of these examples, we will have some uncertainty regarding our estimates of the model parameters, β0 and β1. For instance, we might wonder, should we move the line up or down a little, or should we tilt it more or less?

219

220 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION

●

●

●

●

●

●

●

●

●

●

●

●

Number of Target Corporation stocks to purchase 0 10 20 30

Total cost of the shares (dollars)

0

500

1000

1500

Figure 5.1: Requests from twelve separate buyers were simultaneously placed with a trading company to purchase Target Corporation stock (ticker TGT, April 26th, 2012), and the total cost of the shares were reported. Because the cost is computed using a linear formula, the linear ﬁt is perfect.

−50 0 50

−50

0

50

500 1000 1500

0

10000

20000

0 20 40

−200

0

200

400

Figure 5.2: Three data sets where a linear model may be useful even though the data do not all fall exactly on the line.

As we move forward in this chapter, we will learn diﬀerent criteria for line-ﬁtting, and we will also learn about the uncertainty associated with estimates of model parameters. We will also see examples in this chapter where ﬁtting a straight line to the data, even if there is a clear relationship between the variables, is not helpful. One such case is shown in Figure 5.3 where there is a very strong relationship between the variables even though the trend is not linear. We will discuss nonlinear trends in this chapter and the next, but the details of ﬁtting nonlinear models are saved for a later course.

5.1. LINE FITTING, RESIDUALS, AND CORRELATION 221

●

●

●

●

●

●

●

●

●

●

● ● ● ● ●

●

●

●

●

●

●

●

●

●

●

0 20 40 60 80 Angle of incline (degrees)

Distance traveled (m)

0

5

10

15

Best fitting straight line is flat (!)

Figure 5.3: A linear model is not useful in this nonlinear case. These data are from an introductory physics experiment.

5.1 Line ﬁtting, residuals, and correlation

It is helpful to think deeply about the line ﬁtting process. In this section, we examine criteria for identifying a linear model and introduce a new statistic, correlation.

5.1.1 Beginning with straight lines

Scatterplots were introduced in Chapter 1 as a graphical technique to present two numerical variables simultaneously. Such plots permit the relationship between the variables to be examined with ease. Figure 5.4 shows a scatterplot for the head length and total length of 104 brushtail possums from Australia. Each point represents a single possum from the data.

The head and total length variables are associated. Possums with an above average total length also tend to have above average head lengths. While the relationship is not perfectly linear, it could be helpful to partially explain the connection between these variables with a straight line.

Straight lines should only be used when the data appear to have a linear relationship, such as the case shown in the left panel of Figure 5.6. The right panel of Figure 5.6 shows a case where a curved line would be more useful in understanding the relationship between the two variables.

Caution: Watch out for curved trends We only consider models based on straight lines in this chapter. If data show a nonlinear trend, like that in the right panel of Figure 5.6, more advanced techniques should be used.

222 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION

75 80 85 90 95

85

90

95

100

Total length (cm)

Head length (mm)

●

Figure 5.4: A scatterplot showing head length against total length for 104 brushtail possums. A point representing a possum with head length 94.1mm and total length 89cm is highlighted.

Figure 5.5: The common brushtail possum of Australia. —————————– Photo by wollombi on Flickr: www.ﬂickr.com/photos/wollombi/58499575

5.1. LINE FITTING, RESIDUALS, AND CORRELATION 223

75 80 85 90 95

85

90

95

100

Head length (mm)

Total length (cm)

2000 2500 3000 3500 4000

15

20

25

30

35

40

45

Miles per gallon (city driving)

Weight (pounds)

Figure 5.6: The ﬁgure on the left shows head length versus total length, and reveals that many of the points could be captured by a straight band. On the right, we see that a curved band is more appropriate in the scatterplot for weight and mpgCity from the cars data set.

5.1.2 Fitting a line by eye

We want to describe the relationship between the head length and total length variables in the possum data set using a line. In this example, we will use the total length as the predictor variable, x, to predict a possum’s head length, y. We could ﬁt the linear relationship by eye, as in Figure 5.7. The equation for this line is

ˆ y = 41 + 0.59x (5.2)

We can use this line to discuss properties of possums. For instance, the equation predicts a possum with a total length of 80 cm will have a head length of

ˆ y = 41 + 0.59×80 = 88.2

A “hat” on y is used to signify that this is an estimate. This estimate may be viewed as an average: the equation predicts that possums with a total length of 80 cm will have an average head length of 88.2 mm. Absent further information about an 80 cm possum, the prediction for head length that uses the average is a reasonable estimate.

5.1.3 Residuals

Residuals are the leftover variation in the data after accounting for the model ﬁt:

Data = Fit + Residual

Each observation will have a residual. If an observation is above the regression line, then its residual, the vertical distance from the observation to the line, is positive. Observations below the line have negative residuals. One goal in picking the right linear model is for these residuals to be as small as possible.

224 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION

75 80 85 90 95

85

90

95

100

Total length (cm)

Head length (mm)

Figure 5.7: A reasonable linear model was ﬁt to represent the relationship between head length and total length.

Three observations are noted specially in Figure 5.7. The observation marked by an “×” has a small, negative residual of about -1; the observation marked by “+” has a large residual of about +7; and the observation marked by “4” has a moderate residual of about -4. The size of a residual is usually discussed in terms of its absolute value. For example, the residual for “4” is larger than that of “×” because |−4| is larger than |−1|.

Residual: diﬀerence between observed and expected The residual of the ith observation (xi,yi) is the diﬀerence of the observed response (yi) and the response we would predict based on the model ﬁt (ˆ yi): ei = yi − ˆ yi We typically identify ˆ yi by plugging xi into the model. Example 5.3 The linear ﬁt shown in Figure 5.7 is given as ˆ y = 41 + 0.59x. Based on this line, formally compute the residual of the observation (77.0,85.3). This observation is denoted by “×” on the plot. Check it against the earlier visual estimate, -1. We ﬁrst compute the predicted value of point “×” based on the model: ˆ y× = 41 + 0.59x× = 41 + 0.59×77.0 = 86.4 Next we compute the diﬀerence of the actual head length and the predicted head length: e× = y×− ˆ y× = 85.3−86.4 = −1.1 This is very close to the visual estimate of -1.

5.1. LINE FITTING, RESIDUALS, AND CORRELATION 225

75 80 85 90 95

−5

0

5

Total length (cm)

Residuals

Figure 5.8: Residual plot for the model in Figure 5.7. J Guided Practice 5.4 If a model underestimates an observation, will the residual be positive or negative? What about if it overestimates the observation?1 J Guided Practice 5.5 Compute the residuals for the observations (85.0,98.6) (“+” in the ﬁgure) and (95.5,94.0) (“4”) using the linear relationship ˆ y = 41 + 0.59x. 2

Residuals are helpful in evaluating how well a linear model ﬁts a data set. We often display them in a residual plot such as the one shown in Figure 5.8 for the regression line in Figure 5.7. The residuals are plotted at their original horizontal locations but with the vertical coordinate as the residual. For instance, the point (85.0,98.6)+ had a residual of 7.45, so in the residual plot it is placed at (85.0,7.45). Creating a residual plot is sort of like tipping the scatterplot over so the regression line is horizontal. Example 5.6 One purpose of residual plots is to identify characteristics or patterns still apparent in data after ﬁtting a model. Figure 5.9 shows three scatterplots with linear models in the ﬁrst row and residual plots in the second row. Can you identify any patterns remaining in the residuals?

