ST552 Final Review

Nick Sun

March 16, 2019

State the assumptions required for making inferences in regression

- Linearity (is the underlying relationship of the model linear? $Y = X\beta$)
- Constant variance (errors ϵ_i all have the same variance σ^2)
- Independnece of the errors
- Normality, errors are normally distributed

Also, when dealing with matrix algebra we need our design matrix X to be of rank p

Rank the assumptions in rough order of importance

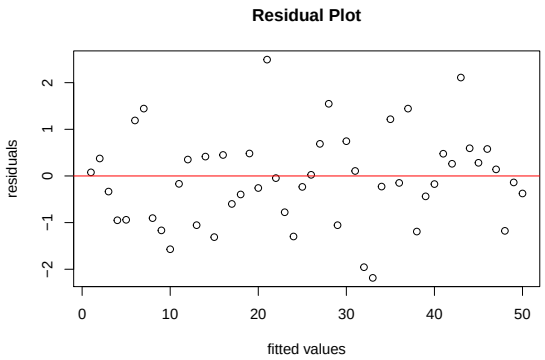
1) Linearity is the most important.
2) Independece of errors
3) Constant variance
4) Normality

Describe the consequences of violating a particular assumption

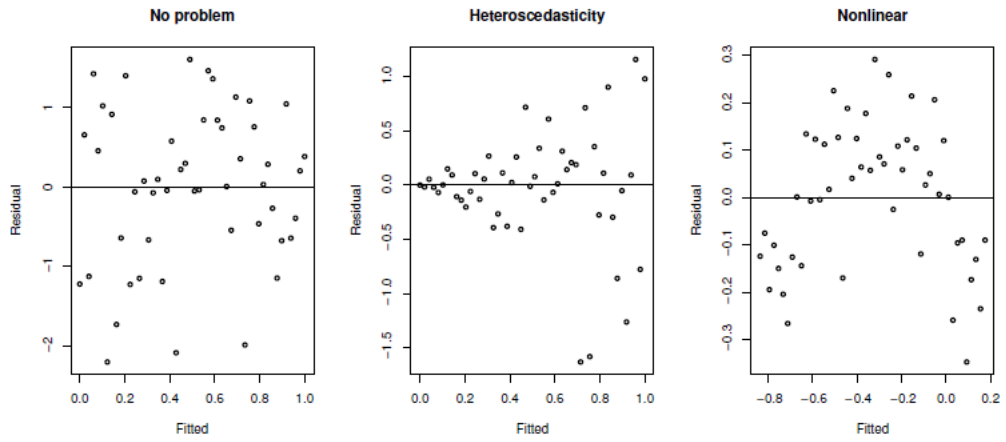
1) If linearity is violated, then our entire model might be meaningless! The parameter estimates will definitely be biased
2) If the errors are not independent, then the parameter estimates are probably fine, but the standard errors of those estimates are going to be incorrect. You will not be able to do inference on those estimates.
3) If constant variance if violated, the variance in predictions is probably not going to be correct.
4) If the errors are not normal, for large datasets we *usually* can get away with it. However, in general for non normal errors we should avoid making prediction intervals.

Sketch residual plots that should be examined to diagnose problems with regression assumptions

Residual plots that deviate from being a constant spread based around 0 should be examined. Basically anything that doesn't roughly look like this might be problem:



Sketch a residual plot that illustrates a violation of a particular assumption (non-linearity, non-constant variance, non-normality)



Check out fig. 1

A non-normal residual plot might appear skewed (more residuals on one side or more extreme residuals on one side).

Given a residual plot, describe evidence you see for violations of the regression assumptions

Should be easy to do, just match with the above.

Suggest a remedy for a particular violation

- For **nonlinearity**: We might want to try transforming the variables, transform the predictors, use splines.
- For **heteroskedastictiy**: You can try transforming the response (log, square root, inverse, Box-Cox, etc.). You can also try using weighted least squares.
- For **nonnormality**: Transform the response. We should try using robust regression. OLS estimates are still BLUE, but they might not be effective.

