5.1: Logistic Regression

Dr. Bean - Stat 5100

1 Why Logistic Regression?

Recall the linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \epsilon \qquad (\epsilon \sim N(0, \sigma^2)).$$

What are some properties of the variable Y that are required for $\epsilon \sim N(0, \sigma^2)$.

- Y must be linearly related to $X_1, \ldots X_{p-1}$.
- Y must be a **continuous, quantitative** variable

1.1 Why not regression on categorical data?

Consider fitting a regression model where we use age to try and predict whether or not a person has a disease (a 0-1 variable).

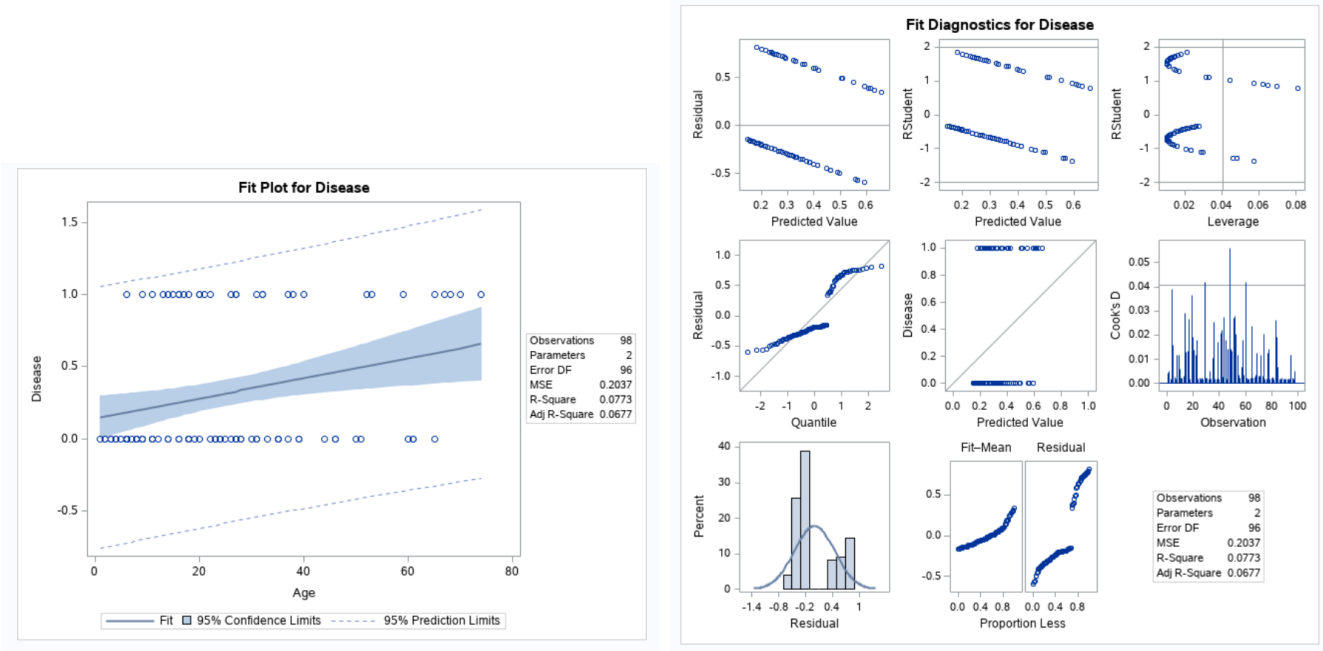


Figure 1: Fit plot and residual diagnostics for regression model that uses age to predict the presence/absence of a disease.

It is for this reason that instead of trying to predict the **value** of a categorical predictor, we should rather try to predict the **probability** of occurrence π_i,

$$\pi_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i \qquad (\epsilon \sim N(0, \sigma^2)). \tag{1}$$

However, based on the previous example, what are some of the issues with trying to predict the probability using (1)?

- We don't actually know π_i.
- Model can predict negative probabilities or probabilities above 1.
- Residual assumptions never satisfied (impossible for residuals to be normally distributed).

2 Transforming Probabilities

Because regression works best with **unconstrained** variables (i.e. variables that can theoretically take on any value). We need to find a transformation that maps $\pi \in [0, 1]$ to $f(\pi) \in (-\infty, \infty)$.

Solution: log-odds ratio.

- $\pi \to [0, 1]$
- $\frac{\pi}{1-\pi} \to [0, \infty)$
- $L = \log\left(\frac{\pi}{1-\pi}\right) \to (-\infty, \infty)$

The **probit** function is another common transformation that achieves similar results.

- Probit: $Q_i = Z_{\pi_i} \to$ Z score (of a standard normal distribution) associated with the percentile π_i.

Other "S" shape curves exist, which tend to reach similar conclusions.

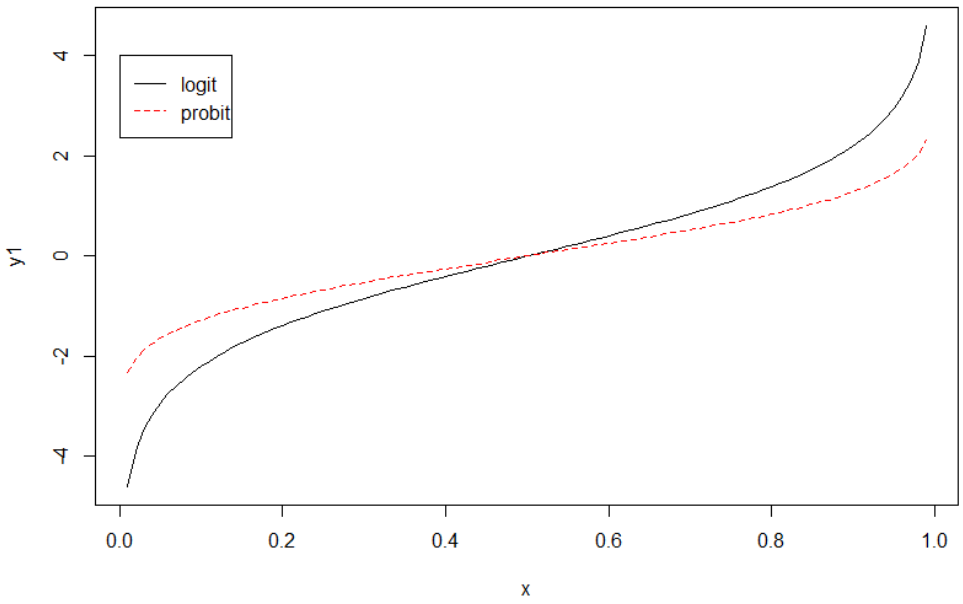


Figure 2: Visualization of logit and probit function for various probabilities.

