Odds and Ends

Prof. Bisbee

Vanderbilt University

Lecture Date: 2023/11/30

Slides Updated: 2023-12-23

Agenda

1. Regression Diagnostics

2. Goodness of Fit

3. Overspecification

4.

Regression Diagnostics

- We have so many assumptions at this point! (How many can you list?)

  1. Linearity
  2. I.i.d. random sample
  3. Non-zero variance / no multicolinearity
  4. Zero-conditional mean
  5. Homoskedasticity / spherical errors
  6. Normally distributed errors (small samples)

- How can we be confident in these assumptions? **Diagnostics** (to an extent)

Regression Diagnostics

- Running example of two DGPs
- DGP 1:

$$Income_i = 15 + 6 * Labor_i + 40 * PhD_i + u_i$$

where $u \sim \mathcal{N}(0, 5)$ and i.i.d. holds

- DGP 2:

$$Income_i = 15 + 6 * Labor_i + 40 * PhD_i + Labor_i^2 + u_i$$

where $u_i = 0.5 * Labor_i * e_i$, $e_i \sim \mathcal{N}(0, 5)$ and i.i.d. holds

- Note that our assumptions hold by construction in DGP 1, but not in DGP 2
  - Specifically, zero conditional mean holds only if *Labor* is mean zero
  - In addition, the errors are not homoskedastic