Describe three ways a point may be considered unusual

- A point can be far away in p-dimensional space from the other data points (high leverage)
- A point can substantially change the model when included or excluded in the model (influential)
- A point can just not fit the model well - their response and predictor variables are really weird (outlier)

Name three case influence statistics and describe conceptually how they can measure usefulness

Leverage

The leverage of an observation $h_i = H_{ii}$ where H is that hat matrix $X(X^TX)^{-1}X^T$ It's just the Mahalnobis distance. If a point has large leverage, it is far away from the mean of all the explanatory variables. Points with high leverage pull the regression line towards it.

Outliers

These points do not fit the regression line well, but this does not necessarily mean that they have high residuals since we try to minimize residuals in OLS regression. We often look at studentized residuals to find outliers which is calculated around

$$\hat{y_{(i)}} - y_i$$

where $\hat{y_{(i)}}$ is the fitted value for the i-th observation from a model fitted to the data excluding the i-th observation.

Influence

Infuence tries to measure how much a model fit changes when that observation is excluded. A big change means big influence. We often measure this using Cook's distance which is based around this quantity:

$$(\hat{y} - \hat{y_{(i)}})^T(\hat{y} - \hat{y_{(i)}})$$

which is essentially squaring the distance between the vector of predicted responses with and without point *i*.

Identify from a scatterplot if a point is likely to be high leverage, influential, or an outlier

We should be able to do this.

Describe a limitation of case influence statistics

Case influence statistics are susceptible to *masking* effects. There might be groups of points that are badly behaved, but because they are grouped together, their effects mask one another. Therefore, we might remove one point thinking we have solved the problem when really we have only removed one of several problematic points.

Describe what is meant by multicollinearity

Multicollinearity or collinearity is when predictor variables are highly correlated to one another.

Describe how multicollinearity might be detected

We can detect multicollinearity using:

- A correlation matrix of the predictors
- R_i^2 which is basically regressing the i-th predictor on the rest of the predictors (closely related to this is VIF which is $(1 - R_j^2)^{-1}$)
- Large condition numbers which are calculated from the eigenvalues λ_i of X^TX

$$\kappa = \sqrt{\frac{\lambda_1}{\lambda_p}}$$

Discuss the consequences of multicollinearity

While linear combination might make X^TX will not be an invertible matrix, closely correlated predictors might lead to imprecise estimates of β. Standard errrors will be inflated and the fit becomes very senstive to measurement errors. It is generally a pretty bad thing to have.

However, Faraway says the following about multicollinearity in relation to the goals of our prediction:

- If our goal is explanation and inference, we can drop collinear predictors since this might help us get better estimates with smaller standard errors for the predictors we actually care about. **However** we should be careful not to assume that the variables we drop are not related to the response. **If** we want to keep all variables in, we should use regularization techniques like ridge regression.
- If our goal is prediction, we *don't necessarily* have to do anything. It depends on if we are predicting for values of x_0 that are far from the rest of our data. In collinear data, this is an *especially* bad thing to do compared to data that is not close to singular or is orthogonal.

Describe the assumption that generalized least squares is designed to relax

In OLS, we generally use the assumption that $\text{var}(\epsilon) = \sigma^2 I$, but it's possible that instead we want to have $\text{var}(\epsilon) = \sigma^2\Sigma$ where Σ is a matrix that could represent some nonconstant variance or correlated errors.

Derive the generalized least squares estimates (for known Σ)

Our assumption here is that $var(\hat{\epsilon}) = \sigma^2\Sigma$ and that we can decompose Σ into SS^T where S is an upper triangular matrix.

