So far from the statistics inference:

* single numerical variable
  + test for single mean, median. T-test, z-test, simulation
* single categorical variable
  + for 2 level categorical variable, test for single proportion, t-test or simulation
  + for more than 2 level categorical variable, chi-square goodness of fit or simulation.
* relationship between a numerical and a categorical variable
  + if categorical variable has two levels, t-test for two means OR t-test for paired mean.
  + if categorical variable has more than two levels, ANOVA
* relationship between two categorical variables
  + if both have 2 levels, test for difference in proportion.
  + if at least one variable has more than 2 levels, chi-square independent test.

Now

* relationship between two numerical variables

**Linear Regression**

Linear models can be used for prediction or to evaluate whether there is a linear relationship between two numerical variables.

Linear regression assumes that the relationship between two variables, x and y, can be modeled by a straight line:

are linear model parameters.

These parameters are estimated using data, and we write their point estimates as and .

**Residuals**

Leftovers from the model fit:

Data = fit + residual

Difference between the observed and predicted y

Each observation will have a residual. If an observation is above the regression line, then its residual, the vertical distance from the observation to the line, is positive. Observations below the line have negative residuals. One goal in picking the right linear model is for these residuals to be as small as possible.

Residuals are helpful in evaluating how well a linear model fits a data set. We often display them in a residual plot.

**Describe linear relationships with correlation**

Correlation defines linear association between two numerical variables.

The correlation is intended to quantify the strength of a linear trend. Nonlinear trends, even when strong, sometimes produce correlations that do not reflect the strength of the relationship.

Two numerical variables can have association but not linear association.

R, correlation coefficient, measures the strength of the linear association between two numerical variables.

Range from -1 to 1 with 0 indicates no linear relationship.

A white square with blue lines and black text

Description automatically generated

The correlation coefficient is unitless and is not affected by changes in the center or scale of either variable. Changing unit does NOT change the correlation coefficient.

The correlation of X with Y is the same as of Y with X. Swap X and Y do not change the correlation coefficient.

The correlation coefficient is sensitive to outliers. See below two plots.

A graph of a graph of a graph

Description automatically generated with medium confidence

A close to zero correlation itself may indicates several scenarios, for example, a weak linear relationship OR a strong association but not linear could both yield a close to zero correlation. Therefore, visualization like scatterplot can help identify the scenario and guide the appropriate approach for data modeling.

Formula

We have observations

**Least square line**

A measure for the best line

Option 1: minimize the sum of magnitudes (absolute values) of residuals.

Option 2: minimize the sum of squared residuals – least squares.

least squares line?

The following are three possible reasons to choose option 1 over option 2:

* It is the most commonly used method.
* Computing the line based on Criterion (7.10) is much easier by hand and in most statistical software.
* In many applications, a residual twice as large as another residual is more than twice as bad. For example, being off by 4 is usually more than twice as bad as being off by 2. Squaring the residuals accounts for this discrepancy.

**Assumptions for the least squares line**

* Linearity: the data should show a linear trend. If there is a nonlinear trend, an advanced regression method from another book or later course should be applied.
* Constant variability. The variability of points around the least squares line remains roughly constant.
* Independent observations. Be cautious about applying regression to time series data, which are sequential observations in time such as a stock price each day. Such data may have an underlying structure that should be considered in a model and analysis.
* Nearly normal residuals. Generally, the residuals must be nearly normal, when this is found to be unreasonable, it is usually because of outliers or concerns about influential points or important predictors, or interaction terms are missing from the model.

**What if assumptions do not meet? What are the alternative approaches?**

**What are the steps when building regression model? Do we first build one and use the output to verify some assumptions that we can not verify before model fitting?**

**For all the assumptions and potential violation, do we have different threshold when it comes to using regression for reference vs. prediction?**

**Finding the least squares line**

and are the parameters of the regression line.

As what we learned in statistics inference, the parameters are estimated using observed data.

, are point estimates of the parameters and for the given sample.

Using the sample data,

Given point is always on the least squares line.