In the ﬁrst data set (ﬁrst column), the residuals show no obvious patterns. The residuals appear to be scattered randomly around the dashed line that represents 0. The second data set shows a pattern in the residuals. There is some curvature in the scatterplot, which is more obvious in the residual plot. We should not use a straight line to model these data. Instead, a more advanced technique should be used.

1If a model underestimates an observation, then the model estimate is below the actual. The residual, which is the actual observation value minus the model estimate, must then be positive. The opposite is true when the model overestimates the observation: the residual is negative. 2(+) First compute the predicted value based on the model:

ˆ y+ = 41 + 0.59x+ = 41 + 0.59×85.0 = 91.15

Then the residual is given by

e+ = y+ − ˆ y+ = 98.6−91.15 = 7.45 This was close to the earlier estimate of 7. (4) ˆ y4 = 41 + 0.59x4 = 97.3. e4 = y4− ˆ y4 = −3.3, close to the estimate of -4.

226 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION

x x y summary(g)$residuals

x

y summary(g)$residuals Figure 5.9: Sample data with their best ﬁtting lines (top row) and their corresponding residual plots (bottom row).

The last plot shows very little upwards trend, and the residuals also show no obvious patterns. It is reasonable to try to ﬁt a linear model to the data. However, it is unclear whether there is statistically signiﬁcant evidence that the slope parameter is diﬀerent from zero. The point estimate of the slope parameter, labeled b1, is not zero, but we might wonder if this could just be due to chance. We will address this sort of scenario in Section 5.4.

5.1.4 Describing linear relationships with correlation

Correlation: strength of a linear relationship Correlation, which always takes values between -1 and 1, describes the strength of the linear relationship between two variables. We denote the correlation by R.

R correlation

We can compute the correlation using a formula, just as we did with the sample mean and standard deviation. However, this formula is rather complex,3 so we generally perform the calculations on a computer or calculator. Figure 5.10 shows eight plots and their corresponding correlations. Only when the relationship is perfectly linear is the correlation either -1 or 1. If the relationship is strong and positive, the correlation will be near +1. If it is strong and negative, it will be near -1. If there is no apparent linear relationship between the variables, then the correlation will be near zero. The correlation is intended to quantify the strength of a linear trend. Nonlinear trends,

3Formally, we can compute the correlation for observations (x1,y1), (x2,y2), ..., (xn,yn) using the formula

R =

1 n−1

n X i=1

xi − ¯ x sx

yi − ¯ y sy

where ¯ x, ¯ y, sx, and sy are the sample means and standard deviations for each variable.

5.2. FITTING A LINE BY LEAST SQUARES REGRESSION 227

R = 0.33 R = 0.69

y

R = 0.98

y

R = 1.00

R = −0.08

y

R = −0.64

y

R = −0.92

y

R = −1.00

Figure 5.10: Sample scatterplots and their correlations. The ﬁrst row shows variables with a positive relationship, represented by the trend up and to the right. The second row shows variables with a negative trend, where a large value in one variable is associated with a low value in the other.

even when strong, sometimes produce correlations that do not reﬂect the strength of the relationship; see three such examples in Figure 5.11. J Guided Practice 5.7 It appears no straight line would ﬁt any of the datasets represented in Figure 5.11. Instead, try drawing nonlinear curves on each plot. Once you create a curve for each, describe what is important in your ﬁt.4

5.2 Fitting a line by least squares regression

Fitting linear models by eye is open to criticism since it is based on an individual preference. In this section, we use least squares regression as a more rigorous approach. This section considers family income and gift aid data from a random sample of ﬁfty students in the 2011 freshman class of Elmhurst College in Illinois.5 Gift aid is ﬁnancial aid that does not need to be paid back, as opposed to a loan. A scatterplot of the data is shown in Figure 5.12 along with two linear ﬁts. The lines follow a negative trend in the data; students who have higher family incomes tended to have lower gift aid from the university. J Guided Practice 5.8 Is the correlation positive or negative in Figure 5.12?6 4We’ll leave it to you to draw the lines. In general, the lines you draw should be close to most points and reﬂect overall trends in the data. 5These data were sampled from a table of data for all freshman from the 2011 class at Elmhurst College that accompanied an article titled What Students Really Pay to Go to College published online by The Chronicle of Higher Education: chronicle.com/article/What-Students-Really-Pay-to-Go/131435 6Larger family incomes are associated with lower amounts of aid, so the correlation will be negative. Using a computer, the correlation can be computed: -0.499.

228 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION

R = −0.23

y

R = 0.31

y

R = 0.50

Figure 5.11: Sample scatterplots and their correlations. In each case, there is a strong relationship between the variables. However, the correlation is not very strong, and the relationship is not linear.

Family income ($1000s)

0 100 200

0

10

20

30

Gift aid from university ($1000s)

Figure 5.12: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College. Two lines are ﬁt to the data, the solid line being the least squares line.

5.2. FITTING A LINE BY LEAST SQUARES REGRESSION 229

5.2.1 An objective measure for ﬁnding the best line

We begin by thinking about what we mean by “best”. Mathematically, we want a line that has small residuals. Perhaps our criterion could minimize the sum of the residual magnitudes: |e1|+|e2|+···+|en| (5.9) which we could accomplish with a computer program. The resulting dashed line shown in Figure 5.12 demonstrates this ﬁt can be quite reasonable. However, a more common practice is to choose the line that minimizes the sum of the squared residuals: e2 1 + e2 2 +···+ e2 n (5.10) The line that minimizes this least squares criterion is represented as the solid line in Figure 5.12. This is commonly called the least squares line. The following are three possible reasons to choose Criterion (5.10) over Criterion (5.9):

1. It is the most commonly used method.

2. Computing the line based on Criterion (5.10) is much easier by hand and in most statistical software.

3. In many applications, a residual twice as large as another residual is more than twice as bad. For example, being oﬀ by 4 is usually more than twice as bad as being oﬀ by 2. Squaring the residuals accounts for this discrepancy.

The ﬁrst two reasons are largely for tradition and convenience; the last reason explains why Criterion (5.10) is typically most helpful.7

5.2.2 Finding the least squares line

For the Elmhurst data, we could write the equation of the least squares regression line as c aid = β0 + β1 ×family income Here the equation is set up to predict gift aid based on a student’s family income, which would be useful to students considering Elmhurst. These two values, β0 and β1, are the parameters of the regression line. As in Chapters 4-6, the parameters are estimated using observed data. In practice, this estimation is done using a computer in the same way that other estimates, like a sample mean, can be estimated using a computer or calculator. However, we can also ﬁnd the parameter estimates by applying two properties of the least squares line: • The slope of the least squares line can be estimated by

b1 =

sy sx

R (5.11)

where R is the correlation between the two variables, and sx and sy are the sample standard deviations of the explanatory variable and response, respectively. • If ¯ x is the mean of the horizontal variable (from the data) and ¯ y is the mean of the vertical variable, then the point (¯ x, ¯ y) is on the least squares line.

b0,b1 Sample estimates of β0, β1

We use b0 and b1 to represent the point estimates of the parameters β0 and β1. 7There are applications where Criterion (5.9) may be more useful, and there are plenty of other criteria we might consider. However, this book only applies the least squares criterion.

