Linear Model Extensions

This chapter covers several techniques for expanding the linear regression framework covered in chapter "Linear Regression" to test hypotheses with more nuance and complexity.

Model Comparisons

Assuming it is warranted by the research objective, it is sometimes helpful to subset data and compare coefficients between models to determine how the strength of associations between predictors and the response compares between cohorts. This is a common approach in pay equity studies, as it clearly highlights differences in how a particular factor such as job level, job profile, or geography impacts compensation for male vs. female employees or across ethnic groups.

To illustrate, let us fit a multiple regression model to understand drivers of YTD sales for salespeople with overtime relative to those without overtime:

```
# Subset employees data frame; leads are only applicable for
↪  those in sales positions
data <- subset(employees, job_title %in% c('Sales Executive',
↪  'Sales Representative'))

# Partition data into overtime and non-overtime groups
data_ot <- subset(data, overtime == 'Yes')
data_nonot <- subset(data, overtime == 'No')

# Regress transformed YTD sales on a combination of predictors
↪  for overtime and non-overtime groups
```


C. Starbuck, *The Fundamentals of People Analytics*,
https://doi.org/10.1007/978-3-031-28674-2_11

```
mlm.fit.ot <- lm(sqrt(ytd_sales) ~ engagement + job_lvl +
↪  stock_opt_lvl + org_tenure, data_ot)
mlm.fit.nonot <- lm(sqrt(ytd_sales) ~ engagement + job_lvl +
↪  stock_opt_lvl + org_tenure, data_nonot)
```

```
##
## Call:
## lm(formula = sqrt(ytd_sales) ~ engagement + job_lvl + stock_opt_lvl +
##     org_tenure, data = data_ot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.927 -22.171  -1.383  19.740 106.769
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    121.815     14.767   8.249 3.27e-13 ***
## engagement      13.171      4.569   2.883  0.00472 **
## job_lvl         35.983      4.754   7.570 1.10e-11 ***
## stock_opt_lvl    7.139      3.342   2.136  0.03481 *
## org_tenure       5.369      0.722   7.437 2.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.78 on 113 degrees of freedom
## Multiple R-squared:  0.688,  Adjusted R-squared:  0.6769
## F-statistic: 62.29 on 4 and 113 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = sqrt(ytd_sales) ~ engagement + job_lvl + stock_opt_lvl +
##     org_tenure, data = data_nonot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -81.952 -19.422   0.136  20.813  96.003
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   132.3391     8.6695  15.265  < 2e-16 ***
## engagement      9.8523     2.4721   3.985 8.56e-05 ***
## job_lvl        33.1396     3.0014  11.042  < 2e-16 ***
## stock_opt_lvl   4.6377     2.1587   2.148   0.0325 *
## org_tenure      6.0435     0.4039  14.964  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.98 on 286 degrees of freedom
## Multiple R-squared:  0.7332, Adjusted R-squared:  0.7295
## F-statistic: 196.5 on 4 and 286 DF,  p-value: < 2.2e-16
```

Since we are comparing two models, we need not scale the variables since comparing a specific predictor's relationship with the response in the overtime model can be juxtaposed against the same predictor in the non-overtime model using the original units of measurement.

Based on the regression output, the model for salespeople who worked overtime explains more variance in square root transformed `ytd_sales` ($R^2 = 0.73$) relative to the model for salespeople without overtime ($R^2 = 0.69$).

We can see that `engagement` has a larger effect on the transformed response among salespeople who worked overtime ($\beta = 13.17$, $t(113) = 2.88$, $p < 0.01$) relative to those who worked no overtime ($\beta = 9.85$, $t(286) = 3.99$, $p < 0.001$). In addition, `job_lvl` has a stronger association with the response in the overtime group ($\beta = 35.98$, $t(113) = 7.57$, $p < 0.01$) relative to the non-overtime group ($\beta = 33.14$, $t(286) = 11.04$, $p < 0.001$). Given that the intercept (average square root of `ytd_sales` when the values of all predictors are set to 0) is higher for the non-overtime group ($\beta = 132.34$, $t(286) = 15.27$, $p < 0.001$) than for the overtime group ($\beta = 121.82$, $t(113) = 8.25$, $p < 0.001$), differences in the coefficients on `job_lvl` may indicate that one's job level is a proxy for skill and capacity to sell more in fewer hours.

