Study Guide for Final Exam

Chapter 2: Simple Regression Model

- β_1 : the *slope parameter*, the relationship between y and x, holding other factors in u fixed.
    - Linearity implies that a one-unit change in x has the same effect on y, regardless of the initial value of x.
    - The estimated function tells us how the expected or average value of y changes with x.
- u : the *error term*, the disturbance in the relationship, “unobserved,” or all factors other than the x that affect y.
- $E(u|x) = 0$, called *zero conditional mean assumption*.
    - The average value of the error term is the same across all slices of the population.
    - Means u is *mean independent* of x.
- The part of y explained by coefficients is the *systematic* part and the error term is the *unsystematic* part of y.
- Deriving the Ordinary Least Squares (OLS) estimates:
    - To estimate the parameters, we obtain a random sample of size n from the population.
    - The parameters give us a *fitted value* of y, and we choose the parameters to minimize the sum of the squared *residuals*: difference between the population value y_i and its fitted value y hat. If residual is positive, the line underpredicts y. The data points must not actually lie on the OLS line.
    - OLS decomposes each y into a fitted value and a residual, which are uncorrelated in the sample.
- Algebraic properties of OLS statistics:
    - $\Sigma\hat{u} = 0$
        - Means that the sum of OLS residuals is 0. Also means the sample average of the residuals is 0.
    - $\Sigma x\hat{u} = 0$
        - Means the sample covariance between the regressors and the OLS residuals is 0.
    - (x bar, y bar) is always on the OLS regression line.
        - Means if you plug in the average x, the predicted value is the average value of y.
- SST = SSE + SSR or SCT = SCE + SCR
    - Means total sum of squares is equal to the sum of the explained sum of squares and the sum of squared residuals.
    - SST measures how spread out the y are in the sample, which can always be expressed as a sum of the explained and unexplained variation.
- Goodness of fit:
    - We need a way to summarize how well the OLS regression line fits the data.
    - Always between zero and one.
        - SSE/SST or 1-SSR/SST
        - Ratio of explained variation to total variation
        - A lower value indicates a poorer fit.
- Statistical properties of OLS:
    - Think of it as choosing n sample values, and given those values, you obtain a sample on y.
    - Unbiased:
        - Means the sampling distribution of the estimator is centered around the population/true value.
        - Four assumptions are needed: linear in parameters, random sampling, variation in the explanatory variable, zero conditional mean. If any of these fail, unbiasedness fails.
        - Zero conditional mean: $E(u|x) = 0$
            - Biggest concern of simple regression analysis is that x may be correlated with u → *spurious correlation*: finding a relationship between y and x that is really due to other unobserved factors.
            - Unbiasedness is a feature of the sampling distribution of the βetas, not the estimate obtained for a given sample.
            - Any variable not accounted for in the form of an X goes to the U.
            - This fails in 3 circumstances:
                - The functional relationship is misspecified.
                - An important factor is omitted (another form of misspecification).
                - Measurement error in an explanatory variable.
            - When this assumption holds, we say we have *exogenous explanatory variables*. When X is correlated with U for any reason, we say X is an *endogenous explanatory variable*.

- Variances of the OLS estimates:
  - Tells you how far we can expect the estimate to be from the population value, on average. Amongst a group of estimators, variance helps us pick the best, because we want the one with the lowest variance.
  - Assume homoscedasticity: Var(u|x) = σ^2 which plays no role in showing that the estimates are unbiased. Also called the error variance, because it is the unconditional expectation of the u.
  - Says the variance of y, given x, is constant.
  - Square root of the variance is the standard deviation.
  - When variance depends on x, the error term is said to exhibit heteroscedasticity.
    - Example: variability in wage is likely not constant across education levels, because people at super low education levels probably all make minimum wage.
  - $\text{Var}(\beta_1) = \sigma^2/\text{SST}_X$
    - The larger the error variance, the more difficult it is to precisely estimate the parameters.
    - More variance in X (example: as the result of an increased sample size) will decrease the variance of the β.
  - Note: σ^2 cannot be known for sure. Have to estimate it. *Errors* are never observed (u), but *residuals* can be computed from the data (u hat). U hat is not equal to u, but the difference between them has an expected value of 0. Unbiased estimate of σ^2 is SSR/(n-2). Square root of this is the *standard error of the regression.*
- Remember, statistical properties have nothing to do with a particular sample but rather the property of estimators when random sampling is done repeatedly.

