1 Goodness of fit and nested models

1.1 Sum of Squares Decomposition-MSE-TSS-ESS-RSS

Mean Squared Error: $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}RSS$

Total Sum of Squares: $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$

- The sum of the squares of the observations from the null (intercept-only, no explanatory variables) model.
- When properly scaled, it is the sample variance of Y, which estimates the population variance of Y.
- Add a new variable, TSS stays the same.

Explained Sum of Squares: $ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

- $\hat{y}_i$ predicts y_i using the LR, while $\bar{y}$ predicts y_i without a model. If our model is better than nothing, this should be large. Measures how much is explained by the additional information given by the LR.

Residual Sum of Squares: $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- This is the sum of the squares of the residuals from the fitted model.
- Our estimated parameters minimize these errors.
- Add a new variable,RSS does not increase, because it makes the model fit better.

If parameters are estimated using least squares and the LR has an intercept: $TSS = ESS + RSS$

The coefficient of determination $R^2 = 1 - \frac{RSS}{TSS}$

Adjusted R^2 Adj $R^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)}$

LR with an intercept and estimated by least squares, it is equivalent to: $R^2 = \frac{ESS}{TSS}$

Interpretations: LR with an intercept and estimated by least squares, the coefficient of determination:

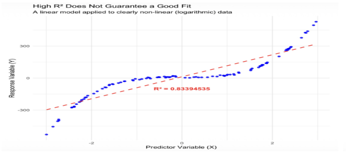
- Proportion of variance of the response (TSS) explained by the model (ESS).
- Lies between 0 and 1, both ESS and RSS are nonnegative.
- Measures the gain in predicting the response using the linear model instead of the sample mean, relative to the total variation in the response.

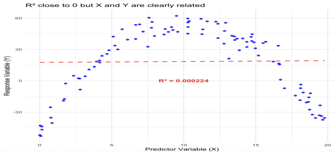
1.2 R-squared R^2

R2=0.051 means that 5.1% of the variability in the response variable (teaching scores) is explained by the LR model.

```
glance(fit2)
y <- dat$score #Manual R-sqr
y.hat <- fitted(fit2)
TSS <- sum((y-mean(y))^2)
RSS <- sum((y-y.hat)^2)
my_Rsq <- 1-RSS/TSS
```

- Computed using **in-sample observations**.
- Does **not** indicate how well the model predicts **out-of-sample (test set)** cases.
- Ranges between **0 and 1** if the LRmodel. Includes an intercept, and Is estimated using least squares (LS).
- **negative** R^2 indicates that the **sample mean** is a better predictor than the estimated linear regression model.
- Used to **compare the size of residuals** from the fitted model with those from the **null model**.
- **Cannot be used to test hypotheses**, since its sampling distribution is not known.
- Adding more variables to a linear regression model **cannot decrease** R^2.
- Cannot compare models of different sizes (Adj R^2 can for dif size) but can for same sizes





1.3 Nested Models - F-test/statistics

To check if reduced model and full model significantly different while simulataneously testing if many parameters are zero! Example: is LR better? Reduced model:(Additive) $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$
Full Model:(Interaction) $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} \times x_{i2} + \varepsilon_i$
Hypothesis:$H_0 : \beta_3 = 0$
With the F-test, we could test the significance of many parameters. For 1 parameter, T and F test same.

```
lm_red <- lm(score ~ age + sex, data=dat)
lm_full <- lm(score ~ age*sex, data=dat)
anova(lm_red,lm_full)
#Note that glance also includes this statistic and p-value
glance(lm_full)
```

If p.value $\leq \alpha$, we have evidence that a model with an interaction term fits the data better than a model without.
sigma in glance is Standard deviation of the error term. tidy gives T-test statistic, Anova gives F-test. $(statistic)^2 =$ F

2 Evaluation Metrics

MSE, the R^2 can be computed for new responses in a test set compared to the predicted values obtained using the trained LR. However, note that it is no longer the coefficient of determination. It measures the correlation between the true and the predicted responses in a test set. The F-statistic and the R^2 both depend on the RSS and the TSS.

2.1 Inference

- **F-Test** to compare and test nested models - anova.
- **T-Test** to test the contribution of individual variables to explain the response. Thus, we can use these tests to evaluate variables one at a time. - lm or tidy
- **R$\hat{2}$ or RSS** The RSS decreases as more variables are included in the model!
- **Adjusted R$\hat{2}$** R$\hat{2}$ is penalized by the number of variables in the model

2.2 Prediction

- **F-Test** can not be used to compare out-of-sample predictions from different models!
- **Estimates of the test MSE** The Mallow's Cp, Akaike information criterion (AIC) and Bayesian information criterion(BIC) add different penalties to the training RSS to adjust for the fact that the training error tends to underestimate the test error

2.3 Step wise model selection RFE

these approaches have many potential issues such as multiple comparison problems resulting in p.values that are too low.

```
dat_train <- sample_n(dat_s, size=nrow(dat_s)*0.75, replace=FALSE)
dat_test <- anti_join(dat_s, dat_train, by="the_geom")
null <- lm(assess_val ~ 1, data = dat_train)
full <- lm(assess_val ~ age+BLDG_METRE+FIREPLACE, dat_train)
forward <- step(null, direction = "forward", scope = formula(full))
fwd_summary <- forward$anova #Note: best model has smallest AIC
coef(forward_model)
```

3 Predictions are Random Variables

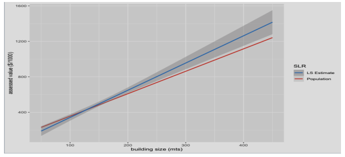
Estimated Prediction and LR change from sample-to-sample. Confidence intervals take into account the sample-to-sample variation of the predictions and regression params.