Hierarchical Regression

A multiple model approach can also be useful for understanding the incremental value a given variable—or set of variables—provides above and beyond a set of control variables. **Hierarchical regression** is a method by which variables are added to the model in steps, and changes in model statistics are evaluated after each step. Let us use hierarchical regression to test the hypothesis below.

H1: Among salespeople who work overtime, engagement has a significant positive relationship with YTD sales after controlling for job level, stock option level, and organization tenure.

```
##
## Call:
## lm(formula = sqrt(ytd_sales) ~ job_lvl + stock_opt_lvl + org_tenure,
##     data = data_ot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.279 -23.803  -0.339  23.017  96.742
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   154.6969     9.6759  15.988  < 2e-16 ***
## job_lvl        37.5715     4.8707   7.714 5.02e-12 ***
## stock_opt_lvl   5.2397     3.3794   1.550    0.124
## org_tenure      5.4935     0.7434   7.389 2.64e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.81 on 114 degrees of freedom
## Multiple R-squared:  0.665,  Adjusted R-squared:  0.6562
## F-statistic: 75.44 on 3 and 114 DF,  p-value: < 2.2e-16

##
```

```
## Call:
## lm(formula = sqrt(ytd_sales) ~ engagement + job_lvl + stock_opt_lvl +
##     org_tenure, data = data_ot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.927 -22.171  -1.383  19.740 106.769
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    121.815     14.767   8.249 3.27e-13 ***
## engagement      13.171      4.569   2.883  0.00472 **
## job_lvl         35.983      4.754   7.570 1.10e-11 ***
## stock_opt_lvl    7.139      3.342   2.136  0.03481 *
## org_tenure       5.369      0.722   7.437 2.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.78 on 113 degrees of freedom
## Multiple R-squared:  0.688,  Adjusted R-squared:  0.6769
## F-statistic: 62.29 on 4 and 113 DF,  p-value: < 2.2e-16
```

Comparing the output from these two regression models, we can determine that the addition of `engagement` to the control set explains an additional 2% of the variance in YTD sales ($\Delta R^2 = 0.69 - 0.67 = 0.02$).

In addition, the controls-only model output shows that without `engagement` in the model, `stock_opt_lvl` is not significant. This is a good reminder that regression does not examine bivariate relationships of each predictor with the response *independent of other variables*; rather, the relationships among all variables in the model impact which predictors emerge as having a statistical association with the response.

Multilevel Models

The models covered thus far have focused only on observation-level effects. That is, there has been an inherent assumption that predictor variables have *fixed* effects on the outcome and these effects do not vary based on group(s) to which the observations belong. These models are sometimes referred to as **fixed effects** models.

It is often the case, however, that the strength and nature of predictors' effects on an outcome vary across categorical dimensions. For example, estimating the number of requisitions that can be filled by a Talent Acquisition team over a certain period may require inputs such as the number of recruiters and position backfill expectations based on attrition assumptions. However, the model should probably account for how these factors impact recruiter productivity at the intersections of group-level factors such as geography, job family, and job level as well. Estimates for recruiters who are focused on filling executive-level positions in geographies with a limited talent pool or fiercely competitive labor market will look quite different relative to recruiters focused on entry-level, low-skilled positions that are

location agnostic. Failure to incorporate these group-level effects may result in inaccurate estimates or incorrectly concluding that variables are not significant in explaining why recruiters vary in the number of requisitions they can fill.

You may wonder how this concept is different from simply including dummy-coded variables in the model to reflect the groups to which individual observations belong. The difference is that the average value of Y when all predictors are set to 0—namely the Y-intercept β_0—does not vary by group with dummy-coded categorical variables. In a multilevel model, the intercept is *random* rather than *fixed* for each group. Group-level effects can also be modeled for select (or all) X variables in addition to varying β_0 for each group.

