Glassdoor Local Pay Reports | METHODOLOGY

by Dr. Andrew Chamberlain, *Chief Economist* | Mario Nuñez, *Lead Data Scientist* November 30, 2016

OVERVIEW

Glassdoor's Local Pay Reports are designed to provide a real-time overview of trends in median base pay for full-time U.S. workers nationally and in metro areas. It applies a proprietary machine-learning statistical model to millions of individual salary reports submitted anonymously to Glassdoor to estimate median base pay by job title, industry and employer size within metro areas across the U.S. This paper provides an overview of the statistical methodology underlying these reports.

OUR APPROACH

- Since 2008, Glassdoor has collected millions of pieces of content including anonymous salary reports from current and former employees, spanning more than 600,000 employers. These salary reports are provided for specific job titles and industries, at specific employers, for metro areas.
- Because Glassdoor data relies on "crowd sourced" information from current and former employees who visit Glassdoor, we do not conduct traditional representative probability samples of the labor market. Instead, the salary data Glassdoor collects reflects the composition of Glassdoor users over time. Although past research has shown the overall distribution of Glassdoor salaries to be similar to U.S. Census Bureau salaries,[1] it may differ in important ways on a month-to-month basis.
- To isolate changes in the composition of Glassdoor users over time from underlying changes in compensation in the U.S. labor market, it's necessary to statistically control for possible composition biases in the types of jobs, companies, industries and areas that are represented over time in our data. By statistically holding those changes in composition steady, Glassdoor salary data can allow us to estimate real underlying economic changes in pay over time.
- Glassdoor's "Local Pay Reports" employ a proprietary machine-learning model that estimates the separate impact of many job features on salary—job title, company, industry, seniority level, company size, metro location, and more. Each of these features is like a "building block" of a salary. We use millions of Glassdoor salaries to estimate the impact of each building block on pay. We then re-assemble those statistical building blocks into localized pay estimates for metro areas, for specific job titles, industries, company sizes, and at different points in time.
- By using *all* of the salary data on Glassdoor—not just the data collected during the previous month—the model allows us to estimate pay for more granular job titles and locations than would be possible with more traditional survey approaches, which simply total up weighted survey responses from the current month.

THE MODEL: ELASTIC NET REGRESSION

Glassdoor Local Pay Reports rely on a statistical model known as an "elastic net regression."[2] It models the relationship between a dependent variable (in this case, salary) and a set of independent variables or "predictors." For each predictor, the model returns "beta coefficients" that quantify the effect of that predictor on salaries, after controlling for the effect of all other predictors in the model. Statisticians refer to this class of models as "generalized linear models" (GLM) and they are widely used in statistical and economic research.

[1] See for example, Figure 1 on page 12 in "Demystifying the Gender Pay Gap," Glassdoor Economic Research, March 2016. Available at https://www.glassdoor.com/research/studies/gender-pay-gap/.

[2] Zou, Hui, and Trevor Hastie (2005). "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society*, Vol. 67, No. 2.

by Dr. Andrew Chamberlain, *Chief Economist* | Mario Nuñez, *Lead Data Scientist* November 30, 2016

Computationally, the elastic net regression produces a set of estimated beta coefficients by solving the minimization problem in Equation (1). For some parameter α strictly between 0 and 1 and some nonnegative parameter λ, it solves:[3]

$$\hat{\beta} = \underset{\beta_0,\beta}{\operatorname{argmin}} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \left[\frac{(1-\alpha)}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right] \quad (1)$$

where y_i is individual base pay reported to Glassdoor by person *i*, and x_i is a vector with the set of *K* predictors in the model, including job title, industry, metro location and more. As is conventional in regressions on salaries, y_i is estimated as the natural log of base salary.

By summing various combinations of beta coefficients from the model, it's possible to produce salary estimates for a given job title, in a given metro location, in a given industry, for a given employer size, at a specific month and year in time. In this way, the beta coefficients from the above model are the basis for Glassdoor's Local Pay Report. In all cases, predicted median base pay from the Local Pay Reports should be interpreted as pay for *annual salaried workers*, not hourly wage earners.

Why Medians?

