Linear Regression

Review of linear regression model

Terms

- X_1 - X_p are predictors. Typically non-random because they assume no variance (i.e., measured 100% exactly right).
- Y_i is response variable.
- e random variable with mean 0 that is the noise (error of other discrepancies) present in a model.

Basic

- If there are multiple quantitative predictors, the model is fitting a hyper-plane in multi-dimensional space.
- We can estimate p + 1 parameters where the "+1" is the intercept.

Least squares

- Least Squares is a method for estimating the unknown parameters β0,β1,…,βp where we minimize the sum of the squared errors (the differences between the estimated values Y^i and the actual values Yi in our *training* data Yi).
  - $$MSE = E(\hat{y}_i - y_i)^2$$
  - With one equation for each parameter, each equation is a restriction which reduces the MSE degrees of freedom to n−(p+1).

f test

- non-simmetrical distribution
- We also use the LINE assumptions to support the F-test which allows us to test multiple variables for significance at once (vs t-test which only lets you test 1)
- Can compare variance in two models (e.g., in a full + reduced model)

T-test

- Now that we have some estimates for our parameters, we need to check if the variables have predictive value in our model. One way to do that is with the t-test.
- you assume null hypothesis of 0 effect bc if we picked a number (e.g., 41) we would need to KNOW the distribution of this variable. But we already know the distribution if the variable is 0.
- normal distribution

Regression output

- the *Estimate* is the β^k for each parameter. This is the average increase in the response variable associated with a one unit increase in the predictor variable, assuming all other predictor variables are held constant.

- the *Std. Error* is the square root of the estimated variance for the parameter.

- the t *value* is the ratio of Estimat.eStd.Error

- Pr>|t| is the result from checking the t distribution with the degrees of freedom n−(p+1). Here that means df = 392−(4+1)=387 since there are four variables plus the slope parameter.

- **Residual standard error**: The average distance observed values are from the regression line. Smaller is better.

- **Multiple R-Squared**: (aka the coefficient of determination). This is the proportion of the variance in the Y that can be explained by the predictors.

- **Adjusted R-squared**: A version of R-squared based on a penalty for the number of predictors. Useful for comparing models with different numbers of variables.

- **F-statistic**: The ratio of MSR/MSE=SSR/pSSE/(n−(p+1).

  - Use this to test the value of the overall model based on a Null hypothesis of all βk=0, i.e., is there at least one variable with a βk≠0.

  - Note SSR is the sum of the squared difference between the y^i and y¯ or SSR=Σ(y^i−y¯)2.

- p-value: This is the p-value for the F-statistic with df of p and n−(p+1).

4 assumptions of linear regression

- **Linearity**: The response and predictor variables have a linear relationship.
- **Independence**: the predictor variables are independent of each other — no multicollinearity.
- **Normal**: The residuals have a Normal distribution with mean 0 and equal variance for all values of X.
- **Uncorrelated Errors**: The residuals are uncorrelated with each other.

Goodness of fit

- Goodness of Fit or Lack-of-Fit is a discussion about over-fitting and under-fitting.

  It is about testing if a linear model (the first LINE assumption) is a good fit to explain the relationship between Y and X as expressed in

  $(3.22)Y\longrightarrow=X\beta\longrightarrow+\epsilon\longrightarrow$

  In Linear Regression, the Null hypothesis $H0:\beta\longrightarrow=0$ asserts there is no relationship between Y and X versus an alternative of there is a linear trend or component to the relationship between Y and X.

- The Alternative model is a **Saturated Regression Model**, the **most** general model we can create, with **maximum flexibility**.

  We will estimate the saturated g(X) based on the values of X.

  - For a **single** value of xi, with **multiple** (k) responses yi,k, the most general model is $yi,k=g(xi)+\epsilon$
  - We can estimate g(xi) by using the average of the yi,k. This is also the least squares estimate, yxi^{-}.

MISC

iid = identically, independently distributed

for 3 level categorical variable: 1 predictor, but uses three degrees of freedom (since one level is reference category)

for Auto data analysis

- If you remove intercept, the estimates are based on if y = 0 (vs y = the intercept). Reduces a degree of freedom and hence making it under-fit the data.

You take the sum of squares of a model

- Then, look at that which can be explained by the model
- Then, look at what is left.
  - Of what is left, some proportion can be explained by a maximally flexible model. This is reducible error.
    - To do this, factor all variables (i.e., turn them all to categorical)
    - Using anova, regular model gets compared w/ factor model and will show that a nonlinear model is better than linear model. f test checks if random noise or can fit non-linear model.
  - Some variance cannot be explained (e.g., multiple y values for one point X) which is "irreducible error"........in those instances, the model estimates the mean of y (ybar).