Consider a linear model constructed to test hypothesized relationships of every X variable with an outcome Y. This is the equivalent of building G independent models, where G is the number of groups, using data subsetted for the respective group:

$$Y_G = \beta_{G0} + \beta_{G1}X_1 + \beta_{G2}X_2 + \ldots + \beta_{Gp}X_p + \epsilon$$

In this case, it is easy to consider wrapping the `lm()` function within a loop that iterates through each G group, filtering to each of the respective group's data in turn. However, if we hypothesize that the effects of only *certain* variables depend on the G group, we need to estimate both group-level *and* observation-level effects within the same model. A multilevel model featuring this mixture of fixed and random effects is known as a **mixed effects** model. This is also known as **Hierarchical Linear Modeling (HLM)**, which is materially different from Hierarchical Regression covered in the prior section, which compared nested regression models.

A model in which group-level effects are hypothesized for β_0 and X_1 and observation-level effects are hypothesized for all other predictors is expressed as:

$$Y_G = \beta_{G0} + \beta_{G1}X_1 + \beta_2X_2 + \ldots + \beta_pX_p + \epsilon$$

To fit a linear mixed effects model in R, we can leverage the `lmer()` function from the `lmerTest` package. Let us demonstrate how to fit a model to understand the random effects of `stock_opt_lvl` and fixed effects of `engagement`, `job_lvl`, and `org_tenure` on `sqrt(ytd_sales)`:

```
# Load library
library(lmerTest)

# Fit linear mixed model
lme.fit <- lmerTest::lmer(sqrt(ytd_sales) ~ engagement +
↪  job_lvl + (1 | stock_opt_lvl) + org_tenure, data_ot)

# Summarize model results
summary(lme.fit)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: sqrt(ytd_sales) ~ engagement + job_lvl + (1 | stock_opt_lvl) +
##     org_tenure
##    Data: data_ot
##
## REML criterion at convergence: 1141.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.52388 -0.63661  0.00411  0.61215  3.13684
##
## Random effects:
##  Groups        Name        Variance Std.Dev.
##  stock_opt_lvl (Intercept)   51.16   7.152
##  Residual                  1069.27  32.700
## Number of obs: 118, groups:  stock_opt_lvl, 4
##
## Fixed effects:
##             Estimate Std. Error       df t value Pr(>|t|)
## (Intercept) 133.6129    14.5677  88.8995   9.172 1.67e-14 ***
## engagement   12.0038     4.5140 113.9662   2.659  0.00896 **
## job_lvl      35.8950     4.7470 112.4054   7.562 1.17e-11 ***
## org_tenure    5.2542     0.7265 113.5918   7.232 5.94e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) enggmn jb_lvl
## engagement -0.727
## job_lvl    -0.445 -0.110
## org_tenure  0.046 -0.054 -0.463
```

The results of `lmer()` contain sections for both fixed and random effects. Consistent with the interpretation of linear regression model output, we can see that the fixed effects of each predictor are statistically significant. The key difference here is that the variance shown for the intercept of the random effects model is large. This indicates that there are meaningful differences in the relationships between predictors and `sqrt(ytd_sales)` across the levels of `stock_opt_lvl` that would be missed without a mixed model that accounts for these group-level effects.

For a more comprehensive treatment on multilevel models, see Gelman and Hill (2006).

Polynomial Regression

Linear regression is a powerful approach to understanding the relative strength of predictors' associations with a response variable. While linear models have advantages in interpretation, inference, and implementation simplicity, the linearity assumption often limits predictive power since this assumption is often a poor

approximation of actual relationships in the data. In this section, we will discuss how to extend the linear regression framework and relax linear model assumptions to handle non-linear relationships.

In a people analytics context, many data sets are cross-sectional and time-invariant, meaning they represent data collected at a single point in time. However, data collected across multiple points in time (time series data) are needed for forecasting future values of a dependent variable (e.g., a workforce planning model that estimates hires by month).