Then the derivation for the generalized least squares estimates is just:

$$
\begin{aligned}
y &= X\beta + \epsilon \\
S^{-1}y &= S^{-1}X\beta + S^{-1}\epsilon \\
y' &= X'\beta + \epsilon' \\
&\text{Now we see that we can get regular old OLS...} \\
var(\epsilon') &= var(S^{-1}\epsilon) \\
&= S^{-1}(var(\epsilon))S^{-T} \\
&= S^{-1}\sigma^2 SS^T S^{-T} \\
&= \sigma^2 I
\end{aligned}
$$

Now we can calculate β

$$
\begin{aligned}
\hat{\beta} &= ((X')^TX')^{-1}(X')^TY' \\
&= ((S^{-1}X)^TS^{-1}X)^{-1}(S^{-1}X)^T(S^{-1}Y) \\
&= (X^T(S^{-1})^TS^{-1}X)^{-1}(X^T(S^{-1})^T)(S^{-1}Y) \\
&= (X^T(SS^T)^{-1}X)^{-1}(X^T(SS^T)^{-1}Y) \\
&= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y
\end{aligned}
$$

and variance of $\hat{\beta} = (X^T\Sigma^{-1}X)^{-1}\sigma^2$

Also notice that $var(\epsilon') = \sigma^2 I$ since:

$$\begin{aligned} var(\epsilon') &= var(S^{-1}\epsilon) \\ &= S^{-1} var(\epsilon)(S^{-1})^T \\ &= S^{-1}\sigma^2\Sigma(S^{-1})^T \\ &= S^{-1}\sigma^2 SS^T(S^{-1})^T \\ &= \sigma^2 S^{-1}SS^T(S^T)^{-1} \\ &= \sigma^2 I \end{aligned}$$

Give an example of data where using weighted least squares is desirable

For generaled least squares, a good example of correlated errors can arise in time series data. Another example would be where the observation are grouped in some way - for example, we have spatial data.

Weighted least squares is a special case of generalized least squares where the errors are uncorrelated but have unequal variance. In this, case Σ is a diagonal matrix of weights. This can happen when the observed responses are actually averages of a bunch of n_i observation (for example, that crawlin' tots data). Errors might also be proportional to some predictor value, for example we might see that there is a positive relationship between ϵ_i and some predictor.

Conduct a lack of fit test

A lack of fit test is based on a simple idea: compare $\hat{\sigma}^2$ to σ^2 to see if we are overfitting or underfitting our model.

Since we usually don't know σ though, we have to make a model free estimate of it. This is possible **if we have repeated values of the response** for one or more values of x. These repeated measures **cannot** just be on the same subjects - we need different subjects with the same predictor levels. This would give us an estimate of the between-subject variability. Values of x with no replication will be fit exactly and will not contribute to our model free estimate of σ^2.

If there are no replicates, we probably just have to resort to using graphical models.

Interpret the result of a lack of fit test

In `R` we can do a lack of fit test by fitting a model where the predictor is split into a factor variable so each unique value is treated as one group. In the end, this just boils down to doing an F-test between our proposed model and the model with the factor variable fit to it.

A low p-value here means that the pure error standard deviation of the factor model is substantially different than the $\hat{\sigma}^2$ in our proposed model. Therefore, there is a lack of fit and we should reconsider our model.

Note that if the null hypothesis is accepted here, that does not mean that we have our true model. We can only say that the data does not contradict our model.

Describe the goal of robust regression techniques

Robust regression is used when the errors fit some distribution that isn't normal. Short-tailed errors usually aren't a big deal, but long-tailed errors can have a big effect on OLS estimates. While it is sometimes fine to

