Regression

- Perhaps the most common goal in statistics is to answer the question: Is the variable X associated with a variable Y, and, if so, what is the relationship and can we use it to predict Y?
- Simple linear regression models the relationship between the magnitude of one variable and that of a second — for example, as X increases, Y also increases. Or as X increases, Y decreases
- Correlation is another way to measure how two variables are related
- The difference is that while correlation measures the strength of an association between two variables, regression quantifies the nature of the relationship.

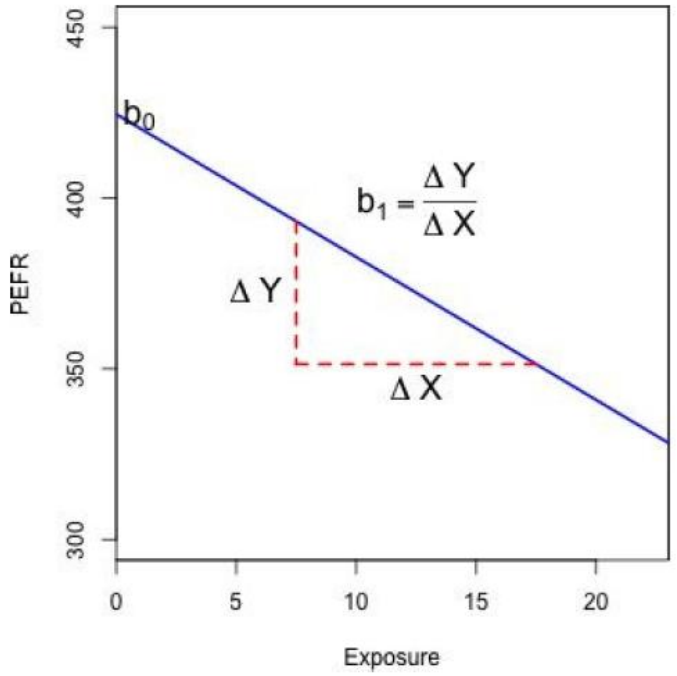
*This presentation is based on the book 'Practical Statistics for Data Scientist' by Peter Bruce & Andrew Bruce

Regression Equation

- Simple linear regression estimates exactly how much Y will change when X changes by a certain amount.

$$Y = b_0 + b_1 X$$

- The symbol b0 is known as the intercept (or constant), and the symbol b1 as the slope for X.
- The Y variable is known as the response or dependent variable since it depends on X.
- The *X* variable is known as the *predictor* or *independent* variable.



Fitted Values and Residuals

- Important concepts in regression analysis are the fitted values and residuals. In general, the data doesn't fall exactly on a line, so the regression equation should include an explicit error term

$$Y_i = b_0 + b_1 X_i + e_i$$

- The fitted values, also referred to as the predicted values, are typically denoted by (Y-hat). These are given by:

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$$

- We compute the residuals by subtracting the predicted values from the original data:

$$\hat{e}_i = Y_i - \hat{Y}_i$$

Least Squares

- In practice, the regression line is the estimate that minimizes the sum of squared residual values, also called the residual sum of squares or RSS:
- The method of minimizing the sum of the squared residuals is termed *least squares* regression, or *ordinary least squares* (OLS) regression
- Least squares, like the mean, are sensitive to outliers, although this tends to be a signicant problem only in small or moderate-sized problems
- A regression model that fits the data well is set up such that changes in X lead to changes in Y. However, by itself, the regression equation does not prove the direction of causation.
  - Conclusions about causation must come from a broader context of understanding about the relationship.

$$RSS = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

$$= \sum_{i=1}^{n} \left(Y_i - \hat{b}_0 - \hat{b}_1 X_i \right)^2$$

Multiple Linear Regression

- When there are multiple predictors, the equation is simply extended to accommodate them:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ \ldots + b_p X_p + e$$

```
house_lm

Call:
lm(formula = AdjSalePrice ~ SqFtTotLiving + SqFtLot + Bathrooms +
    Bedrooms + BldgGrade, data = house, na.action = na.omit)

Coefficients:
  (Intercept)  SqFtTotLiving        SqFtLot      Bathrooms
   -5.219e+05      2.288e+02     -6.051e-02     -1.944e+04
     Bedrooms      BldgGrade
   -4.778e+04      1.061e+05
```

The interpretation of the coefficients is as with simple linear regression: the predicted value $\hat{Y}$ changes by the coefficient b_j for each unit change in X_j assuming all the other variables, X_k for $k \neq j$, remain the same. For example, adding an extra finished square foot to a house increases the estimated value by roughly $229; adding 1,000 finished square feet implies the value will increase by $228,800.

Assessing the Model

- The most important performance metric from a data science perspective is *root mean squared error*, or *RMSE*. RMSE is the square root of the average squared error in the predicted values.

$$\hat{y}_i = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

- Similar to RMSE is the *residual standard error*, or *RSE*. In this case we have *p* predictors, and the RSE is given by:

$$RSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-p-1)}}$$

- The only difference is that the denominator is the degrees of freedom, as opposed to number of records (see “Degrees of Freedom”).
- In practice, for linear regression, the difference between RMSE and RSE is very small, particularly for big data applications.