There is often a seasonality element inherent in the relationship between time and the outcome that is being estimated, which requires accounting for time-variant features (e.g., monthly attrition rate assumptions). **Seasonality** is the variation that occurs at regular intervals within a year. For example, companies with an annual bonus often experience a seasonal spike in voluntary attrition following bonus payouts (beginning in March for many organizations). Accounting for seasonality in models helps reduce error, but it requires estimating a more complex set of model coefficients relative to a more naive linear projection.

The simple linear regression equation, $Y = \beta_0 + \beta_1 X + \epsilon$, can be easily extended to include higher-order polynomial terms and achieve a more flexible fit. This is known as **polynomial regression**.

- Quadratic (2nd Order Polynomial) Regression Equation: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$
- Cubic (3rd Order Polynomial) Regression Equation: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

Figure 1 illustrates how higher-order polynomial functions can fit more curvilinear trends relative to a simple linear projection.

It is important to note that adding higher-order terms to the regression equation usually increases R^2 due to a more flexible fit to the data, but the additional coefficients are not necessarily significant. R^2 will approach 1 as the power of x approaches $n - 1$ since the fit line will connect every data point. However, a model that results in a perfect—or near perfect—fit is likely too flexible to generalize well to other data. This problem is known as overfitting and will be covered in chapter "Predictive Modeling". As a general rule, it is best not to add polynomial terms beyond the second or third orders to protect against overfitting the model.

Comparing the Adjusted R^2 for models with higher-order terms to one with only linear terms will help in determining whether higher-order polynomials add value to the model in explaining incremental variance in the response. Evaluating whether the coefficients on higher-order polynomials are statistically significant is important in determining *which variables* are contributing to any observed increases in Adjusted R^2.

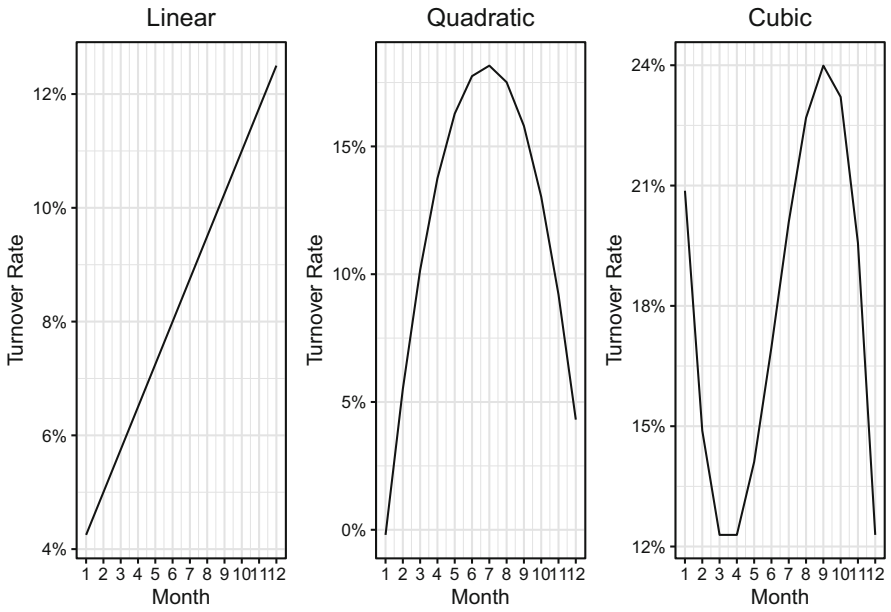

Fig. 1 Left: Linear turnover trend for $y = 0.75x + 3.5$. Middle: Quadratic turnover trend for $y = 7.3x - 0.53x^2 - 6.97$. Right: Cubic turnover trend for $y = -12.48x + 2.47x^2 - 0.13x^3 + 31.01$

Let us demonstrate how to fit a regression model with polynomial terms in R using the `turnover_trends` data set. First, we will subset this data frame to level 4 People Scientists who work remotely, based on the notion that turnover varies by `level` and `remote`, and then visualize the turnover trend to understand month-over-month variation across years.