230 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION J Guided Practice 5.12 Table 5.13 shows the sample means for the family income and gift aid as $101,800 and $19,940, respectively. Plot the point (101.8,19.94) on Figure 5.12 on page 228 to verify it falls on the least squares line (the solid line).8

family income, in $1000s (“x”) gift aid, in $1000s (“y”) mean ¯ x = 101.8 ¯ y = 19.94 sd sx = 63.2 sy = 5.46 R = −0.499 Table 5.13: Summary statistics for family income and gift aid.

J Guided Practice 5.13 Using the summary statistics in Table 5.13, compute the slope for the regression line of gift aid against family income.9

You might recall the point-slope form of a line from math class (another common form is slope-intercept). Given the slope of a line and a point on the line, (x0,y0), the equation for the line can be written as

y−y0 = slope×(x−x0) (5.14)

A common exercise to become more familiar with foundations of least squares regression is to use basic summary statistics and point-slope form to produce the least squares line.

TIP: Identifying the least squares line from summary statistics To identify the least squares line from summary statistics: • Estimate the slope parameter, β1, by calculating b1 using Equation (5.11). • Noting that the point (¯ x, ¯ y) is on the least squares line, use x0 = ¯ x and y0 = ¯ y along with the slope b1 in the point-slope equation: y− ¯ y = b1(x− ¯ x) • Simplify the equation.

8If you need help ﬁnding this location, draw a straight line up from the x-value of 100 (or thereabout). Then draw a horizontal line at 20 (or thereabout). These lines should intersect on the least squares line. 9Apply Equation (5.11) with the summary statistics from Table 5.13 to compute the slope:

b1 =

sy sx

R =

5.46 63.2

(−0.499) = −0.0431

5.2. FITTING A LINE BY LEAST SQUARES REGRESSION 231 Example 5.15 Using the point (101.8,19.94) from the sample means and the slope estimate b1 = −0.0431 from Guided Practice 5.13, ﬁnd the least-squares line for predicting aid based on family income.

Apply the point-slope equation using (101.8,19.94) and the slope b1 = −0.0431: y−y0 = b1(x−x0) y−19.94 = −0.0431(x−101.8) Expanding the right side and then adding 19.94 to each side, the equation simpliﬁes: c aid = 24.3−0.0431×family income Here we have replaced y with c aid and x with family income to put the equation in context. We mentioned earlier that a computer is usually used to compute the least squares line. A summary table based on computer output is shown in Table 5.14 for the Elmhurst data. The ﬁrst column of numbers provides estimates for b0 and b1, respectively. Compare these to the result from Example 5.15.

Estimate Std. Error t value Pr(>|t|) (Intercept) 24.3193 1.2915 18.83 0.0000 family income -0.0431 0.0108 -3.98 0.0002

Table 5.14: Summary of least squares ﬁt for the Elmhurst data. Compare the parameter estimates in the ﬁrst column to the results of Example 5.15. Example 5.16 Examine the second, third, and fourth columns in Table 5.14. Can you guess what they represent? We’ll describe the meaning of the columns using the second row, which corresponds to β1. The ﬁrst column provides the point estimate for β1, as we calculated in an earlier example: -0.0431. The second column is a standard error for this point estimate: 0.0108. The third column is a t test statistic for the null hypothesis that β1 = 0: T = −3.98. The last column is the p-value for the t test statistic for the null hypothesis β1 = 0 and a two-sided alternative hypothesis: 0.0002. We will get into more of these details in Section 5.4. Example 5.17 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?

She may use it as an estimate, though some qualiﬁers on this approach are important. First, the data all come from one freshman class, and the way aid is determined by the university may change from year to year. Second, the equation will provide an imperfect estimate. While the linear equation is good at capturing the trend in the data, no individual student’s aid will be perfectly predicted.

232 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION

5.2.3 Interpreting regression line parameter estimates

Interpreting parameters in a regression model is often one of the most important steps in the analysis. Example 5.18 The slope and intercept estimates for the Elmhurst data are -0.0431 and 24.3. What do these numbers really mean?

Interpreting the slope parameter is helpful in almost any application. For each additional $1,000 of family income, we would expect a student to receive a net diﬀerence of $1,000 × (−0.0431) = −$43.10 in aid on average, i.e. $43.10 less. Note that a higher family income corresponds to less aid because the coeﬃcient of family income is negative in the model. We 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. That is, increasing a student’s family income may not cause the student’s aid to drop. (It would be reasonable to contact the college and ask if the relationship is causal, i.e. if Elmhurst College’s aid decisions are partially based on students’ family income.) The estimated intercept b0 = 24.3 (in $1000s) describes the average aid if a student’s family had no income. The meaning of the intercept is relevant to this application since the family income for some students at Elmhurst is $0. In other applications, the intercept may have little or no practical value if there are no observations where x is near zero.

Interpreting parameters estimated by least squares The slope describes the estimated diﬀerence in the y variable if the explanatory variable x for a case happened to be one unit larger. The intercept describes the average outcome of y if x = 0 and the linear model is valid all the way to x = 0, which in many applications is not the case.

5.2.4 Extrapolation is treacherous

When those blizzards hit the East Coast this winter, it proved to my satisfaction that global warming was a fraud. That snow was freezing cold. But in an alarming trend, temperatures this spring have risen. Consider this: On February 6th it was 10 degrees. Today it hit almost 80. At this rate, by August it will be 220 degrees. So clearly folks the climate debate rages on. Stephen Colbert April 6th, 2010 10

Linear models can be used to approximate the relationship between two variables. However, these models have real limitations. Linear regression is simply a modeling framework. The truth is almost always much more complex than our simple line. For example, we do not know how the data outside of our limited window will behave.

10http://www.colbertnation.com/the-colbert-report-videos/269929/

5.2. FITTING A LINE BY LEAST SQUARES REGRESSION 233

Family income ($1000s)

0 100 200

0

10

20

30

Gift aid from university ($1000s)

Figure 5.15: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College, shown with the least squares regression line. Example 5.19 Use the model c aid = 24.3−0.0431×family income to estimate the aid of another freshman student whose family had income of $1 million. Recall that the units of family income are in $1000s, so we want to calculate the aid for family income = 1000: 24.3−0.0431×family income = 24.3−0.0431×1000 = −18.8 The model predicts this student will have -$18,800 in aid (!). Elmhurst College cannot (or at least does not) require any students to pay extra on top of tuition to attend.

Applying a model estimate to values outside of the realm of the original data is called extrapolation. Generally, a linear model is only an approximation of the real relationship between two variables. If we extrapolate, we are making an unreliable bet that the approximate linear relationship will be valid in places where it has not been explored.