3 Logistic Regression

$$L_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

- b_k estimates obtained from MLE-based iterative procedure (Newton-Raphson, Fisher)
- Transform estimates $\hat{L}_i = b_0 + b_1 X_{i,1} + \cdots + b_{p-1} X_{i,p-1}$ back to probability scale.

$$\hat{\pi}_i = \frac{1}{1 + e^{-\hat{L}_i}} \qquad \hat{Odds}_i = e^{\hat{L}_i}$$

3.1 Interpretation of Estimates

- $X_{i,1} = \cdots = X_{i,p-1} = 0 \implies \hat{L}_i = b_0 \implies \hat{Odds}_i = e^{b_0}$
- Hold $X_{i,2} = \cdots = X_{i,p-1} = 0$, increase $X_{i,1}$ from 0 to 1

$$\implies \hat{L}_i = b_0 + b_1 \implies \hat{Odds}_i = e^{b_0+b_1} = e^{b_0} e^{b_1}$$

- Thus, an increase in one unit in X_j *multiplies the odds* (in favor of $Y = 1$) by a factor of e^{b_j}.
  - Note that it is the *odds* that are multiplied, **not** the probability.
- Alternative Interpretation: the odds of $Y = 1$ change by $100(e^{b_j} - 1)\%$ per unit increase in X_j while holding other predictors constant.
  - Example (Handout 5.1.1): b_j for sector is 1.57 $\implies e^{1.57} = 4.83$.
  - "Holding all other predictors constant, the odds of having disease are $100(4.83 - 1) = 383\%$ greater in Sector 2 than in Sector 1.
- The "Odds Ratio" for X_j (odds of $Y = 1$ when $X_j + 1$ vs odds of $Y = 1$ when X_j)

$$\frac{e^{b_0+b_1X_1+\cdots+\mathbf{b_j}(\mathbf{X_j}+\mathbf{1})+\cdots+b_{p-1}X_{p-1}}}{e^{b_0+b_1X_1+\cdots+\mathbf{b_j}(\mathbf{X_j})+\cdots+b_{p-1}X_{p-1}}} = e^{b_j}$$

3.2 Inference with Estimates

- Single Variable Test:
  - $H_0 : \beta_j = 0$ (X_j has no effect on $P(Y = 1)$).
  - Test statistic: $t = \frac{b_j}{SE\{b_j\}}$ (standard normal for "large" N).
  - $\implies t^2 \sim \chi^2_1$ (obtain confidence intervals from here)
    - This approach is called the "Wald Test"
- Subset variables test:
  - $H_0 : \beta_{p-H} = \cdots = \beta_{p-1} = 0$
    - reorder the X variables so that the subset we are checking for comes last
  - Let L_{full} be the likelihood associated with the full model
  - Test statistics: $\chi^2 = -2 \log \frac{L_{red}}{L_{full}}$
  - Under $H_0 : \chi^2 \sim \chi^2_H$

- Overall model test:

$$\text{Model}\chi^2 = -2\log L_{intercept} + 2\log L_{int\&covariates}$$

  - Often called the **deviance**, DEV or $DEV(X_0, X_1, X_{p-1})$

- Conditional Effect plot: predicted $\hat{\pi}$ vs one predictor X_j

  - While holding all other predictors at some constant level. The default level in SAS is the mean (average) of each variable.

4 Goodness of Fit Measures:

- Pseudo R-square: $\frac{\chi^2}{\chi^2+n}$ (χ^2 from model test)
- Hosmer-Lemeshow Goodness of Fit Test
  - H_0 : logistic regression response function is appropriate
  - Based on sorted $\hat{\pi}$ values, group observations into 5-10 roughly equal sized groups.
  - Within each group, look at the total observed numbers of $Y = 1$ and $Y = 0$
  - Based on the model fit, calculate the total *expected* numbers of $Y = 1$ and $Y = 0$.
  - Test statistic χ^2 is sum (across groups) of $\frac{(observed-expected)^2}{expected}$
- "Concordance" - look at all pairs of observations with different Y
  - Let n_c be the # of "concordant" pairs (observed $Y = 1$ has larger $\hat{\pi}$)
  - Let n_d be the # of "discordant" pairs (observed $Y = 1$ has smaller $\hat{\pi}$)
  - Let n_t be the # of "tied" paired (observed $Y = 1$ and $Y = 0$ have same $\hat{\pi}$ (likely due to identical X-profiles)
  - Define rank correlation indices (larger is better):

$$\text{Somers' } D = \frac{n_c - n_d}{n_c + n_d + n_t}$$

$$\gamma = \frac{n_c - n_d}{n_c + n_d}$$

$$\text{Tau-a} = \frac{n_c - n_d}{0.5(n-1)n}$$

$$\text{AUC} = \frac{n_c + 0.5n_t}{n_c + n_d + n_t}$$

- ROC (Receiver Operating Characteristic) Curve
  - Sort all observations from the smallest to biggest $\hat{\pi}$.
  - At each position in the list:
    - Use $\hat{\pi}$ as threshold for $\hat{Y} = 1$, moving cutoff from the standard 0.5 threshold.

* Calculate sensitivity: (proportion $Y_i = 1$ values with $\hat{Y}_i = 1$).
* Calculate specificity: (proportion $Y = 0$ values with $\hat{Y} = 0$).
    - Sensitivity and Specificity - think smoke alarms.
* False positive rate (prop $Y = 0$ values with $\hat{Y} = 1$) = 1 - specificity
* Plot false positives against true positive rates (sensitivity)
* Calculate the area under the curve.

Given the three ROC curves in Figure 3, which model has the best predictive power and why?