As we can see in Fig. 2, the relationship between month and turnover rate is non-linear, and level 4 People Scientists who work remotely leave at lower rates relative to those who do not work remotely. There is a clear seasonal pattern that is consistent across all five years as well as remote vs. non-remote groups; namely, there is a spike in turnover between March and June as well as later in the year (November/December). Fitting a model to these data will require non-linear terms.

Adding polynomial terms requires an indicator variable `I()` in which the value of x is raised to the desired order (e.g., x^2 = `I(x^2)`). Let us start by fitting linear, quadratic, and cubic regression models (to compare performance) using only `month` as a predictor. Notice that the shape of the trends resemble the cubic function shown in Fig. 1 in that there are two discernible inflection points at which the trend reverses directions:

```
# Fit linear, quadratic, and cubic models to ps_turnover data
ps.lin.fit <- lm(turnover_rate ~ month, data = ps_turnover)
```

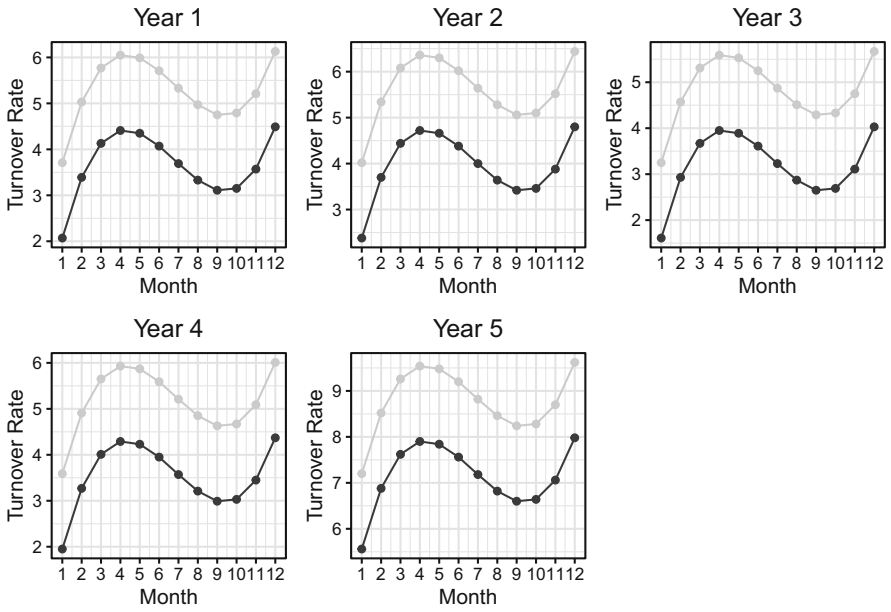

Fig. 2 Year 1–5 turnover trends for level 4 People Scientists, stratified by remote (dark grey line) vs. non-remote (light grey line)

```
ps.quad.fit <- lm(turnover_rate ~ month + I(month^2), data =
↪  ps_turnover)
ps.cube.fit <- lm(turnover_rate ~ month + I(month^2) +
↪  I(month^3), data = ps_turnover)
```

```
##
## Call:
## lm(formula = turnover_rate ~ month, data = ps_turnover)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2807 -1.3007 -0.3407  0.9218  4.5293
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.85067    0.35047   13.84   <2e-16 ***
## month        0.04000    0.04762    0.84    0.403
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.801 on 118 degrees of freedom
## Multiple R-squared:  0.005944,   Adjusted R-squared:  -0.00248
## F-statistic: 0.7056 on 1 and 118 DF,  p-value: 0.4026

##
## Call:
```

```
## lm(formula = turnover_rate ~ month + I(month^2), data = ps_turnover)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9140 -1.2790 -0.3990  0.9535  4.6560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.24400    0.58692   7.231 5.35e-11 ***
## month        0.30000    0.20758   1.445    0.151
## I(month^2)  -0.02000    0.01554  -1.287    0.201
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.796 on 117 degrees of freedom
## Multiple R-squared:  0.01981,    Adjusted R-squared:  0.003058
## F-statistic: 1.182 on 2 and 117 DF,  p-value: 0.3101

##
## Call:
## lm(formula = turnover_rate ~ month + I(month^2) + I(month^3),
##     data = ps_turnover)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.924 -1.464 -0.114  0.486  3.666
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.514000   0.873921   1.732   0.0859 .
## month        2.410000   0.558831   4.313 3.41e-05 ***
## I(month^2)  -0.410000   0.097879  -4.189 5.49e-05 ***
## I(month^3)   0.020000   0.004963   4.030   0.0001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.689 on 116 degrees of freedom
## Multiple R-squared:  0.1402, Adjusted R-squared:  0.1179
## F-statistic: 6.304 on 3 and 116 DF,  p-value: 0.0005334
```