5.2.5 Using R2 to describe the strength of a ﬁt

We evaluated the strength of the linear relationship between two variables earlier using the correlation, R. However, it is more common to explain the strength of a linear ﬁt using R2, called R-squared. If provided with a linear model, we might like to describe how closely the data cluster around the linear ﬁt. The R2 of a linear model describes the amount of variation in the response that is explained by the least squares line. For example, consider the Elmhurst data, shown in Figure 5.15. The variance of the response variable, aid received, is s2aid = 29.8. However, if we apply our least squares line, then this model reduces our uncertainty in predicting aid using a student’s family income. The variability in the residuals describes how much variation remains after using the model: s2RES = 22.4. In short, there was a reduction of s2aid −s2RES s2aid = 29.8−22.4 29.8 = 7.5 29.8 = 0.25

234 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION

0 (used)

1 (new)

30

40

50

60

70

Total Price

price = 42.87 + 10.90 cond\_new

Figure 5.16: Total auction prices for the video game Mario Kart, divided into used (x = 0) and new (x = 1) condition games. The least squares regression line is also shown.

or about 25% in the data’s variation by using information about family income for predicting aid using a linear model. This corresponds exactly to the R-squared value: R = −0.499 R2 = 0.25 J Guided Practice 5.20 If a linear model has a very strong negative relationship with a correlation of -0.97, how much of the variation in the response is explained by the explanatory variable?11

5.2.6 Categorical predictors with two levels

Categorical variables are also useful in predicting outcomes. Here we consider a categorical predictor with two levels (recall that a level is the same as a category). We’ll consider Ebay auctions for a video game, Mario Kart for the Nintendo Wii, where both the total price of the auction and the condition of the game were recorded.12 Here we want to predict total price based on game condition, which takes values used and new. A plot of the auction data is shown in Figure 5.16. To incorporate the game condition variable into a regression equation, we must convert the categories into a numerical form. We will do so using an indicator variable called cond new, which takes value 1 when the game is new and 0 when the game is used. Using this indicator variable, the linear model may be written as d price = β0 + β1 ×cond new 11About R2 = (−0.97)2 = 0.94 or 94% of the variation is explained by the linear model. 12These data were collected in Fall 2009 and may be found at openintro.org.

5.3. TYPES OF OUTLIERS IN LINEAR REGRESSION 235

Estimate Std. Error t value Pr(>|t|) (Intercept) 42.87 0.81 52.67 0.0000 cond new 10.90 1.26 8.66 0.0000

Table 5.17: Least squares regression summary for the ﬁnal auction price against the condition of the game.

The ﬁtted model is summarized in Table 5.17, and the model with its parameter estimates is given as d price = 42.87 + 10.90×cond new Example 5.21 Interpret the two parameters estimated in the model for the price of Mario Kart in eBay auctions.

The intercept is the estimated price when cond new takes value 0, i.e. when the game is in used condition. That is, the average selling price of a used version of the game is $42.87.

The slope indicates that, on average, new games sell for about $10.90 more than used games.

TIP: Interpreting model estimates for categorical predictors. The estimated intercept is the value of the response variable for the ﬁrst category (i.e. the category corresponding to an indicator value of 0). The estimated slope is the average change in the response variable between the two categories.

We’ll elaborate further on this Ebay auction data in Chapter 6, where we examine the inﬂuence of many predictor variables simultaneously using multiple regression. In multiple regression, we will consider the association of auction price with regard to each variable while controlling for the inﬂuence of other variables. This is especially important since some of the predictors are associated. For example, auctions with games in new condition also often came with more accessories.

5.3 Types of outliers in linear regression

In this section, we identify criteria for determining which outliers are important and inﬂuential. Outliers in regression are observations that fall far from the “cloud” of points. These points are especially important because they can have a strong inﬂuence on the least squares line.

236 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION Example 5.22 There are six plots shown in Figure 5.18 along with the least squares line and residual plots. For each scatterplot and residual plot pair, identify any obvious outliers and note how they inﬂuence the least squares line. Recall that an outlier is any point that doesn’t appear to belong with the vast majority of the other points.

(1) There is one outlier far from the other points, though it only appears to slightly inﬂuence the line. (2) There is one outlier on the right, though it is quite close to the least squares line, which suggests it wasn’t very inﬂuential. (3) There is one point far away from the cloud, and this outlier appears to pull the least squares line up on the right; examine how the line around the primary cloud doesn’t appear to ﬁt very well. (4) There is a primary cloud and then a small secondary cloud of four outliers. The secondary cloud appears to be inﬂuencing the line somewhat strongly, making the least square line ﬁt poorly almost everywhere. There might be an interesting explanation for the dual clouds, which is something that could be investigated. (5) There is no obvious trend in the main cloud of points and the outlier on the right appears to largely control the slope of the least squares line. (6) There is one outlier far from the cloud, however, it falls quite close to the least squares line and does not appear to be very inﬂuential.

Examine the residual plots in Figure 5.18. You will probably ﬁnd that there is some trend in the main clouds of (3) and (4). In these cases, the outliers inﬂuenced the slope of the least squares lines. In (5), data with no clear trend were assigned a line with a large trend simply due to one outlier (!).

Leverage Points that fall horizontally away from the center of the cloud tend to pull harder on the line, so we call them points with high leverage.

Points that fall horizontally far from the line are points of high leverage; these points can strongly inﬂuence the slope of the least squares line. If one of these high leverage points does appear to actually invoke its inﬂuence on the slope of the line – as in cases (3), (4), and (5) of Example 5.22 – then we call it an inﬂuential point. Usually we 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. It is tempting to remove outliers. Don’t do this without a very good reason. Models that ignore exceptional (and interesting) cases often perform poorly. For instance, if a ﬁnancial ﬁrm ignored the largest market swings – the “outliers” – they would soon go bankrupt by making poorly thought-out investments.

Caution: Don’t ignore outliers when ﬁtting a ﬁnal model If there are outliers in the data, they 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.

5.3. TYPES OF OUTLIERS IN LINEAR REGRESSION 237

(1) (2) (3)

(4) (5) (6)

Figure 5.18: Six plots, each with a least squares line and residual plot. All data sets have at least one outlier.

238 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION

Caution: Outliers for a categorical predictor with two levels Be cautious about using a categorical predictor when one of the levels has very few observations. When this happens, those few observations become inﬂuential points.

5.4 Inference for linear regression

In this section we discuss uncertainty in the estimates of the slope and y-intercept for a regression line. Just as we identiﬁed standard errors for point estimates in previous chapters, we ﬁrst discuss standard errors for these new estimates. However, in the case of regression, we will identify standard errors using statistical software.

5.4.1 Conditions for the least squares line

When performing inference on a least squares line, we generally require the following:

Linearity. The data should show a linear trend. If there is a nonlinear trend (e.g. left panel of Figure 5.19), an advanced regression method from another book or later course should be applied. Nearly normal residuals. Generally the residuals must be nearly normal. When this condition is found to be unreasonable, it is usually because of outliers or concerns about inﬂuential points, which we will discuss in greater depth in Section 5.3. An example of non-normal residuals is shown in the second panel of Figure 5.19. Constant variability. The variability of points around the least squares line remains roughly constant. An example of non-constant variability is shown in the third panel of Figure 5.19. Independent observations. Be cautious about applying regression to data collected sequentially in what is called a time series. Such data may have an underlying structure that should be considered in a model and analysis. An example of a time series where independence is violated is shown in the fourth panel of Figure 5.19.

For additional information on checking regression conditions, see Section 6.3. Example 5.23 Should we have concerns about applying inference to the Elmhurst data in Figure 5.20? The trend appears to be linear, the data fall around the line with no obvious outliers, the variance is roughly constant. These are also not time series observations. It would be reasonable to analyze the model using inference.

