Regression Models - Notes

Tanner Prestegard

Course taken from 6/1/2015 - 6/28/2015

Introduction

- Questions we may want to answer for this class (considering an example of children's heights and their parents' heights):
  - Use the parents' heights to predict children's heights.
  - To try to find an easily described mean relationship between parent and children heights.
  - To investigate the variation in children's heights that appears unrelated to parents' heights (residual variation).
  - To quantify what impact genotype information has beyond parental height in explaining child height.
  - To figure out how and what assumptions are needed to generalize findings beyond the data in question.
  - Why do children of very tall parents tend to be tall, but a little shorter than their parents, and why do children of very short parents tend to be short, but a little taller than their parents? (regression to the mean)

Background and notation

- Define the empirical (sample) mean as: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- If we subtract the mean from all data points, we are **centering** the random variables: $\tilde{X}_i = X_i - \bar{X}$
  - The mean of $\tilde{X}_i$ is zero.
- The empirical (sample) variance is: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$
  - The data defined by X_i/S have empirical standard deviation 1. This is called "scaling" the data.
- Normalization: centering and scaling the data to have empirical mean 0 and empirical standard deviation 1.
  - $Z_i = \frac{X_i - \bar{X}}{S}$
  - Normalized data have units equal to standard deviations of the original data.
- Empirical covariance
  - Consider a pair of data (X_i, Y_i).
  - Their covariance is $Cov\left(X,Y\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_iY_i - n\bar{X}\bar{Y}\right)$
  - Their correlation is defined as $Cor\left(X,Y\right) = \frac{Cov(X,Y)}{S_xS_y}$, where S_i is the estimate of the standard deviation for variable i.

- $Cor\left(X,Y\right) = Cor\left(Y,X\right)$
- $-1 \leq Cor\left(X,Y\right) \leq 1$
- $Cor\left(X,Y\right) = 1$ and $Cor\left(X,Y\right) = -1$ only when the observations fall perfectly on a positively or negatively sloped line, respectively.
- $Cor\left(X,Y\right)$ measures the strength of the linear relationship between the X and Y data, with stronger relationships as $Cor\left(X,Y\right)$ goes to -1 or +1.
- $Cor\left(X,Y\right)$ implies no linear relationship.

Basic least squares

- Example using children's and parents' heights.
- Consider only the children's heights - how could one describe the "middle"?
  - One definition: let Y_i be the height of child i, for $i = 1,..,n$, then define the middle as the value of μ that minimizes $\sum_{i=1}^{n}\left(Y_i - \mu\right)^2$.
  - This is the physical "center of mass" of the histogram.
  - The end result is $\mu = \bar{Y}$, the mean (and we can do a proof of this).
- Regression through the origin:
  - Suppose the X_i are the parents' heights.
  - Consider picking the slope β that minimizes $\sum_{i=1}^{n}\left(Y_i - X_i\beta\right)^2$.
  - $\beta = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$

Linear least squares

- Let Y_i be the ith child's height and X_i be the ith (average over the pair of) parents' heights.
- Consider finding the "best" line: child's height $=$ β_0 $+$ (parent's height)*β_1
- Use least squares:
  - $\sum_{i=1}^{n}\left[Y_i - \left(\beta_0 + \beta_1 X_i\right)\right]^2$
  - $\hat{\beta}_1 = Cor\left(Y,X\right)\frac{S_y}{S_x}$ (hat indicates that it is an estimator)
    - $\hat{\beta}_1$ has units of Y/X.
  - $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$ (hat indicates that it is an estimator)
    - $\hat{\beta}_0$ has units of Y.
  - The line passes through the point $\left(\bar{X},\bar{Y}\right)$.
  - The slope is the same as what you would get if you centered the data and did regression through the origin.
  - If you normalized the data, the slope is $Cor\left(Y,X\right)$.

Regression to the mean

- Examples:
  - Why are children of tall parents tall, but usually not as tall as their parents?
  - Why do the best-performing athletes from last year usually not perform quite as well this year?
- Imagine simulated pairs of random normal variables.
  - The largest first ones would be the largest by chance, and the probability that there are smaller ones for the second simulation is high.
  - In other words, $P\left(Y < x|X = X\right)$ gets bigger as x heads into the very large values.
  - Similarly, $P\left(Y > x|X = x\right)$ gets bigger as x heads to very small values.
- Suppose that we normalize X (child's height) and Y (parent's height) so that they both have mean 0 and variance 1.
  - Recall that our regression line will pass through $(0, 0)$, the mean of X and Y.
  - The slope of the gression line is $Cor\left(Y, X\right)$, regardless of which variable is the outcome.