Regression Diagnostics

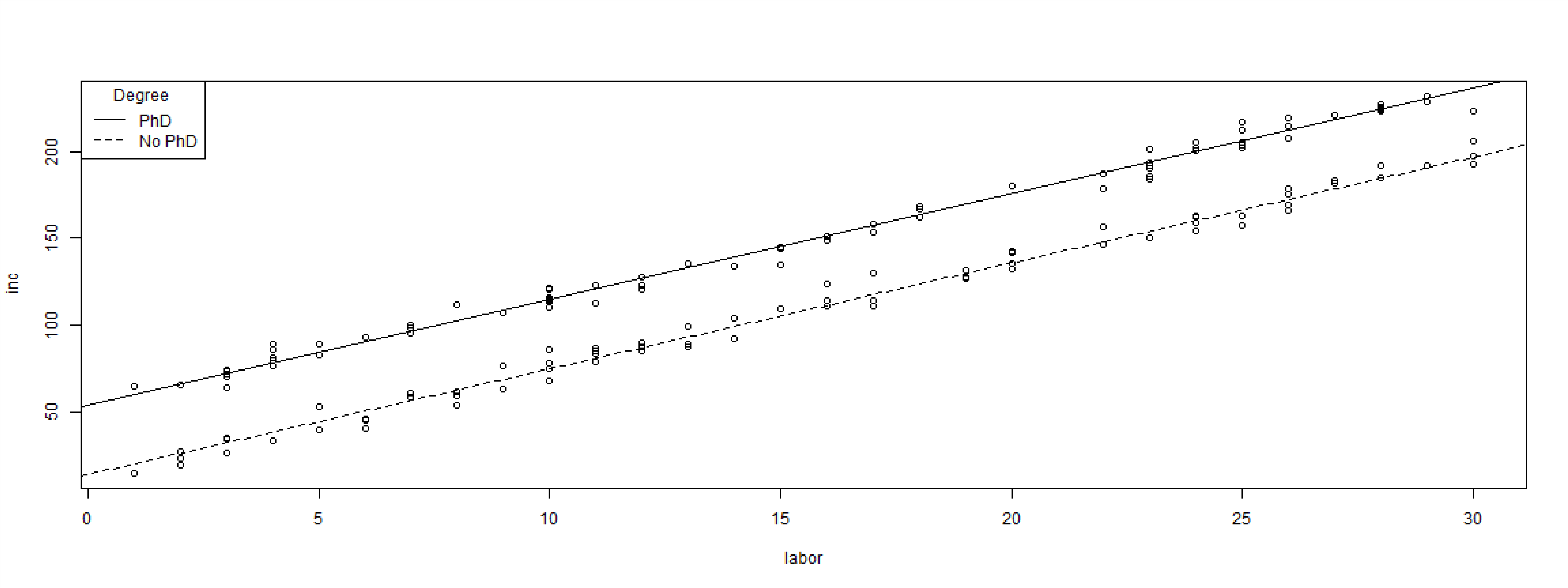
- Estimate with $Income_i = \beta_0 + \beta_1 Labor_i + \beta_2 PhD_i + u_i$

```
m1 <- lm(inc ~ labor + phd, data = data)
## Did we reproduce the truth?
summary(m1)
```

```
##
## Call:
## lm(formula = inc ~ labor + phd, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5052  -3.5249  -0.2899   3.1433  12.7515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.64262    0.94620   14.42   <2e-16 ***
## labor        6.11087    0.04866  125.59   <2e-16 ***
## phd         40.08627    0.83957   47.75   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression Diagnostics

- Visualize the results



Regression Diagnostics

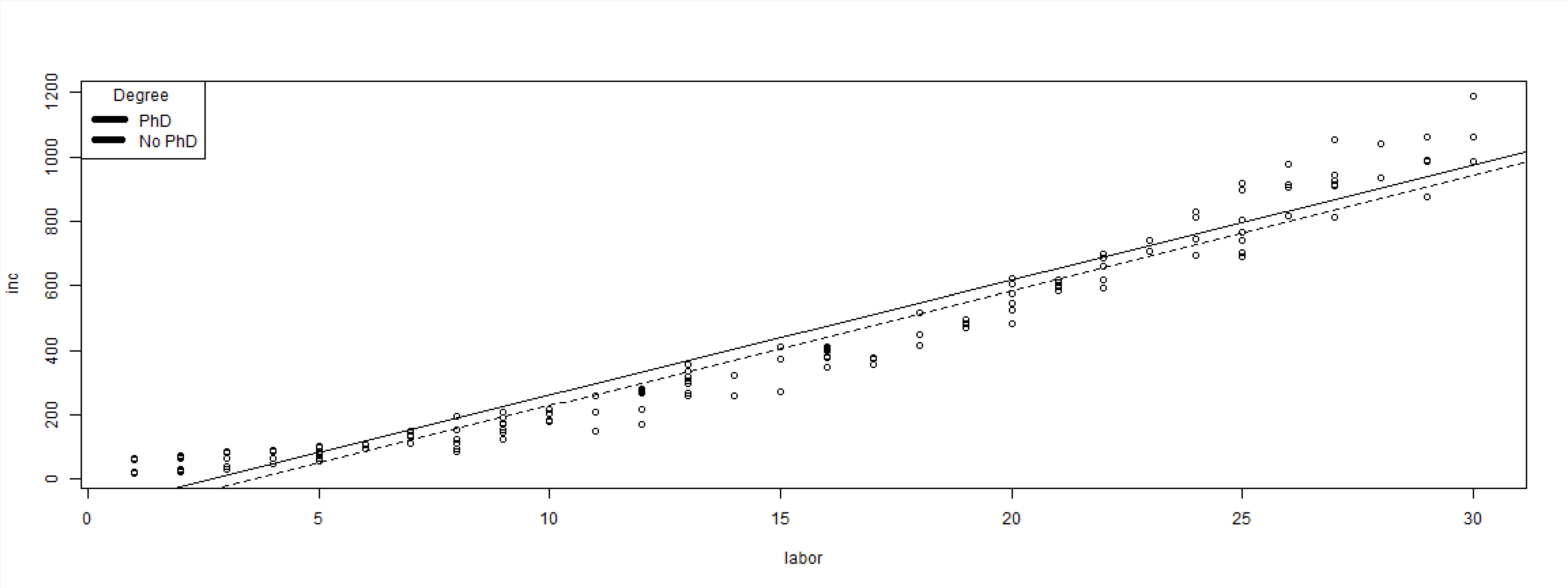
- What if we are in the DGP2 world?