5.4.2 Midterm elections and unemployment

Elections for members of the United States House of Representatives occur every two years, coinciding every four years with the U.S. Presidential election. The set of House elections occurring during the middle of a Presidential term are called midterm elections. In America’s two-party system, one political theory suggests the higher the unemployment rate, the worse the President’s party will do in the midterm elections. To assess the validity of this claim, we can compile historical data and look for a connection. We consider every midterm election from 1898 to 2010, with the exception

5.4. INFERENCE FOR LINEAR REGRESSION 239

x x y g$residuals

x

y g$residuals

x

y g$residuals

Figure 5.19: Four examples showing when the methods in this chapter are insuﬃcient to apply to the data. In the left panel, a straight line does not ﬁt the data. In the second panel, there are outliers; two points on the left are relatively distant from the rest of the data, and one of these points is very far away from the line. In the third panel, the variability of the data around the line increases with larger values of x. In the last panel, a time series data set is shown, where successive observations are highly correlated.

Family income ($1000s)

0 100 200

0

10

20

30

Gift aid from university ($1000s)

Figure 5.20: Gift aid and family income for a random sample of 50 freshman students from Elmhurst College. Two lines are ﬁt to the data, the solid line being the least squares line.

240 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION

Percent change in seats of president's party in House of Rep.

Percent unemployment

●

●

●

●

●

●

●

●

●

●

●

4 8 12

−30

−20

−10

0

10

McKinley 1898

Reagan 1982

Clinton 1998

Bush 2002

Obama 2010

● Democrat Republican

Figure 5.21: The percent change in House seats for the President’s party in each election from 1898 to 2010 plotted against the unemployment rate. The two points for the Great Depression have been removed, and a least squares regression line has been ﬁt to the data.

of those elections during the Great Depression. Figure 5.21 shows these data and the least-squares regression line: % change in House seats for President’s party = −6.71−1.00×(unemployment rate) We consider the percent change in the number of seats of the President’s party (e.g. percent change in the number of seats for Democrats in 2010) against the unemployment rate. Examining the data, there are no clear deviations from linearity, the constant variance condition, or in the normality of residuals (though we don’t examine a normal probability plot here). While the data are collected sequentially, a separate analysis was used to check for any apparent correlation between successive observations; no such correlation was found. J Guided Practice 5.24 The data for the Great Depression (1934 and 1938) were removed because the unemployment rate was 21% and 18%, respectively. Do you agree that they should be removed for this investigation? Why or why not?13

There is a negative slope in the line shown in Figure 5.21. However, this slope (and the y-intercept) are only estimates of the parameter values. We might wonder, is this convincing evidence that the “true” linear model has a negative slope? That is, do the data provide strong evidence that the political theory is accurate? We can frame this

13We will provide two considerations. Each of these points would have very high leverage on any least-squares regression line, and years with such high unemployment may not help us understand what would happen in other years where the unemployment is only modestly high. On the other hand, these are exceptional cases, and we would be discarding important information if we exclude them from a ﬁnal analysis.

5.4. INFERENCE FOR LINEAR REGRESSION 241

investigation into a two-sided statistical hypothesis test. We use a two-sided test since a statistically signiﬁcant result in either direction would be interesting.

H0: β1 = 0. The true linear model has slope zero.

HA: β1 6= 0. The true linear model has a slope diﬀerent than zero. The higher the unemployment, the greater the loss for the President’s party in the House of Representatives, or vice-versa.

We would reject H0 in favor of HA if the data provide strong evidence that the true slope parameter is less than zero. To assess the hypotheses, we identify a standard error for the estimate, compute an appropriate test statistic, and identify the p-value.

5.4.3 Understanding regression output from software

Just like other point estimates we have seen before, we can compute a standard error and test statistic for b1. We will generally label the test statistic using a T, since it follows the t distribution. We will rely on statistical software to compute the standard error and leave the explanation of how this standard error is determined to a second or third statistics course. Table 5.22 shows software output for the least squares regression line in Figure 5.21. The row labeled unemp represents the information for the slope, which is the coeﬃcient of the unemployment variable.

Estimate Std. Error t value Pr(>|t|) (Intercept) -6.7142 5.4567 -1.23 0.2300 unemp -1.0010 0.8717 -1.15 0.2617 df = 25

Table 5.22: Output from statistical software for the regression line modeling the midterm election gains and losses for the President’s party as a response to unemployment.

Example 5.25 What do the ﬁrst and second columns of Table 5.22 represent? The entries in the ﬁrst column represent the least squares estimates, b0 and b1, and the values in the second column correspond to the standard errors of each estimate.

We previously used a t test statistic for hypothesis testing in the context of numerical data. Regression is very similar. In the hypotheses we consider, the null value for the slope is 0, so we can compute the test statistic using the T (or Z) score formula:

T =

estimate−null value SE

= −1.0010−0 0.8717

= −1.15

We can look for the two-tailed p-value – shown in Figure 5.23 – using the probability table for the t distribution in Appendix C.2 on page 342.

242 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION

−2.62 −1.74 −0.87 0 0.87 1.74 2.62

Figure 5.23: The distribution shown here is the sampling distribution for b1, if the null hypothesis was true. The shaded tail represents the p-value for the hypothesis test evaluating whether there is convincing evidence that higher unemployment corresponds to a greater loss of House seats for the President’s party during a midterm election. Example 5.26 Table 5.22 oﬀers the degrees of freedom for the test statistic T: df = 25. Identify the p-value for the hypothesis test.

Looking in the 25 degrees of freedom row in Appendix C.2, we see that the absolute value of the test statistic is smaller than any value listed, which means the tail area and therefore also the p-value is larger than 0.200 (two tails!). Because the p-value is so large, we fail to reject the null hypothesis. That is, the data do not provide convincing evidence that unemployment is a good predictor of how well a president’s party will do in the midterm elections for the House of Representatives.

We could have identiﬁed the t test statistic from the software output in Table 5.22, shown in the second row (unemp) and third column (t value). The entry in the second row and last column in Table 5.22 represents the p-value for the two-sided hypothesis test where the null value is zero.

Inference for regression We usually rely on statistical software to identify point estimates and standard errors for parameters of a regression line. After verifying conditions hold for ﬁtting a line, we can use the methods learned in Section 4.1 for the t distribution to create conﬁdence intervals for regression parameters or to evaluate hypothesis tests.

Caution: Don’t carelessly use the p-value from regression output The last column in regression output often lists p-values for one particular hypothesis: a two-sided test where the null value is zero. If a 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.

5.4. INFERENCE FOR LINEAR REGRESSION 243 Example 5.27 Examine Figure 5.15 on page 233, which relates the Elmhurst College aid and student family income. How sure are you that the slope is statistically signiﬁcantly diﬀerent from zero? That is, do you think a formal hypothesis test would reject the claim that the true slope of the line should be zero?

While the relationship between the variables is not perfect, there is an evident decreasing trend in the data. This suggests the hypothesis test will reject the null claim that the slope is zero. J Guided Practice 5.28 Table 5.24 shows statistical software output from ﬁtting the least squares regression line shown in Figure 5.15. Use this output to formally evaluate the following hypotheses. H0: The true coeﬃcient for family income is zero. HA: The true coeﬃcient for family income is not zero.14

Estimate Std. Error t value Pr(>|t|) (Intercept) 24.3193 1.2915 18.83 0.0000 family income -0.0431 0.0108 -3.98 0.0002 df = 48

Table 5.24: Summary of least squares ﬁt for the Elmhurst College data.

TIP: Always check assumptions If conditions for ﬁtting the regression line do not hold, then the methods presented here should not be applied. The standard error or distribution assumption of the point estimate – assumed to be normal when applying the t test statistic – may not be valid.