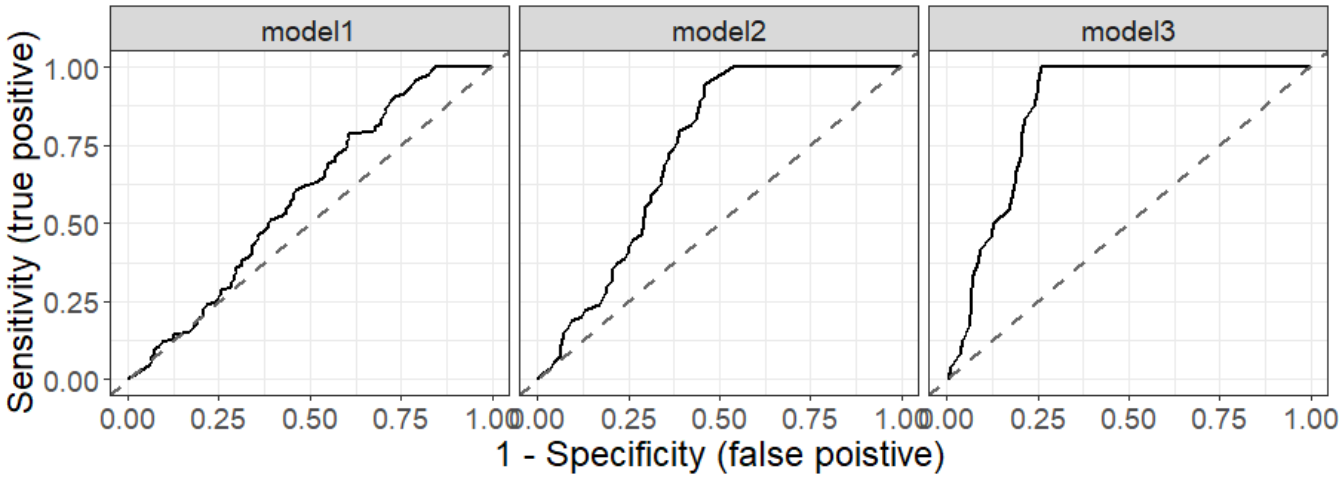


Figure 3: Comparison of three ROC curves.

Model 3 is the most accurate. The model sensitivity increases much faster than the false positive rate.

5 Multicollinearity

Recall that multicollinearity occurs when X variables are highly correlated with each other. It has **nothing** to do with the response variable Y.

As with OLS, multicollinearity inflates the variance of the b_k estimates, making them hard to interpret/test for significance.

As with OLS, stepwise selection and all possible regression methods exist to "score" each combination of explanatory variables and select a best model.

6 Outliers in Logistic Regression

If Y can only take on two values (0 or 1), how are outlier values possible?

An outlier is a point for which the observation strongly disagrees with the predicted probability.

- Define "deviance residual" as

$$dev_i = \text{sign}(Y_i - \hat{\pi}_i)\sqrt{-2\left(Y_i \log \hat{\pi}_i + (1 - Y_i)\log(1 - \hat{\pi}_i)\right)}$$

  - The more certain we are (probability near 0 or 1), the more potential we have to be very wrong.
- $DEV(X_0, \cdots X_{p-1}) = \sum_i dev_i^2$
- "Outliers" are values not well represented by the model
- "Half-normal probability plot - observed $|dev_i|$ vs expected value under normality
  - **However,** since the residuals are not normally distributed, we asses differences from our expectation using simulations based on $\hat{\pi}_i$.
    - Create 19 simulations by generating a "new" response variable where the values of $Y_{new,i} \sim Bernoulli(\hat{\pi}_i)$
  - Simulated envelop (SEE 5.1.1 MACRO ON CANVAS) plots the minimum, maximum, and mean of the 19 simulations
    - Why 19 simulations? - Since our observed deviances represent the 20th observation, the probability that our deviances will fall outside the envelope is less than 5% IF the fitted model is appropriate.
    - Points falling outside in the envelop in the upper right corner of the plot are evidence of outliers/bad fits.

7 Influential Observations

Influential observations have the same effect on model coefficients as they did in OLS.

Diagnostics (similar to Leverage and DFBETAS):

- $\Delta D_i : DEV - DEV_{(i)}$
  - Measures decrease in "misfit" when obs. i is ignored. (essentially measures the "poorness of fit for observation i).
  - "large" $\Delta D_i \implies$ obs. i overly influences model fit
  - SAS: DIFDEV - one step difference in deviance
- ΔB_i
  - Similar to Cook's distance, measures influence of obs. i on the estimates b_j
  - SAS: C - confidence interval displacement C
- $\Delta\chi_i^2$
  - Similar to ΔD_i: "poorness of fit" for obs i
  - SAS: DIFCHISQ - one step difference in Pearson χ^2

Unlike in OLS, there is no consistent numerical rule of thumbs to determine thresholds for the Δ measures.

Instead, we will simply rely on graphical diagnostics.

- $\Delta D_i, \Delta B_i, \Delta X_i^2$ vs Observation Number - look for extreme values
- ΔD_i vs $\hat{\pi}_i$ (or ΔX_i^2 vs $\hat{\pi}_i$)
  - Look for points with low $\hat{\pi}$ but $Y_i = 1$ (upper left corner) OR high $\hat{\pi}$ but $Y = 0$ (upper right corner) which are much different than the overall pattern
  - (Optional) plot different size points where point size is determined by ΔB_i

8 Remedial Measures

Similar to OLS:

- Look for typos in the data
- Consider transformations of the X variables
- Consider dropping problematic points (only if you have a good argument for removing them).

9 Final Thought

If you have a lot of explanatory variables, you should strongly consider classification trees and random forest for classification.

5.2. Nominal/Ordinal Logistic Regression

Dr. Bean - Stat 5100

1 What to do when your categorical data isn't binary?

Recall binary response logistic regression:

- Model framework:

$$Y \in \{0,1\} \qquad \pi = P(Y=1)$$

$$L = \log\frac{\pi}{1-\pi} \quad = \quad \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1}X_{i,p-1}$$

- logit link function:

$$L_i \;=\; \log\frac{P(Y=1|profile_i)}{P(Y=0|profile_i)}$$

Two kinds of multi-class categorical data

- Nominal: no apparent ordering of classes (ex: race, major, color)
- Ordinal: ordering of data makes sense (product rating, pain scale, etc.)

2 Nominal Logistic Regression

- pick one level as reference (say $Y = r$)
- generalized logit (glogit) link function:

$$L_{k|i} \;=\; \log\frac{P(Y=k|profile_i)}{P(Y=r|profile_i)}$$

- coefficient $\beta_{j,k}$ for the (marginal) effect of predictor X_j for $Y = k$ vs. $Y = r$:

$$L_{k|i} \;=\; \beta_{0,k} + \beta_{1,k}X_{i,1} + \ldots + \beta_{p-1,k}X_{i,p-1}$$

- odds ratio interpretation involves the base class r
  - 5.2.1 Example: Coefficient associated with A3 and Importance of 3:
    "Holding all other predictors constant, the odds that a 40+ year old rates AC and power steering as 'very important' *versus not important* are $100(e^{2.9165} - 1) = 1747\%$ greater than for an 18-23 year old."

