Linear Regression

Linear regression means the dependent variable is linear in the model coefficients.

- Linear: $y = b_0 + b_1x_1 + b_2x_2$
- Not linear: $y = b_0x_1^{b_1} + b_3x_2^{b_2}$

`lm` in R

Rentals as a function of temperature and humidity:

```
mod <- lm(data = df, rentals ~ temp
    + rel_humidity)
summary(mod)
```

Consider a "Best Fitting" Line

$$\hat{y} = b_0 + b_1x$$

Slope b_1 - sign is same as the sign of $\text{CORR}(X, Y)$. $\text{CORR}(X, Y)$ is between $[-1, 1]$ and unit-less. But b_1 has units.

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \text{CORR}(X, Y) \cdot \frac{\text{SD}(Y)}{\text{SD}(X)}$$

Intercept - $b_0 = \bar{y} - b_1\bar{x}$

Mean-center data - if both y and x variables are mean-centered, rerun linear regression to get:

$$\hat{y}_i - \bar{y} = b_1(x_i - \bar{x})$$

- new intercept will be 0
- new slope remains b_1

Regression to the mean - If $x = \bar{x}$ (the average of all x values), the predicted $\hat{y} = \bar{y}$ (average of all y values), independent from x.

Independent Variables

Categorical variables ("factor variable" in R) - if encoded with one-hot, one category should be dropped to avoid perfect multicollinearity (a situation where one predictor variable can be perfectly predicted from the others). This omitted category serves as a reference category against which the other categories are compared. This approach is known as creating "dummy variables."

Evaluation

Residuals (error) - $e_i = y_i - \hat{y}_i$

Sum of Squared Residuals -

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

Total Sum of Squares - a measure of the total variability in the observed data.

$$SST = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$$

R^2 - the proportion reduction in sum of squared residuals by the regression model compared to the baseline model (which always predicts the average value of all ys in the data: $\hat{y} = \bar{y}$).

$$R^2 = 1 - \frac{SSR}{SST} = \frac{SST - SSR}{SST}$$

- R^2 of a regression model lies between 0 and 1.
- Adding a new independent variable to a regression model can only increase R^2.
- R^2 on its own cannot judge how good a model is.

R^2 in multiple linear regression - equals the square of the correlation between the actual values y and the predicted values $\hat{y}$.

$$R^2 = (\text{CORR}(y, \hat{y}))^2$$

In the case of a perfect fit where every prediction of y_i is going to be correct, then $R^2 = 1$

R^2 in simple linear regression - equals the square of correlation between dependent variable y and independent variable x.

$$R^2 = (\text{CORR}(y, x))^2$$

Degrees of Freedom - the number of observations minus the number of parameters estimated (including the intercept): $df = n - p - 1$

Residual Variance - $\frac{SSR}{df}$

Residual Standard Error - provides a measure of the average distance that the observed values fall from the regression line.

$$RSE = \sqrt{\frac{SSR}{df}}$$

- Smaller RSE: the model's predictions are closer to the actual data points, suggesting a good fit.
- Larger RSE: the model's predictions are further from the actual data points, suggesting a poor fit.

Troubleshooting

The model's ability to predict the future and its interpretability can be impaired by

- The presence of irrelevant independent variables
- The presence of highly correlated independent variables
- The presence of "too many" variables relative to the size of the dataset

Irrelevant variables - have large p-values. Smaller p-value (R: `Pr(>|t|)`), the better. If lower than 0.05, consider it significant at the 5% significance level; otherwise, consider it non-significant at the 5% significance level. In R, a variable with one or more * is considered significant at 5% level. A variable not significant at the 5% level but at the 10% level can be stated as "less clear".

p-value of a variable x_i - the probability of observing a coefficient estimate as extreme, or more extreme, than the one actually obtained by the regression run in a similar sample assuming $\beta_i = 0$.

Highly correlated variables - Detected by inspecting the correlation matrix. (R: `cor(df)`).