Statistical linear regression models

- Basic regression model with additive Gaussian errors:
  - Least squares is an estimation tool, how do we do inference?
  - Consider developing a probabilistic model for linear regression: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
  - ϵ_i are random Gaussian errors, which are assumed to be IID and drawn from $N\left(0, \sigma^2\right)$.
    - Note: $E\left[Y_i|X_i = x_i\right] = \mu_i = \beta_0 + \beta_1 x_i$.
    - Note: $Var\left(Y_i|X_i = x_i\right) = \sigma^2$. This is the variance around the regression line.
- Interpreting regression coefficients:
  - β_0 is the expected value of the response when the predictor is 0.
    - $E\left[Y|X = 0\right] = \beta_0$
    - Note: this isn't always of interest, for example when $X = 0$ is far outside the range of the data, or physically impossible (X is height, blood pressure, etc.).
    - However, you can shift your X values by some factor a; this will change the intercept, but not the slope.
      - $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i = \beta_0 + a\beta_1 + \beta_1\left(X_i - a\right) + \epsilon_i = \tilde{\beta}_0 + \beta_1\left(X_i - a\right) + \epsilon_i$
      - Often, a is set to $\bar{X}$ so that the intercept is interpreted as the expected response at the average X value.
  - β_1 is the expected change in response for a 1 unit change in the predictor.
    - $E\left[Y|X = x + 1\right] - E\left[Y|X = x\right] = \beta_0 + \beta_1\left(x + 1\right) - \left(\beta_0 + \beta_1 x\right) = \beta_1$
    - Consider the impact of changing the units of X:
      - $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i = \beta_0 + \frac{\beta_1}{a}\left(X_i a\right) + \epsilon_i = \beta_0 + \tilde{\beta}_1\left(X_i a\right) + \epsilon_i$
      - Multiplication of X by a factor a results in dividing the slope by a factor of a.

R code for regression

- Linear regression: `fit <- lm(outcome ~ predictor, data = name_of_data_frame)`
  - To just get the coefficients: `coef(fit)`
- Plotting a fit in ggplot: `g + geom_smooth(method="lm", color="black")`
- If you want to do arithmetic operations inside the `lm` function, use the `I()` function:
  - `fit2 <- lm(outcome ~ I(predictor - mean(predictor)), data=name_of_data_frame)`
- Using a fit to do a prediction: `predict(fit, newdata=data.frame(carat=newx))`

Residuals

- Examples here are using the `diamond` dataset from `UsingR`.
- Variation around a regression line is called **residual variation**.
  - Residuals are just the difference between the actual data point values and those predicted by a regression line.
- Model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where $\epsilon_i \sim N\left(0, \sigma^2\right)$.
- Observed outcome at predictor value X_i is Y_i.
- Predicted outcome is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- The residual is the difference between the observed and predicted outcomes: $e_i = Y_i - \hat{Y}_i$.
  - Least squares minimizes $\sum_{i=1}^n e_i^2$
  - The e_i can be thought of as estimates of the ϵ_i.
- Properties of the residuals
  - $E\left[e_i\right] = 0$.
  - If an intercept is included, $\sum_{i=1}^n e_i = 0$.
  - If a regressor variable, X_i, is included in the model, $\sum_{i=1}^n e_i X_i = 0$.
  - Residuals can be thought of as the outcome Y, with the linear association of the predictor X removed.
  - There is a difference between **residual variation** (variation after removing the predictor) from **systematic variation** (variation explained by the regression model).
- To get residuals in R, use `resid(fit)`, where `fit` is the result of a regression function like `lm`.
- When you plot residuals, you shouldn't see any type of pattern, they should look mostly random. If there is a pattern, there may be some systematic error in your analysis or the assumptions you have made.
- Estimating residual variation:
  - The maximum likelihood estimate of σ^2 is $\frac{1}{n}\sum_{i=1}^n e_i^2$, the average squared residual.
  - Most people use $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n e_i^2$ (you lose two degrees of freedom since you have an intercept and a slope in your fit).
    - The $n-2$ instead of n is so that $E\left[\hat{\sigma}^2\right] = \sigma^2$.

- $R^2 = \frac{\sum_{i=1}^n \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^n \left(Y_i - \bar{Y}\right)^2}$
  - Percentage of total variability that is explained by the linear relationship with the predictor.
  - $0 \leq R^2 \leq 1$
  - R^2 is the sample correlation squared.
  - R^2 can be a misleading summary of model fit.
    - Deleting data can inflate R^2.
    - Adding terms to a regression model always increases R^2.

Inference in regression

- Our model is $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where $\epsilon \sim N\left(0, \sigma^2\right)$.
  - We assume that the true model is known.
- Statistics like $\frac{\hat{\theta}-\theta}{\hat{\sigma}_\theta}$ have the following properties:
  - Normally distributed and have a Student's T distribution if the estimated variance is replaced with a sample estimate (under normality assumptions).
  - Can be used to test $H_0 : \theta = \theta_0$ versus $H_a : \theta >, <, \neq \theta_0$.
  - Can be used to create a confidence interval for θ via $\hat{\theta} \pm Q_{1-\alpha/2}\hat{\sigma}_\theta$, where $Q_{1-\alpha/2}$ is the relevant quantile from either a normal or a T distribution.
    - I think you use $n-2$ degrees of freqedom for a T distribution in this case. Not totally sure, though.
- In the case of regression with IID sampling assumption and normal errors, our inferences will look similar to what we covered in the statistical inference class.
- Variance of the regression slope:
  - $\sigma^2_{\hat{\beta}_1} = Var\left(\hat{\beta}_1\right) = \sigma^2 / \sum_{i=1}^n \left(X_i - \bar{X}\right)^2$
- Variance of the intercept:
  - $\sigma^2_{\hat{\beta}_0} = Var\left(\beta_0\right) = \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}\right)\sigma^2$
- In practice, σ can be replaced by its estimate in the above two equations ($\frac{\sum_{i=1}^n e_i^2}{n-2}$).
- It's probably not surprising that under IID Gaussian errors, $\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}}$ follows a T distribution with $n-2$ degrees of freedom and a normal distribution for large n.
  - This can be used to create confidence intervals and perform hypothesis tests.
- Prediction of outcomes:
  - Consider predicting Y at a value X.
  - The obvious estimate for a prediction at point x_0 is $\hat{\beta}_0 + \hat{\beta}_1 x_0$.
  - A standard error is needed to create a prediction interval.
  - There is a distinction between intervals for the regression line at a particular point x_0 and intervals for the prediction of a y value at a point x_0.

