Sample Material;

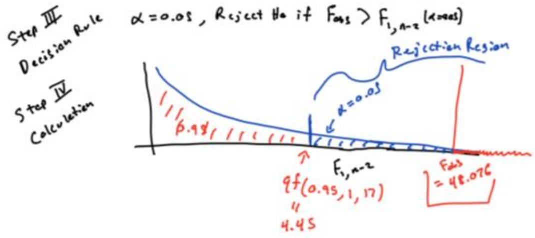
- Qualitative support for causation –
  - *Strength:* We do not mean a high correlation or a small p-value but that β̂ is large in practical terms.
  - *Consistency:* A similar effect has been found for different subjects under different circumstances at different times and places.
  - *Specificity:* The supposed causal factor is associated mostly with a particular response and not with a wide range of other possible responses.
  - *Temporality:* The supposed causal factor is determined or fixed before the outcome or response is generated.
  - *Gradient:* The response increases (or decreases) monotonically as the supposed causal variable increases.
  - *Plausibility:* There is a credible theory suggesting a causal effect.
  - *Natural Experiment:* A natural experiment exists where subjects have apparently been randomly assigned values of the causal variable.
- **Parameter estimation;** Once we have settled upon a model, we must decide how to estimate the parameters of that model. Generally, the method of least-squares is employed for this purpose.
  - Least squares estimation - Specifically, the method of least-squares starts by assigning a distance function (below) which measures the discrepancy between any line and the observed dataset. The least-squares estimates of the true parameters β_0 and β_1 are then those values of b_0 and b_1 which minimize the distance function d;
    - $d(b_0, b_1) = \sum_{i=1}^{n}(Y_i - b_0 - b_1 x_i)^2$
    - In doing the above the resulting fitted regression line, $Y = \beta_0 + \beta_1$ will be "close" to all the observed data points in some sense.
      - Once done, we define the fitted value of each data point as;
        - $\hat{Y}_i = E(Y_i|x) = \beta_0 + \beta_1 x$
      - And associated residual value by;
        - $e_i = Y_i - \hat{Y}_i$
      - We want to minimize the errors;
        - $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - (\beta_0 + \beta_1 x_i))^2$
      - Differentiate in respect to $\beta_0\ and\ \beta_1$ to get the answers of each. These estimators are unbiased which means their sampling distribution will be centred at the true value (population);
        1. $\frac{\partial d}{\partial \beta_0} = -2 \sum(Y_i - \beta_0 - \beta_1 x_i) = 0$
           - $\beta_0 = \bar{y} - \beta_1 \bar{x}$
        2. $\frac{\partial d}{\partial \beta_1} = -2\, x_i \sum(Y_i - \beta_0 - \beta_1) = 0$
           - $\beta_1 = \frac{S_{xy}}{S_{xx}}$
             1. $s_{xy} = \sum x_i y_i - n\bar{x}\bar{y}$ or $s_{xy} = \sum(x_{i-}\bar{x})(y_i - \overline{y)}$
             2. $s_{xx} = \sum x_i^2 - n\bar{x}^2$ or $s_{xx} = \sum(x_{i-}\bar{x}_i)^2$
             3. $Sample\ variance\ of\ X = S_x^2 = \frac{S_{xx}}{n-1}$
             4. $Sample\ variance\ of\ Y = S_y^2 = \frac{S_{yy}}{n-1} = \frac{SST}{n-1}$
             5. $Sample\ covariance\ of\ X\ and\ Y = \frac{S_{xy}}{n-1}$
        3. Therefore we can write the following;
           - $\bar{Y} = \beta_0 + \beta_1 \bar{X}$
        4. Estimated of these are unbiased and thus true for the population which is why we can write β_0 instead of $\widehat{\beta_0}$ because our sample is true for the population
        5. Variances of estimators $\beta_0\ and\ \beta_1$ which are the standard erros in regression ourput under square root;
           - $V(\hat{\beta}_o) = \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}})$
           - $V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$
- **Further discussion of outliers (Faraway chapter 6);**
  1. Two or more outliers next to each other can hide each other.
  2. An outlier in one model may not be an outlier in another when the variables have been changed or transformed. You will usually need to reinvestigate the question of outliers when you change the model.
  3. The error distribution may not be normal and so larger residuals may be expected. For example, day-to-day changes in stock indices seem mostly normal, but larger changes occur from time to time.
  4. Individual outliers are usually much less of a problem in larger datasets. A single point will not have the leverage to affect the fit very much. It is still worth identifying outliers if these types of observations are worth knowing about in the particular application. For large datasets, we need only to worry about clusters of outliers. Such clusters are less likely to occur by chance and more likely to represent actual structure. Finding these clusters is not always easy.
- **Final discussion of assumptions underling the model;** we can order the four assumptions in following order from most important to least important.
  1. The systematic form of the model - If you get this seriously wrong, then predictions will be inaccurate and any explanation of the relationship between the variables may be biased in misleading ways.
  2. Dependence of errors - The presence of strong dependence means that there is less information in the data than the sample size may suggest.
  3. Nonconstant variance - A failure to address this violation of the linear model assumptions may result in inaccurate inferences (testing and confidence interavals)
  4. Normality - This is the least important assumption. For large datasets, the inference will be quite robust to a lack of normality as the central limit theorem will mean that the approximations will tend to be adequate.
- **Interpretation of the Regression Coefficients**
  - Interpretation of $\hat{\beta}_0$;
    - The expected value of Y when x = 0.
    - In some cases (eg heart rate) having x=0 may be stupid because here the person would be dead
  - Interpretation of $\hat{\beta}_1$;
    - The expected change in the value of Y for a unit change in x
- **Model selection;** In general, we will favor models with less explained variation meaning error.
  - Smaller MSE ($\hat{\sigma}^2 = s^2$) - Seen in the ANOVA table (Mean Squared Residual Error)
  - Smaller RSE ($\hat{\sigma} = s$) - Seen in regression summary
- We may also base our decision of the model with the highest adjusted coefficient of determination
- **F-tests;**
  - We can use F-tests to examine the drop is $\hat{\sigma}$ is statistically significant.
  - Note, that $\hat{\sigma}$ is on the same scale as Y , so we cannot compare models on different scales.
  - We cannot compare Y and log(Y) because they are not nested.
- $\boldsymbol{R^2}$**;** we can use R^2 to compare models and even models with different scales because it is a standardized measure

