STATS 210P

Lecture 8

Sevan Koko Gulesserian

University of California, Irvine

Simple Linear Regression: Assumptions

- So far we have learned how to formulate a linear model with explanatory variables X's and used it to predict or explain the response variable Y.

- The simple linear model was of the form $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where ε_i are the independent and identical normally distributed errors.

- We also learned how to estimate the parameters of the model, β_0 and β_1, and other useful quantities such as R-squared, Mean squared error (MSE), and the t^* needed to conduct hypothesis test with the linear model parameters.

- We also extended the simple linear model to the multiple linear model, where we have p many explanatory variables (and thus p many beta, β, coefficients).

Simple Linear Regression: Assumptions

The following assumptions are needed to confirm the validity of using the linear regression model to obtain inference of the association between a continuos quantitative response variable Y and a single or several explanatory variables X's.

- 1. Linearity: There is a linear relationship between the X's and Y variable. In 2 dimensions, fitting a straight line through the data points is appropriate (in multiple dimensions it is a hyper-plane).
- 2. The variance (or standard deviation) of the Y values given the X's is constant for all values of X's in the range of the data.
- 3. Independence: The errors are independent of each other. That is to say knowing the value of one does not help with knowing the value of the others.
- 4. The errors are normally distributed.
- 5. If we want to extend the results to the population, need the sample to be randomly selected from the population.

Simple Linear Regression: Assumptions

- We now come to learn about conducting model assumption checks.
- The checks for now will be done using objective measures on graphical summaries of the model.
- More formal tests will be discussed in the following lectures and remainder of the sequence.

Simple Linear Regression: Assumptions

Standardized residuals.

- Remember the residual for the *i*-th observation was the observed value of Y minus the predicted value of Y
- $r_i = Y_i - \hat{Y}_i$
- The *standardized residual* for unit i is $r_i^* = \frac{r_i}{\sqrt{MSE}}$
- The *externally standardized residual* for unit i, $\tilde{r}_i$, is the same, but MSE is from the model fit without the *i*-th unit's observations.

Simple Linear Regression: Assumptions

Standardized residuals.

- Say we have a simple linear regression model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon$.
- A more precise standardized residual form that we will now use is:
$$r_i^* = \frac{r_i}{\sqrt{MSE(1-h_i)}}$$
- Where $h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum\limits_{i=1}^{n}(X_i - \bar{X})^2}$.
- This result is only for a simple linear regression model.
- h_i is referred to as the *leverage* of the i-th data point (which we will cover more in depth later on).

Simple Linear Regression: Assumptions

Standardized residuals.

- Say we have a multiple linear regression model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + \varepsilon$.
- A standardized residual is now: $r_i^* = \frac{r_i}{\sqrt{MSE(1-h_i)}}$
- Where h_i is the i-th diagonal element of the matrix $X(X^T X)^{-1} X^T$.
- The X matrix was defined in lecture 6 slide 13.
- And X^T is the transpose of X and A^{-1} is the inverse of the matrix A.
- With respect to multiple covariates, h_i is referred to as the *leverage* of the i-th data point.

Simple Linear Regression: Assumptions

Why do we standardize the residuals (also called studentizing)?

- That is to say we subtract off the expectation, then divide by the standard deviation of the residual, $r_i = Y_i - \hat{Y}_i$
- First note that the expectation of r_i is equal to 0.
- That is to say $E(r_i) = 0$.
- The estimated standard deviation of the residuals is $\sqrt{MSE(1-h_i)}$ (since $\widehat{var}(r_i) = MSE(1-h_i)$).
- The intuition is that standardized residual is a measure of the strength of the difference between observed and expected values.
- Thus, the magnitude of the residual alone does not quantify the strength of our model.

Simple Linear Regression: Assumptions

Why we will sometimes standardize the residuals (also called studentizing)?

- Example: A residual of 100,000 does not mean our model is not fitting well. It could be that the true value of Y is 10^9, in which case our model being off by 100,000 is not bad.
- The mathematical reasoning is so that standardized residuals will have expectation 0 and variance (or standard deviation) equal to 1 (and so it is more straightforward to figure out what residuals are extreme values).

Simple Linear Regression: Assumptions

Plot type	**Use to check**
Dotplot, stemplot, histogram of X's	Outliers in X, distribution of X
Residuals (r_i) against X_i's or $\hat{Y}_i$'s	Linearity, constant varince, outliers
r_i^* or $\tilde{r}_i$ against X_i's or $\hat{Y}_i$'s	Linearity, constant variance, outliers
Dotplots, stemplot, or histogram of r_i	Normality assumption
r_i against other explanatory variables	Dependent variables missing
Normal Q-Q plot	Normality assumption
r_i against time (if measured)	Errors are independent

Simple Linear Regression: Assumptions

Independence assumption.