- Quick summary of the *Gauss-Markov assumptions*:
  - Linear in parameters.
  - We have a random sample.
  - Sample variation in the explanatory variables: means the sample outcomes on x are not all the same value.
  - Zero conditional mean: E(u|x) = 0.
    - If the four above hold, we can say we have unbiased estimators of the population parameters.
  - Homoscedasticity: Var(u|x) = σ^2, the error u has the same variance given any value of the x.
    - If the five above hold, we say we have the best linear unbiased estimator (BLUE), which means the estimator has the lowest variance.
- Incorporating nonlinearities in simple regression:
  - Log-level: when the y is log(y).
    - Multiply the coefficient by 100 → x% change in y.
    - Implies a constant percentage return to an increase in x, or an increasing return to x.
    - Sometimes called the *semi-elastic*ity of y with respect to x.
  - Log-log: both x and y are in log form → *constant elasticity model*.
    - 1% increase in x results in β_x % increase in y.
    - β_x is the elasticity of y, with respect to x.
  - Level-log: log(x) and then normal y.
    - An increase in 1% of x results in a $\beta 1/100$% change in y.
- Units of measurement and functional form.
  - Dependent variable:
    - If you multiply the dependent variable by a constant, the OLS intercept and slope estimates are also multiplied by c.
  - Independent variable:
    - Multiplying an independent variable by a nonzero constant c → divide that slope coefficient by c.
    - Generally, does not change the $\beta_{0.}$
- Changing units of measurement does not change measures of goodness of fit.

Chapter 3: Multiple Regression

- Drawback of simple regression: it's hard to draw ceteris paribus conclusions about how x affects y (don't control for anything) → multiple regression allows you to explicitly control for other factors that simultaneously affect y.
  - Any variable you add to the regression is taken out of the u and explicitly put into the equation.
  - With simple regression, anything that is not in the regression explicitly, aka in the u, must not be correlated with the variable in the regression. Otherwise, it is biased.

- Quadratics:
    - Say you have the model $y = \beta_0 + \beta_1X_1 + \beta_2X_1^2 + u$.
    - Now, the change in y caused by increasing X_1 is no longer β_1. The marginal effect depends on the level of X_1 now. So, $\Delta y/\Delta x = \beta_1 + 2\beta_2X_1$.
- Once again, assumption about the error is that, for any values of the x's in the population, the average of the unobserved factors is equal to 0 + any other factors affecting y are not related on average to the explanatory variables: $E(u|X_1 \dots X_K) = 0$.
    - For this to be true, cannot have any key variables be omitted from the equation.
    - If this is true, OLS is unbiased.
- If the multiple regression model contains k independent variables, it contains *k + 1 unknown population parameters*.
- OLS *fitted/predicted values*:
    - Each observation can have a predicted value, which is obtained by plugging the values of the independent variables for the observation into the estimated equation.
    - Several properties worth mentioning (that are just extensions of the simple linear regression):
        - The average of the residuals is 0 → the average population y is equal to the average y hat.
        - The sample covariance between each X and the residuals is 0 → covariance between OLS fitted values and OLS residuals is 0.
        - Point of (all the average X's in the population, population y average), is always on the OLS regression line.
- The only way the β_1 from single regression is equal to the β_1 from multiple regression is if the partial effect of X_2 on y is 0 or X_1 and X_2 are uncorrelated in the sample.
- In considering the "goodness of fit" or R^2, it is worth noting that the R^2 will never decrease when you add a variable to a regression → not the best tool for determining whether a variable should be added or not → instead focus on if the explanatory variable has a nonzero partial effect on y (hypothesis testing in chapter 4).
- New assumption here:
    - No *perfect collinearity*: none of the independent variables is constant (true for the single linear regression too) AND there are no exact linear relationships among the independent variables.
        - Note: can have a substantial amount of correlation.
        - X^2 is fine because, though it is an exact function of X, it is not a linear function of X.
- Including irrelevant variables/*over-specifying the model*:
    - One or more independent variables included in the model has a population coefficient that is not statistically different from 0.
    - After the other variables are controlled for, this variable has no effect on y. May or not may not be correlated with the other explanatory variables (doesn't matter either way).
    - Does not affect the biasedness of any of the estimators but does increase the variances.
- Omitted variable bias/*under-specifying the model*:
    - Can derive the bias caused by the omitted variable through the process of *misspecification analysis* (missing a variable is a form of misspecification bias).
    - We say that we estimate a model with an error term $v = \beta_{O*}$OmittedX + u.
    - The relationship between the included and omitted X is $\beta1_{\text{included variable parameter with bias}} = \beta1_{\text{the beta we would estimate if we were able to include the omitted variable}} + (\beta_{\text{omitted variable}})\delta1$
        - The bias of β_1 is equal to $\beta_2\delta_1$, called the *omitted variable bias*.
        - δ_1 expresses the sample covariance between X_1 and X_2.
            - Has the same sign as the correlation between the variables.
        - β_2 is the theoretical estimate of the parameter on the omitted variable.
        - The bias of β_1 thus depends on the signs of both factors. If both are positive or negative, the β_1 has positive bias. If the signs are not the same direction, the β_1 has negative bias.
            - Positive = *upward bias* = bigger than it should be.
            - Negative = *downward bias* = smaller than it should be.
            - Note: not necessarily true of the specific estimate we obtained. Just means that if we collect many random samples, the average of those estimates will be bigger or smaller than the actual β.

