Lecture 17 – 18

Linear Model Evaluation

QQ Plot (Quantile – Quantile Plot)

A QQ plot is a graphical tool to compare two probability distributions by plotting their quantiles against each other. It is often used to assess whether a dataset follows a theoretical distribution (like the normal distribution) or to compare two empirical samples.

Quantiles are specific values in a dataset or distribution which divide the data into percentages.

Example 1: the 50% quantile (also called the median) is the value below which 50% of the data lie.

Example 2: Consider a normal distribution $N(0,1)$, which is a normal distribution with mean 0 and standard deviation 1. The 50% quantile in this distribution is 0. Which means if we took many samples from this distribution, about half of the samples will be less than 0.

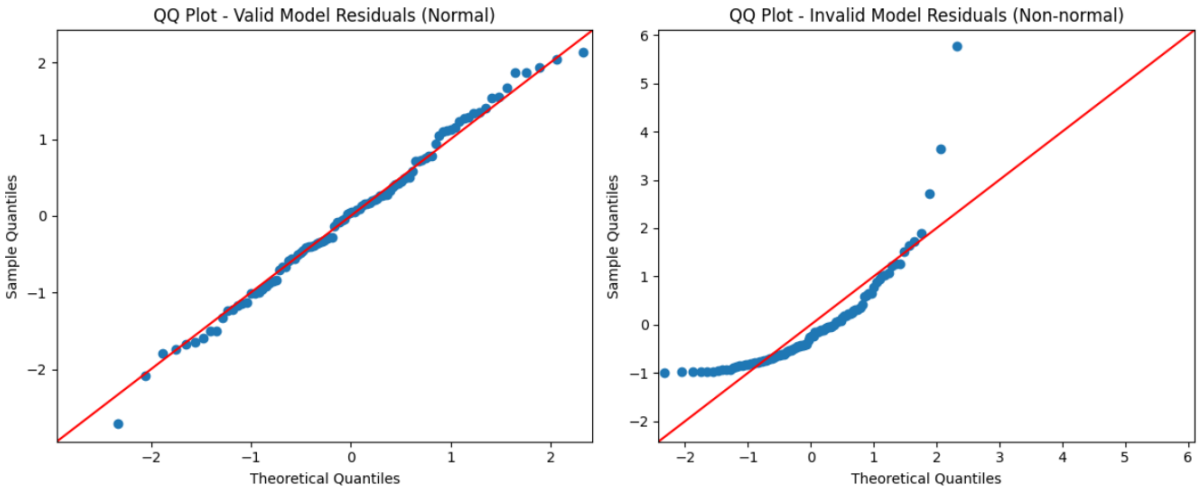
If we plot the quantiles of our sample data ($sample.q$) against the quantiles of a known theoretical distribution ($known\_distribution.q$), and the points lie roughly on a straight line (usually $y = x$ line), it shows that our sample data likely comes from that theoretical distribution.

This is because each quantile in the sample matches the corresponding quantile in the known distribution.

If for all quantiles (q), the quantiles of the sample are equal to the quantiles of the known distribution, then the sample and the known distribution have the same distribution. Therefore, the QQ plot helps us visually check the goodness of fit between our data and a theoretical model.

Importance:

- Residuals (errors) from linear regression models are often assumed to be normally distributed.
- QQ plots help verify if this assumption holds by comparing residuals' distribution with a normal distribution.
- If the residuals follow a normal distribution (points lie on the QQ plot's reference line), the model assumptions hold, and inference made from the model is valid.
- If points deviate significantly from the line, the residuals do not follow the assumed distribution, indicating that the model might be mis-specified or that a different model or transformation may be necessary.



Metrics for Evaluating Linear Regression Model Fit

The loss function is:

$$||y - X\beta||_2^2 = \sum_i (y_i - \hat{y}_i)^2$$

However, this is not enough, because this sum depends on the scale of the data and does not give a relative measure of fit.