By estimating equation (1) using log-transformed base pay, beta coefficients from the model have a "median" interpretation. Summing beta coefficients from the model provide the predicted value of ln(salary). Exponentiating those predicted values yield the predicted *geometric* mean of salary, not the predicted *arithmetic* mean.[4] Since salaries are approximately lognormally distributed, and because under a lognormal distribution the geometric mean is equal to the median, the model outputs should be interpreted as the predicted *median* salary for job titles, industries, company sizes and locations.

TRAINING DATA

The Local Pay Reports re-estimate the above model each month based on the latest salary data for full-time workers collected by Glassdoor. Initially, the model was estimated in October 2016 based on a sample of 2.31 million approved Glassdoor salaries for full-time workers submitted on or after January 1, 2011. Each month, this model is re-estimated using the latest approved salaries submitted to Glassdoor. The model is estimated for annual *base pay* only, and excludes all other forms of compensation such as tips, commissions, bonuses and equity. Additionally, the sample omits outlier salaries below $15,000 per year and above $600,000 per year.

METRO DEFINITIONS

The metro areas in the Local Pay Report are based on "core based statistical areas" or CBSAs defined by the U.S. Census Bureau. These areas contain both "metro" (population of 50,000 or above) and "micro" (population of between 10,000 and 50,000) areas. They include the county of the core urban area, as well as all nearby counties that are closely related as defined by patterns in commuting-to-work data.[5]

---

[3] $\|\beta\|$ is the Euclidean norm operator such that $\|\beta\| = \sqrt{\beta_1^2 + \cdots + \beta_n^2}$.

[4] For more background on interpreting regression coefficients from regressions with log-transformed variables, see Cornell University Statistical Consulting Unit (June 2012), "Interpreting Coefficients in Regression with Log-Transformed Variables," StatNews #83. Available at https://www.cscu.cornell.edu/news/statnews/stnews83.pdf.

[5] U.S. Census Bureau CBSA definitions are available at http://www.census.gov/population/metro/data/def.html. A list of cities and counties contained in U.S. Census Bureau CBSAs is available at http://www.census.gov/population/metro/data/def.html. GIS shape files for U.S. Census Bureau CBSAs are available at https://catalog.data.gov/dataset/core-based-statistical-areas-national.

Glassdoor Local Pay Reports | METHODOLOGY

by Dr. Andrew Chamberlain, *Chief Economist* | Mario Nuñez, *Lead Data Scientist* November 30, 2016

MODEL ACCURACY

Table 1 reports the estimated accuracy of the underlying salary estimates model at the time it was initially developed. To assess accuracy, we examined the "median error" of the model at the job title level when tested on samples of known Glassdoor salary data. In those tests, the model displayed a median error of 10.2 percent, which means that half of the estimates from our model were accurate to within 10.2 percent, while the other half had errors above 10.2 percent.

Additionally, we examined what percentage of Glassdoor salary reports the model was able to accurately predict at the job title level to within different levels of error. In other words, we applied the model to actual Glassdoor salary reports, and tested how often our model was able to accurately predict those individuals' pay. At the time the model was developed, it predicted 75.3 percent of Glassdoor salaries accurately to within 20 percent of the actual reported base pay, 49.1 percent of salaries to within 10 percent, and 27.2 percent of salaries to within 5 percent.

Table 1. Estimated Accuracy of Salary Estimates at the Job Title Level

	Within 5%	Within 10%	Within 20%	Median Error
Model Performance	27.2%	49.1%	75.3%	10.2%

Source: Glassdoor Economic Research (glassdoor.com/research)

COMPARISON TO MEDIAN PAY FROM U.S. BUREAU OF LABOR STATISTICS

As a benchmark of the accuracy of the Local Pay Reports, we compared estimates for median U.S. base pay from our model to official U.S. median wage and salary estimates from the U.S. Bureau of Labor Statistics. For the comparison, we used figures for "median weekly earnings of full-time wage and salary workers" from the U.S Bureau of Labor Statistics (BLS),[6] which relies on a survey of roughly 60,000 U.S. households per month to estimate pay.

For this comparison, it is important to note that there are large differences in coverage and definitions between Glassdoor and BLS data. Glassdoor's Local Pay Reports are more current than what BLS offers, and are updated once a month giving a near real-time view of wages at the local level. Additionally, Glassdoor's data is specific to actual job titles, offering more insight than broad BLS occupational groupings allow. Another key difference is in the definition of wages and salaries: BLS's definition of median wages and salaries includes tips, commissions and other cash bonuses that workers report usually receiving.[7] By contrast, the Glassdoor Local Pay Reports reflect only annual base pay, not these other forms of compensation. By benchmarking against BLS data, we hope to provide an accurate picture of how our data compare to what's been considered the norm in terms of measuring U.S. pay growth.