The linear ($F(1, 118) = 0.71$, $p = 0.40$) and quadratic ($F(2, 117) = 1.18$, $p = 0.31$) models are not significant. However, as expected based on the shape of the turnover trend, the cubic model is significant ($F(3, 116) = 6.30$, $p < 0.001$) and the linear (`month`), quadratic (`I(month^2)`), and cubic (`I(month^3)`) terms all provide significant information in estimating turnover rates ($p < 0.001$).

While the cubic model achieved statistical significance at the $p < 0.001$ level, 86% of the variance in monthly turnover rates remains unexplained ($1-R^2 = 0.86$). To improve the performance of the model, our model needs to reflect the fact that turnover varies as a function of `year` and `remote` in addition to `month`.

As shown in Fig. 3, the multidimensional data vary widely around estimates produced by the two-dimensional models (i.e., `turnover_rate` predicted on the basis of `month`). While the cubic regression model reflects the seasonality in month-over-month turnover, there are notable differences between remote and non-remote turnover rates as well as differences across years.

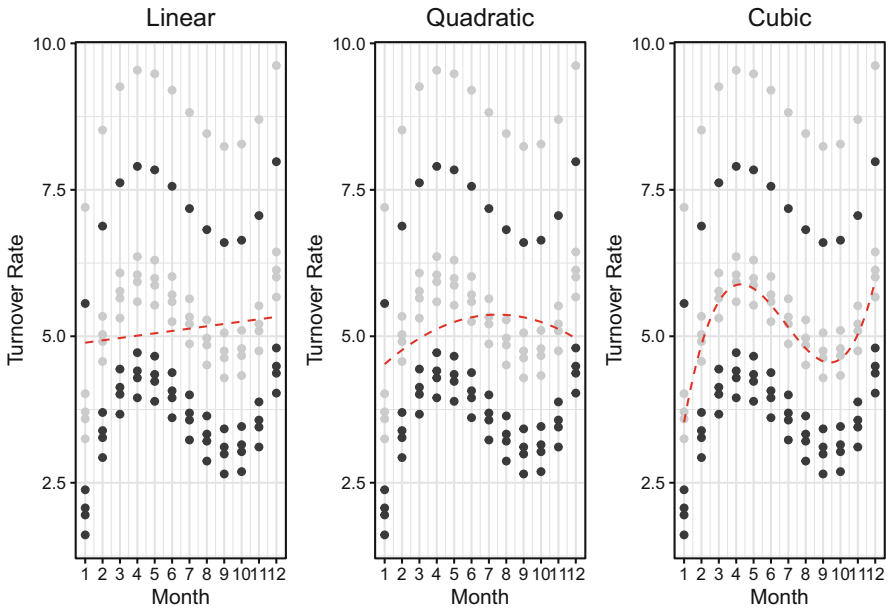

Fig. 3 Linear, quadratic, and cubic models fitted to turnover data (red dashed lines). Remote workers are represented in dark grey points, and non-remote workers in light grey points