Other Comparisons

To compute the log odds ratio for two *non-base* classes simply compute:

$$L_{k_1|k_2} = L_{k_1|i} - L_{k_2|i}$$

The estimated probability of each (non-base) class can be computed as

$$\hat{\pi}_k = \frac{e^{L_{k|i}}}{1 + \sum_{j=1}^{J-1} e^{L_{j|i}}}$$

(Note that $\hat{\pi}_i$ will be fully determined from the estimated probabilities of the other classes.)

3 Ordinal Logistic Regression

- $Y \in \{1, 2, \ldots, r\}$ and $1 < 2 < \ldots < r$
- accumulate probability over lower levels:

$$p_k^c = P(Y \leq k)$$

- logit function accounts for this accumulation ("proportional odds" model):

$$\begin{aligned} L_{k|i} &= \log \frac{p_k^c}{1 - p_k^c} \\ &= \log \frac{P(Y \leq k | profile_i)}{P(Y > k | profile_i)} \end{aligned}$$

- coefficient $\beta_{j,k}$ for the (marginal) effect of predictor X_j for $Y \leq k$ vs. $Y > k$:

$$L_{k|i} = \beta_{0,k} + \beta_{1,k} X_{1,i} + \ldots + \beta_{p-1,k} X_{i,p-1}$$

- odds ratio interpretation involves direction of k:
  - "Holding all other predictors constant, the odds that a 40+ year old rates AC and power steering as either important or very important are $100(e^{2.2322} - 1) = 832\%$ greater than for an 18-23 year old."
- In ordinal logistic regression, coefficient interpretation relies on direction in Y (higher or lower) because we assume the coefficient is the same for all levels of Y:
  - Let $\beta_{j,k}$ be coeff. for predictor X_j in model for $L_{k|i}$

$$L_{k|i} = \beta_{0,k} + \beta_{1,k} X_{1,i} + \ldots + \beta_{p-1,k} X_{i,p-1}$$

$$\begin{aligned} H_0 &: \beta_{j,1} = \beta_{j,2} = \ldots = \beta_{j,r} \\ H_0 &: L_{k|i} = \beta_{0,k} + \beta_1 X_{1,i} + \ldots + \beta_{p-1} X_{i,p-1} \end{aligned}$$

6.1: Introduction to Time Series

Dr. Bean - Stat 5100

1 Why Time Series?

Recall our basic multiple linear regression model:

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \ldots + \beta_{p-1} X_{t,p-1} + \varepsilon_t \qquad (t \text{ index for time})$$
$$\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Previous diagnostics focused on normality and constant variance, but not so much on *independence*.

Violations of independence sometimes detected by **patterns** in residuals over time.

This dependency is often due to auto-correlation ("self-correlation"), which is when the residuals are correlated *with each other*.

Hard to check if we don't know the order in which the data are collected.

What are some examples where you would expect the residuals of a linear model to be auto-correlated over time?

- House prices in Utah (population grows over time, drives prices up)
- Stock prices
- Temperatures

1.1 Autocorrelation, what's the big deal?

- If a random variable is autocorrelated over time, then observations closer in time will tend to be more similar than observations far away in time.
- Thus, repeated samples of the variable *in* time will have **less** variability within the sample than the variability *across* time.
- This means we will **underestimate** the true variance of the random variable, which in OLS causes
    1. The estimates regression coefficients are unbiased, but no longer "best" (i.e. minimum variance)
    2. MSE will underestimate the true residual variance
    3. OLS may also underestimate $s\{b_k\}$, which makes the t-tests unreliable (i.e. destroys inference)

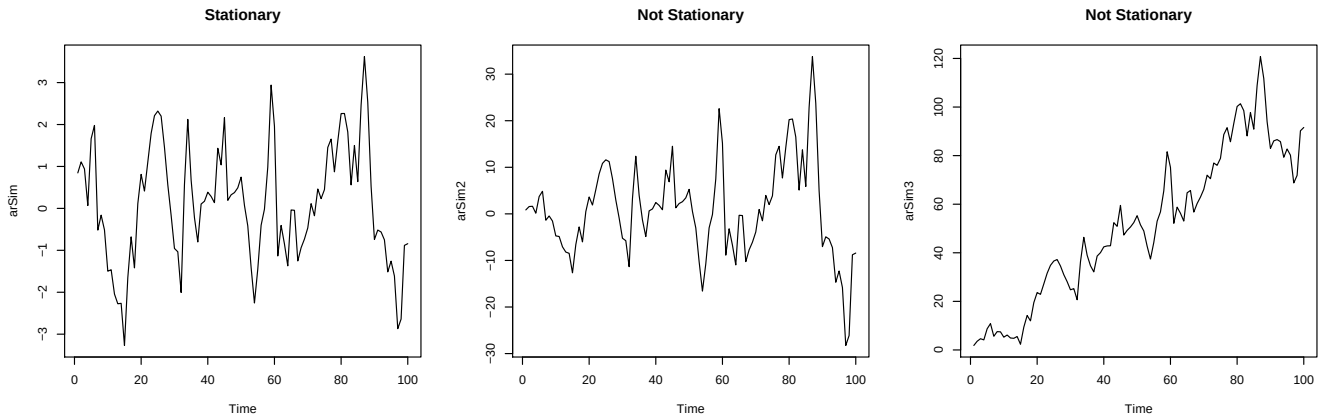


Figure 1: Examples of stationary and non-stationary time series.

2 Time Series Modeling

- autocorrelation means our data contain *structure* over time.
- Accounting for this structure should improve our ability to predict.
- One approach: Box-Jenkins (ARIMA) time series modeling:
  1. Make data stationary
  2. Test for independence
  3. Use sample autocorrelation and sample partial autocorrelation plots to identify potential dependence structures
  4. Fit dependence structures and asses model adequacy
  5. Using adequate model
     - Forecase response variable (w/ confidence interval)
     - Test model terms (incl. predictor variables)

1. **Make data stationary:**

- First Order (constant mean):

$$E[\epsilon_t] = \mu_t \equiv \mu \text{ for all } t$$

- Second Order (constant variance):

$$Var[\epsilon_t] = \sigma_t^2 \equiv \sigma^2 \text{ for all } t$$

- This means that if both conditions are satisfied, the time series will "look" the same no matter what time window (with appropriate scale) that we look at.
  - Graphical check: plot residuals e_t vs t (see Figure 1):
  - SAC (sample autocorrelation; ACF) plot - coming up, a useful diagnostic for stationarity
- Remedial Measures for Non-Stationarity
  - Non constant variance $\rightarrow$ transform Y_t