        - Studying these two factors shows us when an omitted variable will not lead to bias: when the correlation between the two variables is 0, or when the effect of the omitted variable on the estimate (β_2) is 0.
    - When there are more variables in the model, correlation between a single explanatory variable and the error generally results in all estimators being biased.
- Variance of the OLS estimators:
    - In addition to knowing the central tendencies of the βetas, we want a measure of the spread in its sampling distribution. Before we do this, we must impose a homoscedasticity assumption: the error has the same variance given any values of the explanatory variables/for all combinations of variables.
        - $\text{Var}(\beta_1) = \sigma^2/\text{SST}_X(1-R^2_J)$, where R_J is the R-squared from regressing X on all other independent variables. From this equation, you can see what characteristics lead to a larger variance.
            - σ^2 : so, more noise in the equation makes it harder to estimate the partial effect of any one variable. This is unknown so we need to estimate it. Has nothing to do with sample size.
                - Estimate this with SSR/(n-k-1), where n-k-1 is the *degrees of freedom:* number of observations – number of parameters.
                - The square root is the *standard error of the regression*, or the estimate of the standard deviation of the error term.
            - SST_X (total sample variation in X): increasing this (through a larger sample size) will decrease the variance.
            - R^2_J : a higher value indicates that the variables are highly correlated, and results in a larger variance. High, but not perfect, correlation between two or more independent variables is called *multicollinearity.* Don't solve this by just dropping a variable though (leads to bias), just kind of have to increase the sample size and deal with it.
                - Also, keep in mind that the amount of correlation between X_2 and X_3 does not affect the $\text{Var}(\beta_1)$.
    - In misspecified models:
        - Variance will always be smaller in the misspecified model.
        - Variance can shrink as the sample size grows but bias does not.