```
m2 <- lm(inc ~ labor + phd, data = data2)

summary(m2)
```

```
##
## Call:
## lm(formula = inc ~ labor + phd, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -143.12  -55.84  -13.86   58.68  216.07
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -127.7281    13.1058  -9.746   <2e-16 ***
## labor         35.6213     0.6945  51.287   <2e-16 ***
## phd           33.4525    12.1629   2.750   0.0067 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regression Diagnostics



Regression Diagnostics

- We are doing a bad job because:

  1. Income is not a linear function of labor and degree
  2. Errors are not mean zero
  3. Errors are not homoskedastic

- We **know** all this because we simulated these data
- But in reality, we typically never know what the true DGP is...how can we be alerted to the fact something is wrong with our model?

Regression Diagnostics

- We can **look** at our residuals in a number of ways that inform us about our model's fit

1. Residuals vs. Fitted Values: Tells us if the data are roughly linear (smoother is roughly a horizontal line) and if there is heteroskedasticity (residuals are larger for some observations than for others)

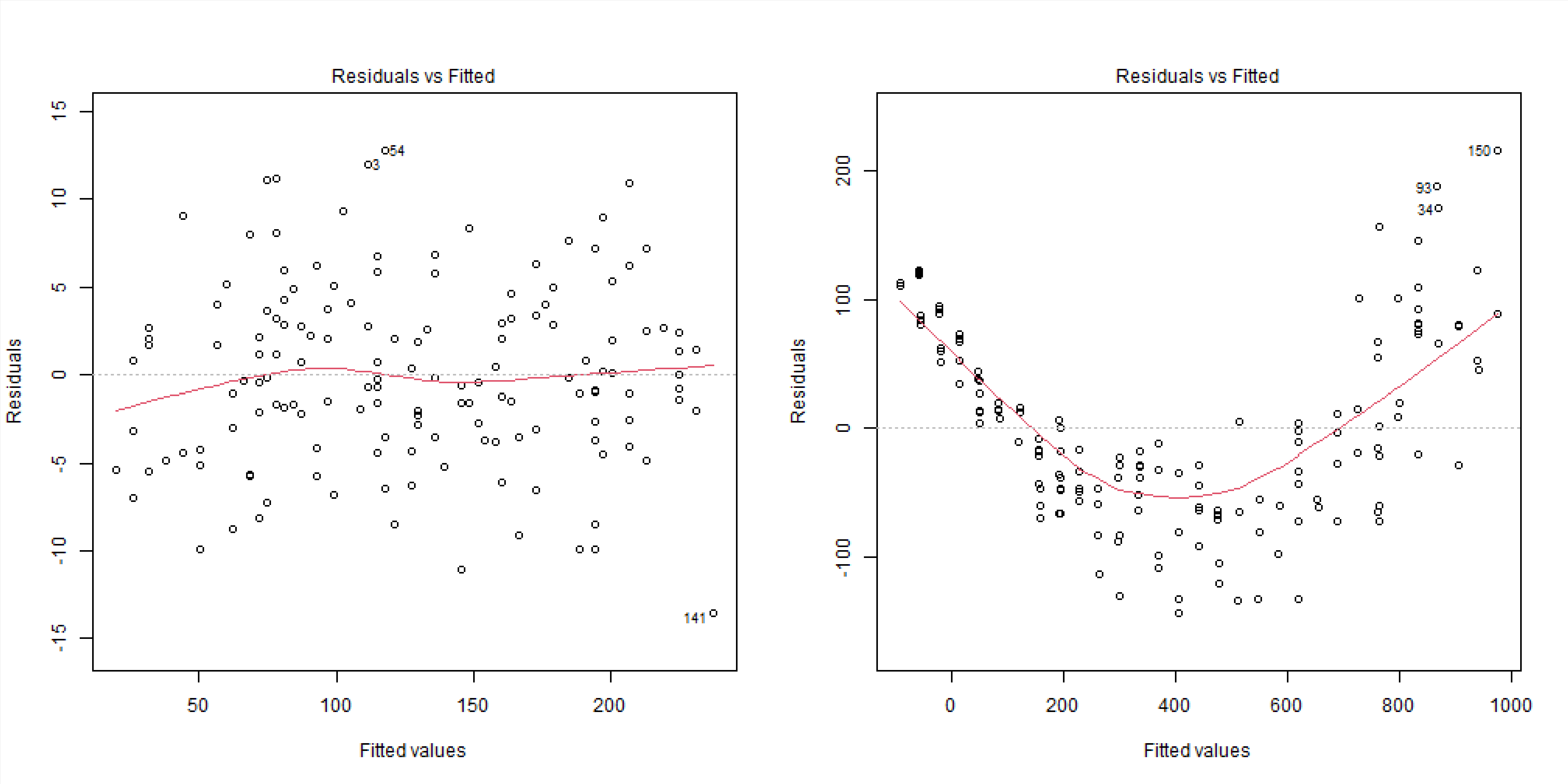
2. Normal Q-Q Plot: Compares quantiles of our observed residuals to quantiles of hypothetical residuals that are normally distributed. If points cling to the 45 degree line, the residuals are normally distributed.