5.4.4 An alternative test statistic

We considered the t test statistic as a way to evaluate the strength of evidence for a hypothesis test in Section 5.4.3. However, we could focus on R2. Recall that R2 described the proportion of variability in the response variable (y) explained by the explanatory variable (x). If this proportion is large, then this suggests a linear relationship exists between the variables. If this proportion is small, then the evidence provided by the data may not be convincing. This concept – considering the amount of variability in the response variable explained by the explanatory variable – is a key component in some statistical techniques. The analysis of variance (ANOVA) technique introduced in Section 4.4 uses this general principle. The method states that if enough variability is explained away by the categories, then we would 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 inclusion of explanatory variables, though we leave these details to a later course. 14We look in the second row corresponding to the family income variable. We see the point estimate of the slope of the line is -0.0431, the standard error of this estimate is 0.0108, and the t test statistic is -3.98. The p-value corresponds exactly to the two-sided test we are interested in: 0.0002. The p-value is so small that we reject the null hypothesis and conclude that family income and ﬁnancial aid at Elmhurst College for freshman entering in the year 2011 are negatively correlated and the true slope parameter is indeed less than 0, just as we believed in Example 5.27.

244 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION

5.5 Exercises

5.5.1 Line ﬁtting, residuals, and correlation

5.1 Visualize the residuals. The scatterplots shown below each have a superimposed regression line. If we were to construct a residual plot (residuals versus x) for each, describe what those plots would look like.

● ●

●

● ●●● ● ●●●

●

● ●●

●

● ●

●

●

●

●

●

●

●●

●

●

● ●

●

●

●

●

●

● ●●

●● ● ● ● ●● ● ●● ●●● ●● ● ● ●

●

● ● ●●

● ●● ●● ● ● ● ● ● ●

● ● ● ● ●●

●

●● ● ● ● ●

●

●

●

●

●

●

● ● ●

●

●

●

●

● ●

(a)

●

●

●

●

●

●

●

●

● ●

●

● ●

●

●

●

●

●●

●

●●

● ● ●

●

●

●

●

●

●

●

●

● ●

● ●

●

●

●

●●

●

●

●● ●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●●

●

● ●●● ●●●●● ●● ● ●●● ● ● ● ● ● ●● ● ● ●●●●●●●

(b)

5.2 Trends in the residuals. Shown below are two plots of residuals remaining after ﬁtting a linear model to two diﬀerent sets of data. Describe important features and determine if a linear model would be appropriate for these data. Explain your reasoning.

● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

●

●

●

● ●

●

●

●

● ● ●

●

● ●

●

●

●

●

● ●

●

●

●

● ● ● ●

●

● ●

● ●

● ●

●

●

● ●

●

● ● ●

●

●

●

●

●

●

● ● ● ● ● ● ● ● ●

●

●

●

●

●

●

●

● ● ●

●

●

●

●

●

●

●

(a)

●

● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

●

●

● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

(b)

5.3 Identify relationships, Part I. For each of the six plots, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether ﬁtting a linear model would be reasonable.

●

●

● ●

●

●

●

●

● ● ●

●

● ●

●

●

● ●

●

●

●

● ●

●

●

●

●

●

●

●

● ●

● ● ●

●

●

●

● ●

●

●

● ● ●

●

● ●

●

●

● ●

●

●

●

● ●

●

● ● ●

●

●

●

● ● ●

●

●

● ●

● ●

●

●

●

●

●

● ● ●

●

● ●

●

●

● ● ●

●

●

● ● ● ●

●

● ● ● ● ● ● ● ● ● ●

●

●

● ●

●

● ●

●

● ●

●

●

● ●

●

(a)

● ● ● ● ●

●

●

●

● ●

● ●

●

●

● ●

●

● ●

● ● ● ● ● ● ● ● ● ●

●

●

●

● ● ●

● ●

●

● ● ● ● ●

●

● ● ●

●

● ●

●

●

● ● ●

●

● ●

● ● ● ● ● ● ● ●

●

● ●

●

●

● ● ● ● ● ●

● ● ●

● ● ●

●

●

●

●

●

●

●

●

● ● ●

●

●

● ●

● ● ● ●

● ●

●

●

●

● ●

● ● ● ● ● ●

●

● ●

● ●

●

(b)

●

●

●

●

●

●

●

●

●

● ●

● ●

●

●

●

●

●

● ●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

● ●

●

● ●

●

●

● ●

●

●

●

●

● ● ●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

● ● ● ●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

● ●

●

●

●

● ●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

(c)

●

●

●

●

● ● ●

●

● ●

● ●

●

●

● ● ●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

● ●

● ●

●

●

●

●

●

● ● ●

● ●

●

●

●

●

●

● ●

●

●

●

● ●

●

●

●

●

●

● ● ●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

● ● ●

● ●

●

●

●

●

●

●

●

● ●

(d)

● ● ●

● ●

●

●

●

●

●

●

●

●

●

●

● ● ● ● ● ● ● ● ●

● ●

●

●

● ● ●

● ●

●

● ● ● ●

●

● ●

● ● ● ● ●

●

● ● ● ● ● ● ● ● ● ● ● ● ● ●

●

●

● ● ● ● ● ●

●

● ● ● ● ●

● ● ●

●

●

●

●

● ● ● ● ● ● ● ● ● ● ● ● ● ●

●

● ● ●

● ● ●

●

●

● ●

●

●

●

● ● ● ● ● ● ● ● ● ● ●

(e)

● ● ●

● ● ● ● ● ● ●

●

●

● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

(f)

5.5. EXERCISES 245

5.4 Identify relationships, Part I. For each of the six plots, identify the strength of the relationship (e.g. weak, moderate, or strong) in the data and whether ﬁtting a linear model would be reasonable.

●

●

●

● ● ● ● ●

● ●

●

● ● ● ● ● ● ● ● ● ● ● ●

●

●

● ● ●

●

●

●

● ●

●

●

● ●

● ●

●

●

● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

● ●

●

●

●

● ● ● ● ●

● ●

● ●

(a)

●

● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

●

●

●

● ● ● ●

● ●

● ● ● ● ● ● ● ● ●

●

● ● ● ●

●

● ● ● ● ● ● ● ● ● ●

●

●

● ● ● ● ● ● ● ● ● ● ●

● ●

● ●

● ●

●

●

●

●

●

● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

● ●

● ●

● ● ● ● ● ● ● ● ● ● ● ●

●

●

●

●

●

●

●

●

●

● ● ● ● ●

●

●

●

● ● ●

●

● ● ●

● ● ●

(b)

●

●

●

●

●

●

● ●

●

●

● ● ●

●

● ●

● ●

●

● ●

●

●

● ●

●

●

● ●

●

●

● ●

● ● ●

●

●

●

●

●

● ●

●

●

●

●

● ● ● ●

● ●

●

● ● ● ● ●

●

●

●

● ●

●

●

●

●

●

●

● ●

● ● ●

● ● ●

●

● ●

● ● ● ●

●

● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

●

●

● ●

● ●

● ● ● ●

●

● ● ●

●

● ●

●

● ● ●

● ●

●

● ●

● ● ●

●

●

●

●

● ● ● ●

●

●

(c)

● ●

●

●

● ● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ● ● ● ● ● ●

● ●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

● ●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

● ●

●

●

● ● ● ● ● ●

●

●

●

●

●

● ●

● ● ●

●

●

●

●

●

●

●

●

●

● ● ●

●

●

●

●

●

●

●

●

●

(d)

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

● ● ● ●

●

●

●

●

●

●

●

●

●

● ● ●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

● ● ●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

● ●

●

(e)

●

● ●

●

●

●

●

●

●

●

● ●

●

●

●

●

● ● ●

●

●

●

● ●

●

●

●

● ●

●

●

●

● ●

●

●

●

●

● ●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

● ●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

● ●

● ● ●

●

● ●

●

●

●

●

●

●

● ●

● ●

●

●

●

●

● ●

●

●

●

●

● ●

● ●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

● ● ●

●

(f)

5.5 Exams and grades. The two scatterplots below show the relationship between ﬁnal and mid-semester exam grades recorded during several years for a Statistics course at a university. (a) Based on these graphs, which of the two exams has the strongest correlation with the ﬁnal exam grade? Explain. (b) Can you think of a reason why the correlation between the exam you chose in part (a) and the ﬁnal exam is higher?

