**Regression Interview Questions**

Taken from Chaeffers Notes (Lecture 7 of stat. inference)

**Why use least squares to estimate parameters in regression?**

1. Maximum likelihood estimator (when errors are normal)
2. Analytically simple to derive B, V(B)
3. Gauss-Markov\_Thoerem -> LS estimates are BLUE (Best Linear Unbiased Estimator) (regardless of dist of errors)

**If predictor and response are interchanged, what effect on regression line?**

If y is resp: B = p \* sigma(y) / sigma(x), Β = /

If x is resp: B = p \* sigma(x) / sigma(y), Β = /

**What is the sum of residuals in OLS?**

0, since errors are orthogonal to all vectors in the 2D space spanned by 1 and x

**If by mistake, each observed pair is included twice in analysis. What effect is on B, t-stat, R^2 ?**

B, R^2 stay the same, but t-stat is inflated by sqrt(2), since t = sigma / sqrt(n)

**Residuals vs fitted plots. What should we see?**

Lack of pattern

**Why residuals vs fitted better than a scatter plot?**

In an n-dimensional regression, fitted vs resids is still one dimension

**If correlation between residuals and fitted = 0, model must be fine. IS this true?**

NO, by construction the correlation of y\_hat and errors should be zero.

**How to deal with heteroskestacity?**

Log transforms, or weighted least squares, weight by 1/sigma

**Why least squares no longer a good choice if errors were not normal (say t dist)?**

LS no longer the MLE

What is MLE?

**selects the set of values of the model parameters that maximizes the**[**likelihood function**](https://en.wikipedia.org/wiki/Likelihood_function)**.**

**Intuitively, this maximizes the "agreement" of the selected model with the observed data,**

From Quora

<https://www.quora.com/What-are-the-interview-questions-on-regression-modeling>

* What are the conditions for Omitted Variable Bias and how does it affect the coefficient estimates? Why? What are some fixes for OVB?

**omitted-variable bias (OVB) occurs when a model created incorrectly leaves out one or more important factors. The "bias" is created when the model compensates for the missing factor by over- or underestimating the effect of one of the other factors.**

* How do you interpret the coefficients in a log-log model? Why?
* What does the Gauss-Markov Theorem say and why is it important?

**the Gauss–Markov theorem - in a**[**linear regression model**](https://en.wikipedia.org/wiki/Linear_regression_model)**in which the errors have expectation zero and are**[**uncorrelated**](https://en.wikipedia.org/wiki/Uncorrelated)**and have equal**[**variances**](https://en.wikipedia.org/wiki/Variance)**, the best linear**[**unbiased**](https://en.wikipedia.org/wiki/Bias_of_an_estimator)[**estimator**](https://en.wikipedia.org/wiki/Estimator)**(BLUE) of the coefficients is given by the**[**ordinary least squares**](https://en.wikipedia.org/wiki/Ordinary_least_squares)**(OLS) estimator.**

**The errors do not need to be**[**normal**](https://en.wikipedia.org/wiki/Normal_distribution)**, nor do they need to be**[**independent and identically distributed**](https://en.wikipedia.org/wiki/Independent_and_identically_distributed)**(only**[**uncorrelated**](https://en.wikipedia.org/wiki/Uncorrelated)**with mean zero and**[**homoscedastic**](https://en.wikipedia.org/wiki/Homoscedastic)

* How does heteroscedasticity affect the coefficient estimates and why? What are some fixes for heteroscedasticity?

**This affects the model in 2 ways:**

1. **The OLS estimator is not efficient (it does not have minimum variance).**
2. **The estimators of the variances are biased. The standard errors reported on the SHAZAM output do not make any adjustment for the heteroskedasticity - so incorrect conclusions may be made if they are used in hypothesis tests.**

Fixes:

* **Weighted Least Squares – weights the observations proportional to the reciprocal of the error variance for that observation**
* Given a regression setting with a binary response variable, what probability model should be used and why?

Logistic regression, gives predictor as a percent chance

* What happens to the errors in the logistic regression function?
* Design a regression model to test the Law of Demand.
* Explain the Central Limit Theorem to a five year-old. (This one gave me the most trouble).
* Why do the residuals from a linear regression add up to 0?

**We fit the line such that the sum of all differences between our fitted values (which are on the regression line) and the actual values that are above the line is exactly equal to the sum of all differences between the regression line and all values below the line. Just as with the arithmetic mean: by constructing our fitted values in this way, it necessarily follows, by construction, that all deviations from that line must sum to zero for otherwise this just wouldnt be an OLS regression.**

* Is this still true if you fit a regression without intercept?
* What's so bad about collinearity?

**If there is collinearity, the design matrix X can become singular, meaning its non-invertible, leading to instability.**

* **We first regress Y on X1 and X2, then regress Y on X1 and Z, where Z=X1−X2. How are the coefficients in the two regressions related?**

**B1 = (B1\* + B2\*), B2 = - B2\***

* We regress Y on categorical data X1,…,Xp. This is a large data set (that's right, big data!) but many of the rows of the design matrix are duplicated. We  can summarize the data by averaging Y for each unique row of the design matrix, and perform a weighted regression with each row of the design matrix weighted by its number of duplicates. This gives us a much smaller data set to work with. How do the regression coefficients and their standard errors compare to regression on the raw data?

Correlation vs linear regression:

* The slope in  a linear regression gives the marginal change in output/target variable by changing the independent variable by unit distance. Correlation has no slope.
* The intercept in a linear regression gives the value of target variable if one of the input/independent variable is set zero. Correlation does not have this information.
* Linear regression can give you a prediction given all the input variables. Correlation analysis does not predict anything.

What if errors are not normally distributed?

Consider using **robust regression, which uses Huber loss function instead of least squares**

Logistic Regression

Logit( P(Y=1)) = X’B

Logit(p) = log odds = log(p / 1-p)

Ridge Regression

Minimize RSS + lamba times sum of Beta squared

Lasso

Minimize RSS + lamba times sum of abs(beta)

LINEAR ALGEBRA

What is trace of a matrix? Sum of diagonals

What is eigenvalue / eigen vector? Eigenvalues are a special set of scalars associated with a [linear system of equations](http://mathworld.wolfram.com/LinearSystemofEquations.html) (i.e., a [matrix equation](http://mathworld.wolfram.com/MatrixEquation.html)) that are sometimes also known as characteristic roots, characteristic values

What is a **GARCH** model?

Basically a EWMA (exponentially weighted moving average) model with an additional long term variance term.

V\_n = (1-a-b)\*w\_0 + B\*v\_n-1 + a\*(v\_n-1)\*(B\_n-2 ^ 2)

w\_o = long term variance

a+b < 1

B\_n is Brownian motion