Let us add `remote` to the cubic regression model to see how performance changes.

```
# Fit linear, quadratic, and cubic models to ps_turnover df
ps.cube.fit <- lm(turnover_rate ~ month + I(month^2) +
↪  I(month^3) + remote, data = ps_turnover)

# Produce model summary
summary(ps.cube.fit)
```

```
##
## Call:
## lm(formula = turnover_rate ~ month + I(month^2) + I(month^3) +
##     remote, data = ps_turnover)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.104 -0.764 -0.644 -0.334  2.846
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.334000   0.775051   3.011   0.0032 **
## month        2.410000   0.488069   4.938 2.70e-06 ***
## I(month^2)  -0.410000   0.085485  -4.796 4.89e-06 ***
## I(month^3)   0.020000   0.004335   4.614 1.03e-05 ***
```

```
## remoteYes   -1.640000   0.269344  -6.089 1.54e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.475 on 115 degrees of freedom
## Multiple R-squared:  0.3498, Adjusted R-squared:  0.3272
## F-statistic: 15.47 on 4 and 115 DF,  p-value: 3.758e-10
```

As shown in the regression output, accounting for remote status increases explained variance by 21% ($\Delta R^2 = 0.35 - 0.14$). In addition to the increase in explained variance, the coefficient on `remote` is statistically significant ($\beta = -1.64$, $t(115) = -6.09$, $p < 0.001$). On average, the turnover rate for remote People Scientists is 1.64% lower than the turnover rate for non-remote People Scientists.

Next, let us include `year` as a linear term in the model since turnover rates also vary along this dimension.

```
# Fit linear, quadratic, and cubic models to ps_turnover df
ps.cube.fit <- lm(turnover_rate ~ year + month + I(month^2) +
↪  I(month^3) + remote, data = ps_turnover)

# Produce model summary
summary(ps.cube.fit)
```

```
##
## Call:
## lm(formula = turnover_rate ~ year + month + I(month^2) + I(month^3) +
##     remote, data = ps_turnover)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.419 -1.104  0.321  0.666  1.536
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.36900    0.63650   0.580    0.563
## year         0.65500    0.07338   8.926 8.71e-15 ***
## month        2.41000    0.37609   6.408 3.43e-09 ***
## I(month^2)  -0.41000    0.06587  -6.224 8.28e-09 ***
## I(month^3)   0.02000    0.00334   5.988 2.53e-08 ***
## remoteYes   -1.64000    0.20755  -7.902 1.90e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.137 on 114 degrees of freedom
## Multiple R-squared:  0.6173, Adjusted R-squared:  0.6005
## F-statistic: 36.77 on 5 and 114 DF,  p-value: < 2.2e-16
```

Explained variance increases to 62% by adding `year` to the model. While the coefficient on `year` is statistically significant ($\beta = 0.66$, $t(114) = 8.93$, $p < 0.001$), the change in attrition by year is not linear. Visualizing the distribution of turnover rates by year will provide evidence that a linear year-over-year growth factor will result in some large residuals since it will not capture the more complex trend present in these data (Fig. 4).