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases}$$

Figure 1: Box Cox formula

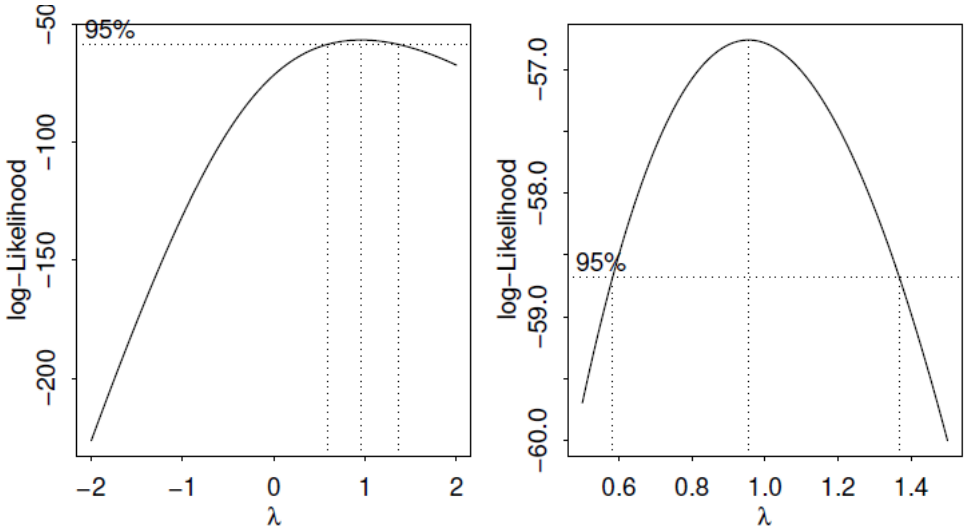


Figure 2: Box cox plot

just remove outliers and the like with Cook's D or something, robust regression is better if you are dealing with more than two outliers.

The basic idea of robust regression is that we use M-estimators to choose β. M-estimators are based on minimizing the sum across all our data points of a function $\rho(x)$ for $\rho(y_i - x_i\beta)$ where:

- $\rho(x) = x^2$ is just OLS
- $\rho(x) = |x|$ is called least absolute deviation regression
- and Huber's method which is basically a compromise between OLS and LAD

Describe why we might transform the response and/or the explanatory variables

If some of our assumptions like nonconstant error variance and non-linearity are violated, it might be possible to transform the predictor to make the relationship more linear or better satisfy our assumptions.

Choose a transform based on a Box-Cox plot

The Box-Cox is a popular method of determining which response transformation is best.

We often determine which λ will be best by using a graph where the x axis are various values of λ and the y-axis are the respective values of the log-likelihood. We want the λ values which maximizes the log-likelihood.

Interpret a parameter estimate based on a regression with a log transformed response

When we use a log transformation, the model is now *multiplicative*. A coefficient β_i on the original scale now means that the predicted response will be multiplied by e^{β_i}.

This is easy to see based on the fact that:

$$ln(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p y = e^{\beta_0} e^{\beta_1 x_1} \ldots e^{\beta_p x_p}$$

State the additional assumption required to make inferences about medians in a regression using a log transformed response

We need to assume that the response is symmetric on the new log-transformed scale.

Give a reason why variable selection might be recommended

If we have a large number of variables, we probably want to remove variables which only add noise to our model. Our model might perform better with a smaller selection of parameters. Additionally, it might be expensive to collect all this data - it would be cheaper to only need to collect a subset of that data.

Charlotte thinks of model selection as:

- A tool for finding predictive models
- A tool for exploratory data analysis

Give a reason why variable selection might be avoided

If our goal is prediction and a model with more parameters performs better in prediction, it might be better to use that model instead of a reduced model.

Also, if we are concerned with finding the explanatory effect of the predictors on the response variable, we probably not want to use variable selection since it will probably not capture the true relationship in the data.

Doing valid inference on a model *after* selection falls into post-selection inference and is an unsolved problem. Classical statistics doesn't hold up after we've done model selection since the inferences become conditional on having selected a particular model.

Describe the process of model selection by forward or backward elimination

Forward selection starts at an intercept model then add variables based on the lowest significant p-value. We continue until no more variables can be added.

Backward selection is the opposite. It starts at the full model and then removes variables based on the highest p-values. IT continues until we can't remove anymore variables.

Name and describe four model selection criteria

- Akaike Information Criterion (small is good)
- Bayesian Information Criterion (small is good)
- Mallow's Cp (small is good)
- Adjusted R^2 (large is good)

Discuss the similarities and differences between model selection criteria

AIC and BIC are both based around the quantity $nlog\left(\frac{RSS}{n}\right)$. They both have slightly different penalty parameters which penalize adding variables to the model.

Mallow's Cp is somewhat similar in that it also uses RSS, but it instead tries to estimate the model with the lowest mean square prediction error.

Adjusted R^2 is similar to regular R^2 but it penalizes adding predictors to the model since it is possible to artificially inflate R^2 by just sticking stuff in the model.