- The first assumption we can check is if the errors are independent of one another.
- We will cover more complicated scenarios in later courses, but for now say we only have a single observation for each sampling unit in the data set.
- We would need our data to be randomly sampled from the population.
- The key is to be randomly sampled.
- This helps ensure that the observations are independent of one another (and hence the errors are independent).

Simple Linear Regression: Assumptions

Independence assumption.

- Example: Say we conduct a study on the effect of meditation on blood pressure (blood pressure is the response).
- Now say our dataset consists of members from the same family.
- Then how one subject's body reacts to meditation will likely be related to how another subject's body reacts to meditation, since they are all related and likely to have same biological/genetic/hereditary features.
- Would want to have subjects that are unrelated to one another.

Simple Linear Regression: Assumptions

Extending sample results to population assumption.

- For us to try and avoid issues with this assumption, we want our sample to be sampled from the population in question.
- Our sample needs to be representative of the population we want to extend our model results to.
- If our study is aimed at obtaining inference about elderly diabetic patients, then our sample should be coming from the population of elderly diabetic patients.
- Another example is obtaining inference on all potential voters in the USA.
- Our sample should include all voting age eligible people of all genders, ethnicities, and religions.

Simple Linear Regression: Assumptions

The first of the plots which we will use is the Residuals vs. Fitted values ($\hat{Y}_i$) plot.

- On the vertical axis (Y- axis) are the residual (r_i) values, and on the horizontal (X- axis) are the fitted $\hat{Y}_i$ values.
- We use this plot to check for the linearity assumption, equal variance assumption, and also to hint at potential outliers.

Simple Linear Regression: Assumptions

Red smoothing line in some of our plots we will see.

- The red line is called a smoothing line.
- If it is presented, it will help us visualize a pattern in a scatterplot.

Simple Linear Regression: Assumptions

- We do not want to observe any pattern in this plot.
- Patterns in this plot would suggest a non-linear relationship between Y and X (or X's in multiple linear regression setting).
- The idea is that if the relationship that is truly non-linear is modeled as linear (used to make prediction), then patterns in the residual plot will emerge.
    - Where the linear fit is ok, the residuals will be similar.
    - But where the linear fit deviates substantially from the truly non-linear relationship, then these residuals will be large (positive or negative).

Simple Linear Regression: Assumptions

The idea is that do not want the linear model to be fitting better/worse for certain values of the fitted value.

Want the points to be equally dispersed about the horizontal line at 0.

The linear line is doing equally well (or bad) for all values of the fitted Y's.

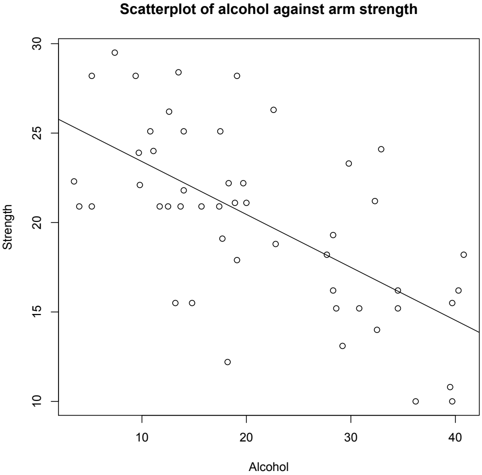
Simple Linear Regression: Assumptions

The following example uses the alcoholarm.txt dataset on the website.

In this long term study, 50 alcoholic men had there alcohol consumption measured and the strength of their deltoid muscle in the non-dominant arm.

The response variable is Y=Strength and X=alcohol consumption is the explanatory variable.

Simple Linear Regression: Assumptions



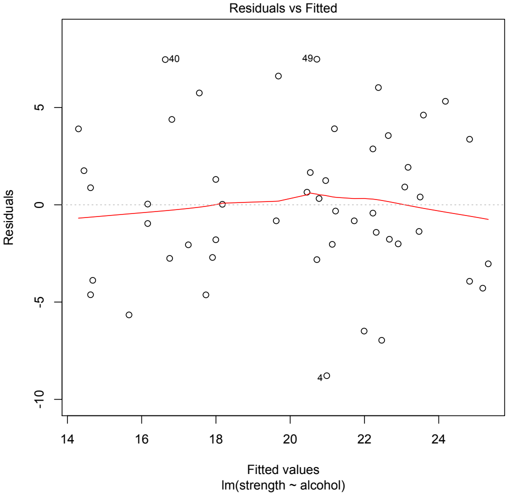
Simple Linear Regression: Assumptions

Here is the model output:

```
lm(formula = strength ~ alcohol, data = alcoholarm)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7847 -2.5450 -0.1477  2.6359  7.4815

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.36954    1.20273  21.925  < 2e-16 ***
alcohol     -0.29587    0.05105  -5.796 5.14e-07 ***
---

Residual standard error: 3.874 on 48 degrees of freedom
Multiple R-squared:  0.4117,Adjusted R-squared:  0.3994
F-statistic: 33.59 on 1 and 48 DF,  p-value: 5.136e-07
```

Simple Linear Regression: Assumptions

Here is the residual vs fitted plot.