**Interpreting regression line parameter estimates**

Interpreting parameters in a regression model is often one of the most important steps in the analysis.

If the given sample data are observational, we cannot interpret a causal connection between the explanatory variables and dependent variables, although the data shows a real association.

The meaning of the intercept may or may not have practical value. If there are no observations where x is near zero, then the intercept would just define the height of the line.

**Extrapolation**

Linear models can be used to approximate the relationship between two variables. However, these models have real limitations. Linear regression is simply a modeling framework. The truth is almost always much more complex than our simple line. For example, we do not know how the data outside of our limited window will behave.

Applying a model estimate to values outside of the realm of the original data is called extrapolation. Generally, a linear model is only an approximation of the real relationship between two variables. If we extrapolate, we are making an unreliable bet that the approximate linear relationship will be valid in places where it has not been analyzed.

**Using to describe the strength of a fit**

We evaluated the strength of the linear relationship between two variables earlier using the correlation, .

To evaluate the strength of a linear fit, we use . It describes the amount of variation in the response that is explained by the least squares line. In other words, by apply the linear model, the model reduces our uncertainty in predicting aid using explanatory variable(s). The variability in the residuals describes how much variation remains after using the model. In short, there was a reduction in the data’s variation by using information about the explanatory variables for predicting y using a linear model.

Categorical predictors with two levels

For categorical predictors with just two levels, the linearity assumption will always be satisfied. However, we must evaluate whether the residuals in each group are approximately normal and have approximately equal variance.

The interpretation

The estimated intercept is the value of the response variable for the first category ( the reference category). The estimated slope is the average change in the response variable between the two categories.

**Outliers in Regression**

Outliers are points that fall away from the cloud of points

**Types of outliers**

* Outliers that fall horizontally away from the center of the cloud but don’t influence the slope of the regression line are called leverage points.
* Outliers that actually influence the slope of the regression line are called influential points.
  + They are usually high leverage points – high leverage are points that fall horizontally away from the center of the cloud
  + If one of these high leverage points also change the slope of the line, we call it an influential point.
  + How to determine if a point is influential?
    - We fit the line without the potential influential points, if those point would have been unusually far from the line, we say a point is influential.
    - We visualize the regression line with and without the point, and ask: does the slope of the line change considerably?
  + The left plot has outliers – leverage point;
  + the right plot has outliers - the influential point
  + A data point can be high leverage point but not necessary influential point (left point), usually influential point appear to be high leverage point (point that fall horizontally away from the center of the cloud)

A graph of a line graph

Description automatically generated with medium confidence A screen shot of a graph

Description automatically generated

*A higher correlation may not indicate strong linear relationship, outliers(influential points) can significantly inflate R. so we would always want to visualize the data to confirm if higher R is truly strong linear relationship or due to existance of outliers.*

*Outliers can reduce R-squared if the remainder of the data show a strong relationship and there is only one or few points that are outside the trajectory of the regression line;*

*Outliers can also increase R-square if the remainder of the data do not show strong relationship (below example).*

A screenshot of a graph

Description automatically generated

*High leverage points (points farther from the canter of the data, horizontal) are more likely to be influential.*

*An influential point does not necessary change the form of relationship between the variables. For example, from linear to non-linear.*

*Existance of outliers(influential points) can dramastically change the model fit. See below plot.*

A graph with a red line and a blue line

Description automatically generated

Don’t ignore outliers without good reasons when fitting a final model; be cautious about using a categorical predictor (with 2 levels) when one of the levels has very few observations. When this happens, those few observations become influential points.

**Inference for linear regression**

Except using linear regression for prediction, we can use linear regression to do inference. For example, hypothesis testing for the significance of a predictor and confidence interval for the slope estimate. There are conditions for regression to be satisfied in order to do inference.

Testing for the slope – hypothesis test on the slope

Questions that can be answered in inference for regression – is the explanatory variable a significant predictor of the response variable?