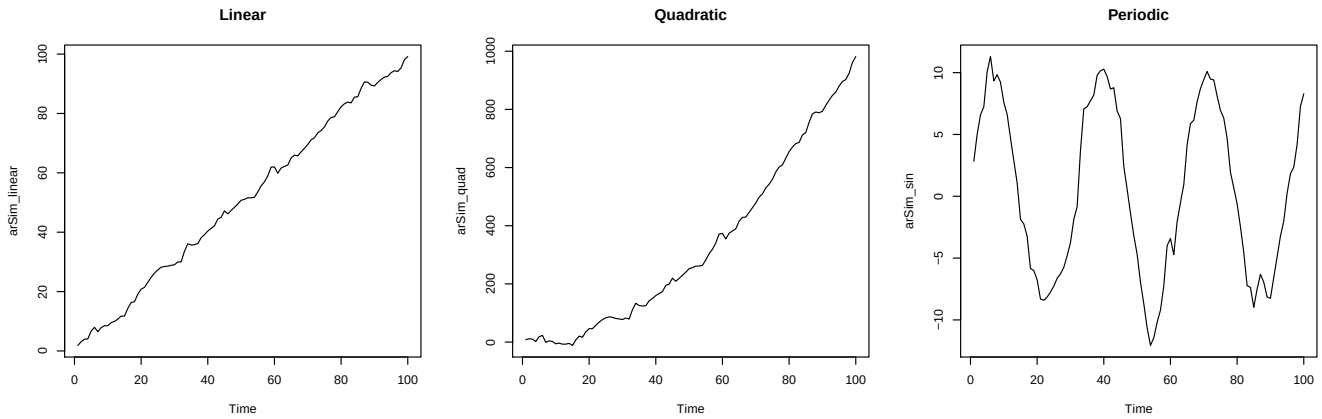


Figure 2: Examples of different trends that may occur in a time series.

- Non-constant mean: "de-trend" the data using a predictive model where time is the explanatory variable.
  - Use a scatter-plot of time vs residuals to determine an appropriate model (see Figure 2).
- "Differencing" for stubborn trends:
  - First differences: $Z_t = Y_t - Y_{t-1}, \quad (t = 2, \ldots, n)$
  - Second differences: $W_t = Z_t - Z_{t-1} = Y_t - 2Y_{t-1} + Y_{t-2} \quad (t = 3, \ldots n)$
- HOWEVER, differencing will make periodic cycles unrecoverable, which can hurt our ability to make forecasts.
- For this reason, differencing is a remedial measure of last resort.

2. **Test for independence**

   There is a difference in a series being a function of time (plus random noise) versus a series that is *correlated* in time.

   Failing to remove time-dependent trends in our data ruins our ability to check for time-dependent correlations, why is this?

   Points will vary together above and below the overall-average, making them look correlated, when they are actually varying randomly about the trend.

   AFTER removing trends, determine if the data are just "white noise" (no dependence structure)
   H_0 : Data are just white noise

   in SAS: χ^2 test for lags 1 through k, (where k is selected by the user).

3. **Identify tentative dependence structures**

   - Notation: Z_t is the stationary time series after "transforming" (including estimating out time trends and other covariates) the original time series $Y_1, \ldots, Y_n$

   - Sample autocorrelation function (ACF or SACF)

     r_m = linear association (correlation) between time series observations separated by a lag of m time units

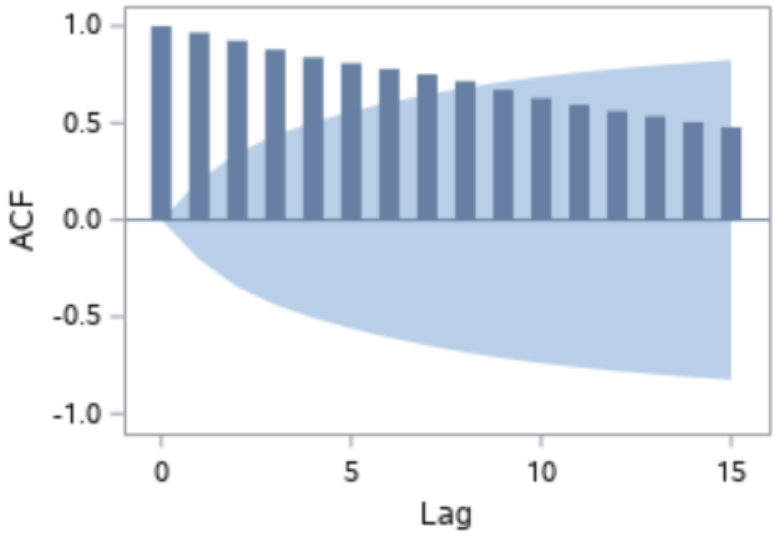


Figure 3: Sample ACF plot for a non-stationary time series.

- PLOT 1: sample autocorrelation plot (or SAC / ACF): check for stationarity and identify tentative dependence structure
  - bar-plot r_m vs. m for various lags m
  - lines often added to represent 2 SE's (rough significance threshold)
  - SAC / ACF terminology:
    - "spike" : r_m is "significant"
    - "cuts off" : no "significant" spikes after r_m
    - "dies down" : decreases in "steady fashion"
  - If Z_t stationary, SAC either cuts off fairly quickly or dies down fairly quickly (sometimes in "damped exponential" fashion)
  - If SAC dies down extremely slowly, Z_t nonstationary (see Figure 3)

- Sample partial autocorrelation function (PACF or SPACF)

$$r_{m,m} = \text{autocorrelation of time series observations separated by a lag of } m \text{ with the effects of the intervening observations eliminated}$$

- PLOT 2: sample partial autocorrelation plot (or SPAC / PACF)
  - bar-plot $r_{m,m}$ vs. m for various lags m
  - lines often added to represent 2 SE's (rough significance threshold)

- Main dependence structures

(a) AR(p) dependence structure: autoregressive process of order p:
  - current time series value depends on past values; common representation for AR(p):

$$Z_t = \delta + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + a_t$$

* ϕ_i are unknown parameters; random shock a_t iid $N(0, \sigma^2)$

– identify using SPAC: first p terms of SPAC will be non-zero, then drop to zero (sketch)

(b) MA(q) dependence structure: moving average process of order q:

– current time series value depends on previous random shocks

– model:

$$Z_t = \delta + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

Z_t : stationary "transformed" time series $\quad$ θ_i : unknown parameters

a_t : random shocks $\quad$ δ : unknown parameter

– identify using SAC: first q terms of SAC will be non-zero, then drop to zero (sketch)

Common Dependence Structures for Stationary Time Series

	SAC	SPAC
MA(1)	cuts off after lag 1	dies down, dominated by damped exponential decay
MA(2)	cuts off after lag 2	dies down, in mixture of damped exp. decay & sine waves
AR(1)	dies down in damped exponential decay	cuts off after lag 1
AR(2)	dies down, in mixture of damped exp. decay & sine waves	cuts off after lag 2
ARMA(1,1)	dies down in damped exp. decay	dies down in damped exp. decay

ARIMA(p,d,q) dependence structure: Autoregressive **Integrated** Moving Average Model

- a very flexible family of models ⇒ useful prediction
- recall first difference: $Z_t = Y_t - Y_{t-1}$, $t = 2, \ldots, n$
  and second difference: $W_t = Z_t - Z_{t-1} = Y_t - 2Y_{t-1} + Y_{t-2}, t = 3, \ldots, n$
- after differencing, AR and MA dependence structures may exist: ARIMA(p, **d** ,q)
  – p : AR(p) – value at time t depends on previous p values)

  - d : # of differences (need to take d^{th} difference to make stationary)
  - q : MA(q) – value at time t depends on previous q random shocks)
- use SAC and SPAC to select p and q – but how to select d?
  - usually look at plots of time series
  - choose lowest d to make stationary (also SAC)
- sometimes see backshift notation: $BY_t = Y_{t-1}$
  - $d = 1$: $Z_t = Y_t - Y_{t-1} = Y_t - BY_t = (1 - B)Y_t$
  - general d: $Z_t = (1 - B)^d Y_t$

- "Fit model" $\rightarrow$ estimates & standard errors for β_j's, ϕ_l's, & θ_l's

- Several approaches exist to estimate ϕ_l's, θ_l's, and β_j's, and deal with initial lag; we'll use ULS (unconditional least squares) for MA(q) & AR(p)

- ARIMA(p,d,q) model rewritten, with $t = 1, \ldots, n$:

$$\begin{aligned} Y_t &= g_1(Y_1, \ldots, Y_{t-1}) + g_2(X_{t,1}, \ldots, X_{t,k-1}) + g_3(a_1, \ldots, a_t) \\ &\text{where} \\ g_1 &= \text{linear combination (LC) of previous observations} \\ g_2 &= \text{LC of predictors at time } t\text{, in terms of parameters } \beta_j \\ g_3 &= \text{function of random shocks in terms of parameters } \phi_l \text{ \& } \theta_l \end{aligned}$$

g_1 differencing
g_2 linear model with predictors

4. **Fit dependence structures and assess model adequacy**

- General SAS code for ARIMA(p,d,q), Y in terms of $X_1, \ldots, X_{k-1}$:

```
proc arima data = a1;
 identify var = Y (d )  crosscorr = (X1 ... Xk-1) ;
 estimate p = p  q = q  input = (X1 ... Xk-1)  method = uls  plot;
 forecast lead = L alpha = a noprint out = fout;
run;
```

option	description
d, p, q	differencing, AR, & MA settings (as before)
`plot`	adds RSAC & RSPAC plots
L	# times after last observed to forecast
a	set confidence limit; a = .10 $\Rightarrow$ 90% conf. limits
`noprint`	optional, suppresses output
`out = fout`	optional, sends forecast data to `fout` data set

- Useful diagnostics for "goodness of fit":
  - Numerical

* Standard Error – measure of "overall fit"; in SAS: `Std Error Estimate`

$$S = \sqrt{\frac{\sum_1^n (Y_t - \hat{Y}_t)^2}{n - n_p}}, \qquad n_p = \#\text{ parameters in model}$$

Note that S is similar to $\sqrt{\text{MSE}}$

* Ljung-Box statistic Q^* (& p-value);
in SAS: lag 6 χ^2 for `Autocorrelation Check of Residuals`
  - basic idea: look at "local" dependence among residuals in first few sample autocorrelations
  - under H_0: "model is adequate", $Q^* \sim \chi^2_{df}$

– Graphical (PLOTS 3 and 4) – focus on residuals
* Residual sample autocorrelation plot (RSAC)
* Residual sample partial autocorrelation plot (RSPAC)

How will we know from these plots if we "succeeded"?

We should see no significant autocorrelations, which would suggest that we have fully accounted for the time dependent structure in the data.

ANALOGY: Mining - we are trying to extract information from data, and unaccounted structure is like knowing we left gold in the ground.

5. **Using adequate model:**

(a) forecast response (*** w/ conf. interval ***) – careful far beyond data

(b) test model terms (incl. predictor variables, but also AR & MA parameters)

7.1: Generalized Additive Models (GAM)

Dr. Bean - Stat 5100

1 Why GAMs?

Up to this point we have assumed models of the form

$$E\left(Y|X_1, X_2, \ldots, X_{p-1}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}$$

which quickly fail if the effects cannot be modeled linearly (with possible variable transformations).

As with LOESS regression, GAMs do not assume a particular model form. GAMs only assume that the effects of each variable are additive (no interactions) i.e.

$$E\left(Y|X_1, X_2, \ldots, X_{p-1}\right) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_{p-1}(X_{p-1})$$

where f_k is some unknown function that relates X_k to Y.

In practice, f_k are approximated using **cubic smoothing splines**, though any smoothing strategy will do. Splines are preferred because they are computationally much more efficient than weighted regression (as is used in LOESS). The key is that the smoothing occurs *over one dimension at a time* after accounting for the effects of the other dimensions.

2 Fitting GAMs

2.1 Splines

Splines are a series of piecewise polynomials that is continuous and differentiable at the break points, called **knots** (see Figure 1).

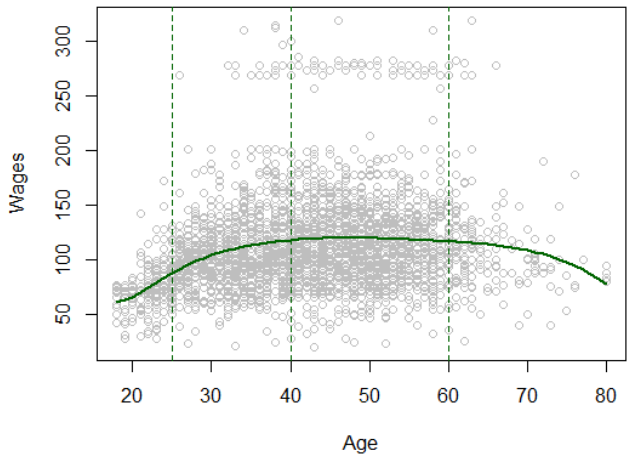


Figure 1: Example of cubic spline with three knots. Code to create image taken from `https://datascienceplus.com/cubic-and-smoothing-splines-in-r/`.

