Motivating example: Sex discrimination in wages

- In 1970's, Harris Trust and Savings Bank was sued for discrimination on the basis of sex.
- Analysis of salaries of employees of one type (skilled, entry-level clerical) presented as evidence by the defense.
- Did female employees tend to receive lower starting salaries than similarly qualified and experienced male employees?

Variables collected

- 93 employees on data file (61 female, 32 male).

bsal: Annual salary at time of hire.
sal77 : Annual salary in 1977.
educ: years of education.
exper: months previous work prior to hire at bank.
fsex: 1 if female, 0 if male
senior: months worked at bank since hired
age: months

Comparison of beginning wages for male and female employees

- T-test shows men started at higher salaries than women ($t=5.8$, $p<.0001$).
- But, it doesn't control for other characteristics.
- From scatter plots, there appear to be associations among gender, salary, and other variables.

Multiple regression model

- For any combination of values of the predictor variables, the average value of the response (bsal) lies on a straight line:

$$\textit{Avg. bsal} = \textit{Intercept} + B_1\,\textit{sex} + B_2\,\textit{senior} + B_3\,\textit{age} + B_4\,\textit{educ} + B_5\,\textit{exper}$$

- Just like in simple regression, assume values of response follow a normal curve within any combination of predictors.

Regression model: Math

- The regression model assumes the following distribution for $Y \mid$ all X.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i,$$

$$\text{where } f(\varepsilon_i) = N(0, \sigma^2)$$

- Another way to write this is:

$$f(y \mid x) = N(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \sigma^2)$$

Regression model: Math

- Estimated coefficients found by taking partial derivatives of

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}))^2$$

- Resulting formulas are too messy to write down, although there is a very nice matrix algebra representation.

Regression model: Math

- Estimated value of regression variance is

$$s_e^2 = \sum_{i=1}^{n} \frac{(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_p x_{pi}))^2}{n - (p+1)}$$

- R-squared has same interpretation.

Matrix Algebra Representation

- Let Y be the n x 1 vector of responses.
- Let X be the n x (p+1) matrix of predictors, such that the ith row of X has the entries

$$(1, x_{1i}, x_{2i}, \ldots, x_{pi})$$

- Then, the OLS estimates of all (p+1) coefficients (intercept plus p slopes) is given by

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

Matrix Algebra Representation

- The variance of the OLS estimates of all (p+1) coefficients (intercept plus p slopes) is

$$Var(\hat{\beta}) = (X^t X)^{-1}\sigma^2$$

- Plug in s_e^2 as estimate of σ^2
- Take square root of the diagonal elements to get the SEs reported by software and used for testing and interval estimation.

Output from regression
(fsex = 1 for females, = 0 for males)

Term	Estimate	Std Error	t Ratio	Prob>|t|
Int.	6277.9	652	9.62	<.0001
Fsex	-767.9	128.9	-5.95	<.0001
Senior	-22.6	5.3	-4.26	<.0001
Age	0.63	.72	.88	.3837
Educ	92.3	24.8	3.71	.0004
Exper	0.50	1.05	.47	.6364

Interpretation of coefficients in multiple regression

- Each estimated coefficient is amount ***Y*** is expected to increase when the value of its corresponding predictor is increased by one, *holding constant the values of the other predictors*.

- Example: estimated coefficient of education equals 92.3.

  For each additional year of education for employee, we expect salary to increase by about 92 dollars, holding all other variables constant.

- Estimated coefficient of fsex equals -767.

  For employees who started at the same time, had the same education and experience, and were the same age, women earned $767 less on average than men.

Which variable is the strongest predictor of the outcome?

- The coefficient that has the strongest linear association with the outcome variable is the one with the largest absolute value of ***T,*** which equals the coefficient over its SE**.**
- It is not size of coefficient. This is sensitive to scales of predictors. The ***T*** statistic is not, since it is a standardized measure.
- Example: In wages regression, seniority is a better predictor than education because it has a larger ***T.***

Hypothesis tests for coefficients

- The reported t-stats (coef. / SE) and p-values are used to test whether a particular coefficient equals 0, given that all other coefficients are in the model.

- Examples:

  1) Test whether coefficient of education equals zero has p-value = .0004. Hence, reject the null hypothesis; it appears that education is a useful predictor of bsal when all the other predictors are in the model.

- 2) Test whether coefficient of experience equals zero has p-value = .6364. Hence, we cannot reject the null hypothesis; it appears that experience is not a particularly useful predictor of bsal when all other predictors are in the model.

Hypothesis tests for coefficients

- The test statistics have the usual form (observed – expected)/SE.
- For p-value, use area under a t-curve with (n-(p+1)) degrees of freedom, where (p+1) is the number of terms in the model.
- In this problem, the degrees of freedom equal (93-6=87).