        - NOTE: If there are no repeated values, just create some duplicate values e.g., divide by 100, round to integer, then multiply by 100.

Use the formula for a regression line** y=β^0+β^1x **to show, in the case of simple linear regression, the le

- I'll begin with an equation for a basic linear model:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{u}_i$$

- We can represent this as the sum of each value of y is equal to the sum of each intercept plus beta 1 times the sum of each value of x:

$$\sum_{i=1}^{n} y = n(\hat{\beta}_0) + \hat{\beta}_1 * \sum_{i=1}^{n} x$$

- From here, we can divide each side by n:

$$\frac{\sum_{i=1}^{n}}{n} y = \hat{\beta}_0 + \hat{\beta}_1 * \frac{\sum_{i=1}^{n} x}{n}$$

- Then, we can cancel out the summation from i=1 to i = n:

$$\frac{\sum_y}{n} = \hat{\beta}_0 + \hat{\beta}_1 * \frac{\sum_x}{n}$$

- Finally, we can cancel out summation/n, in which we are left with ybar is equal to beta0 + beta1 times xbar. This demonstrates that the least squares line passes through ybar and xbar.

$$\bar{y} = \hat{\beta}_0 + \hat{\beta_1}(\bar{x})$$

I collect a set of data ($n=100$ observations) containing a single predictor and a quantitative response. I then f

- Suppose the true relationship between X and Y is linear, i.e., Y=β0+β1X. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.
    - Adding new variables can only reduce the training SSerr, so we expect the training SSerr from the cubic regression model to be smaller.
- Answer (a) using test rather than training RSS.

    - Since the true relationship is linear, the cubic model is expected to provide a larger test SSerr. Using a less flexible linear model results in a lower variance. Also, the linear model is correct, so there is no bias. Therefore, the linear model has a smaller test MSE, and therefore, we expect a smaller test SSerr.

- Suppose the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.
    - Regardless of the true model, the additional terms X2and X3 will explain some additional portion of the total sum of squares SStot, and therefore, will reduce SSerr.
- ***Answer (c) using test rather than training RSS.***
    - Although estimates from a linear model will have a lower variance, they will be biased, because the linear model is incorrect. In this case, we cannot tell, from the given information, which model produces a smaller test SSerr. It depends on the balance of the variance and the squared bias

Example

Consider this regression table:

- `sales` = home price in 000s
- `finished` = square feet finished
- `bedroom` = numb of bedrooms
- `pool` = yes/no pool
- `size` = lot size

```
lm(sales ~ finished + bedroom + pool*size, data = home_prices)
```

- `Intercept` = -0.94e+01
- `Finished` = 1.6e-01
- `bedroom` = -8.6e+00
- `size` = 1.3e-03
- `pool` = 8.9e+01
- `size:pool` = -3.4e-03

Write out two equations, one for a house with a poo, and another for a house without a pool

- the first equation includes pool bc it is `1/0` dichotomous, and then the pool*size interactions because we need to take into account the smaller role of lot size on price
- The second equation does not include pool (bc it is technically `89 x 0`) or the interaction (also `x 0`)

1. sales = -9.4 + 0.16 (finished) - 8.6 (bedrooms) + 0.0013 (size) + 89 (pool) - 0.0034 (size)
2. sales = -9.4 + 0.16 (finished) - 8.6 (bedrooms) + 0.0013 (size)

Suppose the relation between the house price, area, lot size, and number of bedrooms actually does not dep

1. The training SSErr will always increase when you remove a variable from the model, so it will increase when pool is removed.
2. The testing SSErr may increase or decrease but it is expected to decrsae when we use the correct model. Therefore the model w/o pool is likely to be better because it represents the TRUE relationship.

(1) Homeowners are wondering if the sales price of their house is expected to increase if they build a pool. I

(1)

1. Set pool and no-pool equations equal to each other.

    1. -9.4 + 0.16 (finished) - 8.6 (bedrooms) + 0.0013 (size) + 89 (pool) - 0.0034 (size) = -9.4 + 0.16 (finished) - 8.6 (bedrooms) + 0.0013 (size)

2. Subtract one side from the other

    1. change pool/no pool = 89 - 0.0034(size)

3. Solve for breakeven point when change = 0

    1. Change equals 0 when size = 26,176

(2)

1. If the pool costs 10k, we want the change in pool/no pool > 10 to build one
2. Solve 10 = 89 - 0.00034 (size)

    1. size = about 22k. A smaller lot. The more it costs to build the pool, the smaller the lot needed for it to be profitable