Discuss why it is dangerous to use the same data to fit a predictive model and evaluate a model's predictive ability

Since the model was trained on that data, it makes sense that it would predict that same data very well. If we were to calculate a mean square error, it would probably be overinflated.

Describe two regularized regression methods

Regularized regression methods are based around the idea that if we introduce some bias into our estimates, we can reduce the variance in our model.

Instead of minimizing

$$\sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

we instead minimize

$$\sum_{i=1}^{n}(y_i - \hat{y_i})^2 + \lambda \sum_{j=1}^{p} f(\beta_j)$$

There are two main kinds of regularized regression methods: Lasso and Ridge regression.

In Lasso, the second term has $f(x) = |x|$ and in ridge regression the second term has $f(x) = x^2$. Lasso estimates can be shrunk down to 0, so we can actually do model selection using Lasso regression.

According to Faraway: Ridge regression is useful in **collinearity situations** where we want to keep all th predictors we can. Since the typical scenario is that there are a lot of predictors which individually have a non-zero effect on the response, but not necessarily a large effect, this matches pretty well with ridge regression regularization method where we believe the regression coefficients will not be large.

Describe why might we prefer biased estimates

The bias-variance tradeoff! A biased estimate might have smaller variance than an unbiased one.

More complex models might decrease bias, but they generally increase variance. Increasing model complexity only improves performance up to a point.

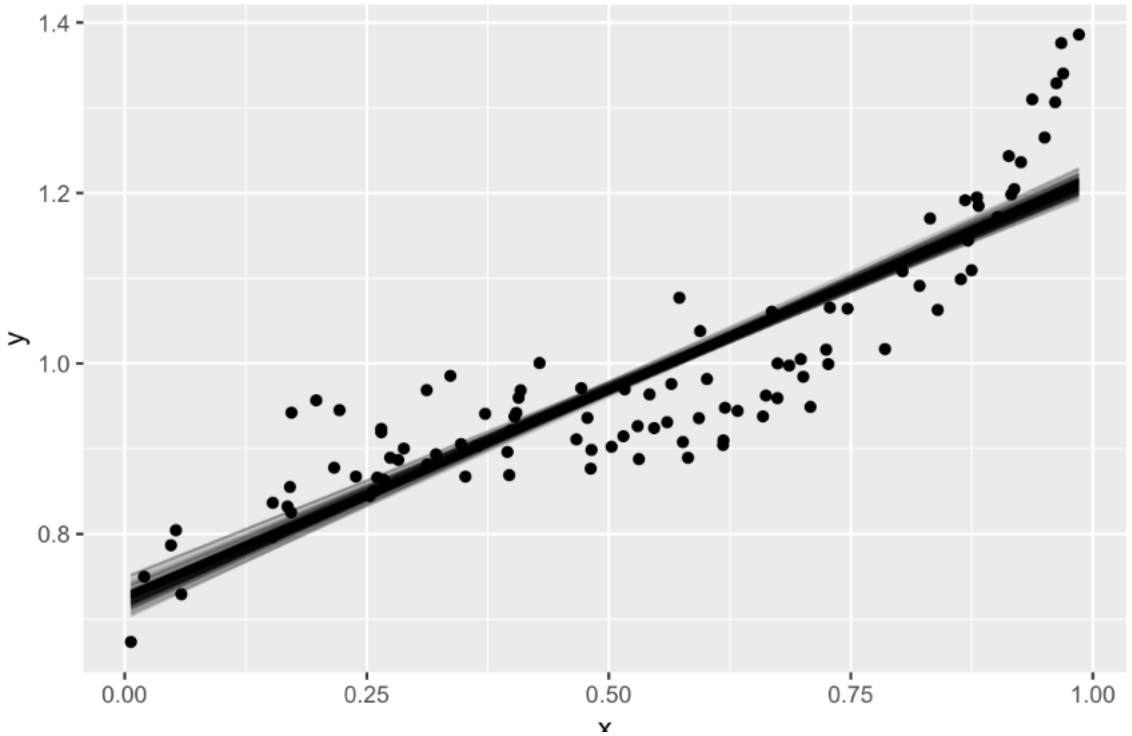


Figure 3: High bias, low variance

Describe/sketch an example of a predictive model that would have low/high variances and low/high bias

Bias captures how far away our prediction are from the true mean over repeated samples. Variance captures how much our prediction vary, again over repeated samples.