TSS: Total Sum of Squares

$$TSS = \sum_i (y_i - \bar{y}_i)^2$$

This is the Total Sum of Squares, measuring the total variability of the data around its mean. It tells us how spread out the observed y values are.

ESS: Explained Sum of Squares

$$ESS = \sum_i (\hat{y}_i - \bar{y}_i)^2$$

The Explained Sum of Squares measures how much of the variability in y our model explains by comparing the model's predictions $\hat{y}_i$ to the mean of y.

RSS: Residual sum of Squares

$$RSS = \sum_i (y_i - \hat{y}_i)^2$$

measures the total squared difference between the actual observed values y_i and the predicted values $\hat{y}_i$ from the regression model. RSS quantifies how much error there is in your model's predictions.

$$\boldsymbol{TSS = RSS + ESS}$$

- $\boldsymbol{y_i}$**:** The observed value of the dependent variable for the i-th data point. This is the actual data point we are trying to model or predict.
- $\boldsymbol{\hat{y}_i}$**:** The predicted (or fitted) value of the dependent variable for the i-th data point, obtained from the regression model.
- $\boldsymbol{\bar{y}_i}$**:** The mean (average) of all observed values of the dependent variable across all data points.

R-squared: Fraction of Variance Explained

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

R^2 measures the fraction of the total variance in the data that is explained by the model predictions.

It Ranges from 0 to 1

- $R^2 = 1$:
  - Perfect fit.
  - Model explains all variance in y.
  - The points lie exactly on the regression line.
  - The model perfectly predicts all y_i, so variability explained is 100%.
- $R^2 = 0$:
  - No explanatory power.
  - Model explains none of the variance; predictions are no better than simply using the mean $\bar{y}_i$.
  - The model predicts just the mean $\bar{y}_i$ for all inputs, so no relationship with X, and no variance explained.

Hypothesis Testing

Hypothesis testing is a method for making statistical decisions about the parameter of a population, using sample data.

We start with two hypotheses for a parameter β, (for example, the coefficient of a variable x in a regression):

- Null Hypothesis (H_0): $\beta = 0$ (There is no effect or no linear relationship between X and Y.
- Alternative Hypothesis (H_0): $\beta \neq 0$ (There is a linear relationship)

We want to know if there is enough evidence in the observed data to reject (H_0). That is, is the coefficient β significantly different from zero?

1. Step 1: Estimation & Normalized Estimate
   a. From the sample, calculate an estimate $\hat{\beta}$

b. Standardize this estimate to form a t-statistic (t-value), which measures how far $\hat{\beta}$ is from zero in terms of standard errors.

$$t = \frac{\hat{\beta} - 0}{SE(\hat{\beta})}$$

c. Here, $SE(\hat{\beta})$ is the standard error of the estimate $\hat{\beta}$.

d. The null hypothesis assumes $\hat{\beta} = 0$.

2. Step 2: Distribution of the Test Statistic Under H_0
   a. Under the null hypothesis, the t-statistic follows a t-distribution parameterized by the degrees of freedom (df), related to the sample size.
   b. The t-distribution looks like a normal distribution but with heavier tails (especially for small samples).
   c. The t-distribution is used because the population variance is unknown and replaced by the sample estimate.
3. Step 3: Calculate the p-value
   a. The p-value measures the probability of observing a t-value as extreme or more extreme than the actual one obtained, assuming the null hypothesis is true.
   b. Formally,

$$p = P(|T| \geq |t_{observed}| \; | H_0)$$

   c. For a two-tailed test (testing if $\beta \neq 0$): The p-value is the sum of probabilities in both tails beyond the observed t-value.
4. Step 4: Decision
   a. Choose a significance level α, typically 0.05.
   b. If the $p - value < \alpha$, reject the null hypothesis (evidence supports $\beta \neq 0$).
   c. If the $p - value > \alpha$, fail to reject H_0 (not enough evidence to say β is different from zero).