Simple Linear Regression: Assumptions

- What we are looking for is no pattern in how the residuals are scattered about the horizontal line at 0.
- Points that are on the line at 0 in the residual plot means that they are on the regression line in the scatterplot. That is to say the predicted value is exactly the true value.
- Can also check for outliers to see what observation has a very large residual. This would imply that the true Y value for this observation deviates substantially from the assumed linear model than the other observations deviate.

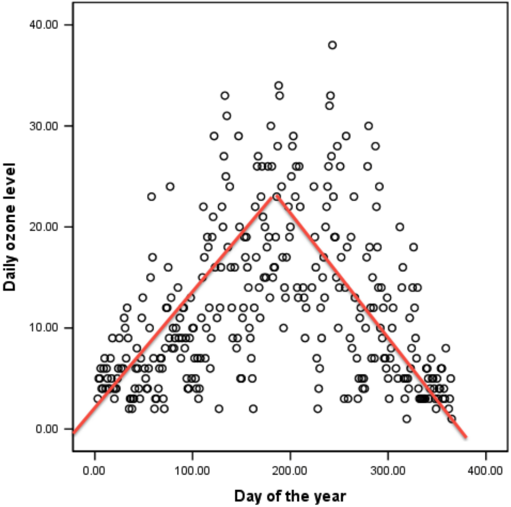
Simple Linear Regression: Assumptions

When the relationship is truly non-linear, we would expect to see patterns in the residual vs fitted plots.

This implies the linear model that we fit to the data is deviating greatly from the true observations at certain points in the range of the predictor/explanatory variable.

When the relationship between Y and X is non linear, then fitting a single linear model to the data will result in low accuracy.

Simple Linear Regression: Assumptions



Simple Linear Regression: Assumptions

In this example, ozone levels are measured from the start of the year (day 1) to the end (day 365).

In the example, it is evident that ozone levels do not have a linear association with the time of year.

Ozone levels peak in the middle of the year.

From the start of the year to summer, levels increase.

Then, once past summer, levels begin to decrease as winter starts.

Simple Linear Regression: Assumptions

Now to examine a non-linear case with residual vs fitted plots.

This example uses the treadwear.txt data on the class site.

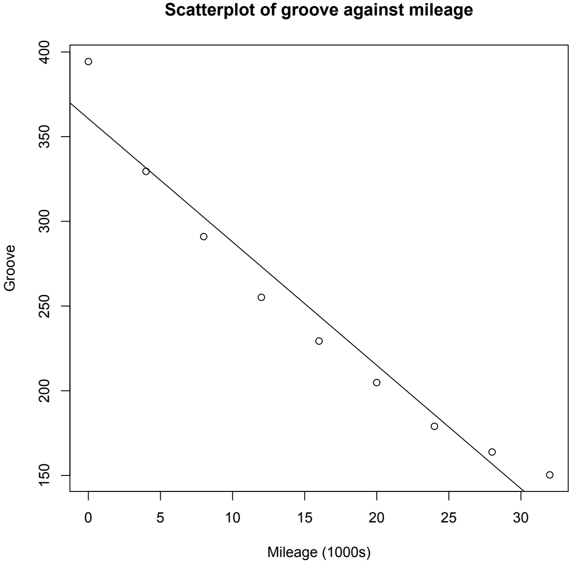
In this study, researchers wanted to investigate if tire tread wear is linearly related to miles travelled.

Tread wear is how much tread (grip) the tire has.