I think Charlotte Wickham's notes describe it best:

Discuss the differences in goals between explanation and prediction

In explanation, we want to test specific parameters and see if there is an actual relationship between those predictors and the response. It requires valid inference with all necessary assumptions.

In prediction, we just want to get a good predictive model. We don't necessarily care about having practically important or unimportant variables in the model so long as we are able to get a low predictive MSE.

Describe the difference between linear and logistic regression

Logistic regression is using a linear model to model a binary response (represented as either a 0 or 1).

The general model makes use of a logit transformation on the response:

$$logit(P(y)) = X\beta$$

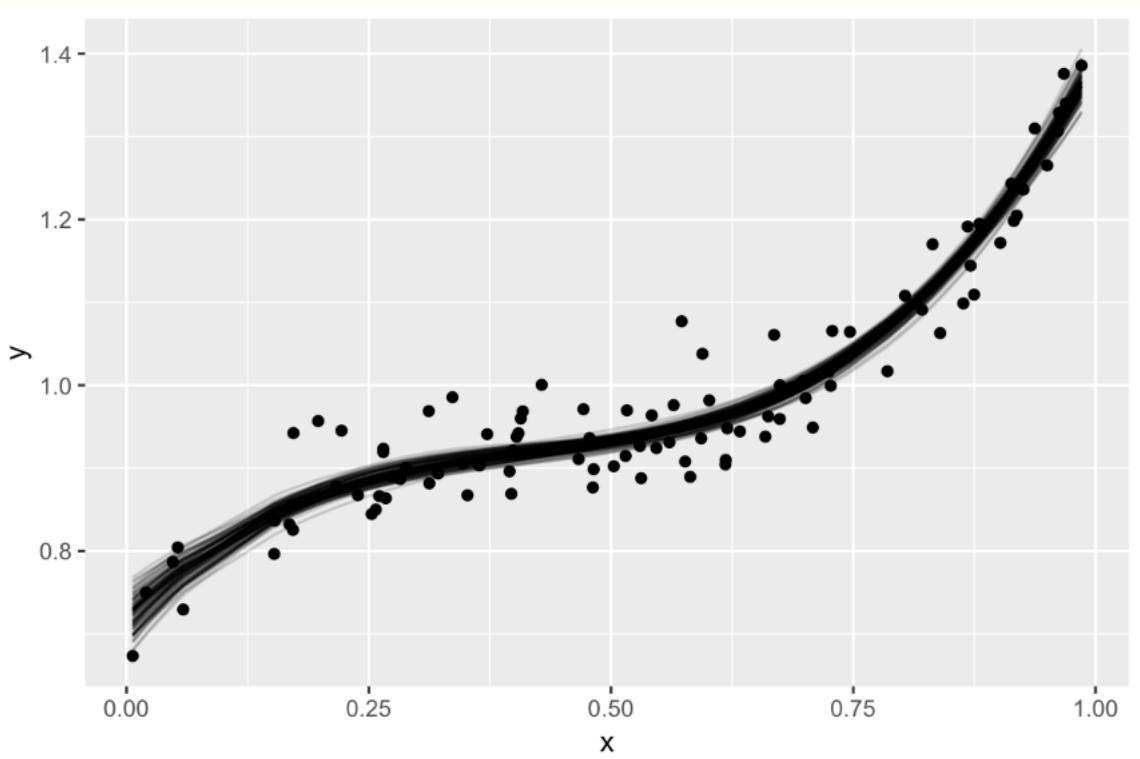


Figure 4: Low Bias Low variance

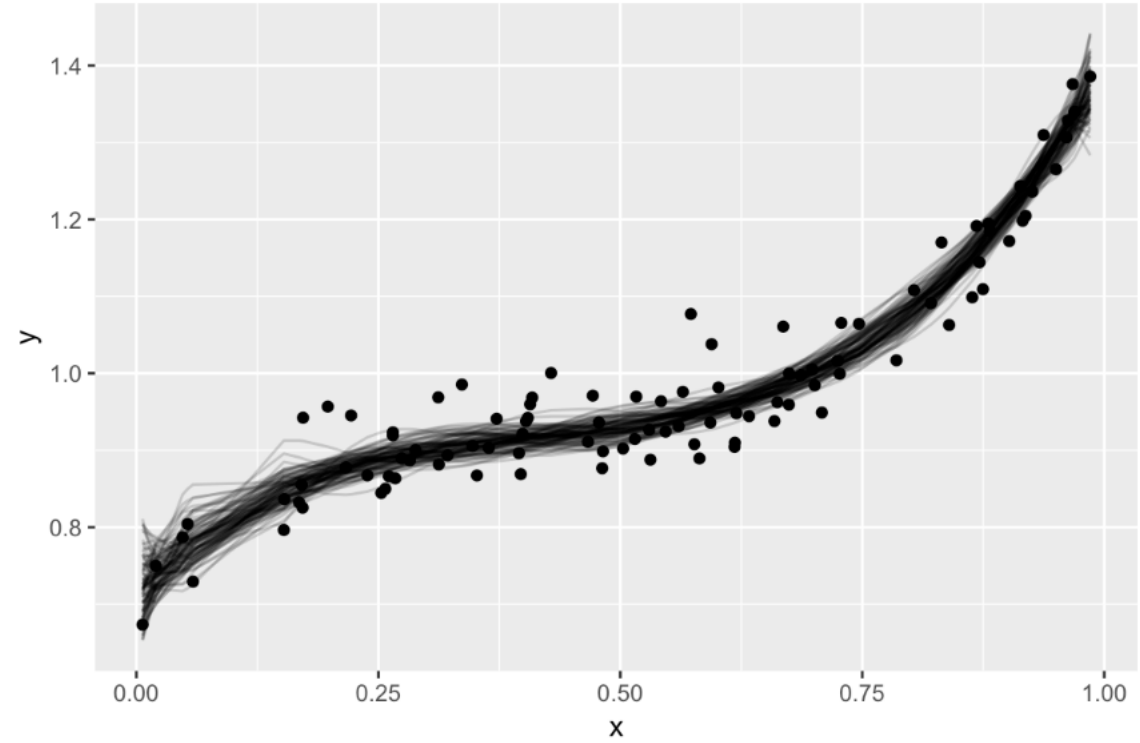


Figure 5: Low bias, high variance

Interpretation of logistic regression coefficients

The basic intuition is that a postive coefficient means that the explanatory increases probability and a negative coefficient decreases probability. To get more precise on the original scale though, we can use a couple techniques:

- Backtransform at or near the center of the data

Using the inverse logit function `1/(1 + exp(-x))`, we can plug our logistic regression output in and get a predicted probability $\hat{p}$. We can then substract our backtransformed model output at different values of our predictor.

- Divide by 4

The logistic function's derivative at its center is $\beta/4$, so we can interpret this as "at most, a one unit change in X is associated with an increase in probability of $\beta/4$"

- Odds Ratio

"A unit increase in x results in a β increase in the log odds ratio of the probability of a success"

This comes from

$$log\left(\frac{P(y=1|x)}{P(y=0|x)}\right) = \beta_0 + \beta_1 x$$

Describe the difference between linear and non-linear regression

Well the big difference is that linear regression is for linear relationships and nonlinear regression can be for any arbitrary smooth function.

The basic set up is that we have some function η s.t.

$$y_i = \eta(x_i, \beta) + \epsilon_i \epsilon_i \sim N(0, \sigma^2)$$

If $\eta = x^T\beta$ then we just have OLS. If η is something else, then we can represent it using basis functions.

Under normal errors, the MLE of β minimizes

$$\sum_{i=1}^{n}(y_i - \eta(x_i, \beta))^2$$

but there's no nice closed form solution for this, so we need to use iterative procedures and provide starting points from the data.

