Richa Shah
HRCM1-GC 3500: Special Topics - Real World Statistics
Professor Kent Bhupathi
May 6, 2025

Is it just me, or does productivity seem to be lacking of late?

Phase 1: Hypothesis Formulation #1

To evaluate the decline in productivity, this project tests the relationship between employees working from home (`wfh`) and their output (`units_completed`). The research hypothesis is:

H_1: Employees who work from home (`wfh` = yes) are less productive (as measured by `units_completed`) than those who do not (`wfh` = no).

Phase 2: Descriptive Statistics

To develop a foundational understanding of the dataset, a series of descriptive statistics were calculated across both continuous and factor variables. The continuous variables analyzed include `units_completed`, `tenure_yrs`, `dist_from_work`, `salary_annual`, and `other_exp_yrs`. The factor variables examined were `wfh`, `gender`, `education`, `level`, and `unit_type`. The `wfh` distribution showed that 84.96% (~85%) of employees do not work from home based on the output of `prop.table(table(Efficiency_Data$wfh))`. This provided values of 0.8496 for ‘no’ and 0.1504 for ‘yes’.

Likewise, the distribution of `education` and `level` showed certain categories (“high school”, “junior team”, and “senior team”) dominating the frequency tables, which also indicates an imbalance to be considered while interpreting model results. Therefore, these insights are critical to later stages of the analysis, as they guide variable selection and overall interpretation of skewness that may violate linear regression assumptions.

Phase 3: Variable Distributions

To assess the underlying structure of key variables, histograms were plotted for `units_completed`, `tenure_yrs`, and `dist_from_work`. Each histogram illustrates a distinct distribution shape that influences the modelling.

Units Completed:

The distribution of `units_completed` (Figure 1) is strongly right-skewed, with a high concentration of employees completing between 0 and 20 units, while most employees complete fewer units. This suggests that while a small subset of workers is highly productive, the majority perform at lower levels. Such skewness and disproportion imply that productivity may overestimate typical employee performance.

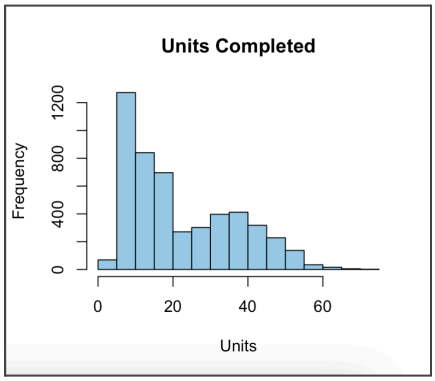

Figure 1: Histogram - Units Completed

Tenure Years:

The histogram for `tenure_yrs` (Figure 2) appears moderately right-skewed, with most employees having between 1 and 6 years of service, meaning that newer employees dominate the sample. While the distribution gradually becomes thinner at higher tenure levels, the frequency never entirely drops off. This indicates a range of experience levels in the dataset.

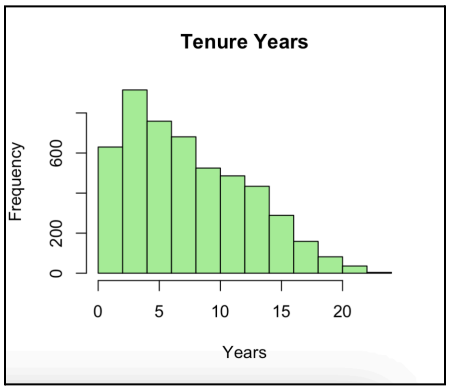

Figure 2: Histogram - Tenure Years

Distance from Work:

The distribution of `dist_from_work` is extremely right-skewed, with a spike around zero and a sharp drop-off thereafter. The skew and outliers are likely due to the presence of remote workers, whose commuting distance may be registered as 0 or a placeholder value.

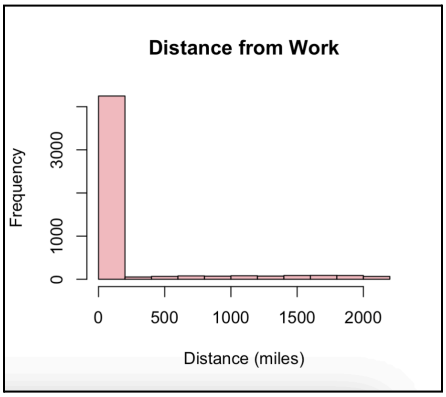

Figure 3: Histogram - Distance from Work

These visualizations validate the decision to use `units_completed` as the outcome variable and show that `tenure_yrs` and `dist_from_work` are useful predictors. Since all three variables are skewed, it is important to check the model's errors in later stages for heteroskedasticity.