- $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = 1 - \frac{SSE}{SST}\ OR\ R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{SSR}{SST}$
- Issues;
    - No obvious point of comparison. How large should R^2 be?
    - Does not protect from overfitting the data. Every additional covariate will increase R^2
- Overfitting; is when the model too closely follows the observed data and thus would be poor for prediction of new data
- As was mentioned, when we add covariates we lower the SSerror, which then increases R^2.
- This suggests that R^2 is useless for choosing between a model with p covariates compared to a model with p + 1 covariates. P can be written as d.
- We want approaches that penalize for adding extra variables.

- **Methods which penalize for adding extra variables;**
    - Adjusted R^2 – Large is good;
        - $R^2_{adj} = 1 - \frac{\frac{SSE}{n-p}}{\frac{SST}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right) \times (1 - R^2)$
    - $Press_p$ – small is good;
        - Consider the press residuals; $\hat{\epsilon}_{(i)} = Y_i - \hat{Y}_{(i)} = \frac{\hat{\epsilon}_{(i)}}{1-h_{ii}}$
        - $Press_p = \sum_{i=1}^{n} \hat{\epsilon}_{(i)}{}^2 = \sum_{i=1}^{n}(\frac{\hat{\epsilon}_{(i)}}{1-h_{ii}})^2$
        - Sadly, the leaps library does not have the PRESS statistic. As is done in the Brick, you could code up all the possible models, fit them and calculate the statistic!
    - Mallows C_p – smaller is better;
        - Based on the idea that mis-specifying the model will create bias in the estimate of σ^2 , and that over-fitting will inflate the variances for predictions.
        - Lengthy arguments and derivation of can be found on pages 35-36 in Chapter 2 of the Brick.
            - $C_p = p + \frac{(n-p)(s^2-\hat{\sigma}^2)}{\hat{\sigma}^2}$
        - This requires some "independent" estimate of $\hat{\sigma}^2$.
        - In practice, we often just use $\hat{\sigma}^2 = s^2$ from the full model.
        - We prefer models, where $C_p = p$, but because we used $s^2 - \hat{\sigma}^2$ from the full model, for the full model $C_p = p$ is guaranteed.
        - This suggests, we want a smaller model, compared to the full model where C_p is close to p
        - You will see that different author's may arrange the algebra differently. For example Faraway has:
            - $C_p = \frac{SSEp}{\hat{\sigma}^2} + 2p - n$
        - Note: $E[SSEp\ ] = (n - p)\hat{\sigma}^2$, so $E[C_p] \approx p$.
        - A model with a bad fit will have Cp much bigger than p

- **Hypothesis test using our F-statistic/F-Test;**

1. $H_0: \frac{\sigma^2_{Y|X}}{\sigma^2} = 1$
1. $H_1: \frac{\sigma^2_{Y|X}}{\sigma^2} > 1$
    a. We are testing to see if the extra variation of the model including x gives us more variation. We set alternative hypothesis to > 1 to see if we are getting any of that extra variance or we are just getting the normal variance
    b. Overall F-test – done from the initial output of the linear model also gives a F-stat which is a quick test to see if any of the overall covariances are useful. In this case both F-stats match as we are using SLR
2. $Test\ statistic = F = \frac{MSregression}{MSresidual/error}$
    a. Where MS regression is our top portion of the hypothesis and MS residual is the bottom part
    b. With degree of freedom for the numerator=K=1 (in this case - SLR)
    c. With degree of freedom of n-p = n-2 (in this case - SLR)
3. Decision rule –
    a. $\alpha = 0.05$
    b. Reject H_0 if $F_{observation} > F_{1,n-2}\ at\ \alpha = 0.05$
4. Calculation –



    a.
    b. We can see that our observed test statistic pf 48.076 is in the rejection region so we reject the null hypothesis
5. Conclusion –
    a. As $F_{observation} = 48.0786 > F_{1,n-2} = 4.45$ we can reject the null hypothesis
    b. Also we can see the $p - value < \alpha = 0.05$ so we can reject the null hypothesis.
6. Interpretation - the proportion of the variance in Y explained by the larger model (including x) is significantly larger than the error variance.
    a. P-value is the probability of seeing your observed tests statistic or something more extreme where more extreme is measured by the alternative hypothesis under the assumption the null hypothesis is true