3. Scale-Location Plot: Similar to Residuals vs. Fitted, except we put the y-axis is now the square root of the standardized residuals. Also informs us about heteroskedasticity (shouldn't see a pattern) and identifies potential outliers

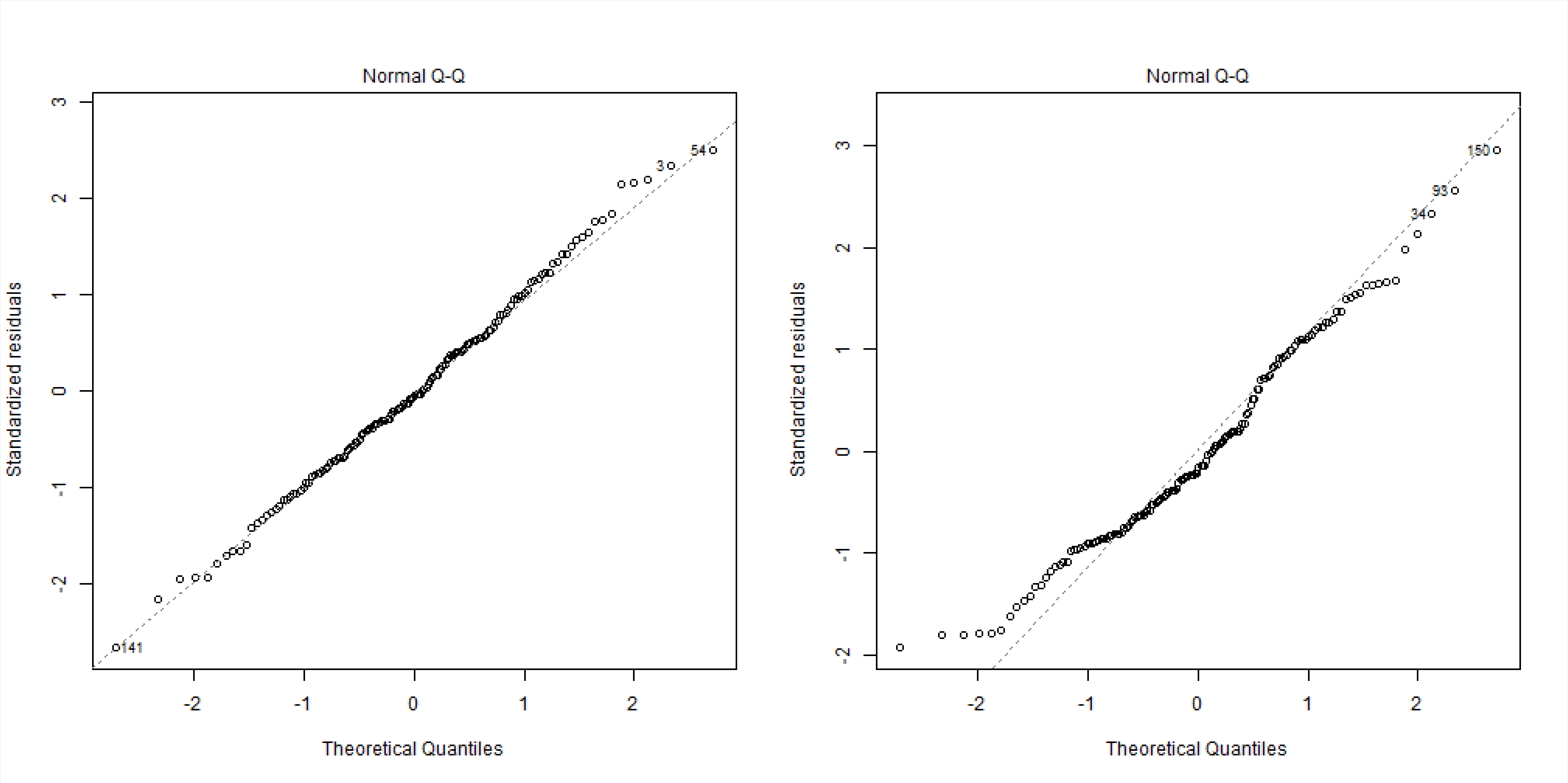
4. Residuals vs.k Leverage: Visualizes the **influence** of each observation on the regression coefficients. Points that are far from other points, especially those that are close to the dashed red lines, are problematic.