[6] Average weekly earnings are converted into annual pay estimates assuming a 52-week work year for full-time workers. Source: U.S. Bureau of Labor Statistics, "Labor Force Statistics from the Current Population Survey: Table 39. Median weekly earnings of full-time wage and salary workers by detailed occupation and sex." Available at http://www.bls.gov/cps/cpsaat39.htm.

[7] Source: U.S. Bureau of Labor Statistics, "Usual Weekly Earnings Technical Note," available at http://www.bls.gov/news.release/wkyeng.tn.htm.

by Dr. Andrew Chamberlain, *Chief Economist* | Mario Nuñez, *Lead Data Scientist* November 30, 2016

Table 2 shows the comparison between Glassdoor's Local Pay Report estimates of median U.S. base pay with BLS estimates of U.S. median wages and salaries for full-time workers for 2014 and 2015. The published Glassdoor Local Pay Reports produce estimates of median base pay for *annual salaried* workers only, not hourly wage workers. For this comparison to BLS data, however, we removed that restriction and report median base pay for *all* full-time workers—hourly wage and annual salaried—to make the data series more comparable.

In 2015, the Glassdoor Local Pay Reports predicted a median base pay for U.S. full-time workers of $44,282, compared to the BLS estimate of $42,068, a difference of 5.3 percent. In 2014, the Local Pay Reports predicted a median base pay of $43,808, compared with the BLS estimate of $41,132, a difference of 6.5 percent. In both years, the Local Pay Report predicted median U.S. pay to within 7 percent of BLS estimates based on large, nationally representative samples of workers.

Table 2. Comparison of Median U.S. Pay from BLS Estimates and Glassdoor's "Local Pay Reports" Model

Year	Glassdoor Local Pay Report Estimate (U.S. Median Base Pay, Full Time Workers)	BLS Estimate (U.S. Median Wages and Salaries, Full Time Workers)	Percentage Error
2015	$44,282	$42,068	5.3%
2014	$43,808	$41,132	6.5%

Source: U.S. Bureau of Labor Statistics, "Labor Force Statistics from the Current Population Survey: Table 39. Median weekly earnings of full-time wage and salary workers by detailed occupation and sex." Available at http://www.bls.gov/cps/cpsaat39.htm.

COMPARISON TO OTHER WAGE GROWTH FIGURES

As a second benchmark of the accuracy, we compared year-over-year wage growth from the Glassdoor Local Pay Reports over the past several years with two widely used official measures of wage growth: The Employment Cost Index for wages and salaries from the U.S. Bureau of Labor Statistics, and the Wage Growth Tracker from the Federal Reserve Bank of Atlanta.[8]

The Employment Cost Index is based on a quarterly survey from BLS of roughly 6,800 business establishments throughout the U.S. The survey asks employers about labor costs in their companies, and the wage and salary component is a widely used measure of pay growth by economists. The Wage Growth Tracker is an analytic data product from the research team at the Atlanta Federal Reserve. It is based on survey data from the U.S. Census Bureau's Current Population Survey, and tracks median hourly pay for workers who've been continuously employed over the past 12 months.

[8] Federal Reserve Bank of Atlanta Wage Growth Tracker is available at https://frbatlanta.org/chcs/wage-growth-tracker/?panel=1). U.S. Bureau of Labor Statistics Employment Cost Index for wages and salaries is available at http://www.bls.gov/web/eci/echistrynaics.txt.