Chapter 4: Hypothesis Testing

- Have to make the assumption of homoscedasticity to estimate the standard deviation of β with the standard error of the estimate because we need to estimate σ.
- Heteroscedasticity does not cause bias in estimation of the β but it biases estimates of the variance → standard errors are invalidated.
- Sampling distributions of the OLS estimators:
    - To perform statistical inference, we need to know the sampling distribution, which depends on the underlying distribution of the errors. We make the *normality assumption*: the population error u is independent of the explanatory variables and is normally distributed with zero mean and variance σ^2.
        - If we make this assumption, assumptions 4 and 5 are automatically true (mean independent and homoscedastic).
        - The normality assumption, in conjunction with the 5 assumptions discussed already (*Gauss-Markov assumptions*), form the *classical linear model (CLM) assumptions*.
        - Normality of the errors is not always a fair assumption, especially when y takes on only a few values. But with large enough sample sizes this isn't really a big problem.
        - If this is true, any linear combinations of the βs are also normally distributed.
- Hypothesis about a single population parameter: *t-test*
    - $(\beta\text{hat} - a_j)/\text{se}(\beta\text{ hat}) \sim t_{n\text{-}k\text{-}1}$
        - a_j is our hypothesized value of β.
    - Means that the difference between the estimate and the predicted value, divided by the standard error of the estimate, is distributed as a t function with n-k-1 degrees of freedom (dof).
    - Typical *null hypothesis*: H_0: β = 0 meaning that once all the other explanatory variables have been accounted for, that X has no effect on the expected value of y. Test against the *alternative hypothesis*:
        - One sided: H_1: β > 0 or H_1: β < 0.
            - Have to decide on a rejection rule → have to decide on a significance level: probability of rejecting H_0 when it is in fact true. Then, we take that value to the t table and look for the

      (1-significacne level)% distribution in a t table with n-k-1 dof. The value on the table is called the *critical value*: if $t > c$, we reject the null.
      - As the significance level falls, the critical value increases, so we need a larger t.
    - Two sided: H_1: $\beta \neq 0$.
      - We do not specific whether the effect of X is believed to be negative or positive → we look at the absolute value of the t statistic.
      - Since we now have a two-tailed test, we need to divide our significance level by two, then go to the 1-significane level percentile distribution. Thus, a 5% significance level requires the 97.5th percentile.
  - Weighs the size of β against its standard error to decide if it is sufficiently far away from 0.
  - Can have a t statistic that passes a one-sided test then fails a two-sided.
  - If you want to test hypothesis besides just β being around 0, you just subtract that value from the β hat first. Still compute the critical value the same and reject the null if $t > c$. Only difference is in how we calculate the t. [So, if $a = -1$, we do: (estimate + 1)/ standard error of the estimate].
  - Computing p-values for t tests:
    - Classical approach requires that we choose significance level ahead of time → a little arbitrary.
    - Instead, look at the *p-value*: given the observed value of the t statistic, what is the smallest significance level at which the null hypothesis would be rejected?
    - Value is always between 0 and 1 because it is a probability: the probability of observing a t statistic as extreme as we did if the null hypothesis is, in fact, true.
    - Small p-values are evidence against the null.
    - Regression packages mostly report two-sided p values → to obtain the one-sided just divide the p-value by 2.
  - Note: a t-stat can indicate statistical significance either because β is large or because the standard error is small, meaning that a variable may be statistically significant with a very small estimated effect (β).
- *Confidence intervals*:
  - Provide a range of likely values for the population parameter, rather than just a point estimate. If random samples were obtained over and over again, with β hat calculated each time, the unknown population value would lie in that interval for 95% of samples.
  - Calculated with: β hat +/- c*se(β hat)
    - c is the 97.5th percentile in a t distribution with n-k-1 degrees of freedom.
    - If dof > 120 we can just use 1.96 as the value.
    - A lower level of confidence → a lower percentile in the t distribution → a narrower confidence interval.
  - We reject the null only if a is not in the 95% confidence interval.
  - Also only works for one β at a time.
- Hypotheses about a single linear combination of the parameters:
  - Say, for example, that we want to test the null that $\beta_1 = \beta_2$. Cannot just use t stat → need to define a new parameter as the difference between the two values.
  - $\theta = \beta_1 - \beta_2$, then plug this value in for one of the betas.
- Testing multiple linear restrictions: *F test*
  - Used to test whether a group of variables has no effect on the dependent variable, once another set of variables has been controlled for.
  - Null hypothesis takes some form along the lines of H_0: $\beta_1 = 0$, $\beta_2 = 0$, $\beta_3 = 0$. This is an example of a set of multiple restrictions → called a *joint hypothesis test*. This test specifically has 3 *exclusion restrictions*.
  - The alternative is simple. H_A: H_0 is not true.
  - If at least one of the β's is different from zero (any or all), the alternative will hold.
  - Constructing the F test:
    - We know that the SSR increases anytime variables are dropped from the model. We want to see if this increase is large enough, relative to the SSR in the model with all the variables, to warrant rejecting the null.
    - Restricted model: model with all of the variables you are testing taken out of it (always has fewer parameters than the original model).
    - Unrestricted model: the original model, has $k + 1$ parameters.