        - Line at x_0 standard error: $\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\left(x_0-\bar{X}\right)^2}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}}$
            - This is usually referred to as a confidence interval.
            - R code:

            ```
            cr <- sigma*sqrt(1/n+(x0-mean(x))^2/ssx)
            y0 + c(-1,1)*qt(0.975,df=n-2)*cr
            ```

    - Prediction interval standard error at x_0: $\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{\left(x_0-\bar{X}\right)^2}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}}$
        - This is usually referred to as a prediction interval.
        - R code:

        ```
        pr <- sigma*sqrt(1+1/n+(x0-mean(x))^2/ssx)
        y0 + c(-1,1)*qt(0.975,df=n-2)*pr
        ```

- Code example

    -
      ```
      library(UsingR); data(diamond)
      y <- diamond$price; x <- diamond$carat; n <- length(y)
      beta1 <- cor(x,y)*sd(y)/sd(x);
      beta0 <- mean(y) - beta1*mean(x);
      e <- y - beta0 - beta1*x;
      sigma <- sqrt(sum(e^2)/(n-2))
      ssx <- sum((x-mean(x))^2) ## this is also equal to var(x)*(n-1)
      seBeta0 <- sqrt(1/n + mean(x)^2/ssx)*sigma
      seBeta1 <- sigma/sqrt(ssx)
      tBeta0 <- beta0/seBeta0;
      tBeta1 <- beta1/seBeta1;
      pBeta0 <- 2*pt(abs(tBeta0), df=n-2, lower.tail=FALSE)
      pBeta1 <- 2*pt(abs(tBeta1), df=n-2, lower.tail=FALSE)
      ```

Multivariable regression

- Sometimes there are several variables which may affect a predictor, or hidden variables affecting the outcome.

- Multivariable analyses attempt to account for other variables that may explain a relationship.

- Example:

    - An insurance company is interested in how last year's claims can predict a person's time in the hospital this year.
    - They want to use an enormous amount of data contained in claims to predict a single number.
    - Simple linear regression is not equipped to handle more than one predictor.
    - How can one generalize simple linear regression to incorporate lots of regressors for the purpose of prediction?
    - What are the consequences of adding lots of regressors?
        - There must be consequences to adding variables that are unrelated to the outcome.
        - There must be consequences to omitting variables that are related to the outcome.

- The linear model for multivariate regression:

  - $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + \epsilon_i = \sum_{k=1}^{p} X_{ki}\beta_k + \epsilon_i$
  - Here, X_{1i} is typically 1 so that an intercept is included.
  - Least squares (and hence ML estimates, under IID Gaussianity of the errors) minimizes $\sum_{i=1}^{n} \left(Y_i - \sum_{k=1}^{p} X_{ki}\beta_k\right)^2$
  - Note: the important linearity here is the linearity in the coefficients!
    - Example: $Y_i = \beta_1 X_{1i}^2 + ...\beta_p X_{pi}^2 + \epsilon_i$ is still a linear model, we've just squared the elements of the predictor variables.

- How to get estimates
  - Recall that the least squares estimate for regression through the origin, $E\left[Y_i\right] = X_{1i}\beta_1$ was $\sum_i X_i Y_i / \sum_i X_i^2$.
  - Let's consider two regressors, $E\left[Y_i\right] = X_{1i}\beta_1 + X_{2i}\beta_2 = \mu_i$.
    - Recall that if $\hat{\mu}_i$ satisfies $\sum_{i=1}^{n} \left(Y_i - \hat{\mu}_i\right)\left(\hat{\mu}_i - \mu_i\right) = 0$, then we've found the least squares estimates.
    - $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} e_{i,Y|X_2} e_{i,X_1|X_2}}{\sum_{i=1}^{n} e_{i,x_1|X_2}^2}$: the regression esimate for β_1 is the regression through the origin estimate, having regressed X_2 out of both the response and the predictor.
    - Similarly, the regression estimate for β_2 is the regression through the origin estimate, having regressed X_1 out of both the response and the predictor.
  - Generally, multivariate regression estimates are exactly those having removed the linear relationship of the other variables from both the regressor and the response.

- Example with two variables, simple linear regression
  - $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i}$, where $X_{2i} = 1$ is an intercept term.
  - Then $\frac{\sum_j X_{2j} X_{1j}}{\sum_j X_{2j}^2} X_{2i} = \frac{\sum_j X_{1j}}{n} = \bar{X}_1$.
  - Thus, $e_{i,X_1|X_2} = X_{1i} - \bar{X}_1$ and $e_{i,Y|X_2} = Y_i - \bar{Y}$.
  - $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} e_{i,Y|X_2} e_{i,X_1|X_2}}{\sum_{i=1}^{n} e_{i,X_1|X_2}^2} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} = Cor\left(X, Y\right) \frac{Sd(Y)}{Sd(X)}$

- Extending to the general case:
  - Least square solution to minimize: $\sum_{i=1}^{n} \left(Y_i - X_{1i}\beta_1 - ... - X_{pi}\beta_p\right)^2$
    - Solving this solution yields the least squares estimates.
    - Obtaining a good, fast, and general solution usually requires linear algebra.
  - The least squares estimate for the coefficient of a multivariate regression model is exactly regression through the origin with the linear relationship with the other regressors removed from both the regressor and outcome by taking residuals.
  - In this sense, multivariate regression "adjusts" a coefficient for the linear impact of the other variables.