Regression Results

OLS Regression Results

Dep. Variable:	y	R-squared:	0.840
Model:	OLS	Adj. R-squared:	0.836
Method:	Least Squares	F-statistic:	254.1
Date:	Sun, 20 Mar 2022	Prob (F-statistic):	2.72e-39
Time:	11:36:16	Log-Likelihood:	-482.37
No. Observations:	100	AIC:	970.7
Df Residuals:	97	BIC:	978.5
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.1912	3.162	0.693	0.490	-4.085	8.467
x1	29.3912	3.274	8.977	0.000	22.893	35.889
x2	78.1391	3.594	21.741	0.000	71.006	85.272

Omnibus:	1.279	Durbin-Watson:	1.824
Prob(Omnibus):	0.527	Jarque-Bera (JB):	1.065
Skew:	0.253	Prob(JB):	0.587
Kurtosis:	2.999	Cond. No.	1.38

1. **Model Information**
    a. *Dependent Variable:* The outcome or response we are trying to predict (here, called y).
    b. *Model:* Indicates the type of model used (OLS regression).
    c. *Method:* The estimation technique used, which is Least Squares here. It means the model finds the line/plane that minimizes the squared difference between observed and predicted values.
    d. *No. Observations:* Number of data points used to fit the model. More data often means more reliable estimates.
2. **Model Fit Statistics**
    a. *R-squared*: Measures how well the independent variables explain the variation in y. It ranges from 0 to 1, with higher values meaning better explanatory power. Here, 0.840 means that about 84% of the variability in y is explained by the variables in the model.
    b. *Adjusted R-squared:* Like R-squared but adjusts for the number of variables. It avoids giving "false credit" for adding unnecessary variables. Slightly lower

than R-squared here, but close i.e. suggesting the variables included are meaningful.

c. *F-statistic and Prob(F-statistic):* The F-test assesses whether the model is statistically significant i.e., whether at least one independent variable has a relationship with y. A very low p-value (probability) associated with this test suggests the model performs better than a model with no predictors.

3. Coefficients Table

a. *coef (Coefficient):* Estimates of the effect size of each variable on y. For example, x_1 has a coefficient ~ 29.4, meaning each unit increase in x_1 is associated with an increase of 29.4 units in the predicted y, holding other variables constant.

b. *std err (Standard Error):* Reflects the uncertainty in the coefficient estimates. Smaller values mean more precise estimates.

c. *t (t-statistic):* A measure of how many standard errors the coefficient is from zero. The larger the absolute t-value, the more confident we can be that the coefficient is truly different from zero.

d. *P>|t| (p-value):* Shows the probability of observing such a coefficient if the true effect were zero (null hypothesis). Low p-values (commonly below 0.05) indicate the effect is statistically significant.

e. *Confidence Interval [0.025, 0.975]:* The range of values in which we are 95% confident the true coefficient lies. If this range includes zero, it suggests that the effect may not be significantly different from zero.

f. For example, const (intercept) has a p-value of 0.490 and confidence interval including zero → not statistically significant, so the intercept may not be reliably different from zero.

g. *Important:* Both x_1 and x_2 have very small p-values and confidence intervals well above zero, indicating strong evidence they influence the dependent variable.

Z – values

A Z-value tells we “How many standard deviations away from the mean" a particular data point or boundary lies in a standard normal distribution i.e. a Gaussian distribution with mean 0 and standard deviation 1.

When we draw many samples from a population, the *Z-value helps identify intervals around the mean that will capture a certain percentage of the data*. For example, the 95% Z-value is about 1.96, meaning 95% of observations lie within 1.96 standard deviations from the mean.

Confidence Intervals

Instead of relying on a single estimate (like the sample mean) to represent a population parameter, a confidence interval provides a range of plausible values based on our data.

Starting with the sample mean (our best guess), we add and subtract a margin of error. The margin depends on the variability of our data (standard error) and the desired confidence level (often 95%, corresponding to a Z-value of 1.96).

A 95% CI means that if we repeated our sampling many times, about 95% of these calculated intervals would contain the true population parameter. It is a way to express uncertainty.