- Another useful metric that you will see in software output is the coefficient of determination, also called the R-squared statistic.
- R-squared ranges from 0 to 1 and measures the proportion of variation in the data that is accounted for in the model.
- It is useful mainly in explanatory uses of regression where you want to assess how well the model fits the data. The denominator is proportional to the variance of *Y*.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

- The t-statistic — and its mirror image, the p-value — measures the extent to which a coefficient is “statistically significant” — that is, outside the range of what a random chance arrangement of predictor and target variable might produce.
- The higher the t-statistic (and the lower the p-value), the more significant the predictor.
- Data scientists primarily focus on the t-statistic as a useful guide for whether to include a predictor in a model or not.
- High t statistics (which go with p-values near 0) indicate a predictor should be retained in a model, while very low t-statistics indicate a predictor could be dropped.

```
summary(house_lm)

Call:
lm(formula = AdjSalePrice ~ SqFtTotLiving + SqFtLot + Bathrooms +
    Bedrooms + BldgGrade, data = house, na.action = na.omit)

Residuals:
     Min       1Q   Median       3Q      Max
-1199508  -118879   -20982    87414  9472982

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -5.219e+05  1.565e+04 -33.349  < 2e-16 ***
SqFtTotLiving  2.288e+02  3.898e+00  58.699  < 2e-16 ***
SqFtLot       -6.051e-02  6.118e-02  -0.989    0.323
Bathrooms     -1.944e+04  3.625e+03  -5.362 8.32e-08 ***
Bedrooms      -4.778e+04  2.489e+03 -19.194  < 2e-16 ***
BldgGrade      1.061e+05  2.396e+03  44.287  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 261200 on 22683 degrees of freedom
Multiple R-squared:  0.5407,    Adjusted R-squared:  0.5406
F-statistic:  5340 on 5 and 22683 DF,  p-value: < 2.2e-16
```

Cross-Validation

- Intuitively, you can see that it would make a lot of sense to set aside some of the original data, not use it to fit the model, and then apply the model to the set-aside (holdout) data to see how well it does.
- Using a holdout sample, though, leaves you subject to some uncertainty that arises simply from variability in the small holdout sample. How different would the assessment be if you selected a different holdout sample?
- Cross-validation extends the idea of a holdout sample to multiple sequential holdout samples. The algorithm for basic *k-fold cross-validation* is as follows:

1. Set aside 1/k of the data as a holdout sample.
2. Train the model on the remaining data.
3. Apply (score) the model to the 1/k holdout, and record needed model assessment metrics.
4. Restore the first 1/k of the data, and set aside the next 1/k (excluding any records that got picked the first time).
5. Repeat steps 2 and 3.
6. Repeat until each record has been used in the holdout portion.
7. Average or otherwise combine the model assessment metrics.

Model Selection and Stepwise Regression

- Adding more variables, however, does not necessarily mean we have a better model. Statisticians use the principle of ***Occam's razor*** to guide the choice of a model: all things being equal, a simpler model should be used in preference to a more complicated model.
- Including additional variables always reduces RMSE and increases . Hence, these are not appropriate to help guide the model choice.
- In the 1970s, Hirotugu Akaike, the eminent Japanese statistician, developed a metric called AIC (Akaike's Information Criteria) that penalizes adding terms to a model.
- How do we find the model that minimizes AIC? One approach is to search through all possible models, called all subset regression. This is computationally expensive and is not feasible for problems with large data and many variables.
- An attractive alternative is to use stepwise regression, which successively adds and drops predictors to find a model that lowers AIC.

- Simpler yet are forward selection and backward selection.
- In forward selection, you start with no predictors and add them one-by-one, at each step adding the predictor that has the largest contribution to R-squared, stopping when the contribution is no longer statistically significant.
- In backward selection, or backward elimination, you start with the full model and take away predictors that are not statistically significant until you are left with a model in which all predictors are statistically significant.
- ***Penalized regression*** is similar in spirit to AIC. Instead of explicitly searching through a discrete set of models, the model-fitting equation incorporates a constraint that penalizes the model for too many variables (parameters).
- Common penalized regression methods are ridge regression and lasso regression.
- Stepwise regression and all subset regression are in-sample methods to assess and tune models. This means the model selection is possibly subject to overfitting and may not perform as well when applied to new data.
- One common approach to avoid this is to use cross-validation to validate the models. In linear regression, overfitting is typically not a major issue, due to the simple (linear) global structure imposed on the data.

The Dangers of Extrapolation

- Regression models should not be used to extrapolate beyond the range of the data.
- The model is valid only for predictor values for which the data has sufficient values.

- As an extreme case, suppose model_lm is used to predict the value of a 5,000-square-foot empty lot.
- In such a case, all the predictors related to the building would have a value of 0 and the regression equation would yield an absurd prediction of –521,900 + 5,000 × –.0605 = –$522,202.
- Why did this happen? The data contains only parcels with buildings — there are no records corresponding to vacant land.
- Consequently, the model has no information to tell it how to predict the sales price for vacant land.