Exam 1

Final Exam

40 60 80 100

40

60

80

100

Exam 2

Final Exam

40 60 80 100

40

60

80

100

246 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION

5.6 Husbands and wives, Part I. The Great Britain Oﬃce of Population Census and Surveys once collected data on a random sample of 170 married couples in Britain, recording the age (in years) and heights (converted here to inches) of the husbands and wives.15 The scatterplot on the left shows the wife’s age plotted against her husband’s age, and the plot on the right shows wife’s height plotted against husband’s height.

Husband's age (in years)

Wife's age (in years)

20 40 60

20

40

60

Husband's height (in inches)

Wife's height (in inches)

60 65 70 75

55

60

65

70

(a) Describe the relationship between husbands’ and wives’ ages.

(b) Describe the relationship between husbands’ and wives’ heights.

(c) Which plot shows a stronger correlation? Explain your reasoning.

(d) Data on heights were originally collected in centimeters, and then converted to inches. Does this conversion aﬀect the correlation between husbands’ and wives’ heights?

5.7 Match the correlation, Part I. Match the calculated correlations to the corresponding scatterplot. (a) R = −0.7 (b) R = 0.45 (c) R = 0.06 (d) R = 0.92

●

● ●

●

● ●

● ● ●

● ●

● ●

●

●

●

●

● ●

●

●

● ●

●

●

●

●

●

●

●

●

● ● ●

● ●

●

●

● ●

●

● ● ● ● ● ● ● ● ●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

● ● ● ● ●

●

●

●

●

● ● ● ●

●

●

● ● ●

● ● ● ● ●

● ●

●

●

●

● ●

●

●

●

● ● ●

●

●

● ● ● ● ● ●

●

●

●

●

●

● ● ●

(1)

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

● ●

●

●

●

● ●

●

● ● ● ● ●

●

●

●

●

●

● ● ● ● ● ● ●

● ●

●

● ●

● ●

● ●

●

●

● ●

●

● ●

●

●

● ● ●

●

●

●

●

● ●

●

●

●

●

● ●

●

●

●

● ● ● ● ● ●

●

●

●

●

●

●

●

●

●

● ● ● ● ● ● ●

●

●

●

●

●

●

●

● ● ● ● ● ● ●

●

● ●

● ● ●

(2)

●

● ●

●

●

● ● ● ●

●

● ●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

● ●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

● ●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

(3)

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

● ● ●

●

●

●

●

●

●

●

● ● ● ●

●

● ●

●

●

●

● ●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

● ● ●

● ●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

● ●

●

●

●

● ●

●

●

●

●

●

●

● ●

●

● ●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

(4)

15D.J. Hand. A handbook of small data sets. Chapman & Hall/CRC, 1994.

5.5. EXERCISES 247

5.8 Match the correlation, Part II. Match the calculated correlations to the corresponding scatterplot. (a) R = 0.49 (b) R = −0.48 (c) R = −0.03 (d) R = −0.85

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

● ● ●

● ●

●

●

● ●

●

● ●

● ●

●

●

●

● ● ●

● ● ● ● ● ●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

● ● ●

● ● ● ●

● ●

●

●

● ●

● ● ● ●

●

●

●

●

●

●

●

●

●

● ● ●

●

●

●

● ●

● ●

●

●

●

●

●

● ● ● ● ● ● ●

● ●

●

● ●

● ●

●

●

●

(1)

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

● ● ●

●

●

●

●

●

●

●

● ●

●

●

● ●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

● ●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

● ●

●

(2)

● ● ● ●

● ● ●

● ● ●

●

●

● ● ● ●

●

● ●

●

●

● ●

●

● ●

●

●

●

●

● ●

●

●

● ● ●

●

●

●

● ●

●

●

●

●

●

●

● ●

●

● ● ●

●

●

●

● ● ● ●

●

● ●

●

●

●

● ● ●

●

●

● ● ● ●

●

●

●

●

●

●

●

●

● ●

● ●

●

● ● ●

● ● ●

●

●

● ● ●

● ●

●

● ● ●

●

● ● ●

●

●

● ●

●

●

● ● ●

●

●

(3)

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

● ●

●

●

●

●

●

●

● ● ●

●

●

● ●

●

●

● ●

●

●

●

●

●

●

●

●

●

● ● ● ●

●

●

●

●

●

●

●

●

●

● ●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

(4) 5.9 Speed and height. 1,302 UCLA students were asked to ﬁll out a survey where they were asked about their height, fastest speed they have ever driven, and gender. The scatterplot on the left displays the relationship between height and fastest speed, and the scatterplot on the right displays the breakdown by gender in this relationship.

Height (in inches)

Fastest speed (in mph)

60 65 70 75

0

50

100

150

Height (in inches)

Fastest speed (in mph)

60 70 80

0

50

100

150

●

female male

(a) Describe the relationship between height and fastest speed. (b) Why do you think these variables are positively associated? (c) What role does gender play in the relationship between height and fastest driving speed?

248 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION

5.10 Trees. The scatterplots below show the relationship between height, diameter, and volume of timber in 31 felled black cherry trees. The diameter of the tree is measured 4.5 feet above the ground.16

●●●

●

● ●

●

●

●

●

●

● ●● ● ●

●

●

●●

●

●

●

●

●

● ●

●

● ●

●

Height (in ft)

Volume (in cubic ft)

65 75 85

10

40

70

●●●

● ● ● ● ● ● ● ● ● ●●

●

●

●

●

● ●

● ●

●

●

●

●●

●

● ●

●

Diameter (in inches)

Volume (in cubic ft)

8 12 16 20

10

40

70

(a) Describe the relationship between volume and height of these trees.

(b) Describe the relationship between volume and diameter of these trees.

(c) Suppose you have height and diameter measurements for another black cherry tree. Which of these variables would be preferable to use to predict the volume of timber in this tree using a simple linear regression model? Explain your reasoning.

5.11 The Coast Starlight, Part I. The Coast Starlight Amtrak train runs from Seattle to Los Angeles. The scatterplot below displays the distance between each stop (in miles) and the amount of time it takes to travel from one stop to another (in minutes).

(a) Describe the relationship between distance and travel time. (b) How would the relationship change if travel time was instead measured in hours, and distance was instead measured in kilometers? (c) Correlation between travel time (in miles) and distance (in minutes) is R = 0.636. What is the correlation between travel time (in kilometers) and distance (in hours)?