- Interpretation of the coefficients
  - $E\left[Y | X_1 = x_1, ..., X_p = x_p\right] = \sum_{k=1}^{p} x_k \beta_k$
  - What if one variable x_1 is incremented by one?
  - $E\left[Y | X_1 = x_1 + 1, ..., X_p = x_p\right] - E\left[Y | X_1 = x_1, ..., X_p = x_p\right] = \beta_1$
  - So the intepretation of a multivariate regression coefficient is the expected change in the response per unit change in the regressor, holding all of the other regressors fixed.

- Fitted values, residuals, and residual variation - all of our simple linear regression quantities can be extended to multivariate linear models.
  - Model: $Y_i = \sum_{k=1}^{p} X_{ik}\beta_k + \epsilon_i$ where $\epsilon_i \sim N\left(0, \sigma^2\right)$
  - Fitted responses: $\hat{Y}_i = \sum_{k=1}^{p} X_{ik}\hat{\beta}_k$
  - Residuals: $e_i = Y_i - \hat{Y}_i$
  - Variance estimate: $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2$
  - To get predicted responses at new values $x_1, ..., x_p$, simply plug these values into the linear model $\sum_{k=1}^{p} x_k\hat{\beta}_k$.
  - Our coefficients have standard errors $\hat{\sigma}_{\hat{\beta}_k}$.
    - * $\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}}$ follows a T distribution with $n - p$ degrees of freedom.
  - Predicted responses have standard errors and we can calculate predicted and expected response intervals.
- Linear models are the single most important applied statistical and machine learning technique **by far**.
- Some amazing things that you can accomplish with linear models:
  - Decompose a signal into its harmonics.
  - Flexibly fit complicated functions.
  - Fit factor variables as predictors.
  - Uncover complex multivariate relationship with the response.
  - Build accurate prediction models.
- Dummy variables
  - The equation for representing the relationship between a particular outcome and several factors contains binary variables, one for each factor. These are called "dummy variables" and each indicates if a particular outcome is associated with a specific factor or category.
  - Consider the linear model $Y_i = \beta_0 + X_{i1}\beta_1 + \epsilon_i$ where each X_{i1} is binary so that it is 1 if measurement i is in a group and 0 otherwise.
  - Then for people in the group, $E\left[Y_i\right] = \beta_0 + \beta_1$.
  - For people not in the group, $E\left[Y_i\right] = \beta_0$
  - β_1 is interpreted as the increase or decrease in the mean comparing those in the group to those who are not.
  - Consider a multi-level factor level. Let's say a three-level factor based on political party affiliation (Republican, Democrat, Independent).
    - * $Y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \epsilon_i$ where X_{i1} is 1 for Republicans and 0 otherwise and X_{i2} is 1 for Democrats and 0 otherwise.
    - * If i is Republican, Democrat, or Independent, $E\left[Y\right] = \beta_0 + \beta_1$, $\beta_0 + \beta_2$, or β_0, respectively.
    - * β_1 compares Republicans to Independents, β_2 compares Democrats to Independents, and $\beta_1 - \beta_2$ compares Republicans to Democrats.
  - The first entry in the Estimate column is labeled as "(Intercept)". That is because sprayA is the first in the alphabetical list of the levels of the factor, and R by default uses the first level as the reference against which the other levels or groups are compared when doing its t-tests (shown in the third column).

  - The estimates in this case are the coefficients of the binary or dummy variables. The Intercept is the mean of the reference group, and the other Estimates are the distances of the other groups' means from the reference mean.
    * If we do an lm(y ~ x-1), this will remove the intercept. Then sprayA will be shown and all of the Estimates (and other quantities) will be absolute, instead of being computed relative to sprayA.
  - To choose the reference group, you can use the `relevel()` function to re-order the factor.
  - To get t-values relative to the reference group, take `(fit$coef[2]-fit$coef[3])/standard error`. The standard error is usually obtained from the one-on-one regressions.

Multivariable regression - notes from swirl()