Factor Variables in Regression

- *Factor* variables, also termed *categorical* variables, take on a limited number of discrete values. For example, a loan purpose can be “debt consolidation,” “wedding,” “car,” and so on.
- The binary (yes/no) variable, also called an *indicator* variable, is a special case of a factor variable.
- Regression requires numerical inputs, so factor variables need to be recoded to use in the model.
- The most common approach is to convert a variable into a set of binary *dummy* variables.

```
prop_type_dummies <- model.matrix(~PropertyType -1, data=house)
head(prop_type_dummies)
  PropertyTypeMultiplex PropertyTypeSingle Family PropertyTypeTownhouse
1                     1                         0                     0
2                     0                         1                     0
3                     0                         1                     0
4                     0                         1                     0
5                     0                         1                     0
6                     0                         0                     1
```

- In the machine learning community, this representation is referred to as one hot encoding

- In the regression setting, a factor variable with P distinct levels is usually represented by a matrix with only P – 1 columns. This is because a regression model typically includes an intercept term.
- With an intercept, once you have defined the values for P – 1 binaries, the value for the Pth is known and could be considered redundant.
- Adding the Pth column will cause a multicollinearity error

```
Call:
lm(formula = AdjSalePrice ~ SqFtTotLiving + SqFtLot + Bathrooms +
    Bedrooms + BldgGrade + PropertyType, data = house)

Coefficients:
              (Intercept)             SqFtTotLiving
                -4.469e+05                 2.234e+02
                   SqFtLot                Bathrooms
                -7.041e-02                -1.597e+04
                  Bedrooms                BldgGrade

                -5.090e+04                 1.094e+05
PropertyTypeSingle Family      PropertyTypeTownhouse
                -8.469e+04                -1.151e+05
```

- There is no coefficient of Multiplex since it is implicitly defined when PropertyTypeSingle Family == 0 and PropertyTypeTownhouse == 0.
- The coefficients are interpreted as relative to Multiplex, so a home that is Single Family is worth almost $85,000 less, and a home that is Townhouse is worth over $150,000 less

Factor Variables with Many Levels

- Some factor variables can produce a huge number of binary dummies — zip codes are a factor variable and there are 43,000 zip codes in the US.
- In such cases, it is useful to explore the data, and the relationships between predictor variables and the outcome, to determine whether useful information is contained in the categories.

```
table(house$ZipCode)

 9800 89118 98001 98002 98003 98004 98005 98006 98007 98008 98010 98011
    1     1   358   180   241   293   133   460   112   291    56   163
98014 98019 98022 98023 98024 98027 98028 98029 98030 98031 98032 98033
   85   242   188   455    31   366   252   475   263   308   121   517
98034 98038 98039 98040 98042 98043 98045 98047 98050 98051 98052 98053
  575   788    47   244   641     1   222    48     7    32   614   499
98055 98056 98057 98058 98059 98065 98068 98070 98072 98074 98075 98077
  332   402     4   420   513   430     1    89   245   502   388   204
98092 98102 98103 98105 98106 98107 98108 98109 98112 98113 98115 98116
  289   106   671   313   361   296   155   149   357     1   620   364
98117 98118 98119 98122 98125 98126 98133 98136 98144 98146 98148 98155
  619   492   260   380   409   473   465   310   332   287    40   358
98166 98168 98177 98178 98188 98198 98199 98224 98288 98354
  193   332   216   266   101   225   393     3     4     9
```

- In some problems, you can consolidate a zip code using the first two or three digits, corresponding to a submetropolitan geographic region. For King County, almost all of the sales occur in 980xx or 981xx, so this doesn't help.

- An alternative approach is to group the zip codes according to another variable, such as sale price.

- Even better is to form zip code groups using the residuals from an initial model.
- The following dplyr code consolidates the 82 zip codes into five groups based on the median of the residual from the house_lm regression:

```
zip_groups <- house %>%
  mutate(resid = residuals(house_lm)) %>%
  group_by(ZipCode) %>%
  summarize(med_resid = median(resid),
            cnt = n()) %>%
  arrange(med_resid) %>%
  mutate(cum_cnt = cumsum(cnt),
         ZipGroup = ntile(cum_cnt, 5))
house <- house %>%
  left_join(select(zip_groups, ZipCode, ZipGroup), by='ZipCode')
```

Ordered Factor Variables

- Ordered factor variables can typically be converted to numerical values and used as is. For example, the variable BldgGrade is an ordered factor variable.

- Treating ordered factors as a numeric variable preserves the information contained in the ordering that would be lost if it were converted to a factor.

- ***Caution***: see that the values are not squenced equally, like 1,2,3,4,5. The distance between the levels are not assumed to be equal in this case.

Table 4-1. A typical data format

Value	Description
1	Cabin
2	Substandard
5	Fair
10	Very good
12	Luxury
13	Mansion

Correlated Predictors

- In multiple regression, the predictor variables are often correlated with each other.

```
step_lm$coefficients
             (Intercept)             SqFtTotLiving
            6.227632e+06              1.865012e+02
                Bathrooms                  Bedrooms
            4.472172e+04             -4.980718e+04
                BldgGrade PropertyTypeSingle Family
            1.391792e+05              2.332869e+04
  PropertyTypeTownhouse          SqFtFinBasement
            9.221625e+04              9.039911e+00
                  YrBuilt
           -3.592468e+03
```

- The coefficient for Bedrooms is negative! This implies that adding a bedroom to a house will reduce its value. How can this be? This is because the predictor variables are correlated: larger houses tend to have more bedrooms, and it is the size that drives house value, not the number of bedrooms. Consider two homes of the exact same size: it is reasonable to expect that a home with more, but smaller, bedrooms would be considered less desirable.
- Having correlated predictors can make it difficult to interpret the sign and value of regression coefficients (and can inflate the standard error of the estimates). The variables for bedrooms, house size, and number of bathrooms are all correlated.