        - Number of exclusion restrictions is called q. This is also the difference in the degrees of freedom for the restricted and unrestricted model.
        - Actual statistic: F = (SSR_r – SSR_{ur})/q all over SSR_{ur}/(n-k-1)
            - Because SSR restricted is always bigger than SSRur, the F is always non-negative.
            - Degrees of freedom in the numerator is equal to q. In the denominator, it is equal to n-k-1, or the unrestricted model's degrees of freedom.
            - F is distributed as an F random variable with (q, n-k-1) degrees of freedom. We reject the null when F is large enough.
            - Note: don't need to divide the significance level by 2 to find the percentile distribution on the F here. No differentiation between one and two-sided tests.
        - The R-squared version:
            - (R^2_{ur} – R^2_r)/q all over (1-R^2_{ur})/(n-k-1)
            - Cannot be applied for testing all linear restrictions.
    - If we reject the null, we say the variables are jointly statistically significant. That does not allow us to say which of the variables has the partial effect on y. If the null is not rejected, you are usually justified in dropping the variables from the model.
    - F is useful for testing exclusion of a group of variables when they are highly correlated (makes it hard to uncover the partial effect of each variable and results in worse individual t-statistics).
    - Computing p-values for an F test:
        - Similar interpretation as before: the probability of observing a value of F at least as large as we did given than the null is true.
    - Relationship between F and t stats:
        - If you calculate an F stat for just excluding one variable, it is equal to the square of the corresponding t stat.
- Testing the *overall significance of a regression*:
    - We can use the F test for the ultimate exclusion restriction: H_0: X_1, …, X_k do not help to explain y. States that none of the explanatory variables has an effect on y. All of the slope parameters are zero.
    - We drop all of the variables from the model and the restricted version is literally just y = β_0 + u. The R^2 is 0. None of the variation in y is being explained because there are no explanatory variables.
    - R^2/k all over (1-R^2)/(n-k-1), where the R^2 is just from the normal regression.

Chapter 6: Quadratics, Interaction Terms, Goodness of Fit, etc.

- More on models with quadratics:
    - If the coefficient on x is positive and negative on x^2: there is a positive value of x where the effect on y changes from positive to negative. Results in a parabolic shape.
    - One cannot simply look at the coefficient on the quadratic term and make decisions about its importance based on its magnitude.
    - If the coefficients on the level and squared terms have the same sign, there is no turning point in the x. Magnitude of the effect (either positive or negative) gets larger as x gets larger.
- Models with interactions terms:
    - When the partial effect of an independent variables on the dependent variable depends on the magnitude of another independent variable, we say that there is an interaction effect.
    - Interaction effects are captured by multiplying the relevant independent variables.
- Adjusted R-squared: Imposes a penalty for adding additional independent variables to a model.
    - Can be used to help us decide between non-nested models (where neither model is a special case of the other model).
    - However, we cannot use this to choose between different functional forms for the dependent variable (e.g., normal y vs log(y)); it does not tell us anything about which model fits better because they are fitting two different dependent variables.

Chapter 7: Dummy Variables and Chow

- Qualitative factors need to be included in a regression as well: gender or race of an individual for example. Often come in the form of binary information → what is known as *dummy variables* (event is assigned either a 0 or a 1).
- A single dummy independent variable:
    - Example: wage= β_0 + $\beta_1$feamle + $\beta_2$education