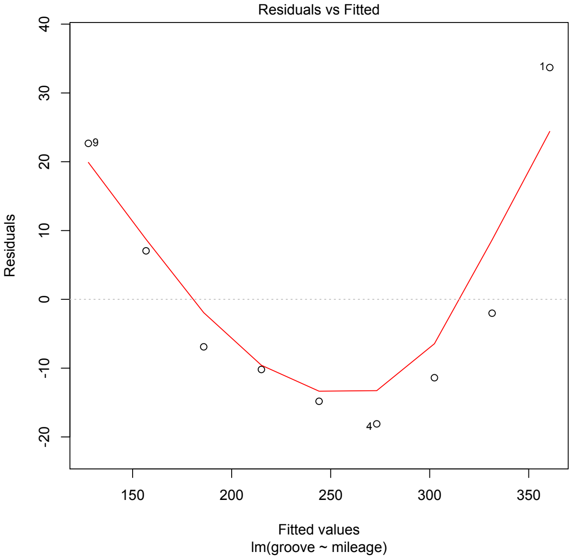
The linear relationship would assume that the tires wear out evenly over the miles traveled.

Here, Y=groove and X=mileage (in 1000s of miles).

Simple Linear Regression: Assumptions



Simple Linear Regression: Assumptions



Simple Linear Regression: Assumptions

- Can see the non-linear shape from the scatterplot is magnified in the residual plot.
- We have some inference on whether there is a linear relationship between response and explanatory variables.
- Potential tools to help alleviate the problem (transform data to try and induce a linear relationship) will be covered in later lectures.

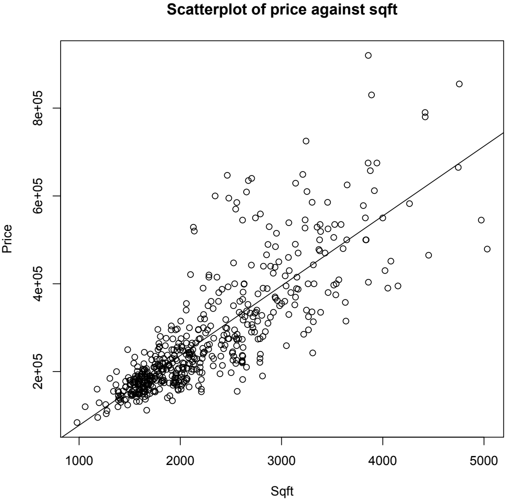
Simple Linear Regression: Assumptions

Constant variance assumption.

- The linear regression model has a constant variance assumption.
- That is that the Y variables have the same variance across all the values of X (all Y's have the same σ_{ε}^2 regardless of what their covariate values are).
- For example the variance of Y when X=100 is the same as the variance of Y when X=1000.
- This assumption needs to be checked as the data may not be conveying this specification (could be the variance changes as the value of X changes).

Simple Linear Regression: Assumptions

Example, MidwestSales data. Y=price and X=square footage.



Simple Linear Regression: Assumptions

- Can see that as Sqft increases, the variance among the response (Price) is increasing.
- As sqft increases, the points become more and more scattered around the linear regression line (which again was the estimated expected value of Y at a given value of X).
- This is to imply that for larger houses (3500+ sqft), the variance of price is greater than smaller houses (<3500sqft) .
- Can check for this in a residual vs fitted plot.

Simple Linear Regression: Assumptions

What are we looking for with respect to the constant variance assumption being violated?

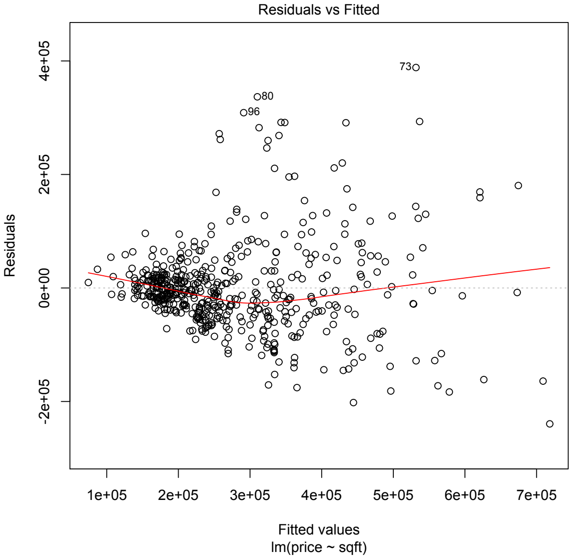
- As with using a residual vs fitted plot to check for non-linear association between X and Y, the same idea can be used to check for non-constance variance of Y over X.
- As the variance changes over X, the linear model will start to have a poorer fit (inaccurate predictions), which will be highlighted in the residual vs fitted plot.

Simple Linear Regression: Assumptions

What are we looking for with respect to the constant variance assumption being violated?

- What we want to see is that the points in the residual vs. fitted plot are equal spread out around the straight horizontal line at intercept 0.
- If the points are not equally scattered around the straight horizontal line at intercept 0, such as the points become more and more scattered as we move left to right along the X-axis, that is indicative of non-constant variance (which means our assumption of a single σ^2 for all the observations is not reasonable).