Multicollinearity

- An extreme case of correlated variables produces multicollinearity — a condition in which there is redundance among the predictor variables.
- Perfect multicollinearity occurs when one predictor variable can be expressed as a linear combination of others.
- Multicollinearity occurs when:
  - A variable is included multiple times by error.
  - P dummies, instead of P – 1 dummies, are created from a factor variable (see “Factor Variables in Regression”).
  - Two variables are nearly perfectly correlated with one another
- Multicollinearity in regression must be addressed — variables should be removed until the multicollinearity is gone. A regression does not have a well-defined solution in the presence of perfect multicollinearity.
- Many software packages, including R, automatically handle certain types of multicolliearity. For example, if SqFtTotLiving is included twice in the regression of the house data, the results are the same as for the house_lm model.

Interactions and Main Effects

- An implicit assumption when only main effects are used in a model is that the relationship between a predictor variable and the response is independent of the other predictor variables. This is often not the case.
- You include interactions between variables in R using the * operator.

- For a home in the lowest ZipGroup, the slope is the same as the slope for the main effect SqFtTotLiving, which is $177 per square foot
- For a home in the highest ZipGroup, the slope is the sum of the main effect plus SqFtTotLiving:ZipGroup5, or $177 + $230 = $447 per square foot.

```
lm(AdjSalePrice ~  SqFtTotLiving*ZipGroup + SqFtLot +
     Bathrooms + Bedrooms + BldgGrade + PropertyType,
   data=house, na.action=na.omit)

Coefficients:
             (Intercept)             SqFtTotLiving
               -4.919e+05                 1.176e+02
                ZipGroup2                 ZipGroup3
               -1.342e+04                 2.254e+04
                ZipGroup4                 ZipGroup5
                1.776e+04                -1.555e+05
                  SqFtLot                 Bathrooms
                7.176e-01                -5.130e+03
                 Bedrooms                 BldgGrade
               -4.181e+04                 1.053e+05
PropertyTypeSingle Family     PropertyTypeTownhouse
                1.603e+04                -5.629e+04
  SqFtTotLiving:ZipGroup2   SqFtTotLiving:ZipGroup3
                3.165e+01                 3.893e+01
  SqFtTotLiving:ZipGroup4   SqFtTotLiving:ZipGroup5
                7.051e+01                 2.298e+02
```

Outliers

- Generally speaking, an extreme value, also called an outlier, is one that is distant from most of the other observations.
- Just as outliers need to be handled for estimates of location and variability, outliers can cause problems with regression models. In regression, an outlier is a record whose actual y value is distant from the predicted value.
- In regression, the standardized residual is the metric that is typically used to determine whether a record is classified as an outlier.

```
house_98105[idx[1], c('AdjSalePrice', 'SqFtTotLiving', 'SqFtLot',
                      'Bathrooms', 'Bedrooms', 'BldgGrade')]

AdjSalePrice SqFtTotLiving SqFtLot Bathrooms Bedrooms BldgGrade
         (dbl)         (int)   (int)     (dbl)    (int)     (int)
1       119748          2900    7276         3        6         7
```

- In this case, the outlier corresonds to a sale that is anomalous and should not be included in the regression. Outliers could also be the result of other problems, such as a “fat-finger” data entry or a mismatch of units

Influential Values

- A value whose absence would significantly change the regression equation is termed an influential observation.
- This data value is considered to have high leverage on the regression.