- Let's look at each in turn (`R` will produce all four by default if you simply run `plot(m1)` on your regression model)

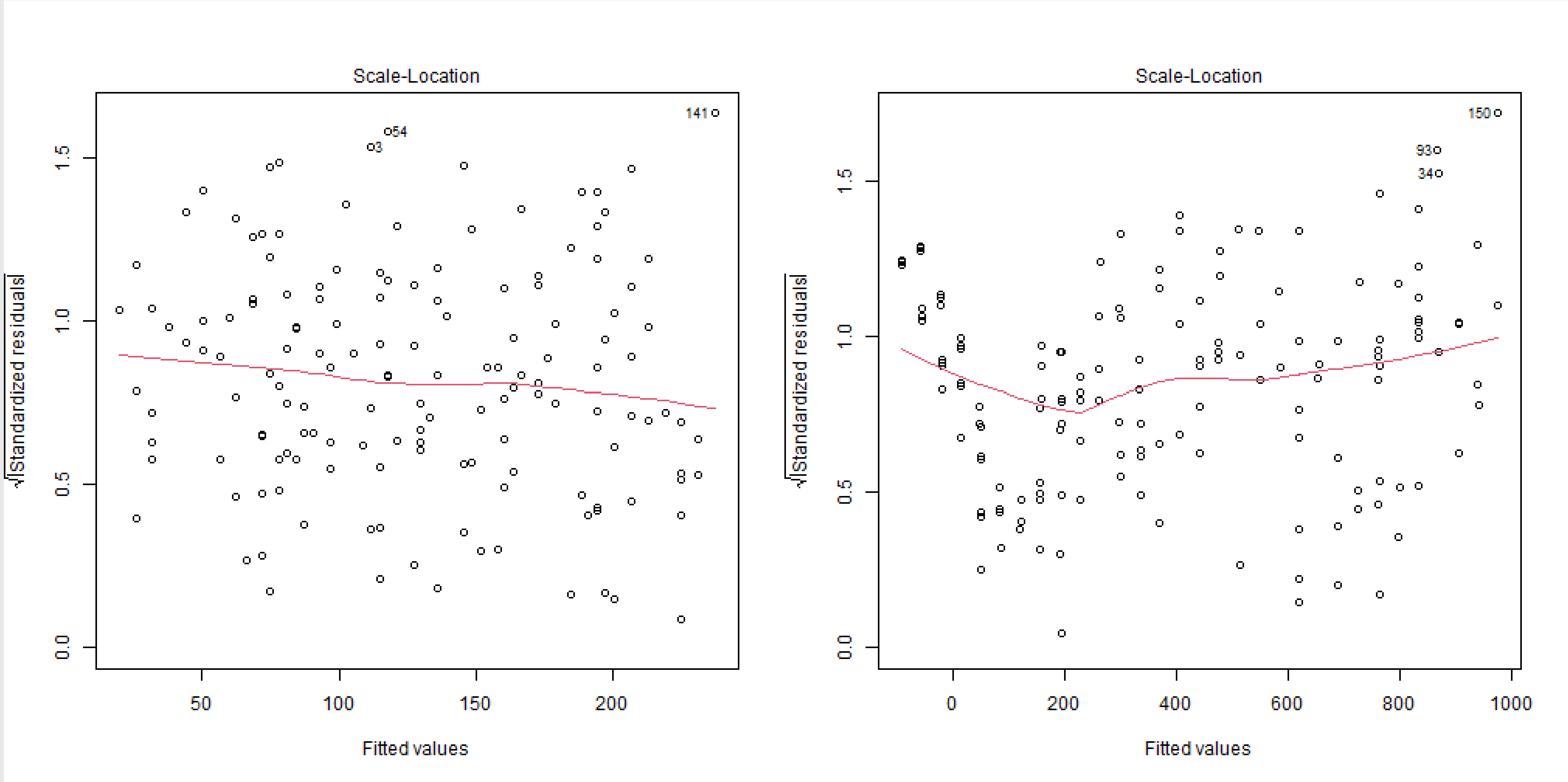
Residuals vs. Fitted



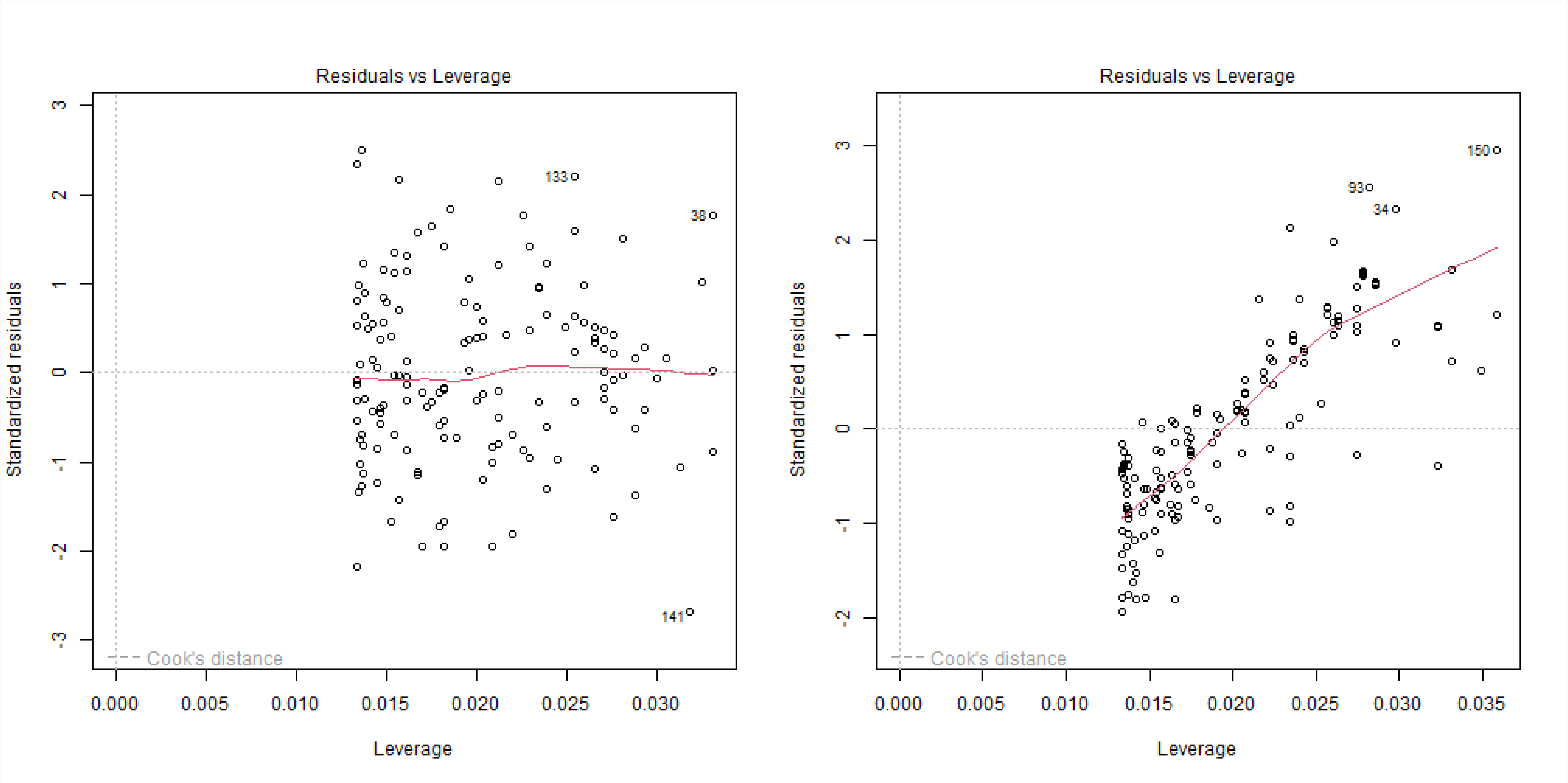
Normal Q-Q Plot



Scale-Location



Residuals vs. Leverage



Regression Diagnostics

- Problematic points: points can be unusual, but not all unusual points are problems
- Consider three types of points:
  1. Outlier Point: an observation with a **large residual**
  2. Leverage Point: an observation with an extreme value for x
  3. Influential Point: an observation that changes the slope of the regression line
- Always good to look at **influential points** to ensure there isn't an error in the measurement
  - But NOT always necessary to blindly throw them out
  - Better to characterize how sensitive the results are to them

Choosing Variables

- With all this in mind, **how do you choose your variables** and **specify your regression equation**?
- We know we want to specify the true relationships, but how do we do this in practice?
- Theory, *theory*, **theory** is essential and should come first