Simple Linear Regression: Assumptions



Simple Linear Regression: Assumptions

- Can see that as fitted values increases (which is from square footage increasing), the residuals begin to vary around 0 much greater than when fitted values are smaller (sqft lower).
- Since there is no pattern to them, just an increase in the variance (how much they are scattered) as fitted value increases is indicative of non-constance variance (variance is increasing as explanatory increases).

Simple Linear Regression: Assumptions

- Note that in the case of simple linear regression (a single covariate), the above plots could have also been the explanatory variable and residual plot.

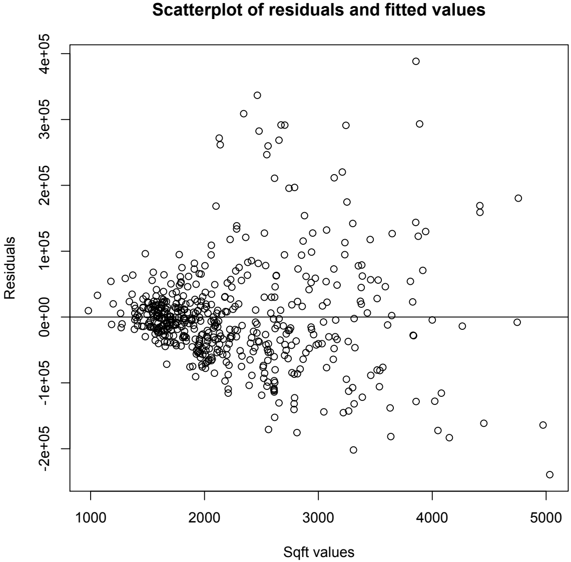
- Instead of the fitted $\hat{Y}_i$ value, we have the explanatory X value.

- Note that since we have a simple linear model, then the plot of fitted $\hat{Y}_i$ value against residual will return the same information as the explanatory X value against residual plot.

- But the plots with the fitted value vs. the residuals is useful for when we have a multiple linear regression setting.

Simple Linear Regression: Assumptions

Now with sqft on the X-axis instead of fitted values.



Simple Linear Regression: Assumptions

Review of the fitted values versus the residual plots.

- Note the plots could have also been with the standardized residuals (instead of the regular residuals).
- When there is a pattern in the plot, then that is indicative of a non-linear relationship between Y and X (or X's in the case of several covariates).
- When the points are not equally scattered around the horizontal line at 0, then that is indicative of non-constant variance (the variance of Y will tend to increase or decrease depending on the X values).

Simple Linear Regression: Assumptions

Normality of the errors (ε's) assumption.

- Now we come to what is known as the *qq plot* (or QQ plot).
- This is a *quantile-quantile plot*. Will plot the quantiles of the standardized residuals against those of a standard normal distribution.
- This can be used to assess the assumption of normal errors in the linear model.

Simple Linear Regression: Assumptions

- If the quantiles of the standardized residuals are close to the quantiles of the standard normal distribution, then the plot should show them on a straight line (in R, by default this line goes through the first and third quartiles).
- If the points scatter about the the 45-degree line (such as at the tails), this implies the errors are not normally distributed.
- Can infer information about the way the errors are distributed based on the way the points differ from the straight line.
- Errors have lighter or heavier tails than a normal distribution, or it is a skewed distribution.

Simple Linear Regression: Assumptions

Return to the alcohol and deltoid muscle arm strength data set.

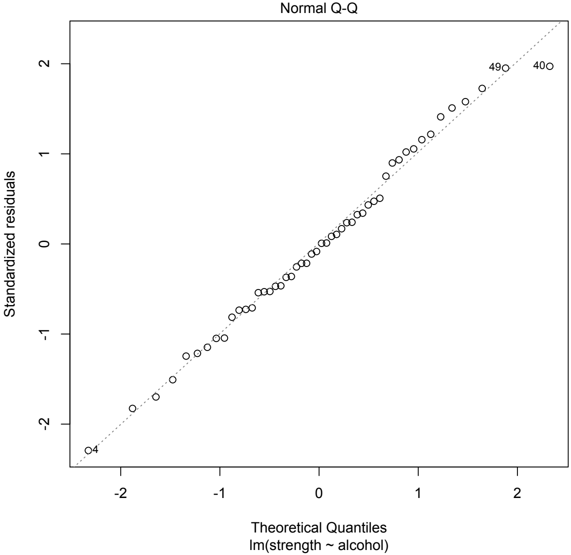
Y=arm strength and X=alcohol consumption.

Fit a model and obtain the residuals.

Using R, we can obtain qq plots easily.