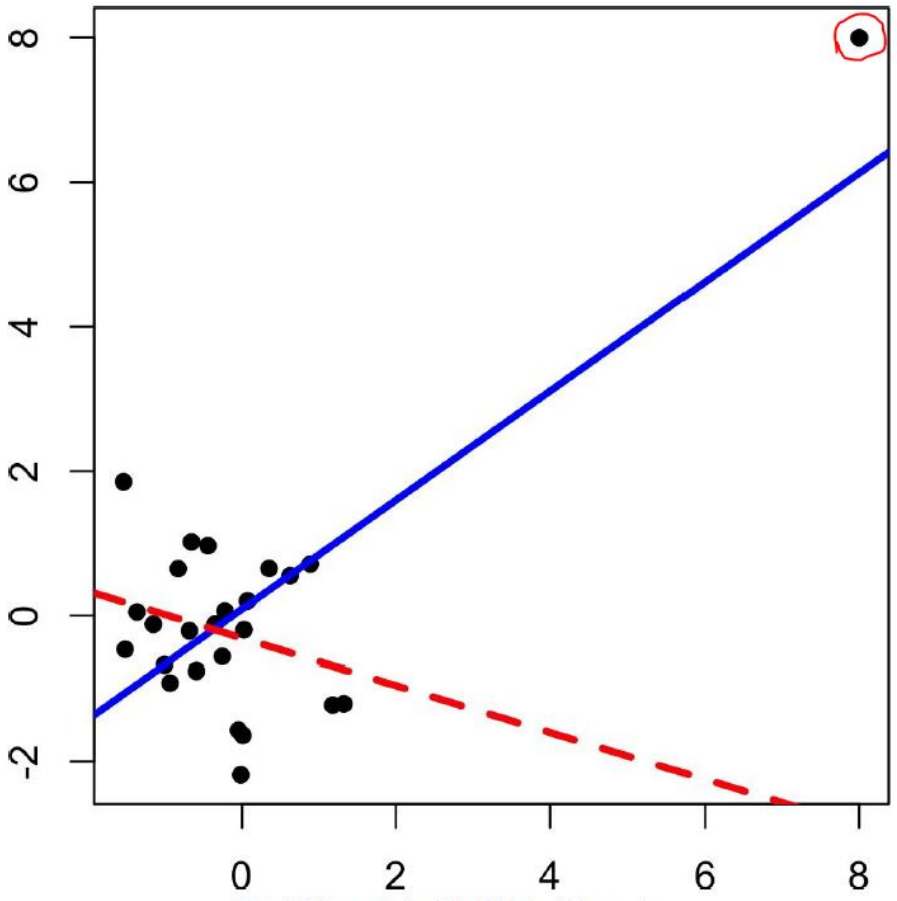
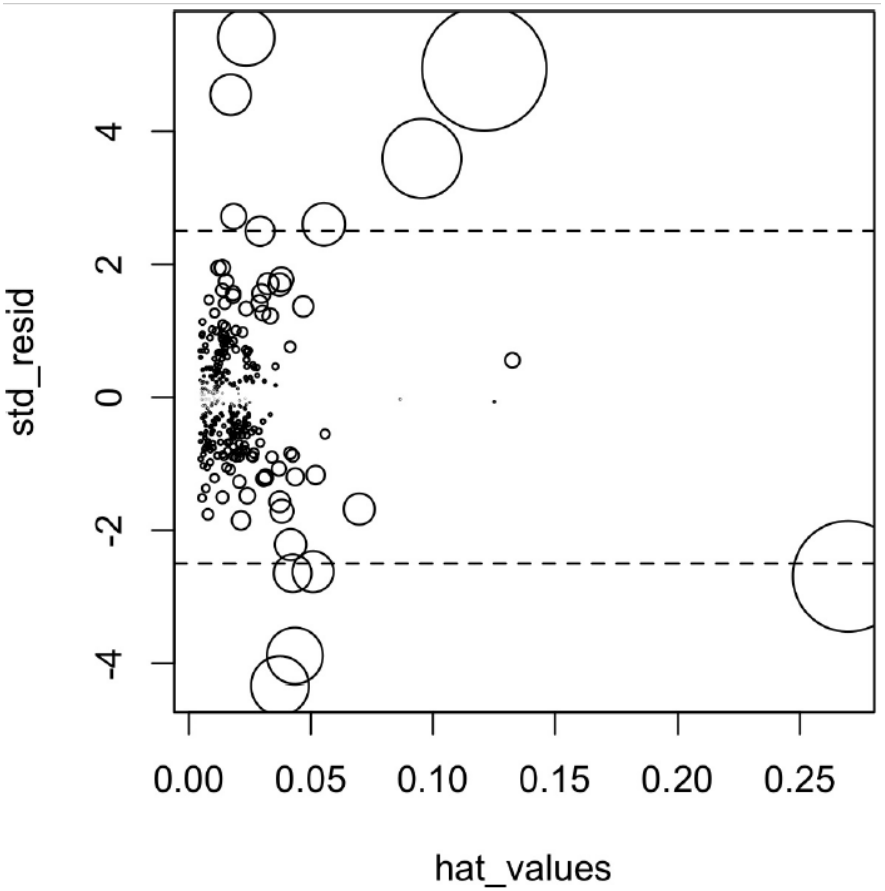


Figure 4-5. An example of an influential data point in regression

- In addition to standardized residuals (see “Outliers”), statisticians have developed several metrics to determine the influence of a single record on a regression. A common measure of leverage is the hat-value.
- Another metric is Cook’s distance, which defines influence as a combination of leverage and residual size.
- An influence plot or bubble plot combines standardized residuals, the hat-value, and Cook’s distance in a single plot.



- Table compares the regression with the full data set and with highly influential data points removed.
- The regression coefficient for Bathrooms changes quite dramatically.