The more knots, the more "wiggly" your data has the chance to become.

Smoothing Splines place a knot at every unique value of X. This will most certainly lead to overfitting, so the splines are constrained by an *effective* degrees of freedom to prevent this. The effective degrees of freedom is often selected via **cross validation**.

The best part about smoothing splines is that it eliminates the need to select the placement of the knots.

2.2 Fitting GAMs

Remember that we have a unique smoothing spline that relates each X to Y after accounting for the effects of all other X's.
Hastie et al. (2002) outlines the algorithm for fitting GAMs

1. Set $b_0 = \bar{y}$ and all $\hat{f}_j \equiv 0$.
2. Cycle through all possible values for j each time updating predictions as:
   - Fit a smoothing spline S_j to the points (x_{ij}, y_j) after accounting for the effect of all other explanatory variables i.e.
   $$\hat{f}_j \leftarrow S_j \left[\left\{ y_i - b_0 - \sum_{k \neq j} \hat{f}_k(x_{ik}) \right\}_1^N \right]$$
   - To account for machine imprecision, recenter the splines around 0 i.e.
   $$\hat{f}_j \leftarrow \hat{f}_j - \frac{1}{N} \sum_{i=1}^{N} \hat{f}_j(x_{ij})$$
3. Repeat Step 2 until the predicted values of $\hat{f}_j$ stop changing (or change very little).

3 Comparisons to Other Methods

3.1 LOESS

- GAMs using splines are much more computationally efficient than LOESS.
  - Splines fit models to "chunks" of data, while LOESS uses weighted least squares on moving neighborhoods of data.
  - The sliding window requires continually recalculating distances between points in LOESS, which does not scale well to large datasets.
- GAMs fit smoothing curves to each variable individually, while LOESS fits one weighted regression surface to all explanatory variables at the same time.
- LOESS is great for one to two dimensional smooths on moderate to small datasets. GAMs are better suited for high dimensional data on large datasets.

3.2 OLS

- Both GAMs and OLS are easy to explain to other people (GAM effects are easy to visualize).
- GAMs do not require linearity like OLS.
- GAMs more computationally expensive to fit compared to OLS.
- GAMs can suffer from instability if the data is sparse at the endpoints.

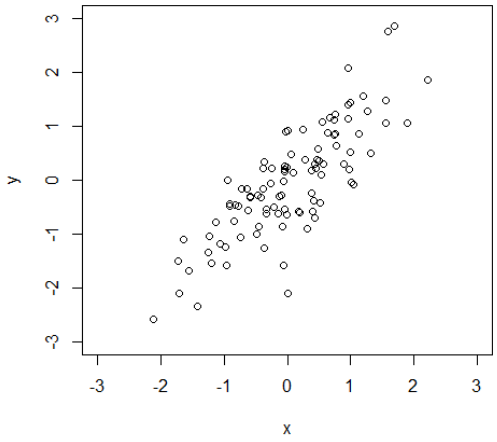
4 Good Resources

- Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome (2001) "The Elements of Statistical Learning" (Chapter 9) `https://web.stanford.edu/~hastie/Papers/ESLII.pdf`

7.2: Principal Components and Quantile Regression

Dr. Bean - Stat 5100

1 Principal Components (PC) Regression



2 Principal Components (PC) Regression

Principal Components is essentially a **re-projection** of the data into a new space where each axis follows the direction of the **highest variance** of the data, in descending order.

- Each component is a linear combination of the original X variables.

$$\begin{aligned} PC_{i,1} &= a_1^{(1)} X_{i,1} + \cdots + a_{p-1}^{(1)} X_{i,p-1} \\ PC_{i,2} &= a_1^{(2)} X_{i,1} + \cdots + a_{p-1}^{(2)} X_{i,p-1} \\ &\vdots \\ PC_{i,p-1} &= a_1^{(p-1)} X_{i,1} + \cdots + a_{p-1}^{(p-1)} X_{i,p-1} \end{aligned}$$

- Components derived from eigenvalues/eigenvectors of the matrix $X^T X$
- Often used as a form of dimensionality reduction.

Nice Mathematical Properties:

- $\rho(PC_j, PC_k) = 0 \quad \forall j \neq k$
- $\sum_j \left(a_j^{(k)}\right)^2 = 1$
- $Var(PC_1) \geq Var(PC_2) \geq \cdots \geq Var(PC_{p-1})$

2.1 Model

$$Y_i = \beta_0 + \beta_1 PC_{i,1} + \cdots + \beta_{p-1} PC_{i,p-1} + \epsilon_i$$

estimated with

$$\hat{Y} = \beta_0 + b_1 PC_{i,1} + \cdots + b_{p-1} PC_{i,p-1}.$$

Note: The PC's only consider relationships among the X variables and do not depend on Y.

Consequently, *all* of the principal components should be considered, not just the first few.

Why might dropping "low variance" principal components hurt our regression model?

Because the principal components have nothing to do with Y, its possible that a "low variance" combination of the X-variables is highly correlated with Y.

2.2 Pros and Cons

- **Pro:** Guaranteed uncorrelated predictors → no multicollinearity → meaningful model *coefficients.*
- **Con:** No guarantee of meaningful *variables.*

3 Quantile Regression

Quantiles: a set of q ranges for which there is an equal probability of an observation falling into each range.

- Example: the median splits the observations into two groups, where 50% of the observations fall in each group.

In ordinary least squares (OLS) our goal was to model the **mean** of the response variable Y.

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

with

$$\hat{Y} = b_0 + b_1 X_1 + \cdots + b_{p-1} X_{p-1}$$

based on the assumption that the model residuals were unbiased, normally distributed, with constant variance.

If the variance was not constant across Y, we were forced to consider variable transformations (Handout 2.2) or weighted least squares regression (Handout 4.2).

In quantile regression, heterskedasticity is seen as an **opportunity** to be pursued, rather than a **problem** to be fixed.

3.1 Motivating Example

Looking at Figure 1, how does the relationship between a Husband's education level and income change *across the quantiles of income*?