In the qq-plot, the Y-axis will be the standardized residuals quantiles and the X-axis will be the normal distribution quantiles.

Simple Linear Regression: Assumptions

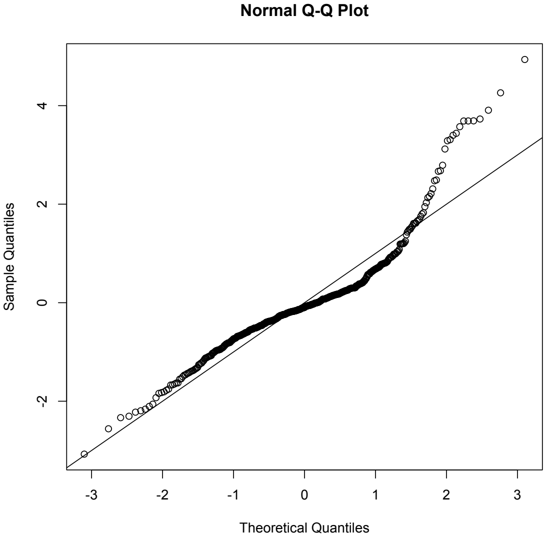


Simple Linear Regression: Assumptions

- Most of the points are on the straight line.
- This would suggest that the standardized errors are approximately standard normally distributed.
- This would further imply that the errors are approximately normally distributed with mean 0 and variance of σ_ε^2 (estimated by MSE).

Simple Linear Regression: Assumptions

Example: MidwestSales data, Y=sales price and X=square footage.

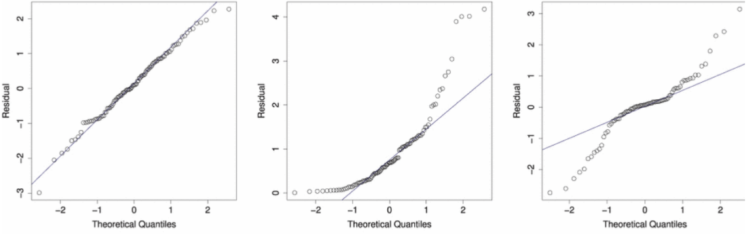


Simple Linear Regression: Assumptions

- Can see the points waiver at the tails of the quantiles (large values) deviate from the straight line.
- This would suggest that the standardized errors are not standard normally distributed.
- Further, the way the points differ at the different ends of the tails would suggest the errors distribution has heavy tails .

Simple Linear Regression: Assumptions

In general, this is how we can assess qq-plots.



(a) Normal residuals

(b) Skewed right residuals

(c) Long-tailed residuals

Simple Linear Regression: Assumptions

- If some assumptions seem to fail, then there are potential fixes.
- The fixes will tend to complicate the model.
- The fixes will try adjust the model to allow for the accurate prediction of Y based on X.
- We will review a few of these fixes.

Simple Linear Regression: Assumptions

- Say we plot the residuals against the fitted values and notice a pattern in the residuals.

- This is to suggest a non-linear relationship between X and Y.

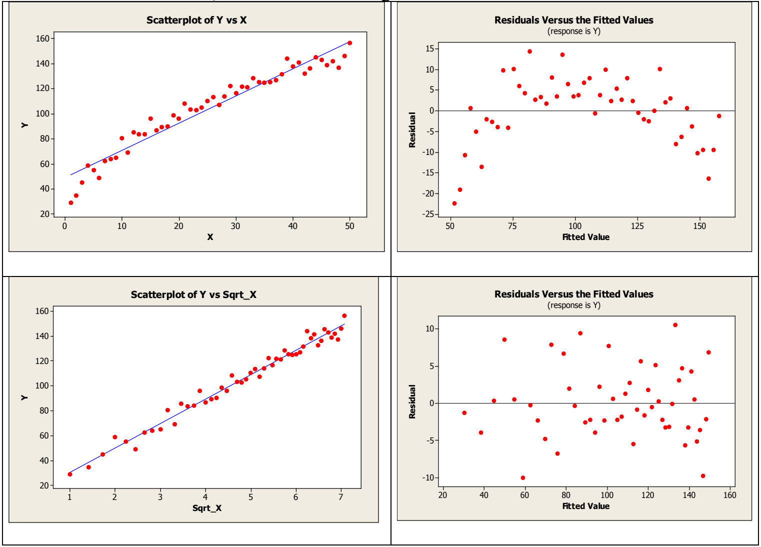
- A suggested fix is to transform the X variable, such as square root ($\sqrt{X}$) or log transform ($log(X)$).

- The idea is that if there is evidence of a non-linear relationship between X and Y, then there could be a linear relationship between a function of X (such as $\sqrt{X}$ or $log(X)$) and Y.