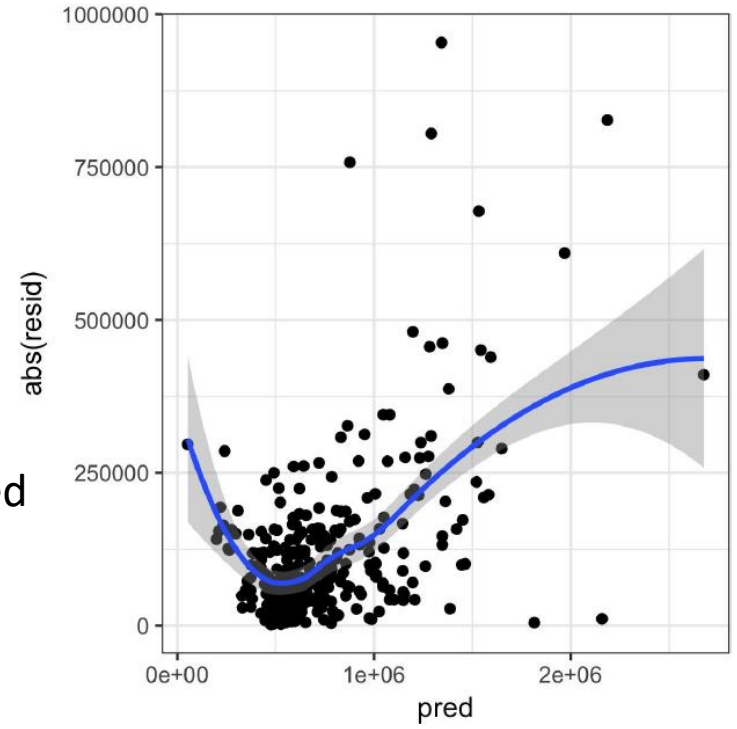
- For purposes of fitting a regression that reliably predicts future data, identifying influential observations is only useful in smaller data sets.
- For regressions involving many records, it is unlikely that any one observation will carry sufficient weight to cause extreme influence on the fitted equation (although the regression may still have big outliers).
- For purposes of anomaly detection, though, identifying influential observations can be very useful.T

	Original	Influential removed
(Intercept)	–772550	–647137
SqFtTotLiving	210	230
SqFtLot	39	33
Bathrooms	2282	–16132
Bedrooms	–26320	–22888
BldgGrade	130000	114871

Heteroskedasticity, Non-Normality

- In most problems, data scientists do not need to be too concerned with the distribution of the residuals.
- One area where this may be of concern to data scientists is the standard calculation of confidence intervals for predicted values, which are based upon the assumptions about the residuals.

- Heteroskedasticity is the lack of constant residual variance across the range of the predicted values.
- In other words, errors are greater for some portions of the range than for others.
- Heteroskedasticity indicates that prediction errors differ for different ranges of the predicted value, and may suggest an incomplete model.



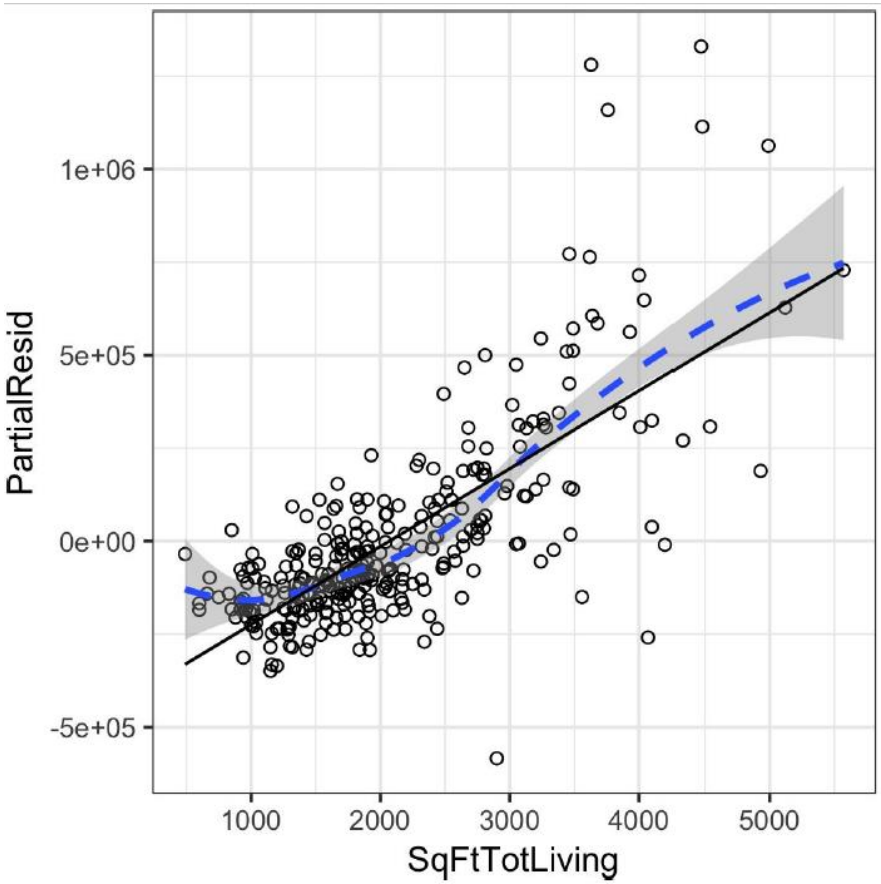
Partial Residual Plots and Nonlinearity

- Partial residual plots are a way to visualize how well the estimated fit explains the relationship between a predictor and the outcome.
- Along with detection of outliers, this is probably the most important diagnostic for data scientists.
- The basic idea of a partial residual plot is to isolate the relationship between a predictor variable and the response, taking into account all of the other predictor variables.
- A partial residual might be thought of as a “synthetic outcome” value, combining the prediction based on a single predictor with the actual residual from the full regression equation.