        - Results in an *intercept shift*, where both men and women have β_0 slope but the intercept for women is $\beta_0 + \beta_1$. Because the difference is not dependent on the amount of education, the two lines are parallel.
        - If $\beta_1 \neq 0$, men will earn a fixed amount different than women.
        - β_1 represents the difference in wage between men and women, given the same amount of education.
        - Here, men are the *base or benchmark group*: the group against which comparisons are made.
        - If we wanted to test for discrimination, we would just check the t-stat on β_1 to see if there is a statistically significant difference in men and women's wages with education held constant.
- Interpreting coefficients on a dummy when the dependent variable is log(y):
    - Just multiply the coefficient by 100 and it's still the %difference in y, holding all other factors fixed.
- Using dummy variables for multiple categories:
    - Say we want to estimate differences among married men, married women, single men, and single women. We need to select a base group that will not be included in the regression.
    - Here, the overall intercept will be common to all groups so we can ignore it in finding differences. All of the other β's will tell the difference between that group and the base group.
- Using dummy variables to incorporate ordinal information:
    - Example: if the category is 0-4, where 0 is worst and 4 is best. These things have ranking meaning, but not a specific quantitative meaning attached to them.
    - If you put each credit ranking into the model as its own dummy variable you allow more flexibility in movement between rankings. Moving from 0 to 1 can have a different effect than moving from 3 to 4.
    - For something like rankings of schools, where there are too many values to include a dummy for each one, you can break down the rank into different categories.
- Interactions among dummy variables:
    - Example: if we didn't want to do four separate categories as we did in the married men and women one above, we could add an interaction term between female and married → $y = \beta_0 + \beta_1 female + \beta_2 married + \beta_3 married*female$.
    - Allows us to explicitly test the null that the gender differential does not depend on marital status (β_3 significance).
    - Only have β_3 as a factor in the regression if both married and female =1.
- Allowing for different slopes:
    - Interacting dummy variables with explanatory variables that are not dummy variables.
    - Example: $y = \beta_0 + \beta_1 female + \beta_2 education + \beta_3 female*education + u$
        - Intercept for women is $\beta_0 + \beta_1$ and the slope is $(\beta_2+\beta_3)$education.
        - Note: the t statistic may go down on female now that you add the interaction term because they are highly correlated.
- Testing for differences in regression functions across groups:
    - Used to test the null hypothesis that two groups follow the same regression function (has k+ 1 restrictions) against the alternative that one or more of the slopes differ across the groups (that can be clearly divided in the data, like men and women).
    - Basically, creates a model where the intercept and every single slope can be different across two groups. The equivalent of putting an interaction term on every single β.
        - If there aren't that many explanatory variables, you can do this and then compute the F statistic.
        - If there are lots of explanatory variables, it is helpful to do a different thing.
    - Compare the SSR of the pooled group (include observations from both subgroups) to the SSR of the two individual regressions (run the regression but only for men in the sample then only for females). The sum of the two separate regressions is the same as the SSR unrestricted (with all the interaction terms and the group dummy variable).
    - Null hypothesis is that the data can be represented with one line.
    - Actual math → *Chow statistic*: $F = [SSR_P - (SSR_1 + SSR_2)]/(SSR_1+SSR_2)$ multiplied by $[n-2(k+1)]/(k+1)$.
        - Only valid under homoscedasticity.
        - Limitation: null allows for no differences at all between the groups (including an intercept difference).
        - To allow for an intercept difference, estimate the SSR_P with a regression that only has an intercept shift and then switch k+1 to k.

- *Linear probability mode*l:
    - Can we use multiple regression to explain a qualitative event, rather than quantitative information?
    - Use y as a binary outcome, equal to 0 if the event does not occur and 1 if it does.
    - Can no longer interpret β as the change in y given a one unit increase in x. Instead, call it the *response probability*: β measures the change in probability of success (y=1) when X changes, holding all other factors fixed.
    - β_0 is now the predicted probability of success when each X is set to 0.
    - Shortcomings:
        - Can get predictions that are above 1 or below 0.
        - Probability cannot be linearly related to the independent variables for all their possible values (because a change from 0 to 10 or whatever could make probability go up by 2).
        - Model usually works best for values of the independent variables that are near the averages in the sample.
        - There must be heteroscedasticity in the model because $\text{Var}(y|x) = p(x)[1-p(x)]$. Does not cause bias but can cause difficulties in justifying the usual t and F statistics.
    - If we have a discrete dependent variable that does not just take on the values of 0 and 1, we can consider the effects of the x on the AVERAGE value of y. So, for example, if β equals .09, we can say that the average change for an individual is .09, or that if all people in a group of 100 increased x by 1, there will be an increase of 9 in y among them.