- Based on the pattern seen in the residual vs fitted plot, a form of the function to be used on X can be suggested.

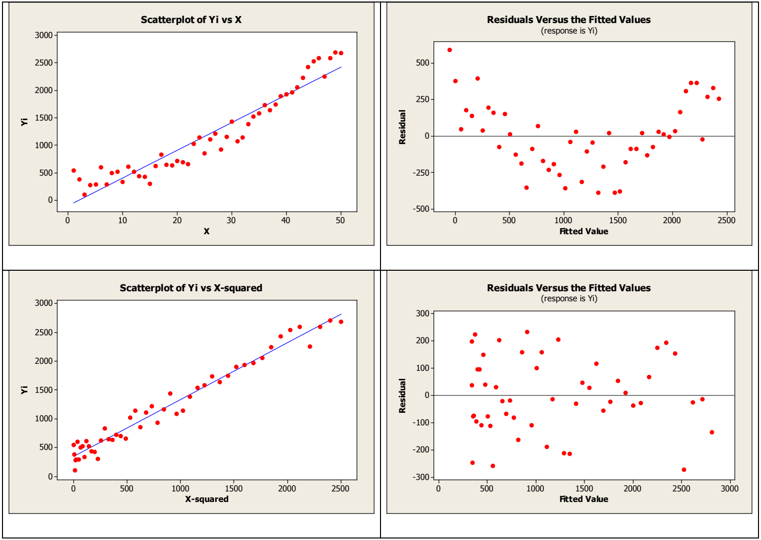
Simple Linear Regression: Assumptions

Residuals are inverted U, use X' = $\sqrt{X}$ or $\log_{10}$ X



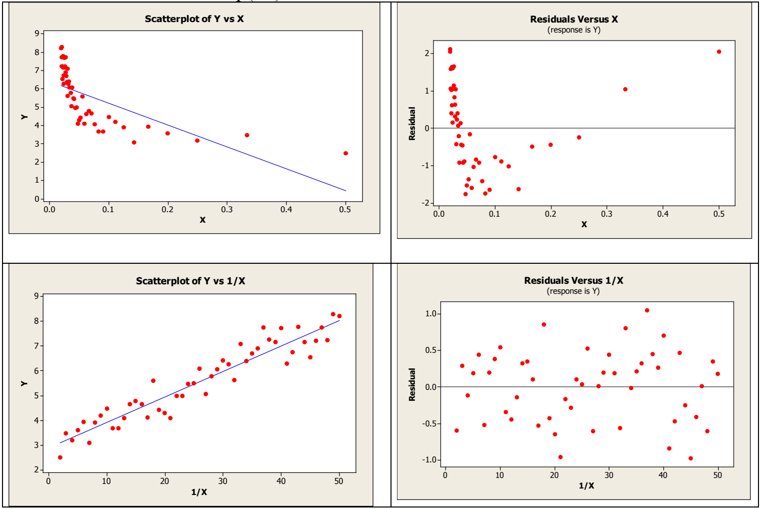
Simple Linear Regression: Assumptions

Residuals are U-shaped and association between X and Y is positive: Use $X' = X^2$



Simple Linear Regression: Assumptions

Residuals are U-shaped and association between X and Y is negative:
Use X' = 1/X   or   X' = exp(-X)



Simple Linear Regression: Assumptions

- As can be seen in some example, by choosing the appropriate transformation of X, a linear relationship was induced.
- What this is doing is modeling a non-linear relationship between X and Y.
- Example: If we do a power transformation of X (X^2), then the predicted values are: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i^2$.
- Have to be careful with how we use the model to interpret the β_1 coefficient when X is transformed.
- Will see in an upcoming lecture how we can have a multiple linear regression model with transformations of the explanatory variables, such as a polynomial model: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i^2$.

Simple Linear Regression: Assumptions

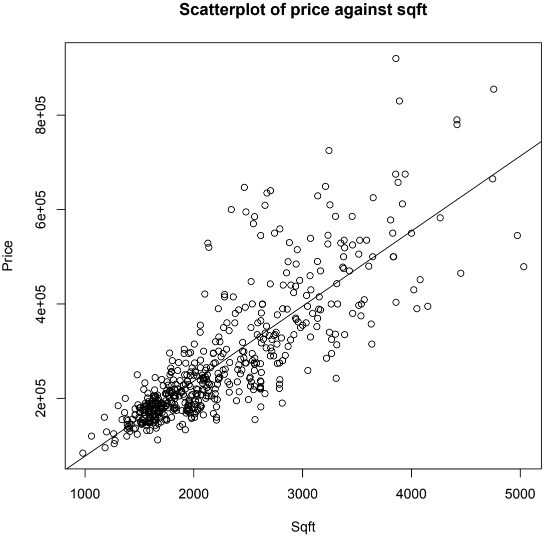
Now about the assumption of constance variance.