$$\text{Partial residual} = \text{Residual} + \hat{b}_i X_i$$

```
df <- data.frame(SqFtTotLiving = house_98105[, 'SqFtTotLiving'],
                 Terms = terms[, 'SqFtTotLiving'],
                 PartialResid = partial_resid[, 'SqFtTotLiving'])
ggplot(df, aes(SqFtTotLiving, PartialResid)) +
  geom_point(shape=1) + scale_shape(solid = FALSE) +
  geom_smooth(linetype=2) +
  geom_line(aes(SqFtTotLiving, Terms))
```

- The partial residual is an estimate of the contribution that SqFtTotLiving adds to the sales price.
- The relationship between SqFtTotLiving and the sales price is evidently nonlinear.
- The regression line underestimates the sales price for homes less than 1,000 square feet and overestimates the price for homes between 2,000 and 3,000 square feet.



- This nonlinearity makes sense in this case: adding 500 feet in a small home makes a much bigger difference than adding 500 feet in a large home.
- This suggests that, instead of a simple linear term for SqFtTotLiving, a nonlinear term should be considered

Polynomial and Spline Regression

- The relationship between the response and a predictor variable is not necessarily linear.
- When statisticians talk about nonlinear regression, they are referring to models that can’t be fit using least squares.
- What kind of models are nonlinear? Essentially all models where the response cannot be expressed as a linear combination of the predictors or some transform of the predictors.
- Nonlinear regression models are harder and computationally more intensive to fit, since they require numerical optimization. For this reason, it is generally preferred to use a linear model if possible.

Polynomial

- Polynomial regression involves including polynomial terms to a regression equation.

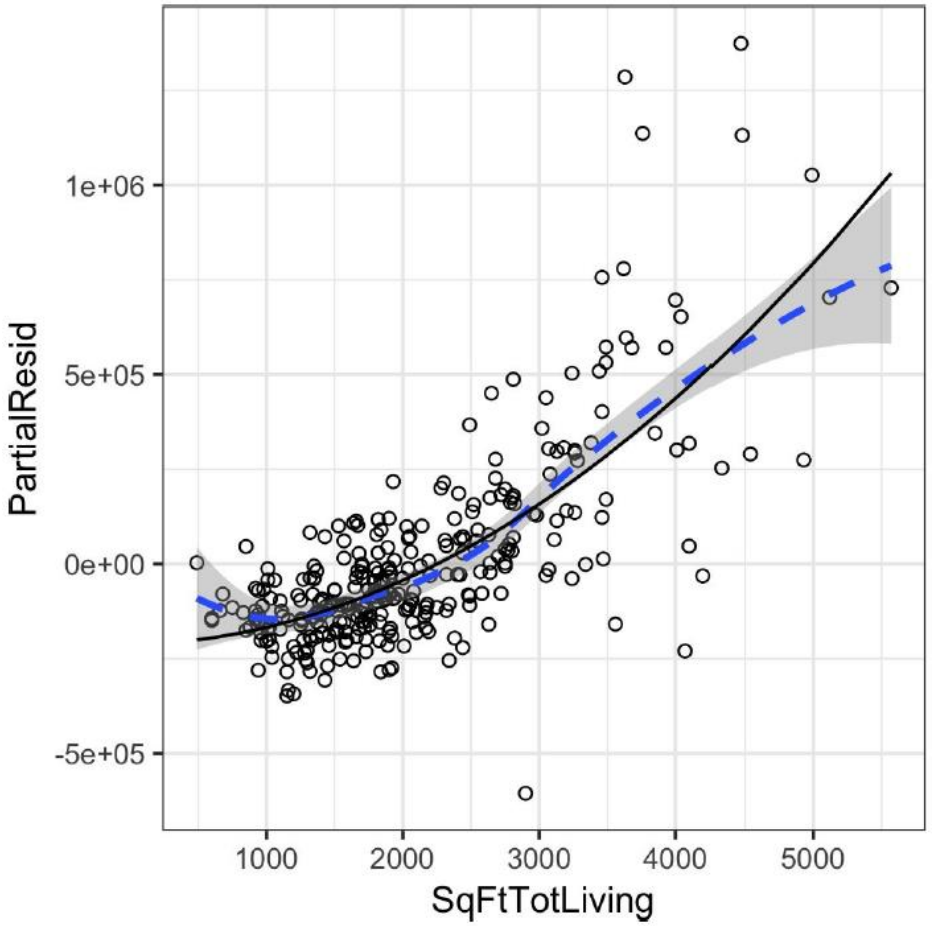
$$Y = b_0 + b_1 X + b_2 X^2 + e$$

```
lm(AdjSalePrice ~  poly(SqFtTotLiving, 2) + SqFtLot +
                BldgGrade +  Bathrooms +  Bedrooms,
                    data=house_98105)

Call:
lm(formula = AdjSalePrice ~ poly(SqFtTotLiving, 2) + SqFtLot +
    BldgGrade + Bathrooms + Bedrooms, data = house_98105)

Coefficients:
            (Intercept)  poly(SqFtTotLiving, 2)1
              -402530.47                3271519.49
poly(SqFtTotLiving, 2)2                   SqFtLot
               776934.02                     32.56
              BldgGrade                 Bathrooms
               135717.06                  -1435.12
               Bedrooms
                -9191.94
```