CIs for regression coefficients

- A 95% CI for the coefficients is obtained in the usual way:

  coef. ± (multiplier) SE

- The multiplier is obtained from the t-curve with (n-(p+1)) degrees of freedom.
- Example: A 95% CI for the population regression coefficient of age equals:

  (0.63 – 1.96*0.72, 0.63 + 1.96*0.72)

CIs for regression coefficients

- Example usage of CIs:

  For employees with the same age, seniority, education, and experience, we expect the average starting salary for female employees to be between $511 and $1024 less than the average starting salary for male employees.

- More succinctly,

  For employees with the same age, seniority, education, and experience, we expect female employees' average starting salary to be around $767 less than male employees' average salary (95% CI: $-1024 to $-511).

Reminder about tests and CIs

- When sample size is large enough, probably reject null hypothesis of $\beta=0$.
  - Consider practical significance, not just statistical significance.

- When sample size is small, may not be enough evidence to reject null hypothesis of $\beta=0$.
  - When you fail to reject null hypothesis, don't be too hasty to say that predictor has no linear association with the outcome. There may be an association, just not strong enough to detect with this sample (or perhaps a nonlinear one).

Output from regression
(fsex = 1 for females, = 0 for males)

Term	Estimate	Std Error	t Ratio	Prob>|t|
Int.	6277.9	652	9.62	<.0001
Fsex	-767.9	128.9	-5.95	<.0001
Senior	-22.6	5.3	-4.26	<.0001
Age	0.63	.72	.88	.3837
Educ	92.3	24.8	3.71	.0004
Exper	0.50	1.05	.47	.6364

Predictions

- Example: Prediction of beginning wages for 25 year old woman with 12 years of education, 10 months of seniority, and two years of experience:

Pred. bsal = 6277.9 - 767.9*1 - 22.6*10
+ .63*300 + 92.3*12 + .50*24
= 6592.6

- Get prediction intervals for individuals or averages as in simple regression using the predict command.

General warnings: Same as for simple regression

- Be even more wary of extrapolation. Because there are several predictors, you can extrapolate in many ways.

- Multiple regression shows association. It does not prove causality. Only a carefully designed observational study or randomized experiment can show causality.

Checking assumptions

- CIs and p-values, as well as predictions, are meaningful only when regression assumptions are plausible.
- Plot the residuals versus the predicted values from the regression line.
- Also plot the residuals versus each of the predictors.
- If systematic patterns in these plots, the assumptions might be violated and results are suspect.

Summary of residual plots

- There appears to be a systematic pattern in the plot of residuals versus experience, and also versus age.
- This model can be improved. Let' s try.... (see the R script for details)
- First, let's introduce mean-centering for predictors, which is a general strategy for improving interpretability of regression output

Advice to aid interpretation

- Intercepts are often hard to interpret, because they represent value when all predictors equal zero. This may be an unrealistic or uninteresting case.
- Instead, for each continuous predictor, subtract its mean from every value. Use these mean-centered predictors in regression
- Intercept interpreted as average value of Y at the average value of X, which is much more interpretable.
- From now on, we mean-center continuous predictors.

Special predictors: Quadratic terms

- Sometimes outcome appears to have quadratic (or even higher order polynomial) trends with some predictors.
- Can include squared terms (or higher order powers) for predictors to capture trends.
- General practice: include all lower order terms when including higher order ones. Aids interpretation.
- Best way to present quadratic trends is to plot predicted average of Y for different values of X.

Special predictors: Indicator (dummy) variables

- Notice how we modeled gender: make a variable equal to one for all females and zero for all males.
- We could have made a variable equal to one for all males and zero for all females, instead.
- The value of that coefficient would be 767 instead of -767. All other statistics stay the same (SE, t-stat, p-value)
- Using the male predictor would not change the estimates for any other coefficient.

Special predictors: Indicator (dummy) variables

- Can we include both the male and female indicator variables in the same model?
- We cannot when we also include the intercept.
- Not possible to estimate all three of these parameters in the same model uniquely.
- Note: no need to mean-center dummy variables, since they have a natural interpretation at zero.

Special predictors: Indicator (dummy) variables

- What if a categorical variable has k > 2 levels?
- Make k dummy variables, one for each level.
- Use only k – 1 of the levels in the regression model, since we cannot uniquely estimate all k at once if we also include an intercept.
- Excluded level is called the baseline.
- Values of coefficients of dummy variables are interpreted as changes in average Y over baseline
- Example with diamonds data in class

Special predictors: Interaction terms

- Sometimes relationship of some predictor with Y depends on values of other predictors.
- This is called an interaction effect.
- Make an interaction predictor: multiply one predictor times the other predictor in the interaction.
- General practice to include all main effects (each variable without interaction) when including interactions.
- Examples in class with ozone data and diamonds data.