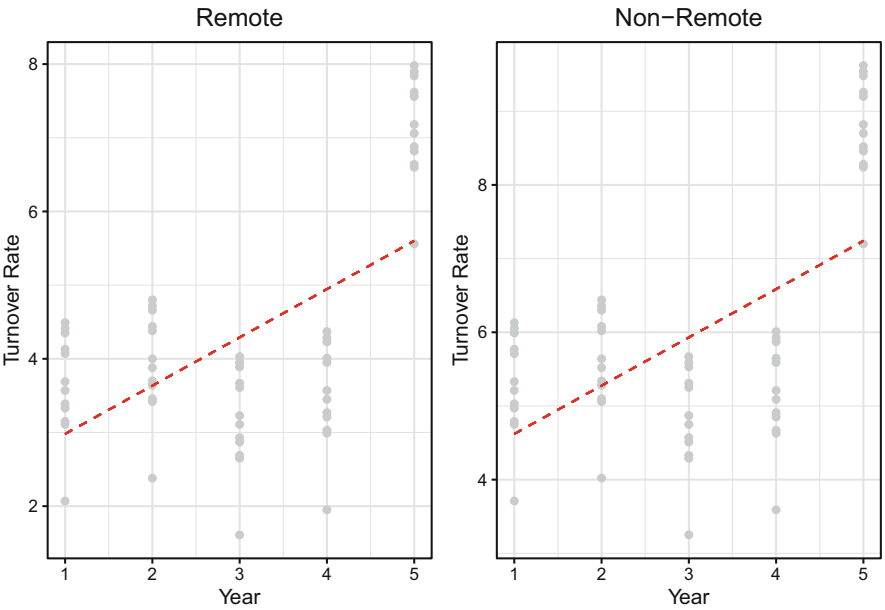

Fig. 4 Turnover rate distribution by year for remote (left) and non-remote (right) groups. Red dashed line reflects linear relationship between year and turnover rate, with y-intercept lowered 1.64% for remote group

Given the cubic nature of the change in turnover year-over-year, let us add quadratic and cubic terms for `year` to examine changes in model performance:

```
##
## Call:
## lm(formula = turnover_rate ~ year + I(year^2) + I(year^3) + month +
##     I(month^2) + I(month^3) + remote, data = ps_turnover)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0025714 -0.0004286 -0.0004286  0.0017143  0.0017143
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.866e+00  1.875e-03  -995.2   <2e-16 ***
## year         5.906e+00  2.179e-03  2710.7   <2e-16 ***
## I(year^2)   -2.712e+00  8.087e-04 -3353.4   <2e-16 ***
## I(year^3)    3.625e-01  8.929e-05  4060.0   <2e-16 ***
## month        2.410e+00  5.491e-04  4388.7   <2e-16 ***
## I(month^2)  -4.100e-01  9.618e-05 -4262.8   <2e-16 ***
## I(month^3)   2.000e-02  4.877e-06  4100.8   <2e-16 ***
## remoteYes   -1.640e+00  3.030e-04 -5411.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00166 on 112 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
```

```
## F-statistic: 1.996e+07 on 7 and 112 DF,  p-value: < 2.2e-16
```

The inclusion of higher-order polynomials on `year` results in a perfect fit to these data ($R^2 = 1$). Albeit a statistical improbability in practice, this indicates that the slope of the relationship between `month` and `turnover_rate` is perfectly consistent across years within remote and non-remote groups.

Our resulting equation for estimating `turnover_rate` on the basis of a combination of linear and non-linear values of `year`, `month`, and `remote` is defined by:

$$\hat{y} = -1.87 + 5.91\ year - 2.71\ year^2 + .36\ year^3 + 2.41\ month - .41\ month^2 + .02\ month^3 - 1.64\ remote + \epsilon$$

The performance of this model may initially seem like a cause for celebration, but the probability is low that this model would estimate future turnover with such a high degree of accuracy. While these data were generated with a goal to simplify illustrations and facilitate a working knowledge of polynomial regression mechanics, data which conform to such a constant pattern of seasonality across multiple years is a highly improbable situation in practice. As stated earlier in this chapter, a model that results in a perfect fit is likely too flexible to generalize well to other data, and methods of evaluating how well models are likely to perform on future data will be covered in chapter "Predictive Modeling".

Review Questions

1. What are some people analytics applications for comparing output from several regression models?
2. What modeling technique is appropriate for understanding an independent variable's contribution to a model's R^2 beyond a set of control variables?
3. In the context of Hierarchical Regression, what is the indicator that ΔR^2 is statistically significant when evaluating whether a particular independent variable provides meaningful information beyond a set of controls?
4. What are some examples of hypotheses that would warrant a linear mixed effects model over a general linear model?
5. What are the differences between Hierarchical Linear Modeling (HLM), which is also referred to as multilevel or mixed effects modeling, and Hierarchical Regression?
6. In what ways does polynomial regression differ from linear regression?
7. Why is it important to evaluate the nature of relationships at various levels of a categorical or time variable?
8. What shape characterizes a quadratic function?

9. If the coefficient on the cubic term is not statistically significant ($p >= 0.05$) in a cubic regression model, but the linear and quadratic terms are statistically significant ($p < 0.05$), what does this indicate about the model's fit to the data?
10. Why might adding higher-order polynomial terms to a model be problematic, even though the additional terms increase the model's R^2?