- The partial residual plot (see “Partial Residual Plots and Nonlinearity”) indicates some curvature in the regression equation associated with SqFtTotLiving.
- The fitted line more closely matches the smooth (see “Splines”) of the partial residuals as compared to a linear fit

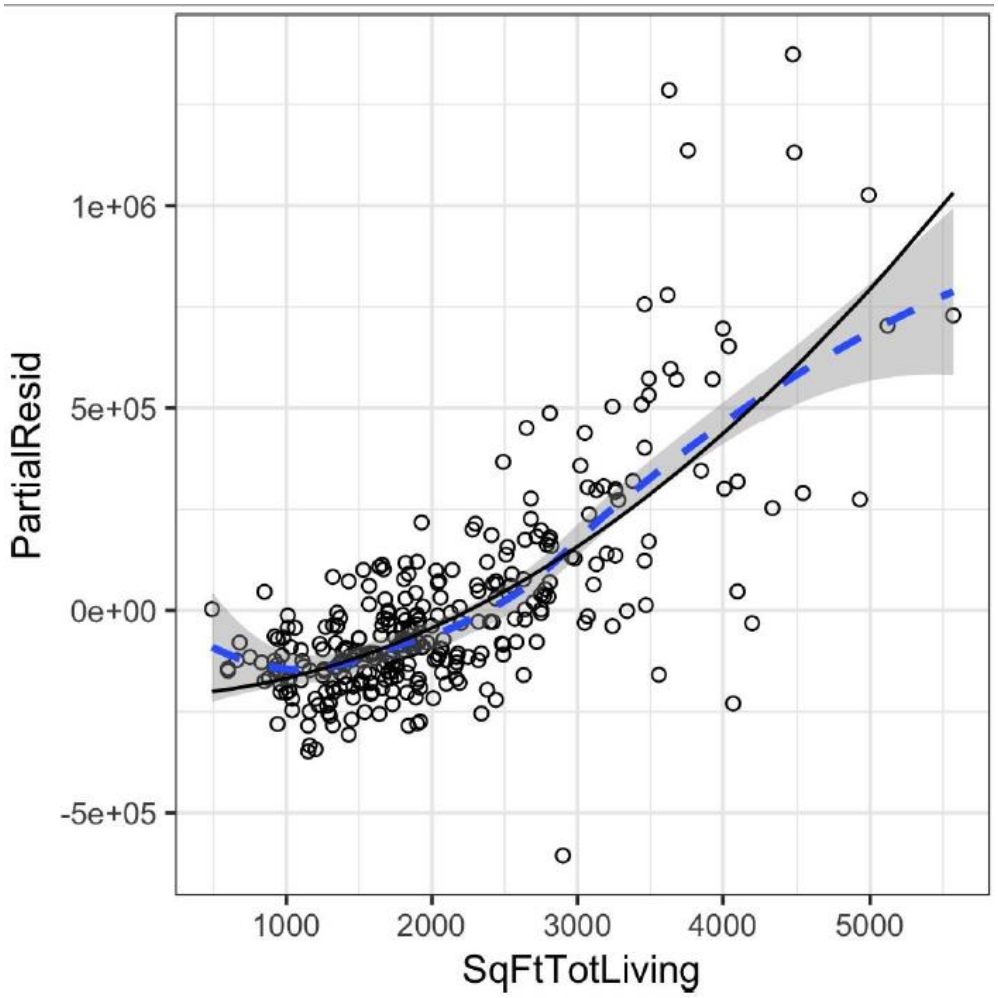


Splines

- Polynomial regression only captures a certain amount of curvature in a nonlinear relationship.
- Adding in higher-order terms, such as a cubic quartic polynomial, often leads to undesirable “wiggliness” in the regression equation.
- An alternative, and often superior, approach to modeling nonlinear relationships is to use splines.
- Splines provide a way to smoothly interpolate between fixed points.
- The technical definition of a spline is a series of piecewise continuous polynomials.
- The polynomial pieces are smoothly connected at a series of fixed points in a predictor variable, referred to as knots.
- Two parameters need to be specified: the degree of the polynomial and the location of the knots.

```
library(splines)
knots <- quantile(house_98105$SqFtTotLiving, p=c(.25, .5, .75))
lm_spline <- lm(AdjSalePrice ~ bs(SqFtTotLiving, knots=knots, degree=3) +
  SqFtLot + Bathrooms + Bedrooms + BldgGrade,  data=house_98105)
```

- In contrast to a linear term, for which the coefficient has a direct meaning, the coefficients for a spline term are not interpretable.
- Instead, it is more useful to use the visual display to reveal the nature of the spline fit.



Generalized Additive Models

- Suppose you suspect a nonlinear relationship between the response and a predictor variable, either by a priori knowledge or by examining the regression diagnostics.
- Polynomial terms may not flexible enough to capture the relationship, and spline terms require specifying the knots.
- Generalized additive models, or GAM, are a technique to automatically fit a spline regression.