Testing if groups of coefficients equal to zero

- With dummy variables, polynomial terms, and interactions, want to test if multiple coefficients equal zero or not.
- Use nested F test (called extra sums of squared in Sleuth).
- Compute

$$F = \frac{(SSE_{small} - SSE_{big})/(p_{big} - p_{small})}{SSE_{big}/(n-(p_{big}+1))}$$

Testing if groups of coefficients equal to zero

- Look for area under the F curve with $p_{big} - p_{small}$ degrees of freedom in the numerator, and $n-(p_{big}+1)$ degrees of freedom in the denominator,

$$F = \frac{(SSE_{small} - SSE_{big})/(p_{big} - p_{small})}{SSE_{big}/(n-(p_{big}+1))}$$

Leverage, Influence, and Standardized Residuals

- Individual observations can have large impact on the estimates of coefficients and SEs.
- Sometimes obvious from scatter plots but often not, especially in multivariate data.
- Concepts and metrics of leverage, influence, and standardized residuals can help identify impactful and unusual points.

Leverage

- Points with outlying predictor values are called leverage points.
- Has nothing to do with values of the response.
- These points POTENTIALLY have large impact on the estimates of coefficients and SEs.
- Special quantity called “leverage” that is large for observations with combinations of predictor values far from typical combinations.

Leverage – matrix algebra (not required for our course)

- For those who know matrix algebra, the leverage for record *i* is defined as the *i*th diagonal element of the matrix,

$$H = I - X\left(X^t X\right)^{-1} X^t$$

where I is a *p x p* identity matrix, and *X* is the *n* x *p* matrix of predictors.

High leverage: What to do?

- Points with high leverage deserve special attention:
  - Make sure that they do not result from data entry errors.
  - Make sure that they are in scope for the types of individuals for which you want to make predictions.
  - Make sure that you look at the impact of those points on estimates, especially when you have interactions in the model.
- Just because a point is a leverage point does not mean it will have large effect on regression. That depends on values of the outcome variable…

Cooks Distance

- What if a point has large impact on estimates of regression coefficients?
  - Dropping that point should change the coefficients lots.
  - Changing coefficients lots should change that point's predicted Y value lots.
- For every point, we could delete it, re-run the regression, and see which points lead to big changes in predicted Y.
- This is time consuming.
- Turns out there is a simple formula that gives the squared change in the predicted Y value after dropping any point. Called the Cooks distance.

Big Cooks Distances: What to do?

- Examine Cooks Distance to look for large values.
  - Make sure there are no data entry errors in those points.
  - For each point with high Cooks distance, fit model with and without that point.
  - If results (predictions or scientific interpretations) do not change much, just report the final model based on all data points and don't reporting anything about the Cooks Distance.
  - If results change a lot, you have several options….

Cooks distance: What to do if large changes in results?

- OK to drop case based on PREDICTOR values if (1) scientifically meaningful to do so and (2) you intended to fit a model over the smaller X range to begin with (and just forgot). You have to mention this in analysis write-up and be careful when making predictions to avoid extrapolation.
- NOT OK to drop point based on its outcome value (assuming no data errors in that value). These are legitimate observations. Dropping them is cheating by changing the data to fit the model.
- Try transformations or collect more data.

Standardized residuals (also called internally studentized residuals)

- How do we best identify outliers, i.e., points that don't fit the pattern implied by the line?
- Look for points with relatively large residuals.
- Would be nice to have a common scale to interpret what a "big" residual is, across all problems.
- As with most metrics in statistics, we look at residual divided by its standard error (hence the term standardized residual).

Standardized residuals (also called internally studentized residuals)

- Turns out that the variance (SE) of any residual (not the ε) depends on the values of the predictors.
- Residuals for high leverage predictors have smaller variance than residuals for low leverage predictors.
- Intuition: the regression line tries to fit high leverage points as closely as possible, which means smaller residuals for those points.
- Standardized residuals have a Normal(0,1) distribution.

Standardized residuals (also called internally studentized residuals)

- Values with large standardized residuals are outliers, in that they don't fit the pattern implied by the line.
- Like any normal distribution, we expect some large standardized values, e.g., 5% of standardized residuals should be outside (-2, 2).
- Values with large standardized residuals not necessarily influential on the regression line. Can be an outlier without impacting the line.
- In fact, really one should do residual plots with standardized residuals instead of regular ones, since they can reflect constant variance assumption when it is true.

Standardized residuals: What to do if large outliers?

- OK to drop case based on PREDICTOR values if (1) scientifically meaningful to do so and (2) you intended to fit a model over the smaller X range to begin with (and just forgot). You have to mention this in analysis write-up and be careful when making predictions to avoid extrapolation.
- NOT OK to drop point based on its outcome value (assuming no data errors in that value). These are legitimate observations. Dropping them is cheating by changing the data to fit the model.
- Try transformations or collect more data.
- Or just do nothing! It's okay to have some outliers.