: the explanatory variable is not a significant predictor of the response variable -> slope of the relationship is 0

: the explanatory variable is a significant predictor of the response variable -> slope of the relationship is different than 0

Use a t-statistic in inference for regression.

is the standard error of our estimates based on the data.

Why degree of freedom is n – 2?

Lose 1 df for each parameter estimated, and in linear regression we estimate 2 parameters, therefore, we lose 2 df.

Confidence interval of

Recap – inference for regression

We can do hypothesis test for the slop and conduct confidence interval for the slope

Conditions for inference

Always be aware of the type of data: random sample, non-random sample or population.

Statistical inference, and the resulting p-values, are meaningless when you already have population data.

If you have a sample that is not -random (biased), the results will be unreliable.

The ultimate goal is to have independent observations – and you know how to check for those by now.

**Variability partitioning**

So far: t-test as a way to evaluate the strength of evidence for a HT for the slope of relationship between x and y.

Alternative: we consider the variability in y explained by x, compared to the unexplained variability. , calculated by squared the correlation coefficient.

Now, partitioning the variability in y to explained and unexplained variability requires ANOVA.

We can get ANOVA type output for our regression model as well.

ANOVA output table offers an alternative way in calculating , besides square the correlation coefficient (R). , here we use sum of square without standardizing by degree freedom

**Why and when we would use ANOVA in regression model?**

ANOVA in regression test the overall significance of the regression model. Specifically, it tests whether there is a linear relationship between the response variable and at least one of the predictor variables. F-statistic caompre the variance explained by the model against the variance unexplained. Null hypothesis: all regression coefficients are equal to zero (no linear relationship). Alternative hypothesis: at least one coefficient is different from zero (indicating a relationship). An f-statistic and its corresponding p-value, indicating whether the regression model provides a better fit to the data than a model with no predictors.

**ANOVA for comparing means across groups**

ANOVA is used to test if there are statistically significant differences in the means of a response variable across multiple groups (categorical independent variable). Null hypothesis: all group means are equal; alternative hypothesis: at least one group mean is different. An F-statistic and p-value, indicating whether there is evidence to reject the null hypothesis of equal means.

**How the usage of ANOVA here differ from ANOVA for comparing means across multiple groups?**

1. Context:

In regression, ANOVA accesss the overall significance of the model, considering all predictors simultaneously.; in group comparison, ANOVA focuses on whether different levels of a single factor (or multiple factors in factorial ANOVA) effect the response variable.

1. Model structure:

Regression ANOVA deals with models that can include continuous predictors, categorical predictors, or both.; group comparison ANOVA typically involves one or more categorical predictors (factors) without considering continuous predictors

1. Interpretation

In regression, a significant F-test suggests that the model explains a significant portion of the variability in the response variable; in group comparison, a significant F-test indicates differences in group means, suggesting that the factor has an effect on the response variable.

1. Assumptions

Both methods assume homoscedasticity and independent observations; for comparing group means, the assumption of normality within each group is also crucial.

1. Formulas for variance partitioning

ANOVA in regression analysis

ANOVA in group comparison

is the j-th observations in i-th group.

In regression ANOVA, SSR measures how well the regression model explains the variability in the response variable. In ANOVA for groups, SSB measures the variability due to the differences in the means of different groups. SSE in both cases represents unexplained variability, but the contest differs – unexplained by the model in regression vs. within-group variability in group comparison)

In summary, ANOVA is a general framework used for variance partitioning, with the goal of determining if the way in which variance is partitioned is statistically significant.

ANOVA in regression and ANOVA for comparison of group means will yield the same results under specific conditions, particularly when analyzing the effect of a categorical variable on a continuous outcome.

The use of residual plots

The use of residual plots is to verify the multiple linear regression model assumption. Major assumption includes, normality of residuals, constant variance in Y or residuals, independence in residuals and linearity between Y and each Xs.

**Residuals vs. Fitted Values plot**:

* **The shape of the scatter plot could imply linear or non**-linear relationship between residuals and the fitted model.
* **The spread of the residuals along with the fitted value (fan shape) could imply constant variance in residuals**.