  - This can be formalized with models, or it can be described with intuition, but no amount of diagnostic plots can replace careful theorizing prior to analysis
- That being said, let's consider some additional tests
- One simple method is to compare two specifications, say one that includes x_2 as a control and another that doesn't
  - How can we compare these models?

Goodness of Fit

- Recall the definition of the R^2 from the simple regression case

$$\begin{aligned} R^2 &= \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \\ &= \frac{SSE}{SST} \\ &= 1 - \frac{SSR}{SST} \end{aligned}$$

- where SSE is the explained sum of squares and SSR is the residual sum of squares
- The R^2 will never decrease as we add additional predictors
  - This is because the denominator doesn't change, but the numerator will either increase or stay the same with additional predictors
- This makes the R^2 a pretty terrible metric for comparing models!

Goodness of Fit

- Instead, we typically use the adjusted R-square value:

$$R^2_{adj} = 1 - \left[\frac{\frac{SSR}{(n-k-1)}}{\frac{SST}{(n-1)}} \right]$$
$$= 1 - \frac{\frac{\hat{\sigma}^2_u}{(n-k-1)}}{\frac{SST}{(n-1)}}$$

- By construction, this will only increase with a new predictor if that variable's t-statistic is greater than 1 in absolute value

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{(n - k - 1)}$$

- On your own, think about when the R^2 and R^2_{adj} will be similar and different?

Too many variables

- Define a variable (denoted W) as **irrelevant** if it has not partial effect on y in the population: $\frac{\partial y}{\partial W} = 0$
- If we include W in our model (aka "overspecifying the model") will not bias the estimates since it does not violate our assumptions 1 through 4
    - In other words, if the true model is $y = \beta_0 + \beta_1 x + u$ but we specify $y = \beta_0 + \beta_1 x + \beta_2 W + u$, all $\hat{\beta}$ will be unbiased
- However, we can still harm our model if W is collinear with x
    - Recall that $Var(\hat{\beta}_j) = \frac{\sigma^2}{n * var(x_j) * (1 - R_j^2)}$ where R_j^2 is the R-squared obtained from regressing x_j on all other independent variables in the model
    - If W is correlated with x, then $Var(\hat{\beta}_1)$ will become inflated, meaning our model is less **efficient**
    - Put a different way, our statistical power decreases, increasing the likelihood of falsely accepting the null