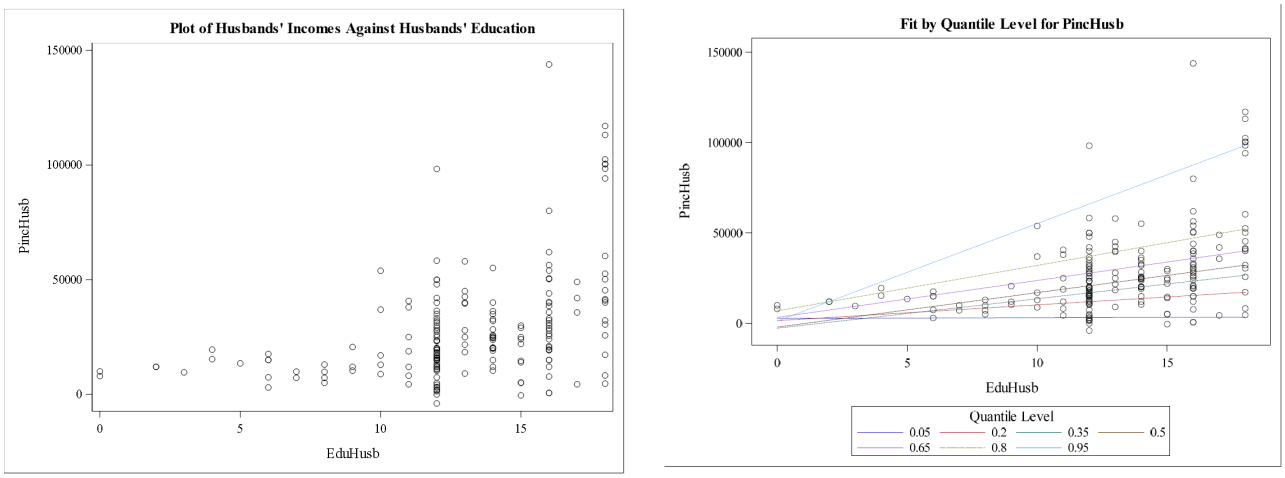


Figure 1: (Left) Plot of education vs income level. (Right) Series of quantile regression lines overlaid on the scatterplot of education and income.

Notice that the positive association between education and income level is much more drastic when comparing only high income earners in each educational group.

Simply trying to model effect of education on *average* income doesn't tell the full story.

3.2 Model

In ordinary least squares (OLS) regression, we assume the model

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$$

where estimated coefficients b_j were selected to minimize $\sum_i \left(Y_i - \hat{Y}_i\right)^2$.

In contrast, Quantile Regression selects $b_k(\tau)$ to minimize $\sum_i \rho_\tau \left(Y_i - Q_\tau(Y_i)\right)$, where

$$Q_\tau(Y_i) = b_0(\tau) + b_1(\tau) X_{i,1} + \cdots + b_{p-1}(\tau) X_{i,p-1}$$

τ	a quantile from 0 to 1
$b_j(\tau)$	estimated coefficient (a function of the quantile)
$Q_\tau(Y_i)$	the estimated τ quantile for the X-profile $X_{i,1}, \ldots, X_{i,p-1}$
$\rho_\tau(r)$	a "check loss" function $\rho_\tau(r) = \max\{\tau r, (\tau - 1)r\}$

For the check loss function in Figure 2, X is the value of the residual and Y is the penalty associated with the residual. Knowing this, how do the "penalties" for $\tau = 0.3$ and $\tau = 0.9$ differ? Why does this difference seem reasonable?

For $\tau = 0.9$ an over-prediction is not penalized nearly as much as an under-prediction. The opposite is true for $\tau = 0.3$. This makes sense because we would expect most of the points to be above the $\tau = 0.3$ line, and below the $\tau = 0.9$ line.

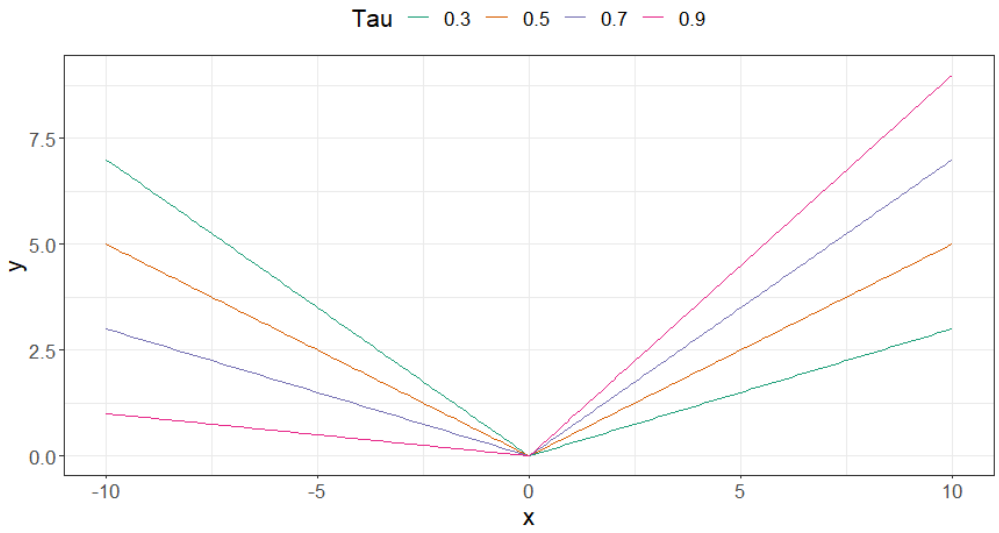


Figure 2: Check loss functions for various quantiles.

3.3 Comparison to OLS

OLS	**Quantile Regression**
Predicts conditional mean $E(Y\|X_1, X_2, \ldots)$	Predicts conditional *distribution* (via quantiles)
Error terms must meet distributional assumptions	No assumptions for error terms
Sensitive to outliers	Robust to outliers (except extreme quantiles)
Effectively applied to small samples	Data-hungry
Computationally cheap	Computationally expensive (no closed-form solution/requires lots of quantile models).

3.4 Good Resources

- Rodriguez, Bob and Yao, Yonggang (2017) "Five Things You Should Know About Quantile Regression" `https://support.sas.com/resources/papers/proceedings17/SAS0525-2017.pdf`
- Rodriguez, Bob (2018) "Three Things you Should Know About Quantile Regression" `https://www.youtube.com/watch?v=CU0ofd3hSOA`