- In this lesson we'll illustrate that regression in many variables amounts to a series of regressions in one.
- Using regression in one variable, we'll show how to eliminate any chosen regressor, thus reducing a regression in N variables, to a regression in N-1.
  - Hence, if we know how to do a regression in 1 variable, we can do a regression in 2.
  - Once we know how to do a regression in 2 variables, we can do a regression in 3, and so on.
- We begin with the galton data and a review of eliminating the intercept by subtracting the means.
- When we perform a regression in one variable, such as lm(child ~ parent, galton), we get two coefficients, a slope and an intercept.
  - The intercept is really the coefficient of a special regressor which has the same value, 1, at every sample. The function, lm, includes this regressor by default.
  - In earlier lessons we demonstrated that the regression line given by lm(child ~ parent, galton) goes through the point x=mean(parent), y=mean(child).
  - We also showed that if we subtract the mean from each variable, the regression line goes through the origin, x=0, y=0, hence its intercept is zero.
  - Thus, by subtracting the means, we eliminate one of the two regressors, the constant, leaving just one, parent. The coefficient of the remaining regressor is the slope.
  - Subtracting the means to eliminate the intercept is a special case of a general technique which is sometimes called Gaussian Elimination.
  - As it applies here, the general technique is to pick one regressor and to **replace all other variables by the residuals of their regressions against that one**.
- Example with swiss dataset:
  - First, use the R function lm to generate the linear model "all" in which Fertility is the variable dependent on all the others. Use the R shorthand "." to represent the five independent variables in the formula passed to lm. Remember the data is "swiss".
  -
    ```
    all <- lm(Fertility ~ ., data=swiss)
    summary(all)
    Call: lm(formula = Fertility ~ ., data = swiss)
    Residuals:       Min        1Q   Median        3Q      Max  -15.2743  -5.2617
    0.5032   4.1198  15.3213
    Coefficients:
                                    Estimate  Std. Error  t value Pr(>|t|)
    (Intercept)       66.91518    10.70604    6.250   1.91e-07
    Agriculture       -0.17211     0.07030   -2.448   0.01873
    ```