●

●

●

●●

●

●

●

●

●

●

●

●

●

●

●

Distance (miles)

Travel Time (minutes)

0 100 200 300

60

120

180

240

300

360

16Source: R Dataset, http://stat.ethz.ch/R-manual/R-patched/library/datasets/html/trees.html.

5.5. EXERCISES 249

5.12 Crawling babies, Part I. A study conducted at the University of Denver investigated whether babies take longer to learn to crawl in cold months, when they are often bundled in clothes that restrict their movement, than in warmer months.17 Infants born during the study year were split into twelve groups, one for each birth month. We consider the average crawling age of babies in each group against the average temperature when the babies are six months old (that’s when babies often begin trying to crawl). Temperature is measured in degrees Fahrenheit (◦F) and age is measured in weeks.

(a) Describe the relationship between temperature and crawling age. (b) How would the relationship change if temperature was measured in degrees Celsius (◦C) and age was measured in months? (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?

●

●

●

●

●

●

●

●

●

● ●

●

30 40 50 60 70

29

30

31

32

33

34

Temperature (in F)

Average crawling age (in weeks)

5.13 Body measurements, Part I. Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender for 507 physically active individuals.18 The scatterplot below shows the relationship between height and shoulder girth (over deltoid muscles), both measured in centimeters.

(a) Describe the relationship between shoulder girth and height. (b) How would the relationship change if shoulder girth was measured in inches while the units of height remained in centimeters?

90 100 110 120 130

150

160

170

180

190

200

Shoulder girth (in cm)

Height (in cm)

17J.B. Benson. “Season of birth and onset of locomotion: Theoretical and methodological implications”. In: Infant behavior and development 16.1 (1993), pp. 69–81. issn: 0163-6383. 18G. Heinz et al. “Exploring relationships in body dimensions”. In: Journal of Statistics Education 11.2 (2003).

250 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION

5.14 Body measurements, Part II. The scatterplot below shows the relationship between weight measured in kilograms and hip girth measured in centimeters from the data described in Exercise 5.13.

(a) Describe the relationship between hip girth and weight. (b) How would the relationship change if weight was measured in pounds while the units for hip girth remained in centimeters?

80 90 100 110 120 130

40

60

80

100

120

Hip girth (in cm)

Weight (in kg)

5.15 Correlation, Part I. What would be the correlation between the ages of husbands and wives if men always married woman who were (a) 3 years younger than themselves? (b) 2 years older than themselves? (c) half as old as themselves?

5.16 Correlation, Part II. What would be the correlation between the annual salaries of males and females at a company if for a certain type of position men always made (a) $5,000 more than women? (b) 25% more than women? (c) 15% less than women?

5.5. EXERCISES 251

5.5.2 Fitting a line by least squares regression

5.17 Tourism spending. 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.

Number of tourists (in thousands)

Spending (in million $)

0 5000 10000 15000 20000 25000

050001000015000 −2000 01000

Residuals −1500 −750 0 750 1500

0

10

20

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?

Calories

Carb (in grams)

100 200 300 400 500

20406080 −40 −2002040

Residuals −30 −15 0 15 30

0

5

10

15

20

25

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.

90 100 110 120 130

150

160

170

180

190

200

Shoulder girth (in cm)

Height (in cm)

(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. ● ● ● ● ● ● ● ● ● ● ● ● 0 20 40 60 80 10 20 30 40 50 60 % receiving reduced−fee lunch % wearing helmets

254 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION

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.

●

● ●

● ●

●

●

●

●

●

●

●

● ●

● ●

●

●

● ● ●

●

●

●

●

● ●

● ●

● ●

●

●

●

●

●

● ● ● ● ●

●

●

●

●

● ●

● ● ●

(a)

● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ● ●

●

(b)

●

●

●

●●

●

●

●

●

●

●

●

●●

●●

●

●

●●●

●

●

●

●

●

●

●●

●

●

●

●

●

●

●

●

●●

●●

●

●

●

●

●

●

●

●●

(c)

5.23 Outliers, Part II. Identify the outliers in the scatterplots shown below and determine what type of outliers they are. Explain your reasoning.

● ●

●

●

●

● ●● ● ●

●

●●

●

● ● ● ● ●● ● ● ● ● ●●

●

●

● ●

●

●

●

●

● ● ●

●

● ●

● ● ●● ●● ●● ●

●

(a)

● ● ● ● ● ●●●● ● ● ●● ● ●●● ● ●● ●●● ● ●● ● ● ●● ● ● ● ● ●●● ● ●● ● ●●●●● ●●●

●

(b)

●

●

●

●

●

●

●●

●

●

●

●●

●

●

●

●

●

●●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●●

●●

●●

●

●

(c)

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?

40 50 60 70 80 90 100

45

50

55

60

65

70

% urban population

% who own home

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.

Calories

Protein (in grams)

100 200 300 400 500

0102030 −20 020

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.

●

●

●

●

●

● ●

●

●

●

●

● ●

●

●

●

●

●

●

●

●

TV (hours)

Grades (out of 100)

0 10 20 30

60708090100

●

●

●

●

●

● ●

●

●

●

●

● ●

●

●

● ●

●

●

●

●

−20−1001020

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.

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

2 4 6 8

0.05

0.10

0.15

Cans of beer

BAC (grams per deciliter)

Estimate Std. Error t value Pr(>|t|) (Intercept) -0.0127 0.0126 -1.00 0.3320 beers 0.0180 0.0024 7.48 0.0000

(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.

Height (in cm)

Weight (in kg)

150 175 200

40

60

80

100

Estimate Std. Error t value Pr(>|t|) (Intercept) -105.0113 7.5394 -13.93 0.0000 height 1.0176 0.0440 23.13 0.0000

(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.

Husband's age (in years)

Wife's age (in years)

20 40 60

20

40

60

Estimate Std. Error t value Pr(>|t|) (Intercept) 1.5740 1.1501 1.37 0.1730 age husband 0.9112 0.0259 35.25 0.0000 df = 168

(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.

Husband's height (in inches)

Wife's height (in inches)

60 65 70 75

55

60

65

70

Estimate Std. Error t value Pr(>|t|) (Intercept) 43.5755 4.6842 9.30 0.0000 height husband 0.2863 0.0686 4.17 0.0000

(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?

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

● ● ●

●

●●

●

●

●

●

●

●

% urban population

% who own home

40 60 80

5560657075

●

●

●

●

●

●

●

●

●

●

●

● ● ●

●

● ● ●●

●

●

●●

●

●

●

●

●

● ● ●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●

●●

●

−10010

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.

−1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0

2.02.53.03.54.04.55.0

beauty

teaching evaluation

−1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0

−1.5−1.0−0.50.00.51.0

beauty

residuals

residuals

Frequency

−2.0 −1.0 0.0 1.0

050100150

23J. Felton et al. “Web-based student evaluations of professors: the relations between perceived quality, easiness and sexiness”. In: Assessment & Evaluation in Higher Education 29.1 (2004), pp. 91–108.

260 CHAPTER 5. INTRODUCTION TO LINEAR REGRESSION

−3 −2 −1 0 1 2 3

−1.5−0.50.5

Normal Q−Q Plot

Theoretical Quantiles

Sample Quantiles

0 100 200 300 400

−1.5−0.50.5

order of data collection

residuals

Estimate Std. Error t value Pr(>|t|) (Intercept) 4.010 0.0255 157.21 0.0000 beauty Cell 1 0.0322 4.13 0.0000

(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.