Glassdoor Local Pay Reports | METHODOLOGY

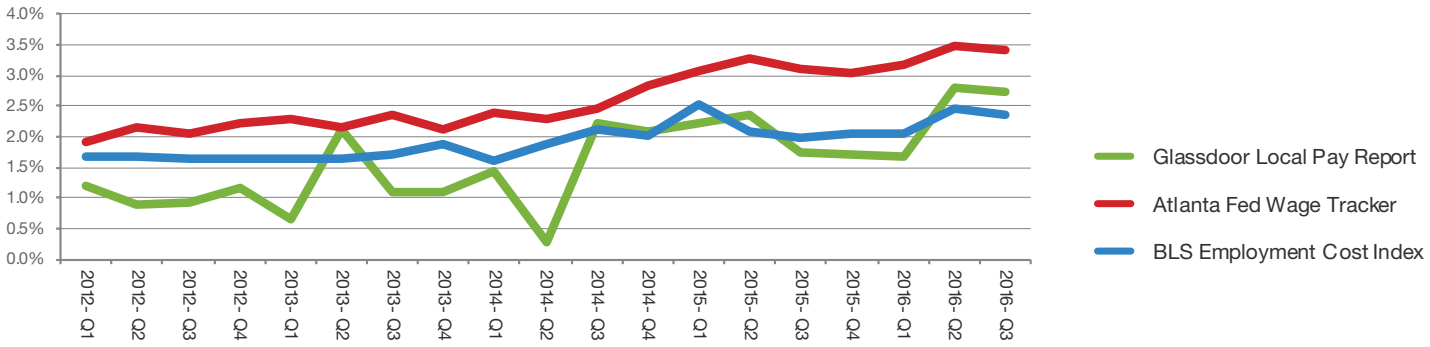
by Dr. Andrew Chamberlain, *Chief Economist* | Mario Nuñez, *Lead Data Scientist* November 30, 2016

Figure 1 shows the comparison of year-over-year wage growth in quarterly average wages from the three sources since 2012. Glassdoor's Local Pay Report is shown in green, the BLS Employment Cost Index for wages and salaries is shown in blue, while the Atlanta Fed's Wage Growth Tracker is shown in red.

Note that there are important differences in definitions and coverage between the three sources that warrant caution in interpreting this comparison. For example, the Atlanta Fed's Wage Growth Tracker follows only workers who have remained continuously employed over a 12-month period, which tend to be older and more educated than the general labor force, which explains the generally higher wage growth in that measure.

Despite many methodological differences, in general estimates of U.S. median base pay from Glassdoor's Local Pay Reports tracks other measures of wage growth quite closely, particularly in recent years. Table 3 shows the raw correlation between Local Pay Reports and these two measures, which is positive and large in both cases: 0.75 correlation with the Atlanta Fed Wage Growth Tracker, and 0.74 correlation with the BLS Employment Cost Index.

Figure 1. Comparison of Year-Over-Year Wage Growth: Glassdoor Local Pay Report, BLS Employment Cost Index, and Atlanta Fed Wage Growth Tracker



Source: Glassdoor Economic Research; Federal Reserve Bank of Atlanta Wage Growth Tracker (available at https://frbatlanta.org/chcs/wage-growth-tracker/?panel=1); U.S. Bureau of Labor Statistics Employment Cost Index, Wages and Salaries (available at http://www.bls.gov/web/eci/echistrynaics.txt).

Table 3. Correlation of Local Pay Report Wage Growth with Government Figures

	Correlation	R-Squared
Atlanta Fed Wage Tracker	0.75	59.8%
BLS Employment Cost Index	0.74	56.7%

Source: Glassdoor Economic Research (glassdoor.com/research)

Table 3 also shows the R-squared value from a simple linear regression of each official measure of wage growth against the estimated wage growth from the Glassdoor Local Pay Reports. This can be interpreted as the percentage of variation in official wage growth measures that is explained by Glassdoor salary data. The R-squared value is 59.8 percent compared to the Atlanta Fed measure, and 56.7 percent compared to the BLS measure. In both cases, wage growth from the Glassdoor Local Pay Reports explains more than half the variation in these two official wage growth measures.

CONTACT INFORMATION
For more information about the methodology behind the Glassdoor Local Pay Reports, please email Glassdoor Economic Research at economics@glassdoor.com.
For media inquiries, contact pr@glassdoor.com.

glassdoor

REFERENCES

Friedman, Jerome H., Trevor Hastie, and Rob Tibshirani (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, Vol. 33, No. 1. Available at www.jstatsoft.org/article/view/v033i01.

Hastie, Trevor and Junyang Qian (2014). "Glmnet Vignette," *CRAN Package Documentation*. Available at https://cran.r-project.org/web/packages/glmnet/vignettes/glmnet_beta.html.

Zou, Hui, and Trevor Hastie (2005). "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society*, Vol. 67, No. 2. Available at https://web.stanford.edu/~hastie/Papers/B67.2%20(2005)%20301-320%20Zou%20&%20Hastie.pdf.