```
Examination        -0.25801     0.25388   -1.016   0.31546
Education          -0.87094     0.18303   -4.758  2.43e-05
Catholic            0.10412     0.03526    2.953   0.00519
Infant.Mortality    1.07705     0.38172    2.822   0.00734
--- Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 7.165 on 41 degrees of freedom Multiple R-squared:
0.7067,    Adjusted R-squared:  0.671  F-statistic: 19.76 on 5 and 41 DF,
p-value: 5.594e-10
```

  - Recall that the Estimates are the coefficients of the independent variables of the linear model (all of which are percentages) and they reflect an estimated change in the dependent variable (fertility) when the corresponding independent variable changes. So, for every 1% increase in percent of males involved in agriculture as an occupation we expect a .17% decrease in fertility, holding all the other variables constant; for every 1% increase in Catholicism, we expect a .10% increase in fertility, holding all other variables constant.

Interactions

- Example: hunger data.
- Linear model: $Hu_i = b_0 + b_1 Y_i + e_i$
  - Hu_i: percent of hungry children.
  - b_0: percent hungry at year 0.
  - b_1: decrease in percent hungry per year.
  - e_i: everything we didn't measure.
- If we consider a model with two lines:
  - Females: $HuF_i = bf_0 + bf_1 YF_i + ef_i$
    - bf_0: percent of girls hungry at year 0.
    - bf_1: decrease in percent of girls hungry per year.
    - ef_i: everything we didn't measure.
  - Males: $HuM_i = bm_0 - bm_1 YM_i + em_i$
    - bm_0: percent of boys hungry at year 0.
    - bn_1: decrease in percent of boys hungry per year.
    - em_i: everything we didn't measure.
  - When we stratify by gender, the slopes and intercepts and residual variances are different for each gender.
- Two lines, same slope, different intercept: $Hu_i = b_0 + b_1 1\left(Sex_i = ''Male''\right) + b_2 Y_i + e_i^*$
  - b_0: percent hungry at year zero for females.
  - $b_0 + b_1$: percent hungry at year zero for males.
  - b_2: change in percent hungry (for either males or females) in one year.
  - e_i^*: everything we didn't measure.
  - How to do this in R:

    ```
    lmBoth <-lm(Numeric ~ Year + Sex, data=hunger)
    ```

  - This says that our slope variable, Year, does not interact with Sex. So only the intercepts will be different.

- Two lines, different slopes (interactions): $Hu_i = b_0 + b_1 1\left(Sex_i = "Male"\right) + b_2 Y_i + b_3 1\left(Sex_i = "Male"\right) Y_i + e_i^+$
  - b_0: percent hungry at year 0 for females.
  - $b_0 + b_1$: percent hungry at year zero for males.
  - b_2: change in percent hungry (females) in one year.
  - $b_2 + b_3$: change in percent hungry (males) in one year.
  - e_i^+: everything we didn't measure.
  - $E\left[H\right] = b_0 + b_1 + \left(b_2 + b_3\right) Y$ for males.
  - $E\left[H\right] = b_0 + b_2 Y$ for females.
  - How to do this in R: `lmInter <- lm(Numeric ~ Year + Sex + Sex*Year, data=hunger)`
- Interpreting a continuous interaction:
  - $E\left[Y_i | X_{1i} = x_1, X_{2i} = x_2\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$
    - The last term here is the interaction term.
  - Holding X_2constant, we have: $E\left[Y_i | X_{1i} = x_1 + 1, X_{2i} = x_2\right] - E\left[Y_i | X_{1i} = x_1, X_{2i} = x_2\right] = \beta_1 + \beta_3 x_2$
    - Thus, the expected change in Y per unit X_1 holding all else constant is not constant.
    - β_1 is the slope when $x_2 = 0$.
  - We can do the same kind of thing to see what happens when both variables change:
    - End result: β_3 is the expected change in Y per unit change in X_1, per unit change in X_2.
    - We can also think of it as the change in the slope relating X_1 and Y per unit change in X_2.
- To investigate residual relationships:
  - Can plot `resid(lm(y ~ x2))` vs. `resid(lm(y ~ x1))`.

Residuals and diagnostics

- Residuals are defined as $e_i = Y_i - \hat{Y}_i$.
- Our estimate of residual variation is $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$ where $n - p$ is chosen such that $E\left[\hat{\sigma}^2\right] = \sigma^2$, i.e., the estimator is unbiased.
- Influential, high leverage, and outlying points
  - Calling a point an outlier is **vague**.
    - Outliers can be the result of spurious or real processes (example: data entry error or a statistical fluctuation).
    - Outliers can have varying degrees of influence.
    - Outliers can conform to the regression relationship (i.e., being marginally outlying in X or Y, but not outlying given the regression relationship).
- Measures of influence
  - Do `?influence.measures` to see the full suite of influence measures in stats.
    - `rstandard`: standardized residuals (residuals divided by their standard deviations).
    - `rstudent`: standardized residuals where the ith data point was deleted in the calculation of the standard deviation for the residual to follow a t-distribution.

    - `hatvalues`: measures of leverage.
    - `dffits`: change in the predicted response when the ith point is deleted in fitting the model.
    - `dfbetas`: change in the individual coefficients when the ith point is deleted in fitting the model.
    - `cooks.distance`: overall change in the coefficients when the ith point is deleted.
    - `resid`: returns the ordinary residuals.
    - `resid(fit)/(1-hatvalues(fit))`: returns the PRESS residuals, i.e., the "leave-one-out" cross-validation residuals - the difference in the response and the predicted response at data point i, where it was not included in the model fitting. Here `fit` is the linear model.
  - How to use these:
    - Be wary of simplistic rules for diagnostic plots and measures. The use of these tools is context-specific.
    - Not all of the measures have meaningful absolute scales. You can look at them relative to the values across the data.
    - They probe your data in different ways to diagnose different problems.
    - Patterns in your residual plots generally indicate some poor aspect of model fit, including:
      - Heteroskedasticity (non-constant variance).
      - Missing model terms.
      - Temporal patterns (plot residuals vs. collection order of data points).
    - Residual QQ plots investigate normality of the errors.
    - Leverage measures (hat values) can be useful for diagnosing data entry errors.
    - Influence measures get to the bottom line - how does deleting or including this point impact a particular aspect of the model?

Model selection

- Choosing a model when you have multiple predictors
  - Machine learning class focuses on prediction, so we'll focus on modeling.
- Prediction has a different set of criteria, needs for interpretability and standard for generalizability.
- In modeling, we want to make a parsimonious, interpretable representation of the data that enhances our understanding of the phenomena under study.
  - Under this philosophy, what's the right model? Whatever model connects the data to a true, parsimonious statement abotu what you're studying.
  - Good modeling decisions are context-dependent.
  - We'll focus on variable inclusion and exclusion.
- "Rumsfeldian triplet"
  - Known knowns: regressors that we know we should include in the model and have included.
  - Known unknowns: regressors that we would like to include in the model but we don't have them.
  - Unknown unknowns: regressors that we don't even know about that we should have included in the model.
- General rules
  - Omitting variables tends to result in bias unless their regressors are uncorrelated with the omitted ones.

        - This is why we randomize treatments - it attempts to uncorrelate our treatment indicator with variables that we don't have and can't include in the model.
        - If there are too many unobserved confounding variables, even randomization won't help us.
    - Including variables that we shouldn't have increases standard errors of the regression variables.
        - Thus, we don't want to idly throw new variables into the model.
    - The model must tend toward perfect fit as the number of non-redundant regressors approaches the number of measurements.
    - R^2 increases monotonically as more regressors are added.
    - The sum of the squared errors decreases monotonically as more regressors are included.

- Variance inflation: variance of model tends to increase as you add uncorrelated predictors (i.e., variables that aren't relevant and you shouldn't add).
    - If the regressors you add are correlated to other regressors, then your variance will become even more inflated!
    - Note: we don't actually know σ, so we can only estimate the increase in the actual standard error of the coefficients for including a regressor.
        - However, σ drops out of the relative standard errors. If you sequentially add variables, you can check the variance inflation for including each one.
    - When the other regressors are actually orthogonal to the regressor of interest, then there is no variance inflation.
    - The variance inflation factor (VIF) is the increase in the variance for the ith regressor compared to the ideal setting where it is orthogonal to the other regressors.
        - Use the `vif(fit)` function in R to get this.
    - Note: variance inflation is only part of the picture. We may still want to include certain variables, even if they dramatically inflate our variance.

- What about residual variance estimation?
    - Assuming that the model is linear with additive IID errors (with finite variance), we can mathematically describe the impact of omitting necessary variables or including unnecessary ones.
    - If we underfit the model, the variance estimate is biased.
    - If we correctly fit or overfit the model, including all necessary covariates and/or uncessary covariates, the variance estimate is biased ($E\left[\hat{\sigma}^2\right] = \sigma^2$).
    - However, the variance of the variance is larger if we include unncessary variables ($Var\left(\hat{\sigma}^2_{unnecessary}\right) \geq Var\left(\hat{\sigma}^2_{necessary}\right)$.

- Covariate model selection
    - Principal components or factor analysis models on covariates are often useful for reducing complex covariate spaces.
    - Good design can often eliminate the need for complex model searches.
    - If the models of interest are nested and without lots of parameters differentiating them, it's fairly uncontroversial to use nested likelihood ratio tests.
    - Suggested approach: given a coefficient of interest, use covariate adjustment and multiple models to probe that effect to evaluate it for robustness and to see what other covariates knock it out. This isn't terribly systematic, but it does tend to teach you a lot about the data.

- How to do nested model testing in R:

-
  ```
  fit1 <- lm(Fertility ~ Agriculture, data=swiss)
  fit2 <- update(fit, Fertility ~ Agriculture + Examination + Education)
  fit5 <- update(fit, Fertility ~ Agriculture + Examination + Education + Catholic +
  anova(fit1, fit3, fit5) ## creates a series of likelihood ratio tests. Output comp
  ```

Generalized linear models

- Linear models are one of the most useful applied statistical techniques, but they do have limitations.
  - Additive response models don't make much sense if the response is discrete or strictly positive.
  - Additive error models often don't make sense. Example: if the outcome has to be positive.
  - Transformations are often hard to interpret.
  - There is value in modeling the data on the scale with which it was collected.
  - Particularly interpretable transformations, specifically natural logarithms, aren't applicable for negative or zero values.
- Generalized linear models (GLMs) involve three components:
  - An exponential family model for the response.
  - A systematic component via a linear predictor.
  - A link function that connects the means of the response to the linear predictor.
- Example:
  - Assume that $Y_i \sim N\left(\mu_i, \sigma^2\right)$ (the Gaussian distribution is an exponential family distribution).
  - Define the linear predictor to be $\eta_i = \sum_{k=1}^{p} X_{ik}\beta_k$
  - The link function is $g\left(\mu\right) = \eta$.
    - For linear models, $g\left(\mu\right) = \mu$ such that $\mu_i = \eta_i$
  - This yields the same likelihood model as our additive error Gausisan linear model $Y_i = \sum_{k=1}^{p} X_{ik}\beta_k + \epsilon_i$
    - As usual, $\epsilon_i \sim N\left(0, \sigma^2\right)$ and are assumed to be IID.
- Example: logistic regression.
  - Assume that $Y_i \sim \mathit{Bernoulli}\left(\mu_i\right)$ such that $E\left[Y_i\right] = \mu_i$ where $0 \le \mu_i \le 1$.
  - Linear predictor: $\eta_i = \sum_{k=1}^{p} X_{ik}\beta_k$
  - Link function: $g\left(\mu\right) = \eta = \log\left(\frac{\mu}{1-\mu}\right)$ is the natural log of the odds, referred to as the logit.
  - Note that we an invert the logit function as $\mu_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$ and $1 - \mu_i = \frac{1}{1+\exp(\eta_i)}$
  - Thus, the likelihood is $\prod_{i=1}^{n} \mu_i^{y_i}\left(1-\mu_i\right)^{1-y_i} = \exp\left(\sum_{i=1}^{n} y_i\eta_i\right)\prod_{i=1}^{n}\left(1+\eta_i\right)^{-1}$
- Example: Poisson regression.
  - Assume that $Y_i \sim \mathit{Poisson}\left(\mu_i\right)$ such that $E\left[Y_i\right] = \mu_i$.
  - Linear predictor: $\eta_i = \sum_{k=1}^{p} X_{ik}\beta_k$
  - Link function: $g\left(\mu\right) = \eta = \log\left(\mu\right)$.
  - Recall that e^x is the inverse of $\log\left(x\right)$ so that $\mu_i = e^{\eta_i}$.
  - Thus, the likelihood is $\prod_{i=1}^{n}\left(y_i!\right)^{-1}\mu_i^{y_i}e^{-\mu_i} \propto \exp\left(\sum_{i=1}^{n} y_i\eta_i - \sum_{i=1}^{n}\mu_i\right)$

- In each case, the only way in which the likelihood depends on the data is through $\sum_{i=1}^{n} y_i \eta_i = \sum_{i=1}^{n} y_i \sum_{k=1}^{p} X_{ik}\beta_k = \sum_{k=1}^{p} \beta_k \sum_{i=1}^{n} X_{ik} y_i$.
  - Thus, we don't really need the full data, only $\sum_{i=1}^{n} X_{ik} y_i$. This simplification is a consequence of choosing so-called "canonical" link functions.
- All models achieve their maximum at the root of the so-called normal equations:
  - $0 = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{Var(Y_i)} W_i$, where W_i are the derivatives of the inverse of the link function.
- Variances
  - For the linear model: $Var\left(Y_i\right) = \sigma^2$ is constant.
  - For the binomial case: $Var\left(Y_i\right) = \mu_i\left(1 - \mu\right)$.
  - For the Poisson case: $Var\left(Y_i\right) = \mu_i$.
  - In the latter cases, it is often relevant to have a more flexible variance model, even if it doesn't correspond to an actual likelhiood.
    - $0 = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{\phi\mu_i(1-\mu_i)} W_i$ and $0 = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{\phi\mu_i} W_i$
    - These are called 'quasi-likelihood' normal equations.
    - ϕ is an extra parameter used to calculate the variance (doesn't depend on i). It can help when there are complicated variance restrictions.
    - The purpose is to relax the very strict assumptions about the relationships between the means and variances in the models.
- Odds and ends
  - The normal equations have to be solved iteratively. Results are $\hat{\beta}_k$ and ϕ (if included).
  - Predicted linear predictor responses can be obtained as $\hat{\eta} = \sum_{k=1}^{p} X_k \hat{\beta}_k$.
  - Predicted mean responses are given by $\hat{\mu} = g^{-1}\left(\hat{\eta}\right)$.
  - Coefficients are interpreted as $g\left(E\left[Y|X_k = x_k + 1, X_{\sim k} = x_{\sim k}\right]\right) - g\left(E\left[Y|X_k = x_k, X_{\sim k} = x_{\sim k}\right]\right) = \beta_k$
    - Difference in a one-unit increase in a particular coefficient, holding the other constant.
    - Done on the link function scale.
  - Variations on the Newton-Raphson method are used to solve iteratively.
  - Asymptotics are used for inference usually.
  - Many of the ideas from linear models can be brought over to GLMs.

More on odds

- Using ANOVA: `anova(logRegRavens, test=”Chisq”)`
  - Want to do this with nested models.
  - Deviance: measure of model fit compared to the previous model.
  - To compare deviances, take a smaller model and subtract the larger model from it.
  - The deviance residuals can be compared to a chi-squared distribution with degrees of freedom equal to that of the difference in degrees of freedom between the two models.
- Interpreting odds ratios
  - They are NOT probabilities!
  - Odds ratio of 1 = log odds ratio of 0 = no difference in odds.
  - Odds ratio < 0.5 or > 2 commonly called a "moderate effect."
  - Relative risk $\frac{Pr(W_i|S_i=10)}{Pr(W_i|S_i=0)}$ is often easier to interpret but harder to estimate.
  - For small probabilities, RR $\approx$ OR, but they are not the same!

GLMs and odds

- Example: playing a game where you flip a coin with success probability p.
  - If it comes up heads, you win X, if it comes up tails, you lose Y.
  - What should we set for X and Y for the game to be fair, i.e. $E\left[earnings\right] = Xp - Y\left(1-p\right) = 0$
  - This implies that $\frac{Y}{X} = \frac{p}{1-p}$.
  - The odds can be stated as "how much you should be willing to pay for a probability p of winning a dollar."
    - If $p > 0.5$, you have to pay more if you lose than you get if you win.
    - If $p < 0.5$, you have to pay less if you lose than you get if you win.
- Ravens logistic regression:

```
logRegRavens <- glm(ravenWinNum ~ ravenScore, family="binomial", data=ravensData)
```

  - Intercept: log odds of Ravens winning when they score zero points.
  - Slope: increase in log odds for every point that they score.
  - Note: $e^x \approx 1 + x$ for small x.
- Odds ratios and confidence intervals
  - `exp(logRegRavens$coeff)`
  - Can exponentiate the confidence interval: `exp(confint(logRegRavens))`.

Poisson regression

- Key ideas:
  - May data take the form of counts: calls to a call center, number of flu cases in an area, number of cars that cross a bridge, etc.
  - Data may also be in the form of rates: percent of children passing a test, percent of hits to a website from a particular country, etc.
  - Linear regression with transformation is an option.
- Uses of the Poisson distribution: counts and rates
  - Examples: web traffic hits, incidence rates, contingency table data, approximating binomial probabilities with small p and large n.
- The Poisson mass function
  - $P\left(X = x\right) = \frac{(t\lambda)^x e^{-t\lambda}}{x!}$ for $x = 0, 1, 2, ...$
  - $E\left[X\right] = Var\left(X\right) = t\lambda$
  - The Poisson distribution tends to a normal distribution as $t\lambda$ gets large.
- Example: number of hits per day on a website
  - Can model as linear regression: $H_i = b_0 + b_1 D_i + e_i$, `lm1 <- lm(visits ~ date)`
    - b_0: number of hits on Julian day 0.
    - b_1: increase in number of hits per day.
    - H_i: number of hits.
    - D_i: Julian day.

  - Taking the log of the outcome has a specific interpretation: $\log(H_i) = b_0 + b_1 D_i + e_i$
    * b_0: log number of hits on Julian day 0.
    * b_1: increase in log number of hits per day.
  - Poisson/log-linear: $\log(E[H_i|D_i, b_0, b_1]) = b_0 + b_1 D_i$
    * Multiplicative differences: $E[H_i|D_i, b_0, b_1] = \exp(b_0 + b_1 D_i) = \exp(b_0)\exp(b_1 D_i)$
    * If D_i is increased by one unit, $E[H_i|D_i, b_0, b_1]$ is multiplied by $\exp(b_1)$.

- Exponentiating coefficients
  - $\exp(E[\log(Y)])$ is the geometric mean of Y.
  - When you take the natural log of outcomes and fit a regression model, your exponentiated coefficients estimate things about geometric means.
  - e^{β_0} estimates the geometric mean hits on day 0.
  - e^{β_1} estimates relative increase or decrease in geometric mean hits per day.
  - There is a problem with logs when you have zero counts, but you can add a constant without affecting your results.

- Poisson regression in R

```
glm1 <- glm(visits ~ julianDay, family="poisson")
## Plotting the fit
lines(julianDay, glm1$fitted)
## confidence intervals
confint(glm1)
## agnostic confidence intervals
confint.agnostic(glm1)
```

- Modeling rates
  - In `glm()`, use an offset like `offset=log(visits+1)`