Phase 4: Correlative Checks

A correlation matrix of the variables, as seen in Figure 4, reveals that there is a weak to moderate positive correlation between `tenure_yrs_` and `units_completed`, suggesting that employees with more tenure may be slightly more productive, though the effect is not strong. Additionally, there is minimal correlation between `dist_from_work` and `units_completed`, indicating that commuting distance does not meaningfully explain variation in productivity. There is a strong positive correlation between `tenure_yrs`, `other_exp_yrs`, and `salary_annual`, which reflects that experience and salary are proportionate and likely increase in tandem over time.

Figure 4: Correlation Matrix

To explore individual associations further, a scatter plot of `tenure_yrs` vs. `units_completed` (Figure 5) reveals a slightly loose, upward trend, supporting its inclusion as a potential explanatory variable.

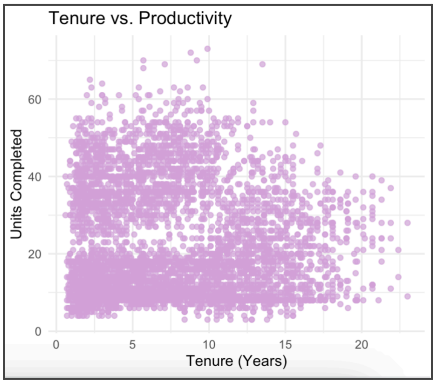

Figure 5: Scatter Plot - Tenure vs. Productivity

In contrast, the scatter plot of `dist_from_work` vs. `units_completed` (Figure 6) shows no discernible pattern, suggesting that this may contribute little explanatory value.

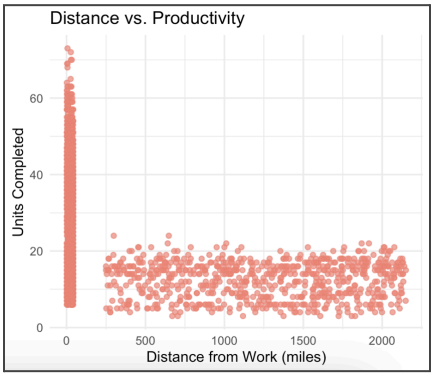

Figure 6: Scatter Plot - Distance vs. Productivity

Lastly, the histogram of `units_completed` segmented by work-from-home status (Figure 7) illustrates that employees who do not work from home (brown bars) tend to complete more units, while work-from-home employees (blue bars) are more clustered in lower productivity bins. This provides early visual evidence in support of the hypothesis.

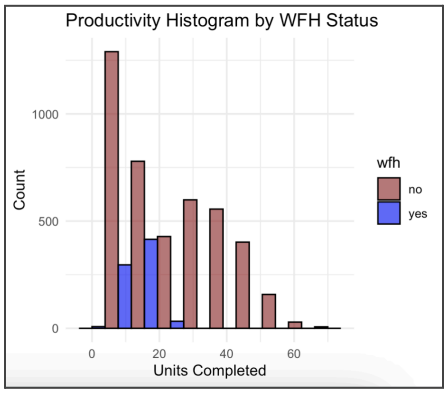

Figure 7: Histogram - Productivity by WFH Status

Phase 5: Choosing a Modelling Paradigm

Given the nature of the dependent variable (`units_completed`, continuous) and the visuals, a linear regression framework is justified. This allows for clear interpretability, stepwise expansion, and model evaluation using residuals and R-squared.

Initial model: `units_completed ~ wfh`

Subsequent models: tenure, experience, demographics, job level, and distance

Phase 6: Hypothesis Formulation #2

1. Null Hypothesis (H_0): The coefficient for `wfh` ($\beta_{wfh} = 0$)
    a. Working from home has no effect on productivity.
2. Alternative Hypothesis (H_1): The coefficient for `wfh` ($\beta_{wfh} \neq 0$)
    a. Working from home does affect productivity.