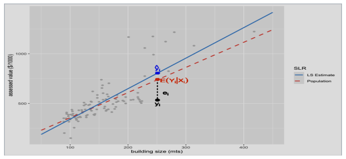
3.1 Confidence intervals for prediction (CIP)

A 95% confidence interval for prediction is a range that with 95% probability contains the average value of a house of this size. If we take a different sample, we get: different estimates, different fitted lines, and different predictions!

```
df_lm <- lm(col1 ~ col2 * col3, df)
CI_col3 <- tidy(df_lm, conf.int = TRUE, conf.level = 0.95) |>
  subset(term == "col2", select = c(conf.low, conf.high))
predict(df_lm, newdata = data.frame(col2 = 45, col3 = "Florida"),
  interval = "confidence", level = 0.95)
dat_cip <- dat_s|>select(col1,col2)|>
  cbind(predict(lm_s,interval="confidence",se.fit=TRUE)$fit)
```

We are 95% confident that the mean value of the population parameter (like mean etc) lies between the lower and upper confidence limits. A 90% confidence interval is narrower than a 95% confidence interval.





3.2 Prediction Intervals (PI) - predicted values CI

A prediction interval is a range that with probability 95% contains the actual value of a house of this size The predicted value $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ also approximates, with uncertainty, an actual observation $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. The uncertainty comes from the estimation and from the error term that generates the data.

```
predict(df_lm, newdata = data.frame(col2 = 45,col3 = "Florida"),
  interval = "prediction",level = 0.95)
dat_pi <- dat_s|> select(col1,col2)|>
  cbind(predict(df_lm,interval="prediction"))
```

With 95% confidence, the value of the response variable for a new observation at $X = x_0$ lies between L and U.

3.3 Conclusion

Confidence intervals for prediction account for the uncertainty given by the estimated LR to predict the conditional expectation of the response.
Prediction intervals account for the uncertainty given by the estimated LR to predict the conditional expectation of the response, plus the error that generates the data.
PI are wider than CIP, both are centered at the fitted value.

4 Potential problems in LR

Linearity By linear in a LR, we mean that variables (or functions of them) are multiplied by a coefficient and then sum together. fitting a quadratic function using LR. If not linear, conclusions are flawed, accuracy compromised.
Normality of the error term Least squares (LS) estimation does not depend on any normality assumption. However, many of the inference results given by lm do.
Quantile-Quantile (Q-Q) plots compare the quantiles of Normal Dist with the empirical quantiles of the standardized residuals. If error term is Normal, we expect most of the quantiles of both distributions to be over the 45 degree line in the plots.

```
dat_s <- sample_n(dat, 1000, replace = FALSE)
lm_large <- lm(assess_val ~ BLDG_METRE, dat_s)
plot(lm_large,2)
```

Equal variance AKA homogeneity, homoscedasticity, or constant variance assumption. In residual plot:you don't want to see a funnel effect: more variation for larger fitted values.
Usually variable transformations can be used to address this issue.
Mulitcollinearity some (or all) of the explanatory variables are linearly related! LS estimators are very "unstable" and the contribution of one variable gets mixed with that of another variable correlated with it. **Diagnosing:** checked using pairwise plots or measured through the variance inflation factors (VIF)
Post-Selection Inference arises when using LASSO for variable selection with the goal of later conducting inference on the selected model. To mitigate these issues, you can perform model selection on the training data, and use the testing data to conduct statistical inference.
Confounding factors refers to a situation in which a variable, not included in the model, is related with both the response and at least one covariate in the model. Not known factors not present

5 Logistic Regression Parameters

Estimated parameter $\hat{\beta}$ in LR by minizing MSE. Normality of errors is not required to estimate coefficients using ordinary least squares (OLS). Each Y_i follows a Bernoulli distribution with success probability p_i Assumptions of homoscedasticity and normality are violated in this case. Solution is to use MLE which leads to logistic regression. We can use the function glm() with the argument family $=$ binomial to get the estimates.

```
log_model <- glm(as.factor(binary_col)~col2,dat,family = binomial)
tidy(log_model, exponentiate = TRUE) #Log -> original scale
odds <- round(exp(predict(log_model,tibble(col1 = 5000), type =
    "link")),3) #gives prediction exponentiate
probability <- odds / (1 + odds)
```

AKA **odds ratio**. **Interpretation:** For a one-unit increase in X_i, the odds that $Y_i = 1$ increase by a factor of $e^{\hat{\beta}_1}$.

5.1 Inference

We can determine whether a regressor is statistically associated with the logarithm of the response's odds through hypothesis testing.

Wald statistic$z_j = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)}, \quad H_0 : \beta_j = 0, \; H_a : \beta_j \neq 0.$

Based on our sample, we reject H_0 (p-value $\approx$ 0), indicating that `fast_food_spend` is statistically associated with the log-odds of `heart_disease`.