The problem of multicollinearity

- Cannot include two variables with a perfect linear association as predictors in regression,
- Ex: Suppose the population line is Avg. $y = 3 + 4x$.
- Suppose we try to include x and $z = x/10$ as predictors in the model,

$$\text{Avg. } y = B_0 + B_1 x + B_2 z$$

and estimate all coefficients. Since $z = x/10$, we have

$$\text{Avg. } y = B_0 + B_1 x + B_2 x / 10 = B_0 + (B_1 + B_2/10)x$$

We could set B_1 and B_2 to ANY two numbers such that $B_1 + B_2/10 = 4$. The data cannot pick from the possible combinations.

Multicollinearity

- Exact same problem arises for any set of predictors such that one is an exact linear combination of the others.
- Ex: Consider a regression model with dummy variables for both males and females, plus an intercept

  Avg. y = B0 + B1 Male + B2 Female
  = B0(1) + B1 Male + B2 Female.
- Note that Male + Female = 1 for all cases. Hence, the intercept variable (always equal to 1) is a perfect linear combination of Male + Female.

Multicollinearity

- In real data, when we get "close" to perfect colinearities we see standard errors inflate, sometimes massively.
- When might we get close:
  - Very high correlations (> .9) among two (or more) predictors in modest sample sizes
  - When one (or more) variables is nearly a linear combination of the others
  - Including quadratic terms as predictors without first mean-centering the values before squaring
  - Including interactions involving continuous variables

How can we diagnose presence of multicollinearity?

- First step is to look at a correlation matrix of all the predictors (including dummy variables). Look for values near -1 or 1,
- If you are suspicious that some predictor is a near linear combination of others, run a regression of that predictor on all other predictors (not including Y) to see if R-squared is near 1.

We see multicollinearity…. so what?

- Multicollinearity only a problem if standard errors for the involved coefficients are too large for useful interpretation– and you care about interpreting those coefficients.
- Harris Saving Bank analysis:
  - Main coefficient of interest is the one for fsex.
  - Rest of variables are "control variables," i.e., variables that might be correlated with bsal and fsex whose effect we want to account for in the model.
  - I would keep age and experience in model, since I want to account for both variables and don't care to interpret age or experience coefficients.
- Another scenario is prediction: including highly correlated predictors can increase prediction uncertainty.

What can we do about multicollinearity?

- What if you do want to interpret the coefficients involved in the multicollinearity, and the SEs are inflated substantially because of it?
- Easiest remedy: remove one of the offending predictors. Keep one that is easiest to explain or that has largest T-statistic.
- Better remedy: Use a Bayesian regression model with an informative prior distribution (take STA 360!)
- Best remedy: go get more data! Multicollinearity tends to be unimportant in large sample sizes.

Which predictors should be in my model?

- This is a very hard question and one of intense statistical research.
- Different people have different opinions on how to answer the question.
- I will talk about the key issues rather than specific methods.
- See Sleuth for details on some specific methods.

What variables to include?

- Depends on the goal of the analysis.
- Goal: prediction
  - Include only variables that are strong predictors of the outcome.
  - Excluding irrelevant variables can reduce the widths of the prediction intervals.
- Goal: interpretation and association
  - Include all variables that you thought a priori were related to the outcome of interest, even if they are not statistically significant
  - Improves interpretation of coefficients of interest.

Common strategies

- Backward selection
  - Fit a large model with all variables of interest
  - Drop variables that are deemed irrelevant according to some criterion. Common ones include
    - Drop if p-value > .10 (possibly all at once)
    - Drop one, if any, that leads to smallest value of the AIC or BIC (see page 356 in Sleuth)
- Forward selection
  - Include variables one-by-one according to some criterion. Common ones include
    - Include one, if any, that leads to the smallest value of the AIC or BIC.

Common strategies

- Stepwise selection
  - Potentially do one forward step to enter a variable in the model, using some criterion to decide if it is worth entering the variable.
  - From the current model, potentially do one backwards step, using some criterion to decide if it is worth dropping one of the variables in the model.
  - Repeat these steps until the model does not change.

Challenges in model selection

- Using an automated procedure might inadvertently lead you to miss key transformations or interaction effects, since you will be tempted to press a button.
- May find scientifically vapid models, since relying on automation rather than science to guide model selection.
- May be many models with very similar values of the criterion you use in model selection.
- Very difficult to interpret standard errors, because really you should account for the randomness in the model selection procedure too.
- If a predictor is potentially important for interpretation, why not let the data estimate its coefficient rather than set it to zero?