In the example from class, we knew the model was of the form: $y_t = \beta_0 + \beta_1 2^{-t/\theta} + \epsilon_t$ so we had to guess good starting values for β_0, β_1, θ.

We did this by figuring out what the practical interpretations were for each parameter and making reasonable guesses based on the graph that we were given.

We guessed $\beta_0 = 100, \beta_1 = 80, \theta = 100$ and that seemed to work fine. Picking different initial guesses can change the estimates so this step is pretty important.

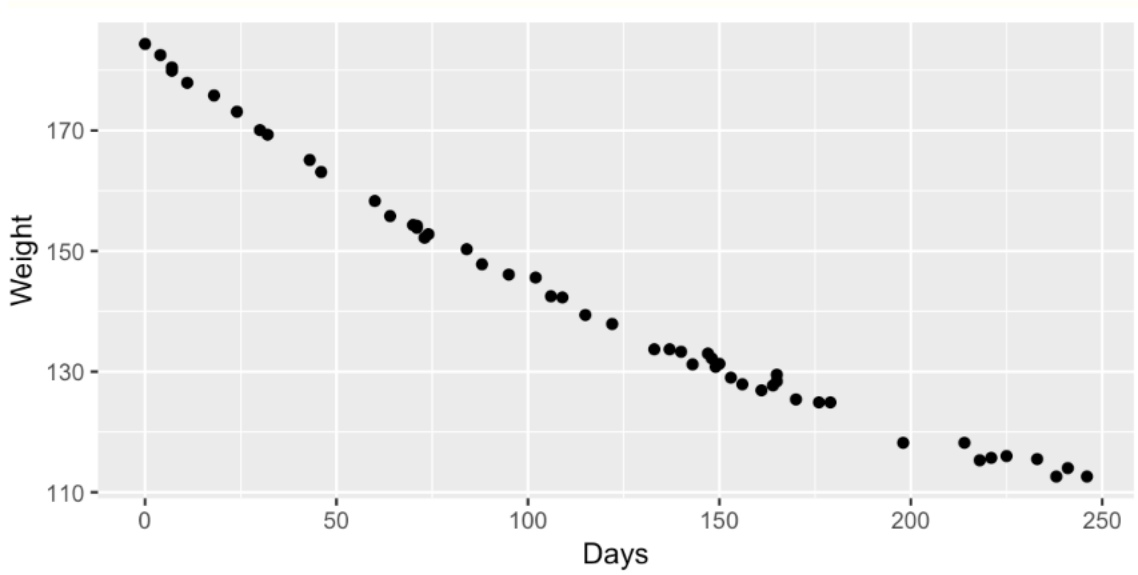


Figure 6: Nonlinear regression class example

Other good tidbits/formulas to know

Derivation of variance of β

$$\begin{aligned} var(\hat{\beta}) &= var((X^TX)^{-1}X^Ty) \\ &= (X^TX)^{-1}X^Tvar(y)((X^TX)^{-1}X^T)^T \\ &= (X^TX)^{-1}X^T\sigma^2I((X^TX)^{-1}X^T)^T \\ &= \sigma^2(X^TX)^{-1}X^T(X(X^TX)^{-1}) \\ &= \sigma^2(X^TX)^{-1}(X^TX)(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1} \end{aligned}$$

Prediction and Confidence intervals

For a particular value of x_0, the confidence interval is:

$$\hat{y_0} \pm t_{n-p}\hat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0}$$

and the prediction interval is:

$$\hat{y_0} \pm t_{n-p}\hat{\sigma}\sqrt{1 + x_0^T(X^TX)^{-1}x_0}$$

Since the MSE for a particular point is $Var(\hat{f}(x_0)) + Bias(\hat{f}(x_0)) + \sigma^2$

Also the confidence interval for a linear combination of parameters is

$$c^T\hat{\beta} = t_{n-p}\sqrt{\hat{\sigma}^2c^T(X^TX)^{-1}c}$$

Standard error of β_i

The formula for $SE(\beta_i) = \hat{\sigma^2}\sqrt{(X^T X)^{-1}_{i+1,i+1}}$