Phase 7: Modelling Journey

The modelling phase consists of five regression iterations, each designed to expand upon the previous model by introducing new covariates to test the strength of `wfh` as a predictor of productivity.

Model 1: Simple Linear Regression

- Formula: `units_completed ~ wfh`
- This base model evaluates the unadjusted relationship between employees working from home and productivity. The effect of `wfh` on `units_completed` is statistically significant with the p-value being far less than 0.05 (Pr(>|t|) = <2e-16). This further indicates that employees who work from home, on average, complete fewer units. However, the R-squared value is low (~0.090), meaning that this model explains only a small portion of the variance in productivity.

Model 2: Add Tenure and Experience

- Formula: `units_completed ~ wfh + tenure_yrs + other_exp_yrs`
- Incorporating these variables improves the model's explanatory power, as `tenure_yrs` and `other_exp_yrs` slightly increase the R-squared value (~0.093). This suggests that experience does play a fair role in shaping productivity. The `wfh` variable, in this case, remains significant.

Model 3: Add Demographics (Gender and Education)

- Formula: `units_completed ~ wfh + tenure_yrs + other_exp_yrs + gender + education`
- Adding the `gender` and `education` variables further adjusts the model. While the significance of these variables varies, some education categories contribute meaningfully. For example, employees with "comm college / some college" and "high school" education levels are associated with significantly higher units completed compared to employees with the "HS drop out" level. These categories have large positive coefficients (~13.76 and ~8.17, respectively) and are statistically significant (Pr(>|t|) = <2e-16). This suggests that intermediate education levels may point towards

higher productivity in this context. Additionally, this reflects heterogeneity in how education level may relate to productivity. The `wfh` variable remains significant.

Model 4: Add Job Level and Distance

- Formula: `units_completed ~ wfh + tenure_yrs + other_exp_yrs + gender + education + level + dist_from_work`
- This iteration improves the model again, primarily due to the level variable being included. The `dist_from_work` variable shows little added value—likely due to its extreme right skew—but it is included for completeness.
- The risk of multicollinearity increases slightly in this model. This is because “HS drop out” (`education`) and “senior team” (`level`) show up as ‘NA’ in the output, indicating that the model was unable to separate and calculate their effects. As a result, this can make some estimates less reliable.

Model 5: Full Model

- Formula: `units_completed ~ .`
- This model uses all the available predictors and gives the best R-squared score so far (~0.86), meaning it fits the data quite well. However, with so many predictors, overfitting becomes a concern.

Train/Test Evaluation:

To address overfitting concerns, the full model is evaluated on a 70/30 train-test split. Predictions made on the test set provide the following performance results:

- RMSE: 5.26
- R-squared: 0.86
- MAE: 4.08

As a result, these metrics confirm that the final model accurately predicts productivity using the `units_completed` variable, with most variation explained and low average error.

Phase 8: Commentary and Final Takeaways

This project set out to investigate whether working from home (`wfh`) is associated with lower productivity (`units_completed`). Across five regression models, the `wfh` variable consistently emerged as statistically significant and negatively associated with `units_completed`, even after testing for factors like tenure, experience, education, gender, job level, and distance from work.

Model 1 showed that `wfh` alone could explain a fair portion of the variation in productivity. As the models expanded to include more variables—especially education and job levels in Models 3 and 4— the R-squared improved progressively, culminating in a final model with an R-squared of approximately 0.86. This indicates that the predictor variables collectively explain 86% of the variation in productivity. The train/test evaluation confirmed that the model generalizes well, with relatively low prediction error.

More importantly, the inclusion of the `education` and `level` variables showed that intermediate education levels, such as high school and community college, and mid-level roles are associated with higher productivity. This suggests that productivity does not only depend on remote status, but also on the alignment of these other factors.

The results of this study reinforce that working from home is not universally detrimental to productivity, but it does show a considerable amount of average decline in output, particularly when not supported by structured roles or experience. These insights have practical implications: if organizations continue enforcing hybrid or remote work arrangements, they must consider job level and overall experience to ensure that productivity is not compromised.

Overall, the analysis supports the hypothesis that productivity has declined under work-from-home conditions. From personal experience in remote teams, these findings resonate quite strongly. Productivity often is a result of clearly defined roles and expectations, as well as access to support or training, more than a physical work environment. Therefore, these results further reinforce that productivity is shaped by a variety of factors, making work-from-home just one of them.