Too many variables

- You can assess this threat by calculating each R_j^2 yourself
  - Also can calculate the **variance inflation factor** (VIF): $VIF(\hat{\beta}_j) = \frac{1}{1-R_j^2}$
  - Can rewrite $Var(\hat{\beta}_j) = \frac{\sigma^2}{n*var(x_j)} VIF(\hat{\beta}_j)$, which is where it gets it's name...the factor by which $Var(\hat{\beta}_j)$ is inflated due to the fact that x_j is correlated with other x's in the model
- However, we will sometimes *want* to include controls that are correlated with y but are **not** correlated with x
  - Note that these are not necessary to recover unbiased estimates (remember the definition of OVB?)
- Why do we want to control for some Z where $\frac{\partial x}{\partial Z} = 0$ but $\frac{\partial y}{\partial Z} \neq 0$?
  - It helps explain variation in y, meaning that σ^2 is lower, meaning $Var(\hat{\beta}_j)$ is also lower
  - In other words, it makes all our estimates more **efficient**

Hypotheses about Parameters

- Thus far, we've always been implicitly interested in a single coefficient, or testing each one at a time
- But we might be interested in how two coefficients related to each other
  - For example, your book has the example where researchers are interested whether the effect on income of an additional year of education at a junior college is as much as the effect of an additional year of education at four-year university.
  - The idea here is that jc's are lower status in the U.S. than universities, so maybe employers value these years of education less.
  - (A complementary hypothesis would be that a jc education may be of lower quality.)
- The model assumed is $\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 work + u$

Hypotheses about Parameters

- If we are interested in whether there is a **difference** in returns to education from junior colleges and universities, what is the appropriate null hypothesis?
- $H_0 : \beta_1 = \beta_2$
- And the alternative?
- $H_A : \beta_1 < \beta_2$
- We can re-write as:

$$H_0 : \beta_1 - \beta_2 = 0$$
$$H_A : \beta_1 - \beta_2 < 0$$

- Thus our quantity of interest is $\beta_1 - \beta_2$ and our test statistic is $\frac{\hat{\beta}_1 - \hat{\beta}_2}{se(\hat{\beta}_1 - \hat{\beta}_2)}$

Hypotheses about Parameters

- What is $se(\hat{\beta}_1 - \hat{\beta}_2)$?

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{var(\hat{\beta}_1 - \hat{\beta}_2)} \text{ and}$$
$$var(\hat{\beta}_1 - \hat{\beta}_2) = var(\hat{\beta}_1) + var(\hat{\beta}_2) - 2cov(\hat{\beta}_1, \hat{\beta}_2) \text{ so}$$
$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{var(\hat{\beta}_1) + var(\hat{\beta}_2) - 2cov(\hat{\beta}_1, \hat{\beta}_2)}$$

- We can just grab these values from the variance-covariance matrix of estimated betas
- Or we can do an even easier trick! Let's denote $\theta = \beta_1 - \beta_2$, meaning that $\beta_1 = \theta + \beta_2$
- Therefore:

$$\begin{aligned}\log(wage) &= \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 work + u \\ &= \beta_0 + (\theta + \beta_2) jc + \beta_2 univ + \beta_3 work + u \\ &= \beta_0 + \theta jc + \beta_2 (univ + jc) + \beta_3 work + u\end{aligned}$$

- So easy! Just create a new variable that is the sum of $univ$ and jc and look at the coefficient on jc!

Multiple Linear Restrictions

- What if we want to know if a **group** of predictors are **jointly** significant?
- Start by defining an **unrestricted** model as $UR : y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$
- Then take the group of variables we are interested in evaluating and move them to the end of the regression
  - I.e., if there are q variables we want to test are jointly significant, denote these as $\beta_{k-q+1}, \beta_{k-q+2}$ etc.
  - Thus our null hypothesis is $H_0 : \beta_{k-q+1} = \beta_{k-q+2} = \cdots = \beta_k = 0$
- We can write a restricted model as $R : y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-q} x_{k-q} + u$
- To test, we use the F-statistic defined as $F \equiv \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$
  - Note that, since $SSR_r \geq SSR_{ur}$, $F > 0$
- This is the ratio of two independent χ^2 random variables, divided by their respective degrees of freedom
- We can therefore conduct hypothesis testing using this: if it is extremely unlikely that we would obtain the observed F-statistic by chance, we reject the null H_0