Multicollinearity - When two variables are highly correlated (typically correlations higher than 0.75 in magnitude), resolve this by removing either of the independent variables and running the linear regression **again**.

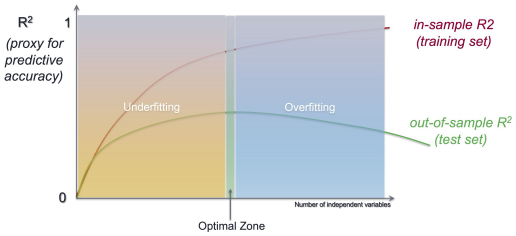
Overfitting ("too many" variables) - leads to poorly predicting future data. Remedy by (1) avoiding "non-significant" variables uising the starts * as a guide and (2) out-of-sample testing.

Out-of-Sample Testing -

1. Partition data set into 70% trainning set and 30% test set before creating the regression model.
2. Generate a set of models (e.g. with different sets of variables) using only the training data.
3. Evaluate models on the test set using out-of-sample R^2.

Out-of-Sample R^2 -

$$\text{OSR}^2 = 1 - \frac{\text{SSR of regression model applied to the test set}}{\text{SSR of baseline model applied to the test set}}$$



Key takeaways -

- Choose final model based on out-of-sample predictive quality metrics.
- Use metrics like R^2 and the standard error of regression in combination because they have different strength and weaknesses.
- Only include significant variables that are not highly correlated.
- Coefficients should "make sense".
- Use for interpolation rather than extrapolation.

Inferring model's coefficient - from the practice: "Correlation is not the same as causation. It could be that individuals with other contributing factors happen to sleep longer. It could be that these other factors, not sleep per se, are causally linked to the condition under study."

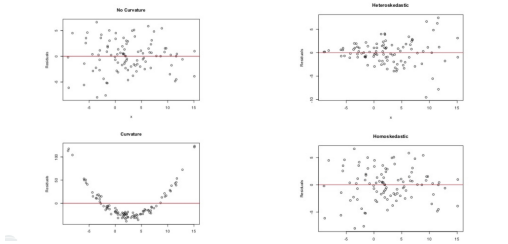
Technical caution - for the p-value to be correct, the "unaccounted for" differences in the regression model ϵ (think of the residuals) needs to have zero mean, constant standard deviation, be independent and follow a normal distribution.

Linearity - there is no curvature

Homoskedasticity - the dispersion of $e_i = y_i - \hat{y}_i$ is not systematically smaller or larger for large x_i values compared to small x_i values.

Nonlinearity or Curvature in Residuals?

Heteroskedasticity in error variances?



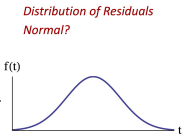
Top-right: hetero. Bottom-right:homo.

Normality - the residuals are approximately normal.

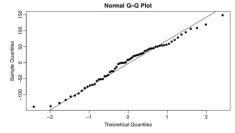
Checking Normality of Residuals.

QQ-Plot (quantile vs quantile plot) is commonly used here.

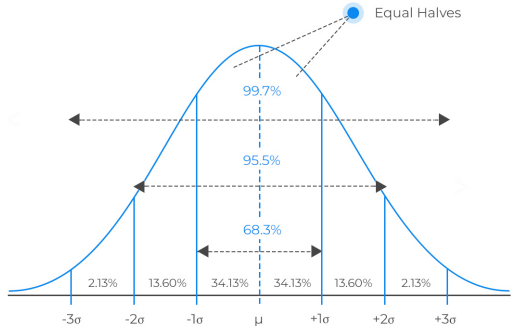
Make a normal quantile plot of the residuals interpretation:




```
qqnorm(resid.both)
qqline(resid.both)
```



Side bar



- about 68% of the total values lie within 1 standard deviation of the mean.
- about 95% lie within 2 standard deviations of the mean
- about 99.7% lie within 3 standard deviations of the mean.