For now, we will discuss how to help alleviate the issue of non-constance variance by transforming the response variable Y.

The issue of non-constance variance was seen when we plot the residuals against the fitted values, and notice the spread of the errors increases (or decreases) as we move along the fitted values (the X-axis).

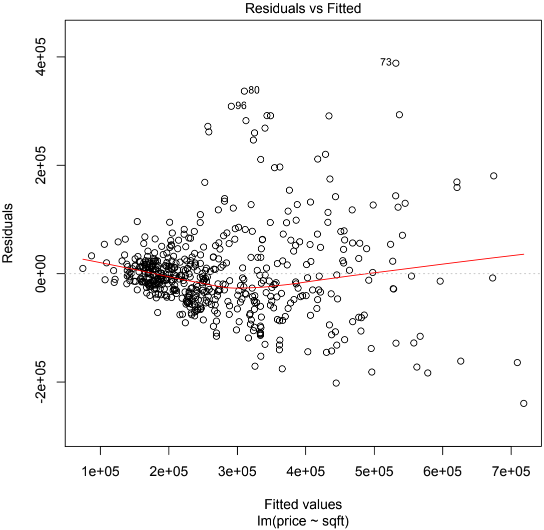
Simple Linear Regression: Assumptions

Return to the MidwestSales data example.



Simple Linear Regression: Assumptions

MidwestSales data example.



Simple Linear Regression: Assumptions

The variance is increasing as the square footage of the house increases.

Just like transforming the explanatory X variable was used to help with the non-linearity of the association, can transform the response Y variable to help with the issue of non-constant variance.

The idea is that transforming the Y variable will induce a new form of the variance.

Simple Linear Regression: Assumptions

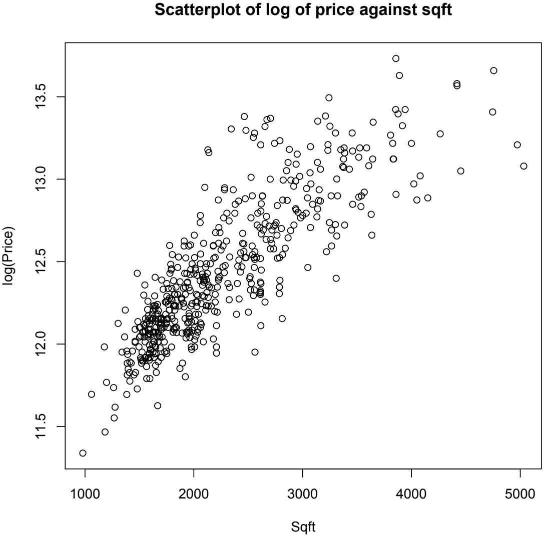
Transform Y onto the natural log scale, $log(Y)$.

Want to decrease the variance, but decrease it more for the larger values of Y.

The new model will be of the form $\widehat{log(Y)} = \hat{\beta}_0 + \hat{\beta}_1 X_i$

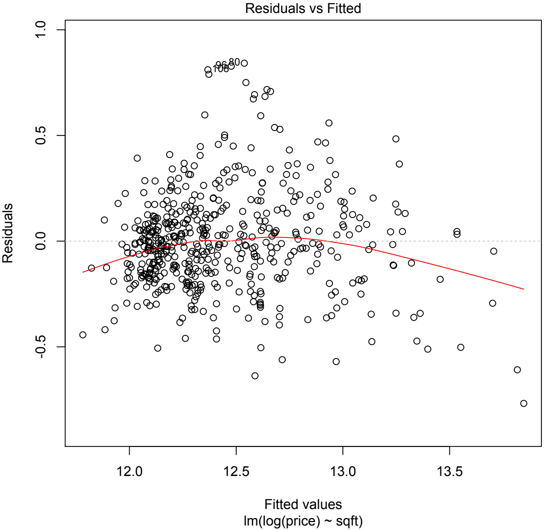
Simple Linear Regression: Assumptions

Return to the MidwestSales data example. Now with log(Y).



Simple Linear Regression: Assumptions

MidwestSales data example. Now with log(Y).



Simple Linear Regression: Assumptions

- The log transformation helped a bit with the non-constant variance.
- Non-constant variance is still present (as sqft increases, the variance of log(price) increase).
- But the non-constance variance is mitigated compared to the untransformed case.
- The change of the transformation can be further seen in the residual vs fitted plot.